An introduction to linear algebra for undergraduate engineering with applications in MATLAB

First made available in 2016. Released under the terms of



I would like to thank Prof. Fred McCourt who asked me to teach the first-year linear algebra course to the nanotechnology engineering class. I am also deeply indebted to Sheldon Axler, the author of the text *Linear Algebra Done Right*. I read this book at some point after completing my graduate studies, and thought "why did I not enjoy matrix and linear algebra during my studies." Afterword, I picked up a more standard text in the subject, and the memories came back. I would use Sheldon's textbook, but—as he puts it—*[t]his text for a second course in linear algebra, aimed at math majors and graduates*…

I would also like to thank Skandan Chandrasekar who reviewed the text book for errors and suggestions prior to the Fall 2019 academic term.

Typographic conventions

This text uses a 10 pt Times New Roman font where *italics* indicates new terms and names of books. `9 pt Consolas` is used for program listings and console commands with output, and within paragraphs for keywords, variables and function names. Section titles are in Constantia.

Disclaimer

This document is intended for the instruction and examination of NE 112 *Linear Algebra for Nanotechnology Engineers* at the University of Waterloo. The material in it reflects the authors' best judgment in light of the information available to them at the time of preparation. Any reliance on this document by any party for any other purpose is the responsibility of such parties. The authors accepts no responsibility for errors or omissions, or for damages, if any, suffered by any party as a result of decisions made or actions based on the contents of this text for any other purpose than that for which it was intended.

This draft is, unfortunately, still incomplete in many respects.

Printed in Canada.

Additional acknowledgments:

William MacDonald, Syed Hasan Ahmed, Anneke van Heuven, Laura Haba, Alex Pezzutto, Brandon Thien Trong Tran, George Magdy Fawzy, Derek Li, Yunheng Zou

A practical introduction to linear algebra for undergraduate engineering with applications in MATLAB

Douglas Wilhelm Harder
University of Waterloo

Version 0.2019.11.08

To Sherry E. Robinson

In memory of David F. Evans

1 Introductory material 10
1.1 Multiplication tables 10
1.2 The Greek alphabet 11
1.3 Finite discrete signals and series 13
1.4 Matlab 25
1.5 A quick introduction to the axiomatic method 39
1.6 Fields and complex numbers 41
1.7 Summary of introductory material 107
2 Vectors and vector spaces 108
2.1 Real finite-dimensional vectors 108
2.2 Finite-dimensional complex vectors 120
2.3 Vector operations 120
2.4 Other vector spaces 138
2.5 Applications of vector arithmetic 145
2.6 Summary of vectors and vector operations 149
3 Subspaces 150
3.1 A review of sets 150
3.2 Determining if a subset is a vector space 151
3.3 Examples of subspaces 158
3.4 Summary of subspaces 167
4 Normed vector spaces 168
4.1 The 2-norm for finite-dimensional vectors 168
4.2 Other norms for finite-dimensional vectors 171
4.3 Unit vectors and normalization of vectors 173
4.4 Norms for other vector spaces 177
4.5 Applications of norms 181
4.6 Summary of norms of vector spaces 192
5 Inner product spaces 193
5.1 Definition of an inner product 193
5.2 The norm *induced* by an inner product 197
5.3 Other inner product spaces 198
5.4 Orthogonality of vectors 200
5.5 Orthogonality in other inner product spaces 200
5.6 Pythagorean theorem 205
5.7 Projections and best approximations 208
5.8 Cauchy–Bunyakovsky–Schwarz inequality 224
5.9 Angle between vectors 226
5.10 The Gram-Schmidt algorithm for the orthogonalization of vectors 229
5.11 Applications of the inner product 237
6 Solutions to systems of linear equations, linear independence and bases 240
6.1 Equations, linear equations and systems of equations 240
6.2 Finding solutions to a system of linear equations 244
6.3 Finding linear of combinations equaling a given vector 251
6.4 A systematic approach to solving a system of linear equations 255
6.5 Linear dependence and independence 333
6.6 Spans and subspaces 341
6.7 Basis and dimension 344
6.8 Vectors as coefficients of a basis 345
7 A digression to real 3-dimensional space 352

7.1 Equations of lines 367
7.2 Finding the line through two points 369
7.3 Planes 370
7.4 The intersection of two planes 375
7.5 The cross product 376
7.6 Applications of the cross product 382
7.7 Finding the plane containing three points 388
8 Linear maps and linear operators 392
8.1 The superposition principle 395
8.2 Definition of linear maps and operators 397
8.3 Properties of linear maps 409
8.4 Special linear maps and operators 411
8.5 Range of a linear map 416
8.6 The null space of a linear mapping 422
8.7 The inverse problem 427
8.8 The vector space of linear maps 432
8.9 Composition of linear maps 440
8.10 Linear operators 454
8.11 Row operations 463
8.12 Gaussian elimination 466
8.13 Summary of linear maps and operators 468
9 Matrix decompositions 469
9.1 Finding *P*, *L* and *U* 471
10 Linear operators mapping a space onto itself 475
10.1 Determinant 475
10.2 Trace 488
10.3 The inverse of a linear operator 490
10.4 Finding the inverse 501
11 The adjoint of a linear operator (transpose and conjugate transpose) 502
11.1 Properties of the adjoint 502
11.2 The adjoint for real finite-dimensional vector spaces 507
11.3 The adjoint for complex finite-dimensional vector spaces 513
11.4 A new definition of the inner product 517
11.5 Self-adjoint and skew-adjoint operators 518
11.6 Results regarding self-adjoint and skew-adjoint linear operators 520
11.7 Normal operators 522
11.8 Isometric operators, orthogonal matrices and unitary matrices 524
11.9 Least squares 527
11.10 The naïve approach 537
11.11 Cholesky factorization 538
11.12 QR factorization 538
11.13 Numerical error 543
11.14 Operator *-algebras 545
11.15 Summary of adjoint operators 545
12 Invariant subspaces 546
12.1 1-dimensional invariant subspaces 547
12.2 Vector space of functions 563
12.3 Diagonalization 565
12.4 Positive-definite matrices 571

13 Change of bases 345
14 Singular-value decomposition (SVD) 573
15 Matrix norms and condition 580
15.1 Matrix 2-norms 581
15.2 Other induced matrix norms 586
15.3 Condition numbers 587
15.4 Operator C*-algebras 588
16 Iterative methods for approximating solutions 590
16.1 The fixed-point theorem and applications 590
16.2 Iterative methods of approximating solutions to systems of linear equations 601
17 Vector and sparse matrix data structures 607
17.1 Data structures for reduced accumulated error 607
17.2 Vector data structures 608
17.3 Sparse matrix data structures 621
17.4 Sparse-matrix–vector operations 632
17.5 Solving systems of linear equations 639
18 Moore-Penrose pseudo-inverse 650
18.1 Finding the Moore-Penrose pseudo-inverse in special cases 650
18.2 The Moore-Penrose pseudo-inverse of a rectangular diagonal matrix 651
18.3 Finding the Moore-Penrose pseuodo-inverse in general 652
18.4 Alternative requirements for the Moore-Penrose pseudo-inverse 652
18.5 Applications 653
18.6 Summary of the Moore-Penrose pseudo-inverse 654
Glossary 655
Appendix A: A quick introduction to integration 656

1 Introductory material

While this is a course on linear algebra, we will cover some introductory material that is necessary for the understanding of the course material. This will include:

1. your multiplication tables,
2. a review of the Greek alphabet,
3. finite and infinite sums,
4. the Matlab programming language and integrated development environment,
5. complex numbers and fields, and
6. an introduction to the axiomatic method.

This will give a solid foundation on which the balance of the course can be taught.

1.1 Multiplication tables

It is unfortunately somewhat embarrassing to have to put this into a university-level text book, but evidence has suggested this is actually necessary. At this point in your academic career, you should know what 6 × 7 is without thinking. Most often in this course, you will be asked to do reasonably simple multiplications such as 6 × 4.2 or to divide 4.2 ÷ 7.

×	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	1	2	3	4	5	6	7	8	9	10	11	12
2	2	4	6	8	10	12	14	16	18	20	22	24
3	3	6	9	12	15	18	21	24	27	30	33	36
4	4	8	12	16	20	24	28	32	36	40	44	48
5	5	10	15	20	25	30	35	40	45	50	55	60
6	6	12	18	24	30	36	42	48	54	60	66	72
7	7	14	21	28	35	42	49	56	63	70	77	84
8	8	16	24	32	40	48	56	64	72	80	88	96
9	9	18	27	36	45	54	63	72	81	90	99	108
10	10	20	30	40	50	60	70	80	90	100	110	120
11	11	22	33	44	55	66	77	88	99	110	121	132
12	12	24	36	48	60	72	84	96	108	120	132	144

Problems

1. Note that 6 × 6 = 36 but 5 × 7 = 35, 7 × 7 = 49 but 6 × 8 = 48, and 8 × 8 = 64 but 7 × 9 = 63. What is the pattern, and justify it.

2. Note that 5 × 5 = 25 but 3 × 7 = 21, 7 × 7 = 49 but 5 × 9 = 45, and 9 × 9 = 81 but 7 × 11 = 77. What is the pattern and justify it.

Solutions

1. The product of $(6-1)(6+1)=6^2-1$. From algebra, we have that $n^2-1=(n-1)(n+1)$.

1.2 The Greek alphabet

In the nanotechnology program, you will be using Greek letters quite regularly, and consequently, we will start with a few of the more common. Those letters that resemble letters from the Latin alphabet are seldom used as Greek letters, so they are grayed out. Those that we will use in class are starred.

A	α	**alpha***
B	β	**beta***
Γ	γ	**gamma***
Δ	δ	**delta***
E	ε	epsilon
Z	ζ	zeta
H	η	eta
Θ	θ	**theta***
I	ι	iota
K	κ	kappa
Λ	λ	**lambda***
M	μ	mu
N	ν	nu
Ξ	ξ	xi
O	o	omicron
Π	π	**pi***
P	ρ	rho
Σ	σ	**sigma***
T	τ	tau
Y	υ	upsilon
Φ	ϕ	phi
X	χ	chi
Ψ	ψ	psi
Ω	ω	omega

You don't have to memorize these now, we will use them often enough that they will become familiar.

In your classes, you should be careful to differentiate between lower- and upper-case theta, phi and psi: compare θ or θ versus Θ, ϕ or ϕ versus Φ, and ψ or ψ versus Ψ, respectively.

Incidentally, the Greeks adopted their alphabet from the Phoenicians with whom they traded. The Phoenician alphabet is similar to the Hebrew alphabet, and more distantly related to the Arabic alphabet, and consequently, there are some similarities; for example: alpha, aleph, alif; delta, daleth, dal; and lambda, lamed, lam are transliterations of the modern Greek, Hebrew and Arabic names for these letters, respectively.

The first two letters give the name of our set of letters: *alpha-beta* or *alphabet*.

Problems

1. Read out the following:

a) $\alpha\mathbf{u} + \beta\mathbf{v}$
b) $\gamma x + \delta y$
c) $\sin^2(\theta)$
d) λA
e) πr^2 and its derivative $2\pi r$
f) $\mu \pm 1.98\sigma$

Solutions

1. These are

a) alpha u plus beta v
b) gamma x plus delta y
c) sine squared theta or sine theta all squared
d) lambda a or lambda times a
e) pi r squared and its derivative two pi r
f) mu plus or minus 1.98 times sigma

1.3 Finite discrete signals and series

Suppose we have either a *finite discrete signal* (also called a *finite sequence*) of *n* values

$$(u_1, u_2, u_3, ..., u_n)$$

and we would like to represent the sum of these *n* values. The sum of a given finite discrete signal is called a *finite series*, and we could represent such a series as

$$u_1 + u_2 + u_3 + \cdots + u_n,$$

but this is imprecise, and you must assume that the interpolated numbers are values u_4 through u_{n-1}. For example, suppose you saw the series

$$2^1 + 2^2 + 2^3 + \cdots + 2^{10};$$

you would need to infer that the ellipsis represent $2^4 + 2^5 + 2^6 + 2^7 + 2^8 + 2^9$. Suppose, however, you saw

$$1 + 1 + \frac{1}{2} + \frac{1}{6} + \cdots + \frac{1}{3628800}.$$

It may be more difficult to infer that the missing summands are $\frac{1}{24} + \frac{1}{120} + \frac{1}{720} + \frac{1}{5040} + \frac{1}{40320} + \frac{1}{362880}$, as you must recognize that these values are $\frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{10!}$; but even if you were able to determine these, the pattern of the series

$$\frac{1}{0!3!}\left(\frac{x}{2}\right)^3 - \frac{1}{1!4!}\left(\frac{x}{2}\right)^5 + \frac{1}{2!5!}\left(\frac{x}{2}\right)^7 - \frac{1}{3!6!}\left(\frac{x}{2}\right)^9 + \cdots + \frac{1}{12!15!}\left(\frac{x}{2}\right)^{27},$$

would be more difficult to infer. In this case, it happens that the individual terms are $\frac{(-1)^k}{k!(k+3)!}\left(\frac{x}{2}\right)^{2k+3}$ where *k* runs from 0 to 12, but it would be more convenient to be able to write "sum the values $\frac{(-1)^k}{k!(k+3)!}\left(\frac{x}{2}\right)^{2k+3}$ for *k* = 0, 1, 2, …, 12"?

A more compact means of writing such a finite series is

$$\sum_{k=1}^{n} u_k,$$

where the Σ is the Greek capital letter "S", signifying that this is a sum. The *k* = 1 below the Σ and the *n* above it indicate that *k* takes on all integer values from 1 to *n*; thus, for example,

$$\sum_{k=1}^{6} u_k = u_1 + u_2 + u_3 + u_4 + u_5 + u_6.$$

Thus, we may write these as

$$\sum_{k=1}^{10} 2^k = 2^1 + 2^2 + 2^3 + \cdots + 2^{10},$$

$$\sum_{k=0}^{10} \frac{1}{k!} = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \cdots + \frac{1}{3628800}, \text{ and}$$

$$\sum_{k=0}^{12} \frac{(-1)^k}{k!(k+3)!}\left(\frac{x}{2}\right)^{2k+3} = \frac{1}{0!3!}\left(\frac{x}{2}\right)^3 - \frac{1}{1!4!}\left(\frac{x}{2}\right)^5 + \frac{1}{2!5!}\left(\frac{x}{2}\right)^7 - \frac{1}{3!6!}\left(\frac{x}{2}\right)^9 + \cdots + \frac{1}{12!15!}\left(\frac{x}{2}\right)^{27}.$$

An discrete signal may also be infinite, such as

$$1, 1, 2, 6, 24, 120, 720, \ldots.$$

Adding the values in this infinite discrete signal yields a *infinite series*,

$$\sum_{k=0}^{\infty} k!,$$

which says continue adding $k!$ for all values of all non-negative integer values of k, or $1+1+2+6+24+120+720+\cdots$.

Sequence	Example	Series	Example	Requirement
finite discrete signal	$u_m, u_{m+1}, \ldots, u_{n-1}, u_n$	finite series	$\sum_{k=m}^{n} u_k$	m and n are integers and $m \le n$
infinite discrete signal	$u_m, u_{m+1}, u_{m+2}, u_{m+3}, \ldots$	infinite series	$\sum_{k=m}^{\infty} u_k$	m is an integer

The series consisting of the factorials of all the non-negative integers is clearly infinite:

$$\sum_{k=0}^{\infty} k! = \infty.$$

It isn't always true that infinite series always equal infinity, however, as you may see with the series

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} + \cdots = \sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k.$$

If you work out the first few finite sums, you come to a pattern:

$$\begin{aligned}
1 &= 1 &&= 2 - 1\\
1 + \frac{1}{2} &= \frac{3}{2} &&= 2 - \frac{1}{2}\\
1 + \frac{1}{2} + \frac{1}{4} &= \frac{7}{4} &&= 2 - \frac{1}{4}\\
1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} &= \frac{15}{8} &&= 2 - \frac{1}{8}\\
1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} &= \frac{31}{16} &&= 2 - \frac{1}{16}
\end{aligned}$$

From this, we may propose the following theorem:

Theorem

$$\sum_{k=0}^{n}\left(\frac{1}{2}\right)^k = 2-\frac{1}{2^n}.$$

Proof:
By induction, we see that this statement is true when $n = 0$, for

$$\sum_{k=0}^{0}\left(\frac{1}{2}\right)^k = \left(\frac{1}{2}\right)^0 = 1 = 2-1 = 2-\frac{1}{2^0}.$$

Assume the statement is true for $n \geq 0$, so we are assuming that $\sum_{k=0}^{n}\left(\frac{1}{2}\right)^k = 2-\frac{1}{2^n}$.

Based on this, we want to determine the value of $\sum_{k=0}^{n+1}\left(\frac{1}{2}\right)^k$. Now,

$$\begin{aligned}\sum_{k=0}^{n+1}\left(\frac{1}{2}\right)^k &= 1+\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^n}+\frac{1}{2^{n+1}}\\ &= \left(1+\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^n}\right)+\frac{1}{2^{n+1}}\\ &= \left(\sum_{k=0}^{n}\left(\frac{1}{2}\right)^k\right)+\frac{1}{2^{n+1}}\end{aligned}$$

By assumption, $\sum_{k=0}^{n}\left(\frac{1}{2}\right)^k = 2-\frac{1}{2^n}$, so

$$\begin{aligned}\sum_{k=0}^{n+1}\left(\frac{1}{2}\right)^k &= \left(\sum_{k=0}^{n}\left(\frac{1}{2}\right)^k\right)+\frac{1}{2^{n+1}}\\ &= \left(2-\frac{1}{2^n}\right)+\frac{1}{2^{n+1}}\\ &= 2+\left(-\frac{1}{2^n}+\frac{1}{2^{n+1}}\right) \text{ finding a common denominator}\\ &= 2+\left(\frac{-2+1}{2^{n+1}}\right)\\ &= 2-\frac{1}{2^{n+1}}\end{aligned}$$

Thus, the statement is true when $n = 0$, and the truth of the statement for an arbitrary n implies that the statement is also true for $n + 1$. Therefore, by the process of mathematical induction, this statement is true for all values of $n \geq 0$.
∎

Consequently, as n gets larger, $\sum_{k=0}^{n}\left(\frac{1}{2}\right)^k$ gets arbitrarily close to 2. In fact, if you want to get closer to 2 than some small value of ε, simply choose $n \geq -\lg(\varepsilon)$.[1] If you want to ensure that the sum is less than 10^{-100} away from 2, simply ensure that n is any integer greater than 664.38, and indeed,

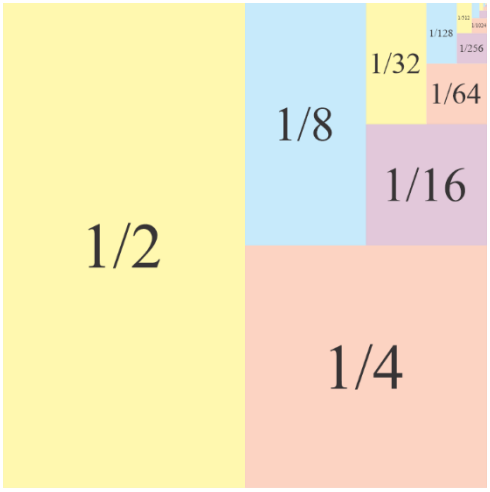
$$\sum_{k=0}^{666}\left(\frac{1}{2}\right)^k = \frac{6123604138321678046184813001752049505652789728277332 45\underbrace{\cdots}_{144\text{ missing digits}}0927}{3061802069160839023092406500876024752826394864138666 22\underbrace{\cdots}_{144\text{ missing digits}}0464}$$

$$\approx 1.9\underbrace{\cdots}_{198\text{ missing 9's}}9673394955842\cdots > 1.9\underbrace{\cdots}_{198\text{ missing 9's}}9 = 1 - 10^{-200}$$

Thus, it is quite reasonable to state that

$$\sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k = 2\,.$$

The following is a fascinating image this author once saw on the Internet that shows that $\sum_{k=1}^{\infty}\left(\frac{1}{2}\right)^k = 1$:



Two well-known series, the value of which you may not be aware, are

$$\sum_{k=0}^{\infty}\frac{1}{k!} = e \text{ and } \sum_{k=0}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6}\,.$$

You will see why these are true in your calculus course.

[1] Here, lg(x) represents the base-2 logarithm $\log_2(x)$.

We will now discuss some properties of series, including:

1. scalar multiplication distributes over a series,
2. adding two series,
3. the linear property of series,
4. arithmetic series, and
5. geometric series.

1.3.1 Distributive property for scalar multiplication

You already know that $3(a + b + c) = 3a + 3b + 3c$. Similarly,

$$\alpha\left(\sum_{k=m}^{n} u_k\right) = \sum_{k=m}^{n} (\alpha u_k) \text{ and } \alpha\left(\sum_{k=m}^{\infty} u_k\right) = \sum_{k=m}^{\infty} (\alpha u_k).$$

For example,

$$\sum_{k=0}^{\infty}\left(\frac{5}{2^k}\right) = 5\sum_{k=0}^{\infty}\left(\frac{1}{2^k}\right) = 5 \cdot 2 = 10.$$

1.3.2 Adding two series

Suppose we have two finite series over the same range:

$$\sum_{k=m}^{n} u_k = s_u \text{ and } \sum_{k=m}^{n} v_k = s_v.$$

You already know that $(a + b + c) + (x + y + z) = (a + x) + (b + y) + (c + z)$, so it should be equally clear that

$$\sum_{k=m}^{n} (u_k + v_k) = \left(\sum_{k=m}^{n} u_k\right) + \left(\sum_{k=m}^{n} v_k\right) = s_u + s_v.$$

For series, it is also true that if

$$\sum_{k=m}^{\infty} u_k = s_u \text{ and } \sum_{k=m}^{\infty} v_k = s_v,$$

are both finite (and not infinite) then

$$\sum_{k=m}^{\infty} (u_k + v_k) = \left(\sum_{k=m}^{\infty} u_k\right) + \left(\sum_{k=m}^{\infty} v_k\right) = s_u + s_v.$$

Thus, two series can always be added so long as their ranges are equal.

1.3.3 The linear property of series

We may now introduce a concept that we will see over and over again in this course: combining these last two properties:

$$\sum_{k=m}^{n} (\alpha u_k + \beta v_k) = \left(\sum_{k=m}^{n} (\alpha u_k)\right) + \left(\sum_{k=m}^{n} (\beta v_k)\right) = \left(\alpha \sum_{k=m}^{n} u_k\right) + \left(\beta \sum_{k=m}^{n} v_k\right) = \alpha s_u + \beta s_v$$

and

$$\sum_{k=m}^{\infty}\left(\alpha u_k+\beta v_k\right)=\left(\sum_{k=m}^{\infty}\left(\alpha u_k\right)\right)+\left(\sum_{k=m}^{\infty}\left(\beta v_k\right)\right)=\left(\alpha\sum_{k=m}^{\infty}u_k\right)+\left(\beta\sum_{k=m}^{\infty}v_k\right)=\alpha s_u+\beta s_v\,.$$

Put more simply,

$$\sum_{k=m}^{n}\left(\alpha u_k+\beta v_k\right)=\alpha\left(\sum_{k=m}^{n}u_k\right)+\beta\left(\sum_{k=m}^{n}v_k\right)$$

and

$$\sum_{k=m}^{\infty}\left(\alpha u_k+\beta v_k\right)=\alpha\left(\sum_{k=m}^{\infty}u_k\right)+\beta\left(\sum_{k=m}^{\infty}v_k\right).$$

This says it doesn't matter whether or not we add the sum-of-products $\alpha u_k+\beta v_k$ and then calculate the series, or if we calculate the series first and then calculate the sum-of-products $\alpha\left(\sum_{k=m}^{\infty}u_k\right)+\beta\left(\sum_{k=m}^{\infty}v_k\right)$.

1.3.4 Finite series of a constant value

If α is a constant, the finite series

$$\sum_{k=m}^{n}\alpha=\underbrace{\alpha+\cdots+\alpha}_{n-m+1\text{ times}}$$

is the simplest finite series one can calculate, as this is simply $n - m + 1$ times the value that is being summed:

$$\sum_{k=m}^{n}\alpha=\alpha\left(n-m+1\right).$$

If the series in infinite, adding the same non-zero constant infinitely many times must—even if it is very small—must be infinite:

$$\sum_{k=m}^{\infty}\alpha=\infty\,,$$

and thus we are generally restricted to discussing such series only when the range is finite.

1.3.5 Modifying the range

Given the sum

$$\sum_{k=m}^{n}u_k\,,$$

it is always possible to either add or remove terms from the sum; for example

$$\sum_{k=m}^{n}u_k=u_m+\left(\sum_{k=m+1}^{n}u_k\right)\text{ and }\sum_{k=m}^{n}u_k=\left(\sum_{k=m}^{n-1}u_k\right)+u_n\,.$$

1.3.6 Arithmetic series

An arithmetic sequence is a discrete signal of the form

$$\alpha, \alpha+\beta, \alpha+2\beta, \alpha+3\beta, \alpha+4\beta, \cdots,$$

where, in general, the kth term in the discrete signal is $u_k = \alpha + k\beta$ starting with $k = 0$. Often it is necessary to calculate

$$\sum_{k=0}^{n} u_k = \sum_{k=0}^{n} (\alpha + k\beta).$$

Using the linear properties of series, we have

$$\sum_{k=0}^{n} (\alpha + k\beta) = \alpha\left(\sum_{k=0}^{n} 1\right) + \beta\left(\sum_{k=0}^{n} k\right).$$

The first sum is $n + 1$, so the first summand is $a(n + 1)$ but to determine the second sum, we must find the value of

$$\sum_{k=0}^{n} k.$$

There are many ways to find the value of this sum, including the observation that

$$\sum_{k=0}^{n} k = \sum_{k=0}^{n} (n-k),$$

so we can add the series to itself, only with the terms in reverse order:

$$\begin{aligned} 2\sum_{k=0}^{n} k &= \sum_{k=0}^{n} k + \sum_{k=0}^{n} (n-k) \\ &= \sum_{k=0}^{n} \big(k + (n-k)\big) \\ &= \sum_{k=0}^{n} n \\ &= n\sum_{k=0}^{n} 1 \\ &= n(n+1) \end{aligned}$$

and dividing both sides by 2 yields the formula

$$\sum_{k=0}^{n} k = \frac{n(n+1)}{2} = \frac{n^2+n}{2}.$$

Pictorially, this may be represented by this image:

$$\begin{array}{rccccccccccccc} & \overbrace{0} & + & 1 & + & 2 & +\cdots+ & n-2 & + & n-1 & + & n} & \text{the sum in order} \\ + & n & + & n-1 & + & n-2 & +\cdots+ & 2 & + & 1 & + & 0 & \text{the sum in reverse order} \\ \hline & n & + & n & + & n & +\cdots+ & n & + & n & + & n & = (n+1)n \end{array}$$

Each sum has $n+1$ terms

As this includes summing the terms twice, we must divide the result by two. Consequently, the value of this arithmetic series is

$$\sum_{k=0}^{n}(\alpha+k\beta)=\alpha(n+1)+\beta\frac{n(n+1)}{2}.$$

The rate of growth of this quantity depends on $\beta\frac{n^2}{2}$, as is shown in Figure 1.

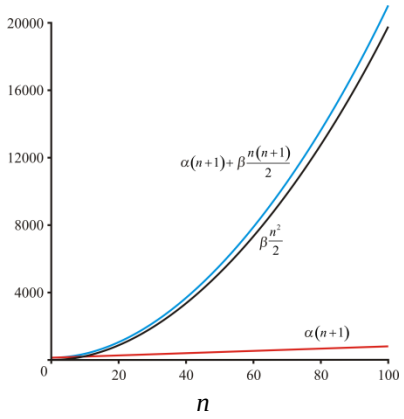


Figure 1. A graph of $\alpha(n+1)$, $\alpha(n+1)+\beta\frac{n(n+1)}{2}$ and $\beta\frac{n^2}{2}$.

If we wanted to prove this as a theorem, we can do so through a proof by induction.

Theorem

Given scalars α, β and a positive integer n, $\sum_{k=0}^{n}(\alpha+k\beta)=\alpha(n+1)+\beta\frac{n(n+1)}{2}$.

Proof:

When $n = 0$, the sum is α, and $\alpha(0+1)+\beta\frac{0(0+1)}{2}=\alpha$.

Assume that $\sum_{k=0}^{n}(\alpha+k\beta)=\alpha(n+1)+\beta\frac{n(n+1)}{2}$ is true for an arbitrary $n \geq 0$, in which case, we note that

$$\sum_{k=0}^{n+1}(\alpha+k\beta)=\left(\sum_{k=0}^{n}(\alpha+k\beta)\right)+\alpha+(n+1)\beta \text{ by Section 1.3.5.}$$

By assumption, the sum on the right-hand side is given

$$\begin{aligned}\sum_{k=0}^{n+1}(\alpha+k\beta) &= \alpha(n+1)+\beta\frac{n(n+1)}{2}+\alpha+(n+1)\beta \\ &= \alpha(n+1)+\alpha+\beta\frac{n(n+1)+2(n+1)}{2} \\ &= \alpha(n+2)+\beta\frac{(n+2)(n+1)}{2}\end{aligned}$$

which equals the sum we would get if we evaluated our formula at $n + 1$.

Thus, the statement is true when $n = 0$, and the truth of the statement for an arbitrary n implies that the statement is also true for $n + 1$. Therefore, by the process of mathematical induction, this statement is true for all values of $n \geq 0$. ∎

1.3.7 Geometric series

Another common series of interest to engineers are both *finite geometric series*, $\sum_{k=0}^{n}\alpha r^k$, and *infinite geometric series*, $\sum_{k=0}^{\infty}\alpha r^k$. The scalar α can be factored out of the sum, so we are left with determining $\sum_{k=0}^{n}r^k$ and, if possible, $\sum_{k=0}^{\infty}r^k$.

The geometric series represents

$$\sum_{k=0}^{n}r^k = 1+r+r^2+r^3+r^4+\cdots+r^n,$$

and so each term is r times the previous, so r is referred to as the *ratio* of the geometric series. We have already seen two geometric series in this chapter: when $r = 2$ and when $r = ½$.

If we define

$$s_n = \sum_{k=0}^{n} r^k = 1 + r + r^2 + r^3 + r^4 + \cdots + r^n ,$$

we note two relationships:

$$\sum_{k=0}^{n+1} r^k = \left(1 + r + r^2 + r^3 + r^4 + \cdots + r^n\right) + r^{n+1} = \left(\sum_{k=0}^{n} r^k\right) + r^{n+1}$$

and so

$$s_{n+1} = s_n + r^{n+1} ,$$

but at the same time,

$$1 + \underbrace{r + r^2 + r^3 + r^4 + \cdots + r^{n+1}}_{\text{factor out an } r \text{ from these}} = 1 + r\left(1 + r + r^2 + r^3 + r^4 + \cdots + r^n\right) .$$

Consequently, $s_{n+1} = 1 + rs_n$. Thus, equating these two, we have that $1 + rs_n = s_n + r^{n+1}$. Solving for $s_{n,}$ we have

$$\begin{aligned} s_n + r^{n+1} = 1 + rs_n \leftrightarrow\ & s_n - rs_n = 1 - r^{n+1} \\ \leftrightarrow\ & s_n\left(1 - r\right) = 1 - r^{n+1} \\ \leftrightarrow\ & s_n = \frac{1 - r^{n+1}}{1 - r} \end{aligned}$$

If we wanted to prove this as a theorem using a proof by induction.

Theorem

Given a scalar α and a positive integer *n*, $\sum_{k=0}^{n} \alpha r^k = \alpha \frac{1 - r^{n+1}}{1 - r}$.

Proof:

When *n* = 0, the sum is α, and $\alpha \frac{1 - r^{0+1}}{1 - r} = \alpha \frac{1 - r}{1 - r} = \alpha$.

Assume that $\sum_{k=0}^{n} \alpha r^k = \alpha \frac{1 - r^{n+1}}{1 - r}$ is true for an arbitrary $n \geq 0$, in which case, we note that

$$\sum_{k=0}^{n+1} \alpha r^k = \left(\sum_{k=0}^{n} \alpha r^k\right) + \alpha r^{n+1} .$$

By assumption, the sum on the right-hand side is given

$$\begin{aligned}\sum_{k=0}^{n+1}\alpha r^k &= \alpha\frac{1-r^{n+1}}{1-r}+\alpha r^{n+1}\\ &= \alpha\frac{1-r^{n+1}+r^{n+1}-r^{n+2}}{1-r}\\ &= \alpha\frac{1-r^{n+2}}{1-r}\end{aligned}$$

which equals the sum we would get if we evaluated our formula at $n + 1$.

Thus, the statement is true when $n = 0$, and the truth of the statement for an arbitrary n implies that the statement is also true for $n + 1$. Therefore, by the process of mathematical induction, this statement is true for all values of $n \geq 0$. ∎

A more intuitive proof that this is indeed the formula for a finite geometric series is the observation that:

$$\begin{aligned}(1-r)\sum_{k=0}^{n} r^k &= \sum_{k=0}^{n} r^k - r\sum_{k=0}^{n} r^k\\ &= \left(1+r+r^2+r^3+\cdots+r^{n-1}+r^n\right)-\left(r+r^2+r^3+\cdots+r^{n-1}+r^n+r^{n+1}\right)\\ &= 1+\left(r+r^2+r^3+\cdots+r^{n-1}+r^n\right)-\left(r+r^2+r^3+\cdots+r^{n-1}+r^n\right)-r^{n+1}\\ &= 1-r^{n+1}\end{aligned}$$

and therefore, dividing both sides by $1 - r$, $\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r}$.

Next, notice that if $|r|<1$, as n becomes larger and larger, r^n becomes smaller and smaller. For example, if $r = 0.9848$, then $r^{100000} \approx 6.368\times10^{-666}$, which is very small indeed. Thus, it is reasonable to assume that

$$\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$$

and therefore

$$\sum_{k=0}^{\infty} \alpha r^k = \frac{\alpha}{1-r}$$

so long as $|r|<1$. In your course on calculus, you will see a more rigorous proof that this is the case. A plot of the ratio $\frac{1}{1-r}$ is shown in Figure 2.

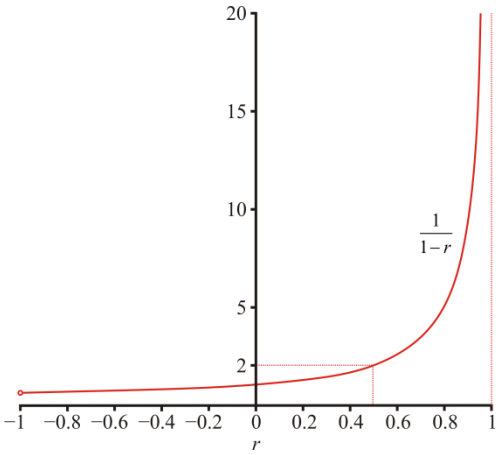


Figure 2. A plot of $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ on (–1, 1).

1.3.8 Summary of discrete signals and series

From this section, you should be aware

$$\sum_{k=m}^{n}(\alpha u_k + \beta v_k) = \alpha\left(\sum_{k=m}^{n} u_k\right) + \beta\left(\sum_{k=m}^{n} v_k\right) \text{ and } \sum_{k=m}^{\infty}(\alpha u_k + \beta v_k) = \alpha\left(\sum_{k=m}^{\infty} u_k\right) + \beta\left(\sum_{k=m}^{\infty} v_k\right);$$

$$\sum_{k=m}^{n} u_k = u_m + \left(\sum_{k=m+1}^{n} u_k\right) \text{ and } \sum_{k=m}^{n} u_k = \left(\sum_{k=m}^{n-1} u_k\right) + u_n;$$

$$\sum_{k=m}^{n} \alpha = \alpha(n-m+1);$$

$$\sum_{k=0}^{n} k = \frac{n(n+1)}{2} \text{ and } \sum_{k=0}^{n}(\alpha + k\beta) = \alpha(n+1) + \beta\frac{n(n+1)}{2}; \text{ and}$$

$$\sum_{k=0}^{n} \alpha r^k = \alpha\frac{1-r^{n+1}}{1-r} \text{ and } \sum_{k=0}^{\infty} \alpha r^k = \frac{\alpha}{1-r}.$$

1.4 Matlab

You will use MATLAB peripherally in this course; however, you will be exposed to MATLAB throughout your undergraduate studies and throughout your professional career, and consequently, we will introduce you to this programming language and associated libraries almost immediately.

For students who have programmed before
If you have already programmed in a language such as C, C++, Java, C# or otherwise used a compiler, you will need to change your frame of reference. For example, in Java, you would have to write a class such as

```
public class MyClass {
    public void main( String[] args ) {
        System.out.println( "Hello world!" );
    }
}
```

Only once you have finished writing this code do you compile it and then execute the resulting program. MATLAB, on the other hand, is an interpreted language. When you launch MATLAB, it presents an interpreter that allows you to enter MATLAB instructions.

In this introduction, we will

1. view the Matlab environment and see how to have Matlab evaluate basic arithmetic expressions,
2. introduce you to some of the built-in mathematical functions,
3. look at Boolean-valued (true-false) operations and functions,
4. introduce constant symbols in Matlab such as π, $\pm\infty$ and $0/0$,
5. see how to control the precision of results displayed to the screen,
6. see how to values assign to variables,
7. supressing output,
8. consider commands that manipulate variables,
9. see that Matlab isn't perfect—using a floating-point representation to approximate real numbers can result in significant errors in computations, and
10. see the available help for Matlab.

We will start with the MATLAB environment.

1.4.1 The MATLAB prompt and basic arithmetic

When you launch MATLAB, you are presented with an *integrated development environment* (IDE) for working in MATLAB.

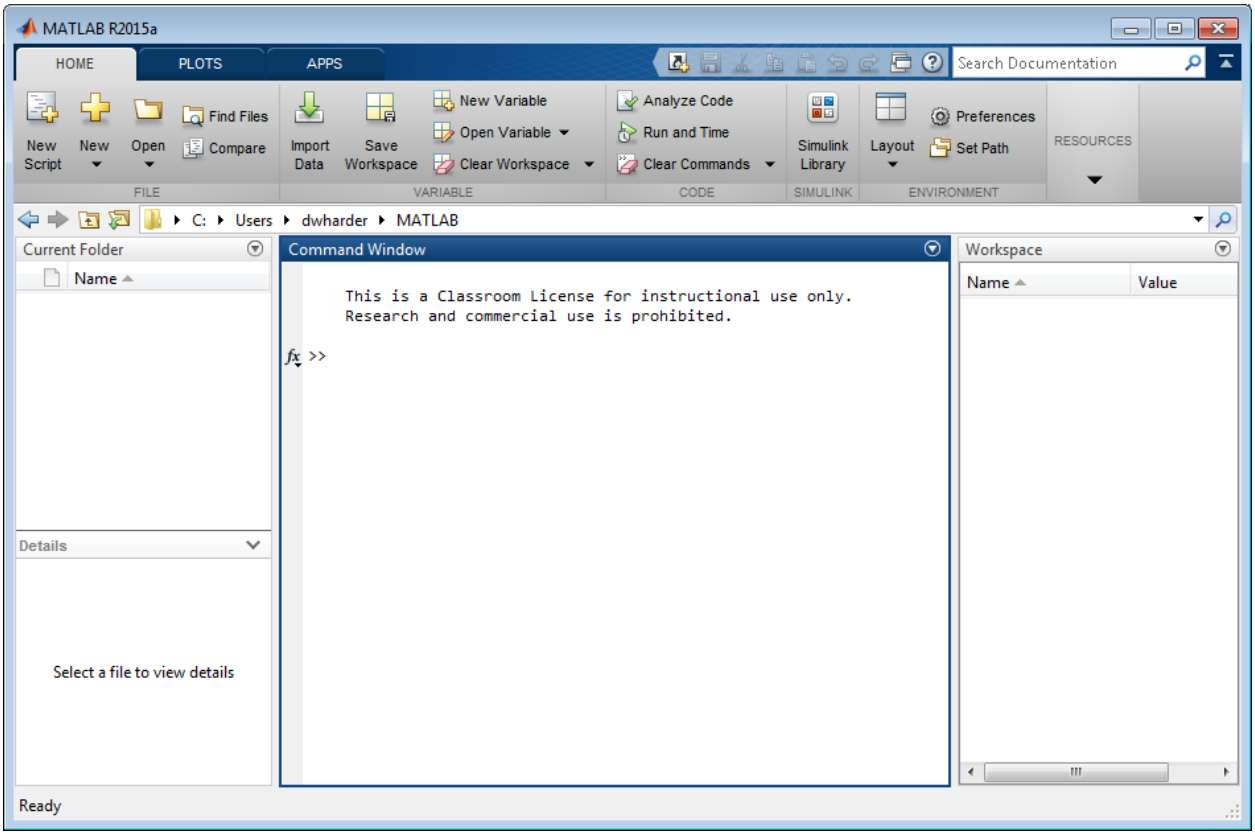


Figure 3. The MATLAB integrated development environment.

For now, we will focus on the central Command Window, where you will be greeted with a prompt and a flashing cursor

```
>>
```

At this prompt, you can now enter a mathematical statement and press Enter to have MATLAB execute your statement. For example, you find that

```
>> 3 + 4*(5.43 + 1/3) - 1.45e-3
    ans =
       26.0519
```

where `1.45e-3` represents 1.45×10^{-3}. We will always show the output of MATLAB in blue to differentiate it from input.

The common arithmetic operations are

Operation	Description
`-x`	negate the value of *x* (*unary minus*)
`x + y`	the sum of *x* and *y*
`x - y`	*y* subtracted from *x* (*binary minus*)
`x*y`	the product of *x* and *y*
`x/y`	*x* divided by *y*
`x^y`	*x* raised to the power of *y*

Operations are performed using the BEDMAS mnemonic: brackets, exponents, division and multiplication, and addition and subtraction, in that order. After that, operations are performed left-to-right. Therefore, if you want to calculate $\frac{2+3}{4\cdot 5}=\frac{1}{4}$, it is necessary to use caution:

```
>> (2+3)/4*5          % incorrect:  this calculates ((2 + 3)/4) * 5
    ans =
       6.2500

>> (2+3)/4/5          % correct
    ans =
       0.2500

>> 2+3/4/5            % incorrect:  this calculates 2 + ((3/4)/5)
    ans =
       2.1500

>> (2+3)/(4*5)        % also correct
    ans =
       0.2500
```

Similarly, $\frac{8+17}{4+6}$, which equals $\frac{5}{2}=2.5$, must be calculated as

```
>> 8 + 17/4 + 6       % incorrect:  this calculates 8 + (17/4) + 6
    ans =
      18.2500

>> (8 + 17)/(4 + 6)   % correct
    ans =
       2.5000
```

If you use brackets, you must still use * for multiplication—multiplication is never implied by juxtaposition (normally, in mathematics, *xy* implies *x* times *y*, however that will not be the case in any programming language you use. Therefore, you must use always explicitly use * for multiplication:

```
>> (3 + 4)(5 + 1)
??? (3 + 4)(5 + 1)
            |
Error: Unbalanced or unexpected parenthesis or bracket.
Did you mean:
>> (3 + 4)*(5 + 1)
    ans =
        42
```

Note that Matlab even suggests a reasonable alternative, which you may accept (by pressing Enter), or edit as you deem appropriate.

1.4.2 Functions in MATLAB

You can also call mathematical function such as sine and cosine:

```
>> sin(2.5)^2 + cos(2.5)^2
    ans =
       1.0000
```

In mathematics class, you may have used notation such as $\sin^2(x)$ or even $\sin^2 x$ (without brackets, implying the next symbol is the argument). These representations are not possible in MATLAB (or most programming language. The argument must always be in parentheses, and it is the value of the function that is squared, hence `sin(2.5)^2`.

Some common functions are:

Routine	Description
`round(x)`	round to the nearest integer
`ceil(x)`	round to the next largest integer (round to infinity)
`floor(x)`	round to the previous smaller integer (round to negative infinity)
`fix(x)`	round toward zero (truncate)
`abs(x)`	$\|x\|$, the absolute value of x
`sqrt(x)`	$\sqrt{x}$, the square root of x
`exp(x)`	e^x, the exponential of x (e raised to the power x)
`log(x)`	$\ln(x)$, the natural logarithm of x
`log10(x)`	$\log_{10}(x)$, the common (or base-10) logarithm of x
`log2(x)`	$\log_2(x)$, the base-2 logarithm of x

A full list of the trigonometric and hyperbolic functions are:

Routine	Description	Routine	Description
`sin(x)`	sine (in radians)	`asin(x)`	inverse sine (in radians)
`cos(x)`	complementary sine (in radians)	`acos(x)`	inverse cosine (in radians)
`tan(x)`	tangent (in radians)	`atan(x)`	inverse tangent (in radians)
`sec(x)`	secant (in radians)	`asec(x)`	inverse secant (in radians)
`csc(x)`	complementary secant (in radians)	`acsc(x)`	inverse cosecant (in radians)
`cot(x)`	complementary tangent (in radians)	`acot(x)`	inverse cotangent (in radians)
`sind(x)`	sine (in degrees)	`asind(x)`	inverse sine (in degrees)
`cosd(x)`	complementary sine (in degrees)	`acosd(x)`	inverse cosine (in degrees)
`tand(x)`	tangent (in degrees)	`atand(x)`	inverse tangent (in degrees)
`secd(x)`	secant (in degrees)	`asecd(x)`	inverse secant (in degrees)
`cscd(x)`	complementary secant (in degrees)	`acscd(x)`	inverse cosecant (in degrees)
`cotd(x)`	complementary tangent (in degrees)	`acotd(x)`	inverse cotangent (in degrees)
`sinh(x)`	hyperbolic sine (in radians)	`asinh(x)`	inverse hyperbolic sine
`cosh(x)`	hyperbolic complementary sine	`acosh(x)`	inverse hyperbolic cosine
`tanh(x)`	hyperbolic tangent	`atanh(x)`	inverse hyperbolic tangent
`sech(x)`	hyperbolic secant	`asech(x)`	inverse hyperbolic secant
`csch(x)`	hyperbolic complementary secant	`acsch(x)`	inverse hyperbolic cosecant
`coth(x)`	hyperbolic complementary tangent	`acoth(x)`	inverse hyperbolic cotangent

Functions related to the trigonometric functions include:

Routine	Description
`atan2(y,x)`	a four-quadrant inverse tangent finding $\tan^{-1}(y/x)$, returning $-\tan^{-1}(y/x)$ if $x < 0$ (in radians)
`atan2d(y,x)`	a four-quadrant inverse tangent (in degrees)
`hypot(x,y)`	the length of the hypotenuse $\sqrt{x^2 + y^2}$
`deg2rad(d)`	convert degrees to radians
`rad2deg(r)`	convert radians to degrees

Some integer-valued functions of integer arguments include:

Routine	Description
`factorial(n)`	calculate $n!$
`gcd(m, n)`	find the greatest common divisor of m and n
`lcm(m, n)`	find the least common multiple of m and n

Note, you will recall from your mathematics courses that often various types of symbolism are used to represent mathematical notations; for example, $|x|$, $n!$, x^y, π, $x \le y$, $x \times y$, $x \div y$ etc. Apart from the most basic operators, such as `+` and `-`, we must use characters found on the keyboard, and other than some of the most common operations, many of these operations are implemented as functions (such as `abs(x)`).

1.4.3 Boolean-valued functions and operations

Finally, a different class of functions, called *queries*, return true or false. In MATLAB (and in C and C++), true is represented by any non-zero value (usually with `1`) and false is represented by `0`. You will meet many more queries throughout this course, but we will start with `isprime`:

`isprime(n)`	return true (1) if *n* is a prime number, and false (0) otherwise

```
>> isprime( 2 )
    ans =
        1

>> isprime( 91 )
    ans =
        0

>> isprime( 9007199254740881 )
    ans =
        1
```

It is also possible to compare two numbers, and again, the value will be either 0 or 1:

Operation	Description
`x == y`	*x* is equal to *y*
`x ~= y`	*x* is not equal to *y*
`x < y`	*x* is less than *y*
`x <= y`	*x* is less than or equal to *y*
`x >= y`	*x* is greater than or equal to *y*
`x > y`	*x* is greater than *y*

For example,

```
>> 3 <= 4
    ans =
         1

>> 4 <= 4
    ans =
         1

>> 4 < 4
    ans =
         0

>> 0.3333 == 1/3
    ans =
         0
```

1.4.4 Constants

MATLAB has a number of variables automatically assigned values, including π. You can also calculate it using the appropriate trigonometric function, e.g., $\sin^{-1}(1) = \frac{\pi}{2}$, $\cos^{-1}(-1) = \pi$ and $\tan^{-1}(1) = \frac{\pi}{4}$:

```
>> pi
    ans =
        3.1416

>> 2*asin(1)
    ans =
        3.1416

>> acos(-1)
    ans =
        3.1416

>> 4*atan(1)
    ans =
        3.1416
```

If you want *e*, you must call the exponential function:

```
>> exp(1)
    ans =
       2.718281828459046
```

Other built-in constants reflect the results of specific operations, such as calculations like $\frac{1}{0}$ and $\frac{0}{0}$:

```
>> 1/0
    ans =
       Inf

>> -1/0
    ans =
      -Inf

>> 0/0
    ans =
       NaN
```

The last stands for *not-a-number*, meaning that the answer is undefined—the number could potentially be any real number depending on the limiting factors. Matlab will always try to give you the most reasonable answer given a specific computation:

```
>> Inf - Inf
    ans =
       NaN

>> Inf - 100
    ans =
       Inf

>> Inf * -2
    ans =
      -Inf
```

```
>> 1/Inf
    ans =
       0
```

There are a few other constants associated with the double-precision floating point representation of real numbers, including `realmax` (the largest non-infinity floating-point number), `realmin` (the smallest positive non-zero full-precision floating-point number) and `eps` (the distance between 1 and the next largest floating-point number). The smallest positive floating-point number with reduced precision is `realmin*eps`:

```
>> realmin*eps
    ans =
        4.940656458412465e-324

>> 2^-1074
    ans =
        4.940656458412465e-324
```

1.4.5 Display

You may note a lack of precision when we evaluate π: only four digits after the decimal point. Internally, at least for our purposes, all numbers are stored as double-precision floating-point numbers (a *double*), storing approximately 16 decimal digits of accuracy. We can see the full precision by issuing the appropriate formatting command:

```
>> pi
    ans =
        3.1416

>> format long              % display all significant digits
>> pi
    ans =
       3.141592653589793

>> format short              % go back to the original formatting
>> pi
    ans =
        3.1416
```

Note that if MATLAB is displaying an integer, it will not print the decimal point, even though internally it is stored as a double.

```
>> 42
    ans =
        42

>> format hex    % the hexadecimal (base 16) representation of the storage
>> 42
    ans =
       4045000000000000
```

Another issue we will come across later is that occasionally, the output will be very large, covering many screens of data. It can sometimes be frustrating to scroll through so much output, so you can require Matlab to display one screen (or *page*) at a time, and it will only continue once you press Enter.

```
>> more on      % Force Matlab to divide large output into pages
>> more off     % Return to the default
```

1.4.6 Assigning to variables

A variable name may be any combination of one or more characters that satisfies the requirements that

1. the first character is a letter or an underscore, and
2. any subsequent characters are letters, numbers or underscores.

Thus, all the following are valid variable names:

```
a  n  n2  n3  top  maximum  max_value  value_  _variable _a_silly_variable_name_
```

Capitalization matters, so `m` is a different variable from `M` and `maxvalue` is different from `maxValue`. Note that if you use _ as a variable name, you may consider asking yourself whether or not you should be in engineering.

A variable may be assigned a value by you using the assignment operator =, for example

```
>> format long
>> x = pi/6
    x =
        0.523598775598299
```

Note: To save yourself from a lot of problems later on (for this programing language, and practically all others), do not read this as "*x* equals pi over six". Instead, always read this statement as "*x* is assigned the value pi over six".

You can now do mathematics with this value:

```
>> x - x^3/factorial(3) + x^5/factorial(5)      % approximate sin(pi/6)
    ans =
        0.500002132588792
```

If your line is too long, use `...` at the end before you hit Enter:

```
>> x - x^3 /factorial(3)  + x^5 /factorial(5) - x^7/factorial(7) + x^9/factorial(9) ...
     - x^11/factorial(11) + x^13/factorial(13)
    ans =
        0.500000000000000
```

If you assign a variable a second time, the original value is lost:

```
>> x
    x =
        0.523598775598299

>> x = 91
    x =
        91
```

If you use a variable before you have assigned it a value, MATLAB will issue an error:

```
>> z
??? Undefined function or variable 'z'.
```

Note that you must use * for multiplication of variables:

```
>> x = 3
    x =
         3
```

```
>> y = 4
    y =
         4

>> xy
??? Undefined function or variable 'xy'.

>> x*y
    ans =
        12
```

1.4.7 Supressing the output

Sometimes, it isn't necessary to see the output of a statement, in which case, you can supress the output by appending a semicolon. For example,

```
>> x = 3;
>> y = 4;
>> x*y
    ans =
        12
```

1.4.8 Commands related to variables

The following commands in Matlab are directly related to variables:

`who`	List all currently assigned variables
`whos`	List all currently assigned variables with additional details
`clear`	Unassign (or *clear*) all variables

For example,

```
>> x = 4;
>> y = 5;
>> who
    Your variables are:
    x  y

>> whos
      Name      Size            Bytes  Class     Attributes
      x         1x1                 8  double
      y         1x1                 8  double

>> clear
>> x
??? Undefined function or variable 'x'.
```

1.4.9 MATLAB isn't perfect: numerical error

Unfortunately, Matlab doesn't always give the correct answer:

```
>> cos(pi/2)    % This should be 0--Matlab uses radians
    ans =
         6.123233995736766e-17
```

Normally, you expect addition to be associative: $(a + b) + c$ should give the same result as $a + (b + c)$, however:

```
>> 0.1 + (16 - 16)  % This gives the right answer
```

```
    ans =
       0.100000000000000

>> (0.1 + 16) - 16  % This gives the wrong answer
    ans =
       0.100000000000001
```

You will recall that Matlab adds numbers from left to right, and even this can cause issues:

```
>> 1+1+2^53   % This gives the right answer
    ans =
         9.007199254740994e+15

>> 2^53+1+1   % This gives the wrong answer
    ans =
         9.007199254740992e+15
```

You will remember the formula for the roots of a quadratic:

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Consider the quadratic polynomial

$$\left(x - 100000000\right)\left(x - \frac{1}{100000000}\right) = x^2 - 100000000.00000001x + 1$$

which has the two roots 10^8 and 10^{-8}. We will store this in the form $ax^2 + bx + c$ with

```
>> a = 1;
>> b = -100000000.00000001;
>> c = 1;
```

Let's find the roots using the quadratic formula:

```
>> (-b + sqrt(b^2 - 4*a*c))/(2*a)
    ans =
       100000000

>> (-b - sqrt(b^2 - 4*a*c))/(2*a)
    ans =
         1.490116119384766e-08
```

The first one is very exact, but the second is off by 50%. To fix this, we could try a different formula: instead of rationalizing the denominator, as is so often done in high school to remove any radicals from the denominator, we can rationalize the numerator by multiplying by $\frac{-b \mp \sqrt{b^2 - 4ac}}{-b \mp \sqrt{b^2 - 4ac}}$ to get the formula

$$\frac{2c}{-b \mp \sqrt{b^2 - 4ac}}.$$

Trying this new formula, we get

```
>> (2*c)/(-b - sqrt(b^2 - 4*a*c))
    ans =
        67108864

>> (2*c)/(-b + sqrt(b^2 - 4*a*c))
    ans =
         1.000000000000000e-08
```

Thus, in each case, the formula performs extremely poorly at determining one of the two roots.

MATLAB stores representations of real numbers in binary, but this isn't the problem. For example, the binary representation of 0.5 is 0.1_2, while the binary representation of 0.3 is $0.0\overline{1001}_2$. The small subscript "2" is there to remind us that this is the binary representation and not 1/10 and 91/9090, respectively. The reason for this is that Matlab cannot store an infinite number of digits. Thus, 0.3 is stored as $0.0100110011001100110011001100110011001100110011001100110011_2$, truncated to only 53 binary digits, or "bits". Similarly, π is stored as $11.00100100001111110110101010001000100001011010001100_2$. This is just like in elementary school where you used 3.14 as a "good enough" approximation to π; after all, the value 6.28 m approximates the circumference of a circle of radius 1 m to within an accuracy of 3.2 mm.

This is not a course on numerical error, but in subsequent courses, you will see how steps can be taken to either avoid entirely or mitigate the effects of numerical error.

1.4.10 Getting help

There are three primary sources of help for Matlab:

1. Within Matlab, you can always type help *function name* and a text-based help page will appear. Within these help pages, Matlab commands are written in either bold or in all upper case letters; for example,

```
>> help exp
 exp    Exponential.
    exp(X) is the exponential of the elements of X, e to the X.
    For complex Z=X+i*Y, exp(Z) = exp(X)*(COS(Y)+i*SIN(Y)).

    See also expm1, log, log10, expm, expint.

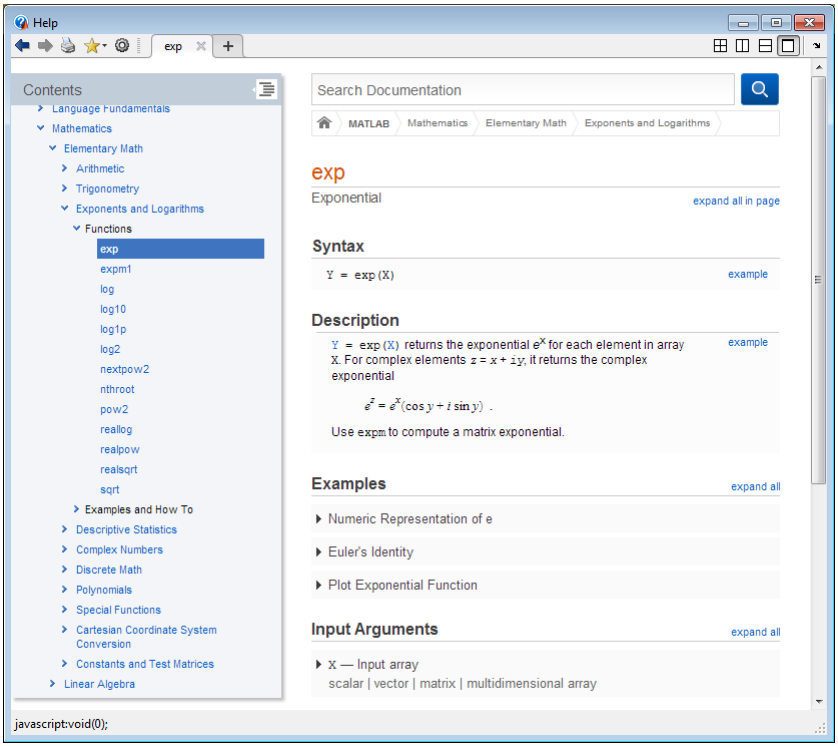
    Other functions named exp

    Reference page in Help browser
       doc exp
```

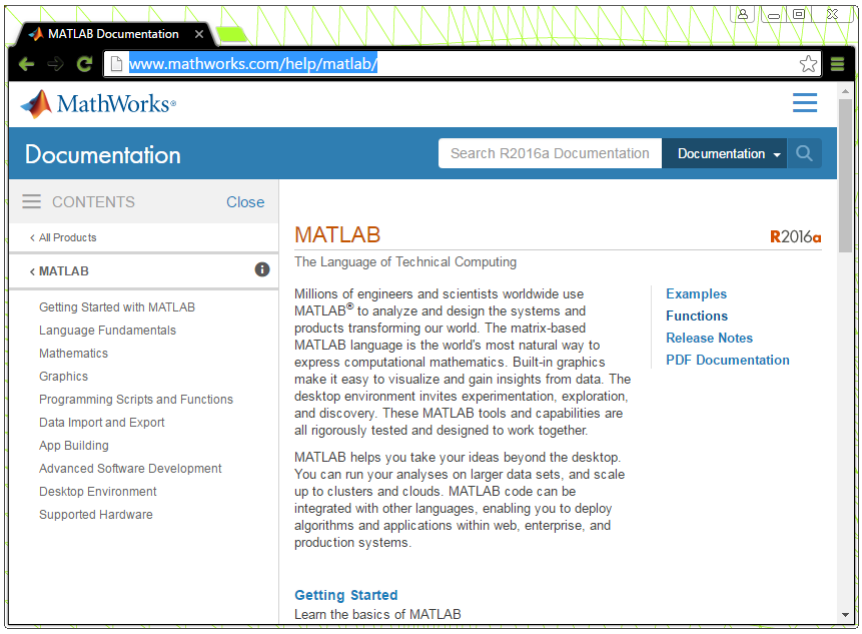
2. The Matlab help browser, which can be accessed from the ribbon by selecting the question mark icon or searching the documentation, both of which appear in the title bar:



   This brings up the help browser:


Help
exp
Contents
Language Fundamentals
Mathematics
Elementary Math
Arithmetic
Trigonometry
Exponents and Logarithms
Functions
exp
expm1
log
log10
log1p
log2
nextpow2
nthroot
pow2
reallog
realpow
realsqrt
sqrt
Examples and How To
Descriptive Statistics
Complex Numbers
Discrete Math
Polynomials
Special Functions
Cartesian Coordinate System Conversion
Constants and Test Matrices
Linear Algebra
Search Documentation
MATLAB
Mathematics
Elementary Math
Exponents and Logarithms
exp
Exponential
expand all in page
Syntax
Y = exp(X)
example
Description
Y = exp(X) returns the exponential e^X for each element in array X. For complex elements z = x + iy, it returns the complex exponential
example
e^Z = e^X(cos y + i sin y) .
Use expm to compute a matrix exponential.
Examples
expand all
Numeric Representation of e
Euler's Identity
Plot Exponential Function
Input Arguments
expand all
X — Input array
scalar | vector | matrix | multidimensional array
javascript:void(0);

3. The Matlab documentation website, available at http://www.mathworks.com/help/matlab/:



1.4.11 Summary of MATLAB

We have briefly introduced MATLAB, including how to perform basic arithmetic, calling functions, the idea of Boolean-valued operators and operations (and how 0 represents false and 1 represents true), built-in constants such as pi, controlling the display, and assigning to variables. We also showed how to suppress output and how to view a list of all variables that have been assigned values. We have also seen that MATLAB does not always give the correct or even close answer.

1.5 A quick introduction to the axiomatic method

In elementary or secondary school, you were made aware of Euclid's five axioms for geometry and you were asked to deduce additional results from this information. In addition to his axioms, Euclid also included definitions and *common notions*. His definitions included:

1. A point is that which has no part.
2. A line is breadthless length.
3. The ends of a line are points.
4. A straight line is a line which lies evenly with the points on itself.
5. A surface is that which has length and breadth only.
6. The edges of a surface are lines.
7. A plane surface is a surface which lies evenly with the straight lines on itself.
8. A plane angle is the inclination to one another of two lines in a plane which meet one another and do not lie in a straight line.
9. And when the lines containing the angle are straight, the angle is called rectilinear.
10. When a straight line standing on a straight line makes the adjacent angles equal to one another, each of the equal angles is right, and the straight line standing on the other is called a perpendicular to that on which it stands.
11. An obtuse angle is an angle greater than a right angle.
12. An acute angle is an angle less than a right angle.
13. A boundary is that which is an extremity of anything.
14. A figure is that which is contained by any boundary or boundaries.
15. A circle is a plane figure contained by one line such that all the straight lines falling upon it from one point among those lying within the figure equal one another.
16. And the point is called the center of the circle.
17. A diameter of the circle is any straight line drawn through the center and terminated in both directions by the circumference of the circle, and such a straight line also bisects the circle.
18. A semicircle is the figure contained by the diameter and the circumference cut off by it. And the center of the semicircle is the same as that of the circle.
19. Rectilinear figures are those which are contained by straight lines, trilateral figures being those contained by three, quadrilateral those contained by four, and multilateral those contained by more than four straight lines.
20. Of trilateral figures, an equilateral triangle is that which has its three sides equal, an isosceles triangle that which has two of its sides alone equal, and a scalene triangle that which has its three sides unequal.
21. Further, of trilateral figures, a right-angled triangle is that which has a right angle, an obtuse-angled triangle that which has an obtuse angle, and an acute-angled triangle that which has its three angles acute.
22. Of quadrilateral figures, a square is that which is both equilateral and right-angled; an oblong that which is right-angled but not equilateral; a rhombus that which is equilateral but not right-angled; and a rhomboid that which has its opposite sides and angles equal to one another but is neither equilateral nor right-angled. And let quadrilaterals other than these be called trapezia.
23. Parallel straight lines are straight lines which, being in the same plane and being produced indefinitely in both directions, do not meet one another in either direction.

His common notions were:

1. Things which equal the same thing also equal one another.
2. If equals are added to equals, then the wholes are equal.
3. If equals are subtracted from equals, then the remainders are equal.
4. Things which coincide with one another equal one another.
5. The whole is greater than the part.

Finally, his axioms (he called them *postulates*) were:

1. To draw a straight line from any point to any point.
2. To produce a finite straight line continuously in a straight line.
3. To describe a circle with any center and radius.
4. That all right angles equal one another.
5. That, if a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.

The first observation that many had was that the fifth axiom was significantly more complex than the first four, and consequently, for many millennia, it was wondered whether the fifth could be deduced from the first four, in which case, the fifth axiom would simply be a theorem. As it turns out, the fifth cannot be deduced from the first four, but there are numerous other theorems (*postulates*) that Euclid attempted to prove using his axiomatic system. We will look at his first:

Theorem
To construct an equilateral triangle on a given finite straight line.

Proof:
1. Let AB be the given finite straight line.
2. By Axiom 3, describe the circle BCD with center A and radius AB. Again describe the circle ACE with center B and radius BA.
3. By Axiom 1, join the straight lines CA and CB from the point C at which the circles cut one another to the points A and B.
4. Now, since the point A is the center of the circle CDB, therefore AC equals AB. By Definition 15, since the point B is the center of the circle CAE, therefore BC equals BA.
5. But AC was proved equal to AB, therefore each of the straight lines AC and BC equals AB.
6. By Common Notion 1, and things which equal the same thing also equal one another, therefore AC also equals BC.
7. Therefore the three straight lines AC, AB, and BC equal one another.
8. By Definition 20, therefore the triangle ABC is equilateral, and it has been constructed on the given finite straight line AB. ∎

This construction is shown in Figure 4.

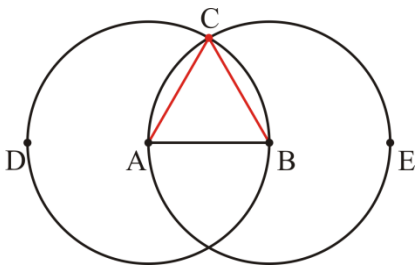


Figure 4. The construction of an equilateral triangle.

The issue with this proof is that it is assumed that the circles centred at A and B intersect at C. It is not possible to deduce this from the definitions, common notions or axioms listed by Euclid. Consequently, at least one more axiom is required. This was left unaddressed for over two thousand years until David Hilbert proposed 20 axioms for Euclidean geometry in his book *Grundlagen der Geometrie* (The Foundations of Geometry). From these 20 axioms, all the theorems proposed in Euclid's Elements could be deduced.

In secondary school, you have already been exposed to what we will call *finite-dimensional vectors*. You have added vectors together, you have multiplied vectors by a scalar value, you have taken the inner product (dot product) of two vectors, and yet there are other objects that behave the same way, as we will see. Consequently, we will base our approach on looking at what are the fundamental properties, or *axioms*, of vectors and the inner product, and take it from there. This will be very relevant as soon as your second year, at which point you will have your introductory course on quantum mechanics, and if you ask any upper year nanotechnology student, knowledge of and the ability to apply what you learn from linear algebra will be crucial to your success in that course.

1.6 Fields and complex numbers

In secondary school, you would have been introduced to *real numbers*; that is, numbers that have

1. at least one digit (possibly zero) but at most a finite number of digits to the left of the decimal point, and
2. infinitely many digits, possibly repeating, to the right of the decimal point.

For example,

$$\begin{array}{rrr} 981641.5193139816415\cdots & 981641.982418241824\cdots & 981641.9\overline{8241} \\ 9312.4275520565869\cdots & 9312.400000000000\cdots & 9312.4 \\ 1.0224241704684\cdots & 1.022420555555\cdots & 1.022420\overline{5} \\ 0.0018748445952\cdots & 0.000006231212\cdots & 0.000006231\overline{2} \end{array}$$

If at some point, all digits to the right of a digit are zero, we simply terminate the digits, and you would have seen that, for example, $14.257\overline{9}$ is equivalent to $14.258\overline{0}$ or simply 14.258. A rational number equals a real number if the division of the numerator by the denominator using long division produces the digits of the real number. For example, $\frac{1}{42} = 0.0\overline{238095}$ as

$$\begin{array}{r} 0.0238095 \\ 42\overline{)\,1\,0000000} \\ -84 \\ \hline 160 \\ -126 \\ \hline 340 \\ -336 \\ \hline 400 \\ -378 \\ \hline 220 \\ -210 \\ \hline 10 \end{array}$$

You would have been introduced to numbers like π and $\sqrt{2}$, numbers that cannot be written as a ratio of two integers (rational) and are thus classified as *irrational*.

As an aside, the proof that $\sqrt{2}$ is irrational is quite straight-forward. Any rational number can be written in the form $\frac{n}{d}$ where n and d have no common factors, and if there is a common factor, then we need only divide it out of both

the numerator and denominator. Assume that $\sqrt{2}$ is rational. Therefore $\sqrt{2} = \frac{n}{d}$ where *n* and *d* have no common factors. Therefore, $2 = \frac{n^2}{d^2}$ and thus $2d^2 = n^2$. Thus, n^2 must be divisible by 2, and if n^2 is divisible by 2, then *n* must be divisible by 2. Therefore, $n = 2m$ for some integer value of *m*. Thus we have that $\sqrt{2} = \frac{n}{d} = \frac{2m}{d}$, and therefore $2 = \frac{(2m)^2}{d^2} = \frac{4m^2}{d^2}$, and thus $2d^2 = 4m^2$ and so $d^2 = 2m^2$. This means that *d* is also divisible by 2, which contradicts our original assumption that we could write $\sqrt{2}$ as $\frac{n}{d}$ where *n* and *d* have no common factors.

To fully understand this proof, try it again, but this time to demonstrate that $\sqrt{4}$ is rational.

1.6.1 Field axioms

In secondary school, you would have been exposed to both the rational numbers and the real numbers. We will represent these two collections of numbers by the bold letters **Q** (from *quotients*) and **R**, respectively. While you learned that the irrational numbers were all those real numbers that were not rational, you never focused on them—instead, if you needed the irrational numbers (because you were computing, for example $\sqrt{2}$), you considered them as a subset of the real numbers.

The reason we focus on the rational and real numbers is because they have nice properties:

1. The rationals and reals are *closed* under addition and multiplication; that is, if α and β are both rational or real, then so are $\alpha + \beta$ and $\alpha \cdot \beta$.
2. Addition and multiplication are *associative*, meaning it doesn't what order you apply three consecutive operations, so $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$ and $(\alpha \cdot \beta) \cdot \gamma = \alpha \cdot (\beta \cdot \gamma)$.
3. Addition and multiplication are *commutative*, meaning that order does not matter: $\alpha + \beta = \beta + \alpha$ and $\alpha \cdot \beta = \beta \cdot \alpha$.
4. Both have 0 which is an additive identity; namely, $\alpha + 0 = \alpha$.
5. Both have 1 as a multiplicative identity, namely, $\alpha \cdot 1 = \alpha$.
6. Every number has an additive inverse: given α, we can find a $-\alpha$ such that $\alpha + (-\alpha) = 0$.
7. Every non-zero number has a multiplicative inverse: given α, we can find an α^{-1} such that $\alpha \cdot \left(\alpha^{-1}\right) = 1$.
8. Multiplication distributes across addition: $\alpha \cdot (\beta + \gamma) = (\alpha \cdot \beta) + (\alpha \cdot \gamma)$.

If we consider only the collection of all irrational numbers, we have one serious flaw:

Theorem
The irrationals are not closed under addition and multiplication.

Proof:
For this, we need find only one counter-example for each: note that $\sqrt{2}$, $-\sqrt{2}$ and $\sqrt{8}$ are all irrational, but $\sqrt{2} + \left(-\sqrt{2}\right) = 0$ and $\sqrt{2} \cdot \sqrt{8} = 4$, neither of which are irrational. ∎

We generally write $\alpha + (-\beta)$ as $\alpha - \beta$ and $\alpha \cdot \left(\beta^{-1}\right)$ as α / β.

You should now see why we never focus on just the irrational numbers: it is possible to add two irrational numbers and get a number that is not irrational, the easiest example of which is $\sqrt{2} - \sqrt{2} = 0$. Another issue is that neither 0 nor 1 are irrational.

As an aside, there is another field between the rational numbers and the real numbers: the field of *algebraic numbers*. These include all numbers that are roots of polynomials with integer coefficients. These include all rational numbers, as each rational number *a*/*b* is the root of the polynomial $bx - a$, but it also includes numbers such as $\sqrt{2}$ but to does not include irrational numbers such as π and *e*.

The integers, however, are not a field, for there is no multiplicative inverse of 2 that is also an integer. The only integers that have multiplicative inverses are 1 and –1.

The term *real numbers* suggests that numbers like 0, 1, $\frac{1}{3}$, π and $\sqrt{2}$ "exist", while anything that may be called an *imaginary number* does not. In reality, however, both are constructs that are used to model the real world, just like perfect triangles do not exist outside text books and mathematical constructions, but are still useful when building a wood frame for a house. Thus, engineers have found that imaginary numbers are a very appropriate and convenient tool for modelling the real world. We will look at some examples after we combine real and imaginary numbers to create *complex numbers*.

Problems

1. The integers are not a field. Which of the eight field properties are not satisfied by the integers.

2. For those interested in cryptography, suppose we define the modular integer arithmetic, where we take all operations "modulo N". Thus, we calculate $m + n$ and mn, but then we add or subtract N sufficiently often until the result is between 0 and $N - 1$, inclusive. For example, if we consider taking arithmetic modulo 5, then we can create the following tables

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

×	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

From these two tables, does it appear that the integers modulo 5 form a field?

3. Carrying on from Problem 2, are the integers modulo 6 a field?

4. Show that the sum of two irrational numbers can be a rational number other than zero.

Solutions

1. Only the 7th property fails. This is important: this is the only property that differs between the real numbers and the

3. Unlike the integers module 5, the integers module 6 are not a field. This can most easily be shown by observing that $1 \times 2 = 2$, $2 \times 2 = 4$, $3 \times 2 = 0$, $4 \times 2 = 2$ and therefore there is no multiplicative inverse of 2 modulo 6; that is, there is no integer n such that $n \times 2 = 1$.

1.6.2 Introducing *j*

The *integers*, denoted by the symbol $\mathbb{Z}$ (from the German word for integers, *Zahlen*) are *closed* under addition, subtraction, and multiplication, but not division: ½ is not an integer, even though both 1 and 2 are. To find closure under division, we must define the *rational numbers*: The rational numbers, denoted by the symbol $\mathbb{Q}$ (from quotients), are closed under addition, subtraction, multiplication and division by a non-zero rational number). Unfortunately, consider the sequence of numbers defined by

$$a_n = \sum_{k=0}^{n} \frac{1}{k!} ,$$

which results in the sequence of rational numbers

$$1, 2, \frac{5}{2}, \frac{8}{3}, \frac{64}{24}, \frac{163}{60}, \frac{1957}{720}, \frac{685}{252}, \frac{109601}{40320}, \frac{98641}{36288}, \frac{9864101}{3628800}, \ldots,$$

and if you were to write these in their decimal representation

$$1,\ 2,\ 2.5,\ 2.\overline{6},\ 2.708\overline{3}, 2.71\overline{6},\ 2.7180\overline{5},\ 2.7\overline{1825396}, 2.718278\overline{7698412},$$
$$2.71828152557319223985890 6,\ 2.7182818011463844797178130 5, \ldots$$

you will notice that they appear to be converging to a number close to 2.718281828. In your calculus course, however, you will find that this converges to a special number *e*, which is not a rational number, or *irrational*, meaning that has an infinite non-repeating decimal representation. Consequently, there are well-defined sequences of rational numbers that do not *converge* to a rational numbers, and it is hence necessary to introduce the *real numbers*. We will denote the set of real numbers by the symbol **R**.

Unfortunately, even the real numbers are insufficient to describe solutions to simple mathematical equations, and they are even ill-suited for subjects such as quantum mechanics and electromagnetism, which you will see in your 2nd year. The some weakness can be summarized as follows:

1. some quadratic polynomials with real coefficients have two real roots
2. some have only one (a double root), and
3. some have no real roots.

The simplest examples of this are the three polynomials

$$x^2 - 1 , \ x^2 \text{ and } x^2 + 1 ,$$

respectively. Now, in high school, you learned that the roots of a quadratic polynomial $ax^2 + bx + c$ are given by

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

where the term under the square root, $b^2 - 4ac$, is called the *discriminant*.

Remember that the square root *s* of a positive real number *r* is the positive number *s* such that $s^2 = r$. Thus, while both $x = 2$ and $x = -2$ have the property that $x^2 = 4$, we will define $\sqrt{4} = 2$.

Thus,

1. if $b^2 - 4ac > 0$, we have two real roots,
2. if $b^2 - 4ac = 0$, we have a repeated root at $-\frac{b}{2a}$, and
3. if $b^2 - 4ac < 0$, we have no real roots.

If we blindly apply the formula to $x^2 + 4$ and use the fact that $\sqrt{xy} = \sqrt{x}\sqrt{y}$, we find that the two roots are

$$\pm\frac{\sqrt{-4\cdot 4}}{2} = \pm\frac{\sqrt{-1\cdot 16}}{2} = \pm\frac{\sqrt{-1}\sqrt{16}}{2} = \pm 2\sqrt{-1}\,.$$

Recalling that $(xy)^2 = x^2y^2$ and that $\left(\sqrt{x}\right)^2 = x$ for $x > 0$, if we plug either of these into the original polynomial, we see that each is a root:

$$\begin{aligned}\left(2\sqrt{-1}\right)^2 + 4 &= \left(\sqrt{-1}\right)^2 (2)^2 + 4\\ &= \left(\sqrt{-1}\right)^2 4 + 4\\ &= -4 + 4 = 0\end{aligned}$$

and

$$\begin{aligned}\left(-2\sqrt{-1}\right)^2 + 4 &= (-1)^2\left(\sqrt{-1}\right)^2 (2)^2 + 4\\ &= \left(\sqrt{-1}\right)^2 4 + 4\\ &= -4 + 4 = 0\end{aligned}$$

These only make sense if we define $\left(\sqrt{-1}\right)^2 = -1$, but if we do this, then note also that

$$\left(-\sqrt{-1}\right)^2 = (-1)^2\left(\sqrt{-1}\right)^2 = \left(\sqrt{-1}\right)^2$$

where

$$\sqrt{-1} + \left(-\sqrt{-1}\right) = \left(1 + (-1)\right)\left(\sqrt{-1}\right) = 0\,,$$

and so that there is both a *positive* $\sqrt{-1}$ and a *negative* $-\sqrt{-1}$ (we will call the second its *additive inverse*, just like the value –2 is the additive inverse of 2, and 2 is the additive inverse of –2).

Unfortunately, always writing $\sqrt{-1}$ will become very tedious very quickly, and thus we will, instead, just define the symbol

$$j \overset{\text{def}}{=} \sqrt{-1}\,.$$

The notation $\overset{\text{def}}{=}$ indicates that the left-hand side is, by definition, equal to the right-hand side.

Why *j* and not *i*? Very simple: engineering is an applied science, and *i* had been used to represent current before complex numbers were applied to the discipline of engineering (think $V = IR$); therefore, the use of *i* for current and the imaginary unit would lead to significant confusion and error. The use of *i* for current goes back to Ampère, who referred to electric current as "*l'intensité du courant électrique*". For example, see his publication *Recuil d'Observations Électro-dynamiques*, Paris, Chez Crochard Libraire, 1822.

Now we can find, for example, the root of the polynomial $9x^2 + 16$:

$$\begin{aligned}\frac{-0 \pm \sqrt{0 - 4 \cdot 9 \cdot 16}}{2 \cdot 9} &= \pm \frac{\sqrt{-576}}{18} \\ &= \pm \frac{\sqrt{576}\sqrt{-1}}{18} \\ &= \pm \frac{24j}{18} \\ &= \pm \frac{4j}{3}\end{aligned}$$

and therefore the roots are $-\frac{4}{3}j$ and $\frac{4}{3}j$. If you substitute these back into the polynomial (using *foil* to multiply two complex numbers), you get the expected result:

$$\begin{aligned}9\left(\frac{4}{3}j\right)^2 + 16 &= 9\left(\frac{4}{3}\right)^2 j^2 + 16 \\ &= -16 + 16 = 0\end{aligned}$$

and

$$\begin{aligned}9\left(-\frac{4}{3}j\right)^2 + 16 &= 9\left(-\frac{4}{3}\right)^2 j^2 + 16 \\ &= -16 + 16 = 0\end{aligned}$$

Thus, we see that both are $-\frac{4}{3}j$ and $\frac{4}{3}j$ are roots of the polynomial, but only if we understand that $j^2 = -1$. Note that for any real number *x*, $(xj)^2 = x^2 j^2 = -x^2$.

Notice that we can treat *j* just like any other variable, only if we raise *j* to an integer exponent, we can calculate its value:

$$\begin{aligned}j^2 &= -1 \\ j^3 &= -j \\ j^4 &= 1 \\ j^5 &= j\end{aligned}$$

This is because the multiplicative inverse of *j* is $-j$: $j \cdot (-j) = -j^2 = -(-1) = 1$. Thus, we also have that $j^{-1} = -j$, so in general, we can say that $j^{4n} = 1$, $j^{4n+1} = j$, $j^{4n+2} = -1$ and $j^{4n+3} = -j$, or

$$j^n = j^{n \bmod 4}$$

where $n \bmod 4$ is the remainder when you divide *n* by four, so 7 mod 4 is 3.

We will define the set of complex numbers as all numbers of the form

$$\alpha + j\beta$$

where α and β are real numbers. We will denote this set of numbers by **C**.

In some cases, we may write $\alpha + j\beta$ and in others we may write $\alpha + \beta j$. We could even write a complex number as
$j\beta + \alpha$ or $\beta j + \alpha$, but this will usually not be the case.

MATLAB application

If you launch MATLAB, you will be met with a prompt

```
>>
```

You can now start typing commands, so we will begin with entering a complex number:

```
>> 3 + 4j
  ans =
       3.0000 + 4.0000i
```

Note first that while you used "`j`", MATLAB continues to display complex numbers with the more usual notation of "`i`", and second that MATLAB is storing these numbers as floating-point numbers, not as integers.

You can assign a complex number to a variable of your choice:

```
>> z = 3 + 4j
  z =
       3.0000 + 4.0000i
```

We will continue using this example in subsequent sections.

Notice: For students who have little programming experience: the statement `z = 3 + 4j` looks like an equation; however, in almost all programming languages, rather than saying

"z equals 3 + 4*j*,"

you should read this as

"z is assigned the value 3 + 4*j*."

If you get into the habit of saying this in your mind, you will save yourself significant stress later on in life. Later, we will see that many programming languages use `==` for a Boolean-valued operator that returns true if both sides are equal and false otherwise.

Some programming languages take a different approach: Maple, for example, uses `:=` for assignment and `=` for equality testing.

Problems

1. Write $\sqrt{-4}$, $\sqrt{-10}$ and $\sqrt{-20}$ using *j*.

2. Write $\sqrt{-7.29}$ and $\sqrt{-\frac{9}{25}}$ using *j*.

3. What are the values of j^{1001}, j^{1002}, $j^{1003}$and j^{1004}?

4. What are the roots of the polynomial $5x^2 + 12$?

5. What is so special about the roots of a polynomial $ax^2 + c$?

6. What are the roots of the polynomial $10x^2 + 7x + 1$? How do these compare to the roots of $10x^2 + 6x + 1$?

7. Given the polynomial $ax^2 + bx + c$, what are the position of the roots relative to $-\frac{b}{2a}$?

Solutions

1. $2j$, $\sqrt{10}j$ and $2\sqrt{5}j$.

3. $j^4 = 1$, and thus $j^{1001} = j^{4\times 250+1} = \left(j^4\right)^{250} j^1 = (1)^{250} j = j$, so $j^{1002} = -1$, $j^{1003} = -j$ and $j^{1004} = 1$.

5. If $b = 0$, the roots are $\frac{\pm\sqrt{-4ac}}{2a} = \pm\sqrt{-\frac{c}{a}}$, so if both a and c are both positive or both negative, both roots are a real number times j, and if one of a and c is positive and the other negative, both are real.

7. The roots are equally spaced around this point, either with $\pm r$ or $\pm jr$ for some real number r.

1.6.3 Complex numbers real and their components

We will define the field of complex numbers as the collection of all numbers of the form:

$$\mathbf{C} = \{\alpha + j\beta : \alpha \in \mathbf{R} \wedge \beta \in \mathbf{R}\}.$$

Given a complex number $z = \alpha + \beta j$, we define the *real component* or *real part* of the complex number as *a* and denote it by $\Re\mathrm{e}(z) = \Re\mathrm{e}(\alpha + \beta j) \overset{\text{def}}{=} \alpha$, and we define the *imaginary component* or *imaginary part* of the complex number as *b* and denote it by $\Im\mathrm{m}(z) = \Im\mathrm{m}(\alpha + \beta j) \overset{\text{def}}{=} \beta$. A complex number *z* is said to be *real* if $\Im\mathrm{m}(z) = 0$ and all other complex numbers are said to be *imaginary*. If a complex number has zero real part, that is, $\Re\mathrm{e}(z) = 0$, then we say that it is *purely imaginary*. Note that zero is simultaneously real and purely imaginary, but not imaginary. ☺

When a complex number is written in the form $\alpha + j\beta$, we will call this the *rectangular representation*.

Complex numbers *z* and *w* are *equal* if and only if $\Re\mathrm{e}(z) = \Re\mathrm{e}(w)$ and $\Im\mathrm{m}(z) = \Im\mathrm{m}(w)$.

MATLAB application

The routines in MATLAB that return the real and imaginary components of a complex number are `real(…)` and `imag(…)`, respectively:

```
>> z = 3 + 4j
  z =
       3.0000 + 4.0000i
>> real( z )
  ans =
       3

>> imag( z )
  ans =
       4
```

We will now introduce a query (that is, routines that return true (`1`) or false (`0`), also known as a Boolean-valued routine). The function `isreal` returns true if the imaginary component is zero, and false otherwise.

```
>> isreal( z )
  ans =
       0
>> isreal( Re( z ) )
  ans =
       1
```

A complex number *z* is *purely imaginary* if *zj* is real:

```
>> w = 0.0 + 5.2j
  w =
       0.0000 + 5.2000j
>> isreal( z*1j )          % remember to use 1j for the complex number 0 + j
  ans =
    1
```

Problems

1. What are the real and imaginary components of $3 + 4j$?

2. What are the real and imaginary components of $-2 - 5j$?

3. If $z = 3.24 - 2.59j$, what is $\Re\mathrm{e}(z)$ and what is $\Im\mathrm{m}(z)$?

4. If $w = -12.4 - 1.13j$, what is $\Re\mathrm{e}(w)$ and what is $\Im\mathrm{m}(w)$?

5. What is the complex number such that $\Re\mathrm{e}(z) = 2.54$ and $\Im\mathrm{m}(z) = 7.13$?

6. What is the complex number such that $\Re\mathrm{e}(z) = -7.35$ and $\Im\mathrm{m}(z) = -5.04$?

7. According to the definition of a complex number, is $-5.04 + \infty j$ a complex number?

8. According to the definition of a complex number, is $-\infty + 0.2j$ a complex number?

9. Is $0 + 0j$ different from 0? Is 1 different from either $1 + 0j$, or is $1 + 1j$ different from $1 + j$? Is $0 + 2j$ different from either $2j$ or $j2$? You may ask yourself, is $\frac{1}{2}$ different from 0.5, and is $0.\bar{3}$ different from $\frac{1}{3}$?

10. In MATLAB, how do the results of the first two statements differ from the results of the last two?

```
>> 3 + 4j
>> 3 + 4*j
>> j = 5              % assignment in Matlab
>> 3 + 4j
>> 3 + 4*j
```

Solutions

1. The real component of $3 + 4j$ is 3 and the imaginary component is 4.

3. $\Re\mathrm{e}(z) = 3.24$ and $\Im\mathrm{m}(z) = -2.59$.

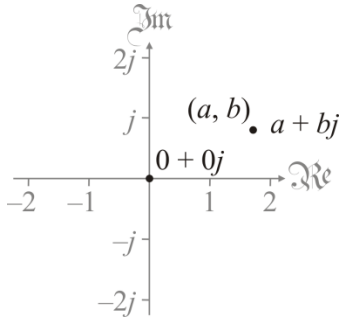
5. $2.54 + 7.13j$

7. No, because the definition of a complex number requires that the real component is a real number, and ∞ is not a real number.

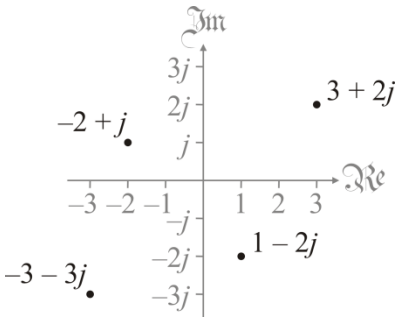
9. No, they can be considered to be equal. Therefore, it is not wrong to write that $3.2 = 3.2 + 0j = 3.2 - 0j$, although both require significantly more writing. Similarly, it is easier to write $2j$ as opposed to either $0 + 2j$ or $-0 + 2j$.

1.6.4 Geometric interpretation of a complex number

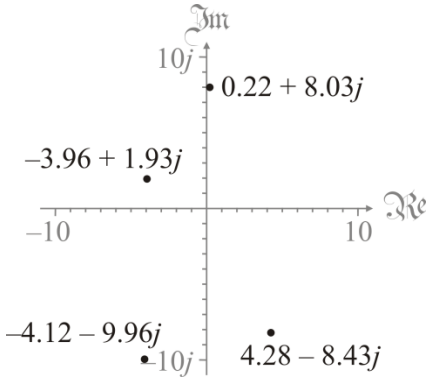
Given a complex number $\alpha + \beta j$, we can represent that number by plotting the point (*a*, *b*) on the Cartesian plane, as shown in . The abscissa[2] and ordinate[3] (horizontal and vertical axes, respectively) will be labeled as The origin represents 0 = 0 + 0*j*.



For example, shows the complex numbers 3 + 2*j*, –2 + *j*, –3 – 3*j* and 1 – 2*j*.



shows the points 0.22 + 8.03*j*, –3.96 + 1.93*j*, –4.12 – 9.96*j* and 4.28 – 8.43*j*.



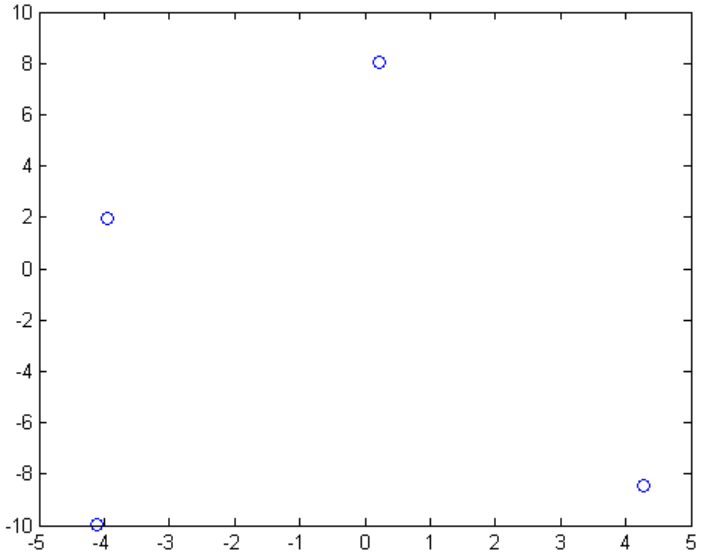
[2] From Latin: the noun *ōrdinātus* meaning orderly, regular, regulated. Reference: OED.com

[3] From Latin: the transitive verb *abscindere* meaning to tear or cut off, to separate. *Ibid*.

We will now look at plotting complex numbers in MATLAB

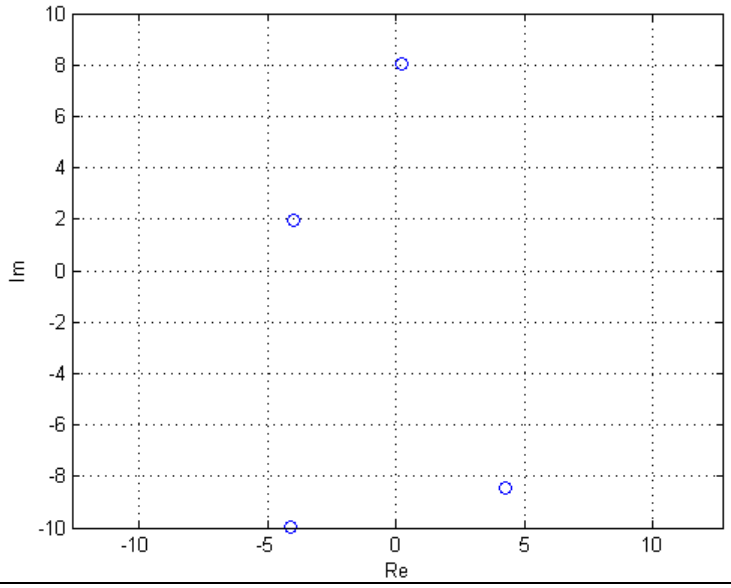
```
>> plot( [0.22+8.03j, -3.96 + 1.93j, -4.12-9.96j, 4.28-8.43j], 'o' )
```

We give plot an array of four values and the second argument is the MATLAB representation of a string, in this case, a single character little-o indicating that the points should be plotted with circles.



There are a few things that are unsatisfying about this plot, and thus we can modify it:

```
>> ylim( [-10, 10] )    % set the y-axis to span from -10 to 10
>> axis equal           % require that the spacing in the x- and y-axes is the same
>> grid on              % impose a grid on the plot
>> xlabel( 'Re' )       % give the x-axis the label 'Re' (again, here 'Re' is a string)
>> ylabel( 'Im' )       % give the y-axis the label 'Im'
```

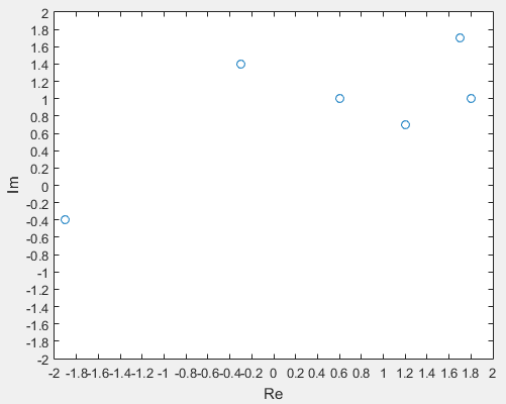


Problems

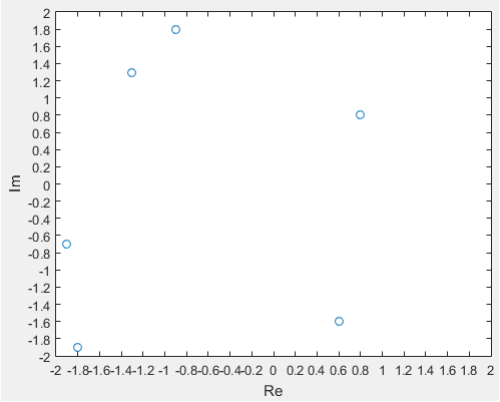
1. Plot the points $1 + 3j$, $2 - j$, $-3j$, $-0.5 + 1.5j$ and -2.5 on the complex plane.

2. Plot the points 3.4, $-2.3 + 1.4j$, $2.1j$, $-1.9j$, $-0.7 + 1.7j$ and -2.1 on the complex plane.

3. What complex numbers, to the nearest 0.1, are shown in this plot?

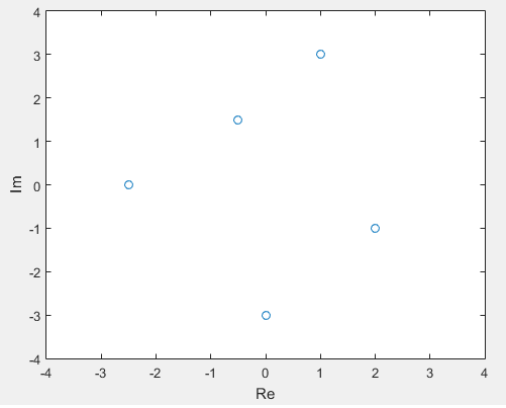


4. What complex numbers, to the nearest 0.1, are shown in this plot?



Solutions

1. The plotted points are shown here (although, you may have the axes in the middle).



3. –0.3 + 1.4*j*, 1.7 + 1.7*j*, 1.2 + 0.7*j*, 1.8 + *j*, 0.6 + *j* and –1.9 – 0.4*j*

1.6.5 Magnitude or absolute value of a complex number

Given our geometric interpretation of a complex number, it makes sense to define the absolute value of a complex number $z = a + bj$ as being the *distance* from the origin (0, 0) to the point (*a*, *b*), in which case,

$$|z| = \sqrt{a^2 + b^2} \,,$$

as shown in Figure 5.

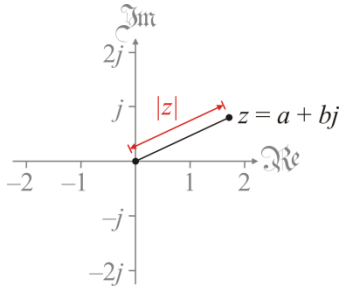


Figure 5. Magnitude of a complex number.

Thus, $|j| = |-j| = 1$ and $|bj| = |b|$.

MATLAB application

The absolute value of a complex (or real) number is found using the `abs` routine:

```
>> abs( z )
   ans =
        5.0000

>> abs( 1j )
   ans =
        1.0000

>> abs( -52.4j )
   ans =
       52.4000
```

Note: We could use `j` instead of `1j`, but in MATLAB, `j` can also be used as a variable. Similarly, we could use

```
>> z = 3 + 13*j
  z =
        3.0000 + 13.0000i
```

but, again, if `j` is assigned a value, you may get some weird results:

```
>> j = 51;
  z =
       51
>> z = 3 + 13*j
  z =
      666
>> z = 3 + 13j
  z =
        3.0000 + 13.0000i
```

Problems

1. What are the magnitudes of the complex numbers $2 + j$, $1 - 2j$, $-2 - j$ and $-1 + 2j$?

2. What are the magnitudes of the complex numbers $1-\frac{4}{3}j$, $-3+\frac{5}{4}j$ and $4-\frac{15}{2}j$?

3. What are the magnitudes of the complex numbers 3, $2j$, -4 and $-5j$?

4. The square root of a real number can be imaginary. Why is it not possible for the absolute value (or magnitude) of a complex number to be imaginary?

5. Find two different complex numbers z such that $\Re\mathrm{e}(z) = 2$ and $|z| = 3$.

6. Find two different complex numbers z such that $\Im\mathrm{m}(z) = -1$ and $|z| = 2$.

7. What is the magnitude or absolute value of $0 + 0j$?

8. Can any complex number not equal to zero have magnitude or absolute value equal to zero? Justify your answer.

9. Plot all points on the complex plane that satisfy the equation $|z| = 2$.

10. Plot all points on the complex plane that have the same absolute value or magnitude of $z = -1 - 3j$.

Solutions

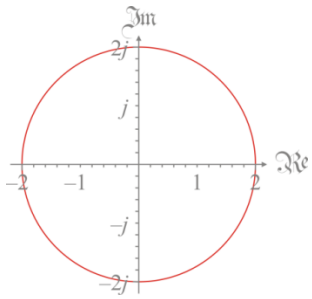
1. They all have magnitude $\sqrt{5}$.

3. 3, 2, 4 and 5, respectively.

5. If $\Re\mathrm{e}(z) = 2$ then $z = 2 + j\beta$, and $|z| = \sqrt{4+\beta^2}$ so if $\sqrt{4+\beta^2} = 3$ then $4 + \beta^2 = 9$ or $\beta = \pm\sqrt{5}$, so two different complex numbers that satisfy these requirements are $2 + j\sqrt{5}$ and $2 - j\sqrt{5}$; or, to write both at the same time, $2 \pm j\sqrt{5}$.

7. 0

9. All points with an absolute value or magnitude of 2 exist on a circle with radius 2 centered at the origin in the complex plane.



1.6.6 The angle or argument of a complex number and polar representations

Given our geometric interpretation, we note that we can consider each complex number to have an angle between it and the positive real axis, as shown in Figure 6.

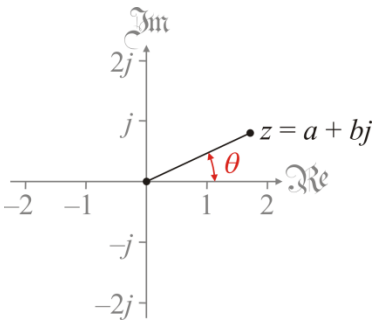


Figure 6. The *argument* of a complex number.

We will call this value the *argument* of the complex number. If we measure the angle as θ, then $\theta + 2\pi n$ is also a measure of the angle for every integer value of *n*. Consequently, we define the *principal argument* to be the angle restricted restrict the angle to the interval $(-\pi, \pi]$. Thus, every non-zero complex number may be represented uniquely by a pair (r, θ) where r is the absolute value of the complex number and θ is the argument. When a complex number is written as an absolute-value—argument pair, it is usually written as $r\angle\theta$, and this is read as "r phase θ". Many times in engineering, the argument will be written in degrees, and so the following are all equivalent:

$$
\begin{array}{ccccc}
1 & = & 1\angle 0 & = & 1\angle 0° \\
j & = & 1\angle \frac{\pi}{2} & = & 1\angle 90° \\
-1 & = & 1\angle \pi & = & 1\angle 180° \\
3+4j & \approx & 5\angle 0.9272 & \approx & 5\angle 53.13° \\
-2-2j & = & 2\sqrt{2}\angle -\frac{3}{4}\pi & = & 2\sqrt{2}\angle -135° \\
1-\sqrt{3}j & = & 2\angle -\frac{1}{3}\pi & = & 2\angle -60°
\end{array}
$$

If $z = r\angle\theta$, we will write that arg(z) = θ..

Note that $\angle$ is not the less-than sign. The phase symbol represents an angle. Many engineers and instructors will use $\angle$; e.g., $2\angle -60°$.

MATLAB application

The (principal) argument complex (or real) number is found using the `angle` routine:

```
>> angle( z )
   ans =
       0.9273

>> angle( 1j )
   ans =
       1.5708

>> angle( -52.4j )
   ans =
      -1.5708
```

In your Calculus course, you will also find that $r\angle\theta = re^{j\theta}$; that is, it is possible to raise $e = 2.7182818284590\cdots$ to a complex exponent, and so $e^{j\theta} = 1\angle\theta$. In some of your engineering courses, you will see $r\angle\theta$ while in others you will see $re^{j\theta}$. Sorry, that's life.

Problems

1. What is the argument of 1, 1 + *j*, *j*, –1 + *j*, –1, –1 – *j*, –*j* and 1 – *j*?

2. Given that the argument of 1 + 2*j* is approximately $63.435°$, what is a complex number that has an argument approximately equal to $26.565°$? What is a complex number with an argument approximately equal to $-153.435°$?

3. How would you describe all complex numbers that have an argument of $60°$ on the complex plane?

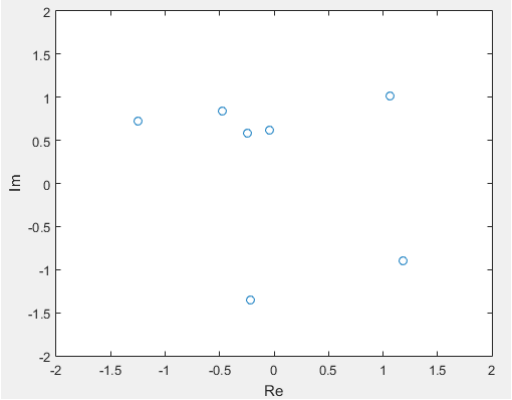
4. How would you describe all complex numbers with an argument of -0.9π?

5. Is it fair to say that the phase of 0 is any number in $(-\pi, \pi]$ and we only choose to represent 0 by $0\angle 0$ out of convenience?

6. Using your calculator, find the complex number that has a phase of -0.3π and a magnitude equal to 2.

7. You have the following representation of seven complex numbers, together with the plot of those numbers, but all of the lists got mixed up. Find the representations that represent the same points in the plot without doing any mathematics.

–0.2450 + 0.5853*j*	1.4822∠–0.6491	0.6345∠112.7172°
–0.4738 + 0.8375*j*	0.6345∠ 1.9673	1.4822∠–37.1884°
1.0621 + 1.0187*j*	1.4717∠ 0.7646	0.6217∠ 93.7757°
1.1808 – 0.8959*j*	1.3670∠–1.7307	0.9622∠ 119.4976°
–1.2525 + 0.7188*j*	0.6217∠ 1.6367	1.4441∠ 150.1486°
–0.0409 + 0.6204*j*	0.9622∠ 2.0856	1.4717∠ 43.8073°
–0.2177 – 1.3496*j*	1.4441∠ 2.6206	1.3670∠–99.1617°



Solutions

1. 0, $\frac{\pi}{4}$, $\frac{\pi}{2}$, $\frac{3\pi}{4}$, π, $-\frac{3\pi}{4}$, $-\frac{\pi}{2}$ and $-\frac{\pi}{4}$, respectively.

3. All complex numbers with an argument of $60°$ would be all complex numbers on the complex plane extending out from 0 in a line that is $60°$ above the positive real axis.

5. Yes.

7. The point in the top-right corner must be 1.0621 + 1.0187*j*. It has the smallest positive angle, so it must also have the representations of $1.4717\angle 0.7646$ and $1.4717\angle 43.8073°$, and—of course—have the same magnitude or absolute value. Which point has the next smallest argument?

1.6.7 Switching between representations

To convert from rectangular coordinates to polar coordinates, if $z = a + jb$, it follows that $z = |z| \angle \arg(z)$, where $\tan(\theta) = \frac{b}{a}$, but as the tangent function is not one-to-one, we must be careful with our selection: the arctangent function returns a value on the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, but we require a value on the range $(-\pi, \pi]$, and thus we define:

$$\text{angle}(z) = \text{angle}(a + jb) = \begin{cases} \arctan\left(\frac{b}{a}\right) & a > 0 \\ \frac{\pi}{2} & a = 0 \wedge b > 0 \\ 0 & a = 0 \wedge b = 0 \\ -\frac{\pi}{2} & a = 0 \wedge b < 0 \\ \arctan\left(\frac{b}{a}\right) + \pi & a < 0 \wedge b \geq 0 \\ \arctan\left(\frac{b}{a}\right) - \pi & a < 0 \wedge b < 0 \end{cases}$$

You should read the $\wedge$ operator as the logical operator AND—both conditions must be true.

To go from polar coordinates to rectangular coordinates, if $z = r \angle \theta$, it follows that $z = r\cos(\theta) + r\sin(\theta)\, j$.

The numbers $0.22 + 8.03j$, $-3.96 + 1.93j$, $-4.12 - 9.96j$ and $4.28 - 8.43j$ have polar representations of

$$\begin{aligned} 0.22 + 8.03\,j &\approx 8.033 \angle 1.543 \approx 8.033 \angle 88.43° \\ -3.96 + 1.93\,j &\approx 4.405 \angle 2.688 \approx 4.405 \angle 154.02° \\ -4.12 - 9.96\,j &\approx 10.778 \angle -1.963 \approx 10.778 \angle -112.47° \\ 4.28 - 8.43\,j &\approx 9.454 \angle -1.101 \approx 9.454 \angle -63.08° \end{aligned}$$

MATLAB application

In MATLAB, the `abs` and `angle` functions map onto each entry of an array:

```
>> abs( [0.22+8.03j, -3.96+1.93j, -4.12-9.96j, 4.28-8.43j] )
    ans =
        8.0330    4.4053   10.7785    9.4543

>> angle( [0.22+8.03j, -3.96+1.93j, -4.12-9.96j, 4.28-8.43j] )
    ans =
        1.5434    2.6881   -1.9630   -1.1010
```

This leads to Euler's formula, a formula that you that you will see again in your calculus course:

$$1 \angle \theta = e^{j\theta} = \cos(\theta) + \sin(\theta)\, j\,.$$

From this it follows that $r \angle \theta = re^{j\theta} = r\cos(\theta) + r\sin(\theta)\, j$.

For those of you who have seen Taylor series before, you may be wondering how is this at all possible? Remember that

$(aj)^n = a^n j^n$, so given the Taylor series expansion of e^x,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \frac{x^7}{7!} + \cdots,$$

if we substitute $x = j\theta$, we have

$$e^{j\theta} = 1 + j\theta + \frac{(j\theta)^2}{2!} + \frac{(j\theta)^3}{3!} + \frac{(j\theta)^4}{4!} + \frac{(j\theta)^5}{5!} + \frac{(j\theta)^6}{6!} + \frac{(j\theta)^7}{7!} + \cdots.$$

We can now rewrite the powers:

$$e^{j\theta} = 1 + j\theta + \frac{j^2\theta^2}{2!} + \frac{j^3\theta^3}{3!} + \frac{j^4\theta^4}{4!} + \frac{j^5\theta^5}{5!} + \frac{j^6\theta^6}{6!} + \frac{j^7\theta^7}{7!} + \cdots.$$

Recall that $j^2 = -1$, $j^3 = -j$, $j^4 = 1$, etc., so

$$e^{j\theta} = 1 + j\theta - \frac{\theta^2}{2!} - j\frac{\theta^3}{3!} + \frac{\theta^4}{4!} + j\frac{\theta^5}{5!} - \frac{\theta^6}{6!} - j\frac{\theta^7}{7!} + \cdots.$$

If we combine all those terms not multiplied by j and all those terms multiplied by j, we get

$$e^{j\theta} = \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots\right) + j\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots\right).$$

You may recognize that the first parenthesized series is the Taylor series of $\cos(\theta)$ while the second is the Taylor series of $\sin(\theta)$, so $e^{j\theta} = \cos(\theta) + j\sin(\theta)$.

MATLAB application

In MATLAB, we can observe this property:

```
>> format long
>> exp(3.2j)
    ans =
       -0.998294775794753 - 0.058374143427580i
>> cos(3.2) + 1j*sin(3.2)
    ans =
       -0.998294775794753 - 0.058374143427580i
```

1.6.8 Complex arithmetic

The next step is to define arithmetic for complex numbers. We will describe addition, the additive inverse, subtraction, multiplication, complex conjugation which can be used to find multiplicative inverses, division and exponentiation.

1.6.8.1 Addition

If $z = \alpha + j\beta$ and $w = \gamma + \delta j$, then

$$\begin{aligned} z + w &= (\alpha + \beta j) + (\gamma + \delta j) \\ &= (\alpha + \gamma) + (\beta j + \delta j) \\ &= (\alpha + \gamma) + (\beta + \delta) j \end{aligned}$$

For example, (3.2 + 4.3*j*) + (5.2 – 2.1*j*) equals (3.2 + 5.2) + (4.3 – 2.1)*j* = 8.4 + 2.2*j*. The geometric interpretation of complex addition may be visualized by considering the two complex number *z* and *w* as arrows originating from the origin. The sum is determined by placing the tail of one of the two arrows at the head of the other, as shown in Figure 7.

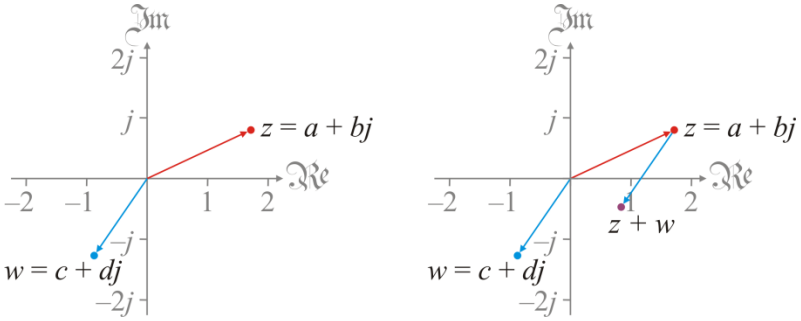


Figure 7. A geometric interpretation of complex addition.

MATLAB application

The + operator can be used to add two complex numbers

```
>> z + 3 - 1j - 1 + 4j
   z =
       5.0000 + 7.0000i

>> 2 + 4 + 6 + 8
   ans =
       20

>> ans + 3j
   ans =
       20.0000 + 3.0000i
```

We just introduced a new novelty of MATLAB: if you do not assign a result to anything, that result is assigned to the variable `ans`. The value of `ans` remains unchanged until your next statement that has an unassigned result.

```
>> 3
    ans =
        3
>> v = ans + 2
    v =
        5
>> ans
    ans =
        3
```

Problems

1. Calculate the sums $(-5.2 + 2.1j) + (8.9 - 5.4j)$, $(3.2 - 5.4j) + (2.7 + 5.4j)$, $(2.1 - 4.2j) + (-2.1 + 4.2j)$ and $(7.2 + 2.9j) + (-7.2 - 5.5j)$.

2. Is complex arithmetic commutative? That is, if $z = \alpha + \beta j$ and $w = \gamma + \delta j$, does $z + w = w + z$?

3. Is complex arithmetic associative? That is, if $z_1 = \alpha_1 + \beta_1 j$, $z_2 = \alpha_2 + \beta_2 j$ and $z_1 = \alpha_1 + \beta_1 j$.

4. Add the following 10 complex numbers together:

$$2 - 2j \quad -4 + 5j \quad 3 + 2j \quad 4 + 7j \quad 5 - 5j \quad -8 - j \quad -5 + 9j \quad 8 - 9j \quad -2 + j \quad -3 - 7j$$

5. For complex numbers w and z, is $\Re\mathrm{e}(w+z) = \Re\mathrm{e}(w) + \Re\mathrm{e}(z)$?

6. For complex numbers w and z, is $\Im\mathrm{m}(w+z) = \Im\mathrm{m}(w) + \Im\mathrm{m}(z)$?

7. In adding n complex numbers, can you add the real and imaginary parts separately?

Solutions

1. $3.7 - 3.3j$, 5.9, 0, $-2.6j$

3. Yes, as

$$\begin{aligned} z_1 + (z_2 + z_3) &= (\alpha_1 + \beta_1 j) + ((\alpha_2 + \beta_2 j) + (\alpha_3 + \beta_3 j)) \\ &= (\alpha_1 + \beta_1 j) + ((\alpha_2 + \alpha_3) + (\beta_2 + \beta_3) j) \\ &= (\alpha_1 + (\alpha_2 + \alpha_3)) + (\beta_1 + (\beta_2 + \beta_3)) j \\ &= (\alpha_1 + \alpha_2 + \alpha_3) + (\beta_1 + \beta_2 + \beta_3) j \end{aligned}$$

because all of α_1, α_2, α_3, β_1, β_2 and β_3 are real, and if we do the same thing with $(z_1 + z_2) + z_3$, you get the same result.

5. Yes, for if $z = \alpha + \beta j$ and $w = \gamma + \delta j$, then $z + w = (\alpha + \gamma) + (\beta + \delta) j$, and $\Re\mathrm{e}(w + z) = \alpha + \gamma$, while on the other hand, $\Re\mathrm{e}(w) + \Re\mathrm{e}(z) = \alpha + \gamma$.

7. That we can do this follows from the previous result.

1.6.8.2 The additive inverse

If $z = a + bj$, the *additive inverse*, or that number $-z$ such that $z + (-z) = 0$, is

$$\begin{aligned} -z &= -(a+bj) \\ &= -a-bj \end{aligned}$$

Thus, the additive inverse of $8.2 - 2.3j$ is $-8.2 + 2.3j$, and we note that $(8.2 - 2.3j) + (-8.2 + 2.3j) = 0$. The geometric interpretation of the additive inverse is a reflection through the origin, as shown in Figure 8.

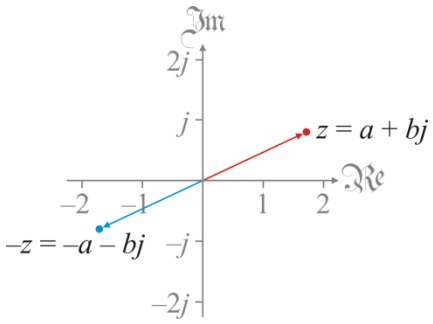


Figure 8. A geometric interpretation of the additive inverse.

MATLAB application

The – operator can be used either as a unary operator, or as a binary operator. MATLAB is one of the few programming languages where $--z = -(-z) = z$.

```
>> -z
   z =
      -3.0000 - 4.0000i
>> -z + 3
    ans =
             0 - 4.0000i
>> --z
    ans =
       3.0000 + 4.0000i
```

Problems

1. What are the additive inverses of $3.7 - 4.7j$, $-3.2j$, 42.0 and 0?

2. Is $-(w+z) = (-w)+(-z)$? That is, is the additive inverse of a sum the sum of the additive inverses?

3. What the statement $\Re e(-z) = -\Re e(z)$ say? Recall that z is a complex number while $\Re e(z)$ is a real number.

4. How do the angles of z and $-z$ differ?

5. If $z = -z$, what does this say about z?

Solutions

1. $-3.7 + 4.7j$, $3.2j$, -42.0 and 0.

3. The real part of the additive inverse of z (as a complex number) equals the additive inverse of the real part of z.

5. If $z = \alpha + j\beta$ and $z = -z$ then $\alpha + j\beta = -\alpha - j\beta$ so $\alpha = -\alpha$ and $\beta = -\beta$, so $\alpha = \beta = 0$, so $z = 0$. That is, zero is the only number that is its own additive inverse.

1.6.8.3 Subtraction

Complex subtraction is simply the addition of the additive inverse of the second argument onto the first:

$$\begin{aligned} z - w &= z + (-w) \\ &= (a + bj) + (-c - dj) \\ &= (a - c) + (bj - dj) \\ &= (a - c) + (b - d)\, j \end{aligned}$$

A geometric interpretation of subtraction is most easily interpreted as the addition of the additive inverse of the second argument, as shown in Figure 9.

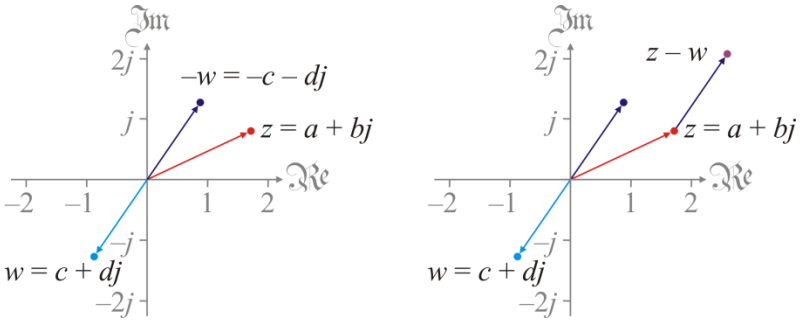


Figure 9. A geometric interpretation of complex subtraction.

Matlab application

As a binary operator, – subtracts the second from the first. Parentheses, however, must be used to completely identify the right-hand operand:

```
>> z - 1+2j
   z =
       2.0000 + 6.0000i

>> z - (1 + 2j)
    ans =
       2.0000 + 2.0000i
```

What is the solution to the following statement (without entering it into MATLAB)?

```
>> -2--4---8----16-----32------64
```

Problems

1. What is (3 + 4*j*) – (7 + 6*j*)?

2. How would you write $z - \Re\mathrm{e}(z)$?

3. How would you write $\Im\mathrm{m}(z)\, j - z$?

4. How would you describe $\Re\mathrm{e}(z-w) = \Re\mathrm{e}(z) - \Re\mathrm{e}(w)$?

5. Given two complex numbers *w* and *z*, does $|w-z| = |z-w|$?

6. Given two complex numbers *w* and *z*, what does $|w-z|$ describe in the complex plane?

Solutions

1. –4 – 2*j*

3. $-\Re\mathrm{e}(z)$ or $\Re\mathrm{e}(-z)$

5. Yes, for if $z = \alpha + \beta j$ and $w = \gamma + \delta j$ then

$$\begin{aligned} |w-z| &= |(\alpha-\gamma) + j(\beta-\delta)| \\ &= \sqrt{(\alpha-\gamma)^2 + (\beta-\delta)^2} \end{aligned}$$

and because $(\alpha-\gamma)^2 = \big((-1)(\gamma-\alpha)\big)^2 = (-1)^2(\gamma-\alpha)^2 = (\gamma-\alpha)^2$, this means that

$$\begin{aligned} |w-z| &= \sqrt{(\gamma-\alpha)^2 + (\delta-\beta)^2} \\ &= |(\gamma-\alpha) + j(\delta-\beta)| \\ &= |z-w| \end{aligned}$$

1.6.8.4 Multiplication

To multiply two complex numbers, first we define what we mean to multiply a complex number z by a real value γ. If $z=\alpha+j\beta$ then $\gamma z=(\gamma\alpha)+j(\gamma\beta)$, so $3.2\cdot(4.5-2.9j)=14.4-9.28j$. This has the effect of stretching the length of the vector by $|\gamma|$ and reflecting it through the origin if $\gamma<0$. In polar coordinates, it is even easier to calculate:

$$\gamma\cdot(r\angle\theta)=\begin{cases}\gamma r\angle\theta & \text{if } \gamma>0\\ 0 & \text{if } \gamma=0\\ |\gamma|r\angle(\theta+\pi) & \text{if } \gamma<0\end{cases}$$

Next, if $z=\alpha+j\beta$ and $w=\gamma+j\delta$, then complex multiplication is defined as

$$\begin{aligned}zw&=(\alpha+j\beta)(\gamma+j\delta)\\&=\alpha\gamma+\alpha(j\delta)+(j\beta)\gamma+(j\beta)(j\delta)\\&=\alpha\gamma+j\alpha\delta+j\beta\gamma+j^2\beta\delta\\&=(\alpha\gamma-\beta\delta)+j(\alpha\delta+\beta\gamma)\end{aligned}$$

You may recognize this from *"foil"*, or *first*, *outside*, *inside* and *last*, as shown in Figure 10.

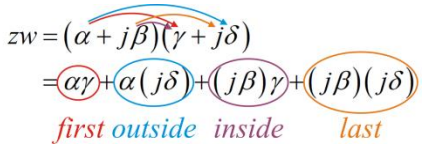


Figure 10. Application of *foil* (*first*, *outside*, *inside*, *last*) to the multiplication of two complex numbers.

The geometric interpretation of complex multiplication is more difficult.

Theorem

If *z* and *w* are complex numbers, then $|zw|=|z||w|$.

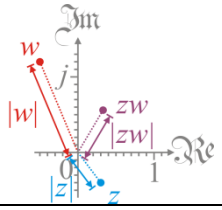
Proof:

$$\begin{aligned}|zw|^2&=(ac-bd)^2+(ad+bc)^2\\&=(ac)^2-2abcd+(bd)^2+(ad)^2+2abcd+(bc)^2\\&=(ac)^2+(bd)^2+(ad)^2+(bc)^2\\&=(a^2+b^2)(c^2+d^2)\\&=|z|^2|w|^2\end{aligned}$$

and therefore $|zw|=\sqrt{|zw|^2}=\sqrt{|z|^2|w|^2}=\sqrt{|z|^2}\sqrt{|w|^2}=|z||w|$. ∎

Example of this theorem

If $z = 0.3 - 0.4j$ and $w = -0.5 + 1.2j$, $|z| = \sqrt{0.3^2 + (-0.4)^2} = 0.5$ and $|w| = \sqrt{(-0.5)^2 + 1.2^2} = 1.3$, so $|z||w| = 0.5 \cdot 1.3 = 0.65$, but at the same time, $zw = (0.3 - 0.4j)(-0.5 + 1.2j) = -0.15 + 0.48 + 0.36j + 0.20j = 0.33 + 0.56j$ and $|zw| = \sqrt{0.33^2 + 0.56^2} = 0.65$.



Theorem

If z and w are complex numbers, then $\arg(zw) = \arg(z) + \arg(w)$.

Proof:

Before we begin, it is important to recall the *double-angle formulas*:

$$\cos(\theta + \phi) = \cos(\theta)\cos(\phi) - \sin(\theta)\sin(\phi)$$
$$\sin(\theta + \phi) = \cos(\theta)\sin(\phi) + \sin(\theta)\cos(\phi)$$

For the proof, it is easiest to use the polar representation. Assume $z = r\angle\theta$ and $w = s\angle\phi$. Then

$$\begin{aligned} zw &= \left(r\cos(\theta) + jr\sin(\theta)\right)\left(s\cos(\phi) + js\sin(\theta)\right) \\ &= rs\cos(\theta)\cos(\phi) + jrs\cos(\theta)\sin(\phi) + jrs\sin(\theta)\cos(\phi) + j^2 rs\sin(\theta)\sin(\phi) \\ &= rs\left(\cos(\theta)\cos(\phi) - \sin(\theta)\sin(\phi) + j\left(\cos(\theta)\sin(\phi) + \sin(\theta)\cos(\phi)\right)\right) \\ &= rs\left(\cos(\theta + \phi) + j\sin(\theta + \phi)\right) \end{aligned}$$

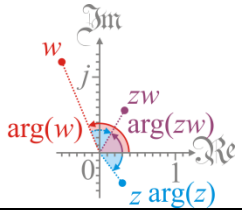
and therefore $\arg(zw) = \theta + \phi$ or $\arg(zw) = \arg(z) + \arg(w)$. ∎

Demonstration of this theorem

If $z = 3 - 4j$ and $w = -5 + 12j$, $\arg(z) \approx -0.9272952179$ and $\arg(w) \approx 1.965587447$, so

$$\arg(z) + \arg(w) \approx 1.038292229,$$

but at the same time, $zw = (3 - 4j)(-5 + 12j) = -15 + 48 + 36j + 20j = 33 + 56j$ and $\arg(wz) \approx 1.038292229$.



Thus, we could also write that $|zw|\angle\arg(zw) = |z||w|\angle\left(\arg(z) + \arg(w)\right)$. Fortunately, if z and w are real, this reduces to simple real-valued multiplication. Note that we may have to adjust $\arg(z) + \arg(w)$ to fall into the interval $(-\pi, \pi]$.

Geometrically, the product is that line extending from the origin that forms an angle equal to the sum of the two angles and the length of the line is the length of the products of the two lines, as shown in Figure 11.

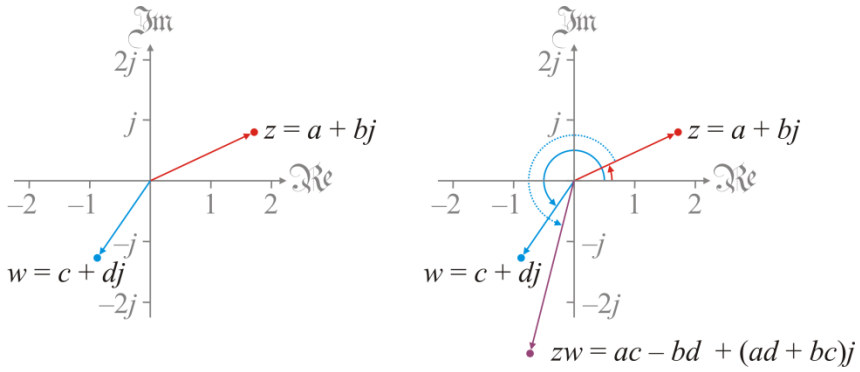


Figure 11. A geometric interpretation of complex multiplication.

Like the integers, rational numbers and the real numbers, complex multiplication distributes over complex addition; that is, if z, w_1 and w_2 are complex numbers than

$$z(w_1 + w_2) = zw_1 + zw_2.$$

The proof of this is left to the reader.

MATLAB application

We can verify these properties in MATLAB:

```
>> format long
>> z = 3.54 - 4.71j;
>> w = -0.29 + 1.54j;
>> z*w
    ans =
      6.226800000000000 + 6.817500000000000i

>> abs( z*w )
    ans =
       9.233165464238144

>> angle( z*w )
    ans =
       0.830651394487295

>> abs( z ) * abs( w )
    ans =
       9.233165464238146

>> angle( z ) + angle( w )  % you may have to add or subtract 2*pi to get it in (-pi, pi]
    ans =
       0.830651394487295
```

You must remember parentheses—multiplication and division come before addition and subtraction

```
>> (3.54 - 4.71j)*(9.02 + 0.425j)
    ans =
     33.932549999999999 -40.979699999999994i

>> 3.54 - 4.71j * 9.02 + 0.425j
    ans =
      3.540000000000000 -42.059199999999997i
```

Problems

1. Multiply $(3 + 2j)(4 – 5j)$, $3.2(2.5 + 6.1j)$ and $(3.2j)(2.5 + 6.1j)$.

2. Multiply $(5.4 – 1.7j)(–j)$.

3. Multiply $(5.3\angle 1.2)(2.0\angle -0.8)$, $(3\angle 1)(2\angle 2)$ and $(3\angle 2)(2\angle 2)$

4. Verify that $(–1 + j)(3 – 3j) = 6j$ by

   a) multiplying the two directly, and
   b) converting each into polar coordinates, multiplying and then converting back to rectangular coordinates.

5. If the real and imaginary parts of two complex number are integers, will the real and imaginary parts of the product also be integers? If the real and imaginary parts of two complex numbers are rational numbers, will the real and imaginary parts of the product also be rational numbers?

Solutions

1. $22 – 7j$, $8 + 19.52j$ and $–19.52 + 8j$

3. $10.6\angle 0.4$, $6\angle 3$ and $6\angle(4-2\pi)$ because $4 > \pi$ but $4-2\pi \approx -2.28$ and thus $-\pi < 4-2\pi \le \pi$.

5. In both cases, yes.

1.6.8.5 Complex conjugates

First, a review of "conjugates" from secondary school: Given an expression such as $\frac{a+b\sqrt{c}}{d+e\sqrt{f}}$ where *a*, *b*, *d* and *e* are integers and *c* and *f* are positive integers, you learned that you could multiply both the numerator and denominator by the *conjugate* of the denominator; that is, $d-e\sqrt{f}$ (your teacher may not have used the term, but that is the conventional description). You would then calculate

$$\frac{a+b\sqrt{c}}{d+e\sqrt{f}}\frac{d-e\sqrt{f}}{d-e\sqrt{f}}=\frac{ad-ae\sqrt{f}+eb\sqrt{c}-be\sqrt{cf}}{d^2-e^2f}=\frac{ad}{d^2-e^2f}+\frac{e}{d^2-e^2f}\left(a\sqrt{f}+b\left(\sqrt{c}-\sqrt{cf}\right)\right)$$

to remove the square root from the denominator. For example, to rationalize the denominator of $\frac{5-2\sqrt{3}}{12+4\sqrt{2}}$, we calculate

$$\frac{5-2\sqrt{3}}{12+4\sqrt{2}}\frac{12-4\sqrt{2}}{12-4\sqrt{2}}=\frac{60-20\sqrt{2}-24\sqrt{3}+8\sqrt{6}}{112}=\frac{15}{28}-\frac{5}{28}\sqrt{2}-\frac{3}{14}\sqrt{3}+\frac{1}{14}\sqrt{6}\,.$$

Recall that $j=\sqrt{-1}$, so the conjugate of $\alpha+j\beta=\alpha+\beta\sqrt{-1}$ would similarly be $\alpha-\beta\sqrt{-1}=\alpha-j\beta$.

Next, let us return to the roots of a quadratic: if $\alpha + j\beta$ is a root of a quadratic polynomial with real coefficients, the other root is of the form $\alpha - j\beta$. This is because if the roots of $ax^2 + bx + c$ are

$$\frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

and $b^2-4ac<0$, then $\sqrt{b^2-4ac}=\sqrt{(-1)\left(4ac-b^2\right)}=\sqrt{-1}\sqrt{4ac-b^2}=j\sqrt{4ac-b^2}$, so the roots are

$$-\frac{b}{2a}+j\frac{\sqrt{4ac-b^2}}{2a}\text{ and }-\frac{b}{2a}-j\frac{\sqrt{4ac-b^2}}{2a}\,.$$

Because we almost universally need to refer to such complex numbers in pairs, given the complex number $z = a + bj$, we will define its *complex conjugate* to be $z^* = a - bj$. A geometric interpretation of the complex conjugate is a reflection in the real axis, as shown in Figure 12.

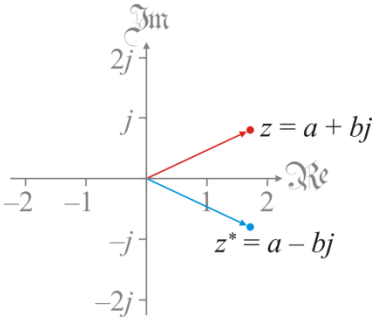


Figure 12. A geometric interpretation of the complex conjugate.

In polar form, the complex conjugate of $z = r\angle\theta = re^{j\theta}$ is $z^* = r\angle -\theta = re^{-j\theta}$. Note that

$$|z|^2 = zz^*$$

as

$$\begin{aligned} zz^* &= (a+bj)(a-bj) \\ &= a^2 + abj - abj - (bj)^2 \\ &= a^2 - b^2 j^2 \\ &= a^2 + b^2 \\ &= |z|^2 \end{aligned}$$

Thus, we may deduce that, as $zz^* = |z|^2$, it follows that $|zz^*| = |z||z^*| = |z|^2$ and therefore $|z^*| = |z|$. This is an automatic consequence of the polar form. Also from the polar representation that

$$(r\angle\theta)(r\angle -\theta) = r^2\angle(\theta - \theta) = r^2\angle 0,$$

which also equals $|z|^2$.

Theorem

If z is a complex number, then $z + z^* = 2\Re\mathrm{e}(z)$ and $z - z^* = 2j\Im\mathrm{m}(z)$.

Proof:

If $z = \alpha + j\beta$ then $z^* = \alpha - j\beta$ and therefore

$$z + z^* = (\alpha + j\beta) + (\alpha - j\beta) = 2\alpha = 2\Re\mathrm{e}(z)$$

and

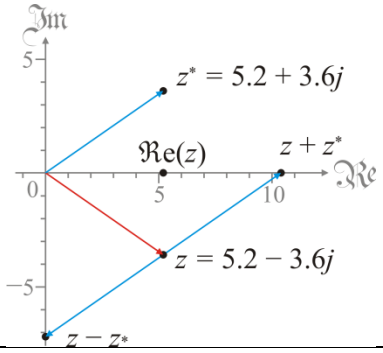
$$z - z^* = (\alpha + j\beta) - (\alpha - j\beta) = 2j\beta = 2j\Im\mathrm{m}(z). \blacksquare$$

Demonstration of this theorem

If $z = 5.2 - 3.6j$, then $\Re\mathrm{e}(z) = 5.2$, $\Im\mathrm{m}(z) = -3.6$ and $z^* = 5.2 + 3.6j$.

Thus, $z + z^* = 5.2 - 3.6j + 5.2 + 3.6j = 10.4$ and equals $2\Re\mathrm{e}(z) = 10.4$.

Similarly, $z - z^* = 5.2 - 3.6j - (5.2 + 3.6j) = -7.2j$ and equals $2j\Im\mathrm{m}(z) = -7.2j$.



MATLAB application

To calculate the complex conjugate in MATLAB, you must use the `conj` routine:

```
>> format long
>> z = 3.54 - 4.71j;
>> conj( z )
    ans =
      3.540000000000000 + 4.710000000000000i

>> angle( z )
    ans =
      -0.926276888414713

>> angle( -z )
    ans =
       2.215315765175080

>> angle( conj( z ) )
    ans =
       0.926276888414713

>> abs( z )^2
    ans =
      34.715699999999998

>> z*conj( z )
    ans =
      34.715699999999998
```

You may wonder why it is that the answer is 34.715699999999998 and not 34.7157. This has to do with numbers being stored in base 2 (or *binary*) instead of base 10 (what you know as *decimal*) and there is no finite binary representation of 34.7157. As a simpler example, the representation of 0.3 in binary is $0.01001\overline{10011}_2$, where the subscripted '2' indicates that that number is base 2.

Problems

1. What are $(3 - 2j)^*$, 5^*, $(-4)^*$, j^*, $(-3j)^*$, $(1.7 + 3.1j)^*$ and $(-9.8 - 7.6j)^*$?

2. Is it true that $(w+z)^* = w^* + z^*$? That is, is the complex conjugate of a sum equal to the sum of the complex conjugates?

3. Is it true that $(wz)^* = w^* z^*$? That is, is the complex conjugate of a product equal to the product of the complex conjugates?

4. If you are given a complex number z and $z = z^*$, what does this say about z?

5. If you are given a complex number w and $w = -w^*$, what does this say about w?

6. Is $(-z)^* = -(z^*)$? That is, is the complex conjugate of the additive inverse equal to the additive inverse of the complex conjugate of that complex number?

7. What are $(3\angle 2)^*$, $(1.5\angle -0.54)^*$ and $(6.3\angle 45°)^*$?

Solutions

1. $3 + 2j$, 5, -4, $-j$, $3j$, $1.7 - 3.1j$ and $-9.8 + 7.6j$

3. Yes, for if $z = \alpha + \beta j$ and $w = \gamma + \delta j$ then $zw = (\alpha\gamma - \beta\delta) + j(\alpha\delta - \beta\gamma)$ so

$$(zw)^* = (\alpha\gamma - \beta\delta) - j(\alpha\delta + \beta\gamma) = (\alpha\gamma - \beta\delta) + j(-\alpha\delta - \beta\gamma)$$

and $z^* w^* = (\alpha - j\beta)(\gamma - j\delta) = (\alpha\gamma - \beta\delta) + j(-\alpha\delta - \beta\gamma)$, and this equals $(zw)^*$.

5. If $w = \gamma + \delta j$ and $w = w^*$ then $\gamma + \delta j = -(\gamma - \delta j) = -\gamma + \delta j$ so $\gamma = -\gamma$, and therefore $\gamma = 0$, so w must be purely imaginary.

7. $3\angle -2$, $1.5\angle 0.54$ and $6.3\angle -45°$

1.6.8.6 The multiplicative inverse

The *multiplicative inverse* (or *reciprocal*) of a complex number z is that number z^{-1} such that $zz^{-1}=1$. To find the multiplicative inverse of a complex number, it is easier to use the polar coordinate representation, in which case, our requirement may be restated as requiring that

$$\left|zz^{-1}\right| \measuredangle \arg\left(zz^{-1}\right) = |z|\left|z^{-1}\right| \measuredangle \left(\arg(z)+\arg\left(z^{-1}\right)\right) = 1 \measuredangle 0,$$

from which it follows that

$$|z|\left|z^{-1}\right| = 1 \text{ so } \left|z^{-1}\right| = \frac{1}{|z|}$$

and

$$\arg(z)+\arg\left(z^{-1}\right) = 0 \text{ so } \arg\left(z^{-1}\right) = -\arg(z).$$

This can be interpreted geometrically as shown in Figure 13.

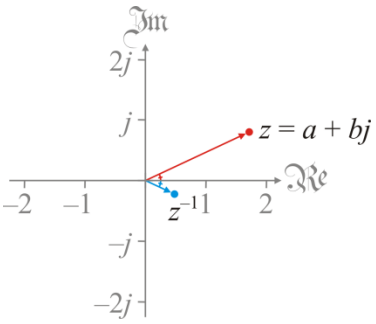


Figure 13. A geometric interpretation of the multiplicative inverse.

Let us, however, return to the rectangular representation and demonstrate that we get the same result: we require that $ac-bd=1$ and $ad+bc=0$. This is a system of two linear equations in two unknowns (c and d), and therefore we may solve this. We begin with

$$\begin{aligned} ac-bd &= 1 \\ ad+bc &= 0 \end{aligned}$$

we rewrite the second to get

$$\begin{aligned} ac-bd &= 1 \\ bc+ad &= 0 \end{aligned}$$

multiply the first equation by b and the second by a to get

$$\begin{aligned} abc-b^2d &= b \\ abc+a^2d &= 0 \end{aligned}$$

We may now subtract the two to get

$$b^2 d - a^2 d = b$$

and solving for *d* we get

$$d = -\frac{b}{a^2 + b^2}.$$

By substituting this into the second equation, we get

$$bc + a\left(-\frac{b}{a^2 + b^2}\right) = 0$$

$$bc - \frac{ab}{a^2 + b^2} = 0$$

$$c = \frac{a}{a^2 + b^2}$$

Thus, the inverse is

$$\frac{a}{a^2 + b^2} - \frac{b}{a^2 + b^2} j.$$

We may also note that

$$(a + bj)(a - bj) = a^2 + b^2$$

and therefore

$$(a + bj)\frac{a - bj}{a^2 + b^2} = 1,$$

and thus

$$z^{-1} = \frac{a - bj}{a^2 + b^2},$$

and that

$$z^{-1} = \frac{z^*}{|z|^2},$$

which is, as we shall see shortly, a very useful result.

It is left as an exercise to the reader to demonstrate that $\left|\frac{z^*}{|z|^2}\right| = \frac{1}{|z|}$ and $\arg\left(\frac{z^*}{|z|^2}\right) = -\arg(z)$.

As an alternate approach, let us recall in secondary school, you could rationalize a denominator. For example, to rationalize $\frac{1}{3 + 2\sqrt{3}}$ you multiplied this by $1 = \frac{3 - 2\sqrt{3}}{3 - 2\sqrt{3}}$ to get

$$\frac{1}{3+2\sqrt{3}}\cdot\frac{3-2\sqrt{3}}{3-2\sqrt{3}}=\frac{3-2\sqrt{3}}{3^2-2^2\cdot 3}=\frac{3-2\sqrt{3}}{9-12}=\frac{-3+2\sqrt{3}}{3}=-1+\frac{2}{3}\sqrt{3}\,.$$

Recall that $j=\sqrt{-1}$, so in order to *rationalize* the denominator of, for example, $\frac{1}{-3-2j}$, let's do the same thing:

$$\frac{1}{-3-2j}=\frac{1}{-3-2j}\cdot\frac{-3+2j}{-3+2j}=\frac{-3+2j}{(-3)^2-2^2j^2}=\frac{-3+2j}{9+4}=\frac{-3+2j}{13}=-\frac{3}{13}+\frac{2}{13}j\,.$$

Therefore, the multiplicative inverse of $-3-2j$ is $-\frac{3}{13}+\frac{2}{13}j$, and we see that this is true, as

$$(-3-2j)\left(-\frac{3}{13}+\frac{2}{13}j\right)=\frac{9}{13}-\frac{\cancel{6}}{\cancel{13}}j+\frac{\cancel{6}}{\cancel{13}}j-\frac{4}{13}j^2=\frac{9}{13}+\frac{4}{13}=1.$$

MATLAB application

Finding the inverse is easy to do in MATLAB:

```
>> format long
>> z = 3.54 - 4.71j;
>> 1/z
    ans =  0.101971154261617 + 0.135673484907405i

>> z^-1
    ans =  0.101971154261617 + 0.135673484907405i

>> conj( z )/abs( z )^2
    ans =  0.101971154261617 + 0.135673484907405i

>> z*ans
    ans =  1

>> 1/abs( z )          % this and the next value should be the same
    ans =  0.169721568483108

>> abs( z^-1 )
    ans =  0.169721568483108

>> angle( z )          % the next value should be the negative of this one
    ans = -0.926276888414713

>> angle( z^-1 )
    ans =  0.926276888414713
```

MATLAB cannot, in most cases, calculate the exact inverse, but it is always close

```
>> format long
>> z = 2.3 + 0.1j;
>> one = z*(1/z);
>> real( one - 1 )  // Ideally, this should be 0, but 0.00000000000000011 is close
    ans = -1.110223024625157e-16
>> imag( one )      // Ideally, this should also be 0, but 0.000000000000000069 is close
    ans = -6.938893903907228e-18
```

Problems

1. What are the multiplicative inverses of $1 + j$, $-3 - 4j$, 7, $-3j$ and $2 - 3j$?

2. What are the multiplicative inverses of $2\angle 2$, $0.2\angle -0.3$ and $2.5\angle -160°$?

3. Use the polar representation to argue that $\left(z^{-1}\right)^{*} = \left(z^{*}\right)^{-1}$.

4. Explicitly multiply $-6 - 7j$ and $-\frac{6}{85} + \frac{7}{85} j$ to see that their product equals 1.

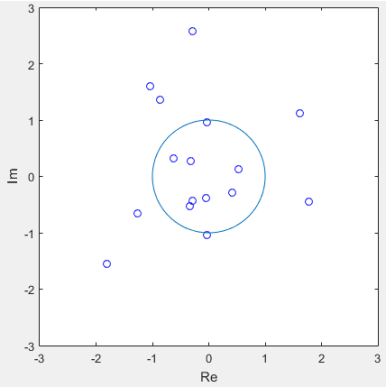
5. Explicitly multiply $2\angle 2$ and $\frac{1}{2}\angle -2$ to see that their product equals $1\angle 0 = 1$.

6. Is it true that $\Re\mathrm{e}\left(z^{-1}\right) = \frac{1}{\Re\mathrm{e}\left(z\right)}$? Hint, what if z is purely imaginary?

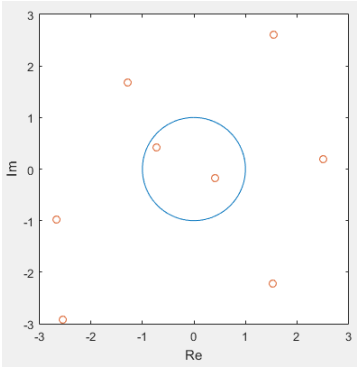
7. Show that if $|z| > 1$ then $0 < \left|z^{-1}\right| < 1$, and if $0 < |z| < 1$ then $\left|z^{-1}\right| > 1$ using whatever representation you wish.

8. Find the multiplicative inverse of $z = -\frac{1}{4} - \frac{1}{3} j$ and find $|z|$ and $\left|z^{-1}\right|$.

9. This plot shows eight complex numbers together with their inverses. Identify each pair of complex numbers that are each others inverse.



10. With a pen, estimate where the multiplicative inverse of each of these complex numbers is.



11. Show that $\left(z^{-1}\right)^{-1} = z$.

Solutions

1. $\frac{1}{2} - \frac{1}{2}j$, $-\frac{3}{25} + \frac{4}{25}j$, $\frac{1}{7}$, $\frac{1}{3}j$ and $\frac{2}{13} - \frac{3}{13}j$.

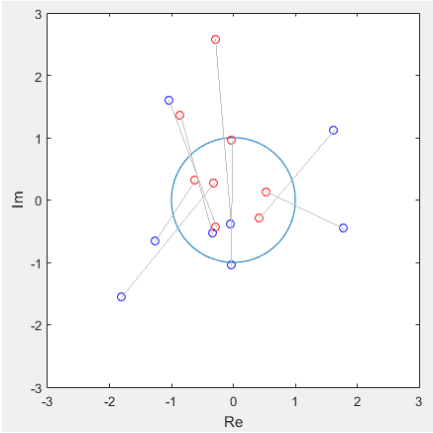
3. Given $r\angle\theta$, $\left(r\angle\theta\right)^{-1} = \frac{1}{r}\angle -\theta$ and so $\left(\left(r\angle\theta\right)^{-1}\right)^{*} = \frac{1}{r}\angle\theta$. Similarly, $\left(r\angle\theta\right)^{*} = r\angle -\theta$ so

$$\left(\left(r\angle\theta\right)^{*}\right)^{-1} = \frac{1}{r}\angle -\left(-\theta\right) = \frac{1}{r}\angle\theta .$$

5. $\left(2\angle 2\right)\left(\frac{1}{2}\angle -2\right) = \left(2\cdot\frac{1}{2}\right)\angle\left(2-2\right) = 1\angle 0$.

7. The easiest is the polar representation, where $|r\angle\theta| = r$, so $\left|\left(r\angle\theta\right)^{-1}\right| = \left|\frac{1}{r}\angle -\theta\right| = \frac{1}{r}$. Alternatively, if $z = \alpha + j\beta$, then $|z| = \sqrt{\alpha^2 + \beta^2}$ and $\left|z^{-1}\right| = \left|\frac{z^*}{zz^*}\right| = \frac{\left|z^*\right|}{\left|zz^*\right|} = \frac{|z|}{|z|^2} = \frac{1}{|z|} = \frac{1}{\sqrt{\alpha^2 + \beta^2}}$, so again, the result follows.

9. The pairs of complex numbers that are multiplicative inverses of each other are shown here with a connected line.



11. If $z = r\angle\theta$, then $z^{-1} = \frac{1}{r}\measuredangle(-\theta)$ so $\left(z^{-1}\right)^{-1} = \frac{1}{\frac{1}{r}}\measuredangle -(-\theta) = r\measuredangle\theta = z$.

1.6.8.7 Division

Complex division of z/w can be now be reduced to calculating $zw^{-1} = \frac{zw^*}{ww^*} = \frac{zw^*}{|w|^2}$. As may be expected,

$$\left|zw^{-1}\right| = \left|\frac{zw^*}{|w|^2}\right| = \frac{\left|zw^*\right|}{|w|^2} = \frac{|z|}{|w|}.$$

Similarly, in our polar representation, $\frac{r\angle\theta}{s\angle\phi} = (r\angle\theta)\left(\frac{1}{s}\angle-\phi\right) = \frac{r}{s}\angle(\theta-\phi)$.

Problems

1. Calculate $\frac{4+2j}{-2-j}$, $\frac{4+2j}{-2+j}$ and having calculated the second, find $\frac{4+2j}{2-j}$.

2. Calculate $\frac{4+3j}{3-4j}$, $\frac{4+3j}{3+4j}$ and having calculated the second, find $\frac{4+3j}{-4+3j}$.

3. Should all ratios in Question 1 have the same magnitude? Why?

Solutions

1. There are two ways to do the first: $\frac{4+2j}{-2-j}\cdot\frac{-2+j}{-2+j} = \frac{-10}{5} = -2$ or realize that $\frac{4+2j}{-2-j} = \frac{2\cancel{(2+j)}}{-\cancel{(2+j)}} = \frac{2}{-1} = -2$. The second requires more $\frac{4+2j}{-2+j}\cdot\frac{-2-j}{-2-j} = \frac{6-8j}{5} = \frac{6j}{5} - \frac{8j}{5} = 1.2-1.6j$. The third can be calculated by realizing that $\frac{4+2j}{2-j} = \frac{4+2j}{-(-2+j)} = -\frac{4+2j}{-2+j} = -1.2+1.6j$.

3. Yes, they should have all the same magnitude, for $|w| = |w^*|$, and $|-w| = |w|$ as does $|jw| = |w|$.

1.6.8.8 Integer exponentiation

Given an integer n, $z^n = z \cdot z^{n-1}$ and therefore $z = z^1 = z \cdot z^0$ so $z^0 = 1$. If z is not zero, this is even true for negative integers. While integer exponentiation of the rectangular representation is difficult to calculate, it is trivial to calculate in the polar representation: $(r \angle \theta)^n = r^n \angle (n\theta)$. For example, $1 + j\sqrt{3} = 2 \angle \frac{\pi}{6}$, so $\left(1 + j\sqrt{3}\right)^{10} = 2^{10} \angle 10 \cdot \frac{\pi}{6} = 1024 \angle -\frac{\pi}{3}$. From this, we may also conclude that $\left|z^n\right| = |z|^n$.

Now, if $|z| > 1$, it follows that the magnitude of z^n will grow exponentially large, while if $0 < |z| < 1$ then z^n will converge towards zero. For example, calculating powers of 0.1 – 0.3*j*, we get the shrinking sequence

$(0.1 - 0.3j)^1 = 0.1 - 0.3j$
$(0.1 - 0.3j)^2 = -0.08 - 0.06j$
$(0.1 - 0.3j)^3 = -0.026 + 0.018j$
$(0.1 - 0.3j)^4 = 0.0028 + 0.0096j$
$(0.1 - 0.3j)^5 = 0.00316 + 0.00012j$
$(0.1 - 0.3j)^6 = 0.000352 - 0.000936j$
$(0.1 - 0.3j)^7 = -0.0002456 - 0.0001992j$
$(0.1 - 0.3j)^8 = -0.00008432 + 0.00005376j$
$(0.1 - 0.3j)^9 = 0.000007696 + 0.000030672j$
$(0.1 - 0.3j)^{10} = 0.0000099712 + 0.0000007584j$

Problems

1. Calculate $(1+j)^8$ in two ways: first using the rectangular representation, and second using the polar representation.

2. If r is a real number and $r > 1$, then r^n grows monotonically towards infinity. What happens if you have a complex number z where $|z| > 1$ and $\arg(z) \approx 0$, and you calculate z^n for progressively larger n?

3. If r is a real number and $r < -1$, then r^n grows alternates between being positive and negative, but increases in magnitude towards infinity. What happens if you have a complex number z where $|z| > 1$ and $\arg(z) \approx \pi$, and you calculate z^n for progressively larger n?

4. What happens if you have a complex number z where $|z| = 1$ and you calculate z^n for progressively larger n?

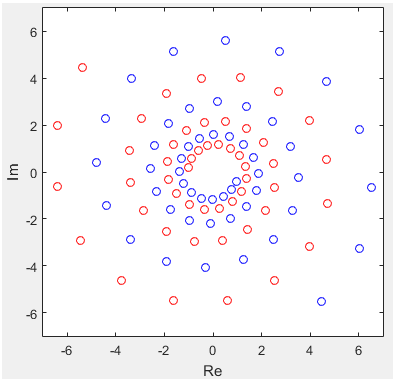
Solutions

1. $(1+j)^8 = \left((1+j)^4\right)^2 = \left(\left((1+j)^2\right)^2\right)^2$, so $(1+j)^2 = 2j$, $(2j)^2 = -4$ and $(-4)^2 = 16$. If you wanted to be more direct, you could just calculate

$$(1+j)^2 = 2j,\ 2j(1+j) = -2+2j,\ (-2+2j)(1+j) = -4,\ -4(1+j) = -4-4j,$$
$$(-4-4j)(1+j) = -8j,\ -8j(1+j) = 8-8j \text{ and } (8-8j)(1+j) = 16$$

Using the other approach, we have $1+j = \sqrt{2}\angle 45°$ so $(1+j)^8 = \left(\sqrt{2}\right)^8 \angle(8 \cdot 45°) = 16\angle 180° = 16\angle 0°$.

3. It will alternate outward in a spiral pattern. For example, if we plot $(-1 + 0.2j)^n$, then the powers alternate between the red dots and the blue dots in this plot, always increasing in magnitude.



1.6.9 Complex numbers are a field

Note that, like the real numbers, we have:

Closure under addition and multiplication	$w+z \in \mathbf{C}$	$wz \in \mathbf{C}$
Commutativity for both addition and multiplication	$w + z = z + w$	$wz = zw$
Associativity for addition and multiplication	$(x + y) + z = x + (y + z)$	$(xy)z = x(yz)$
The existence of an identity element	$0 + 0j = 0$	$1 + 0j = 1$

In addition,

1. every complex number has an additive inverse, and every complex number except for the additive identity (0) has a multiplicative inverse, and
2. multiplication distributes over addition: $x(y+z) = xy + xz$.

Note: these are all the same properties that you have come to expect from real numbers. In fact, they are not only the properties of the real numbers, but also of the rational numbers. These are not, however, the properties of the integers: only 1 and –1 have multiplicative inverses within the integers.

We can divide complex numbers into nine overlapping categories:

Category	Definition	Property	MATLAB test
The real line	$z = a + 0j$	$z = z^*$	`imag( z ) == 0`
The imaginary line	$z = 0 + \beta j$	$z = -z^*$	`real( z ) == 0`
The unit circle	$\lvert z\rvert = 1$ or $z = \cos(\theta) + j\sin(\theta)$	$z^{-1} = z^*$, so $zz^* = 1$	`abs( z ) == 1`
The unit disk	$\lvert z\rvert \le 1$	$0 \le zz^* \le 1$	`abs( z ) <= 1`
The open unit disk	$\lvert z\rvert < 1$	$0 \le zz^* < 1$	`abs( z ) < 1`
Open left-hand plane	$z = \alpha + \beta j$ with $\alpha < 0$	$\Re e(z) < 0$	`real( z ) < 0`
Closed left-hand plane	$z = \alpha + \beta j$ with $\alpha \le 0$	$\Re e(z) \le 0$	`real( z ) <= 0`
Closed right-hand plane	$z = \alpha + \beta j$ with $\alpha \ge 0$	$\Re e(z) \ge 0$	`real( z ) >= 0`
Open right-hand plane	$z = \alpha + \beta j$ with $\alpha > 0$	$\Re e(z) > 0$	`real( z ) > 0`

These are shown in Figure 14.

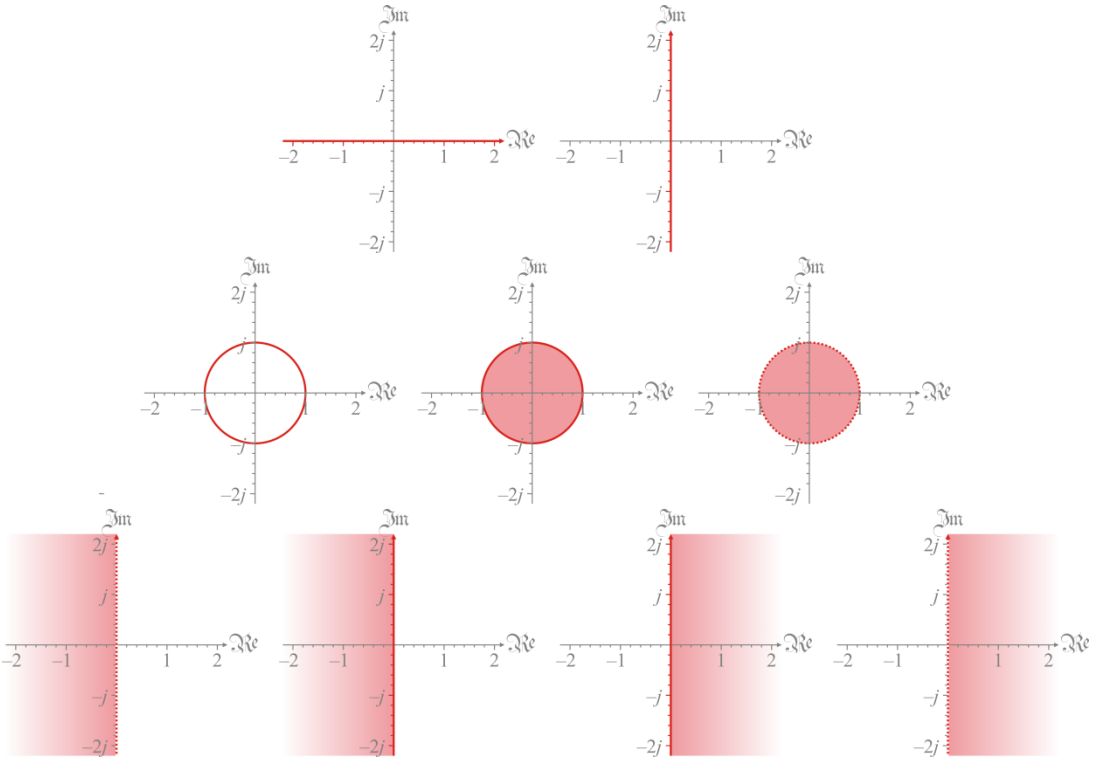


Figure 14. The real and imaginary axes; the unit circle, unit disk and open unit disk;
and the open left-, close left-, closed right- and open right-hand planes.

These become relevant as we can say that if z is in the open left-hand plane, then $|e^z| < 1$; z is on the unit circle if and only if z^{-1} is also on the unit circle; and $z \neq 0$ is in the open unit disk if and only if z^{-1} is outside the unit circle. This will become more relevant when you discuss

MATLAB application

In MATLAB, it uses the double-precision floating-point representation specified in the IEEE-754 standard. Thus, not all the properties of complex numbers hold in MATLAB; for example:

1. Numbers greater than or equal to 2^{1024} cannot be represented, and therefore are represented by a value of `Inf`:

```
>> x = 1e308;              % x = 10^308
>> x + x
    ans =  Inf
```

2. Numbers smaller than 2^{-1074} cannot be represented and are therefore replaced with 0:

```
>> x = 2^-1074
    x =  4.940656458412465e-324

>> x/2
    ans =  0
```

however, the standard guarantees that if $x \neq y$, then $x - y \neq 0$.

3. Addition is no longer necessarily associative:

```
>> (-7.701508841452665e-12 + 3.141592653589793) - 3.141592653595635
    ans = -1.354338863279736e-011

>> -7.701508841452665e-12 + (3.141592653589793 - 3.141592653595635)
    ans = -1.354350239703024e-011
```

When you do numerical methods for approximating solutions to various equations, you will have to take steps to avoid situations that lead to such undesirable computations.

1.6.10 Three inequalities

We will describe three inequalities:

1. two relative inequalities, and
2. the triangle inequality.

These are covered in the next sections.

1.6.10.1 Two relative inequalities

The real and imaginary components of a complex number are always less than the absolute value of the complex number. For a complex number $z = \alpha + j\beta$, $|z|^2 = \alpha^2 + \beta^2 \geq \alpha^2 = \Re\mathrm{e}(z)^2$, and therefore

$$|z| \geq |\Re\mathrm{e}(z)| \geq \Re\mathrm{e}(z).$$

Similarly, we may deduce that

$$|z| \geq |\Im\mathrm{m}(z)| \geq \Im\mathrm{m}(z).$$

Problems

1. Demonstrate these two inequalities when $z = 3 + 4j$.

2. Demonstrate these two inequalities when $z = -7 + 2j$.

3. Under what conditions is $|z| = |\Re\mathrm{e}(z)| = \Re\mathrm{e}(z)$?

4. Find an example when $|z| > |\Re\mathrm{e}(z)| > \Re\mathrm{e}(z)$ and another example when $|z| > |\Im\mathrm{m}(z)| > \Im\mathrm{m}(z)$.

Solutions

1. $|z| = 5 > |\Re\mathrm{e}(z)| = 3 = \Re\mathrm{e}(z)$.

3. If $|z| = |\Re\mathrm{e}(z)|$ then $\sqrt{\alpha^2 + \beta^2} = |\alpha| = \sqrt{\alpha^2}$, so $\beta = 0$, so thus the number must be real. For $|\alpha| = \alpha$, it must be that $\alpha \geq 0$, so this inequality is equal if and only if *z* is a non-negative real number.

1.6.10.2 The triangle inequality

The triangle inequality says that any one side of a triangle must always be less than or equal to the sum of the lengths of the two other sides, and equality holds only when the triangle is *degenerate*. For example, in Figure 15, we see that for triangle ABC,

$$AB < AC + BC,\ AC < AB + BC \text{ and } BC < AB + AC,$$

with the same holding true for triangle DEF and for the degenerate triangle, GI = GH + HI.

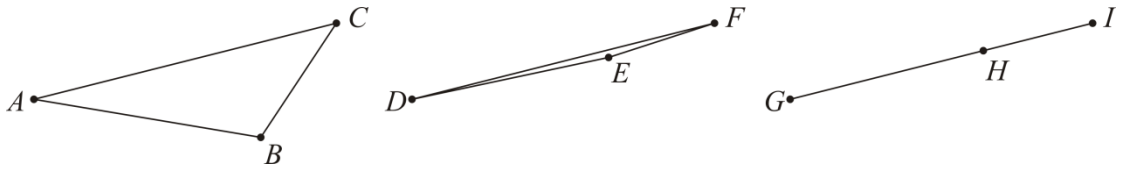


Figure 15. Three triangles, the third of which is degenerate.

With complex addition, you can think of $w + z$ as one side of a triangle, as shown in Figure 16.

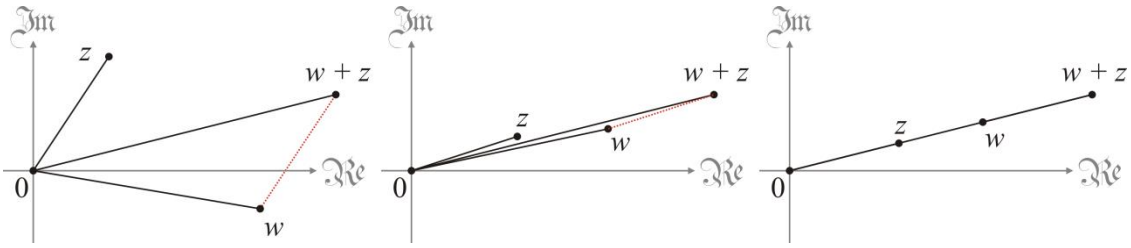


Figure 16. Three sums of complex numbers, where in the third, one is a real multiple of the other.

Consequently, we have the following theorem:

Theorem

Given two complex numbers w and z, then

$$|w+z| \le |w| + |z| \ .$$

While this is obvious geometrically speaking, we can also prove this analytically.

Proof:

$$
\begin{aligned}
|w+z|^2 &= (w+z)(w+z)^* \\
&= (w+z)(w^*+z^*) \\
&= ww^* + wz^* + zw^* + zz^* \\
&= ww^* + wz^* + \left((zw^*)^*\right)^* + zz^* \quad \text{because } z = (z^*)^* \\
&= ww^* + wz^* + (z^*w)^* + zz^* \quad \text{because } (zw)^* = w^*z^* \text{ or, in this case, } (zw^*)^* = z^*(w^*)^* = z^*w \\
&= ww^* + wz^* + (wz^*)^* + zz^* \quad \text{because } wz = zw \text{ or, in this case, } z^*w = wz^* \\
&= |w|^2 + 2\Re\mathrm{e}(wz^*) + |z|^2 \quad \text{because } z + z^* = 2\Re\mathrm{e}(z) \\
&\le |w|^2 + 2|wz^*| + |z|^2 \quad \leftarrow \text{note the inequality} \\
&= |w|^2 + 2|w||z^*| + |z|^2 \quad \text{because } |z| = |z^*| \\
&= |w|^2 + 2|w||z| + |z|^2 \quad \text{but } a^2 + 2ab + b^2 = (a+b)^2 \\
&= (|w| + |z|)^2
\end{aligned}
$$

and therefore, as both objects being squared are positive, $|w+z| \le |w| + |z|$. ∎

Further examples of the triangle inequality can be seen visually in Figure 17.

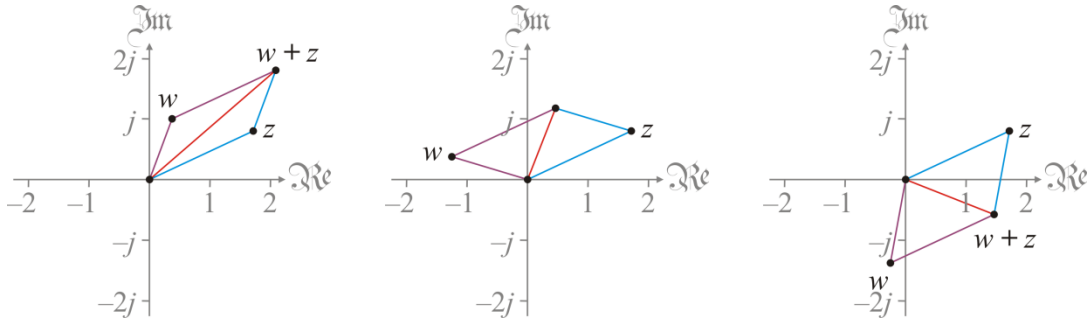


Figure 17. A graphical representation of the triangle inequality.

MATLAB application

We can use MATLAB to demonstrate this the `abs` and `angle` functions map onto each entry of an array:

```
>> w = -8.106 + 7.449j;
>> z =  4.572 - 9.531j;
>> abs( w + z )
    ans =
       4.101692333659364
>> abs( w ) + abs( z )
    ans =
      21.579715685600995
>> z =  -4.571 + 4.083j;    % Now 'z' is pointing in the same direction as 'w'
>> abs( w + z )
    ans =
      17.137483858489844
>> abs( w ) + abs( z )

ans =

  17.137876241409174
```

Problems

1. Show the triangle inequality is true with $z = 2 + j$ and $w = 4 - 2j$.

2. Show that the triangle inequality is actually an equality when $z = 3 - 5j$ and $w = 9 - 15j$.

3. Based on your calculations in Question 2, what conditions are required for $|w+z| = |w| + |z|$?

Solutions

1. $|w+z| = |6-j| = \sqrt{37} < \sqrt{5} + \sqrt{20} = |w| + |z|$ for $37 < \left(\sqrt{5} + \sqrt{20}\right)^2 = 5 + 2\sqrt{5}\sqrt{20} + 2 = 45$.

3. The triangle inequality holds as equal if and only if w is a positive scalar multiple of z; that is, $w = \gamma z$ where $\gamma \geq 0$.

1.6.11 The fundamental theorem of algebra

The *fundamental theorem of algebra* states that:

Theorem

Every polynomial of degree n has n complex roots (some of which may be real, too) when we count multiplicity of roots.

The term *multiplicity* indicates how many times a root is multiplied into the polynomial. The following are equivalent:

1. A polynomial p has a root r of multiplicity m.
2. $p(r) = 0$ and $\left.\frac{d^k}{dz^k} p(z)\right|_{z=r} = 0$ for $k = 1, 2, \ldots, m-1$ but $\left.\frac{d^m}{dz^m} p(z)\right|_{z=r} \neq 0$, or, in English, the polynomial and the first $m - 1$ derivatives of that polynomial when evaluated at the root all equal zero, but the mth derivative evaluated at that root is non-zero.
3. A polynomial p may be written as $(z-r)^m q(z)$ where q is a polynomial and $q(r) \neq 0$.

We will not prove this result, but it is similar to the prime factorization theorem that says that each integer can be written as a product of prime numbers. As an example, consider

$$z^7 - 2z^6 + 2z^5 - 24z^4 + 41z^3 - 22z^2 + 140z - 200$$

which has the seven roots 2, 2, 2, $-1 - 2j$, $-1 - 2j$, $-1 + 2j$, and $-1 + 2j$, it can be written both as

$$(z-2)^3 (z+1-2j)^2 (z+1+2j)^2$$

and

1. the derivative $7z^6 - 12z^5 + 10z^4 - 96z^3 + 123z^2 - 44z + 140$ evaluated at 2, $-1 - 2j$ and $-1 + 2j$ are all zero,
2. the second derivative $42z^5 - 60z^4 + 40z^3 - 288z^2 + 246z - 44$ evaluated at 2 is zero, but evaluated at $-1 \pm 2j$ evaluates to $-288 \mp 1472j$, respectively, and
3. the third derivative $210z^4 - 240z^3 + 120z^2 - 576z + 246$ evaluated at 2 is 1014.

We will now look at the roots of the very simple polynomial $z^n = 1$.

Sketch of the proof—significantly beyond the scope of this course and not even required reading
From the study of infinitely differentiable functions, a result from the Little Picard (Émile, not Jean-Luc) Theorem is that a non-constant polynomial must take on every possible complex number for some argument. Therefore, every polynomial $p(z)$ of degree n is not constant, must have some point z_1 such that $p(z_1) = 0$. This point is a root of the polynomial, and we may now use polynomial division to write the polynomial $p(z) = (z - z_1)p_1(z)$ where $p_1(z)$ is a polynomial of degree $n - 1$. We may apply this a total of n times until finally we have n roots.

As a simple example, consider the polynomial

$$(0.7-1.3j)z^3+(3.93-1.07j)z^2+(2.526+4.714j)z+(-1.5896-3.4384j).$$

From the Little Picard Theorem, this polynomial must have a root, and searching for one, we have one such root is $0.2 - 1.8j$. We may now apply polynomial division to find that this polynomial equals

$$(z-0.2+1.8j)\left((0.7-1.3j)z^2+(1.73-2.59j)z+(-1.79+1.082j)\right).$$

We may now reapply the theorem to find that the polynomial $(0.7-1.3j)z^2+(1.73-2.59j)z+(-1.79+1.082j)$ has a root, until we ultimately determine that the polynomial may be written as

$$(0.7-1.3j)(z-0.2+1.8j)(z-0.5-0.2j)(z+2.6+0.4j).$$

Note that the factor out front is the coefficient of z^3 and the three roots are $0.2 - 1.8j$, $0.5 + 0.2j$ and $-2.6 - 0.4j$.

If you are not sure how to do polynomial long division, you are welcome to read the corresponding Wikipedia page:
https://en.wikipedia.org/wiki/Polynomial_long_division.

Note that the result from the Little Picard theorem does not have an analogous result for polynomials on the real line with real coefficients. For example, on **R**, the polynomial $x^2 - 2x - 5$ has the range $[-6,\infty)$, and therefore there is no real value x such that $x^2 - 2x - 5 = -6.00001$.

Problems

1. If you multiply two polynomials together, how many roots, counting multiplicity, does the product have?

2. If you add two polynomials together, and the degrees of the polynomials are different, how many roots will the sum of the polynomials be?

3. Show that $x^3 - 7x^2 + 11x - 5$ has a root of multiplicity one at $x = 5$.

4. Show that $x^3 - 7x^2 + 11x - 5$ has a root of multiplicity two at $x = 1$.

Solutions

1. The product will have as many roots as the sum of the roots of each of the two polynomials. For example, multiplying a polynomial of degree five and a polynomial of degree four will produce a polynomial of degree nine, which has nine roots.

3. If we evaluate the polynomial at $x = 5$, we get $5^3 - 7\cdot5^2 + 11\cdot5 - 5 = 125-175+55-5 = 0$. Differentiating the polynomial, we get $3x^2 - 14x + 11$, and evaluating this at $x = 5$, we get $3\cdot5^2 - 14\cdot5 + 11 = 75-70+11 = 16$, which is non-zero and therefore the multiplicity of the root is 1.

1.6.12 Roots of unity (or the Roots of 1)

We know that if $z^2 = 1$, then $z = \pm 1$, and if $z^4 = 1$, with a little thought, it should be clear that $z = \pm 1$ or $z = \pm j$. All of these solutions lie on the unit circle, and in general, the solutions to $z^n = 1$ are n values that are equally spaced on the unit circle, each with an angle of $\frac{2\pi}{n}$ between them. For example, Figure 18 shows the 5th roots of unit, the 8th roots of unity, and the 13th roots of unity, respectively.

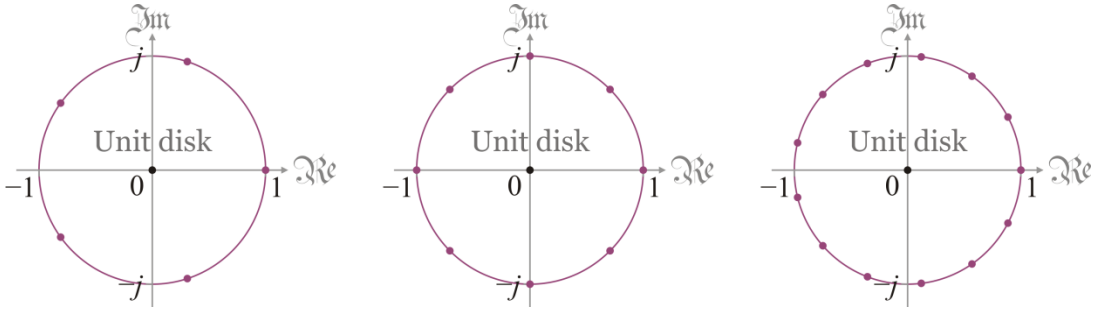


Figure 18. The 5th, 8th and 13th-roots of 1. All the points z in the first image have the property $z^5 = 1$, all the points z in the second have the property that $z^8 = 1$, and all the points z in the third have the property that $z^{13} = 1$.

These numbers have the following properties:

1. the form of the nth roots of unity are $\cos\left(k\frac{2\pi}{n}\right) + j\sin\left(k\frac{2\pi}{n}\right)$ for $k = 0,\ldots,n-1$,
2. the product of two n^{th} roots of unity is an nth root of unity, and
3. the multiplicative inverse of an nth root of unity is an nth root of unity.

The nth root of unity that has the smallest angle is referred to as the *principal* n^{th} *root of unity*. Thus, we have that the 2nd through 8th principal roots of unity are

$$-1, -\frac{1}{2} + j\frac{\sqrt{3}}{2}, j, \cos\left(\frac{2\pi}{5}\right) + j\sin\left(\frac{2\pi}{5}\right), \frac{1}{2} + j\frac{\sqrt{3}}{2}, \cos\left(\frac{2\pi}{7}\right) + j\sin\left(\frac{2\pi}{7}\right), \frac{\sqrt{2}}{2} + j\frac{\sqrt{2}}{2}.$$

Problems

1. Using the polar representation, what are the fifth roots of unity?

2. Using rectangular coordinates, what are the eighth roots of unity?

3. Argue that if z is an nth root of unity, then so is z^*.

Solutions

1. The five roots of unity are $1\angle 0°, 1\angle \pm 72°$ and $1\angle \pm 144°$ or, using radians, $1\angle 0, 1\angle \pm\frac{2\pi}{5}$ and $1\angle \pm\frac{4\pi}{5}$.

3. If $z^n = 1$, then $\left(z^*\right)^n = \left(z^n\right)^* = 1^* = 1$. Alternatively, you could argue that if $z = r\angle\theta$ is an nth root of unity, then $n\theta$ must be a multiple of 2π, in which case, $-n\theta$ is also a multiple of 2π.

1.6.13 Roots of polynomials with real coefficients

We now look at a very useful result.

Theorem

A polynomial with real coefficients has roots that are either real or come in complex conjugate pairs.

Proof:

Assume that all the coefficients of a polynomial of degree n are real. In this case, it is necessary to show that if r is a root of the polynomial, then so is r^*. In this case, we know that

$$\sum_{k=0}^{n} a_k r^k = 0 .$$

For example, the polynomial $3x^2 - 5x + 6$ has $n = 2$ with $a_0 = 6$, $a_1 = -5$ and $a_2 = 3$, so the polynomial equals $\sum_{k=0}^{2} a_k x^k = a_2 x^2 + a_1 x + a_0$ with these values of a_0, a_1 and a_2. This polynomial has a complex root $\frac{5 + j\sqrt{35}}{6}$, so $3\left(\frac{5 + j\sqrt{35}}{6}\right)^2 - 5\left(\frac{5 + j\sqrt{35}}{6}\right) + 6 = 0$. Our goal will be to show that $\frac{5 - j\sqrt{35}}{6}$ must also be a root because all the coefficients a_0, a_1 and a_2 are real.

Upon taking the complex conjugate of both sides (where $0^* = 0$), we have

$$\begin{aligned}\left(\sum_{k=0}^{n} a_k r^k\right)^* &= \sum_{k=0}^{n} \left(a_k r^k\right)^* \\ &= \sum_{k=0}^{n} a_k^* \left(r^k\right)^* \\ &= \sum_{k=0}^{n} a_k^* \left(r^*\right)^k\end{aligned}$$

and as the coefficients are real, we have that $a_k^* = a_k$, so that

$$\left(\sum_{k=0}^{n} a_k r^k\right)^* = \sum_{k=0}^{n} a_k \left(r^*\right)^k = 0 .$$

Therefore, r^* is also a root. ∎

Demonstration of this theorem

You are told that $x = 2 + j$ and $x = -4 - 2j$ are roots of the polynomial

$$2x^5 + 8x^4 - 14x^3 - 80x^2 + 200x .$$

First, because the constant term of this polynomial is zero, $x = 0$ is a root. Because both $2 + j$ and $-4 - 2j$ are roots, so are $2 - j$ and $-4 + 2j$, and thus we have found five roots of this polynomial.

Conversely, we also have the following theorem:

Theorem
A polynomial where all roots are either real or come in complex conjugate pairs has real coefficients whenever the coefficient of the leading term is real.

Proof:
Assume that a polynomial has all roots as either real or as complex conjugate pairs and that the leading term has the coefficient c_0. In this case, the polynomial may be written as

$$p(x) = c_0 \left(\prod_{k=1}^{n_r} (x - r_k) \right) \left(\prod_{k=1}^{n_c} (x - c_k)(x - c_k^*) \right)$$

where the first product includes the real roots and the second product includes all complex conjugate pairs. In the case where a pair of roots come as a complex conjugate pair, if we multiply the pairwise products, we get

$$\begin{aligned} (x - c_k)(x - c_k^*) &= x^2 - (c_k + c_k^*)x + c_k c_k^* \\ &= x^2 - 2\Re e(c_k) + |c_k|^2 \end{aligned}$$

and each coefficient of this quadratic is real. As the original polynomial is the product of either linear or quadratic polynomials all multiplied by c_0, the product must itself have real coefficients. ∎

Demonstration of this theorem
If you multiply the polynomial $3(x - 5)(x - 4 + j)(x - 4 - j) = 3x^3 - 39x^2 + 171x - 255$, where the coefficients of which are all real.
Similarly, $6(x - 3 + 2j)(x - 3 - 2j)(x + 7 + 5j)(x + 7 - 5j) = 6x^4 + 48x^3 + 18x^2 - 1572x + 5772$, again where the coefficients of which are all real.

In MATLAB, the vector `[a b c d e f]` represents the polynomial $ax^5 + bx^4 + cx^3 + dx^2 + ex + f$ (the constant coefficient is always the last entry, and each previous entry represents the coefficient of the next highest term) and the roots routine will return a column vector of the roots (both real and complex) of the polynomial.

```
>> format long
>> roots( [1 2] )
    ans =
        -2

>> roots( [1 2 3] )
    ans =
     -1.000000000000000 + 1.414213562373095i
     -1.000000000000000 - 1.414213562373095i

>> roots( [1 2 3 4] )
    ans =
     -1.650629191439386
     -0.174685404280305 + 1.546868887231397i
     -0.174685404280305 - 1.546868887231397i

>> roots( [1 2 3 4 5] )
    ans =
      0.287815479557649 + 1.416093080171911i
      0.287815479557649 - 1.416093080171911i
     -1.287815479557648 + 0.857896758328490i
     -1.287815479557648 - 0.857896758328490i

>> roots( [1 2 3 4 5 6] )
    ans =
      0.551685463458981 + 1.253348860277207i
      0.551685463458981 - 1.253348860277207i
     -1.491797988139899
     -0.805786469389030 + 1.222904713374409i
     -0.805786469389030 - 1.222904713374409i
```

Problems

1. If 5, $1 + j$ and $2 - 3j$ are roots of a polynomial with real coefficients, what is the minimum possible value of the degree of the polynomial?

2. What is the simplest polynomial that has the root $1 + j$? By simplest, the polynomial has the lowest possible degree and the coefficient of the leading term is 1.

3. What is the simplest polynomial with real coefficients that has $-2 - 3j$ and 1 as roots?

4. If a polynomial is of the form $x^2 + bx + c$, what is the relationship between b and c that results in there being two complex roots?

Solutions

1. As one root is real and the other two are non-real complex numbers, their complex conjugates must also be roots, and thus the degree of the polynomial must be at least five.

3. As 2 is a root, the polynomial must be of the form $(x - 1)p(x)$ where $p(x)$ is another polynomial. As $-2 - 3j$ is a root, so is $-2 + 3j$, so $p(x) = (x + 2 - 3j)(x + 2 + 3j) = x^2 + 4x + 13$, so the full polynomial is $(x - 1)(x^2 + 4x + 13) = x^3 + 3x^2 + 9x - 13$.

1.6.14 Geometric series

You will recall from secondary school the theorem that

Theorem

Given any real number r,

$$\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r},$$

and if $|r|<1$ then

$$\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}.$$

For example, you may have seen that $\sum_{k=0}^{\infty}\left(\frac{2}{3}\right)^k = \frac{1}{1-\frac{2}{3}} = \frac{1}{\frac{1}{3}} = 3$ and $\sum_{k=0}^{n} 2^k = 2^{n+1}-1$; for example, $1 + 2 + 4 = 8 – 1 = 7$.

Proof 1:

$$\begin{aligned}
(1-r)\sum_{k=0}^{n} r^k &= \left(\sum_{k=0}^{n} r^k\right) - r\left(\sum_{k=0}^{n} r^k\right) \\
&= \left(\sum_{k=0}^{n} r^k\right) - \left(\sum_{k=0}^{n} r^{k+1}\right) \\
&= \left(\sum_{k=0}^{n} r^k\right) - \left(\sum_{k=1}^{n+1} r^k\right) \\
&= \left(1+\sum_{k=1}^{n} r^k\right) - \left(\left(\sum_{k=1}^{n} r^k\right) + r^{n+1}\right) \\
&= 1+\left(\sum_{k=1}^{n} r^k\right) - \left(\sum_{k=1}^{n} r^k\right) - r^{n+1} = 1-r^{n+1}
\end{aligned}$$

and therefore dividing both sides by $1 – r$, we get

$$\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r}.$$

If $|r| < 1$ then r^{n+1} goes to zero as n becomes large, so this becomes

$$\sum_{k=0}^{n} r^k = \frac{1}{1-r}. \blacksquare$$

If you don't like the sums, you can more easily see this with

Proof 2:

$$\begin{aligned}(1-r)\sum_{k=0}^{n} r^k &= \left(\sum_{k=0}^{n} r^k\right) - r\left(\sum_{k=0}^{n} r^k\right) \\ &= \left(1+r+r^2+\cdots+r^{n-1}+r^n\right) - r\left(1+r+r^2+\cdots+r^{n-1}+r^n\right) \\ &= \left(1+r+r^2+\cdots+r^{n-1}+r^n\right) - \left(r+r^2+r^3+\cdots+r^n+r^{n+1}\right) \\ &= 1+\left(r+r^2+\cdots+r^{n-1}+r^n\right) - \left(r+r^2+\cdots+r^{n-1}+r^n\right) - r^{n+1} \\ &= 1-r^{n+1}\end{aligned}$$

and, again, divide both sides by $1 - r$ to get our result. ∎

In a sense, the simplified proof is more heuristic—the first proof with sums, changes of indices, etc., is more rigorous. The second is best for comprehension, the first is better for certainty.

If we try this with complex numbers, we get the same result:

$$\begin{aligned}(1+2j)^0+(1+2j)^1+(1+2j)^2+(1+2j)^3+(1+2j)^4 &= 1+(1+2j)+(-3+4j)+(-11-2j)+(-7-24j) \\ &= -19-20j\end{aligned}$$

and

$$\begin{aligned}\frac{1-(1+2j)^5}{1-(1+2j)} &= \frac{1-(41-38j)}{-2j} \\ &= \frac{-40+38j}{-2j} \\ &= \frac{(-40+38j)(2j)}{-2j(2j)} \\ &= \frac{-76-80j}{4} \\ &= -19-20j\end{aligned}$$

Consequently, we also know that $|0.3-0.4\,\mathrm{j}| = 0.5 < 1$, and therefore we may deduce that

$$\begin{aligned}\sum_{k=0}^{\infty}(0.3+0.4j)^k &= \frac{1}{1-(0.3+0.4j)} \\ &= \frac{1}{0.7-0.4j} \\ &= \frac{1}{0.7-0.4j}\cdot\frac{0.7+0.4j}{0.7+0.4j} \\ &= \frac{0.7+0.4j}{0.65} \\ &\approx 1.076923077+0.6153846154j\end{aligned}$$

In Maple, we can calculate such infinite sums exactly:

```
> interface( imaginaryunit = j ): # use "j" instead of "I" to represent the square root of -1.
> s := sum( (3/10 - 4/10*j)^k, k = 0..infinity );
```

$$s := \frac{14}{13} - \frac{8}{13} j$$

```
> # the 'evalf' routine 'eval'uates the argument to a 'f'loating-point number
> evalf( s );
```

$$1.076923077 - 0.6153846154\, j$$

Problems

1. What does the infinite sum of $1+\left(\frac{1}{2}+\frac{1}{2}j\right)+\left(\frac{1}{2}+\frac{1}{2}j\right)^2+\left(\frac{1}{2}+\frac{1}{2}j\right)^3+\left(\frac{1}{2}+\frac{1}{2}j\right)^4+\cdots$ equal?

2. What does the infinite sum $\left(-\frac{1}{3}-\frac{1}{4}j\right)+\left(-\frac{1}{3}-\frac{1}{4}j\right)^2+\left(-\frac{1}{3}-\frac{1}{4}j\right)^3+\left(-\frac{1}{3}-\frac{1}{4}j\right)^4+\cdots$ equal?

3. Given that $(1+2j)^{11} = 6469 - 2642j$, what does the sum $1+(1+2j)+(1+2j)^2+(1+2j)^3+\cdots+(1+2j)^9+(1+2j)^{10}$ equal?

4. Given that $(2-j)^5 = -38-41j$, what does the sum $1+(2-j)+(2-j)^2+(2-j)^3+(2-j)^4$ equal?

Solutions

1. $1 + j$

3. $-1321 - 3234j$

1.6.15 The exponential function

Essential to the engineer is the complex exponential function. While we will give this definition without proof, it can be shown that—and you will see this in your calculus course—that the exponential of a complex number may be found by calculating:

$$e^z = e^{\Re\mathrm{e}(z)}\left(\cos\left(\Im\mathrm{m}(z)\right) + j\sin\left(\Im\mathrm{m}(z)\right)\right).$$

Specifically, if α is real, then $\Re\mathrm{e}(\alpha) = \alpha$ and $\Im\mathrm{m}(\alpha) = 0$, so $e^{\alpha} = e^{\alpha}\left(\cos(0) + j\sin(0)\right) = e^{\alpha}\cdot 1$, and if $j\beta$ is purely imaginary, then $\Re\mathrm{e}(\beta) = 0$ and $\Im\mathrm{m}(\beta) = \beta$, so $e^{j\beta} = e^0\left(\cos(\beta) + j\sin(\beta)\right) = \cos(\beta) + j\sin(\beta)$.

It is very important to notice that $\left|\cos\left(\Im\mathrm{m}(z)\right) + j\sin\left(\Im\mathrm{m}(z)\right)\right| = \sqrt{\cos^2\left(\Im\mathrm{m}(z)\right) + \sin^2\left(\Im\mathrm{m}(z)\right)} = 1$ so therefore it follows that because $\left|e^z\right| = \left|e^{\Re\mathrm{e}(z)}\right|$.

1.6.16 Fields in this course

For the majority of this course, we will focus on vectors that have real entries; however, on occasion, we will venture into looking at vectors the entries of which are complex numbers. This will be relevant to future courses such as quantum mechanics. As we have already noted, we will denote the field of real numbers by **R**, the field of complex numbers by **C**, and if the specific field is irrelevant, we will use **F** to denote any field. As an aside, the field of rational numbers is usually represented by **Q**.

We will represent entries in these fields by lowercase Greek letters.

1.6.17 An application to electrical engineering

One example of where mathematics can be used to model the real world is in alternating current. In secondary school mathematics, you would have learned the tools that could, for example, simplify

$$3.2\sin(377t)+6.5\sin(377t+1)-4.7\sin(377t+2)$$

to $8.7500\cdot\sin(377t+0.1371)$ with trigonometric identities. In your circuits course, you will see that it each of these can be represented by a complex number, as shown in this table:

Trigonometric function	Complex representation	Approximate floating-point value
$3.2\sin(377x)$	$3.2\angle 0 = 3.2(\cos(0)+j\sin(0))$	$=3.2$
$6.5\sin(377x+1)$	$6.5\angle 1 = 6.5(\cos(1)+j\sin(1))$	$\approx 3.5120+5.4696j$
$-4.7\sin(377x+2)$	$-4.7\angle 2 = -4.7(\cos(2)+j\sin(2))$	$\approx 1.9559-4.2737j$

Now, the resulting signal has the complex representation equal to the sum of these values:

$$z = 3.2+3.5120+5.4696+1.9559j-4.2737j = 8.6679+1.1959j$$

and the corresponding sinusoid is $|z|\sin(377t+\angle z)\approx 8.750\sin(377t+\angle 0.1371)$.

1.6.18 Complex numbers in this course

From secondary school, you are aware of vectors: for example, the following is a 2-dimensional vector

$$\mathbf{v}=\begin{pmatrix}3.2\\-5.7\end{pmatrix}.$$

In this case, the entries of the vector are restricted to real values, and therefore we will say that this is a real vector. If, however, we all the entries to be complex, then an example of such a vector is

$$\mathbf{v}=\begin{pmatrix}3.2-9.7j\\-5.7+1.5j\end{pmatrix},$$

and we will call this a *complex vector*. Again, these complex vectors will become useful in circuits with alternating current, but also in quantum mechanics. Sometimes, if we don't care if a vector is real or complex, we will simply call it a *vector over a field*.

1.6.19 Applications of complex numbers

Alternating current may be described by a formula of the form $V = V_{max}\cos(2\pi ft + \phi)$ where f is the frequency in cycles per second or Hz and ϕ is the *phase shift*. In North America, 60 Hz is used, while Europe uses 50 Hz. It happens that $60\pi = 376.99111843077\cdots \approx 377$, so it is common to write the voltage as $V_{max}\cos(377t)$, where 377 is the *angular frequency* measured in radians per second. Power sources may have different phases, so combining two or more voltage sources with different phase shifts requires one to compute, for example,

$$V_1\cos(377t+\phi_1)+V_2\cos(377t+\phi_2).$$

For example, you may be combining two voltage sources with

$$109\cos(377t+1)+134\cos(377t+2).$$

Using double-angle formulas, you will know that you can write this as

$$109\cos(377t)\cos(1)-109\sin(377t)\sin(1)+134\cos(377t)\cos(2)-134\sin(377t)\sin(2)$$

this approximately equals

$$\begin{aligned}109\cos(377t)\cos(1)-109\sin(377t)\sin(1)+134\cos(377t)\cos(2)-134\sin(377t)\sin(2)\\ = 3.129\cos(377t)-213.566\sin(377t)\end{aligned}$$

At this point, we may solve the two non-linear equations

$$\begin{aligned}A\cos(\phi)&=3.129\\ A\sin(\phi)&=213.566\end{aligned}$$

so $A^2 = 3.129^2 + 213.566^2 = 45620.310$ so $A = 213.589$ and thus $\tan(\phi) = \dfrac{213.566}{3.129}$ so $\phi = 1.556$, and therefore

$$109\cos(377t+1)+134\cos(377t+2)+213.589\cos(377t+1.556).$$

Now, consider the effort required to combine more signals; for example,

$$103\cos(377t+0.032)+109\cos(377t+1.053)+134\cos(377t+1.893)+116\cos(377t+3.014)?$$

However, let's look at this a different way. We are going to assume that the frequency association with each sinusoid being added is the same, and that they only differ in their amplitude and their phase shift. We will therefore represent each sinusoid as a complex number, where for

1. $109\cos(377t+1)$, let $z_1 = 109\angle 1$,
2. $134\cos(377t+2)$, let $z_2 = 134\angle 2$,

and, in general, for $V\cos(377t+\phi)$, let $z=V\angle\phi$. Note that $V\sin(377t+\phi)=V\cos(377t+\phi-\frac{\pi}{2})$, so the representation of $V\sin(377t+\phi)$ is $z=V\angle(\phi-\frac{\pi}{2})$. Now, note that $z_1=58.893+91.720\,j$, $z_2=-55.764+121.846\,j$ and

$$z_1+z_2=3.129+213.566\,j=213.589\angle 1.556\,.$$

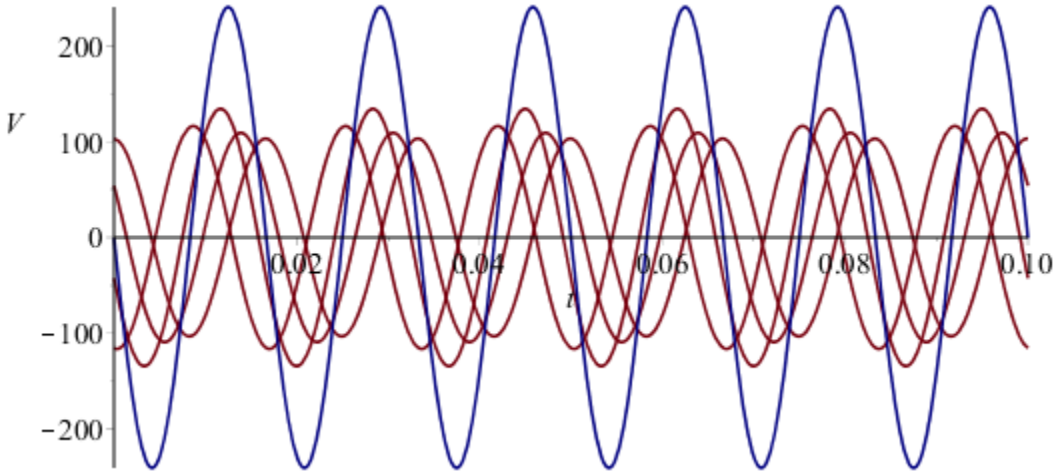
This complex number $213.589\angle 1.556$ represents $213.589\cos(377t+1.556)$, and this is the answer we were looking for. Therefore, for our more difficult problem

$$103\cos(377t+0.032)+109\cos(377t+1.053)+134\cos(377t+1.893)+116\cos(377t+3.014),$$

we simply calculate

$$103\angle 0.032+109\angle 1.053+134\angle 1.893+116\angle 3.014=-0.591+239.872\,j=239.873\angle(1.573),$$

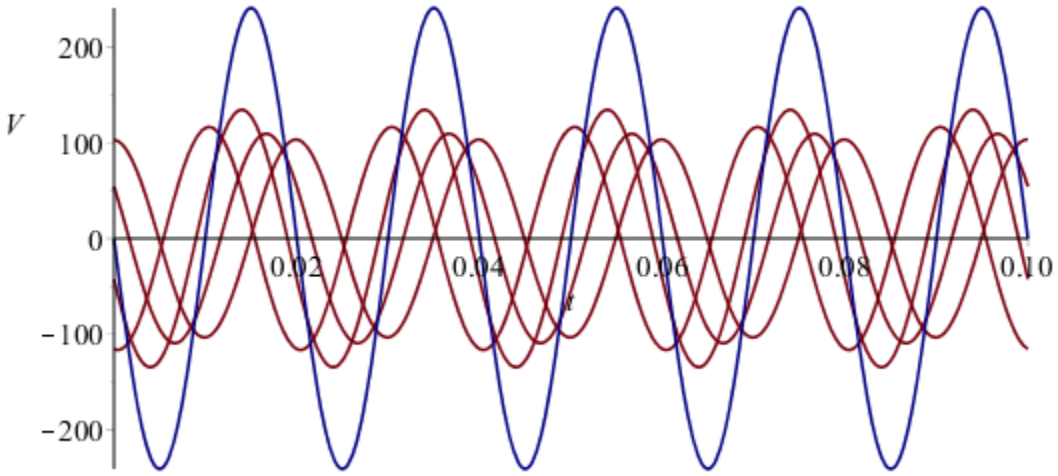
and the consequentially, the sum of these sinusoidal functions is $239.873\cos(377t+1.573)$. Note that all that matters is that the frequencies are the same. For your interest, the following is a plot of the first 100 ms of theses sinusoids and their sum:



In Europe, electrical power is supplied at 50 Hz, and therefore

$$103\cos(100\pi t+0.032)+109\cos(100\pi t+1.053)+134\cos(100\pi t+1.893)+116\cos(100\pi t+3.014)$$
$$=239.873\cos(100\pi t+1.573)$$

Graphically, this sum is shown

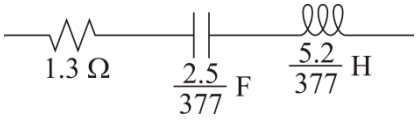


Of course, if the frequencies of the various signals differ, none of this works.

Similarly, when we have alternating currents, the linear circuit elements of resistors, capacitors and inductors all have complex *impedances*:

	Resistor	**Capacitor**	**Inductor**
Impedance	$V=RI$	$V=\frac{1}{377\,jC}I$	$V=(377\,jL)I$

For an alternating current, any linear circuit elements in series have an additive effect, so the effect of the following four circuit elements is

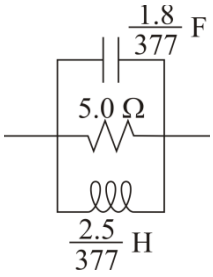


is the complex impedance

$$R+\frac{1}{377\,jC}+377\,jL=1.3+\frac{1}{2.5\,j}+5.2\,j=1.3-0.4\,j+5.2\,j=1.3+4.8\,j\ .$$

You may note that the effects of the capacitors and inductors cancel each other.

Similarly, if they are in parallel,



the complex impedance is

$$\frac{1}{\frac{1}{R}+\frac{1}{\frac{1}{377\,jC}}+\frac{1}{377\,jL}}=\frac{1}{\frac{1}{5.0}+1.8j+\frac{1}{2.5j}}=\frac{1}{0.2+1.8j-0.4j}=\frac{1}{0.2+1.4j}=\frac{0.2-1.4j}{2}=0.1-0.7j\,.$$

Problems

1. Find the complex impedance for circuit elements in series if the resistor is 3.7 Ω, the capacitor is 4/377 F and the inductor has 8.5/377 H.

2. Find the complex impedance for circuit elements in series if the resistor is 1.7 Ω, the capacitor is 5/377 F and the inductor has 9.2/377 H.

3. Find the complex impedance for circuit elements in parallel if the resistor is 2.0 Ω, the capacitor is 1.9/377 F and the inductor has 2.5/377 H.

4. Find the complex impedance for circuit elements in parallel if the resistor is 8.0 Ω, the capacitor is 1.5/377 F and the inductor has 1.6/377 H.

Solutions

1. $3.7 + 8.25j$

3. $0.2 - 0.6j$

1.6.20 Glossary

conjugate	Given a complex number $z = \alpha + j\beta$, the conjugate of z has the same real part but the negated imaginary part and is represented by $z^* = \alpha - j\beta$.
magnitude	Given a complex number $z = \alpha + j\beta$, the magnitude of z is that positive real number $\lvert z\rvert = \sqrt{\alpha^2 + \beta^2}$.

1.6.21 Summary of fields

In this chapter, we have looked at the fields of real numbers and introduced complex numbers. While most operations are similar between the real numbers and complex numbers, we introduce the concept of the argument and complex conjugate. We also saw that there are just as many integers as there are rational numbers, but there are significantly more real numbers than there are complex numbers.

1.7 Summary of introductory material

In this introduction, we have covered the Greek alphabet, a brief introduction to MATLAB, a very brief review of the axiomatic method, and an introduction to the imaginary unit *j* and complex numbers. For complex numbers, we saw that we could write them both in rectangular and polar forms; for example, $12 - 5j$ versus $13\angle -0.39479$ versus $13\angle -22.6199°$. We discussed complex arithmetic and saw that many of the properties of the real numbers are represented by the complex numbers, as well, with the one exception being that the complex numbers cannot be ordered from "smallest" to "largest". We described how every complex polynomial of degree *n* has exactly *n* roots counting multiplicity, and saw the specific *n* roots of unity. Formulas such as geometric sums still apply to complex numbers, and in your Calculus course, you will also see that so do Taylor series representations of trigonometric and exponential functions. Finally, we looked at an application and indicated that if the field doesn't matter, we will use **F**, but if we must restrict ourselves to the real numbers or the complex numbers, we will use **R** or **C**, respectively.

2 Vectors and vector spaces

You are already familiar with vectors from secondary school; however, we will see that there are a great many more mathematical objects that have essentially the same properties, at least, the properties that matter. Mathematics simplifies the work of those who use it by abstracting out those ideas that are important. For example, you do not study 5th-degree polynomials separately from 6th-degree polynomials. Similarly, we will see that while the vectors you have previously learned have interesting and useful properties, we will also see that different objects may also have parallel—and equally useful—properties.

Therefore, we will begin by considering the real finite-dimensional vectors you are already familiar with. We will the proceed to considering such vectors but with complex numbers. Then, in both cases, we will be able to define the basic operations of vector addition and scalar multiplication for both of these classes of vectors. We will conclude by seeing that, despite their apparent differences, that discrete signals (a.k.a. infinite sequences) can also be considered vectors, as can polynomials and as can even more general functions of a real variable.

What you will have to do to really be able to use vectors is to learn to ignore the differences that are superfluous and recognize the similarities that are important, in short:

1. you can add two objects together,
2. you can multiply any object by a scalar value (either a real number or a complex number), and
3. there is a special *zero* vector that has properties similar to the 0 of both real and complex numbers.

Thus, we begin with real finite-dimensional vectors.

2.1 Real finite-dimensional vectors

For finite dimensional vectors containing real entries, we will:

1. define real *n*-dimensional vectors and define a real *n*-dimensional vector space,
2. define the zero vector,
3. consider a geometric interpretation of vectors,
4. describe when two vectors may be considered equal, and
5. consider some applications of vectors.

We will also see that two-dimensional vectors should not be thought of as being equivalent to complex numbers.

2.1.1 Definition of real *n*-dimensional vectors

In secondary school, you would have been exposed to vectors. For example, the following are two-dimensional real vectors:

$$\begin{pmatrix} 3.1 \\ 4.2 \end{pmatrix}, \begin{pmatrix} -3.7 \\ 14.5 \end{pmatrix}, \begin{pmatrix} 18.3 \\ -14.9 \end{pmatrix} \text{ and } \begin{pmatrix} -4.19 \\ -23.53 \end{pmatrix}.$$

We call them *real vectors* because the entries are real numbers. We can interpret these as points in a plane, but be very careful, these are not complex numbers—we will not be multiplying vectors.

The collection of all 2-dimensional vectors will be represented by $\mathbf{R}^2$ and we will write

$$\mathbf{R}^2 = \left\{ \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} : v_1, v_2 \in \mathbf{R} \right\}$$

and describe it as the *vector space* $\mathbf{R}^2$. That is, $\mathbf{R}^2$ is the set of all two-dimensional vectors where the entries are both real numbers (i.e., are both in the set of reals). Please recall that $\pm\infty$ are not real numbers, and therefore we will never consider $\begin{pmatrix} 3 \\ \infty \end{pmatrix}$ to be a real 2-dimensional vector.

You will also have been exposed to $\mathbf{R}^3$, the vector space of all 3-dimensional real vectors, which includes vectors such as

$$\begin{pmatrix} 2.5 \\ 3.9 \\ -8.6 \end{pmatrix} \text{ and } \begin{pmatrix} -4.9 \\ 13.7 \\ -8.1 \end{pmatrix},$$

and again, the set is written as

$$\mathbf{R}^3 = \left\{ \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} : v_1, v_2, v_3 \in \mathbf{R} \right\}.$$

As you may suspect, we can easily generalize this to define a vector for a dimension equal to any positive integer value. Consequently, we will define the vector space $\mathbf{R}^n$ as the set of all *n*-dimensional real vectors, and write

$$\mathbf{R}^n = \left\{ \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} : v_1, v_2, \ldots, v_n \in \mathbf{R} \right\}.$$

For example, the following are both 8-dimensional real vectors:

$$\begin{pmatrix} 1.3 \\ -2.5 \\ -8.4 \\ 0.9 \\ 14.3 \\ 5.8 \\ 1.6 \\ -2.5 \end{pmatrix} \text{ and } \begin{pmatrix} 4.7 \\ 6.2 \\ -1.5 \\ 9.0 \\ 24.5 \\ 4.7 \\ 0.5 \\ 1.3 \end{pmatrix}.$$

Rather than writing vectors in full, we will usually represent vectors by bold lower-case letters; for example, **u**, **v** and **w**. On the blackboard, where it is difficult to create bold characters, we will use superscript arrows, such as $\vec{u}$, $\vec{v}$ and $\vec{w}$. We will usually, but no exclusively select letters near the end of the alphabet for vectors. The individual entries of the *n*-dimensional vector **u** will be represented by italicised letters with a subscript indicating the position from 1 to *n*. Thus, we may say that the entries of **u** are u_1, u_2, through to u_n.

2.1.2 The zero vector

In every vector space, there is one vector of particular importance: the vector where all the entries are zero. We will write the *n*-dimensional zero vector as $\mathbf{0}_n$, and if the actual dimension is not important, we will just write $\mathbf{0}$. For example,

$$\mathbf{0}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

On the blackboard, we will write $\vec{0}_n$ or just $\vec{0}$.

Variable names

So far, we have used numerous variable names, such as `u`, `v1`, `v2`, `z1`, `z2`, etc. In most programming languages, a variable name is any contiguous sequence of characters where:

1. the first character is either a letter of the alphabet or an underscore, and
2. all subsequent characters (if any) are either letters of the alphabet, numbers or underscores.

Most programming languages are also *case-sensitive*, so the variable name `m` is different from the variable name `M`.

You may ask yourself, why can you not use, for example, `3rd_test` as a variable name?

1. First, this ensures that each variable name has at least one letter, and it is easier to distinguish a variable name from a number if all one need do is look at the first character. After all, is `3141S9` a variable name or a number?
2. Second, scientific notation requires that we must interpret `3e5` as 3×10^5. It would be very awkward if we were to have rules like: it can start with a number so long as it doesn't have a single "e", etc.

Suggestions: don't start variable names with just a single character `l` or `O` followed by numbers.

Which of the following are valid variable names?

```
m1  _32  3_r  str$  min_value  max-value  #cat  error_code!  l32  abs_error
```

Whether you are coding in C, C++, Java, C# or most other programming languages, the variable names you can use will be restricted to those defined here.

The zero vector can be easily generated by the constructor

```
>> zero3 = [0 0 0]'
    zero3 =
         0
         0
         0
```

Note that because we cannot start a variable name with a number, we will spell out the word.

2.1.3 Geometric interpretation real 2- and 3-dimensional vectors

A geometric interpretation of a 2-dimensional vector is a point on the *xy*-plane—very similar to the geometric interpretation of a complex number. For example, we will plot the four vectors

$$\begin{pmatrix}1.7\\1.2\end{pmatrix}, \begin{pmatrix}-1.6\\0.9\end{pmatrix}, \begin{pmatrix}-0.8\\-1.1\end{pmatrix} \text{ and } \begin{pmatrix}0.2\\-1.4\end{pmatrix}.$$

on an *xy*-plane where the first entry gives the offset in the *x* direction and the second entry gives the offset in the *y* direction. In Figure 19, we show these four vectors together with the zero vector.

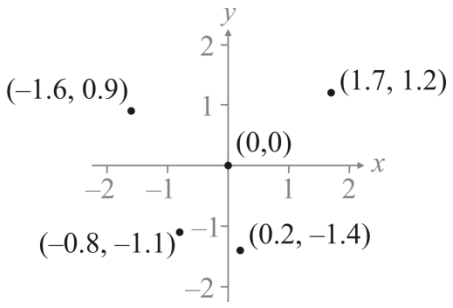


Figure 19. Four vectors in the plane together with the zero vector.

Notice that we may sometimes write vectors as a row, as opposed to our usual column representation, but usually we will use the column representation. In some cases, it is desirable to demonstrate that a vector represents a direction. In this case, the vector can be displayed as an arrow going from the origin to the point on the plane, as shown in Figure 20.

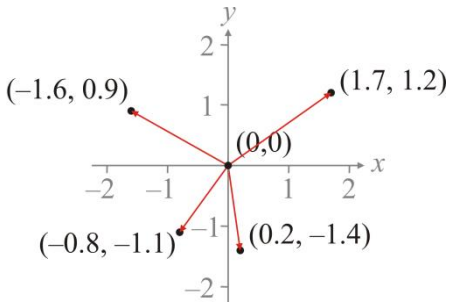


Figure 20. Two-dimensional vectors displayed using arrows to indicate a position relative to the origin.

It becomes more difficult to visualize vectors in three dimensions. The for example, the vectors $\mathbf{u} = \begin{pmatrix} -0.3 \\ 1.0 \\ 0.4 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 0.8 \\ 1.8 \\ -1.3 \end{pmatrix}$ could be shown as Figure 21.

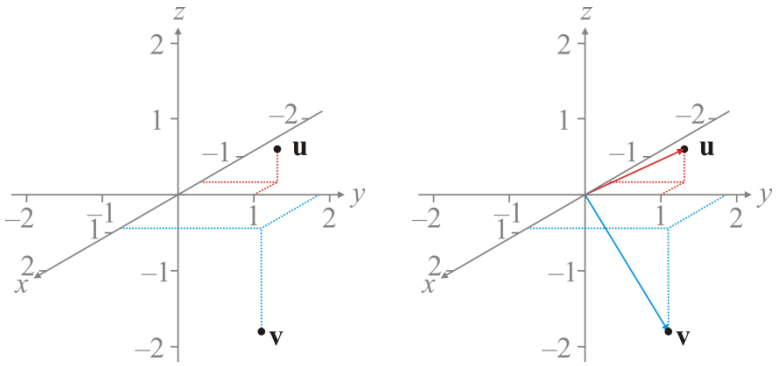


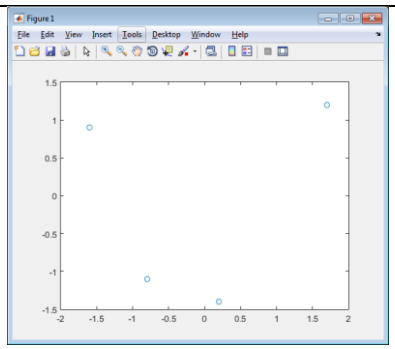
Figure 21. Two three-dimensional vectors shown either as points or as arrows..

Often the first generalization a student will attempt is to interpret higher-dimensional vectors graphically. In a word, **don't**. Some of you may be able to visualize 4-dimensional vectors; however, in applications, you will be often be using vectors with over one million entries. It is easier to think of a real *n*-dimensional vector as a ordered collection of *n* real numbers. There is no intuitive benefit to attempting to visual vectors that are higher than dimension three.

To plot 2-dimensional vectors as points in MATLAB, you can use the `plot` routine. The first argument is a list of the first entries of each of the vectors you mean to plot (to be plotted along the *x*-axis or *abscissa*), and the second argument is a list of the second entries (to be plotted along the *y*-axis or *ordinate*). Of course, both lists must have the same number of entries. The third argument indicates formatting. We will use `'o'` for now to indicate that the points should be drawn as circles. The apostrophe is used to indicate a string in MATLAB.

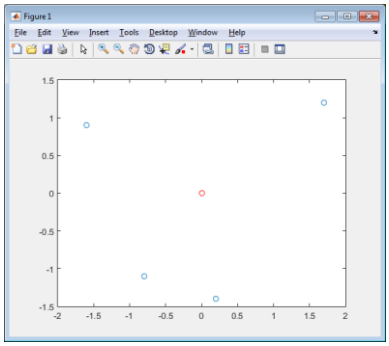
In order to plot the above four vectors, we use

```
>> plot( [-1.6 1.7 -0.8, 0.2], [0.9, 1.2, -1.1, -1.4], 'o' );
```



We note that this does not include the origin. If we issue another plot command, it will erase the current plot, unless we tell MATLAB to hold the current plot.

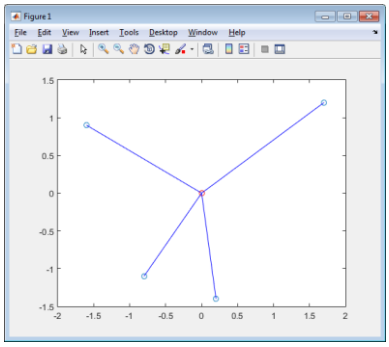
```
>> hold on
>> plot( [0], [0], 'ro' )
```



Here, the "r" in `'ro'` indicates that it should be drawn in red.

If you wanted to draw lines from the origin, each must be drawn separately:

```
>> plot( [0 -1.6], [0, 0.9], '-b' )      % the '-b' indicates a blue line
>> plot( [0  1.7], [0, 1.2], '-b' )
>> plot( [0 -0.8], [0, -1.1], '-b' )
>> plot( [0  0.2], [0, -1.4], '-b' )
```



That MATLAB has no easy way of drawing vectors as arrows, this should indicate the significance of this representation.

2.1.4 Geometric interpretation of *n*-dimensional vectors

Given a 4-dimensional vector, a student may try to interpret this as a position in space-time, so

$$\mathbf{v} = \begin{pmatrix} 3.2 \\ 2.7 \\ -1.5 \\ 4.8 \end{pmatrix}$$

is a point 3.2 m ahead, 2.7 m to the left, 1.5 m down, and 4.8 s ahead in time. For example, if you were standing on the balcony of the United States Capitol on February 28th, 2017 and considered that point to be the origin, then John F. Kennedy would appear at location $\mathbf{v} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1770595200 \text{ s} \end{pmatrix}$. The futility of such representations becomes more difficult when you attempt to visualize 5-dimensional or 235-dimensional vectors. Instead, you can think of an *n*-dimensional vector as a function of the first *n* integers. For example, the vector **v** above could be represented as the four points in Figure 22.

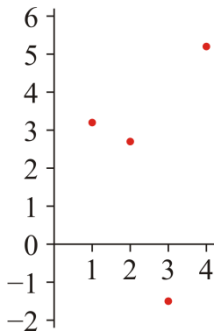


Figure 22. A graphical representation of a 4-dimensional vector.

Consequently, a 32-dimensional vector $\mathbf{v} = \begin{pmatrix} 1.41 \\ 3.52 \\ \vdots \\ 0.76 \end{pmatrix}$ may be represented as shown in Figure 41.

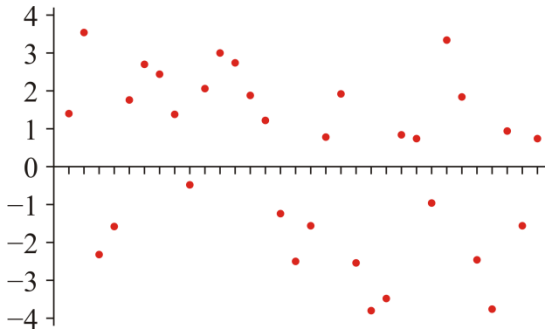


Figure 23. A 32-dimensional vector.

Each entry of this vector may represent, for example, a voltage at a node within a circuit, in which case, there may be no real relationship between value in a vector and another. As an example of another vector, is the number of atoms in a specific molecule. For example, the vectors shown in Figure 24. Rather than numbering the indices, they are represented by the corresponding symbols.

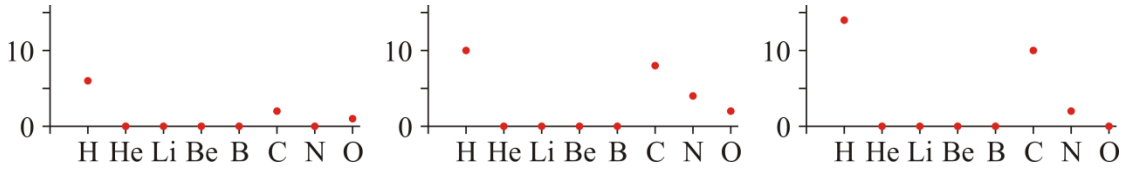


Figure 24. Three 8-dimensional vectors representing the number of atoms from hydrogen through oxygen in ethanol, caffeine and nicotine.

2.1.5 Equality of vectors

Two vectors are said to be *equal* if all the entries are equal. Thus, we will write **u** = **v** if and only if $u_1 = v_1$, $u_2 = v_2$ and all the way up to $u_n = v_n$. If even only one pair of entries differ, the vectors will be considered to be different. Thus, the following are different vectors:

$$\begin{pmatrix} 5.2 \\ -4.3 \\ 6.8 \\ -0.1 \end{pmatrix}, \begin{pmatrix} 5.2 \\ -4.3 \\ 6.8 \\ -0.11 \end{pmatrix}, \begin{pmatrix} 5.2000000001 \\ -4.3 \\ 6.8 \\ -0.1 \end{pmatrix}, \begin{pmatrix} 5.2 \\ -4.3 \\ 6.8 \\ -0.099999999 \end{pmatrix}, \begin{pmatrix} -4.3 \\ 5.2 \\ -0.1 \\ 6.8 \end{pmatrix}.$$

Due to reasons that will be discussed later, we generally will not compare two vectors directly. Given two vectors, we can compare them entry-wise by using the `==` operator.

```
>> u = [1.3 -1.4 0.3 0.5]';
>> v = [1.3 -1.4 0.29 0.5]';
>> u == v    % 1 indicates 'true', while 0 indicates 'false'
    ans =
         1
         1
         0
         1
```

We can ask if *any* of the entries are equal, or if *all* of the entries are equal:

```
>> any( u == v )   % at least one entry is 'true' (1)
    ans =
         1

>> all( u == v )   % they are not all 'true'--the 3rd entry is false
    ans =
         0
```

Problems

1. Are the following pairs of vectors equal? $\begin{pmatrix} 5.2 \\ -4.3 \end{pmatrix}$ and $\begin{pmatrix} 5.2 \\ -4.30 \end{pmatrix}$, $\begin{pmatrix} 5.2 \\ -4.3 \end{pmatrix}$ and $\begin{pmatrix} -4.3 \\ 5.2 \end{pmatrix}$, $\begin{pmatrix} 5.2 \\ -4.4 \end{pmatrix}$ and $\begin{pmatrix} \frac{26}{5} \\ -\frac{22}{5} \end{pmatrix}$, and $\begin{pmatrix} \pi \\ -\sqrt{2} \end{pmatrix}$ and $\begin{pmatrix} 3.141592654 \\ -1.414213562 \end{pmatrix}$.

1. Are the following pairs of vectors equal? $\begin{pmatrix} 3.2-4.5j \\ 6.1+1.7j \end{pmatrix}$ and $\begin{pmatrix} 3.2+4.5j \\ 6.1-1.7j \end{pmatrix}$, $\begin{pmatrix} 1-j \\ 2\sqrt{2}+2\sqrt{2}j \end{pmatrix}$ and $\begin{pmatrix} \sqrt{2}\angle -45° \\ 4\angle 45° \end{pmatrix}$, and $\begin{pmatrix} 2\angle 180° \\ 1\angle 2\pi \end{pmatrix}$ and $\begin{pmatrix} -2 \\ 1 \end{pmatrix}$.

3. Are the following vectors equal? $\begin{pmatrix} 5.2 \\ -4.3 \end{pmatrix}$ and $\begin{pmatrix} 5.2 \\ -4.30 \\ 0 \end{pmatrix}$?

Solutions

1. Because –4.3 = –4.30, the first pair are equal; the corresponding entries are not equal in the second, so they are unequal; 26/5 = 5.2 and –22/5 = –4.4, so the next pair are equal; and finally, the two numbers in the second vector are only approximations of the entries in the first, so the vectors are unequal.

3. No. Comparing vectors from different vector spaces is like comparing apples and oranges. They are different beasts. Later, we will see how we can embed $\mathbf{R}^2$ into $\mathbf{R}^3$, but this embedding is not unique.

2.1.6 Example applications of vectors

Previously, we saw how vectors can be interpreted graphically to represent points or directions in space. This is, however, a very restricted view of vectors. For example, given *n* objects, an *n*-dimensional vector may represent

1. the mass of the objects,
2. the speeds of the objects in a single direction, or
3. the accelerations of the objects in a single direction.

If we are dealing with finances, given *n* stocks, an *n*-dimensional vector may represent

1. the number of the stock that are currently held, or
2. the value per stock.

Given a chemical reaction, given *n* different molecules, an *n*-dimensional vector may represent

1. the amount of each of the molecules (be it in moles or an explicit count of the atoms involved), or
2. the number of carbon atoms per molecule.

In industry, given *n* products that could be manufactured, each requiring specific resources, some of which may be raw materials, others of which may be components that are manufactured separately, then for a specific raw material or component, an *n*-dimensional vector may represent

1. the amount of the raw material or component required by each of the products, or
2. the number of each product that is to be produced.

Alternatively, an *n* dimensional vector could also indicate the profit per product produced. You may note that in many of these cases, the entries are not real, but either integers or discrete values. We can, however, never-the-less, consider these to be real vectors.

2.1.7 Two-dimensional vectors versus complex numbers

We have previously considered how we can represent complex numbers as points on the plane and we have now described how two-dimensional vectors can also be considered as points on a plane. Thus, this begs the question, can we not consider

$$z = 3.2 + 4.5\,j \quad \text{and} \quad \mathbf{u} = \begin{pmatrix} 3.2 \\ 4.5 \end{pmatrix}$$

to be equivalent? After all, they are both represented by the same point on the plane. If you think about it, complex numbers have all the same properties as two-dimensional vectors, for example

$$2 \cdot z = 6.4 + 9.0\,j \quad \text{and} \quad 2\mathbf{u} = \begin{pmatrix} 6.4 \\ 9.0 \end{pmatrix},$$

however, it doesn't make any sense to multiply vectors, and there is no way to multiply three-dimensional vectors in any reasonable sense (and no, the cross product does not count). Similarly, while 1 is a very important complex number, the vector $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ isn't that important. Thus, think of two-dimensional vectors as a specific type of vector, and think of the complex numbers as a two-dimensional analog to real numbers.

As a historical aside, in the mid-1800s, William Rowan Hamilton introduced a more general four-dimensional analog of complex numbers, which he called *quaternions*. He defined a multiplication that had most of the properties associated with the real and complex numbers, only it was no longer commutative—it now mattered whether or not you calculated *xy* or *yx*. It was even possible to define an 8-dimensional variation, which unfortunately, lost the other useful property of associativity: (*xy*)*z* need not equal *x*(*yz*). With further work on vectors and vector analysis, the *vectorialists* finally set the stage for vectors to dominate the world of science and engineering, while quaternions were relegated to applications in graphics.

In MATLAB, the transpose operator is the apostrophe.

```
>> u = [1 2 3 4]     % create a row vector
    u =
         1     2     3     4

>> v = u'                % define v to be the transpose of u
    v =
         1
         2
         3
         4

>> u''
    ans =
         1     2     3     4
```

This gives us a very convenient means of defining a column vector. Instead of using the semi-colon to separate the rows, we can just create a row vector with the same entries, and then just transpose the result, as shown here:

```
>> w = [3.2; -4.5; 8.2; 9.1; 0.4; 9.7]
    w =
        3.2000
       -4.5000
        8.2000
        9.1000
        0.4000
        9.7000

>> w = [3.2 -4.5 8.2 9.1 0.4 9.7]'    % this is much easier and cleaner
    w =
        3.2000
       -4.5000
        8.2000
        9.1000
        0.4000
        9.7000
```

2.2 Finite-dimensional complex vectors

A complex finite-dimensional vector is essentially the same as the vectors we have previously described; however, the entries are complex. We will represent the collection of all *n*-dimensional vectors as the vector space $\mathbf{C}^n$ where, for example,

$$\mathbf{C}^3 = \left\{ \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} : v_1, v_2, v_3 \in \mathbf{C} \right\}.$$

Examples of 2-dimensional complex vectors include

$$\begin{pmatrix} 3.2 - 4.0j \\ -0.6 - 5.4j \end{pmatrix} \text{ and } \begin{pmatrix} 0.5 + 2.3j \\ 6.7 + 8.1j \end{pmatrix}.$$

The zero vector is the same for complex vector spaces as it is for real vector spaces. You cannot visualize complex vectors even in two dimensions, as each dimension is itself a plane. Applications of complex vector spaces include modelling RLC[4] circuits that are supplied by alternating current (AC) source. Here, resistors are positive real, capacitors are positive purely imaginary numbers and inductors are negative purely imaginary numbers.

As we go through the next chapter on vector operations, we will see that the operations that we will define on vectors apply to both real and complex finite-dimensional vectors.

If we do not care whether or not we use real or complex finite dimensional vector spaces, we will use the notation $\mathbf{F}^n$.

2.3 Vector operations

There are two operations we will consider on vectors, be they real or complex, as these operations can be easily defined on any vector of any dimension:

1. scalar multiplication, and
2. vector addition.

For real vectors, we will consider scalar multiplication by real numbers, and for complex vectors we will consider multiplication by complex numbers. We will already begin our step towards abstraction by not caring whether or not we are in a real vector space or a complex vector space. Instead, where it does not matter, we will represent the field to be **F**. In a real vector space, the field is the reals, **R**, while in a complex vector space, the field will be the complex numbers, **C**. Thus, if you see $\mathbf{F}^3$, this means that the example could apply to either $\mathbf{R}^3$ or $\mathbf{C}^3$. In the first case, the scalars would be reals, and in the second, they would be complex numbers.

2.3.1 Scalar multiplication

The most straight-forward operation is scalar multiplication. If we think of a vector as an array or list of entries, multiplying that vector by a scalar $\alpha \in \mathbf{F}$ multiplies each entry by that value. Thus, if **u** is the *n*-dimensional vector

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix},$$

[4] A circuit consisting of resistors, capacitors and inductors.

we will define the scalar multiple $\alpha\mathbf{u}$ as the product

$$\alpha\mathbf{u} = \begin{pmatrix} \alpha u_1 \\ \alpha u_2 \\ \vdots \\ \alpha u_n \end{pmatrix}.$$

If there is the possibility of an ambiguity, we may place a small dot to indicate scalar multiplication. For example, while $\pi\mathbf{u}$ does not pose any visual difficulties, we may prefer to write $3.541\cdot\mathbf{u}$ as opposed to writing $3.541\mathbf{u}$. In a real vector space, the scalar multiples α are restricted to the real numbers, while in a complex vector space, the scalars may also be complex valued.

For example, the following are four vectors together with various scalar multiples of those vectors:

$$\mathbf{u}_1 = \begin{pmatrix} 1.7 \\ 1.2 \end{pmatrix}, \text{ so } 0.3\cdot\mathbf{u}_1 = \begin{pmatrix} 0.51 \\ 0.36 \end{pmatrix},$$

$$\mathbf{u}_2 = \begin{pmatrix} -1.6 \\ 0.9 \end{pmatrix}, \text{ so } 0\cdot\mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$\mathbf{u}_3 = \begin{pmatrix} -0.8 \\ -1.1 \end{pmatrix}, \text{ so } 2.1\cdot\mathbf{u}_3 = \begin{pmatrix} -1.68 \\ -2.31 \end{pmatrix}, \text{ and}$$

$$\mathbf{u}_4 = \begin{pmatrix} 0.2 \\ -1.4 \end{pmatrix}, \text{ so } -1.25\cdot\mathbf{u}_4 = \begin{pmatrix} -0.25 \\ 1.75 \end{pmatrix}.$$

These four vectors, together with their scalar multiples, are shown in Figure 25.

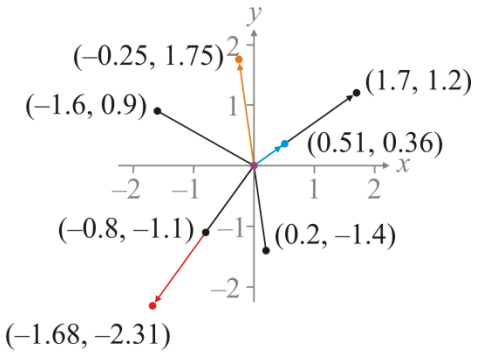


Figure 25. Four vectors and scalar multiples of those vectors.

Geometrically speaking, the vectors are stretched by the scalar value, and if the scalar is negative, the direction of the vector changes.

If **v** is the 6-dimensional complex vector

$$\mathbf{v} = \begin{pmatrix} 0.3 - 4.1j \\ -3.2 + 0.8j \\ 0.0 - 0.1j \\ 0.6 + 5.3j \\ 1.1 - 2.9j \\ -7.9 + 6.0j \end{pmatrix},$$

then $(1.3 - 0.1j)\mathbf{v}$ is the vector

$$(1.3 - 0.1j)\mathbf{v} = \begin{pmatrix} -0.02 - 5.36j \\ -4.08 + 1.36j \\ -0.01 - 1.30j \\ 1.31 + 6.83j \\ 1.14 - 3.88j \\ -9.67 + 8.59j \end{pmatrix}$$

where each entry of the vector **v** is multiplied by 1.3 – 0.1*j*.

One observation we will quickly note is that given an *n*-dimensional vector **v**, then $1 \cdot \mathbf{v} = \mathbf{v}$ and $0 \cdot \mathbf{v} = \mathbf{0}_n$. That is, a vector multiplied by 1 leaves the vector unchanged, and any vector multiplied by 0 produces the zero vector of the same dimension.

In Matlab, we can perform scalar multiplication using the * operator by multiplying a vector by :

```
>> u1 = [1.7; 1.2]
    u =
         1.7000
         1.2000
>> 0.3*u1
    ans =
         0.5100
         0.3600

>> v = [0.3 - 4.1i; -3.2+0.8j; -0.1j; 0.6+5.3j; 1.1-2.9j; -7.9+6j]
    v =

         0.3000 - 4.1000i
        -3.2000 + 0.8000i
              0 - 0.1000i
         0.6000 + 5.3000i
         1.1000 - 2.9000i
        -7.9000 + 6.0000i

>> (1.3 - 0.1j)*v
    ans =
        -0.0200 - 5.3600i
        -4.0800 + 1.3600i
        -0.0100 - 0.1300i
         1.3100 + 6.8300i
         1.1400 - 3.8800i
        -9.6700 + 8.5900i
```

Problems

1. What is the scalar multiplication $2.3 \cdot \begin{pmatrix} 2 \\ 4 \\ -5 \end{pmatrix}$?

2. Is there a scalar multiple of $\begin{pmatrix} 2 \\ 4 \\ -5 \end{pmatrix}$ that gives us $\begin{pmatrix} -1 \\ -2 \\ 2.5 \end{pmatrix}$?

3. Is there a scalar multiple of $\begin{pmatrix} 4.5 \\ -2.1 \\ -0.3 \end{pmatrix}$ that gives $\begin{pmatrix} -9.0 \\ 4.2 \\ 0.9 \end{pmatrix}$?

4. What is the scalar multiplication of $j \begin{pmatrix} 1-j \\ 2+j \\ 3-4j \\ 4+2j \\ -3+2j \end{pmatrix}$?

5. What is the scalar multiplication of $-2j \begin{pmatrix} 2\angle 20° \\ 3\angle -50° \\ 1\angle 120° \\ 3\angle -100° \\ 1.5\angle 170° \end{pmatrix}$?

Solutions

1. $\begin{pmatrix} 4.6 \\ 9.2 \\ -11.5 \end{pmatrix}$

3. No, because to get $4.5\alpha = -9.0$ and $-2.1\alpha = 4.2$, we require that $\alpha = -2$, but $(-2)\cdot(-0.3) = 0.6$ which does not equal 0.9, the third entry in the second vector.

5. First, we note that $-2j = 2\angle -90°$, so we must multiply each absolute value by 2, and subtract 90 degrees from each of the angles, giving us $\begin{pmatrix} 4\angle -70° \\ 6\angle -140° \\ 2\angle 30° \\ 6\angle -190° \\ 3\angle 80° \end{pmatrix}$ but –190 < –180, so we should represent this with $\begin{pmatrix} 4\angle -70° \\ 6\angle -140° \\ 2\angle 30° \\ 6\angle 170° \\ 3\angle 80° \end{pmatrix}$.

2.3.2 Vector addition

Given two *n*-dimensional vectors from the same vector space (either $\mathbf{R}^n$ or $\mathbf{C}^n$), we can add the two vectors together to produce a new vector. For example, if

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix},$$

we will then define

$$\mathbf{u} + \mathbf{v} = \begin{pmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \vdots \\ u_n + v_n \end{pmatrix}.$$

For example, in $\mathbf{R}^4$,

$$\begin{pmatrix} 3.2 \\ 4.7 \\ -1.5 \\ 0.2 \end{pmatrix} + \begin{pmatrix} 7.3 \\ -2.5 \\ -7.2 \\ -4.4 \end{pmatrix} = \begin{pmatrix} 10.5 \\ 2.2 \\ -8.7 \\ -4.2 \end{pmatrix}$$

and in $\mathbf{C}^4$,

$$\begin{pmatrix} 7.2 - 1.2j \\ 4.6 + 2.3j \\ -8.5 - 0.6j \\ 4.2 + 9.7j \end{pmatrix} + \begin{pmatrix} 4.7 + 4.5j \\ 2.9 - 3.7j \\ 0.3 + 0.8j \\ -4.9 - 8.3j \end{pmatrix} = \begin{pmatrix} 11.9 + 3.3j \\ 7.5 - 1.4j \\ -8.2 + 0.2j \\ -0.7 + 1.4j \end{pmatrix}.$$

In two or three dimensions, you can visualize the addition of real vectors by shifting the tail of the one arrow to the head of the other. For example, the addition of $\mathbf{u} = \begin{pmatrix} -1.6 \\ 0.9 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 0.2 \\ -1.4 \end{pmatrix}$, where $\mathbf{u} + \mathbf{v} = \begin{pmatrix} -1.4 \\ -0.5 \end{pmatrix}$ is shown in Figure 26.

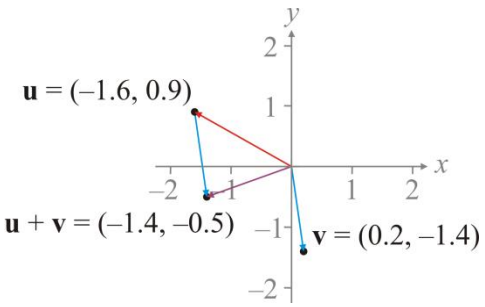


Figure 26. A geometric interpretation of the sum of two 2-dimensional real vectors **u** and **v**.

The addition of $\mathbf{u} = \begin{pmatrix} -0.3 \\ 1.0 \\ 0.4 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 0.8 \\ 1.8 \\ -1.3 \end{pmatrix}$, where $\mathbf{u} + \mathbf{v} = \begin{pmatrix} 0.5 \\ 2.8 \\ -0.9 \end{pmatrix}$, is shown geometrically in Figure 27.

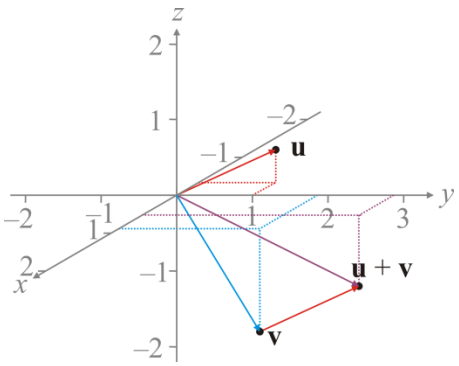


Figure 27. A geometric interpretation of the sum of two 3-dimensional real vectors **u** and **v**.

If both operands of an addition operation are vectors of the same orientation and dimension, MATLAB will add the two; otherwise, it will throw an exception.

```
>> u3 = [1;  2;  3];       % 3-dimensional vector
>> v3 = [4; -3;  5];       % 3-dimensional vector
>> v2 = [5; 4];            % 2-dimensional vector
>> u3 + v3
    ans =
         5
        -1
         8

>> u3 + v2
??? Error using ==> plus
Matrix dimensions must agree.
```

Problems

1. Calculate the following sums of real vectors: $\begin{pmatrix} 3.2 \\ -4.5 \end{pmatrix} + \begin{pmatrix} -5.6 \\ -0.3 \end{pmatrix}$, $\begin{pmatrix} -9.2 \\ 1.5 \\ -7.3 \end{pmatrix} + \begin{pmatrix} 8.9 \\ -9.3 \\ -2.5 \end{pmatrix}$ and $\begin{pmatrix} 0.3 \\ 1.7 \\ 6.4 \\ 2.5 \end{pmatrix} + \begin{pmatrix} -0.3 \\ 2.5 \\ 1.7 \\ 4.9 \end{pmatrix}$.

2. What real vector must be added to $\begin{pmatrix} -9.2 \\ 1.5 \\ -7.3 \end{pmatrix}$ to get the vector $\begin{pmatrix} 8.2 \\ 5.7 \\ -9.5 \end{pmatrix}$?

3. Calculate the sum of the complex vectors $\begin{pmatrix} 3.7+4.6j \\ 5.2-2.9j \end{pmatrix} + \begin{pmatrix} 2.2+3.7j \\ -3.9-1.6j \end{pmatrix}$.

4. What complex vector must be added to $\begin{pmatrix} 3-4j \\ 2+2j \\ 0-3j \\ -5+2j \\ 4+5j \end{pmatrix}$ to get $\begin{pmatrix} 2+2j \\ -1+3j \\ 1-3j \\ -5-6j \\ 3+6j \end{pmatrix}$?

5. In Section 2.1.4, we introduced vectors containing the number of atoms of each of the first eight elements in three various chemicals. Suppose now you had a rum and coke, in which case, there may be 0.00014 moles of caffeine, 1.15 moles of alcohol and 40 moles of water. Calculate a vector using these.

$$0.00014\begin{pmatrix} 6 \\ \vdots \\ 2 \\ 0 \\ 1 \end{pmatrix} + 1.15\begin{pmatrix} 10 \\ \vdots \\ 8 \\ 4 \\ 2 \end{pmatrix} + 40.0\begin{pmatrix} 2 \\ \vdots \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 91.50084 \\ \vdots \\ 9.20028 \\ 4.6 \\ 42.30014 \end{pmatrix}.$$

6. Why rephrase Question 5 as a vector question? Why not

Solutions

1. $\begin{pmatrix} -2.4 \\ -4.8 \end{pmatrix}$, $\begin{pmatrix} -0.3 \\ -7.8 \\ -9.8 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 4.2 \\ 8.1 \\ 7.4 \end{pmatrix}$.

3. $\begin{pmatrix} 5.9+8.3j \\ 1.3-4.5j \end{pmatrix}$

5. $0.00014\begin{pmatrix}6\\ \vdots\\ 2\\ 0\\ 1\end{pmatrix}+1.15\begin{pmatrix}10\\ \vdots\\ 8\\ 4\\ 2\end{pmatrix}+40.0\begin{pmatrix}2\\ \vdots\\ 0\\ 0\\ 1\end{pmatrix}=\begin{pmatrix}91.50084\\ \vdots\\ 9.20028\\ 4.6\\ 42.30014\end{pmatrix}$

2.3.3 Properties of vector operations

First, we will observe that there are some properties that hold for vector addition and scalar multiplication, regardless as to whether we are considering real or complex finite-dimensional vectors. These properties include:

1. Scalar multiplication of a vector by 1 leaves that vector unchanged.
2. Vector addition is commutative.
3. Vector addition is associative.
4. Scalar multiplication distributes over vector addition.
5. Scalar multiplication distributes over addition in the field.
6. Scalar multiplication associates with multiplication within the field in question.
7. The zero vector is an identity element for vector addition.
8. There is an additive inverse for each vector.

These are the properties of vector operations that are considered to be essential. Later, we will see that any set of objects upon which we can define operations of scalar multiplication and vector addition can be considered to be vector spaces. Now, however, we will investigate each of these properties.

2.3.3.1 Scalar multiplication of a vector by 1 leaves that vector unchanged.

This may seem trivial, but this is an important property necessary for vector spaces, and if we modify or change the definition of scalar multiplication, we must never-the-less be sure that this is the case. By definition of scalar multiplication,

$$1\cdot\mathbf{u}=1\cdot\begin{pmatrix}u_1\\ u_2\\ \vdots\\ u_n\end{pmatrix}=\begin{pmatrix}1\cdot u_1\\ 1\cdot u_2\\ \vdots\\ 1\cdot u_n\end{pmatrix}=\begin{pmatrix}u_1\\ u_2\\ \vdots\\ u_n\end{pmatrix}=\mathbf{u}$$

because each u_k is in the field, and 1 is the multiplicative identity, so $1\cdot u_k=u_k$ for each entry. Thus, $1\mathbf{u}=\mathbf{u}$ for all vectors **u**.

2.3.3.2 Vector addition is commutative

Recall that the addition of both real numbers and of complex numbers is commutative, meaning $\alpha+\beta=\beta+\alpha$. Consequently, we note that for vector addition,

$$\mathbf{u}+\mathbf{v}=\begin{pmatrix}u_1+v_1\\u_2+v_2\\\vdots\\u_n+v_n\end{pmatrix}=\begin{pmatrix}v_1+u_1\\v_2+u_2\\\vdots\\v_n+u_n\end{pmatrix}=\mathbf{v}+\mathbf{u}.$$

This is because each of the pairs of entries u_k and v_k belong to a field (real or complex) and in a field addition is commutative, so $u_k + v_k = v_k + u_k$. To emphasize this commutative property, vector addition is sometimes represented using a diamond shape, as is shown in Figure 28.

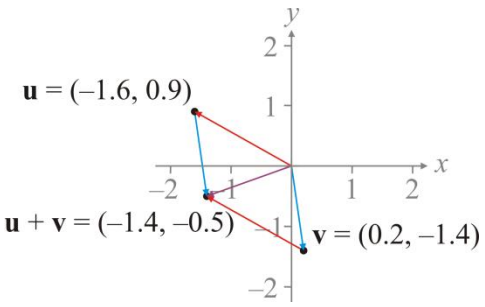


Figure 28. The geometric interpretation of the sums **u** + **v** and **v** + **u**.

This property says that we do not care which order we add two vectors together—you will always get the same result.

Problems

1. Just because **u** + **v** = **v** + **u**, does this mean that 3.2**u** + 1.5**v** = 1.5**v** + 3.2**u**?

Solutions

1. Yes, because any two vectors commute, and 3.2**u** and 1.5**v** are both vectors.

2.3.3.3 Vector addition is associative

Another property of both real and imaginary numbers is that they are associative, meaning that $\alpha+(\beta+\gamma)=(\alpha+\beta)+\gamma$, or if you are adding three numbers together, it doesn't matter if you add the first two first, and then add the third, or add the last two first, and then add the first. Consequently, this property must also hold for vector addition:

$$\mathbf{u}+(\mathbf{v}+\mathbf{w})=\begin{pmatrix}u_1\\u_2\\\vdots\\u_n\end{pmatrix}+\begin{pmatrix}v_1+w_1\\v_2+w_2\\\vdots\\v_n+w_n\end{pmatrix}=\begin{pmatrix}u_1+(v_1+w_1)\\u_2+(v_2+w_2)\\\vdots\\u_n+(v_n+w_n)\end{pmatrix}=\begin{pmatrix}(u_1+v_1)+w_1\\(u_2+v_2)+w_2\\\vdots\\(u_n+v_n)+w_n\end{pmatrix}=\begin{pmatrix}u_1+v_1\\u_2+v_2\\\vdots\\u_n+v_n\end{pmatrix}+\begin{pmatrix}w_1\\w_2\\\vdots\\w_n\end{pmatrix}=(\mathbf{u}+\mathbf{v})+\mathbf{w}$$

Together, these two properties say that given any list of *m* vectors $\mathbf{u}_1,\ldots,\mathbf{u}_m$, if we want to add them, it doesn't matter in what order we add them in, the result will always be the same. What this says is that we can write down **u** + **v** + **w** without any possibility of ambiguity—we do not need to indicate which operation occurs first.

Problems

1. Show that **u** + **v** + **w** = **w** + **v** + **u**.

Solutions

1. (**u** + **v**) + **w** = (**v** + **u**) + **w** = **v** + (**u** + **w**) = **v** + (**w** + **u**) = (**v** + **w**) + **u** = (**w** + **v**) + **u** = **w** + **v** + **u**. In some cases, we use commutativity and in others we use associativity.

2.3.3.4 Scalar multiplication distributes over vector addition

Recall that for real numbers, multiplication distributes over addition, so $\alpha(\beta+\gamma)=\alpha\beta+\alpha\gamma$. It would be desirable that this property hold for scalar multiplication and vector addition, meaning that $\alpha(\mathbf{u}+\mathbf{v})=\alpha\mathbf{v}+\alpha\mathbf{u}$. As expected, it does, for

$$\alpha(\mathbf{u}+\mathbf{v})=\alpha\begin{pmatrix}u_1+v_1\\u_2+v_2\\\vdots\\u_n+v_n\end{pmatrix}=\begin{pmatrix}\alpha(u_1+v_1)\\\alpha(u_2+v_2)\\\vdots\\\alpha(u_n+v_n)\end{pmatrix}=\begin{pmatrix}\alpha u_1+\alpha v_1\\\alpha u_2+\alpha v_2\\\vdots\\\alpha u_n+\alpha v_n\end{pmatrix}=\begin{pmatrix}\alpha u_1\\\alpha u_2\\\vdots\\\alpha u_n\end{pmatrix}+\begin{pmatrix}\alpha v_1\\\alpha v_2\\\vdots\\\alpha v_n\end{pmatrix}=\alpha\mathbf{u}+\alpha\mathbf{v}\ .$$

The next result is similar.

Problems

1. Suppose you were asked to calculate $\frac{1}{3}\begin{pmatrix}2\\4\\2\\4\end{pmatrix}+\frac{1}{3}\begin{pmatrix}3\\5\\2\\6\end{pmatrix}+\frac{1}{3}\begin{pmatrix}5\\2\\4\\8\end{pmatrix}$. How would you perform this calculation efficiently?

Solutions

1. If we multiplied all three vectors first, this would require 12 multiplications, followed by eight additions. If, however, we rewrite it as $\frac{1}{3}\left(\begin{pmatrix}2\\4\\2\\4\end{pmatrix}+\begin{pmatrix}3\\5\\2\\6\end{pmatrix}+\begin{pmatrix}5\\2\\4\\8\end{pmatrix}\right)=\frac{1}{3}\begin{pmatrix}10\\11\\8\\18\end{pmatrix}=\begin{pmatrix}\frac{10}{3}\\\frac{11}{3}\\\frac{8}{3}\\6\end{pmatrix}$, this requires eight additions followed by four multiplications.

2.3.3.5 Scalar multiplication distributes over addition in the field

Similar to the last result, we would like $(\alpha+\beta)\mathbf{u}=\alpha\mathbf{u}+\beta\mathbf{u}$, meaning that, for example, 7**u** has the same result as 3**u** + 4**u**. Again, expanding the definition, we see this is true:

$$(\alpha+\beta)\mathbf{u}=\begin{pmatrix}(\alpha+\beta)u_1\\(\alpha+\beta)u_2\\\vdots\\(\alpha+\beta)u_n\end{pmatrix}=\begin{pmatrix}\alpha u_1+\beta u_1\\\alpha u_2+\beta u_2\\\vdots\\\alpha u_n+\beta u_n\end{pmatrix}=\begin{pmatrix}\alpha u_1\\\alpha u_2\\\vdots\\\alpha u_n\end{pmatrix}+\begin{pmatrix}\beta u_1\\\beta u_2\\\vdots\\\beta u_n\end{pmatrix}=\alpha\mathbf{u}+\beta\mathbf{u}\,.$$

Problems

1. If $\alpha_1+\alpha_2+\alpha_3+\alpha_4+\cdots+\alpha_n=1$, does it follow that $\alpha_1\mathbf{u}+\alpha_2\mathbf{u}+\alpha_3\mathbf{u}+\alpha_4\mathbf{u}+\cdots+\alpha_n\mathbf{u}=\mathbf{u}$?

2. Why would you prefer to calculate $(\alpha_1+\alpha_2+\alpha_3)\mathbf{u}$ instead of $\alpha_1\mathbf{u}+\alpha_2\mathbf{u}+\alpha_3\mathbf{u}$?

3. Which is more efficient to calculate, and why?

Solutions

1. Generalizing this result, we have that $\alpha_1\mathbf{u}+\alpha_2\mathbf{u}+\alpha_3\mathbf{u}+\alpha_4\mathbf{u}+\cdots+\alpha_n\mathbf{u}=(\alpha_1+\alpha_2+\alpha_3+\alpha_4+\cdots+\alpha_n)\mathbf{u}$, and in this case, because $\alpha_1+\alpha_2+\alpha_3+\alpha_4+\cdots+\alpha_n=1$, it follows that

$$\alpha_1\mathbf{u}+\alpha_2\mathbf{u}+\alpha_3\mathbf{u}+\alpha_4\mathbf{u}+\cdots+\alpha_n\mathbf{u}=(\alpha_1+\alpha_2+\alpha_3+\alpha_4+\cdots+\alpha_n)\mathbf{u}=1\cdot\mathbf{u}=\mathbf{u}\,.$$

3. The left-hand side requires one scalar addition and *n* multiplications, while the right-hand side requires 2*n* multiplications.

2.3.3.6 Scalar multiplication associates with multiplication within the field in question

Like the above properties, we would like that $\alpha(\beta\mathbf{u}) = (\alpha\beta)\mathbf{u}$, so for example, it would be desirable that the scalar multiplication of 3 by the scalar multiple 4**u** equals the scalar multiple 12**u**. Again, this is based on the fact that multiplication in the field is associative, meaning $\alpha(\beta\gamma) = (\alpha\beta)\gamma$.

$$(\alpha\beta)\mathbf{u} = \begin{pmatrix} (\alpha\beta)u_1 \\ (\alpha\beta)u_2 \\ \vdots \\ (\alpha\beta)u_n \end{pmatrix} = \begin{pmatrix} \alpha(\beta u_1) \\ \alpha(\beta u_2) \\ \vdots \\ \alpha(\beta u_n) \end{pmatrix} = \alpha \begin{pmatrix} \beta u_1 \\ \beta u_2 \\ \vdots \\ \beta u_n \end{pmatrix} = \alpha(\beta\mathbf{u}).$$

Problems

1. Why can we say that –2(–0.5**v**) = **v**?

2. Suppose that **u** is an *n*-dimensional vector. How many multiplications are required for $(\alpha_1\alpha_2\alpha_3\alpha_4\cdots\alpha_m)\mathbf{u}$ and how many multiplications are required for $\alpha_1\left(\alpha_2\left(\alpha_3\left(\alpha_4\cdots\alpha_{m-1}(\alpha_m\mathbf{u})\right)\right)\right)$?

3. Which is more efficient and why?

Solutions

1. –2(–0.5**v**) = ((–2)(–0.5))**v** = 1**v** = **v**.

3. The left-hand calculation is more efficient requiring $n + 1$ multiplications, while the right-hand side requires $2n$ multiplications.

2.3.3.7 The zero vector is an identity element for vector addition

Note that if we add the zero vector onto any vector, we get that vector back, for

$$\mathbf{u}+\mathbf{0}_n=\begin{pmatrix}u_1+0\\u_2+0\\\vdots\\u_n+0\end{pmatrix}=\begin{pmatrix}u_1\\u_2\\\vdots\\u_n\end{pmatrix}=\mathbf{u}\,.$$

Thus, whether we are considering $\mathbf{R}^n$ or $\mathbf{C}^n$, $\mathbf{0}+\mathbf{u}=\mathbf{u}+\mathbf{0}=\mathbf{u}$ for all $\mathbf{u}$ in that vector space. To this point, we have implicitly assumed that there is only a single zero vector. You may ask yourself, is there a second vector $\mathbf{0}'\neq\mathbf{0}$ such that $\mathbf{0}'+\mathbf{u}=\mathbf{u}$ for all vectors $\mathbf{u}$? This is the first proof where we only need to consider the properties of the zero vector, and we don't even have to consider the representation as an *n*-dimensional vector.

Theorem
The zero vector in a vector space is unique.

Proof:
Suppose that there are two vectors $\mathbf{0}$ and $\mathbf{0}'$, both of which satisfy the conditions of the zero element. Both vectors must satisfy the property of the identity element, and therefore

$$\begin{aligned}\mathbf{0}&=\mathbf{0}+\mathbf{0}' && \text{because } \mathbf{u}=\mathbf{u}+\mathbf{0}' \text{ for all } \mathbf{u}\\ &=\mathbf{0}' && \text{because } \mathbf{u}=\mathbf{u}+\mathbf{0} \text{ for all } \mathbf{u}\end{aligned}$$

Thus, the identity element is unique. ∎

Demonstration of this theorem

In $\mathbf{R}^2$, there is only one zero vector, namely, $\mathbf{0}_2=\begin{pmatrix}0\\0\end{pmatrix}$. In the vector space of all discrete signals, the signal $z = (0, 0, 0, 0, \ldots)$ is the zero vector. The zero function $\hat{0}:\mathbf{R}\to\mathbf{R}$ defined by $\hat{0}(x)=0$ both the zero vector in the vector space of all polynomials and the vector space of all functions.

Finding the zero vector is also trivial: it is simply the scalar multiplication of any vector by the scalar 0.

Theorem
For any vector $\mathbf{u}$, $0\cdot\mathbf{u} = \mathbf{0}$.

Proof:
We note that $0\cdot\mathbf{u}+\mathbf{u}=(0+1)\cdot\mathbf{u}=\mathbf{u}$, and therefore, $0\mathbf{u}$ must be the zero vector; that is, $\mathbf{0}$. ∎

Demonstration of this theorem

In $\mathbf{R}^3$, note that if , $\mathbf{u}=\begin{pmatrix}-2\\1\\5\end{pmatrix}$, then $0\cdot\mathbf{u}=\begin{pmatrix}0\cdot(-2)\\0\cdot1\\0\cdot5\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}$, which equals the zero vector $\mathbf{0}_3$ of $\mathbf{R}^3$. Multiplying the polynomial $x^3 - 4x^2 + 5x + 1$ by 0 gives the zero polynomial $\hat{0}$.

2.3.3.8 There exists an additive inverse for each vector

Recall that for each real or complex number x, we may define a $-x$ so that $x + (-x) = 0$, and subtraction can be defined as

$$x - y = x + (-y).$$

Note that, given a vector $\mathbf{u}$, we can define a new vector $-\mathbf{u}$ by

$$-\mathbf{u} = (-1)\cdot\mathbf{u} = \begin{pmatrix} -u_1 \\ -u_2 \\ \vdots \\ -u_n \end{pmatrix}$$

and this new vector has the property that $-\mathbf{u} + \mathbf{u} = \mathbf{u} + (-\mathbf{u}) = \mathbf{0}$:

$$\mathbf{u} + (-\mathbf{u}) = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} + \begin{pmatrix} -u_1 \\ -u_2 \\ \vdots \\ -u_n \end{pmatrix} = \begin{pmatrix} u_1 - u_1 \\ u_2 - u_2 \\ \vdots \\ u_n - u_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}_n .$$

This vector $-\mathbf{u}$ is called the *additive inverse* of $\mathbf{u}$. Again, for example, given the same vector $\mathbf{u} = \begin{pmatrix} 3.2 \\ -4.5 \\ 1.7 \end{pmatrix}$ above,

$$-\mathbf{u} = \begin{pmatrix} -3.2 \\ 4.5 \\ -1.7 \end{pmatrix}$$

and

$$\mathbf{u} + (-\mathbf{u}) = \begin{pmatrix} 3.2 \\ -4.5 \\ 1.7 \end{pmatrix} + \begin{pmatrix} -3.2 \\ 4.5 \\ -1.7 \end{pmatrix} = \begin{pmatrix} 3.2 - 3.2 \\ -4.5 + 4.5 \\ 1.7 - 1.7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \mathbf{0}_3 .$$

As we do with real and complex numbers, we will write $\mathbf{v} + (-\mathbf{u})$ as $\mathbf{v} - \mathbf{u}$. We can also show that, like the uniqueness of a zero vector, the additive inverse for a specific vector is unique.

Theorem
The additive inverse of a given vector **u** is unique.

Proof:
Suppose that given a vector **u**, there are two different additive inverses, call them **v** and **v′**. In this case,

$$\begin{aligned}
\mathbf{v} &= \mathbf{v} + \mathbf{0} && \text{because } \mathbf{u} = \mathbf{u} + \mathbf{0} \text{ for all } \mathbf{u} \\
&= \mathbf{v} + (\mathbf{u} + \mathbf{v}') && \text{because a vector plus its additive inverse is } \mathbf{0}, \text{ so } \mathbf{0} = \mathbf{u} + \mathbf{v}' \\
&= (\mathbf{v} + \mathbf{u}) + \mathbf{v}' && \text{because vector addition is associative} \\
&= \mathbf{0} + \mathbf{v}' && \text{because a vector plus its additive inverse is } \mathbf{0} \\
&= \mathbf{v}' && \text{because } \mathbf{u} = \mathbf{u} + \mathbf{0} \text{ for all } \mathbf{u}
\end{aligned}$$

Therefore, **v** = **v′**. Therefore **u** must have a unique additive inverse. ∎

Demonstration of this theorem
Given the polynomial p defined as $p(x) = x^3 - 4x^2 - 17x + 60$, if you work out the algebra, you will note that

$$(x^3 - 4x^2 - 17x + 60) + (3 - x)(x + 4)(x - 5) = 0$$

and

$$(x^3 - 4x^2 - 17x + 60) + (-x^3 + 4x^2 + 17x - 60) = 0,$$

it follows that both $(3 - x)(x + 4)(x - 5)$ and $-x^3 + 4x^2 + 17x - 60$ are inverses of p and that both must be equal, namely,

$$(x - 3)(x + 4)(x - 5) = -x^3 + 4x^2 + 17x - 60.$$

We also have a very simple rule for finding the additive inverse.

Theorem
The additive inverse of a vector **u** is (–1)**u**.

Proof:
We note that $(-1)\mathbf{u} + \mathbf{u} = (-1+1)\mathbf{u} = 0\mathbf{u} = \mathbf{0}$, and therefore, $(-1)\mathbf{u}$ must be the additive inverse of **u**, that is, $-\mathbf{u} = (-1)\mathbf{u}$. ∎

As you may also suspect, the additive inverse of the additive inverse of a vector is the vector itself.

Demonstration of this theorem
The additive inverse of the real vector $\mathbf{u} = \begin{pmatrix} 2.3 \\ -1.4 \\ 7.6 \end{pmatrix}$ is $-\mathbf{u} = (-1)\cdot\mathbf{u} = \begin{pmatrix} -2.3 \\ 1.4 \\ -7.6 \end{pmatrix}$.

The additive inverse of the complex vector $\mathbf{v} = \begin{pmatrix} 2.1 - 4.7j \\ -0.6 + 8.9j \end{pmatrix}$ is $-\mathbf{v} = (-1)\cdot\mathbf{v} = \begin{pmatrix} -2.1 + 4.7j \\ 0.6 - 8.9j \end{pmatrix}$.

The additive inverse of the polynomial $p(x) = x^3 - 4x^2 + 6x + 2$ is

$$(-p)(x) = (-1)\,p(x) = (-1)\left(x^3 - 4x^2 + 6x + 2\right) = -x^3 + 4x^2 - 6x - 2.$$

The additive inverse of the function $f(x) = 2\sin(x) - 3\cos(x)$ is

$$(-f)(x) = (-1)\,f(x) = (-1)\left(2\sin(x) - 3\cos(x)\right) = -2\sin(x) + 3\cos(x).$$

Theorem
The additive inverse of the vector –**u** is **u**.

Proof:
Using the result of the previous theorem, $-(-\mathbf{u}) = (-1)((-1)\mathbf{u}) = ((-1)(-1))(\mathbf{u}) = 1\mathbf{u} = \mathbf{u}$. ∎

Demonstration of this theorem

If $\mathbf{u} = \begin{pmatrix} 2.4 \\ -3.9 \\ 1.5 \end{pmatrix}$, then $-\mathbf{u} = -1 \cdot \begin{pmatrix} 2.4 \\ -3.9 \\ 1.5 \end{pmatrix} = \begin{pmatrix} -2.4 \\ 3.9 \\ -1.5 \end{pmatrix}$, and the additive inverse of –**u** is $-(-\mathbf{u}) = -1 \cdot \begin{pmatrix} -2.4 \\ 3.9 \\ -1.5 \end{pmatrix} = \begin{pmatrix} 2.4 \\ -3.9 \\ 1.5 \end{pmatrix} = \mathbf{u}$.

2.3.3.9 Summary of the properties of vector addition and scalar multiplication
We have seen that there are some standard properties that we should come to expect from vectors. What is more important is that it has been determined that these are the essential properties of vectors, at least the properties from which we may deduce subsequent results. We will now look at other such vector spaces.

2.3.4 Linear combinations
Given n vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and n scalars $\alpha_1, \ldots, \alpha_n$, we can calculate the following sum of products of the given vectors with the given scalars:

$$\alpha_1\mathbf{v}_1 + \cdots + \alpha_n\mathbf{v}_n \ .$$

This, too, must be a vector, and it is such a common operation that we will call any such sum of scalar products of vectors with a given set of scalars a *linear combination* of those vectors.

For example, given $\mathbf{u}_1 = \begin{pmatrix} 5 \\ 2 \\ -1 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} -1 \\ 4 \\ 3 \end{pmatrix}, \mathbf{u}_3 = \begin{pmatrix} 2 \\ 3 \\ -2 \end{pmatrix}$, then $4\mathbf{u}_1 + 2\mathbf{u}_2 - 3\,\mathbf{u}_3$ is a linear combination of these three vectors, and it equals

$$4\mathbf{u}_1 + 2\mathbf{u}_2 - 3\mathbf{u}_3 = 4\begin{pmatrix} 5 \\ 2 \\ -1 \end{pmatrix} + 2\begin{pmatrix} -1 \\ 4 \\ 3 \end{pmatrix} - 3\begin{pmatrix} 2 \\ 3 \\ -2 \end{pmatrix} = \begin{pmatrix} 20 \\ 8 \\ -4 \end{pmatrix} + \begin{pmatrix} -2 \\ 8 \\ 6 \end{pmatrix} + \begin{pmatrix} -6 \\ -9 \\ 6 \end{pmatrix} = \begin{pmatrix} 12 \\ 7 \\ 8 \end{pmatrix}.$$

2.3.5 Summary of vector operations
This topic has described both scalar multiplication and vector addition, and we have observed some of the properties that have been recognized as being essential the properties of vector spaces. We have also defined a linear combination of vectors as being a sum of those vectors where each is multiplied by a scalar. We will now look at other vector spaces that also have these properties.

2.4 Other vector spaces

We will now see that there are many other spaces that *behave* in the same way that finite-dimensional vectors do: it is possible to add two such objects within the spaces, it is possible to multiply each by a scalar, there are zero vectors, there are additive inverses, and they all have the other seven properties we described previously. These include, but are not limited to:

1. the vector space of discrete signals,
2. the vector space of polynomials of a single variable, and
3. the vector space of functions of a single variable.

In each case, we will examine both real and complex variations.

2.4.1 Vector space of discrete signals (a.k.a. infinite sequences)

Suppose you have a sensor from which of which you are reading data, and you have an analog-to-digital converter that converts the signal into samples. For example, a temperature sensor in a room on a warm day may produce a sequence such as

$$y = (32.5, 32.4, 32.4, 32.4, 32.5, 32.7, 32.6, 32.6, 32.5, 32.6, 32.6, 32.6, 32.5, 32.6, 33.5, 42.0, \ldots).$$

Under the assumption that the sensor is not being turned off, you can consider such a sequence of readings to be an infinite in length. In your courses on signals and linear systems, you will become familiar with such signals.

Unlike finite-dimensional vectors that are represented using bold roman letters such as **u** and **v**, discrete signals are generally represented by lower-case italicized letters, often with subscripts. If we want to represent the nth value in the sequence y, we will use the notation $y[n]$ (the *zeroeth* entry comes first). Thus, in the above signal, $y[0] = 32.5$, $y[1] = y[2] = y[3] = 32.4$ and $y[4] = 32.5$. We can now define the various vector-space operations:

2.4.1.1 Equality

Two discrete signals x and y are said to be equal if all of the corresponding entries are equal; that is, if $x[k] = y[k]$ for $k = 0,1,2,\ldots$.

2.4.1.2 Scalar multiplication

The kth entry of the discrete signal αy is $\alpha y[k]$; that is, $(\alpha y)[k] = \alpha \cdot y[k]$ for $k = 0,1,2,\ldots$.

For example, given the above discrete signal y, we see that

$$0.5y = (16.25, 16.2, 16.2, 16.2, 16.25, 16.35, 16.3, 16.3, 16.25, 16.3, 16.3, 16.3, 16.25, 16.3, 16.75, 21.0, \ldots).$$

Remember that αy is the discrete signal produced by multiplying each entry of the discrete signal y by α, and $(\alpha y)[k] = \alpha \cdot y[k]$ is the real value that results from multiplying the kth entry of the signal y by α.

2.4.1.3 Vector addition

Given two discrete signals x and y, the kth entry of the discrete signal $x + y$ is $x[k] + y[k]$; that is, $(x+y)[k] = x[k] + y[k]$ for $k = 0,1,2,\ldots$.

Given these two discrete signals,

y_1 = (32.5, 32.4, 32.4, 32.4, 32.5, 32.7, 32.6, 32.6, 32.5, 32.6, 32.6, 32.6, 32.5, 32.6, 33.5, 42.0, …),
y_2 = (32.3, 32.4, 32.3, 32.4, 32.4, 32.5, 32.4, 32.5, 32.5, 32.3, 32.4, 32.4, 32.4, 32.7, 33.3, 40.3, …)

we could calculate 0.5 y_1 + 0.5 y_2, which would give the average of the two discrete signals:

0.5 y_1 + 0.5 y_2
= (32.4, 32.4, 32.35, 32.4, 32.45, 32.6, 32.5, 32.55, 32.5, 32.45, 32.5, 32.5, 32.45, 32.65, 33.4, 41.45, …)

2.4.1.4 Additive inverses

The kth entry of the additive inverse of the discrete signal y is $-y[k]$, meaning that $(-y)[k] = -y[k]$. Again, from our example, the additive inverse of the signal y is

$-y$ = (–32.5, –32.4, –32.4, –32.4, –32.5, –32.7, –32.6, –32.6, –32.5, –32.6, –32.6, –32.6, –32.5, –32.6, –33.5, …).

The reader is asked to determine whether or not the other properties of vectors described above are also valid for discrete signals. For example: Is the addition of discrete signals associative? Does scalar multiplication distribute across the addition of two discrete signals? With a little work, you will see that all seven properties hold.

As an aside, it is possible to define signals explicitly; for example,

$x[k] = 2^{-k}$ so x = (1, 0.5, 0.25, 0.125, 0.0625, …).

You will notice that for this signal approaches zero as n goes to infinity. These are the kinds of discrete signals you will consider in your future signals and systems course.

2.4.1.5 Complex discrete signals

In the above discussion and examples, we have assumed that the entries of these discrete signals are real. There are, however, situations where it is appropriate to have discrete signals of complex numbers. Again, we could add such discrete signals together; we could multiply such a discrete signal by a complex number; and we are left, in each case, with a vector space. For example, the complex discrete signal x defined by $x[k] = (0.5 + 0.5j)^k$ forms a bounded discrete signal of values that spiral to zero in the complex plane:

x = (1, 0.5 + 0.5j, 0.5j, –0.25 + 0.25j, –0.25, –0.125 – 0.125j, –0.125j, 0.0625 – 0.0625j, 0.0625, …).

2.4.2 Vector space of polynomials

We can begin by considering the collection of all polynomials with real coefficients $\mathbf{R}[x]$. This includes all polynomials in x, but it does not include trigonometric, exponential or logarithmic functions. We may add two polynomials with real coefficients together, and that sum is a polynomial—this should be obvious. We may multiply a polynomial p by a real scalar αp so that if

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \cdots + a_2 x^2 + a_1 x + a_0$$

where deg(p) = n, then

$$\alpha p(x) = \alpha a_n x^n + \alpha a_{n-1} x^{n-1} + \alpha a_{n-2} x^{n-2} + \cdots + \alpha a_2 x^2 + \alpha a_1 x + \alpha a_0 .$$

The zero polynomial is the constant polynomial 0, and the additive inverse of a polynomial is that polynomial with all of its coefficients negated; that is,

$$(-p)(x) = -1 \cdot p(x) = -a_n x^n - a_{n-1} x^{n-1} - a_{n-2} x^{n-2} - \cdots - a_2 x^2 - a_1 x - a_0 .$$

As you may suspect, the collection $\mathbf{R}[x]$ defines a vector space, and you are welcome to demonstrate that all the seven properties are satisfied by the above definitions.

MATLAB makes use of this correspondence between vectors and polynomials by allowing you to define polynomials using vectors. For example, the vector

```
>> p = [3 -2.1 5.48 -2.76]    % this can be either a row or column vector
    p =
        3.0000   -2.1000    5.4800   -2.7600
```

defines the polynomial

$$3x^3 - 2.1x^2 + 5.48x - 2.76$$

You can tell MATLAB to interpret a vector as a polynomial and evaluate it at a point using the `polyval` routine:

```
>> polyval( p, 0.6 )    % evaluate the polynomial at the point 0.6
    ans =
        0.4200
```

which is correct, as

$$\begin{aligned} 3(0.6)^3 - 2.1(0.6)^2 + 5.48(0.6) - 2.76 &= 3(0.216) - 2.1(0.36) + 5.48(0.6) - 2.76 \\ &= 0.648 - 0.756 + 3.288 - 2.76 \\ &= 0.42 \end{aligned}$$

2.4.2.1 Polynomials with complex coefficients

We can also define the space of all polynomials with complex coefficients, $\mathbf{C}[z]$. Like polynomials with real confidents, these polynomials can be added together and multiplied by a complex scalar. For example, two polynomials in $\mathbf{C}[z]$ include

$$0.1+0.4j+(-1+1.5j)z+(0.5-0.8j)z^2+(0.2-0.3j)z^3$$
$$2-0.1j+(0.3-1.8j)z+(-0.6-1.1j)z^2+(0.2+0.4j)z^3+(-0.5+0.7j)z^4$$

The sum of these two polynomials is the quartic polynomial

$$2.1+0.3j+(-0.7-0.3j)z+(-0.1-1.9j)z^2+(0.4+0.1j)z^3+(-0.5+0.7j)z^4$$

and the first polynomial multiplied by $0.3 - 1.5j$ is the cubic polynomial

$$0.63-0.03j+(1.95+1.95j)z+(-1.05-0.990j)z^2+(-0.390-0.390j)z^3.$$

Note it is sometimes preferable to ensure that the real coefficient is positive, so the second polynomial may be written as

$$2-0.1j+(0.3-1.8j)z-(0.6+1.1j)z^2+(0.2+0.4j)z^3-(0.5-0.7j)z^4.$$

MATLAB makes use of this correspondence between vectors and polynomials by allowing you to define polynomials using vectors. For example, the vector

```
>> q = [3.2-4.5j 2.5-1.3j -0.2+0.5j] % this can be either a row or column vector
    q =
       3.2000 - 4.5000i   2.5000 - 1.3000i  -0.2000 + 0.5000i
```

defines the polynomial

$$(3.2-4.5j)z^2+(2.5-1.3j)z+(-0.2+0.5j)$$

You can tell MATLAB to interpret a vector as a polynomial and evaluate it at a point using the `polyval` routine:

```
> polyval( q, 1+1j )    % evaluate the polynomial at the point 1 + j
    ans =
      12.6000 + 8.1000i
```

which is correct, as

$$\begin{aligned}(3.2-4.5j)(1+j)^2+(2.5-1.3j)(1+j)+(-0.2+0.5j) &= (3.2-4.5j)2j+(2.5-1.3j)(1+j)+(-0.2+0.5j)\\ &= (9+6.4j)+(3.8+1.2j)+(-0.2+0.5j)\\ &= 12.6+8.1j\end{aligned}$$

2.4.3 Vector space of functions of a single variable (analog signals)

Starting with any domain *D*, where *D* could be the real line **R**, the infinite interval $[0,\infty)$ or just a finite interval $[a,b]$, we can consider the collection of all functions $f: D \to \mathbf{R}$. For a fixed domain *D*, it is possible to

1. multiply such a function *f* by a real scalar α defining a new function $(\alpha f): D \to \mathbf{R}$ where $(\alpha f)(t) = \alpha f(t)$,
2. add two such functions *f* and *g* defining a new function $(f+g): D \to \mathbf{R}$ where $(f+g)(t) = f(t) + g(t)$,
3. define the zero function as that function that is 0 on the domain *D*, and
4. defining the additive inverse of a function *f* by the function $(-f)(t) = -f(t)$.

For example, Figure 29 contains the decaying sinusoid with the graph $e^{-t}\cos(3t)$ (in red), together with two scalar multiples, $2.54e^{-t}\cos(3t)$ (in blue) and $-0.3e^{-t}\cos(3t)$ (in green).

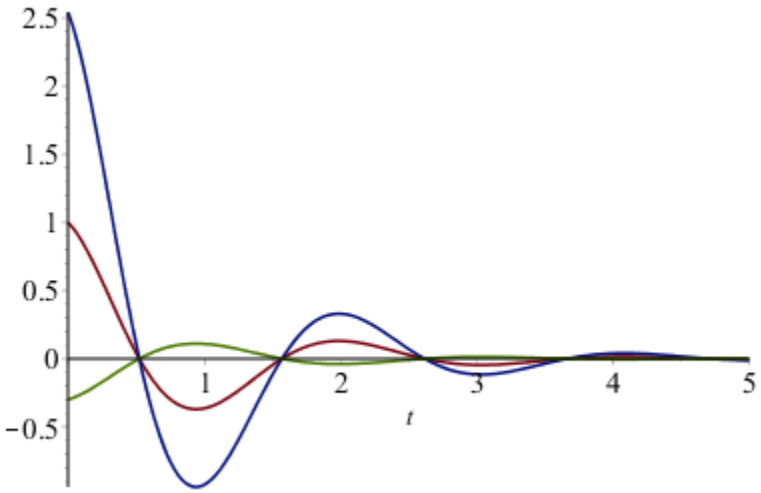


Figure 29. The function $e^{-t}\cos(3t)$ shown in red, together with $2.54e^{-t}\cos(3t)$ and $-0.3e^{-t}\cos(3t)$ in blue and green, respectively.

Similarly, we can add two functions together. For example, Figure 30 shows the sum $-1.9e^{-t}\cos(3x) + 3.7e^{-t}\sin(3t)$.

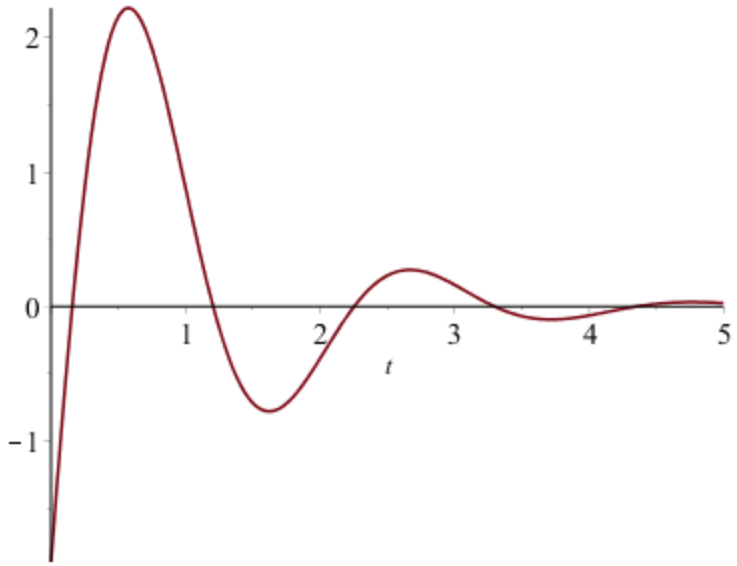


Figure 30. The function resulting from the sum of $-1.9e^{-t}\cos(3x)$ and $3.7e^{-t}\sin(3t)$.

As you may suspect, all the properties of vector space operations are satisfied by these definitions of scalar multiplication and vector addition as long as, in each case, we restrict ourselves to real-valued functions defined on a domain *D*.

These are sometimes called *function spaces*, as opposed to *vector spaces*, but they both satisfy the vector space properties.

2.4.3.1 Complex-valued function spaces

In the same way that we can define the space of real-valued functions of a single variable, we can also define the space of complex-valued functions of a single variable.

2.4.4 Space of images

A gray-scale image is an *m* × *n* grid of values between 0 (black) and 1 (white). While some of you may want to think of this as a *matrix*—something we will see later on, it is best thought of as a vector. Multiplying an image by a value $0<\alpha<1$ can be understood as darkening an image, while multiplying an image by any value $\alpha>1$ can be seen as brightening the image; however, images, while they can be approximately described by vectors, are not true vectors: if **u** is an image and it is being brightened by 20 % by the scalar multiplication 1.2**u**, any entry of **u** that is greater than 1 must be capped at 1.

For example, the following image



Figure 31. A grayscale image of Lena Söderberg (second), multiplied by 0.8 on the left and 1.5 on the right. Multiplying by 1.5 causes some pixels to become capped at 1; these are shown in the last image as black. Thus, images are ideal vectors.

Because the human eye has three different types of receptors, it is possible to approximate full color by using a 3-dimensonal vector. Consequently, a color image is an $m \times n \times 3$ array of values which, again, can be described as a vector. In some cases, such as the cyan-magenta-yellow-black color space (CMYK) used by color printers, the image would be $m \times n \times 4$ array of values. This allows the printer to use black ink as opposed the mixing large equal quantities of cyan, magenta and yellow which should give black but in practice can only produce a very dark grey.

2.4.5 Summary of other vector spaces

In this section, we have described a number of different collections of objects on which we can define addition and scalar multiplication in such a way that the operations have the same properties as those of our finite-dimensional vector spaces. You, as engineers, will use many of these vector spaces in your future courses. While this course will focus on finite-dimensional vectors, each time we introduce a concept, we will see how that concept may be extended to

2.5 Applications of vector arithmetic

We will consider two applications of vector arithmetic:

1. calculating the mean of a set of points,
2. calculating the center of mass, and
3. the *k*-mean algorithm,
4. additive noise.

2.5.1 Calculating the mean or average of a set of vectors

The *mean* or *average value* of a collection of points $\{x_1, x_2, \ldots, x_n\}$ can be found by calculating

$$\bar{x} \overset{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n} x_k \ ,$$

so for example, the mean of {4.88, 0.78, 4.8, 7.99, 3.61} is 4.412. The average of a set of vectors is identical: given a collection of vectors $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$, we define the average or mean $\bar{\mathbf{u}}$ of these vectors as

$$\bar{\mathbf{u}} \overset{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n} \mathbf{u}_k \ .$$

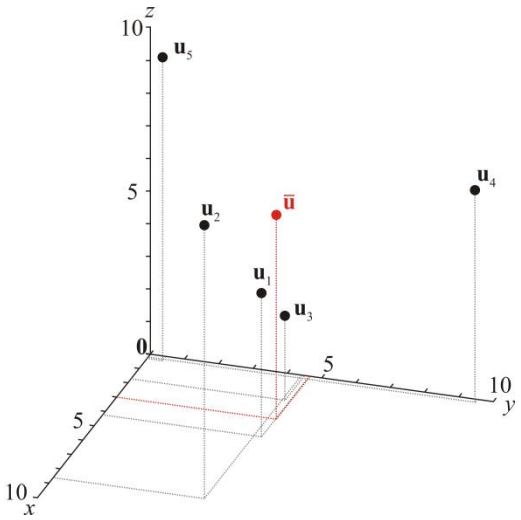
For example, given the following collection of 3-dimensional vectors

$$\left\{ \mathbf{u}_1 = \begin{pmatrix} 4.15 \\ 4.65 \\ 4.59 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 8.69 \\ 4.42 \\ 8.40 \end{pmatrix}, \mathbf{u}_3 = \begin{pmatrix} 1.80 \\ 4.50 \\ 2.65 \end{pmatrix}, \mathbf{u}_4 = \begin{pmatrix} 0.23 \\ 9.46 \\ 6.57 \end{pmatrix}, \mathbf{u}_5 = \begin{pmatrix} 0.03 \\ 0.29 \\ 9.22 \end{pmatrix} \right\} \ ,$$

the average is

$$\bar{\mathbf{u}} = \frac{1}{5}\sum_{k=1}^{5} \mathbf{u}_k = \frac{1}{5}\begin{pmatrix} 14.90 \\ 23.32 \\ 31.43 \end{pmatrix} = \begin{pmatrix} 2.980 \\ 4.664 \\ 6.286 \end{pmatrix},$$

as can be seen in this figure.



2.5.2 Calculating the center of mass

Suppose we have n objects, each with a position $\mathbf{v}_k$ and a mass m_k. The apparent center of mass of these objects may be calculated in a manner very similar to calculating the average, only now we will be calculating a *weighted average*:

$$\mathbf{v} \overset{\text{def}}{=} \frac{\sum_{k=1}^{n} m_k \mathbf{v}_k}{\sum_{k=1}^{n} m_k}$$

As a different such calculation, what is the center of mass of the star system containing Alpha Centauri?

Now, information about the stars is given here:

Name	Alternate	Right ascension	Declination	Distance	Mass
Alpha Centauri A	Rigil Kentaurus	$14^{h}\ 39^{m}\ 36.49400^{s}$ 1.2216781019π	$-60°\ 50'\ 02.3737''$ -0.3379666261π	41.300 Pm	$1.100\ M_{\odot}$ 2.187×10^{31} kg
Alpha Centauri B		$14^{h}\ 39^{m}\ 35.06311^{s}$ 1.2216449794π	$-60°\ 50'\ 15.0992''$ -0.3379862642π	41.304 Pm[5]	$0.907\ M_{\odot}$ 1.804×10^{31} kg
Alpha Centauri C	Proxima Centauri	$14^{h}\ 29^{m}\ 42.94853^{s}$ 1.2079386234π	$-62°\ 40'\ 46.1631''$ -0.3371082764π	40.140 Pm	$0.123\ M_{\odot}$ 2.446×10^{30} kg

Thus, their rectangular coordinates with respect to the vernal equinox is

$$\mathbf{v}_A = \begin{pmatrix} -15.440 \times 10^{15} \\ -12.911 \times 10^{15} \\ -36.064 \times 10^{15} \end{pmatrix}, \mathbf{v}_B = \begin{pmatrix} -15.442 \times 10^{15} \\ -12.909 \times 10^{15} \\ -36.068 \times 10^{15} \end{pmatrix}, \mathbf{v}_C = \begin{pmatrix} -15.609 \times 10^{15} \\ -11.947 \times 10^{15} \\ -34.998 \times 10^{15} \end{pmatrix}$$

The center of mass can therefore be calculated as

[5] Thus assumes Alpha Centauri A is closer than Alpha Centauri B.

$$\frac{m_A\mathbf{v}_A + m_B\mathbf{v}_B + m_C\mathbf{v}_C}{m_A + m_B + m_C\mathbf{v}_C} = \begin{pmatrix} -15.451\times10^{15} \\ -12.855\times10^{15} \\ -36.004\times10^{15} \end{pmatrix}.$$

We can convert this back to standard coordinates to get that the distance is 41.23 Pm, a declination of -0.3287467940π or −60° 49′ 40.8830″, and a right ascension of $14^{h}\ 39^{m}\ 2.3724^{s}$. You will note that, as the ratio $\frac{m_C}{m_A + m_B + m_C\mathbf{v}_C} = 0.05775$ and that, as the angles are so close to each other, each is approximately 5% the distance from the average of the two larger stars towards Proxima Centauri.

2.5.3 Noisy signals

As described, vectors describe discrete (digital) or continuous (analog) flows of information. Such information must often be communicated from one location to another, but interference from other sources may add to the signal. Often, the noise can be described as being white Gaussian noise and the behavior is that this noise is added to the desired signal. For example, the vector

$$\mathbf{u} = (1,0,1,0,1,0,1,0,0,0,1,0,1)$$

may represent the signal you wish to send, but when the signal is received, you get

$$\mathbf{v} = (0.363,-0.151,0.146,-0.393,0.444,-0.573,-0.534,-0.404,-1.472,-0.719,0.162,-0.377,0.685)$$

$$\mathbf{u} + \mathbf{v} = (1.363,-0.151,1.146,-0.393,1.444,-0.573,0.466,-0.404,-1.472,-0.719,1.162,-0.377,1.685).$$

In this case, to remove the noise, you could try to reconstruct the original signal by rounding each entry to whichever value is closer, either 0 or 1, but in this case, the noise overpowers one entry:

$$\tilde{\mathbf{u}} = (1,0,1,0,1,0,0,0,0,0,1,0,1).$$

If you tune your radio to an unused FM radio frequency and turn up the volume, you will hear continuous white Gaussian noise. In some cases, when a receiver does not detect a signal in the noise at a frequency, it mutes the output, a processes known as *squelch*.

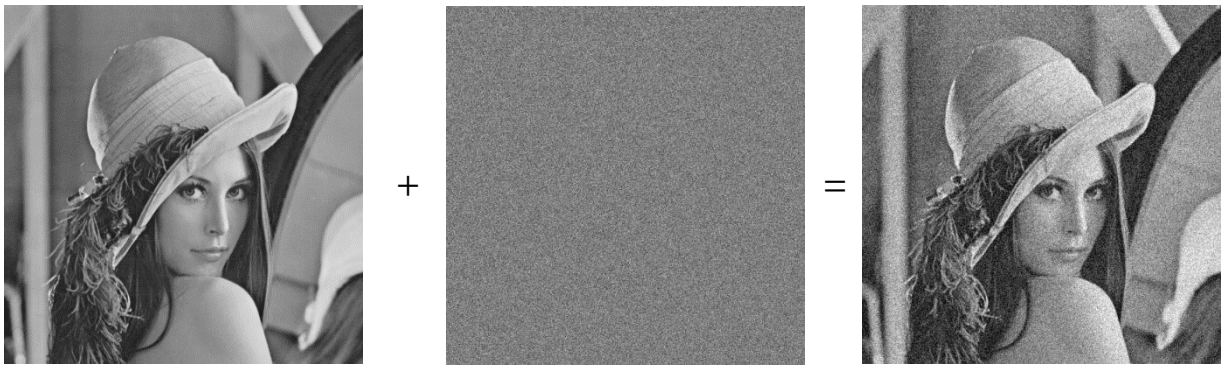
2.5.4 Noisy images

When you take a digital photograph, there is an ideal image you are attempting to capture, one that is a faithful representation of what you see with your eye. If, however, you take that photograph at a very high ISO setting (a high *film speed*), making the image sensor significantly more sensitive, this will add noise to the final image. For example,

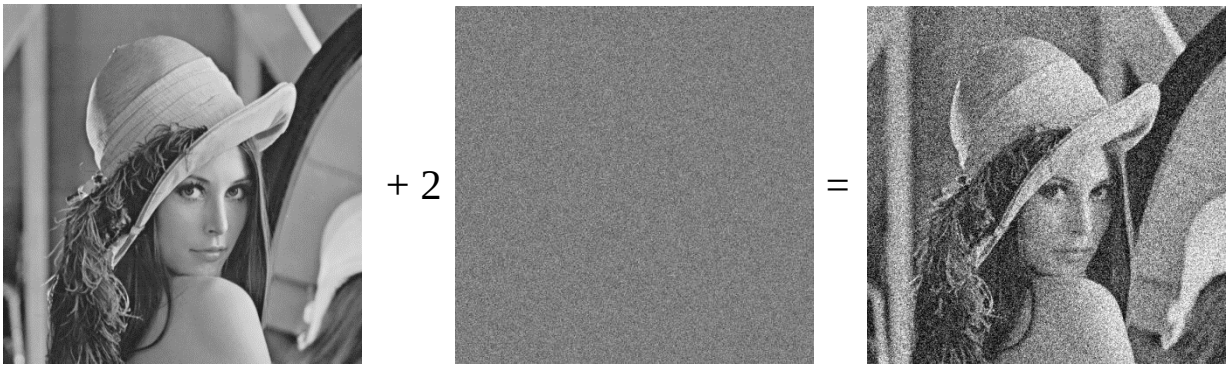


Figure 32. Ref: An image taken under usual lighting and the same image taken User:Ravedave and User:Janke.

As an example of how noise **v** can be added to an image **u**, the following image of Lena has white Gaussian noise added to it creating the final noisy image. The white Gaussian noise is generated by sampling from a random normal distribution with a standard deviation of 0.1. The noise is shown as being relative to an image of 50% gray or 0.5. Of course, any pixel to which the added noise produces a value < 0 or > 1 will be set to the values 0 and 1, respectively.



As before, we may consider **u** = 2**v** to get a noisier image.



Courses in image processing will look at algorithms to reduce the effect of noise on images.

2.6 Summary of vectors and vector operations

In this first section on vectors, we have described vectors and vector operations, the properties of vector operations, and then examined, in addition to $\mathbf{R}^n$ and $\mathbf{C}^n$, other spaces that appear to have the same properties, including polynomials, discrete signals and functions. We concluded by looking at some applications of vector arithmetic.

3 Subspaces

Given a vector space *V*, one important question is: Under what conditions is a subset of *V* a vector space in its own right? To understand this, we will describe sets and subsets, and the look at how vector operations work within various subsets of a given vector space.

3.1 A review of sets

In secondary school, you would have been exposed to sets. Given a set $A = \{a, b, c\}$, the set $B = \{a, c\}$ is said to be a *subset* of *A* (written as $B \subseteq A$), as every element in *B* is also in *A*. On the other hand, the set $C = \{a, d\}$ is not a subset of either sets *A* or *B* (and we write $C \nsubseteq A$). In order to refresh your memory, the entries of a set are called *elements*, so the aforementioned set *A* has three elements *a*, *b* and *c*. We can then write that $b \in A$ but we can also write that $d \notin A$, as *d* is not an element of *A*. Set operations can include $A \cap B$, which is the intersection of the sets *A* and *B* containing exactly those elements that are common to both *A* and *B*, and $A \cup B$, which is the union of the sets *A* and *B* containing exactly those elements that are either in *A* or in *B* or both.

Sets may be described either explicitly, such as {Jadrian, Syed, Waleed, Hashem}, where each item in the sent is explicitly listed, or the set may be described through some formula: $\{\alpha : \alpha \in \mathbf{R} \wedge \alpha > 0\}$. This describes the set of all numbers that are real numbers and that are positive. On the other hand, $\{\alpha j : \alpha \in \mathbf{R} \wedge \alpha > 0\}$ describes the set of numbers of the form αj where α is a real number and positive. Similarly, we could describe a set as $\left\{\begin{pmatrix} \alpha \\ \alpha^2 \end{pmatrix} : \alpha \in \mathbf{R}\right\}$. This is the set of all two-dimensional vectors where the second entry is the square of the first, so this set includes all of these vectors:

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} -3 \\ 9 \end{pmatrix}, \begin{pmatrix} 1.2 \\ 1.44 \end{pmatrix}, \begin{pmatrix} \sqrt{2} \\ 2 \end{pmatrix},$$

but not, for example, $\begin{pmatrix} 2 \\ 5 \end{pmatrix}$.

Consider, for example, the set of the English alphabet, $A = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z\}$. The set of English vowels, $V = \{a, e, i, o, u\}$ is a subset of *A*, so we may write $V \subseteq A$. The set of Russian vowels, however, is $V_{\text{Cyrillic}} = \{а, э, е, и, о, ю, я, у, ы\}$. You will note that some vowels overlap, but $V_{\text{Cyrillic}} \nsubseteq A$, $V_{\text{Cyrillic}} \nsubseteq V$ and $V \nsubseteq V_{\text{Cyrillic}}$.

3.2 Determining if a subset is a vector space

If *S* is a subset of *V*, then if *S* is itself a vector space, we will say that *S* is a *subspace* of *V*. There is no special notation for subspaces as there is for subsets.

For example, given a vector space *V* with a non-zero vector $\mathbf{v} \in V$, the set $\{\mathbf{v}\} \subseteq V$ is not a vector space in its own right. For example, given $\mathbf{R}^2$, the set $\left\{\begin{pmatrix}1\\0\end{pmatrix}\right\} \subseteq \mathbf{R}^2$ but $2\cdot\begin{pmatrix}1\\0\end{pmatrix}=\begin{pmatrix}2\\0\end{pmatrix} \notin \left\{\begin{pmatrix}1\\0\end{pmatrix}\right\}$, so in general, a subset containing only a single vector does not make a vector space. To show that an arbitrary set of objects together with two operations called *element addition* and *scalar multiplication* actually forms a vector space requires us to show that all the properties of a vector space hold. What's important is that it is always possible to define element addition and scalar multiplication is such a way that all but one of the properties we indicate hold, but that one property fails. If, however, we already know that a set is a subset of a vector space, our job is much simpler:

Theorem

If *V* is a vector space and $S \subseteq V$, *S* is a vector space if and only if $\mathbf{u}, \mathbf{v} \in S$ implies that $\alpha\mathbf{u} + \beta\mathbf{v} \in S$.

Proof:

If *S* is a vector space and $\mathbf{u}, \mathbf{v} \in S$, then by definition $\alpha\mathbf{u} + \beta\mathbf{v} \in S$.

Alternatively, if $\mathbf{u}, \mathbf{v} \in S$ implies that $\alpha\mathbf{u} + \beta\mathbf{v} \in S$ then all other properties of vector spaces must hold for vectors is *S*. ∎

Demonstration of this theorem

Consider all vectors of the form $\begin{pmatrix}a\\b\\c\end{pmatrix}$ where c = 2*a* – 5*b*. If **u** and **v** are of this form, then $\mathbf{u}=\begin{pmatrix}u_1\\u_2\\2u_1-5u_2\end{pmatrix}$ and $\mathbf{v}=\begin{pmatrix}v_1\\v_2\\2v_1-5v_2\end{pmatrix}$. In this case,

$$\begin{aligned}\alpha\mathbf{u}+\beta\mathbf{v} &= \alpha\begin{pmatrix}u_1\\u_2\\2u_1-5u_2\end{pmatrix}+\beta\begin{pmatrix}v_1\\v_2\\2v_1-5v_2\end{pmatrix}\\ &= \begin{pmatrix}\alpha u_1\\\alpha u_2\\2\alpha u_1-5\alpha u_2\end{pmatrix}+\begin{pmatrix}\beta v_1\\\beta v_2\\2\beta v_1-5\beta v_2\end{pmatrix}\\ &= \begin{pmatrix}\alpha u_1+\beta v_1\\\alpha u_2+\beta v_2\\2\alpha u_1-5\alpha u_2+2\beta v_1-5\beta v_2\end{pmatrix}\\ &= \begin{pmatrix}\alpha u_1+\beta v_1\\\alpha u_2+\beta v_2\\2(\alpha u_1+\beta v_1)-5(\alpha u_2-\beta v_2)\end{pmatrix}\end{aligned}$$

and thus, $\alpha\mathbf{u}+\beta\mathbf{v}$ is also of the desired form. Thus, all vectors of the given form define a subset of $\mathbf{R}^3$ that is also a vector space, and therefore a subspace of $\mathbf{R}^3$. Note that $\begin{pmatrix}1\\1\\1\end{pmatrix}$ is not of this form, so the subspace is a proper subspace of $\mathbf{R}^3$, meaning it is a subset and not equal to the entire space.

Next, consider all the vectors of the form $\begin{pmatrix}a\\b\\c\end{pmatrix}$ where c = 2*ab*. Two such vectors are $\mathbf{u}=\begin{pmatrix}1\\1\\2\end{pmatrix}$ and $\mathbf{v}=\begin{pmatrix}1\\-1\\-2\end{pmatrix}$. In this case, $\alpha\mathbf{u}+\beta\mathbf{v}=\alpha\begin{pmatrix}1\\1\\2\end{pmatrix}+\beta\begin{pmatrix}1\\-1\\-2\end{pmatrix}=\begin{pmatrix}\alpha\\\alpha\\2\alpha\end{pmatrix}+\begin{pmatrix}\beta\\-\beta\\-2\beta\end{pmatrix}=\begin{pmatrix}\alpha+\beta\\\alpha-\beta\\2\alpha-2\beta\end{pmatrix}$. You will note that $(\alpha+\beta)(\alpha-\beta)\alpha^2-\beta^2\neq 2\alpha-2\beta$, and therefore $\alpha\mathbf{u}+\beta\mathbf{v}$ is not in the desired form, and therefore all vectors of the given form do not define a subspace of $\mathbf{R}^3$.

A related theorem splits the condition into two parts.

Theorem
If *V* is a vector space and $S\subseteq V$, *S* is a vector space if and only if $\mathbf{u},\mathbf{v}\in S$ implies that $\mathbf{u}+\mathbf{v}\in S$ and $\alpha\mathbf{u}\in S$.

Proof:
If $\mathbf{u},\mathbf{v}\in S$, then $\alpha\mathbf{u},\beta\mathbf{v}\in S$, and so therefore $\alpha\mathbf{u}+\beta\mathbf{v}\in S$. On the other hand, if $\alpha\mathbf{u}+\beta\mathbf{v}\in S$, then setting $\alpha=\beta=1$, it follows that $\mathbf{u}+\mathbf{v}\in S$ and setting $\beta=0$, it follows that $\alpha\mathbf{u}+0\cdot\mathbf{v}=\alpha\mathbf{u}+\mathbf{0}=\alpha\mathbf{u}\in S$. ∎

Demonstration of this theorem

Consider all vectors of the form $\begin{pmatrix}a\\b\\c\end{pmatrix}$ where c = 2*a* – 5*b*. If **u** and **v** are of this form, then $\mathbf{u}=\begin{pmatrix}u_1\\u_2\\2u_1-5u_2\end{pmatrix}$ and $\mathbf{v}=\begin{pmatrix}v_1\\v_2\\2v_1-5v_2\end{pmatrix}$. In this case,

$$\mathbf{u}+\mathbf{v}=\begin{pmatrix}u_1\\u_2\\2u_1-5u_2\end{pmatrix}+\begin{pmatrix}v_1\\v_2\\2v_1-5v_2\end{pmatrix}\qquad \alpha\mathbf{u}=\alpha\begin{pmatrix}u_1\\u_2\\2u_1-5u_2\end{pmatrix}$$

$$=\begin{pmatrix}u_1+v_1\\u_2+v_2\\2u_1-5u_2+2v_1-5v_2\end{pmatrix}\quad\text{and}\quad=\begin{pmatrix}\alpha u_1\\\alpha u_2\\2\alpha u_1-5\alpha u_2\end{pmatrix}$$

$$=\begin{pmatrix}u_1+v_1\\u_2+v_2\\2(u_1+v_1)-5(u_2-v_2)\end{pmatrix}\qquad=\begin{pmatrix}\alpha u_1\\\alpha u_2\\2(\alpha u_1)-5(\alpha u_2)\end{pmatrix}$$

and thus, $\mathbf{u}+\mathbf{v}$ and $\alpha\mathbf{u}$ are both of the desired form. Thus, all vectors of the given form define a subset of $\mathbf{R}^3$ that is also a vector space, and therefore a subspace of $\mathbf{R}^3$.

Next, consider all the vectors of the form $\begin{pmatrix} a \\ b \\ c \end{pmatrix}$ where c = 2*ab*. One such vectors is $\mathbf{u} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$ and $\alpha\mathbf{u} = \alpha\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha \\ 2\alpha \end{pmatrix}$. You will note that $\alpha^2 \neq 2\alpha^2$, and therefore $\alpha\mathbf{u}$ is not in the desired form, and therefore all vectors of the given form do not define a subspace of $\mathbf{R}^3$. Because we have already found one counter-example, we don't even have to test **u** + **v**.

Sometimes it is easier to show that $\alpha\mathbf{u}+\beta\mathbf{v} \in S$ and sometimes it is easier to show $\mathbf{u}+\mathbf{v} \in S$ and $\alpha\mathbf{u} \in S$. On the other hand, if *S* is not a vector space, it is only necessary to show that one of the two fail. For example, the set of all pairs of integers $\begin{pmatrix} m \\ n \end{pmatrix}$ is not a real vector space because $\frac{1}{2}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}$ is not a pair of integers. Similarly, the set of all pairs of real numbers $\begin{pmatrix} x \\ y \end{pmatrix}$ such that $x^2 + y^2 \leq 2$ is not a vector space because $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is in this space, but $2 \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ is not. Previously, we pointed out that $\{\mathbf{v}\} \subseteq V$ is not a vector space if **v** is not the zero vector. The following, however, shows that the set containing only the zero vector is itself a vector space.

Theorem
If *V* is a vector space, then {**0***v*} is a subspace of *V*.

Proof:
First, because *V* is a vector space, it must have a zero vector, so $\{\mathbf{0}_V\} \subseteq V$. Thus, $\mathbf{0}_V + \mathbf{0}_V = \mathbf{0}_V \in \{\mathbf{0}_V\}$ and $\alpha\mathbf{0}_V = \mathbf{0}_V \in \{\mathbf{0}_V\}$ for all $\alpha \in \mathbf{F}$. Therefore, $\{\mathbf{0}_V\}$ is a subspace of *V*. ∎

Demonstration of this theorem

In $\mathbf{R}^2$, the set containing only the zero vector, $\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right\}$, is a subspace of $\mathbf{R}^2$. In $\mathbf{C}^3$, the set $\left\{\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\right\}$. Given the set of all polynomials **R**[*x*], the set containing only the zero polynomial, $\{\hat{0}\}$ is the a subspace of all polynomials.

The smallest subspace containing a vector **v** is the one-dimensional subspace of all scalar multiples of that vector.

Theorem
If *V* is a vector space and $\mathbf{v} \in V$, then the set $S_{\mathbf{v}} = \{\alpha\mathbf{v} : \alpha \in \mathbf{F}\}$ is a subspace.

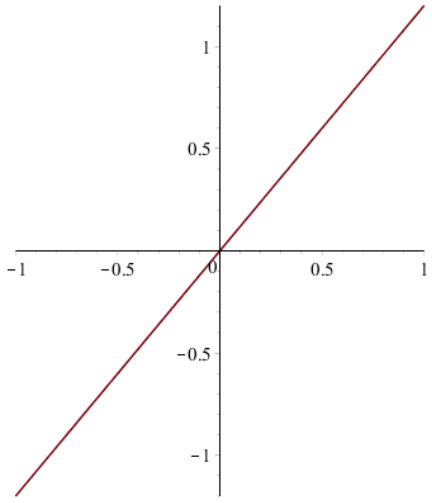
Proof:
If $\mathbf{u}_1, \mathbf{u}_2 \in S_{\mathbf{v}}$, then $\mathbf{u}_1 = \gamma_1\mathbf{v}$ and $\mathbf{u}_2 = \gamma_2\mathbf{v}$, so

$$\alpha\mathbf{u}_1 + \beta\mathbf{u}_2 = \alpha(\gamma_1\mathbf{v}) + \beta(\gamma_2\mathbf{v}) = (\alpha\gamma_1)\mathbf{v} + (\beta\gamma_2)\mathbf{v} = (\alpha\gamma_1 + \beta\gamma_2)\mathbf{v}.$$

but because **F** is a field, $\alpha\gamma_1 + \beta\gamma_2 \in \mathbf{F}$, so $(\alpha\gamma_1 + \beta\gamma_2)\mathbf{v}$ is a scalar multiple of **v** so $(\alpha\gamma_1 + \beta\gamma_2)\mathbf{v} = S_{\mathbf{v}}$. ∎

Demonstration of this theorem

If we consider all scalar multiples of the vector $\mathbf{u} = \begin{pmatrix} 6 \\ 5 \end{pmatrix}$, these form a line in $\mathbf{R}^2$, as shown here.



We note that $0.5 \cdot \mathbf{u} = \begin{pmatrix} 3 \\ 2.5 \end{pmatrix} \in S_{\mathbf{u}}$ and $-2.4 \cdot \mathbf{u} = \begin{pmatrix} -14.4 \\ -12 \end{pmatrix} \in S_{\mathbf{u}}$, and $\begin{pmatrix} 3 \\ 2.5 \end{pmatrix} + \begin{pmatrix} -14.4 \\ -12 \end{pmatrix} = \begin{pmatrix} -11.4 \\ -9.5 \end{pmatrix} = -1.9 \cdot \mathbf{u} \in S_{\mathbf{u}}$.

If we consider the polynomial

Given two subspaces, it is possible to consider their intersection. This intersection is, itself, also a vector space.

Theorem

If *V* is a vector space, then the intersection of two subspaces *V* is itself a subspace of *V*.

Proof:

Suppose that *S* and *T* are subspaces of *V*. Assume that $\mathbf{u}, \mathbf{v} \in S \cap T$. Then because both *S* and *T* are vector spaces, it follows that $\alpha\mathbf{u} + \beta\mathbf{v} \in S$ and $\alpha\mathbf{u} + \beta\mathbf{v} \in T$, so $\alpha\mathbf{u} + \beta\mathbf{v} \in S \cap T$. Therefore, $S \cap T$ is a vector space. ∎

Demonstration of this theorem

Consider for example the two subspaces of $\mathbf{R}^3$ defined first by all those vectors that satisfy $\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$ where $\gamma = 2\alpha - 5\beta$, and second by all those vectors that satisfy $\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$ where $\gamma = 4\alpha$. Now, if we consider the intersection of these two subspaces, that is all vectors that satisfy both $\gamma = 2\alpha - 5\beta$ and $\gamma = 4\alpha$, then $2\alpha - 5\beta = 4\alpha$ or $\beta = -0.4\alpha$. Thus, the intersection consists of all vectors of the form $\begin{pmatrix} \alpha \\ -0.4\alpha \\ 4\alpha \end{pmatrix}$, and we note that this is the vector space of all scalar multiples of $\begin{pmatrix} 1 \\ -0.4 \\ 4 \end{pmatrix}$, which is itself a vector space.

For example, suppose that **u** and **v** are two vectors in *V*. Then $S_{\mathbf{u}}$ and $S_{\mathbf{v}}$ are both vector spaces. If $\mathbf{u} = \alpha\mathbf{v}$, then $S_{\mathbf{u}}$ = $S_{\mathbf{v}}$; otherwise, $S_{\mathbf{u}} \cap S_{\mathbf{v}} = \{\mathbf{0}_V\}$, which we have already shown is a vector space.

Problems

1. Demonstrate whether or not the set $S_1 = \left\{ \begin{pmatrix} \alpha \\ 0 \end{pmatrix} : \alpha \in \mathbf{R} \right\}$ (that is, all vectors that have a value in the first entry and 0 in the second), forms a subspace of $\mathbf{R}^2$.

2. Demonstrate whether or not the set $S_2 = \left\{ \begin{pmatrix} \alpha \\ \alpha \\ \beta \end{pmatrix} : \alpha, \beta \in \mathbf{R} \right\}$ (that is, all vectors that have the same entry in the first two entries, and a possibly different number in the third), forms a subspace of $\mathbf{R}^3$.

3. Demonstrate whether or not the set $S_3 = \left\{ \begin{pmatrix} \alpha \\ \alpha + 1 \end{pmatrix} : \alpha \in \mathbf{R} \right\}$ is a subset of $\mathbf{R}^2$.

4. Demonstrate whether or not the set $S_4 = \left\{ \begin{pmatrix} \alpha \\ \beta \\ (\alpha - \beta)^2 \end{pmatrix} : \alpha, \beta \in \mathbf{R} \right\}$ is a subset of $\mathbf{R}^3$.

5. Demonstrate whether or not the set $S_5 = \left\{ \begin{pmatrix} \alpha \\ \alpha + \beta \\ \alpha - 3\beta \end{pmatrix} : \alpha, \beta \in \mathbf{R} \right\}$ is a subset of $\mathbf{R}^3$.

6. Demonstrate whether or not the set $S_6 = \left\{ \begin{pmatrix} 3\alpha \\ 4\alpha - \beta \\ -2\alpha + 3\beta \end{pmatrix} : \alpha, \beta \in \mathbf{R} \right\}$ is a subset of $\mathbf{R}^3$.

7. Demonstrate whether or not the set of all polynomials of degree less than or equal to three is a subspace of the vector space of all polynomials.

8. Demonstrate whether or not the set of all polynomials that have a root at $x = 3$ is a subspace of the vector space of polynomials.

9. Demonstrate whether or not the set of all polynomials p such that $p(4) = 1$ is a subspace of the vector space of polynomials.

10. Demonstrate whether or not the set of all polynomials p such that the slope at $x = 0$ is 1 is a subspace of the vector space of polynomials.

Solutions

1. If $\mathbf{u} \in S_1$ then $\mathbf{u}$ must be of the form $\mathbf{u} = \begin{pmatrix} \alpha \\ 0 \end{pmatrix}$ for some α, and if $\mathbf{v} \in S_1$ then $\mathbf{v}$ must be of the form $\mathbf{v} = \begin{pmatrix} \beta \\ 0 \end{pmatrix}$ for some β. If we calculate $\mathbf{u} + \mathbf{v} = \begin{pmatrix} \alpha \\ 0 \end{pmatrix} + \begin{pmatrix} \beta \\ 0 \end{pmatrix} = \begin{pmatrix} \alpha + \beta \\ 0 \end{pmatrix}$, then we note that this is also of the form $\begin{pmatrix} \gamma \\ 0 \end{pmatrix}$ where $\gamma = \alpha + \beta$, so $\mathbf{u} + \mathbf{v} \in S_1$. Similarly, if $\delta \in \mathbf{R}$, then $\delta\mathbf{u} + \mathbf{v} = \delta \cdot \begin{pmatrix} \alpha \\ 0 \end{pmatrix} = \begin{pmatrix} \delta\alpha \\ 0 \end{pmatrix}$, and again, this is of the form of a vector in S_1, so $\delta\mathbf{u} \in S_1$. Therefore S_1 is a subspace of $\mathbf{R}^2$.

3. There are many ways to show that this is not a subspace of $\mathbf{R}^2$, and you could choose any:

a) It is impossible to write the zero vector in the form $\begin{pmatrix} \alpha \\ \alpha + 1 \end{pmatrix}$, so $\mathbf{0} \notin S_3$, so S_3 is not a subspace.

b) We see that $\mathbf{u} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \in S_3$ (by letting $\alpha = 0$), but $2\mathbf{u} = 2 \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$ cannot be written in the form $\begin{pmatrix} \alpha \\ \alpha + 1 \end{pmatrix}$, so $2\mathbf{u} \notin S_3$, so S_3 is not a subspace.

c) We note that $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \in S_3$ and that $\mathbf{v} = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \in S_3$, but $\mathbf{u} + \mathbf{v} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$ cannot be written in the form $\begin{pmatrix} \alpha \\ \alpha + 1 \end{pmatrix}$, so $\mathbf{u} + \mathbf{v} \notin S_3$, so S_3 is not a subspace.

5. If $\mathbf{u}_1, \mathbf{u}_2 \in S_5$, then these two vectors must be of the form $\mathbf{u}_1 = \begin{pmatrix} \alpha_1 \\ \alpha_1 + \beta_1 \\ \alpha_1 - 3\beta_1 \end{pmatrix}$ and $\mathbf{u}_2 = \begin{pmatrix} \alpha_2 \\ \alpha_2 + \beta_2 \\ \alpha_2 - 3\beta_2 \end{pmatrix}$ for some $\alpha_1, \alpha_2, \beta_1, \beta_2$. In this case, $\mathbf{u}_1 + \mathbf{u}_2 = \begin{pmatrix} \alpha_1 \\ \alpha_1 + \beta_1 \\ \alpha_1 - 3\beta_1 \end{pmatrix} + \begin{pmatrix} \alpha_2 \\ \alpha_2 + \beta_2 \\ \alpha_2 - 3\beta_2 \end{pmatrix} = \begin{pmatrix} \alpha_1 + \alpha_2 \\ \alpha_1 + \beta_1 + \alpha_2 + \beta_2 \\ \alpha_1 - 3\beta_1 + \alpha_2 - 3\beta_2 \end{pmatrix}$, and we note we can write

$$\begin{pmatrix} (\alpha_1+\alpha_2) \\ (\alpha_1+\alpha_2)+(\beta_1+\beta_2) \\ (\alpha_1+\alpha_2)-3(\beta_1+\beta_2) \end{pmatrix},$$

so this is also of the required form for vectors in S_5. Similarly, $\gamma\mathbf{u}_1 = \gamma \cdot \begin{pmatrix} \alpha_1 \\ \alpha_1+\beta_1 \\ \alpha_1-3\beta_1 \end{pmatrix} = \begin{pmatrix} \gamma\alpha_1 \\ \gamma(\alpha_1+\beta_1) \\ \gamma(\alpha_1-3\beta_1) \end{pmatrix}$ can also be written in the form $\begin{pmatrix} (\gamma\alpha_1) \\ (\gamma\alpha_1)+(\gamma\beta_1) \\ (\gamma\alpha_1)-3(\gamma\beta_1) \end{pmatrix}$, so this is also in S_5. Therefore, S_5 is a subspace of $\mathbf{R}^3$.

7. If p and q are polynomials of degree less than or equal to three, then we must be able to write $p(x) = \alpha_3 x^3 + \alpha_2 x^2 + \alpha_1 x + \alpha_0$ and $q(x) = \beta_3 x^3 + \beta_2 x^2 + \beta_1 x + \beta_0$.

In this case, $\gamma p(x) = (\gamma\alpha_3)x^3 + (\gamma\alpha_2)x^2 + (\gamma\alpha_1)x + (\gamma\alpha_0)$ is still a polynomial of degree less than or equal to three (in fact, if $\gamma = 0$, then γp is the zero polynomial, and otherwise γp must have the same degree as p.

Similarly, $(p+q)(x) = (\alpha_3+\beta_3)x^3 + (\alpha_2+\beta_2)x^2 + (\alpha_1+\beta_1)x + (\alpha_0+\beta_0)$ is also a polynomial of degree no more than three. Therefore, the set of all polynomials of degree less than or equal to three represents a subspace.

9. There are many ways to show that this is not a vector space:

a) The zero polynomial at $x = 0$ is equal to 0, not 1, so the zero polynomial is not in this set, and therefore the set is not a subspace.

b) The constant polynomial $p(x) \overset{def}{=} 1$ is a polynomial such that $p(4) = 1$, but $(2p)(4) = 2\,p(4) = 2$, so $2p$ is not in this set, and therefore this set is not a subspace.

c) The polynomials $p(x) \overset{def}{=} 1$ and $q(x) \overset{def}{=} x - 3$ both satisfy $p(4) = q(4) = 1$, but $(p+q)(x) = 1 + x - 3 = x - 2$ has the value $(p+q)(4) = 4 - 2 = 2$, so $(p+q)(4) \neq 1$ and therefore the sum is not in this set, so this set is not a subspace.

3.3 Examples of subspaces

Previously we looked at discrete signals, polynomials and function spaces. We will now consider, without proof in many cases, various subspaces of these spaces. We will, however, start with subspaces of $\mathbf{R}^3$.

Notice: Do not memorize any of these vector spaces described in this section. The point is to demonstrate that there are many different kinds of possible subspaces, significantly beyond the finite-dimensional vector spaces we have already seen. On an examination, such subspaces would be defined for you.

3.3.1 Subspaces of R^2

Without proof, the subspaces of $\mathbf{R}^2$ include

1. the zero vector $\left\{\begin{pmatrix}0\\0\end{pmatrix}\right\}$,
2. all straight lines containing the zero vector, and
3. all of $\mathbf{R}^2$ itself.

We will now look at many examples of subspaces of $\mathbf{R}^2$ that are not subspaces. Now, from the above, we already have that all scalar multiples of $\mathbf{u}=\begin{pmatrix}3\\-0.5\end{pmatrix}$ are a subspace of $\mathbf{R}^2$. One might ask, is $\mathbf{v}_1=\begin{pmatrix}17.4\\-2.8\end{pmatrix}$ or $\mathbf{v}_2=\begin{pmatrix}-19.2\\3.2\end{pmatrix}$ is a scalar multiple of $\mathbf{u}$?

For the first to be a scalar multiple of $\mathbf{u}$, there must be an α_1 such that $\mathbf{v}_1=\alpha_1\mathbf{u}$; that is, we ask, is there an α_1 such that

$$\alpha_1\begin{pmatrix}3\\-0.5\end{pmatrix}=\begin{pmatrix}3\alpha_1\\-0.5\alpha_1\end{pmatrix}=\begin{pmatrix}17.4\\-2.8\end{pmatrix}?$$

Thus, we have that

$$\begin{aligned}3\alpha_1&=17.4\\-0.5\alpha_1&=-2.8\end{aligned}$$

The first says that $\alpha_1=\dfrac{17.4}{3}=5.8$ while the second says that $\alpha_1=\dfrac{-2.8}{-0.5}=5.6$. The scalar α_1 cannot take on two different values, so thus we conclude that $\mathbf{v}_1$ is not a scalar multiple of $\mathbf{u}$.

For the second to be a scalar multiple of $\mathbf{u}$, there must be an α_2 such that $\mathbf{v}_2=\alpha_2\mathbf{u}$; we ask, is there an α_2 such that

$$\alpha_2\begin{pmatrix}3\\-0.5\end{pmatrix}=\begin{pmatrix}3\alpha_2\\-0.5\alpha_2\end{pmatrix}=\begin{pmatrix}-19.2\\3.2\end{pmatrix}?$$

Thus, we have that

$$\begin{aligned}3\alpha_2&=-19.2\\-0.5\alpha_2&=3.2\end{aligned}$$

The first says that $\alpha_2 = \frac{-19.2}{3} = -6.4$ and, similarly, the second says that $\alpha_2 = \frac{3.2}{-0.5} = -6.4$. Thus, we have that $\mathbf{v}_2 = -6.4 \cdot \mathbf{u}$.

Example 1

Consider a vector **v** and all positive scalar multiples. For example, in Figure 33, we see the set of all positive scalar multiples of the vector $\mathbf{v} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$. Is this a subspace? We note that if two vectors are positive scalar multiples of **v**, then they must be of the form $\alpha_1 \mathbf{v}$ and $\alpha_2 \mathbf{v}$ where $\alpha_1, \alpha_2 \geq 0$. In this case, $\alpha_1 \mathbf{v} + \alpha_2 \mathbf{v} = (\alpha_1 + \alpha_2)\mathbf{v}$ where $\alpha_1 + \alpha_2 \geq 0$, so the sum of the two vectors is a positive scalar of **v**. Unfortunately, this set is not closed under scalar multiplication. The vector $\mathbf{v} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ is in this set, but $(-1)\mathbf{v} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$ is not in this set, and therefore this set is not closed under scalar multiplication. Thus, this set is not a subspace of $\mathbf{R}^2$.

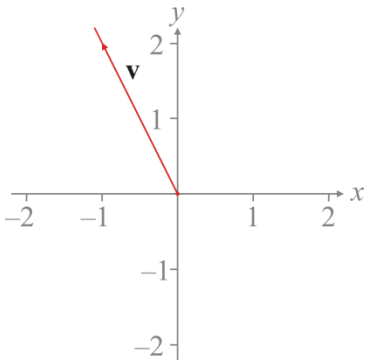


Figure 33. All positive scalar multiples of the vector $\mathbf{v} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$.

Example 2

Consider all vectors that are either a scalar multiple of $\mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ or $\mathbf{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Now, any vector in this subset is either of the form $\alpha \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ or in the form $\alpha \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. In either case, multiplying any such vector by a second scalar β will result in either the vector $\beta\left(\alpha \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = (\beta\alpha)\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ or $\beta\left(\alpha \begin{pmatrix} 1 \\ -1 \end{pmatrix}\right) = (\beta\alpha)\begin{pmatrix} 1 \\ -1 \end{pmatrix}$. In either case, the result is a vector that is either a scalar multiple of **u** or a scalar multiple of **v**. Unfortunately, this set is not a subspace: $\mathbf{u} + \mathbf{v} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ is not a scalar multiple of either **u** or **v**. Thus, this subset is not a subspace.

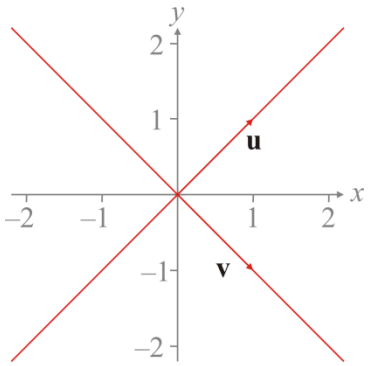


Figure 34. All vectors that are scalar multiples of $\mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ or $\mathbf{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

Example 3

Consider all vectors that are of the form $\mathbf{v} + \alpha\mathbf{u}$ where $\mathbf{v} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\mathbf{u} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$, as shown in Figure 35. In this case, it is easy to see that this set is not closed under either scalar multiplication or vector addition, for $\mathbf{v}$ is in this collection, but $2\mathbf{v} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ is not. Similarly, $\mathbf{v} + \mathbf{v} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, which is, again, not in this set. Thus, the vectors on that line are not a subspace.

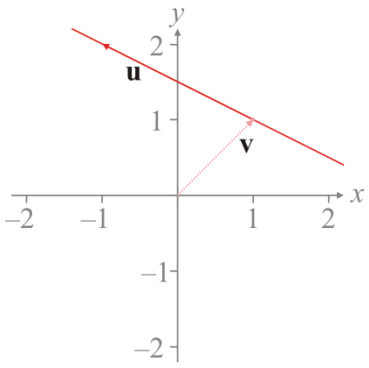


Figure 35. All vectors that are of the form $\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \alpha \begin{pmatrix} -2 \\ 1 \end{pmatrix}$.

Example 4

Next, consider all vectors in the cone defined by the vectors $\mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$, which includes all vectors in the pink region and along the boundaries, as shown in Figure 36. For example, this includes $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 7 \\ -2 \end{pmatrix}$. If we add any two vectors in this cone, the result must still be within this cone.

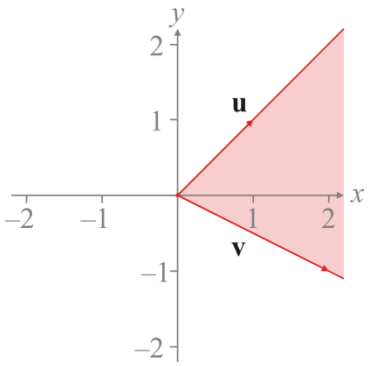


Figure 36. A cone defined by the two vectors $\mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$.

For example, if we added the vector $\begin{pmatrix} 1 \\ 0.5 \end{pmatrix}$ to any vector within this cone, we would get a vector within the blue region shown in Figure 37. Thus, this cone is closed under vector addition.

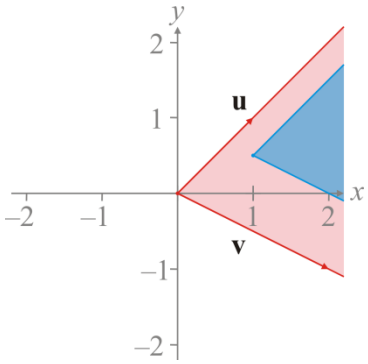


Figure 37. The vector $\begin{pmatrix} 1 \\ 0.5 \end{pmatrix}$ added to every point in the cone in Figure 36 shown using a blue region.

The cone, however, is not closed under scalar multiplication: while $\mathbf{u}$ is in the cone, $(-1)\mathbf{u} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$ is not.

Example 4

As our last example, let us reflect the cone in Figure 36 through the origin (the zero vector), as shown in Figure 38.

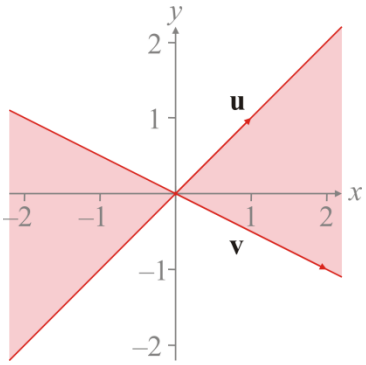


Figure 38. The cone in Figure 36 reflected through **0**.

This subset of $\mathbf{R}^2$ is now closed under scalar multiplication: if **u** is in the double cone, then so is $\alpha\mathbf{u}$. The double cone, however, is now no longer closed under vector addition: $\mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $(-1)\mathbf{v} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$ are both in the cone, but $\mathbf{u} + (-1)\mathbf{v} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ is not.

Consequently, you may deduce that in $\mathbf{R}^2$, the only subspaces that is "larger" than a line passing through the origin is all of $\mathbf{R}^2$.

3.3.2 Subspaces of $\mathbf{R}^3$

Without proof, the subspaces of $\mathbf{R}^3$ include

4. the zero vector $\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right\}$,
5. all straight lines containing the zero vector,
6. all planes containing the zero vector, and
7. all of $\mathbf{R}^3$ itself.

As you may observe, there are infinitely many different subspaces. It is necessary for the lines and planes to pass through the origin, for if a line or plane does not pass through the origin and **v** is a vector on that line or plane, then $0\mathbf{v} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ is a scalar multiple of **v**, but as the line or plane does not pass through the origin, that point is not contained within line or plane.

Problems

1. Show that a line of the form $\mathbf{u} + \alpha\mathbf{v}$ where $\alpha \in \mathbf{R}$ passes through the origin (**0**) if and only if **u** is a scalar multiple of **v**.

2. Show that if a line or plane does not pass through the origin and **v** is a vector in the line or plane, then no scalar multiple of **v** other than $1 \cdot \mathbf{v}$ lies in the line or plane, respectively.

Solutions

1. If the line $\mathbf{u}+\alpha\mathbf{v}$ passes through the origin, $\mathbf{u}+\alpha\mathbf{v}=\mathbf{0}$, in which case, $\mathbf{u}=\mathbf{0}-(\alpha\mathbf{v})=(-\alpha)\mathbf{v}$, and thus **u** is a scalar multiple of **v**.

3.3.3 Subspaces of $\mathbf{C}^2$

All subspaces of $\mathbf{C}^2$ include the trivial subspace of just the zero vectors {**0**}, the subspaces containing all complex multiples of a complex vector $\alpha\mathbf{u}$, and, of course, all of $\mathbf{C}^2$. Suppose, for example, you consider all of the scalar multiples of the vector $\mathbf{u}=\begin{pmatrix}1-j\\2+3j\end{pmatrix}$. This includes

$$2\mathbf{u}=\begin{pmatrix}2-2j\\4+6j\end{pmatrix} \text{ and } -0.5\mathbf{u}=\begin{pmatrix}-0.5+0.5j\\-1-1.5j\end{pmatrix},$$

but also contains multiples such as

$$(4+5j)\mathbf{u}=\begin{pmatrix}9+j\\-7+22j\end{pmatrix} \text{ and } (-2-0.5j)\mathbf{u}=\begin{pmatrix}-2.5+1.5j\\-2.5-7j\end{pmatrix}.$$

Suppose we wanted to determine if these two vectors

$$\mathbf{v}_1=\begin{pmatrix}-0.5+2.1j\\-5-2.3j\end{pmatrix} \text{ and } \mathbf{v}_2=\begin{pmatrix}-1.1+6.7j\\17.8+0.7j\end{pmatrix}$$

are complex scalar multiples of **u**.

For the first to be a scalar multiple of **u**, there must be an α_1 such that $\mathbf{v}_1=\alpha_1\mathbf{u}$; that is, we ask, is there an α_1 such that

$$\alpha_1\begin{pmatrix}1-j\\2+3j\end{pmatrix}=\begin{pmatrix}\alpha_1(1-j)\\\alpha_1(2+3j)\end{pmatrix}=\begin{pmatrix}-0.5+2.1j\\-5-2.3j\end{pmatrix}?$$

If this were true, we would have that

$$\alpha_1(1-j)=-0.5+2.1j$$
$$\alpha_1(2+3j)=-5-2.3j$$

The first says that $\alpha_1=\dfrac{-0.5+2.1j}{1-j}$ and thus $\alpha_1=\dfrac{-0.5+2.1j}{1-j}\cdot\dfrac{1+j}{1+j}=\dfrac{-2.6+1.6j}{2}=-1.3+0.8j$, and, similarly, the second says that $\alpha_1=\dfrac{-5-2.3j}{2+3j}=\dfrac{-5-2.3j}{2+3j}\cdot\dfrac{2-3j}{2-3j}=\dfrac{-16.9+10.4j}{13}=-1.3+0.8j$, and thus, $\mathbf{v}_1=(-1.3+0.8j)\mathbf{u}$.

For the second to be a scalar multiple of **u**, there must be an α_2 such that $\mathbf{v}_2=\alpha_2\mathbf{u}$; that is, we ask, is there an α_2 such that

$$\alpha_2\begin{pmatrix}1-j\\2+3j\end{pmatrix}=\begin{pmatrix}\alpha_2(1-j)\\\alpha_2(2+3j)\end{pmatrix}=\begin{pmatrix}-1.1-6.7j\\17.8+0.7j\end{pmatrix}?$$

Thus, we have that

$$\alpha_2(1-j) = -1.1-6.7j$$
$$\alpha_2(2+3j) = 17.8+0.7j$$

The first says that $\alpha_2 = \frac{-1.1-6.7j}{1-j}$ and thus $\alpha_2 = \frac{-1.1-6.7j}{1-j}\cdot\frac{1+j}{1+j} = \frac{5.6-7.8j}{2} = 2.8-3.9j$, yet the second says that $\alpha_2 = \frac{17.8+0.7j}{2+3j} = \frac{17.8+0.7j}{2+3j}\cdot\frac{2-3j}{2-3j} = \frac{37.7-52j}{13} = 2.9-4j$. Therefore, $\mathbf{v}_2$ cannot be a scalar multiple of **u**.

3.3.4 Subspaces of the vector space of discrete signals

We will now look at subspaces of real and then complex discrete signals. These will include discrete signals that are

1. bounded,
2. square summable, and
3. eventually zero.

We will show that each of these is a subspace of the vector space of discrete signals.

3.3.4.1 Bounded discrete signals

A discrete signal x is said to be *bounded* if $|x[k]| \le M$ for some value M. For example, discrete signal x defined by $x[k] = 2^{-k}$ is *bounded*, as $|x[k]| \le 1$, but the discrete signal y defined by $x[k] = 2^k$ is *unbounded*, as its values grow infinitely large. If we consider the collection of all bounded discrete signals, we note that

1. if we multiply a discrete signal x bounded by M, then αx will be bounded by $|\alpha|M$, and
2. if we add two discrete signals x and y bounded by M_x and M_y, respectively, the sum $x + y$ must be bounded by $M_x + M_y$.

Thus, the collection of all bounded discrete signals forms a subspace.

3.3.4.2 Square-summable discrete signals

A discrete signal $x = (x[0], x[1], x[2], \ldots)$ is said to be square-summable if the sum of the squares of the entries is finite, or

$$\sum_{k=0}^{\infty} |x[k]|^2 < \infty \ .$$

Note that every square-summable discrete signal must therefore be bounded, but not every bounded discrete signal is square summable; consider, for example, the bounded discrete signal x where $x[k] = 1$ for each $k = 0, 1, \ldots$.

Given a square-summable discrete signal x, it follows that $\sum_{k=0}^{\infty} |x[k]|^2 = M$ for some value of M. It follows therefore that

1. if we multiply a square-summable discrete signal x with square sum M, then αx will sum to $|\alpha|^2 M$, and
2. if we add two square-summable discrete signals x and y with square sums M_x and M_y, respectively, the sum $x + y$ must be bounded by $(M_x + M_y)^2$.

Such discrete signals are also said to have *finite energy*.

3.3.4.3 Discrete signals that are eventually 0

A discrete signal $x = (x[0], x[1], x[2], \ldots)$ is said to be *eventually 0* if there exists an integer M such that $x[k] = 0$ for $k = M, M+1, M+2, \ldots$. Again, every discrete signal that is eventually zero must be square summable, but not all square summable discrete signal are eventually zero.

3.3.5 Subspaces of the vector space of polynomials

We have already considered the vector space of all polynomials with real coefficients, $\mathbf{P}(\mathbf{R})$. Next, we will consider the collection of all polynomials of degree less than or equal to n. We will represent this by $\mathbf{P}_n(\mathbf{R})$, and as you may suspect, each of these, too, is itself a vector space:

1. if you multiply a polynomial of degree n by α, the scalar multiple is still a polynomial of degree n unless $\alpha = 0$, in which case, the scalar multiple is the zero polynomial (of degree 0), and
2. if you add two polynomials of degree m and n, the sum must be a polynomial of degree $\max(m, n)$ if $m \neq n$ and the sum when both polynomials have the same degree $m = n$ must be of degree less-than-or-equal-to m.

The set $\mathbf{P}_0(\mathbf{R})$ is the vector space of all constant functions, the set $\mathbf{P}_1(\mathbf{R})$ is the vector space of all affine functions, and so on. Note that

$$\mathbf{P}_0(\mathbf{R}) \subset \mathbf{P}_1(\mathbf{R}) \subset \mathbf{P}_2(\mathbf{R}) \subset \cdots \subset \mathbf{P}(\mathbf{R}),$$

and each of these is itself its own distinct vector space.

Questions:

1. Is the collection of all polynomials such that $p(0) = 0$ a vector space?
2. Is the collection of all polynomials such that $p(0) = 1$ a vector space?

MATLAB makes use of this correspondence between vectors and polynomials by allowing you to define polynomials using vectors. For example, the vector

```
>> p = [3 -2.1 5.48 -2.76]    % this can be either a row or column vector
    p =
        3.0000   -2.1000    5.4800   -2.7600
```

defines the polynomial

$$3x^2 - 2.1x^2 + 5.48 - 2.76$$

You can tell MATLAB to interpret a vector as a polynomial and evaluate it at a point using the `polyval` routine:

```
>> polyval( p, 0.6 )    % evaluate the polynomial at the point 0.6
    ans =
        0.4200
```

3.3.6 Subspaces of the vector space of functions of a single variable (analog signals)

We will now consider other vector spaces of functions, all of which will be useful to you in future courses, including the vector spaces of

1. all continuous functions defined on *D*,
2. all infinitely differentiable functions defined on *D*,
3. all bounded functions defined on *D*,
4. all functions *f* defined on **R** such that $\lim_{t\to\infty} f(t) = 0$ and $\lim_{-\infty\leftarrow t} f(t) = 0$.

3.3.6.1 *Continuous functions defined on D*

A function is *continuous* if (as a first approximation), it its graph can be drawn without lifting the pencil from the paper while it is being drawn. You will learn a rigorous definition of a continuous function in your calculus course, but it is sufficient to say that if you add two continuous functions together, the result is continuous, and if you multiply a continuous function by a scalar, the result is still continuous. The collection of all continuous functions is often written as $C^0(D)$.

3.3.6.2 *Infinitely differentiable functions defined on D*

A function is infinitely differentiable if it is continuous, and every derivative of the function is also continuous. This collection of functions is often written as $C^\infty(D)$. Notice that this includes the class of all polynomials, but also includes the trigonometric functions without asymptotes (including sine and cosine), the exponential functions, and the hyperbolic functions without asymptotes (including sinh, cosh, tanh and sech). This does not include the absolute value function, as the function is itself continuous, but its derivative is not defined at $t = 0$.

3.3.6.3 *Bounded functions defined on D*

A function f is bounded if there exists a real $M \geq 0$ such that $|f(t)| \leq M$ for all t in the domain. Clearly, the zero function is bounded. It also follows that $|(-f)(t)| = |-f(t)| = |f(t)| \leq M$, $|(\alpha f)(t)| = |\alpha f(t)| = |\alpha||f(t)| \leq |\alpha| M$ and if $|f(t)| \leq M_f$ and $|g(t)| \leq M_g$, by the triangle inequality, it follows that

$$|(f+g)(t)| = |f(t)+g(t)| \leq |f(t)| + |g(t)| \leq M_f + M_g .$$

3.3.6.4 *Functions on zero limits at ±∞*

In your course on quantum mechanics, you will look at differentiable functions defined on **R** with the addition property that in the limit, the function goes to 0 as the argument goes to ±∞. As you may suspect, if you multiply such a function by a scalar, it is still differentiable and the limits at ±∞ are still zero, and if you add two such functions together, they continue to have the same properties. Consequently, the collection of all such functions defines a vector space.

3.3.6.5 *Complex-valued function spaces*

For each function space we have looked at in this section, we can also consider the complex-valued functions. Again, in each case, if we define the addition of functions and the scalar multiplication of functions in the intuitive manner, the resulting space is a vector space.

3.4 Summary of subspaces

This chapter has described how to determine when a subset of a vector space is itself a vector space in its own right. Given a vector space *V* and a subset *S*, to determine if *S* is itself a vector space, we need only determine that if $\mathbf{u} \in S$, then $\alpha\mathbf{u} \in S$ for all possible scalar values α, and if $\mathbf{u}, \mathbf{v} \in S$, then $\mathbf{u} + \mathbf{v} \in S$.

4 Normed vector spaces

One question that one might ask is *how long is a vector*? After, all, one thinks of a vector as an offset from the origin, and for two- and three-dimensional vectors, we can use the Euclidean norm. There are, however, other norms that are equally useful, and we can also define norms on discrete signals and functions. In this topic, we will describe the 2-norm for finite-dimensional vectors, which you may have previously called the Euclidean norm or Euclidean length of a vector. We will then look at two other norms of finite-dimensional vectors, define *unit vectors* and the process of normalization, as well as some applications of norms.

4.1 The 2-norm for finite-dimensional vectors

As we have already described vectors to be points on a plane or points in space, one might ask how we can measure the length of such a vector. The length is usually measured by what is commonly referred to as the *Euclidean length* or the *Euclidean norm*. For example, the Euclidean norm of the two-dimensional vector $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ is $\|\mathbf{u}\|_{\text{Euclidean}} = \sqrt{u_1^2 + u_2^2}$ and the Euclidean norm of the three-dimensional vector $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$ is $\|\mathbf{v}\|_{\text{Euclidean}} = \sqrt{v_1^2 + v_2^2 + v_3^2}$. Notice that the absolute value of a scalar x is $|x|$, the norm of a vector is $\|\mathbf{u}\|$. For convenience, however, instead of using the term *Euclidean norm*, we will refer to this norm as the *2-norm* and the generalization to n-dimensional vectors is straight-forward: if $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$, we will define

$$\|\mathbf{u}\|_2 = \sqrt{\sum_{k=1}^{n} |u_k|^2}.$$

In MATLAB, the 2-norm of a vector is computed using the `norm` routine. If the `norm` routine is called with only a single argument, that being a vector, it automatically computes the 2-norm; however, we will see that it is possible to define different types of norms, and therefore you can explicitly state that you wish to compute the 2-norm by passing a second argument `2` to the `norm` routine.

```
>> format long
>> v = [1 -2 3]';
>> norm( v )
    ans =
       3.741657386773941

>> norm( v, 2 )
    ans =
       3.741657386773941

>> sqrt( 1^2 + (-2)^2 + 3^2 )
    ans =
       3.741657386773941
```

We will also define the distance between two vectors **u** and **v** as

$$\text{dist}(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2.$$

You will note that the 2-norm has a number of properties:

1. $\|\mathbf{u}\|_2 \geq 0$ and $\|\mathbf{u}\|_2 = 0$ if and only if $\mathbf{u} = \mathbf{0}_n$,
2. $\|\alpha\mathbf{u}\|_2 = |\alpha|\|\mathbf{u}\|_2$, and
3. $\|\mathbf{u}+\mathbf{v}\|_2 \leq \|\mathbf{u}\|_2 + \|\mathbf{v}\|_2$.

We will describe each of these properties.

4.1.1 Non-negativity and point separation

The first property essentially says, "the length of a vector is always positive, and the length of a vector is zero if and only if the vector is the zero vector." Consequently, if we ever deduce that $\|\mathbf{u}\|_2 = 0$, it immediately follows that $\mathbf{u} = \mathbf{0}$. The phrase *point separation* says that $\mathbf{u} = \mathbf{v}$ if and only if $\text{dist}(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2 = 0$, and consequently, if the distance between two vectors is zero, they are the same vector; and if the distance is non-zero, the vectors must be different.

That this is the case for the 2-norm may be deduced as follows: If $\|\mathbf{u}\|_2 \neq 0$, then there exists an index *i* such that $u_i \neq 0$, and therefore

$$\|\mathbf{u}\|_2^2 = \sum_{k=1}^{n} |u_k|^2 \geq |u_i|^2 > 0,$$

consequently, as the square root function is monotonic,

$$\|\mathbf{u}\|_2 = \sqrt{\sum_{k=1}^{n} |u_k|^2} \geq |u_i| > 0.$$

4.1.2 Absolute scalability

The second property is true, as

$$\begin{aligned}
\|\alpha\mathbf{u}\|_2 &= \sqrt{\sum_{k=1}^{n} |\alpha u_k|^2} \\
&= \sqrt{\sum_{k=1}^{n} |\alpha|^2 |u_k|^2} \\
&= \sqrt{|\alpha|^2 \sum_{k=1}^{n} |u_k|^2} \\
&= \sqrt{|\alpha|^2} \sqrt{\sum_{k=1}^{n} |u_k|^2} \\
&= |\alpha| \|\mathbf{u}\|_2
\end{aligned}$$

4.1.3 Triangle inequality for vectors

You will recall the triangle inequality for real or complex numbers: $|w + z|$ $|w+z| \le |w|+|z|$. The same property holds for the length of vectors, and thus has the same name. What this means can be much more easily illustrated in an image: the distance from Waterloo to Guelph is always less-than-or-equal-to the distance from Waterloo to, say, Cambridge plus the distance from Cambridge to Guelph.



Similarly, given two vectors **u** and **v**, $\|\mathbf{u}+\mathbf{v}\|_2 \le \|\mathbf{u}\|_2 + \|\mathbf{v}\|_2$. The proof is somewhat intricate, but essentially has you showing that $\|\mathbf{u}+\mathbf{v}\|_2^2 \le (\|\mathbf{u}\|_2 + \|\mathbf{v}\|_2)^2$, so

$$\|\mathbf{u}+\mathbf{v}\|_2^2 = (u_1+v_1)^2 + \cdots + (u_n+v_n)^2 = u_1^2 + 2u_1v_1 + v_1^2 + \cdots + u_n^2 + 2u_nv_n + v_n^2 = \|\mathbf{v}\|_2^2 + 2u_1v_1 + \cdots + 2u_nv_n + \|\mathbf{v}\|_2^2$$

while

$$(\|\mathbf{u}\|_2 + \|\mathbf{v}\|_2)^2 = \|\mathbf{u}\|_2^2 + 2\|\mathbf{u}\|_2\|\mathbf{v}\|_2 + \|\mathbf{v}\|_2^2 .$$

Thus, it only necessary to show that

$$2u_1v_1 + \cdots + 2u_nv_n \le 2\|\mathbf{u}\|_2\|\mathbf{v}\|_2$$

or

$$u_1v_1 + \cdots + u_nv_n \le \|\mathbf{u}\|_2\|\mathbf{v}\|_2 .$$

We will prove that this is true in a future topic on the Cauchy-Bunyakovsky-Schwarz inequality which we will see in Section 5.8.

Problems

1. Give an example of two vectors such that $\|\mathbf{u} + \mathbf{v}\|_2 = \|\mathbf{u}\|_2 + \|\mathbf{v}\|_2$.

2. Give an example of two vectors such that $\|\mathbf{u} + \mathbf{v}\|_2 = 0$.

Solutions

1. If **u** = **v**, then $\|\mathbf{u} + \mathbf{u}\|_2 = \|2\mathbf{u}\|_2 = 2\|\mathbf{v}\|_2 = \|\mathbf{u}\|_2 + \|\mathbf{u}\|_2$.

4.2 Other norms for finite-dimensional vectors

In the next section, we will see why the 2-norm is so important, as it is intimately connected to the dot product[6] or, as we will call it, the inner product. There are, however, other ways of measuring the length of a vector, and we will look at two of them:

1. the 1-norm or "Manhattan norm", and
2. the infinity-norm.

Both have applications in different areas of engineering.

4.2.1 The 1-norm

Suppose you are in downtown Manhattan, and a friend tells you it is 267 m from the door of St. Patrick's Cathedral to the Museum of Modern Art.



Looking at a map, you quickly realize that the actual distance is appears to be approximately $\sqrt{2}$ longer, at 382 m, at which point, you ask your friend: "Do I look more like a crow or a fox?"[7] Of course, as the crow is a dinosaur and foxes and humans are not, the answer is rather obvious, and so we need a different means of measuring the length of vectors in Manhattan. The 1-norm of a vector is defined as

$$\|\mathbf{u}\|_1 = \sum_{k=1}^{n} |u_k| ,$$

and this measures the length of the vector between two points in Manhattan as shown in .



[6] You should have seen the dot product in secondary school.

[7] "Five-and-forty leagues as the crow flies we have come, though many long miles further our feet have walked." from J.R.R. Tolkien's *The Lord of the Rings*.

To calculate the 1-norm, one need simply pass a "1" as a second argument to the norm function.

```
>> format long
>> v = [1 -2 3]';
>> norm( v, 1 )
    ans =
       6
```

If you consider the three properties of the 2-norm, you will see that all three properties are still satisfied by this norm:

1. $\|\mathbf{u}\|_1 \geq 0$ and $\|\mathbf{u}\|_1 = 0$ if and only if $\mathbf{u} = \mathbf{0}_n$,
2. $\|\alpha\mathbf{u}\|_1 = |\alpha|\|\mathbf{u}\|_1$, and
3. $\|\mathbf{u}+\mathbf{v}\|_1 \leq \|\mathbf{u}\|_1 + \|\mathbf{u}\|_1$.

Next, we will look at the infinity norm.

4.2.2 The infinity-norm

Consider a 3D-printer, as shown in Figure 39.

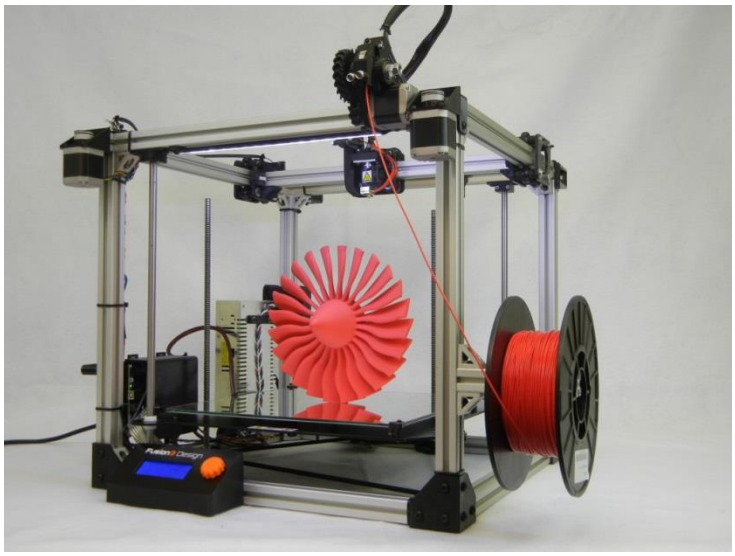
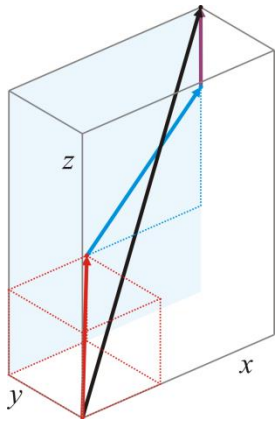


Figure 39. The Fusion 3 3D-printer from www.3dprint.com.

The head of the printer is moved in each of three dimensions by one of three motors, and assume each motor moves the head at the same rate. Suppose a vector defines the change in position of the head of the printer. Each motor will turn on only as long as it needs to. For example, consider the vector in .



Here, initially, all three motors turn on until we get to point *A*, after which the *y*-motor turns off. The *x*- and *z*-motors continue to move the head to point *B*, after which, the *x*-motor turns off. In this case, the time it takes the head of the 3d-printer to move to the new location depends entirely on the largest entry of the vector. Thus, we may define the *infinity*-norm:

$$\|\mathbf{u}\|_\infty = \max_{1\le k\le n}\{|u_k|\}.$$

To calculate the infinity-norm, one need simply pass a "Inf" as a second argument to the norm function.

```
>> format long
>> v = [1 -2 3]';
>> norm( v, Inf )
    ans =
        3
```

Again, it is left as an exercise to see that the infinity norm satisfies all three properties.

As an aside, you may wonder why it is not call the maximum-norm. We can define a more general norm called a *p*-norm, which is defined as

$$\|\mathbf{u}\|_p = \sqrt[p]{\sum_{k=1}^{n}|u_k|^p}.$$

If you substitute $p = 1$ into this formula, you get the 1-norm, and if you substitute $p = 2$, you get the 2-norm. If you begin substituting larger and larger values of *p* into this formula, you will find that the norms approach the maximum value, and so in the limit,

4.2.3 Summary of other norms

The 1- and infinity-norms are used in many applications in engineering, as has been suggested. When you are referring to a norm, it is critical to specify which norm you are using. While the 2-norm is the most common, it is not the exclusive tool of the engineer.

4.3 Unit vectors and normalization of vectors

For a given norm, if the norm of a vector is 1, we call that vector a *unit vector*. For 2-dimensional real-valued vectors, all unit vectors for the 2-norm are of the form

$$\hat{\mathbf{u}} = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}.$$

If we know that a vector is a unit vector (because it is defined as such), we will denote it by a *cap* over the symbol; e.g., $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$. Given a non-zero vector **u**, we can define its associated *normalized unit vector* as

$$\hat{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|}.$$

Of course, the unit vector may change depending on which norm you choose. For the most part in this course, we will use the 2-norm.

For example, if $\mathbf{u} = \begin{pmatrix} 3 \\ -4 \end{pmatrix}$, then

	$\|\mathbf{u}\|$	**Normalized vector**
1-norm	7	$\hat{\mathbf{u}} = \begin{pmatrix} \frac{3}{7} \\ -\frac{4}{7} \end{pmatrix}$
2-norm	5	$\hat{\mathbf{u}} = \begin{pmatrix} \frac{3}{5} \\ -\frac{4}{5} \end{pmatrix}$
infinity-norm	4	$\hat{\mathbf{u}} = \begin{pmatrix} \frac{3}{4} \\ -1 \end{pmatrix}$

Similarly, if $\mathbf{u} = \begin{pmatrix} 3.1 \\ -4.2 \\ 1.7 \end{pmatrix}$,

	$\|\mathbf{u}\|$	**Normalized vector**
1-norm	9	$\hat{\mathbf{u}} = \begin{pmatrix} 0.3\overline{4} \\ -0.6\overline{6} \\ 0.1\overline{8} \end{pmatrix}$
2-norm	$\sqrt{30.14}$	$\hat{\mathbf{u}} \approx \begin{pmatrix} 0.564663960412256 \\ -0.765028591526282 \\ 0.309654429903495 \end{pmatrix}$

infinity-norm	4.2	$\hat{\mathbf{u}} = \begin{pmatrix} 0.3\overline{4} \\ -0.6\overline{6} \\ 0.1\overline{8} \end{pmatrix}$

Note that we use the approximately-equal-to symbol $\approx$, as the 2-norm is not exact, as

$$\sqrt{30.14} \approx 5.489990892524321847611280754754801593383877751590197408993271528600\cdots.$$

The only vector that cannot be normalized is the zero vector, as $\frac{0}{0}$ is undefined.

In MATLAB, normalizing a vector is straight-forward:

```
>> u = [2 5 -2 -4]'
    u =
         2
         5
        -2
        -4

>> u/norm(u)
    ans =
       0.285714285714286
       0.714285714285714
      -0.285714285714286
      -0.571428571428571
```

In MATLAB, if you attempt to divide `0.0` by `0.0`, you get a special number displayed as `NaN`, a placeholder for "not a number". This differs from integer arithmetic, where attempting to divide an integer `0` by `0` results in a software interrupt that terminates the execution of a program.

```
>> 1/0
    ans =
       Inf
>> -1/0
    ans =
      -Inf
>> 0/0
    ans =
       NaN
```

This gives us our first opportunity to write a MATLAB routine: write a function that takes a vector as an argument and returns that vector normalized:

```
function u_hat = normalize( u )
    norm_u = norm( u );   % norm( u ) == norm( u, 2 )

    if norm_u == 0
        u_hat = u;
    else
        u_hat = u/norm_u;
    end
end
```

Problems

1. What are the 2-norms of the vectors $\mathbf{u}_1 = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$, $\mathbf{u}_2 = \begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix}$ and $\mathbf{u}_3 = \begin{pmatrix} 2 \\ -3 \\ -1 \\ 5 \end{pmatrix}$?

2. What are the 2-norms of $\mathbf{v}_1 = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$

3. Demonstrate that if $-\mathbf{u}$ is the additive inverse of $\mathbf{u}$, then $\|-\mathbf{u}\|_2 = \|\mathbf{u}\|_2$.

4. Is $\|\mathbf{v}-\mathbf{u}\|_2 = \|\mathbf{u}-\mathbf{u}\|_2$?

Solutions

1. The norms are $\|\mathbf{u}_1\|_2 = \sqrt{29}$, $\|\mathbf{u}_2\|_2 = \sqrt{14}$ and $\|\mathbf{u}_3\|_2 = \sqrt{39}$.

3. $\|-\mathbf{u}\|_2 = \|(-1)\mathbf{u}\|_2 = |-1|\|\mathbf{u}\|_2 = \|\mathbf{u}\|_2$

4.4 Norms for other vector spaces

Previously, we have defined other vector spaces, and now we will see that there are perfectly useful norms that can be defined on those vector spaces, at least, under the right conditions.

4.4.1 Normed vector space of discrete signals

The norm of a digital signal, unfortunately, may be infinite, and consequently, we cannot simply define the 2-norm of a digital signal *y* as the infinite sum

$$\|y\|_2 = \sqrt{\sum_{k=0}^{\infty} |y[k]|^2},$$

as there are signals that have an infinite sum; for example, $y = (1,1,1,\ldots)$. Consequently, we must restrict ourselves to one of two types of signals.

1. The normed vector space of *finite-energy discrete* signals, where
$$\|y\|_2 = \sqrt{\sum_{k=0}^{\infty} |y[k]|^2} < \infty.$$
2. The normed vector space of *finite-power discrete signals*, where the signal is periodic, where
$$y[k+T] = y[k]$$
for some integer $T > 1$ and thus we may define
$$\|y\|_2 = \frac{1}{T}\sqrt{\sum_{k=0}^{T-1} |y[k]|^2},$$
which must be finite.

The reason we can call these vector spaces, for if *x* and *y* are both

1. finite-energy discrete signals, or
2. finite-power discrete signals

then so is $\alpha x + \beta y$.

Also, these two vector spaces intersect only at the zero discrete signal. Both of these spaces play a significant role in signal processing.

We may also define alternative norms:

1. The 1-norm of a discrete signal is defined as $\|y\|_1 = \sum_{k=0}^{\infty} |y[k]|$, while
2. the infinity norm of a discrete signal is defined as $\|y\|_\infty = \sup_{k \geq 0} \{|y[k]|\}$.

Here, sup indicates the *supremum* as opposed to *maximum* as a discrete signal may grow toward, but never achieve, a specific value. For example, consider the following:

1. The discrete signal $y[k] \stackrel{def}{=} \sin(k)$ is such that there is no *k* such that $|\sin(k)| = 1$ (because, if this was true, then π would be rational), but it approaches 1 arbitrarily closely as, $\sin(699) \approx 0.99999047$ and $\sin(573204) \approx 0.99999999999995681$.
2. Similarly, $y[k] \stackrel{def}{=} \tanh(k)$ approaches 1, but never equals 1, as $\tanh(11514) \approx 0.9\underbrace{\cdots}_{10000\ 9s}97668$ and $\tanh(115130) \approx 0.\underbrace{9\cdots}_{100000\ 9s}95496$.

Similarly, we may define the vector space those discrete signals where $\|y\|_1 < \infty$ and those where $\|y\|_\infty < \infty$. Note that every vector with $\|y\|_1 < \infty$ has $\|y\|_2 < \infty$ (but not vice versa; consider *y* = (1/2, 1/3, 1/4, 1/5, 1/6, …)) and every vector with $\|y\|_2 < \infty$ also has $\|y\|_\infty < \infty$ (but, again, not vice versa; consider *y* = (1, 1, 1, 1, 1, …)).

Problems

1. Find the 1-, 2- and infinity-norms of the discrete signals

$$x = \left(1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \ldots\right) \text{ and } y = \left(1, -\frac{1}{3}, \frac{1}{9}, -\frac{1}{27}, \frac{1}{81}, \ldots\right).$$

2. Find the 1-, 2- and infinity-norms of the discrete signals

$$x = \left(-12, -6, -3, -\frac{3}{2}, -\frac{3}{4}, \ldots\right) \text{ and } y = \left(-6, 2, -\frac{2}{3}, \frac{2}{9}, -\frac{2}{27}, \ldots\right).$$

3. Find the 1-, 2- and infinity-norms of the discrete signals

$$x = \left(1, \frac{1}{2}j, -\frac{1}{4}, -\frac{1}{8}j, \frac{1}{16}, \ldots\right) \text{ and } y = \left(1, \frac{1}{3}\angle -45°, \frac{1}{9}\angle -90°, \frac{1}{27}\angle -135°, \frac{1}{81}\angle 180°, \ldots\right).$$

4. Find the 1-, 2- and infinity-norms of the discrete signals

$$x = \left(-12, -6, -3, -\frac{3}{2}, -\frac{3}{4}, \ldots\right) \text{ and } y = \left(-6, 2, -\frac{2}{3}, \frac{2}{9}, -\frac{2}{27}, \ldots\right).$$

Solutions

1. These two discrete signals are $\frac{1}{2^n} = \left(\frac{1}{2}\right)^n$ and $\frac{1}{(-3)^n} = \left(-\frac{1}{3}\right)^n$, and thus we may calculate the norms as

$$\|x\|_1 = \sum_{k=0}^{\infty}\left|\frac{1}{2^k}\right| = \sum_{k=0}^{\infty}\left(\frac{1}{2}\right)^k = \frac{1}{1-\frac{1}{2}} = 2 \text{ and } \|y\|_1 = \sum_{k=0}^{\infty}\left|\frac{1}{(-3)^k}\right| = \sum_{k=0}^{\infty}\left(\frac{1}{3}\right)^k = \frac{1}{1-\frac{1}{3}} = \frac{3}{2},$$

and

$$\|x\|_2^2 = \sum_{k=0}^{\infty}\left(\frac{1}{2^k}\right)^2 = \sum_{k=0}^{\infty}\left(\left(\frac{1}{2}\right)^k\right)^2 = \sum_{k=0}^{\infty}\left(\left(\frac{1}{2}\right)^2\right)^k = \sum_{k=0}^{\infty}\left(\frac{1}{4}\right)^k = \frac{1}{1-\frac{1}{4}} = \frac{4}{3} \text{ and}$$

$$\|y\|_2^2 = \sum_{k=0}^{\infty}\left(\frac{1}{(-3)^k}\right)^2 = \sum_{k=0}^{\infty}\left(\left(-\frac{1}{3}\right)^k\right)^2 = \sum_{k=0}^{\infty}\left(\left(-\frac{1}{3}\right)^2\right)^k = \sum_{k=0}^{\infty}\left(\frac{1}{9}\right)^k = \frac{1}{1-\frac{1}{9}} = \frac{9}{8}$$

so $\|x\|_2 = \frac{2}{\sqrt{3}}$ and $\|y\|_2 = \frac{3}{2\sqrt{2}}$, and in both cases, the maximum entry in absolute value is 1, so $\|x\|_\infty = \|y\|_\infty = 1$.

3. These two discrete signals have $\left|\frac{j}{2}\right| = \frac{1}{2}$ and $\left|\frac{1}{3}\angle -45°\right| = \frac{1}{3}$, and thus the norms are equal to the norms of Question 1.

4.4.2 Norms of polynomials

Given the polynomial $p(x) = ax^2 + bx + c$, you could argue that possible norms include $\|p\|_2 = \sqrt{a^2+b^2+c^2}$ or $\|p\|_1 = |a|+|b|+|c|$ or $\|p\|_\infty = \max\{|a|,|b|,|c|\}$, and while these satisfy the properties of a norm, they have absolutely no significance—they do not convey any useful information about the polynomial, and therefore defining them is essentially useless. There are norms for polynomials, but they are more esoteric and they only have applications in very specialized fields. These are:

$$\|p\|_1 = \sqrt{\int_0^1 \left|p\left(e^{j2\pi t}\right)\right| dt}\,,\ \|p\|_2 = \sqrt{\int_0^1 \left|p\left(e^{j2\pi t}\right)\right|^2 dt} \text{ and } \|p\|_\infty = \max_{0\le t\le 1}\left\{\left|p\left(e^{j2\pi t}\right)\right|\right\}.$$

Essentially, these either integrate over or find the maximum on the unit complex circle. These will never be used in this course; however, they do satisfy all the properties of a norm and are intended to demonstrate the various non-obvious definitions of norms that may be applied.

4.4.3 Normed vector space of functions of a real variable

Similarly, we could define the norm of a function as

$$\|f\|_2 = \sqrt{\int_{-\infty}^{\infty} |f(x)|^2\, dx}\,,$$

but as before, we must deal with the fact that some functions have an infinite area under the square of the absolute value. Consequently, as before, we must restrict ourselves to those signals with finite area.

1. The normed vector space of *finite-energy functions of a real variable*, where
$$\|f\|_2 = \sqrt{\int_{-\infty}^{\infty} |f(x)|^2\, dx} < \infty\,.$$
2. The normed vector space of *finite-power functions of a real variable*, where the signal is periodic, where
$$f(t+T) = f(T)$$
for $T > 0$ and thus we may define

$$\|f\|_2 = \frac{1}{T}\sqrt{\int_0^T |f(x)|^2\,dx},$$

which must be finite.

The reason we can call these vector spaces, for if *f* and *g* are both

1. finite-energy discrete signals, or
2. finite-power discrete signals

then so is $\alpha f + \beta g$. As we have already done before, we can also define additional norms on the vector space of functions of a real variable, including

$$\|f\|_1 = \int_{-\infty}^{\infty} |f(x)|\,dx \text{ and } \|f\|_\infty = \sup_{-\infty<x<\infty} \{|f(x)|\}.$$

In quantum mechanics, they are especially interested in functions with unit area, $\|f\|_2 = 1$; however, these do not form a vector space.

It is also possible to restrict oneself to functions defined on a specific interval, [*a*, *b*]?

1. The area under the absolute value of the function: $\|f\|_1 = \int_a^b |f(x)|\,dx$.
2. The square root of the area under the square of the function: $\|f\|_2 = \sqrt{\int_a^b |f(x)|^2\,dx}$

   The maximum the absolute value of the function appears to achieve: $\|f\|_\infty = \sup_{a\le x\le b} |f(x)|$.

As you may

4.5 Applications of norms

The most obvious application of a norm to engineering is determining distance: what is the distance between two vectors. This information can be used to determine the closest point to another.

4.5.1 Root-mean square of a vector and the standard deviation

Engineers often use a norm related to the 2-norm known as the root mean-square of a vector. This is defined as the square root of the average value of the squares of the entries of the vectors, or

$$\operatorname{rms}(\mathbf{v}) \overset{\text{def}}{=} \sqrt{\frac{1}{n}\sum_{k=1}^{n}|v_k|^2} = \frac{\|\mathbf{v}\|_2}{\sqrt{n}}.$$

If you check, you will see that this function has all the properties of a norm:

1. $\operatorname{rms}(\mathbf{v}) \geq 0$ and $\operatorname{rms}(\mathbf{v}) = 0$ if and only if $\mathbf{v} = \mathbf{0}$,
2. $\operatorname{rms}(\alpha\mathbf{v}) = \frac{\|\alpha\mathbf{v}\|_2}{\sqrt{n}} = |\alpha|\frac{\|\mathbf{v}\|_2}{\sqrt{n}} = |\alpha|\operatorname{rms}(\mathbf{v})$, and
3. $\operatorname{rms}(\mathbf{u}+\mathbf{v}) = \frac{\|\mathbf{u}+\mathbf{v}\|_2}{\sqrt{n}} \leq \frac{\|\mathbf{u}\|_2 + \|\mathbf{v}\|_2}{\sqrt{n}} = \frac{\|\mathbf{u}\|_2}{\sqrt{n}} + \frac{\|\mathbf{v}\|_2}{\sqrt{n}} = \operatorname{rms}(\mathbf{u}) + \operatorname{rms}(\mathbf{v})$.

Note that the RMS of a constant-valued vector is the absolute value of that constant. For example,

$$\operatorname{rms}(c\mathbf{1}_n) \overset{\text{def}}{=} \sqrt{\frac{1}{n}\sum_{k=1}^{n}|c|^2} = \sqrt{\frac{1}{n}n|c|^2} = |c|.$$

We will define the *standard deviation* std(**v**) of a vector **v** to be that value such that

$$\operatorname{rms}^2(\mathbf{v}) = \operatorname{mean}^2(\mathbf{v}) + \operatorname{std}^2(\mathbf{v}).$$

For a constant-valued vector, the standard deviation is 0.

4.5.2 Digital sound, data compression and the root-mean-squared error

Sound is the result of vibrations in the air, and this pressure is recorded as being relative to the current air pressure. The Compact Disc digital audio format allows this pressure to be sampled 44 100 times per second, or at 44.1 kHz. For example, a digitized sampling of a cello playing an A above middle C (440 Hz) at 44.1 kHz may appear as is shown in Figure 40.

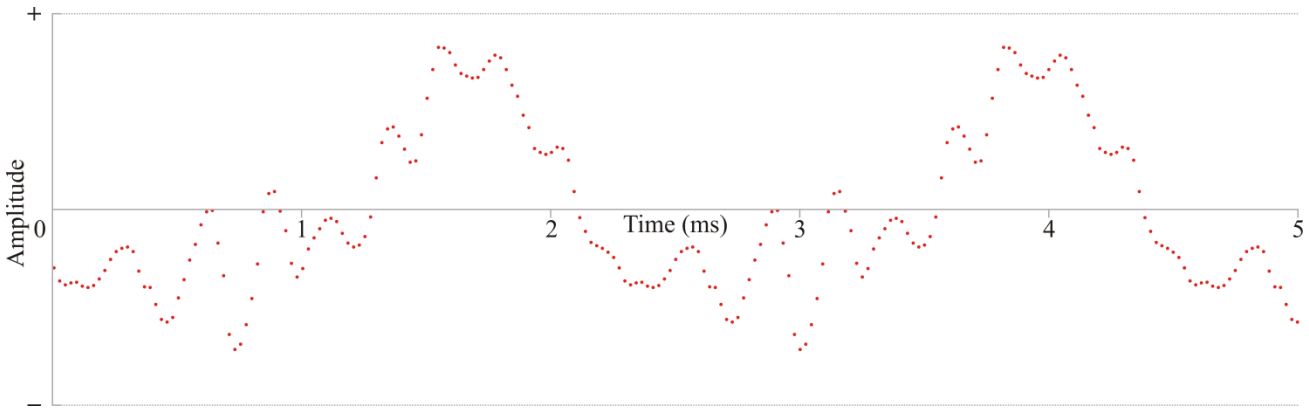


Figure 40. A digitized recording of a cello playing a 440 Hz A.[8]

Each sample is converted into 16 bits (or two bytes), and therefore a one-minute sample would occupy $44100 \cdot 2 \cdot 60 = 5292000$ bytes or approximately 5.05 MiB. A lossy compression format like MP3 attempts to store this information using significantly less than the full 5.05 MiB, but it is in general impossible to perfectly reconstruct the original data—there will be errors. Both the original bytes and the reconstructed bytes can be represented as vectors, and we can compare them using vector norms. For example, a one-minute recording would be a 2.646-million-dimensional vector, so the actual and reconstructed versions may be represented by $\mathbf{v}$ and $\tilde{\mathbf{v}}$, respectively. Suppose that we determine that

$$\left\|\tilde{\mathbf{v}} - \mathbf{v}\right\|_2 = 38766017 .$$

What is the significance of such an error? Also, if the recording was only a 30-second recording instead of a one-minute recording, you would expect the 2-norm to be approximately half, whereas if it was a 10-minute recording, you would expect the 2-norm to increase by approximately a factor of ten.

When measuring the error of a vector $\mathbf{v}$, engineers often use the *root mean-squared error* or rmse($\mathbf{v}$). Since we are assuming that we are comparing an approximation with an original vector $\mathbf{v}$, we usually only write $\operatorname{rmse}\left(\tilde{\mathbf{v}}\right)$ and define

$$\operatorname{rmse}\left(\tilde{\mathbf{v}}\right) \overset{\text{def}}{=} \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left|v_k - \tilde{v}_k\right|^2} = \operatorname{rms}\left(\mathbf{v} - \tilde{\mathbf{v}}\right).$$

You will note that $\operatorname{rmse}\left(\tilde{\mathbf{v}}\right) \overset{\text{def}}{=} \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left|v_k - \tilde{v}_k\right|^2} = \sqrt{\frac{1}{n}}\sqrt{\sum_{k=1}^{n}\left(v_k - \tilde{v}_k\right)^*\left(v_k - \tilde{v}_k\right)} = \frac{\left\|\mathbf{v} - \tilde{\mathbf{v}}\right\|_2}{\sqrt{n}}$. This describes the error per entry in the vector. Suppose the error is approximately constant throughout the approximation, so $v_k - \tilde{v}_k \approx \varepsilon$. In this case,

[8] The original sound wave comes from the document *Sound* by R. Luke DuBois and Wilm Thoben available at https://processing.org/tutorials/sound/, a tutorial in support of the text *Processing: A Programming Handbook for Visual Designers and Artists*, 2nd Ed., by Casey Reas and Ben Fry, published by MIT Press, 2014.

$$\text{rmse}(\tilde{\mathbf{v}}) = \sqrt{\frac{1}{n}\sum_{k=1}^{n}|v_k - \tilde{v}_k|^2} \approx \sqrt{\frac{1}{n}\sum_{k=1}^{n}|\varepsilon|^2} = \sqrt{\frac{1}{n}n|\varepsilon|^2} = |\varepsilon|,$$

so the error does not depend on the dimension of the vector.

Note that most definitions of the RMSE assume that the entries are real, so they simply use $\sqrt{\frac{1}{n}\sum_{k=1}^{n}(v_k - \tilde{v}_k)^2}$, but there is no reason the definition cannot also be applied to calculating the error in complex vectors.

4.5.3 RMSE applied to images

The RMSE can, of course, be applied to other approximations, including the approximation of images. For example, a grey-scale digital image may be an $m \times n$ grid of pixels with a value between 0 (black) and 1 (white). If the vector $\mathbf{u}$ represents the actual digital image, and $\tilde{\mathbf{u}}$ represents an approximation, then

$$\text{RMSE}(\tilde{\mathbf{u}}) \overset{\text{def}}{=} \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}|u_{i,j} - \tilde{u}_{i,j}|^2}.$$

For the images of Lenna in Section 2.5.4, the white Gaussian noise was sampled from a normal distribution with a standard deviation of 0.1. One property of the RMSE is that the error of white Gaussian noise is equal to the standard deviation. Consequently, if you were to calculate the rmse of both noisy images, you will get the values of 0.1 and 0.2, respectively.



Figure 41. The image of Lenna with white Gaussian noise with standard deviations 0.1 and 0.2 added to it.

4.5.4 Infinity-norm and 3D printing

Any time that a 3D printer (or any additive manufacturing process) is not depositing binding material unnecessarily adds to the manufacturing time. Suppose the printer head is currently at the location

$$\mathbf{u} = \begin{pmatrix} 0.2 \\ 0.7 \\ 0.4 \end{pmatrix}$$

and the printer can continue to any one of three possible points. A simple *greedy* algorithm would dictate that the head should move to the closest point at which it can continue printing. As most 3D printers have three separate motors, each controlling one of the three directions the head can be moved, the infinity norm is the best means of approximating the time it will take to get the printer head to the next location. Suppose there are three possible locations where the head could continue printing:

$$\mathbf{v}_1 = \begin{pmatrix} 0.5 \\ 0.7 \\ 0.4 \end{pmatrix},\ \mathbf{v}_2 = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.4 \end{pmatrix} \text{ and } \mathbf{v}_3 = \begin{pmatrix} 0.5 \\ 0.4 \\ 0.1 \end{pmatrix}.$$

	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$
$\lVert \mathbf{v}_k - \mathbf{u} \rVert_1$	0.3	0.5	0.9
$\lVert \mathbf{v}_k - \mathbf{u} \rVert_2$	0.3	0.36	0.52
$\lVert \mathbf{v}_k - \mathbf{u} \rVert_\infty$	0.3	0.3	0.3

If either the 1- or 2-norms were used to calculate the distance to each of these points, they would both agree that the closest point is the first, $\mathbf{v}_1$. The infinity-norm, however, would recognize that all three points are the same distance away—regardless of which point you pick, the printer head will get to all three points in the same amount of time. Consequently, the infinity norm would give the greatest choice as to where the printer head is to proceed to next.

Suppose, however, that rather than placing emphasis on minimizing the time required to finish the manufacturing process, your goal is to minimize to total amount of time that the motors are being used; that is, minimizing the energy used to complete the manufacturing process. Which norm would be most appropriate now?

Suppose we have a digital recording and we compress this by using only 15 bits instead of the original 16 bits used in the CD recording by truncating the last bit. This would be equivalent to approximately a 16-digit integer by setting the last digit always to zero and storing only the first fifteen. Consequently, on average, half the points will be recorded perfectly (the last bit was a 0 anyway), and for the other half, the error would be 1. What would the RMSE be?

$$\text{RMSE}(\tilde{\mathbf{v}}) = \sqrt{\frac{1}{n}\sum_{k=1}^{n} |v_k - \tilde{v}_k|^2} = \sqrt{\frac{1}{n}\left(\frac{n}{2}\cdot 0 + \frac{n}{2}\cdot 1\right)} = \frac{1}{\sqrt{2}} \approx 0.707.$$

Suppose now we use 14 bits, in which case, approximately one quarter will have an error of two, one half will have an error of one, and the remaining one quarter would be stored precisely. In this case, the RMSE will likely be approximately equal to

$$\text{RMSE}(\tilde{\mathbf{v}}) = \sqrt{\frac{1}{n}\sum_{k=1}^{n} |v_k - \tilde{v}_k|^2} \approx \sqrt{\frac{1}{n}\left(4\cdot\frac{n}{4} + 1\cdot\frac{n}{2} + 0\cdot\frac{n}{4}\right)} = \sqrt{3} \approx 1.732.$$

Thus, the RMSE, a scaled version of the 2-norm, is often used in industry to calculate the error between an approximation of a digital signal. The value estimates the average error at each point, as opposed to calculating the total error.

4.5.5 The *k*-mean algorithm

Suppose you have *n* vectors, and it appears they represent *k* different clusters. How do you best determine these clusters? The *k*-mean clustering algorithm says:

1. Estimate the *k* different means $\bar{\mathbf{u}}_1, \ldots, \bar{\mathbf{u}}_k$.
2. Iterate the following until the sets no longer change:
    a. For each vector $\mathbf{u}_i$, place it into a set U_j if $\mathbf{u}_i$ is closer to $\bar{\mathbf{u}}_j$ than to any other of the *k* means.
    b. Redefine *k* new means by letting $\bar{\mathbf{u}}_j$ be the average of the entries in U_j.

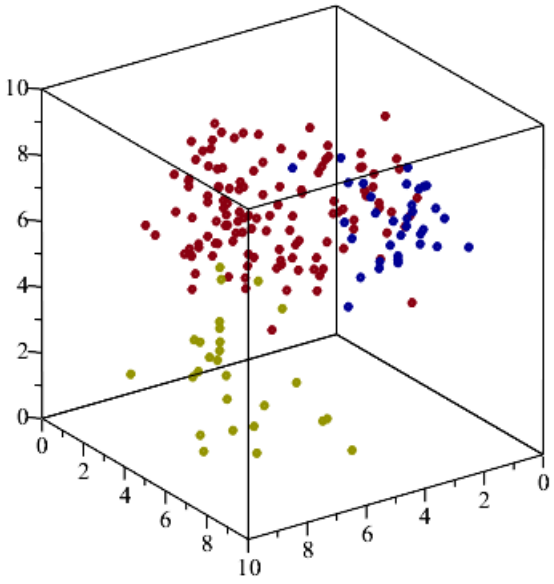
This is a reasonable fast algorithm that quickly partitions the initial *n* points into *k* collections. For example, the following 200 points were estimated to appear in three clusters with means

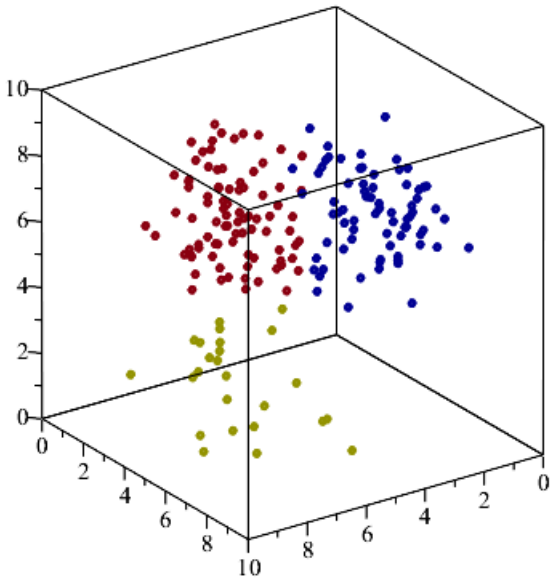
$$\left\{\begin{pmatrix} 0 \\ 0 \\ 10 \end{pmatrix}, \begin{pmatrix} 0 \\ 10 \\ 10 \end{pmatrix}, \begin{pmatrix} 10 \\ 10 \\ 0 \end{pmatrix}\right\}.$$

After seven iterations, these two hundred points were partitioned into three clusters with means

$$\left\{\begin{pmatrix} 4.95 \\ 1.94 \\ 5.76 \end{pmatrix}, \begin{pmatrix} 1.99 \\ 4.79 \\ 6.05 \end{pmatrix}, \begin{pmatrix} 6.73 \\ 4.86 \\ 1.69 \end{pmatrix}\right\}.$$

The following two images show the clusters as determined by the first three estimators of the means of the clusters together with the clusters as determined by the last set of means.


10
8
6
4
2
0
0
2
4
6
8
10
0
2
4
6
8
10


10
8
6
4
2
0
0
2
4
6
8
10
0
2
4
6
8
10

The following is the Matlab code for generating vectors, and then using the *k*-mean algorithm to partition the vectors into *k* sets.

```
% Define a function that maps an integer m onto 1, 2, ..., n cyclically
%  - The 'mod( m, n )' routine mas an integer m onto 0, 1, ..., n and
%    thus we must shift the argument and return value appropriately.
mod1 = @(m, n)(mod( m - 1, n ) + 1 );

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Constants                                                               %
%   I am using all capital letters to help you remember these are         %
%   are constant values                                                   %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Number of vectors in each initial set
VECTORS_PER_SET = 20;

% Number of sets
NUM_OF_SETS = 6;

% Next, we define styles and colors. This will help
% us differentiate between the vectors in the sets
%  - We could just count and assign NUM_OF_SYMBOLS and NUM_OF_COLORS, but
%    then if we changed the colors we are using, we would also have to update
%    the next statement, as well. If we forgot, this would be a bug.
%  - For example, some users might not like using yellow as a color.
%  - Ideally, the greatest common divisor (gcd) of NUM_OF_SYMBOLS and
%    NUM_OF_COLORS should be 1, in which case all combinations of symbols
%    and colors will be used before any repetition.

SYMBOLS = '+xod.v*ps^<>h';
NUM_OF_SYMBOLS = length( SYMBOLS );
COLORS = 'rgbcmky';
NUM_OF_COLORS = length( COLORS );

if gcd( NUM_OF_SYMBOLS, NUM_OF_COLORS ) ~= 1
    warning( 'Not all possible combinations of colors and symbols will be used.' );
end
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Generating the vectors                                                 %
%   I am using all capital letters to help you remember these are        %
%   are constant values                                                  %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Create an array of NUM_OF_SETS x VECTORS_PER_SET vectors
%  - The ith vector is accessed by vectors(:, i)
vectors = zeros( 3, NUM_OF_SETS*VECTORS_PER_SET );

% Define NUM_OF_SETS mean vectors
%  - The kth mean vector is accessed by means(:, k)
means   = [ 1.3       0.2       0.7      0.5     -1.5    0.2
            0.3       1.7      -0.1      1.7      0.4    1.3
           -0.1       0.4       1.3      0.1      2.1    0.2];
% You could also uncomment this to create NUM_OF_SETS vectors in [0,1]x[0,1]x[0,1]
% means = rand( 3, NUM_OF_SETS );

% Our initial collection of vectors will be created by having
% VECTORS_PER_SET vectors normally distributed around each of the
% centroinds vectors.
%  - This vector allows each cluster of vectors to have a different
%    standard deviation.
%  - The smaller the standard deviation, the tighter the vectors
%    will be around the mean vector.
%  - If the standard deviation is large enough, vectors may actually
%    end up closer to other means.
std_devs = [ 0.6       0.7       0.5      0.6      1.0    2.3];

% You could also uncomment this to create NUM_OF_SETS standard deviations
% std_devss = 0.25 + 0.5*rand( 1, NUM_OF_SETS );

% Now we generate VECTORS_PER_SETS vectors for each of the mean
% vectors by generating random normal vectors multiplied by the
% corresponding standard deviations.
%  - The ... means that the command goes to the next line
for k = 1:NUM_OF_SETS
    for i = 1:VECTORS_PER_SET
        vectors(:, VECTORS_PER_SET*(k - 1) + i) = ...
                        means(:,k) + std_devs(k)*randn(3,1);
    end
end

% You could also uncomment this to create these random vectors more quickly
% vectors = zeros( 3, 0 );
% for k = 1:NUM_OF_SETS
%     vectors = [vectors ...
%                repmat( means(:,k) ) + std_devs(k)*randn( 3, VECTORS_PER_SET )];
% end
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Plot the vectors in their original sets                                     %
%  - The 'hold on' and 'hold off' control whether or not the next             %
%    plot is added to the current plot, or creates a new plot                 %
%    altogether, respectively.                                                %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
hold off
for k = 1:NUM_OF_SETS
    rng = ((k - 1)*VECTORS_PER_SET + 1):k*VECTORS_PER_SET;
    plot3( vectors(1,rng), vectors(2,rng), vectors(3,rng), ...
           [SYMBOLS(mod1( k, NUM_OF_SYMBOLS )) ...
            COLORS(mod1( k, NUM_OF_COLORS ))] );
    hold on
end
hold off

% Pause for two seconds
pause( 2 );
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Execute the k-mean algorithm                                          %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% We will define an array that stores which mean each vector is
% associated with
k_set = zeros( 1, NUM_OF_SETS*VECTORS_PER_SET );

% This variable will store the number of times the following loop is run
iterations = 0;

while ( 1 )
    iterations = iterations + 1;
    old_k_set = k_set;

    % For each vector, find which mean it is closest to it by
    %  - First assuming that the first mean is closest
    %  - Check the 2nd through last means, and if any are closer,
    %    update this assumption
    %  - Finally, record the closest index for the ith vector
    for i = 1:NUM_OF_SETS*VECTORS_PER_SET
        closest_distance = norm( means(:,1) - vectors(:,i), 2 );
        closest_index = 1;

        for k = 2:NUM_OF_SETS
            next_distance = norm( means(:,k) - vectors(:,i), 2 );

            if next_distance < closest_distance
                closest_distance = next_distance;
                closest_index = k;
            end
        end

        k_set(i) = closest_index;
    end

    % If nothing has changed (that is, if all the entries
    % in old_k_set equal the entries in k_set) we are finished
    if all( old_k_set == k_set )
        break;
    end

    % We will calculate the average of all the vectors closest to
    % the ith mean by
    %  - Summing up the vectors and counting the number of each
    %  - Dividing the sums by the count to get the average
    means = zeros( 3, NUM_OF_SETS );
    count = zeros( 1, NUM_OF_SETS );

    for k = 1:NUM_OF_SETS*VECTORS_PER_SET
        means(:, k_set(k)) = means(:, k_set(k)) + vectors(:, k);
        count(k_set(k))    = count(k_set(k))    + 1;
    end

    for k = 1:NUM_OF_SETS
        means(:, k) = means(:, k)/count(k);
    end
end

% Print the number of iterations to the screen
iterations
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Plot the vectors in their new set defined by the k-mean algorithm  %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
hold off
for i = 1:NUM_OF_SETS*VECTORS_PER_SET
    plot3( vectors(1,i), vectors(2,i), vectors(3,i), ...
           [SYMBOLS(mod1( k_set(i), NUM_OF_SYMBOLS )) ...
            COLORS(mod1( k_set(i), NUM_OF_COLORS ))] );
    hold on
end
hold off
```

In executing this algorithm, one example is shown in Figure 42. On the left is the original set of vectors that are randomly generated, and on the right are the vectors divided into the six sets. Many of the vectors originally indicated by black inverted triangles have been dispersed between other groups, and the centers of the of the black, cyan and green groups have shifted significantly.

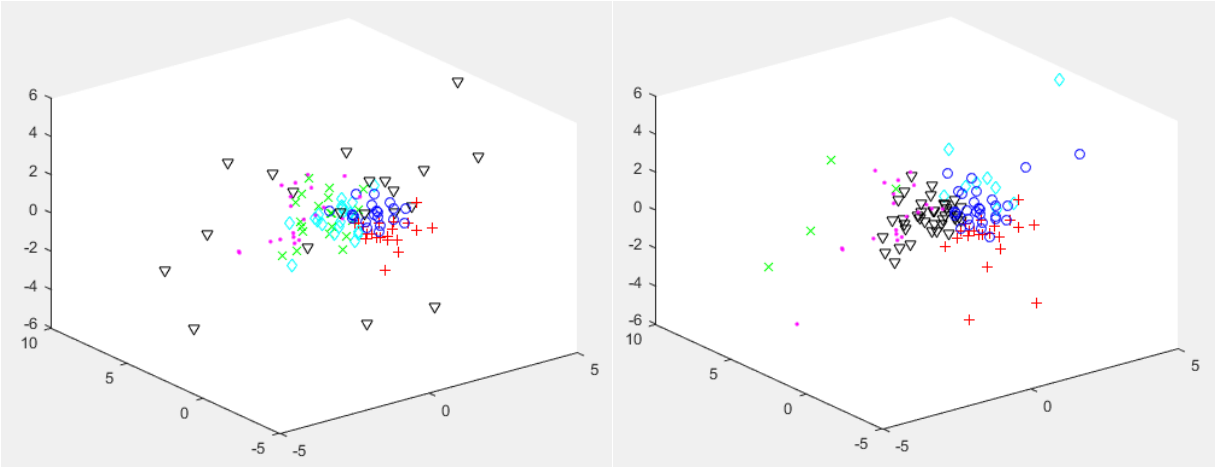


Figure 42. Applying the *k*-mean algorithm to an initial set of 120 points.

Initially, the six means around which the randomly generated vectors are distributed were:

$$+\begin{pmatrix}1.3\\0.3\\-0.1\end{pmatrix},\times\begin{pmatrix}0.2\\1.7\\0.4\end{pmatrix},\bigcirc\begin{pmatrix}0.7\\-0.1\\1.3\end{pmatrix},\lozenge\begin{pmatrix}0.5\\1.7\\0.1\end{pmatrix},\bullet\begin{pmatrix}-1.5\\0.4\\2.1\end{pmatrix},\nabla\begin{pmatrix}0.2\\1.3\\0.2\end{pmatrix}.$$

After the *k*-mean algorithm, the means shifted slightly in some cases, but significantly in others.

$$+\begin{pmatrix}1.3761\\0.5132\\-0.7575\end{pmatrix},\times\begin{pmatrix}-1.2493\\6.0520\\-0.2085\end{pmatrix},\bigcirc\begin{pmatrix}0.7623\\-0.2996\\1.5665\end{pmatrix},\lozenge\begin{pmatrix}2.5456\\2.7957\\1.3252\end{pmatrix},\bullet\begin{pmatrix}-1.8851\\0.4643\\1.8312\end{pmatrix},\nabla\begin{pmatrix}-0.0030\\1.7980\\0.5346\end{pmatrix}$$

As a second example, Figure 43 shows a set of randomly generated vectors that have 20 sets with 50 vectors each. It took 18 iterations of the *k*-mean algorithm to converge to the partitioning on the right.

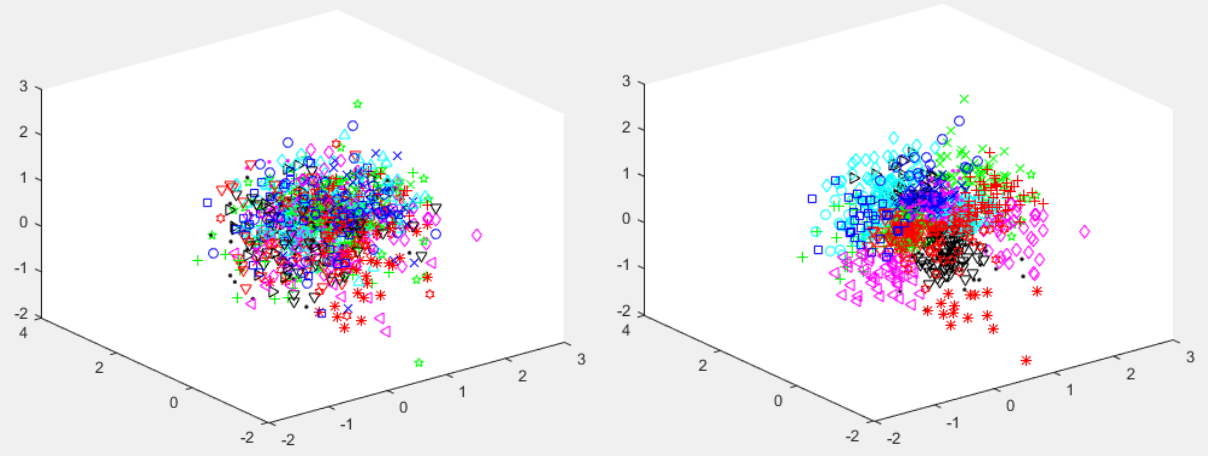


Figure 43. Applying the *k*-mean algorithm to 20 sets of 50 points each.

4.6 Summary of norms of vector spaces

To summarize, we have looked at many different norms on various types of vectors. You will, however, notice the similarities between all of them, as is shown in Table 1.

Table 1. The 1-, 2- and infinity-norms of finite-dimensional vectors, discrete signals, polynomials and functions.

	1-norm	**2-norm**	**∞-norm**
Finite-dimensional vectors	$\|\mathbf{u}\|_1 = \sum_{k=1}^{n} \|u_k\|$	$\|\mathbf{u}\|_2 = \sqrt{\sum_{k=1}^{n} \|u_k\|^2}$	$\|\mathbf{u}\|_\infty = \max_{1\le k\le n} \{\|u_k\|\}$
Discrete signals	$\|y\|_1 = \sum_{k=0}^{\infty} \|y[k]\|$	$\|y\|_2 = \sqrt{\sum_{k=0}^{\infty} \|y[k]\|^2}$	$\|y\|_\infty = \sup_{k\ge 0} \{\|y[k]\|\}$
Polynomials	$\|p\|_1 = \sqrt{\int_0^1 \|p(e^{j2\pi t})\| dt}$	$\|p\|_2 = \sqrt{\int_0^1 \|p(e^{j2\pi t})\|^2 dt}$	$\|p\|_\infty = \max_{0\le t\le 1} \{\|p(e^{j2\pi t})\|\}$
Functions	$\|f\|_1 = \int_{-\infty}^{\infty} \|f(x)\| dx$	$\|f\|_2 = \sqrt{\int_{-\infty}^{\infty} \|f(x)\|^2 dx}$	$\|f\|_\infty = \sup_{-\infty<x<\infty} \{\|f(x)\|\}$

All of these norms have their applications in various fields of engineering. The most important, however, are the 2-norms, as they are all intimately related to another concept in linear algebra: that of the inner product.

5 Inner product spaces

This chapter is likely the most significant to all engineers: the *inner product* or *dot product* as it is likely called in secondary school. We will revisit the definition of the inner product and then consider other forms of inner product on other vector spaces.

5.1 Definition of an inner product

First, we will review the *inner product*, or as you may have learned it, the *dot product* of two vectors. If

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

are real-valued vectors then the inner product is defined as

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{k=1}^{n} u_k v_k \, .$$

Note that we will be using the angled-bracket notation instead of the—perhaps more familiar—version $\mathbf{u} \bullet \mathbf{v}$. When you take a course in quantum mechanics, this will become the usual notation for taking the inner product of two vectors.

The use of the dot to represent the inner product will be seen again in calculus and in courses related to, for example, electricity and magnetism. For example, Gauss's laws may be written as

$$\vec{\nabla} \bullet \mathbf{E} = \frac{\rho}{\varepsilon_0} \text{ and } \vec{\nabla} \bullet \mathbf{B} = 0 \, ,$$

or, in integral form,

$$\oiint_{\partial V} \mathbf{E} \bullet \mathrm{d}\mathbf{S} = \frac{1}{\varepsilon_0} \iiint_V \rho \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}z \text{ and } \oiint_{\partial \Omega} \mathbf{B} \bullet \mathrm{d}\mathbf{S} = 0 \, .$$

Similarly, you will see the divergence theorem and Stokes's theorem:

$$\oiint_{\partial V} \mathbf{f} \bullet \mathrm{d}\mathbf{s} = \frac{1}{\varepsilon_0} \iiint_V \left(\vec{\nabla} \bullet \mathbf{f} \right) \mathrm{d}v \text{ and } \oint_{\partial R} \mathbf{f} \bullet \mathrm{d}\mathbf{r} = \frac{1}{\varepsilon_0} \iint_R \left(\vec{\nabla} \times \mathbf{f} \right) \bullet \mathrm{d}\mathbf{s} \, .$$

This operation has a number of characteristics. The inner product for real vectors the properties that it is

1. *symmetric* as $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle$,
2. *bilinear* as $\langle \alpha \mathbf{u} + \beta \mathbf{v}, \mathbf{w} \rangle = \alpha \langle \mathbf{u}, \mathbf{w} \rangle + \beta \langle \mathbf{v}, \mathbf{w} \rangle$ and $\langle \mathbf{u}, \alpha \mathbf{v} + \beta \mathbf{w} \rangle = \alpha \langle \mathbf{u}, \mathbf{v} \rangle + \beta \langle \mathbf{u}, \mathbf{w} \rangle$, and
3. *positive definite*, $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$ and $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ if and only if $\mathbf{u} = \mathbf{o}$.

You will note that the 2-norm may be defined as $\|\mathbf{u}\|_2 = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$, for

$$\langle \mathbf{u}, \mathbf{u} \rangle = \sum_{k=1}^{n} u_k u_k = \sum_{k=1}^{n} u_k^2 \ .$$

Unfortunately, if we try the same with a complex vector space, we run into issues, for if $\mathbf{u} = \begin{pmatrix} 1 \\ j \end{pmatrix}$ then $\langle \mathbf{u}, \mathbf{u} \rangle = 1 \cdot 1 + j \cdot j = 1 - 1 = 0$, so we lose the positive definite character—the inner product of **u** with itself is zero even though **u** is not the zero vector. Consequently, we need an alternative definition of the complex inner product. If **u** and **v** are complex-valued vectors, the inner product can be defined as

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{k=1}^{n} u_k^* v_k \ .$$

Whether we take the complex conjugate of the first vector entries or the second is a matter of preference. Mathematicians generally choose the second; however, for your quantum mechanics course, their preference is to take the complex conjugate of the first. Note now that the 2-norm is still defined using the this inner product:

$$\langle \mathbf{u}, \mathbf{u} \rangle = \sum_{k=1}^{n} u_k^* u_k = \sum_{k=1}^{n} |u_k|^2 \ .$$

There are, however, some small changes to the characteristics:

1. conjugate symmetric, as $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle^*$ and
2. sesquilinear[9] as $\langle \mathbf{u}, \alpha \mathbf{v} + \beta \mathbf{w} \rangle = \alpha \langle \mathbf{u}, \mathbf{v} \rangle + \beta \langle \mathbf{u}, \mathbf{w} \rangle$ but $\langle \alpha \mathbf{u} + \beta \mathbf{v}, \mathbf{w} \rangle = \alpha^* \langle \mathbf{u}, \mathbf{w} \rangle + \beta^* \langle \mathbf{v}, \mathbf{w} \rangle$,

but it is still positive definite, so $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$ and $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ if and only if $\mathbf{u} = \mathbf{o}$.

In Matlab, an inner product is expressed as a row vector multiplied by a column vector. As most of the vectors that we shall use are column vectors, the usual representation will be `u'*v`.

```
>> u = [1 2 3]';
>> v = [2 -1 4]';
>> u' * v
    ans =
        12

>> norm( u )
    ans =
       3.741657386773941

>> sqrt( u'*u )
    ans =
       3.741657386773941
```

Definition
An *inner product space* is a vector space on which is defined an inner product for all vectors in that space.

[9] The prefix "sesqui" is derived from one-and-a-half in Latin, so a 150-year anniversary is a sesquicentennial event. Here, it indicates that it is linear in the second term, but *not really* linear in the first.

In this course, we will only discuss the *pure* form of the inner product; however, there are other inner products, as well, all of which satisfy the properties of an inner product. If **w** is a vector where each entry is greater than zero (that is, $w_k > 0$ for $k = 1, \ldots, n$, then the *inner product weighted by* ***w*** is defined as

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{w}} = \sum_{k=1}^{n} u_k^* v_k w_k \ .$$

This operation is symmetric, sesquilinear and positive definite. The entries of **w** must be strictly positive for the resulting weighted inner product to be positive definite.

Problems

1. What is the inner product of the vector $\mathbf{u} = \begin{pmatrix} 3 \\ 2 \\ -1 \end{pmatrix}$ with each of the thee vectors $\begin{pmatrix} 3 \\ 2 \\ -1 \end{pmatrix}$, $\begin{pmatrix} -3 \\ -2 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 2 \\ 0 \\ 6 \end{pmatrix}$, $\begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ -2 \\ -3 \end{pmatrix}$?

2. Given your answers in Question 1, what are the inner product of the given vector **u** and the vectors $\begin{pmatrix} -4 \\ -6 \\ -8 \end{pmatrix}$ and $\begin{pmatrix} 3 \\ -6 \\ -9 \end{pmatrix}$? Do not compute these directly.

3. What is the inner product of the vectors $\begin{pmatrix} -1 \\ -12 \\ -17 \end{pmatrix}$ and $\begin{pmatrix} -7 \\ 0 \\ 1 \end{pmatrix}$ with the vector **u** based on the results in Question 2?

4. Suppose that all the entries in **v** are smaller that 10^{-10} in absolute value (that is, $\|\mathbf{v}\|_\infty \le 10^{-10}$). Argue that $-10^{-10} \|\mathbf{u}\|_1 \le \langle \mathbf{u}, \mathbf{v} \rangle \le 10^{-10} \|\mathbf{u}\|_1$. Why are we using the 1-norm instead of the 2-norm?

5. What vector **v** containing only +1 or –1 in each entry maximizes the inner product $\langle \mathbf{u}, \mathbf{v} \rangle$ when $\mathbf{u} = \begin{pmatrix} 3 \\ -2 \\ 4 \\ -1 \\ 5 \\ 2 \end{pmatrix}$?

6. What is the inner product of the two complex vectors $\mathbf{u} = \begin{pmatrix} 3+j \\ -2-2j \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1-3j \\ 1-4j \end{pmatrix}$?

7. What is the inner product $\langle \mathbf{v}, \mathbf{u} \rangle$ based on your answer to the previous question?

8. What is the inner product of the two complex vectors $\mathbf{u} = \begin{pmatrix} 1-2j \\ 3+4j \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 4-3j \\ 2-j \end{pmatrix}$?

9. Based on the results from the last question, what are the inner products $\left\langle \begin{pmatrix} 2+j \\ -4+3j \end{pmatrix}, \begin{pmatrix} 4-3j \\ 2-j \end{pmatrix} \right\rangle$, $\left\langle \begin{pmatrix} 1-2j \\ 3+4j \end{pmatrix}, \begin{pmatrix} 8-6j \\ 4-2j \end{pmatrix} \right\rangle$ and $\left\langle \begin{pmatrix} 2+j \\ -4+3j \end{pmatrix}, \begin{pmatrix} 8-6j \\ 4-2j \end{pmatrix} \right\rangle$?

Solutions

1. The inner products are 14, –14, 0, 8 and 2. -16 and 6

3. We note that the first vector is the sum of the vectors in Question 2, so the inner product is the sum of the two given inner products, so the inner product is –10. The second vector is the first vector in Question 2 minus the second, and therefore the inner product is –22.

5. The vector that maximizes the inner product is the one that has the same signs as the entries of **u**: $\mathbf{v} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \\ 1 \\ 1 \end{pmatrix}$.

7. Because the complex inner product is conjugate symmetric, it follows that the inner product is also 6.

9. The complex inner products is conjugate linear in the first argument, and the first vector is $j\mathbf{u}$, so the inner product is therefore $(-j)(12-6j) = -6-12j$, it is linear in the second argument, and the second vector is $2\mathbf{v}$, so the inner product is $2(12-6j) = 24-12j$, and the last result combines both of these, so the inner product is $2(-j)(12-6j) = -12-24j$.

5.2 The norm *induced* by an inner product

In the last chapter, we introduced the 1-, 2- and infinity norms. We will now see why the 2-norm is of such significance. We have defined the 2-norm and the inner product on finite dimensional vectors and we noted that

$$\|\mathbf{v}\|_2 = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} \ .$$

Whenever we define any inner product, we will always be able to define a norm based on that definition.

Problems

1. Express $\|\mathbf{u}+\mathbf{v}\|_2^2$ in terms of inner products. How can you simplify the result if you know that **u** and **v** are real vectors? How can you simplify the result if the vectors are complex vectors?

2. What property of the inner product ensures that this value can never be zero?

3. If $\langle \mathbf{u}, \mathbf{v} \rangle = 0$, how can we express $\|\mathbf{u}+\mathbf{v}\|_2^2$ in terms of the 2-norms of **u** and **v**?

Solutions

1. $\|\mathbf{u}+\mathbf{v}\|_2^2 = \langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u}+\mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u}+\mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle$. If the vectors are real, we know that $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$ and thus $\|\mathbf{u}+\mathbf{v}\|_2^2 = \langle \mathbf{u}, \mathbf{u} \rangle + 2\langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle$, but if the vectors are complex then $\langle \mathbf{u}, \mathbf{v} \rangle^* = \langle \mathbf{v}, \mathbf{u} \rangle$, so $\|\mathbf{u}+\mathbf{v}\|_2^2 = \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle^* + \langle \mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + 2\Re\mathrm{e}\left(\langle \mathbf{u}, \mathbf{v} \rangle\right) + \langle \mathbf{v}, \mathbf{v} \rangle$.

3. If $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ then $\|\mathbf{u}+\mathbf{v}\|_2^2 = \langle \mathbf{u}, \mathbf{u} \rangle + \cancelto{0}{\langle \mathbf{u}, \mathbf{v} \rangle} + \cancelto{0}{\langle \mathbf{v}, \mathbf{u} \rangle} + \langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2$.

5.3 Other inner product spaces

We will look at various inner products defined on other vector spaces we have examined. In each case, we will see that there is a 2-norm induced by the inner product.

5.3.1 Inner products on discrete signals

If we have two real discrete signals, we can define an inner product

$$\langle x, y \rangle = \sum_{k=0}^{\infty} x[k]^* \, y[k] .$$

Unfortunately, it may turn out that for certain discrete signals this inner product is infinite; however, if we restrict our choice of discrete signals to those such that $\|y\|_2 < \infty$, then this inner product must also be finite. As with inner products on finite-dimensional vectors, the 2-norm of discrete signals can be defined as $\|y\|_2 = \sqrt{\langle y, y \rangle}$.

As an example, we can compute the inner product of the two signals defined by $x[n] = 2^{-n}$ and $y[n] = 3^{-n}$:

$$\begin{aligned}
\langle x, y \rangle &= \sum_{k=0}^{\infty} x[k] y[k] \\
&= \sum_{k=0}^{\infty} 2^{-n} 3^{-n} \\
&= \sum_{k=0}^{\infty} 6^{-n} \\
&= \sum_{k=0}^{\infty} \left(\frac{1}{6} \right)^n \\
&= \frac{6}{5}
\end{aligned}$$

You will recall from secondary school that $\sum_{k=0}^{n} r^k = \frac{1 - r^{n+1}}{1 - r}$ and if $|r| < 1$ then $\sum_{k=0}^{\infty} r^k = \frac{1}{1 - r}$

5.3.2 Inner product on polynomials

Suppose we have polynomials $p, q : \mathbf{R} \to \mathbf{R}$ where we can define

$$\langle p, q \rangle = \int_0^1 p\left(e^{j2\pi t}\right) q\left(e^{j2\pi t}\right) dt .$$

This integral has all the properties of an inner product, including

1. It is symmetric, as $\langle p, q \rangle = \int_0^1 p\left(e^{j2\pi t}\right) q\left(e^{j2\pi t}\right) dt = \int_0^1 q\left(e^{j2\pi t}\right) p\left(e^{j2\pi t}\right) dt = \langle q, p \rangle$,
2. linear, as

$$\begin{aligned}
\langle \alpha p + \beta q, r \rangle &= \int_0^1 \left(\alpha p\left(e^{j2\pi t}\right) + \beta q\left(e^{j2\pi t}\right) \right) r\left(e^{j2\pi t}\right) dt \\
&= \alpha \int_0^1 p\left(e^{j2\pi t}\right) r\left(e^{j2\pi t}\right) dx + \beta \int_0^1 q\left(e^{j2\pi t}\right) r\left(e^{j2\pi t}\right) dx , \\
&= \alpha \langle p, r \rangle + \beta \langle q, r \rangle
\end{aligned}$$

and,

3. it is positive definite as $\langle p, p\rangle = \int_0^1 p\left(e^{j2\pi t}\right)^2 dt \geq 0$ and $\langle p, p\rangle = 0$ if and only if p is the zero polynomial.

Note that we *could* define the inner product of two quadratic polynomials $p : x \mapsto ax^2 + bx + c$ and $q : x \mapsto dx^2 + ex + f$ as $\langle p, q\rangle = ad + be + cf$, and this would continue to hold all of the properties of inner products, but it is essentially meaningless, as we have already discussed with respect to the norm of a polynomial.

If the polynomials $p, q : \mathbf{R} \to \mathbf{C}$ (and therefore with possibly complex coefficients), we could expand the definition of the inner product to

$$\langle p, q\rangle = \int_0^1 p^*\left(e^{j2\pi t}\right) q\left(e^{j2\pi t}\right) dt,$$

and based on the properties of the integral, all the properties of the inner product for complex vector spaces continues to hold.

5.3.3 Inner products on functions

Suppose we have two analog signals $f, g : D \to \mathbf{C}$, we could define a similar integral

$$\langle f, g\rangle = \int_D f^*(t) g(t) dt$$

where, if we restrict ourselves to those functions such that $\|f\|_2 < \infty$, then this inner product, too, must be finite. The properties of the inner product follow from the linearity of the integral.

5.4 Orthogonality of vectors

Two vectors **u** and **v** are said to be *orthogonal*, that is, at right angles, if the inner product is zero:

$$\langle \mathbf{u}, \mathbf{v} \rangle = 0 \cdot$$

If two vectors are orthogonal, we may also write this as $\mathbf{u} \perp \mathbf{v}$, indicating that the vectors are *at right angles* to each other.

Two vectors are orthogonal if they essentially contain no information about the other with respect to the inner product.

Note: the zero vector is orthogonal to all vectors (including itself).

Problems

1. Which of the following pairs of vectors are orthogonal? $\mathbf{u}_1 = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix}$, $\mathbf{u}_2 = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$, $\mathbf{u}_3 = \begin{pmatrix} 3 \\ 3 \\ 1 \end{pmatrix}$, $\mathbf{u}_4 = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}$ and $\mathbf{u}_5 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$.

2. If **u** is orthogonal to both **v** and **w**, is **u** orthogonal to $\alpha\mathbf{v} + \beta\mathbf{w}$?

3. If **u** is orthogonal to **v** and **v** is orthogonal to **w**, is **u** necessarily orthogonal to **w**?

Solutions

1. $\mathbf{u}_1$ is orthogonal to $\mathbf{u}_3$, $\mathbf{u}_3$ is orthogonal to $\mathbf{u}_5$, and $\mathbf{u}_2$ is orthogonal to $\mathbf{u}_4$.

3. None of the properties of the inner product allow you deduce that if $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ and $\langle \mathbf{v}, \mathbf{w} \rangle = 0$, then $\langle \mathbf{u}, \mathbf{w} \rangle = 0$, and actually, suppose that **u** and **v** were orthogonal. In this case, both $\langle \mathbf{u}, \mathbf{v} \rangle = 0$ and $\langle \mathbf{v}, \mathbf{u} \rangle = 0$, then if orthogonality was transitive, it would follow that $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ for all **u**, which is false.

5.5 Orthogonality in other inner product spaces

We will look at orthogonal polynomials and orthogonal functions, specifically looking at the bases for Fourier series.

5.5.1 Orthogonal polynomials as functions

A very important area of research in mathematics and of significant use to science and engineering is the concept of orthogonal polynomials. Suppose, for example, we consider the interval $[-1,1]$. We note that

$$\begin{aligned}\langle 1, t\rangle &= \int_{-1}^{1} (1\cdot t)\, dt \\ &= \int_{-1}^{1} t\, dt \\ &= \frac{1}{2}t^2\Big|_{-1}^{1} \\ &= \frac{1}{2} - \frac{1}{2} = 0\end{aligned}$$

and

$$\begin{aligned}\left\langle t, t^2\right\rangle &= \int_{-1}^{1} t\cdot t^2 dx \\ &= \int_{-1}^{1} t^3 dx \\ &= \frac{1}{4}t^4\Big|_{-1}^{1} \\ &= \frac{1}{4} - \frac{1}{4} = 0\end{aligned}$$

and therefore the polynomials 1 and t are orthogonal, as are the polynomials t and t^2. This can be seen, for example, in Figure 44.

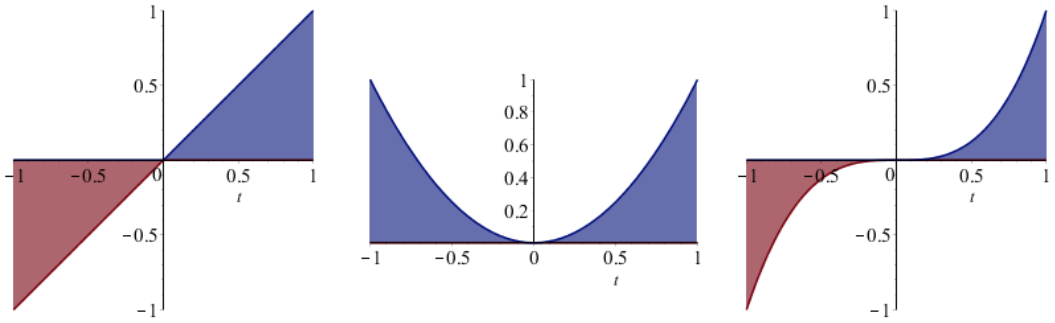


Figure 44. Plots of the curves t, t^2 and t^3; the first showing that 1 and t are orthogonal, the second showing that 1 and t^2 are not orthogonal, and the third showing that the pair 1 and t^3 and the pair t and t^2 are orthogonal on the given interval [–1, 1].

In a few chapters, we shall see how we can impose orthogonality on at least some sets of non-mutually orthogonal vectors.

One group of mutually orthogonal polynomials on $[-1,1]$ are called the *Chebyshev polynomials* with a weighting function of $\frac{1}{\sqrt{1-x^2}}$ and the first eight of which are

1, t, t^2-1, $4t^3-3t$, $8t^4-8t^2+1$, $16t^5-20t^3+5t$, $32t^6-48t^4+18t^2-1$, and $64t^7-112t^5+56t^3-7t$.

For example,

$$\begin{aligned}
\left\langle 4t^3-3t, 8t^4-8t^2+1 \right\rangle_{\frac{1}{\sqrt{1-x^2}}} &= \int_{-1}^{1} \frac{\left(4t^3-3t\right)\left(8t^4-8t^2+1\right)}{\sqrt{1-t^2}}\,dt \\
&= \int_{-1}^{1} \frac{32t^7-56t^5+28t^3-3t}{\sqrt{1-t^2}}\,dt \\
&= \left(\frac{37t^8-71t^6+54t^4-t^2-19}{7\sqrt{1-t^2}} \right|_{-1}^{1} \\
&= \lim_{t \nearrow 1} \frac{37t^8-71t^6+54t^4-t^2-19}{7\sqrt{1-t^2}} - \lim_{t \searrow -1} \frac{37t^8-71t^6+54t^4-t^2-19}{7\sqrt{1-t^2}} \\
&= 0-0=0
\end{aligned}$$

That this integral is zero can also be seen by examining the plot of $\frac{\left(4t^3-3t\right)\left(8t^4-8t^2+1\right)}{\sqrt{1-t^2}}$ in Figure 45, which suggests that the corresponding positive and negative areas cancel each other out.

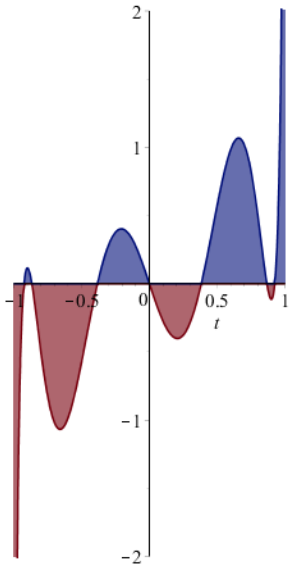


Figure 45. A plot of $\frac{\left(4t^3-3t\right)\left(8t^4-8t^2+1\right)}{\sqrt{1-t^2}}$ on the interval [–1, 1].

The coefficients are chosen so that the maxima and minima achieve values of ± 1, as shown in Figure 46 which shows the first six Chebyshev orthogonal polynomials.

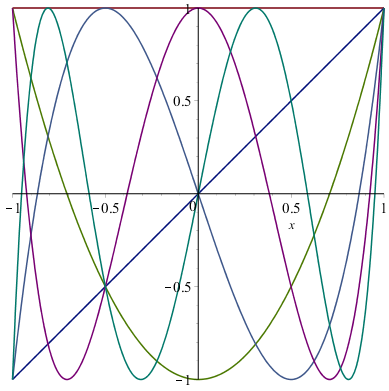
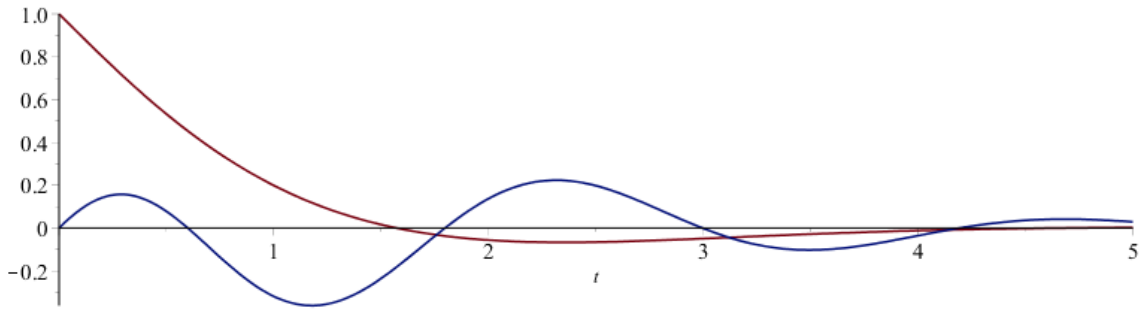


Figure 46. The Chebyshev polynomials of degrees 0 through 5.

5.5.2 Orthogonal functions

As for examples of functions that are orthogonal on the semi-infinite range $[0,\infty)$, consider the two functions $e^{-t}\cos(t)$ and $te^{-t}\cos(2.6222141407t)$. The actual angular frequency is only approximated by 2.6222⋯, but as one may see, it is more difficult to find orthogonal functions on a semi-infinite interval.



One important class of orthogonal functions are

$$1, \cos(t), \sin(t), \cos(2t), \sin(2t), \cos(3t), \sin(3t), \ldots$$

on the interval $[0, 2\pi]$. It is beyond the scope of this course, but

$$\int_0^{2\pi} (1\cdot 1)\,dt = 2\pi \quad \text{and} \quad \int_0^{2\pi} \cos^2(nt)\,dt = \int_0^{2\pi} \sin^2(nt)\,dt = \pi$$

but for positive integers m and n, if $m \neq n$,

$$\int_0^{2\pi} \cos(mt)\cos(nt)\,dt = \int_0^{2\pi} \sin(mt)\sin(nt)\,dt = 0$$

and for all positive integral values of m and n,

$$\int_0^{2\pi} \left(\cos(mt)\cdot 1\right) dt = \int_0^{2\pi} \left(\sin(mt)\cdot 1\right) dt = \int_0^{2\pi} \cos(mt)\sin(nt)\, dt = 0 .$$

Consequently, this collection forms an orthogonal collection of functions on the interval [0, 2π]. This is a rather important collection of orthogonal functions, as it will define a Fourier series in second year when you approximate periodic functions by this collection of orthogonal functions.

Another collection of orthogonal complex-valued functions [0, 2π] are the exponential functions

$$e^{j2\pi nt}$$

for n = …, –3, –2, –1, 0, 1, 2, 3, … . These are even more elegant that

$$\int_0^{2\pi} \left(e^{j2\pi mt}\right)^* e^{j2\pi nt}\, dt = \int_0^{2\pi} e^{-j2\pi mt} e^{j2\pi nt}\, dt = \begin{cases} 0 & m \neq n \\ 2\pi & m = n \end{cases}.$$

5.6 Pythagorean theorem

We will now demonstrate a generalization of the familiar Pythagorean theorem:

Pythagorean theorem

If **u** and **v** are orthogonal with respect to an inner product (that is, $\langle \mathbf{u}, \mathbf{v} \rangle = 0$) and $\|\bullet\|_2$ is the norm induced by the inner product, then $\|\mathbf{u}\|_2^{\,2} + \|\mathbf{v}\|_2^{\,2} = \|\mathbf{u} + \mathbf{v}\|_2^{\,2}$.

Proof:

Using the properties of the inner product:

$$\begin{aligned} \|\mathbf{u} + \mathbf{v}\|_2^{\,2} &= \langle \mathbf{u} + \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} + \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} + \mathbf{v} \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle \end{aligned}$$

but by assumption, $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle = 0$. Therefore

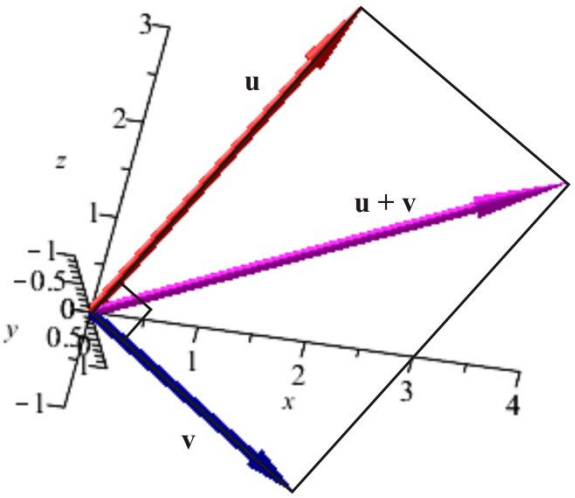
$$\|\mathbf{u} + \mathbf{v}\|_2^{\,2} = \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{u}\|_2^{\,2} + \|\mathbf{v}\|_2^{\,2}. \blacksquare$$

Demonstration of this theorem

For example, the vectors $\mathbf{u} = \begin{pmatrix} 2 \\ -1 \\ 3 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix}$ are orthogonal, and $\mathbf{u}+\mathbf{v} = \begin{pmatrix} 4 \\ 0 \\ 2 \end{pmatrix}$. We see that

$$\|\mathbf{u}\|_2^2 = 2^2 + (-1)^2 + 3^2 = 14,\ \|\mathbf{v}\|_2^2 = 2^2 + 1^2 + (-1)^2 = 6 \text{ and } \|\mathbf{u}+\mathbf{v}\|_2^2 = 4^2 + 0^2 + 2^2 = 20.$$

When oriented correctly, you can see that these **u** and **v** are orthogonal and that the length of **u** + **v** must be equal to $\sqrt{\|\mathbf{u}\|_2^{\,2} + \|\mathbf{v}\|_2^{\,2}}$.



If you know integration, you will know that sine and cosine are orthogonal with respect to the inner product defined by $\langle f, g \rangle = \int_0^{2\pi} f(x) g(x) dx$. Thus, we note that

$$\|\sin - 2\cos\|_2^2 = \int_0^{2\pi} (\sin(x) - 2\cos(x))^2 dx = 5\pi$$

while it is also true that

$$\|\sin\|_2^2 = \int_0^{2\pi} \sin^2(x) dx = \pi \text{ and } \|-2\cos\|_2^2 = \int_0^{2\pi} (-2\cos(x))^2 dx = 4\int_0^{2\pi} \cos^2(x) dx = 4\pi.$$

Recall that mathematicians often write $(\sin(x))^2$ by $\sin^2(x)$.

Problems

1. Show that the Pythagorean theorem is true with the orthogonal vectors $\mathbf{u} = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 3 \\ 3 \\ 1 \end{pmatrix}$.

2. Show that the Pythagorean theorem is true with the orthogonal vectors $\mathbf{u} = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}$.

3. Determine whether or not the following two vector are orthogonal by using the Pythagorean theorem: $\mathbf{u} = \begin{pmatrix} 3 \\ 3 \\ -1 \end{pmatrix}$ and $\mathbf{u} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$.

4. Determine whether or not the following two vector are orthogonal by using the Pythagorean theorem: $\mathbf{u} = \begin{pmatrix} 3 \\ 3 \\ -1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}$.

Solutions

1. $\mathbf{u} + \mathbf{v} = \begin{pmatrix} 4 \\ -1 \\ 4 \end{pmatrix}$ and $\|\mathbf{u} + \mathbf{v}\|_2^2 = 33$ while $\|\mathbf{u}\|_2^2 = 14$ and $\|\mathbf{v}\|_2^2 = 19$.

3. Well, $\mathbf{u} + \mathbf{v} = \begin{pmatrix} 4 \\ 2 \\ -1 \end{pmatrix}$, so $\|\mathbf{u} + \mathbf{v}\|_2^2 = 21$ and this equals $\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 = 19 + 2$, so they must be orthogonal.

5.7 Projections and best approximations

In the previous section, we discussed how a multiple of one vector may be a reasonable approximation to another. The question is, given two real vectors **u** and **v**, what scalar multiple $\alpha\mathbf{u}$ is the "best" approximation to **v**? We will claim that it is the vector that is "closest", that is, the vector that minimizes

$$\|\mathbf{v}-\alpha\mathbf{u}\|_2 \cdot$$

We will consider this both for real vector spaces and for complex vector spaces, and while we will see that the complex case is analogous to the real case, we will never-the-less examine the real case first as the solution is more intuitive.

5.7.1 For real vector spaces

From our definition of the 2-norm, this equals

$$\sqrt{\langle \mathbf{v}-\alpha\mathbf{u}, \mathbf{v}-\alpha\mathbf{u}\rangle}\,.$$

Because the square root function is strictly monotonic increasing, minimizing this value is equivalent to minimizing

$$\langle \mathbf{v}-\alpha\mathbf{u}, \mathbf{v}-\alpha\mathbf{u}\rangle$$

and from the properties of the inner product,

$$\begin{aligned}\langle \mathbf{v}-\alpha\mathbf{u}, \mathbf{v}-\alpha\mathbf{u}\rangle &= \langle \mathbf{v}, \mathbf{v}-\alpha\mathbf{u}\rangle + \langle -\alpha\mathbf{u}, \mathbf{v}-\alpha\mathbf{u}\rangle \\ &= \langle \mathbf{v}, \mathbf{v}-\alpha\mathbf{u}\rangle - \alpha\langle \mathbf{u}, \mathbf{v}-\alpha\mathbf{u}\rangle \\ &= \langle \mathbf{v}, \mathbf{v}\rangle + \langle \mathbf{v}, -\alpha\mathbf{u}\rangle - \alpha\langle \mathbf{u}, \mathbf{v}\rangle - \alpha\langle \mathbf{u}, -\alpha\mathbf{u}\rangle \\ &= \langle \mathbf{v}, \mathbf{v}\rangle - \alpha\langle \mathbf{v}, \mathbf{u}\rangle - \alpha\langle \mathbf{u}, \mathbf{v}\rangle + \alpha^2\langle \mathbf{u}, \mathbf{u}\rangle\end{aligned}$$

and because the real inner product is symmetric, we have

$$\langle \mathbf{v}-\alpha\mathbf{u}, \mathbf{v}-\alpha\mathbf{u}\rangle = \langle \mathbf{v}, \mathbf{v}\rangle - \alpha\langle \mathbf{v}, \mathbf{u}\rangle - \alpha\langle \mathbf{u}, \mathbf{v}\rangle + \alpha^2\langle \mathbf{u}, \mathbf{u}\rangle = \langle \mathbf{v}, \mathbf{v}\rangle - 2\alpha\langle \mathbf{u}, \mathbf{v}\rangle + \alpha^2\langle \mathbf{u}, \mathbf{u}\rangle \cdot$$

Now, the inner products are constant, and therefore this is simply finding the minimum of a polynomial; however, we already know this from secondary school: the minimum of a polynomial ax^2+bx+c is the point $x=-\frac{b}{2a}$ [10], so in this case, the minimum is at

$$\alpha = -\frac{-2\langle \mathbf{u}, \mathbf{v}\rangle}{2\langle \mathbf{u}, \mathbf{u}\rangle} = \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle},$$

and therefore the "best" approximation of **v** is the vector

[10] Recall that a local maximum or minimum of a quadratic polynomial occurs when the derivative is zero, so we must solve $\frac{d}{dx}\left(ax^2+bx+c\right) = 2ax+b = 0$, or $2ax = -b$ and thus $x=-\frac{b}{2a}$. It is a minimum if $a < 0$ and a maximum if $a > 0$.

$$\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} .$$

Alternate derivation

An alternate derivation is to find that value of α that makes $\mathbf{v} - \alpha\mathbf{u}$ perpendicular to $\mathbf{u}$. In this case, we have that

$$\langle \mathbf{u}, \mathbf{v} - \alpha\mathbf{u} \rangle = 0 ,$$

so $\langle \mathbf{u}, \mathbf{v} \rangle - \alpha \langle \mathbf{u}, \mathbf{u} \rangle = 0$ and so solving for α yields

$$\alpha = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle}$$

and so $\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u}$ is the best approximation of $\mathbf{v}$.

We define this best approximation of $\mathbf{v}$ by a multiple of $\mathbf{u}$ to be the *projection* of $\mathbf{v}$ onto $\mathbf{u}$, and we shall write it as

$$\text{proj}_{\mathbf{u}} (\mathbf{v}) \overset{def}{=} \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} .$$

Of course, this is not a valid formula if $\langle \mathbf{u}, \mathbf{u} \rangle = 0$, which implies that $\|\mathbf{u}\|_2 = 0$, and therefore $\mathbf{u} = \mathbf{o}$. The projection of any vector onto the zero vector **0** is **0**.

The vector between the projection and $\mathbf{v}$ is orthogonal to $\mathbf{u}$:

$$\begin{aligned} \langle \mathbf{v} - \text{proj}_{\mathbf{u}} (\mathbf{v}), \mathbf{u} \rangle &= \left\langle \mathbf{v} - \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u}, \mathbf{u} \right\rangle \\ &= \langle \mathbf{v}, \mathbf{u} \rangle - \left\langle \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u}, \mathbf{u} \right\rangle \\ &= \langle \mathbf{v}, \mathbf{u} \rangle - \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \langle \mathbf{u}, \mathbf{u} \rangle \\ &= \langle \mathbf{v}, \mathbf{u} \rangle - \langle \mathbf{u}, \mathbf{v} \rangle = 0 \end{aligned}$$

For example, given the two vectors in Figure 47, we can find both the projection of $\mathbf{u}$ onto $\mathbf{v}$ (left) and the projection of $\mathbf{v}$ onto $\mathbf{u}$. In both cases, the projection is the black vector.

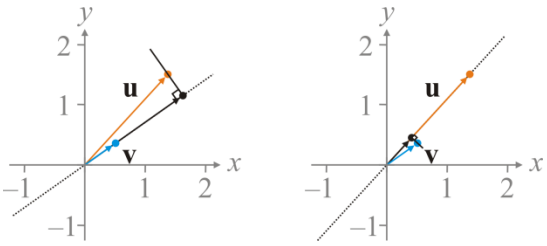


Figure 47. The projection of **u** onto **v** and the projection of **v** onto **u**.

As a real example, let us consider the two vectors $\mathbf{u} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} -5 \\ 1 \end{pmatrix}$. Thus,

$$\begin{aligned} \text{proj}_{\mathbf{u}}(\mathbf{v}) &= \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} \\ &= \frac{3 \cdot (-5) + 2 \cdot 1}{3 \cdot 3 + 2 \cdot 2} \begin{pmatrix} 3 \\ 2 \end{pmatrix} \\ &= -1 \cdot \begin{pmatrix} 3 \\ 2 \end{pmatrix} \\ &= \begin{pmatrix} -3 \\ -2 \end{pmatrix} \end{aligned}$$

and therefore $\mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v}) = \mathbf{v} - \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} = \begin{pmatrix} -5 \\ 1 \end{pmatrix} - \begin{pmatrix} -3 \\ -2 \end{pmatrix} = \begin{pmatrix} -2 \\ 3 \end{pmatrix}$, as shown in Figure 48.

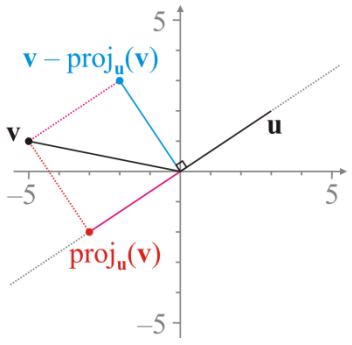


Figure 48. The projection of **v** onto **u** and $\mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v})$.

The perpendicular component of **v** is $\text{perp}_{\mathbf{u}}(\mathbf{v}) \overset{def}{=} \mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v})$, and therefore $\mathbf{v} = \text{perp}_{\mathbf{u}}(\mathbf{v}) + \text{proj}_{\mathbf{u}}(\mathbf{v})$ where $\text{perp}_{\mathbf{u}}(\mathbf{v}) \perp \text{proj}_{\mathbf{u}}(\mathbf{v})$. Like the projection, the perpendicular component is also linear in its argument.

Problems

1. Find the projection of the vector $\mathbf{u} = \begin{pmatrix} 2 \\ 5 \\ 4 \end{pmatrix}$ onto the vector $\mathbf{v} = \begin{pmatrix} -3 \\ 1 \\ 2 \end{pmatrix}$.

2. Find the projection of the vector $\mathbf{v} = \begin{pmatrix} -2 \\ 3 \\ -4 \end{pmatrix}$ onto the vector $\mathbf{u} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

3. Find the perpendicular component of the projection of the vector **u** onto the vector **v** in Question 1.

4. Find the perpendicular component of the projection of the vector **v** onto the vector **u** in Question 2.

5. Show that the projection and the perpendicular component satisfy the Pythagorean theorem with the vectors in Question 1.

6. Show that the projection and perpendicular component satisfy the Pythagorean theorem with the vectors in Question 2.

7. Demonstrate that the Pythagorean theorem must be true for the projection and perpendicular components of a vector **v** projected onto a vector **u**.

8. Draw a diagram explaining the property shown in Question 7.

Solutions

1. The projection is $\text{proj}_{\mathbf{v}}(\mathbf{u}) = \dfrac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v} = \dfrac{-6+5+8}{9+1+4} \begin{pmatrix} -3 \\ 1 \\ 2 \end{pmatrix} = \dfrac{7}{14} \begin{pmatrix} -3 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -1.5 \\ 0.5 \\ 1 \end{pmatrix}$.

3. $\text{perp}_{\mathbf{v}}(\mathbf{u}) \overset{def}{=} \mathbf{v} - \text{proj}_{\mathbf{v}}(\mathbf{u}) = \begin{pmatrix} 2 \\ 5 \\ 4 \end{pmatrix} - \begin{pmatrix} -1.5 \\ 0.5 \\ 1 \end{pmatrix} = \begin{pmatrix} 3.5 \\ 4.5 \\ 3 \end{pmatrix}$.

5. $\|\text{proj}_{\mathbf{v}}(\mathbf{u})\|_2^2 = (-1.5)^2 + (0.5)^2 + 1^2 = 3.5$, $\|\text{perp}_{\mathbf{v}}(\mathbf{u})\|_2^2 = 3.5^2 + 4.5^2 + 3^2 = 41.5$, and $\|\mathbf{u}\|_2^2 = 2^2 + 5^2 + 4^2 = 45$; and 3.5 + 41.5 = 45.

7. $\|\text{proj}_{\mathbf{u}}(\mathbf{v})\|_2^2 = \left\| \dfrac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} \right\|_2^2 = \dfrac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{\langle \mathbf{u}, \mathbf{u} \rangle^2} \|\mathbf{u}\|_2^2 = \dfrac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{\langle \mathbf{u}, \mathbf{u} \rangle}$ because $\|\mathbf{u}\|_2^2 = \langle \mathbf{u}, \mathbf{u} \rangle$. Similarly,

$$\|\text{perp}_{\mathbf{u}}(\mathbf{v})\|_2^2 = \|\mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v})\|_2^2 = \langle \mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v}), \mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v}) \rangle = \langle \mathbf{v}, \mathbf{v} \rangle - 2\langle \mathbf{v}, \text{proj}_{\mathbf{u}}(\mathbf{v}) \rangle + \langle \text{proj}_{\mathbf{u}}(\mathbf{v}), \text{proj}_{\mathbf{u}}(\mathbf{v}) \rangle,$$

and substituting in the definition of the projection, we have

$$\|\text{perp}_{\mathbf{u}}(\mathbf{v})\|_2^2 = \langle \mathbf{v}, \mathbf{v} \rangle - 2\left\langle \mathbf{v}, \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} \right\rangle + \left\langle \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u}, \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} \right\rangle = \langle \mathbf{v}, \mathbf{v} \rangle - 2\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \langle \mathbf{v}, \mathbf{u} \rangle + \left(\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \right)^2 \langle \mathbf{u}, \mathbf{u} \rangle = \langle \mathbf{v}, \mathbf{v} \rangle - \frac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{\langle \mathbf{u}, \mathbf{u} \rangle}$$

.

Adding these two together, we get $\dfrac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{\langle \mathbf{u}, \mathbf{u} \rangle} + \langle \mathbf{v}, \mathbf{v} \rangle - \dfrac{\langle \mathbf{u}, \mathbf{v} \rangle^2}{\langle \mathbf{u}, \mathbf{u} \rangle} = \langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{v}\|_2^2$, so $\|\mathbf{v}\|_2^2 = \|\text{proj}_{\mathbf{u}}(\mathbf{v})\|_2^2 + \|\text{perp}_{\mathbf{u}}(\mathbf{v})\|_2^2$.

5.7.2 For complex vector spaces

The inner product for real vectors is symmetric, and thus, $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$; however, the inner product for complex numbers is conjugate symmetric and sesquilinear, so

$$\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle^*, \ \langle \mathbf{v}, \alpha \mathbf{u} \rangle = \alpha \langle \mathbf{v}, \mathbf{u} \rangle \text{ but } \langle \alpha \mathbf{v}, \mathbf{u} \rangle = \alpha^* \langle \mathbf{v}, \mathbf{u} \rangle .$$

Thus,

$$\begin{aligned}
\langle \mathbf{v} - \alpha \mathbf{u}, \mathbf{v} - \alpha \mathbf{u} \rangle &= \langle \mathbf{v}, \mathbf{v} \rangle - \langle \mathbf{v}, \alpha \mathbf{u} \rangle - \langle \alpha \mathbf{u}, \mathbf{v} \rangle + \langle \alpha \mathbf{u}, \alpha \mathbf{u} \rangle \\
&= \langle \mathbf{v}, \mathbf{v} \rangle - \alpha \langle \mathbf{v}, \mathbf{u} \rangle - \alpha^* \langle \mathbf{u}, \mathbf{v} \rangle + \alpha^* \alpha \langle \mathbf{u}, \mathbf{u} \rangle \\
&= \langle \mathbf{v}, \mathbf{v} \rangle - \alpha \langle \mathbf{v}, \mathbf{u} \rangle - \alpha^* \langle \mathbf{v}, \mathbf{u} \rangle^* + |\alpha|^2 \langle \mathbf{u}, \mathbf{u} \rangle \\
&= \langle \mathbf{v}, \mathbf{v} \rangle - \alpha \langle \mathbf{v}, \mathbf{u} \rangle - \left(\alpha \langle \mathbf{v}, \mathbf{u} \rangle \right)^* + |\alpha|^2 \langle \mathbf{u}, \mathbf{u} \rangle \\
&= \langle \mathbf{v}, \mathbf{v} \rangle - 2\Re\mathrm{e}\left(\alpha \langle \mathbf{v}, \mathbf{u} \rangle \right) + |\alpha|^2 \langle \mathbf{u}, \mathbf{u} \rangle \\
&= \langle \mathbf{v}, \mathbf{v} \rangle - 2\Re\mathrm{e}(\alpha) \Re\mathrm{e}\left(\langle \mathbf{v}, \mathbf{u} \rangle \right) + 2\Im\mathrm{m}(\alpha) \Im\mathrm{m}\left(\langle \mathbf{v}, \mathbf{u} \rangle \right) + \Re\mathrm{e}(\alpha)^2 \langle \mathbf{u}, \mathbf{u} \rangle + \Im\mathrm{m}(\alpha)^2 \langle \mathbf{u}, \mathbf{u} \rangle
\end{aligned}$$

This is of the form $a - 2bx + 2cy + dx^2 + ey^2$ where $x = \Re\mathrm{e}(\alpha)$ and $y = \Im\mathrm{m}(\alpha)$. As there is no cross term (that is, there is no *xy* term), such a polynomial is minimized when it is minimized in both *x* and *y* independently.

$$\Re\mathrm{e}(\alpha) = \frac{\Re\mathrm{e}\left(\langle \mathbf{v}, \mathbf{u} \rangle \right)}{\langle \mathbf{u}, \mathbf{u} \rangle} \text{ and } \Im\mathrm{m}(\alpha) = -\frac{\Im\mathrm{m}\left(\langle \mathbf{v}, \mathbf{u} \rangle \right)}{\langle \mathbf{u}, \mathbf{u} \rangle}$$

Thus, $\alpha = \dfrac{\Re\mathrm{e}\left(\langle \mathbf{v}, \mathbf{u} \rangle \right) - j\Im\mathrm{m}\left(\langle \mathbf{v}, \mathbf{u} \rangle \right)}{\langle \mathbf{u}, \mathbf{u} \rangle} = \dfrac{\langle \mathbf{v}, \mathbf{u} \rangle^*}{\langle \mathbf{u}, \mathbf{u} \rangle} = \dfrac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle}$.

You may ask yourself: are $\dfrac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u}$ and $\dfrac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u}$ so different, and does it even matter? We can consider, for example,

$$\mathbf{u} = \begin{pmatrix} 1 + j \\ -2 - j \\ -2j \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} -3 - 2j \\ -1 + 4j \\ 1 - 3j \end{pmatrix},$$

where

$$\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} = \begin{pmatrix} \frac{5}{11} - j\frac{7}{11} \\ -\frac{4}{11} + j\frac{13}{11} \\ -\frac{12}{11} + j\frac{2}{11} \end{pmatrix} \text{ and } \left\| \mathbf{v} - \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} \right\|_2 = \sqrt{\frac{403}{11}}$$

and

$$\frac{\langle \mathbf{v},\mathbf{u}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u} = \begin{pmatrix} -\frac{7}{11} + j\frac{5}{11} \\ \frac{8}{11} - j \\ \frac{12}{11} + j\frac{2}{11} \end{pmatrix} \text{ and } \left\| \mathbf{v} - \frac{\langle \mathbf{v},\mathbf{u}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u} \right\|_2 = \sqrt{\frac{547}{11}}.$$

.

To demonstrate that $\text{proj}_{\mathbf{u}}(\mathbf{v}) \overset{def}{=} \frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u}$ is the correct choice, if we plot $\left\| \mathbf{v} - \left(\frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle} + \alpha + \beta j \right)\mathbf{u} \right\|_2$ for $-0.1 \le \alpha \le 0.1$ and $-0.1 \le \beta \le 0.1$, we see an apparent minimum when $\alpha = \beta = 0$, as seen in Figure 49 where $\sqrt{\frac{403}{11}} \approx 6.052798$.

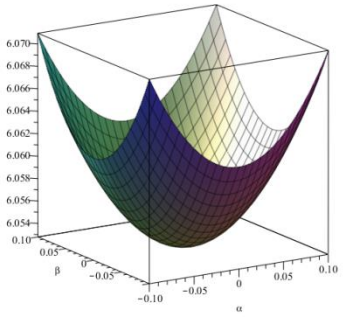


Figure 49. A plot of $\left\| \mathbf{v} - \left(\frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle} + \alpha + \beta j \right)\mathbf{u} \right\|_2$ for $-0.1 \le \alpha \le 0.1$ and $-0.1 \le \beta \le 0.1$.

5.7.3 Properties of projections

We will look at four properties of projections: they are idempotent and linear in its argument, the norm of the projection is bounded by the norm of the vector being projected, and the formula simplifies significantly for projections onto unit vectors.

Definition

A mapping is idempotent if $f(f(x)) = f(x)$ for all possible arguments x.

Theorem

The projection onto a vector **u** is idempotent, meaning $\text{proj}_{\mathbf{u}}(\text{proj}_{\mathbf{u}}(\mathbf{v})) = \text{proj}_{\mathbf{u}}(\mathbf{v})$.

Proof:

$$\begin{aligned}
\text{proj}_{\mathbf{u}}\left(\text{proj}_{\mathbf{u}}\left(\mathbf{v}\right)\right) &= \text{proj}_{\mathbf{u}}\left(\frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u}\right) \\
&= \frac{\left\langle \mathbf{u}, \frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u}\right\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u} \qquad \text{but } \frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle} \text{ is a scalar and the inner product is linear in the second operand} \\
&= \frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\frac{\langle \mathbf{u},\mathbf{u}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u} \\
&= \frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u} \\
&= \text{proj}_{\mathbf{u}}\left(\mathbf{v}\right)
\end{aligned}$$

Thus, the projection is idempotent. ∎

Demonstration of this theorem

Given $\mathbf{u} = \begin{pmatrix} 2 \\ -2 \\ 1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 4 \\ 5 \\ -1 \end{pmatrix}$, we see that $\text{proj}_{\mathbf{u}}\left(\mathbf{v}\right) = \frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u} = \frac{8-10-1}{9}\begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix} = -\frac{1}{3}\begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix} = \begin{pmatrix} -\frac{2}{3} \\ \frac{2}{3} \\ -\frac{1}{3} \end{pmatrix}$, but if we define $\mathbf{w} = \text{proj}_{\mathbf{u}}\left(\mathbf{v}\right) = \begin{pmatrix} -\frac{2}{3} \\ \frac{2}{3} \\ -\frac{1}{3} \end{pmatrix}$, the we also see that

$$\text{proj}_{\mathbf{u}}\left(\mathbf{w}\right) = \frac{\langle \mathbf{u},\mathbf{w}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u} = \frac{2\cdot\left(-\frac{2}{3}\right)+\left(-2\right)\cdot\frac{2}{3}+1\cdot\left(-\frac{1}{3}\right)}{9}\begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix} = -\frac{1}{3}\begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix} = \begin{pmatrix} -\frac{2}{3} \\ \frac{2}{3} \\ -\frac{1}{3} \end{pmatrix} = \mathbf{w}\cdot$$

Theorem

The projection onto a vector **u** is linear, meaning $\text{proj}_{\mathbf{u}}(\alpha\mathbf{v}+\beta\mathbf{w}) = \alpha\,\text{proj}_{\mathbf{u}}(\mathbf{v}) + \beta\,\text{proj}_{\mathbf{u}}(\mathbf{w})$.

Proof:

$$\begin{aligned}\text{proj}_{\mathbf{u}}(\alpha\mathbf{v}+\beta\mathbf{w}) &= \frac{\langle \mathbf{u}, \alpha\mathbf{v}+\beta\mathbf{w}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u} \quad \text{but the inner product is linear in its second operand}\\ &= \frac{\alpha\langle \mathbf{u},\mathbf{v}\rangle + \beta\langle \mathbf{u},\mathbf{w}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u}\\ &= \alpha\frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u} + \beta\frac{\langle \mathbf{u},\mathbf{w}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u}\\ &= \alpha\,\text{proj}_{\mathbf{u}}(\mathbf{v}) + \beta\,\text{proj}_{\mathbf{u}}(\mathbf{w})\end{aligned}$$

Therefore, the projection is linear. ∎

Theorem

The projection satisfies $\|\text{proj}_{\mathbf{u}}(\mathbf{v})\|_2 \le \|\mathbf{v}\|_2$ and $\|\text{proj}_{\mathbf{u}}(\mathbf{v})\|_2 = \|\mathbf{v}\|_2$ if and only if $\mathbf{v} = \alpha\mathbf{u}$.

Proof:

By the Pythagorean theorem,

$$\|\mathbf{v}\|_2^2 = \|\text{proj}_{\mathbf{u}}(\mathbf{v})\|_2^2 + \|\text{perp}_{\mathbf{u}}(\mathbf{v})\|_2^2.$$

Therefore, $\|\mathbf{v}\|_2^2 \le \|\text{proj}_{\mathbf{u}}(\mathbf{v})\|_2^2$. If these are equal, the perpendicular component is zero, so $\mathbf{v} = \frac{\langle \mathbf{u},\mathbf{v}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u}$. ∎

Demonstration of this theorem

Consider $\mathbf{u} = \begin{pmatrix}6\\8\end{pmatrix}$. We see that $\mathbf{v}_1 = \begin{pmatrix}-8\\1\end{pmatrix}$ is not a scalar multiple of **u**, and that $\text{proj}_{\mathbf{u}}(\mathbf{v}_1) = \begin{pmatrix}-2.4\\-3.2\end{pmatrix}$ and that $\|\text{proj}_{\mathbf{u}}(\mathbf{v}_1)\|_2 = \sqrt{(-4.2)^2 + 0^2 + 1.4^2} \approx 4.427$ while $\|\mathbf{v}_1\|_2 = \sqrt{(-4)^2 + 1^2 + 2^2} = \sqrt{21} \approx 4.583$ and we note that

Consider $\mathbf{u} = \begin{pmatrix}3\\0\\-1\end{pmatrix}$. We see that $\mathbf{v}_1 = \begin{pmatrix}-4\\1\\2\end{pmatrix}$ is not a scalar multiple of **u**, and that $\text{proj}_{\mathbf{u}}(\mathbf{v}_1) = \begin{pmatrix}-4.2\\0\\1.4\end{pmatrix}$ and that $\|\text{proj}_{\mathbf{u}}(\mathbf{v}_1)\|_2 = \sqrt{(-4.2)^2 + 0^2 + 1.4^2} \approx 4.427$ while $\|\mathbf{v}_1\|_2 = \sqrt{(-4)^2 + 1^2 + 2^2} = \sqrt{21} \approx 4.583$ and we note that

Theorem

If $\hat{\mathbf{v}}$ is a unit vector, then $\text{proj}_{\hat{\mathbf{v}}}(\mathbf{u}) = \langle \hat{\mathbf{v}}, \mathbf{u} \rangle \hat{\mathbf{v}}$.

Proof:

If $\hat{\mathbf{v}}$ is a unit vector, then by definition $\|\hat{\mathbf{v}}\|_2 = 1$, so $\langle \hat{\mathbf{v}}, \hat{\mathbf{v}} \rangle = \|\hat{\mathbf{v}}\|_2^2 = 1^2 = 1$, hence

$$\text{proj}_{\hat{\mathbf{v}}}(\mathbf{u}) = \frac{\langle \hat{\mathbf{v}}, \mathbf{u} \rangle}{\langle \hat{\mathbf{v}}, \hat{\mathbf{v}} \rangle} \hat{\mathbf{v}} = \frac{\langle \hat{\mathbf{v}}, \mathbf{u} \rangle}{1} \hat{\mathbf{v}} = \langle \hat{\mathbf{v}}, \mathbf{u} \rangle \hat{\mathbf{v}} . \blacksquare$$

Given a unit vector in $\mathbf{F}^n$, the calculation of $\langle \hat{\mathbf{v}}, \mathbf{u} \rangle \hat{\mathbf{v}}$ requires only $2n$ multiplications and $n - 1$ additions, whereas if the vector $\mathbf{v}$ is not normalized, then $\frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v}$ requires $3n + 1$ multiplications and $2n - 2$ additions. Consequently, the execution time is approximation 40 % faster if $\mathbf{v}$ is a unit vector.

Demonstration of this theorem

Consider $\hat{\mathbf{v}} = \begin{pmatrix} -0.6 \\ 0.8 \end{pmatrix}$ and $\mathbf{u} = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$. As this is a unit vector,

$$\text{proj}_{\hat{\mathbf{v}}}(\mathbf{u}) = \langle \hat{\mathbf{v}}, \mathbf{u} \rangle \hat{\mathbf{v}} = (-0.6 \cdot 3 + 0.8 \cdot 6) \hat{\mathbf{v}} = 3\hat{\mathbf{v}} = \begin{pmatrix} -1.8 \\ 2.4 \end{pmatrix}.$$

Problems

1. Demonstrate that the projection is idempotent by explicitly calculating $\text{proj}_{\mathbf{u}}(\mathbf{v})$ and $\text{proj}_{\mathbf{u}}\left(\text{proj}_{\mathbf{u}}(\mathbf{v})\right)$ for the vectors $\mathbf{u} = \begin{pmatrix} 2 \\ -3 \\ 4 \\ 1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 3 \\ 4 \\ -1 \\ -2 \end{pmatrix}$.

2. Demonstrate that the projection is idempotent by explicitly calculating $\text{proj}_{\mathbf{u}}(\mathbf{v})$ and $\text{proj}_{\mathbf{u}}\left(\text{proj}_{\mathbf{u}}(\mathbf{v})\right)$ for the vectors $\mathbf{u} = \begin{pmatrix} 2 \\ 4 \\ -3 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}$.

3. Demonstrate that the projection is linear by calculating the projection of $2 \cdot \begin{pmatrix} 2 \\ 3 \end{pmatrix} + 3 \cdot \begin{pmatrix} 4 \\ 5 \end{pmatrix}$ onto the vector $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ in two ways.

4. Demonstrate that the projection is linear by calculating the projection of $4 \cdot \begin{pmatrix} -2 \\ 1 \end{pmatrix} + 5 \cdot \begin{pmatrix} 3 \\ -1 \end{pmatrix}$ onto the vector $\begin{pmatrix} 2 \\ 3 \end{pmatrix}$ in two ways.

5. Demonstrate that the projection of a vector is shorter with respect to the 2-norm by finding the norms of $\mathbf{u} = \begin{pmatrix} 2 \\ 4 \\ -6 \end{pmatrix}$ and the projection of $\mathbf{u}$ onto the vectors $\mathbf{v}_1 = \begin{pmatrix} 2 \\ 4 \\ 1 \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} -1 \\ -2 \\ 3 \end{pmatrix}$.

6. Demonstrate that the projection of a vector is shorter with respect to the 2-norm by finding the norms of $\mathbf{u} = \begin{pmatrix} 1 \\ 3 \\ -2 \end{pmatrix}$ and the projection of $\mathbf{u}$ onto the vectors $\mathbf{v}_1 = \begin{pmatrix} -2 \\ -6 \\ 4 \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$.

Solutions

1. $\text{proj}_{\mathbf{u}}(\mathbf{v}) = \dfrac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} = \dfrac{6 - 12 - 4 - 2}{4 + 9 + 16 + 1} \mathbf{u} = -\dfrac{12}{30} \begin{pmatrix} 2 \\ -3 \\ 4 \\ 1 \end{pmatrix} = -\dfrac{2}{5} \begin{pmatrix} 2 \\ -3 \\ 4 \\ 1 \end{pmatrix} = -\dfrac{2}{5} \begin{pmatrix} -\frac{4}{5} \\ \frac{6}{5} \\ -\frac{8}{5} \\ -\frac{2}{5} \end{pmatrix}$.

3. $2\cdot\begin{pmatrix}2\\3\end{pmatrix}+3\cdot\begin{pmatrix}4\\5\end{pmatrix}=\begin{pmatrix}16\\21\end{pmatrix}$, thus $\text{proj}_{\mathbf{u}}\left(\begin{pmatrix}2\\3\end{pmatrix}\right)=\frac{2+6}{1+5}\mathbf{u}=\frac{4}{3}\begin{pmatrix}1\\2\end{pmatrix}=\begin{pmatrix}\frac{4}{3}\\\frac{8}{3}\end{pmatrix}$, $\text{proj}_{\mathbf{u}}\left(\begin{pmatrix}4\\5\end{pmatrix}\right)=\frac{4+10}{1+5}\mathbf{u}=\frac{7}{3}\begin{pmatrix}1\\2\end{pmatrix}=\begin{pmatrix}\frac{7}{3}\\\frac{14}{3}\end{pmatrix}$ and finally $\text{proj}_{\mathbf{u}}\left(\begin{pmatrix}16\\21\end{pmatrix}\right)=\frac{16+42}{1+5}\mathbf{u}=\frac{29}{3}\begin{pmatrix}1\\2\end{pmatrix}=\begin{pmatrix}\frac{29}{3}\\\frac{58}{3}\end{pmatrix}$ and $2\cdot\begin{pmatrix}\frac{4}{3}\\\frac{8}{3}\end{pmatrix}+3\cdot\begin{pmatrix}\frac{7}{3}\\\frac{14}{3}\end{pmatrix}=\begin{pmatrix}\frac{29}{3}\\\frac{58}{3}\end{pmatrix}$.

5. $\text{proj}_{\mathbf{v}_1}(\mathbf{u})=\frac{4+16-6}{4+16+1}\cdot\begin{pmatrix}2\\4\\1\end{pmatrix}=\frac{2}{3}\cdot\begin{pmatrix}2\\4\\1\end{pmatrix}=\begin{pmatrix}\frac{4}{3}\\\frac{8}{3}\\\frac{2}{3}\end{pmatrix}$ $\mathbf{u}=\begin{pmatrix}2\\4\\-6\end{pmatrix}$ and $\text{proj}_{\mathbf{v}_2}(\mathbf{u})=\frac{-2-8-18}{1+4+9}\cdot\begin{pmatrix}-1\\-2\\3\end{pmatrix}=-\frac{28}{14}\cdot\begin{pmatrix}-1\\-2\\3\end{pmatrix}=\begin{pmatrix}2\\4\\-6\end{pmatrix}$. In the first case, $\|\mathbf{u}\|_2=\sqrt{4+16+36}=\sqrt{56}$ and $\|\text{proj}_{\mathbf{v}_1}(\mathbf{u})\|_2=\sqrt{\frac{16}{9}+\frac{64}{9}+\frac{4}{9}}=\frac{\sqrt{84}}{3}<\sqrt{56}$; and in the second case, $\mathbf{u}\,||\,\mathbf{v}_2$.

5.7.4 Projections in other inner product spaces

Let us consider the space of square-integrable functions on the interval $[a,b]$. The projection of any function f onto the constant function is

$$\begin{aligned}\text{proj}_1(f) &= \frac{\langle 1, f\rangle}{\langle 1,1\rangle}1 \\ &= \frac{\int_a^b 1\cdot f(x)\,dx}{\int_a^b 1^2\,dx}\cdot 1 \\ &= \frac{1}{b-a}\int_a^b f(x)\,dx\end{aligned}$$

which is the average value of the function on the interval, usually written as $\bar{f}$. This means, the average value of a function is that value that minimizes $\left\|f-\bar{f}\right\|_2 = \sqrt{\int_a^b \left|f(t)-\bar{f}\right|^2 dt}$. You will also notice that this projection is orthogonal to the perpendicular component:

$$\begin{aligned}\left\langle f-\bar{f}, \bar{f}\right\rangle &= \int_a^b \left(f-\bar{f}\right)(t)\,\bar{f}dt \\ &= \int_a^b \left(f(t)-\bar{f}\right)\bar{f}dt \\ &= \int_a^b \left(\bar{f}\,f(t)-\bar{f}^2\right)dt \\ &= \int_a^b \bar{f}\,f(t)\,dt - \int_a^b \bar{f}^2 dt \\ &= \bar{f}\int_a^b f(t)\,dt - \bar{f}^2(b-a)\end{aligned}$$

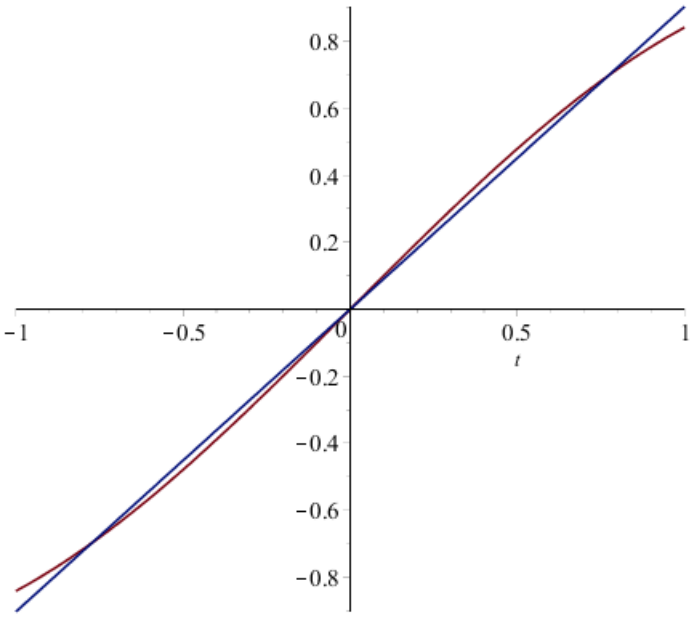
But, by definition, $\bar{f} = \frac{1}{b-a}\int_a^b f(t)\,dt$ so $\int_a^b f(t)\,dt = \bar{f}(b-a)$, so replacing this, we have

$$\begin{aligned}&= \bar{f}\cdot\bar{f}(b-a) - \bar{f}^2(b-a) \\ &= 0\end{aligned}$$

Alternatively, what is the best approximation of the sine function by a multiple of the polynomial $p : t \mapsto t$ t on the interval $[-1,1]$?

$$\begin{aligned}
\text{proj}_p(\sin) &= \frac{\langle p, \sin \rangle}{\langle p, p \rangle} \cdot p \\
&= \frac{\int_{-1}^{1} t \cdot \sin(t)\, dt}{\int_{-1}^{1} t^2 dt} \cdot p \\
&= \frac{2(\sin(1) - \cos(1))}{\frac{2}{3}} \cdot p \\
&= 3(\sin(1) - \cos(1))\, p \\
&\approx 0.9035\, p
\end{aligned}$$

So, the best approximation is approximately $t \mapsto 0.9035t$. To see this, consider the next image which shows both the sine function and the projection onto *p*.



One very important projection you will use in your courses are projections onto the

We will now write a Matlab function to do the projection. We must

1. check that the arguments are numeric vectors of the same length,
2. allow a second argument 'unit' that indicates that the vector being projected onto is a unit vector, thereby removing the requirement to divide through by $\|\mathbf{v}\|_2$,
3. throw an exception if there is anything other than 2 or 3 arguments,
4. calculate the projection, and
5. if there is a second output argument, calculate the perpendicular component.

```
function [pro, perp] = proj( v, u, opts )
% PROJ  Project the vector V onto the vector U.
%   PRO = PROJ(V, U)              project V onto U.
%   PRO = PROJ(V, U, 'unit')      project V onto U where U is a unit vector
%   [PRO, PERP] = PROJ(V, U, ...) assign to PERP = V - PRO

    if    ~isvector( v ) || ~isnumeric( u ) ...
       || ~isvector( v ) || ~isnumeric( u ) ...
       ||    length( v ) ~= length( u )
            throw( MException( 'linalg:proj', ['The arguments U and V '...
                       'should be numeric vectors of the same length'] ) );
    end

    % If there are two input arguments, it is the usual projection;
    % if there are three input arguments, and the third is the string
    % 'unit', it is implied that

    if nargin == 2
        uu = norm( u, 2 )^2;   % This is faster than u'*u

        if norm2u == 0
            pro = u;
        else
            pro = ((u'*v)/uu) * u;
        end
    elseif nargin == 3
        validatestring( opts, {'unit'} );

        pro = (u'*v) * u;
    else
        throw( MException( 'linalg:proj', ...
                      'Expecting 2 or 3 arguments, but got %d', nargin ) );
    end

    % If there is a second output argument, assign to it the perpendicular
    % component, namely, u - proj( u, v )
    if nargout == 2
        perp = v - pro;
    elseif nargout > 2
        throw( MException( 'linalg:proj', 'Too many output arguments.' ) );
    end
end
```

As before, we must save this to the file named `proj.m`.

Recall that the complex exponential functions on [0, 2π] defined as

$$e^{j2\pi nt}$$

for $n = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$ are orthogonal. We could project an arbitrary function defined on this interval onto each of these, so

$$\begin{aligned}
\text{proj}_{e^{j2\pi n\bullet}}(f)(t) &= \frac{\left\langle e^{j2\pi n\bullet}, \sin \right\rangle}{\left\langle e^{j2\pi n\bullet}, e^{j2\pi n\bullet} \right\rangle} \cdot e^{j2\pi nt} \\
&= \frac{\int_0^{2\pi} \left(e^{j2\pi nt}\right)^* \sin(t)\, dt}{\int_0^{2\pi} \left(e^{j2\pi nt}\right)^* e^{j2\pi nt}\, dt} \cdot e^{j2\pi nt} \\
&= \frac{\int_0^{2\pi} e^{-j2\pi nt} \sin(t)\, dt}{\int_0^{2\pi} 1\, dt} \cdot e^{j2\pi nt} \\
&= \frac{\int_0^{2\pi} e^{-j2\pi nt} \sin(t)\, dt}{2\pi} \cdot e^{j2\pi nt}
\end{aligned}$$

The coefficient $\dfrac{\int_0^{2\pi} e^{-j2\pi nt} \sin(t)\, dt}{2\pi}$ is the complex Fourier coefficient for $e^{j2\pi nt}$.

5.7.5 Best approximations a constant-valued vector

A constant-valued vector is any vector that is any vector that has the same value in each entry. All such vectors are scalar multiples of that vector containing the multiplicative identity of the field in question. For finite-dimensional vector spaces, all such vectors are scalar multiples of

$$\mathbf{1}_n = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

For finite-dimensional vector spaces, this equals:

$$\text{proj}_{\mathbf{1}_n}(\mathbf{v}) = \frac{\langle \mathbf{1}_n, \mathbf{v} \rangle}{\langle \mathbf{1}_n, \mathbf{1}_n \rangle} \mathbf{1}_n = \frac{\sum_{k=1}^{n} 1 \cdot v_k}{\sum_{k=1}^{n} 1 \cdot 1} \mathbf{1}_n = \left(\frac{1}{n} \sum_{k=1}^{n} v_k \right) \mathbf{1}_n .$$

The scalar $\text{mean}(\mathbf{v}) \stackrel{def}{=} \frac{1}{n} \sum_{k=1}^{n} v_k$ is the *average value* or *mean value* of the entries of the vector. Note that for functions, if a function is defined on a finite interval [*a*, *b*] with the corresponding inner product, the average value becomes:

$$\text{proj}_{\vec{1}}(f) = \frac{\langle \vec{1}, f \rangle}{\langle \vec{1}, \vec{1} \rangle} \vec{1} = \frac{\int_a^b 1 \cdot f(x)\,dx}{\int_a^b 1 \cdot 1\,dx} \vec{1} = \left(\frac{1}{b-a} \int_a^b f(x)\,dx \right) \vec{1} .$$

The scalar $\text{mean}(f) \stackrel{def}{=} \frac{1}{b-a} \int_a^b f(x)\,dx$ is the *average value* or *mean value* of the function *f*.

5.8 Cauchy–Bunyakovsky–Schwarz inequality

Next, we will demonstrate that there is an upper bound to the inner product, and that upper bound is only satisfied if the vectors are scalar multiples of each other. This theorem is called the Cauchy-Bunyakovsky-Schwarz inequality, which says that inner product is bounded by the magnitudes of the vectors:

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 .$$

Proof:

If either $\mathbf{u} = \mathbf{o}$ or $\mathbf{v} = \mathbf{o}$, then $|\langle \mathbf{u}, \mathbf{v} \rangle| = 0 = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2$. Otherwise, we can always write one as the projection onto and the perpendicular component with respect to the other:

$$\mathbf{u} = \text{proj}_{\mathbf{v}}(\mathbf{u}) + \text{perp}_{\mathbf{v}}(\mathbf{u}),$$

where the two components are perpendicular to each other. Thus, by the Pythagorean theorem, it follows that

$$\|\mathbf{u}\|_2^2 = \|\text{proj}_{\mathbf{v}}(\mathbf{u})\|_2^2 + \|\text{perp}_{\mathbf{v}}(\mathbf{u})\|_2^2 .$$

Because $\left\|\text{perp}_{\mathbf{v}}(\mathbf{u})\right\|_2^2 \geq 0$, it follows that $\|\mathbf{u}\|_2^2 \geq \left\|\text{proj}_{\mathbf{v}}(\mathbf{u})\right\|_2^2$. Now, using the definition of the projection, we have that

$$\|\mathbf{u}\|_2^2 \geq \left\|\frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\langle \mathbf{v}, \mathbf{v}\rangle}\mathbf{v}\right\|_2^2 = \left|\frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\langle \mathbf{v}, \mathbf{v}\rangle}\right|^2 \|\mathbf{v}\|_2^2 = \frac{\left|\langle \mathbf{u}, \mathbf{v}\rangle\right|^2}{\langle \mathbf{v}, \mathbf{v}\rangle^2}\|\mathbf{v}\|_2^2,$$

as $\langle \mathbf{v}, \mathbf{v}\rangle \geq 0$. However, because $\langle \mathbf{v}, \mathbf{v}\rangle = \|\mathbf{v}\|_2^2$, it follows that

$$\|\mathbf{u}\|_2^2 \geq \frac{\left|\langle \mathbf{u}, \mathbf{v}\rangle\right|^2}{\|\mathbf{v}\|_2^2},$$

and multiplying both sides by $\|\mathbf{v}\|_2^2$, we have

$$\|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2 \geq \left|\langle \mathbf{u}, \mathbf{v}\rangle\right|^2,$$

and taking the square root of both sides, we have the desired result:

$$\|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \geq \left|\langle \mathbf{u}, \mathbf{v}\rangle\right| . \blacksquare$$

Demonstration of this theorem
This theorem essentially says that the inner product is bounded by the product of the norms of the two vectors. Consequently, if $\|\mathbf{u}\|_2 = 3$ and $\|\mathbf{v}\|_2 = 2$, we are guaranteed that $\left|\langle \mathbf{u}, \mathbf{v}\rangle\right| \leq 3 \cdot 2 = 6$.

Problems

1. Demonstrate the Cauchy–Bunyakovsky–Schwarz inequality with the vectors $\mathbf{u} = \begin{pmatrix} 2 \\ -3 \\ 4 \\ 1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 3 \\ 4 \\ -1 \\ -2 \end{pmatrix}$.

2. Demonstrate the Cauchy–Bunyakovsky–Schwarz inequality with the vectors $\mathbf{u} = \begin{pmatrix} 2 \\ 4 \\ -3 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}$.

3. Prove that $\|\text{proj}_{\mathbf{v}}(\mathbf{u})\|_2 \leq \|\mathbf{u}\|_2$.

4. Under what conditions does $\|\mathbf{u}\|_2 \|\mathbf{v}\|_2 = |\langle \mathbf{u}, \mathbf{v} \rangle|$?

Solutions

1. $\langle \mathbf{u}, \mathbf{v} \rangle = 6 - 12 - 4 - 2 = -12$, $\|\mathbf{u}\|_2 = \sqrt{4+9+16+1} = \sqrt{30}$ and $\|\mathbf{v}\|_2 = \sqrt{9+16+1+4} = \sqrt{30}$, and we note that $|\langle \mathbf{u}, \mathbf{v} \rangle| = 12 < 30 = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2$.

3. $\left\|\text{proj}_{\mathbf{v}}(\mathbf{u})\right\|_2 = \left\| \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\langle \mathbf{v}, \mathbf{v} \rangle} \mathbf{v} \right\|_2 = \frac{|\langle \mathbf{v}, \mathbf{u} \rangle|}{\|\mathbf{v}\|_2^2} \|\mathbf{v}\|_2 = \frac{|\langle \mathbf{v}, \mathbf{u} \rangle|}{\|\mathbf{v}\|_2}$ but $\|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \geq |\langle \mathbf{u}, \mathbf{v} \rangle|$ so $\|\mathbf{u}\|_2 \geq \frac{|\langle \mathbf{u}, \mathbf{v} \rangle|}{\|\mathbf{v}\|_2}$, so $\|\text{proj}_{\mathbf{v}}(\mathbf{u})\|_2 \leq \|\mathbf{u}\|_2$.

5.9 Angle between vectors

You will recall from secondary school that for 2- and 3-dimensional vectors the angle between two vectors can be found from

$$\langle \mathbf{u}, \mathbf{v} \rangle = \|\mathbf{u}\|_2 \|\mathbf{v}\|_2 \cos(\theta)$$

where θ is the angle between the two vectors. Because of the Cauchy-Bunyakovsky-Schwarz inequality, we are guaranteed that, so long as we use the norm induced by the inner product, it will always be true that

$$-1 \leq \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2} \leq 1.$$

Consequently, we will define for any vectors and inner product

$$\theta = \measuredangle(\mathbf{u}, \mathbf{v}) = \arccos\left(\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2} \right),$$

which returns a value in $[0, \pi]$ where

1. 0 indicates that the vectors are positive scalar multiples of each other,
2. values between 0 and $\frac{\pi}{2}$ indicate that the vectors are *pointing* in approximately the same direction,

3. $\frac{\pi}{2}$ indicates that the two vectors are orthogonal,
4. values between $\frac{\pi}{2}$ and π indicate that the vectors are *pointing* in approximately opposite directions, and
5. π indicates the vectors are negative scalar multiples of each other.

If the angle between two vectors is a right angle (that is, $\frac{\pi}{2}$ or 90°), we will say that the vectors are *orthogonal*[11].

Application of the angle

If **u** and **v** are vectors on the sphere of radius *r*, then $r\measuredangle(\mathbf{u},\mathbf{v})$ is the distance between the two points along the surface of the sphere. For example, on Earth, the city of Ottawa is at 45° 25′ N 75° 41′ W while the city of Mecca is at 21° 25′ N 39° 49′ E. As the radius of the earth is 6371 km, this places the two cities at

$$\mathbf{v}_{\text{Ottawa}} = 6371\begin{pmatrix}\cos(-75°\ 41')\cos(45°\ 25')\\ \sin(-75°\ 41')\cos(45°\ 25')\\ \sin(45°\ 25')\end{pmatrix} \approx \begin{pmatrix}1106\\ -4333\\ 4538\end{pmatrix}$$

and

$$\mathbf{v}_{\text{Mecca}} = 6371\begin{pmatrix}\cos(39°\ 49')\cos(21°\ 25')\\ \sin(39°\ 49')\cos(21°\ 25')\\ \sin(21°\ 25')\end{pmatrix} \approx \begin{pmatrix}4556\\ 3797\\ 2326\end{pmatrix}.$$

The angle between the two cities is therefore

$$\theta = \measuredangle(\mathbf{v}_{\text{Ottawa}}, \mathbf{v}_{\text{Mecca}}) = \arccos\left(\frac{\langle \mathbf{v}_{\text{Ottawa}}, \mathbf{v}_{\text{Mecca}} \rangle}{\|\mathbf{v}_{\text{Ottawa}}\|_2 \|\mathbf{v}_{\text{Mecca}}\|_2}\right) = \arccos\left(\frac{-858077}{6371^2}\right) = 1.5919$$

and thus the distance is $1.5919 \cdot 6371\,\text{km} = 10142\,\text{km}$.

[11] The word orthogonal comes from Greek where *orth-* means "right" and *gonia* means "angle". You may recall the title of the "Greek Orthodox Church" where *doxa* means "belief", and thus "Orthodox" means "right belief". Similarly, the profession of *orthodontistry* deals with the adjustment of teeth so that they grow straight, or "right teeth". On the other hand, for example, a *pentagon* is a shape with "five angles".

There is no Matlab routine to calculate the angle between two vectors, so you must do this explicitly. The arccosine function in Matlab (and almost all standard mathematical libraries for most programming languages) is `acos`.

```
>> u = [1 2 3]';
>> v = [2 -1 4]';
>> acos( u'*v/norm(u)/norm(v) )
    ans =
       0.795602953484535
```

We shall now see the first signs that using floating-point numbers may not always give us the desired result.

From our definition, **u** and 2**u** should have an angle of 0, but when we calculate this, we get a complex number

```
>> acos( u'*(2*u)/norm(u)/norm(2*u) )
    ans =
      0 + 2.107342425544702e-008i
```

The first thing you must note is that this is not 0 + 2.10*j*, but 0 + 2.10 × 10^{-8} *j* or 0.0000000210*j*. Many novice Matlab users will see the mantissa but miss the exponent `e-008`. Thus, this number is close to, but not equal to zero. This happens because there is a small error in the calculation of the argument; the calculation of

```
>> u'*(2*u)/norm(u)/norm(2*u)
```

does not produce 1, but rather, the next largest floating-point number, approximately 1.000000000000000222. While complex analysis is beyond the scope of this course, the inverse cosine function is not real for arguments outside the range [–1, 1]. A similar issue occurs when we try to find the angle between **u** and –2**u**, which should be π, but is, for the same reason, slightly complex.

```
>> acos( u'*(-2*u)/norm(u)/norm(-2*u) )
ans =
      3.141592653589793 - 0.000000021073424i
```

A better solution, suggested by Roger Stafford, is

```
 angle = atan2( norm( cross( a, b ) ), u'*v )
```

This formula works, because

$$\frac{\|\mathbf{u}\times\mathbf{v}\|}{\langle\mathbf{u},\mathbf{v}\rangle}=\frac{\|\mathbf{u}\|\|\mathbf{v}\|\sin(\theta)}{\|\mathbf{u}\|\|\mathbf{v}\|\cos(\theta)}=\tan(\theta),$$

and because of the properties of the `atan2` function require that if the first argument is positive, the angle must be in [0, π].

This is a difficult formula to continually write, so we will take this as our first opportunity to introduce functions in Matlab.

```
function [theta] = vangle( u, v )
    theta = atan2( norm( cross( a, b ) ), u'*v );
end
```

We give the function a different name than just angle as there is already such a function in Matlab, and *vector angle* is a very Matlab approach to naming functions. If we were programming in Java, the appropriate name would be `vectorAngle(…)`.

For this function to work, it must be saved to the file `vangle.m` and it must exist in the directory **TO BE COMPLETED**.

5.10 The Gram-Schmidt algorithm for the orthogonalization of vectors

In this section, we will look at how we can take *n* vectors that are not necessarily mutually orthogonal, and where those vectors may not be normalized (norm equal to one), we will describe an algorithm that converts these *n* vectors into a set of *n* mutually orthogonal normalized vectors, or a set of *orthonormal* vectors. Recall that two vectors are orthogonal if neither vector contains information about the other.

You will recall that we may write a vector $\mathbf{u}$ as the sum of the its projection onto a vector $\mathbf{v}_1$ together with the perpendicular component of that projection, as shown in Figure 50.

$$\mathbf{u} = \text{proj}_{\mathbf{v}_1}(\mathbf{u}) + \text{perp}_{\mathbf{v}_1}(\mathbf{u}),$$

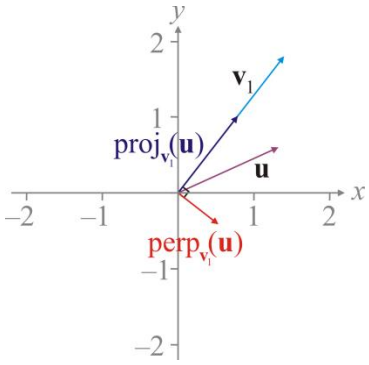


Figure 50. The projection of the vector $\mathbf{u}$ onto $\mathbf{v}_1$ and its perpendicular component.

Now, if we were to first normalize $\mathbf{v}_1$, the projection and the perpendicular components of $\mathbf{u}$ would remain unchanged, as shown in Figure 51, however, now, the formula for calculating the projection is simpler:

$$\mathbf{u} = \langle \hat{\mathbf{v}}_1, \mathbf{u} \rangle \hat{\mathbf{v}}_1 + \underbrace{\mathbf{u} - \langle \hat{\mathbf{v}}_1, \mathbf{u} \rangle \hat{\mathbf{v}}_1}_{\text{perp}_{\hat{\mathbf{v}}_1}(\mathbf{u})}.$$

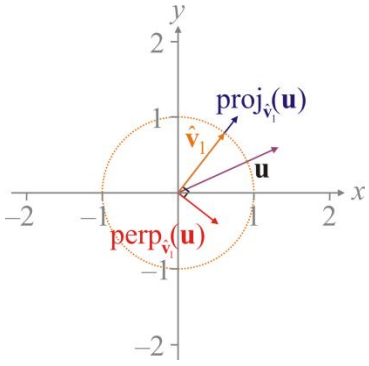


Figure 51. The projection of the vector $\mathbf{u}$ onto the unit vector $\hat{\mathbf{v}}_1$ and its perpendicular component.

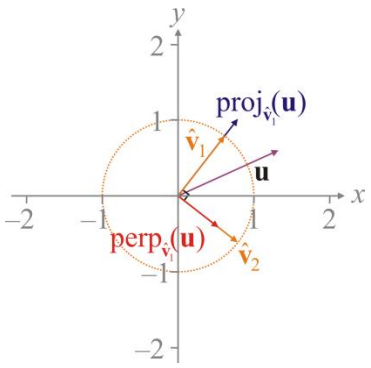
We can now normalize the perpendicular component, and let us call it

$$\hat{\mathbf{v}}_2 = \frac{\mathbf{u} - \langle \hat{\mathbf{v}}_1, \mathbf{u} \rangle \hat{\mathbf{v}}_1}{\left\| \mathbf{u} - \langle \hat{\mathbf{v}}_1, \mathbf{u} \rangle \hat{\mathbf{v}}_1 \right\|_2}.$$

Now, we can simply write

$$\mathbf{u} = \langle \hat{\mathbf{v}}_1, \mathbf{u} \rangle \hat{\mathbf{v}}_1 + \langle \hat{\mathbf{v}}_2, \mathbf{u} \rangle \hat{\mathbf{v}}_2 ,$$

as both are unit vectors.



Suppose now we have two vectors that are perpendicular, $\{\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2\}$. Now, these could be any two vectors that are perpendicular, not the ones we just defined. Suppose then that we have a third vector **u** that may not be parallel to either vector. You will note that we can now write

$$\mathbf{u} = \langle \hat{\mathbf{v}}_1, \mathbf{u} \rangle \hat{\mathbf{v}}_1 + \langle \hat{\mathbf{v}}_2, \mathbf{u} \rangle \hat{\mathbf{v}}_2 + \underbrace{\mathbf{u} - \left(\langle \hat{\mathbf{v}}_1, \mathbf{u} \rangle \hat{\mathbf{v}}_1 + \langle \hat{\mathbf{v}}_2, \mathbf{u} \rangle \hat{\mathbf{v}}_2 \right)}_{\text{Component of } \mathbf{u} \text{ perpendicular to both } \hat{\mathbf{v}}_1 \text{ and } \hat{\mathbf{v}}_2 .}$$

Our goal will be able to take any collection of vectors, and from this derive a collection of vectors that are mutually orthogonal to each other, all of which are unit vectors. We will call this set *orthonormal*[12]. Thus, given a set of *n* vectors,

$$\{\mathbf{u}_1, \ldots, \mathbf{u}_n\} ,$$

we require a set of vectors such that

$$\{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_n\}$$

forms a mutually orthogonal collection of vectors. Thus, given another vector **v**, we may find the best approximation of **v** as a linear combination of the vectors $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_n$ by calculating

$$\mathbf{v} \approx \langle \hat{\mathbf{u}}_1, \mathbf{v} \rangle \hat{\mathbf{u}}_1 + \cdots + \langle \hat{\mathbf{u}}_n, \mathbf{v} \rangle \hat{\mathbf{u}}_n .$$

Here is the algorithm in detail. Because of the danger of associating *equality* with *assignment*, we will use a right arrow ($\leftarrow$) to indicate assignment. Thus, you should read $a \leftarrow b$ as *"the value b is being assigned to the symbol (or variable) a*". Hence, we have the algorithm:

1. If $\mathbf{u}_1 = \mathbf{0}$, discard it, otherwise define $\hat{\mathbf{u}}_1 \leftarrow \dfrac{\mathbf{u}_1}{\|\mathbf{u}_1\|_2}$ (normalize $\mathbf{u}_1$).
2. Update $\mathbf{u}_2 \leftarrow \mathbf{u}_2 - \langle \hat{\mathbf{u}}_1, \mathbf{u}_2 \rangle \hat{\mathbf{u}}_1$ (make $\mathbf{u}_2$ perpendicular to $\hat{\mathbf{u}}_1$), and

   If $\mathbf{u}_2 = \mathbf{0}$, discard it, otherwise define $\hat{\mathbf{u}}_2 \leftarrow \dfrac{\mathbf{u}_2}{\|\mathbf{u}_2\|_2}$ (normalize $\mathbf{u}_2$).

[12] Both *ortho*gonal and *normal*ized.

3. Update $\mathbf{u}_3 \leftarrow \mathbf{u}_3 - \langle \hat{\mathbf{u}}_1, \mathbf{u}_3 \rangle \hat{\mathbf{u}}_1$ (make $\mathbf{u}_3$ perpendicular to $\hat{\mathbf{u}}_1$),
   Update $\mathbf{u}_3 \leftarrow \mathbf{u}_3 - \langle \hat{\mathbf{u}}_2, \mathbf{u}_3 \rangle \hat{\mathbf{u}}_2$ (make $\mathbf{u}_3$ perpendicular to $\hat{\mathbf{u}}_2$), and
   If $\mathbf{u}_3 = \mathbf{0}$, discard it, otherwise define $\hat{\mathbf{u}}_3 \leftarrow \dfrac{\mathbf{u}_3}{\|\mathbf{u}_3\|_2}$ (normalize $\mathbf{u}_3$).

In general, the kth step will be as follows:

k. Update $\mathbf{u}_k \leftarrow \mathbf{u}_k - \langle \hat{\mathbf{u}}_1, \mathbf{u}_k \rangle \hat{\mathbf{u}}_1$ (make $\mathbf{u}_k$ perpendicular to $\hat{\mathbf{u}}_1$),
   Update $\mathbf{u}_k \leftarrow \mathbf{u}_k - \langle \hat{\mathbf{u}}_2, \mathbf{u}_k \rangle \hat{\mathbf{u}}_2$ (make $\mathbf{u}_k$ perpendicular to $\hat{\mathbf{u}}_2$),
   ⋮
   Update $\mathbf{u}_k \leftarrow \mathbf{u}_k - \langle \hat{\mathbf{u}}_{k-1}, \mathbf{u}_k \rangle \hat{\mathbf{u}}_{k-1}$ (make $\mathbf{u}_k$ perpendicular to $\hat{\mathbf{u}}_{k-1}$), and
   If $\mathbf{u}_k = \mathbf{0}$, discard it, otherwise define $\hat{\mathbf{u}}_k \leftarrow \dfrac{\mathbf{u}_k}{\|\mathbf{u}_k\|_2}$ (normalize $\mathbf{u}_k$).

In general, if $\mathbf{u}_k$ ever equals the zero vector in this process, this indicates that all the information in that vector is already contained in the vectors $\mathbf{u}_1$ through $\mathbf{u}_{k-1}$.

Performing this process *n* times will produce a collection of up to *n* vectors that form an orthonormal set.

As an example of the Gram-Schmidt process, let us start with the four vectors

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 2 \\ 2 \\ 0 \\ 0 \end{pmatrix}, \mathbf{u}_3 = \begin{pmatrix} 2 \\ -1 \\ 3 \\ 6 \end{pmatrix}, \mathbf{u}_4 = \begin{pmatrix} 2 \\ 0 \\ 1 \\ -5 \end{pmatrix}.$$

These vectors are neither normalized nor are they mutually orthogonal. We begin with the first vector, which we need only normalize:

$$\hat{\mathbf{u}}_1 \leftarrow \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}$$

Next, we subtract from $\mathbf{u}_2$ the projection of $\mathbf{u}_2$ onto $\hat{\mathbf{u}}_1$:

$$\mathbf{u}_2 \leftarrow \mathbf{u}_2 - \langle \hat{\mathbf{u}}_1, \mathbf{u}_2 \rangle \hat{\mathbf{u}}_1 = \begin{pmatrix} 2 \\ 2 \\ 0 \\ 0 \end{pmatrix} - \left\langle \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \\ 0 \\ 0 \end{pmatrix} \right\rangle \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 0 \\ 0 \end{pmatrix} - 2 \cdot \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}$$

Now we normalize this vector:

$$\hat{\mathbf{u}}_2 \leftarrow \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix}.$$

Now, $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$ form a mutually orthogonal pair of vectors.

We will now proceed to $\mathbf{u}_3$. We subtract from $\mathbf{u}_3$ the projection of $\mathbf{u}_3$ onto $\hat{\mathbf{u}}_1$:

$$\mathbf{u}_3 \leftarrow \mathbf{u}_3 - \langle \hat{\mathbf{u}}_1, \mathbf{u}_3 \rangle \hat{\mathbf{u}}_1 = \begin{pmatrix} 2 \\ -1 \\ 3 \\ 6 \end{pmatrix} - \left\langle \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 2 \\ -1 \\ 3 \\ 6 \end{pmatrix} \right\rangle \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 3 \\ 6 \end{pmatrix} - 5 \cdot \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} -0.5 \\ -3.5 \\ 0.5 \\ 3.5 \end{pmatrix}.$$

Next, from this updated $\mathbf{u}_3$, subtract from it the projection onto $\hat{\mathbf{u}}_2$:

$$\mathbf{u}_3 \leftarrow \mathbf{u}_3 - \langle \hat{\mathbf{u}}_2, \mathbf{u}_3 \rangle \hat{\mathbf{u}}_2 = \begin{pmatrix} -0.5 \\ -3.5 \\ 0.5 \\ 3.5 \end{pmatrix} - \left\langle \begin{pmatrix} 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix}, \begin{pmatrix} -0.5 \\ -3.5 \\ 0.5 \\ 3.5 \end{pmatrix} \right\rangle \begin{pmatrix} 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix} = \begin{pmatrix} -0.5 \\ -3.5 \\ 0.5 \\ 3.5 \end{pmatrix} - (-4) \cdot \begin{pmatrix} 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix} = \begin{pmatrix} 1.5 \\ -1.5 \\ -1.5 \\ 1.5 \end{pmatrix}.$$

Now we normalize this vector:

$$\hat{\mathbf{u}}_3 \leftarrow \frac{1}{3}\begin{pmatrix} 1.5 \\ -1.5 \\ -1.5 \\ 1.5 \end{pmatrix} = \begin{pmatrix} 0.5 \\ -0.5 \\ -0.5 \\ 0.5 \end{pmatrix}.$$

Now, $\hat{\mathbf{u}}_1$, $\hat{\mathbf{u}}_2$ and $\hat{\mathbf{u}}_3$ form a mutually orthogonal set of three vectors.

Finally, we proceed to $\mathbf{u}_4$. We subtract from $\mathbf{u}_4$ the projection of $\mathbf{u}_4$ onto $\hat{\mathbf{u}}_1$:

$$\mathbf{u}_4 \leftarrow \mathbf{u}_4 - \langle \hat{\mathbf{u}}_1, \mathbf{u}_4 \rangle \hat{\mathbf{u}}_1 = \begin{pmatrix} 2 \\ 1 \\ 0 \\ -5 \end{pmatrix} - \left\langle \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \\ -5 \end{pmatrix} \right\rangle \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \\ -5 \end{pmatrix} - (-1) \cdot \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 0.5 \\ 1.5 \\ -4.5 \end{pmatrix}.$$

Next, from this updated $\mathbf{u}_4$, subtract from it the projection onto $\hat{\mathbf{u}}_2$:

$$\mathbf{u}_3 \leftarrow \mathbf{u}_3 - \langle \hat{\mathbf{u}}_2, \mathbf{u}_3 \rangle \hat{\mathbf{u}}_2 = \begin{pmatrix} 2.5 \\ 0.5 \\ 1.5 \\ -4.5 \end{pmatrix} - \left\langle \begin{pmatrix} 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix}, \begin{pmatrix} 2.5 \\ 0.5 \\ 1.5 \\ -4.5 \end{pmatrix} \right\rangle \begin{pmatrix} 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 0.5 \\ 1.5 \\ -4.5 \end{pmatrix} - 3 \cdot \begin{pmatrix} 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -3 \\ 3 \end{pmatrix}.$$

Next, from this updated $\mathbf{u}_4$, subtract from it the projection onto $\hat{\mathbf{u}}_3$:

$$\mathbf{u}_4 \leftarrow \mathbf{u}_4 - \langle \hat{\mathbf{u}}_3, \mathbf{u}_4 \rangle \hat{\mathbf{u}}_3 = \begin{pmatrix} 1 \\ -1 \\ -3 \\ 3 \end{pmatrix} - \left\langle \begin{pmatrix} 0.5 \\ -0.5 \\ -0.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ -3 \\ 3 \end{pmatrix} \right\rangle \begin{pmatrix} 0.5 \\ -0.5 \\ -0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -3 \\ 3 \end{pmatrix} - (-2) \cdot \begin{pmatrix} 0.5 \\ -0.5 \\ -0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \\ 2 \\ -2 \end{pmatrix}.$$

Now we normalize this vector:

$$\hat{\mathbf{u}}_4 \leftarrow \frac{1}{4} \begin{pmatrix} 2 \\ -2 \\ 2 \\ -2 \end{pmatrix} = \begin{pmatrix} 0.5 \\ -0.5 \\ 0.5 \\ -0.5 \end{pmatrix}.$$

Thus, we now have a collection of four orthonormal vectors, meaning that each is normalized, and any pair of vectors are orthogonal to each other.

$$\hat{\mathbf{u}}_1 = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix}, \hat{\mathbf{u}}_2 = \begin{pmatrix} 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{pmatrix}, \hat{\mathbf{u}}_3 = \begin{pmatrix} 0.5 \\ -0.5 \\ -0.5 \\ 0.5 \end{pmatrix}, \hat{\mathbf{u}}_4 = \begin{pmatrix} 0.5 \\ -0.5 \\ 0.5 \\ -0.5 \end{pmatrix}.$$

Problems

1. Perform the Gram-Schmidt algorithm on the vectors $\begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}$ and $\begin{pmatrix} 3 \\ 21 \\ 0 \end{pmatrix}$.

2. Perform the Gram-Schmidt algorithm on the vectors $\begin{pmatrix} 1 \\ 12 \\ 12 \end{pmatrix}$ and $\begin{pmatrix} 15 \\ 32 \\ 39 \end{pmatrix}$.

3. Perform the Gram-Schmidt algorithm on the vectors $\begin{pmatrix} 1 \\ 4 \\ 8 \end{pmatrix}$, $\begin{pmatrix} 6 \\ 15 \\ 12 \end{pmatrix}$ and $\begin{pmatrix} -19 \\ -22 \\ -17 \end{pmatrix}$.

4. Perform the Gram-Schmidt algorithm on the vectors $\begin{pmatrix} -9 \\ 2 \\ -6 \end{pmatrix}$, $\begin{pmatrix} 2 \\ 24 \\ 5 \end{pmatrix}$ and $\begin{pmatrix} -3 \\ 8 \\ -13 \end{pmatrix}$.

5. Perform the Gram-Schmidt algorithm on the vectors $\begin{pmatrix} -1 \\ 2 \\ -2 \\ 4 \end{pmatrix}, \begin{pmatrix} 12 \\ 1 \\ -6 \\ 12 \end{pmatrix}, \begin{pmatrix} 14 \\ -13 \\ 13 \\ 4 \end{pmatrix}$ and $\begin{pmatrix} -15 \\ 15 \\ 15 \\ 15 \end{pmatrix}$.

6. Perform the Gram-Schmidt algorithm on the following two sets of vectors:

$$\left\{\begin{pmatrix}4\\1\\-2\\2\end{pmatrix},\begin{pmatrix}-8\\-11\\-1\\8\end{pmatrix},\begin{pmatrix}4\\-2\\13\\-19\end{pmatrix},\begin{pmatrix}-19\\20\\17\\-5\end{pmatrix}\right\} \text{ and } \left\{\begin{pmatrix}-11\\8\\-6\\2\end{pmatrix},\begin{pmatrix}-9\\12\\-5\\0\end{pmatrix},\begin{pmatrix}9\\-12\\10\\15\end{pmatrix},\begin{pmatrix}-9\\-3\\-26\\-3\end{pmatrix}\right\}.$$

7. All the calculations in this section have relatively *nice* answers. Is this always the case?

8. Is this the sort of algorithm that should be implemented on a computer?

Solutions

1. $\mathbf{v}_1 \leftarrow \begin{pmatrix}1\\2\\-2\end{pmatrix}$ and $\|\mathbf{v}_1\|_2 = 3$ so $\hat{\mathbf{v}}_1 \leftarrow \begin{pmatrix}\frac{1}{3}\\\frac{2}{3}\\-\frac{2}{3}\end{pmatrix}$.

Next, $\mathbf{v}_2 \leftarrow \begin{pmatrix}3\\21\\0\end{pmatrix}$, so $\mathbf{v}_2 \leftarrow \mathbf{v}_2 - \text{proj}_{\hat{\mathbf{v}}_1}(\mathbf{v}_2) = \begin{pmatrix}3\\21\\0\end{pmatrix} - 15\begin{pmatrix}\frac{1}{3}\\\frac{2}{3}\\-\frac{2}{3}\end{pmatrix} = \begin{pmatrix}-2\\11\\10\end{pmatrix}$ and $\|\mathbf{v}_2\|_2 = 15$ so $\hat{\mathbf{v}}_2 \leftarrow \begin{pmatrix}-\frac{2}{15}\\\frac{11}{15}\\\frac{2}{3}\end{pmatrix}$.

3. $\mathbf{v}_1 \leftarrow \begin{pmatrix}1\\4\\8\end{pmatrix}$ and $\|\mathbf{v}_1\|_2 = 9$ so $\hat{\mathbf{v}}_1 \leftarrow \begin{pmatrix}\frac{1}{9}\\\frac{4}{9}\\\frac{8}{9}\end{pmatrix}$.

Next, $\mathbf{v}_2 \leftarrow \begin{pmatrix}6\\15\\12\end{pmatrix}$, so $\mathbf{v}_2 \leftarrow \mathbf{v}_2 - \text{proj}_{\hat{\mathbf{v}}_1}(\mathbf{v}_2) = \begin{pmatrix}6\\15\\12\end{pmatrix} - 18\begin{pmatrix}\frac{1}{9}\\\frac{4}{9}\\\frac{8}{9}\end{pmatrix} = \begin{pmatrix}4\\7\\-4\end{pmatrix}$ and $\|\mathbf{v}_2\|_2 = 9$ so $\hat{\mathbf{v}}_2 \leftarrow \begin{pmatrix}\frac{4}{9}\\\frac{7}{9}\\-\frac{4}{9}\end{pmatrix}$.

Finally, $\mathbf{v}_3 \leftarrow \begin{pmatrix}-19\\-22\\-17\end{pmatrix}$, so $\mathbf{v}_3 \leftarrow \mathbf{v}_3 - \text{proj}_{\hat{\mathbf{v}}_1}(\mathbf{v}_3) = \begin{pmatrix}-19\\-22\\-17\end{pmatrix} + 27\begin{pmatrix}\frac{1}{9}\\\frac{4}{9}\\\frac{8}{9}\end{pmatrix} = \begin{pmatrix}-16\\-10\\7\end{pmatrix}$,

$\mathbf{v}_3 \leftarrow \mathbf{v}_3 - \text{proj}_{\hat{\mathbf{v}}_2}(\mathbf{v}_3) = \begin{pmatrix}-16\\-10\\7\end{pmatrix} + 18\begin{pmatrix}\frac{4}{9}\\\frac{7}{9}\\-\frac{4}{9}\end{pmatrix} = \begin{pmatrix}-8\\4\\-1\end{pmatrix}$ and $\|\mathbf{v}_3\|_2 = 9$ so $\hat{\mathbf{v}}_3 \leftarrow \begin{pmatrix}-\frac{8}{9}\\\frac{4}{9}\\-\frac{1}{9}\end{pmatrix}$.

5. Without stepping through the algorithm, the solutions are $\hat{\mathbf{v}}_1 = \begin{pmatrix}-\frac{1}{5}\\\frac{2}{5}\\-\frac{2}{5}\\\frac{4}{5}\end{pmatrix}$, $\hat{\mathbf{v}}_2 = \begin{pmatrix}\frac{14}{15}\\-\frac{1}{5}\\-\frac{2}{15}\\\frac{4}{15}\end{pmatrix}$, $\hat{\mathbf{v}}_3 = \begin{pmatrix}-\frac{2}{15}\\-\frac{2}{5}\\\frac{11}{15}\\\frac{8}{15}\end{pmatrix}$ and $\hat{\mathbf{v}}_4 = \begin{pmatrix}\frac{4}{15}\\\frac{4}{5}\\\frac{8}{15}\\-\frac{1}{15}\end{pmatrix}$.

7. Definitely not.

In order to do this in Matlab, we must learn about iteration. Suppose we have a vector with *n* entries, in which case, the `for` statement allows you to execute the same statements, with each

```
s = 0;
for x = [2 3 5 7 11]
    s = s + x
end
    s =
        2

    s =
        5

    s =
        10

    s =
        17

    s =
        28
```

Each time the loop runs, x takes on the next value. The most common loop is to simply do something *n* times

```
for x = [1 2 3 4 5 6 7 8 9]
    // Do something
end
```

This, however, requires us to hard-code the array, and thus we introduce our first *vector constructor* in Matlab:

```
>> v = 1:10
    v =
        1   2   3   4   5   6   7   8   9  10

>> w = -3:3
    w =
       -3  -2  -1   0   1   2   3

>> x = 3.4:8.9
    x =
        3.4000   4.4000   5.4000   6.4000  7.4000   8.4000
```

From the examples, `m:n` creates a vector of length $\lfloor n-m+1 \rfloor$ with the entries

$$n, n+1, n+2, \ldots, n+\lfloor n-m \rfloor .$$

Let us assume that the columns of a matrix **U** represent the vectors we would like to orthogonalize. We will ensure that the arguments and the return values are correct and of the correct number.

```
function [V] = gramschmidt( U )
% GRAMSCHMIDT  Perform the Gram-Schmidt process on the columns of U
%   V = GRAMSCHMIDT(U)  The columns of the matrix V are the normalized
%   and orthogonal vectors resulting from the Gram Schmidt process applied
%   to the columns of the matrix U.

    if nargin ~= 1
        throw( MException( 'linalg:gramschmidt', ...
                           'expecting one argument, but got %d', nargin ) );
    elseif nargin >= 2
        throw( MException( 'linalg:gramschmidt', ...
                           'Too many output arguments.' ) );
    elseif ~ismatrix( U ) || ~isnumeric( U )
        throw( MException( 'linalg:gramschmidt', ...
                           'The argument U should be a numeric matrix' ) );
    end

    V = U;

    for k = 1:size( V, 2 )
        for j = 1:(k - 1)
            % Find the perpendicular component of V(:,k) relative to each
            % of the previous k - 1 normalized orthogonal vectors.
            [~, V(:,k)] = proj( V(:,k), V(:,j), 'unit' );
        end

        normVk = norm( V(:,k) );

        % If the k'th column is insignificantly small,
        %     do not normalize it,
        %     rather issue a warning and leave the column unchanged;
        % otherwise, normalize the k'th column.
        if normVk < size( V, 1 )*eps
            warning( 'linalg:gramschmidt', ...
                     ['Column %d appears to be within the span ' ...
                      'of columns 1 through %d'], k, k - 1 );
        else
            V(:,k) = V(:,k)/normVk;
        end
    end
end
```

At the end of this algorithm, the return value will be a matrix of vectors that are reasonably close to orthogonal. We say *reasonably close to orthogonal*, as numerical error may result in small errors so that the mutual inner products are not precisely zero.

5.11 Applications of the inner product

The most important application of the inner product is to determine how *similar* two vectors are to each other. If the inner product is positive, the vectors are at least pointing in the same direction; if the inner product is negative, the vectors are pointing in opposing directions; and if the inner product is zero, the vectors are orthogonal.

5.11.1 Mean and variance

Given a vector **x**, the *average value* or *mean* of the vector $\mu_{\mathbf{x}}$ is the sum of the entries divided by the number of entries:

$$\mu_{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n} x_k \ .$$

More generally, you can always define the mean as

$$\mu_{\mathbf{u}} = \frac{\left\langle \hat{\mathbf{1}}, \mathbf{u} \right\rangle}{\left\langle \hat{\mathbf{1}}, \hat{\mathbf{1}} \right\rangle},$$

where $\hat{\mathbf{1}}$ is the vector of all entries equal to 1. Thus, you can, in some cases, define the average value or mean of a function:

$$\mu_f = \frac{\left\langle \hat{1}, f \right\rangle}{\left\langle \hat{1}, \hat{1} \right\rangle} = \frac{\int_a^b f(x)\,dx}{\int_a^b 1\,dx} = \frac{1}{b-a}\int_a^b f(x)\,dx$$

Subtracting off the mean from a vector is

$$\mathbf{u} - \frac{\left\langle \mathbf{1}, \mathbf{u} \right\rangle}{\left\langle \mathbf{1}, \mathbf{1} \right\rangle}\mathbf{1}$$

produces a vector that is orthogonal to the vector **1**; that is, it has an average value or mean of 0. The *variance* of a vector is defined as the inner product of this no

$$\left\langle \mathbf{u}-\frac{\langle \mathbf{1},\mathbf{u}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}\mathbf{1},\mathbf{u}-\frac{\langle \mathbf{1},\mathbf{u}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}\mathbf{1}\right\rangle = \langle \mathbf{u},\mathbf{u}\rangle - \left\langle \mathbf{u},\frac{\langle \mathbf{1},\mathbf{u}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}\mathbf{1}\right\rangle - \left\langle \frac{\langle \mathbf{1},\mathbf{u}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}\mathbf{1},\mathbf{u}\right\rangle + \left\langle \frac{\langle \mathbf{1},\mathbf{u}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}\mathbf{1},\frac{\langle \mathbf{1},\mathbf{u}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}\mathbf{1}\right\rangle$$

$$= \langle \mathbf{u},\mathbf{u}\rangle - \frac{\langle \mathbf{1},\mathbf{u}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}\langle \mathbf{u},\mathbf{1}\rangle - \frac{\langle \mathbf{1},\mathbf{u}\rangle^*}{\langle \mathbf{1},\mathbf{1}\rangle^*}\langle \mathbf{1},\mathbf{u}\rangle + \frac{\langle \mathbf{1},\mathbf{u}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}\frac{\langle \mathbf{1},\mathbf{u}\rangle^*}{\langle \mathbf{1},\mathbf{1}\rangle^*}\langle \mathbf{1},\mathbf{1}\rangle$$

$$= \langle \mathbf{u},\mathbf{u}\rangle - \frac{\langle \mathbf{1},\mathbf{u}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}\langle \mathbf{u},\mathbf{1}\rangle - \frac{\langle \mathbf{u},\mathbf{1}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}\langle \mathbf{1},\mathbf{u}\rangle + \frac{\langle \mathbf{1},\mathbf{u}\rangle\langle \mathbf{u},\mathbf{1}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}$$

$$= \langle \mathbf{u},\mathbf{u}\rangle - \frac{\langle \mathbf{1},\mathbf{u}\rangle\langle \mathbf{u},\mathbf{1}\rangle}{\langle \mathbf{1},\mathbf{1}\rangle}$$

$$= \|\mathbf{u}\|_2^2 - \frac{|\langle \mathbf{1},\mathbf{u}\rangle|^2}{\langle \mathbf{1},\mathbf{1}\rangle}$$

As another example, suppose we have *n* stocks and that **q** is an *n*-dimensional vector of the quantities of shares of each stock held in our portfolio, while **v** is an *n*-dimensional vector that stores the corresponding price per share. Thus, in this case, $\langle \mathbf{q},\mathbf{v}\rangle$ is total value of our portfolio.

The next most important application of the inner product is as a compact representation of linear equations. For example, consider the linear equation

$$3x + 4y - 5z = 6.$$

If we define $\mathbf{c} = \begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix}$ and $\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$, then we compactly write this linear equation as $\langle \mathbf{c},\mathbf{x}\rangle = 6$.

As a third example, suppose we have *n* objects and **m** is an *n*-dimensional vector of the mass of each of these objects, and **s** is a vector of the speed of these masses in a specified direction. In this case, $\langle \mathbf{m},\mathbf{s}\rangle$ is the total momentum in the specified direction.

As a fourth example, suppose that we have a chemical reaction that sees sucrose and water (in the presence of catalysts) converted into alcohol and carbon dioxide.

$$a\ C_{12}H_{22}O_{11} + b\ H_2O \rightarrow c\ C_2H_5OH + d\ CO_2.$$

This needs to be thought of as

$$a\ C_{12}H_{22}O_{11} + b\ H_2O + c\ C_2H_5OH + d\ CO_2 = 0,$$

with $a < 0$ and $b < 0$ to indicate that sucrose and water are reactants, and $c > 0$ and $d > 0$ to indicate that ethanol and carbon dioxide are products.

Then, the vector $\mathbf{q} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}$ is the quantity of each substance (where $a < 0$ and $b < 0$ indicating that those substances has not reacted yet) and $\mathbf{p}_\text{C} = \begin{pmatrix} 12 \\ 0 \\ 2 \\ 1 \end{pmatrix}$, $\mathbf{p}_\text{O} = \begin{pmatrix} 11 \\ 1 \\ 1 \\ 2 \end{pmatrix}$ and $\mathbf{p}_\text{H} = \begin{pmatrix} 22 \\ 2 \\ 6 \\ 0 \end{pmatrix}$ then the inner products $\langle \mathbf{q}, \mathbf{p}_\text{C} \rangle = \langle \mathbf{q}, \mathbf{p}_\text{O} \rangle = \langle \mathbf{q}, \mathbf{p}_\text{H} \rangle = 0$ may be employed to determine $\mathbf{q}$. If we begin with $\langle \mathbf{q}, \mathbf{p}_\text{H} \rangle = 0$, or $22a + 2b + 6c = 0$, which gives $b = -11a - 3c$. Then, from $\langle \mathbf{q}, \mathbf{p}_\text{O} \rangle = 0$, or $11a + b + c + 2d = 0$, we obtain $-2c + 2d = 0$ after having substituted the previous result. From this, we deduce that $c = d$. Finally, we use $\langle \mathbf{q}, \mathbf{p}_\text{C} \rangle = 0$ and, simplifying, $4a + c = 0$. The simplest solution is to set $a = -1$, in which case, $b = -1$ and $c = d = 4$, thus obtaining the balanced chemical equation

$$\text{C}_{12}\text{H}_{22}\text{O}_{11} + \text{H}_2\text{O} \rightarrow 4\ \text{C}_2\text{H}_5\text{OH} + 4\ \text{CO}_2.$$

As a fifth example, if the entries of $\mathbf{p}$ represent is the profit per product manufactured, then $\langle \mathbf{p}, \mathbf{q} \rangle$ is the total profit made having produced the number of each of the products.

Finally, as a sixth example, in industry, if each of the entries of an n-dimensional vector is associated with a specific product that is to be manufactured, then for a given resource x (a raw material or required component), the vector $\mathbf{r}_x$ could represent the amount or number required of the given resource for each of the n products. Additionally, a vector $\mathbf{q}$ could represent the quantity of each product produced. In this case, the inner product $\langle \mathbf{r}_x, \mathbf{q} \rangle$ represents the total amount or number of the specified resource that is required. In the previous example of balancing a chemical reaction, the inner product was equated to zero; however, in this case, there may be a limit as to the amount or number of a given resource that is available. If we call this limit ℓ_x, then it is absolutely necessary that

$$\langle \mathbf{r}_x, \mathbf{q} \rangle \leq \ell_x,$$

otherwise there will not be a sufficient amount or number of the resource available to make the required products.

5.11.2 Correlation

Given two sets of data, you may ask whether or not one set of data is *correlated* to another.

6 Solutions to systems of linear equations, linear independence and bases

In this chapter, we will look at solving systems of linear equations, look at the possible number of solutions and determining those solutions, looking in depth at the Gaussian elimination algorithm, which you may have seen in secondary school, but introducing a variation described as partial pivoting to reduce numerical error. From this, we will describe linear dependence, the span of a set of vectors, linear independence and the concepts of base and dimension. To start, however, we will describe equations, linear equations, and systems of linear equations.

6.1 Equations, linear equations and systems of equations

An equation is the equating of two mathematical expressions that have one or more *variables* or *unknowns*, where the goal is find values of the variables that satisfy the equality. For example, you may have an equation like finding all the values (if any) where the polynomial $x^2 + 2x - 5$ equals 4, or

$$x^2 + 2x - 5 = 4.$$

Alternatively, you may wish to find where the polynomial $x^2 + 2x - 5$ has the same values as the polynomial $x^3 - 3x + 1$, or

$$x^2 + 2x - 5 = x^3 - 3x + 1.$$

These are often written in a *standard form*, whereby one side is subtracted from the other; in these two cases, these would be rewritten in the forms

$$x^2 + 2x - 9 = 0 \text{ and } -x^3 + x^2 + 5x - 6 = 0,$$

respectively. When we are solving for when an expression equals zero, we refer to that as a *root-finding problem*. A *solution* to an equation is any values to the variables or unknowns that satisfy the equation. For example, solutions to the first equation are $x = -1 \pm \sqrt{10}$ and the solutions to the second are $x = 2$ and $x = -\frac{1}{2} \pm \frac{1}{2}\sqrt{13}$. There is no reason to restrict an equation to just one variable; for example, $xy = 1$ has infinitely many solutions, $(x, y) = (r, \frac{1}{r})$ for any non-zero real value of r if we are considering only real numbers, or $(x, y) = (z, \frac{1}{z})$ for any non-zero complex value z if we are also considering complex numbers.

A *system of equations* is when there are two or more equations where any solution must simultaneously satisfy all of the equations in the system. For example, consider the following two equations:

$$\begin{aligned} x^2 - 2xy - 2 &= 0 \\ xy^2 + y - 1 &= 0 \end{aligned}$$

If we consider only real numbers, there is only one solution: $(x, y) = (2, \frac{1}{2})$; however, if we also consider complex numbers, we have two additional solutions: $(x, y) = (1 + j, j)$ and $(x, y) = (1 - j, -j)$. In general, finding exact solutions to nonlinear equations is exceptionally difficult; for example, try finding the solutions to the system of equations

$$\begin{aligned} x^3 + x^2 y + xy^2 + y^3 - 1 &= 0 \\ x^2 + xy - y^2 - 1 &= 0 \end{aligned}$$

has one possibly obvious solution: $(x, y) = (1, 0)$, but as a plot of these two relations shows, as shown in Figure 52, shows that there must be at least one more solution in the vicinity of $(x, y) = (0.93, 0.19)$, a solution that cannot be represented algebraically. If we consider complex solutions, there are another four.

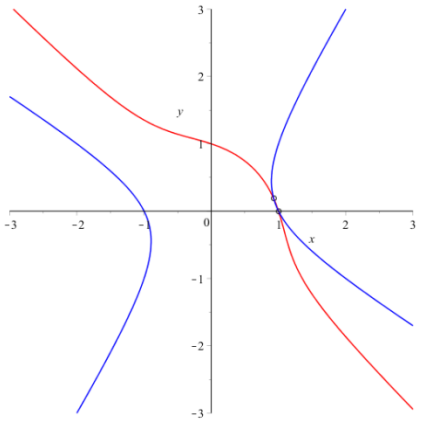


Figure 52. The relations $x^3 + x^2 y + xy^2 + y^3 - 1 = 0$ and $x^2 + xy - y^2 - 1 = 0$.

A *linear equation* is an equation that is the sum of either scalars or the product of a scalar and a variable or unknown; for example, the following are examples of linear equations:

$$\begin{aligned} 1 + 3x - 4y + z &= 2x + y + 4z - 2 \\ x + y + z + 1 &= 0 \\ 2x - 5y &= 7 + x + 3y \end{aligned}$$

Any equation that is not linear is said to be a *non-linear equation*.

Unlike non-linear equations where the standard form is to write each expression equated to zero, it is standard practice to write linear equations in the form of all scalars multiplied by an unknown on the left-hand side, and all scalars are brought to the right-hand side. Consequently, the above linear equations would be written in standard form are

$$\begin{aligned} x - 5y - 3z &= -3 \\ x + y + z &= -1 \\ x - 8y &= 7 \end{aligned}$$

At times, we will indicate the number of unknown variables. The first two are linear equations *in three unknown variables* or just *in three unknowns*, while the third is a linear equation *in two unknowns*.

A *system of linear equations* is a collection of linear equations, the solution to which (if one even exists) must simultaneously satisfies all of the equations. For example, if we consider the above three equations as a system, there is only one solution:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\frac{7}{15} \\ -\frac{14}{15} \\ \frac{12}{5} \end{pmatrix}.$$

Note that because the variables x, y and z are ordered, we can represent the solution unambiguously using a vector in $\mathbf{R}^3$.

In general, we will be looking to finding an algorithm to solve a system of *m* linear equations in *n* unknowns, and we will often write this as

$$\begin{aligned}
a_{1,1}x_1 + a_{1,2}x_2 + a_{1,3}x_3 + \cdots + a_{1,n}x_n &= b_1 \\
a_{2,1}x_1 + a_{2,2}x_2 + a_{2,3}x_3 + \cdots + a_{2,n}x_n &= b_2 \\
a_{3,1}x_1 + a_{3,2}x_2 + a_{3,3}x_3 + \cdots + a_{3,n}x_n &= b_3 \\
&\vdots \\
&\vdots \\
a_{m,1}x_1 + a_{m,2}x_2 + a_{m,3}x_3 + \cdots + a_{m,n}x_n &= b_m
\end{aligned}$$

where the *n* unknowns are $x_1, x_2, x_3, \ldots, x_n$ and all other values ($a_{1,1}$ through $a_{m,n}$ and b_1 through b_n) are known. The solution will be represented using an *n*-dimensional vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$.

Problems

1. Determine which of the following are linear equations

$$3x + 4y = 4 - 3z,\ \ xy = 3x + 4y - 3,\ \ x + y = x - y,\ \ xyz = 1.$$

For those that are, write them in standard form.

2. Determine which of the following are linear equations

$$5x + z = 4y - 4 - 3z,\ \ 3x + 3xy - 2z = 3,\ \ x^2 + 2xy + y^2 = 1,\ \ 5 + x - 4 = 2 - 3x + 1$$

For those that are, write them in standard form.

3. Given the following system of linear equations,

$$\begin{aligned}
x - 3y &= 1 \\
x + y &= 2
\end{aligned}$$

which of the following are solutions to this system of linear equations?

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix},\ \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{9} \end{pmatrix},\ \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{7}{4} \\ \frac{1}{4} \end{pmatrix} \text{ and } \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{11}{5} \\ -\frac{1}{5} \end{pmatrix}$$

4. Given the following system of linear equations,

$$\begin{aligned} x + y + z &= 1 \\ x + 2y + 4z &= 2 \\ x + 3y + 9z &= -1 \end{aligned}$$

which of the following are solutions to this system of linear equations?

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 4 \\ -5 \\ 2 \end{pmatrix}, = \begin{pmatrix} 5 \\ -5 \\ 1 \end{pmatrix}, = \begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix} \text{ or } = \begin{pmatrix} -4 \\ 7 \\ -2 \end{pmatrix}.$$

Solutions

1. The first and third are linear equations: $3x + 4y + 3z = 4$ and $2y = 0$.

3. The first does not satisfy either equation ($1 - 3 \cdot 2 = -5 \neq 1$ and $1 + 2 = 3 \neq 2$); the second only satisfies the first linear equation, but not the second ($\frac{1}{3} + \left(-\frac{2}{9}\right) = \frac{1}{9} \neq 2$); the third satisfies both equations; and the fourth only satisfies the second equation ($\frac{11}{5} - 3\left(-\frac{1}{5}\right) = \frac{14}{5} \neq 1$).

6.2 Finding solutions to a system of linear equations

We will see that given a system of linear equations, there is only ever one of three possibilities:

1. there are no solutions,
2. there is exactly one solution, and
3. there are infinitely many solutions.

We will start by looking at a single linear equation, and then proceed to looking at examples with a system of two linear equations and then a system of three linear equations.

6.2.1 Linear equations in a single unknown

The linear equation in a single unknown

$$2x = 1$$

has exactly one solution: $x = 0.5$. If a single linear equation has more than one variable, then there will be an infinite number of solutions; for example,

$$20x - 110y = 1$$

allows us choose any value for y, and that value of y fixes the value of x: given y, $x = \frac{1+110y}{20} = \frac{1}{20} + \frac{11}{2}y = 0.05 + 5.5y$. Thus, if $y = 1$, $x = 5.55$; if $y = 0$, $x = 0.05$, and if $y = 0.112$, $y = 0.666$.

Suppose now we have a system of two linear equations:

1. The system of two linear equations in one unknown $2x = 4$ and $3x = 6$ has a unique solution: $x = 2$.
2. The system of two linear equations $2x = 4$ and $3x = 9$ has no solutions, for the first requires that $x = 2$ and the second requires that $x = 3$.

With a little thought, it should be obvious that a system with two or more linear equations in a single unknown has a solution if and only if all the equations are scalar multiples of each other.

Problems

1. Solve the linear equation $(3-2j)x=-4+j$.

2. Solve the linear equation $(5+3j)x=-6+2j$.

3. Find all solutions to the linear equation $(3-2j)x+(1+3j)y=-4+j$.

4. Find all solutions to the linear equation $(5+3j)x+(7-4j)y=-6+2j$.

5. In finding all the solutions to the linear equation in Questions 3 and 4, do we let *y* take on all real values, or can *y* also take on all possible complex values?

6. Does the solution to $x + y = 1$ change in any way if we are solving for either real values of *x* and *y* or complex values of *x* and *y*?

7. What is the solution to the linear equation in Question 3 when $y = 3 - 2j$?

8. What is the solution to the linear equation in Question 4 when $y = -5 - 3y$?

Solutions

1. $x=\dfrac{-4+j}{3-2j}=\dfrac{-4+j}{3-2j}\cdot\dfrac{3+2j}{3+2j}=\dfrac{-14-5j}{13}=-\dfrac{14}{13}-\dfrac{5}{13}j$.

3. Using the solution to Question 1,

$$x=\frac{-4+j-(1+3j)y}{3-2j}=\frac{-4+j}{3-2j}+\frac{-1-3j}{3-2j}y=-\frac{14}{13}-\frac{5}{13}j+\frac{-1-3j}{3-2j}\cdot\frac{3+2j}{3+2j}y$$
$$=-\frac{14}{13}-\frac{5}{13}j+\frac{3-11j}{13}y=-\frac{14}{13}-\frac{5}{13}j+\left(\frac{3}{13}-\frac{11}{13}j\right)y$$

5. If are finding complex solutions to our linear equation, then *y* could take on any complex value.

7. Using the solution to Question 1, $(3-2j)x+(1+3j)(3-2j)=-\dfrac{14}{13}-\dfrac{5}{13}j-1-3j=-\dfrac{27}{13}-\dfrac{44}{13}j$

6.2.2 A system of two linear equations

In secondary school, you would have also attempted to find a solution to a system of two linear equations in two unknowns:

$$\begin{aligned} x + 2y &= 1 \\ 3x - 2y &= 9 \end{aligned}$$

You would have learned that you can add a multiple of one equation onto the other. In this case, we can add –3 times Equation 1 onto Equation 2 to get

$$\begin{aligned} x + 2y &= 1 \\ -8y &= 6 \end{aligned}$$

If two systems of linear equations have the same solution, we will say that they are *equivalent* and we will write this as

$$\begin{aligned} x + 2y &= 1 \\ 3x - 2y &= 9 \end{aligned} \quad \sim \quad \begin{aligned} x + 2y &= 1 \\ -8y &= 6 \end{aligned}$$

From this operation, we may conclude that $y = -0.75$, and substituting this into the first equation, we get that

$$x + 2(-0.75) = 1$$

or, solving for x,

$$x = 1 - 2(-0.75) = 2.5.$$

Thus, this system of two linear equations in two unknowns has one unique solution $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -0.75 \\ 2.5 \end{pmatrix}$. You may, however, had experiences with systems such as

$$\begin{aligned} x + 2y &= 1 \\ -3x - 6y &= 4 \end{aligned}$$

where, when adding 3 times Equation 1 onto Equation 2 yields the particular odd second equation

$$\begin{aligned} x + 2y &= 1 \\ 0 &= 7 \end{aligned}$$

As $0 = 7$ is a contradiction, this indicates that no solution exists to this system of two linear equations.

On the other hand,

$$\begin{aligned} x-2y&=-1\\ -3x+6y&=3 \end{aligned}$$

has infinitely many solutions, for if you add –3 times Equation 1 onto Equation 2, you get

$$\begin{aligned} x-2y&=-1\\ 0&=0 \end{aligned}$$

and so we are left with the single linear equation in two unknowns $x - 2y = -1$, which, as before, allows us to assign y any value we wish, and x may then be determined based on the value of y. Thus, the solution is

$$\begin{pmatrix} x\\ y \end{pmatrix}=\begin{pmatrix} -1+2y\\ y \end{pmatrix}=\begin{pmatrix} -1\\ 0 \end{pmatrix}+y\begin{pmatrix} 2\\ 0 \end{pmatrix}.$$

Note that the solution can be written as a vector plus an arbitrary scalar multiple of another.

Problems

1. Find all solutions to each of the following systems of two linear equations:

$$\begin{aligned} 2x+3y&=5\\ x+2y&=3 \end{aligned}, \quad \begin{aligned} 2x+4y&=6\\ x+2y&=3 \end{aligned} \text{ and } \begin{aligned} 2x+4y&=7\\ x+2y&=3 \end{aligned}.$$

2. Find all solutions to the following systems of two linear equations:

$$\begin{aligned} 2x+3y+z&=6\\ x+2y+3z&=6 \end{aligned}, \quad \begin{aligned} -2x+4y-6z&=12\\ x-2y+3z&=6 \end{aligned} \text{ and } \begin{aligned} -2x+4y-6z&=12\\ x-2y+3z&=-6 \end{aligned}.$$

3. Find all solutions to the system of two linear equations

$$\begin{aligned} jx+3y&=2\\ 2x+jy&=3j \end{aligned}.$$

4. Find all solutions to the system of two linear equations

$$\begin{aligned} x+jy&=j\\ jx-y&=1 \end{aligned}.$$

Solutions

1. First, adding –0.5 times Equation 1 onto Equation 2 yields

$$\begin{aligned} 2x+3y&=5\\ 0.5y&=0.5 \end{aligned}$$

This gives us that $y = 1$, and thus substituting $y = 1$ into the first equation gives us that $2x = 5 – 3$ or $x = 1$, so the solution is $\begin{pmatrix} x\\ y \end{pmatrix}=\begin{pmatrix} 1\\ 1 \end{pmatrix}$.

Second, adding –0.5 times Equation 1 onto Equation 2 yields

$$\begin{aligned} 2x+4y &= 6 \\ 0 &= 0 \end{aligned}$$

The second equation is trivially true, and thus we may let y take on any real value, after which $x = 3-2y$, or

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3-2y \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \end{pmatrix} - y\begin{pmatrix} -2 \\ 0 \end{pmatrix}.$$

Third, adding –0.5 times Equation 1 onto Equation 2 yields

$$\begin{aligned} 2x+4y &= 7 \\ 0 &= -0.5 \end{aligned}$$

The second equation implies that there is a contradiction, and thus no solution exists.

3. Adding $\frac{-2}{j} = 2j$ times Equation 1 onto Equation 2 yields

$$\begin{aligned} jx+3y &= 2 \\ 7jy &= 7j \end{aligned}.$$

This gives us that $y = 1$, and thus substituting $y = 1$ into the first equation gives us that $jx = 2-3$ or $x = j$, so the solution is $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} j \\ 1 \end{pmatrix}$.

6.2.3 Solving higher systems of linear equations

If we are given more linear equations or more unknowns, we continue the process of eliminating, where possible, variables from the various equations. For example,

$$\begin{aligned} 2x+3y+z &= -0.7 \\ x+2y+4z &= 0.1 \\ -3x+4y+2z &= -2.5 \end{aligned}$$

We can add –0.5 times Equation 1 onto Equation 2, and we can add 1.5 times Row 1 onto Row 3, to get

$$\begin{aligned} 2x+3y+z &= -0.7 \\ 0.5y+3.5z &= 0.45 \\ 8.5y+3.5z &= -3.55 \end{aligned}$$

At this point, we see that Equation 1 is the only equation that contains x, while Equations 2 and 3 contain information about the relationship between y and z. Focusing on these, we see that we can eliminate z from the third equation by adding –17 times Equation 2 onto Equation 3:

$$\begin{aligned} 2x+3y+z &= -0.7 \\ 0.5y+3.5z &= 0.45 \\ -56z &= -11.2 \end{aligned}$$

The last equation is now one we can solve: $-56z = -11.2$ implies that $z = 0.2$. Substituting this value back into Equation 2, we get that $0.5y + 3.5 \cdot 0.2 = 0.45$ so $0.5y = 0.45 - 0.7$ or $y = -0.5$. Finally, substituting back these values of *y* and *z* into Equation 1 yields that $2x + 3 \cdot (-0.5) + 0.2 = -0.7$ or $2x = -0.7 + 1.3$ so $x = 0.3$. The solution vector is thus $\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0.3 \\ -0.5 \\ 0.2 \end{pmatrix}$.

In general, we may have a system of *m* linear equations in *n* unknowns:

$$\begin{aligned} \alpha_{1,1}x_1 + \alpha_{1,2}x_2 + \cdots + \alpha_{1,n}x_n &= \beta_1 \\ \alpha_{2,1}x_1 + \alpha_{2,2}x_2 + \cdots + \alpha_{2,n}x_n &= \beta_2 \\ &\vdots \\ \alpha_{m,1}x_1 + \alpha_{m,2}x_2 + \cdots + \alpha_{m,n}x_n &= \beta_m \end{aligned}$$

where the *x*-values indicate unknowns while the α- and β-values are all either real numbers or complex numbers. The solution vector will be the *n*-dimensional vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$.

It may be computationally more difficult, but we could use Equation 1 to eliminate x_1 from Equations 2 through *n*. Next, we use Equation 2 to eliminate x_2 from Equations 3 through *n*, and so forth. At some point, we will then be able to stop and determine

1. if a unique solution exists, in which case, we use backward substitution to find the solution;
2. if no solution exists, in which case, we are finished; or
3. if an infinite number of solutions exist, in which case, we find that we can define all such solutions based on the values of independent or *free* variables. We use the term *free* because we are allowed to assign these variables any values we wish, and the result will be one of the infinitely many solutions.

However, before we describe an algorithm for finding which is the case, we will look at what appears to be a different problem.

Problems

1. The system of three linear equations

$$\begin{aligned} x + y + z &= 3 \\ x + 2y + 3z &= 6 \\ y + 2z &= 3 \end{aligned}$$

has two solutions: first $x = 2$, $y = -1$ and $z = 2$ and second $x = -1$ $y = 5$, $z = -1$. Show that

$$\alpha \begin{pmatrix} 2 \\ -1 \\ 2 \end{pmatrix} + (1-\alpha) \begin{pmatrix} -1 \\ 5 \\ -1 \end{pmatrix}$$

is also a solution for any value of α. Because $\alpha + (1-\alpha) = 1$, this specific linear combination is sometimes called an *affine* combination of these two vectors.

2. Another solution to the system of three linear equations and three unknowns is $x = y = z = 1$. Find the value of α that yields this solution as an affine combination of the two solutions above.

3. Do all solutions to this system of linear equations form a subspace of $\mathbf{R}^3$?

Solutions

1. Simplifying these, we get

$$\begin{aligned} x &= 2\alpha + \alpha - 1 = 3\alpha - 1 \\ y &= -\alpha + 5 - 5\alpha = 5 - 6\alpha \\ z &= 2\alpha - 1 + \alpha = 3\alpha - 1 \end{aligned}$$

Substituting these

$$\begin{aligned} 3\alpha - 1 + 5 - 6\alpha + 3\alpha - 1 &= 3 \\ 3\alpha - 1 + 10 - 12\alpha + 9\alpha - 3 &= 6 \\ 5 - 6\alpha + 6\alpha - 2 &= 3 \end{aligned}$$

and we see that all the terms with α cancel and we are left with the three true statements that

$$\begin{aligned} -1 + 5 - 1 &= 3 \\ -1 + 10 - 3 &= 6 \\ 5 - 2 &= 3 \end{aligned}$$

3. No, because the solutions are of the form $\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -1 \\ 5 \\ -1 \end{pmatrix} + \alpha \begin{pmatrix} 3 \\ -6 \\ 3 \end{pmatrix}$, and the zero vector is not of this form.

6.3 Finding linear of combinations equaling a given vector

We have now described solving systems of linear equations. The next problem may initially appear to be significantly different, but we will see that they are essentially the same problem; thus, an algorithm for solving one solves the other. The problem is: given an ordered collection of *m*-dimensional vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ and a target vector $\mathbf{v}$, is there a linear combination of the collection of vectors equaling the target vector? That is, do there exists $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that

$$\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n = \mathbf{v}.$$

We would represent such a solution by a vector $\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}$. Suppose that vectors have the entries $\mathbf{u}_k = \begin{pmatrix} u_{1,k} \\ u_{2,k} \\ \vdots \\ u_{m,k} \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix}$. In this case, we have that our question becomes

$$\alpha_1 \begin{pmatrix} u_{1,1} \\ u_{2,1} \\ \vdots \\ u_{m,1} \end{pmatrix} + \alpha_2 \begin{pmatrix} u_{1,2} \\ u_{2,2} \\ \vdots \\ u_{m,2} \end{pmatrix} + \cdots + \alpha_n \begin{pmatrix} u_{1,n} \\ u_{2,n} \\ \vdots \\ u_{m,n} \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix}.$$

Recalling that the scalar multiple of a vector simply has each entry of the vector multiplied by the scalar, so this is identical to the question of finding

$$\begin{pmatrix} u_{1,1}\alpha_1 \\ u_{2,1}\alpha_1 \\ \vdots \\ u_{m,1}\alpha_1 \end{pmatrix} + \begin{pmatrix} u_{1,2}\alpha_2 \\ u_{2,2}\alpha_2 \\ \vdots \\ u_{m,2}\alpha_2 \end{pmatrix} + \cdots + \begin{pmatrix} u_{1,n}\alpha_n \\ u_{2,n}\alpha_n \\ \vdots \\ u_{m,n}\alpha_n \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix}.$$

Normally, we write $\alpha \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix} = \begin{pmatrix} \alpha u_1 \\ \alpha u_2 \\ \vdots \\ \alpha u_m \end{pmatrix}$, however, there is no reason we cannot write it as $\alpha \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{pmatrix} = \begin{pmatrix} u_1\alpha \\ u_2\alpha \\ \vdots \\ u_m\alpha \end{pmatrix}$. Next, the sum of *m* vectors is simply the sum of the individual entries, so we simply have

$$\begin{pmatrix} u_{1,1}\alpha_1 + u_{1,2}\alpha_2 + \cdots + u_{1,n}\alpha_n \\ u_{2,1}\alpha_1 + u_{2,2}\alpha_2 + \cdots + u_{2,n}\alpha_n \\ \vdots \\ u_{m,1}\alpha_1 + u_{m,2}\alpha_2 + \cdots + u_{m,n}\alpha_n \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix}.$$

Thus, if these two vectors are equal, each of the entries must be equal, and thus we have

$$\begin{aligned} u_{1,1}\alpha_1 + u_{1,2}\alpha_2 + \cdots + u_{1,m}\alpha_m &= v_1 \\ u_{2,1}\alpha_1 + u_{2,2}\alpha_2 + \cdots + u_{2,m}\alpha_m &= v_2 \\ &\vdots \\ u_{m,1}\alpha_1 + u_{m,2}\alpha_2 + \cdots + u_{m,n}\alpha_n &= v_m \end{aligned}$$

For example, suppose we want to find if the vector $\begin{pmatrix} 2.1 \\ 5.3 \\ -4.8 \\ 6.7 \end{pmatrix}$ can be written as a linear combination of the three vectors $\begin{pmatrix} 6.2 \\ -0.5 \\ -0.2 \\ 1.2 \end{pmatrix}$, $\begin{pmatrix} 0.3 \\ 7.1 \\ 0.2 \\ -0.5 \end{pmatrix}$ and $\begin{pmatrix} -1.3 \\ 0.4 \\ 4.8 \\ 0.9 \end{pmatrix}$. We would need to find α_1, α_2 and α_3 such that

$$\alpha_1 \begin{pmatrix} 6.2 \\ -0.5 \\ -0.2 \\ 1.2 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0.3 \\ 7.1 \\ 0.2 \\ -0.5 \end{pmatrix} + \alpha_3 \begin{pmatrix} -1.3 \\ 0.4 \\ 4.8 \\ 0.9 \end{pmatrix} = \begin{pmatrix} 2.1 \\ 5.3 \\ -4.8 \\ 6.7 \end{pmatrix},$$

or, using the technique above, we would need to solve

$$\begin{aligned} 6.2\alpha_1 + 0.3\alpha_2 - 1.3\alpha_3 &= 2.1 \\ -0.5\alpha_1 + 7.1\alpha_2 + 0.4\alpha_3 &= 5.3 \\ -0.2\alpha_1 + 0.2\alpha_2 + 4.8\alpha_3 &= -4.8 \\ 1.2\alpha_1 + -0.5\alpha_2 + 0.9\alpha_3 &= 6.7 \end{aligned}$$

This is a system of four linear equations in three unknowns. Similarly, we can transform any system of linear equations into a question of finding a linear combination of vectors equaling a given vector. Given our previous example of a system of three linear equations in three unknowns

$$\begin{aligned} 2x + 3y + z &= -0.7 \\ x + 2y + 4z &= 0.1 \\ -3x + 4y + 2z &= -2.5 \end{aligned}$$

this is equivalent to finding

$$\begin{pmatrix} 2x + 3y + z \\ x + 2y + 4z \\ -3x + 4y + 2z \end{pmatrix} = \begin{pmatrix} -0.7 \\ 0.1 \\ -2.5 \end{pmatrix},$$

and we can write the left-hand vector as a sum of three vectors:

$$\begin{pmatrix} 2x \\ x \\ -3x \end{pmatrix} + \begin{pmatrix} 3y \\ 2y \\ 4y \end{pmatrix} + \begin{pmatrix} z \\ 4z \\ 2z \end{pmatrix} = \begin{pmatrix} -0.7 \\ 0.1 \\ -2.5 \end{pmatrix}.$$

Now, each of the entries in the three vectors is multiplied by a variable—this is the effect of scalar multiplication, so we can rewrite this as

$$x\begin{pmatrix}2\\1\\-3\end{pmatrix}+y\begin{pmatrix}3\\2\\4\end{pmatrix}+z\begin{pmatrix}1\\4\\2\end{pmatrix}=\begin{pmatrix}-0.7\\0.1\\-2.5\end{pmatrix}.$$

Thus, solving the system of three linear equations in three unknowns is equivalent to asking: does there exist a linear combination of the three vectors $\begin{pmatrix}2\\1\\-3\end{pmatrix},\begin{pmatrix}3\\2\\4\end{pmatrix},\begin{pmatrix}1\\4\\2\end{pmatrix}$ that equals $\begin{pmatrix}-0.7\\0.1\\-2.5\end{pmatrix}$?

Problems

1. Write this system of two linear equations as a question finding a linear combination of a given collection of vectors equaling a target vector:

$$\begin{aligned}2x+3y-z&=5\\x+2y&=3\end{aligned}.$$

2. Write this system of two linear equations as a question finding a linear combination of a given collection of vectors equaling a target vector:

$$\begin{aligned}2x+3y+z&=6\\x+2y+3z&=6\end{aligned}\quad\text{and}\quad\begin{aligned}2x+3y+z&=-0.7\\x+2y+4z&=0.1\\-3x+4y+2z&=-2.5\end{aligned}.$$

3. Write this problem of finding a target vector as a linear combination of a given set of vectors as a system of linear equations:

$$\alpha_1\begin{pmatrix}2.1\\0.5\\-1.2\end{pmatrix}+\alpha_2\begin{pmatrix}6.7\\9.0\\4.3\end{pmatrix}+\alpha_3\begin{pmatrix}8.5\\4.9\\-1.2\end{pmatrix}+\alpha_4\begin{pmatrix}-1.2\\-4.5\\-8.8\end{pmatrix}=\begin{pmatrix}2.3\\-5.3\\-1.0\end{pmatrix}$$

4. Write this problem of finding a target vector as a linear combination of a given set of vectors as a system of linear equations:

$$\alpha_1\begin{pmatrix}2\\1\\-4\\3\\7\end{pmatrix}+\alpha_2\begin{pmatrix}4\\9\\3\\-1\\6\end{pmatrix}+\alpha_3\begin{pmatrix}3\\6\\3\\4\\1\end{pmatrix}=\begin{pmatrix}8\\12\\-3\\5\\8\end{pmatrix}$$

5. Write this system of four linear equations as a question finding a linear combination of a given collection of vectors equaling a target vector:

$$\begin{aligned} 3w &= 5 \\ 4x &= -3 \\ 2y &= 9 \\ -3z &= 8 \end{aligned}$$

Solutions

1. $x\begin{pmatrix}2\\1\end{pmatrix}+y\begin{pmatrix}3\\2\end{pmatrix}+z\begin{pmatrix}-1\\0\end{pmatrix}=\begin{pmatrix}5\\3\end{pmatrix}$.

3. $\begin{aligned} 2.1\alpha_1 + 6.7\alpha_2 + 8.5\alpha_3 - 1.2\alpha_4 &= 2.3 \\ 0.5\alpha_1 + 9.0\alpha_2 + 4.9\alpha_3 - 4.5\alpha_4 &= -5.3 \\ -1.2\alpha_1 + 4.3\alpha_2 - 1.2\alpha_3 - 8.8\alpha_4 &= -1.0 \end{aligned}$.

5. $w\begin{pmatrix}3\\0\\0\\0\end{pmatrix}+x\begin{pmatrix}0\\4\\0\\0\end{pmatrix}+y\begin{pmatrix}0\\0\\2\\0\end{pmatrix}+z\begin{pmatrix}0\\0\\0\\-3\end{pmatrix}=\begin{pmatrix}5\\-3\\9\\8\end{pmatrix}$

6.4 A systematic approach to solving a system of linear equations

We have seen that solving a system of m linear equations in n unknowns and finding whether a linear combination of a given collection of n vectors in $\mathbf{F}^m$ equals a given vector in $\mathbf{F}^m$ are equivalent: if you can solve one, you can solve the other. We have also seen that we can determine if there exists a solution to a system of linear equations by performing a sequence of operations where we add a multiple of one equation onto another. What we want to do is to create an algorithm that will allow us to do this methodically. To do this, we will

1. define matrices,
2. given a system of linear equations or finding whether a linear combination of a collection of vectors equals a given target vector, we will define the *augmented matrix*,
3. describe the possible row operations,
4. define leading zeros and row-echelon form, and
5. conclude with defining a numerically stable Gaussian elimination algorithm on the augmented matrix.

6.4.1 Matrices, indexing and dimensions

First, given either a system of linear equations or a collection of vectors, we will associate with these equations a *matrix*. Suppose we are attempting to determine if there exists a linear combination of m vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$, all of which are in $\mathbf{R}^m$, we will define the matrix A to be

$$A = \begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \end{pmatrix}.$$

For example, if we are finding which linear combination of vectors $\begin{pmatrix} 6.2 \\ -0.5 \\ -0.2 \\ 1.2 \end{pmatrix}, \begin{pmatrix} 0.3 \\ 7.1 \\ 0.2 \\ -0.5 \end{pmatrix}, \begin{pmatrix} -1.3 \\ 0.4 \\ 4.8 \\ 0.9 \end{pmatrix}$ equals the vector $\begin{pmatrix} 2.1 \\ 5.3 \\ -4.8 \\ 6.7 \end{pmatrix}$, we will define the corresponding matrix, which we will call A, to be

$$A = \begin{pmatrix} 6.2 & 0.3 & -1.3 \\ -0.5 & 7.1 & 0.4 \\ -0.2 & 0.2 & 4.8 \\ 1.2 & -0.5 & 0.9 \end{pmatrix}.$$

Note that the matrix does not include vector we are attempting to find. Now, suppose we are attempting to find a solution to the system of linear equations

$$\begin{aligned} 2x + 3y + z &= -0.7 \\ x + 2y + 4z &= 0.1 \\ -3x + 4y + 2z &= -2.5 \end{aligned}$$

The corresponding matrix, which we will call B, will be

$$B = \begin{pmatrix} 2 & 3 & 1 \\ 1 & 2 & 4 \\ -3 & 4 & 2 \end{pmatrix}$$

while the target vector is $\mathbf{v} = \begin{pmatrix} 0.7 \\ 0.1 \\ -2.5 \end{pmatrix}$.

Now, given a 5-dimensional vector $\mathbf{v} = \begin{pmatrix} 0.3 \\ -4.5 \\ 2.8 \\ 7.6 \\ -9.1 \end{pmatrix}$, it is an unambiguous statement that 7.6 is the fourth entry in the vector and we would write $v_4 = 7.6$.

How, however, do we describe the location of '1.2' in the matrix A or '3' in B? First, we will always represent a matrix with an italicized capital letter; for example, A, B, C, etc. In some cases, we may write a matrix with a subscript, so if we had three matrices, we may label them A_1, A_2 and A_3. Like vectors, the entries of a matrix will be the corresponding italisized lower-case letters, but now indexed by two variables; for example, an entry of the matrix A may be the entry $a_{i,j}$ where i and j are indices.

Warning: you will have to get used to the fact that j is sometimes used as an index, and other times as the imaginary unit equal to $j \stackrel{def}{=} \sqrt{-1}$. You will have to get used to this. As the indices of a matrix or vector are strictly positive integers, it makes no sense for the "j" in v_j or $a_{i,j}$ to represent the square root of –1. Similarly, if you see $a_{i,j} = 3.2 - 4.5j$, the j on the left-hand side of the equation must be an index, while the j on the right-hand side is obviously the square root of –1.

This still, however, begs the question, in the given matrix

$$C = \begin{pmatrix} 1.3 & -2.4 & 7.6 & 8.0 & 9.5 \\ 2.7 & 1.8 & 0.9 & 6.5 & 7.4 \\ 9.4 & 7.3 & 8.5 & 1.2 & 0.6 \\ 0.5 & 9.6 & 1.3 & 4.7 & 2.8 \end{pmatrix},$$

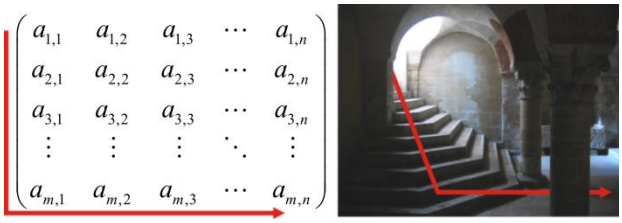
does $c_{1,3}$ refer to 7.6 or 9.4? We will adopt the convention that the first index indicates the row, and the second the column of the entry; thus, $c_{1,3} = 7.6$ and $c_{1,3} = -9.4$. Similarly, we will describe C as a 4×5 matrix, where "4" indicates the number of rows, and "5" the number of columns. We will say that these are the *dimensions* of the matrix.

In general, if A is an $m \times n$ matrix, we may write it as

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots & a_{2,n} \\ a_{3,1} & a_{3,2} & a_{3,3} & \cdots & a_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & a_{m,3} & \cdots & a_{m,n} \end{pmatrix}$$

The best *aide-memoire* this author has come up with to remember which comes first, the rows or the columns, is that many first-year students positively hate the subject of linear algebra. Consequently, consider a crypt in a graveyard, which is underground. First you go "down the stairs, and into the crypt". The first number tells you which row you are in (or the number of rows) and the second tells you the column (or the number of columns). As an aside, this

lecture, when taught in the fall academic term, tends to be given close to the chidlren's holiday of Hallowe'en. This *aide-memoire* is shown here:



You can also think of $u_{i,j}$ as the "i^{th} entry of the jth column of the matrix A."

For example, given a 3 × 3 matrix $B = \begin{pmatrix} 2 & 3 & 1 \\ 1 & 2 & 4 \\ -3 & 4 & 2 \end{pmatrix}$, we can think of this matrix as being composed of three column vectors

$$\mathbf{b}_1 = \begin{pmatrix} 2 \\ 1 \\ -3 \end{pmatrix},\ \mathbf{b}_2 = \begin{pmatrix} 3 \\ 2 \\ 4 \end{pmatrix} \text{ and } \mathbf{b}_3 = \begin{pmatrix} 1 \\ 4 \\ 2 \end{pmatrix},$$

and so some times we may write $B = (\mathbf{b}_1\, \mathbf{b}_2\, \mathbf{b}_3)$ or sometimes $B = (\mathbf{b}_{:,1}\, \mathbf{b}_{:,2}\, \mathbf{b}_{:,3})$ to emphasize that these are the columns of the matrix *B*.

Similarly, we may think of the matrix $A = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots & a_{2,n} \\ a_{3,1} & a_{3,2} & a_{3,3} & \cdots & a_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & a_{m,3} & \cdots & a_{m,n} \end{pmatrix}$ as being comprised of *n* vectors in $\mathbf{R}^m$ where

$A = (\mathbf{a}_1 \quad \mathbf{a}_2 \quad \mathbf{a}_3 \quad \cdots \quad \mathbf{a}_n)$ or $A = (\mathbf{a}_{:,1} \quad \mathbf{a}_{:,2} \quad \mathbf{a}_{:,3} \quad \cdots \quad \mathbf{a}_{:,n})$ where $\mathbf{a}_k = \mathbf{a}_{:,k} = \begin{pmatrix} a_{1,k} \\ a_{2,k} \\ a_{3,k} \\ \vdots \\ a_{m,k} \end{pmatrix}$.

Definition

The entries of a matrix that have equal indices ($a_{1,1}$, $a_{2,2}$, etc.) are described as being the *diagonal entries* of the matrix. These are shown visually in Figure 53.

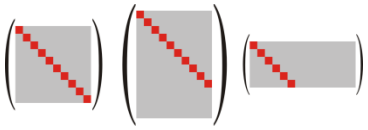


Figure 53. A visual representation of the diagonal entries, shown in red, of three matrices.

Definition

The entries of a matrix that have unequal indices ($a_{1,1}$, $a_{2,2}$, etc.) are described as being the *off-diagonal entries* of the matrix. All entries that are not diagonal entries are off-diagonally entries. These are shown visually in Figure 54.

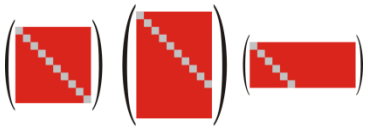


Figure 54. A visual representation of the off-diagonal entries, shown in red, of three matrices.

The diagonal entries of these three matrices are highlighted in red, and all off-diagonal entries are black.

$$\begin{pmatrix} 3.1 & 0.7 & 0.9 \\ 0.7 & 4.2 & -0.5 \\ 0.9 & -0.5 & 5.9 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0.1 & 0.2 & 0.3 & 0.4 & 0.5 & 0.6 \\ 0.01 & 0.04 & 0.09 & 0.16 & 0.25 & 0.36 \end{pmatrix}.$$

The *diagonal* entries of the matrix are usually expressed as the ordered list $(a_{1,1}, a_{2,2}, \ldots, a_{\min(m,n),\ \min(m,n)})$. Thus, the diagonal entries of the above three matrices are (3.1, 4.2, 5.9), (1, 4) and (1, 0.2, 0.09), respectively.

Definition

The *super-diagonal* entries[13] of a matrix are those entries directly above/to the right of the diagonal entries. The super-diagonal entries are off-diagonal entries. These are shown visually in Figure 55.

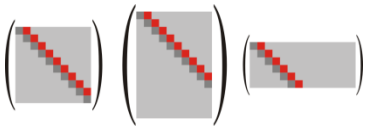


Figure 55. A visual representation of the super-diagonal entries, shown in red, immediately above/to the right of the diagonal entries (shown in darker gray) of three matrices.

The *super-diagonal* entries of the matrix are usually expressed as the ordered list $(a_{1,2}, a_{2,3}, \ldots, a_{\min(m,n-1),\ \min(m+1,n)})$.[14] Thus, the diagonal entries of the above three matrices are (0.7, –0.5); (2); and (1, 0.3, 0.16), respectively.

Definition

The *sub-diagonal* entries[15] of a matrix are those entries directly below/to the left of the diagonal entries. The super-diagonal entries are off-diagonal entries. These are shown visually in Figure 56.

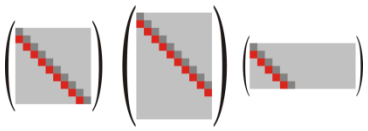


Figure 56. A visual representation of the sub-diagonal entries, shown in red, immediately below/to the left of the diagonal entries (shown in darker gray) of three matrices.

[13] The prefix is borrowed from the Latin preposition *super* meaning "above, on top or beyond".

[14] Do not memorize this formula; simply appreciate that we can explicitly describe the relationship between the super- and sub-diagonal entries mathematically.

[15] The prefix is borrowed from the Latin preposition *sub* meaning "under or beneath".

The *sub-diagonal* entries of the matrix are usually expressed as the ordered list $(a_{2,1}, a_{3,2}, \ldots, a_{\min(m,n+1),\ \min(m-1,n)})$.[16] Thus, the diagonal entries of the above three matrices are (0.7, –0.5); (3, 6); and (0.1, 0.04), respectively.

At this point, you are probably wondering why we are so focused on the columns of the matrix, as opposed to the rows. This will become clear in time, and this intimate relationship between a matrix and its columns is why we have been representing vectors "vertically"; that is, we have always written the 5-dimensional vector $\mathbf{u}$ as $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{pmatrix}$ and never as $\mathbf{u} = \begin{pmatrix} u_1 & u_2 & u_3 & u_4 & u_5 \end{pmatrix}$.

Problems

1. Given the matrix

$$A = \begin{pmatrix} 4.7 & -2.4 & 2.7 & -5.0 & 5.9 \\ 2.1 & 8.3 & 1.3 & -0.2 & 5.7 \\ 2.5 & -1.2 & 7.3 & 8.8 & 2.1 \\ 5.6 & 1.1 & 4.5 & 0.9 & -0.9 \end{pmatrix},$$

what are the entries $a_{3,2}$, $a_{1,6}$ and $a_{2,5}$?

2. Given the matrix

$$B = \begin{pmatrix} 2.8 & -5.4 & -7.1 \\ -0.3 & 4.5 & 2.9 \\ 3.8 & 2.7 & -9.9 \\ 5.7 & 6.1 & 4.8 \\ -5.9 & -3.2 & 0.7 \\ 2.3 & 0.5 & -0.2 \end{pmatrix},$$

what are the entries $b_{1,3}$, $b_{6,2}$ and $b_{5,1}$?

3. What are the diagonal, super-diagonal and sub-diagonal entries of the matrix A in Problem 1?

4. What are the diagonal, super-diagonal and sub-diagonal entries of the matrix B in Problem 2?

[16] Same as footnote 14.

Solutions

1. –1.2, does not exist and 5.7.

3. The diagonal entries are (4.7, 8.3, 7.3, 0.9), the super-diagonal entries are (–2.4, 1.3, 8.8, –0.9) and the sub-diagonal entries are (2.1, –1.2, 4.5).

6.4.2 Classification of matrices

Matrices may be described by their shape and by the non-zero entries.

Definition
If A is an $n \times n$ matrix (that is, the number of rows equals the number of columns), we say the matrix is *square*. A matrix that is not square is sometimes described as being *rectangular*.

Definition
If A is an $m \times n$ matrix where all the entries are zero, A is said to be the *zero matrix* and is often written as $\emptyset_{m,n}$.

Definition
If A is an $m \times n$ matrix where all off-diagonal entries are zero, A is said to be a *diagonal matrix*. That is, $a_{i,j} = 0$ if $i \neq j$.

Rather than writing zeros in all the off-diagonal locations, it is common to place a single large zero both to the right of and below the diagonal to indicate the matrix is a diagonal matrix, as shown in Figure 57. The zero matrix is a diagonal matrix, and it is possible for some of the diagonal entries of a diagonal matrix to be zero.

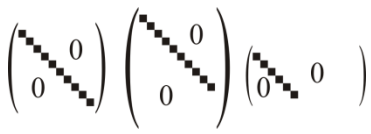


Figure 57. Three diagonal matrices, the first of which is square and the other two are rectangular matrices. The entries on the diagonal may be zero or non-zero.

Definition
If A is an $m \times n$ matrix where all entries below the diagonal are zero, A is said to be an *upper-triangular matrix*. That is, $a_{i,j} = 0$ if $i > j$.

Rather than writing zeros in all the entries below the diagonal, it is common to place a single large zero below the diagonal to indicate the matrix is an upper-triangular, as shown in Figure 58. Every diagonal matrix is also an upper-triangular matrix.

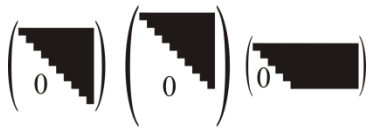


Figure 58. Three upper-triangular matrices, the first of which is square and the other two are rectangular matrices. The entries to the right of the diagonal may be zero or non-zero.

Definition
If A is an $m \times n$ matrix where all entries to the right of the diagonal are zero, A is said to be an *lower-triangular matrix*. That is, $a_{i,j} = 0$ if $i < j$.

To emphasize that a matrix is lower triangular, we may write it as $A_{\searrow}$.

Rather than writing zeros in all the entries to the right of the diagonal, it is common to place a single large zero to the right of the diagonal to indicate the matrix is a lower-triangular, as shown in Figure 59. Every diagonal matrix is also a lower-triangular matrix.

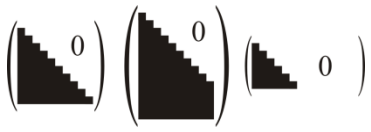


Figure 59. Three lower-triangular matrices, the first of which is square and the other two are rectangular matrices. The entries below the diagonal may be zero or non-zero.

Definition
If A is an $m \times n$ matrix where all entries to the right of the super-diagonal and all the entries below the sub-diagonal are zero, A is said to be a *tri-diagonal matrix*. That is, $a_{i,j} = 0$ if $i < j - 1$ or $i > j + 1$.

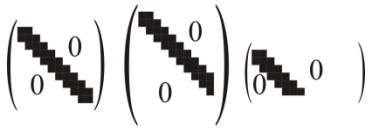


Figure 60. Three tri-diagonal matrices where, the first of which is square and two that are rectangular matrices. The entries on the diagonal, super-diagonal and sub-diagonal may be zero or non-zero.

For example, the matrix

$$\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}$$

is a tri-diagonal matrix used in finding solutions to Laplace's equation in one dimension. Laplace's equation describes systems affected by natural forces that are in *steady state*, meaning that the conditions are unchanging. You will note that occasionally, rather than writing out all the zeros, we will leave them as blanks, with the assumption that the entry at that location is zero: it is much easier to identify the important features of the matrix.

6.4.3 Augmented matrices

Now that we have defined the matrix associated with an ordered collection of vectors, our next step is to use this to devise a systematic algorithm for determining the solution or solutions (if any) to a system of linear equations. Given the system of m linear equations in n unknowns or linear combination of an ordered collection of n vectors in $\mathbf{R}^m$ equaling a target vector,

$$\begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,m}x_m &= y_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,m}x_m &= y_2 \\ &\vdots \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,m}x_m &= y_n \end{aligned} \quad \text{or} \quad x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_m\mathbf{a}_m = \mathbf{y},$$

we will define the associated *augmented matrix* as the matrix associated with the left-hand sides augmented by a column of the right-hand sides:

$$\left(\begin{array}{ccccc:c} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,n} & y_1 \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots & a_{2,n} & y_2 \\ a_{3,1} & a_{3,2} & a_{3,3} & \cdots & a_{3,n} & y_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m,1} & a_{m,2} & a_{m,3} & \cdots & a_{m,n} & y_m \end{array}\right).$$

We may, in some cases, represent the augmented matrix by $\left(A \;\vdots\; \mathbf{y}\right)$, indicating that the augmented matrix is the concatenation of the $m \times n$ matrix A and the n-dimensional vector $\mathbf{y}$. The result of this concatenation is an $m \times (n + 1)$ dimensional augmented matrix. It makes no sense to concatenate anything other than an m-dimensional vector to an $m \times n$ matrix. Note also that as the unknowns are also ordered, there is no reason we cannot represent the solution as a vector itself:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}.$$

Note that the column vectors are in $\mathbf{F}^m$ while the solution vector $\mathbf{x}$ is in $\mathbf{F}^n$.

For example, given the system of linear equations or linear combination of an ordered collection of vectors,

$$\begin{aligned} 2x + 3y + z &= -0.7 \\ x + 2y + 4z &= 0.1 \\ -3x + 4y + 2z &= -2.5 \end{aligned} \quad \text{or} \quad x\begin{pmatrix} 2 \\ 1 \\ -3 \end{pmatrix} + y\begin{pmatrix} 3 \\ 2 \\ 4 \end{pmatrix} + z\begin{pmatrix} 1 \\ 4 \\ 2 \end{pmatrix} = \begin{pmatrix} -0.7 \\ 0.1 \\ -2.5 \end{pmatrix},$$

respectively, the corresponding augmented matrix is therefore

$$\left(\begin{array}{ccc:c} 2 & 3 & 1 & -0.7 \\ 1 & 2 & 4 & 0.1 \\ -3 & 4 & 2 & -2.5 \end{array}\right).$$

The solution vector could be represented as $\mathbf{w} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$. Solving such a system of linear equations is not immediately obvious, especially if we are dealing with significantly larger systems of linear equations, say 1000 linear equations in 1000 unknowns. There is, however, one system of linear equations that can be solved easily enough, one represented by an augmented matrix that has a *shape* like:

$$\left(\begin{array}{ccc|c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ 0 & 0 & -56 & -11.2 \end{array}\right).$$

The steps in solving this are:

1. The third row says that $-56z = -11.2$, or $z = 0.2$.
2. We then look at the second row, which says $0.5y + 3.5z = 0.45$, but $z = 0.2$, so $0.5y + 3.5 \cdot 0.2 = 0.45$, so solving this for y yields $y = -0.5$.
3. Finally, the first row says $2x + 3y + z = -0.7$, as given the two previous findings, it follows that $x = 0.3$.

The solution can thus be represented by the vector

$$\mathbf{w} = \begin{pmatrix} 0.3 \\ -0.5 \\ 0.2 \end{pmatrix}.$$

A much easier example is when we have the target vector being the zero vector, as in

$$\left(\begin{array}{ccc|c} 2 & 3 & 1 & 0 \\ 0 & 0.5 & 3.5 & 0 \\ 0 & 0 & -56 & 0 \end{array}\right)$$

With a little thought, it becomes obvious that the only solution is $x = y = z = 0$, with the solution vector being

$$\mathbf{w} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Note that the zero vector is a subspace of $\mathbf{R}^3$.

As a different example,

$$\left(\begin{array}{ccc|c} 2 & 3 & -1 & 4.9 \\ 0 & 2 & 1 & 0.6 \end{array}\right).$$

The steps in solving this are:

1. The second row says $2y + z = 0.6$. We can solve for either y in terms of z or z in terms of y. We will choose to solve for y in terms of z and thus we will declare z to be a free variable—it can take on any value we wish, and the value of y now depends on the value we assign z. Thus, $y = 0.3 - 0.5z$.

2. The first row says $2x + 3y - z = 4.9$. We know $y = 0.3 - 0.5z$ and z is still a free variable, so substituting these into the equation gives us that $2x + 3(0.3 - 0.5z) - z = 4.9$, and solving for x yields $x = 1.25z + 2$.

Thus, the solution vector is

$$\mathbf{w} = \begin{pmatrix} 1.25z + 2 \\ 0.3 - 0.5z \\ z \end{pmatrix}.$$

Here is something interesting now: recall the definitions of vector addition and scalar multiplication. We can rewrite this vector as

$$\mathbf{w} = \begin{pmatrix} 1.25z + 2 \\ 0.3 - 0.5z \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 0.3 \\ 0 \end{pmatrix} + \begin{pmatrix} 1.25z \\ -0.5z \\ z \end{pmatrix} = \begin{pmatrix} 2 \\ 0.3 \\ 0 \end{pmatrix} + z \begin{pmatrix} 1.25 \\ -0.5 \\ 1 \end{pmatrix}.$$

Thus, $\mathbf{w} = \begin{pmatrix} 2 \\ 0.3 \\ 0 \end{pmatrix} + z \begin{pmatrix} 1.25 \\ -0.5 \\ 1 \end{pmatrix}$ is one solution, but so is $\mathbf{w} = \begin{pmatrix} 7 \\ -1.7 \\ 4 \end{pmatrix}$ when $z = 4$, as is $\mathbf{w} = \begin{pmatrix} 1.6625 \\ 0.435 \\ -0.27 \end{pmatrix}$ when $z = -0.27$. In this case, there are infinitely many solutions, and the collection of all solutions forms a line in $\mathbf{R}^3$, as shown in Figure 61. You will note that the line does not go through the origin, so this is not a subspace of $\mathbf{R}^3$.

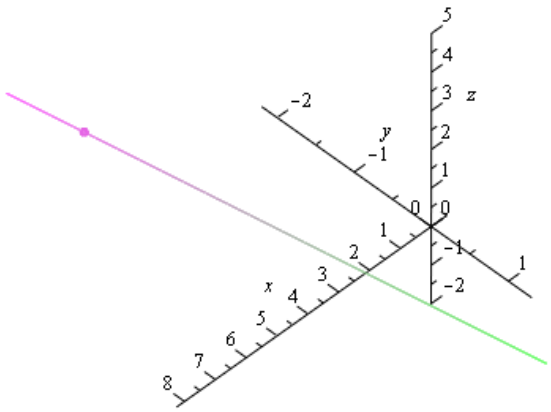


Figure 61. All solutions to the system of linear equations $2x + 3y - z = 4.9$ and $2y + z = 0.6$.

An important concept: In this example, we allowed z to be the free variable. It does not matter if we picked y or z to be the free variable. For the purposes of this course, we will choose the free variable via a descriptive algorithm, thus guaranteeing that if you follow the algorithm, you will find all solutions. For example, we could have chosen y to be the free variable instead of z, so $z = 0.6 - 2y$, and thus solving for x, we have $x = 2.75 - 2.5y$, so all solutions are in the form

$$\mathbf{w} = \begin{pmatrix} 2.75 - 2.5y \\ y \\ 0.6 - 2y \end{pmatrix} \text{ instead of } \mathbf{w} = \begin{pmatrix} 1.25z + 2 \\ 0.3 - 0.5z \\ z \end{pmatrix}.$$

Does it matter which representation we use? No: both define all operations. For example, if we let $z = 17.5$ in the second case, we would get the solution

$$\mathbf{w} = \begin{pmatrix} 1.25 \cdot 17.5 + 2 \\ 0.3 - 0.5 \cdot 17.5 \\ 17.5 \end{pmatrix} = \begin{pmatrix} 23.875 \\ -8.45 \\ 17.5 \end{pmatrix}.$$

With a little algebra, you will note that you get this exact same solution in the first case when $y = -8.45$, as

$$\mathbf{w} = \begin{pmatrix} 2.75 - 2.5(-8.45) \\ -8.45 \\ 0.6 - 2(-8.45) \end{pmatrix} = \begin{pmatrix} 23.875 \\ -8.45 \\ 17.5 \end{pmatrix}.$$

Thus, if there is a choice of free variables, it does not matter which variable we allow to be free, it is just that we will have an algorithm that autmatocially picks one.

If we take the same system of linear equations, but set the target vector to be the zero vector, that is,

$$\left(\begin{array}{ccc|c} 2 & 3 & -1 & 0 \\ 0 & 2 & 1 & 0 \end{array}\right),$$

the collection of all solutions may be described by

$$\mathbf{w} = \begin{pmatrix} 1.25z \\ -0.5z \\ z \end{pmatrix} = z\begin{pmatrix} 1.25 \\ -0.5 \\ 1 \end{pmatrix}.$$

You will notice that now all of the solutions form a line passing through the origin, or a subspace of $\mathbf{R}^3$, as shown in Figure 62.

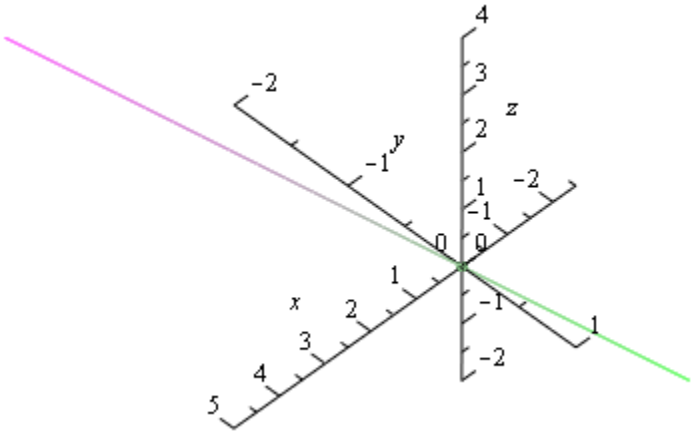


Figure 62. All solutions to the system of linear equations 2x + 3y – z = 0 and 2y + z = 0.

Finally, consider the absurd system of linear equations represented by

$$\left(\begin{array}{cc|c} 2 & 3 & -1 \\ 0 & 0 & 5 \end{array}\right).$$

The last line says $0 \cdot x + 0 \cdot y = 5$, which is of course absurd—such a system of linear equations cannot have a solution. Consequently, we would say that such a system of linear equations has no solution.

In the next few sections, we will define *row-echelon form* for an augmented matrix, and the algorithm of backward substitution for solving a system of linear equations when they are in this form. Next, we will see how we can convert any system of linear equations to row-echelon form using the algorithm of *Gaussian elimination with partial pivoting*.

Problems

1. Write the following system of linear equations as an augmented matrix:

$$\begin{aligned} 3w + 4y + z &= 3 \\ 2x + z &= 5 \\ y + z &= 3 \end{aligned}$$

2. Write the following system of linear equations as an augmented matrix:

$$\begin{aligned} -x + y - 3z &= 4 \\ -3x + 2z &= -4 \\ 2y + 5z &= 12 \\ x + y + z &= 5 \end{aligned}$$

3. Write the following augmented matrix as a system of linear equations and as a problem of finding a target vector as a linear combination of a given collection of vectors:

$$\left(\begin{array}{ccc:c} 2 & 3 & 1 & -0.7 \\ 1 & 2 & 4 & 0.1 \\ -3 & 4 & 2 & -2.5 \end{array}\right)$$

4. Write the following augmented matrix as a system of linear equations and as a problem of finding a target vector as a linear combination of a given collection of vectors

$$\left(\begin{array}{cccccc:c} 2 & -1 & 0 & 0 & 0 & 0 & 0.3 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0.5 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0.6 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0.8 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0.9 \\ 0 & 0 & 0 & 0 & -1 & 2 & 1.1 \end{array}\right)$$

6.4.4 Leading zeros and leading non-zero entries of a matrix

To define row-echelon form, we must first define various properties of rows within a matrix or augmented matrix.

Definition

Given a matrix A, the ith row has k *leading zeros* if $a_{ij} = 0$ for $j = 1, \ldots, k$ and $a_{i,k+1} \neq 0$.

Definition

If a row within a matrix has k leading zeros, the entry $a_{i,k+1}$ is described as the *leading non-zero entry* of the ith row.

Definition

A row that consists entirely of zeros is simply said to be a *zero row*.

Consider the matrix

$$A = \begin{pmatrix} 0 & 0 & 0 & 1.3 & 0 & 0 \\ 3.8 & 2.3 & 3.4 & 0 & 0 & 0 \\ 0 & 0.3 & 4.2 & 4.7 & 5.8 & -0.3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -6.4 & 0 & 2.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4.9 \end{pmatrix}.$$

In this matrix A,

1. Row 1 has three leading zeros and $a_{1,4} = 1.3$ is the leading non-zero entry,
2. Row 2 has no leading zeros and $a_{2,1} = 3.8$ is the leading non-zero entry,
3. Row 3 has one leading zero and $a_{3,2} = 0.3$ is the leading non-zero entry,
4. Row 4 is a zero row,
5. Row 5 has one leading zero and $a_{5,2} = -6.4$ is the leading non-zero entry, and
6. Row 6 has five leading zeros and $a_{6,6} = 4.9$ is the leading non-zero entry.

Next, we will describe a property of a column within a matrix with respect to the leading non-zero entries:

Definition

A column j in a matrix has a leading non-zero entry if at least one row i within the matrix has the entry $a_{i,j}$ being a leading non-zero entry of that row.

In the above matrix:

1. Column 1 has one leading non-zero entry, namely $a_{2,1} = 3.8$,
2. Column 2 has two leading non-zero entries, namely $a_{3,2} = 0.3$ and $a_{5,2} = -6.4$,
3. Column 3 has no leading non-zero entries,
4. Column 4 has one leading non-zero entry, namely $a_{1,4} = 1.3$,
5. Column 5 has no leading non-zero entries, and
6. Column 6 has one leading non-zero entry, namely $a_{6,6} = 4.9$.

It is important to realize that it does not matter what comes after the first non-zero entry; we are counting only those zeros in the row that appear to the left of the first non-zero entry in that row. For example, the rows of the matrices

$$B = \begin{pmatrix} 0.1 & 0 & 0 & 0 & 0 \\ 0 & 2.3 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 8.7 & 0 \\ 0 & 0 & 0 & 0 & -4.5 \end{pmatrix} \text{ and } C = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 6 & 7 & 8 & 9 \\ 0 & 0 & 10 & 11 & 12 \\ 0 & 0 & 0 & 13 & 14 \\ 0 & 0 & 0 & 0 & 15 \end{pmatrix}$$

have 0, 1, 2, 3 and 4 leading zeros in their respective rows. The leading non-zero entries in the matrix B above are 0.1, 2.3, 0.3, 8.7 and –4.5, and the leading non-zero entries of the matrix C are 1, 6, 10, 13 and 15. Each column has exactly one leading non-zero entry.

The augmented matrix

$$D = \left(\begin{array}{cccccc|c} 0.5 & 0.7 & 0.3 & 0.1 & -0.3 & 0.4 & -1.2 \\ 0 & 0 & 1.2 & 0 & 0 & 0 & -2.3 \\ 0 & 0 & 0 & 0 & 0 & 0.9 & -0.8 \end{array}\right)$$

has 0, 2 and 5 leading zeros in each of the rows, respectively. The leading non-zero entries are $d_{1,1} = 0.5$, $d_{2,3} = 1.2$ and $d_{3,6} = 0.9$. Three columns that contain leading non-zero entries: the first, third and sixth.

Problems

1. Describe the properties of the matrices with respect to leading zeros and columns with leading non-zero entries.

$$\left(\begin{array}{ccc|c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ 0 & 0 & -56 & -11.2 \end{array}\right), \left(\begin{array}{ccc|c} 2 & 3 & -1 & 4.9 \\ 0 & 2 & 1 & 0.6 \end{array}\right) \text{ and } \left(\begin{array}{cc|c} 2 & 3 & -1 \\ 0 & 0 & 5 \end{array}\right).$$

2. Describe the properties of the matrices with respect to leading zeros and columns with leading non-zero entries.

$$\left(\begin{array}{ccc|c} 2 & 3 & 1 & 0 \\ 0 & 0.5 & 3.5 & 0 \\ 0 & 0 & -56 & 0 \end{array}\right), \left(\begin{array}{ccc|c} 2 & 3 & -1 & 0 \\ 0 & 2 & 1 & 0 \end{array}\right) \text{ and } \left(\begin{array}{cc|c} 2 & 3 & 0 \\ 0 & 0 & 0 \end{array}\right).$$

Solutions

1. The first matrix has 0, 1 and 2 leading non-zero entries, the second has 0 and 1 leading non-zero entries, and the third has 0 and 2 leading non-zero entries. The first matrix has columns 1, 2 and 3 with leading non-zero entries, the second has columns 1 and 2 have leading non-zero entries, and the third has columns 1 and 3 with leading non-zero entries.

6.4.5 Row-echelon form and rank

Having defined leading zero entries, we will now describe a very specific *shape* of a matrix. In the navy, a squadron of ships is said to be sailing in "echelon" if each subsequent ship is behind and to the same side of the ship in front of it.



Figure 63. Ships maneuvering in echelon form.

For example, during the Battle of the Nile (Nelson versus Napoleon), the Leander, Swiftsure, Alexander and Majestic were sailing in echelon form, as shown in Figure 64.



Figure 64. Formation of ships within the English fleet during the prelude to the Battle of the Nile.

Definition

A matrix or augmented matrix is in *row-echelon form* if:

1. all the zero rows are below any rows that contain at least one non-zero entry, and
2. the number of leading zeros in any row containing at least one non-zero entry has more leading zeros than the row immediately preceding it.

For example, the following augmented matrices are all in row-echelon form:

$$\left(\begin{array}{ccc|c} 3 & 2 & 5 & 0 \\ 0 & 9 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{ccccc|c} 3 & 2 & 5 & 4 & 1 & 9 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 9 \end{array}\right), \left(\begin{array}{cccc|c} 3 & 0 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \end{array}\right).$$

Definition

The *rank* of a matrix that is in row-echelon form is the number of non-zero rows.

Given an augmented matrix, one may consider separately the rank of the matrix (those entries to the left of the dotted line, or those columns other than the target vector) and the rank of the augmented matrix. In the above three examples, the ranks of the matrices are 2, 3 and 4, respectively, while the ranks of the augmented matrices are 3, 3 and 4, respectively.

Given an augmented matrix that is in row-echelon form, the rank of the matrix itself may either be equal to or one less than the rank of the augmented matrix. In the following three examples,

$$\left(\begin{array}{ccc:c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ 0 & 0 & -56 & -11.2 \end{array}\right), \left(\begin{array}{ccc:c} 2 & 3 & -1 & 4.9 \\ 0 & 2 & 1 & 0.6 \end{array}\right) \text{ and } \left(\begin{array}{cc:c} 2 & 3 & -1 \\ 0 & 0 & 5 \end{array}\right),$$

both the matrix and the augmented matrix have rank 3 in the first case, and both the matrix and the augmented matrix have rank 2 in the second. In the third case, however, the matrix has rank 1 while the augmented matrix has rank 2. It is not possible for the augmented matrix to have a rank more than one more than the matrix itself, as for this to occur, you would have to have a situation like

$$\left(\begin{array}{cc:c} 3 & 2 & 5 \\ 0 & 9 & 0 \\ 0 & 0 & 8 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{array}\right),$$

but this matrix is not in row-echelon form: the third and fourth rows both have two leading non-zero entries.

Problems

1. Which of the following augmented matrices are in row-echelon form, and if they are in row-echelon form, what are the ranks of the both the matrix and the augmented matrix?

$$\left(\begin{array}{cc:c} 3 & 2 & 5 \\ 2 & 9 & 0 \\ 0 & 5 & 8 \\ 0 & 3 & 6 \\ 0 & 0 & 4 \end{array}\right), \left(\begin{array}{cccc:c} 0 & 2 & 2 & 4 & 2 \\ 0 & 0 & 4 & 1 & 3 \\ 0 & 0 & 0 & 5 & 4 \\ 0 & 0 & 0 & 0 & 6 \end{array}\right), \left(\begin{array}{ccccc:c} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{cccc:c} 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 4 & 1 & 3 \\ 0 & 2 & 0 & 5 & 4 \\ 5 & 4 & 3 & 0 & 6 \end{array}\right), \left(\begin{array}{cc:c} 3 & 0 & 0 \\ 0 & 2 & -16 \\ 0 & -1 & 8 \\ 0 & 0 & 0 \end{array}\right), \left(\begin{array}{ccc:c} 3 & 3 & 4 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 5 \end{array}\right)$$

Solutions

1. The solutions are:

a. The first matrix is not in row-echelon form because the second row does not have more leading zeros than the first.
b. The second matrix is in row-echelon form, and the rank of the matrix is 3 while the rank of the augmented matrix is 4.
c. The third matrix is in row-echelon form, and both the matrix and augmented matrix have rank 2.
d. The forth matrix is not in row-echelon form as the second row does not have more leading zeros than the first row.
e. The fifth matrix is not in row-echelon form as the third row does not have more leading zeros than the second.
f. The sixth matrix is not in row-echelon form as the zero row is not in the last row of the matrix.

6.4.6 Backward substitution

In general, we cannot easily find a linear combination of a given collection of vector $\mathbf{u}_1,\ldots,\mathbf{u}_n$ in $\mathbf{F}^m$ that equals a target vector $\mathbf{v}$ also in $\mathbf{F}^m$; that is, it is *difficult* if there exist scalars $\alpha_1,\ldots,\alpha_n$ such that

$$\alpha_1\mathbf{u}_1+\cdots+\alpha_n\mathbf{u}_n=\mathbf{v}\,.$$

We will represent the solution as an n-dimensional vector in $\mathbf{F}^n$,

$$\boldsymbol{\alpha}=\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_n\end{pmatrix}.$$

The first point we will note is that there are in general to classes of problems:

1. when the target vector is the zero vector, $\mathbf{v}=\mathbf{0}_m$, and
2. when the target vector is not the zero vector, $\mathbf{v}\neq\mathbf{0}_m$.

If the target vector is $\mathbf{v}=\mathbf{0}_m$, we are guaranteed that there is at least one solution: $\boldsymbol{\alpha}=\mathbf{0}_n$, for

$$0\cdot\mathbf{u}_1+\cdots+0\cdot\mathbf{u}_n=\mathbf{0}_m\,.$$

If the target vector is not the zero vector ($\mathbf{v}\neq\mathbf{0}_m$), however, it is may be possible that no solutions exist. However, if a solution exists, it may either be

1. a unique solution of a vector $\boldsymbol{\alpha}$, or
2. an infinite number of solution.

Finding out whether or not there is a solution, and if a solution exists, is it unique or describing all possible solutions if there are infinitely many is, in general, very difficult. There is one special case, however, where it is possible to find all solutions to this problem in a straight-forward manner: if the matrix $U=(\mathbf{u}_1\cdots\mathbf{u}_n)$ is defined as the $m \times n$ matrix with the vectors $\mathbf{u}_k$ as its columns, and the augmented matrix $(U\ \vdots\ \mathbf{v})$ is already in row-echelon form, we can easily find any solutions using *backward substitution*.

For example, if we wanted to see if the vector $\mathbf{v}=\begin{pmatrix}1.5\\3.5\\5.2\\2.1\end{pmatrix}$ can be written as a linear combination of $\left\{\begin{pmatrix}3\\0\\0\\0\end{pmatrix},\begin{pmatrix}-1\\4\\0\\0\end{pmatrix},\begin{pmatrix}2\\5\\8\\0\end{pmatrix},\begin{pmatrix}-3\\6\\4\\7\end{pmatrix}\right\}$, the resulting augmented matrix would be

$$\left(\begin{array}{cccc|c} 3 & -1 & 2 & -3 & 1.5 \\ 0 & 4 & 5 & 6 & 3.5 \\ 0 & 0 & 8 & 4 & 5.2 \\ 0 & 0 & 0 & 7 & 2.1 \end{array}\right),$$

which is in row-echelon form. Starting with the last row,

1. the fourth row indicates that $7\alpha_4 = 2.1$ or $\alpha_4 = \frac{2.1}{7} = 0.3$,
2. the third row indicates that $8\alpha_3 + 4\alpha_4 = 5.2$ or $\alpha_3 = \frac{5.2 - 4 \cdot 0.3}{8} = \frac{4}{8} = 0.5$,
3. the second row indicates that $4\alpha_2 + 5\alpha_3 + 6\alpha_4 = 3.5$ or $\alpha_2 = \frac{3.5 - 5 \cdot 0.5 - 6 \cdot 0.3}{4} = \frac{-0.8}{4} = -0.2$, and
4. the first row indicates that $4\alpha_1 - \alpha_2 + 2\alpha_3 - 3\alpha_4 = 1.2$ or $\alpha_1 = \frac{1.5 + (-0.2) - 2 \cdot 0.5 + 3 \cdot 0.3}{3} = \frac{1.2}{3} = 0.4$.

Thus, the solution vector is $\boldsymbol{\alpha} = \begin{pmatrix} 0.4 \\ -0.2 \\ 0.5 \\ 0.3 \end{pmatrix}$.

One problem we will often need to solve is when the target vector is the zero vector; that is, we want to see if the vector $\mathbf{v} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ can be written as a linear combination of $\left\{\begin{pmatrix} 3 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ 4 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 5 \\ 8 \\ 0 \end{pmatrix}, \begin{pmatrix} -3 \\ 6 \\ 4 \\ 7 \end{pmatrix}\right\}$. This is so important in linear algebra that we describe such a system of linear equations as a *homogeneous system of linear equations*. The resulting augmented matrix would be

$$\left(\begin{array}{cccc|c} 3 & -1 & 2 & -3 & 0 \\ 0 & 4 & 5 & 6 & 0 \\ 0 & 0 & 8 & 4 & 0 \\ 0 & 0 & 0 & 7 & 0 \end{array}\right),$$

which, too, is in row-echelon form. The solution here, however, is easier: $\alpha_4 = \alpha_3 = \alpha_2 = \alpha_1 = 0$ and therefore the solution vector is $\boldsymbol{\alpha} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \mathbf{0}_4$.

Now, it is not always the case that there is one unique solution as is the case in the these two examples. If an augmented matrix $(U \mid \mathbf{v})$ is in row-echelon form, the backward substitution algorithm will either tell us

1. if there are no solutions,
2. if there is a unique solution in which case the algorithm will find that solution, or
3. if there are infinitely many solutions in which case the algorithm will describe all possible solutions.

You may observe that the augmented matrix representing a systems of linear equations is not in row-echelon form; however, we will see that when the augmented matrix is in row-echelon form, it is relatively easy to find the solution or solutions, if any, and then we will introduce the concept of row operations that allow us to transform a general augmented matrix into one that is in row-echelon form. This is a general engineering practice, and is shown in .

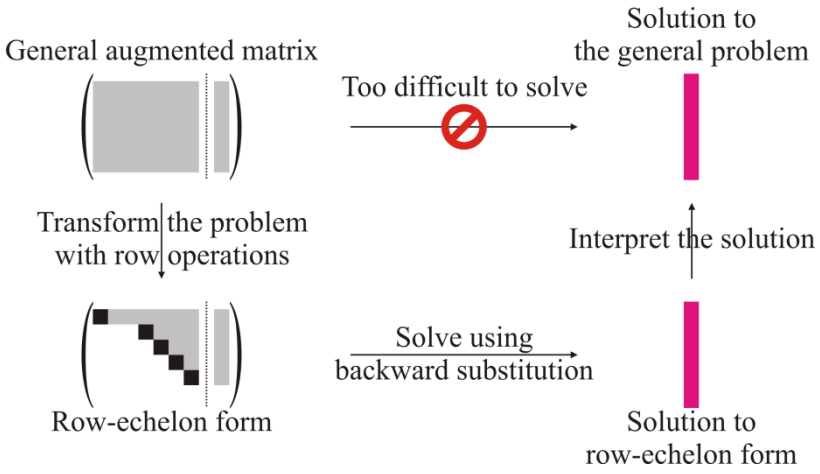


Figure 65. Finding solutions to a system of linear equations by transforming the augmented matrix into one in row-echelon form.

As an example of how this is done, consider the problem of adding hundreds of numbers. It is too inefficient to add so many numbers effectively using the human brain, but it is possible enter those numbers into a computer. The computer converts the numbers to binary, the numbers are then added using binary arithmetic, and the answer is then converted back to a decimal number, as shown in .

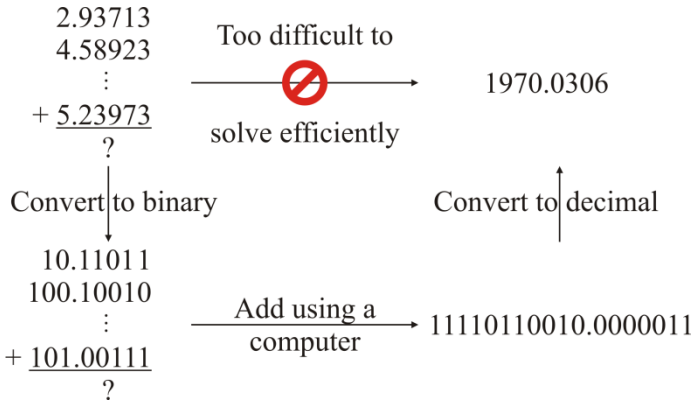


Figure 66. The process of adding numbers by converting them to binary, adding them using binary arithmetic in a computer, and then converting the solution back to a decimal number.

There are five possible cases:

1. A homogeneous system of linear equations, or when the target vector is **0**, there is either
    a. one unique solution, or
    b. infinitely many solutions forming a subspace of $\mathbf{R}^n$; and
2. a general system of linear equations when the target vector is not **0**, there are either
    a. no solutions,
    b. one unique solution, or
    c. infinitely many solutions.

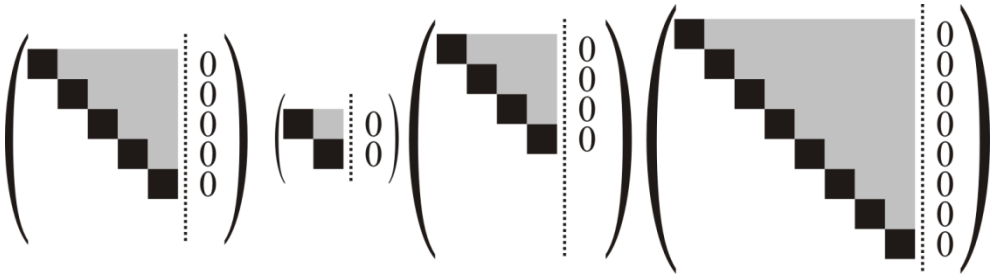
We will look at each of these five cases.

6.4.6.1 One unique solution when v = 0

If we want to find if there exists a linear combination of n vectors $\mathbf{u}_1,\ldots,\mathbf{u}_n$ in $\mathbf{F}^m$ that equals the zero vector, that is,

$$\alpha_1\mathbf{u}_1+\cdots+\alpha_n\mathbf{u}_n=\mathbf{0},$$

and the augmented matrix $(U \mid \mathbf{0})$ is in row-echelon form, the solution is unique if and only if $\operatorname{rank}(U)=\operatorname{rank}(U \mid \mathbf{0})=n$. For example, all of the following augmented matrices, where black squares indicate leading non-zero entries, satisfy this condition:



For example, there is only one way to add the vectors

$$\left\{\begin{pmatrix}3\\0\\0\\0\end{pmatrix},\begin{pmatrix}-1\\4\\0\\0\end{pmatrix},\begin{pmatrix}2\\5\\8\\0\end{pmatrix}\right\}$$

to equal the zero vector, for the augmented matrix is

$$\left(\begin{array}{ccc|c}3&-1&2&0\\0&4&5&0\\0&0&8&0\\0&0&0&0\end{array}\right).$$

To prove that this is the only solution if $\text{rank}\left(U \;\vdots\; \mathbf{0}\right) = n$, we note that

n. *Starting* with n, $u_{n,n}\alpha_n = 0$ implies that $\alpha_n = 0$,

$n-1$. Next, $u_{n-1,n-1}\alpha_{n-1} + u_{n-1,n}\cdot 0 = 0$ implies that $\alpha_{n-1} = 0$,

$n-2$. Next, $u_{n-2,n-2}\alpha_{n-2} + u_{n-2,n-1}\cdot 0 + u_{n-2,n}\cdot 0 = 0$ implies that $\alpha_{n-2} = 0$,

and repeating this another $n-3$ times, we see that $\alpha_1 = \alpha_2 = \alpha_3 = \cdots = \alpha_{n-1} = \alpha_n = 0$, and thus the solution vector is

$$\boldsymbol{\alpha} = \mathbf{0}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

We call this solution the *trivial solution* because it is so obviously correct. Note that the zero vector is a subspace of $\mathbf{R}^n$.

All of the following are equivalent ways of saying that there is a unique solution to a homogeneous system of linear equations or finding when $\alpha_1\mathbf{u}_1 + \cdots + \alpha_n\mathbf{u}_n = \mathbf{0}$ when the matrix U is in row-echelon form:

1. rank(U) = n,
2. the diagonal entries of U are all non-zero,
3. each column has one row containing a non-zero entry in that column, and
4. there are no columns that do not contain a leading non-zero entry of some row.

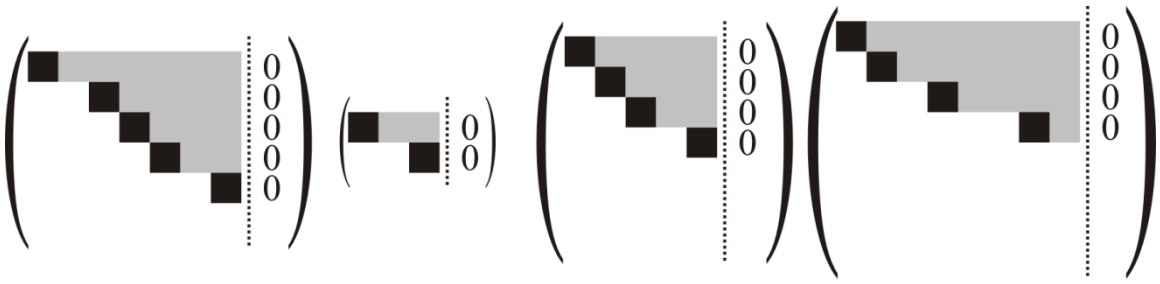
If any of these is true, there is a unique solution.

6.4.6.2 Infinitely many unique solution when v = 0

Because the rank of an augmented matrix $\text{rank}\left(U\right) \le n$, if $\text{rank}\left(U\right) \ne n$, the only other possibility is that $\text{rank}\left(U\right) < n$. In this case, the number of solutions to finding a linear combination of n vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ in $\mathbf{F}^m$ that equals the zero vector, that is,

$$\alpha_1\mathbf{u}_1 + \cdots + \alpha_n\mathbf{u}_n = \mathbf{0},$$

and the augmented matrix $\left(U \;\vdots\; \mathbf{0}\right)$ is in row-echelon form, there must be an infinite number of solutions. For example, all of the following augmented matrices, where black squares indicate leading non-zero entries, satisfy this condition:



In the first case, the rank is five while $n = 7$; in the second, the rank is two while $n = 3$; in the third, the rank is four while $n = 5$, and in the last the rank is also four, but $n = 8$.

Each time a row has a leading non-zero entry $u_{i,j}$, we can solve for the corresponding α_j by observing that this means

$$u_{i,j}\alpha_j + u_{i,j+1}\alpha_{j+1} + u_{i,j+2}\alpha_{j+2} + \cdots + u_{i,n}\alpha_n = 0,$$

and so

$$\alpha_j = \frac{-u_{i,j+1}\alpha_{j+1} - u_{i,j+2}\alpha_{j+2} - \cdots - u_{i,n}\alpha_n}{u_{i,j}}.$$

The problem is, we cannot solve for an α_k if there isn't a row that has a leading non-zero entry in the kth column. For example, consider

$$\left(\begin{array}{ccc|c} 2 & 4 & 1 & 0 \\ 0 & 0 & 5 & 0 \end{array}\right).$$

The second line says $5\alpha_3 = 0$ so $\alpha_3 = 0$. The first line, once we substitute the value of $\alpha_3 = 0$, we get $2\alpha_1 + 4\alpha_2 = 0$. If we choose to solve for α_1, we have that $\alpha_1 = -2\alpha_2$. Regardless of the value of α_2, if we set α_1 appropriately, we will have a solution. Thus, all solutions can be written as

$$\begin{pmatrix} -2\alpha_2 \\ \alpha_2 \\ 0 \end{pmatrix} = \alpha_2 \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}.$$

Notice that this describes all scalar multiples of the vector $\begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}$, which is a subspace of $\mathbf{R}^3$.

Similarly, consider the example

$$\left(\begin{array}{ccccc|c} 4 & -4 & 2 & -3 & 2 & 0 \\ 0 & 0 & 5 & 4 & 6 & 0 \\ 0 & 0 & 0 & 2 & -7 & 0 \end{array}\right)$$

where columns 2 and 5 do not have leading non-zero entries. Thus,

3. the third row says $2\alpha_4 - 7\alpha_5 = 0$ so $\alpha_4 = 3.5\alpha_5$,
2. the second row says $5\alpha_3 + 4\alpha_4 + 6\alpha_5 = 0$, but $\alpha_4 = 3.5\alpha_5$ so $5\alpha_3 + 4(3.5\alpha_5) + 6\alpha_5 = 0$ so $\alpha_3 = -4\alpha_5$, and
1. the first row says $4\alpha_1 - 4\alpha_2 + 2\alpha_3 - 3\alpha_4 + 2\alpha_5 = 0$, and substituting in our values of $\alpha_4 = 3.5\alpha_5$ and $\alpha_3 = -4\alpha_5$, we get that $4\alpha_1 - 4\alpha_2 + 2(-4\alpha_5) - 3(3.5\alpha_5) + 2\alpha_5 = 0$ so $\alpha_1 = \alpha_2 + 4.125\alpha_5$.

Thus, our solution vector is

$$\begin{pmatrix} \alpha_2 + 4.125\alpha_5 \\ \alpha_2 \\ -4\alpha_5 \\ 3.5\alpha_5 \\ \alpha_5 \end{pmatrix} = \begin{pmatrix} \alpha_2 \\ \alpha_2 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 4.125\alpha_5 \\ 0 \\ -4\alpha_5 \\ 3.5\alpha_5 \\ \alpha_5 \end{pmatrix} = \alpha_2 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \alpha_5 \begin{pmatrix} 4.125 \\ 0 \\ -4 \\ 3.5 \\ 1 \end{pmatrix}.$$

You will note that all linear combinations of the two vectors $\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4.125 \\ 0 \\ -4 \\ 3.5 \\ 1 \end{pmatrix}$ forms a subspace of $\mathbf{R}^5$.

For example, if we let $\alpha_2 = 1, \alpha_5 = 0$, we have that $\begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} -4 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$. Similarly, if we let $\alpha_2 = 0, \alpha_5 = 1$, we have that $4.125\begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix} - 4\begin{pmatrix} 2 \\ 5 \\ 0 \end{pmatrix} + 3.5\begin{pmatrix} -3 \\ 4 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 6 \\ -7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, but also if we let $\alpha_2 = 3, \alpha_5 = -4$, we have the solution vector is

$$3\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} - 4\begin{pmatrix} 4.125 \\ 0 \\ -4 \\ 3.5 \\ 1 \end{pmatrix} = \begin{pmatrix} -13.5 \\ 3 \\ 16 \\ -14 \\ -4 \end{pmatrix}$$

and we note $-13.5\begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix} + 3\begin{pmatrix} -4 \\ 0 \\ 0 \end{pmatrix} + 16\begin{pmatrix} 2 \\ 5 \\ 0 \end{pmatrix} - 14\begin{pmatrix} -3 \\ 4 \\ 2 \end{pmatrix} - 4\begin{pmatrix} 2 \\ 6 \\ -7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$. Indeed, any linear combination of the two vectors $\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4.125 \\ 0 \\ -4 \\ 3.5 \\ 1 \end{pmatrix}$ is a solution to the homogeneous system of linear equations posed by these three equations in five unknowns.

In general, all coefficients associated with columns with no leading non-zero entries will be considered to be free variables, and we will then solve for all coefficients associated with columns with leading non-zero entries in terms of these. The result will can always be written as a linear combination of solution vectors. We start with the last non-zero row, and solve for the coefficient associated with the leading non-zero entry. Having done so, we move the previous row and, after substituting the last coefficient, we solve for the coefficient associated with the leading non-zero entry, and so on up until we have solved for the first coefficient.

At this point, writing down an algorithm is rather difficult, and so therefore it is much simpler to describe it. However, we can make some observations:

Theorem
If we are trying to solve a system of *m* homogeneous linear equations in *n* unknowns, or equivalently, finding a linear combination of *n* vectors in $\mathbf{R}^m$ to equal the zero vector $\mathbf{0}_m$, and $m < n$, then there must be infinitely many solutions.

Proof:
The maximum rank of $(U \mid \mathbf{0})$ is *m*; that is, $\text{rank}(U \mid \mathbf{0}) \le m$, so if $m < n$, it follows that $\text{rank}(U \mid \mathbf{0}) < n$. Thus, there are always an infinite number of linear combinations of *n* vectors in $\mathbf{F}^m$. ∎

All of the following are equivalent ways of saying that there are infinitely many solutions to a homogeneous system of linear equations or finding when $\alpha_1\mathbf{u}_1 + \cdots + \alpha_n\mathbf{u}_n = \mathbf{0}$ when the matrix *U* is in row-echelon form:

1. rank(*U*) < *n*,
2. at least one diagonal entries of *U* is zero,
3. there exists at least one column that does not have a leading non-zero entry of one row.

If any of these is true, there is a unique solution.

6.4.6.3 No solutions when v ≠ 0

When the target vector is **0**, there is always at least one solution to a system of linear equations; however, if the target vector is not the zero vector, it is possible that there may be no solutions. If the augmented matrix $(U \mid \mathbf{v})$ is in row-echelon form, it is quite easy to describe when a system of linear equations has no solutions:

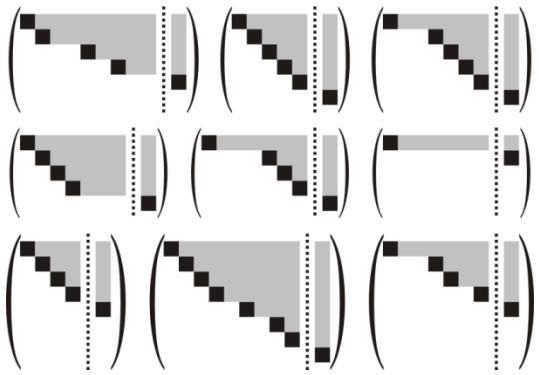
Theorem
If an augmented matrix $(U \mid \mathbf{v})$ is in row-echelon form and the rank of $(U \mid \mathbf{v})$ is greater than the rank of the matrix *U*, there are no solutions.

Proof:
If a row within the augmented matrix has all zeros except in the right-most column, this represents an equation of the form

$$0 \cdot \alpha_1 + 0 \cdot \alpha_2 + \cdots + 0 \cdot \alpha_m = \beta$$

where β is not zero. This is, however, a contradiction, for the left-hand side must be zero for any real or complex values of $\alpha_1, \alpha_2, \ldots, \alpha_m$. Thus, no solution exists. ∎

For example, all of the following augmented matrices are in row-echelon form, where the black squares represent a leading-non-zero entry and the grey squares represent numbers that may or may not be zero. In each case, the rank of the augmented matrix is one greater than the rank of the matrix itself. Thus, no solution can exist.



As an actual example, consider the system of linear equations represented by the augmented matrix in row-echelon form

$$\left(\begin{array}{ccc|c} 1 & 4 & 7 & 1 \\ 0 & -3 & -6 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right).$$

Here we see the rank of the matrix is two, while the rank of the augmented matrix is three; thus, no solution exists: it is impossible to write the vector $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$ as a linear combination of the vectors $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 \\ -3 \\ 0 \end{pmatrix}, \begin{pmatrix} 7 \\ -6 \\ 0 \end{pmatrix}$.

6.4.6.4 *A unique solution when v ≠ 0*

A unique solution exists under very specific circumstances similar to that when there is a unique solution when the target vector is the zero vector.

Theorem

The problem of finding $\alpha_1\mathbf{u}_1 + \cdots + \alpha_n\mathbf{u}_n = \mathbf{v}$ has a unique solution when the augmented matrix $\left(U \mid \mathbf{v}\right)$ is in row-echelon form has a unique solution if $\operatorname{rank}\left(U \mid \mathbf{v}\right) = \operatorname{rank}\left(U\right) = n$. Equivalently, there is a unique solution if the ranks are equal and the diagonal entries of U are all non-zero.

The proof of this theorem will be an algorithm that is defined to find the unique solution. If A is an $m \times m$ matrix, and therefore $\left(A \mid \mathbf{v}\right)$ is an $m \times (m + 1)$ augmented matrix and $\mathbf{x}$ is the solution vector, we will find the entries of $\mathbf{x}$ in reverse order:

$$n. \quad x_n = \frac{v_n}{a_{n,n}},$$

$$n-1. \quad x_{n-1} = \frac{v_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}},$$

$n-2.\quad x_{n-2} = \dfrac{v_{n-2} - a_{n-2,n-1}x_{n-1} - a_{n-2,n}x_n}{a_{n-2,n-2}}$,

and so on, so that in general, one calculates

$i.\quad x_i = \dfrac{v_i - \sum_{j=i+1}^{m} a_{i,j}x_j}{a_{i,i}}$.

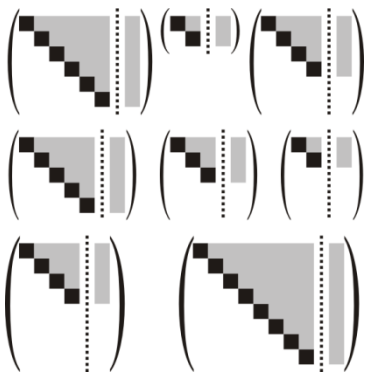
Because we assumed that $a_{i,i} \neq 0$ for $i = 1, \ldots, m$, none of these calculations can fail. This process is repeated until we have calculated the value of x_1.

If we change any one value x_i of the solution vector **x**, the linear equation represented by the ith row cannot be satisfied, and thus the modified vector cannot be a solution. If we change more than one entry, we need look only at the last value that was changed.

Proof of the theorem

If all the diagonal entries are non-zero, there are no operations in the backward substitution algorithm that can produce a division-by-zero, and this algorithm produces the unique solution. ∎

For example, the following all represent augmented matrices in row-echelon form where one unique solution exists. The black square indicates a leading non-zero entry, and all gray squares could represent either zero or non-zero values. In each case, however, the rank of the matrix is equal to the rank of the augmented matrix.



The simplest backward substitution is when the matrix is diagonal; for example, consider solving

$$\left(\begin{array}{ccc|c} 2 & 0 & 0 & -7 \\ 0 & 4 & 0 & 10 \\ 0 & 0 & 10 & -12 \end{array}\right).$$

In this case, the solution is $x_3 = -1.2$, $x_2 = 2.5$ and $x_1 = -3.5$, or $\mathbf{x} = \begin{pmatrix} -1.2 \\ 2.5 \\ -3.5 \end{pmatrix}$.

All of the following are equivalent ways of saying that there is a unique solution to a non-homogeneous system of linear equations or finding when $\alpha_1\mathbf{u}_1 + \cdots + \alpha_n\mathbf{u}_n = \mathbf{v}$ when $\mathbf{v} \neq \mathbf{0}$ and the matrix U is in row-echelon form:

1. $\text{rank}\left(U \mid \mathbf{v}\right) = \text{rank}\left(U\right) = n$,
2. $\text{rank}\left(U \mid \mathbf{v}\right) = \text{rank}\left(U\right)$ and the diagonal entries of U are all non-zero,
3. $\text{rank}\left(U \mid \mathbf{v}\right) = \text{rank}\left(U\right)$ and each column has one row containing a non-zero entry in that column, and
4. $\text{rank}\left(U \mid \mathbf{v}\right) = \text{rank}\left(U\right)$ and there are no columns that do not contain a leading non-zero entry of some row.

If any of these is true, there is a unique solution.

6.4.6.5 Infinitely many solutions when v ≠ 0

Finally, we come to the case where there are an infinite number of solutions. As with the homogenous case,

Theorem

If the rank of augmented matrix $\left(A \mid \mathbf{v}\right)$ in row-echelon form is equal to the rank of the matrix A and at least one diagonal entry of the matrix A is zero, there is an infinity of solutions.

Again, we will construct an algorithm that finds these solutions. We will divide the columns into

1. those columns with leading non-zero entries, and
2. those columns without.

If a column does not have a leading non-zero entry, we will decide that the variable corresponding to that column may be freely chosen to be any value what-so-ever; that is, we will call that entry a *free variable*. For example, in the augmented matrix in row-echelon form

$$\left(\begin{array}{rrrrr|r} 2 & 2 & 4 & -7 & 4 & -3.2 \\ 0 & 4 & 3 & 10 & 3 & 11.1 \\ 0 & 0 & 10 & -12 & 6 & -10.6 \end{array}\right),$$

Columns 4 and 5 do not have leading non-zero entries, and thus x_4 and x_5 will be free variables.

Important: The last row says $10x_3 - 12x_4 + 6x_5 = -10.6$, so we are not forced to solve for x_3 in terms of x_4 and x_5; instead, we could solve for x_4 in terms of x_3 and x_5 or also x_5 in terms of x_3 and x_4. However, we require an algorithmic approach, and thus we will choose to solve for the column with the leading non-zero entry in terms of others.

In the augmented matrix in row-echelon form

$$\left(\begin{array}{rrrrr|r} 2 & 2 & 4 & -7 & 4 & 1 \\ 0 & 0 & 3 & 10 & 3 & -0.4 \\ 0 & 0 & 0 & 0 & 6 & 3.6 \end{array}\right)$$

Columns 2 and 4 do not have leading non-zero entries, and thus x_2 and x_4 will be free variables.

Finally, in the augmented matrix in row-echelon form

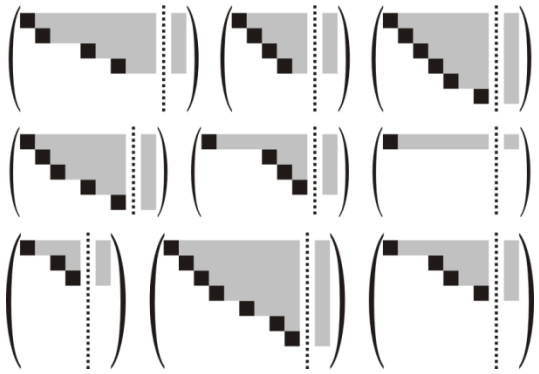
$$\left(\begin{array}{ccccc|c} 2 & 2 & 4 & -7 & 4 & 1 \\ 0 & 0 & 0 & 10 & 3 & 3.8 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

Columns 2, 3 and 5 do not have leading non-zero entries, so x_2, x_3 and x_5 will be free variables. For each variable depending on a free variable, you will start with the largest, working you way down solving for smaller. The formula is similar, but

$$x_i = \frac{v_i - \sum_{k=j+1}^{n} a_{i,k} x_k}{a_{i,j}}$$

Proof:
When we apply the backwards-substitution algorithm, if the entry $a_{i,i}$ is the first entry on the diagonal that is non-zero, we are free to give the ith coefficient any value we wish, and consequently, there will be an infinity of solutions. ∎

For example, the following all represent augmented matrices in row-echelon form where an infinity of solutions exist. The black square indicates a leading non-zero entry, and all gray squares could represent either zero or non-zero values. In each case, however, the rank of the matrix is equal to the rank of the augmented matrix.



As an example of this, consider

$$\left(\begin{array}{ccc|c} 1 & 4 & 7 & 3 \\ 0 & -3 & -6 & -3 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

Here, we immediately note that we can allow $-3\alpha_2 - 6\alpha_3 = -3$, and so we can allow α_3 to be any value we wish, in which case $\alpha_2 = 1 - 2\alpha_3$. Substituting this into the first equation, this says $\alpha_1 + 4\alpha_2 + 7\alpha_3 = 3$ but $\alpha_2 = 1 - 2\alpha_3$, so we have that $\alpha_1 = 3 - 4(1 - 2\alpha_3) + 7\alpha_3 = -1 + \alpha_3$. Thus, we have that

$$(-1+\alpha_3)\begin{pmatrix}1\\2\\3\end{pmatrix} + (1-2\alpha_3)\begin{pmatrix}4\\5\\6\end{pmatrix} + \alpha_3\begin{pmatrix}7\\8\\9\end{pmatrix} = \begin{pmatrix}3\\3\\3\end{pmatrix}.$$

The solution vector may be written as

$$\mathbf{a} = \begin{pmatrix} -1+\alpha_3 \\ 1-2\alpha_3 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

You will notice this defines a line in $\mathbf{R}^3$ that does not pass through the origin.

For example, we could allow $\alpha_3 = 0.5$ and we note that

$$-0.5\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + 0 \cdot \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + 0.5\begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}.$$

As a second example, suppose that our augmented matrix is in row-echelon form as follows:

$$\left(\begin{array}{ccccc|c} 2 & 2.4 & 0.4 & -1 & -1.6 & 16.2 \\ 0 & 0 & 4.2 & -10.5 & 2.1 & 18.9 \\ 0 & 0 & 0 & 3 & -2.4 & -7.2 \end{array}\right).$$

The leading non-zero entries are 2, 4.2 and 3 and therefore Columns 1, 3 and 4 contain leading non-zero entries. Consequently, Columns 2 and 5 do not contain leading non-zero entries and therefore we will consider the variables associated with those columns, α_2 and α_5, to be free; that is, we are free to assign those variables any values we wish. Consequently, we start with the third row:

$$3\alpha_4 - 2.4\alpha_5 = -7.2,$$

so we may deduce that $\alpha_4 = -2.4 + 0.8\alpha_5$. The second row says that

$$4.2\alpha_3 - 10.5\alpha_4 + 2.1\alpha_5 = 18.9,$$

and as $\alpha_4 = -2.4 + 0.8\alpha_5$, we have

$$4.2\alpha_3 - 10.5(-2.4 + 0.8\alpha_5) + 2.1\alpha_5 = 18.9.$$

Expanding and solving for α_3, we have $\alpha_3 = -1.5 + 1.5\alpha_5$. Finally, the first row says that

$$2\alpha_1 + 2.4\alpha_2 + 0.4\alpha_3 - \alpha_4 - 1.6\alpha_5 = 16.2$$

and given $\alpha_4 = -2.4 + 0.8\alpha_5$ and $\alpha_3 = -1.5 + 1.5\alpha_5$, we have

$$2\alpha_1 + 2.4\alpha_2 + 0.4(-1.5 + 1.5\alpha_5) - (-2.4 + 0.8\alpha_5) - 1.6\alpha_5 = 16.2$$

and solving for α_1, we have $\alpha_1 = 7.2 - 1.2\alpha_2 + 0.9\alpha_5$. Thus, the solution vector is of the form

$$\boldsymbol{\alpha} = \begin{pmatrix} 7.2 - 1.2\alpha_2 + 0.9\alpha_5 \\ \alpha_2 \\ -1.5 + 1.5\alpha_5 \\ -2.4 + 0.8\alpha_5 \\ \alpha_5 \end{pmatrix} = \begin{pmatrix} 7.2 \\ 0 \\ -1.5 \\ -2.4 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} -1.2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \alpha_5 \begin{pmatrix} 0.9 \\ 0 \\ 1.5 \\ 0.8 \\ 1 \end{pmatrix}.$$

With a little thought, you can see that because there are two free variables, this defines a 2-dimensional plane in $\mathbf{R}^5$.

As a slightly different example, consider the following:

$$\left(\begin{array}{cccc:c} 4 & 2.4 & 0.5 & 2 & 2.9 \\ 0 & 0 & 3 & 2.5 & 2.1 \\ 0 & 0 & 0 & 2 & 1.2 \end{array}\right)$$

In this case, $\alpha_4 = 0.6$ and $3\alpha_3 + 2.5\alpha_4 = 2.1$, so $\alpha_3 = 0.2$. The first row has one free variable, α_2, and thus

$$4\alpha_1 + 2.4\alpha_2 + 0.5\alpha_3 + 2\alpha_4 = 2.9$$

and thus $4\alpha_1 + 2.4\alpha_2 + 0.5 \cdot 0.2 + 2 \cdot 0.6 = 2.9$, so solving for α_1 yields $\alpha_1 = 0.4 - 0.6\alpha_2$. Writing this as a solution vector, we get

$$\boldsymbol{\alpha} = \begin{pmatrix} 0.4 - 0.6\alpha_2 \\ \alpha_2 \\ 0.2 \\ 0.6 \end{pmatrix} = \begin{pmatrix} 0.4 \\ 0 \\ 0.2 \\ 0.6 \end{pmatrix} + \alpha_2 \begin{pmatrix} -0.6 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

Again, the solution is a line that is not through the origin.

All of the following are equivalent ways of saying that there are infinitely many solutions to a non-homogeneous system of linear equations or finding when $\alpha_1\mathbf{u}_1 + \cdots + \alpha_n\mathbf{u}_n = \mathbf{v}$ when $\mathbf{v} \neq \mathbf{0}$ and the matrix U is in row-echelon form:

1. $\text{rank}\left(U \;\vdots\; \mathbf{v}\right) = \text{rank}\left(U\right) < n$,
2. $\text{rank}\left(U \;\vdots\; \mathbf{v}\right) = \text{rank}\left(U\right)$ and at least one diagonal entry of U is zero,
3. $\text{rank}\left(U \;\vdots\; \mathbf{v}\right) = \text{rank}\left(U\right)$ and there exists at least one column that does not contain a leading non-zero value of a row.

If any of these is true, there is are infinitely many solutions.

6.4.6.6 Summary of backward substitution for the general case

For backward substitution in finding a linear combination of n vectors in $\mathbf{R}^m$, there are three cases:

1. no solution exists,
2. one unique solution where that solution is a vector in $\mathbf{R}^m$, and
3. there are infinitely many solutions, and those solutions are parameterized by k free variables and the solution is of the form $\mathbf{v}_\text{p} + \alpha_{i_1}\mathbf{v}_{i_1} + \alpha_{i_2}\mathbf{v}_{i_2} + + \cdots + \alpha_{i_k}\mathbf{v}_{i_k}$

If the target vector is the zero vector $\mathbf{0}_m$, then there is guaranteed to be at least one solution, and the collection of all solutions is always a subspace of $\mathbf{R}^n$. If the target vector is not zero, there may be no solutions, one solution or infinitely many solutions, but if there are infinitely many solutions,

Conditions		The target vector v	
		$\mathbf{v} = \mathbf{0}_m$	$\mathbf{v} \neq \mathbf{0}_m$
$\text{rank}(U \mid \mathbf{v}) > \text{rank}(U)$		n/a	No solution
$\text{rank}(U \mid \mathbf{v}) = \text{rank}(U)$	$\text{rank}(U) = n$	The zero vector $\boldsymbol{\alpha} = \mathbf{0}_n$	A unique solution $\boldsymbol{\alpha} \neq \mathbf{0}_n$
	$\text{rank}(U) < n$	A subspace of $\mathbf{F}^n$	A subset of $\mathbf{F}^n$ not containing $\mathbf{0}_n$

6.4.7 Row operations

We will now describe three operations we can perform on augmented matrices: operations that modify one row at a time on the augmented matrix that preserve the solution to the corresponding system of linear equations. Such operations are called *row operations*, and based on this, we will define the concept of row equivalence. We will begin by defining the row operation of adding a multiple of one row onto another and defining row equivalence. This will be followed by defining the row operation of swapping two rows, and finally multiplying a single row by a non-zero scalar. In each case, we will see that row operations can always be *undone*.

6.4.7.1 Adding a scalar multiple of a row onto another

Given an augmented matrix, the *row operation* of adding a scalar multiple α of Row i onto Row j where $i \neq j$ will be represented by the symbol $R_{\alpha;i\to j}$. Thus, we will write

$$\left(\begin{array}{ccc|c} 2 & 3 & 1 & -0.7 \\ 1 & 2 & 4 & 0.1 \\ -3 & 4 & 2 & -2.5 \end{array}\right) \overset{R_{-0.5;1\to 2}}{\sim} \left(\begin{array}{ccc|c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ -3 & 4 & 2 & -2.5 \end{array}\right) \overset{R_{1.5;1\to 3}}{\sim} \left(\begin{array}{ccc|c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ 0 & 8.5 & 3.5 & -3.55 \end{array}\right) \overset{R_{-17;2\to 3}}{\sim} \left(\begin{array}{ccc|c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ 0 & 0 & 56 & 11.2 \end{array}\right)$$

with the understanding that we are moving from left to right. We will say that two augmented matrices are *row equivalent* if there exists a sequence of row operations that change one matrix into the other, and we will write this as

$$\left(\begin{array}{ccc|c} 2 & 3 & 1 & -0.7 \\ 1 & 2 & 4 & 0.1 \\ -3 & 4 & 2 & -2.5 \end{array}\right) \sim \left(\begin{array}{ccc|c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ 0 & 0 & 56 & 11.2 \end{array}\right).$$

The solution to the system of linear equations represented by the left-hand augmented matrix is the same as the solution to the system of linear equations represented by the right-hand augmented matrix.

Now, if we add α times Row i onto Row j, we can undo this operation by adding $-\alpha$ times Row i onto Row j, and thus we can say that the inverse of the row operation $R_{\alpha;i\to j}$, denoted $R^{-1}_{\alpha;i\to j}$, is the row operation, $R_{-\alpha;i\to j}$, and so we will write

$$R^{-1}_{\alpha;i\to j} = R_{-\alpha;i\to j}.$$

Note that if we apply such a row operation onto a matrix, and then its inverse, the combination leaves the matrix unchanged. Note therefore that if we start with the last matrix in the expression above and perform the inverse row operations in the reversed order, we end up with the original matrix:

$$\left(\begin{array}{cccc:c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ 0 & 0 & 56 & 11.2 \end{array}\right) \overset{R_{17;2\to3}}{\sim} \left(\begin{array}{cccc:c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ 0 & 8.5 & 3.5 & -3.55 \end{array}\right) \overset{R_{-1.5;1\to3}}{\sim} \left(\begin{array}{cccc:c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ -3 & 4 & 2 & -2.5 \end{array}\right) \overset{R_{0.5;1\to2}}{\sim} \left(\begin{array}{cccc:c} 2 & 3 & 1 & -0.7 \\ 1 & 2 & 4 & 0.1 \\ -3 & 4 & 2 & -2.5 \end{array}\right).$$

Problems

1. Given the matrix

$$A = \left(\begin{array}{cccc:c} 5 & 4 & 2 & 9 & 6 \\ -1.5 & 2.8 & -8.6 & 2.3 & -1.8 \\ 3.5 & 3.2 & 2.6 & 9.8 & 13.2 \\ 2 & 0.4 & 4.0 & 11.3 & 8 \end{array}\right),$$

perform the row operations $R_{0.3;1\to2}$, $R_{-0.7;1\to3}$ and $R_{-0.4;1\to4}$, in that order.

2. Given the matrix you finished with in Question 1, perform the row operations $R_{-0.1;2\to3}$, $R_{0.3;2\to4}$ and $R_{-0.4;3\to4}$.

3. Does it affect the final result if you perform the row operations in Question 1 in a different order? Why or why not?

4. Does it affect the final result if you perform the row operations in Question 2 in a different order? Why or why not?

5. If $R_{3.2;4\to5}$ is a row operation of adding 3.2 times Row 4 onto Row 5, and $R_{-2.9;4\to5}$ is the row operation of adding –2.9 times Row 4 onto Row 5, how could you write the combined operation of first performing $R_{3.2;4\to5}$ and then $R_{-2.9;4\to5}$? We will write this in the order $R_{-2.9;4\to5}R_{3.2;4\to5} = R_{?;4\to5}$.

6. Given the matrix A in Question 1, which row operation places a zero in the $(2, 1)^{\text{th}}$ entry,

$$R_{\frac{a_{1,1}}{a_{2,1}};1\to2},\ R_{-\frac{a_{1,1}}{a_{2,1}};1\to2},\ R_{\frac{a_{2,1}}{a_{1,1}};1\to2} \text{ or } R_{-\frac{a_{2,1}}{a_{1,1}};1\to2}?$$

7. What is the inverse operation of $R_{5.7;1\to2}$?

Solutions

1. You should end up with

$$A \sim \left(\begin{array}{cccc:c} 5 & 4 & 2 & 9 & 6 \\ 0 & 4 & -8 & 5 & 0 \\ 0 & 0.4 & 1.2 & 3.5 & 9 \\ 0 & -1.2 & 3.2 & 7.7 & 5.6 \end{array}\right)$$

3. Because you are adding multiples of the same row onto three other rows, the order should not matter.

5. Adding 3.2 times Row 4 onto Row 5 and then adding –2.9 times Row 4 again onto Row 5 would be equivalent to adding 0.3 times Row 4 onto Row 5, or $R_{0.3;4\to5}$.

7. The inverse operation is $R_{-5.7;1\to2}$.

6.4.7.2 Swapping two rows

A second row operation that leaves two augmented matrices equivalent is that of swapping two rows. Recall that if we swap two equations, this does not effect the solution:

$$\begin{aligned} 2x+3y+z&=-0.7 \\ x+2y+4z&=0.1 \\ -3x+4y+2z&=-2.5 \end{aligned} \quad \sim \quad \begin{aligned} -3x+4y+2z&=-2.5 \\ 2x+3y+z&=-0.7 \\ x+2y+4z&=0.1 \end{aligned}$$

For augmented matrices, we will represent this with the symbol $R_{i\leftrightarrow j}$ which indicates that we swapped Rows *i* and *j* where $i \neq j$. Thus, given our example

$$\left(\begin{array}{ccc|c} 2 & 3 & 1 & -0.7 \\ 1 & 2 & 4 & 0.1 \\ -3 & 4 & 2 & -2.5 \end{array}\right) \overset{R_{1\leftrightarrow 2}}{\sim} \left(\begin{array}{ccc|c} 1 & 2 & 4 & 0.1 \\ 2 & 3 & 1 & -0.7 \\ -3 & 4 & 2 & -2.5 \end{array}\right) \overset{R_{-2;1\to 2}}{\sim} \left(\begin{array}{ccc|c} 1 & 2 & 4 & 0.1 \\ 0 & -1 & -7 & -0.9 \\ -3 & 4 & 2 & -2.5 \end{array}\right)$$

$$\overset{R_{3;1\to 3}}{\sim} \left(\begin{array}{ccc|c} 1 & 2 & 4 & 0.1 \\ 0 & -1 & -7 & -0.9 \\ 0 & 10 & 14 & -2.2 \end{array}\right) \overset{R_{10;2\to 3}}{\sim} \left(\begin{array}{ccc|c} 1 & 2 & 4 & 0.1 \\ 0 & -1 & -7 & -0.9 \\ 0 & 0 & -56 & -11.2 \end{array}\right)$$

The inverse operation of swapping two rows is to, again, swap those two rows, and therefore

$$R_{i\leftrightarrow j}^{-1} = R_{i\leftrightarrow j}\,.$$

We will look at one further row operation—an operation that, on an augmented matrix leaves the solution unchanged—that of multiplying a row by a non-zero scalar.

Questions

1. Is swapping rows *i* and *j* and then swapping rows *j* and *k* the same as performing these operations in the reverse order?

Solutions

1. Performing these operations in the reverse order only produces the same result if *i* = *k*. Otherwise, we note that: Row *i* ends up in the kth row, Row *j* ends up in the ith row and Row *k* ends up in the jth row. If we swap rows *j* and *k* first, Row *k* is in the jth column and thus in the correct location, but Row *j* has moved to the kth row, and we must now swap rows *i* and *k*.

6.4.7.3 Multiplying a row by a non-zero scalar

If you look at Rows 2 and 3 of both matrices in the previous example, each is simply a non-zero scalar multiple of the corresponding row in the other augmented matrix. This introduces our third row operation, and its inverse: the symbol $R_{\alpha;i}$ indicates the row operation of multiplying Row i by a non-zero scalar α. To demonstrate that these two matrices are indeed row equivalent, we could simply calculate:

$$\left(\begin{array}{ccc|c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ 0 & 0 & 56 & 11.2 \end{array}\right) \overset{R_{0.5;1}}{\sim} \left(\begin{array}{ccc|c} 1 & 1.5 & 0.5 & -0.35 \\ 0 & 0.5 & 3.5 & 0.45 \\ 0 & 0 & 56 & 11.2 \end{array}\right) \overset{R_{-2.0;2}}{\sim} \left(\begin{array}{ccc|c} 1 & 1.5 & 0.5 & -0.35 \\ 0 & -1 & -7 & 0.9 \\ 0 & 0 & 56 & 11.2 \end{array}\right) \overset{R_{-1.0;3}}{\sim} \left(\begin{array}{ccc|c} 1 & 1.5 & 0.5 & -0.35 \\ 0 & -1 & -7 & 0.9 \\ 0 & 0 & -56 & -11.2 \end{array}\right)$$

Finally, we a little deduction, we can observe that if we add –0.5 times Row 2 onto Row 1, we have the right-hand side:

$$\left(\begin{array}{ccc|c} 2 & 3 & 1 & -0.7 \\ 0 & 0.5 & 3.5 & 0.45 \\ 0 & 0 & 56 & 11.2 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 1.5 & 0.5 & -0.35 \\ 0 & -1 & -7 & 0.9 \\ 0 & 0 & -56 & -11.2 \end{array}\right) \overset{R_{-0.5;2\to 1}}{\sim} \left(\begin{array}{ccc|c} 1 & 2 & 4 & 0.1 \\ 0 & -1 & -7 & -0.9 \\ 0 & 0 & -56 & -11.2 \end{array}\right).$$

Note, in general, we will never attempt to demonstrate that two matrices are row equivalent; rather, we will always find that there exists a sequence of row operations that demonstrates that a matrix is row equivalent to a standard form.

The inverse of multiplying Row i by a non-zero scalar α is, rather obviously, multiply that row by $1/\alpha$:

$$R_{\alpha;i}^{-1} = R_{\frac{1}{\alpha};i}.$$

6.4.7.4 Row equivalence properties

Remember that the properties of equality are as follows:

1. identity, as $a = a$ for all items we are inspecting,
2. symmetry, as $a = b$ if and only if $b = a$, and
3. transitivity, for if $a = b$ and $b = c$, then $a = c$.

Not just equality has this property. Suppose we say that two people are "equivalent" if they are the same age. Consequently, we note that:

1. everyone is the same age as themselves,
2. Ali is the same age as Bailey if and only if Baily is the same age as Ali, and
3. if Ali is the same age as Bailey and Bailey is the same age as Casey, then Ali is the same age as Casey.

Row equivalence has the same properties:

1. Every matrix is trivially row equivalent to itself,
2. A is row equivalent to B if and only if B is row equivalent to A, and
3. if A is row equivalent to B and B is row equivalent to C, then A is row equivalent to C.

We can write this as

1. $A \sim A$ for all matrices,
2. $A \sim B$ if and only if $B \sim A$, and
3. if $A \sim B$ and $B \sim C$, then $A \sim C$.

The proof of each of these is quite straight-forward:

Theorem
$A \sim B$ if and only $B \sim A$.

Proof:
If $A \sim B$, then there is a sequence of row operations that transforms the matrix A to B. Each of those row operations has an inverse operation, and if those inverse operations are performed in the reverse order on the matrix B, the result must be A. Therefore, $B \sim A$. ∎

Theorem
If $A \sim B$ and $B \sim C$, then $A \sim C$.

Proof:
If $A \sim B$, then there is a sequence of row operations that transforms the matrix A to B; and, similarly, if $B \sim C$, there is a sequence of row operations that transforms the matrix . ∎

Any relationship that satisfies these three properties is said to be an *equivalence relation*.

1. Which of the following is an equivalence relation?

a. The relationship that two people are *genetic siblings* if they share the same parents.
b. The relationship that two people are *siblings* if they share at least one parent.

6.4.7.5 A summary of row operations and row equivalence

There are three row operations we have been described, each with its restrictions and representations. We have also noted that each operation has an inverse operation, an associated row operation that undoes the operation performed by the described row operation. We will now only add three additional terms that, later, will describe the operation, namely shearing, reflecting and scaling.

Operation	Restriction	Representation	Inverse	Description
Add α times Row i on Row j	$i \neq j$	$R_{\alpha;i\to j}$	$R_{-\alpha;i\to j}$	shear
Swap Rows i and j	$i \neq j$	$R_{i\leftrightarrow j}$	$R_{i\leftrightarrow j}$	reflection
Multiply Row i by the scalar α	$\alpha \neq 0$	$R_{\alpha;i}$	$R_{\frac{1}{\alpha};i}$	scale

Two matrices are row equivalent if there exists a sequence of row operations that transforms one matrix into the other.

6.4.8 The Gaussian elimination algorithm with *partial pivoting*

The two row operations of adding a multiple of one row onto another and swapping two rows is sufficient to convert any matrix to row-echelon form, after which we can use backward substitution to find the solution or solutions, if any. Engineers, however, must deal the fact that we are solving these problems with computers.

Consider the following example:

$$\left(\begin{array}{cc|c} 0.00001 & 1 & 1 \\ 1 & 1 & 2 \end{array}\right).$$

With a little thought, it should be obvious that a solution to this particular problem is $\alpha_1 \approx \alpha_2 \approx 1$, where $a \approx b$ says that *a* is approximately equal to *b*. The exact solution is $\alpha_1 = 1.\overline{00001}$ and $\alpha_2 = 0.\overline{99998}$. Suppose, however, the computer stores numbers in scientific notation with no more than four digits in the mantissa. In this case, the above problem would be written as

$$\left(\begin{array}{cc|c} 1.000\times10^{-5} & 1.000\times10^{0} & 1.000\times10^{0} \\ 1.000\times10^{0} & 1.000\times10^{0} & 2.000\times10^{0} \end{array}\right)$$

Applying row operations to reduce this to row-echelon form, we would add 100 000 times Row 1 onto Row 2, to yield

$$1-1=0,\ 1-100\,000=-99\,999 \text{ and } 2-100\,000=-99\,998.$$

The problem is, we cannot store neither –99 999 nor –99 998 in scientific notation using only four digits, as these would be represented by -9.9999×10^{4} and -9.9998×10^{4}, respectively; thus we must round them to four significant digits: 1.000×10^{5} and 1.000×10^{5}. Thus, our system of linear equations becomes

$$\left(\begin{array}{cc|c} 1.000\times10^{-5} & 1.000\times10^{0} & 1.000\times10^{0} \\ 0 & 1.000\times10^{5} & 1.000\times10^{5} \end{array}\right).$$

Using backwards substitution, we now have $\alpha_2 = 1$ and therefore $\alpha_1 = 0$. While it may be argued that 1 is a reasonably good approximation of $1.\overline{00001}$, it is unlikely that anyone would argue that 0 is a good approximation of 1 under any circumstances. If we were, however, to swap Rows 1 and 2 first, we would end up with the system

$$\left(\begin{array}{cc|c} 1.000\times10^{-5} & 1.000\times10^{0} & 1.000\times10^{0} \\ 1.000\times10^{0} & 1.000\times10^{0} & 2.000\times10^{0} \end{array}\right) \overset{R_{1\leftrightarrow2}}{\sim} \left(\begin{array}{cc|c} 1.000\times10^{0} & 1.000\times10^{0} & 2.000\times10^{0} \\ 1.000\times10^{-5} & 1.000\times10^{0} & 1.000\times10^{0} \end{array}\right).$$

Now in adding –0.00001 times Row 1 onto Row 2 would yield the calculations 1 – 0.00001 = 0.99999 and 2 – 0.00001 = 1.99999, both of which rounded to four digits would remain 1 and 2, respectively. Thus, we would have

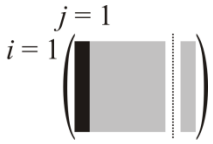
$$\left(\begin{array}{cc|c} 1.000\times10^{-5} & 1.000\times10^{0} & 1.000\times10^{0} \\ 1.000\times10^{0} & 1.000\times10^{0} & 2.000\times10^{0} \end{array}\right) \sim \left(\begin{array}{cc|c} 1.000\times10^{0} & 1.000\times10^{0} & 2.000\times10^{0} \\ 0 & 1.000\times10^{0} & 1.000\times10^{0} \end{array}\right)$$

and thus applying backward substitution, we would get the approximate solution $\alpha_1 = \alpha_2 = 1$. The problem occurred because we were adding a large scalar multiple of one row onto another. This can be solved by swapping rows. Thus, we have the more general algorithm of *Gaussian elimination with partial pivoting*.

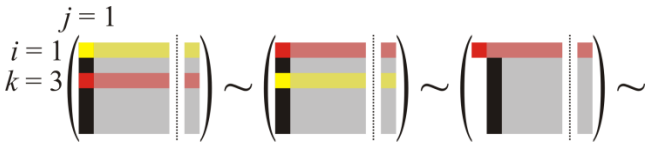
Given an $m \times n$ matrix A or $m \times (n + 1)$ augmented matrix $(A|\mathbf{v})$ representing a system of linear equations:

1. Set $i \leftarrow 1$ (that is, starting with the first row)
2. For each column $j = 1, 2, \ldots, n$:
    i. If there are no rows containing exactly $j - 1$ leading zeros, we are done; go to the next step column.
    ii. Otherwise,
        a. of all rows containing $j - 1$ leading zeros, find that Row k that has the largest leading entry in absolute value and swap that row with Row i; that is, perform the operation $R_{i \leftrightarrow k}$,
        b. next, for each Row $k \neq i$ that also has exactly $j - 1$ leading zeros, add an appropriate multiple of Row i onto Row k to zero that entry; namely, perform the operation $R_{-\frac{a_{k,j}}{a_{i,j}};i \to k}$, and
        c. finally, set $i \leftarrow i + 1$.

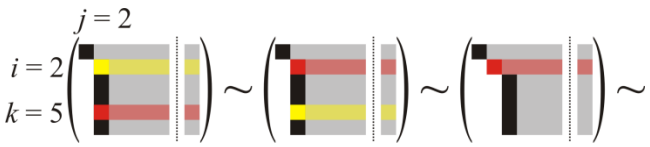
At the end of this algorithm, you will have found a matrix in row-echelon form that is equivalent to the original matrix or augmented matrix. As an example of this algorithm, suppose we a system of five linear equations in five unknowns that has a unique solution. Initially, let us assume that each row has no leading zeros:



Our algorithm begins by examining the first column ($j = 1$). With $i = 1$, we find that row that has the largest entry in absolute value (so –2.7 would be larger in absolute value than 2.5). Assume that entry is in the third row, so we swap Rows 1 and 3 and with the largest entry at location (1, 1), we add appropriate multiples of the first row to zero out all the leading non-zero entries below it. The result is a second column with mostly leading non-zero entries, and we increment both j and i so that now we dealing with the 2nd column.



For this second column, we find the largest leading non-zero entry in the second column. We are, of course ignoring anything in the first row, for if there is a value in that first column, it isn't a leading non-zero entry. Suppose we find that the largest entry in absolute value is in Row 5. In this case, we swap Rows 2 and 5. Having done so, we proceed add appropriate multiples of Row 2 onto Rows 3 through 6 to zero out those entries. This will ensure that Row 2 is the only row with a leading non-zero entry in the second column. Thus, we proceed to the next column and we increase i to three as well.



Problems

1. Apply Gaussian elimination with partial pivoting each of the following matrices:

$$\begin{pmatrix} 2.0 & -3.6 & 4.4 \\ -5.0 & -1.0 & -4.0 \\ 4.0 & -4.2 & -0.8 \end{pmatrix} \text{ and } \begin{pmatrix} 2.0 & 0.4 & 4.2 \\ 5.0 & -4.0 & 2.0 \\ -3.0 & -2.6 & -2.2 \end{pmatrix}.$$

2. Apply Gaussian elimination with partial pivoting to solve each of the following systems of linear equations:

$$\begin{aligned} 2x - 3.6y + 4.4z &= -9.6 \\ -5 - y - 4z &= -3 \\ 4x - 4.2y - 0.8z &= -3.6 \end{aligned} \quad \text{and} \quad \begin{aligned} 2x - 3.6y + 4.4z &= 12.8 \\ -5 - y - 4z &= -18 \\ 4x - 4.2y - 0.8z &= 6.4 \end{aligned}.$$

What do you notice about the operations that you're applying?

3. Apply Gaussian elimination with partial pivoting to each of the following matrices:

$$\begin{pmatrix} 1.2 & 5.8 & 0 & 8.2 \\ 6.0 & 5.0 & -1.0 & 3.0 \\ 1.2 & 7.0 & -2.2 & 3.6 \\ 0.6 & -1.3 & 3.5 & 1.4 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & -5.0 & -4.0 & 2.0 \\ 4.0 & 2.1 & 3.7 & 4.4 \\ 5.0 & -3.0 & 2.0 & 3.0 \\ 0 & 3.0 & 7.4 & 2.8 \end{pmatrix}.$$

4. Apply Gaussian elimination with partial pivoting to solve each of the following systems of linear equations:

$$\begin{aligned} 4.2w + 6.7x + 8.5y + 3.5z &= 4.3 \\ 4.8w + 6.2x + 7.9y + 9.7z &= -1.5 \\ 6w + x + 5y + 5z &= 9 \\ -5.4w + 2.7x + 4.5y - 1.5z &= -0.3 \end{aligned} \quad \text{and} \quad \begin{aligned} -1.8w + 4.8x - 5.5y + 4.4z &= 11.4 \\ -2.4w - 5.2x + 1.6y + 3.2z &= -9.8 \\ -6w - 4x - 5y - 2z &= 3 \\ 0.6w + 4x - 7.9y - 1.8z &= 14.5 \end{aligned}.$$

5. In each of these examples, the matrices were row-equivalent to an integer matrix. Do you expect that this will always be the case?

6. Apply Gaussian elimination with partial pivoting to the matrix

$$\begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}.$$

7. Apply Gaussian elimination with partial pivoting to the matrix

$$\begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix}.$$

Solutions

1. $\begin{pmatrix} -5 & -1 & -4 \\ 0 & -5 & -4 \\ 0 & 0 & 6 \end{pmatrix}$ and $\begin{pmatrix} 5 & -4 & 2 \\ 0 & -5 & -1 \\ 0 & 0 & 3 \end{pmatrix}$.

3. $\begin{pmatrix} 6 & 5 & -1 & 3 \\ 0 & 6 & -2 & 3 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & 0 & 4 \end{pmatrix}$ and $\begin{pmatrix} 5 & -3 & 2 & 3 \\ 0 & -5 & -4 & 2 \\ 0 & 0 & 5 & 4 \\ 0 & 0 & 0 & 5 \end{pmatrix}$.

5. Absolutely not! These matrices are created to ensure that they are reasonable to be done by hand.

7. $\begin{pmatrix} 4 & -1 & -1 & 0 \\ 0 & \frac{15}{4} & -\frac{1}{4} & -1 \\ 0 & 0 & \frac{56}{15} & -\frac{16}{15} \\ 0 & 0 & 0 & \frac{24}{7} \end{pmatrix}$

6.4.8.1 Exactly one solution to the homogeneous case

For example, the following system of linear equations is homogeneous and has

$$\begin{array}{l} 5x - 2y - 5z = 0 \\ 3x + 0.8y + z = 0 \\ -1x - 0.6y - 4z = 0 \end{array} \text{ or } \alpha_1 \begin{pmatrix} 5 \\ 3 \\ -1 \end{pmatrix} + \alpha_2 \begin{pmatrix} -2 \\ 0.8 \\ -0.6 \end{pmatrix} + \alpha_3 \begin{pmatrix} -5 \\ 1 \\ -4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

is row equivalent to the augmented matrix in row-echelon form

$$\left(\begin{array}{cccc} 5 & -2 & -5 & 0 \\ 3 & 0.8 & 1 & 0 \\ -1 & -0.6 & -4 & 0 \end{array}\right) \sim \left(\begin{array}{cccc} 5 & -2 & -5 & 0 \\ 0 & 2 & 4 & 0 \\ 0 & 0 & -3 & 0 \end{array}\right).$$

Looking at the last equation in the augmented matrix in row-echelon form, this says that $-3\alpha_3 = 0$ or $\alpha_3 = 0$. If we substitute this into the equation $2\alpha_2 + 4\alpha_3 = 0$, we have that

$$2\alpha_2 + 4 \cdot 0 = 0,$$

from which we must again conclude that $2\alpha_2 = 0$ or $\alpha_2 = 0$. Finally, if we substitute both of these into the first equation $5\alpha_1 - 2\alpha_2 - 5\alpha_3 = 0$, we have that

$$5\alpha_1 - 2 \cdot 0 - 5 \cdot 0 = 0$$

and so we again conclude that $5\alpha_1 = 0$ or $\alpha_1 = 0$. Thus, the solution may be written as the vector $\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

Because this solution is so easy to find, it is consequently called the *trivial solution*.

Theorem

If the rank of the augmented matrix representing a system of linear homogeneous equations equals the number of unknowns, then the homogeneous equation has exactly one solution, the trivial solution.

Proof:

A proof would be constructive using mathematical induction. If the augmented matrix has a rank equal to the number of unknowns, this means that for the row-equivalent augmented matrix in row-echelon form, $a_{i,i} \neq 0$ for each $i = 1, \ldots, n$ with $a_{i,j} = 0$ when $j < i$, then $a_{n,n} \neq 0$ and $a_{n,n}\alpha_n = 0$ imply that $\alpha_n = 0$ is the only possible value. Assume that $\alpha_n = \alpha_{n-1} = \cdots = \alpha_k = 0$. In this case, the previous equation indicates that

$$a_{k-1,k-1}\alpha_{k-1} + \cancelto{0}{a_{k-1,k}\alpha_k} + \cancelto{0}{a_{k-1,k+1}\alpha_{k+1}} + \cdots + \cancelto{0}{a_{k-1,n-1}\alpha_{n-1}} + \cancelto{0}{a_{k-1,n}\alpha_n} = 0,$$

and therefore $a_{k-1,k-1} \neq 0$ implies that $\alpha_{k-1} = 0$. Therefore, by the process of mathematical induction, $\alpha_k = 0$ for $k = n, n-1, \ldots, 2, 1$. ∎

6.4.8.2 Infinitely many solutions to the homogeneous case

Suppose that the rank of the augmented matrix is less than the number of unknowns. For example, suppose we have the homogeneous linear equation

$$2\alpha_1 - 3\alpha_2 = 0.$$

In this case, the augmented matrix $\left(\begin{array}{cc|c} 2 & -3 & 0 \end{array}\right)$ is already in row-echelon form. Any variable associated with a column that does not contain a leading non-zero value is consequently not constrained as to any value it could take. In this case, *y* is such a variable, and we describe such a variable as the *free variable* α_2. Thus, when it comes to solving for *x*, we have that $\alpha_1 = 1.5\alpha_2$. We could write the solutions as the vector $\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1.5\alpha_2 \\ \alpha_2 \end{pmatrix} = \alpha_2 \begin{pmatrix} 1.5 \\ 1 \end{pmatrix}$, and thus all scalar multiples of $\begin{pmatrix} 1.5 \\ 1 \end{pmatrix}$ are solutions to the linear equation.

Consequently, we have the following algorithm:

1. Set *i* equal to the rank of the matrix.
2. For each $j = n, n-1, \ldots, 2, 1$ perform the following:
   a. If the entry $a_{i,j}$ is a leading non-zero entry ($a_{i,j} \neq 0$ and $a_{i,k} = 0$ for $k = 1, \ldots, j-1$),
      i. Set $\alpha_i \leftarrow 0$.
      ii. For each column $k = j+1, j+2, \ldots, n$,
          set $\alpha_i \leftarrow \alpha_i - a_{i,k}\alpha_k$.
      iii. Set $\alpha_i \leftarrow \dfrac{\alpha_i}{a_{i,j}}$.
      iv. Set $i \leftarrow i - 1$.
   b. Otherwise, α_j is a *free variable*; it can take on any value it wishes, so leave it unassigned.

When the algorithm is finished, there will be rank(A_{aug}) variables assigned values and n – rank(A_{aug}) free variables.

If $r = \text{rank}\left(A_{\text{aug}}\right)$ and there are $n - r$ free variables, where $j_1, \ldots, j_{n-r}$ are the values of *j* that equaled the indices corresponding to the free variables, the solution will always be of the form

$$\begin{pmatrix} \beta_{1,1}\alpha_{j_1} + \cdots + \beta_{1,n-r}\alpha_{j_{n-r}} \\ \beta_{2,1}\alpha_{j_1} + \cdots + \beta_{2,n-r}\alpha_{j_{n-r}} \\ \vdots \\ \beta_{n,1}\alpha_{j_1} + \cdots + \beta_{n,n-r}\alpha_{j_{n-r}} \end{pmatrix} = \alpha_{j_1}\begin{pmatrix} \beta_{1,1} \\ \beta_{2,1} \\ \vdots \\ \beta_{n,1} \end{pmatrix} + \cdots + \alpha_{j_{n-r}}\begin{pmatrix} \beta_{1,n-r} \\ \beta_{2,n-r} \\ \vdots \\ \beta_{n,n-r} \end{pmatrix}.$$

More succinctly, we can always write all solutions as a linear combination of $n - r$ vectors. We will look at a number of such examples.

Example 1

Consider the system of equations

$$\begin{aligned} 5x+\ \ 3y+\ \ 2z&=0\\ 1.5x+5.9y+3.6z&=0 \end{aligned} \quad \text{or} \quad \alpha_1\begin{pmatrix}5\\1.5\end{pmatrix}+\alpha_2\begin{pmatrix}3\\5.9\end{pmatrix}+\alpha_3\begin{pmatrix}2\\3.6\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}$$

The augmented matrix and a row-equivalent augmented matrix in row-echelon form are

$$\left(\begin{array}{ccc:c} 5 & 3 & 2 & 0\\ 1.5 & 5.9 & 3.6 & 0 \end{array}\right) \sim \left(\begin{array}{ccc:c} 5 & 3 & 2 & 0\\ 0 & 5 & 3 & 0 \end{array}\right).$$

Inspecting the matrix in row-echelon form, we see that the rank is two and the number of unknown variables is three, and thus there must be one free variable. In this case, applying the algorithm, we immediately note that the third column does not contain a leading non-zero value, so $z = \alpha_3$ is a free variable. Next, the second column is associated with a leading non-zero value, so we may solve

$$5\alpha_2 + 3\alpha_3 = 0$$

or $\alpha_2 = -0.6\alpha_3$. We may now substitute this into the first equation $5\alpha_1 + 3\alpha_2 + 2\alpha_3 = 0$ to get

$$5\alpha_1 + 3(-0.6\alpha_3) + 2\alpha_3 = 0$$

and solving this for α_1, we have that $\alpha_1 = \dfrac{1.8\alpha_3 - 2\alpha_3}{5} = -0.04\alpha_3$, and thus our solution may be represented by the vector $\begin{pmatrix}-0.04\alpha_3\\-0.6\alpha_3\\\alpha_3\end{pmatrix} = \alpha_3\begin{pmatrix}-0.04\\-0.6\\1\end{pmatrix}$. You will note that $x = -0.04$, $y = -0.6$ and $z = 1$ satisfy both equations, but so do all scalar multiples of these values; for example, so does $x = 8$, $y = 120$ and $z = -200$.

Example 2

Consider the system of equations

$$\begin{array}{r} 4w- \ 2x+ \ 3y+ \ 2z=0 \\ 2.8w-1.4x+7.1y+8.4z=0 \end{array} \text{ or } \alpha_1\begin{pmatrix}4\\2.8\end{pmatrix}+\alpha_2\begin{pmatrix}-2\\-1.4\end{pmatrix}+\alpha_3\begin{pmatrix}3\\7.1\end{pmatrix}+\alpha_4\begin{pmatrix}2\\8.4\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}$$

The augmented matrix and a row-equivalent augmented matrix in row-echelon form are

$$\left(\begin{array}{cccc|c} 4 & -2 & 3 & 2 & 0 \\ 2.8 & -1.4 & 7.1 & 8.4 & 0 \end{array}\right) \sim \left(\begin{array}{cccc|c} 4 & -2 & 3 & 2 & 0 \\ 0 & 0 & 5 & 7 & 0 \end{array}\right).$$

Inspecting the augmented matrix in row-echelon form, we see that the rank is two and the number of unknown variables is four, and thus there must be two free variables. In applying the algorithm, we note that the fourth column does not contain a leading non-zero, so $z = \alpha_4$ is a free variable. Next, the third column is associated with a leading non-zero value, so we solve

$$5\alpha_3 + 7\alpha_4 = 0$$

to get that $\alpha_3 = -1.4\alpha_4$. Next, we note that the second column is not associated with a leading non-zero entry, so $x = \alpha_2$ is our second free variable. Finally, the first column is associated with a leading non-zero value, so we substitute the values we have already found into the equation $4\alpha_1 - 2\alpha_2 + 3\alpha_3 + 2\alpha_4 = 0$ to get that

$$4\alpha_1 - 2\alpha_2 + 3(-1.4\alpha_4) + 2\alpha_4 = 0$$

and solving this for α_1, we have that $\alpha_1 = \dfrac{2\alpha_2 + 4.2\alpha_4 - 2\alpha_4}{4} = 0.5\alpha_2 + 0.55\alpha_4$. Therefore, the solution can be written as the vector $\begin{pmatrix} 0.5\alpha_2 + 0.55\alpha_4 \\ \alpha_2 \\ -1.4\alpha_4 \\ \alpha_4 \end{pmatrix} = \alpha_2\begin{pmatrix}0.5\\1\\0\\0\end{pmatrix} + \alpha_4\begin{pmatrix}0.55\\0\\-1.4\\1\end{pmatrix}$. You will note that the values in both vectors satisfy both equations, but so do the values in any linear combination of these vectors; for example, $\begin{pmatrix}0\\-22\\-28\\20\end{pmatrix}$ when $\alpha_2 = -22$ and $\alpha_4 = 20$.

Example 3

Consider the system of equations

$$\begin{aligned} 5w-\ 4x-\ 3y+\ 8z&=0\\ 2w-5.6x+4.8y+4.2z&=0\\ -1.5w+3.2x-2.1y+6.1z&=0\\ -0.5w-0.4x+1.5y-3.3z&=0 \end{aligned} \quad \text{or} \quad \alpha_1\begin{pmatrix}5\\2\\-1.5\\-0.5\end{pmatrix}+\alpha_2\begin{pmatrix}-4\\-5.6\\3.2\\-0.4\end{pmatrix}+\alpha_3\begin{pmatrix}-3\\4.8\\-2.1\\1.5\end{pmatrix}+\alpha_4\begin{pmatrix}8\\4.2\\6.1\\-3.3\end{pmatrix}=\begin{pmatrix}0\\0\\0\\0\end{pmatrix}$$

The corresponding augmented matrix is row equivalent to the following augmented matrix in row-echelon form:

$$\left(\begin{array}{cccc|c} 5 & -4 & -3 & 8 & 0\\ 2 & -5.6 & 4.8 & 4.2 & 0\\ -1.5 & 3.2 & -2.1 & 6.1 & 0\\ -0.5 & -0.4 & 1.5 & -3.3 & 0 \end{array}\right) \sim \left(\begin{array}{cccc|c} 5 & -4 & -3 & 8 & 0\\ 0 & -4 & 6 & 1 & 0\\ 0 & 0 & 0 & 9 & 0\\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Thus, the rank is three, and as the number of unknown variables is four, there must be one free variable. Starting with the fourth column, it contains a leading non-zero value, so we solve $9\alpha_4 = 0$ or $\alpha_4 = 0$. Next, we note that the third column does not contain a leading non-zero value so $y = \alpha_3$ is a free variable. The second equation has a leading non-zero value, so substituting both of these into the second equation $-4\alpha_2 + 6\alpha_3 + \alpha_4 = 0$, we have that

$$-4\alpha_2 + 6\alpha_3 + 0 = 0,$$

so solving for α_2, we get that $\alpha_2 = -\dfrac{6\alpha_3}{-4} = 1.5\alpha_3$. Finally, the first column is associated with a leading-non zero value, so we substitute the known and free variables into the equation $5\alpha_1 - 4\alpha_2 - 3\alpha_3 + 8\alpha_4 = 0$ to get

$$5\alpha_1 - 4(1.5\alpha_3) - 3\alpha_3 + 8\cdot 0 = 0$$

and solving for α_1, we get that $\alpha_1 = \dfrac{4(1.5\alpha_3)+3\alpha_3}{5} = 1.8\alpha_3$. Therefore, the solution may be written as the vector $\begin{pmatrix}1.8\alpha_3\\1.5\alpha_3\\\alpha_3\\0\end{pmatrix} = \alpha_3\begin{pmatrix}1.8\\1.5\\1\\0\end{pmatrix}$. Thus, not only does the vector represent a solution to the homogenous equation, but so does every scalar multiple of that vector; for example, $w = -30.6$, $y = -25.5$, $y = -17$ and $z = 0$.

Note: not all systems of linear equations will have integer solutions. These examples were designed to ensure that you can easily verify that these are indeed solutions.

6.4.8.3 Summary of the homogenous case

Given a system of m linear equations in n unknowns, if the rank of the system is r, then

1. if $r = n$, the only solution is represented by the n-dimensional zero vector $\mathbf{0}_n$;
2. otherwise $r < n$, in which case there are $n - r$ free variables and the solution may be written as a linear combination of $n - r$ n-dimensional vectors $\alpha_{j_1}\mathbf{v}_{j_1} + \cdots + \alpha_{j_{n-r}}\mathbf{v}_{j_{n-r}}$.

Next, we will look at the non-homogeneous case; however, we will see that the solution to the homogenous case plays a significant role in the solution of the non-homogeneous case.

Problems

1. Suppose you have found that the augmented matrices corresponding to given homogenous systems of linear equations are row-equivalent to the augmented matrices

$$\left(\begin{array}{cc|c} -9 & 3 & 0 \\ 0 & 7 & 0 \end{array}\right), \left(\begin{array}{cc|c} 5 & -8 & 0 \\ 0 & 0 & 0 \end{array}\right).$$

Find all solutions and represent them as a linear combination of

2. Repeat Question 1 for the following two augmented matrices:

$$\left(\begin{array}{cc|c} 8 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right), \left(\begin{array}{cc|c} -5 & -7 & 0 \\ 0 & -3 & 0 \end{array}\right).$$

3. Repeat Question 1 for the following four augmented matrices:

$$\left(\begin{array}{ccc|c} 2 & 3 & 4 & 0 \\ 0 & 5 & 2 & 0 \\ 0 & 0 & -7 & 0 \end{array}\right), \left(\begin{array}{ccc|c} 5 & 4 & 1 & 0 \\ 0 & -2 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{ccc|c} 2 & -5 & 3 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{ccc|c} 2 & -3 & 7 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

4. Repeat Question 1 for the following three augmented matrices:

$$\left(\begin{array}{ccc|c} 4 & 1 & -7 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{ccc|c} -1 & -4 & -3 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right), \left(\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 6 & 0 \end{array}\right).$$

5. Repeat Question 1 for the following two augmented matrices:

$$\left(\begin{array}{cccc|c} 4 & 0 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 \end{array}\right), \left(\begin{array}{cccc|c} 4 & 0 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

6. Repeat Question 1 for the following two augmented matrices:

$$\left(\begin{array}{cccc:c} 2 & 0 & 0 & 2 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & -3 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 \end{array}\right), \left(\begin{array}{cccc:c} 4 & 0 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

Solutions

1. $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 1.6\alpha_2 \\ \alpha_2 \end{pmatrix}$.

3. $\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, $\alpha_3 \begin{pmatrix} -1.4 \\ 1.5 \\ 1 \end{pmatrix}$, $\alpha_2 \begin{pmatrix} 2.5 \\ 1 \\ 0 \end{pmatrix}$, $\alpha_2 \begin{pmatrix} 1.5 \\ 1 \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} -3.5 \\ 0 \\ 1 \end{pmatrix}$

6.4.8.4 The non-homogeneous (or general) case

In the non-homogeneous case, where at least one linear equation is equated to a non-zero value, if an augmented matrix is row-equivalent to a matrix that is in row-echelon form

1. are no solutions,
2. is exactly one solution, or
3. are infinitely many solutions.

If the rank of the augmented matrix is greater than the rank of the unaugmented matrix, there are no solutions.

No solutions

The most significant difference between homogeneous and non-homogeneous systems of linear equations is that the latter may have no solutions, while the solutions of all zeros (the zero vector) is always a solution to a homogeneous system of linear equations.

For example, we see that in these three cases involving systems of two linear equations (or finding linear combinations of 2-dimensional vectors), the rank of the augmented matrix is two while the rank of the unaugmented matrix is one, and therefore no solutions exist:

Linear equations	Linear combination of vectors	Row-equivalent row-echelon form
$2x = 3$ $x = 4$	$\alpha_1\begin{pmatrix}2\\1\end{pmatrix}=\begin{pmatrix}3\\4\end{pmatrix}$	$\left(\begin{array}{cc\|c}2&3\\1&4\end{array}\right)\sim\left(\begin{array}{cc\|c}2&3\\0&2.5\end{array}\right)$
$4x - 2y = 6$ $-2x + y = 5$	$\alpha_1\begin{pmatrix}4\\-2\end{pmatrix}+\alpha_2\begin{pmatrix}-2\\1\end{pmatrix}=\begin{pmatrix}6\\5\end{pmatrix}$	$\left(\begin{array}{cc\|c}4&-2&6\\-2&1&5\end{array}\right)\sim\left(\begin{array}{cc\|c}4&-2&6\\0&0&8\end{array}\right)$
$3.5x + y + 2z = 1.0$ $-4.2x - 1.2y - 1.4z = 4.8$	$\alpha_1\begin{pmatrix}3.5\\-4.2\end{pmatrix}+\alpha_2\begin{pmatrix}1\\-1.2\end{pmatrix}+\alpha_3\begin{pmatrix}2\\-1.4\end{pmatrix}=\begin{pmatrix}1\\4.8\end{pmatrix}$	$\left(\begin{array}{ccc\|c}3.5&1&2&1\\-4.2&-1.2&-1.4&4.8\end{array}\right)$ $\sim\left(\begin{array}{ccc\|c}3.5&1&2&1\\0&0&0&6\end{array}\right)$

In these next three cases involving systems of three linear equations (or finding linear combinations of 3-dimensional vectors), the rank of the unaugmented matrix is always one less than that of the augmented matrix.

Linear equations	Linear combination of vectors	Row-equivalent row-echelon form
$3x + y + 2z = 4$ $-1.8x + 1.4y + 2.8z = -1.4$ $0.6x - 0.6y - 1.2z = 2.4$	$\alpha_1\begin{pmatrix}3\\-1.8\\0.6\end{pmatrix}+\alpha_2\begin{pmatrix}1\\1.4\\-0.6\end{pmatrix}+\alpha_3\begin{pmatrix}2\\2.8\\-1.2\end{pmatrix}=\begin{pmatrix}4\\-1.4\\2.4\end{pmatrix}$	$\left(\begin{array}{ccc\|c}3&1&2&4\\-1.8&1.4&2.8&-1.4\\0.6&-0.6&-1.2&2.4\end{array}\right)\sim\left(\begin{array}{ccc\|c}3&1&2&4\\0&2&4&1\\0&0&0&2\end{array}\right)$
$4x + 2y + z = 3$ $1.6x + 0.8y + 0.4z = 6.2$ $-2x - y - 0.5z = 0.5$	$\alpha_1\begin{pmatrix}4\\1.6\\-2\end{pmatrix}+\alpha_2\begin{pmatrix}2\\0.8\\-1\end{pmatrix}+\alpha_3\begin{pmatrix}1\\0.4\\-0.5\end{pmatrix}=\begin{pmatrix}3\\6.2\\0.5\end{pmatrix}$	$\left(\begin{array}{ccc\|c}4&2&1&3\\1.6&0.8&0.4&6.2\\-2&-1&-0.5&0.5\end{array}\right)\sim\left(\begin{array}{ccc\|c}4&2&1&3\\0&0&0&5\\0&0&0&0\end{array}\right)$
$2.5x + y = 3.5$ $2x + 2.8y = 3.8$ $1.5x + 0.2y = 6.9$	$\alpha_1\begin{pmatrix}2.5\\2\\1.5\end{pmatrix}+\alpha_2\begin{pmatrix}1\\2.8\\0.2\end{pmatrix}=\begin{pmatrix}3.5\\3.8\\6.9\end{pmatrix}$	$\left(\begin{array}{cc\|c}2.5&1&3.5\\2&2.8&3.8\\1.5&0.2&6.9\end{array}\right)\sim\left(\begin{array}{cc\|c}2.5&1&3.5\\0&2&1\\0&0&5\end{array}\right)$
$4x + 2y + z = 3$ $1.6x + 0.8y + 0.4z = 6.2$ $-2x - y - 0.5z = 0.5$	$\alpha_1\begin{pmatrix}3\\-0.3\\0.9\end{pmatrix}+\alpha_2\begin{pmatrix}2\\1.8\\0.2\end{pmatrix}+\alpha_3\begin{pmatrix}-4\\0.4\\-1.2\end{pmatrix}+\alpha_4\begin{pmatrix}5\\4.5\\0.5\end{pmatrix}=\begin{pmatrix}3\\-3.3\\3.5\end{pmatrix}$	$\left(\begin{array}{cccc\|c}3&2&-4&5&3\\-0.3&1.8&0.4&4.5&-3.3\\0.9&0.2&-1.2&0.5&3.5\end{array}\right)\sim\left(\begin{array}{cccc\|c}3&2&-4&5&3\\0&2&0&5&-3\\0&0&0&0&2\end{array}\right)$

Theorem

Given any system of linear equations in n unknowns $\alpha_1,\ldots,\alpha_n$, if the rank of the unaugmented matrix is less than the rank of the augmented matrix, then no solution exists.

Proof:

A short overview of the proof is that if the rank of the augmented matrix is one greater than the rank of the unaugmented matrix, this says that there is a linear combination of the linear equations that implies that

$$0\cdot\alpha_1+0\cdot\alpha_2+\cdots+0\cdot\alpha_n\neq 0,$$

which is impossible, and therefore no solution can exist. ∎

Exactly one solution: backward substitution

If the ranks of both the unaugmented and the augmented matrices equals the number of unknowns, there is exactly one solution. Consequently, if there are more unknowns than there are equations, it is impossible for there to be a unique solution. It is possible, however, for there to be more equations than unknowns and still have a unique solution so long as additional equations are simply linear combinations of an existing set of equations.

Linear equations	**Linear combination of vectors**	**Row-equivalent row-echelon form**
$5x = 7$	$\alpha_1 \begin{pmatrix} 5 \end{pmatrix} = \begin{pmatrix} 7 \end{pmatrix}$	$\left(\begin{array}{c} 5 \end{array} \middle\vdots \begin{array}{c} 7 \end{array}\right)$
$2x = 3$ $x = 1.5$	$\alpha_1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 1.5 \end{pmatrix}$	$\left(\begin{array}{c} 2 \\ 1 \end{array} \middle\vdots \begin{array}{c} 3 \\ 1.5 \end{array}\right) \sim \left(\begin{array}{c} 2 \\ 0 \end{array} \middle\vdots \begin{array}{c} 3 \\ 0 \end{array}\right)$
$2x + y = 4$ $x - y = 2$	$\alpha_1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$	$\left(\begin{array}{cc} 2 & 1 \\ 1 & -1 \end{array} \middle\vdots \begin{array}{c} 4 \\ 2 \end{array}\right) \sim \left(\begin{array}{cc} 2 & 1 \\ 0 & 1 \end{array} \middle\vdots \begin{array}{c} 4 \\ 0 \end{array}\right)$
$3x - 2y = 4$ $-1.2x + 1.2y = 5$ $0.6x - 0.2y = -1.1$	$\alpha_1 \begin{pmatrix} 3 \\ -1.2 \\ 0.6 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ 1.2 \\ -0.2 \end{pmatrix} = \begin{pmatrix} 4 \\ 5 \\ -1.1 \end{pmatrix}$	$\left(\begin{array}{cc} 3 & 2 \\ -1.2 & 1.2 \\ 0.6 & -0.2 \end{array} \middle\vdots \begin{array}{c} 4 \\ 5 \\ -1.1 \end{array}\right) \sim \left(\begin{array}{cc} 3 & 2 \\ 0 & 2 \\ 0 & 0 \end{array} \middle\vdots \begin{array}{c} 4 \\ 7 \\ 0 \end{array}\right)$
$5x - 3y - 7z = 8$ $2x + 1.8y - 6.8z = 5.2$ $1.5x - 1.8y + 1.1z = -7.2$	$\alpha_1 \begin{pmatrix} 5 \\ 2 \\ 1.5 \end{pmatrix} + \alpha_2 \begin{pmatrix} -3 \\ 1.8 \\ -1.8 \end{pmatrix} + \alpha_3 \begin{pmatrix} -7 \\ -6.8 \\ 1.1 \end{pmatrix} = \begin{pmatrix} 8 \\ 5.2 \\ -7.2 \end{pmatrix}$	$\left(\begin{array}{ccc} 5 & -3 & -7 \\ 2 & 1.8 & -6.8 \\ 1.5 & -1.8 & 1.1 \end{array} \middle\vdots \begin{array}{c} 8 \\ 5.2 \\ -7.2 \end{array}\right) \sim \left(\begin{array}{ccc} 5 & -3 & -7 \\ 0 & 3 & -4 \\ 0 & 0 & 2 \end{array} \middle\vdots \begin{array}{c} 8 \\ 2 \\ -9 \end{array}\right)$

At this point, we may find the solution by apply backward substitution on the row-equivalent matrix in row-echelon form. Recall that if a matrix is in row-echelon form, such as

$$\left(\begin{array}{cccc} 2 & 1 & -2 & 4 \\ 0 & 5 & 2 & -3 \\ 0 & 0 & 4 & -2 \\ 0 & 0 & 0 & 5 \end{array} \middle\vdots \begin{array}{c} -1.4 \\ 1.2 \\ 2.8 \\ -2 \end{array}\right)$$

essentially says:

$$\begin{aligned} 2\alpha_1 + \alpha_2 - 2\alpha_3 + 4\alpha_4 &= -1.4 \\ 5\alpha_2 + 2\alpha_3 - 3\alpha_4 &= 1.2 \\ 4\alpha_3 - 2\alpha_4 &= 2.8 \\ 5\alpha_4 &= -2 \end{aligned}$$

To solve such a system of linear equations, we begin by solving for α_4 and, having found that, we substitute that value into the previous equation to find α_3. Now, having found both α_3 and α_4, we substitute both of those into the

previous equation to now find α_2, and so on. Following this algorithm, we find a solution for each of the unknowns. For example, in this case, the last equation

$$5\alpha_4 = -2$$

gives us that $\alpha_4 = -0.4$. Now, substituting this value into the third equation $4\alpha_3 - 2\alpha_4 = 2.8$, we now have that

$$4\alpha_3 - 2(-0.4) = 2.8,$$

and consequently, $\alpha_3 = \frac{2.8 - 0.8}{4} = 0.5$. Now, substituting both values into the second equation $5\alpha_2 + 2\alpha_3 - 3\alpha_4 = 1.2$, so we now have that

$$5\alpha_2 + 2 \cdot 0.5 - 3(-0.4) = 1.2$$

and consequently, $\alpha_2 = \frac{1.2 - 1 - 1.2}{5} = -0.2$, and finally, we substitute all three of these values into the first equation $2\alpha_1 + \alpha_2 - 2\alpha_3 + 4\alpha_4 = -1.4$, so we finally get that

$$2\alpha_1 + (-0.2) - 2 \cdot 0.5 + 4(-0.4) = -1.4$$

and thus $\alpha_1 = \frac{-1.4 + 0.2 + 1 + 1.6}{2} = 0.7$.

Note that we can now write the general solution as a vector: $\begin{pmatrix} 0.7 \\ -0.2 \\ 0.5 \\ -0.4 \end{pmatrix}$.

If A is an $m \times n + 1$ augmented matrix where $m \geq n$ and $a_{i,i} \neq 0$ for $i = 1, \ldots, n$ then the algorithm of backward substitution may be written as

1. For each $i = n, n - 1, \ldots, 2, 1$ perform the following:
    a. Set $\alpha_i \leftarrow a_{i,n+1}$, the augmented entry in the ith row.
    b. For each column $j = i + 1, i + 2, \ldots, n$, set $\alpha_i \leftarrow \alpha_i - a_{i,j}\alpha_j$.
    c. Set $\alpha_i \leftarrow \frac{\alpha_i}{a_{i,i}}$.

At the end of this algorithm, you will have computed values for α_1 through α_n. The number of floating-point operations required for this algorithm are $\frac{n(n-1)}{2}$ additions and $\frac{n(n+1)}{2}$ multiplications, which gives a total of exactly n^2 floating-point operations.

Infinitely many solutions: finding all solutions

If the rank of the unaugmented matrix equals the rank of the augmented matrix, and the number of unknowns is greater than the rank of the matrix, then there are infinitely many solutions. We will not be able to solve for the value of an α_k that corresponds to a column that does not contain a leading non-zero entry, for a total of n – rank(A) separate values. We will call these *free variables*. Let's look at two simple examples:

Suppose we are given one equation in two unknowns:

$$2\alpha_1 - 3\alpha_2 = 5\,.$$

The augmented matrix, which is already in row-echelon form, is

$$\left(\begin{array}{cc|c} 2 & -3 & 5 \end{array}\right)$$

As identified, α_2 corresponds to a column that does not have a leading non-zero value, and thus it is a free variable. Consequently, we cannot solve for it, we can simply bring it to the right-hand side when solving for α_1:

$$\alpha_1 = \frac{5+3\alpha_2}{2}\,.$$

Note that we can write this as

$$\alpha_1 = 2.5 + 1.5\alpha_2\,.$$

Note that we can now write the general solution as a vector

$$\begin{pmatrix} 2.5+1.5\alpha_2 \\ \alpha_2 \end{pmatrix}$$

where α_2 can be any number in our field (either **R** or **C**). You will recall also that we can also use vector addition and scalar multiplication to rewrite this as

$$\begin{pmatrix} 2.5+1.5\alpha_2 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 0 \end{pmatrix} + \begin{pmatrix} 1.5\alpha_2 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 1.5 \\ 1 \end{pmatrix}.$$

You will note that this is of the form $\mathbf{v}_0 + \alpha_2 \mathbf{v}_2$ where $\begin{pmatrix} 2.5 \\ 0 \end{pmatrix}$ is a solution to equation $2\alpha_1 - 3\alpha_2 = 5$ but where $\begin{pmatrix} 1.5 \\ 1 \end{pmatrix}$ is a solution to the corresponding *homogeneous* linear equation $2\alpha_1 - 3\alpha_2 = 0$ (see Section 6.4.8.2 above).

Consequently, we have the following algorithm:

1. Set i equal to the rank of the matrix.
2. For each $j = n, n-1, \ldots, 2, 1$ perform the following:
   a. If the entry $a_{i,j}$ is a leading non-zero entry ($a_{i,j} \neq 0$ and $a_{i,k} = 0$ for $k = 1, \ldots, j-1$),
      i. Set $\alpha_i \leftarrow a_{i,n+1}$, the augmented entry in the ith row.
      ii. For each column $k = j+1, j+2, \ldots, n$, set $\alpha_i \leftarrow \alpha_i - a_{i,k}\alpha_k$.
      iii. Set $\alpha_i \leftarrow \dfrac{\alpha_i}{a_{i,j}}$.
      iv. Set $i \leftarrow i-1$.
   b. Otherwise, α_j is a *free variable*; it can take on any value it wishes, so leave it unassigned.

When the algorithm is finished, there will be rank(A_{aug}) variables assigned values and n – rank(A_{aug}) free variables.

If $r = \text{rank}\left(A_{\text{aug}}\right)$ and there are $n-r$ free variables, where $j_1, \ldots, j_{n-r}$ are the values of j that equaled the indices corresponding to the free variables, the solution will always be of the form

$$\begin{pmatrix} \beta_{1,0} + \beta_{1,1}\alpha_{j_1} + \cdots + \beta_{1,n-r}\alpha_{j_{n-r}} \\ \beta_{2,0} + \beta_{2,1}\alpha_{j_1} + \cdots + \beta_{2,n-r}\alpha_{j_{n-r}} \\ \vdots \\ \beta_{n,0} + \beta_{n,1}\alpha_{j_1} + \cdots + \beta_{n,n-r}\alpha_{j_{n-r}} \end{pmatrix} = \begin{pmatrix} \beta_{1,0} \\ \beta_{2,0} \\ \vdots \\ \beta_{n,0} \end{pmatrix} + \alpha_{j_1}\begin{pmatrix} \beta_{1,1} \\ \beta_{2,1} \\ \vdots \\ \beta_{n,1} \end{pmatrix} + \cdots + \alpha_{j_{n-r}}\begin{pmatrix} \beta_{1,n-r} \\ \beta_{2,n-r} \\ \vdots \\ \beta_{n,n-r} \end{pmatrix}.$$

We will call the first vector the *particular solution* for this system of linear equations. You will notice, however, that any other vector must be a solution to the corresponding homogeneous system of linear equations. Consequently, we have the following theorem:

Theorem
Every solution to a non-homogeneous system of linear equations can be written as a particular solution to that non-homogeneous system summed with a general solution to the corresponding homogeneous system of linear equations.

More succinctly, we can always write all solutions as a particular solution plus a linear combination of solutions to the corresponding homogeneous system of linear equations. We will look at a number of such examples.

Example 1

Consider the system of equations

$$\begin{aligned} 5x+\ \ 3y+\ \ 2z&=\ \ 4\\ 1.5x+5.9y+3.6z&=-9.8 \end{aligned} \quad \text{or} \quad \alpha_1\begin{pmatrix}5\\1.5\end{pmatrix}+\alpha_2\begin{pmatrix}3\\5.9\end{pmatrix}+\alpha_3\begin{pmatrix}2\\3.6\end{pmatrix}=\begin{pmatrix}4\\-9.8\end{pmatrix}$$

The augmented matrix and a row-equivalent augmented matrix in row-echelon form are

$$\left(\begin{array}{cccc}5 & 3 & 2 & 4\\ 1.5 & 5.9 & 3.6 & -9.8\end{array}\right)\sim\left(\begin{array}{cccc}5 & 3 & 2 & 4\\ 0 & 5 & 3 & -11\end{array}\right).$$

Inspecting the matrix in row-echelon form, we see that the rank is two and the number of unknown variables is three, and thus there must be one free variable. In this case, applying the algorithm, we immediately note that the third column does not contain a leading non-zero value, so $z = \alpha_3$ is a free variable. Next, the second column is associated with a leading non-zero value, so we may solve

$$5\alpha_2 + 3\alpha_3 = -11$$

or $\alpha_2 = -2.2 - 0.6\alpha_3$. We may now substitute this into the first equation $5\alpha_1 + 3\alpha_2 + 2\alpha_3 = 4$ to get

$$5\alpha_1 + 3\left(-2.2 - 0.6\alpha_3\right) + 2\alpha_3 = 4$$

and solving this for α_1, we have that $\alpha_1 = \dfrac{10.6 - 1.8\alpha_3 - 2\alpha_3}{5} = 2.12 - 0.04\alpha_3$, and thus our solution may be represented by the vector $\begin{pmatrix}2.12 - 0.04\alpha_3\\ -2.2 - 0.6\alpha_3\\ \alpha_3\end{pmatrix} = \begin{pmatrix}2.12\\-2.2\\0\end{pmatrix} + \alpha_3\begin{pmatrix}-0.04\\-0.6\\1\end{pmatrix}$. You will note that

1. $x = 2.12$, $y = -2.2$ and $z = 0$ satisfies both equations,
2. $x = -0.04$, $y = -0.6$ and $z = 1$ satisfies both corresponding homogeneous equations, and
3. any multiple of the second solution added to the first is also a solution to the two equations above.

That is, the first vector represents a particular solution to the given system of linear equations, while the second vector is a solution to the corresponding homogeneous system of linear equations. For example, if we let $\alpha_3 = 3$, another solution is

$$\begin{pmatrix}2.12\\-2.2\\0\end{pmatrix} + 3\begin{pmatrix}-0.04\\-0.6\\1\end{pmatrix} = \begin{pmatrix}2\\-4\\3\end{pmatrix}.$$

You will notice that if you refer back to Example 1 in Section 6.4.8.2, you will see that the solution to the homogeneous is as we found it.

Example 2

Consider the system of equations

$$\begin{array}{rrrrr} 4w- & 2x+ & 3y+ & 2z= & -3 \\ 2.8w- & 1.4x+ & 7.1y+ & 8.4z= & -2.1 \end{array} \text{ or } \alpha_1\begin{pmatrix}4\\2.8\end{pmatrix}+\alpha_2\begin{pmatrix}-2\\-1.4\end{pmatrix}+\alpha_3\begin{pmatrix}3\\7.1\end{pmatrix}+\alpha_4\begin{pmatrix}2\\8.4\end{pmatrix}=\begin{pmatrix}-3\\-2.1\end{pmatrix}$$

The augmented matrix and a row-equivalent augmented matrix in row-echelon form are

$$\left(\begin{array}{cccc|c} 4 & -2 & 3 & 2 & -3 \\ 2.8 & -1.4 & 7.1 & 8.4 & -2.1 \end{array}\right) \sim \left(\begin{array}{cccc|c} 4 & -2 & 3 & 2 & -3 \\ 0 & 0 & 5 & 7 & 0 \end{array}\right).$$

Inspecting the augmented matrix in row-echelon form, we see that the rank is two and the number of unknown variables is four, and thus there must be two free variables. In applying the algorithm, we note that the fourth column does not contain a leading non-zero, so $z = \alpha_4$ is a free variable. Next, the third column is associated with a leading non-zero value, so we solve

$$5\alpha_3 + 7\alpha_4 = 0$$

to get that $\alpha_3 = -1.4\alpha_4$. Next, we note that the second column is not associated with a leading non-zero entry, so $x = \alpha_2$ is our second free variable. Finally, the first column is associated with a leading non-zero value, so we substitute the values we have already found into the equation $4\alpha_1 - 2\alpha_2 + 3\alpha_3 + 2\alpha_4 = -3$ to get that

$$4\alpha_1 - 2\alpha_2 + 3(-1.4\alpha_4) + 2\alpha_4 = -3$$

and solving this for α_1, we have that $\alpha_1 = \dfrac{-3 + 2\alpha_2 + 4.2\alpha_4 - 2\alpha_4}{4} = -0.75 + 0.5\alpha_2 + 0.55\alpha_4$. Therefore, the solution can be written as the vector $\begin{pmatrix}-0.75 + 0.5\alpha_2 + 0.55\alpha_4\\ \alpha_2 \\ -1.4\alpha_4 \\ \alpha_4\end{pmatrix} = \begin{pmatrix}-0.75\\0\\0\\0\end{pmatrix} + \alpha_2\begin{pmatrix}0.5\\1\\0\\0\end{pmatrix} + \alpha_4\begin{pmatrix}0.55\\0\\-1.4\\1\end{pmatrix}$. You will note that the first vector is a particular solution to the given system of linear equations, and the next two vectors are solutions to the corresponding homogeneous system of linear equations. Consequently, the first vector summed with any linear combination of the next two vectors will also be a solution to the system of linear equations. For example, when $\alpha_2 = 1.5$ and $\alpha_4 = 0$, this gives us $\begin{pmatrix}0\\1.5\\0\\0\end{pmatrix}$; and when $\alpha_2 = 3$ and $\alpha_4 = -5$, this gives us $\begin{pmatrix}-2\\3\\7\\-5\end{pmatrix}$; both of which are solutions to the given system of linear equations.

Example 3

Consider the system of equations

$$\begin{aligned} 5w-\ 4x-\ 3y+\ 8z&=-24\\ 2w-5.6x+4.8y+4.2z&=\ 24.4\\ -1.5w+3.2x-2.1y+6.1z&=-45.8\\ -0.5w-0.4x+1.5y-3.3z&=\ 20 \end{aligned} \quad \text{or} \quad \alpha_1\begin{pmatrix}5\\2\\-1.5\\-0.5\end{pmatrix}+\alpha_2\begin{pmatrix}-4\\-5.6\\3.2\\-0.4\end{pmatrix}+\alpha_3\begin{pmatrix}-3\\4.8\\-2.1\\1.5\end{pmatrix}+\alpha_4\begin{pmatrix}8\\4.2\\6.1\\-3.3\end{pmatrix}=\begin{pmatrix}-24\\24.4\\-45.8\\20\end{pmatrix}$$

The corresponding augmented matrix is row equivalent to the following augmented matrix in row-echelon form:

$$\left(\begin{array}{rrrr|r} 5 & -4 & -3 & 8 & -24\\ 2 & -5.6 & 4.8 & 4.2 & 24.4\\ -1.5 & 3.2 & -2.1 & 6.1 & -45.8\\ -0.5 & -0.4 & 1.5 & -3.3 & 20 \end{array}\right) \sim \left(\begin{array}{rrrr|r} 5 & -4 & -3 & 8 & -24\\ 0 & -4 & 6 & 1 & 34\\ 0 & 0 & 0 & 9 & -36\\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Thus, the rank is three, and as the number of unknown variables is four, there must be one free variable. Starting with the fourth column, it contains a leading non-zero value, so we solve $9\alpha_4 = -36$ or $\alpha_4 = -4$. Next, we note that the third column does not contain a leading non-zero value so $y = \alpha_3$ is a free variable. The second equation has a leading non-zero value, so substituting both of these into the second equation $-4\alpha_2 + 6\alpha_3 + \alpha_4 = 34$, we have that

$$-4\alpha_2 + 6\alpha_3 - 4 = 34,$$

so solving for α_2, we get that $\alpha_2 = \dfrac{38 - 6\alpha_3}{-4} = -9.5 + 1.5\alpha_3$. Finally, the first column is associated with a leading-non zero value, so we substitute the known and free variables into the equation $5\alpha_1 - 4\alpha_2 - 3\alpha_3 + 8\alpha_4 = -24$ to get

$$5\alpha_1 - 4(-9.5 + 1.5\alpha_3) - 3\alpha_3 + 8\cdot(-4) = -24$$

and solving for α_1, we get that $\alpha_1 = \dfrac{-30 + 4(1.5\alpha_3) + 3\alpha_3}{5} = -6 + 1.8\alpha_3$. Therefore, the solution may be written as the vector $\begin{pmatrix}-6+1.8\alpha_3\\-9.5+1.5\alpha_3\\\alpha_3\\-4\end{pmatrix} = \begin{pmatrix}-6\\-9.5\\0\\-4\end{pmatrix} + \alpha_3\begin{pmatrix}1.8\\1.5\\1\\0\end{pmatrix}$. The first vector is a particular solution to the given system of linear equations while the second vector is a solution to the corresponding homogenous system of linear equations. In this case, we see that every solution must have $z = -4$, while if we let $\alpha_3 = 5$, another solution is $\begin{pmatrix}3\\-2\\5\\-4\end{pmatrix}$.

Note: not all systems of linear equations will have integer solutions. These examples were designed to ensure that you can easily verify that these are indeed solutions.

6.4.8.5 Summary of the non-homogenous case

Given a system of m linear equations in n unknowns, if the rank of the system is r, then

3. if $r = n$, the only solution is represented by the n-dimensional zero vector $\mathbf{0}_n$;
4. otherwise $r < n$, in which case there are $n - r$ free variables and the solution may be written as a linear combination of $n - r$ n-dimensional vectors $\alpha_{j_1}\mathbf{v}_{j_1} + \cdots + \alpha_{j_{n-r}}\mathbf{v}_{j_{n-r}}$.

Next, we will look at the non-homogeneous case; however, we will see that the solution to the homogenous case plays a significant role in the solution of the non-homogeneous case.

6.4.9 Solutions to a homogeneous system of linear equations

There is one special case we must investigate: when we are finding a linear combination of vectors that equals the zero vector:

$$\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n = \mathbf{0},$$

or, alternatively, a system of *m* linear equations in *n* unknowns where each equation is equated to zero:

$$\begin{aligned} u_{1,1}\alpha_1 + u_{1,2}\alpha_2 + \cdots + u_{1,n}\alpha_n &= 0 \\ u_{2,1}\alpha_1 + u_{2,2}\alpha_2 + \cdots + u_{2,n}\alpha_n &= 0 \\ &\vdots \\ u_{m,1}\alpha_1 + u_{m,2}\alpha_2 + \cdots + u_{m,n}\alpha_n &= 0 \end{aligned}$$

In this case, there is always at least one solution, when all the coefficients equal zero:

$$\alpha_1 = \alpha_2 = \cdots + \alpha_n = 0.$$

This is called the *trivial solution* to a homogeneous system of linear equations. But just as in the more general case, there may also be an infinite number of solutions.

Definition

Given a system of linear equations, the corresponding homogeneous system of linear equations is that system of linear equations where each of the linear combinations of unknowns is equated to zero.

For example, the system of linear equations

$$\begin{aligned} 2x + 3y + z &= -0.7 \\ x + 2y + 4z &= 0.1 \\ -3x + 4y + 2z &= -2.5 \end{aligned}$$

the corresponding homogenous system of linear equations is

$$\begin{aligned} 2x + 3y + z &= 0 \\ x + 2y + 4z &= 0 \\ -3x + 4y + 2z &= 0 \end{aligned}$$

Given a previous problem we looked at, consider its corresponding homogeneous problem:

$$\left(\begin{array}{ccc|c} 1 & 4 & 7 & 0 \\ 2 & 5 & 8 & 0 \\ 3 & 6 & 9 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 4 & 7 & 0 \\ 0 & -3 & -6 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right).$$

Thus, the second row gives us that $-3\alpha_2 - 6\alpha_3 = 0$ so $\alpha_2 = -2\alpha_3$ and from this we may deduce that $\alpha_1 = -4(-2\alpha_3) + 7\alpha_3 = \alpha_3$. Therefore,

$$\alpha_3 \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} - 2\alpha_3 \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + \alpha_3 \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Theorem
Given a solution to a system of linear equations, and a solution to the corresponding homogeneous system of linear equations, the sum of those two solutions is also a solution to the system of linear equations.

We will start with an example before we go to the proof. One solution to the system of linear equations represented by the augmented matrix

$$\left(\begin{array}{ccc:c} 1 & 4 & 7 & 3 \\ 2 & 5 & 8 & 3 \\ 3 & 6 & 9 & 3 \end{array}\right)$$

is $\alpha_1 = -1$, $\alpha_2 = 1$ and $\alpha_3 = 0$. From above, we have that a solution to the corresponding homogeneous system of linear equations is given by $\alpha_2 = -2\alpha_3$ and $\alpha_1 = \alpha_3$ where α_3 is a free variable. Thus if we let $\alpha_3 = 1.2$, a solution to the homogeneous equation is $\alpha_1 = 1.2$, $\alpha_2 = -2.4$ and $\alpha_3 = 1.2$.

Theorem
A system of linear equations has a unique solution if and only if the corresponding homogeneous system of linear equations has only the trivial solution.

6.4.10 Applications

We will look at how systems of linear equations can be applied to actual problems, including

1. balancing chemical reactions,
2. conservation of flow with directional forces,
3. conservation of flow with alternating forces, and
4. solving differential equations.

6.4.10.1 Balancing chemical reactions

Suppose you know that a reaction either converts or leads to a state of equilibrium .

Example 1

For example, yeast takes glucose and converts it to ethyl alcohol and carbon dioxide.

$$\alpha_1 C_6H_{12}O_6 \rightarrow \alpha_2 C_2H_5OH + \alpha_3 CO_2 .$$

There are three elements, for whatever quantity of elements on one side, there must be the same amount on the other, so we have the following table:

Element	Appearance in the reaction	Homogeneous linear equation
carbon	$\alpha_1 C_6H_{12}O_6 \rightarrow \alpha_2 C_2H_5OH + \alpha_3 CO_2$	$6\alpha_1 = 2\alpha_2 + \alpha_3$
oxygen	$\alpha_1 C_6H_{12}O_6 \rightarrow \alpha_2 C_2H_5OH + \alpha_3 CO_2$	$6\alpha_1 = \alpha_2 + 2\alpha_3$
hydrogen	$\alpha_1 C_6H_{12}O_6 \rightarrow \alpha_2 C_2H_5OH + \alpha_3 CO_2$	$12\alpha_1 = (5+1)\alpha_2 + 0\cdot\alpha_3$

This gives us a homogeneous system of linear equations:

$$\begin{aligned} -6\alpha_1 + 2\alpha_2 + \alpha_3 &= 0 \\ -6\alpha_1 + \alpha_2 + 2\alpha_3 &= 0 \\ -12\alpha_1 + 6\alpha_2 &= 0 \end{aligned}$$

From this, we may determine the corresponding augmented matrix, which we can then convert to row-echelon form:

$$\left(\begin{array}{ccc|c} -6 & 2 & 1 & 0 \\ -6 & 1 & 2 & 0 \\ -12 & 6 & 0 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|c} -12 & 6 & 0 & 0 \\ -6 & 1 & 2 & 0 \\ -6 & 2 & 1 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|c} -12 & 6 & 0 & 0 \\ 0 & -2 & 2 & 0 \\ 0 & -1 & 1 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|c} -12 & 6 & 0 & 0 \\ 0 & -2 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

We note that the rank of the augmented matrix equals the rank of the unaugmented matrix, so at least one solution exists, but the rank is less than the number of unknowns by one; thus, there is one free variable. We will solve this system and interpret that free variable. Thus, $\alpha_2 = \alpha_3$ and $-12\alpha_1 + 6\alpha_3 = 0$ so $\alpha_1 = 0.56\alpha_3$. Thus, our solution vectors is

$$\begin{pmatrix} 0.5\alpha_3 \\ \alpha_3 \\ \alpha_3 \end{pmatrix} = \alpha_3 \begin{pmatrix} 0.5 \\ 1 \\ 1 \end{pmatrix} .$$

For example, if we have 0.27 mol of glucose, this would produce 0.54 mol of ethyl alcohol and 0.54 mol of carbon dioxide. It is common, however, and not entirely necessary but convenient, to choose α_3 to be the smallest possible value such that the vector is entirely of integers, in this case, $\alpha_3 = 2$ and to substitute this value into the equation:

$$C_6H_{12}O_6 \rightarrow 2C_2H_5OH + 2CO_2$$

so one glucose molecule is required to produce two molecules of ethyl alcohol with a by-product of two carbon dioxides.

Example 2

Consider the following, taken from the web site "Methods of balancing chemical equations" by Mark E. Tuckerman at New York University.[17]

$$\alpha_1 \mathrm{NaCl} + \alpha_2 \mathrm{SO_2} + \alpha_3 \mathrm{H_2O} + \alpha_4 \mathrm{O_2} \rightarrow \alpha_5 \mathrm{Na_2SO_4} + \alpha_6 \mathrm{HCl}$$

Here, table salt, sulfur dioxide, water and oxygen are combined to produce sodium sulfate and hydrochloric acid.

Element	Symbol	Appearance in the reaction	Homogeneous linear equation
sodium	Na	$\alpha_1 \mathrm{NaCl} + \cdots \rightarrow \alpha_5 \mathrm{Na_2SO_4} + \cdots$	$\alpha_1 = 2\alpha_5$
chlorine	Cl	$\alpha_1 \mathrm{NaCl} + \cdots \rightarrow \cdots + \alpha_6 \mathrm{HCl}$	$\alpha_1 = \alpha_6$
sulfur	S	$\cdots + \alpha_2 \mathrm{SO_2} + \cdots \rightarrow \alpha_5 \mathrm{Na_2SO_4} + \cdots$	$\alpha_2 = \alpha_5$
hydrogen	H	$\cdots + \alpha_3 \mathrm{H_2O} + \cdots \rightarrow \cdots + \alpha_6 \mathrm{HCl}$	$2\alpha_3 = \alpha_6$
oxygen	O	$\cdots + \alpha_2 \mathrm{SO_2} + \alpha_3 \mathrm{H_2O} + \alpha_4 \mathrm{O_2} \rightarrow \alpha_5 \mathrm{Na_2SO_4} + \cdots$	$2\alpha_2 + \alpha_3 + 2\alpha_4 = 4\alpha_5$

Now, looking at this table, you immediately see that $\alpha_1 = \alpha_6$ and $\alpha_2 = \alpha_5$, and thus $\alpha_1 = 2\alpha_2$ and $\alpha_3 = \frac{1}{2}\alpha_1$, etc., but let us see how this would appear as a system of linear equations and how solving such a system could be done algorithmically. The system of linear equations may be represented by the following augmented matrix, which is row equivalent to the given augmented matrix in row-echelon form:

$$\left(\begin{array}{cccccc|c} 1 & 0 & 0 & 0 & -2 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & -1 & 0 \\ 0 & 2 & 1 & 2 & -4 & 0 & 0 \end{array}\right) \sim \left(\begin{array}{cccccc|c} 1 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 2 & 1 & 2 & -4 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & 1 & -0.25 & 0 \\ 0 & 0 & 0 & 0 & 2 & -1 & 0 \end{array}\right)$$

Thus, we see that α_6 is a free variable, $\alpha_5 = 0.5\alpha_6$, $\alpha_4 = 0.5\alpha_6 - 0.25\alpha_6 = 0.25\alpha_6$, $\alpha_3 = 0.5\alpha_6$, $\alpha_2 = \frac{1}{2}\left(-0.5\alpha_6 - 0.5\alpha_6 + 2\alpha_6\right) = 0.5\alpha_6$ and $\alpha_1 = \alpha_6$. The free variable α_6 represents the amount of hydrochloric acid that is produced. To make all other coefficients integers, we can let $\alpha_6 = 4$ and substitute these into the equation:

$$4\mathrm{NaCl} + 2\mathrm{SO_2} + 2\mathrm{H_2O} + \mathrm{O_2} \rightarrow 2\mathrm{Na_2SO_4} + 4\mathrm{HCl},$$

however, in general letting $\alpha_6 = 4n$,

$$4n\mathrm{NaCl} + 2n\mathrm{SO_2} + 2n\mathrm{H_2O} + n\mathrm{O_2} \rightarrow 2n\mathrm{Na_2SO_4} + 4n\mathrm{HCl}$$

which says that 4*n* moles of table salt, 2*n* moles of both sulfur dioxide and water and *n* moles of oxygen will produce 2*n* moles of sodium sulfate and 4*n* moles of hydrochloric acid.

Question: Can you say that 4 g of table salt, 2 g of sulfur dioxide, 2 g of water and 1 g of oxygen produces 2 g of sodium sulfate and 4 g of hydrochloric acid?

[17] http://www.nyu.edu/classes/tuckerman/adv.chem/lectures/lecture_2/node3.html

Example 3

Consider the following, taken from the web site "What are some examples of reactions that involve catalysts" by Fred Senese.[18]

$$\alpha_1 C_2H_4 + \alpha_2 C_3H_6 + \alpha_3 H_2 \xrightarrow{Pt} \alpha_4 C_2H_6 + \alpha_5 C_3H_8$$

Here, ethane and cyclopropane, in the presence of platinum together with hydrogen forms ethane and propane.

Element	Homogeneous linear equation
hydrogen	$4\alpha_1 + 6\alpha_2 + 2\alpha_3 = 6\alpha_4 + 8\alpha_5$
carbon	$2\alpha_1 + 3\alpha_2 = 2\alpha_4 + 3\alpha_5$

This gives us two equations and five unknowns, and

$$\left(\begin{array}{ccccc|c} -4 & -6 & -2 & 6 & 8 & 0 \\ -2 & -3 & 0 & 2 & 3 & 0 \end{array}\right) \sim \left(\begin{array}{ccccc|c} -4 & -6 & -2 & 6 & 8 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 \end{array}\right).$$

Thus, α_4, α_5 are free variables and the second equation gives us that $\alpha_3 = \alpha_4 + \alpha_5$. Also, α_2 is a free variable and substituting these into the first equation, we get that

$$-4\alpha_1 - 6\alpha_2 - 2(\alpha_4 + \alpha_5) + 6\alpha_4 + 8\alpha_5 = 0,$$

or, solving for α_1, $\alpha_1 = \dfrac{6\alpha_2 + 2(\alpha_4 + \alpha_5) - 6\alpha_4 - 8\alpha_5}{-4} = -1.5\alpha_2 + \alpha_4 + 1.5\alpha_5$. What is important here, however, is properly interpreting all these free variables. There is an interesting negative sign on the left-hand side, which should be moved to the right, and if we assume that no ethane is produced, we have the following reaction:

$$(-1.5\alpha_2 + \alpha_4 + 1.5\alpha_5) C_2H_4 + \alpha_2 C_3H_6 + (\alpha_4 + \alpha_5) H_2 \xrightarrow{Pt} \alpha_4 C_2H_6 + \alpha_5 C_3H_8$$

$$1.5\alpha_5 C_2H_4 + \alpha_2 C_3H_6 + \alpha_5 H_2 \xrightarrow{Pt} 1.5\alpha_2 C_2H_4 + \alpha_5 C_3H_8$$

Now, if $\alpha_5 \neq \alpha_2$, this suggests that it is possible to produce ethane from cyclopropane and hydrogen. While such a reaction may be possible, it is not possible simply in the presence of platinum, and thus, we must conclude that $\alpha_5 = \alpha_2$, or

$$\alpha_4 C_2H_4 + \alpha_5 C_3H_6 + (\alpha_4 + \alpha_5) H_2 \xrightarrow{Pt} \alpha_4 C_2H_6 + \alpha_5 C_3H_8.$$

You will note that this is essentially two separate reactions: in the presence of platinum, hydrogen can be combined with ethane to produce ethane, and cyclopropane to produce propane.

[18] http://antoine.frostburg.edu/chem/senese/101/reactions/faq/examples-of-catalysts.shtml

Example 4

If you are balancing a reduction–oxidation (*redox*) reaction, you have one further constraint: the ions must also match. For example, given the redox reaction

$$\alpha_1 \mathrm{IO_3^-} + \alpha_2 \mathrm{I^-} + \alpha_3 \mathrm{H^+} \rightarrow \alpha_4 \mathrm{H_2O} + \alpha_5 \mathrm{I_2}\,,$$

if you simply balance the atoms, you will find that there are three linear equations but five unknowns. One solution that balances the atoms is

$$\mathrm{IO_3^-} + \mathrm{I^-} + 6\mathrm{H^+} \rightarrow \mathrm{H_2O} + \mathrm{I_2}\,,$$

however, this leaves the charges unbalanced. In this case, we must add one more linear equation:

$$-\alpha_1 - \alpha_2 + \alpha_3 = 0\,.$$

When this equation is added to the three, this leads to the correct solution

$$\mathrm{IO_3^-} + 5\mathrm{I^-} + 6\mathrm{H^+} \rightarrow 3\mathrm{H_2O} + 3\mathrm{I_2}\,.$$

As an example, consider

$$\mathrm{Mn_2^+} + \mathrm{BiO_3^-} + \mathrm{H^+} \rightarrow \mathrm{MnO_4^-} + \mathrm{Bi_3^+} + \mathrm{H_2O}\,.$$

The system of linear equations is representable by the augmented matrix

$$\left(\begin{array}{cccccc|c} -2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 3 & 0 & 0 \\ 0 & -3 & 0 & 4 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ -1 & 1 & -1 & -1 & 1 & 0 & 0 \end{array}\right) \sim \left(\begin{array}{cccccc|c} -2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -3 & 0 & 4 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1.\overline{3} & 3 & -0.\overline{3} & 0 \\ 0 & 0 & 0 & 0 & 0.625 & -0.625 & 0 \end{array}\right)$$

where the last linear equation balances the charges. This has a solution

$$\mathrm{Mn_2^+} + 3\mathrm{BiO_3^-} + \mathrm{H^+} \rightarrow 2\mathrm{MnO_4^-} + \mathrm{Bi_3^+} + \mathrm{H_2O}\,.$$

Problems

Many of these problems are taken with permission from ChemBuddy at https://www.chembuddy.com/ and should not be considered under the Creative Commons licence of this text.[19]

1. Balance the reaction $Li_2O + H_2O \rightarrow LiOH$.

2. Balance the reaction $CO_2 + NH_3 + H_2O \rightarrow NH_4HCO_3$.

3. Is it possible to create alcohol from carbon monoxide and water? $CO + H_2O \rightarrow C_2H_5OH$

4. Is it possible to create methane and alcohol from water and carbon dioxide?

$$CO_2 + H_2O \rightarrow CH_4 + C_2H_5OH$$

5. Balance the reactions:

$$FeS_2 + O_2 \rightarrow Fe_2O_3 + SO_2$$
$$HNO_3 + P_2O_5 \rightarrow N_2O_5 + HPO_3$$
$$Mn_2O_3 + Al \rightarrow Al_2O_3 + Mn$$
$$P_4O_{10} + HCl \rightarrow POCl_3 + HPO_3$$
$$Fe_2O_3 + CO \rightarrow Fe + CO_2$$
$$Ca_3(PO_4)_2 + SiO_2 \rightarrow P_4O_{10} + CaSiO_3$$
$$Hg(OH)_2 + H_3PO_4 \rightarrow Hg_3(PO_4)_2 + H_2O$$

6. Balance the following reactions:

$$Fe_2O_3 + CO \rightarrow Fe + CO_2$$
$$Ca_3(PO_4)_2 + SiO_2 \rightarrow P_4O_{10} + CaSiO_3$$
$$Hg(HO)_3 + H_3PO_4 \rightarrow Hg_3(PO_4)_2 + H_2O$$
$$FeS_2 + O_2 \rightarrow Fe_2O_3 + SO_2$$
$$Mg_3N_2 + H_2O \rightarrow Mg(OH)_2 + NH_3$$
$$Al(NO_3)_3 + Na_2CO_3 \rightarrow Al_2(CO_3)_3 + NaNO_3$$
$$HClO_4 + P_3O_{10} \rightarrow H_3PO_4 + Cl_2O_7$$

7. Balance the reaction $KNO_3 + C_{12}H_{22}O_{11} \rightarrow N_2 + CO_2 + H_2O + K_2CO_3$.

8. Balance the reaction $Al + NH_4ClO_4 \rightarrow Al_2O_3 + AlCl_3 + NO + H_2O$.

9. Balance $K_4Fe(CN)_6 + KMnO_4 + H_2SO_4 \rightarrow KHSO_4 + Fe_2(SO_4)_3 + MnSO_4 + HNO_3 + CO_2 + H_2O$.

10. Balance

$$K_3[Fe(SCN)_6] + Na_2Cr_2O_7 + H_2SO_4 \rightarrow Fe(NO_3)_3 + Cr2(SO_4)_3 + CO_2 + H_2O + Na_2SO_4 + KNO_3.$$

[19] http://www.chembuddy.com/?left=balancing-stoichiometry-questions&right=balancing-questions

Solutions

1. $\mathrm{Li_2O + H_2O \rightarrow 2LiOH}$

3. No, as the rank of the matrix $\left(\begin{array}{ccc|c} 1 & 0 & -2 & 0 \\ 0 & 2 & -6 & 0 \\ 1 & 1 & -1 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 0 & -2 & 0 \\ 0 & 2 & -6 & 0 \\ 0 & 0 & 4 & 0 \end{array}\right)$ is three, so there is only the trivial solution.

5. The balanced reactions are

$$\mathrm{4FeS_2 + 11O_2 \rightarrow 2Fe_2O_3 + 8SO_2}$$
$$\mathrm{2HNO_3 + P_2O_5 \rightarrow N_2O_5 + 2HPO_3}$$
$$\mathrm{Mn_2O_3 + 2Al \rightarrow Al_2O_3 + 2Mn}$$
$$\mathrm{P_4O_{10} + 3HCl \rightarrow POCl_3 + 3HPO_3}$$
$$\mathrm{Fe_2O_3 + 3CO \rightarrow 2Fe + 3\,CO_2}$$
$$\mathrm{2Ca_3(PO_4)_2 + 6SiO_2 \rightarrow P_4O_{10} + 6CaSiO_3}$$
$$\mathrm{3Hg(OH)_2 + 2H_3PO_4 \rightarrow Hg_3(PO_4)_2 + 6H_2O}$$

7. $\mathrm{48KNO_3 + 5C_{12}H_{22}O_{11} \rightarrow 24N_2 + 36CO_2 + 55H_2O + 24K_2CO_3}$

9. $\mathrm{10K_4Fe(CN)_6 + 122KMnO_4 + 299H_2SO_4}$
$\mathrm{\rightarrow 162KHSO_4 + 5Fe_2(SO_4)_3 + 122MnSO_4 + 60HNO_3 + 60CO_2 + 188H_2O}$

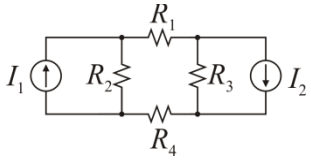
6.4.10.2 Conservation of flow with directional forces

The idea of conservation of current and forces is ubiquitous throughout engineering:

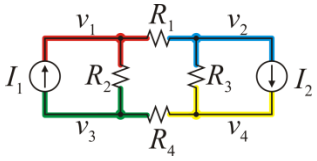
1. For civil engineering, in traffic flow analysis, the vehicles moving into an intersection must equal the vehicles moving out of an intersection.
2. For mechanical engineering, in fluid flow analysis, the flow of an incompressible liquid flowing into a junction of pipes must equal the flow out of those pipes.
3. For electrical engineering, in circuit analysis, the electrical current flowing into any connection must equal the current flowing out of that connection.

Example 1

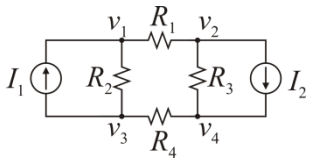
We will look at an example in electrical engineering. Consider the following circuit.



One property of an ideal conductor is that the voltage is the same at any point on that conductor. Thus, we can break this circuit into four segments of equal voltage. We don't know what the voltages are, so we will label these as v_1, v_2, v_3 and v_4.



It is customary, but not necessary, to identify one point or *node* on each segment as having the voltage in question, but it should be understood that the voltage at any point on that segment.



Now, the voltage across any resistor determines the current according to Ohm's law,

$$V = IR.$$

Thus, Kirchhoff's current law says that any node, the currents flowing into the node equal the currents flowing out of the node, or put another way, the sum of the currents flowing into a node is zero. Looking at the first node, the voltage drop across the resistor R_1 is $v_2 - v_1$, and thus the current flowing from v_2 to v_1 is $\frac{v_2 - v_1}{R_1}$. Similarly, the current flowing through the resistor R_2 is $\frac{v_3 - v_1}{R_2}$. Thus, we have the linear equation

$$I_1 + \frac{v_2 - v_1}{R_1} + \frac{v_3 - v_1}{R_2} = 0 .$$

We can repeat this at nodes 2, 3 and 4 to get

$$-I_2 + \frac{v_1 - v_2}{R_1} + \frac{v_4 - v_2}{R_3} = 0$$
$$-I_1 + \frac{v_1 - v_3}{R_2} + \frac{v_4 - v_3}{R_4} = 0$$
$$I_2 + \frac{v_2 - v_4}{R_3} + \frac{v_3 - v_4}{R_4} = 0$$

To simplify the equations, we can multiply each equation by a common denominator and write them in standard form:

$$-\left(\frac{1}{R_1} + \frac{1}{R_2}\right)v_1 + \frac{1}{R_1}v_2 + \frac{1}{R_2}v_3 = -I_1$$
$$\frac{1}{R_1}v_1 - \left(\frac{1}{R_1} + \frac{1}{R_3}\right)v_2 + \frac{1}{R_3}v_4 = I_2$$
$$\frac{1}{R_2}v_1 - \left(\frac{1}{R_2} + \frac{1}{R_4}\right)v_3 + \frac{1}{R_4}v_4 = I_1$$
$$\frac{1}{R_3}v_2 + \frac{1}{R_4}v_3 - \left(\frac{1}{R_3} + \frac{1}{R_4}\right)v_4 = -I_2$$

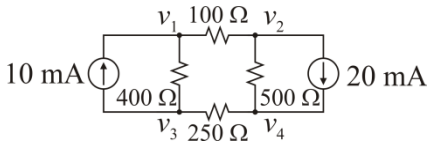
which we can represent using the augmented matrix

$$\left(\begin{array}{cccc|c} -\left(\frac{1}{R_1} + \frac{1}{R_2}\right) & \frac{1}{R_1} & \frac{1}{R_2} & 0 & -I_1 \\ \frac{1}{R_1} & -\left(\frac{1}{R_1} + \frac{1}{R_3}\right) & 0 & \frac{1}{R_3} & I_2 \\ \frac{1}{R_2} & 0 & -\left(\frac{1}{R_2} + \frac{1}{R_4}\right) & \frac{1}{R_4} & I_1 \\ 0 & \frac{1}{R_3} & \frac{1}{R_4} & -\left(\frac{1}{R_3} + \frac{1}{R_4}\right) & -I_2 \end{array}\right)$$

Let us see what happens when we substitute specific values for these resistances and current sources. Suppose that the first current source provides 10 mA while the second provides 20 mA, or

$$I_1 = 0.01 \text{ A} \text{ and } I_2 = 0.02 \text{ A} .$$

Suppose, also, that the resistors have $R_1 = 100\ \Omega$, $R_2 = 400\ \Omega$, $R_3 = 500\ \Omega$ and $R_4 = 250\ \Omega$ of resistance, respectively, as is shown here:



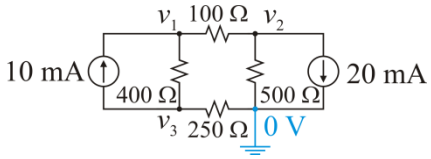
If we were to generate the matrix using these numbers, we would have

$$\left(\begin{array}{cccc|c} -0.0125 & 0.01 & 0.0025 & 0 & -0.01 \\ 0.01 & -0.012 & 0 & 0.002 & 0.02 \\ 0.0025 & 0 & -0.0065 & 0.004 & 0.01 \\ 0 & 0.002 & 0.004 & -0.006 & -0.02 \end{array}\right) \sim \left(\begin{array}{cccc|c} -0.0125 & 0.01 & 0.0025 & 0 & -0.01 \\ 0 & -0.004 & 0.002 & 0.002 & 0.012 \\ 0 & 0 & -0.005 & 0.005 & 0.014 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

This is potentially dangerous, as this suggests that there are infinitely many solutions. Does this mean that the behavior of the circuit may depend on the relative voltage of the circuit itself? For example, would this circuit behave differently if it was on an aircraft that had just been struck by lightning? Fortunately, if we solve this system of linear equations, we note that we get

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = \begin{pmatrix} -3.28 + v_4 \\ -4.4 + v_4 \\ -2.8 + v_4 \\ v_4 \end{pmatrix} = \begin{pmatrix} -3.28 \\ -4.4 \\ -2.8 \\ 0 \end{pmatrix} + v_4 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

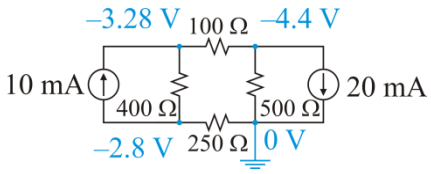
Because the currents only depend on the differences in voltages, the additional factor cancels, and thus, we may choose any voltage for the last node. Standard practice is to *ground* one of the nodes; that is, assume the voltage at one of the nodes is zero:



in which case, the system of linear equations becomes

$$\left(\begin{array}{ccc|c} -0.0125 & 0.01 & 0.0025 & -0.01 \\ 0.01 & -0.012 & 0 & 0.02 \\ 0.0025 & 0 & -0.0065 & 0.01 \end{array}\right) \sim \left(\begin{array}{ccc|c} -0.0125 & 0.01 & 0.0025 & -0.01 \\ 0 & -0.004 & 0.002 & 0.012 \\ 0 & 0 & -0.005 & 0.014 \end{array}\right)$$

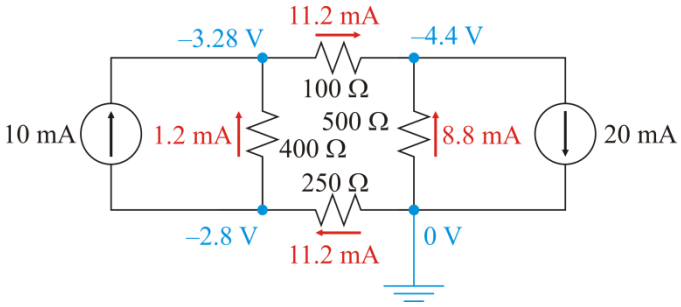
with solutions $\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} -3.28 \\ -4.4 \\ -2.8 \end{pmatrix}$.



As current flows from high voltage to low voltage, the current flowing through the

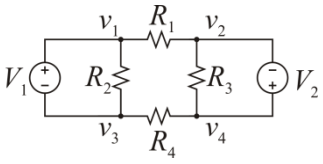
1. 100 Ω resistor is $\dfrac{-3.28\text{ V}-(-4.4\text{ V})}{100\ \Omega}=0.0112\text{ A}=11.2\text{ mA}$ from v_1 to v_2,
2. 400 Ω resistor is $\dfrac{-2.8\text{ V}-(-3.28\text{ V})}{400\ \Omega}=0.0012\text{ A}=1.2\text{ mA}$ from v_3 to v_1,
3. 500 Ω resistor is $\dfrac{0\text{ V}-(-4.4\text{ V})}{500\ \Omega}=0.0088\text{ A}=8.8\text{ mA}$ from v_4 to v_2, and
4. 250 Ω resistor is $\dfrac{0\text{ V}-(-2.8\text{ V})}{250\ \Omega}=0.0112\text{ A}=11.2\text{ mA}$ from v_4 to v_3.

We note that indeed, the currents all satisfy Kirchhoff's current law.



Example 2

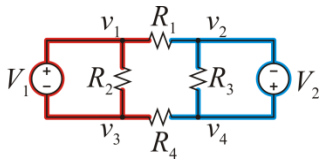
Suppose we have voltage sources instead of current sources.



Now, because we have explicit voltage sources, we know that the voltage drop between v_1 and v_3 must be V_1 V and the voltage drop between v_4 and v_3 must be V_2 V. Thus, we have two equations:

$$v_1 = v_3 + V_1$$
$$v_4 = v_2 + V_2$$

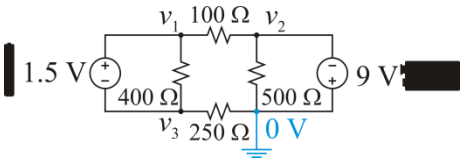
Now, we cannot determine the current that will pass through the voltage source, and thus, we consider the voltage source as a short circuit, so we have



From this, we essentially have two *supernodes*, and the currents flowing into these nodes must be zero:

$$\frac{v_4 - v_3}{R_4} + \frac{v_2 - v_1}{R_1} = 0$$
$$\frac{v_3 - v_4}{R_4} + \frac{v_1 - v_2}{R_1} = 0$$

Thus, we have four equations and four unknowns. As you no doubt have already guessed, the above two equations are identical, and therefore there will be infinitely many equations, and thus we will again ground one of the nodes. Additionally, if we use the same resistances as before, and put a AAAA 1.V batter at the left-hand side and a 9 V batter at the right-hand side, as shown here,



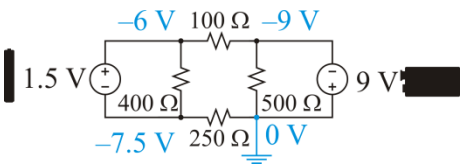
we have the following system of three equations and three unknowns:

$$v_1 - v_3 = V_1$$
$$-v_2 = V_2$$
$$-\frac{1}{R_1}v_1 + \frac{1}{R_1}v_2 - \frac{1}{R_4}v_3 = 0$$

which has the corresponding augmented matrix

$$\left(\begin{array}{ccc|c} 1 & 0 & -1 & 1.5 \\ 0 & -1 & 0 & 9 \\ -0.01 & 0.01 & -0.004 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 0 & -1 & 1.5 \\ 0 & -1 & 0 & 9 \\ 0 & 0 & -0.014 & 0.105 \end{array}\right)$$

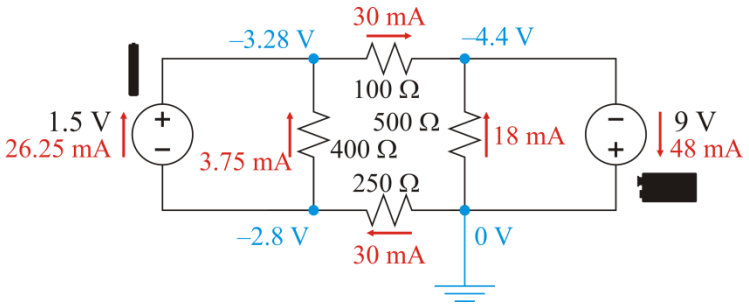
and, solving for the voltages yields that $\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} -6 \\ -9 \\ -7.5 \end{pmatrix}$



As current flows from high voltage to low voltage, the current flowing through the

1. 100 Ω resistor is $\frac{-6\text{ V}-(-9\text{ V})}{100\ \Omega}=0.03\text{ A}=30\text{ mA}$ from v_1 to v_2,
2. 400 Ω resistor is $\frac{-6\text{ V}-(-7.5\text{ V})}{400\ \Omega}=0.00375\text{ A}=3.75\text{ mA}$ from v_3 to v_1,
3. 500 Ω resistor is $\frac{0\text{ V}-(-9\text{ V})}{500\ \Omega}=0.018\text{ A}=18\text{ mA}$ from v_4 to v_2, and
4. 250 Ω resistor is $\frac{0\text{ V}-(-7.5\text{ V})}{250\ \Omega}=0.03\text{ A}=30\text{ mA}$ from v_4 to v_3.

We can also use Kirchhoff's current law to determine the current flowing through each of the voltage sources: 26.25 mA must be flowing through the 1.5 V AAAA battery and 48 mA must be flowing through the 9 V battery.



It is either a nine-volt battery or a 9 V battery, but you do not hyphenate units. Similarly, it is a 1.5 V AAAA battery or a one-and-a-half–volt AAAA battery, if you must use a hyphen. See IEEE 260.1-2004.

Summary of conservation of flow with direct current

We have analyzed direct current flow for both current and voltage sources. We could consider similar situations with fluid flow, where you have current sources and pressure sources.

What is being moved		Rate of movement		Force per object moved	
Description	**Unit**	**Description**	**Unit**	**Description**	**Unit**
charge	C	current	C/s or A	newtons per charge	N/C or V/m
volume	m^3	volume flow	m^3/s	newtons per volume	N/m^3 or Pa/m

Energy per unit moved		Resistance to movement	
Description	**Unit**	**Description**	**Unit**
joules per charge (electric potential)	J/C or V	joules per charge per current	J/C/A or V/A or Ω
joules per volume (pressure)	J/m^3 or Pa	pressure per volume flow	$Pa/(m^3/s)$

For traffic analysis, you may recall that certain highways have metered entrances. For a given on-ramp, a traffic light meters the entrance of traffic at a fixed rate per minute. This allows civil engineers to model traffic flow using constant vehicle sources, allowing more consistent, and therefore faster on-average, traffic flows.

Problems

1.

6.4.10.3 Conservation of flow with alternating forces

In the previous section, we discussed circuits with direct current. With alternating current, there are actually significant simplifications that can be made using complex arithmetic that allows us to include not only resistors, but also capacitors and inductors. For those of you who are not familiar with capacitors and inductors:

1. a capacitor is a physical circuit element that approximate the integration of current passing through it, and
2. an inductor is a physical circuit element that approximates the differentiation of current passing through it.

When alternating current is used, you may note that if $i(t) = i_0 \cos(377t)$, then the voltage across a resistor with R Ω of resistance is

$$v(t) = Ri_0 \cos(377t),$$

while the voltage across a capacitor with capacitance of C farads (F) that is initially free of any charge is

$$v(t) = \frac{1}{C}\int_0^t i_0 \cos(377t)\,dt = \frac{1}{377}\frac{i_0}{C}\sin(377t) = \frac{1}{377}\frac{i_0}{C}\cos\left(377t - \frac{\pi}{2}\right)$$

or

$$v(t) = \frac{1}{377}\frac{i_0}{C}\cos\left(377t - \frac{\pi}{2}\right),$$

and the voltage across an inductor with an inductance of L henrys (H) is

$$v(t) = L\frac{d}{dt}\left(i_0 \cos(377t)\right) = -377Li_0 \sin(377t) = 377Li_0 \cos\left(377t + \frac{\pi}{2}\right)$$

or

$$v(t) = 377Li_0 \cos\left(377t + \frac{\pi}{2}\right).$$

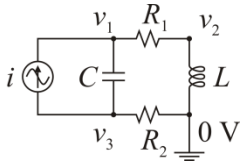
You may recall that in Section 1.6.19, we discussed how the sinusoidal current $i(t) = i_0 \cos(377t)$ may be represented by phasor $I = i_0 \angle 0$, and in general, $i(t) = i_0 \cos(377t + \phi)$ may be represented as $I = i_0 \angle \phi$.

Thus, if the current of $I = i_0 \angle 0$ passes through a resistor, capacitor or inductor, the voltage across the circuit element is $V = Ri_0 \angle 0 = RI$, $V = \frac{i_0}{377C}\angle(-90°) = -\frac{I}{377C}j = \frac{I}{377jC}$, and $V = 377Li_0\angle(90°) = 377LIj$.

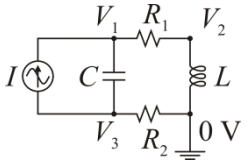
Equivalently, we may now find the currents with respect to the voltages:

	Resistor	Capacitor	Inductor
Voltage	$V = RI$	$V = \frac{1}{377jC}I$	$V = (377jL)I$
Current	$I = \frac{1}{R}V$	$I = (377jC)V$	$I = \frac{1}{377Lj}V$

Thus, for the circuit, while we previously found the voltages at the nodes, as here



the issue is now that the voltages change at each of these nodes, and thus, we will find the phasors at each of the nodes:



Thus, applying Kirchhoff's current law at each node, we have that

$$\begin{aligned} I + \frac{V_2 - V_1}{R_1} + 377jC(V_3 - V_1) &= 0 \\ \frac{V_1 - V_2}{R_1} + j\frac{0 - V_2}{377jL} &= 0 \\ -I + 377jC(V_1 - V_3) + \frac{0 - V_3}{R_2} &= 0 \end{aligned}$$

or, writing this as an augmented matrix,

$$\left(\begin{array}{ccc|c} -\left(\frac{1}{R_1} + 377Cj\right) & \frac{1}{R_1} & 377Cj & -I \\ \frac{1}{R_1} & -\left(\frac{1}{R_1} + \frac{1}{377jL}\right) & 0 & 0 \\ 377Cj & 0 & -\left(\frac{1}{R_2} + 377Cj\right) & I \end{array}\right)$$

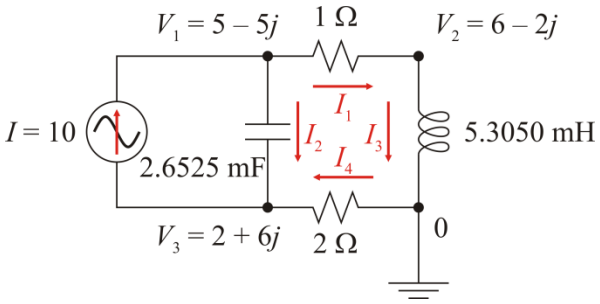
Thus, while for direct current, resistance had a multiplicative effect on the current to get the voltage, for alternating current, resistance, capacitance and inductance all have a multiplicative effect on the current to get the voltage; that is, they are *linear*. Thus, for example, if we assume that *i* = 10 cos(377*t*) A (so $I = 10\angle 0$), $R_1 = 1\,\Omega$, $R_2 = 2\,\Omega$, $C = \frac{1}{377}\text{ F} \approx 2.6525\text{ mF}$ and $L = \frac{2}{377}\text{ H} \approx 5.3050\text{ mH}$, in which case, our augmented matrix becomes

$$\left(\begin{array}{ccc|c} -1-j & 1 & j & -10 \\ 1 & -1+0.5j & 0 & 0 \\ j & 0 & -0.5-j & 10 \end{array}\right)$$

Solving this, we have that

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 5-5j \\ 6-2j \\ 2+6j \end{pmatrix}$$

Thus, for example, the voltage at Node 1 is $v_1(t) = 5\sqrt{2}\cos\left(377t - \frac{\pi}{4}\right)$. However, we would like to find the currents between the nodes. We will label the currents and specify directions:



Next, using the three properties that $I = \frac{V}{R}$, $I = 377CjV$ and $I = \frac{1}{377\,jL}V$ across resistors, capacitors and inductors, respectively, we have

$$\begin{aligned}
I_1 &= \frac{V_1 - V_2}{R_1} && = 5 - 5j - (6 - 2j) && = -1 - 3j && \approx 3.16\angle -1.89 \\
I_2 &= 377\,jC(V_1 - V_3) && = j(5 - 5j - (2 + 6j)) && = 11 - 3j && \approx 11.40\angle -0.27 \\
I_3 &= \frac{1}{377\,jL}(V_2 - 0) && = \frac{1}{2j}(6 - 2j - 0) && = -1 - 3j && \approx 3.16\angle -1.89 \\
I_4 &= \frac{0 - V_3}{R_2} && = \frac{0 - (2 + 6j)}{2} && = -1 - 3j && \approx 3.16\angle -1.89
\end{aligned}$$

Thus, as expected, the currents flowing through the right-hand loop are all identical.

$$
\begin{aligned}
i_1(t) &= \quad 7.37564\cos(377t - 0.86217) \\
i_2(t) &= \quad 7.64199\cos(377t + 0.82242) \\
i_3(t) &= 13.88524\cos(377t - 0.84655) \\
i_4(t) &= \quad 7.09366\cos(377t - 0.24360)
\end{aligned}
$$

6.4.10.4 Solving differential equations

The power of linear algebra will be applied in numerous other courses in your undergraduate studies. Consider, for example, the non-homogeneous linear ordinary differential equation (LODE) and the corresponding homogenous LODE:

$$\frac{d^2}{dt^2}y(t)+y(t)=e^{-t} \text{ and } \frac{d^2}{dt^2}y(t)+y(t)=0.$$

When we look at the homogeneous LODE, we immediately note that if $y=\hat{0}$ (that is, *y* is the zero function), then $\frac{d^2}{dt^2}\hat{0}(t)+\hat{0}(t)=0$ for all values of *t*. Notice also, that if *y* = sin or *y* = cos then these, too, satisfies the homogeneous LODE, for $\frac{d^2}{dt^2}\sin(t)+\sin(t)=-\sin(t)+\sin(t)=0$ for all values of *t*, and the same for the cosine function. You will therefore also note that any linear combination of the functions sin and cos $y_0(t)\overset{def}{=}\alpha\sin(t)+\beta\cos(t)$ is also a solution of the homogeneous LODE, for

$$\begin{aligned}\frac{d^2}{dt^2}y_0(t)+y_0(t)&=\frac{d^2}{dt^2}\left(\alpha\sin(t)+\beta\cos(t)\right)+\left(\alpha\sin(t)+\beta\cos(t)\right)\\&=\alpha\frac{d^2}{dt^2}\sin(t)+\beta\frac{d^2}{dt^2}\cos(t)+\alpha\sin(t)+\beta\cos(t)\\&=-\alpha\sin(t)-\beta\cos(t)+\alpha\sin(t)+\beta\cos(t)\\&=0\end{aligned}$$

Notice also that the function $y_p(t)\overset{def}{=}\frac{1}{2}e^{-t}$ is a particular solution to the non-homogeneous LODE, for

$$\begin{aligned}\frac{d^2}{dt^2}y_p(t)+y_p(t)&=\frac{d^2}{dt^2}\left(\frac{1}{2}e^{-t}\right)+\frac{1}{2}e^{-t}\\&=\frac{1}{2}e^{-t}+\frac{1}{2}e^{-t}\\&=e^{-t}\end{aligned}.$$

Note, however, if you add to the particular solution any solution to the homogeneous LODE, that too is a solution to the non-homogeneous LODE; for example, if you define $y(t)\overset{def}{=}\frac{1}{2}e^{-t}+3.7\sin(t)-9.1\cos(t)$, you will note that

$$\begin{aligned}\frac{d^2}{dt^2}y(t)+y(t)&=\frac{d^2}{dt^2}\left(\frac{1}{2}e^{-t}+3.7\sin(t)-9.1\cos(t)\right)+\left(\frac{1}{2}e^{-t}+3.7\sin(t)-9.1\cos(t)\right)\\&=\frac{1}{2}\frac{d^2}{dt^2}e^{-t}+3.7\frac{d^2}{dt^2}\sin(t)-9.1\frac{d^2}{dt^2}\cos(t)+\frac{1}{2}e^{-t}+3.7\sin(t)-9.1\cos(t)\\&=\frac{1}{2}e^{-t}-3.7\sin(t)+9.1\cos(t)+\frac{1}{2}e^{-t}+3.7\sin(t)-9.1\cos(t)\\&=e^{-t}\end{aligned}$$

6.5 Linear dependence and independence

In this section, we will introduce concepts of linear dependence. We will discuss this both in the context of vectors and polynomials.

6.5.1 Linear dependence and independence of vectors

If one vector is a scalar multiple of another, they both contain the same information up to a scalar multiple. If one vector is a linear combination of two or more other vectors, it does not contain any additional information that is not already available in the other vectors. Thus, given a collection of vectors, we may ask whether or not it contains redundant information; that is, are there more vectors than necessary to describe all linear combinations of those vectors.

Definition

A vector **u** is said to be *linearly dependent* on a set of *n* vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ if it can be written as a linear combination of these vectors.

Definition

If a vector **u** is not linearly dependent on a set of *m* vectors, we say that **u** is *linearly independent* of those vectors.

For example, every 3-dimensional vector linearly depends on $\left\{\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\1\\0\end{pmatrix},\begin{pmatrix}0\\0\\1\end{pmatrix}\right\}$ because any vector $\begin{pmatrix}\alpha\\\beta\\\gamma\end{pmatrix}$ can be written as the linear combination $\begin{pmatrix}\alpha\\\beta\\\gamma\end{pmatrix}=\alpha\begin{pmatrix}1\\0\\0\end{pmatrix}+\beta\begin{pmatrix}0\\1\\0\end{pmatrix}+\gamma\begin{pmatrix}0\\0\\1\end{pmatrix}$. Similarly, the vector $\begin{pmatrix}1\\1\\1\end{pmatrix}$ depends on the three vectors $\left\{\begin{pmatrix}1\\2\\3\end{pmatrix},\begin{pmatrix}3\\4\\1\end{pmatrix},\begin{pmatrix}-1\\3\\5\end{pmatrix}\right\}$ because

$$\left(\begin{array}{ccc|c}1&3&-1&1\\2&4&3&1\\3&1&5&1\end{array}\right)\sim\left(\begin{array}{ccc|c}1&3&-1&1\\0&-2&5&-1\\0&-8&8&-2\end{array}\right)\sim\left(\begin{array}{ccc|c}1&3&-1&3\\0&-2&5&-1\\0&0&-12&2\end{array}\right)$$

and if **a** represents the coefficient vector, $-12a_3=2\rightarrow a_3=-\frac{1}{6}$, $-2a_2+5a_3=-1\rightarrow a_2=\frac{1}{12}$ and $a_1+3a_2-a_3=1\rightarrow a_1=\frac{7}{12}$, and therefore

$$\begin{pmatrix}3\\4\\1\end{pmatrix}=\frac{7}{12}\begin{pmatrix}1\\2\\3\end{pmatrix}+\frac{1}{12}\begin{pmatrix}3\\4\\1\end{pmatrix}-\frac{1}{6}\begin{pmatrix}-1\\3\\5\end{pmatrix}.$$

Similarly, you will notice that the polynomial $q : t \mapsto t^2 + t + 1$ linearly depends on the three polynomials $p_1 : t \mapsto t^2 + 2t + 3$, $p_2 : t \mapsto 3t^2 + 4t + 1$ and $p_3 : t \mapsto -t^2 + 3t + 5$ because, as in our previous example, $q = \frac{7}{12} p_1 + \frac{1}{12} p_2 - \frac{1}{6} p_3$.

As a different example, you will notice that we can write $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ as a linear combination of $\left\{ \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}, \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} \right\}$ because

$$\left(\begin{array}{ccc|c} 1 & 4 & 7 & 1 \\ 2 & 5 & 8 & 1 \\ 3 & 6 & 9 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 4 & 7 & 1 \\ 0 & -3 & -6 & -1 \\ 0 & -6 & -12 & -2 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 4 & 7 & 1 \\ 0 & -3 & -6 & -1 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

and therefore, assuming $a_3 = 0$, $a_2 = \frac{1}{3}$ and thus $a_1 = -\frac{1}{3}$, so $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = -\frac{1}{3}\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \frac{1}{3}\begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}$; however, if you try to write $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ as a linear combination of these same three vectors, you will note that no combination matches the vector:

$$\left(\begin{array}{ccc|c} 1 & 4 & 7 & 1 \\ 2 & 5 & 8 & 1 \\ 3 & 6 & 9 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 4 & 7 & 1 \\ 0 & -3 & -6 & -2 \\ 0 & -6 & -12 & -3 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & 4 & 7 & 1 \\ 0 & -3 & -6 & -2 \\ 0 & 0 & 0 & 1 \end{array}\right)$$

for the last line requires that 0 = 1; a contradiction. Thus, the vector $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ is linearly independent of $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}, \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix}$.

Likewise, you will notice that the two functions $\sin^2(t)$ and $\cos^2(t)$ both depend on the two functions 1 and $\cos(2t)$, as you are aware from trigonometry,

$$\sin^2(t) = \frac{1}{2} - \frac{1}{2}\cos(2t) \text{ and } \cos^2(t) = \frac{1}{2} + \frac{1}{2}\cos(2t).$$

On the other hand, the functions $\sin(t)$ and $\cos(t)$ are linearly independent on the functions $\sin^2(t)$ and $\cos^2(t)$; that is, it is not possible to write $\sin(t) = \alpha \sin^2(t) + \beta \cos^2(t)$.

Note that the zero vector is linearly dependent on any collection of vectors, for $\mathbf{0} = 0 \cdot \mathbf{v}_1 + \cdots + 0 \cdot \mathbf{v}_n$.

Theorem
Any set of vectors containing the zero vector is linearly dependent.

Proof:

Given a set of n vectors where the kth vector is the zero vector, then

$$0\cdot\mathbf{v}_1+\cdots+0\cdot\mathbf{v}_{k-1}+1\cdot\mathbf{0}+0\cdot\mathbf{v}_{k+1}\cdots+0\cdot\mathbf{v}_n=\mathbf{0}$$

and therefore there is a non-trivial solution to the equation $\alpha_1\mathbf{v}_1+\cdots+\alpha_n\mathbf{v}_n=\mathbf{0}$. ∎

A vector can be either linearly dependent or linearly independent of a given collection of vectors. On the other hand, we can also use this terminology to describe a collection of vectors.

Definition
A collection of vectors is *linearly dependent* if any one vector can be written as a linear combination of the others.

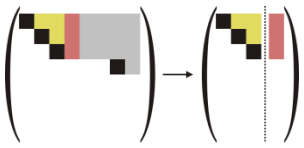
Note, it doesn't matter how many vectors depend on others; it could be as simple as one vector being a scalar multiple of another. A collection containing just the zero vector will still be considered to be a linearly dependent set.

Definition
A collection of vectors is *linearly independent* if each vector in the collection is linearly independent of the others. That is, a collection of vectors is linearly independent if the collection of vectors is not linearly dependent.

Theorem
A collection of n non-zero vectors in $\mathbf{R}^m$ is linearly dependent if the rank of the row equivalent row-echelon form is less than n.

Proof:
If a matrix in row-echelon form V is row-equivalent to the matrix of vectors ($\mathbf{v}_1$ … $\mathbf{v}_n$) and the rank is not equal to n, there must be a first column of this matrix that does not contain a row with a leading non-zero entry. Let that be the kth row. In this case, if we were to solve for the kth vector in terms of the first $k-1$ vectors, the rank of the matrix of the first $k-1$ vectors $\begin{pmatrix}\mathbf{v}_1 & \cdots & \mathbf{v}_{k-1}\end{pmatrix}$ and the rank of the augmented matrix $\left(\begin{array}{ccc|c}\mathbf{v}_1 & \cdots & \mathbf{v}_{k-1} & \mathbf{v}_k\end{array}\right)$ must both be $k-1$. Therefore, there exists a unique solution were we can write $\mathbf{v}_k$ as a linear combination of $\mathbf{v}_1,\ldots,\mathbf{v}_{k-1}$. ∎

To visualize this, note that in the following example, the first three columns in of this matrix in row-echelon form have leading non-zero entries, while the fourth does not. Thus, if we were to attempt to solve for the fourth column as a linear combination of the first three, the resulting augmented matrix would have rank 3, thus, a unique solution exists, and thus, the fourth column can be written as a linear combination of the first three.



Thus, if the rank of the matrix of n vectors ($\mathbf{v}_1$ … $\mathbf{v}_n$) has a rank equal to n, the collection of vectors $\mathbf{v}_1$, …, $\mathbf{v}_n$ is linearly independent. Another way of saying this is the following theorem:

Theorem
A collection of n vectors $\mathbf{v}_1,\ldots,\mathbf{v}_n$ is linearly independent if and only if the only linear combination of those vectors that equals the zero vector (that is $\alpha_1\mathbf{v}_1+\cdots+\alpha_n\mathbf{v}_n=\mathbf{0}$) is the trivial solution $\alpha_1=\cdots=\alpha_n=0$.

While this follows from the previous claim, we will prove this separately for interest.

In proving a theorem of the form $p \leftrightarrow q$ (that is, the statement p is true if and only if the statement q is true), it is necessary to prove one of the following four cases:

1. $p \to q$ and $q \to p$,
2. $p \to q$ and $\neg p \to \neg q$,
3. $\neg q \to \neg p$ and $\neg p \to \neg q$, or
4. $\neg q \to \neg p$ and $q \to p$.

You could not, however, show that

5. $p \to q$ and $\neg p \to \neg q$, or
6. $q \to p$ and $\neg q \to \neg p$.

This is because "If it is raining, there are clouds in the sky" is equivalent to saying "If there are no clouds in the sky, it is not raining." Both of these statements are true and state the same proposition, so if you proved the first, you also proved the second. If, however, you say "If it is raining, there are clouds in the sky" is equivalent to saying "If there are clouds in the sky, it is raining,", you only have to go outside to see that this statement is false. For example, if you wanted to prove that "$x > 0$ if and only if $x^3 > 0$", you could show any of the following:

1. If $x > 0$, it follows that $x^3 > 0$, and if $x^3 > 0$, it follows that $x > 0$.
2. If $x > 0$, it follows that $x^3 > 0$, and if $x^3 \le 0$, it follows that $x \le 0$.
3. If $x \le 0$, it follows that $x^3 \le 0$, and if $x^3 \le 0$, it follows that $x \le 0$.
4. If $x \le 0$, it follows that $x^3 \le 0$, and if $x^3 > 0$, it follows that $x > 0$.

You could not, however, show that

5. If $x > 0$, it follows that $x^3 > 0$, and if $x^3 \le 0$, it follows that $x \le 0$.
6. If $x^3 \le 0$, it follows that $x \le 0$, and if $x^3 > 0$, it follows that $x > 0$.

In the first case, this does not rule out the possibility that $x \le 0$ but $x^3 > 0$; for example, suppose you tried to prove that "$x > 0$ if and only if $x^2 > 0$". It is true that if $x > 0$, $x^2 > 0$. It is also true that if $x^2 \le 0$ the $x \le 0$. This does not prove that $x > 0$ if and only if $x^2 > 0$, for we know this statement is false: when $x = -1$, $x < 0$, but $x^2 = 1 > 0$.

Proof:
In this proof, we will show that

1. if n vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is linearly dependent, there are infinitely many linear combination of those vectors that equals the zero vector; and
2. if there are infinitely many linear combination of n vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ that equals the zero vector, the vectors are linearly dependent.

First, assume the vectors are linearly dependent. This means that there is at least one vector that can be written as a linear combination of the others:

$$\mathbf{v}_k = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_{k-1} \mathbf{v}_{k-1} + \alpha_{k+1} \mathbf{v}_{k+1} + \cdots + \alpha_n \mathbf{v}_n \ .$$

If this is true, then

$$\alpha_1 \mathbf{v}_1 + \cdots + \alpha_{k-1} \mathbf{v}_{k-1} - \mathbf{v}_k + \alpha_{k+1} \mathbf{v}_{k+1} + \cdots + \alpha_n \mathbf{v}_n = \mathbf{0} \,,$$

and thus there is more than one solution finding a linear combination of the vectors equaling the zero vector.

Next, assume that there are infinitely many linear combinations the n vectors that equals the zero vector. In this case, there must be at least one solution where

$$\alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n = \mathbf{0}$$

where not all of the coefficients are zero. Let $\mathbf{v}_k$ be that vector that has a non-zero coefficient. In this case, we have that

$$\alpha_1 \mathbf{v}_1 + \cdots + \alpha_{k-1} \mathbf{v}_{k-1} + \alpha_k \mathbf{v}_k + \alpha_{k+1} \mathbf{v}_{k+1} + \cdots + \alpha_n \mathbf{v}_n = \mathbf{0} .$$

We may subtract $\alpha_k \mathbf{v}_k$ from both sides to get

$$\alpha_1 \mathbf{v}_1 + \cdots + \alpha_{k-1} \mathbf{v}_{k-1} + \alpha_{k+1} \mathbf{v}_{k+1} + \cdots + \alpha_n \mathbf{v}_n = -\alpha_k \mathbf{v}_k$$

and then divide both sides by $-\alpha_k$, and thus we see that

$$\mathbf{v}_k = -\frac{\alpha_1}{\alpha_k} \mathbf{v}_1 - \cdots - \frac{\alpha_{k-1}}{\alpha_k} \mathbf{v}_{k-1} - \frac{\alpha_{k+1}}{\alpha_k} \mathbf{v}_{k+1} - \cdots - \frac{\alpha_n}{\alpha_k} \mathbf{v}_n .$$

Thus, the set of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is linearly dependent. ∎

This generalizes to the following statement theorem, which we will use in subsequent chapters:

Theorem

Given a collection of *n* vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ that are linearly dependent, there is a first vector $\mathbf{v}_k$ with $2 \le k \le n$ such that $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}\}$ forms a linearly independent set and $\mathbf{v}_k$ depends on those first $k - 1$ vectors.

Proof:

The set $\{\mathbf{v}_1\}$ forms an linearly independent set. Suppose it was always true that given a set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}\}$, that the addition of a vector $\mathbf{v}_k$ always formed a linearly independent set $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$. In this case, by induction, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ would form a linearly independent set, which is, by our assumption, a contradiction. Therefore, there must have been a first *k* such that $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}\}$ formed a linearly independent set but where $\mathbf{v}_k$ depends on those first $k - 1$ vectors. ∎

Demonstration of this theorem

Given four vectors in $\mathbf{R}^3$, it follows that because 4 > 3 that the vectors must be linearly dependent.

Consider the set of vectors $\left\{ \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}, \begin{pmatrix} -2 \\ -4 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix} \right\}$. In this case, the second vector is a scalar multiple of the first, and therefore the second vector depends on the first.

Consider the set of vectors $\left\{\begin{pmatrix}1\\2\\-2\end{pmatrix},\begin{pmatrix}2\\3\\-1\end{pmatrix},\begin{pmatrix}1\\0\\4\end{pmatrix},\begin{pmatrix}2\\4\\6\end{pmatrix}\right\}$. The first two vectors are not linearly dependent, but the third vector is linearly dependent on the first two, as $\begin{pmatrix}1\\0\\4\end{pmatrix}=-3\begin{pmatrix}1\\2\\-2\end{pmatrix}+2\begin{pmatrix}2\\3\\-1\end{pmatrix}$. Thus, the third vector is the first to linearly depend on those before it.

Consider the set of vectors $\left\{\begin{pmatrix}1\\2\\-2\end{pmatrix},\begin{pmatrix}2\\3\\-1\end{pmatrix},\begin{pmatrix}1\\2\\-3\end{pmatrix},\begin{pmatrix}-3\\-3\\-4\end{pmatrix}\right\}$. The first three vectors are linearly independent, and thus, the fourth vector must linearly depend on the first three, namely, $\begin{pmatrix}-3\\-3\\-4\end{pmatrix}=2\begin{pmatrix}1\\2\\-2\end{pmatrix}-3\begin{pmatrix}2\\3\\-1\end{pmatrix}+\begin{pmatrix}1\\2\\-3\end{pmatrix}$.

For example, the collection of vectors $\left\{\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\1\\0\end{pmatrix},\begin{pmatrix}0\\0\\1\end{pmatrix}\right\}$ is linearly independent because

$$\alpha_1\begin{pmatrix}1\\0\\0\end{pmatrix}+\alpha_2\begin{pmatrix}0\\1\\0\end{pmatrix}+\alpha_3\begin{pmatrix}0\\0\\1\end{pmatrix}=\begin{pmatrix}\alpha_1\\\alpha_2\\\alpha_3\end{pmatrix},$$

and $\begin{pmatrix}\alpha_1\\\alpha_2\\\alpha_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}$ if and only if $\alpha_1=\alpha_2=\alpha_3=0$.

Alternatively, if we consider $\left\{\begin{pmatrix}1\\2\end{pmatrix},\begin{pmatrix}3\\4\end{pmatrix}\right\}$, are these linearly independent? If we try to solve

$$\alpha_1\begin{pmatrix}1\\2\end{pmatrix}+\alpha_2\begin{pmatrix}3\\4\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix},$$

we get a system of two equations in two unknowns:

$$\begin{aligned}\alpha_1+3\alpha_2&=0\\2\alpha_1+4\alpha_2&=0\end{aligned}$$

Previously, we saw how to solve such a system of equations, and therefore we get

$$\left(\begin{array}{cc|c} 1 & 3 & 0 \\ 2 & 4 & 0 \end{array}\right) \to \left(\begin{array}{cc|c} 2 & 4 & 0 \\ 1 & 3 & 0 \end{array}\right) \to \left(\begin{array}{cc|c} 2 & 4 & 0 \\ 0 & 1 & 0 \end{array}\right),$$

and therefore $\alpha_2 = 0$ and therefore $2\alpha_1 + 4\alpha_2 = 0 \to 2\alpha_1 = 0 \to \alpha_1 = 0$ and so the only solution is $\alpha_1 = \alpha_2 = 0$.

Now, consider the collection of three vectors $\left\{\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}, \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix}\right\}$. With a little trial-and-error, we may deduce that we may write this as

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} - 2\begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

and thus, the three vectors are linearly dependent, as we can write, for example, $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = 2\begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} - \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix}$.

We could, however, also try to solve the system of linear equations:

$$\left(\begin{array}{ccc|c} 1 & 4 & 7 & 0 \\ 2 & 5 & 8 & 0 \\ 3 & 6 & 9 & 0 \end{array}\right) \to \left(\begin{array}{ccc|c} 3 & 6 & 9 & 0 \\ 2 & 5 & 8 & 0 \\ 1 & 2 & 3 & 0 \end{array}\right) \to \left(\begin{array}{ccc|c} 3 & 6 & 9 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

The last non-zero row says that $\alpha_2 + 2\alpha_3 = 0$, and thus we may designate α_3 to be a free variable and so $\alpha_2 = -2\alpha_3$. We can now let $\alpha_3 = 1$, and therefore $\alpha_2 = -2$, and substituting this into the first equation, we get $3\alpha_1 + 6\cdot(-2) + 9\cdot 1 = 0 \to 3\alpha_1 = 3 \to \alpha_1 = 1$, and so

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} - 2\begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} + \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

or, if we let $\alpha_3 = -2$, we would also get that

$$-2\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + 4\begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} - 2\begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

6.5.2 Linear dependence and independence of polynomials

Now, we have already shown that the set of all polynomials is a vector space, and we can multiply a polynomial by a scalar, and add two polynomials together. Now, however, we can continue to ask if a set of polynomials is independent.

For example, the polynomials $1, t, t^2, t^3$ are linearly independent, for it is impossible to write, for example, the constant polynomial 1 as a linear combination $\alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3$, for if any one coefficient was non-zero, the result would be a non-constant polynomial, but if all the coefficients are zero, the linear combination is the zero polynomial.

There are some circumstances where you may not want to represent a polynomial of degree less-than-or-equal-to three in the form $\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3$. You may, for example, want to write it as a linear combination of

$$\alpha_0 + \alpha_1 (t - 0.5) + \alpha_2 (t - 0.5)^2 + \alpha_3 (t - 0.5)^3 .$$

The questions we may ask, however, are $1, t - 0.5, (t - 0.5)^2, (t - 0.5)^3$ linearly independent? Thus, we are asking, is the only possible solution to

$$\alpha_0 + \alpha_1 (t - 0.5) + \alpha_2 (t - 0.5)^2 + \alpha_3 (t - 0.5)^3 = 0$$

when all the confidents are zero? If we expand this out, we get:

$$\alpha_0 + \alpha_1 (t - 0.5) + \alpha_2 \left(t^2 - t + 0.25\right) + \alpha_3 \left(t^3 - 1.5t^2 + 0.75t - 0.125\right) = 0$$

or

$$\alpha_0 + \alpha_1 t - \alpha_1 0.5 + \alpha_2 t^2 - \alpha_2 t + \alpha_2 0.25 + \alpha_3 t^3 - \alpha_3 1.5t^2 + \alpha_3 0.75t - \alpha_3 0.125 = 0 .$$

Let group terms with the same power of *t*:

$$(\alpha_0 - \alpha_1 0.5 + \alpha_2 0.25 - \alpha_3 0.125) + t(\alpha_1 - \alpha_2 + \alpha_3 0.75) + t^2 (\alpha_2 - \alpha_3 1.5) + t^3 (\alpha_3) = 0 .$$

For this to be equal to zero, each of the coefficients must equal zero, so we now have four equations and four unknowns:

$$\begin{aligned} \alpha_0 - \alpha_1 0.5 + \alpha_2 0.25 - \alpha_3 0.125 &= 0 \\ \alpha_1 - \alpha_2 + \alpha_3 0.75 &= 0 \\ \alpha_2 - \alpha_3 1.5 &= 0 \\ \alpha_3 &= 0 \end{aligned}$$

The corresponding augmented matrix is

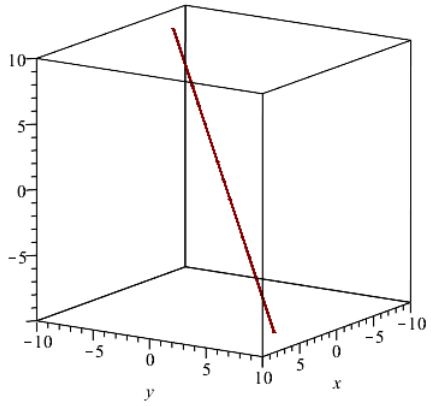
$$\left(\begin{array}{cccc:c} 1 & -0.5 & 0.25 & -0.125 & 0 \\ 0 & 1 & -1 & 0.75 & 0 \\ 0 & 0 & 1 & -1.5 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right)$$

Immediately, we see that this is in row-echelon form, and the rank is four, so there is a unique solution to the homogeneous system of linear equations. Thus, the four polynomials $1, t - 0.5, (t - 0.5)^2, (t - 0.5)^3$ are linearly independent.

6.6 Spans and subspaces

Consider the vector $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix}$ and now consider the set of all scalar multiples of **u**: $U = \{\alpha\mathbf{u} : \alpha \in \mathbf{R}\}$ (read this as "*U* is the set of all vectors of the form $\alpha\mathbf{u}$ such that α is a real number.") Notice that if we consider any two vectors **v** and **w** that are in this set, then $\mathbf{v} = \alpha\mathbf{u}$ and $\mathbf{w} = \beta\mathbf{u}$ for some real values of α and β. Thus, we note that $\mathbf{v} + \mathbf{w} = \alpha\mathbf{u} + \beta\mathbf{u} = (\alpha + \beta)\mathbf{u}$ is also in *U*, and $\gamma\mathbf{v} = \gamma(\alpha\mathbf{u}) = (\gamma\alpha)\mathbf{u}$ are also in *U*, and thus the set *U* is itself closed under the operations of vector addition and scalar multiplication.

Thus, *U* is just as much a vector space as $\mathbf{R}^3$ in which it lies, but not every vector in $\mathbf{R}^3$ is in *U*, and so we will call *U* a *subspace* of $\mathbf{R}^3$.



In general, a *subspace* $\mathcal{U}$ of a vector space $\mathcal{V}$ is any subset of $\mathcal{V}$ that is closed under vector addition and scalar multiplication; that is, if $\mathbf{u}, \mathbf{v} \in U$ then $\mathbf{u} + \mathbf{v} \in U$ and $\alpha\mathbf{u} \in U$ for any scalar value α.

The most trivial subspace of any vector space is $\{\mathbf{0}\}$, as $\mathbf{0} + \mathbf{0} = \mathbf{0}$ and $\alpha\mathbf{0} = \mathbf{0}$ for all values of α.

Theorem
If *U* is a subspace of *V*, then $\mathbf{0} \in U$.

Proof:
As $\alpha\mathbf{u} \in U$, for all values of α, in particular, $0\mathbf{u} = \mathbf{0} \in U$. ∎

Note that the empty set is not a vector space—a vector space must have an additive identity element **0**, so {**0**} is the smallest possible vector space (and therefore smallest possible subspace).

Demonstration of this theorem

Consider the set of all vectors of the form $\begin{pmatrix} \alpha \\ \beta \\ \alpha+\beta-1 \end{pmatrix}$ where $\alpha, \beta \in \mathbf{F}$. In this case, we note that there are no values of α and β that allow $\begin{pmatrix} \alpha \\ \beta \\ \alpha+\beta-1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$. Consequently, the set of all such vectors cannot form a subspace of $\mathbf{R}^3$.

On the other hand, if you consider all vectors of the form $\begin{pmatrix} 3\alpha+\beta-\gamma \\ 5\alpha+2\beta-3\gamma \\ 2\alpha-\beta+2\gamma \end{pmatrix}$ where $\alpha, \beta, \gamma \in \mathbf{F}$, we see that letting $\alpha = \beta = \gamma = 0$ gives us the zero vector, and therefore, we must investigate further as to whether or not all vectors of this form a subspace. (It is.)

If you consider the collection of all vectors of the form $\begin{pmatrix} \sin(\alpha) \\ \cos(\beta) \end{pmatrix}$, we see that when $\alpha = 0$ and $\beta = \frac{\pi}{2}$ that $\begin{pmatrix} \sin(\alpha) \\ \cos(\beta) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. Thus, we cannot immediately reject that this forms a subspace. However, while $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is in this collection, $2 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ is not, so it turns out that it is never-the-less not a subspace.

If you consider all polynomials such that $p(1) = 0$, we see that the zero polynomial satisfies this requirement, and therefore $\hat{0}(1) = 0$. However, if you consider all polynomials such that $p(1) = 1$, we see that the zero polynomial does not satisfy this requirement, and therefore the collection of all such polynomials cannot define a subspace.

6.6.1 The span of a set of vectors

Given a collection of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$, the *span* of those vectors is denoted as $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and is defined as all possible linear combinations of the vectors $\mathbf{v}_1$ through $\mathbf{v}_n$; that is,

$$\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \{\alpha_1\mathbf{v}_1 + \cdots + \alpha_n\mathbf{v}_n : \alpha_1, ..., \alpha_n \in \mathbf{F}\}.$$

Note that the span is itself a vector space, for if $\mathbf{u}_1$ and $\mathbf{u}_2$ are vectors in $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, then

$$\mathbf{u}_1 = \alpha_{1,1}\mathbf{v}_1 + \cdots + \alpha_{1,n}\mathbf{v}_n$$

and

$$\mathbf{u}_2 = \alpha_{2,1}\mathbf{v}_1 + \cdots + \alpha_{2,n}\mathbf{v}_n$$

for appropriate values of $\alpha_{1,1}$ through $\alpha_{1,n}$ and $\alpha_{2,1}$ through $\alpha_{2,n}$, so

$$\begin{aligned}\mathbf{u}_1+\mathbf{u}_2 &= \left(\alpha_{1,1}\mathbf{v}_1+\cdots+\alpha_{1,n}\mathbf{v}_n\right)+\left(\alpha_{2,1}\mathbf{v}_1+\cdots+\alpha_{2,n}\mathbf{v}_n\right)\\ &= \left(\alpha_{1,1}\mathbf{v}_1+\alpha_{2,1}\mathbf{v}_1\right)+\cdots+\left(\alpha_{1,n}\mathbf{v}_n+\alpha_{2,n}\mathbf{v}_n\right)\\ &= \left(\alpha_{1,1}+\alpha_{2,1}\right)\mathbf{v}_1+\cdots+\left(\alpha_{1,n}+\alpha_{2,n}\right)\mathbf{v}_n\end{aligned}$$

and so $\mathbf{u}_1+\mathbf{u}_2$ is also in $\text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$, and it should be straightforward to see that if $\mathbf{u}$ is in the span, then so is $\beta\mathbf{u}$.

Given a collection of vectors $\mathbf{v}_1,\ldots,\mathbf{v}_n$ in a vector space, if every vector in the corresponding vector space is in $\text{span}\{\mathbf{v}_1,\ldots,\mathbf{v}_n\}$, we will say that the collection $\mathbf{v}_1,\ldots,\mathbf{v}_n$ *spans* the entire vector space.

Definition
If a collection of vectors in a vector space *V* is empty, then the span of that set is defined to be the zero vector of *V*; that is, $\text{span}\{\ \} = \{\mathbf{0}_V\}$.

Theorem
Two spans $\text{span}\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_m\}$ and $\text{span}\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n\}$ are equal if and only if $\mathbf{v}_k \in \text{span}\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_m\}$ for each $k=1,\ldots,n$ and $\mathbf{u}_j \in \text{span}\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n\}$ for each $j=1,\ldots,m$.

Proof:
If $\text{span}\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_m\} = \text{span}\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n\}$, then the second component follows by definition.

If $\mathbf{v}_k \in \text{span}\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_m\}$ for each $k=1,\ldots,n$ and $\mathbf{u}_j \in \text{span}\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n\}$ for each $j=1,\ldots,m$, then suppose that $\mathbf{w} \in \text{span}\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_m\}$. Therefore $\mathbf{w} = \alpha_1\mathbf{u}_1+\alpha_2\mathbf{u}_2+\cdots+\alpha_m\mathbf{u}_m$; however, by definition, for each $\mathbf{u}_j = \beta_{j,1}\mathbf{v}_1+\beta_{j,2}\mathbf{v}_2+\cdots+\beta_{j,n}\mathbf{v}_n$, and therefore

$$\begin{aligned}\mathbf{w} &= \alpha_1\mathbf{u}_1+\cdots+\alpha_m\mathbf{u}_m\\ &= \alpha_1\left(\beta_{1,1}\mathbf{v}_1+\cdots+\beta_{1,n}\mathbf{v}_n\right)+\cdots+\alpha_m\left(\beta_{m,1}\mathbf{v}_1+\cdots+\beta_{m,n}\mathbf{v}_n\right)\\ &= \left(\alpha_1\beta_{1,1}+\cdots+\alpha_m\beta_{m,1}\right)\mathbf{v}_1+\cdots+\left(\alpha_1\beta_{1,n}+\cdots+\alpha_m\beta_{m,n}\right)\mathbf{v}_n\end{aligned}$$

and therefore $\mathbf{w} \in \text{span}\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n\}$. We could similarly show that each vector in $\text{span}\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_n\}$. ∎

What this says is that in order to determine if two spans are equal, we need not consider any other vectors other than those defining the span.

6.6.2 The relationship between spans and dependence

Note that saying $\mathbf{v} \in \text{span}\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_m\}$ is equivalent to saying that **v** depends on $\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_m\}$.

6.7 Basis and dimension

In the previous section, we talked about the span of a set of vectors. Clearly,

$$\text{span}\left\{\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\1\\0\end{pmatrix},\begin{pmatrix}0\\0\\1\end{pmatrix}\right\}=\mathbf{R}^3,$$

but also

$$\text{span}\left\{\begin{pmatrix}1\\0\\0\end{pmatrix},\begin{pmatrix}0\\1\\0\end{pmatrix},\begin{pmatrix}0\\0\\1\end{pmatrix},\begin{pmatrix}2\\3\\-5\end{pmatrix}\right\}=\mathbf{R}^3.$$

On the other hand, you may be able to convince yourself that

$$\text{span}\left\{\begin{pmatrix}1\\1\\0\end{pmatrix},\begin{pmatrix}0\\1\\1\end{pmatrix},\begin{pmatrix}1\\0\\-1\end{pmatrix}\right\}\neq\mathbf{R}^3,$$

as it is never possible to write

$$a\begin{pmatrix}1\\1\\0\end{pmatrix}+b\begin{pmatrix}0\\1\\1\end{pmatrix}+c\begin{pmatrix}1\\0\\-1\end{pmatrix}=\begin{pmatrix}1\\1\\1\end{pmatrix},$$

for when we try to solve the inverse problem $\begin{pmatrix}1&0&1\\1&1&0\\0&1&-1\end{pmatrix}\begin{pmatrix}a\\b\\c\end{pmatrix}=\begin{pmatrix}1\\1\\1\end{pmatrix}$, we perform Gaussian elimination on the augmented matrix to find that there is no solution:

$$\left(\begin{array}{ccc|c}1&0&1&1\\1&1&0&1\\0&1&-1&1\end{array}\right)\rightarrow\left(\begin{array}{ccc|c}1&0&1&1\\0&1&-1&0\\0&1&-1&1\end{array}\right)\rightarrow\left(\begin{array}{ccc|c}1&0&1&1\\0&1&-1&0\\0&0&0&1\end{array}\right).$$

One goal in engineering and mathematics is to minimize the amount of information we require to describe a vector space, so given a set of vectors, can we find a minimal set of vectors such that

$$\text{span}\left\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_m\right\}=\text{span}\left\{\mathbf{u}_{k_1},\mathbf{u}_{k_2},\ldots,\mathbf{u}_{k_n}\right\}?$$

Because every vector in the vector space can be written

If $\left\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_n\right\}$ is a finite collection of linearly independent vectors, then $\left\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_n\right\}$ forms a basis for $\text{span}\left\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_m\right\}$ and the dimension of this subspace is *n*.

Every basis has the same number of vectors. This number is the *dimension* of the vector space.

6.7.1 Orthogonal and orthonormal bases

Given a basis $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$, then finding a linear combination of those basis vectors that equals a given vector $\mathbf{v}$ requires us to solve

$$\mathbf{v} = \alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \cdots + \alpha_n \mathbf{u}_n .$$

This is a slow process requiring $O(n^3)$ operations. In some cases, a poor choice of basis can result in unexpectedly large coefficients. For example, consider the basis

$$\left\{ \begin{pmatrix} 2 \\ 0.1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 0.1 \end{pmatrix}, \begin{pmatrix} 2 \\ 0.1 \\ 0.1 \end{pmatrix} \right\}.$$

Thus, to find the linear combination of these basis vectors that equals the three unit vectors, we must solve

$$\left(\begin{array}{ccc|c} 2 & 2 & 2 & 1 \\ 0.1 & 0 & 0.1 & 0 \\ 0 & 0.1 & 0.1 & 0 \end{array}\right), \left(\begin{array}{ccc|c} 2 & 2 & 2 & 0 \\ 0.1 & 0 & 0.1 & 1 \\ 0 & 0.1 & 0.1 & 0 \end{array}\right) \text{ and } \left(\begin{array}{ccc|c} 2 & 2 & 2 & 0 \\ 0.1 & 0 & 0.1 & 0 \\ 0 & 0.1 & 0.1 & 1 \end{array}\right)$$

to get that

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = 0.5 \begin{pmatrix} 2 \\ 0.1 \\ 0 \end{pmatrix} + 0.5 \begin{pmatrix} 2 \\ 0 \\ 0.1 \end{pmatrix} - 0.5 \begin{pmatrix} 2 \\ 0.1 \\ 0.1 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = -10 \begin{pmatrix} 2 \\ 0 \\ 0.1 \end{pmatrix} + 10 \begin{pmatrix} 2 \\ 0.1 \\ 0.1 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = -10 \begin{pmatrix} 2 \\ 0.1 \\ 0 \end{pmatrix} + 10 \begin{pmatrix} 2 \\ 0.1 \\ 0.1 \end{pmatrix}$$

If, however, the basis vectors are orthogonal, it is no longer necessary to solve a system of linear equations. Instead, we may

$$\begin{aligned} \langle \mathbf{v}, \mathbf{u}_k \rangle &= \langle \alpha_1 \mathbf{u}_1 + \cdots + \alpha_k \mathbf{u}_k + \cdots + \alpha_n \mathbf{u}_n, \mathbf{u}_k \rangle \\ &= \langle \alpha_1 \mathbf{u}_1, \mathbf{u}_k \rangle + \cdots + \langle \alpha_k \mathbf{u}_k, \mathbf{u}_k \rangle + \cdots + \langle \alpha_n \mathbf{u}_n, \mathbf{u}_k \rangle \\ &= \alpha_1 \cancelto{0}{\langle \mathbf{u}_1, \mathbf{u}_k \rangle} + \cdots + \alpha_k \langle \mathbf{u}_k, \mathbf{u}_k \rangle + \cdots + \alpha_n \cancelto{0}{\langle \mathbf{u}_n, \mathbf{u}_k \rangle} \\ &= \alpha_k \langle \mathbf{u}_k, \mathbf{u}_k \rangle \end{aligned}$$

and therefore, we see that $\alpha_k = \dfrac{\langle \mathbf{v}, \mathbf{u}_k \rangle}{\langle \mathbf{u}_k, \mathbf{u}_k \rangle}$, in other words,

$$\mathbf{v} = \text{proj}_{\mathbf{u}_1}(\mathbf{v}) + \text{proj}_{\mathbf{u}_2}(\mathbf{v}) + \cdots + \text{proj}_{\mathbf{u}_n}(\mathbf{v}).$$

This reduces the complexity to O(n^2) down from O(n^3)—a significant savings.

If the basis vectors are also normalized (an *orthonormal basis*), it is no longer necessary to divide through by the norm of the basis vectors during the computation of the projection, so this simplifies further to

$$\mathbf{v} = \langle \mathbf{v}, \hat{\mathbf{u}}_1 \rangle \hat{\mathbf{u}}_1 + \langle \mathbf{v}, \hat{\mathbf{u}}_2 \rangle \hat{\mathbf{u}}_2 + \cdots + \langle \mathbf{v}, \hat{\mathbf{u}}_n \rangle \hat{\mathbf{u}}_n .$$

Suppose we have an orthonormal basis, so $\langle \hat{\mathbf{u}}_i, \hat{\mathbf{u}}_j \rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$. Suppose we define the matrix where each row is one of these vectors:

$$\mathbf{U} = \begin{pmatrix} \hat{\mathbf{u}}_1^T \\ \hat{\mathbf{u}}_2^T \\ \vdots \\ \hat{\mathbf{u}}_n^T \end{pmatrix}.$$

Note that the multiplication $\mathbf{Uv}$ finds the coefficients.

```
>> U = rand( 3, 3 )
    U =
        0.1067    0.7749    0.0844
        0.9619    0.8173    0.3998
        0.0046    0.8687    0.2599

>> % Perform the modified Gram Schmidt process on the rows of U
>> for i=1:3
       for j = 1:(i - 1)
           U(i,:) = U(i,:) - (U(j,:)*U(i,:)')*U(j,:);
       end

       % Normalize the ith row
       U(i,:) = U(i,:)/norm(U(i,:));
   end
>> U
    U =
        0.1356    0.9849    0.1073
        0.9295   -0.1639    0.3304
       -0.3430   -0.0550    0.9377

>> ui = U*[1 0 0]'
    ans =
        0.1356
        0.9295
       -0.3430

>> uj = U*[0 1 0]'
    ans =
        0.9849
       -0.1639
       -0.0550

>> uk = U*[0 0 1]'
    ans =
        0.1073
        0.3304
        0.9377
```

6.7.2 Bases for discrete signals

If we define the *delta impulse signal* as

$$\delta : n \mapsto \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases}$$

then the collection of *shifted* delta impulse signals functions $\delta[\bullet]$, $\delta[\bullet - 1]$, $\delta[\bullet - 2]$, $\delta[\bullet - 3]$, … form a basis, as the signal *x* can be written as the linear combination of the basis functions

$$x = \sum_{k=0}^{\infty} x[k]\delta[\bullet - k] \ .$$

Note that $\delta[n-k] = \begin{cases} 1 & n = k \\ 0 & n \neq k \end{cases}$. The number of bases signals in this set is countable, meaning, for each integer, there is a basis signal. This basis has the additional properties that it is both orthogonal and normalized.

Next, consider the set of exponential signals where $E_z : n \mapsto z^n$. There is an exponential signal for each complex number. We may ask if this forms a basis, but very quickly we see it is not so: there are uncountably many complex numbers, and therefore there are more basis signals in this collection than there are basis signals in the set of shifted delta impulse signals.

6.7.3 Bases for polynomials

If we consider the set of all polynomials, you may believe that the following is a basis for those polynomials:

$$1, x, x^2, x^3, \cdots ,$$

However, you will see in calculus that this is not true, as there are linear combinations where each coefficient is non-zero and yet the resulting sum is not a polynomial. In your calculus course, you will see that

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \frac{z^5}{5!} + \cdots \ .$$

Thus, instead, if we define $\mathbf{P}_n$ to be the vector space of polynomials of degree less than or equal to *n*, then a basis for that vector space is the monomials

$$1, x, x^2, x^3, \cdots , x^n$$

and the dimension of this vector space is *n* + 1. We cannot discuss an orthogonal basis, because we do not have an inner product, in general, for polynomials. If, however, we define the inner product on the interval [–1, 1] with

$$\langle p, q \rangle = \int_{-1}^{1} p(x) q(x) dx \ ,$$

then we can take the basis vectors $1, x, x^2, x^3, \cdots$ and create an orthogonal basis.

6.7.3.1 Hermite polynomials (optional)

Up to now, we have defined the inner product as a simple sum or integral; however, it is also possible to include a weighting vector, weighting signal or weighting function which has the property that it is always positive. One of the most important weighting functions, used in physics where they relate to the quantum harmonic oscillator and finite element methods where they are used to shape beams, is the function $w: x \mapsto e^{-x^2}$ and thus we define our inner product as

$$\langle p, q \rangle = \int_{-\infty}^{\infty} p(x) q(x) e^{-x^2} dx .$$

If you consider all of the properties of an inner product, you will note that this modified inner product also satisfies all of them. Without proof or derivation, these polynomials are

$$\begin{aligned} H_0 &: x \mapsto 1 \\ H_1 &: x \mapsto 2x \\ H_2 &: x \mapsto 4x^2 - 2 \\ H_3 &: x \mapsto 8x^3 - 12x \\ H_4 &: x \mapsto 16x^4 - 48x^2 + 12 \end{aligned}$$

In your Calculus course, you will see how to do this, but in Maple, you can see that this is true:

```
> int( (4*x^2 - 2)*(8*x^3 - 12*x)*exp(-x^2), x = -infinity..infinity );
```

$$0$$

In your Calculus course, you will see how to do this, but in Maple, you

6.8 Vectors as coefficients of a basis

To this point, we have considered vectors as a representation of a point in space; for example, $(2,3)$ is the point going out 2 in the *x* direction and three up in the *y* direction. Similarly, $(1,-2,4)$ defines a point or vector in 3-space going one unit in the *x*-direction, backwards 2 units in the *y* direction and 4 units up. However, we should really think of the entries of a vector *with respect to a given basis*. The standard basis is, of course, the usual basis, so in two dimensions, the basis is $\{\mathbf{e}_1, \mathbf{e}_2\}$ where $\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and, in general, the basis for an *n*-dimensional vector space is $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ where $\mathbf{e}_k = \begin{pmatrix} e_{1,k} \\ \vdots \\ e_{n,k} \end{pmatrix}$ is such that $e_{i,k} = \begin{cases} 1 & i = k \\ 0 & i \neq k \end{cases}$. The easiest observation, however, may be that suppose that a vector **u** represents the offset of a point in space relative to an origin where the basis is

$$B_{\mathrm{m}} = \left\{ \begin{pmatrix} 1\mathrm{m} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1\mathrm{m} \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1\mathrm{m} \end{pmatrix} \right\},$$

and so the vector $\mathbf{u} = \begin{pmatrix} 0.32 \\ 1.54 \\ -0.76 \end{pmatrix}$ represents a point 32 cm in the *x*-direction, 154 cm in the *y*-direction and 76 cm down.

In this case, what are the coefficients of the point with respect to the basis

$$B_{\text{mm}} = \left\{ \begin{pmatrix} 1\text{mm} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1\text{mm} \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1\text{mm} \end{pmatrix} \right\} \text{ or } B_{\text{ft}} = \left\{ \begin{pmatrix} 1\text{ft} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1\text{ft} \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1\text{ft} \end{pmatrix} \right\} ?$$

What are the coordinates with respect to the first basis, or the second? We will represent the original vector as $\mathbf{u}_{\text{m}}$ and coordinates with respect to the latter two as $\mathbf{u}_{\text{cm}}$ and $\mathbf{u}_{\text{ft}}$. In this case, it should be obvious that the transformation matrix $A_{\text{mm}\leftarrow\text{m}} = \begin{pmatrix} 1000 & 0 & 0 \\ 0 & 1000 & 0 \\ 0 & 0 & 1000 \end{pmatrix}$ and $A_{\text{ft}\leftarrow\text{m}} = \begin{pmatrix} \frac{1250}{381} & 0 & 0 \\ 0 & \frac{1250}{381} & 0 \\ 0 & 0 & \frac{1250}{381} \end{pmatrix}$ (as one foot is defined as 0.3048 m). Therefore,

$$\begin{aligned} \mathbf{u}_{\text{mm}} &= A_{\text{mm}\leftarrow\text{m}}\mathbf{u}_{\text{m}} \\ &= \begin{pmatrix} 1000 & 0 & 0 \\ 0 & 1000 & 0 \\ 0 & 0 & 1000 \end{pmatrix} \begin{pmatrix} 0.32 \\ 1.54 \\ -0.76 \end{pmatrix} \\ &= \begin{pmatrix} 320 \\ 1540 \\ -760 \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned} \mathbf{u}_{\text{ft}} &= A_{\text{ft}\leftarrow\text{m}}\mathbf{u}_{\text{m}} \\ &= \begin{pmatrix} \frac{1250}{381} & 0 & 0 \\ 0 & \frac{1250}{381} & 0 \\ 0 & 0 & \frac{1250}{381} \end{pmatrix} \begin{pmatrix} 0.32 \\ 1.54 \\ -0.76 \end{pmatrix} \\ &\approx \begin{pmatrix} 1.0499 \\ 5.0525 \\ -2.4934 \end{pmatrix} \end{aligned}$$

In all three cases, the bases

In Calculus, you will learn of *non-linear coordinate* systems. For example, suppose you are controlling a drone: You don't care that an object is

1. 500 m in front of you (your direction of travel),
2. 30 m to your left (parallel to the Earth's surface and perpendicular to the direction of travel), and
3. 10 m up in the air (perpendicular to the Earth's surface).

Indeed, it may be very difficult to determine this information. Instead, a much more *natural* basis may be:

1. how far away is the object,
2. how many degrees to the left (positive) or the right (negative) is it, and
3. how many degrees up (positive) or down (negative) is it.

In this case, each object has a unique *spherical coordinate* in terms of one distance and two angles, (r,θ,ϕ).

In this case, however, the relationship between the standard basis and the spherical basis is *non-linear*:

$$r=\sqrt{x^2+y^2+z^2} \qquad \theta=\cos^{-1}\left(\frac{z}{r}\right) \qquad \phi=\tan^{-1}\left(\frac{y}{x}\right)$$

and

$$x=r\sin(\theta)\cos(\phi) \qquad y=r\sin(\theta)\sin(\phi) \qquad z=r\cos(\theta)$$

Such non-linear bases will be the subject of your calculus course.

At this point, it should be obvious that the transformation matrix

Consider for example the Fourier transform. The complex exponential functions of the form

$$e^{-j2\pi\omega t}$$

form an orthogonal basis, of sorts, as

$$\begin{aligned}\int_{-\infty}^{\infty}\left(e^{j2\pi\omega_1 t}\right)^* e^{j2\pi\omega_2 t}dt &= \int_{-\infty}^{\infty} e^{-j2\pi\omega_1 t}e^{j2\pi\omega_2 t}dt \\ &= \int_{-\infty}^{\infty} e^{j2\pi(\omega_2-\omega_1)t}dt \\ &= \begin{cases}\infty & \omega_1=\omega_2 \\ 0 & \omega_1\neq\omega_2 \text{ on average}\end{cases}\end{aligned}$$

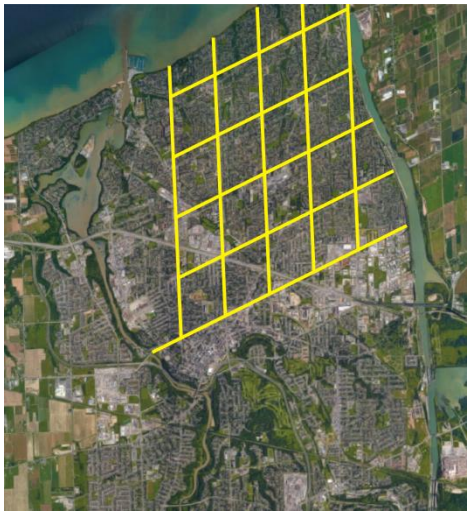
.

6.9 Change of bases

To this point in the course, we have always considered a vector as information related to the canonical basis. For example, the vector

$$\mathbf{v} = \begin{pmatrix} 33.1 \\ 29.2 \\ -13.2 \end{pmatrix}$$

could represent a point 33.1 m North, 29.2 m West and –13.2 m into the ground from a given location. Similarly, in New York, you may give directions such as "Go three blocks East and two blocks South." In St. Catharines, however, the major arteries do not cross at right angles. In this case, it would be more natural to give directions with respect to the coordinates of the angular block.



Thus, you may give directions like "go one block NORTH and two blocks ENE." In reality, this would take you 900 m north plus an addition 400 m NORTH and 1600 m EAST; however, if you were to give direction such as "go 1.3 km NORTH and 1.6 km EAST, you would be considered absurd. The natural basis to discuss distance is in terms of the existing city blocks.

In general, you can indicate that a vector is a coordinate with respect to a given basis. By default, we assume the canonical basis

$$\hat{\mathbf{e}}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \hat{\mathbf{e}}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \hat{\mathbf{e}}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \text{ and } \hat{\mathbf{e}}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

so the vector

$$\mathbf{u} = \begin{pmatrix} -2.3 \\ 1.6 \\ 4.7 \\ -9.0 \end{pmatrix}$$

represents the actual position

$$\sum_{k=1}^{n} u_k \hat{\mathbf{e}}_k = -2.3 \cdot \hat{\mathbf{e}}_1 + 1.6 \cdot \hat{\mathbf{e}}_2 + 4.7 \cdot \hat{\mathbf{e}}_3 - 9.0 \cdot \hat{\mathbf{e}}_4 = -2.3 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + 1.6 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + 4.7 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} - 9.0 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -2.3 \\ 1.6 \\ 4.7 \\ -9.0 \end{pmatrix},$$

however, you could also state that that vector represents coordinates with respect to the basis $B = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4\}$ where

$$\mathbf{b}_1 = \begin{pmatrix} 3 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{b}_2 = \begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{b}_3 = \begin{pmatrix} -2 \\ 1 \\ -4 \\ 0 \end{pmatrix} \text{ and } \mathbf{b}_4 = \begin{pmatrix} 4 \\ 1 \\ 2 \\ -2 \end{pmatrix},$$

in which case, the vector **u** would represent the actual position

$$\sum_{k=1}^{n} u_k \mathbf{b}_k = -2.3 \cdot \mathbf{b}_1 + 1.6 \cdot \mathbf{b}_2 + 4.7 \cdot \mathbf{b}_3 - 9.0 \cdot \mathbf{b}_4 = -2.3 \begin{pmatrix} 3 \\ 0 \\ 0 \\ 0 \end{pmatrix} + 1.6 \begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \end{pmatrix} + 4.7 \begin{pmatrix} -1 \\ 1 \\ -4 \\ 0 \end{pmatrix} - 9.0 \begin{pmatrix} 4 \\ 1 \\ 2 \\ -2 \end{pmatrix} = \begin{pmatrix} -46 \\ -1.1 \\ -36.8 \\ 18 \end{pmatrix}.$$

Now, we could rewrite this as a matrix-vector product. If the vector **u** is the coordinates with respect to the standard basis, then the actual position is

$$Id\ \mathbf{u} = \mathbf{u},$$

but if the vector **u** is the coordinates with respect to the basis *B*, then if we define the matrix $B = (\mathbf{b}_1 \quad \mathbf{b}_2 \quad \mathbf{b}_3 \quad \mathbf{b}_4)$, the vector **u** represents the actual position

$$B\mathbf{u} = \mathbf{v}.$$

To indicate that a vector represents coordinates with respect to a specific basis, we will write the vector as $\mathbf{u}_B$, and therefore we will represent the actual position as $\mathbf{u}_{Id}$.

Now, in general, given an actual position $\mathbf{u}_{Id}$, finding the coordinates with respect to a given basis *B* requires us to solve the system of linear equations

$$B\mathbf{u}_B = \mathbf{u}_{Id},$$

as while we could compute the inverse of *B* and find that $\mathbf{u}_B = B^{-1}\mathbf{u}_{Id}$, calculating the inverse can be a numerically unstable operation, so normally, this requires us to solve a system of linear equations. This can be simplified using an LU-decomposition

$$B\mathbf{u}_B = PLU\mathbf{u}_B = \mathbf{u}_{Id}.$$

Because $PP^* = Id$, the permutation matrix *P* is unitary, and thus finding the inverse is trivial—we need only calculate the adjoint. For $\mathbf{R}^n$, the terminology used is that *P* is orthogonal and we need only calculate its transpose.

Unfortunately, L and U are still lower- and upper-triangular, respectively, and thus we must perform first forward than then backward substitution. The only time that finding the inverse of B is trivial is if B is itself unitary (or, for real matrices, orthogonal). If B is unitary, then to find the coordinates of an actual position with respect to the basis B, we need only calculate

$$\mathbf{u}_B = B^* \mathbf{u}_{Id} .$$

Now, suppose we have a linear operator A acting on a vector $\mathbf{u}$ and we would like to calculate $A\mathbf{u}$. Normally, this is an expensive operation, requiring n^2 multiplications. Such a computation is less expensive if A is, for example, diagonal. In this case, the computation requires only n multiplications, for

$$A\mathbf{u} = \begin{pmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n,n} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} a_{1,1}u_1 \\ a_{2,2}u_2 \\ \vdots \\ a_{n,n}u_n \end{pmatrix},$$

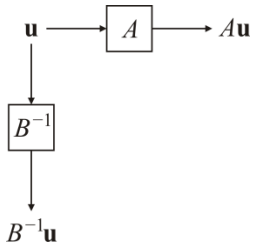
This matrix simply stretches or contracts the ith coordinate by the factor $a_{i,i}$.

In general, if we wanted to calculate $A^m\mathbf{u}$, this would require us to calculate $A\left(A\left(\cdots A\left(A\mathbf{u}\right)\right)\right)$ with m matrix multiplications, requiring therefore mn^2 multiplications. If A is diagonal, however, this becomes much easier, as

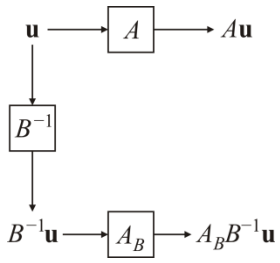
$$A^m\mathbf{u} = \begin{pmatrix} a_{1,1} & 0 & \cdots & 0 \\ 0 & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n,n} \end{pmatrix}^m \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} a_{1,1}^m u_1 \\ a_{2,2}^m u_2 \\ \vdots \\ a_{n,n}^m u_n \end{pmatrix}$$

requires only n exponentiations and n multiplications.

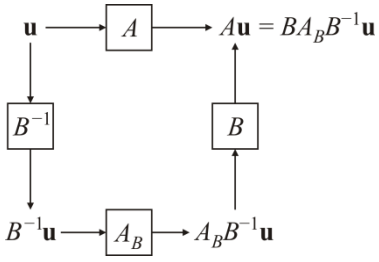
Now, suppose we have a matrix A that is not diagonal, but that there is a basis such that the action of the matrix is nothing more than an expansion or contraction with respect to each basis vector. In this case, rather than computing $A\mathbf{u}$, the first step is to find the representation of $\mathbf{u}$ with respect to this basis B, and we do so by multiplying by B^{-1}, as shown here:



$B^{-1}\mathbf{u}$ are the coordinates of $\mathbf{u}$ with respect to the basis B. With respect to the basis B, the operation of A is that of a diagonal matrix A_B, and thus we multiply by A_B, as shown here:



This gives us the action of the matrix A with respect to the basis B, but we need to find the coordinates with respect to the original basis vectors, and thus we must multiply the result by B, to get:



Now, if we wanted to calculate $A^m\mathbf{u}$, this is now much easier, as with respect to the given basis, $A\mathbf{u} = BA_B{}^mB^{-1}\mathbf{u}$. The question is, however, when is possible to find such a diagonal matrix?

Theorem

A matrix $A:\mathbf{F}^n \to \mathbf{F}^n$ is diagonalizable ($A = BA_B{}^mB^{-1}$) if and only if the matrix has n linearly independent eigenvectors.

Proof:

If A has n linearly independent eigenvectors, then $A\mathbf{v}_k = \lambda_k \mathbf{v}_k$ for $k = 1, \ldots, n$. In this case, we may define the matrix

$$B = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{pmatrix},$$

in which case, $B^{-1}\mathbf{u}$ yields a vector $B^{-1}\mathbf{u} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$ such that

$$\mathbf{u} = \sum_{k=1}^{n} \alpha_k \mathbf{v}_k \ .$$

Note that now

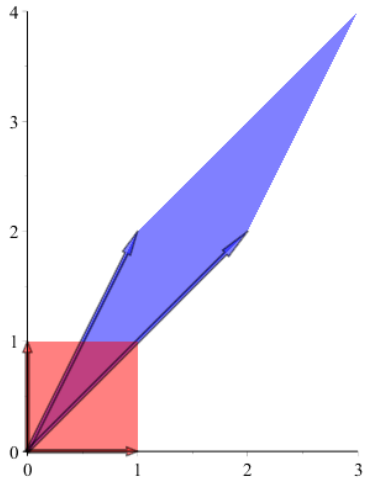
$$A\mathbf{u} = \sum_{k=1}^{n} A(\alpha_k \mathbf{v}_k) = \sum_{k=1}^{n} \alpha_k A\mathbf{v}_k = \sum_{k=1}^{n} \alpha_k \lambda_k \mathbf{v}_k \ .$$

This, however, is nothing more than $BA_B{}^mB^{-1}\mathbf{u}$.

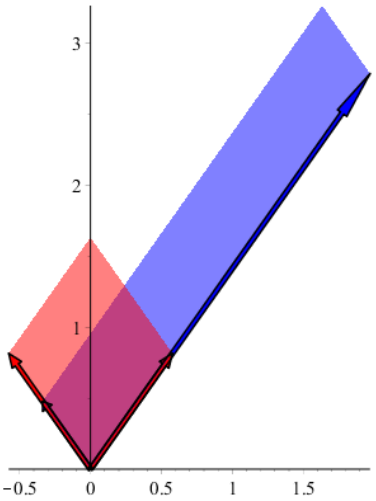
Proving this in reverse is equally straight-forward, if such a matrix B exists, then its columns must be linearly independent eigenvectors, and the corresponding entry in the diagonal matrix must be the eigenvalue. ∎

For example, consider the matrix $A=\begin{pmatrix}2 & 1\\ 2 & 2\end{pmatrix}$, which is invertible. It has a determinant equal to 2, and its eigenvalues are $\lambda_1,\lambda_2=2\pm\sqrt{2}$ with corresponding eigenvectors $\mathbf{v}_1=\begin{pmatrix}\frac{1}{\sqrt{2}}\\ 1\end{pmatrix}$ and $\mathbf{v}_2=\begin{pmatrix}-\frac{1}{\sqrt{2}}\\ 1\end{pmatrix}$, respectively. We also note that $\left(2+\sqrt{2}\right)\left(2-\sqrt{2}\right)=4-2=2$. We note that these two eigenvectors are not orthogonal. We also note that matrix of these eigenvectors $B=\begin{pmatrix}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\\ 1 & 1\end{pmatrix}$, but as this matrix is not orthogonal, we must calculate its inverse explicitly to find that $\left(\begin{array}{cc|cc}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 1 & 0\\ 1 & 1 & 0 & 1\end{array}\right)\sim\left(\begin{array}{cc|cc}1 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{2}\\ 0 & 1 & -\frac{1}{\sqrt{2}} & \frac{1}{2}\end{array}\right)$, so $B^{-1}=\begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{1}{2}\\ -\frac{1}{\sqrt{2}} & \frac{1}{2}\end{pmatrix}$.

Now, if we draw the unit square and its image under *A*, we see that the ratio of the unit square and its image is 1:2. Because the canonical basis vectors are not eigenvectors, their images are not scalar multiples of the canonical basis vectors. All of this is summarized in the following image.



If, however, we draw the normalized unit vectors and their images, we see that the first eigenvector stretched, while the second is shrunk. While the calculation of the area of the rhombus and parallelograms is more complicated, the ratio of the areas is still two.



If we wanted to calculate the image of a vector such as $\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, we may calculate this directly to get that

$A\mathbf{e}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ and $A\mathbf{e}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, or we could determine that

$$\begin{aligned} A\mathbf{e}_1 &= B\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} B^{-1}\mathbf{e}_1 \\ &= \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 2+\sqrt{2} & 0 \\ 0 & 2-\sqrt{2} \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{2} \\ -\frac{1}{\sqrt{2}} & \frac{1}{2} \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 2+\sqrt{2} & 0 \\ 0 & 2-\sqrt{2} \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1+\sqrt{2} \\ 1-\sqrt{2} \end{pmatrix} \\ &= \begin{pmatrix} 2 \\ 2 \end{pmatrix} \end{aligned}$$

and that

$$
\begin{aligned}
A\mathbf{e}_2 &= B\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}B^{-1}\mathbf{e}_2\\
&= \begin{pmatrix}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\\ 1 & 1\end{pmatrix}\begin{pmatrix}2+\sqrt{2} & 0\\ 0 & 2-\sqrt{2}\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{1}{2}\\ -\frac{1}{\sqrt{2}} & \frac{1}{2}\end{pmatrix}\begin{pmatrix}0\\ 1\end{pmatrix}\\
&= \begin{pmatrix}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\\ 1 & 1\end{pmatrix}\begin{pmatrix}2+\sqrt{2} & 0\\ 0 & 2-\sqrt{2}\end{pmatrix}\begin{pmatrix}\frac{1}{2}\\ \frac{1}{2}\end{pmatrix}\\
&= \begin{pmatrix}\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\\ 1 & 1\end{pmatrix}\begin{pmatrix}1+\frac{1}{\sqrt{2}}\\ 1-\frac{1}{\sqrt{2}}\end{pmatrix}\\
&= \begin{pmatrix}1\\ 2\end{pmatrix}
\end{aligned}
$$

Your first thought, at this point, is "why?" Certainly the direct calculation is more straight-forward than the round-about calculation. However, suppose you simply wanted to calculate $A^{20}\mathbf{e}_1$ or $A^{20}\mathbf{e}_2$. In this case, it is much easier to calculate, for example,

$$
A^{20}\mathbf{e}_2 = B\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}^{20}B^{-1}\mathbf{e}_2 = B\begin{pmatrix}\lambda_1^{20} & 0\\ 0 & \lambda_2^{20}\end{pmatrix}B^{-1}\mathbf{e}_2 .
$$

Unfortunately, the given matrix has an inverse that must be explicitly calculated—an operation that we have already suggested is undesirable as it is potentially numerically unstable. It would be much nicer if matrix of basis vectors also happened to be unitary, so that calculating its inverse is equivalent to determining the adjoint.

In this case, if $B^{-1} = B^*$, it follows that $A^* = \left(BA_BB^*\right)^* = \left(B^*\right)^* A_B^*B^* = BA_B^*B^*$. If the eigenvalues of A are real, we see that A is self-adjoint.

Another issue is, as we have seen, not all matrices in $L(\mathbf{R}^n)$ have n eigenvectors; for example,

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

have one and no eigenvectors, respectively.

Thus, this technique is only useful if:

1. all of the eigenvalues are real, and
2. there are n eigenvectors.

Thus, we have our next theorem:

Theorem

A matrix $A: \mathbf{F}^n \to \mathbf{F}^n$ is unitarily diagonalizable (that is, orthogonally for real matrices) ($A = BA_BB^*$) if and only if the matrix commutes with its adjoint (that is, $AA^* = A^*A$). We call such matrices *normal*.

Clearly, if A is real and symmetric, it must commute with its transpose; however, real skew-symmetric matrices, orthogonal and others, too, are also normal. For example, $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ is normal but has none of the properties listed, nor is it a scalar multiple of an orthogonal matrix. However, the only normal matrices that have all real eigenvalues are the self-adjoint matrices, and the only real self-adjoint matrices are symmetric matrices. Thus, we may conclude with the theorem:

Theorem

A matrix $A: \mathbf{R}^n \to \mathbf{R}^n$ is orthogonally diagonalizable ($A = BA_BB^{\mathrm{T}}$) if and only if the matrix is symmetric.

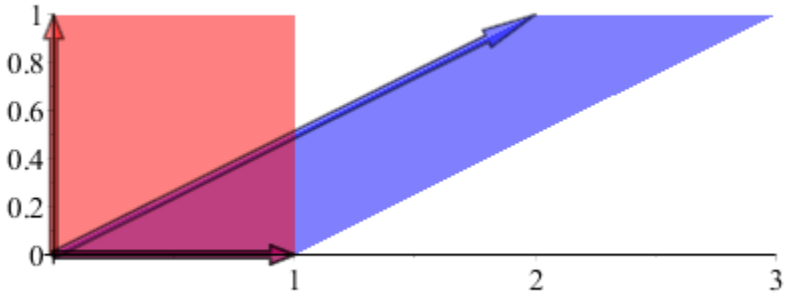
The eigenvectors of A form the column of B, and the corresponding eigenvalues form the diagonal entries of the diagonal matrix A_B.

We will look at one example. Consider the symmetric matrix $A = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$ which has eigenvalues $\lambda_{1,2} = 1 \pm \sqrt{2}$. Again, we note that the determinant is –1 and $\lambda_1\lambda_2 = \left(1+\sqrt{2}\right)\left(1-\sqrt{2}\right) = 1-2 = -1$. Two eigenvectors associated with these two eigenvalues are $\mathbf{v}_{1,2} = \begin{pmatrix} 1 \\ -1 \pm \sqrt{2} \\ 1 \end{pmatrix}$. Normalizing these two eigenvectors results in an expression that is of no benefit to this course, so we will use a numerical approximation:

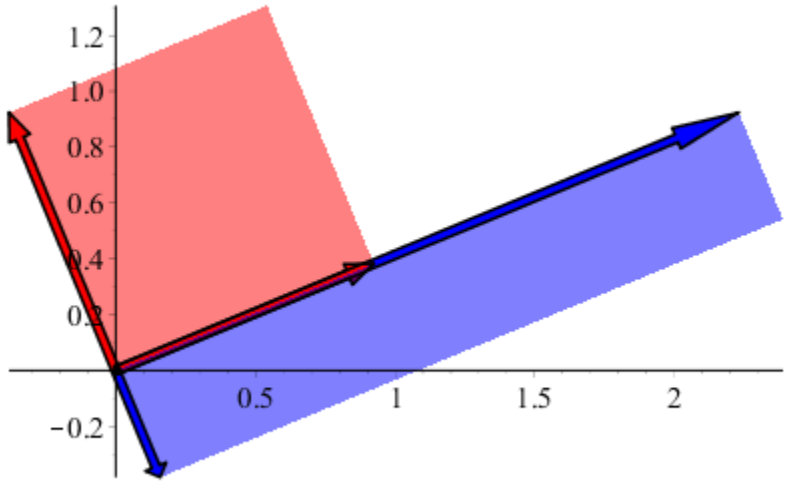
$$B = \begin{pmatrix} 0.9238795325 & -0.3826834325 \\ 0.3826834325 & 0.9238795325 \end{pmatrix}.$$

By inspection, we see that the column vectors are orthonormal, and thus this matrix may be described as being orthonormal, and therefore its inverse is its transpose.

Now, if we look of the unit square, we may think that one eigenvector is the vector $\mathbf{e}_1$, but on inspection, we see that $A\mathbf{e}_2 = \mathbf{e}_1$.



If we plot the normalized eigenvectors, we see that each is mapped onto a scalar multiple of itself:



The benefit, here, however, is that the two eigenvectors are orthogonal, and thus, finding the inverse of B is calculating its transpose:

$$B^{-1} = B^{\mathrm{T}} = \begin{pmatrix} 0.9238795325 & 0.3826834325 \\ -0.3826834325 & 0.9238795325 \end{pmatrix}.$$

This operation is numerically stable.

Problems

1. Find the eigenvalues and normalized eigenvectors of the matrices

$$A = \begin{pmatrix} -4.6 & 7.2 \\ 7.2 & -0.4 \end{pmatrix},\ B = \begin{pmatrix} \frac{59}{5} & -\frac{168}{5} \\ -\frac{168}{5} & \frac{586}{5} \end{pmatrix} \text{ and } C = \begin{pmatrix} 1 & 1 \\ -0.125 & 0.25 \end{pmatrix}$$

and find a basis that diagonalizes A. Only normalize the eigenvectors if they are orthogonal.

2. Find the eigenvalues and normalized eigenvectors of the matrices

$$A=\begin{pmatrix}-5.392 & -1.344\\ -1.344 & 9.608\end{pmatrix},\ B=\begin{pmatrix}6.8 & -2.4\\ -2.4 & 8.2\end{pmatrix} \text{ and } C=\begin{pmatrix}2 & 1\\ -1.5 & -0.5\end{pmatrix}.$$

and find an orthogonal matrix that diagonalizes *A*.

3. Find the eigenvalues and normalized eigenvectors of the matrix

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

and find an orthogonal matrix that diagonalizes *A*.

5. Find the eigenvalues and normalized eigenvectors of the matrix

$$A=\begin{pmatrix}\frac{3}{5} & 2 & -\frac{4}{5}\\ 2 & 8 & -6\\ -\frac{4}{5} & -6 & \frac{47}{5}\end{pmatrix}$$

and find an orthogonal matrix that diagonalizes *A*.

Solutions

1. For $A=\begin{pmatrix}-4.6 & 7.2\\ 7.2 & -0.4\end{pmatrix}$, $\det(A-\lambda Id)=\lambda^2+5\lambda-50$ which has roots $\lambda_1 = 5$ and $\lambda_2 = -10$.

To find the null space of $A-5\cdot Id$, we note that $A-5Id=\begin{pmatrix}-9.6 & 7.2\\ 7.2 & -5.4\end{pmatrix}\sim\begin{pmatrix}-9.6 & 7.2\\ 0 & 0\end{pmatrix}$ and therefore α_2 is a free variable (as it does not correspond to any leading non-zero entry). The first equation thus gives us that $-9.6\alpha_1+7.2\alpha_2=0$ so $\alpha_1=0.75\alpha_2$. Thus, a corresponding eigenvector is $\mathbf{u}_1=\begin{pmatrix}0.75\\ 1\end{pmatrix}$, and normalized, this is $\hat{\mathbf{u}}_1=\begin{pmatrix}0.6\\ 0.8\end{pmatrix}$.

To find the null space of $A-(-10)\cdot Id$, we note that $A+10\cdot Id=\begin{pmatrix}5.4 & 7.2\\ 7.2 & 9.6\end{pmatrix}\sim\begin{pmatrix}7.2 & 9.6\\ 0 & 0\end{pmatrix}$, and therefore, again, α_2 is a free variable. The first equation thus gives us that $7.2\alpha_1+9.6\alpha_2=0$ so $\alpha_1=-1.\overline{3}\alpha_2$. Thus, a corresponding eigenvector is $\mathbf{u}_2=\begin{pmatrix}-1.\overline{3}\\ 1\end{pmatrix}$, and normalized, this is $\hat{\mathbf{u}}_2=\begin{pmatrix}-0.8\\ 0.6\end{pmatrix}$.

Therefore, *A* may be diagonalized by $E=\begin{pmatrix}0.6 & -0.8\\ 0.8 & 0.6\end{pmatrix}$ and $A_E=\begin{pmatrix}5 & 0\\ 0 & -10\end{pmatrix}$. We note that

$$EA_E E^{\mathrm{T}} = \begin{pmatrix} 0.6 & -0.8 \\ 0.8 & 0.6 \end{pmatrix}\begin{pmatrix} 5 & 0 \\ 0 & -10 \end{pmatrix}\begin{pmatrix} 0.6 & 0.8 \\ -0.8 & 0.6 \end{pmatrix} = \begin{pmatrix} -4.6 & 7.2 \\ 7.2 & -0.4 \end{pmatrix} = A.$$

For $B = \begin{pmatrix} \frac{59}{5} & -\frac{168}{5} \\ -\frac{168}{5} & \frac{586}{5} \end{pmatrix}$, $\det(B - \lambda Id) = \lambda^2 - 129\lambda + 254$ which has roots $\lambda_1 = 2$ and $\lambda_2 = 127$.

To find the null space of $B - 2 \cdot Id$, we note that $B - 2Id = \begin{pmatrix} \frac{49}{5} & -\frac{168}{5} \\ -\frac{168}{5} & \frac{576}{5} \end{pmatrix} \sim \begin{pmatrix} -\frac{168}{5} & \frac{576}{5} \\ 0 & 0 \end{pmatrix}$ and therefore α_2 is a free variable. The first equation thus gives us that $-\frac{168}{5}\alpha_1 + \frac{576}{5}\alpha_2 = 0$ so $\alpha_1 = \frac{24}{7}\alpha_2$. Thus, a corresponding eigenvector is $\mathbf{u}_1 = \begin{pmatrix} \frac{24}{7} \\ 1 \end{pmatrix}$, and normalized, this is $\hat{\mathbf{u}}_1 = \begin{pmatrix} \frac{24}{25} \\ \frac{7}{25} \end{pmatrix}$.

To find the null space of $B - 127 \cdot Id$, we note that $B - 127Id = \begin{pmatrix} -\frac{576}{5} & -\frac{168}{5} \\ -\frac{168}{5} & -\frac{49}{5} \end{pmatrix} \sim \begin{pmatrix} -\frac{576}{5} & -\frac{168}{5} \\ 0 & 0 \end{pmatrix}$ and therefore α_2 is a free variable. The first equation thus gives us that $-\frac{576}{5}\alpha_1 - \frac{168}{5}\alpha_2 = 0$ so $\alpha_1 = -\frac{7}{24}\alpha_2$. Thus, a corresponding eigenvector is $\mathbf{u}_2 = \begin{pmatrix} -\frac{7}{24} \\ 1 \end{pmatrix}$, and normalized, this is $\hat{\mathbf{u}}_2 = \begin{pmatrix} -\frac{7}{25} \\ \frac{24}{25} \end{pmatrix}$.

Therefore, *A* may be diagonalized by $E = \begin{pmatrix} \frac{24}{25} & -\frac{7}{25} \\ \frac{7}{25} & \frac{24}{25} \end{pmatrix}$ and $B_E = \begin{pmatrix} 2 & 0 \\ 0 & 127 \end{pmatrix}$. We note that

$$EB_E E^{\mathrm{T}} = \begin{pmatrix} \frac{24}{25} & -\frac{7}{25} \\ \frac{7}{25} & \frac{24}{25} \end{pmatrix}\begin{pmatrix} 2 & 0 \\ 0 & 127 \end{pmatrix}\begin{pmatrix} \frac{24}{25} & \frac{7}{25} \\ -\frac{7}{25} & \frac{24}{25} \end{pmatrix} = \begin{pmatrix} \frac{59}{5} & -\frac{168}{5} \\ -\frac{168}{5} & \frac{586}{5} \end{pmatrix} = B.$$

For the matrix $C = \begin{pmatrix} 1 & 1 \\ -0.125 & 0.25 \end{pmatrix}$, $\det(C - \lambda Id) = \lambda^2 - 1.25\lambda + 0.375$ which has roots $\lambda_1 = 0.5$ and $\lambda_2 = 0.75$.

To find the null space of $A - 0.5 \cdot Id$, we note that $A - 0.5 \cdot Id = \begin{pmatrix} 0.5 & 1 \\ -0.125 & -0.25 \end{pmatrix} \sim \begin{pmatrix} 0.5 & 1 \\ 0 & 0 \end{pmatrix}$ and therefore α_2 is a free variable. The first equation thus gives us that $0.5\alpha_1 + \alpha_2 = 0$ so $\alpha_1 = -2\alpha_2$. Thus, a corresponding eigenvector is $\mathbf{u}_1 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$.

To find the null space of $A - 0.75 \cdot Id$, we note that $A - 0.5 \cdot Id = \begin{pmatrix} 0.25 & 1 \\ -0.125 & -0.5 \end{pmatrix} \sim \begin{pmatrix} 0.25 & 1 \\ 0 & 0 \end{pmatrix}$ and therefore α_2 is a free variable. The first equation thus gives us that $0.25\alpha_1 + \alpha_2 = 0$ so $\alpha_1 = -4\alpha_2$. Thus, a corresponding eigenvector is $\mathbf{u}_2 = \begin{pmatrix} -4 \\ 1 \end{pmatrix}$.

Now, either by inspection, or by noting that the matric C is not symmetric, $\mathbf{u}_1$ and $\mathbf{u}_2$ are not orthogonal, and therefore, we can simply write $E = \begin{pmatrix} -2 & -4 \\ 1 & 1 \end{pmatrix}$ and $C_E = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.75 \end{pmatrix}$, and calculate the inverse as $E^{-1} = \begin{pmatrix} 0.5 & 2 \\ -0.5 & -1 \end{pmatrix}$ and we see that

$$EC_E E^{-1} = \begin{pmatrix} -2 & -4 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 0.5 & 0 \\ 0 & 0.75 \end{pmatrix}\begin{pmatrix} 0.5 & 2 \\ -0.5 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ -0.125 & 0.25 \end{pmatrix} = C .$$

3. First, the characteristic polynomial is found by finding the determinant

$$\begin{aligned} \det\left(A - \lambda \cdot Id_3\right) = \det\left(\begin{pmatrix} 1-\lambda & 1 & 0 \\ 1 & 1-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{pmatrix}\right) &= \left(1-\lambda\right)^3 - \left(1-\lambda\right) \\ &= 1 - 3\lambda + 3\lambda^2 - \lambda^3 - 1 + \lambda \\ &= -\lambda^3 + 3\lambda^2 - 2\lambda \end{aligned}$$

We can factor out a $-\lambda$ to get $-\lambda\left(\lambda^2 - 3\lambda + 2\right)$, either by inspection or by the quadratic formula,

$$\frac{3 \pm \sqrt{9 - 4 \cdot 1 \cdot 2}}{2} = \frac{3 \pm 1}{2} = 1, 2 ,$$

and therefore the three eigenvalues are 0, 1 and 2.

We now know that $A - \lambda \cdot Id_3$ is non-invertible only when λ = 0, 1 or 2. If you substitute in any other value of λ, you will find that the result is invertible; for example, $A - 5 \cdot Id_3 = \begin{pmatrix} 1-5 & 1 & 0 \\ 1 & 1-5 & 0 \\ 0 & 0 & 1-5 \end{pmatrix} = \begin{pmatrix} -4 & 1 & 0 \\ 1 & -4 & 0 \\ 0 & 0 & -4 \end{pmatrix}$ is invertible with inverse $\left(A - 5 \cdot Id_3\right)^{-1} = \begin{pmatrix} -\frac{4}{15} & -\frac{1}{15} & 0 \\ -\frac{1}{15} & -\frac{4}{15} & 0 \\ 0 & 0 & -\frac{1}{4} \end{pmatrix}$.

Now, to find the eigenvectors corresponding to each of these three eigenvalues, we must find the null space of

$$\begin{pmatrix} 1-\lambda & 1 & 0 \\ 1 & 1-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{pmatrix}$$

for each of $\lambda_1 = 0$, $\lambda_2 = 1$ and $\lambda_3 = 2$, respectively.

Thus, to begin, to find the null space when $\lambda = 0$, we note

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

Thus, the only free variable is α_2, as it is the only coefficient without a corresponding leading non-zero entry; therefore the dimension of the null space is one. Thus, backward substitution gives us that $\alpha_3 = 0$, and therefore $\alpha_1 + \alpha_2 = 0$, so $\alpha_1 = -\alpha_2$. Thus, the null space of this matrix is

$$\alpha_2 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix},$$

and thus, the dimension of the null space is $\lambda = 0$ and a basis vector for it is $\mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$. We can normalize this vector to get $\hat{\mathbf{v}}_1 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}$.

Next, to find the null space when $\lambda_2 = 1$, we note that

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

and therefore α_3 is a free variable so the dimension of the null space is one. The second equation gives us that $\alpha_2 = 0$ and the first gives us that $\alpha_1 = 0$, and thus all solutions are of the form

$$\alpha_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix},$$

and thus a normalized basis vector for this null space is $\hat{\mathbf{v}}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$.

Finally, to find the null space when $\lambda_3 = 2$, we note that

$$\begin{pmatrix} 1-2 & 1 & 0 \\ 1 & 1-2 & 0 \\ 0 & 0 & 1-2 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \sim \begin{pmatrix} -1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} \sim \begin{pmatrix} -1 & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}$$

and therefore α_2 is, again, a free variable so the dimension of the null space is one. The second equation gives us that $-\alpha_3 = 0$, and so $\alpha_3 = 0$, and the first equation gives us that $-\alpha_1 + \alpha_2 = 0$, so $\alpha_1 = \alpha_2$. Thus, all solutions are of the form

$$\alpha_2 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix},$$

and thus a normalized basis vector for this null space is $\hat{\mathbf{v}}_3 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}$. Thus, our orthogonal matrix is

$$B = \left(\hat{\mathbf{v}}_1\ \hat{\mathbf{v}}_2\ \hat{\mathbf{v}}_3\right) = \begin{pmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{pmatrix}.$$

Because the original matrix is symmetric, it follows that because B is orthogonal (meaning, its column vectors are orthogonal and normalized), it follows that $B^{-1} = B^{\mathrm{T}}$. The corresponding diagonal matrix is

$$A_B = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

We note that

$$BA_BB^{\mathrm{T}} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}\begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ \sqrt{2} & \sqrt{2} & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = A.$$

If we chose a different order of the eigenvalues, this would simply mean rearranging the columns of B, the rows of B^{T}, and the entries of A_B.

5. First, the characteristic polynomial is

$$\det\left(A - \lambda \cdot Id_3\right) = -\lambda^3 + 18\lambda^2 - 45\lambda$$

from which we may deduce that the eigenvalues are 0, 3 and 15. Next, we find the null space of each of the matrices

$$A - 0Id_3 = \begin{pmatrix} \frac{3}{5} & 2 & -\frac{4}{5} \\ 2 & 8 & -6 \\ -\frac{4}{5} & -6 & \frac{47}{5} \end{pmatrix},\ A - 3Id_3 = \begin{pmatrix} -\frac{12}{5} & 2 & -\frac{4}{5} \\ 2 & 5 & -6 \\ -\frac{4}{5} & -6 & \frac{32}{5} \end{pmatrix} \text{ and } A - 15Id_3 = \begin{pmatrix} -\frac{72}{5} & 2 & -\frac{4}{5} \\ 2 & -7 & -6 \\ -\frac{4}{5} & -6 & -\frac{28}{5} \end{pmatrix},$$

and solving these, we get that

$$A \sim \begin{pmatrix} 2 & 8 & -6 \\ 0 & -\frac{14}{5} & 7 \\ 0 & 0 & 0 \end{pmatrix},\ A - 3Id_3 \sim \begin{pmatrix} -\frac{12}{5} & 2 & -\frac{4}{5} \\ 0 & \frac{20}{3} & -\frac{20}{3} \\ 0 & 0 & 0 \end{pmatrix} \text{ and } A - 15Id_3 = \begin{pmatrix} -\frac{72}{5} & 2 & -\frac{4}{5} \\ 2 & -7 & -6 \\ -\frac{4}{5} & -6 & -\frac{28}{5} \end{pmatrix}$$

Finding the null spaces of each of these, we get the vectors

$$\mathbf{v}_1 = \alpha_3 \begin{pmatrix} -7 \\ \frac{5}{2} \\ 1 \end{pmatrix},\ \mathbf{v}_2 = \alpha_3 \begin{pmatrix} \frac{1}{2} \\ 1 \\ 1 \end{pmatrix} \text{ and } \mathbf{v}_3 = \alpha_3 \begin{pmatrix} -\frac{2}{11} \\ -\frac{10}{11} \\ 1 \end{pmatrix}.$$

Normalizing the corresponding eigenvectors, we get

$$\hat{\mathbf{v}}_1 = \begin{pmatrix} -\frac{14}{15} \\ \frac{1}{3} \\ \frac{2}{15} \end{pmatrix},\ \hat{\mathbf{v}}_2 = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{pmatrix} \text{ and } \mathbf{v}_3 = \begin{pmatrix} -\frac{2}{15} \\ -\frac{2}{3} \\ \frac{11}{15} \end{pmatrix},$$

and thus our orthogonal matrix is

$$B = \begin{pmatrix} -\frac{14}{15} & \frac{1}{3} & -\frac{2}{15} \\ \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ \frac{2}{15} & \frac{2}{3} & \frac{11}{15} \end{pmatrix}$$

and the diagonal matrix corresponding to this orthogonal matrix is

$$A_B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 15 \end{pmatrix}.$$

We note that

$$BA_B B^{\mathrm{T}} = \begin{pmatrix} -\frac{14}{15} & \frac{1}{3} & -\frac{2}{15} \\ \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ \frac{2}{15} & \frac{2}{3} & \frac{11}{15} \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 15 \end{pmatrix} \begin{pmatrix} -\frac{14}{15} & \frac{1}{3} & \frac{2}{15} \\ \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ -\frac{2}{15} & -\frac{2}{3} & \frac{11}{15} \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & 2 & -\frac{4}{5} \\ 2 & 8 & -6 \\ -\frac{4}{5} & -6 & \frac{47}{5} \end{pmatrix} = A.$$

Note, if you choose a different order of the eigenvalues, you will simply rearrange the columns of B.

7 A digression to real 3-dimensional space

Having described the inner product and defined orthogonality, we will now look at one very specific application of these concepts to that of discuss lines and planes in 3-space, or $\mathbf{R}^3$. We will see how we can define vectors perpendicular to a plane, and planes with all points perpendicular to a line. We will also introduce the cross product.

First, the standard basis in $\mathbf{F}^n$ is usually represented by $\{\hat{\mathbf{e}}_1,\ldots,\hat{\mathbf{e}}_n\}$; however, in $\mathbf{R}^3$, it is common to use $\{\hat{\mathbf{i}},\hat{\mathbf{j}},\hat{\mathbf{k}}\}$ or $\{\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{z}}\}$. We will use the *ijk*-representation in this chapter but be prepared in other courses to see different representations.

7.1 Equations of lines

Every line in 3-space may be written as

$$\mathbf{u}+\alpha\mathbf{v} \text{ or } \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}+\alpha\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

Now, consider, for example, the line

$$\begin{pmatrix} 3.3 \\ 4.9 \\ 72.7 \end{pmatrix}+\alpha\begin{pmatrix} 0 \\ 0 \\ 143.2 \end{pmatrix}.$$

First, we should note that the direction vector could be normalized, as follows,

$$\begin{pmatrix} 3.3 \\ 4.9 \\ 72.7 \end{pmatrix}+\alpha\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

At this point, too, we note that the 2-norm of the vector $\begin{pmatrix} 3.3 \\ 4.9 \\ 72.7 \end{pmatrix}$ is quite large at approximately 72.94. It would be preferable to have this vector as small as possible, and in fact, it would be preferable to have this vector be the closest point to the origin on this line. In this case, that would be the vector $\begin{pmatrix} 3.3 \\ 4.9 \\ 0 \end{pmatrix}$, but in general, this would be the perpendicular component of **u** with respect to **v**.

Thus, for any line $\mathbf{u}+\alpha\mathbf{v}$, we could define a new line $\mathbf{u}_\perp+\alpha\hat{\mathbf{v}}$ where $\mathbf{u}_\perp=\text{perp}_{\mathbf{v}}(\mathbf{u})=\mathbf{u}-\text{proj}_{\mathbf{v}}(\mathbf{u})$ and $\hat{\mathbf{v}}$ is the normalized vector **v**. This vector $\mathbf{u}_\perp$ represents the closest point to the origin of the line.

For example, suppose a line is defined by

$$\begin{pmatrix} 2 \\ 4 \\ -2 \end{pmatrix} + \alpha \begin{pmatrix} -3 \\ 2 \\ -6 \end{pmatrix}.$$

First, the two vectors are not perpendicular, as the inner product is $2(-3)+4\cdot 2+(-2)(-6)=14$. Consequently, we find

$$\mathbf{u}_{\perp} = \text{perp}_{\mathbf{v}}(\mathbf{u}) = \begin{pmatrix} 2 \\ 4 \\ -2 \end{pmatrix} - \frac{14}{49}\begin{pmatrix} -3 \\ 2 \\ -6 \end{pmatrix} = \begin{pmatrix} \frac{20}{7} \\ \frac{24}{7} \\ -\frac{2}{7} \end{pmatrix},$$

and if we normalize **v**, we have

$$\hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2} = \frac{1}{7}\begin{pmatrix} -3 \\ 2 \\ -6 \end{pmatrix} = \begin{pmatrix} -\frac{3}{7} \\ \frac{2}{7} \\ -\frac{6}{7} \end{pmatrix}.$$

Thus, the simplified equation of our line is

$$\begin{pmatrix} \frac{20}{7} \\ \frac{24}{7} \\ -\frac{2}{7} \end{pmatrix} + \alpha \begin{pmatrix} -\frac{3}{7} \\ \frac{2}{7} \\ -\frac{6}{7} \end{pmatrix},$$

and we note that the two vectors are orthogonal: $\frac{20}{7}\left(-\frac{3}{7}\right)+\frac{24}{7}\cdot\frac{2}{7}+\left(-\frac{2}{7}\right)\left(-\frac{6}{7}\right)=\frac{-60+48+12}{7}=0$. Additionally, the shortest distance to the line from the origin is $\sqrt{\left(\frac{20}{7}\right)^2+\left(\frac{24}{7}\right)^2+\left(-\frac{2}{7}\right)^2}=\sqrt{20}\approx 4.47$.

7.2 Finding the line through two points

Given two different points **u** and **v**, the equation of the line running through those two points is

$$\mathbf{u}+\alpha(\mathbf{u}-\mathbf{v}).$$

If $\mathbf{u}-\mathbf{v}=\mathbf{0}$, the points are *coincident* (the same point), in which case, we cannot find such a line.

Consequently, the line passing through the two points $\begin{pmatrix}-3\\0\\-1\end{pmatrix}$ and $\begin{pmatrix}3\\3\\-7\end{pmatrix}$ may be described by

$$\begin{pmatrix}-3\\0\\1\end{pmatrix}+\alpha\begin{pmatrix}-6\\-3\\6\end{pmatrix},$$

and this may be simplified to

$$\begin{pmatrix}-\frac{34}{100}\\1\\\frac{238}{100}\end{pmatrix}+\alpha\begin{pmatrix}-\frac{7}{\sqrt{50}}\\0\\-\frac{1}{\sqrt{50}}\end{pmatrix},$$

where the first vector is perpendicular to the second, and the second vector is normalized and the shortest distance to this line is the norm of the constant vector, in this case, approximately 2.60.

7.3 Planes

You will recall from secondary school that the equation of a plane is of the form

$$ax + by + cz + d = 0$$

where a, b, c and d are constants. For example, $x - 2y + 2z + 5 = 0$ is one such plane.

If $\mathbf{v} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ is a point in $\mathbf{R}^3$, you will note that the first half of the equation is the inner product of the vector $\mathbf{x}$ and the constant vector $\mathbf{n} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$, in which case, we may rewrite this plane as

$$\langle \mathbf{v}, \mathbf{n} \rangle = -d \, .$$

The vector $\mathbf{n}$ is *normal*[20] to the plane. As with the previous problem, there is no reason we should not choose a normal vector has unit length. Replacing $\mathbf{n} = \|\mathbf{n}\|_2 \dfrac{\mathbf{n}}{\|\mathbf{n}\|_2} = \|\mathbf{n}\|_2 \, \hat{\mathbf{n}}$ in the equation yields

$$\left\langle \mathbf{v}, \|\mathbf{n}\|_2 \, \hat{\mathbf{n}} \right\rangle = \|\mathbf{n}\|_2 \langle \mathbf{v}, \hat{\mathbf{n}} \rangle = -d$$

so

$$\langle \mathbf{v}, \hat{\mathbf{n}} \rangle = -\frac{d}{\|\mathbf{n}\|_2}$$

is an equation of the line that is now normalized. For example,

The question now is, what is the significance of the constant $-\dfrac{d}{\|\mathbf{n}\|_2}$? Recall that $\langle \hat{\mathbf{n}}, \hat{\mathbf{n}} \rangle = 1$ so $\left\langle -\dfrac{d}{\|\mathbf{n}\|_2} \hat{\mathbf{n}}, \hat{\mathbf{n}} \right\rangle = -\dfrac{d}{\|\mathbf{n}\|_2} \langle \hat{\mathbf{n}}, \hat{\mathbf{n}} \rangle = -\dfrac{d}{\|\mathbf{n}\|_2}$, so $-\dfrac{d}{\|\mathbf{n}\|_2} \hat{\mathbf{n}}$ is a point on the plane.

Thus,

$$\langle \mathbf{x}, \hat{\mathbf{n}} \rangle = \left\langle -\frac{d}{\|\mathbf{n}\|_2} \hat{\mathbf{n}}, \hat{\mathbf{n}} \right\rangle$$

for every point on the plane.

---

[20] Unfortunately, the word "normal" here has nothing to do with either of the words we have previously used: "norm" or "normalize". Any vector that is perpendicular to a plane is said to be normal to that plane, and a carpenter's square (for making perpendicular lines) was called a *norma*.

Now, let **x** be an arbitrary point on the plane. Thus,

$$\mathbf{x} = \mathbf{x} - \left(-\frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}} \right) + \left(-\frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}} \right).$$

Notice that **x** is the sum of two vectors: $\mathbf{x} + \frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}}$ and $-\frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}}$. If we calculate the inner product of these two vectors, we get

$$\begin{aligned}\left\langle \mathbf{x} + \frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}}, -\frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}} \right\rangle &= \left\langle \mathbf{x}, -\frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}} \right\rangle + \left\langle \frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}}, -\frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}} \right\rangle \\ &= -\frac{d}{\|\mathbf{n}\|_2}\langle \mathbf{x}, \hat{\mathbf{n}} \rangle + \left(\frac{d}{\|\mathbf{n}\|_2} \right) \langle \hat{\mathbf{n}}, \hat{\mathbf{n}} \rangle \\ &= -\left(\frac{d}{\|\mathbf{n}\|_2} \right)^2 + \left(\frac{d}{\|\mathbf{n}\|_2} \right)^2 = 0\end{aligned}$$

Thus, these two vectors are perpendicular, and therefore by applying Pythagorus's theorem, we have

$$\|\mathbf{x}\|_2^2 = \left\| \mathbf{x} - \left(-\frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}} \right) \right\|_2^2 + \left\| -\frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}} \right\|_2^2 \geq \left\| -\frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}} \right\|_2^2 = \frac{d^2}{\|\mathbf{n}\|_2^2},$$

and therefore $\|\mathbf{x}\|_2 \geq \frac{|d|}{\|\mathbf{n}\|_2}$. It follows from this that the point $-\frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}}$ is the point on the plane that is closest to the origin, and thus the shortest distance from the origin to the plane is $\frac{|d|}{\|\mathbf{n}\|_2}$. Given the original equation of the plane, $ax + by + cz + d = 0$, the shortest distance from the origin to the plane is therefore $\frac{|d|}{\sqrt{a^2 + b^2 + c^2}}$.

If $d > 0$, normal vector $\hat{\mathbf{n}}$ points from the plane towards the origin, while if $d < 0$, the normal vector points from the plane away from the origin.

Now, given $\mathbf{n} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$, because **n** is not the zero vector, at least one entry is non-zero. Consequently, the following three vectors are orthogonal to **n** and must therefore parallel to the plane:

$$\begin{pmatrix} b \\ -a \\ 0 \end{pmatrix}, \begin{pmatrix} c \\ 0 \\ -a \end{pmatrix}, \begin{pmatrix} 0 \\ c \\ -b \end{pmatrix}.$$

Note that:

1. If only one of a, b or c is non-zero, one of these vectors is the zero vector and the other two will be orthogonal. For example, if $b = c = 0$, the three vectors are $\begin{pmatrix} 0 \\ -a \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ -a \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.
2. If one of a, b or c is zero, one vector will be a scalar multiple of the another, and this will be orthogonal to the third. For example, if $c = 0$, the three vectors are $\begin{pmatrix} b \\ -a \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ -a \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ -b \end{pmatrix}$.
3. If a, b and c are non-zero, we see that only the first two are linearly independent:

$$\begin{pmatrix} b & c & 0 \\ -a & 0 & c \\ 0 & -a & -b \end{pmatrix} \sim \begin{pmatrix} b & c & 0 \\ 0 & \frac{ac}{b} & c \\ 0 & -a & -b \end{pmatrix} \sim \begin{pmatrix} b & c & 0 \\ 0 & \frac{ac}{b} & c \\ 0 & 0 & 0 \end{pmatrix}.$$

The the first two, $\begin{pmatrix} b \\ -a \\ 0 \end{pmatrix}$ and $\begin{pmatrix} c \\ 0 \\ -a \end{pmatrix}$, are not orthogonal, but $\begin{pmatrix} c \\ 0 \\ -a \end{pmatrix} - \frac{cb}{a^2+b^2}\begin{pmatrix} b \\ -a \\ 0 \end{pmatrix} = \begin{pmatrix} c - \frac{b^2 c}{a^2+b^2} \\ \frac{abc}{a^2+b^2} \\ -a \end{pmatrix}$ is orthogonal to $\begin{pmatrix} b \\ -a \\ 0 \end{pmatrix}$. As a scalar multiple of an orthogonal vector is still orthogonal, we could multiply this vector by $\frac{a^2+b^2}{a}$ to get $\begin{pmatrix} ac \\ bc \\ -a^2-b^2 \end{pmatrix}$ and we note that $\begin{pmatrix} a \\ b \\ c \end{pmatrix}$, $\begin{pmatrix} b \\ -a \\ 0 \end{pmatrix}$ and $\begin{pmatrix} ac \\ bc \\ -a^2-b^2 \end{pmatrix}$ are three mutually orthogonal vectors.

Let us now consider the plane we presented as an example: $x - 2y + 2z + 5 = 0$. The normal vector is $\mathbf{n} = \begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}$ and the normalized normal is $\hat{\mathbf{n}} = \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{2}{3} \end{pmatrix}$. Thus, another representation of the plane is $\frac{1}{3}x - \frac{2}{3}y + \frac{2}{3}z + \frac{5}{3} = 0$ or all vectors that satisfy

$$\left\langle \mathbf{x} + \begin{pmatrix} \frac{5}{9} \\ -\frac{10}{9} \\ \frac{10}{9} \end{pmatrix}, \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{2}{3} \end{pmatrix} \right\rangle = 0 .$$

We note that the closest point on the plane to the origin is $-\frac{d}{\|\mathbf{n}\|_2}\hat{\mathbf{n}} = \begin{pmatrix} -\frac{5}{9} \\ \frac{10}{9} \\ -\frac{10}{9} \end{pmatrix}$, and the distance to the plane is $\frac{|d|}{\|\mathbf{n}\|_2} = \frac{5}{\sqrt{1^2+(-2)^2+2^2}} = \frac{5}{3}$. We note that $\left(-\frac{5}{9}\right)-2\left(\frac{10}{9}\right)+2\left(-\frac{10}{9}\right)+5 = -\frac{5}{9}-\frac{20}{9}-\frac{20}{9}+\frac{45}{9} = 0$, so indeed, this point is on the plane. Note that as $d > 0$, the closest point to the plain is in the opposite direction of the normal vector, and thus, the normal vector points from the plane towards the origin.

The normal vector and two perpendicular vectors in the plane are

$$\begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix}, \begin{pmatrix} -2 \\ -1 \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} 2 \\ -4 \\ -5 \end{pmatrix}.$$

Thus, all vectors in the plane can be written in the form $\begin{pmatrix} -\frac{5}{9} \\ \frac{10}{9} \\ -\frac{10}{9} \end{pmatrix} + \alpha \begin{pmatrix} -\frac{2}{\sqrt{5}} \\ -\frac{1}{\sqrt{5}} \\ 0 \end{pmatrix} + \beta \begin{pmatrix} \frac{2}{3\sqrt{5}} \\ -\frac{4}{3\sqrt{5}} \\ -\frac{5}{3\sqrt{5}} \end{pmatrix}$.

7.4 The intersection of two planes

Two planes are parallel to each other if two normal vectors are scalar products of each other. Suppose that we have two planes that are not perpendicular to each other. In this case, they must intersect at a line, so we would like to find the equation of that line. Suppose that two planes are defined by

$$a_1x + b_1y + c_1z = d_1$$
$$a_2x + b_2y + c_2z = d_2$$

This is a system of two equations and two unknowns. If we solve such a system, it will necessarily be underdetermined, so there will be either zero or an infinite number of solutions. If there are no solutions, the planes are parallel but different, but if there are infinitely many solutions, there are two possibilities:

1. All the points are the same; that is, the two planes are identical, or
2. A line of points are the same, which defines the line of intersection.

For example, we will consider the three planes:

$$\begin{aligned} x - 4y + z &= 2 \\ -2x + 8y - 2z &= -4 \end{aligned} \qquad \begin{aligned} x - 4y + z &= 2 \\ -2x + 8y - 2z &= 7 \end{aligned} \qquad \begin{aligned} x - 4y + z &= 2 \\ -2x + 6y - 2z &= 7 \end{aligned}$$

Writing these in matrix form, we have:

$$\left(\begin{array}{ccc|c} 1 & -4 & 1 & 2 \\ -2 & 8 & -2 & -4 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & -4 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{array}\right) \qquad \left(\begin{array}{ccc|c} 1 & -4 & 1 & 2 \\ -2 & 8 & -2 & 7 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & -4 & 1 & 2 \\ 0 & 0 & 0 & 11 \end{array}\right) \qquad \left(\begin{array}{ccc|c} 1 & -4 & 1 & 2 \\ -2 & 6 & -2 & 7 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & -4 & 1 & 2 \\ 0 & -2 & 0 & 11 \end{array}\right)$$

In the first case, the remaining equation is the equation of the plane: $x - 4y + z = 2$. In the second, the two planes are parallel and distinct. In the third, the line of intersection is defined by $-2y = 11$ or $y = -\frac{11}{2}$ and thus $x + 22 + z = 2$, or

$x = -z = -20$, so the solution is $\begin{pmatrix} -20 \\ -5.5 \\ 0 \end{pmatrix} + z\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$. Converting this into the normal form, $\begin{pmatrix} -20 \\ -5.5 \\ 0 \end{pmatrix} - \frac{20}{2}\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -10 \\ -5.5 \\ -10 \end{pmatrix}$, so the most economical representation is $\begin{pmatrix} -10 \\ -5.5 \\ -10 \end{pmatrix} + \alpha\begin{pmatrix} -\frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{pmatrix}$. You will note that $\left\|\begin{pmatrix} -20 \\ -5.5 \\ 0 \end{pmatrix}\right\|_2 \approx 20.74$ while $\left\|\begin{pmatrix} -10 \\ -5.5 \\ -10 \end{pmatrix}\right\|_2 \approx 15.17$.

7.5 The cross product

Given any two vectors in 3-space that are not parallel, these two define a plane, or a 2-dimensional subspace of 3-space. One goal may be to find a vector that is perpendicular to this plane. Thus, given two vectors $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$, we require a vector $\mathbf{w}$ such that $\langle \mathbf{u}, \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle = 0$, or, in other words,

$$\begin{aligned} u_1 w_1 + u_2 w_2 + u_3 w_3 &= 0 \\ v_1 w_1 + v_2 w_2 + v_3 w_3 &= 0 \end{aligned}.$$

As you can see, this defines a system of two equations and three unknowns, and so we may

$$\begin{aligned} u_1 w_1 \qquad + u_2 w_2 \qquad\qquad + u_3 w_3 &= 0 \\ \left(v_2 - \frac{u_2 v_1}{u_1} \right) w_2 + \left(v_3 - \frac{u_3 v_1}{u_1} \right) w_3 &= 0 \end{aligned}$$

We can multiply the denominator by u_1

$$\begin{aligned} u_1 w_1 \qquad\qquad + u_2 w_2 \qquad\qquad + u_3 w_3 &= 0 \\ \left(u_1 v_2 - u_2 v_1 \right) w_2 + \left(u_1 v_3 - u_3 v_1 \right) w_3 &= 0 \end{aligned}$$

If you look at the last line, it is of the form $\beta w_2 + \gamma w_3 = 0$, so if we let $w_2 = -\alpha\gamma$ and $w_3 = \alpha\beta$, then

$$\beta w_2 + \gamma w_3 = -\beta\alpha\gamma + \gamma\alpha\beta = -\alpha\beta\gamma + \alpha\beta\gamma = 0.$$

Thus, let us simply assume that

$$\begin{aligned} w_2 &= -\alpha \left(u_1 v_3 - u_3 v_1 \right) \\ &= \alpha u_3 v_1 - \alpha u_1 v_3 \\ w_3 &= \alpha \left(u_1 v_2 - u_2 v_1 \right) \\ &= \alpha u_1 v_2 - \alpha u_2 v_1 \end{aligned}$$

in which case, the second equation is satisfied. Whether we choose to negate the assignment of w_2 or w_3 is a choice we make that will ultimately affect the orientation, but for now, we will continue and substitute these values into the first equation to get

$$\begin{aligned} u_1 w_1 + u_2 \left(\alpha u_3 v_1 - u_1 v_3 \right) + u_3 \left(\alpha u_1 v_2 - u_2 v_1 \right) &= 0 \\ \cancel{u_1} w_1 + \cancel{\alpha v_1 u_2 u_3} - \alpha \cancel{u_1} u_2 v_3 + \alpha \cancel{u_1} v_2 u_3 - \cancel{\alpha v_1 u_2 u_3} &= 0 \end{aligned}$$

and therefore $w_1 = \alpha u_2 v_3 - \alpha v_2 u_3$. Because we can chose to be any real value, we will choose $\alpha = 1$, and therefore will define this to be the cross product of the two vectors **u** and **v**:

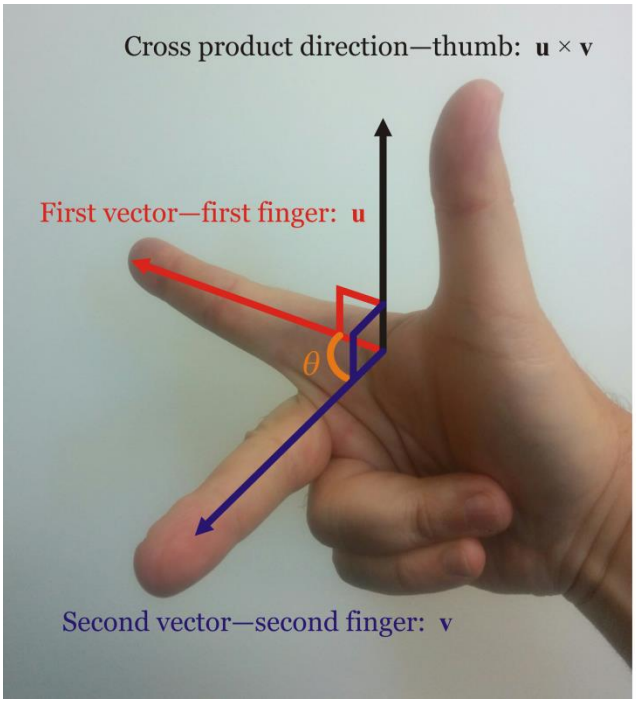
$$\mathbf{u}\times\mathbf{v}=\begin{pmatrix}u_2v_3-u_3v_2\\u_3v_1-u_1v_3\\u_1v_2-u_2v_1\end{pmatrix}.$$

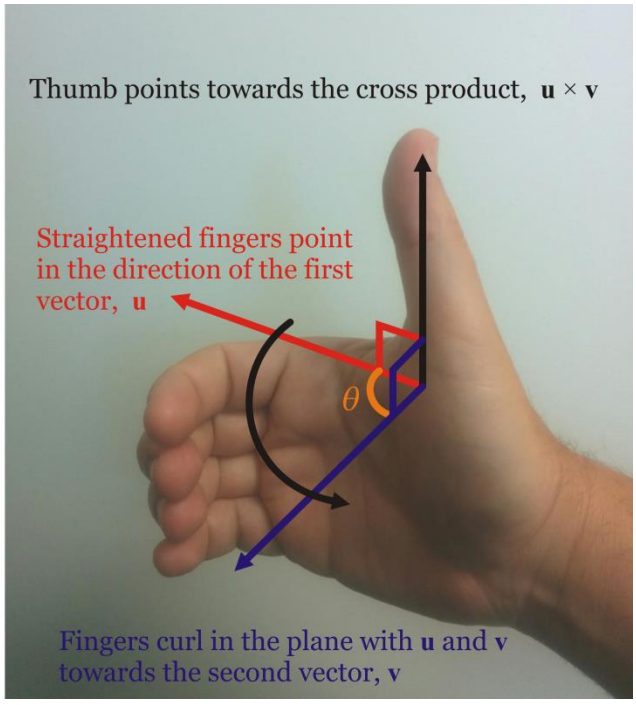
This leads us to what is used to describe the orientation of the cross product relative to the two vectors as the *right-hand rule*. The direction of the cross product can be determined relative to the two vectors by using the right hand where either

1. the first (or index) finger points in the direction of **u**,
2. the second (or middle) finger points in the direction of **v**, and
3. the thumb points in the direction of the cross product, $\mathbf{u}\times\mathbf{v}$,

or

1. the straightened fingers point in the direction of **u**,
2. the fingers curl in the direction of **v**, and
3. the cross product, $\mathbf{u}\times\mathbf{v}$, is in the direction of the thumb.





Had we chosen any other value of α , we would have found

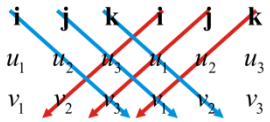
An easy way to memorize this is through the following mnemonic: define the unit vectors

$$\hat{\mathbf{i}}=\begin{pmatrix}1\\0\\0\end{pmatrix},\hat{\mathbf{j}}=\begin{pmatrix}0\\1\\0\end{pmatrix},\hat{\mathbf{k}}=\begin{pmatrix}0\\0\\1\end{pmatrix}$$

and thus create a matrix-like grid where

1. these three unit vectors appear in the first row twice,
2. the entries of **u** appear on in the second row twice, and
3. the entries of **v** appear in the third row twice.

Then, add the product of the diagonals in blue, and subtract from this the product of the reverse diagonals in red:



or

$$(u_2v_3-u_3v_2)\hat{\mathbf{i}}+(u_2v_3-u_3v_2)\hat{\mathbf{j}}+(u_2v_3-u_3v_2)\hat{\mathbf{k}}=\begin{pmatrix}u_2v_3-u_3v_2\\u_3v_1-u_1v_3\\u_1v_2-u_2v_1\end{pmatrix}.$$

Theorem

Given two 3-dimensinal vectors **u** and **v**, $\|\mathbf{u}\times\mathbf{v}\|_2=\|\mathbf{u}\|_2\|\mathbf{v}\|_2\sin(\theta)$ where θ is the angle between the two vectors.

Proof:

$$\begin{aligned}
\|\mathbf{u}\times\mathbf{v}\|_2^2 &= (u_2v_3-u_3v_2)^2+(u_3v_1-u_1v_3)^2+(u_1v_2-u_2v_1)^2\\
&= (u_2^2v_3^2-2u_2v_3u_3v_2+u_3^2v_2^2)+(u_1^2v_3^2-2u_1v_3u_3v_1+u_3^2v_1^2)+(u_1^2v_2^2-2u_1v_2u_2v_1+u_2^2v_1^2)\\
&= u_1^2v_1^2+u_1^2v_2^2+u_1^2v_3^2+u_2^2v_1^2+u_2^2v_2^2+u_2^2v_3^2+u_3^2v_1^2+u_3^2v_2^2+u_3^2v_3^2\\
&\quad -(u_1^2v_1^2+u_1v_1u_2v_2+u_1v_1u_3v_3+u_2v_2u_1v_1+u_2^2v_2^2+u_2v_2u_3v_3+u_3v_3u_1v_1+u_3v_3u_2v_2+u_3^2v_3^2)\\
&= (u_1^2+u_2^2+u_3^2)(v_1^2+v_2^2+v_3^2)-(u_1v_1+u_2v_2+u_3v_3)^2\\
&= \|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2-\langle\mathbf{u},\mathbf{v}\rangle^2\\
&= \|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2-\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2\cos^2(\theta)\\
&= \|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2(1-\cos^2(\theta))\\
&= \|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2\sin^2(\theta)
\end{aligned}$$

Taking the square root of both sides gives us our desired result. ∎

Demonstration of this theorem

If two vectors are perpendicular, the cross product is zero. If both **u** and **v** are unit vectors, then the magnitude of the cross product is $\sin(\theta)$. Suppose that

$$\mathbf{u} = \begin{pmatrix} 1 \\ -3 \\ 2 \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} -2 \\ 5 \\ -4 \end{pmatrix}.$$

In this case, $\mathbf{u} \times \mathbf{v} = \begin{pmatrix} 2 \\ 0 \\ -1 \end{pmatrix}$, and therefore the sine of the angle between the two vectors **u** and **v** is $\frac{\sqrt{5}}{\sqrt{14}\sqrt{45}} = \frac{\sqrt{14}}{42} \approx 0.089087$ and thus the angle is approximately 0.089205. You will recall that $\sin(\theta) \approx \theta$ for small θ. You can verify this with the inner product: $\langle \mathbf{u}, \mathbf{v} \rangle = -25$ and $\frac{25}{\sqrt{14}\sqrt{45}} = \frac{5\sqrt{5}\sqrt{14}}{42} \approx 0.996024$ and you will note that

$$\left(\frac{\sqrt{14}}{42}\right)^2 + \left(\frac{5\sqrt{5}\sqrt{14}}{42}\right)^2 = \frac{14 + 1750}{1764} = 1.$$

Theorem

The cross product is bilinear.

Proof:

$$(\alpha\mathbf{u}) \times \mathbf{v} = \begin{pmatrix} \alpha u_2 v_3 - \alpha u_3 v_2 \\ \alpha u_3 v_1 - \alpha u_1 v_3 \\ \alpha u_1 v_2 - \alpha u_2 v_1 \end{pmatrix} = \alpha \begin{pmatrix} u_2 v_3 - u_3 v_2 \\ u_3 v_1 - u_1 v_3 \\ u_1 v_2 - u_2 v_1 \end{pmatrix} = \alpha(\mathbf{u} \times \mathbf{v})$$

$$(\mathbf{u} + \mathbf{v}) \times \mathbf{w} = \begin{pmatrix} (u_2 + v_2) w_3 - (u_3 + v_3) w_2 \\ (u_3 + v_3) w_1 - (u_1 + v_1) w_3 \\ (u_1 + v_1) w_2 - (u_2 + v_2) w_1 \end{pmatrix} = \begin{pmatrix} u_2 w_3 - u_3 w_2 \\ u_3 w_1 - u_1 w_3 \\ u_1 w_2 - u_2 w_1 \end{pmatrix} + \begin{pmatrix} v_2 w_3 - v_3 w_2 \\ v_3 w_1 - v_1 w_3 \\ v_1 w_2 - v_2 w_1 \end{pmatrix} = (\mathbf{u} \times \mathbf{w}) + (\mathbf{v} \times \mathbf{w})$$

The proof that the cross product is linear in its second term is similar. ∎

Demonstration of this theorem

If you know that $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \times \mathbf{u} = \begin{pmatrix} 2 \\ -1 \\ 2 \end{pmatrix}$ and that $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \times \mathbf{u} = \begin{pmatrix} 1 \\ 3 \\ 4 \end{pmatrix}$, then $\begin{pmatrix} 3 \\ -2 \\ 0 \end{pmatrix} \times \mathbf{u} = 3\begin{pmatrix} 2 \\ -1 \\ 2 \end{pmatrix} - 2\begin{pmatrix} 1 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 4 \\ -9 \\ -2 \end{pmatrix}$.

Similarly, if you have determined that $\mathbf{u} \times \mathbf{v} = \begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix}$, then $(-2\mathbf{u}) \times (3\mathbf{v}) = (-2)3(\mathbf{u} \times \mathbf{v}) = -6\begin{pmatrix} 1 \\ -2 \\ 3 \end{pmatrix} = \begin{pmatrix} -6 \\ 12 \\ -18 \end{pmatrix}$

.

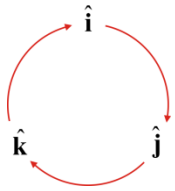
Theorem

The cross product is anti-symmetric, or $\mathbf{u}\times\mathbf{v}=-(\mathbf{v}\times\mathbf{u})$.

Proof:

$$\mathbf{u}\times\mathbf{v}=\begin{pmatrix}u_2v_3-u_3v_2\\u_3v_1-u_1v_3\\u_1v_2-u_2v_1\end{pmatrix}=\begin{pmatrix}-(-u_2v_3+u_3v_2)\\-(-u_3v_1+u_1v_3)\\-(-u_1v_2+u_2v_1)\end{pmatrix}=-\begin{pmatrix}u_3v_2-u_2v_3\\u_1v_3-u_3v_1\\u_2v_1-u_1v_2\end{pmatrix}=-(\mathbf{u}\times\mathbf{v})\,.\ \blacksquare$$

For the unit vectors, we have that $\hat{\mathbf{i}}\times\hat{\mathbf{j}}=\hat{\mathbf{k}}$, $\hat{\mathbf{j}}\times\hat{\mathbf{k}}=\hat{\mathbf{i}}$ and $\hat{\mathbf{k}}\times\hat{\mathbf{i}}=\hat{\mathbf{j}}$, and therefore also $\hat{\mathbf{j}}\times\hat{\mathbf{i}}=-\hat{\mathbf{k}}$, $\hat{\mathbf{k}}\times\hat{\mathbf{j}}=-\hat{\mathbf{i}}$ and $\hat{\mathbf{i}}\times\hat{\mathbf{k}}=-\hat{\mathbf{j}}$. This relationship is often summarized by the graphic



where if you follow the direction of the arrows, the cross product is the third, while if you go in the opposite direction, the cross product is the additive inverse of the third.

Demonstration of this theorem

For $\mathbf{u}=\begin{pmatrix}1\\-3\\2\end{pmatrix}$ and $\mathbf{v}=\begin{pmatrix}-2\\5\\-4\end{pmatrix}$, $\mathbf{u}\times\mathbf{v}=\begin{pmatrix}2\\0\\-1\end{pmatrix}$ while $\mathbf{v}\times\mathbf{u}=\begin{pmatrix}-2\\0\\1\end{pmatrix}$.

Theorem

The cross product is not associative, meaning $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})\neq(\mathbf{u}\times\mathbf{v})\times\mathbf{w}$ in general.

Proof:

For this proof, we need only one counterexample. Consider the three unit vectors:

$$\hat{\mathbf{i}}\times(\hat{\mathbf{i}}\times\hat{\mathbf{j}})=\hat{\mathbf{i}}\times\hat{\mathbf{k}}=-\hat{\mathbf{j}}\ \text{ but }\ (\hat{\mathbf{i}}\times\hat{\mathbf{i}})\times\hat{\mathbf{j}}=\mathbf{0}\times\hat{\mathbf{j}}=\mathbf{0}\,.\ \blacksquare$$

Note that the cross product is only applicable in three dimensions—there is no analogous definition of such a product in two or more than three dimensions; however, given $n-1$ n-dimensional vectors, one could find a vector perpendicular each of the $n-1$ vectors.

7.6 Applications of the cross product

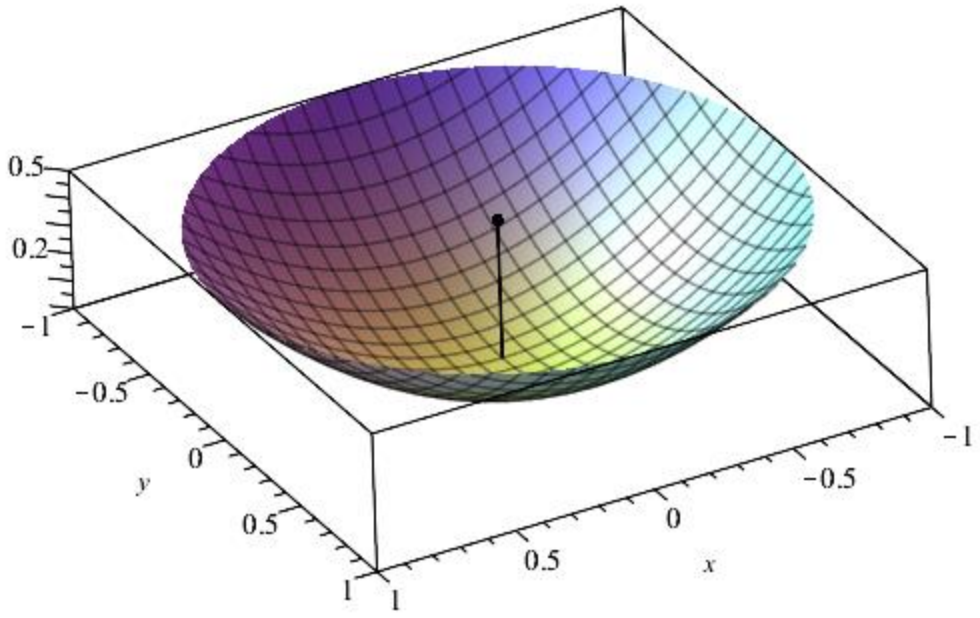
The force acting on a charged particle with a charge of q coulombs that is moving at a velocity **v** in the presence of an electric field of **E** newtons per coulomb (N/C or V/m) and a magnetic field **B** tesla (T, N/C/(m/s) or N/m/A) is given by

$$\mathbf{F} = q\left(\mathbf{E} + \mathbf{v} \times \mathbf{B}\right).$$

Suppose that the electric field at a point is given by $\mathbf{A} = 3\begin{pmatrix} \frac{2}{3} \\ \frac{1}{3} \\ -\frac{2}{3} \end{pmatrix}$ N/C , the velocity $\mathbf{v} = 0.01\begin{pmatrix} \frac{1}{9} \\ -\frac{8}{9} \\ \frac{4}{9} \end{pmatrix}$ m/s and the magnetic field $\mathbf{B} = 0.3\begin{pmatrix} \frac{6}{7} \\ \frac{3}{7} \\ -\frac{2}{7} \end{pmatrix}$ T . What is the force on a charge of q = 500 nC?

$$\mathbf{F} = q\left(3\begin{pmatrix} \frac{2}{3} \\ \frac{1}{3} \\ -\frac{2}{3} \end{pmatrix} + 0.01\begin{pmatrix} \frac{1}{9} \\ -\frac{8}{9} \\ \frac{4}{9} \end{pmatrix} \times 0.3\begin{pmatrix} \frac{6}{7} \\ \frac{3}{7} \\ -\frac{2}{7} \end{pmatrix}\right) \text{N} =$$

Another application in engineering is finding a *unit normal vector* to a surface. When modelling a surface such as the surface of a car body or a wing or a building, it is customary to divide the body into *patches*, where each patch is described by an appropriate vector-valued function of two variables. For example, one of the easiest non-trivial surfaces to model is that of a parabolic dish antenna. Suppose we have a parabolic dish that is two metres across and pointing directly to the sky where the feed antenna (the receiver) is at the focus is 50 cm above the surface . In this case, we may define the *patch* by mapping the unit circle to the coordinate $\mathbf{s}(x, y) = \begin{pmatrix} x \\ y \\ 0.5\left(x^2 + y^2\right) \end{pmatrix}$.



If you partially differentiate each entry of this vector with respect to *x* (that is, treating *y* as if it were a constant), we have the vector $\frac{\partial}{\partial x}\mathbf{s}(x,y)=\begin{pmatrix}1\\0\\x\end{pmatrix}$, and if we partially differentiate each entry with respect to *y*, we get the vector $\frac{\partial}{\partial y}\mathbf{s}(x,y)=\begin{pmatrix}0\\1\\y\end{pmatrix}$.[21] At any point, these two vectors are *tangent* to the surface, one in the direction of *x*, the other in the direction of *y*. Consequently, the cross product gives a vector that must be normal to the surface:

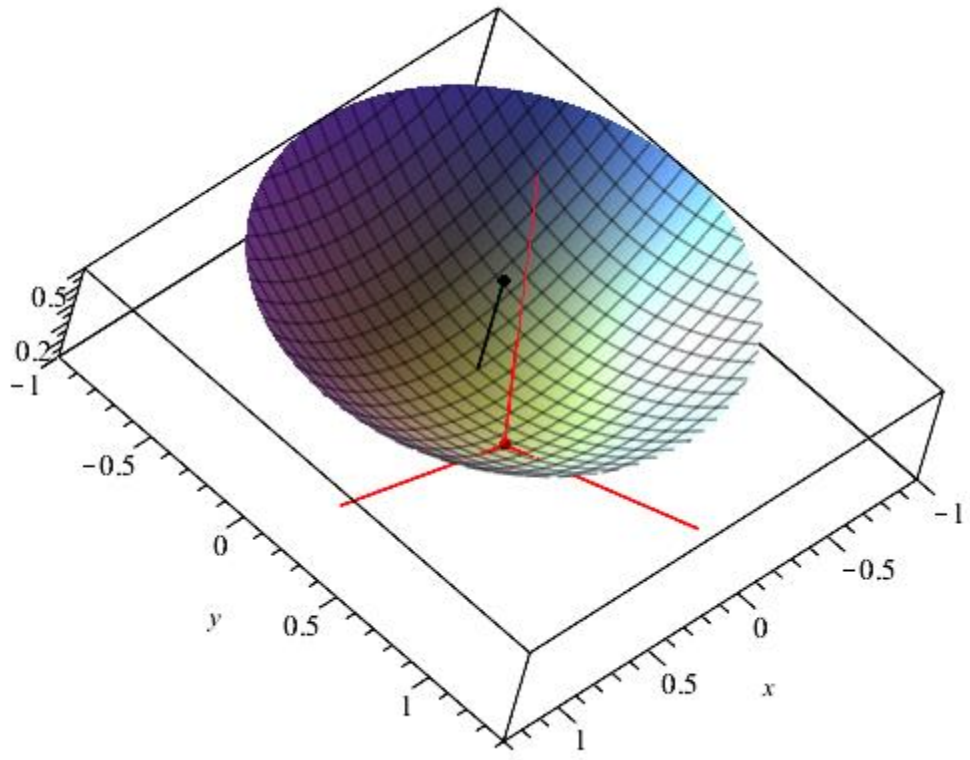
$$\begin{pmatrix}1\\0\\x\end{pmatrix}\times\begin{pmatrix}0\\1\\y\end{pmatrix}=\begin{pmatrix}-x\\-y\\1\end{pmatrix}.$$

Therefore, a vector perpendicular to the surface at the point $\begin{pmatrix}x\\y\\0.5\left(x^2+y^2\right)\end{pmatrix}$ is the vector connecting this point to the point $\begin{pmatrix}x\\y\\0.5\left(x^2+y^2\right)\end{pmatrix}+\begin{pmatrix}-x\\-y\\1\end{pmatrix}=\begin{pmatrix}0\\0\\0.5\left(x^2+y^2\right)+1\end{pmatrix}$. The normal vector is not *normalized* and therefore a *unit normal vector* would be

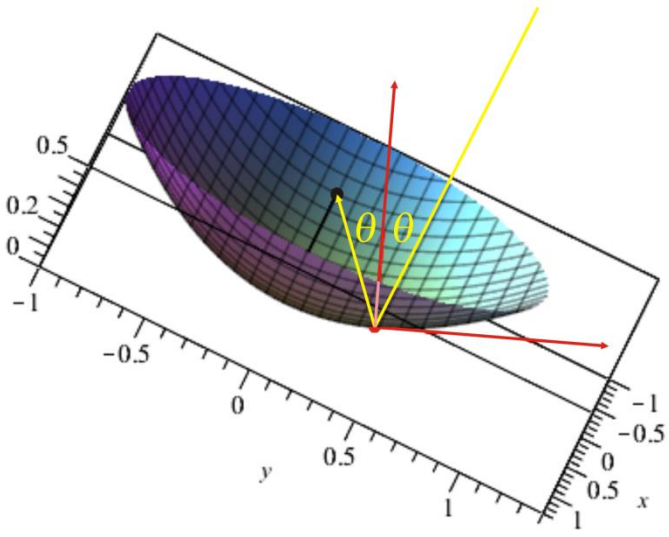
[21] You will learn the finer points and more rigorous definition of partial differentiation in your calculus courses.

$$\begin{pmatrix} -\dfrac{x}{\sqrt{x^2+y^2+1}} \\ -\dfrac{y}{\sqrt{x^2+y^2+1}} \\ \dfrac{1}{\sqrt{x^2+y^2+1}} \end{pmatrix}$$

For example, at the point on the surface $\mathbf{p} = \begin{pmatrix} 0.3 \\ 0.4 \\ 0.125 \end{pmatrix}$, the tangent vectors to the surface of the parabolic dish are $\mathbf{t}_1 = \begin{pmatrix} 1 \\ 0 \\ 0.3 \end{pmatrix}$ and $\mathbf{t}_2 = \begin{pmatrix} 0 \\ 1 \\ 0.4 \end{pmatrix}$ and a vector normal to the surface of the parabola is $\mathbf{n} = \mathbf{t}_1 \times \mathbf{t}_2 = \begin{pmatrix} -0.3 \\ -0.4 \\ 1 \end{pmatrix}$. This is shown in .



Visually, we can see it appears that an incoming photon will be reflected to the focus, as shown in .



We can, however, test this mathematically, as well. The normal vector is

$$\mathbf{n} = \begin{pmatrix} -0.3 \\ -0.4 \\ 1 \end{pmatrix}$$

and the angle between this vector and an incoming vector $\mathbf{v}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ is

$$\cos^{-1}\left(\frac{\langle \mathbf{n}, \mathbf{v}_1 \rangle}{\|\mathbf{n}\|_2 \|\mathbf{v}_1\|_2} \right) = \cos^{-1}\left(\frac{1}{\sqrt{1.25}} \right).$$

The vector from the point $\mathbf{p} = \begin{pmatrix} 0.3 \\ 0.4 \\ 0.125 \end{pmatrix}$ to the focus at $\mathbf{f} = \begin{pmatrix} 0 \\ 0 \\ 0.5 \end{pmatrix}$, so the vector from **p** to **f** is the vector

$$\mathbf{f} - \mathbf{p} = \begin{pmatrix} 0 \\ 0 \\ 0.5 \end{pmatrix} - \begin{pmatrix} 0.3 \\ 0.4 \\ 0.125 \end{pmatrix} = \begin{pmatrix} -0.3 \\ -0.4 \\ 0.375 \end{pmatrix}$$

and the angle between the normal and this vector is

$$\cos^{-1}\left(\frac{\langle \mathbf{n}, \mathbf{f} - \mathbf{p} \rangle}{\|\mathbf{n}\|_2 \|\mathbf{f} - \mathbf{p}\|_2} \right) = \cos^{-1}\left(\frac{0.625}{\sqrt{1.25} \cdot 0.625} \right) = \cos^{-1}\left(\frac{1}{\sqrt{1.25}} \right).$$

Consequently, without evaluating the inverse cosine functions, we see that the two angles are equal; however, for interest, the angle is approximately $\theta \approx 26.565°$.

To understand the strength of the electromagnetic force, it is necessary to consider the two equations:

$$F = G\frac{m_1 m_2}{r^2} \text{ and } F = k_e \frac{q_1 q_2}{r^2}$$

where $G = 6.674 \times 10^{-11}$ Nm^2/kg^2 and $k_e = 8.988 \times 10^9$ Nm^2/C^{-2}. There are twenty orders of magnitude between these two coefficients, but that doesn't tell the entire story, for one is measured in kilograms while the other is measure in coulombs. To understand the actual difference, let us find the gravitational force and electrostatic force between two electrons and two protons.

$$F = k_e \frac{ee}{r^2} = 8.988 \times 10^9 \frac{\left(1.602 \times 10^{-19}\right)^2}{r^2} = \frac{2.307 \times 10^{-28}}{r^2}$$

while

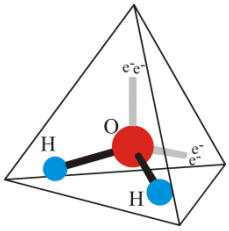
$$F = G\frac{m_e m_e}{r^2} = 6.674 \times 10^{-11} \frac{\left(9.109 \times 10^{-31}\right)^2}{r^2} = \frac{5.538 \times 10^{-71}}{r^2}$$

while

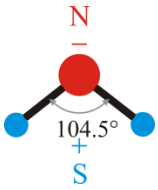
$$F = G\frac{m_p m_p}{r^2} = 6.674 \times 10^{-11} \frac{\left(1.673 \times 10^{-27}\right)^2}{r^2} = \frac{1.868 \times 10^{-64}}{r^2}.$$

Consequently, the electromagnetic repulsion between two protons is 38 orders of magnitude larger than the corresponding gravitational force, and for electrons it is 45 orders of magnitude greater. To understand how much this this is, 38 orders of magnitude translates to the difference between the mass of five average human cells and the mass of the Earth. Equivalently, 38 orders of magnitude is the difference between the average volume of a drop from an inkjet printer in comparison to the volume of the Earth. It is simply impossible to

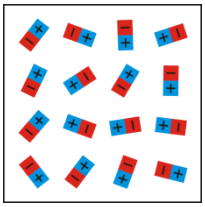
A stationary electron or static charge is not affected by a magnetic field. To understand this, we will quickly give an introduction to magnetism. To begin, we must understand the concept of a *dipole*. The easiest way to understand this is to visualize the water molecule H_2O. Each hydrogen atom shares an electron with the oxygen molecule, and with the two shared electrons from the hydrogen and the six from the outer shell of oxygen, they naturally try to arrange themselves in approximately a tetrahedral shape.



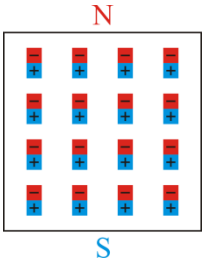
Due to other factors, the angle is slightly greater, at approximately 104.5°. In chemistry, you will learn that the oxygen atom attracts the electrons slightly more than the hydrogen atoms. Consequently, in the direction of the oxygen atom, the electron cloud is slightly denser. This creates a *dipole* or miniature magnet.



but the oxygen atom more strongly attracts the electrons. It is the attraction between the slightly more negative oxygen and the more positive hydrogen atoms that creates the beautiful lattices found in ice and why ice is less dense than liquid water. If all the dipoles lined up, these small dipoles would amplify each other; however, this is generally not the case, and so ice does not form a magnet.



If, however, it is possible to line up each molecule, it is possible for the dipoles to reinforce each other, thus producing a magnet:



Freezing water in an electric field will produce a very weak magnet, but not one that is useful. To understand how iron magnets work, we must now look at some more fundamental properties of atomic particles. Protons, neutrons and electrons all have magnetic moments, just like electrons and protons have charge. Pairs of electrons within a

Due to quantum effects caused by unpaired electrons in iron, each iron atom is a dipole and under certain circumstances these can be lined up to form a magnet. For example, this alignment occurs naturally in *lodestone* due in part to the crystalline structure of Fe_3O_4.

Your first thought may be that an electron can now be attracted to the south end of a magnet, but because the positive and negative ends of each dipoles are so close to each other, from a distance, the magnet appears electrically neutral:

a stationary electron in a magnetic field has no forces acting on it. Thus, even the attractive force of two magnets is very weak relative to the electromagnetic force. Magnets, however, will induce a torque on each other so that their magnetic fields align.

Given two magnets, however, there will be a tendency for the two magnets to line up with each other with the more negative

The force on an electron moving in a magnetic field is given by $q(\mathbf{v}\times\mathbf{B})$ where q is the charge of the electron, **v** is the velocity in m/s and **B** is the strength of the magnetic field in tesla (T). From this, it may be deduced that a tesla is $\mathrm{T}=\frac{\mathrm{N}}{\mathrm{Cm}/s}$. The magnetic field strength on the surface of the Earth is between 25 μT and 65 μT. For this example, we will use 50 μT. Now, in an alternating current, the voltage is v = 170 cos(377 t) V. The root-mean-square (RMS) of this alternating voltage is $\frac{170}{\sqrt{2}}\mathrm{V}\approx 120\,\mathrm{V}$. Thus, if we have a light bulb with an rms of 60 W of power, we have from

$$V = IR$$
$$P = I^2R$$

Substituting the first into the second yields $P=IV$. Thus, the rms current is $0.5\mathrm{A}$, and so the current may be described as i = 0.707 cos(377 t) A. Now, the density of copper is 8.96 g/cm^3 and the atomic weight of copper is 63.546 g/mol, and as each copper atom has one free electron, we therefore have 0.141 mol/cm^3. Now, a mole of charge equals 96485.33289 C (this is the Faraday constant). Thus, copper has 13604 C/cm^3 of free electrons. If the width of a wire is 1 mm, we must find the length of wire that holds 0.5 C, and thus the length will indicate the speed of the electrons. Thus, we have that

$$h\pi 0.05^2\cdot 13604 = 0.5\,,$$

and solving this yields h = 0.00468 cm or h = 46.8 μm, so the electrons flow at 46.8 μm/s. Consequently, the actual speed may be described as

$$v = -66 \times 10^{-6} \cos(377\,t) \text{ m/s}$$

as electron flow is opposite to current flow.

Consequently, the magnitude of a force one electron is $5.3 \times 10^{-28} \sin(\theta)$ N where θ is the angle between the magnetic field vector and the velocity vector. As force is mass times acceleration, and the mass of an electron is 9.10938356(11)×10−31

7.7 Finding the plane containing three points

The equation of a plane containing three points $\mathbf{u}_1$, $\mathbf{u}_2$ and $\mathbf{u}_3$ may be found as follows:

$$\mathbf{v} = (\mathbf{u}_1 - \mathbf{u}_3)\times(\mathbf{u}_2 - \mathbf{u}_3)\,.$$

If this cross product is **0**, the points are either collinear (on the same line) or coincident (all the same point).

This vector **v** is perpendicular to the plane, and thus, the equation of the plane are all points that satisfy

$$\langle \mathbf{x}-\mathbf{u}_1, \mathbf{v} \rangle = 0 \text{ or } \langle \mathbf{x}, \mathbf{v} \rangle = \langle \mathbf{u}_1, \mathbf{v} \rangle,$$

although we could pick any of the three points, $\mathbf{u}_1$, $\mathbf{u}_2$ or $\mathbf{u}_3$ in this calculation, and the result should be always the same.

For example, the plane that passes through the three points $\begin{pmatrix} -2 \\ 4 \\ 3 \end{pmatrix}$, $\begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} -4 \\ 2 \\ -1 \end{pmatrix}$ is found by

$$\mathbf{v} = \left(\begin{pmatrix} -2 \\ 4 \\ 3 \end{pmatrix} - \begin{pmatrix} -4 \\ 2 \\ -1 \end{pmatrix} \right) \times \left(\begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} -4 \\ 2 \\ -1 \end{pmatrix} \right) = \begin{pmatrix} 2 \\ 2 \\ 4 \end{pmatrix} \times \begin{pmatrix} 7 \\ -2 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 24 \\ -18 \end{pmatrix}.$$

For simplicity, we can chose any scalar multiple, and thus we will choose $\mathbf{v} = \begin{pmatrix} 2 \\ 4 \\ -3 \end{pmatrix}$, and thus the equation of the plane are all points

$$2x + 4y - 3z = 3$$

as $\left\langle \begin{pmatrix} 2 \\ 4 \\ -3 \end{pmatrix}, \begin{pmatrix} 3 \\ 0 \\ 1 \end{pmatrix} \right\rangle = 6 - 3 = 3$. If we substitute the three points into this equation, we find that they all confirm that these points are on the plane.

Suppose we wish to write a plane in the form

$$\mathbf{u} + \alpha \mathbf{v}_1 + \beta \mathbf{v}_2 \ .$$

In this case, we may first wish to find that **u** that is smallest in magnitude; that is, we want to find that point in the plane that is closest to the origin. Thus, we want to minimize

$$\| \mathbf{u} + \alpha \mathbf{v} + \beta \mathbf{w} \|_2$$

which is equivalent to minimizing

$$\begin{aligned} \langle \mathbf{u} + \alpha \mathbf{v}_1 + \beta \mathbf{v}_2, \mathbf{u} + \alpha \mathbf{v}_1 + \beta \mathbf{v}_2 \rangle &= \langle \mathbf{u}, \mathbf{u} \rangle + \alpha \langle \mathbf{u}, \mathbf{v}_1 \rangle + \beta \langle \mathbf{u}, \mathbf{v}_2 \rangle + \alpha \langle \mathbf{v}_1, \mathbf{u} \rangle + \alpha^2 \langle \mathbf{v}_1, \mathbf{v}_1 \rangle + \alpha\beta \langle \mathbf{v}_1, \mathbf{v}_2 \rangle \\ &\quad + \beta \langle \mathbf{v}_2, \mathbf{u} \rangle + \alpha\beta \langle \mathbf{v}_2, \mathbf{v}_1 \rangle + \beta^2 \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \\ &= \langle \mathbf{u}, \mathbf{u} \rangle + 2\alpha \langle \mathbf{v}_1, \mathbf{u} \rangle + 2\beta \langle \mathbf{v}_2 \mathbf{u} \rangle + 2\alpha\beta \langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \alpha^2 \langle \mathbf{v}_1, \mathbf{v}_1 \rangle + \beta^2 \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \end{aligned}$$

In differential calculus, you may have learned already that minimizing the an equation in one variable *x* is equivalent to differentiating with respect to that variable *x* and solving for the derivative equalling zero. In this case, we must differentiate first with respect to α, and then a second time with respect to β:

$$\frac{\partial}{\partial\alpha}\left(\langle\mathbf{u},\mathbf{u}\rangle+2\left(\alpha\langle\mathbf{u},\mathbf{v}_1\rangle+\beta\langle\mathbf{u},\mathbf{v}_2\rangle+\alpha\beta\langle\mathbf{v}_1,\mathbf{v}_2\rangle\right)+\alpha^2\langle\mathbf{v}_1,\mathbf{v}_1\rangle+\beta^2\langle\mathbf{v}_2,\mathbf{v}_2\rangle\right)$$
$$=2\langle\mathbf{u},\mathbf{v}_1\rangle+2\beta\langle\mathbf{v}_1,\mathbf{v}_2\rangle+2\alpha\langle\mathbf{v}_1,\mathbf{v}_1\rangle=0$$
$$\frac{\partial}{\partial\beta}\left(\langle\mathbf{u},\mathbf{u}\rangle+2\alpha\langle\mathbf{u},\mathbf{v}_1\rangle+2\beta\langle\mathbf{u},\mathbf{v}_2\rangle+2\alpha\beta\langle\mathbf{v}_1,\mathbf{v}_2\rangle+\alpha^2\langle\mathbf{v}_1,\mathbf{v}_1\rangle+\beta^2\langle\mathbf{v}_2,\mathbf{v}_2\rangle\right)$$
$$=2\langle\mathbf{u},\mathbf{v}_2\rangle+2\alpha\langle\mathbf{v}_1,\mathbf{v}_2\rangle+2\beta\langle\mathbf{v}_2,\mathbf{v}_2\rangle=0$$

You will note that this is a system of two equations and two unknowns:

$$\alpha\langle\mathbf{v}_1,\mathbf{v}_1\rangle+\beta\langle\mathbf{v}_1,\mathbf{v}_2\rangle=-\langle\mathbf{u},\mathbf{v}_1\rangle$$
$$\alpha\langle\mathbf{v}_1,\mathbf{v}_2\rangle+\beta\langle\mathbf{v}_2,\mathbf{v}_2\rangle=-\langle\mathbf{u},\mathbf{v}_2\rangle$$

If the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal, this simplifies to

$$\alpha\langle\mathbf{v}_1,\mathbf{v}_1\rangle=-\langle\mathbf{v}_1,\mathbf{u}\rangle$$
$$\beta\langle\mathbf{v}_2,\mathbf{v}_2\rangle=-\langle\mathbf{v}_2,\mathbf{u}\rangle,$$

so

$$\alpha=-\frac{\langle\mathbf{v}_1,\mathbf{u}\rangle}{\langle\mathbf{v}_1,\mathbf{v}_1\rangle}$$
$$\beta=-\frac{\langle\mathbf{v}_2,\mathbf{u}\rangle}{\langle\mathbf{v}_2,\mathbf{v}_2\rangle}$$

In other words, the optimal choice of vector **u** is

$$\mathbf{u}' = \mathbf{u} - \frac{\langle \mathbf{v}_1, \mathbf{u} \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 - \frac{\langle \mathbf{v}_2, \mathbf{u} \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle} \mathbf{v}_2 ,$$

or

$$\mathbf{u}' = \mathbf{u} - \text{proj}_{\mathbf{v}_1} (\mathbf{u}) - \text{proj}_{\mathbf{v}_2} (\mathbf{u}) .$$

If the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are not orthogonal, it becomes more difficult, but as long as the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are not parallel,

$$\mathbf{u}' = \mathbf{u} - \frac{\langle \mathbf{v}_1, \mathbf{u} \rangle \langle \mathbf{v}_2, \mathbf{v}_2 \rangle - \langle \mathbf{v}_2, \mathbf{u} \rangle \langle \mathbf{v}_1, \mathbf{v}_2 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle \langle \mathbf{v}_2, \mathbf{v}_2 \rangle - \langle \mathbf{v}_1, \mathbf{v}_2 \rangle^2} \mathbf{v}_1 - \frac{\langle \mathbf{v}_2, \mathbf{u} \rangle \langle \mathbf{v}_1, \mathbf{v}_1 \rangle - \langle \mathbf{v}_1, \mathbf{u} \rangle \langle \mathbf{v}_1, \mathbf{v}_2 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle \langle \mathbf{v}_2, \mathbf{v}_2 \rangle - \langle \mathbf{v}_1, \mathbf{v}_2 \rangle^2} \mathbf{v}_2 .$$

In either case, we may now write the plane as all combinations of vectors in the form

$$\mathbf{u}' + \alpha \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|_2} + \beta \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|_2} ,$$

where preferably, but not necessarily the vectors are orthonormal. We can always apply Gram-Schmidt to make them orthonormal.

8 Linear maps and linear operators

We have seen many different types of vectors: finite-dimensional vectors, discrete signals, and functions. In engineering, any signal, be it voltage or intensity or sound, or global positioning system (GPS) coordinates, be either discrete or continuous, and therefore may be represented by a vector. In engineering, a significant component is taking such a signal and generating, modifying, or extracting information from a signal. We shall describe any such process as an *operation* on the signal, and we can use a block diagram to show this, as shown in Figure 67.



Figure 67. The mapping of a vector.

Consider the following three examples: Given an input vector, we may wish to

1. calculate the average value,
2. reduce the noise within the vector, and
3. increase the resolution (displaying an analog TV signal on a 4K television).

In the first case, the vector we are working on may be finite-dimensional, discrete signals or continuous functions, but the output is a scalar value; that is, a value in a 1-dimensional vector space. In the second, the output is likely the same space as the input, and in the third, the output has a higher dimension than the input. Therefore, a mapping is a transformation from one vector space to another, as shown in Figure 68. Engineers often refer to such a mapping as a *system*.



Figure 68. A *system*; that is, a mapping or transformation *A* from one vector space, *U* to another, *V*.

Suppose we have two vector spaces $\mathcal{U}$ and $\mathcal{V}$, and we have a such a mapping $A:\mathcal{U}\to\mathcal{V}$. We will say that a vector $\mathbf{u}\in\mathcal{U}$ is *mapped* onto the vector $A\mathbf{u}\in\mathcal{V}$. Occasionally, we may also represent this mapping as $A(\mathbf{u})$, especially when the mapping is applied to a linear combination of vectors (e.g., $A(\alpha\mathbf{u}_1 + \beta\mathbf{u}_2)$). We have three definitions:

Definition
Given an map $A:\mathcal{U}\to\mathcal{V}$, $\mathcal{U}$ is referred to as the *domain* while $\mathcal{V}$ is referred to as the *codomain*.

If the domain and codomain are the same, we will refer to the mapping as an *operator*:

Definition
Given an map $A:\mathcal{U}\to\mathcal{U}$, where the domain and co-domain are the same, we will describe the mapping as an *operator*.

Definition
We will say that $A\mathbf{u}$ is the *image* of the vector $\mathbf{u}$ of the mapping A, and we will say that $\mathbf{u}$ is a *pre-image* of $A\mathbf{u}$ under the mapping A.

Examples of maps and operators

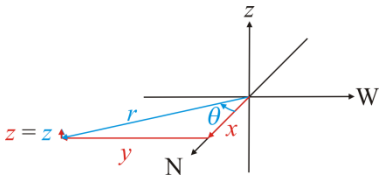
Here are two common operators between vector spaces that you will use in calculus and modeling systems.

1. Suppose you are in a vehicle and you have the GPS coordinates of a location relative to your own. The three-dimensional vector $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ indicates how many metres north, how many metres west, and how many metres up the location is relative to your own vehicle, with negative values indicating a southerly, easterly or downwards offsets. Assuming you are in an all-terrain vehicle, you would prefer to simple know which direction to travel and how far to travel in that direction (ignoring any change in elevation). We can write this description of the location as:
   a. the distance in the *xy* plane (the *radius*),
   b. the direction relative to north to the location (the *azimuth*), and
   c. the change in elevation

   using the mapping

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} \sqrt{x^2 + y^2} \\ \tan^{-1}\left(\frac{y}{x}\right) \\ z \end{pmatrix}.$$

   This new vector describes the location relative to your own using *cylindrical coordinates*, and the three coordinates are referred to as r, θ and z. For example, if the offset was given as $\begin{pmatrix} 179 \text{ m} \\ -452 \text{ m} \\ 19 \text{ m} \end{pmatrix}$, we could find the image under this operation as $\begin{pmatrix} 486.2 \text{ m} \\ -1.19 \text{ rad} \\ 19 \text{ m} \end{pmatrix}$. This would tell us that we would have to proceed at an angle of approximately 68.4 degrees east of north for a distance of 486 m. Visually, we may interpret these two coordinate systems as shown in, with the rectangular coordinates of 179 m, –452 m and 19 m shown in red, while the cylindrical coordinates including the radius 486.2 m, an angle of 1.19 radians east of north, and an unchanged 19 m up shown in blue.

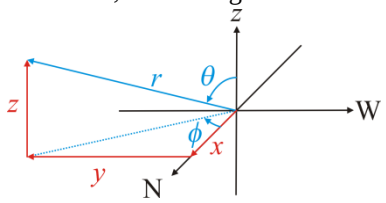


2. Suppose, instead, you are controlling a drone and you have the GPS coordinates of a location relative to your drone. In this case, you may want to travel to that location in a straight line. For this, you must know:
    a. the distance (the *radius*),
    b. the direction relative to north to the location (the *azimuth*), and
    c. the angle relative to the *xy*-plane that you must travel (the *inclination*)

    using the mapping

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} \sqrt{x^2+y^2+z^2} \\ \tan^{-1}\left(\frac{y}{x}\right) \\ \cos^{-1}\left(\frac{z}{\sqrt{x^2+y^2+z^2}}\right) \end{pmatrix}$$

    This new vector describes the location relative to your own using *spherical coordinates*, and the three coordinates are referred to as r, θ and ϕ. For example, if the offset was given as $\begin{pmatrix} 179\text{ m} \\ -452\text{ m} \\ 267\text{ m} \end{pmatrix}$, we could find the image under this operation as $\begin{pmatrix} 562.5\text{ m} \\ -1.19\text{ rad} \\ 1.07\text{ rad} \end{pmatrix}$. This would tell us that we would have to proceed at an angle of approximately 68.4 degrees east of north, at an angle of 61 degrees down from straight up for a distance of 562 m. Visually, we may interpret these two coordinate systems as shown in, with the rectangular coordinates of 179 m, –452 m and 267 m shown in red, while the spherical coordinates including the radius 486.2 m, an angle of 1.19 radians east of north, and an angle of 1.07 radians from straight up shown in blue.



Two other mappings of a 4-dimensional vector onto a one-dimensional vector include

1. the average $\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} \mapsto \left(\frac{1}{4}\sum_{k=1}^{4} u_k\right)$, and
2. the standard deviation $\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} \mapsto \left(\sqrt{\frac{1}{4}\sum_{k=1}^{4}\left(u_k - \bar{u}\right)^2}\right)$.

8.1 The superposition principle

The *superposition property* is an observation from physics that there are many systems where the net response at a given place and time caused by two or more stimuli is the sum of the responses which would have been caused by each stimulus individually.[22] For example, suppose you throw two stones into a pool—while this example uses water waves, the same holds for electromagnetic waves—after which the waves begin to collide. When they waves collide with each other, waves will either constructively or destructively interfere; however, that interference will be additive.

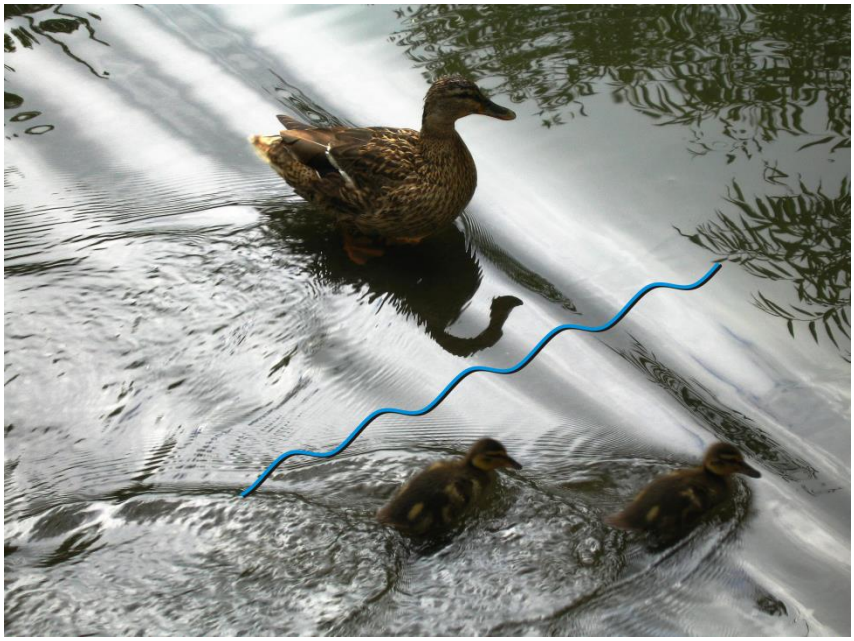


Figure 69. Demonstration of the superposition of water waves. Photograph by Flickr user Spiralz.

The technology in noise-cancelling headphones works similarly. Normally, when you wear a set of headphones, the dominant sound you hear is the music generated by the speaker, but no headphones are perfectly insulated from ambient noise (traffic, conversations, etc.). For example, suppose you are listening to Middle C (261.6 Hz) being played for 100 ms. The you hear is the generated sound and the noise superimposed, as shown in Figure 70.

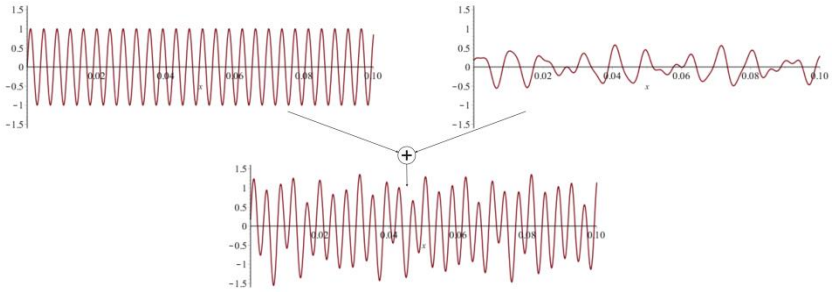


Figure 70. Middle C superimposed with lower frequency noise.

A noise-cancelling headphone has a microphone that listens to the ambient noise, negates it and adds that negated noise to the music being generated. Now, when the generated music superimposed with the actual ambient noise, the noises cancel each other leaving the music as it was intended to be heard, as is shown in Figure 71.

[22] Definition from Wikipedia.

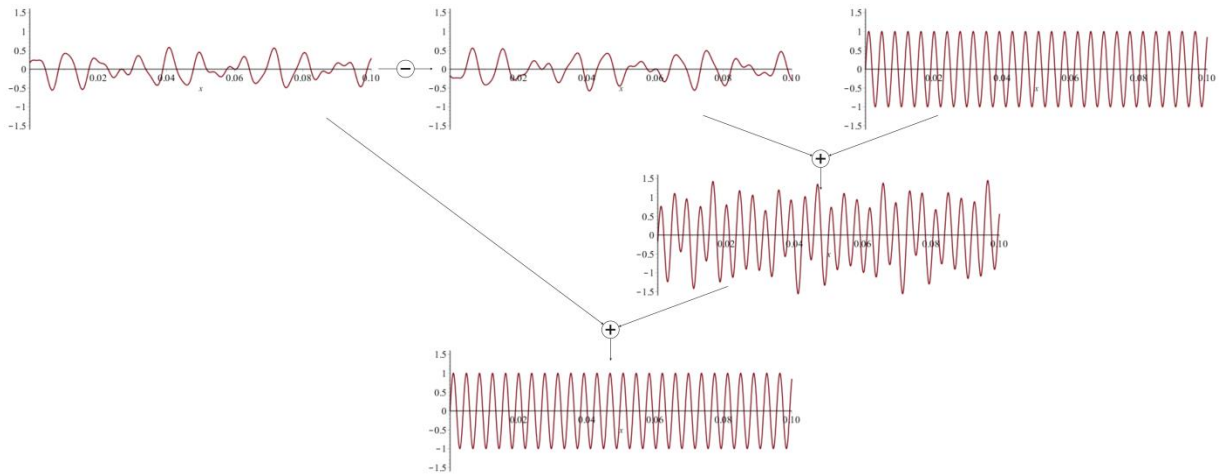


Figure 71. Sound cancellation of the noise introduced in Figure 70.

If sound did not have the superposition property, such sound cancellation would be significantly more difficult to perform.

Forces also obey the superposition principle: the force on an object is equal to the sum of the forces

$$\begin{aligned} F_{\text{gravitational}} &= m_0 \mathbf{a}_k = \sum_{k=1}^{n} G \frac{m_0 m_k}{\left\| \mathbf{x}_0 - \mathbf{x}_k \right\|_2^3} \left(\mathbf{x}_0 - \mathbf{x}_k \right) \\ &= \left(G m_0 \sum_{k=1}^{n} \frac{m_k}{\left\| \mathbf{x}_0 - \mathbf{x}_k \right\|_2^3} \right) \mathbf{x}_0 - G m_0 \sum_{k=1}^{n} \frac{m_k \mathbf{x}_k}{\left\| \mathbf{x}_0 - \mathbf{x}_k \right\|_2^3} \end{aligned}$$

and

$$\begin{aligned} F_{\text{electric}} &= m_0 \mathbf{a}_k = \sum_{k=1}^{n} k_e \frac{q_0 q_k}{\left\| \mathbf{x}_0 - \mathbf{x}_k \right\|_2^3} \left(\mathbf{x}_0 - \mathbf{x}_k \right) \\ &= \left(k_e q_0 \sum_{k=1}^{n} \frac{q_k}{\left\| \mathbf{x}_0 - \mathbf{x}_k \right\|_2^3} \right) \mathbf{x}_0 - k_e q_0 \sum_{k=1}^{n} \frac{q_k \mathbf{x}_k}{\left\| \mathbf{x}_0 - \mathbf{x}_k \right\|_2^3} \end{aligned}$$

Similarly, the force on a particle with electric charge q_0 and velocity $\mathbf{v}$ in the presence of numerous electric fields $\mathbf{E}_k$ and magnetic fields $\mathbf{B}_k$ also obeys the superposition principle:

$$\begin{aligned} F_{\text{electric}} &= q_0 \left(\sum_{k=1}^{n_{\text{electric}}} \mathbf{E}_k + \mathbf{v} \times \sum_{k=1}^{n_{\text{magnetic}}} \mathbf{B}_k \right) \\ &= \sum_{k=1}^{n_{\text{electric}}} q_0 \mathbf{E}_k + \sum_{k=1}^{n_{\text{magnetic}}} q_0 \mathbf{v} \times \mathbf{B}_k \end{aligned}$$

Consequently, being able to model systems that obey, at least approximately, the superposition principle is one of the primary goals of physicists and engineers, and thus, we will define *linear* maps.

8.2 Definition of linear maps and operators

We will now define the properties of operators that ensure that they satisfy the superposition property, or in the terminology of mathematics, that they satisfy *linearity*. First, let us look at some operators between vector spaces:

1. $A:\mathbf{R}^2 \to \mathbf{R}^3$ defined by $A\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} u_1 + u_2 \\ u_1 + 1 \\ u_1 - u_2 \end{pmatrix}$,

2. $B:\mathbf{R}^2 \to \mathbf{R}^4$ defined by $B\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} u_1^3 \\ 3u_1^2 u_2 \\ 3u_1 u_2^2 \\ u_2^3 \end{pmatrix}$,

3. $C:\mathbf{R}^2 \to \mathbf{R}^3$ defined by $C\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \cos(u_1)\cos(u_2) \\ \cos(u_1)\sin(u_2) \\ \sin(u_1) \end{pmatrix}$,

4. $D:\mathbf{R}^3 \to \mathbf{R}^2$ defined by $D\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} u_2 \\ u_3 \end{pmatrix}$, and

5. $E:\mathbf{R}^3 \to \mathbf{R}^3$ defined by $E\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 2u_1 - 6u_2 + u_3 \\ 4u_1 + 2u_2 + 5u_3 \\ -u_1 - u_2 - 3u_3 \end{pmatrix}$.

In each case, all vectors in the domain are mapped onto some vector in the range. When an engineer works with a system that operates on an input signal, there are some desirable properties:

Suppose we have a system A and we amplify or attenuate an input $\mathbf{u}$ by a scalar α, that is, we calculate the output $A(\alpha\mathbf{u})$, then the response of the system should be identical if we simply determine the output $A\mathbf{u}$ and multiply that output by α. Diagrammatically, this is shown in

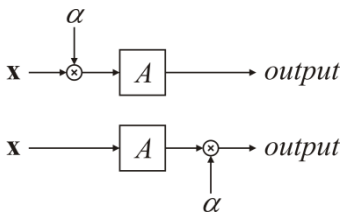


Figure 72. $A(\alpha\mathbf{u}) = \alpha A(\mathbf{u})$.

We now have the theorem of interest.

Theorem

If an operator $A:\mathcal{U} \to \mathcal{V}$ satisfies the property that $A(\alpha\mathbf{u}) = \alpha A(\mathbf{u})$, then $A(\mathbf{0}_U) = \mathbf{0}_V$.

Proof:

Recall that if $A:\mathcal{U} \to \mathcal{V}$, then $\mathbf{u} \in \mathcal{U}$ and $A\mathbf{u} \in \mathcal{V}$. Because $\mathcal{U}$ is a vector space, $0\mathbf{u} = \mathbf{0}$ for any vector $\mathbf{u} \in \mathcal{U}$. Thus, if this property is satisfied, $A(\mathbf{0}_u) = A(0\mathbf{u}) = 0A\mathbf{u}$, but $A\mathbf{u} \in \mathcal{V}$ and so $0A\mathbf{u} = \mathbf{0}_v$. Therefore, $A(\mathbf{0}_u) = \mathbf{0}_v$. ∎

If we consider our five mappings above, we note that

$$A(\mathbf{0})=\begin{pmatrix}0\\1\\0\end{pmatrix},\ B(\mathbf{0})=\begin{pmatrix}0\\0\\0\\0\end{pmatrix},\ C(\mathbf{0})=\begin{pmatrix}1\\0\\0\end{pmatrix},\ D(\mathbf{0})=\begin{pmatrix}0\\0\end{pmatrix}=\mathbf{0}\ \text{ and }\ E(\mathbf{0})=\begin{pmatrix}0\\0\\0\end{pmatrix}.$$

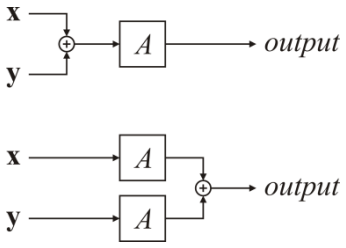
Consequently, at the very least *A* and *C* do not have this property and therefore are undesirable from this point of view.

Note, however, that

$$B\begin{pmatrix}\alpha u_1\\ \alpha u_2\end{pmatrix}=\begin{pmatrix}(\alpha u_1)^3\\ 3(\alpha u_1)^2(\alpha u_2)\\ 3(\alpha u_1)(\alpha u_2)^2\\ (\alpha u_2)^3\end{pmatrix}=\begin{pmatrix}\alpha^3u_1^3\\ 3\alpha^3u_1^2u_2\\ 3\alpha^3u_1u_2^2\\ \alpha^3u_2^3\end{pmatrix}=\alpha^3\begin{pmatrix}u_1^3\\ 3u_1^2u_2\\ 3u_1u_2^2\\ u_2^3\end{pmatrix}=\alpha^3B\begin{pmatrix}u_1\\u_2\end{pmatrix}\neq\alpha B\begin{pmatrix}u_1\\u_2\end{pmatrix},$$

so doubling the input does not double the output—int increases the output by a factor of 8. Thus, *B* is not a desirable mapping, either.

Another desirable property of a mapping is that the response to the sum of two input vectors equals the sum of the two responses, or $A(\mathbf{u}+\mathbf{v})=A\mathbf{u}+A\mathbf{v}$. That is, if you have a system, and you give it two inputs, and then sum the outputs, this should be the same as if you first added the two inputs and then found the output of the system.



Is it possible that a function that satisfies the first property $A(\alpha\mathbf{u})=\alpha A(\mathbf{u})$ but not satisfy the property $A(\mathbf{u}+\mathbf{v})=A(\mathbf{u})+A(\mathbf{v})$? Consider the operator $F:\mathbf{R}^2\to\mathbf{R}^2$ defined by

$$F\begin{pmatrix}u_1\\u_2\end{pmatrix}=\begin{pmatrix}\sqrt[3]{u_1^3+u_2^3}\\ \sqrt[3]{u_1^3-u_2^3}\end{pmatrix}.$$

We note that $F(\mathbf{0})=\mathbf{0}$ and $F(\alpha\mathbf{u})=\alpha F(\mathbf{u})$, but

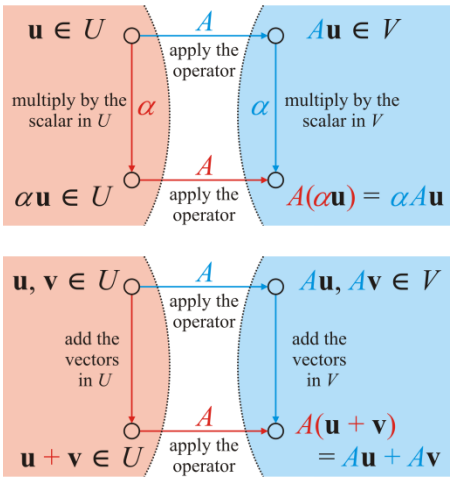
$$F\left(\begin{pmatrix}1\\0\end{pmatrix}+\begin{pmatrix}0\\1\end{pmatrix}\right)=F\left(\begin{pmatrix}1\\1\end{pmatrix}\right)=\left(\begin{pmatrix}\sqrt[3]{1+1}\\ \sqrt[3]{1-1}\end{pmatrix}\right)=\left(\begin{pmatrix}\sqrt[3]{2}\\0\end{pmatrix}\right),$$

but

$$F\begin{pmatrix}1\\0\end{pmatrix}+F\begin{pmatrix}0\\1\end{pmatrix}=\begin{pmatrix}\sqrt[3]{1+0}\\ \sqrt[3]{1-0}\end{pmatrix}+\begin{pmatrix}\sqrt[3]{0+1}\\ \sqrt[3]{0-1}\end{pmatrix}=\begin{pmatrix}1\\1\end{pmatrix}+\begin{pmatrix}1\\-1\end{pmatrix}=\begin{pmatrix}0\\2\end{pmatrix},$$

so $F\left(\begin{pmatrix}1\\0\end{pmatrix}+\begin{pmatrix}0\\1\end{pmatrix}\right)\neq F\begin{pmatrix}1\\0\end{pmatrix}+F\begin{pmatrix}0\\1\end{pmatrix}$. Again, the operator possesses a behavior we would rather not see.

Visually, you can think of the preservation of scalar multiplication and the preservation of vector addition graphically through the next two images:



A mapping $A:\mathcal{U}\to\mathcal{V}$ is linear if it does not matter whether or not you perform the vector operation first in $\mathcal{U}$, then apply the operator or apply the operator first and then perform the vector operation in $\mathcal{V}$.

If we look at the last two operations:

$$D\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}=\begin{pmatrix}u_2\\u_3\end{pmatrix}\text{ and }E\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}=\begin{pmatrix}2u_1-6u_2+u_3\\4u_1+2u_2+5u_3\\-u_1-u_2-3u_3\end{pmatrix},$$

For the first, we see that

$$D\left(\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}+\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}\right)=D\begin{pmatrix}u_1+v_1\\u_2+v_2\\u_3+v_3\end{pmatrix}=\begin{pmatrix}u_2+v_2\\u_3+v_3\end{pmatrix}\text{ while }D\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}+D\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}=\begin{pmatrix}u_2\\u_3\end{pmatrix}+\begin{pmatrix}v_2\\v_3\end{pmatrix}=\begin{pmatrix}u_2+v_2\\u_3+v_3\end{pmatrix},$$

and therefore $D(\mathbf{u}+\mathbf{v})=D\mathbf{u}+D\mathbf{v}$, and for the second

$$E\left(\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}+\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}\right)=E\begin{pmatrix}u_1+v_1\\u_2+v_2\\u_3+v_3\end{pmatrix}=\begin{pmatrix}2(u_1+v_1)-6(u_2+v_2)+(u_3+v_3)\\4(u_1+v_1)+2(u_2+v_2)+5(u_3+v_3)\\-(u_1+v_1)-(u_2+v_2)-3(u_3+v_3)\end{pmatrix}=\begin{pmatrix}2u_1+2v_1-6u_2-6v_2+u_3+v_3\\4u_1+4v_1+2u_2+2v_2+5u_3+5v_3\\-u_1-v_1-u_2-v_2-3u_3-3v_3\end{pmatrix}$$

while

$$E\begin{pmatrix}u_1\\u_2\\u_3\end{pmatrix}+E\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}=\begin{pmatrix}2u_1-6u_2+u_3\\4u_1+2u_2+5u_3\\-u_1-u_2-3u_3\end{pmatrix}+\begin{pmatrix}2v_1-6v_2+v_3\\4v_1+2v_2+5v_3\\-v_1-v_2-3v_3\end{pmatrix}=\begin{pmatrix}2u_1-6u_2+u_3+2v_1-6v_2+v_3\\4u_1+2u_2+5u_3+4v_1+2v_2+5v_3\\-u_1-u_2-3u_3-v_1-v_2-3v_3\end{pmatrix}$$

and because field addition is commutative, we see that, again $E(\mathbf{u} + \mathbf{v}) = E\mathbf{u} + E\mathbf{v}$.

Recall when we found subspaces that we had two choices, we could either prove both properties separately (if **u** and **v** are in the subspace, then **u** + **v** is in the subspace, and if **u** is in the subspace then $\alpha\mathbf{u}$ is in the subspace for all scalars α), or we could simply prove it in one step (if **u** and **v** are in the subspace, then $\alpha\mathbf{u} + \beta\mathbf{v}$ is in the subspace for all scalars α and β. An analogous condition works here. An operator has both these properties described above if and only if $A(\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha A\mathbf{u} + \beta A\mathbf{v}$, shown in Figure 73.

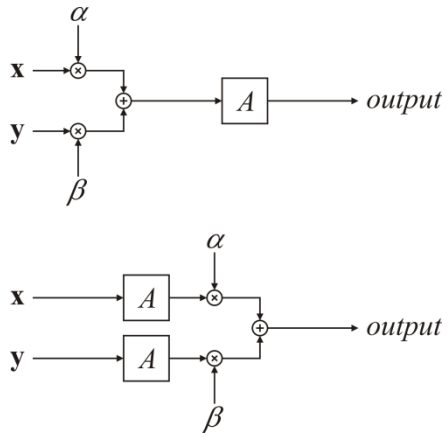


Figure 73. If the system *A* is linear, then both outputs will be the same.

Definition
We will say that the map $A: \mathcal{U} \to \mathcal{V}$ is a *linear map* if it is true that

$$A(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2) = \alpha_1 A\mathbf{u}_1 + \alpha_2 A\mathbf{u}_2$$

for all $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}$ and for all for all $\alpha_1, \alpha_2 \in \mathbf{F}$.[23]

Definition
If the domain and co-domain of a linear map are equal, or $A: \mathcal{U} \to \mathcal{U}$, we call *A* a *linear operator*.

Definition
Given two linear maps $A, B: \mathcal{U} \to \mathcal{V}$, we will say that if and only if

$$A\mathbf{u} = B\mathbf{u}$$

for all $\mathbf{u} \in \mathcal{U}$.

Theorem
Given a mapping $A: \mathcal{U} \to \mathcal{V}$, that $A(\alpha\mathbf{u}_1 + \beta\mathbf{u}_2) = \alpha A\mathbf{u}_1 + \beta A\mathbf{u}_2$ for all $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}$ and for all for all $\alpha, \beta \in \mathbf{F}$ is equivalent saying that both

1. $A(\mathbf{u}_1 + \mathbf{u}_2) = A\mathbf{u}_1 + A\mathbf{u}_2$ and
2. $A(\alpha\mathbf{u}_1) = \alpha A\mathbf{u}_1$

[23] Recall that, in general, **F** will be either the real numbers or the complex numbers.

are true for all $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}$ and for all for all $\alpha, \beta \in \mathbf{F}$. ∎

Proof:
This is an if-and-only-if statement, so we must prove it both ways.

First, assume that $A(\alpha\mathbf{u}_1 + \beta\mathbf{u}_2) = \alpha A\mathbf{u}_1 + \beta A\mathbf{u}_2$ for all $\mathbf{u}_1, \mathbf{u}_2 \in \mathcal{U}$ and for all for all $\alpha, \beta \in \mathbf{F}$. In this case, we may let $\alpha = \beta = 1$, in which case, we get the first statement $A(\mathbf{u}_1 + \mathbf{u}_2) = A\mathbf{u}_1 + A\mathbf{u}_2$, and in the second, if we choose $\mathbf{u}_2 = \mathbf{0}$, we get the second statement $A(\alpha\mathbf{u}_1) = \alpha A\mathbf{u}_1$.

Second, assume that the two individual statements are true for a mapping *A*. We must now show that our original definition is also correct. Thus, given $A(\alpha\mathbf{u}_1 + \beta\mathbf{u}_2)$, we know that both $\alpha_1\mathbf{u}_1, \alpha_2\mathbf{u}_2 \in \mathcal{U}$, so we may apply the first statement:

$$A(\alpha\mathbf{u}_1 + \beta\mathbf{u}_2) = A(\alpha\mathbf{u}_1) + A(\beta\mathbf{u}_2).$$

Next, we may apply the second statement to both operands of the right-hand sum, to get that

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

Therefore, the two individual requirements imply the first. ∎

Definition
Two linear mappings $A, B : \mathcal{U} \to \mathcal{V}$ are said to be *equal* whenever

$$A\mathbf{u} = B\mathbf{u}$$

for all $\mathbf{u} \in \mathcal{V}$.

Problems

1. Find counter examples that demonstrate that the following operators are not linear:

a. $A : \mathbf{R}^2 \to \mathbf{R}^2$ by $A\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} u_1 + 1 \\ u_2 + 1 \end{pmatrix}$, and

b. $B : \mathbf{R}^5 \to \mathbf{R}^2$ by $B\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{pmatrix} = \begin{pmatrix} \min\{u_1, u_2, u_3, u_4, u_5\} \\ \max\{u_1, u_2, u_3, u_4, u_5\} \end{pmatrix}$.

2. Find counter examples that demonstrate that the following operators are not linear:

a. $C:\mathbf{R}^3 \to \mathbf{R}^3$ by $C\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} |u_1| \\ |u_2| \\ |u_3| \end{pmatrix}$, and

b. $D:\mathbf{R}^3 \to \mathbf{R}$ by $D\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \sqrt{\frac{1}{3}\left(u_1^2+u_2^2+u_3^2\right)-\frac{1}{9}\left(u_1+u_2+u_3\right)^2}$.

Solutions

1. We only need to find one counter example:

a. $A\mathbf{0} = A\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \mathbf{0}$.

b. If $\mathbf{v} = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 2 \\ 3 \end{pmatrix}$, then $B(-\mathbf{v}) = \begin{pmatrix} -3 \\ 1 \end{pmatrix}$ but $-B(\mathbf{v}) = \begin{pmatrix} 1 \\ -3 \end{pmatrix}$ and these are unequal.

8.2.1 Linear mappings between finite-dimensional vector spaces

Now, one important question is: can we, in some way, describe all linear operators? In finite-dimensional vector spaces this is quite easy. We will first define matrix-vector multiplication.

Definition

If A is an $n \times m$ matrix and $\mathbf{u}$ is an m-dimensional vector, then the matrix-vector product $\mathbf{v} = A\mathbf{u}$ is the n-dimensional vector $\mathbf{v}$ defined as

$$v_i = \sum_{j=1}^{n} a_{i,j} u_j \ .$$

Alternatively, if $A = \left(\mathbf{a}_{:,1}\ \mathbf{a}_{:,2} \cdots \mathbf{a}_{:,m}\right)$ where $\mathbf{a}_{:,k}$ is the vector representing the kth column of the matrix A, we can also define the matrix-vector product as the linear combination of the column vectors

$$A\mathbf{u} = \left(\mathbf{a}_{:,1}\ \mathbf{a}_{:,2} \cdots \mathbf{a}_{:,m}\right)\mathbf{u} = u_1\mathbf{a}_{:,1} + u_2\mathbf{a}_{:,2} + \cdots + u_m\mathbf{a}_{:,m} .$$

Note: This is the definition is true whether or not we are dealing with a real or complex vector spaces. Compare this with the difference between the definition of the inner product for real and complex vector spaces.

For example,

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x+2y+3z \\ 4x+5y+6z \\ 7x+8y+9z \end{pmatrix} \text{ or } \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = x\begin{pmatrix} 1 \\ 4 \\ 7 \end{pmatrix} + y\begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix} + z\begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix}$$

and

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x+2y+3z \\ 4x+5y+6z \\ 7x+8y+9z \\ 10x+11y+12z \end{pmatrix} \text{ or } \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = x\begin{pmatrix} 1 \\ 4 \\ 7 \\ 10 \end{pmatrix} + y\begin{pmatrix} 2 \\ 5 \\ 8 \\ 11 \end{pmatrix} + z\begin{pmatrix} 3 \\ 6 \\ 9 \\ 12 \end{pmatrix}$$

and

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix}\begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} w+2x+3y+4z \\ 5w+6x+7y+8z \\ 9w+10x+11y+12z \end{pmatrix} \text{ or } \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix}\begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = w\begin{pmatrix} 1 \\ 5 \\ 9 \end{pmatrix} + x\begin{pmatrix} 2 \\ 6 \\ 10 \end{pmatrix} + y\begin{pmatrix} 3 \\ 7 \\ 11 \end{pmatrix} + z\begin{pmatrix} 4 \\ 8 \\ 12 \end{pmatrix}.$$

Theorem

If A is an $m \times n$ matrix, then, by matrix-vector multiplication, $A: \mathbf{F}^n \to \mathbf{F}^m$ and from the properties of matrix-vector multiplication, A must be a linear map.

Proof:

If $\mathbf{u}, \mathbf{v} \in \mathbf{F}^n$, then the entries of $\alpha\mathbf{u} + \beta\mathbf{v}$ are $\alpha u_k + \beta v_k$, and therefore if we consider the ith entry of $A(\alpha\mathbf{u} + \beta\mathbf{v})$, may be note that by the definition of matrix-vector multiplication,

$$\begin{aligned} \left(A(\alpha\mathbf{u} + \beta\mathbf{v})\right)_i &= \sum_{j=1}^{n} a_{i,j}\left(\alpha u_k + \beta v_k\right) \\ &= \sum_{j=1}^{n}\left(a_{i,j}\alpha u_k + a_{i,j}\beta v_k\right) \\ &= \sum_{j=1}^{n} a_{i,j}\alpha u_k + \sum_{j=1}^{n} a_{i,j}\beta v_k \\ &= \alpha\sum_{j=1}^{n} a_{i,j}u_k + \beta\sum_{j=1}^{n} a_{i,j}v_k \\ &= \alpha\left(A\mathbf{u}\right)_i + \beta\left(A\mathbf{v}\right)_i \end{aligned}$$

Thus, the ith entry of $A(\alpha\mathbf{u} + \beta\mathbf{v})$ equals the ith entry of $\alpha A\mathbf{u} + \beta A\mathbf{v}$, thus, the mapping A is linear. ∎

In fact, every matrix represents a linear transformation between appropriate vector spaces, and every linear transformation between two finite-dimensional vector spaces may be represented by a matrix.

8.2.1.1 Matrix-vector multiplication as sum of products

Given a an $m \times n$ matrix $A: \mathbf{F}^n \to \mathbf{F}^m$, matrix-vector multiplication $A\mathbf{u} = \mathbf{v}$ is mapping of a vector $\mathbf{u} \in \mathbf{F}^n$ onto a vector in $\mathbf{F}^m$. Finding the kth entry of this vector $\mathbf{v}$ in $\mathbf{F}^m$ can be visualized graphically as taking the element-wise product of the kth row of A and the vector $\mathbf{u}$. Each row has n columns, and thus this is always defined. You can visualize matrix-vector multiplication as shown in the following series of figures.

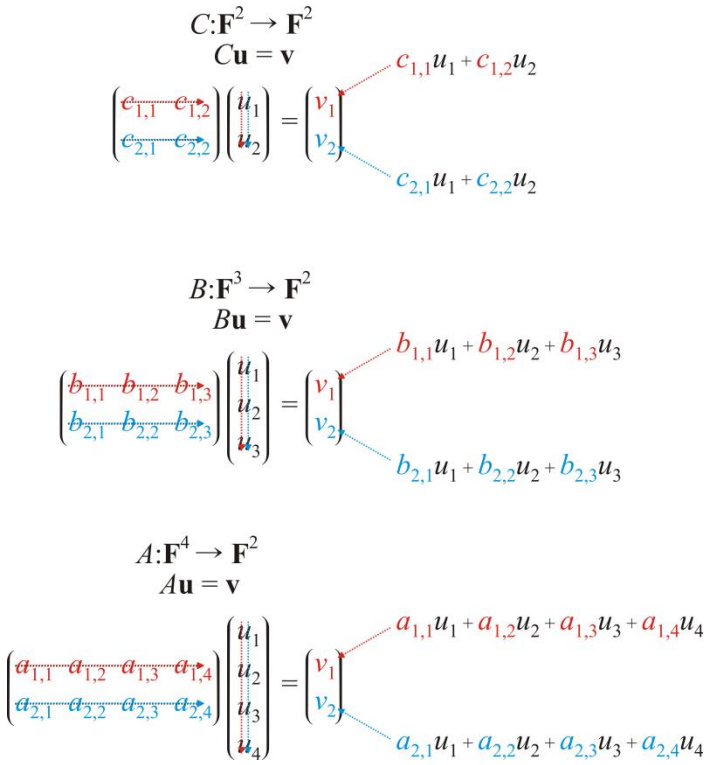


Figure 74. Mapping onto the vector space $\mathbf{R}^2$.

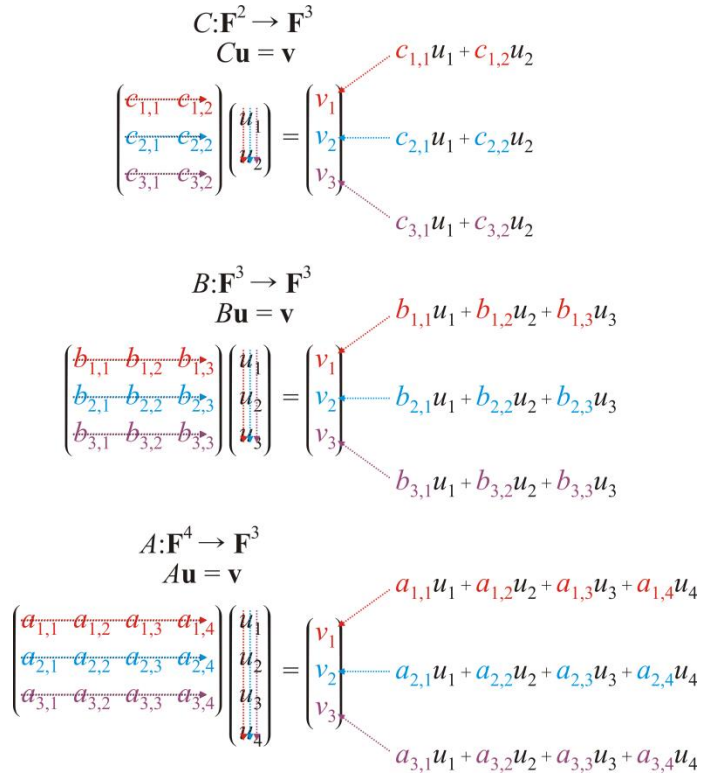


Figure 75. Mapping onto the vector space $\mathbf{R}^3$.

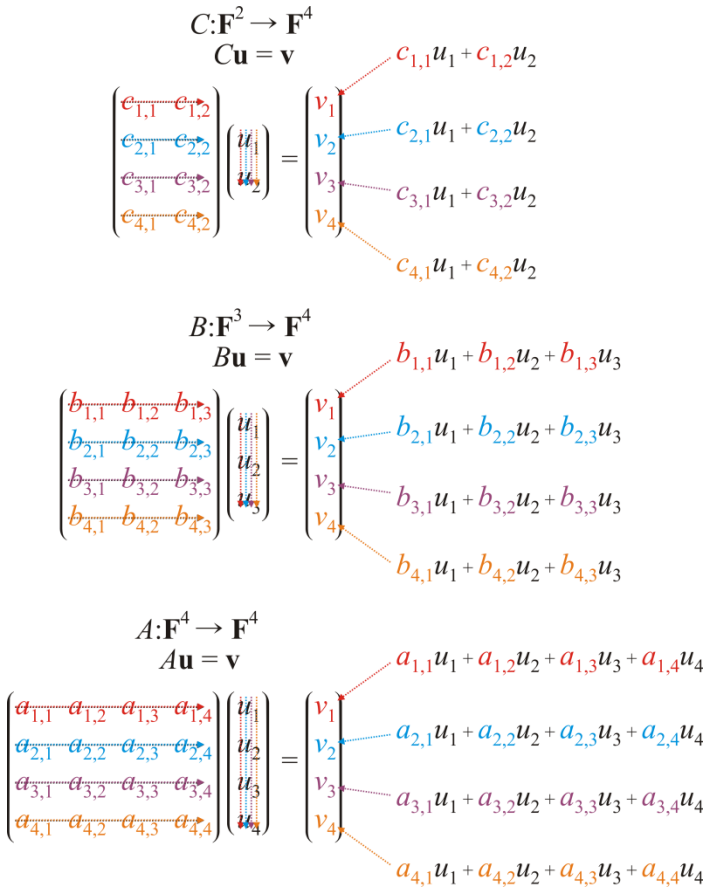


Figure 76. Mapping onto the vector space $\mathbf{R}^3$.

Important: Even though these sums of products look like inner products, if these matrices contain complex entries (that is, they map C^n onto C^m) do not take the complex conjugate of the entries in the matrix.

Problems

1. Describe the vector space from which these matrices maps vectors from and to which they map those vectors to.

$$A = \begin{pmatrix} 2 & 1 & -2 & 3 & 4 \\ -1 & 2 & 0 & 5 & 2 \\ 3 & 2 & 4 & 5 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 2.1 & -0.3 & -0.5 & 0.6 \\ 0 & 3.2 & 0.3 & 1.5 \\ 0 & 0 & 4.2 & -2.1 \\ 0 & 0 & 0 & 5.9 \end{pmatrix} \text{ and } C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

2. Describe the vector spaces from which these matrices map vectors from and to which they map vectors to.

$$A = \begin{pmatrix} 2 & 1 & -2 & 3 & 4 \end{pmatrix},\ B = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \text{ and } C = \begin{pmatrix} \frac{1}{3} \\ 0 \\ \frac{1}{3} \\ 0 \\ \frac{1}{3} \\ 0 \end{pmatrix}.$$

3. Calculate the following matrix-vector products:

$$\begin{pmatrix} 2 & 1 & -2 \\ -1 & 2 & 0 \end{pmatrix}\begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix},\ \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} -2 \\ 5 \end{pmatrix} \text{ and } \begin{pmatrix} 4 & -1 \\ -1 & 4 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

4. Calculate the following matrix-vector products:

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}\begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix},\ \begin{pmatrix} 3 & -2 \\ -1 & 5 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 2 \\ -1 \end{pmatrix} \text{ and } \begin{pmatrix} 3 & 1 \\ 2 & 6 \\ 0 & 4 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

5. Calculate the following matrix-vector products:

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 1 \\ 4 \end{pmatrix},\ \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}\begin{pmatrix} 2 \\ -1 \\ 5 \end{pmatrix} \text{ and } \begin{pmatrix} 2 & 1 & 0 \\ 3 & 2 & 1 \\ 0 & -1 & 2 \\ -2 & 2 & 3 \end{pmatrix}\begin{pmatrix} -2 \\ 3 \\ 1 \end{pmatrix}.$$

6. Calculate the following matrix-vector products:

$$\begin{pmatrix} 3 & 3 & 0 \\ 4 & 2 & 1 \\ 0 & 2 & 2 \\ 2 & -1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix},\ \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 1 & -2 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 4 \end{pmatrix}.$$

Solutions

1. The matrix $A: \mathbf{R}^5 \to \mathbf{R}^3$, while $B: \mathbf{R}^4 \to \mathbf{R}^4$ and $C: \mathbf{R}^2 \to \mathbf{R}^6$.

3. The products are $\begin{pmatrix} 1 \\ -4 \end{pmatrix}$, $\begin{pmatrix} 8 \\ 14 \end{pmatrix}$ and $\begin{pmatrix} 5 \\ 10 \\ -3 \end{pmatrix}$.

5. The products are $\begin{pmatrix} 2 \\ -2 \\ -2 \end{pmatrix}$, $\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}$ and $\begin{pmatrix} -1 \\ 1 \\ -1 \\ 13 \end{pmatrix}$.

8.2.1.2 Matrix-vector multiplication is a linear combination of the column vectors

An alternate interpretation of the mapping of one vector space onto another is to consider the mapping to be a linear combination of vectors in the codomain. Thus, the matrix multiplication can be interpreted as shown here:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = w \begin{pmatrix} 1 \\ 5 \\ 9 \end{pmatrix} + x \begin{pmatrix} 2 \\ 6 \\ 10 \end{pmatrix} + y \begin{pmatrix} 3 \\ 7 \\ 11 \end{pmatrix} + z \begin{pmatrix} 5 \\ 8 \\ 12 \end{pmatrix}$$

To understand this better, the following three images show the interpretation of how a linear mapping represented as a matrix can be interpreted as a linear combination of the columns of the matrix.

$$C:\mathbf{F}^2 \rightarrow \mathbf{F}^2$$
$$C\mathbf{u} = \mathbf{v}$$
$$\begin{pmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = u_1 \begin{pmatrix} c_{1,1} \\ c_{2,1} \end{pmatrix} + u_2 \begin{pmatrix} c_{1,2} \\ c_{2,2} \end{pmatrix} = \begin{pmatrix} c_{1,1}u_1 + c_{1,2}u_2 \\ c_{2,1}u_1 + c_{2,2}u_2 \end{pmatrix}$$

$$B:\mathbf{F}^3 \rightarrow \mathbf{F}^2$$
$$B\mathbf{u} = \mathbf{v}$$
$$\begin{pmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = u_1 \begin{pmatrix} b_{1,1} \\ b_{2,1} \end{pmatrix} + u_2 \begin{pmatrix} b_{1,2} \\ b_{2,2} \end{pmatrix} + u_3 \begin{pmatrix} b_{1,3} \\ b_{2,3} \end{pmatrix} = \begin{pmatrix} b_{1,1}u_1 + b_{1,2}u_2 + b_{1,3}u_3 \\ b_{2,1}u_1 + b_{2,2}u_2 + b_{2,3}u_3 \end{pmatrix}$$

$$A:\mathbf{F}^4 \rightarrow \mathbf{F}^2$$
$$A\mathbf{u} = \mathbf{v}$$
$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = u_1 \begin{pmatrix} a_{1,1} \\ a_{2,1} \end{pmatrix} + u_2 \begin{pmatrix} a_{1,2} \\ a_{2,2} \end{pmatrix} + u_3 \begin{pmatrix} a_{1,3} \\ a_{2,3} \end{pmatrix} + u_4 \begin{pmatrix} a_{1,4} \\ a_{2,4} \end{pmatrix} = \begin{pmatrix} a_{1,1}u_1 + a_{1,2}u_2 + a_{1,3}u_3 + a_{1,4}u_4 \\ a_{2,1}u_1 + a_{2,2}u_2 + a_{2,3}u_3 + a_{2,4}u_4 \end{pmatrix}$$

Figure 77. Linear mappings to $\mathbf{R}^2$ represented as linear combinations of the column vectors of the matrix.

$$C:\mathbf{F}^2 \to \mathbf{F}^3$$
$$C\mathbf{u} = \mathbf{v}$$
$$\begin{pmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \\ c_{3,1} & c_{3,2} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = u_1 \begin{pmatrix} c_{1,1} \\ c_{2,1} \\ c_{3,1} \end{pmatrix} + u_2 \begin{pmatrix} c_{1,2} \\ c_{2,2} \\ c_{3,2} \end{pmatrix} = \begin{pmatrix} c_{1,1}u_1 + c_{1,2}u_2 \\ c_{2,1}u_1 + c_{2,2}u_2 \\ c_{3,1}u_1 + c_{3,2}u_2 \end{pmatrix}$$

$$B:\mathbf{F}^3 \to \mathbf{F}^3$$
$$B\mathbf{u} = \mathbf{v}$$
$$\begin{pmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \\ b_{3,1} & b_{3,2} & b_{3,3} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = u_1 \begin{pmatrix} b_{1,1} \\ b_{2,1} \\ b_{3,1} \end{pmatrix} + u_2 \begin{pmatrix} b_{1,2} \\ b_{2,2} \\ b_{3,2} \end{pmatrix} + u_3 \begin{pmatrix} b_{1,3} \\ b_{2,3} \\ b_{3,3} \end{pmatrix}$$
$$= \begin{pmatrix} b_{1,1}u_1 + b_{1,2}u_2 + b_{1,3}u_3 \\ b_{2,1}u_1 + b_{2,2}u_2 + b_{2,3}u_3 \\ b_{3,1}u_1 + b_{3,2}u_2 + b_{3,3}u_3 \end{pmatrix}$$

$$A:\mathbf{F}^4 \to \mathbf{F}^3$$
$$A\mathbf{u} = \mathbf{v}$$
$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = u_1 \begin{pmatrix} a_{1,1} \\ a_{2,1} \\ a_{3,1} \end{pmatrix} + u_2 \begin{pmatrix} a_{1,2} \\ a_{2,2} \\ a_{3,2} \end{pmatrix} + u_3 \begin{pmatrix} a_{1,3} \\ a_{2,3} \\ a_{3,3} \end{pmatrix} + u_4 \begin{pmatrix} a_{1,4} \\ a_{2,4} \\ a_{3,4} \end{pmatrix}$$
$$= \begin{pmatrix} a_{1,1}u_1 + a_{1,2}u_2 + a_{1,3}u_3 + a_{1,4}u_4 \\ a_{2,1}u_1 + a_{2,2}u_2 + a_{2,3}u_3 + a_{2,4}u_4 \\ a_{3,1}u_1 + a_{3,2}u_2 + a_{3,3}u_3 + a_{3,4}u_4 \end{pmatrix}$$

Figure 78. Linear mappings to $\mathbf{R}^3$ represented as linear combinations of the column vectors of the matrix.

$$C:\mathbf{F}^2 \to \mathbf{F}^4$$
$$C\mathbf{u} = \mathbf{v}$$
$$\begin{pmatrix} c_{1,1} & c_{1,2} \\ c_{2,1} & c_{2,2} \\ c_{3,1} & c_{3,2} \\ c_{4,1} & c_{4,2} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = u_1 \begin{pmatrix} c_{1,1} \\ c_{2,1} \\ c_{3,1} \\ c_{4,1} \end{pmatrix} + u_2 \begin{pmatrix} c_{1,2} \\ c_{2,2} \\ c_{3,2} \\ c_{4,2} \end{pmatrix} = \begin{pmatrix} c_{1,1}u_1 + c_{1,2}u_2 \\ c_{2,1}u_1 + c_{2,2}u_2 \\ c_{3,1}u_1 + c_{3,2}u_2 \\ c_{4,1}u_1 + c_{4,2}u_2 \end{pmatrix}$$

$$B:\mathbf{F}^3 \to \mathbf{F}^4$$
$$B\mathbf{u} = \mathbf{v}$$
$$\begin{pmatrix} b_{1,1} & b_{1,2} & b_{1,3} \\ b_{2,1} & b_{2,2} & b_{2,3} \\ b_{3,1} & b_{3,2} & b_{3,3} \\ b_{4,1} & b_{4,2} & b_{4,3} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = u_1 \begin{pmatrix} b_{1,1} \\ b_{2,1} \\ b_{3,1} \\ b_{4,1} \end{pmatrix} + u_2 \begin{pmatrix} b_{1,2} \\ b_{2,2} \\ b_{3,2} \\ b_{4,2} \end{pmatrix} + u_3 \begin{pmatrix} b_{1,3} \\ b_{2,3} \\ b_{3,3} \\ b_{4,3} \end{pmatrix}$$
$$= \begin{pmatrix} b_{1,1}u_1 + b_{1,2}u_2 + b_{1,3}u_3 \\ b_{2,1}u_1 + b_{2,2}u_2 + b_{2,3}u_3 \\ b_{3,1}u_1 + b_{3,2}u_2 + b_{3,3}u_3 \\ b_{3,1}u_1 + b_{3,2}u_2 + b_{3,3}u_3 \end{pmatrix}$$

$$A:\mathbf{F}^4 \to \mathbf{F}^4$$
$$A\mathbf{u} = \mathbf{v}$$
$$\begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & a_{4,4} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = u_1 \begin{pmatrix} a_{1,1} \\ a_{2,1} \\ a_{3,1} \\ a_{4,1} \end{pmatrix} + u_2 \begin{pmatrix} a_{1,2} \\ a_{2,2} \\ a_{3,2} \\ a_{4,2} \end{pmatrix} + u_3 \begin{pmatrix} a_{1,3} \\ a_{2,3} \\ a_{3,3} \\ a_{4,3} \end{pmatrix} + u_4 \begin{pmatrix} a_{1,4} \\ a_{2,4} \\ a_{3,4} \\ a_{4,4} \end{pmatrix}$$
$$= \begin{pmatrix} a_{1,1}u_1 + a_{1,2}u_2 + a_{1,3}u_3 + a_{1,4}u_4 \\ a_{2,1}u_1 + a_{2,2}u_2 + a_{2,3}u_3 + a_{2,4}u_4 \\ a_{3,1}u_1 + a_{3,2}u_2 + a_{3,3}u_3 + a_{3,4}u_4 \\ a_{4,1}u_1 + a_{4,2}u_2 + a_{4,3}u_3 + a_{4,4}u_4 \end{pmatrix}$$

Figure 79. Linear mappings to $\mathbf{R}^4$ represented as linear combinations of the column vectors of the matrix.

8.3 Properties of linear maps

We will now look at some properties of linear maps:

Theorem

A linear mapping $A:\mathcal{U}\to\mathcal{V}$ maps $\mathbf{0}_u\in\mathcal{U}$ onto $\mathbf{0}_v\in\mathcal{V}$.

Proof:

If $A:\mathcal{U}\to\mathcal{V}$ is linear, then it must satisfy $A(\alpha\mathbf{u})=\alpha A\mathbf{u}$. Therefore, if we choose $\alpha=0$, then for any $\mathbf{u}\in\mathcal{U}$

$$A\mathbf{0}_u=A(0\cdot\mathbf{u})=0\cdot A\mathbf{u}=\mathbf{0}_v.\ \blacksquare$$

As an example of this theorem, if $A:\mathbf{R}^3\to\mathbf{R}^2$ is represented by the matrix, $A=\begin{pmatrix}2.3 & 0.2 & -0.4\\ -0.3 & 4.5 & 0.5\end{pmatrix}$ and we see that

$$A\begin{pmatrix}0\\0\\0\end{pmatrix}=\begin{pmatrix}2.3 & 0.2 & -0.4\\ -0.3 & 4.5 & 0.5\end{pmatrix}\begin{pmatrix}0\\0\\0\end{pmatrix}=\begin{pmatrix}2.3\cdot0+0.2\cdot0-0.4\cdot0\\ -0.3\cdot0+4.5\cdot0+0.5\cdot0\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}.$$

Second, the differential operator $\frac{\mathrm{d}}{\mathrm{d}x}:\mathbf{R}[x]\to\mathbf{R}[x]$ is linear, and $\frac{\mathrm{d}}{\mathrm{d}x}0=0$; that is, the derivative of the zero polynomial is the zero polynomial.

Theorem

A linear map $A:\mathcal{U}\to\mathcal{V}$ maps lines onto lines or onto a single point.

Proof:

Recall that a line in $\mathcal{U}$ is defined as $\mathbf{u}_1+\alpha\mathbf{u}_2$ where $\alpha\in\mathbf{F}$. Therefore,

$$A(\mathbf{u}_1+\alpha\mathbf{u}_2)=A\mathbf{u}_1+\alpha A\mathbf{u}_2,$$

and if $A\mathbf{u}_2=\mathbf{0}_v$, then the line is mapped to a single point $A\mathbf{u}_1$, otherwise, it is mapped onto a line in $\mathcal{V}$. ∎

As an example of this theorem, if $A:\mathbf{R}^3\to\mathbf{R}^2$ is represented by the matrix, $A=\begin{pmatrix}2.3 & 0.2 & -0.4\\ -0.3 & 4.5 & 0.5\end{pmatrix}$, we know that $\begin{pmatrix}1.2\\0.3\\-0.7\end{pmatrix}+\alpha\begin{pmatrix}0.5\\-2.1\\1.7\end{pmatrix}$ is a line in 3-space. Hence,

$$A\left(\begin{pmatrix}1.2\\0.3\\-0.7\end{pmatrix}+\alpha\begin{pmatrix}0.5\\-2.1\\1.7\end{pmatrix}\right)=A\begin{pmatrix}1.2\\0.3\\-0.7\end{pmatrix}+\alpha A\begin{pmatrix}0.5\\-2.1\\1.7\end{pmatrix}=\begin{pmatrix}2.3 & 0.2 & -0.4\\ -0.3 & 4.5 & 0.5\end{pmatrix}\begin{pmatrix}1.2\\0.3\\-0.7\end{pmatrix}+\alpha\begin{pmatrix}2.3 & 0.2 & -0.4\\ -0.3 & 4.5 & 0.5\end{pmatrix}\begin{pmatrix}0.5\\-2.1\\1.7\end{pmatrix}$$

$$=\begin{pmatrix}2.3\cdot1.2+0.2\cdot0.3-0.4\cdot(-0.7)\\ -0.3\cdot1.2+4.5\cdot0.3+0.5\cdot(-0.7)\end{pmatrix}+\alpha\begin{pmatrix}2.3\cdot0.5+0.2\cdot(-2.1)-0.4\cdot1.7\\ -0.3\cdot0.5+4.5\cdot(-2.1)+0.5\cdot1.7\end{pmatrix}$$

$$=\begin{pmatrix}3.1\\0.64\end{pmatrix}+\alpha\begin{pmatrix}0.05\\-8.75\end{pmatrix}$$

which is the equation of a line in the plane.

On the other hand, $\begin{pmatrix}1.2\\0.3\\-0.7\end{pmatrix}+\alpha\begin{pmatrix}-1.9\\1.03\\-10.41\end{pmatrix}$ is also a line in 3-space, but

$$A\left(\begin{pmatrix}1.2\\0.3\\-0.7\end{pmatrix}+\alpha\begin{pmatrix}-1.9\\1.03\\-10.41\end{pmatrix}\right)=A\begin{pmatrix}1.2\\0.3\\-0.7\end{pmatrix}+\alpha A\begin{pmatrix}-1.9\\1.03\\-10.41\end{pmatrix}=\begin{pmatrix}2.3 & 0.2 & -0.4\\ -0.3 & 4.5 & 0.5\end{pmatrix}\begin{pmatrix}1.2\\0.3\\-0.7\end{pmatrix}+\alpha\begin{pmatrix}2.3 & 0.2 & -0.4\\ -0.3 & 4.5 & 0.5\end{pmatrix}\begin{pmatrix}-1.9\\1.03\\-10.41\end{pmatrix}$$

$$=\begin{pmatrix}2.3\cdot1.2+0.2\cdot0.3-0.4\cdot(-0.7)\\ -0.3\cdot1.2+4.5\cdot0.3+0.5\cdot(-0.7)\end{pmatrix}+\alpha\begin{pmatrix}2.3\cdot(-1.9)+0.2\cdot1.03-0.4\cdot(-10.41)\\ -0.3\cdot(-1.9)+4.5\cdot1.03+0.5\cdot(-10.41)\end{pmatrix}$$

$$=\begin{pmatrix}3.1\\0.64\end{pmatrix}+\alpha\begin{pmatrix}0\\0\end{pmatrix}=\begin{pmatrix}3.1\\0.64\end{pmatrix}$$

and thus all of the points on this line in 3-space are mapped onto the point $\begin{pmatrix}3.1\\0.64\end{pmatrix}$.

8.4 Special linear maps and operators

We will now look at three linear maps and operators of interest: the zero map, the identity operator, and the delay and advance operators. The last is only defined in discrete vector spaces such as finite dimensional vector spaces and the vector space of discrete signals.

8.4.1 The zero mapping

Definition

The zero map $O:\mathcal{U}\to\mathcal{V}$ maps every vector in $\mathcal{U}$ onto $\mathbf{0}_v$; that is, $O\mathbf{u}=\mathbf{0}_v$ for all $\mathbf{u}\in\mathcal{U}$.

Theorem

The zero map $O:\mathcal{U}\to\mathcal{V}$ which maps every vector $\mathbf{u}\in\mathcal{U}$ onto $\mathbf{0}_v$, that is, $O\mathbf{u}=\mathbf{0}_v$, is linear.

Proof:

$$\mathbf{0}_v = O(\alpha_1\mathbf{u}_1+\alpha_2\mathbf{u}_2)=\alpha_1 O(\mathbf{u}_1)+\alpha_2 O(\mathbf{u}_2)=\alpha_1\mathbf{0}_v+\alpha_2\mathbf{0}_v=\mathbf{0}_v\,.\ \blacksquare$$

The zero map $O:\mathbf{F}^n\to\mathbf{F}^m$ is represented by the *m* × *n* matrix of zeros $O_{m,n}$, or an *m* × *n zero matrix*. For example, the zero map $O:\mathbf{F}^5\to\mathbf{F}^3$ is represented by the matrix

$$O_{3,5}=\begin{pmatrix}0&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\end{pmatrix}.$$

In MATLAB, an *m* × *n* zero matrix may be generated by

```
>> zeros( 3, 5 )
    ans =
         0     0     0     0     0
         0     0     0     0     0
         0     0     0     0     0
```

8.4.2 The identity operator

Definition

Given any vector space $\mathcal{U}$, we define the identity operator $I:\mathcal{U}\to\mathcal{U}$ where $I\mathbf{u} = \mathbf{u}$ for all $\mathbf{u}\in\mathcal{U}$.

Theorem

The identity operator is linear.

Proof:

Given $I:\mathcal{U}\to\mathcal{U}$ as defined and $\mathbf{u}_1,\mathbf{u}_2\in\mathcal{U}$ and $\alpha_1,\alpha_2\in\mathbf{F}$, we have

$$\alpha_1\mathbf{u}_1+\alpha_2\mathbf{u}_2 = I\left(\alpha_1\mathbf{u}_1+\alpha_2\mathbf{u}_2\right) = \alpha_1 I\mathbf{u}_1+\alpha_2 I\mathbf{u}_2 = \alpha_1\mathbf{u}_1+\alpha_2\mathbf{u}_2 \ . \blacksquare$$

The identity operator $I_n:\mathbf{F}^n\to\mathbf{F}^n$ is defined by the $n \times n$ matrix of all zeros except for ones on the diagonal.

For example, the identity matrix $I_3:\mathbf{F}^3\to\mathbf{F}^3$ is the matrix

$$I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

as

$$I_3\mathbf{u} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 1u_1+0u_2+0u_3 \\ 0u_1+1u_2+0u_3 \\ 0u_1+0u_2+1u_3 \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}.$$

Thus, $I_3\mathbf{u} = \mathbf{u}$ for all vectors $\mathbf{u}\in\mathbf{F}^3$. Note that it doesn't matter whether or not we use **R** or **C** for our vector space—in either case, the multiplicative identity element is 1.

Similarly, for example,

$$Id_5 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Another means of defining the identity matrix is to define the matrix to be the entries

$$I_n = \left(\delta_{i,j}\right) \text{ where } \delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

Note that it is impossible to define an identity matrix that maps one vector space into another vector space.

In MATLAB, an $m \times n$ zero matrix may be generated by the `eye` routine:

```
>> eye( 4 )
    ans =
         1     0     0     0
         0     1     0     0
         0     0     1     0
         0     0     0     1
```

8.4.3 The diagonal operator and diagonal matrices

A linear operator A is *diagonal operator* for a given basis $B = \{\mathbf{u}_1, \mathbf{u}_2, ...\}$ if the action of the linear operator is $A\mathbf{u}_k = \lambda_k \mathbf{u}_k$. Thus, if we know that $\mathbf{v} = \alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \cdots$, then

$$A\mathbf{v} = A(\alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \cdots) = \alpha_1 A\mathbf{u}_1 + \alpha_2 A\mathbf{u}_2 + \cdots = \alpha_1 \lambda_1 \mathbf{u}_1 + \alpha_2 \lambda_2 \mathbf{u}_2 + \cdots.$$

In a finite dimensional vector space, a linear operator A is diagonal if A has a matrix representation of the form

$$A = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \cdots \\ 0 & 0 & \lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

We call such a matrix a *diagonal* matrix.

For example, the diagonal matrix

$$A = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & -4 \end{pmatrix}$$

maps the vector $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}$ onto $A\mathbf{v} = \begin{pmatrix} 3v_1 \\ 2v_2 \\ 5v_3 \\ -4v_4 \end{pmatrix}$.

For a diagonal matrix A, the system of linear equations defined by $A\mathbf{u} = \mathbf{v}$ is defined by

$$\begin{aligned} \lambda_1 u_1 &= v_1 \\ \lambda_2 u_2 &= v_2 \\ \lambda_3 u_3 &= v_3 \\ &\vdots \end{aligned}$$

Given a general matrix A, we will say that the diagonal entries are the entries $a_{1,1}$, $a_{2,2}$, ….

Problems

1. Assume that $A\mathbf{v} = \begin{pmatrix} -2v_1 \\ 4v_2 \\ 3v_3 \end{pmatrix}$, what is the matrix representation of A?

2.

3. What are the number of multiplications and additions required for the matrix-vector multiplication of a general $n \times n$ matrix as compared to the number required for the matrix-vector multiplication of a diagonal $n \times n$ matrix?

Solutions

1. $A = \begin{pmatrix} -2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{pmatrix}$.

3. For a general $n \times n$ matrix, a matrix-vector multiplication requires n^2 multiplications and $n(n-1) = n^2 - n$ additions, while for a diagonal matrix, it requires only n multiplications.

8.4.4 The super- and sub-diagonal matrices

A linear operator A is a *shift operator* for a given basis $B = \{\mathbf{u}_1, \mathbf{u}_2, ...\}$ if the action of the linear operator is described by $A\mathbf{u}_k = \lambda_k \mathbf{u}_{k+1}$. Thus, if we know that $\mathbf{v} = \alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \cdots$, then

$$A\mathbf{v} = A(\alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \cdots) = \alpha_1 A\mathbf{u}_1 + \alpha_2 A\mathbf{u}_2 + \cdots = \alpha_1 \lambda_1 \mathbf{u}_1 + \alpha_2 \lambda_2 \mathbf{u}_2 + \cdots.$$

In a finite dimensional vector space, a linear operator A is a forward-shift operator for the canonical basis $B = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \end{pmatrix}, ... \right\}$ if A has a matrix representation of the form

$$A = \begin{pmatrix} 0 & \lambda_1 & 0 & 0 & \cdots \\ 0 & 0 & \lambda_2 & 0 & \cdots \\ 0 & 0 & 0 & \lambda_3 & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & & \ddots \end{pmatrix}.$$

We say that the entries are on the *super-diagonal* of the matrix A. Given a general matrix A, the entries on the *super-diagonal* are the entries $a_{1,2}$, $a_{2,3}$, $a_{3,4}$, ….

In a finite dimensional vector space, a linear operator A is a backward-shift operator for the canonical basis $B = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ \vdots \end{pmatrix}, ... \right\}$ if A has a matrix representation of the form

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots \\ \lambda_1 & 0 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & 0 & \cdots \\ 0 & 0 & \lambda_3 & 0 & \cdots \\ \vdots & \vdots & \vdots & & \ddots \end{pmatrix}.$$

We say that the entries are on the *sub-diagonal* of the matrix A. Given a general matrix A, the entries on the *sub-diagonal* are the entries $a_{2,1}$, $a_{3,2}$, $a_{4,3}$, ….

In MATLAB, a matrix with entries 5, 6 and 7 on the super-diagonal may be created by calling the routine

```
>> diag( [5 6 7], 1 )
    ans =
         0     5     0     0
         0     0     6     0
         0     0     0     7
         0     0     0     0
```

while a matrix with entries 2, 4 and 5 on the sub-diagonal may be created by calling the routine

```
>> diag( [2 4 5], -1 )
    ans =
         0     0     0     0
         2     0     0     0
         0     4     0     0
         0     0     5     0
```

Note that, in general, `diag( v, n )` creates a square (dim(**v**) + |*n*|) × (dim(**v**) + |*n*|) matrix with the entries of **v** on a line parallel to the diagonal.

Note that the reverse works, as well: calling diag on a matrix extracts a vector of the appropriate dimension containing those entries either above or below the diagonal:

```
>> A = rand( 5, 4 )
    A =
        0.8147    0.2785    0.9572    0.7922
        0.9058    0.5469    0.4854    0.9595
        0.1270    0.9575    0.8003    0.6557
        0.9134    0.9649    0.1419    0.0357
        0.6324    0.1576    0.4218    0.8491

>> diag( A, 1 )
    ans =
        0.2785
        0.4854
        0.6557

>> diag( A, -1 )
    ans =
        0.9058
        0.9575
        0.1419
        0.8491
```

8.4.5 Block diagonal matrices

While diagonal matrices are the simplest of matrices, if the only non-zero entries of a matrix appear in non-overlapping square matrices where the diagonals of those square matrices coincide with the diagonal of the overall matrix. The smaller non-overlapping square matrices are described as *blocks*. Of course, every square matrix is a trivial block-

diagonal matrix with only a single block, but in some cases, there may be more than one block. Some examples of non-trivial block-diagonal matrices are

$$\begin{pmatrix} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ 0 & 0 & 5 & 6 \\ 0 & 0 & 7 & 8 \end{pmatrix}$$ with two blocks and $$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 4 \\ 0 & 0 & 5 & 6 \end{pmatrix}$$ with three blocks (two being 1 × 1 matrices).

8.4.6 The delay and advance operators for discrete signals

For the vector space of discrete signals, the delay operator maps the discrete signal

$$x = \left(x[0], x[1], x[2], x[3], x[4], \cdots\right)$$

then the delay of this discrete signal is

$$Dx = \left(0, x[0], x[1], x[2], x[3], \cdots\right)$$

and the advance of this discrete signal is

$$Ex = \left(x[1], x[2], x[3], x[4], x[5], \cdots\right).$$

Similarly here we say that this is a delay operator because if $x[k]$ represents the value of a sensor after k seconds, then $(Dx)[k]$ is the value of the sensor at the previous second, namely $x[k-1]$.

8.5 Range of a linear map

Recall that given the linear map $A: \mathcal{U} \to \mathcal{V}$, the domain is $\mathcal{U}$ and the codomain is $\mathcal{V}$. It may happen, however, that there are vectors in $\mathcal{V}$ which is not the image of $\mathbf{u}$ for all $\mathbf{u} \in \mathcal{U}$. For example, the zero map O maps all vectors in $\mathcal{U}$ onto the origin $\mathbf{0}_{\mathcal{V}}$, and therefore no vector in $\mathcal{U}$ is ever mapped to any other vector in $\mathcal{V}$.

We will define the *range* of a mapping $A: \mathcal{U} \to \mathcal{V}$ as the collection of all vectors in $\mathbf{v} \in \mathcal{V}$ such that there exists a vector in $\mathbf{u} \in \mathcal{U}$ such that $A\mathbf{u} = \mathbf{v}$; that is, the range is the subset of V defined by

$$\text{range}(A) \overset{def}{=} \{A\mathbf{u} : \mathbf{u} \in \mathcal{U}\}.$$

We may also represent the range as $A\mathcal{U}$, that is, because $\mathcal{U}$ is a collection of vectors, $A\mathcal{U}$ is the image of all vectors

For example, our non-linear mapping $A: \mathbf{R}^2 \to \mathbf{R}$ defined by $A\mathbf{u} = u_1^2 - u_2^2$ has $\text{range}(A) = \mathbf{R}$ (for if $x > 0$, then $A\begin{pmatrix} \sqrt{x} \\ 0 \end{pmatrix} = x$ and if $x < 0$ then $A\begin{pmatrix} 0 \\ \sqrt{-x} \end{pmatrix} = x$), but the non-linear mapping $B: \mathbf{R}^2 \to \mathbf{R}$ defined by $B\mathbf{u} = u_1^2 + u_2^2$ has $\text{range}(B) = [0, \infty)$. If the mapping is linear, however, we then say something special about the range.

Theorem
The range (represented as either range(A) or $A\mathcal{U}$) of a linear mapping $A:\mathcal{U}\to\mathcal{V}$ is a subspace of $\mathcal{V}$.

Proof:
Let us assume that $\mathbf{v}_1,\mathbf{v}_2\in\mathcal{V}$. By our definition, there must therefore exist two vectors $\mathbf{u}_1,\mathbf{u}_2\in\mathcal{U}$ such that

$$A\mathbf{u}_1=\mathbf{v}_1 \text{ and } A\mathbf{u}_2=\mathbf{v}_2.$$

We must now show that $\alpha\mathbf{v}_1+\beta\mathbf{v}_2\in\text{range}(A)$. The obvious candidate for the pre-image of $\alpha\mathbf{u}_1+\beta\mathbf{u}_2$, and we see that

$$A(\alpha\mathbf{u}_1+\beta\mathbf{u}_2)=\alpha A\mathbf{u}_1+\beta A\mathbf{u}_2=\alpha\mathbf{v}_1+\beta\mathbf{v}_2,$$

and since $\alpha\mathbf{u}_1+\beta\mathbf{u}_2\in\mathcal{U}$, it follows that $A(\alpha\mathbf{u}_1+\beta\mathbf{u}_2)=\alpha\mathbf{v}_1+\beta\mathbf{v}_2\in\text{range}(A)$. Therefore, the range of a linear mapping is a subspace of the codomain. ∎

Consequently, given any linear mapping $A:\mathcal{U}\to\mathcal{V}$, it naturally defines a subspace of $\mathcal{V}$.

Definition
If $\mathcal{S}\subseteq\mathcal{U}$ is a subspace of the vector space $\mathcal{U}$ and $A:\mathcal{U}\to\mathcal{V}$, we will define the *image of* $\mathcal{S}$ to be the set of all

$$\{A\mathbf{u}:\mathbf{u}\in\mathcal{S}\}$$

and we will denote this image as $A\mathcal{S}$. Basically, if **u** is a vector, then A**u** is the image of the vector **u**, and if $\mathcal{S}$ is a set, the $A\mathcal{S}$ is the set of all images of vectors in $\mathcal{S}$.

Theorem
If $A:\mathcal{U}\to\mathcal{V}$ and $\mathcal{S}\subseteq\mathcal{U}$ is a subspace of $\mathcal{U}$, then $A\mathcal{S}\subseteq\mathcal{V}$ is a subspace of $\mathcal{V}$.

Proof:
At this point, you should be able to see that the proof will be identical to the proof that the image of $\mathcal{U}$ is a subspace of $\mathcal{V}$.

Let us assume that $\mathbf{v}_1,\mathbf{v}_2\in A\mathcal{S}\subseteq\mathcal{V}$. By our definition, there must therefore exist two vectors $\mathbf{s}_1,\mathbf{s}_2\in\mathcal{S}$ such that

$$A\mathbf{s}_1=\mathbf{v}_1 \text{ and } A\mathbf{s}_2=\mathbf{v}_2.$$

We must now show that $\alpha\mathbf{v}_1+\beta\mathbf{v}_2\in A\mathcal{S}$. The obvious candidate for the pre-image of $\alpha_1\mathbf{s}_1+\alpha_2\mathbf{s}_2$, and we see that

$$A(\alpha_1\mathbf{s}_1+\alpha_2\mathbf{s}_2)=\alpha_1A\mathbf{s}_1+\alpha_2A\mathbf{s}_2=\alpha_1\mathbf{v}_1+\alpha_2\mathbf{v}_2,$$

and since $\alpha_1\mathbf{s}_1+\alpha_2\mathbf{s}_2\in\mathcal{S}$, it follows that $A(\alpha_1\mathbf{s}_1+\alpha_2\mathbf{s}_2)=\alpha_1\mathbf{v}_1+\alpha_2\mathbf{v}_2\in A$. Therefore, the image of a subspace of the domain of a linear mapping is a subspace of the codomain. ∎

Definition

If the range of $A:\mathcal{U}\to\mathcal{V}$ equals all of $\mathcal{V}$, we say that *A* is *onto* (in that *A* maps *onto* the entire vector space $\mathcal{V}$). This is also sometimes called *surjective*.

You may recall the term *surcharge*, a cost that is over or on top of an already existing payment. Similarly, here, a linear mapping is surjective if it maps *on top of* $\mathcal{V}$, meaning, on all vectors of $\mathcal{V}$.

Theorem

If $A:\mathcal{U}\to\mathcal{V}$ is a linear map and if for two vectors $\mathbf{u}_1,\mathbf{u}_2\in\mathcal{U}$, both map onto the same vector $\mathbf{v}\in\mathcal{V}$ under *A* (that is, $A\mathbf{u}_1 = A\mathbf{u}_2 = \mathbf{v}$) then any weighted average of the vectors $\mathbf{u}_1,\mathbf{u}_2\in\mathcal{U}$ also maps onto **v**.

Proof:

We say that $c_1\mathbf{u}_1 + c_2\mathbf{u}_2$ forms a weighted average of $\mathbf{u}_1,\mathbf{u}_2\in\mathcal{U}$ if $c_1 + c_2 = 1$. In this case,

$$\begin{aligned} A\left(c_1\mathbf{u}_1 + c_2\mathbf{u}_2\right) &= c_1A\mathbf{u}_1 + c_2A\mathbf{u}_2 \\ &= c_1\mathbf{v} + c_2\mathbf{v} \\ &= \left(c_1 + c_2\right)\mathbf{v} \\ &= 1\cdot\mathbf{v} = \mathbf{v} \end{aligned}$$

Thus, *A* also maps the weighted average of $\mathbf{u}_1,\mathbf{u}_2$ onto **v**. ∎

Note that the coefficients c_1 and c_2 do not necessarily need to be positive, and that this generalizes to taking a weighted average of an arbitrary number of vectors, too.

The weighted average could be any combination; for example, consider the linear mapping represented by the matrix $A = \begin{pmatrix} 2 & -4 \\ -1 & 2 \end{pmatrix}$. Here we see that $A\begin{pmatrix}1\\1\end{pmatrix} = A\begin{pmatrix}-2\\-0.5\end{pmatrix} = \begin{pmatrix}-2\\1\end{pmatrix}$. Consequently, all three of

$$0.5\begin{pmatrix}1\\1\end{pmatrix} + 0.5\begin{pmatrix}-2\\-0.5\end{pmatrix} = \begin{pmatrix}-0.5\\0.25\end{pmatrix}, \quad 0.1\begin{pmatrix}1\\1\end{pmatrix} + 0.9\begin{pmatrix}-2\\-0.5\end{pmatrix} = \begin{pmatrix}-1.7\\-0.35\end{pmatrix} \quad \text{and} \quad 3\begin{pmatrix}1\\1\end{pmatrix} - 2\begin{pmatrix}-2\\-0.5\end{pmatrix} = \begin{pmatrix}7\\4\end{pmatrix},$$

and of course, many more, all map onto $\begin{pmatrix}-2\\1\end{pmatrix}$. For two points, all points on the line passing through them in $\mathcal{U}$ map onto the same point.

Notice, however, that if $\mathbf{v}\neq\mathbf{0}$, then the collection of all vectors mapping onto **v** does not form a subspace, as it can never happen that $\mathbf{0}_{\mathcal{U}}$ maps onto **v**. If, however, $\mathbf{v} = \mathbf{0}$, we have another interesting subspace; the *null space*, which we will investigate later.

Problems

1. Consider the vector space of all polynomials with real coefficients of degree less than or equal to 5; that is, $\mathbf{P}_5(\mathbf{R})$. What is the range of the differential operator?

Solutions

1. Except for the zero polynomial, the derivative of a degree n polynomial is a polynomial of degree $n - 1$, and therefore if we take the derivative of all polynomials of degree 5 or less, these must all be of degree 4 or less, and therefore the image of the differential operator is $\mathbf{P}_4(\mathbf{R})$, which is, of course, a subspace of $\mathbf{P}_5(\mathbf{R})$.

8.5.1 Range of a finite-dimensional linear mapping

Finding the range of the matrix representation of a finite-dimensional linear mapping is quite straight-forward. Recall that matrix-vector multiplication is simply a linear combination of all of the column vectors of the matrix representation of the linear mapping. Therefore, the image of a linear mapping represented by a matrix A must be the span of the column vectors forming the matrix A.

Theorem

If A is the matrix representation of a finite-dimensional linear mapping, then a basis of the range of A may be found by either

1. applying the Gram-Schmidt process to the columns of A, or
2. applying Gaussian elimination to the matrix A and in the row-echelon form, selecting the columns of A for each column containing a leading non-zero entry within a row.

Proof:

Every image of a linear mapping A is the image of some linear combination of the standard unit basis vectors $\left\{\hat{\mathbf{u}}_1,...,\hat{\mathbf{u}}_m\right\}$. The image of the kth standard unit basis vector is the kth column of A. Therefore,

$$\text{range}\left(A\right) = \text{span}\left\{A\hat{\mathbf{u}}_1,\ldots,A\hat{\mathbf{u}}_m\right\} = \text{span}\left\{A_{:,1},\ldots,A_{:,m}\right\}.$$

Thus, all we need do is apply Gram Schmidt onto the column vectors of A. ∎

Theorem

Given a finite-dimensional mapping $A:\mathbf{R}^n \to \mathbf{R}^m$, the rank of any matrix representation equals the dimension of the range. That is, dim(range(A)) = rank(A).

This is a consequence of the previous theorem.

Suppose now, given a linear operator $M:V \to W$ that two vectors in $\mathbf{v}_1,\mathbf{v}_2 \in V$ both map to the same vector $\mathbf{w} \in W$. What can we say about any other vectors that map onto $\mathbf{w}$?

For example, consider the matrix $A = \begin{pmatrix} 2 & 1 & -2 & 3 \\ 4 & 2 & -1 & 7 \\ -6 & -3 & 0 & -11 \\ -2 & -1 & 11 & 0 \end{pmatrix}$. If we apply Gaussian elimination, we see that $A \sim \begin{pmatrix} 2 & 1 & -2 & 3 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$, and therefore, a basis for the range are the 1st and 3rd columns (dividing the 1st column by 2), or $\left\{\begin{pmatrix} 1 \\ 2 \\ -3 \\ -1 \end{pmatrix}, \begin{pmatrix} -2 \\ -1 \\ 0 \\ 11 \end{pmatrix}\right\}$, and the rank equals the dimension the range space, which is 2.

Problems

1. Find a basis for the range of the linear mappings represented by the matrices

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 3 \\ 0 & -1.2 \end{pmatrix}, \begin{pmatrix} -2.4 & -1.6 \\ 3 & 2 \end{pmatrix}, \begin{pmatrix} 0.9 & 4.6 \\ -3 & -2 \end{pmatrix}.$$

find the dimension of the range, and state the domain and the codomain of these matrices.

2. Do the same as in Question 1 for these matrices:

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1.2 \\ 0 & 3 \\ 0 & -0.9 \end{pmatrix}, \begin{pmatrix} -1.2 & -0.6 \\ 4 & 2 \\ 2.8 & 1.4 \end{pmatrix}, \begin{pmatrix} -1.2 & -0.6 \\ 4 & 2 \\ 2.8 & 1.4 \end{pmatrix}.$$

3. Do the same as in Question 1 for these matrices:

$$\begin{pmatrix} -2 & -1.2 & 0.8 \\ 5 & 3 & -2 \end{pmatrix}, \begin{pmatrix} 3 & 2 & -4 \\ 0.6 & 1.4 & 1.2 \end{pmatrix}, \begin{pmatrix} -2.8 & 0.7 & 6.1 \\ 4 & -1 & -3 \end{pmatrix}, \begin{pmatrix} 0 & -2.4 & 2.8 \\ 0 & -3 & 1 \end{pmatrix}.$$

4. Do the same as in Question 1 for these matrices:

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0.4 & -0.6 \\ 0 & 2 & -3 \\ 0 & -1.8 & 2.7 \end{pmatrix}, \begin{pmatrix} 3.2 & 0.4 & -2.8 \\ 4 & 2 & -1 \\ -0.8 & 2.6 & 5.2 \end{pmatrix}, \begin{pmatrix} -0.3 & 0.5 & -0.6 \\ -0.6 & 1 & -1.6 \\ 3 & -5 & -2 \end{pmatrix}, \begin{pmatrix} 2.5 & -0.9 & 3.1 \\ -5 & 3 & 0 \\ -1 & 2.6 & -3 \end{pmatrix},$$

Solutions

1. The four matrices $\mathbf{R}^2$ to $\mathbf{R}^2$, and the dimensions of the ranges are 0, 1, 1 and 2. Bases for the ranges of these include the second column vector of the second matrix, the first column vector of the third matrix, and both column vectors for the fourth.

3. The four matrices $\mathbf{R}^3$ to $\mathbf{R}^2$, and the dimensions of the ranges are 1, 2, 2 and 2. Bases for the ranges of these include the first column of the first matrix, the first and second columns of the second matrix, the first and third columns of the third matrix, and the second and third columns of the fourth matrix.

8.6 The null space of a linear mapping

Given a linear mapping $A:\mathcal{U}\to\mathcal{V}$, we have already seen that there is at least one vector that must always map onto the zero vector of $\mathcal{V}$; that is, the zero vector of V: $A\mathbf{0}_U=\mathbf{0}_V$. There may, however, be additional vectors that map onto the zero vector of $\mathcal{V}$. For example, if we consider the linear mapping represented by $A=\begin{pmatrix}2 & -4\\ -1 & 2\end{pmatrix}$ where, for example, $A\begin{pmatrix}2\\1\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}$. We can now make a statement about all vectors of U that map onto the zero vector of V.

Theorem

Given a linear mapping $A:\mathcal{U}\to\mathcal{V}$, the collection of all vectors that map onto the origin of $\mathcal{V}$ forms a subspace of $\mathcal{U}$.

Proof:

We have already noted that $\mathbf{0}_{\mathcal{U}}$ is in this collection, as $A\mathbf{0}_{\mathcal{U}}=\mathbf{0}_{\mathcal{V}}$. Suppose now that $\mathbf{u}_1,\mathbf{u}_2\in\mathcal{U}$ both map onto the zero vector. In this case,

$$\begin{aligned}A(\alpha\mathbf{u}_1+\beta\mathbf{u}_2)&=\alpha A\mathbf{u}_1+\beta A\mathbf{u}_2\\&=\mathbf{0}_{\mathcal{V}}+\mathbf{0}_{\mathcal{V}}=\mathbf{0}_{\mathcal{V}}\end{aligned}$$

Consequently, any linear combination of vectors $\alpha\mathbf{u}_1+\beta\mathbf{u}_2$ also maps onto the zero vector of $\mathcal{V}$, and therefore the collection of all vectors that map onto the zero vector of $\mathcal{V}$ forms a subspace of $\mathcal{U}$. ∎

We will define the collection of all vectors that map onto the origin as the *null space* of A and write it as

$$\text{null}(A)=\{\mathbf{u}\in\mathcal{U}:A\mathbf{u}=\mathbf{0}_{\mathcal{V}}\}.$$

Next, we will look at some theorems that describe how the null space interacts with vectors in $\mathcal{U}$.

Theorem

If $A:\mathcal{U}\to\mathcal{V}$ is a linear mapping and suppose that for an arbitrary $\mathbf{u}\in\mathcal{U}$, $A\mathbf{u}=\mathbf{v}$, Then for any $\mathbf{u}_0\in\text{null}(A)$, it follows that $A(\mathbf{u}+\mathbf{u}_0)=\mathbf{v}$.

Proof:

By linearity,

$$\begin{aligned}A(\mathbf{u}+\mathbf{u}_0)&=A(\mathbf{u})+A(\mathbf{u}_0)\\&=\mathbf{v}+\mathbf{0}_V=\mathbf{v}\end{aligned}$$

and thus the result follows. ∎

Theorem

If $A:\mathcal{U}\to\mathcal{V}$ is a linear mapping and suppose that for $\mathbf{u}_1,\mathbf{u}_2\in\mathcal{U}$, $A\mathbf{u}_1=A\mathbf{u}_2$, it follows therefore that $\mathbf{u}_1-\mathbf{u}_2\in\text{null}(A)$.

Proof:

Suppose that $A\mathbf{u}_1=A\mathbf{u}_2=\mathbf{v}$, then by linearity,

$$\begin{aligned}A(\mathbf{u}_1-\mathbf{u}_2)&=A\mathbf{u}_1-A\mathbf{u}_2\\&=\mathbf{v}-\mathbf{v}=\mathbf{0}_V\end{aligned}$$

and thus $\mathbf{u}_1-\mathbf{u}_2\in\text{null}(A)$. ∎

Definition

We will say that a mapping $A:\mathcal{U}\to\mathcal{V}$ is *one-to-one* if for each $\mathbf{u}\in\mathcal{U}$, $A\mathbf{u}\in\mathcal{V}$ is unique. Such a mapping is also said to be *injective*. The exponential function is one-to-one for the real numbers. We can now find a condition under which a linear mapping is one-to-one.

Theorem

If $A:\mathcal{U}\to\mathcal{V}$ is a linear mapping, then A is injective if and only if $\text{null}(A)=\{\mathbf{0}_\mathcal{U}\}$ (or, equivalently, if $\dim(\text{null}(A))=0$).

Note: to prove an if-and-only-if statement ($p\leftrightarrow q$), we have multiple options:

1. show that if the left-hand side is assumed to be true that the right-hand side must also be true, and if the right-hand side is assumed to be true that the left-hand side must also be true (that is, show $(p\to q)\wedge(q\to p)$); or
2. show that if the left-hand side is assumed to be true that the right-hand side must also be true, and if the left-hand side is assumed to be false that the right-hand side must also be true (that is, show $(p\to q)\wedge(\neg p\to\neg q)$).

There are two other variations on this, but this is sufficient.

Proof:

Assume that *A* is one-to-one. Therefore, as $A\mathbf{0}_\mathcal{U}=\mathbf{0}_\mathcal{V}$, it follows that $\text{null}(A)=\{\mathbf{0}_\mathcal{U}\}$.

Assume that $\text{null}(A)=\{\mathbf{0}_\mathcal{U}\}$, then if $A\mathbf{u}_1=A\mathbf{u}_2$, then $\mathbf{0}_\mathcal{U}=A\mathbf{u}_1-A\mathbf{u}_2=A(\mathbf{u}_1-\mathbf{u}_2)$, and as the null space contains only the zero vector, therefore $\mathbf{0}_\mathcal{U}=\mathbf{u}_1-\mathbf{u}_2$, and so $\mathbf{u}_1=\mathbf{u}_2$, so *A* is one-to-one.

Alternatively, assume that *A* is not one-to-one. Therefore there is a vector **v** such that $A\mathbf{u}_1=A\mathbf{u}_2=\mathbf{v}$ with $\mathbf{u}_1\neq\mathbf{u}_2$. It follows that $\mathbf{u}_1-\mathbf{u}_2\neq\mathbf{0}_U$ and yet $A(\mathbf{u}_1-\mathbf{u}_2)=\mathbf{v}-\mathbf{v}=\mathbf{0}_V$, and therefore $\mathbf{u}_1-\mathbf{u}_2\in\text{null}(A)$. Thus, $\text{null}(A)\neq\{\mathbf{0}_U\}$. ∎

We will now see the relationship between $\text{null}(A)$, $\text{range}(A)$ and $\dim(U)$.

For operators in other vector spaces, it is still possible to deduce the null spaces. For example, the null space of the differential operator is the set of all constant-valued functions, and the dimension of this space is 1:

$$\text{null}\left(\frac{d\bullet}{dt}\right)=\{t\mapsto\alpha:\alpha\in\mathbf{R}\}\quad\text{so}\quad\dim\left(\text{null}\left(\frac{d\bullet}{dt}\right)\right)=1.$$

8.6.1 Null space of a finite-dimensional linear mapping

Given the matrix representation of a finite-dimensional linear mapping $A:\mathcal{U}\to\mathcal{V}$, to find the null space, we need only solve for

$$A\mathbf{x}=\mathbf{0}_V\ .$$

For example, given the matrix $A=\begin{pmatrix}1&2&3\\4&5&6\\7&8&9\\10&11&12\end{pmatrix}$, we note that $A:\mathbf{R}^3\to\mathbf{R}^4$. Thus, applying Gaussian elimination, we have that

$$\left(\begin{array}{ccc|c}1&2&3&0\\4&5&6&0\\7&8&9&0\\10&11&12&0\end{array}\right)\sim\left(\begin{array}{ccc|c}1&2&3&0\\0&-3&-6&0\\0&0&0&0\\0&0&0&0\end{array}\right).$$

This says that u_3 may be chosen arbitrarily (it is not constrained), and therefore $-3u_2-6u_3=0$ or $u_2=-2u_3$, and therefore $u_1+2u_2+3u_3=0$ or $u_1+2(-2u_3)+3u_3=0$, so $u_1=4u_3-3u_3=u_3$. Therefore, all solutions are of the form

$$\begin{pmatrix}u_3\\-2u_3\\u_3\end{pmatrix},$$

and as there is only one free variable, $\dim\left(\text{null}\left(A\right)\right)=1$ and a basis for the null space is $\left\{\begin{pmatrix}1\\-2\\1\end{pmatrix}\right\}$. To verify, we note that $A\begin{pmatrix}1\\-2\\1\end{pmatrix}=\begin{pmatrix}0\\0\\0\\0\end{pmatrix}$, and we also note that based on the row-echelon form, $\left\{\begin{pmatrix}1\\4\\7\\10\end{pmatrix},\begin{pmatrix}2\\5\\8\\11\end{pmatrix}\right\}$ form a linearly independent set for a basis of range(A).

As a second example, given the matrix $A=\begin{pmatrix}1&2&3&4\\5&6&7&8\\9&10&11&12\end{pmatrix}$, we note that $A:\mathbf{R}^4\to\mathbf{R}^3$. Applying Gaussian elimination, we note that this is equivalent to

$$\left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 0 \\ 5 & 6 & 7 & 8 & 0 \\ 9 & 10 & 11 & 12 & 0 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & 2 & 3 & 4 & 0 \\ 0 & -4 & -8 & -12 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

Consequently, u_3 and u_4 may be chosen arbitrarily, $-4u_2 - 8u_3 - 12u_4 = 0$, so $u_2 = -2u_3 - 3u_4$ and $u_1 + 2u_2 + 3u_3 + 4u_4 = 0$, thus $u_1 - 4u_3 - 6u_4 + 3u_3 + 4u_4 = 0$, so $u_1 - 4u_3 - 6u_4 + 3u_3 + 4u_4 = u_3 + 2u_4$. Thus, all solutions are of the form $\begin{pmatrix} u_3 + 2u_4 \\ -2u_3 - 3u_4 \\ u_3 \\ u_4 \end{pmatrix}$. Thus, we may choose as our basis of the null space $\left\{\begin{pmatrix} 1 \\ -2 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ -3 \\ 0 \\ 1 \end{pmatrix}\right\}$. We can see that these are linearly independent, and checking, we see that $A\begin{pmatrix} 1 \\ -2 \\ 1 \\ 0 \end{pmatrix} = A\begin{pmatrix} 2 \\ -3 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

Theorem

If $A: \mathbf{R}^n \to \mathbf{R}^m$ and $n > m$, it follows that the null space must have dimension greater than or equal to $n - m$.

Proof:

If $n > m$, as the maximum rank of the matrix A is m, it follows that because $\dim(\text{range}(A)) + \dim(\text{null}(A)) = n$, therefore

$$\begin{aligned} \dim\left(\text{null}\left(A\right)\right) &= n - \dim\left(\text{range}\left(A\right)\right) \\ &\geq n - m \\ &> 0 \end{aligned}$$

Therefore, the null space must be non-empty. ∎

Theorem

If $A: \mathbf{R}^n \to \mathbf{R}^m$ and $n > m$, it follows that the linear mapping is not one-to-one.

Proof:

From the previous theorem, the null space has dimension greater than or equal to 1, and therefore it is not one-to-one. ∎

Problems

1. Suppose that the matrix A is row equivalent to the row-echelon matrix

$$\begin{pmatrix} 0 & a_{1,2} & * & * & * & * & * \\ 0 & 0 & 0 & 0 & a_{2,5} & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where $a_{1,2} \neq 0$ and $a_{2,5} \neq 0$ and where all other entries marked * are any number in the field **F**. What is the dimension of the null space?

Solutions

1. This matrix $A: \mathbf{R}^7 \to \mathbf{R}^3$, and the rank of the matrix is 2. Therefore dim(null(A)) = 7 – 2 = 5.

8.6.2 The dimension of the null space and range

An interesting result that ties the domain of a linear mapping $A: \mathcal{U} \to \mathcal{V}$ and the null space and range together is the following:

Theorem

If $A: \mathcal{U} \to \mathcal{V}$ is a linear mapping and $\mathcal{U}$ is a finite-dimensional vector space, then

$$\dim\left(\text{null}\left(A\right)\right) + \dim\left(\text{range}\left(A\right)\right) = \dim\left(\mathcal{U}\right).$$

Sketch of a proof:

If a matrix $A: \mathbf{F}^n \to \mathbf{F}^m$, then A is row-equivalent to a matrix that is in row-echelon form. Each column of that matrix in row-echelon form therefore either has an entry that is a leading non-zero entry in some row, or not. If a column has a leading non-zero entry, then the corresponding column in A adds another vector into the basis for the range of A; otherwise, if a column does not have a leading non-zero entry, then that entry adds an additional free variable. Each shuch free variable adds an additional dimension to the null space. Thus, as there are n columns, each columns increases either the dimension of the null space or the dimension of the range, and thus, $\dim\left(\text{null}\left(A\right)\right) + \dim\left(\text{range}\left(A\right)\right) = n$. ∎

8.7 The inverse problem

A related problem is the question of finding all solutions **u** to $A\mathbf{u} = \mathbf{v}$ for some target vector **v**. One way to find all solutions is perform Gaussian elimination with partial pivoting; however, there is another technique:

Theroem

If $A:\mathcal{U}\to\mathcal{V}$, then finding the image of a vector $\mathbf{u}\in U$ is straight-forward. For example, if $A:\mathbf{R}^n\to\mathbf{R}^m$, then all one must do is multiply the vector by the $m \times n$ matrix A. Similarly, the rules for differentiation are quite straight-forward, most importantly, we use linearity:

$$\frac{d}{dt}\left(\alpha f(t)+\beta g(t)\right)=\alpha\frac{d}{dt}f(t)+\beta\frac{d}{dt}g(t).$$

Then there are other rules:

1. $\frac{d}{dt}\left(f\left(g(t)\right)\right)=\left(\frac{d}{dt}f(t)\right)\Big|_{t=g(t)}\frac{d}{dt}g(t)$,
2. $\frac{d}{dt}\left(f(t)g(t)\right)=f(t)\frac{d}{dt}g(t)+g(t)\frac{d}{dt}f(t)$, and
3. $\frac{d}{dt}\frac{1}{f(t)}=-\frac{\frac{d}{dt}f(t)}{f^2(t)}$;

and rules for specific functions such as

1. $\frac{d}{dt}t^n=nt^{n-1}$,
2. $\frac{d}{dt}\sin(t)=\cos(t)$ and $\frac{d}{dt}\cos(t)=-\sin(t)$, and
3. $\frac{d}{dt}e^{\lambda t}=\lambda e^{\lambda t}$.

A more difficult problem is given a target vector **v**, then find all **u** such that $A\mathbf{u} = \mathbf{v}$. For matrices, we may apply Guassian elimination with partial pivoting to the augmented matrix, but that is a lot more work than finding **v** given the vector **u**.

If $A:\mathcal{U}\to\mathcal{V}$ is a linear mapping and suppose we are trying to find all solutions to $A\mathbf{u} = \mathbf{v}$ for a target vector **v**. Suppose that we have found one particular solution $\mathbf{u}_p$ such that $A\mathbf{u}_p = \mathbf{v}$. Then all solutions may be found by adding to $\mathbf{u}_p$ any vector in the null space of A:

Theorem

Given $A:\mathcal{U}\to\mathcal{V}$ and a vector $\mathbf{v}\in\mathcal{V}$, if $\mathbf{u}_p\in\mathcal{U}$ is one particular vector such that $A\mathbf{u}_p=\mathbf{v}$ and $\mathbf{u}_0\in U$ is any solution to $A\mathbf{u}=\mathbf{0}_{\mathcal{V}}$, then $\mathbf{u}_p+\alpha\mathbf{u}_0$ is also a solution to $A\mathbf{u}=\mathbf{v}$.

Proof:

$$\begin{aligned}A\left(\mathbf{u}_p+\alpha\mathbf{u}_0\right)&=A\mathbf{u}_p+\alpha A\mathbf{u}_0\\&=\mathbf{v}+\mathbf{0}_{\mathcal{V}}=\mathbf{v}\end{aligned}$$

Thus, $\mathbf{u}_p+\alpha\mathbf{u}_0$ is also a solution to $A\mathbf{u}=\mathbf{v}$. ∎

Specifically, if $\mathbf{u}_{0,1}, \ldots, \mathbf{u}_{0,\dim(\text{null}(A))}$ is a basis for the null space, then all solutions may be written as

$$\mathbf{u}_p + \alpha_1 \mathbf{u}_{0,1} + \cdots + \alpha_{\dim(\text{null}(A))} \mathbf{u}_{0,\dim(\text{null}(A))}.$$

As finding all solutions is straight-forward for finite-dimensional matrices, we will look at an example for differential equations.

From previous examples, we know that $\frac{\mathrm{d}^2}{\mathrm{d}t^2} + 3\frac{\mathrm{d}}{\mathrm{d}t} + 2I$ is a linear operator for differentiable functions. For example, $\left(\frac{\mathrm{d}^2}{\mathrm{d}t^2} + 3\frac{\mathrm{d}}{\mathrm{d}t} + 2I\right)\left(e^t\right) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}e^t + 3\frac{\mathrm{d}}{\mathrm{d}t}e^t + 2e^t = 6e^t$. We would like to find functions $y(t)$ such that that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} y(t) + 3\frac{\mathrm{d}}{\mathrm{d}t} y(t) + 2y(t) = \sin(t).$$

The homogeneous equation is

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} y(t) + 3\frac{\mathrm{d}}{\mathrm{d}t} y(t) + 2y(t) = 0$$

and thus a basis for the null space is e^{-t} and e^{-2t}. Thus, we only need find one particular solution to

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} y(t) + 3\frac{\mathrm{d}}{\mathrm{d}t} y(t) + 2y(t) = \sin(t),$$

after which we can then describe all solutions. We will guess that a particular solution must be a combination of sines and cosines, and thus $y_p(t) = \beta_1 \sin(t) + \beta_2 \cos(t)$. Plugging this into the differential equation, we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\beta_1 \sin(t) + \beta_2 \cos(t)\right) + 3\frac{\mathrm{d}}{\mathrm{d}t}\left(\beta_1 \sin(t) + \beta_2 \cos(t)\right) + 2\left(\beta_1 \sin(t) + \beta_2 \cos(t)\right) = \sin(t)$$

and thus expanding this, we have

$$-\beta_1 \sin(t) - \beta_2 \cos(t) + 3\beta_1 \cos(t) - 3\beta_2 \sin(t) + 2\beta_1 \sin(t) + 2\beta_2 \cos(t) = \sin(t).$$

Now, $\sin(t)$ and $\cos(t)$ are linearly independent (one is not a scalar multiple of the other), and thus, this says

$$\left(\beta_1 - 3\beta_2 - 1\right)\sin(t) + \left(3\beta_1 + \beta_2\right)\cos(t) = 0.$$

The only way for a linear combination of linearly independent vectors (in this case, functions) to equal the zero function is for both coefficients to be zero, so we therefore have:

$$\begin{aligned} \beta_1 - 3\beta_2 - 1 &= 0 \\ 3\beta_1 + \beta_2 &= 0 \end{aligned}$$

As you can see, this defines a system of two linear equations in two unknowns, which we can solve:

$$\left(\begin{array}{cc|c} 1 & -3 & 1 \\ 3 & 1 & 0 \end{array}\right) \sim \left(\begin{array}{cc|c} 3 & 1 & 0 \\ 1 & -3 & 1 \end{array}\right) \sim \left(\begin{array}{cc|c} 3 & 1 & 0 \\ 0 & -\frac{10}{3} & 1 \end{array}\right);$$

and thus, the solution vector is $\boldsymbol{\beta} = \begin{pmatrix} 0.1 \\ -0.3 \end{pmatrix}$. Thus, a particular solution to the non-homogeneous 2nd-order differential equation is $y_p(t) = 0.1\sin(t) - 0.3\cos(t)$, so the general solution is the particular solution plus a a linear combination of the

$$y(t) = 0.1\sin(t) - 0.3\cos(t) + \alpha_1 e^{-t} + \alpha_2 e^{-2t}.$$

Now, you may be wondering how you may constrain the unknowns α_1 and α_2. These come from what are described as *initial conditions*. For example, for this solution, we may know that at time $t = 0$, that $y(0) = -1$ and that the solution is at rest (that is, the derivative is zero or $y^{(1)}(0) = 0$). Plugging these into our solution, we have:

$$\begin{aligned} y(0) &= 0.1\sin(0) - 0.3\cos(0) + \alpha_1 e^{-0} + \alpha_2 e^{-2\cdot 0} \\ &= -0.3 + \alpha_1 + \alpha_2 = -1 \end{aligned}$$

and

$$\begin{aligned} y^{(1)}(0) &= 0.1\cos(0) + 0.3\sin(0) - \alpha_1 e^{-0} - 2\alpha_2 e^{-2\cdot 0} \\ &= 0.1 - \alpha_1 - 2\alpha_2 = 0 \end{aligned}$$

As you can see, this gives us, yet again, a system of two linear euqations in two unknowns:

$$\begin{aligned} \alpha_1 + \alpha_2 &= -0.7 \\ -\alpha_1 - 2\alpha_2 &= -0.1 \end{aligned}$$

Again, solving this, we have

$$\left(\begin{array}{cc|c} 1 & 1 & -0.7 \\ -1 & -2 & -0.1 \end{array}\right) \sim \left(\begin{array}{cc|c} 1 & 1 & -0.7 \\ 0 & -1 & -0.8 \end{array}\right);$$

and thus the solution vector is $\boldsymbol{\alpha} = \begin{pmatrix} -1.5 \\ 0.8 \end{pmatrix}$. Therefore, the solution to the non-homogeneous differential equation with the given initial conditions is

$$y(t) = 0.1\sin(t) - 0.3\cos(t) - 1.5e^{-t} + 0.8e^{-2t}.$$

A plot of the solution to this non-homogeneous differential equation with the given initial conditions is shown in Figure 80.

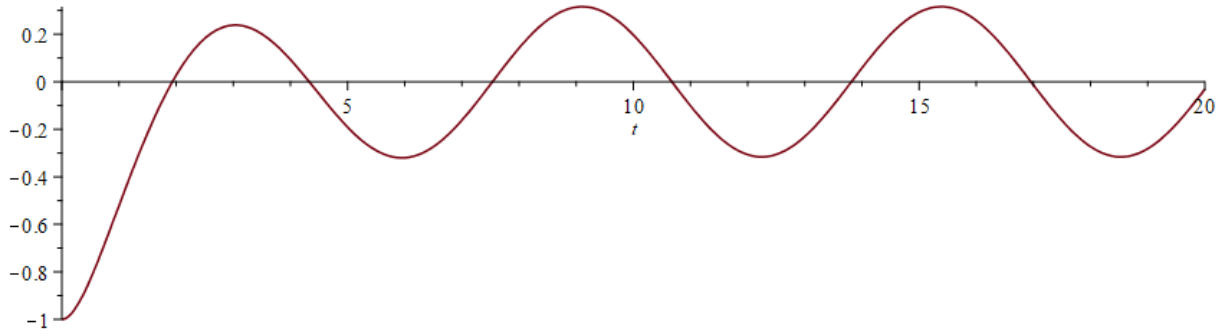


Figure 80. The unique solution to the differential equation $\frac{d^2}{dt^2}y(t)+3\frac{d}{dt}y(t)+2y(t)=\sin(t)$ with initial conditions $y(0) = -1$ with $y^{(1)}(0) = 0$, namely $y(t)=0.1\sin(t)-0.3\cos(t)-1.5e^{-t}+0.8e^{-2t}$.

Problems

1. Find the general solution to the differential equation $\frac{d^2}{dt^2}y(t)+3\frac{d}{dt}y(t)+2y(t)=2t^2$. You should assume that the particular solution is a polynomial of degree two.

2. Find the general solution to the differential equation $\frac{d^2}{dt^2}y(t)+3\frac{d}{dt}y(t)+2y(t)=t-1$. You should assume that the particular solution is a polynomial of degree one.

3. Find the solution to the differential equation in Question 1 when the initial conditions are $y(0) = 1$ and $y^{(1)}(0) = 1$.

4. Find the solution to the differential equation in Question 2 when the initial conditions are $y(0) = 1$ and $y^{(1)}(0) = 1$.

5. Find the general solution to the differential equation $\frac{d^2}{dt^2}y(t)+5\frac{d}{dt}y(t)+4y(t)=2\cos(t)$. You should assume that the particular solution is a linear combination of sines and cosines.

6. Find the general solution to the differential equation $\frac{d^2}{dt^2}y(t)+4\frac{d}{dt}y(t)+3y(t)=-\cos(2t)$. You should assume that the particular solution is a linear combination of sines and cosines with the same frequency as the target.

Solutions

2. $y(t)=0.5t-1.25+\alpha_1 e^{-t}+\alpha_2 e^{-2t}$.

4. $y(t)=0.5t-1.25+0.5e^{-t}-2.75e^{-2t}$.

6. $y(t)=-0.2\cos(t)-0.4\sin(t)+\alpha_1 e^{-3t}+\alpha_2 e^{-t}$.

8.8 The vector space of linear maps

We will see now that we can define both addition and scalar multiplication for linear maps, and we will also see that the collection of all linear maps $A:\mathcal{U}\to\mathcal{V}$ also forms a vector space. We will therefore

1. describe scalar multiplication and addition of linear maps,
2. describe scalar multiplication and addition of finite-dimensional matrices,
3. demonstrate that operations define a vector space of linear maps,
4. demonstrate that the operations we have defined on matrices satisfies these properties,
5. describe scalar multiplications and additions of the differential operator.

8.8.1 Scalar multiplication and addition of linear maps

Suppose we have a linear map $A:\mathcal{U}\to\mathcal{V}$, we can define scalar multiplication of this map as follows:

Definition

Given $A:\mathcal{U}\to\mathcal{V}$, we define $(\alpha A):\mathcal{U}\to\mathcal{V}$ where

$$(\alpha A)\mathbf{u}=\alpha(A\mathbf{u}).$$

for all vectors $\mathbf{u}\in\mathcal{U}$.

That is, given a linear map A, αA scales a the result of $A\mathbf{u}$, or $\mathbf{u}\mapsto\alpha(A\mathbf{u})$. Scalar multiplication may be viewed as either an amplification or attenuation of the output of a linear system described by A, as shown in Figure 81.

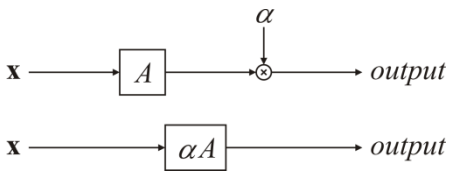


Figure 81. Finding a single linear system that has the same response as the original system attenuated or amplified.

Similarly, given two linear maps between the same vector spaces, $A,B:\mathcal{U}\to\mathcal{V}$, we can define the sum of these two maps as follows:

Definition

Given $A,B:\mathcal{U}\to\mathcal{V}$, we define $(A+B):\mathcal{U}\to\mathcal{V}$ where

$$(A+B)\mathbf{u}=(A\mathbf{u})+(B\mathbf{u}).$$

for all vectors $\mathbf{u}\in\mathcal{U}$. For example, suppose we have two linear systems, the output of which is summed. In this case we may wish to find a single linear system that has the same output, as shown in Figure 82.

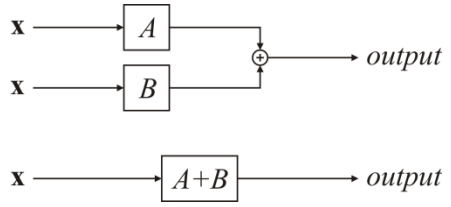


Figure 82. Finding a single linear system that replaces the sum of the responses to two different linear systems.

Now, you may ask yourself, are these not saying the same thing? Actually, no: the first says that we are defining the linear mapping $A+B$ (a new mapping) as one that maps $\mathbf{u} \mapsto A\mathbf{u} + B\mathbf{u}$, and in the second, we are defining a new linear mapping αA that maps $\mathbf{u} \mapsto \alpha(A\mathbf{u})$.

8.8.2 Scalar multiplication and addition of *m* × *n* matrices

If $A: \mathbf{F}^n \to \mathbf{F}^m$, we define $(\alpha A)\mathbf{u} = \alpha(A\mathbf{u})$, it follows that $(\alpha A): \mathbf{R}^n \to \mathbf{R}^m$ and the ith entry of $(\alpha A)\mathbf{u}$ must be

$$\begin{aligned}\big((\alpha A)\mathbf{u}\big)_i &= \alpha\big(A\mathbf{u}\big)_i \\ &= \alpha\sum_{j=1}^{n} a_{i,j}u_j \\ &= \sum_{j=1}^{n}\big(\alpha a_{i,j}\big)u_j\end{aligned}$$

and consequently, the $(i, j)^{\text{th}}$ entry of αA must be $\alpha a_{i,j}$. Given an $m \times n$ matrix A, calculating αA requires mn multiplications and therefore mn FLOPs. For example, if $A: \mathbf{R}^n \to \mathbf{R}^m$ and

$$A = \begin{pmatrix} 2 & 4 & 1 & 0 \\ -3 & 5 & 2 & -1 \\ 4 & -3 & 5 & -2 \end{pmatrix}$$

then $-3A$ has the representation

$$-3A = \begin{pmatrix} -6 & -12 & -3 & 0 \\ 9 & -15 & -6 & 3 \\ -12 & 9 & -15 & 6 \end{pmatrix}.$$

If $A, B: \mathbf{F}^n \to \mathbf{F}^m$, then each of these linear maps has a matrix representation, and since we defined

$$(A+B)\mathbf{u} = A\mathbf{u} + B\mathbf{u},$$

it follows that the ith entry of $(A+B)\mathbf{u}$ must be

$$\begin{aligned}\big((A+B)\mathbf{u}\big)_i &= \big(A\mathbf{u}\big)_i + \big(B\mathbf{u}\big)_i \\ &= \sum_{j=1}^{n} a_{i,j}u_j + \sum_{j=1}^{n} b_{i,j}u_j \\ &= \sum_{j=1}^{n} a_{i,j}u_j + b_{i,j}u_j \\ &\sum_{j=1}^{n}\big(a_{i,j} + b_{i,j}\big)u_j\end{aligned}$$

Thus, $A + B$ must have the representation where the $(i, j)^{\text{th}}$ entry is

$$\big(A+B\big)_{i,j} = a_{i,j} + b_{i,j}\ .$$

Given two $m \times n$ matrices A and B, calculating $A + B$ requires mn additions and therefore mn FLOPs.

For example, if we have two linear mappings $A, B : \mathbf{R}^4 \to \mathbf{R}^3$ where

$$A = \begin{pmatrix} 2 & 4 & 1 & 0 \\ -3 & 5 & 2 & -1 \\ 4 & -3 & 5 & -2 \end{pmatrix} \text{ and } = B = \begin{pmatrix} 3 & 1 & 0 & 4 \\ 0 & -2 & 1 & -3 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

then $(A+B) : \mathbf{R}^4 \to \mathbf{R}^3$ has the representation

$$A + B = \begin{pmatrix} 5 & 5 & 1 & 4 \\ -3 & 3 & 3 & -4 \\ 4 & -3 & 6 & -2 \end{pmatrix}.$$

Finally, we note that $-A = (-1)A$, or

$$-A = \begin{pmatrix} -2 & -4 & -1 & 0 \\ 3 & -5 & -2 & 1 \\ -4 & 3 & -5 & 2 \end{pmatrix}.$$

We now see that $A + (-A) = O$.

Consequently, the collection of $m \times n$ matrices represents the space $\mathcal{L}(\mathbf{F}^n, \mathbf{F}^m)$ (representing all $\mathcal{L}$inear mappings from $\mathbf{F}^n$ to $\mathbf{F}^m$) and it seems that the collection of all such matrices themselves form a vector space. We will prove this in general in the next section.

Problems

1. Find the results of the following matrix operations:

$$\begin{pmatrix} 2 & -2 & 1 & 3 \\ 1 & 2 & -3 & 2 \\ 1 & 2 & 4 & 2 \end{pmatrix} + \begin{pmatrix} 1 & 3 & -2 & 4 \\ -2 & 1 & 5 & 2 \\ -1 & 3 & 2 & 4 \end{pmatrix} \text{ and } -3 \cdot \begin{pmatrix} 2 & -2 & 1 & 3 \\ 1 & 2 & -3 & 2 \\ 1 & 2 & 4 & 2 \end{pmatrix}.$$

2. Find the results of the following matrix operations:

$$\begin{pmatrix} 3 & 2 & -1 \\ 2 & 4 & 2 \\ 3 & -1 & 2 \\ 3 & 2 & 0 \\ 0 & 3 & 1 \end{pmatrix} + \begin{pmatrix} 3 & 0 & 1 \\ 0 & -2 & 0 \\ 0 & -1 & 3 \\ 0 & 3 & 1 \\ 0 & 2 & -4 \end{pmatrix} \text{ and } 4 \cdot \begin{pmatrix} 3 & 2 & -1 \\ 2 & 4 & 2 \\ 3 & -1 & 2 \\ 3 & 2 & 0 \\ 0 & 3 & 1 \end{pmatrix} - 3 \cdot \begin{pmatrix} 3 & 0 & 1 \\ 0 & -2 & 0 \\ 0 & -1 & 3 \\ 0 & 3 & 1 \\ 0 & 2 & -4 \end{pmatrix}$$

3. What is the additive inverse of the matrix $\begin{pmatrix} 2 & -2 & 1 & 3 \\ 1 & 2 & -3 & 2 \\ 1 & 2 & 4 & 2 \end{pmatrix}$?

4. What is the additive inverse of the matrix $-4 \cdot \begin{pmatrix} 3 & 2 & -1 \\ 2 & 4 & 2 \\ 3 & -1 & 2 \\ 3 & 2 & 0 \\ 0 & 3 & 1 \end{pmatrix}$?

Solutions

1. $\begin{pmatrix} 3 & 1 & -1 & 7 \\ -1 & 3 & 2 & 4 \\ 0 & 5 & 6 & 6 \end{pmatrix}$ and $\begin{pmatrix} -6 & 6 & -3 & -9 \\ -3 & -6 & 9 & -6 \\ -3 & -6 & -12 & -6 \end{pmatrix}$.

3. The additive inverse is $\begin{pmatrix} -2 & 2 & -1 & -3 \\ -1 & -2 & 3 & -2 \\ -1 & -2 & -4 & -2 \end{pmatrix}$.

8.8.3 The vector space of all linear mappings

Given two arbitrary vector spaces $\mathcal{U}$ and $\mathcal{V}$, the collection of all linear mappings $A:\mathcal{U}\to\mathcal{V}$ where the mapping addition and scalar multiplication of a mapping are as defined above forms a vector space. Recall that the properties of a vector space are:

1. Scalar multiplication of a vector by 1 leaves that vector unchanged.
2. Vector addition is commutative.
3. Vector addition is associative.
4. Scalar multiplication distributes over vector addition.
5. Scalar multiplication distributes over addition in the field.
6. Scalar multiplication associates with multiplication within the field in question.
7. The zero vector is an identity element for vector addition.
8. There is an additive inverse for each vector.

Now, given that these are linear mappings, we can now show that all the properties of a vector space are satisfied:

1. Multiplying a linear map by 1 leaves it unchanged, for if $A:\mathcal{U}\to\mathcal{V}$ and $\mathbf{u}\in\mathcal{U}$,
$$(1\cdot A)\mathbf{u}=1\cdot(A\mathbf{u})=A\mathbf{u},$$
for $A\mathbf{u}\in\mathcal{V}$ and $\mathcal{V}$ is a vector space, so $1\cdot A=A$.
2. To demonstrate associativity for addition of linear maps, we must show that for $A,B,C:\mathcal{U}\to\mathcal{V}$ that $A+(B+C)=(A+B)+C$. If $A,B,C:\mathcal{U}\to\mathcal{V}$ and $\mathbf{u}\in\mathcal{U}$,
$$\begin{aligned}(A+(B+C))\mathbf{u}&=A\mathbf{u}+(B+C)\mathbf{u}\\&=A\mathbf{u}+(B\mathbf{u}+C\mathbf{u})\\&=(A\mathbf{u}+B\mathbf{u})+C\mathbf{u}\\&=(A+B)\mathbf{u}+C\mathbf{u}\\&=((A+B)+C)\mathbf{u}\end{aligned}$$
for $A\mathbf{u},B\mathbf{u},C\mathbf{u}\in\mathcal{V}$ and vector addition is associative, so $A+(B+C)=(A+B)+C$.
3. Addition of linear maps is communitive, for if $A,B:\mathcal{U}\to\mathcal{V}$ and $\mathbf{u}\in\mathcal{U}$,
$$\begin{aligned}(A+B)\mathbf{u}&=A\mathbf{u}+B\mathbf{u}\\&=B\mathbf{u}+A\mathbf{u}\\&=(B+A)\mathbf{u}\end{aligned}$$
for $A\mathbf{u},B\mathbf{u},C\mathbf{u}\in\mathcal{V}$ and vector addition is commutative, so $A + B = B + A$.

4. Scalar multiplication distributes over addition of linear maps, for if $A, B : \mathcal{U} \to \mathcal{V}$, $\alpha \in \mathbf{F}$ and $\mathbf{u} \in \mathcal{U}$,

$$\begin{aligned}\big(\alpha(A+B)\big)\mathbf{u} &= \alpha\big((A+B)\mathbf{u}\big)\\ &= \alpha\big(A\mathbf{u}+B\mathbf{u}\big)\\ &= \alpha\big(A\mathbf{u}\big)+\alpha\big(B\mathbf{u}\big)\\ &= (\alpha A)\mathbf{u}+(\alpha B)\mathbf{u}\\ &= (\alpha A+\alpha B)\mathbf{u}\end{aligned}$$

as $A\mathbf{u}, B\mathbf{u} \in \mathcal{V}$ and scalar multiplication distributes over vector addition, so $\alpha(A+B) = \alpha A + \alpha B$.

5. Scalar multiplication distributes over addition in the field, for if $A : \mathcal{U} \to \mathcal{V}$, $\alpha, \beta \in \mathbf{F}$ and $\mathbf{u} \in \mathcal{U}$,

$$\begin{aligned}\big((\alpha+\beta)A\big)\mathbf{u} &= (\alpha+\beta)\big(A\mathbf{u}\big)\\ &= \alpha\big(A\mathbf{u}\big)+\beta\big(A\mathbf{u}\big)\\ &= (\alpha A)\mathbf{u}+(\beta A)\mathbf{u}\\ &= (\alpha A+\beta A)\mathbf{u}\end{aligned}$$

as $A\mathbf{u} \in \mathcal{V}$ and scalar multiplication distributes over addition in the field for a vector space, so $(\alpha+\beta)A = \alpha A + \beta A$.

6. Scalar multiplication associates with multiplication within the field, for if $A : \mathcal{U} \to \mathcal{V}$, $\alpha, \beta \in \mathbf{F}$ and $\mathbf{u} \in \mathcal{U}$,

$$\begin{aligned}\big(\alpha(\beta A)\big)\mathbf{u} &= \alpha\big((\beta A)\mathbf{u}\big)\\ &= \alpha\big(\beta\big(A\mathbf{u}\big)\big)\\ &= (\alpha\beta)\big(A\mathbf{u}\big)\\ &= \big((\alpha\beta)A\big)\mathbf{u}\end{aligned}$$

as $A\mathbf{u} \in \mathcal{V}$ and scalar multiplication associates with multiplication within the field for a vector space, so $\alpha(\beta A) = (\alpha\beta)A$.

7. The zero map is an identity element for linear map addition, for if $A : \mathcal{U} \to \mathcal{V}$,

$$\begin{aligned}(A+O)\mathbf{u} &= A\mathbf{u}+O\mathbf{u}\\ &= A\mathbf{u}+\mathbf{0}_v\\ &= A\mathbf{u}\end{aligned}$$

as $A\mathbf{u} \in \mathcal{V}$ and the zero vector is the identity element for a vector space, so $A + O = A$.

8. There is an additive inverse for each linear map, for if $A:\mathcal{U}\to\mathcal{V}$, we can define $-A:\mathcal{U}\to\mathcal{V}$ by the linear map that maps $\mathbf{u}\mapsto -A\mathbf{u}$, then if $\mathbf{u}\in\mathcal{U}$,

$$\begin{aligned}\big(A+(-A)\big)\mathbf{u} &= A\mathbf{u}+(-A)\mathbf{u}\\ &= A\mathbf{u}+\big(-(A\mathbf{u})\big)\\ &= \mathbf{0}_v\end{aligned}$$

as $A\mathbf{u}\in\mathcal{V}$ and here is an additive inverse for each vector in a vector space, so $A+(-A)=O$.

Thus, the collection of all linear mappings $A:\mathcal{U}\to\mathcal{V}$ produces a vector space.

8.8.4 The vector space of all *m* × *n* matrices

Since we have already proved that in general the collection of all linear maps $A:\mathcal{U}\to\mathcal{V}$ forms a vector space, consequently, all linear maps $A:\mathbf{R}^n\to\mathbf{R}^m$ must also form a vector space. Thus, all $m \times n$ matrices together with scalar multiplication of matrices and matrix addition must form a vector space. You will remember that we proved a number of theorems based on the properties of vector spaces, and consequently, these, too must hold up for matrices:

1. The zero matrix $O_{m,n}:\mathbf{R}^n\to\mathbf{R}^m$ is the unique zero linear map such that $A+O_{m,n}=A$ for all $m \times n$ matrices A.
2. The zero matrix may be found by multiply any $m \times n$ matrices A by 0: $0\cdot A = O_{m,n}$.
3. The additive inverse of an $m \times n$ matrices A is unique.
4. The additive inverse of an $m \times n$ matrices A is found by multiplying each entry in the matrix by –1.

For all the vector spaces we have discussed until now, we have also seen that we can define both norms and inner products. For matrices, there is no inner product that provides any useful information as to the similarity of two matrices. We will see, however, that it is possible to define norms on matrix based on the maximum the matrix stretches the length of a vector:

$$|||A|||_1=\max_{\mathbf{u}\neq\mathbf{0}_n}\left\{\frac{\|A\mathbf{u}\|_1}{\|\mathbf{u}\|_1}\right\},\quad |||A|||_2=\max_{\mathbf{u}\neq\mathbf{0}_n}\left\{\frac{\|A\mathbf{u}\|_2}{\|\mathbf{u}\|_2}\right\}\quad\text{and}\quad |||A|||_\infty=\max_{\mathbf{u}\neq\mathbf{0}_n}\left\{\frac{\|A\mathbf{u}\|_\infty}{\|\mathbf{u}\|_\infty}\right\}.$$

For the vast majority of this course, however, we will simply consider linear maps and matrices as vector spaces.

8.8.5 Vector space of differential operators

Recall that we have seen that on the vector space of all polynomials with coefficients in **F**, namely, **F**[*x*], and on this space, both the identity operator and the various derivatives are linear. It is therefore possible to define a linear differential operator $G : \mathbf{F}[x] \to \mathbf{F}[x]$ by taking linear combinations of known operators; for example,

$$G = \frac{\mathrm{d}^2}{\mathrm{d}x^2} + 3\frac{\mathrm{d}}{\mathrm{d}x} + 2I\,,$$

in which case, $G(p)(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2}p(x) + 3\frac{\mathrm{d}}{\mathrm{d}x}p(x) + 2p(x)$. In your course on circuits, you will see how linear circuits (those including resistors, inductors and capacitors) has a response defined by such an operator of derivatives and that such a response is linear for alternating current.

We will now proceed to see that the space of linear operator is in fact *richer* than other vector spaces, as we may also define additional operations that bring us closer to having the properties of fields such as the real and complex numbers.

8.9 Composition of linear maps

Suppose we have one linear mapping $A:\mathcal{U}\to\mathcal{V}$ and a second linear mapping $B:\mathcal{V}\to\mathcal{W}$, we can define the *composition* of these linear mappings as $(BA):\mathcal{U}\to\mathcal{W}$ where $(BA)\mathbf{u}=B(A\mathbf{u})$.[24] The first observation we will make is that the composition of two linear mappings is itself a linear map.

Theorem

If $A:\mathcal{U}\to\mathcal{V}$ and $B:\mathcal{V}\to\mathcal{W}$ are both linear maps, then $(BA):\mathcal{U}\to\mathcal{W}$ must also be a linear map.

Proof:

If $A:\mathcal{U}\to\mathcal{V}$ and $B:\mathcal{V}\to\mathcal{W}$ are linear maps and $\mathbf{u}_1,\mathbf{u}_2\in\mathcal{U}$ and $\alpha_1,\alpha_2\in\mathbf{F}$

$$\begin{aligned}(BA)(\alpha_1\mathbf{u}_1+\alpha_2\mathbf{u}_2) &= B\big(A(\alpha_1\mathbf{u}_1+\alpha_2\mathbf{u}_2)\big)\\ &= B(\alpha_1A\mathbf{u}_1+\alpha_2A\mathbf{u}_2)\\ &= \alpha_1B(A\mathbf{u}_1)+\alpha_2B(A\mathbf{u}_2)\\ &= \alpha_1(BA)\mathbf{u}_1+\alpha_2(BA)\mathbf{u}_2\end{aligned}$$

as first $\mathbf{u}_1,\mathbf{u}_2\in\mathcal{U}$ and A is linear and second $A\mathbf{u}_1, A\mathbf{u}_2\in\mathcal{V}$ and B is linear; therefore, BA is also linear. ∎

From a systems point-of-view, if the output of A is fed as input into B, we want to find a single linear system BA that has the same output, as shown in Figure 83. The question we would like to ask is: can we find a single linear map BA that has the same action as the composition?

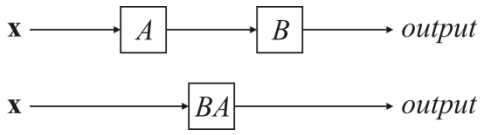


Figure 83. Finding a single system BA that has the same output as when the output of system A becomes the input for system B.

We will now look at some examples, describe matrix-matrix multiplication and then see that linear mappings themselves form a vector space.

[24] Note that the composition may also be written as $B\circ A$, but we will almost exclusively use juxtaposition, or BA, to denote composition.

8.9.1 Properties of composition of linear maps

We will now look at some properties of compositions.

Theorem

Given $A:\mathcal{U}\to\mathcal{V}$, $B:\mathcal{V}\to\mathcal{W}$ and $C:\mathcal{W}\to\mathcal{X}$, composition is associative, or $(CB)A = C(BA)$.

Proof:

Given $\mathbf{u}\in\mathcal{U}$,

$$\begin{aligned}\big(C(BA)\big)\mathbf{u} &= C\big((BA)\mathbf{u}\big)\\ &= C\big(B(A\mathbf{u})\big)\\ &= (CB)(A\mathbf{u})\\ &= \big((CB)A\big)\mathbf{u}\end{aligned}$$

as $A\mathbf{u}\in\mathcal{V}$ and therefore $C(B\mathbf{v}) = (CB)\mathbf{v}$, where in this case, $\mathbf{v} = A\mathbf{u}$. ∎

Theorem

Given $A:\mathcal{U}\to\mathcal{V}$ and $B:\mathcal{V}\to\mathcal{W}$ with $\alpha\in\mathbf{F}$, composition commutes with scalar multiplication, or $\alpha(BA) = (\alpha B)A = B(\alpha A)$.

Proof:

Given $\mathbf{u}\in\mathcal{U}$,

$$\begin{aligned}\big(\alpha(BA)\big)\mathbf{u} &= \alpha\big((BA)\mathbf{u}\big)\\ &= \alpha\big(B(A\mathbf{u})\big)\\ &= (\alpha B)(A\mathbf{u})\\ &= \big((\alpha B)A\big)\mathbf{u}\end{aligned}$$

and

$$\begin{aligned}\big(\alpha(BA)\big)\mathbf{u} &= \alpha\big((BA)\mathbf{u}\big)\\ &= \alpha\big(B(A\mathbf{u})\big)\\ &= B\big(\alpha(A\mathbf{u})\big)\\ &= B\big((\alpha A)\mathbf{u}\big)\\ &= \big(B(\alpha A)\big)\mathbf{u}\end{aligned}$$

as $A\mathbf{u}\in\mathcal{V}$ and therefore $\alpha(B\mathbf{v}) = (\alpha B)\mathbf{v}$ and $\alpha(B\mathbf{v}) = B(\alpha\mathbf{v})$, where in this case, $\mathbf{v} = A\mathbf{u}$. Thus, $\alpha(BA) = (\alpha B)A = B(\alpha A)$. ∎

Theorem

Given $A:\mathcal{U}\to\mathcal{V}$ with $1\in\mathbf{F}$, $1A = A$.

Proof:

Given $\mathbf{u}\in\mathcal{U}$, and thus $(1A)\mathbf{u} = A(1\mathbf{u}) = A\mathbf{u}$, and therefore $1A = A$. ∎

Theorem
Composition of linear maps distributes over addition of linear maps.

Proof:
Suppose that $A_1, A_1 : \mathcal{U} \to \mathcal{V}$ and $B : \mathcal{V} \to \mathcal{W}$. What we must do is go back to the definitions to see that $B(A_1 + A_2)$ and $BA_1 + BA_2$ have the same value for all vectors $\mathbf{u} \in U$:

$$\begin{aligned}\big(B(A_1 + A_2)\big)\mathbf{u} &= B\big((A_1 + A_2)\mathbf{u}\big) \\ &= B(A_1\mathbf{u} + A_2\mathbf{u}) \\ &= B(A_1\mathbf{u}) + B(A_2\mathbf{u}) \\ &= (BA_1)\mathbf{u} + (BA_2)\mathbf{u} \\ &= (BA_1 + BA_2)\mathbf{u}\end{aligned}$$

as $A_1\mathbf{u}, A_2\mathbf{u} \in \mathcal{V}$ and B is linear. Thus, the composition of linear maps distributes over addition of linear maps. ∎

Theorem
The commutation of linear mappings may not be defined, unequatable or not commutative.

Proof:
If $A : \mathcal{U} \to \mathcal{V}$ and $B : \mathcal{V} \to \mathcal{W}$, then while $(BA) : \mathcal{U} \to \mathcal{W}$ is defined, AB is not, for if $\mathbf{v} \in \mathcal{V}$, then $B\mathbf{v} \in \mathcal{W}$ and thus $B\mathbf{v} \notin \mathcal{U}$, so we cannot even define $(AB)\mathbf{v}$.

If $A : \mathcal{U} \to \mathcal{V}$ and $B : \mathcal{V} \to \mathcal{U}$, then $(BA) : \mathcal{U} \to \mathcal{U}$ while $(AB) : \mathcal{V} \to \mathcal{V}$, and thus we cannot even ask if AB and BA are equal.

If $A : \mathcal{U} \to \mathcal{U}$ and $B : \mathcal{U} \to \mathcal{U}$, then both $(BA) : \mathcal{U} \to \mathcal{U}$ and $(AB) : \mathcal{U} \to \mathcal{U}$; however, to show that they do not commute in general, we need only one counter example. Consider, for example, the linear operators $\text{proj}_{\begin{pmatrix}1\\0\end{pmatrix}}, \text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}} : \mathbf{R}^2 \to \mathbf{R}^2$. We saw that these were both linear operators, for

$$\text{proj}_{\mathbf{v}}(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2) = \alpha_1 \text{proj}_{\mathbf{v}}(\mathbf{u}_1) + \alpha_2 \text{proj}_{\mathbf{v}}(\mathbf{u}_2).$$

Thus, we note that

$$\left(\text{proj}_{\begin{pmatrix}1\\0\end{pmatrix}} \circ \text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}}\right)\begin{pmatrix}0\\1\end{pmatrix} = \text{proj}_{\begin{pmatrix}1\\0\end{pmatrix}}\left(\text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}}\begin{pmatrix}0\\1\end{pmatrix}\right) = \text{proj}_{\begin{pmatrix}1\\0\end{pmatrix}}\left(\frac{1}{2}\begin{pmatrix}1\\1\end{pmatrix}\right) = \text{proj}_{\begin{pmatrix}1\\0\end{pmatrix}}\begin{pmatrix}\frac{1}{2}\\\frac{1}{2}\end{pmatrix} = \frac{\frac{1}{2}}{1}\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}\frac{1}{2}\\0\end{pmatrix},$$

but

$$\left(\text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}} \circ \text{proj}_{\begin{pmatrix}1\\0\end{pmatrix}}\right)\begin{pmatrix}0\\1\end{pmatrix} = \text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}}\left(\text{proj}_{\begin{pmatrix}1\\0\end{pmatrix}}\begin{pmatrix}0\\1\end{pmatrix}\right) = \text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}}\left(\frac{0}{1}\begin{pmatrix}1\\0\end{pmatrix}\right) = \text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}}\begin{pmatrix}0\\0\end{pmatrix} = \frac{0}{2}\begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix},$$

and therefore $\left(\text{proj}_{\begin{pmatrix}1\\0\end{pmatrix}} \circ \text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}}\right) = \begin{pmatrix}\frac{1}{2}\\0\end{pmatrix} \neq \begin{pmatrix}0\\0\end{pmatrix} = \left(\text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}} \circ \text{proj}_{\begin{pmatrix}1\\0\end{pmatrix}}\right)\begin{pmatrix}0\\1\end{pmatrix}$, and thus $\text{proj}_{\begin{pmatrix}1\\0\end{pmatrix}} \circ \text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}} \neq \text{proj}_{\begin{pmatrix}1\\1\end{pmatrix}} \circ \text{proj}_{\begin{pmatrix}1\\0\end{pmatrix}}$ and therefore even if both $(AB),(BA): \mathcal{U} \to \mathcal{U}$, it may happen that $AB \neq BA$. ■

8.9.2 Composition of the differential operator

As a first example, in your calculus course, you are already aware that the differential operator is a linear operator mapping the space of polynomials onto itself; that is, $\frac{\mathrm{d}}{\mathrm{d}x} : \mathbf{F}[x] \to \mathbf{F}[x]$. If we compose the differential operator with itself, we get the second derivative: $\left(\frac{\mathrm{d}}{\mathrm{d}x} \circ \frac{\mathrm{d}}{\mathrm{d}x}\right) = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}}{\mathrm{d}x} f(x)\right) = \frac{\mathrm{d}^2}{\mathrm{d}x^2} f(x)$. After all, from your calculus course, you know that the derivative is the limit

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x) = \lim_{h \to 0} \frac{f\left(x + \frac{h}{2}\right) - f\left(x - \frac{h}{2}\right)}{h},$$

and therefore the second derivative is the derivative of this, so

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}x^2} f(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}}{\mathrm{d}x} f(x)\right) \\
&= \lim_{h \to 0} \frac{\dfrac{f\left(\left(x + \frac{h}{2}\right) + \frac{h}{2}\right) - f\left(\left(x - \frac{h}{2}\right) + \frac{h}{2}\right)}{h} - \dfrac{f\left(\left(x + \frac{h}{2}\right) - \frac{h}{2}\right) - f\left(\left(x - \frac{h}{2}\right) - \frac{h}{2}\right)}{h}}{h} \\
&= \lim_{h \to 0} \frac{f(x+h) - f(x) - f(x) + f(x-h)}{h^2} \\
&= \lim_{h \to 0} \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}
\end{aligned}$$

In general, however, calculating the composition higher derivatives is done iteratively: first you calculate the first derivative, then the second, and so on. While you are expected to memorize that $\frac{\mathrm{d}}{\mathrm{d}x} x^n = n x^{n-1}$, you are not expected to memorize that $\frac{\mathrm{d}^2}{\mathrm{d}x^2} x^n = n(n-1) x^{n-2}$, as it is easy to calculate the derivative of nx^{n-1}. Note, however, that it is instructive to realize that $\frac{\mathrm{d}^k}{\mathrm{d}x^k} x^n = \frac{n!}{(n-k)!} x^{n-k}$.

8.9.3 Composition of linear mappings between finite-dimensional vector spaces

Suppose we have two linear mappings A and B where

$$A:\mathbf{F}^n \to \mathbf{F}^m \text{ and } B:\mathbf{F}^m \to \mathbf{F}^\ell,$$

then A is an $m \times n$ matrix and B is an $\ell \times m$ matrix, and if we define $(BA)\mathbf{u} = B(A\mathbf{u})$, then

$$(BA):\mathbf{F}^n \to \mathbf{F}^\ell.$$

Recall that if $\mathbf{u} \in \mathbf{F}^n$, then $A\mathbf{u} \in \mathbf{F}^m$, and thus $B(A\mathbf{u}) \in \mathbf{F}^\ell$. Now, because the composition of linear mappings is itself linear, it follows that BA must be representable by an $\ell \times n$ matrix. As matrices, we see that the dimensions of the matrices are as follows:

B	A	BA
$\ell \times m$	$m \times n$	$\ell \times n$

We can multiply BA because the column dimension of B equals the row dimension of A. Thus, you can multiply a 7×5 and a 5×8 matrix, and the result will be a 7×8 matrix. You cannot multiply a 5×8 and a 7×5 matrix. What is that matrix BA? We will look at two approaches for finding this product, we will examine the time complexity, and then ask under which circumstances it may happen that $AB = BA$.

8.9.3.1 BA = (Ba₁ ... Baₙ)

Remember that if $A = (\mathbf{a}_1 \cdots \mathbf{a}_n)$, where $\mathbf{a}_k$ is the kth column of A, then

$$A\mathbf{u} = (\mathbf{a}_1 \cdots \mathbf{a}_n)\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \mathbf{a}_1 u_1 + \cdots + \mathbf{a}_n u_n \ .$$

Thus, $B(A\mathbf{u}) = B(\mathbf{a}_1 u_1 + \cdots + \mathbf{a}_n u_n) = B(\mathbf{a}_1 u_1) + \cdots + B(\mathbf{a}_n u_n) = (B\mathbf{a}_1)u_1 + \cdots + (B\mathbf{a}_n)u_n$. But the latter is the matrix-vector product of $\mathbf{u}$ with a vary simple matrix. Thus, we have:

Theorem

If $A:\mathbf{F}^n \to \mathbf{F}^m$ and $B:\mathbf{F}^m \to \mathbf{F}^\ell$, then

$$BA = B(\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_n) = (B\mathbf{a}_1\ B\mathbf{a}_2\ \cdots\ B\mathbf{a}_n).$$

Proof:

Let $A:\mathbf{F}^n \to \mathbf{F}^m$ and $B:\mathbf{F}^m \to \mathbf{F}^\ell$ and let $\mathbf{u} \in \mathbf{F}^n$. Then

$$
\begin{aligned}
B(A\mathbf{u}) &= B(\mathbf{a}_1 u_1 + \mathbf{a}_2 u_2 + \cdots + \mathbf{a}_n u_n) \\
&= \left(B(\mathbf{a}_1 u_1) + B(\mathbf{a}_2 u_2) + \cdots + B(\mathbf{a}_n u_n)\right) \\
&= \left((B\mathbf{a}_1)u_1 + (B\mathbf{a}_2)u_2 + \cdots + (B\mathbf{a}_n)u_n\right) \\
&= \begin{pmatrix} B\mathbf{a}_1 & B\mathbf{a}_2 & \cdots & B\mathbf{a}_n \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} \\
&= (BA)\mathbf{u}
\end{aligned}
$$

Thus, $B(A\mathbf{u}) = (BA)\mathbf{u}$ for all $\mathbf{u}$, and therefore $BA = \begin{pmatrix} B\mathbf{a}_1 & B\mathbf{a}_2 & \cdots & B\mathbf{a}_n \end{pmatrix}$. ∎

Recall that this makes sense, for if $A: \mathbf{F}^n \to \mathbf{F}^m$ then $(BA): \mathbf{F}^n \to \mathbf{F}^\ell$, so BA must continue to have the same number of columns as A.

For example, if we multiply $BA = \begin{pmatrix} 2 & 1 \\ -1 & 0 \\ 3 & -2 \\ 0 & -3 \end{pmatrix} \begin{pmatrix} 2 & 5 & -1 \\ 3 & 0 & 4 \end{pmatrix}$, we see that BA is the matrix

$$\begin{aligned} BA &= \begin{pmatrix} B\mathbf{a}_1 & B\mathbf{a}_2 & B\mathbf{a}_3 \end{pmatrix} \\ &= \left(\begin{pmatrix} 2 & 1 \\ -1 & 0 \\ 3 & -2 \\ 0 & -3 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 \\ -1 & 0 \\ 3 & -2 \\ 0 & -3 \end{pmatrix} \begin{pmatrix} 5 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 \\ -1 & 0 \\ 3 & -2 \\ 0 & -3 \end{pmatrix} \begin{pmatrix} -1 \\ 4 \end{pmatrix} \right) \\ &= \left(\begin{pmatrix} 7 \\ -2 \\ 0 \\ -9 \end{pmatrix} \quad \begin{pmatrix} 10 \\ -5 \\ 15 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 2 \\ 1 \\ -11 \\ -12 \end{pmatrix} \right) = \begin{pmatrix} 7 & 10 & 2 \\ -2 & -5 & 1 \\ 0 & 15 & -11 \\ -9 & 0 & -12 \end{pmatrix} \end{aligned}$$

where each of the three columns of the result is the matrix-vector product of B with the columns of A.[25]

8.9.3.2 Fiding the $(i, k)^{\text{th}}$ entry of BA

We can also use the formulas for the jth entry of a matrix-vector product to find the $(i,k)^{\text{th}}$ entry of BA as follows:

Recall that the jth entry of $A\mathbf{u}$ is:

$$\left(A\mathbf{u}\right)_j = \sum_{k=1}^{n} a_{j,k} u_k \ ,$$

and therefore the ith entry of $B(A\mathbf{u})$ is

$$\begin{aligned} \left(B\left(A\mathbf{u}\right)\right)_i &= \sum_{j=1}^{m} b_{i,j} \left(A\mathbf{u}\right)_j \\ &= \sum_{j=1}^{m} b_{i,j} \left(\sum_{k=1}^{n} a_{j,k} u_k \right) \\ &= \sum_{j=1}^{m} \sum_{k=1}^{n} b_{i,j} a_{j,k} u_k \\ &= \sum_{k=1}^{n} \sum_{j=1}^{m} b_{i,j} a_{j,k} u_k \\ &= \sum_{k=1}^{n} \left(\sum_{j=1}^{m} b_{i,j} a_{j,k} \right) u_k \end{aligned}$$

consequently, the $(i, k)^{\text{th}}$ entry of BA is

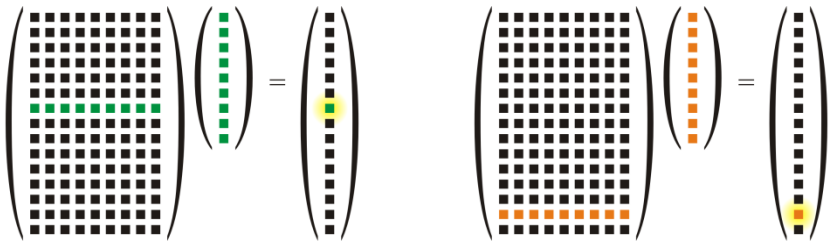
$$\left(BA\right)_{i,k} = \sum_{j=1}^{m} b_{i,j} a_{j,k} \ .$$

[25] Thanks to Yunheng Zou who made this observation.

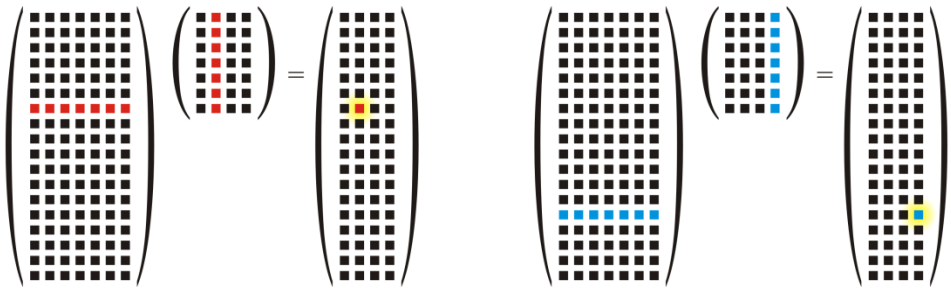
Visually, this is akin to taking the sum of the entry-wise products of the ith row of B and the kth column of A, each of which has m entries:

$$\begin{pmatrix} b_{1,1} & b_{1,2} & b_{1,3} & \cdots & b_{1,m-1} & b_{1,m} \\ b_{2,1} & b_{2,2} & b_{2,3} & \cdots & b_{2,m-1} & b_{2,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ b_{i,1} & b_{i,2} & b_{i,3} & \cdots & b_{i,m-1} & b_{i,m} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ b_{n-1,1} & b_{n-1,2} & b_{n-1,3} & \cdots & b_{n-1,m-1} & b_{n-1,m} \\ b_{n,1} & b_{n,2} & b_{n,3} & \cdots & b_{n,m-1} & b_{n,m} \end{pmatrix} \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,k} & \cdots & a_{1,\ell-1} & a_{1,\ell} \\ a_{2,1} & b_{2,2} & \cdots & a_{2,k} & \cdots & a_{2,\ell-1} & a_{2,\ell} \\ a_{3,1} & a_{3,2} & \cdots & a_{3,k} & \cdots & a_{3,\ell-1} & a_{3,\ell} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ a_{m-1,1} & a_{m-1,2} & \cdots & a_{m-1,k} & \cdots & a_{m-1,\ell-1} & a_{m-1,\ell} \\ a_{m,1} & a_{m,2} & \cdots & a_{m,k} & \cdots & a_{m,\ell-1} & a_{m,\ell} \end{pmatrix}.$$

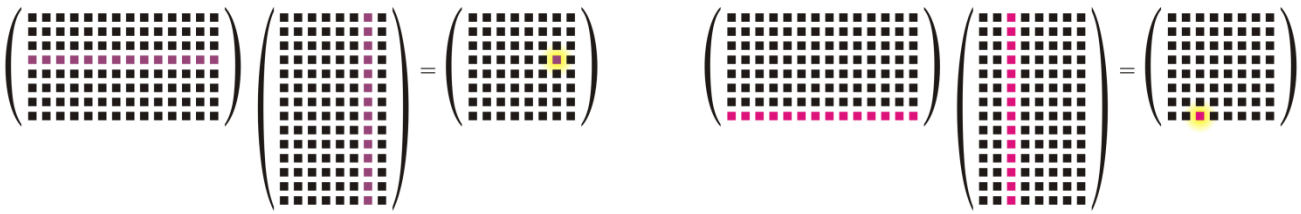
Suppose we have $A: \mathbf{F}^1 \to \mathbf{F}^9$ and $B: \mathbf{F}^9 \to \mathbf{F}^{15}$, and thus A is represented by a 9 × 1 matrix, and B is represented by a 15 × 9 matrix. The composition $(BA): \mathbf{F}^1 \to \mathbf{F}^{15}$ and thus must be represented by a 15 × 1 matrix. The product of a 15 × 9 and a 9 × 1 matrix is a 15 × 1 matrix. In the following image, you can see how a 15 × 9 matrix multiplied by a 9 × 1 matrix produces a 15 × 1 matrix, and the operation is similar to that of matrix-vector multiplication. The $(7,1)^{\text{th}}$ entry of the result is found by taking the real inner product the 7th row of the first matrix and the 1st column of the second. The $(14,1)^{\text{th}}$ entry of the result is found taking the real inner product of the 14th row of the first matrix and the 1st column of the second.



Similarly, if we multiply an 18 × 9 matrix with a 9 × 4 matrix, the result is an 18 × 4 matrix. This represents the finding the composition of one linear mapping $\mathbf{F}^{18} \to \mathbf{F}^9$ and a second linear mapping $\mathbf{F}^9 \to \mathbf{F}^4$ producing a linear mapping $\mathbf{F}^{18} \to \mathbf{F}^4$. The $(7,2)^{\text{th}}$ entry of the result is found by taking the real inner product of the 7th row of the first matrix and the 2nd column of the second. The $(14, 4)^{\text{th}}$ entry is found by taking the real inner product of the 14th row of the first matrix and



Finally, if we multiply a 8 × 14 and a 14 × 8 matrix, the result is a 8 × 8 matrix. The (4, 7)th entry of the result is found by taking the real inner product of the 4th row of the first matrix and the 7th column of the second. Similarly, the (8, 3)rd entry of the result is found by taking the real inner product of the 8th row of the first matrix and the 3rd column of the second.



8.9.3.3 Time complexity of finding BA

Calculating the product of an $\ell \times m$ and an $m \times n$ matrix requires ℓmn multiplications and $\ell(m-1)n$ additions, and therefore approximately on the order of $2\ell mn$ FLOPs.

You may ask yourself: why do we distinguish between matrices and vectors and scalars? After all, Matlab treats vectors as $n \times 1$ or $1 \times n$ matrices, and scalars as 1×1 matrices. You must always refer back to the original concepts:

1. a scalar is a quantity,
2. a vector is an element in a vector space, and
3. a matrix is a representation of a linear mapping between vector spaces.

The justification is as follows:

1. In a vector space, we are comparing related items of data. The similarity between two vectors in a vector space is calculated by the inner product, and this is a scalar value expressing the amount of similarity.
2. A vector or a matrix can be multiplied by a scalar.
3. If $A:\mathbf{R}^n \to \mathbf{R}$ is a linear mapping, then A is a $1 \times n$ matrix, and $A\mathbf{u}$ is a vector in the codomain of the mapping. Even though A looks like a vector, and even though $A\mathbf{u}$ is calculated in a manner similar to that of the inner product, there is a very significant distinction:
   a. it makes no sense to discuss how *similar* a linear mapping is to a vector, and
   b. matrix-vector multiplication does not involve taking the complex conjugate of the first argument, so it cannot be a description of how similar the two components are.

   For example, given the matrix $A = (¼, ¼, ¼, ¼)$, then $A\mathbf{u}$ maps $\mathbf{u}$ onto a vector containing the average of the four entries. To ask whether or not $\mathbf{u}$ is *similar* to A is of less significance.

8.9.3.4 Can AB = BA?

If $A:\mathbf{R}^n \to \mathbf{R}^m$ and $B:\mathbf{R}^m \to \mathbf{R}^\ell$, we can find $BA:\mathbf{R}^n \to \mathbf{R}^\ell$; however, AB is only defined if $\ell = n$. Thus, for AB to even be defined, we must have that $B:\mathbf{R}^m \to \mathbf{R}^n$. In this case, the dimensions of BA is thus $n \times n$ while the dimensions of AB is $m \times m$, and so the only circumstance that there is even the possibility of asking if $AB = BA$ is if $\ell = m = n$, so $A:\mathbf{R}^n \to \mathbf{R}^n$ and $B:\mathbf{R}^n \to \mathbf{R}^n$. In general, however, matrix multiplication is not commutative. There are a few special circumstances where they do commute, and we will examine this here for the case when $n = 2$.

It is quite easy to find two matrices $A, B:\mathbf{R}^2 \to \mathbf{R}^2$ that do not commute; for example,

$$\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix} \text{ but } \begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix}\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}=\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix},$$

and

$$\begin{pmatrix}1 & 2\\ 3 & 4\end{pmatrix}\begin{pmatrix}5 & 6\\ 7 & 8\end{pmatrix}=\begin{pmatrix}19 & 22\\ 43 & 50\end{pmatrix} \text{ but } \begin{pmatrix}5 & 6\\ 7 & 8\end{pmatrix}\begin{pmatrix}1 & 2\\ 3 & 4\end{pmatrix}=\begin{pmatrix}23 & 34\\ 31 & 46\end{pmatrix}.$$

In general, 2×2 matrices do not commute, and there are only very specific circumstances where 2×2 matrices do commute. We will examine the cases where 2×2 matrices commute. Let us start with two generic matrices and calculate their products:

$$\begin{pmatrix}a & b\\ c & d\end{pmatrix}\begin{pmatrix}e & f\\ g & h\end{pmatrix}=\begin{pmatrix}ae+bg & af+bh\\ ce+dg & cf+dh\end{pmatrix} \text{ yet } \begin{pmatrix}e & f\\ g & h\end{pmatrix}\begin{pmatrix}a & b\\ c & d\end{pmatrix}=\begin{pmatrix}ea+fc & eb+fd\\ ga+hc & gb+hd\end{pmatrix}=\begin{pmatrix}ae+cf & be+df\\ ag+ch & bg+dh\end{pmatrix}.$$

If these are to be equal, this requires that

$$\begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix} = \begin{pmatrix} ae+cf & be+df \\ ag+ch & bg+dh \end{pmatrix}$$

and thus

$$\begin{aligned} ae+bg &= ae+cf \\ af+bh &= be+df \\ ce+dg &= ag+ch \\ cf+dh &= bg+dh \end{aligned}$$

The first and last equations require that $bg = cf$, and thus either $\frac{b}{c} = \frac{f}{g}$ in which case $f = \lambda b$ and $g = \lambda c$, or both sides are zero, in which case, at least one variable on each side is zero, yielding the possibilities:

1. $b = 0$ and $f = 0$,
2. one of $b = 0$ or $f = 0$ and $a = d$,
3. $c = 0$ and $g = 0$, or
4. one of $c = 0$ or $g = 0$ and $e = h$.

We will consider each of these in turn:

1. If $b = c = 0$, then $af = df, dg = ag$ so either $a = d$ so $\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}\begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae & af \\ ag & ah \end{pmatrix} = \begin{pmatrix} e & f \\ g & h \end{pmatrix}\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$ commute or $f = g = 0$ so $\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}\begin{pmatrix} e & 0 \\ 0 & h \end{pmatrix} = \begin{pmatrix} ae & 0 \\ 0 & dh \end{pmatrix} = \begin{pmatrix} e & 0 \\ 0 & h \end{pmatrix}\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}$. These say that scalar multiples of the identity commute with all matrices and that diagonal matrices commute, respectively.
2. If $b = f = 0$, then $ce+dg = ag+ch$ or $c(e-h) = g(a-d)$, so again we have five possibilities:
    a. If $c = g = 0$, this reduces to diagonal matrices commuting.
    b. If $c = 0$ and $a = d$, this reduces to a special case of scalar multiples of the identity matrix commuting with all matrices.
    c. If $g = 0$ and $e = h$, this again reduces to a special case of scalar multiples of the identity matrix commuting with all matrices.
    d. If $e = h$ and $a = d$, then $\begin{pmatrix} a & 0 \\ c & a \end{pmatrix}\begin{pmatrix} e & 0 \\ g & e \end{pmatrix} = \begin{pmatrix} ae & 0 \\ ce+ag & ae \end{pmatrix} = \begin{pmatrix} e & 0 \\ g & e \end{pmatrix}\begin{pmatrix} a & 0 \\ c & a \end{pmatrix}$, which says that lower-triangular matrices with equal entries on the diagonal commute.
    e. Otherwise, $c(e-h) = g(a-d)$ so $\frac{c}{g} = \frac{a-d}{e-h}$ and thus $c = \lambda(a-d)$ and $g = \lambda(e-h)$, so
    $$\begin{pmatrix} a & 0 \\ \lambda(a-d) & d \end{pmatrix}\begin{pmatrix} e & 0 \\ \lambda(e-h) & h \end{pmatrix} = \begin{pmatrix} ae & 0 \\ \lambda(ae-dh) & dh \end{pmatrix} = \begin{pmatrix} e & 0 \\ \lambda(e-h) & h \end{pmatrix}\begin{pmatrix} a & 0 \\ \lambda(a-d) & d \end{pmatrix}$$
    and thus if the diagonal entries are not equal for lower triangular matrices, the off-diagonal entries must be scalar multiples of the difference of the diagonal entries.
3. If $c = g = 0$, then $b(e-h) = f(a-d)$. If $b = 0$ or $f = 0$, it reduces to Possibilities 2a, 2b or 2c, so we are left with:

a. If $a = d$ and $e = h$, this reduces to upper triangular matrices with equal entries on the diagonal commuting.

b. Otherwise, $\frac{b}{f} = \frac{a-d}{e-h}$ and thus again $b = \lambda(a-d)$ and $f = \lambda(e-h)$, and so

$$\begin{pmatrix} a & \lambda(a-d) \\ 0 & d \end{pmatrix}\begin{pmatrix} e & \lambda(e-h) \\ 0 & h \end{pmatrix} = \begin{pmatrix} ae & \lambda(ae-dh) \\ 0 & dh \end{pmatrix} = \begin{pmatrix} e & \lambda(e-h) \\ 0 & h \end{pmatrix}\begin{pmatrix} a & \lambda(a-d) \\ 0 & d \end{pmatrix}$$

4. If $g = f = 0$, this again reduces to the two possible cases we saw in Possibility 1.

5. Finally, $\frac{b}{c} = \frac{f}{g}$ or $f = \lambda b$ and $g = \lambda c$, so $\lambda(a-d) = e-h$. If either $\lambda = 0$ or $a = d$ and $e = h$, we are once again left with the previous possibilities, and thus $\lambda = \frac{e-h}{a-d}$

$$\begin{pmatrix} a & \lambda(a-d) \\ \mu(a-d) & d \end{pmatrix}\begin{pmatrix} e & \lambda(e-h) \\ \mu(e-h) & h \end{pmatrix} = \begin{pmatrix} ae+\lambda\mu(a-d)(e-h) & \lambda(ae-dh) \\ \mu(ae-dh) & dh+\lambda\mu(a-d)(e-h) \end{pmatrix}$$
$$= \begin{pmatrix} e & \lambda(e-h) \\ \mu(e-h) & h \end{pmatrix}\begin{pmatrix} a & \lambda(a-d) \\ \mu(a-d) & d \end{pmatrix}$$

Note that the upper- and lower-triangular matrices that commute are special cases of the final case where $\lambda = 0$ or $\mu = 0$ and the diagonal matrices commuting is a special case when $\lambda = 0$ and $\mu = 0$, while if $a = d$ or $e = h$, one matrix degenerates into a scalar multiple of the identity matrix.

Now, if you are given a specific matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, there are two possible cases, when the diagonal entries are equal, and when they differ. If the diagonal entries are equal, then $\begin{pmatrix} a & b \\ c & a \end{pmatrix}$ commutes with any matrix of the form $\begin{pmatrix} e & bf \\ cf & e \end{pmatrix}$.

$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ commutes with any matrix $\begin{pmatrix} e & b\frac{h-e}{d-a} \\ c\frac{h-e}{d-a} & h \end{pmatrix}$

Thus, an example of matrices that commute are

$$\begin{pmatrix} 4.3 & -2.8\lambda \\ -2.8\mu & 7.1 \end{pmatrix}\begin{pmatrix} 2.3 & 3.7\lambda \\ 3.7\mu & -1.4 \end{pmatrix}$$

for any possible values of λ and μ; e.g.,

$$\begin{pmatrix} 4.3 & -2.8 \\ 5.6 & 7.1 \end{pmatrix}\begin{pmatrix} 2.3 & 3.7 \\ -7.4 & -1.4 \end{pmatrix}$$

8.9.3.5 Summary of composition of mappings between finite-dimensional vector spaces

Now, while it is usual to speak of the composition of linear maps in general, we will describe the calculation of BA as a *matrix product* or *matrix multiplication*. The product BA may be found with $BA = \begin{pmatrix} B\mathbf{a}_1 & B\mathbf{a}_2 & \cdots & B\mathbf{a}_n \end{pmatrix}$ but it is also possible to calculate the $(i, k)^{\text{th}}$ entry by summing the entry-wise product of Row i of B and Column k of A, both of which must have equal dimension. It is expensive to find such products in general.

Problems

1. Given the two matrices $A = \begin{pmatrix} 1 & -2 \\ 2 & 0 \\ -1 & 3 \end{pmatrix}$ and $B = \begin{pmatrix} -1 & 2 & 0 \\ 2 & 1 & 3 \\ -2 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}$, find the composition BA. Demonstrate this this is correct by taking $\mathbf{u} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$ and calculating $B(A\mathbf{u})$ and $(BA)\mathbf{u}$.

Solutions

1. Multiplying out the entries, we have

$$\begin{pmatrix} -1 & 2 & 0 \\ 2 & 1 & 3 \\ -2 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 2 & 0 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & \circ \\ \circ & \circ \\ \circ & \circ \\ \circ & \circ \end{pmatrix} \quad (-1)\cdot 1 + 2\cdot 2 + 0\cdot(-1) = 3, \quad \begin{pmatrix} -1 & 2 & 0 \\ 2 & 1 & 3 \\ -2 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 2 & 0 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ \circ & \circ \\ \circ & \circ \\ \circ & \circ \end{pmatrix} \quad (-1)\cdot(-2) + 2\cdot 0 + 0\cdot 3 = 2$$

$$\begin{pmatrix} -1 & 2 & 0 \\ 2 & 1 & 3 \\ -2 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 2 & 0 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 1 & \circ \\ \circ & \circ \\ \circ & \circ \end{pmatrix} \quad 2\cdot 1 + 1\cdot 2 + 3\cdot(-1) = 1, \quad \begin{pmatrix} -1 & 2 & 0 \\ 2 & 1 & 3 \\ -2 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 2 & 0 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 1 & 5 \\ \circ & \circ \\ \circ & \circ \end{pmatrix} \quad 2\cdot(-2) + 1\cdot 0 + 3\cdot 3 = 5$$

$$\begin{pmatrix} -1 & 2 & 0 \\ 2 & 1 & 3 \\ -2 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 2 & 0 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 1 & 5 \\ -3 & \circ \\ \circ & \circ \end{pmatrix} \quad (-2)\cdot 1 + 0\cdot 2 + 1\cdot(-1) = -3, \quad \begin{pmatrix} -1 & 2 & 0 \\ 2 & 1 & 3 \\ -2 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 2 & 0 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 1 & 5 \\ -3 & 7 \\ \circ & \circ \end{pmatrix} \quad (-2)\cdot(-2) + 0\cdot 0 + 1\cdot 3 = 7$$

$$\begin{pmatrix} -1 & 2 & 0 \\ 2 & 1 & 3 \\ -2 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 2 & 0 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 1 & 5 \\ -3 & 7 \\ 0 & \circ \end{pmatrix} \quad 0\cdot 1 + 1\cdot 2 + 2\cdot(-1) = 0 \quad \begin{pmatrix} -1 & 2 & 0 \\ 2 & 1 & 3 \\ -2 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 2 & 0 \\ -1 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 1 & 5 \\ -3 & 7 \\ 0 & 6 \end{pmatrix} \quad 0\cdot(-2) + 1\cdot 0 + 2\cdot 3 = 6$$

Therefore $BA = \begin{pmatrix} 3 & 2 \\ 1 & 5 \\ -3 & 7 \\ 0 & 6 \end{pmatrix}$, and $(BA)\mathbf{u} = \begin{pmatrix} 3 & 2 \\ 1 & 5 \\ -3 & 7 \\ 0 & 6 \end{pmatrix}\begin{pmatrix} -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 9 \\ 17 \\ 12 \end{pmatrix}$, while

$$B(A\mathbf{u}) = \begin{pmatrix} -1 & 2 & 0 \\ 2 & 1 & 3 \\ -2 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix} \left(\begin{pmatrix} 1 & -2 \\ 2 & 0 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} -1 \\ 2 \end{pmatrix} \right) = \begin{pmatrix} -1 & 2 & 0 \\ 2 & 1 & 3 \\ -2 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} -5 \\ -2 \\ 7 \end{pmatrix} = \begin{pmatrix} 1 \\ 9 \\ 17 \\ 12 \end{pmatrix}.$$

8.10 Linear operators

In the previous section, we described the idea of composition of linear maps, so if $A:\mathcal{U}\to\mathcal{V}$ and $B:\mathcal{V}\to\mathcal{W}$, we may define a new mapping $(BA):\mathcal{U}\to\mathcal{W}$. If we restrict ourselves to those linear operators that map a vector space onto itself, that is, $A,B:\mathcal{U}\to\mathcal{U}$, then with composition defined as

$$(BA)\mathbf{u} = B(A\mathbf{u})$$

it follows that $(BA):\mathcal{U}\to\mathcal{U}$, as well. We will call the collection of all linear operators $A:\mathcal{U}\to\mathcal{U}$ as the *operator algebra of linear operators on the vector space* $\mathcal{U}$ and we will represent this as $\mathcal{L}(\mathcal{U})$ (the "$\mathcal{L}$' representing linear maps while the single $\mathcal{U}$ indicating that $\mathcal{U}$ is both the domain and the co-domain of these maps). For mathematicians, an *algebra* is somewhere between a vector space and a field, so there is a form of *multiplication* defined and it is called composition, but we cannot call it multiplication because not every non-zero operator is necessarily invertible, nor is composition necessarily commutative. By the end of this section, you will understand why this course is called *linear algebra*.

8.10.1 Properties of linear mappings

In addition to the properties associated with the composition of linear maps, we have properties specific to linear operators. We already know that composition of linear maps is associative, and therefore the following theorem is quite trivial.

Theorem
Composition of linear operators is associative; that is, $C(BA) = (CB)A$.

Proof:
Linear operators are linear maps, and we have previously seen that composition of linear maps is associative. ∎

Theorem
Composition of linear operators distributes over addition of linear operators; that is, $B(A_1 + A_2) = BA_1 + BA_2$.

Proof:
Linear operators are linear maps, and we have previously seen that composition of linear maps distributes over addition of linear maps. ∎

Theorem
For all operators $A\in\mathcal{L}(\mathcal{U})$, $A\circ I = I\circ A = A$.

Proof:
We have that $(A\circ I)\mathbf{u} = A(I\mathbf{u}) = A\mathbf{u}$ and $(I\circ A)\mathbf{u} = I(A\mathbf{u}) = A\mathbf{u}$. ∎

Thus, the space of operators seems to be very similar to fields, as I behaves very similarly to the multiplicative identity 1. However, we have already seen that that composition in operator algebras is not always commutative; that is, it is usually the case that $AB\neq BA$.

For finite-dimensional vector spaces, the matrix representation of the identity operator is the identity matrix. For example, the identity operator in $\mathcal{L}\left(\mathbf{R}^3\right)$ is $Id_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

Another property of fields is that multiplication distributes over addition; that is, $\alpha(\beta+\gamma) = \alpha\beta + \alpha\gamma$. We can also see this is also the case for operator algebras.

8.10.2 Powers of linear operators

Now that we have operator composition defined, we may also define operator powering.

Definition

Given a linear operator $A: \mathcal{U} \to \mathcal{U}$, we will define $A^0 = I$ and then $A^n = AA^{n-1}$ for all integers $n = 1, 2, 3, \ldots$.

Theorem

If $A: \mathcal{U} \to \mathcal{U}$, $A^m A^n = A^{m+n}$.

Proof:

In the first case, we have that $A^m A^n = A^m\left(AA^{n-1}\right) = \left(A^m A\right)A^{n-1} = A^{m+1}A^{n-1}$. Repeat this n times and we have $A^m A^n = A^{m+n}A^0 = A^{m+n}I = A^{m+n}$. ∎

Theorem

If $A: \mathcal{U} \to \mathcal{U}$, $(A^m)^n = A^{mn}$.

Proof:

Using the previous theorem and the definition of integer powers of matrix, we have

$$\left(A^m\right)^n = A^m\left(A^m\right)^{n-1} = A^m\left(A^m\left(A^m\right)^{n-2}\right) = \left(A^m A^m\right)\left(A^m\right)^{n-2} = A^{2m}\left(A^m\right)^{n-2}$$

and repeating this n times, we have $\left(A^m\right)^n = A^{mn}\left(A^m\right)^0 = A^{mn}I = A^{mn}$. ∎

Theorem

If $A:\mathcal{U}\to\mathcal{U}$ and $\alpha\in\mathbf{F}$, $(\alpha A)^n=\alpha^n A^n$.

Proof:

Finally, using the definition of integer powers of a matrix of

$$\begin{aligned}(\alpha A)^n &= (\alpha A)(\alpha A)^{n-1}\\ &= \alpha A(\alpha A)^{n-1}\\ &= \alpha A\left((\alpha A)(\alpha A)^{n-2}\right)\\ &= \alpha A\left(\alpha\left(A(\alpha A)^{n-2}\right)\right)\\ &= \alpha^2 A\left(A(\alpha A)^{n-2}\right)\\ &= \alpha^2 (AA)(\alpha A)^{n-2}\\ &= \alpha^2 A^2 (\alpha A)^{n-2}\end{aligned}$$

and repeating this n times yields that $(\alpha A)^n=\alpha^n A^n(\alpha A)^0=\alpha^n A^n I=\alpha^n A^n$. ∎

8.10.3 Efficient calculation of matrix powers

Suppose that $A:\mathbf{F}^n\to\mathbf{F}^n$ and we are attempting to calculate A^m. In this case, we must perform $m-1$ products of $n\times n$ matrices. As each matrix multiplication requires n^3 multiplications and $n^2(n-1)$ additions, for a total of approximately $2n^3$ FLOPs, performing this $m-1$ times yields a total of approximately $2mn^3$ FLOPs.

A faster means of calculating this is to recall that every positive integer can be written uniquely as a sum of powers of two. For example,

$$1 = 2^0,\ 2 = 2^1,\ 3 = 2^1 + 2^0,\ 4 = 2^2,\ 5 = 2^2 + 2^0,\ 6 = 2^2 + 2^1,$$

$42 = 32 + 8 + 2 = 2^5 + 2^3 + 2^1$ and

$$1970 = 1024 + 512 + 256 + 128 + 32 + 16 + 2 = 2^{10} + 2^9 + 2^8 + 2^7 + 2^5 + 2^4 + 2^1.$$

Beyond the scope of this course, this is the basis for the binary representation of numbers.

Now, suppose you want to calculate A^{1970} without performing $2\cdot 1970n^3$ FLOPs. Consider the following strategy:

1. Calculate the following:

$$A^2 = AA$$
$$A^4 = A^2 A^2$$
$$A^8 = A^4 A^4$$
$$A^{16} = A^8 A^8$$
$$A^{32} = A^{16} A^{16}$$
$$A^{64} = A^{32} A^{32}$$
$$A^{128} = A^{64} A^{64}$$
$$A^{256} = A^{128} A^{128}$$
$$A^{512} = A^{256} A^{256}$$
$$A^{1024} = A^{512} A^{512}$$

where this required 10 matrix-matrix products, or a total of $2 \cdot 10n^3$ FLOPs. We can choose the number of steps by calculating $\lfloor \lg(m) \rfloor$ where $\lg(m) \overset{\text{def}}{=} \log_2(m)$ is the base-2 logarithm.

2. Recalling that 1970 = 1024 + 512 + 256 + 128 + 32 + 16 + 2, so we may now calculate

$$A^{1970} = A^{1024 + 512 + 256 + 128 + 32 + 16 + 2}$$
$$= A^{1024} A^{512} A^{256} A^{128} A^{32} A^{16} A^2$$

As there are at most 10 terms we can multiply, this will require at most $2 \cdot 9n^3$ FLOPs.

In this example, we perform approximately $2 \cdot 19n^3$ FLOPs = $38n^3$ FLOPs and you may compare this to having to perform $3940n^3$ FLOPs if you were to approximate this iteratively. This is approximately 100 times fewer floating-point operations required to calculate the same result.

Thus, we will require approximately $2\lfloor \lg(m) \rfloor n^3 + 2\left(\lfloor \lg(m) \rfloor - 1\right) n^3$ FLOPs or approximately $4\lg(m) n^3$.

For example, if $A=\begin{pmatrix}1 & -0.1 & 0.1\\ 0 & 1 & -0.1\\ 0 & 0 & 1\end{pmatrix}$, then

$$A^2 = AA = \begin{pmatrix}1 & -0.1 & 0.1\\ 0 & 1 & -0.1\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & -0.1 & 0.1\\ 0 & 1 & -0.1\\ 0 & 0 & 1\end{pmatrix} = \begin{pmatrix}1 & -0.2 & 0.21\\ 0 & 1 & -0.2\\ 0 & 0 & 1\end{pmatrix}$$

$$A^4 = A^2A^2 = \begin{pmatrix}1 & -0.2 & 0.21\\ 0 & 1 & -0.2\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & -0.2 & 0.21\\ 0 & 1 & -0.2\\ 0 & 0 & 1\end{pmatrix} = \begin{pmatrix}1 & -0.4 & 0.46\\ 0 & 1 & -0.4\\ 0 & 0 & 1\end{pmatrix}$$

$$A^8 = A^4A^4 = \begin{pmatrix}1 & -0.4 & 0.46\\ 0 & 1 & -0.4\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & -0.4 & 0.46\\ 0 & 1 & -0.4\\ 0 & 0 & 1\end{pmatrix} = \begin{pmatrix}1 & -0.8 & 1.08\\ 0 & 1 & -0.8\\ 0 & 0 & 1\end{pmatrix}$$

$$A^{16} = A^8A^8 = \begin{pmatrix}1 & -0.8 & 1.08\\ 0 & 1 & -0.8\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & -0.8 & 1.08\\ 0 & 1 & -0.8\\ 0 & 0 & 1\end{pmatrix} = \begin{pmatrix}1 & -1.6 & 2.8\\ 0 & 1 & -1.6\\ 0 & 0 & 1\end{pmatrix}$$

$$A^{32} = A^{16}A^{16} = \begin{pmatrix}1 & -1.6 & 2.8\\ 0 & 1 & -1.6\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & -1.6 & 2.8\\ 0 & 1 & -1.6\\ 0 & 0 & 1\end{pmatrix} = \begin{pmatrix}1 & -3.2 & 8.16\\ 0 & 1 & -3.2\\ 0 & 0 & 1\end{pmatrix}$$

so

$$A^{42} = A^{32+8+2} = \begin{pmatrix}1 & -3.2 & 8.16\\ 0 & 1 & -3.2\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & -0.8 & 1.08\\ 0 & 1 & -0.8\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}1 & -0.2 & 0.21\\ 0 & 1 & -0.2\\ 0 & 0 & 1\end{pmatrix} = \begin{pmatrix}1 & -4.2 & 12.81\\ 0 & 1 & -4.2\\ 0 & 0 & 1\end{pmatrix},$$

which is the correct answer. Similarly, we may continue to find

$$A^{64} = \begin{pmatrix}1 & -6.4 & 26.56\\ 0 & 1 & -6.4\\ 0 & 0 & 1\end{pmatrix}, A^{128} = \begin{pmatrix}1 & -12.8 & 94.08\\ 0 & 1 & -12.8\\ 0 & 0 & 1\end{pmatrix}, A^{256} = \begin{pmatrix}1 & -25.6 & 352\\ 0 & 1 & -25.6\\ 0 & 0 & 1\end{pmatrix},$$

$$A^{512} = \begin{pmatrix}1 & -51.2 & 1359.36\\ 0 & 1 & -51.2\\ 0 & 0 & 1\end{pmatrix} \text{ and } A^{1024} = \begin{pmatrix}1 & -102.4 & 5340.16\\ 0 & 1 & -102.4\\ 0 & 0 & 1\end{pmatrix}$$

and so

$$
\begin{aligned}
A^{1970} &= A^{1024\,+\,512\,+\,256\,+\,128\,+\,32\,+\,16\,+\,2} \\
&= A^{1024}A^{512}A^{256}A^{128}A^{32}A^{16}A^{2} \\
&= \begin{pmatrix} 1 & -102.4 & 5340.16 \\ 0 & 1 & -102.4 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -51.2 & 1359.36 \\ 0 & 1 & -51.2 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -25.6 & 352 \\ 0 & 1 & -25.6 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -12.8 & 94.08 \\ 0 & 1 & -12.8 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -3.2 & 8.16 \\ 0 & 1 & -3.2 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -1.6 & 2.8 \\ 0 & 1 & -1.6 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & -0.2 & 0.21 \\ 0 & 1 & -0.2 \\ 0 & 0 & 0 \end{pmatrix} \\
&= \begin{pmatrix} 1 & -197 & 19591.65 \\ 0 & 1 & -197 \\ 0 & 0 & 0 \end{pmatrix}
\end{aligned}
$$

You may consider this calculation to be quite tedious, but this is 100 times less work than would be required if you were to calculate A^{1970} iteratively.

8.10.4 Polynomials of linear operators

Given that powers of A are linear operators, it follows that we could take a linear combination of powers of A to produce a subsequent linear operator; for example,

$$A^2 + 3A + 2I\,.$$

From algebra, you know that the polynomial $x^2 + 3x + 2$ can be factored as $(x + 1)(x + 2)$, as is the case here:

$$
\begin{aligned}
(A+I)(A+2I) &= A(A+2I) + I(A+2I) \\
&= A^2 + A + 2A + 2I \\
&= A^2 + 3A + 2I
\end{aligned}
$$

It is important to understand that $x^2 + 3x + 2$ is to be interpreted as $x^2 + 3x + 2x^0$ when we are evaluating a polynomial at a matrix. Consequently, any polynomial p with coefficients in $\mathbf{F}$ can be evaluated at a matrix $A : \mathbf{F}^n \to \mathbf{F}^n$ producing a new operator $p(A) : \mathbf{F}^n \to \mathbf{F}^n$.

Similarly, we have seen that $\frac{\mathrm{d}}{\mathrm{d}x} : \mathbf{F}[x] \to \mathbf{F}[x]$ and thus, given a polynomial p, we could evaluate it at $\frac{\mathrm{d}}{\mathrm{d}x}$ to get a new operator; for example, if $p : x \mapsto x^2 + 3x + 2$,

$$p\left(\frac{\mathrm{d}}{\mathrm{d}x}\right) = \left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^2 + 3\frac{\mathrm{d}}{\mathrm{d}x} + 2I = \frac{\mathrm{d}^2}{\mathrm{d}x^2} + 3\frac{\mathrm{d}}{\mathrm{d}x} + 2I\,.$$

In your calculus course, you will see that trigonometric and exponential functions can be described as infinite series, or never-ending polynomials. For example, you will see that:

$$
\begin{aligned}
e^z &= 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \frac{z^5}{5!} + \cdots \\
\sin(z) &= z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \frac{z^9}{9!} + \cdots \\
\cos(z) &= 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \frac{z^8}{8!} + \cdots
\end{aligned}
$$

While you are not yet aware of why such series converge to transcendental functions such as the exponential and trigonometric functions, you can already appreciate the formula that $e^{jx} = \cos(x) + j\sin(x)$, for

$$\begin{aligned}
e^{jx} &= 1 + jx + \frac{(jx)^2}{2!} + \frac{(jx)^3}{3!} + \frac{(jx)^4}{4!} + \frac{(jx)^5}{5!} + \frac{(jx)^6}{6!} + \frac{(jx)^7}{7!} + \frac{(jx)^8}{8!} + \frac{(jx)^9}{9!} + \cdots \\
&= 1 + jx + j^2\frac{x^2}{2!} + j^3\frac{x^3}{3!} + j^4\frac{x^4}{4!} + j^5\frac{x^5}{5!} + j^6\frac{x^6}{6!} + j^7\frac{x^7}{7!} + j^8\frac{x^8}{8!} + j^9\frac{x^9}{9!} + \cdots \\
&= 1 + jx - \frac{x^2}{2!} - j\frac{x^3}{3!} + \frac{x^4}{4!} + j\frac{x^5}{5!} - \frac{x^6}{6!} - j\frac{x^7}{7!} + \frac{x^8}{8!} + j\frac{x^9}{9!} + \cdots \\
&= \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} + \cdots\right) + \left(jx - j\frac{x^3}{3!} + j\frac{x^5}{5!} - j\frac{x^7}{7!} + j\frac{x^9}{9!} + \cdots\right) \\
&= \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} + \cdots\right) + j\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} + \cdots\right) \\
&= \cos(x) + j\sin(x)
\end{aligned}$$

Consequently, as we can now evaluate, for example, exponential functions at a linear operator; for example, later we will see that

$$e^{\begin{pmatrix} 1 & -2 \\ -2 & 4 \end{pmatrix}} = \begin{pmatrix} \frac{4}{5} + \frac{1}{5}e^5 & \frac{2}{5} - \frac{2}{5}e^5 \\ \frac{2}{5} - \frac{2}{5}e^5 & \frac{1}{5} + \frac{4}{5}e^5 \end{pmatrix}, \; e^{\begin{pmatrix} 0.6 & -0.8 \\ 0.8 & 0.6 \end{pmatrix}} = \begin{pmatrix} e^{0.6}\cos(0.8) & -e^{0.6}\sin(0.8) \\ e^{0.6}\sin(0.8) & e^{0.6}\cos(0.8) \end{pmatrix} \text{ and } e^{\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}} \approx \begin{pmatrix} 51.969 & 74.737 \\ 111.105 & 164.074 \end{pmatrix}.$$

Definition

An operator $A: \mathcal{U} \to \mathcal{U}$ is said to be invertible if there exists an operator $A^{-1} \in \mathcal{L}(\mathcal{U})$ such that $A^{-1} \circ A = A \circ A^{-1} = Id$.

Theorem

If $A: \mathcal{U} \to \mathcal{U}$ is a linear operator, then A^{-1} as defined above is also a linear operator.

Proof:

If $\mathbf{v}_1 = A\mathbf{u}_1$ and $\mathbf{v}_2 = A\mathbf{u}_2$, we must show that $A^{-1}(\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2) = \alpha_1 A^{-1}\mathbf{v}_1 + \alpha_2 A^{-1}\mathbf{v}_2$. Now,

$$\alpha_1 A^{-1}\mathbf{v}_1 + \alpha_2 A^{-1}\mathbf{v}_2 = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2,$$

and yet

$$A(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2) = \alpha_1 A\mathbf{u}_1 + \alpha_2 A\mathbf{u}_2 = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2.$$

Thus, $A^{-1}(\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2) = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2$; but by definition $\mathbf{u}_1 = A^{-1}\mathbf{v}_1$ and $\mathbf{u}_2 = A^{-1}\mathbf{v}_2$, so

$$A^{-1}(\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2) = \alpha_1 A^{-1}\mathbf{v}_1 + \alpha_2 A^{-1}\mathbf{v}_2.$$

Thus, if A is linear, then so is A^{-1}. ∎

Theorem

If $A:\mathcal{U}\to\mathcal{U}$ is a linear operator, then $(A^{-1})^{-1} = A$.

Proof:

Theorem

If $A:\mathcal{U}\to\mathcal{U}$ is a linear operator and $\alpha\in\mathbf{F}$ where $\alpha\neq 0$, then $(\alpha A)^{-1} = \frac{1}{\alpha}A^{-1}$.

Proof:

From the properties of matrix composition and scalar multiplication:

$$(\alpha A)\left(\frac{1}{\alpha}A^{-1}\right) = \frac{\alpha}{\alpha}\left(AA^{-1}\right) = 1\cdot I = I\,.$$

Thus, the inverse of αA is $\frac{1}{\alpha}A^{-1}$. ∎

8.10.5 Properties of linear operators for finite-dimensional vector spaces (square matrices)

If $A:\mathbf{F}^n\to\mathbf{F}^n$, then A may be represented as a square matrix. Recall that a matrix is one-to-one if the rank equals the dimension of the domain, and onto if the rank equals the dimension of the co-domain. Because the domain and co-domain are the same, it follows that a square $n \times n$ matrix is

1. one-to-one if rank(A) = n, and
2. onto if rank(A) = n.

Thus, for square matrices, either a matrix is both one-to-one and onto if rank(A) = n, otherwise it is neither. We will see that for $A:\mathbf{F}^n\to\mathbf{F}^n$, then an inverse exists if and only if rank(A) = n. After all, if **v** is in the co-domain of A, then there must be a unqiue vector **u** in A such that $A\mathbf{u} = \mathbf{v}$. Let A^{-1} be that operator defined as $A^{-1}\mathbf{v} = \mathbf{u}$. The question is, is A^{-1} a linear map?

8.10.6 Summary of linear operators

We can summarize and highlight the similarities and differences between fields, vector spaces and algebras in the next table.

	Addition	Scalar multiplication	Composition	Comments
Fields	addition	multiplication	*not applicable*	Addition and multiplication are associative and commutative. There is an additive identity 0 and a multiplicative identity 1. All elements have additive inverses, and all non-zero elements have multiplicative inverses.
Vector spaces	vector addition	scalar multiplication	*not defined*	Vector addition is associative and commutative. There is an additive identity **0**. All vectors have additive inverses.
Algebras	operator addition	scalar multiplication	operator composition	Operator addition and composition are associative, but only operator addition is necessarily commutative. There is an additive identity O and a composition identity I. All operators have additive inverses, but not all operators have inverses for composition.

Notice that we offer proofs of certain ideas, but we cannot prove, for example, that composition is not commutative or that not every operator has an inverse. For this, we must actually look at concrete examples.

For example, let us consider $A \in \mathcal{L}\left(\mathbf{R}^2\right)$. We already know that not all matrices are invertible. We already know that this is representable by the collection of all 2 × 2 matrices. First, matrix-multiplication is not commutative, for

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \text{ but } \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

This essentially says that the order in which you apply linear systems may affect the output, summarized in .

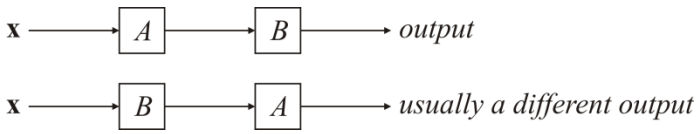


Figure 84. Composition, and therefore matrix multiplication, is not necessarily commutative.

You can now understand why this course is called *linear algebra*—it is the study of algebras of linear operators on vector spaces.

8.11 Row operations

In finite-dimensional vector spaces, a row operation can be represented by a matrix (i.e., a linear mapping), and each operation can be interpreted as a generalization of a physical effect. In each case, the matrix representing the row operation can be found by performing the row operation on the identity matrix.

8.11.1 Swapping two rows

Given a matrix $A : \mathbf{F}^n \to \mathbf{F}^m$, the row operation of swapping two rows, represented by $R_{i \leftrightarrow j}$, is equivalent to multiplying the matrix on the left by the $m \times m$ matrix consisting of the identity matrix with the rows i and j swapped. That is,

$$R_{i \leftrightarrow j} = \begin{array}{c} \\ 1 \\ \vdots \\ \\ i \\ \\ \vdots \\ \\ j \\ \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{cccccccccccc} 1 & \cdots & & i & & \cdots & & j & & \cdots & m \end{array} \\ \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array}.$$

For example, suppose we wish to swap the 1st and 3rd rows of the matrix

$$A = \begin{pmatrix} 2 & 1 & -2 & 3 & -1 & 0 \\ 1 & 2 & 4 & -2 & 1 & -2 \\ 5 & -2 & 3 & 1 & 2 & 3 \\ -3 & 4 & 2 & -1 & 3 & 1 \end{pmatrix},$$

in which case, we would multiply on the left by the matrix

$$R_{1 \leftrightarrow 3} A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & -2 & 3 & -1 & 0 \\ 1 & 2 & 4 & -2 & 1 & -2 \\ 5 & -2 & 3 & 1 & 2 & 3 \\ -3 & 4 & 2 & -1 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 5 & -2 & 3 & 1 & 2 & 3 \\ 1 & 2 & 4 & -2 & 1 & -2 \\ 2 & 1 & -2 & 3 & -1 & 0 \\ -3 & 4 & 2 & -1 & 3 & 1 \end{pmatrix}.$$

If we wanted to swap the 3rd and 6th entries of the vector

$$\mathbf{v} = \begin{pmatrix} 3.2 \\ -1.2 \\ 0.5 \\ 0.7 \\ -2.3 \\ 4.2 \end{pmatrix},$$

we could multiply it on the left by

$$R_{3\leftrightarrow 6}\mathbf{v} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 3.2 \\ -1.2 \\ 0.5 \\ 0.7 \\ -2.3 \\ 4.2 \end{pmatrix} = \begin{pmatrix} 3.2 \\ -1.2 \\ 4.2 \\ 0.7 \\ -2.3 \\ 0.5 \end{pmatrix}.$$

This can be seen as reflecting the space in the plane defined by $u_i = u_j$.

8.11.2 Multiplying a row by a scalar

Given a matrix $A: \mathbf{F}^n \to \mathbf{F}^m$, the row operation of multiplying a row by a non-zero scalar α, represented by $R_{\alpha;i}$, is equivalent to multiplying the matrix on the left by the $m \times m$ matrix consisting of the identity matrix with the ith diagonal entry set to α. That is,

$$R_{\alpha;i} = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & \cdots & & i & & \cdots & m \end{array} \\ \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \alpha & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array}.$$

For example, multiplying the 3rd row of the matrix A above by 2.5 would be performed by

$$R_{0.2;3}A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & -2 & 3 & -1 & 0 \\ 1 & 2 & 4 & -2 & 1 & -2 \\ 5 & -2 & 3 & 1 & 2 & 3 \\ -3 & 4 & 2 & -1 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 & -2 & 3 & -1 & 0 \\ 1 & 2 & 4 & -2 & 1 & -2 \\ 1 & -0.4 & 0.6 & 0.2 & 0.4 & 0.6 \\ -3 & 4 & 2 & -1 & 3 & 1 \end{pmatrix}.$$

Similarly, if we wanted to multiply the 2nd entry of the above vector $\mathbf{v}$ by –3.5, we would multiply on the left by

$$R_{-3.5;2}\mathbf{v} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -3.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3.2 \\ -1.2 \\ 0.5 \\ 0.7 \\ -2.3 \\ 4.2 \end{pmatrix} = \begin{pmatrix} 3.2 \\ 4.2 \\ 0.5 \\ 0.7 \\ -2.3 \\ 4.2 \end{pmatrix}.$$

This can be interpreted as stretching, contracting or reflecting the vector space in the ith dimension.

8.11.3 Adding a multiple of one row onto another

Finally, the last row operation was adding a multiple α of one Row i onto Row j. This is represented by the identity matrix with the $(j, i)^{\text{th}}$ entry set to α. You can remember this by simply performing the row operation on the identity matrix:

$$R_{\alpha;i\to j} = \begin{array}{c} \\ 1 \\ \vdots \\ \\ i \\ \\ \vdots \\ \\ j \\ \\ \vdots \\ m \end{array} \begin{array}{c} \begin{array}{ccccccccccc} 1 & \cdots & & i & & \cdots & & j & & \cdots & m \end{array} \\ \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \alpha & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array}$$

$$R_{-2.5;1\to 3}A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2.5 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & -2 & 3 & -1 & 0 \\ 1 & 2 & 4 & -2 & 1 & -2 \\ 5 & -2 & 3 & 1 & 2 & 3 \\ -3 & 4 & 2 & -1 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 & -2 & 3 & -1 & 0 \\ 1 & 2 & 4 & -2 & 1 & -2 \\ 0 & -4.5 & 8 & -6.5 & 4.5 & 3 \\ -3 & 4 & 2 & -1 & 3 & 1 \end{pmatrix}.$$

Similarly, adding

$$R_{2.5;2\to 5}\mathbf{v} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1.2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 3.2 \\ -1.2 \\ 0.5 \\ 0.7 \\ -2.3 \\ 4.2 \end{pmatrix} = \begin{pmatrix} 3.2 \\ -1.2 \\ 4.2 \\ 0.7 \\ -5.3 \\ 0.5 \end{pmatrix}$$

This can be seen as a shear in the jth dimension.

Problems

1. Given the matrix $A = \begin{pmatrix} 1 & 2 & -2 \\ 3 & 1 & 2 \\ 4 & -2 & 1 \\ 4 & 2 & 1 \\ 4 & 6 & -2 \end{pmatrix}$, what are the matrices corresponding to $R_{2\leftrightarrow 3}$, $R_{-5.9;3}$ and $R_{4.7;2\to 3}$.

2. Given the matrix $B = \begin{pmatrix} 2 & -3 & -4 & 2 & 0 & 5 & 2 \\ -1 & 1 & 3 & -2 & -1 & 4 & -2 \\ 4 & -2 & 1 & 5 & 4 & -1 & 2 \end{pmatrix}$, what are the matrices corresponding to $R_{1\leftrightarrow 2}$, $R_{4.8;2}$ and $R_{6.3;1\to 2}$.

3. Identify the following matrices corresponding to row operations, and give as much information as you can regarding the matrices to which these operations can apply.

$$\begin{pmatrix} 1 & 0 & 2.7 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 8.1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

4. Identify the following matrices corresponding to row operations, and give as much information as you can regarding the matrices to which these operations can apply.

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -0.5 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 9.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Solutions

1. As A is 5×3, the matrices are

$$R_{2\leftrightarrow 3} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, R_{-5.9;3} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -5.9 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } R_{4.7;2\to 3} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 4.7 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

3. Add 2.7 times Row 3 onto Row 1, swap Rows 1 and 4, and multiply Row 1 by 8.1. As these matrices are 4×4, they apply to row operations applied to $4 \times n$ matrices.

8.12 Gaussian elimination

Recall now that Gaussian elimination is a sequence of row operations. For example, performing Gaussian elimination on the matrix $A = \begin{pmatrix} 3 & 6 & 9 \\ 2 & 5 & 8 \\ 1 & 4 & 7 \end{pmatrix}$ is as follows:

$$R_{1\leftrightarrow 3}A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix} = \begin{pmatrix} 3 & 6 & 9 \\ 2 & 5 & 8 \\ 1 & 4 & 7 \end{pmatrix}.$$

Adding $-\frac{2}{3}$ times Row 1 onto Row 2 of the result is like multiplying each of these matrices by $R_{-\frac{2}{3};1\to 2}$, and thus

$$R_{-\frac{2}{3};1\to 2}R_{1\leftrightarrow 3}A=\begin{pmatrix}1&0&0\\-\frac{2}{3}&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}\begin{pmatrix}1&4&7\\2&5&8\\3&6&9\end{pmatrix}=\begin{pmatrix}1&0&0\\-\frac{2}{3}&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}3&6&9\\2&5&8\\1&4&7\end{pmatrix}=\begin{pmatrix}3&6&9\\0&1&2\\1&4&7\end{pmatrix}.$$

Now, adding $-\frac{1}{3}$ times Row 1 onto Row 3 of the result is like multiplying each of these matrices by $R_{-\frac{1}{3};1\to 3}$, and thus

$$R_{-\frac{1}{3};1\to 3}R_{-\frac{2}{3};1\to 2}R_{1\leftrightarrow 3}A=\begin{pmatrix}1&0&0\\0&1&0\\-\frac{1}{3}&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\-\frac{2}{3}&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}\begin{pmatrix}1&4&7\\2&5&8\\3&6&9\end{pmatrix}=\begin{pmatrix}1&0&0\\0&1&0\\-\frac{1}{3}&0&1\end{pmatrix}\begin{pmatrix}3&6&9\\0&1&2\\1&4&7\end{pmatrix}=\begin{pmatrix}3&6&9\\0&1&2\\0&2&4\end{pmatrix}$$

.

Next, we swap Row 2 and Row 3:

$$\begin{aligned}R_{2\leftrightarrow 3}R_{-\frac{1}{3};1\to 3}R_{-\frac{2}{3};1\to 2}R_{1\leftrightarrow 3}A&=\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\-\frac{1}{3}&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\-\frac{2}{3}&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}\begin{pmatrix}1&4&7\\2&5&8\\3&6&9\end{pmatrix}\\&=\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}\begin{pmatrix}3&6&9\\0&1&2\\0&2&4\end{pmatrix}=\begin{pmatrix}3&6&9\\0&2&4\\0&1&2\end{pmatrix}\end{aligned}$$

and finally, we add $-\frac{1}{2}$ times Row 2 onto Row 3, so

$$\begin{aligned}R_{-\frac{1}{2};2\to 3}R_{2\leftrightarrow 3}R_{-\frac{1}{3};1\to 3}R_{-\frac{2}{3};1\to 2}R_{1\leftrightarrow 3}A&=\begin{pmatrix}1&0&0\\0&1&0\\0&-\frac{1}{2}&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\-\frac{1}{3}&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\-\frac{2}{3}&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}\begin{pmatrix}1&4&7\\2&5&8\\3&6&9\end{pmatrix}\\&=\begin{pmatrix}1&0&0\\0&1&0\\0&-\frac{1}{2}&1\end{pmatrix}\begin{pmatrix}3&6&9\\0&2&4\\0&1&2\end{pmatrix}=\begin{pmatrix}3&6&9\\0&2&4\\0&0&0\end{pmatrix}\end{aligned}$$

Thus, we conclude that

$$R_{-\frac{1}{2};2\to 3}R_{2\leftrightarrow 3}R_{-\frac{1}{3};1\to 3}R_{-\frac{2}{3};1\to 2}R_{1\leftrightarrow 3}A=\begin{pmatrix}3&6&9\\0&2&4\\0&0&0\end{pmatrix}.$$

8.13 Summary of linear maps and operators

In this chapter, we have looked at linear maps and operators in general. We have defined the range and null space of linear mappings, and seen that for linear mappings on finite-dimensional vector spaces, the dimension of the null space plus the dimension of the range must equal the dimension of the domain. We saw that all linear mappings between finite-dimensional vectors spaces are representable by matrices, and the operation is matrix-vector multiplication. We saw that you could add and multiply linear mappings by scalars, thus creating a vector space of linear mappings, but we can also define the composition of linear mappings, this being matrix-matrix multiplication for the composition of linear mappings on finite-dimensional vector spaces. Finally, if we consider only those linear operators mapping a vector space onto itself, we can discuss concepts such as polynomials of operators and operator inverses.

9 Matrix decompositions

You are already aware of prime decompositions of integers: every integer can be written as a product of prime numbers, so for example, 15 = 3·5. Note that if we are trying to solve a problem like:

$$(ab)x = y,$$

one approach is to simply find the multiplicative inverse of *ab*, and to then multiply both sides by that multiplicative inverse:

$$x = \frac{1}{ab} y .$$

However, another approach may be to rewrite the problem as

$$a(bx) = y,$$

to find the multiplicative inverse of *a*,

$$bx = \frac{1}{a} y ,$$

and having found *bx*, now solve both sides by multiplying both sides by the inverse of *b*:

$$x = \frac{1}{b}\frac{1}{a} y .$$

Both approaches work the same way. With matrices, this is similar: if

$$A\mathbf{u} = \mathbf{v} ,$$

then multiplying both sides on the left by A^{-1} yields

$$\mathbf{u} = A^{-1}\mathbf{v} .$$

Alternatively, we could perform Gaussian elimination and backward substitution on $A\mathbf{u} = \mathbf{v}$. If we're very fortunate and *A* is already in row-echelon form, then solving $A\mathbf{u} = \mathbf{v}$ is actually very fast, as we need only perform backward substitution.

There is a second case where solving a system of linear equations is exceptionally fast: when a matrix is in *reverse row-echelon form*. That is, when the shape of the matrix is similar to

$$\begin{pmatrix} * & 0 & 0 & 0 & 0 \\ * & * & * & 0 & 0 \\ * & * & * & * & 0 \\ * & * & * & * & * \end{pmatrix} \text{ or } \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ * & 0 & 0 & 0 & 0 \\ * & * & * & 0 & 0 \\ * & * & * & * & * \end{pmatrix}.$$

In this case, we may use forward substitution to solve for the system. For example, if we are trying to solve $L\mathbf{u} = \mathbf{v}$ where

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.2 & 1 & 0 & 0 \\ -0.1 & 0.4 & 1 & 0 \\ 0.3 & 0.5 & 0.1 & 1 \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} 2 \\ -2.6 \\ -0.4 \\ 3.2 \end{pmatrix},$$

we could write this as an augmented matrix and immediately solve:

$$\left(\begin{array}{cccc|c} 1 & 0 & 0 & 0 & 2 \\ 0.2 & 1 & 0 & 0 & -2.6 \\ -0.1 & 0.4 & 1 & 0 & -0.4 \\ 0.3 & 0.5 & 0.1 & 1 & 3.2 \end{array}\right)$$

which yields that $u_1 = 2$, substituting this into the second equation, $0.2 \cdot u_1 + u_2 = -2.6$, so $u_2 = -3$, substituting these into the third equation yields $-01 \cdot u_1 + 0.4 \cdot u_2 + u_3 = -0.4$, so $u_3 = 1$, and finally, substituting these three values into the last equation yields that $03 \cdot u_1 + 0.5 \cdot u_2 + 0.1 \cdot u_3 + u_4 = 3.2$, so $u_4 = 4$. Thus, the solution to $L\mathbf{u} = \mathbf{v}$ is

$$\mathbf{u} = \begin{pmatrix} 2 \\ -3 \\ 1 \\ 4 \end{pmatrix}.$$

Again, this does not require any of the difficult process of performing a *reverse* Gaussian elimination, as the matrix is already in the appropriate form. Fortunately, there is a theorem that says that every $m \times n$ matrix can be written as the product of

1. a permutation matrix,
2. a lower triangular $m \times m$ matrix with ones on the diagonal, and
3. a matrix that is in row-echelon form.

We will not actually go through the algorithm of proving this, but recall that every single matrix can be converted to row-echelon form through a series of row operations, and as each row operation can be represented by a matrix, we can therefore write

$$R_N R_{N-1} \cdots R_3 R_2 R_1 A = U$$

where A is the original matrix, U is in row-echelon form (*upper triangular*) and each R is a row operation matrix. Note also that

$$R_{j \leftrightarrow k} R_{\alpha; i \to j} = R_{\alpha; i \to k} R_{j \leftrightarrow k}$$

if $i < j$ and therefore, we may rewrite this sequence of row operations as sequence

$$\left(R_{\alpha_1; i_1' \to j_1'} \cdots R_{\alpha_{N_1}; i_{N_1}' \to j_{N_1}'}\right)\left(R_{i_1'' \leftrightarrow j_1''} \cdots R_{i_{N_2}'' \leftrightarrow j_{N_2}''}\right) A = U,$$

and then applying the inverse of each of these operations,

$$A = \underbrace{\left(R_{i''_{N_2} \leftrightarrow j''_{N_2}} \cdots R_{i''_1 \leftrightarrow j''_1}\right)}_{P}\underbrace{\left(R_{-\alpha_{N_1};i'_{N_1} \to j'_{N_1}} \cdots R_{-\alpha_1;i'_1 \to j'_1}\right)}_{L}U \ .$$

The product of the first N_2 matrices is a permutation matrix, and the product of the remaining matrices is a lower triangular matrix with all ones on the diagonal. Because all of the row operations of adding a multiple of one row onto another involve adding a multiple of less than or equal to one in absolute value, all entries below the diagonal will be less than or equal to one in absolute value.

9.1 Finding *P*, *L* and *U*

Finding these three matrices is, again, a systematic algorithm. We will define a special augmented matrix composed of one matrix with

$$P^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \ L_{\blacktriangle} = \begin{pmatrix} * & 0 & \cdots & 0 \\ 0 & * & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & * \end{pmatrix}, \text{ and } U = A.$$

Apply Gaussian elimination with partial pivoting to the matrix A, and for each operation applied to the matrix U:

1. if it is a row swap operation, apply it to both P^{-1} and to L, not touching the starred entries, and
2. if it is an operation adding α times Row i onto j, change the $(j, i)^{\text{th}}$ entry of L to $-\alpha$.

Once U has been converted to row-echelon form, switch each diagonal entry of L to 1, and this will give the three matrices P^{-1}, L and U such that $A = PLU$.

For example, given the matrix

$$A = \begin{pmatrix} 1.5 & 6.6 & -2.1 \\ -1 & 2 & 2.2 \\ 5 & 2 & 3 \end{pmatrix},$$

we would define

$$P^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \ L_{\blacktriangle} = \begin{pmatrix} * & 0 & 0 \\ 0 & * & 0 \\ 0 & 0 & * \end{pmatrix}, \text{ and } U_{\blacktriangledown} = \begin{pmatrix} 1.5 & 6.6 & -2.1 \\ -1 & 2 & 2.2 \\ 5 & 2 & 3 \end{pmatrix}.$$

Apply the rules of Gaussian elimination and partial pivoting, we begin by swapping Rows 1 and 3:

$$P^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \ L_{\blacktriangle} = \begin{pmatrix} * & 0 & 0 \\ 0 & * & 0 \\ 0 & 0 & * \end{pmatrix}, \text{ and } U_{\blacktriangledown} = \begin{pmatrix} 5 & 2 & 3 \\ -1 & 2 & 2.2 \\ 1.5 & 6.6 & -2.1 \end{pmatrix}.$$

Next, add 0.2 times Row 1 onto Row 2, storing –0.2 in the entry (2, 1) of L:

$$P^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \ L_{\blacktriangle} = \begin{pmatrix} * & 0 & 0 \\ -0.2 & * & 0 \\ 0 & 0 & * \end{pmatrix}, \text{ and } U = \begin{pmatrix} 5 & 2 & 3 \\ 0 & 2.4 & 2.8 \\ 1.5 & 6.6 & -2.1 \end{pmatrix}.$$

Next, add –0.3 times Row 1 onto Row 3, storing 0.3 in the entry (3,1) of L:

$$P^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \; L = \begin{pmatrix} * & 0 & 0 \\ -0.2 & * & 0 \\ 0.3 & 0 & * \end{pmatrix}, \text{ and } U = \begin{pmatrix} 5 & 2 & 3 \\ 0 & 2.4 & 2.8 \\ 0 & 6 & -3 \end{pmatrix}.$$

Next, we swap Rows 2 and 3—in all three matrices, but not touching the diagonal entries of L:

$$P^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \; L = \begin{pmatrix} * & 0 & 0 \\ 0.3 & * & 0 \\ -0.2 & 0 & * \end{pmatrix}, \text{ and } U = \begin{pmatrix} 5 & 2 & 3 \\ 0 & 6 & -3 \\ 0 & 2.4 & 2.8 \end{pmatrix}.$$

Finally, add –0.4 times Row 2 onto Row 3, storing 0.4 in the entry (3,2) of L:

$$P^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \; L = \begin{pmatrix} * & 0 & 0 \\ 0.3 & * & 0 \\ -0.2 & 0.4 & * \end{pmatrix}, \text{ and } U = \begin{pmatrix} 5 & 2 & 3 \\ 0 & 6 & -3 \\ 0 & 0 & 4 \end{pmatrix}.$$

You will now note that

$$PLU = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0.3 & 1 & 0 \\ -0.2 & 0.4 & 1 \end{pmatrix} \begin{pmatrix} 5 & 2 & 3 \\ 0 & 6 & -3 \\ 0 & 0 & 4 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 5 & 2 & 3 \\ 1.5 & 6.6 & -2.1 \\ -1 & 2 & 2.2 \end{pmatrix} = \begin{pmatrix} 1.5 & 6.6 & -2.1 \\ -1 & 2 & 2.2 \\ 5 & 2 & 3 \end{pmatrix}.$$

Problems

1. Find the PLU decomposition of the matrix

$$\begin{pmatrix} -1.5 & -0.5 & -1.3 & -6.9 \\ 5 & 2 & 3 & 4 \\ 1 & 1.4 & -3.4 & 3.8 \end{pmatrix}.$$

2. Find the PLU decomposition of the matrix

$$\begin{pmatrix} -1.5 & -0.3 & 3.5 \\ -0.5 & 1 & -1.7 \\ 1 & 3.4 & -1.8 \\ 5 & 2 & 1 \end{pmatrix}.$$

3. How could you simplify the PLU decomposition of the previous question?

4. Find the PLU decomposition of the matrix

$$\begin{pmatrix} 0.4 & 3.3 & 9 & 20.5 \\ 1 & 2 & 3 & 4 \\ 0.1 & 5.2 & 6.3 & 7.4 \\ 0.2 & 1.9 & 10.4 & 11.9 \end{pmatrix}.$$

5. What is the PLU decomposition of an upper-triangular matrix A?

6. What is a condition for the PLU decomposition of a lower-triangular matrix A to have an upper-triangular matrix that is simply a matrix with the diagonal entries of A on the diagonal of U.

Solutions

1. Starting with the three candidate matrices for P^{-1}, $L_\blacktriangleleft$ and $U = A$, we apply the rules of Gaussian elimination with partial pivoting, storing the inverse of the sheer operations in $L_\blacktriangleleft$ and applying row swaps to all three matrices.

$$\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},\begin{pmatrix}*&0&0\\0&*&0\\0&0&*\end{pmatrix},\begin{pmatrix}-1.5&-0.5&-1.3&-6.9\\5&2&3&4\\1&1.4&-3.4&3.8\end{pmatrix}$$

$$\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix},\begin{pmatrix}*&0&0\\0&*&0\\0&0&*\end{pmatrix},\begin{pmatrix}5&2&3&4\\-1.5&-0.5&-1.3&-6.9\\1&1.4&-3.4&3.8\end{pmatrix}$$

$$\begin{pmatrix}0&1&0\\1&0&0\\0&0&1\end{pmatrix},\begin{pmatrix}*&0&0\\-0.3&*&0\\0.2&0&*\end{pmatrix},\begin{pmatrix}5&2&3&4\\0&0.1&-0.4&-5.7\\0&1&-4&3\end{pmatrix}$$

$$\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix},\begin{pmatrix}*&0&0\\0.2&*&0\\-0.3&0&*\end{pmatrix},\begin{pmatrix}5&2&3&4\\0&1&-4&3\\0&0.1&-0.4&-5.7\end{pmatrix}$$

$$\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix},\begin{pmatrix}*&0&0\\0.2&*&0\\-0.3&0.1&*\end{pmatrix},\begin{pmatrix}5&2&3&4\\0&1&-4&3\\0&0&0&-6\end{pmatrix}.$$

Therefore,

$$P^{-1}=\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix},L_\blacktriangleleft=\begin{pmatrix}1&0&0\\0.2&1&0\\-0.3&0.1&1\end{pmatrix},U=\begin{pmatrix}5&2&3&4\\0&1&-4&3\\0&0&0&-6\end{pmatrix}.$$

2. The solution is

$$P^{-1}=\begin{pmatrix}0&0&1&0\\0&0&0&1\\0&1&0&0\\1&0&0&0\end{pmatrix},L_\blacktriangleleft=\begin{pmatrix}1&0&0&0\\0.2&1&0&0\\-0.3&0.1&1&0\\-0.1&0.4&-0.2&1\end{pmatrix},U=\begin{pmatrix}5&2&1\\0&3&-2\\0&0&4\\0&0&0\end{pmatrix}.$$

3. Note that L is 4 × 4 and U is 4 × 3, but as the last row of U is all zeros, it makes no contribution to the product LU. Thus, stripping off the last column of L and the last row of U, as in

$$P^{-1} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, L = \begin{pmatrix} 1 & 0 & 0 \\ 0.2 & 1 & 0 \\ -0.3 & 0.1 & 1 \\ -0.1 & 0.4 & -0.2 \end{pmatrix}, U = \begin{pmatrix} 5 & 2 & 1 \\ 0 & 3 & -2 \\ 0 & 0 & 4 \end{pmatrix},$$

gives the same matrix decomposition as the previously calculated result.

4. The solution is

$$P^{-1} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}, L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.1 & 1 & 0 & 0 \\ 0.2 & 0.3 & 1 & 0 \\ 0.4 & 0.5 & 0.6 & 1 \end{pmatrix}, U = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 5 & 6 & 7 \\ 0 & 0 & 8 & 9 \\ 0 & 0 & 0 & 10 \end{pmatrix}.$$

5. The PLU decomposition of an upper triangular matrix is $P^{-1} = Id_n, L = Id_n, U = A$.

6. The largest entries in each column of *A* must be on the diagonal.

10 Linear operators mapping a space onto itself

In the previous section, we described general linear mappings $A:\mathcal{U}\to\mathcal{V}$ where $\mathcal{U}$ and $\mathcal{V}$ are both vector spaces. There are, however, many cases where we will consider linear operators that map a vector space onto itself; that is, $A:\mathcal{U}\to\mathcal{U}$. In this case, we may be able to more succinctly describe some of the properties of that linear operator. We will consider:

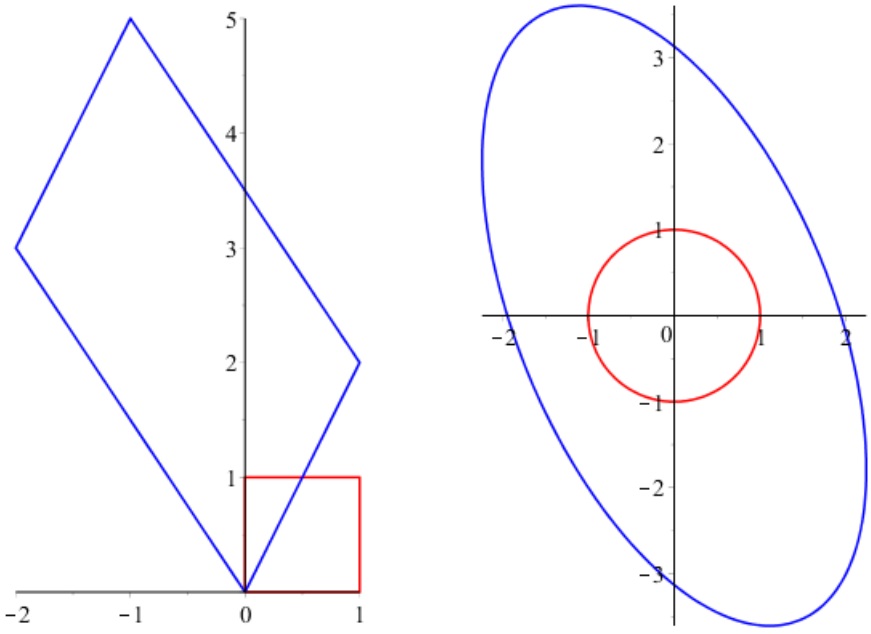
1. the determinant,
2. the trace, and
3. the inverse.

10.1 Determinant

One of our goals is to characterize the behavior of a linear operator in simpler terms. We have already described eigenvalues and eigenvectors, but suppose we have a linear operator $A:\mathbf{R}^n\to\mathbf{R}^n$. If we take any region *R* that has a non-zero *n*-dimensional volume[26], and we consider the image of that region, often written as *AR*, we will start to see a trend: no matter what *n*-volume we choose, the ratio between the *n*-volumes is always the same.

10.1.1 Defining the determinant

As a concrete example, consider the matrix $A=\begin{pmatrix}1 & -2\\ 2 & 3\end{pmatrix}$. If we take a unit square, this gets mapped into a parallelogram of volume 7. If we take the unit circle, this gets mapped into an ellipse of area 7π. In all cases, given a region *R* with a volume vol(*R*), then if find the image of each point in the region (i.e., multiply each vector in the region *R* by the matrix), the volume of the image is always seven times larger.



[26] In $\mathbf{R}^1$, the 1-dimensional volume of an interval is the *width*; in $\mathbf{R}^2$, the 2-dimensional volume of a region is the *area*, and in $\mathbf{R}^3$, the 3-dimesnional volume of a region is what is normally referred to as the *volume*. We will, however, in general simply refer to all such measures as the "volume".

Another characteristic of a linear operator is that it may either preserve or invert orientation. For example, you see a linear transformation when you look in a mirror—that is, a refection—and yet, in the mirror, an object that is in your right hand would appear to be in your left. If you view the reflection in the mirror through a second mirror, the object appears—once again—to be in your right hand.

Definition

The determinant of a linear operator $A:\mathbf{R}^n \to \mathbf{R}^n$ is that number such that

1. the absolute value of the determinant is the change in volume of the operator, and
2. the sign of the determinant is positive or negative depending on whether or not the linear operator preserves or inverts orientation.

The determinant of a linear operator A will be written as det(A).

Theorem

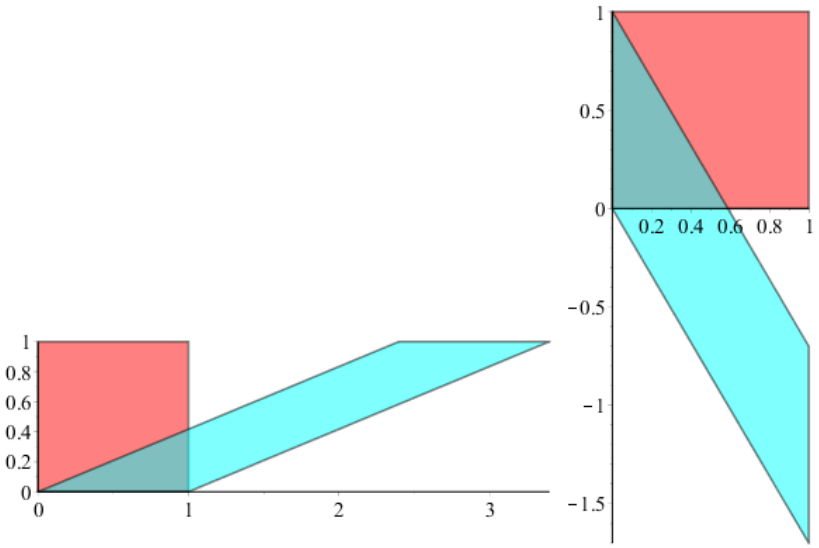
The determinant must satisfy the property that if $A:\mathbf{R}^n \to \mathbf{R}^n$ and $B:\mathbf{R}^n \to \mathbf{R}^n$, then det($BA$) = det($B$) det($A$).

Theorem

For row operations, $\det\left(R_{\alpha;i\to j}\right)=1$, $\det\left(R_{i\leftrightarrow j}\right)=-1$ and $\det\left(R_{\alpha;i}\right)=\alpha$.

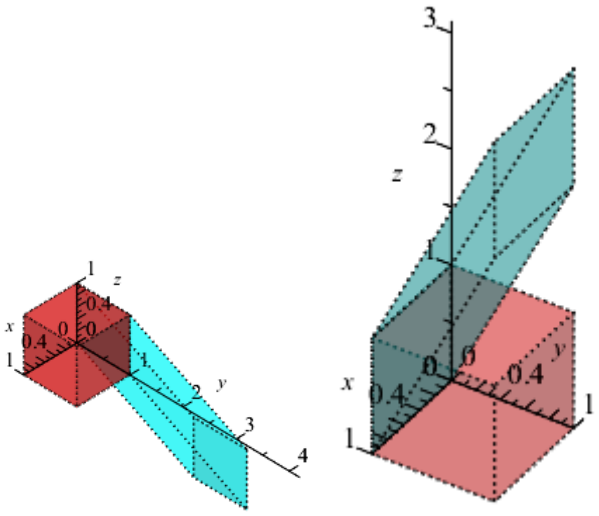
Sketch of a proof:

Adding a multiple of one row onto another does not stretch a volume in any direction. For example, the unit square (in red) and the image of the unit square after being multiplied by the matrices $A=\begin{pmatrix}1 & 2.4\\ 0 & 1\end{pmatrix}$ and $B=\begin{pmatrix}1 & 0\\ -1.7 & 1\end{pmatrix}$ are shown, respectively, in the following two plots:



In both cases, the image of the unit square is a parallelogram that also has an area equal to 1.

Similarly, finding the image of the unit cube in $\mathbf{R}^3$ under the mappings



Recall that with the PLU decomposition, we were able to, through row operations, write a matrix A as a product of a permutation matrix P, a lower-triangular matrix L and an upper triangular matrix U. Thus, if a determinant exists, it must be true that

$$\det(A) = \det(PLU) = \det(P)\det(L)\det(U).$$

The permutation matrix is made up of a composition of reflections, so if n_P reflections were required, $\det(P) = (-1)^{n_P}$. The lower-triangular matrix, however, is the product of n_L sheer matrices, and thus

$$\begin{aligned}\det(L) &= \det\left(R_{\alpha_1;i_1\to j_1}R_{\alpha_2;i_2\to j_2}\cdots R_{\alpha_{n_L};i_{n_L}\to j_{n_L}}\right)\\ &= \det\left(R_{\alpha_1;i_1\to j_1}\right)\det\left(R_{\alpha_2;i_2\to j_2}\right)\cdots\det\left(R_{\alpha_{n_L};i_{n_L}\to j_{n_L}}\right)\\ &= 1^{n_L} = 1\end{aligned}$$

Now, without proof, we know that U is an upper-triangular matrix. It is possible to write U as

$$U = R_{\alpha_1;i_1\to j_1}R_{\alpha_2;i_2\to j_2}\cdots R_{\alpha_{n_L};i_{n_U}\to j_{n_U}}D$$

where D is a diagonal matrix with all of the same entries of U; that is, $d_{i,i} = u_{i,i}$ for $i = 1, \ldots, n$. Thus,

$$\begin{aligned}\det(U) &= \det\left(R_{\alpha_1;i_1\to j_1}R_{\alpha_2;i_2\to j_2}\cdots R_{\alpha_{n_L};i_{n_U}\to j_{n_U}}D\right)\\ &= \det\left(R_{\alpha_1;i_1\to j_1}\right)\det\left(R_{\alpha_2;i_2\to j_2}\right)\cdots\det\left(R_{\alpha_{n_L};i_{n_U}\to j_{n_U}}D\right)\det(D)\\ &= \det(D)\end{aligned}$$

but the determinant of a diagonal matrix is the product of the diagonals.

1. swapping rows *j* and *k*: $\left|\mathbf{R}_{j\leftrightarrow k}\right| = -1$,
2. adding a multiple *c* of row *j* onto row *k*: $\left|\mathbf{R}_{c,\, j\to k}\right| = 1$ (a sheer operation), and
3. multiplying a row *j* by a scalar *c*: $\left|\mathbf{R}_{c,\, j}\right| = c$.

Recall that when we reduce a matrix to row-echelon form, we apply the first two row operations, and thus $\mathbf{M}_{\text{REF}} = \mathbf{R}_n \mathbf{R}_{n-1} \cdots \mathbf{R}_2 \mathbf{R}_1 \mathbf{M}$, where $n_{\text{shear}} + n_{\text{swap}} = n$.

Theorem
Given a matrix $\mathbf{M}$, if we apply only the operations of adding a multiple of one row onto another (a sheer operation) and n_{swap} swaps of two rows, and $\mathbf{M}_{\text{REF}}$, then $\left|\mathbf{M}\right| = (-1)^{n_{\text{swap}}} \left|\mathbf{M}_{\text{REF}}\right|$.

The determinant of a matrix is zero if and only if the column vectors of the matrix are linearly dependent.

Recall that a set of *n* vectors is linearly independent if and only if the only solution to

Thus, recall from our previous exercise that

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 0 \end{pmatrix}$$

and thus, the determinant of both these matrices are zero.

Thus, recall from our previous exercise that

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 3 \\ 0 & -3 & -6 \\ 0 & 0 & 0 \end{pmatrix}$$

and thus, the determinant of both these matrices are zero.

Theorem
The determinant of a diagonal matrix is the product of the diagonal entries.

Proof:
The tridiagonal matrix **D** is the product of the row multiplication matrices

$$\mathbf{D} = \prod_{k=1}^{n} \mathbf{R}_{d_{k,k};k}$$

and therefore

$$\left|\mathbf{D}\right| = \left|\prod_{k=1}^{n} \mathbf{R}_{d_{k,k};k}\right| = \prod_{k=1}^{n} \left|\mathbf{R}_{d_{k,k};k}\right| = \prod_{k=1}^{n} d_{k,k} \text{ . } \blacksquare$$

Theorem
The determinant of an upper-triangular matrix with non-zero entries on the diagonal is the product of the diagonal entries.

Proof:
Every upper-triangular matrix may be transformed to a diagonal matrix through a series of sheer operations, and as these leave the diagonal unchanged, we may consequently deduce that the determinant of an upper-triangular matrix is equal to the determinant of a matrix containing just the diagonal entries. From the previous theorem, we may therefore deduce that the determinant equals the determinant of the diagonal entries. ∎

Theorem
The determinant of an upper-triangular matrix with a zero on the diagonal is zero.

Proof:
If one of the entries on the diagonal of an upper-triangular matrix is zero, the product of the diagonal entries must be zero. ∎

Notice that as the determinant of a sheer operation is always 1, it follows that because TBW.

The determinant is the product of the eigenvalues of a matrix.

We may draw a quick conclusion from this: If one eigenvalue is zero, the determinant is zero; otherwise, if all the eigenvalues are non-zero, then the determinant will be non-zero. In a sense, the determinant can be used to *determine* whether or not the matrix is invertible:

Theorem
A matrix **M** is invertible if and only if the determinant is non-zero.

Warning: most linear algebra texts focus almost exclusively on the determinant, using it to calculate the inverses of matrices, solving systems of linear equations, etc. Calculating the determinant is numerically unstable and calculating the inverse of a matrix is also numerically unstable; consequently, you, as an engineer, should never calculate the determinant unless you know it is safe to do so. One place where calculating the determinant is safe is in a change of coordinates requiring the calculation of the Jacobian of the transformation, a topic you will see in your course on vector calculus.

Because the determinant is the change in area, it also follows that $|\mathbf{MN}| = |\mathbf{M}||\mathbf{N}|$, and therefore we have another corollary.

Theorem
A matrix product **MN** is invertible if and only if both matrices **M** and **N** are invertible.

By our definition, the easiest way to calculate the determinant is to find the eigenvalues. Recall that when we perform Gaussian elimination using only pivoting and adding a multiple of one row onto another, and therefore

$$\mathbf{U} = \mathbf{R}_m \cdots \mathbf{R}_1 \mathbf{M}$$

where the matrixes $\mathbf{R}_k$ are either swapping two rows, $\mathbf{R}_{j\leftrightarrow k}$; or adding a multiple *c* of row *j* onto row *k*, $\mathbf{R}_{c,j\to k}$. The eigenvalues of both of these basic row operations is 1, so $\left|\mathbf{R}_{j\leftrightarrow k}\right| = \left|\mathbf{R}_{c,j\to k}\right| = 1$ so long as $j \neq k$. Thus, we note that

$$\begin{aligned} \left|\mathbf{U}\right| &= \left|\mathbf{R}_m \cdots \mathbf{R}_1 \mathbf{M}\right| \\ \left|\mathbf{M}\right| &= \left|\mathbf{R}_m\right| \cdots \left|\mathbf{R}_1\right| \left|\mathbf{M}\right| \\ &= 1 \cdots 1 \cdot \left|\mathbf{M}\right| = \left|\mathbf{M}\right| \end{aligned}$$

and therefore the determinant of **M** is the same as the determinant of the upper-triangular matrix **U** resulting from the process of Gaussian elimination to row-echelon form, but the diagonal entries of an upper-triangular matrix are the eigenvalues, and therefore the determinant is the product of the diagonal entries of **U**.

For example, consider the matrix

$$\mathbf{M} = \begin{pmatrix} 3 & 0 & 1 & -1 \\ -0.3 & 4 & 0.9 & 0.1 \\ -0.9 & 0.4 & 4.8 & 2.3 \\ 0.6 & -2 & 0.95 & 4.3 \end{pmatrix}.$$

When you perform Gaussian elimination on this matrix to bring it to row-echelon form, the resulting matrix is

$$\mathbf{U} = \begin{pmatrix} 3 & 0 & 1 & -1 \\ 0 & 4 & 1 & 0 \\ 0 & 0 & 5 & 2 \\ 0 & 0 & 0 & 4 \end{pmatrix}$$

and the eigenvalues of **U** are 3, 4, 5 and 4, so $\left|\mathbf{M}\right| = \left|\mathbf{U}\right| = 3 \cdot 4 \cdot 5 \cdot 4 = 240$. Note, however, that the eigenvalues of **M** are not the same as the eigenvalues of **U**, as the eigenvalues of **M** are approximately

$$\begin{aligned} &5.197885262728804 \pm 0.659453811064522\, j \\ &2.852114737271196 \pm 0.779544963504550\, j \end{aligned}$$

where, as expected,

$$\left(5.197885262728804^2 + 0.659453811064522^2\right)\left(2.852114737271196^2 + 0.779544963504550^2\right) \approx 240 .$$

10.1.2 Calculating the determinant

We will look at algorithms for calculating the determinant for 2 × 2, 3 × 3 and general $n \times n$ matrices.

10.1.2.1 The determinant of a 2 × 2 matrix

You will, at some point, need to calculate the determinant of a 2 × 2 or 3 × 3 matrix. These can be found much more easily using a simple rule-of-thumb that you must memorize:

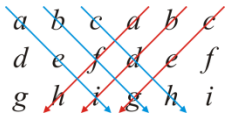
If $\mathbf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then $|\mathbf{M}| = ad - bc$, and this can be added by memorizing the following image:



For example, the determinant of $\mathbf{M} = \begin{pmatrix} 3.8 & 2.1 \\ 7.5 & 4.6 \end{pmatrix}$ is $|\mathbf{M}| = 3.8 \cdot 4.6 - 2.1 \cdot 7.5 = 1.73$.

10.1.2.2 The determinant of a 3 × 3 matrix

If $\mathbf{M} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}$, then $|\mathbf{M}| = aei + bfg + cdh - afh - bdi - ceg$, and this can be added by memorizing the following image:



For example, the determinant of $\mathbf{M} = \begin{pmatrix} 1.2 & 5.3 & 8.1 \\ 7.4 & 6.9 & 4.7 \\ 2.8 & 3.5 & 5.6 \end{pmatrix}$ where we first duplicate the matrix

$$\begin{matrix} 1.2 & 5.3 & 8.1 & 1.2 & 5.3 & 8.1 \\ 7.4 & 6.9 & 4.7 & 7.4 & 6.9 & 4.7 \\ 2.8 & 3.5 & 5.6 & 2.8 & 3.5 & 5.6 \end{matrix}$$

and then multiply and add and subtract the appropriate diagonals:

$$\begin{aligned} |\mathbf{M}| &= 1.2 \cdot 6.9 \cdot 5.6 + 5.3 \cdot 4.7 \cdot 2.8 + 8.1 \cdot 7.4 \cdot 3.5 - 1.2 \cdot 4.7 \cdot 3.5 - 5.3 \cdot 7.4 \cdot 5.6 - 8.1 \cdot 6.9 \cdot 2.8 \\ &= -69.958 \end{aligned}.$$

And now you must also memorize the following warning:

Warning: this pattern does not naturally extend to 4 × 4 matrices and larger.

10.1.2.3 The determinant of an n × n matrix

There is no efficient means of calculating the determinant of a general $n \times n$ matrix. A reasonable approach is to use the rule $\det(AB) = \det(A)\det(B)$; specifically, if we find the LU-decomposition of a matrix A, then

$$A = PLU$$

and therefore $\det(A) = \det(PLU) = \det(P)\det(L)\det(U)$. As P is a permutation matrix, its determinant is either 1 or –1, as L is a lower-triangular matrix and the diagonal entries of L are all 1, so $\det(L) = 1$, and finally as U is an upper-triangular matrix, its determinant is the product of its diagonal entries. For example, if

$$A = \begin{pmatrix} 1.5 & -6.6 & 3.9 & 3.2 \\ -0.5 & -4 & 5.8 & 2.4 \\ 5 & -2 & 3 & 4 \\ -1 & -2 & -4.4 & 2 \end{pmatrix}$$

then we may find the lu-decomposition is

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.3 & 1 & 0 & 0 \\ -0.2 & 0.4 & 1 & 0 \\ -0.1 & 0.7 & -0.8 & 1 \end{pmatrix} \begin{pmatrix} 5 & -2 & 3 & 4 \\ 0 & -6 & 3 & 2 \\ 0 & 0 & -5 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

and there were three row-swaps involved, and therefore $\det(P) = -1$ and $\det(U) = \prod_{j=1}^{4} u_{jj}$, and thus the determinant is $\det(A) = (-1) \cdot 450 = -450$.

10.1.3 The determinant and singular matrices

Where do we use the determinant? So often, it is observed that a matrix is invertible if and only if its determinant is non-zero. Unfortunately, numeric error prevents this, we will ask Matlab what the determinants of the matrices

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{pmatrix} \text{ and } \begin{pmatrix} 10^{-8} & 0 & 0 & 0 \\ 0 & 10^{-8} & 0 & 0 \\ 0 & 0 & 10^{-8} & 0 \\ 0 & 0 & 0 & 10^{-8} \end{pmatrix}.$$

The first is non-invertible, for

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & -4 & -8 & -12 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

but the inverse of the second matrix is

$$\begin{pmatrix} 10^8 & 0 & 0 & 0 \\ 0 & 10^8 & 0 & 0 \\ 0 & 0 & 10^8 & 0 \\ 0 & 0 & 0 & 10^8 \end{pmatrix}$$

Yet, if we ask Matlab, numeric error gives the determinant of the first matrix as being larger than that of the second:

```
>> M1 = [1 2 3 4; 5 6 7 8; 9 10 11 12; 13 14 15 16];
>> det( M1 )
    ans =
         4.733165431326070e-30

>> M2 = [1e-8 0 0 0; 0 1e-8 0 0; 0 0 1e-8 0; 0 0 0 1e-8];
>> det( M2 )
    ans =
         1.000000000000000e-32
```

From this, it appears that the second matrix is *closer* to being non-invertible; however, in reality, the second matrix has a very simple and stable inverse. It is the first matrix that is non-invertible.

10.1.4 Application of the determinant

There are many useless applications of the determinant that are advocated on many books and web pages, including:

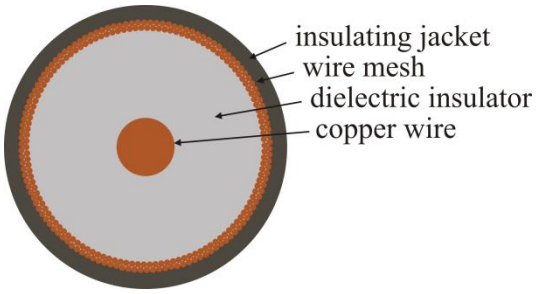
1. finding the area of a triangle (don't we already know how to do this?),
2. finding the area of a parallelogram or parallelepiped (again, don't we already know how to do this?),
3. Kramer's rule (a rule that is in running for the worst algorithm in the world), and
4. determining if a matrix is invertible, or, similarly, determining if a system of n equations in n unknowns has a unique solution (both of which are numerically unstable applications of the determinant).

10.1.4.1 Change-of-basis and the Jacobian determinant

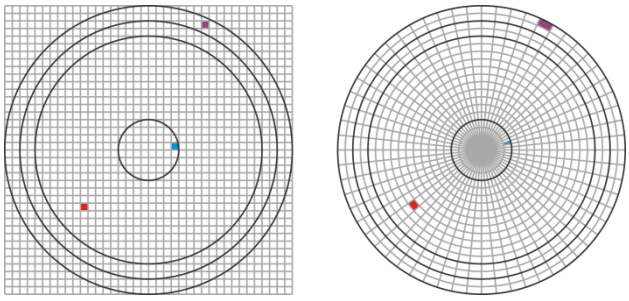
In engineering, due to symmetries you will often perform non-linear changes of bases in vector calculus. For example, on the plane, each point is identified by a Cartesian coordinate (x, y), but you could also represent it as a radius and an angle (r, θ). As we have already seen this with the rectangular and polar representations of a complex number, where

$$x = r\cos(\theta)$$
$$y = r\sin(\theta)$$

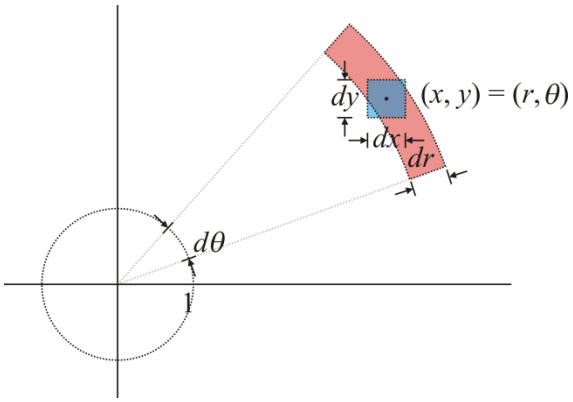
In vector calculus, however, often integrate functions over regions, and it may be easier to express a region in polar coordinates as opposed to Cartesian coordinates. For example, the cross section of a coaxial cable is circular, and describing the regions using x and y is more difficult than describing them using polar coordinates.



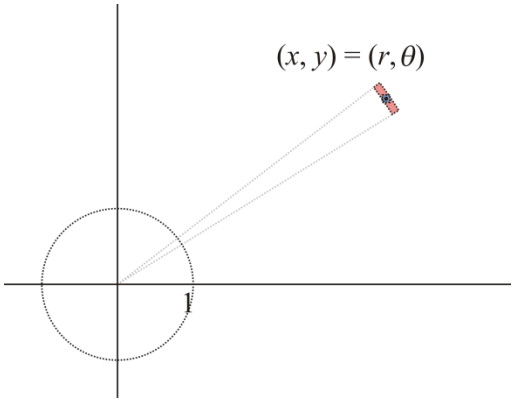
In vector calculus, you will integrate functions over these regions, functions that describe, for example, the electromagnetic potential. In either case, you must divide the plane into regions with equally spaced x and y regions, or equally spaced r and θ regions. With Cartesian coordinates, each square is equal in size, but with polar coordinates, regions closer to the origin are smaller than those further away, as can be shown here.



Now, picking one particular point and looking at the region around it with $dx = dy = d\theta = dr$, it is in shows that the regions may be very different in size:



What we need to know, however, is the ratio between these two areas. If we choose a slightly smaller region,



you may begin to see that the region in polar coordinates is approximately r times larger than that of the region in Cartesian coordinates. In your course on vector calculus, you will learn about the Jacobian matrix, defined as

$$\begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix}$$

where, for example, $\frac{\partial x}{\partial r}$ is the partial derivative of x with respect to r (that is, differentiate x with respect to r while treating θ as a constant). You will learn that ratio of the regions is given by the determinant of the Jacobian, and in this case, that determinant is

$$\begin{aligned} \cos(\theta)\cdot r\cos(\theta) - \left(-r\sin(\theta)\right)\cdot\sin(\theta) &= r\cos^2(\theta) + r\sin^2(\theta) \\ &= r\left(\cos^2(\theta) + \sin^2(\theta)\right) \\ &= r \end{aligned}$$

Changes of coordinates in three dimensions involve three non-linear equations. For example, cylindrical coordinates map (x, y, z) to (r, θ, z) via

$$x = r\cos(\theta)$$
$$y = r\sin(\theta)$$
$$z = z$$

In this case,

$$\begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} & \frac{\partial x}{\partial z} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial z} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \theta} & \frac{\partial z}{\partial z} \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) & 0 \\ \sin(\theta) & r\cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This matrix is block diagonal, and thus, the determinant of the matrix is the determinant of the product of the diagonal blocks, so the determinants of product of $\begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix}$ and (1), which again is $r \cdot 1 = r$.

Finally, as one last example, spherical coordinates map (x, y, z) to (r, θ, ϕ) where

$$x = r\sin(\theta)\cos(\phi)$$
$$y = r\sin(\theta)\sin(\phi)$$
$$z = r\cos(\theta)$$

In this case, the Jacobian matrix is

$$\begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} & \frac{\partial x}{\partial \phi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial \phi} \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial \theta} & \frac{\partial z}{\partial \phi} \end{pmatrix} = \begin{pmatrix} \sin(\theta)\cos(\phi) & r\cos(\theta)\cos(\phi) & -r\sin(\theta)\sin(\phi) \\ \sin(\theta)\sin(\phi) & r\cos(\theta)\sin(\phi) & r\sin(\theta)\cos(\phi) \\ \cos(\theta) & -r\sin(\theta) & 0 \end{pmatrix},$$

the determinant of which is

$$\begin{aligned} & r\cos(\theta)\cos(\phi)\cdot r\sin(\theta)\cos(\phi)\cdot\cos(\theta) + r\sin(\theta)\sin(\phi)\cdot\sin(\theta)\sin(\phi)\cdot r\sin(\theta) \\ & \quad + \sin(\theta)\cos(\phi)\cdot r\sin(\theta)\cos(\phi)\cdot r\sin(\theta) + r\sin(\theta)\sin(\phi)\cdot r\cos(\theta)\sin(\phi)\cdot\cos(\theta) \\ & = r^2\sin(\theta)\left(\cos^2(\theta)\cos^2(\phi) + \sin^2(\theta)\sin^2(\phi) + \cos^2(\phi)\sin^2(\theta) + \sin^2(\phi)\cos^2(\theta)\right) \\ & = r^2\sin(\theta)\left(\cos^2(\theta) + \sin^2(\theta)\right)\left(\cos^2(\phi) + \sin^2(\phi)\right) \\ & = r^2\sin(\theta)\cdot 1\cdot 1 \\ & = r^2\sin(\theta) \end{aligned}$$

In engineering,

Here is a more interesting question: through a lot of linear algebra, we note that

$$\begin{pmatrix} x^2 & x & 1 \\ y^2 & y & 1 \\ z^2 & z & 1 \end{pmatrix} \sim \begin{pmatrix} x^2 & x & 1 \\ 0 & \dfrac{y(x-y)}{x} & \dfrac{x^2-y^2}{x^2} \\ 0 & 0 & \dfrac{(x-z)(y-z)}{xy} \end{pmatrix}$$

We note that the determinant is the product of the diagonal entries, and thus $(x-y)(x-z)(y-z)$, and thus the determinant non zero if all three values *x*, *y* and *z* are different, and zero if any two, or all three, are the same.

Thus, if $\mathbf{M}: V \to V$, and $U \subset V$ with volume $\text{vol}(U)$, then if we represent $\mathbf{M}U$ as the image of each point in *U* under the mapping **M**, what is $\text{vol}(\mathbf{M}U)$? It turns out that this is a constant regardless of the shape or volume of *U*, so this number *determines* the change volume, and thus we will call this the *determinant* of **M**, and it will be denoted as either $\det(\mathbf{M})$ or as $|\mathbf{M}|$. Thus, we may say that

$$\text{vol}(\mathbf{M}U) = |\mathbf{M}| \, \text{vol}(U).$$

10.2 Trace

Another description of a finite-dimensional linear operator is the trace, defined as the sum of the eigenvalues. Without proof, the trace of a linear operator $A \in L\left(\mathbf{R}^n\right)$ can be found using the formula

$$\text{tr}(A) = \sum_{i=1}^{n} a_{i,i} \ .$$

For example, the trace of the matrix $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ is 1 + 5 + 9 = 15. The eigenvalues of the matrix are 0 and $\frac{15 \pm 3\sqrt{33}}{2}$, and the sum of these three is 15. The trace is a linear operator, for $\text{tr}(\alpha A + \beta B) = \alpha\,\text{tr}(A) + \beta\,\text{tr}(B)$.

Theorem
The set of all linear operators in $L(\mathbf{R}^n)$ with a trace of zero form a subspace of $L(\mathbf{R}^n)$.

Proof:

Because $L\left(\mathbf{R}^n\right)$ is already a vector space, we need only show that if $A, B \in L\left(\mathbf{R}^n\right)$, then so is $\alpha A + \beta B$. By the linearity property, we have that

$$\text{tr}\left(\alpha A + \beta B\right) = \alpha\,\text{tr}\left(A\right) + \beta\,\text{tr}\left(B\right) = 0 + 0 = 0\,,$$

and therefore any linear combination of linear operators with trace zero also has trace zero. Therefore, they form a subspace of $L\left(\mathbf{R}^n\right)$. ∎

In one case, the trace is actually very useful in approximating the determinant. You know from calculus that

$$x(t_0 + \varepsilon) \approx f\left(t_0\right) + \varepsilon \dot{x}(t_0)$$

where $\dot{x}(t)$ is the time-derivative of *x*. So, for example, we know $x : t \mapsto e^{-t}\cos(t)$ and thus $\dot{x} : t \mapsto -e^{-t}\cos(t) - e^{-t}\sin(t)$, so

$$x(0 + 0.1) \approx x(0) + 0.1\dot{x}(0) = 1 + 0.1 \cdot (-1) = 0.9\,,$$

and this actually a very reasonable approximation, as $x(0.1) \approx 0.900317000$, so the relative error of the approximation is less than 0.0352% or one part in 2840.

Theorem
If $A \in \mathcal{L}\left(\mathbf{R}^n\right)$, then $\det\left(Id + \varepsilon A\right) \approx 1 + \varepsilon\,\text{tr}\left(A\right)$.

Proof:
We will describe the proof for when *A* is upper triangular, in which case,

$$\begin{aligned}\det\left(Id+\varepsilon A\right) &= \left(1+\varepsilon a_{1,1}\right)\left(1+\varepsilon a_{2,2}\right)\cdots\left(1+\varepsilon a_{n,n}\right)\\ &= 1+\varepsilon a_{1,1}+\varepsilon a_{2,2}+\cdots+\varepsilon a_{n,n}+\varepsilon^2\times\text{lower-order terms}\\ &\approx 1+\varepsilon\left(a_{1,1}+a_{2,2}+\cdots+a_{n,n}\right)\\ &= 1+\varepsilon\,\mathrm{tr}\left(A\right)\end{aligned}$$

Thus, assuming that ε is small enough so that ε^2 is very small, we may ignore the lower-order terms. ∎

Example of this theorem:

Suppose $A=\begin{pmatrix}3 & 2 & -2\\ 0 & 4 & 3\\ 0 & 0 & -2\end{pmatrix}$, in which case,

$\det\left(Id_3+0.01A\right)=\det\begin{pmatrix}1.03 & 0.02 & -0.02\\ 0 & 1.04 & 0.03\\ 0 & 0 & 0.98\end{pmatrix}=1.03\cdot 1.04\cdot 0.98=1.049776$, which is reasonably close to the approximation of $1+0.01\mathrm{tr}\left(A\right)=1+0.01\cdot 5=1.05$.

Suppose $A=\begin{pmatrix}2 & 3 & 4\\ 2 & -4 & 3\\ -3 & 4 & -2\end{pmatrix}$, in which case, $\det\left(Id_3-0.01A\right)=\det\begin{pmatrix}0.98 & -0.03 & -0.04\\ -0.02 & 1.04 & -0.03\\ 0.03 & -0.04 & 1.02\end{pmatrix}=1.039039$, which is reasonably close to the approximation of $1-0.01\mathrm{tr}\left(A\right)=1-0.01\cdot\left(-4\right)=1.04$; and

$\det\left(Id_3+0.001A\right)=\det\begin{pmatrix}1.002 & 0.003 & 0.004\\ 0.002 & 0.996 & 0.003\\ -0.003 & 0.004 & 0.998\end{pmatrix}=0.995989961$, which is even more closely approximated by to the approximation of $1+0.001\mathrm{tr}\left(A\right)=1+0.001\cdot\left(-4\right)=0.996$.

In Matlab, the trace can be calculated using the `trace` routine.

```
>> A = rand( 4, 4 )
    A =
       0.655740699156587   0.678735154857773   0.655477890177557   0.276922984960890
       0.035711678574190   0.757740130578333   0.171186687811562   0.046171390631154
       0.849129305868777   0.743132468124916   0.706046088019609   0.097131781235848
       0.933993247757551   0.392227019534168   0.031832846377421   0.823457828327293

>> trace( A )
    ans =
       2.942984746081821

>> eigs( A )
    ans =
      -0.163041887858380
       0.543363582317600
       0.756416663841249
       1.806246387781351

>> sum( ans )
    ans =
       2.942984746081821
```

10.3 The inverse of a linear operator

Given a function such as x^2, you know it does not have an inverse because it is not one-to-one; both –1 and 1 map to 1. Similarly, e^x does not have an inverse for every real number because it not onto; there is no real value of x such that $e^x = -1$. On the other hand, some functions are invertible; for example,

1. the inverse of $y = ax + b$ is $x = \frac{y}{a} - \frac{b}{a}$, and
2. the inverse of $y = ax^3 + b$ is $x = \sqrt[3]{\frac{y}{a} - \frac{b}{a}}$.

For a function like $y = e^x$, we can find an inverse that maps its range back onto the domain, namely, $x = \ln(y)$, but one cannot, at least for the real numbers, find an inverse of $y = -1$.

For an inverse to exist, a mapping must be both one-to-one and onto. We will now look at aspects related to properties of the inverse in a general vector space, and then we will look at explicitly finding the inverse of a matrix in $L\left(\mathbf{R}^n\right)$.

10.3.1 The inverse of a linear operator

Given a linear operator (that is, $A:U \to U$), the operator is said to be *invertible* if there is an operator $A^{-1} \in L(U)$ such that $A^{-1}A\mathbf{u} = \mathbf{u}$ for each vector $\mathbf{u} \in U$; that is to say, the composition of the two operators is the identity operator, or

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

where $Id\mathbf{u} = \mathbf{u}$ for each $\mathbf{u} \in U$. We will look at properties of the inverse, and how to find the inverse matrix of the matrix representation of a linear operator. If the inverse of a matrix exists, we will then define $A^{-n} = A^{-1}A^{-n+1}$ for integers $n = -1, -2, -3, \ldots$.

10.3.1.1 Properties of the inverse

There are a number of properties of inverses that derive naturally from the definition of the inverse, however, some only apply for finite dimensional vector spaces $\mathbf{F}^n$.

Theorem

If $A \in L(U)$ is not one-to-one, it is not invertible.

Proof:

If *A* is not one-to-one, this means there is at least one vector $\mathbf{v} \in U$ for which there exist at least two different vectors $\mathbf{u}_1, \mathbf{u}_2 \in U$ such that $A\mathbf{u}_1 = A\mathbf{u}_2 = \mathbf{v}$. If an inverse existed, then $A^{-1}\mathbf{v} = \mathbf{u}_1$ and $A^{-1}\mathbf{v} = \mathbf{u}_2$, in which case, $\mathbf{u}_1 = \mathbf{u}_2$, which contradicts our assumption. ∎

Theorem

If $A \in L(U)$ is not onto, it is not invertible.

Proof:

If *A* is not onto, this means there is at least one vector $\mathbf{v} \in U$ such that there does not exist a $\mathbf{u} \in U$ such that $A\mathbf{u} = \mathbf{v}$. Consequently, there cannot be any linear mapping A^{-1}, as then $A^{-1}\mathbf{v} = \mathbf{u}$, and so therefore, $\mathbf{v} = AA^{-1}\mathbf{v} = A\mathbf{u}$, which contradicts our assumption that **v** is not in the range of *A*. ∎

For finite dimensional vector spaces, we have a straight-forward description of inverses: a matrix $A \in L(\mathbf{R}^n)$ is invertible if and only if *A* is both one-to-one and onto.

Theorem

If $A \in L(\mathbf{R}^n)$ is one-to-one and onto, then $A^{-1}A = Id$ if and only if $AA^{-1} = Id$.

Proof:

Assuming that *A* is one-to-one and onto and that $A^{-1}A = Id$, we have that

$$\begin{aligned}
\mathbf{u} = \mathbf{u} &\Leftrightarrow Id\,\mathbf{u} = \mathbf{u} \\
&\Leftrightarrow \left(A^{-1}A\right)\mathbf{u} = \mathbf{u} \\
&\Leftrightarrow A\left(\left(A^{-1}A\right)\mathbf{u}\right) = A\mathbf{u} \\
&\Leftrightarrow A\left(A^{-1}\left(A\mathbf{u}\right)\right) = A\mathbf{u} \\
&\Leftrightarrow \left(AA^{-1}\right)\left(A\mathbf{u}\right) = A\mathbf{u}
\end{aligned}$$

But this last statement says that $AA^{-1} = Id$. ∎

Theorem

If $A \in L(U)$ is invertible, then the inverse is unique.

Proof:

If *A* is invertible, assume that there is a second matrix *B* have the property that $BA = Id$. In this case, we may multiply both sides by A^{-1} on the right to get that

$$(BA)A^{-1} = Id \cdot A^{-1},$$

but because matrix multiplication is associative, we may write both sides as

$$B\left(AA^{-1}\right) = A^{-1} \Leftrightarrow B \cdot Id = A^{-1} \Leftrightarrow B = A^{-1}$$

Thus, both sides are equal. ∎

Theorem

If $A \in L(U)$ is invertible, then $(\alpha A)^{-1} = \frac{1}{\alpha} A^{-1}$ for $\alpha \neq 0$.

Proof:
If we compose these two together, we get that

$$(\alpha A)\left(\frac{1}{\alpha} A^{-1}\right) = \frac{\alpha}{\alpha} AA^{-1} = Id\ ,$$

and therefore $\frac{1}{\alpha} A^{-1}$ must be the inverse of αA. ∎

Theorem

If $A, B \in L(U)$ are both invertible, then $(BA)^{-1} = A^{-1}B^{-1}$.

Proof:

Given a vector $\mathbf{u} \in U$, we note that

$$\left(A^{-1}B^{-1}\right)(BA) = A^{-1}\left(B^{-1}B\right)A = A^{-1}Id\ A = A^{-1}A = Id\ ,$$

and therefore, by our previous uniqueness proof, the inverse of BA must be $A^{-1}B^{-1}$. ∎

Note that if *A* and *B* are invertible, it is not necessarily true that *A* + *B* is invertible. For example, we note that *Id* is invertible with $Id^{-1} = Id$, and $(-Id)^{-1} = -Id$, so both *Id* and *–Id* are invertible, but $Id - Id = O$ is not invertible.

Theorem

If $A \in L(U)$ is invertible, then $\left(A^n\right)^{-1} = A^{-n}$.

Proof:

We will show this by induction. If *n* = 1, then $\left(A^1\right)^{-1} = A^{-1}$. Now, suppose that the statement is true for all positive integers up to and equal to *n*. In this case, $\left(A^{n+1}\right)^{-1} = \left(AA^n\right)^{-1} = \left(A^n\right)^{-1} A^{-1}$ from above. Thus, by assumption

$$\left(A^{n+1}\right)^{-1} = \left(AA^n\right)^{-1} = A^{-n}A^{-1} = A^{-n-1} = A^{-(n+1)}\ .$$ ∎

10.3.1.2 Finding the inverse of a matrix representation

Notice that these theorems make no reference what-so-ever to matrices—they simply use the properties of what an invertible linear operator. We would now like to find the inverse of a matrix. To do this, we will build the Previously, we have seen that for each row operator, there is an inverse row operator that restores the matrix back to its original state, no matter the original state:

Row operator	**Description**	**Inverse row operator**	**Description**
$R_{j\leftrightarrow k}$	Swapping rows *j* and *k*	$R_{j\leftrightarrow k}$	Swapping rows *j* and *k*
$R_{\alpha,\,j\to k}$	Adding α times Row *j* onto *k*	$R_{-\alpha,\,j\to k}$	Adding $-\alpha$ times Row *j* onto *k*
$R_{\alpha,\,j}$	Multiply row *j* by α	$R_{1/\alpha,\,j}$	Multiply row *j* by $\frac{1}{\alpha}$ when $\alpha \neq 0$.

For example, $R_{1/\alpha,\,j}R_{\alpha,\,j}A = A$. This says that each row operation is invertible, and $R_{j\leftrightarrow k}^{-1} = R_{j\leftrightarrow k}$, $R_{\alpha,\,j\to k}^{-1} = R_{1/\alpha,\,j\to k}$ and $R_{\alpha,\,j}^{-1} = R_{1/\alpha,\,j}$. Now, previously, we described row-echelon form. If all the entries on the diagonal of the row-echelon form are non-zero, we could then further multiply each row by a scalar to make the diagonal entries all equal to 1. Following this, we could then perform another sequence of row operations to make the matrix the identity matrix; that is, if the row-echelon form has all non-zero entries on the diagonal, the matrix is itself row equivalent to the identity matrix.

For example, given the matrix $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, we have that

$$R_{-3,1\to 2}A = \begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix},$$

and therefore we can multiply the second row by –0.5, to get

$$R_{-\frac{1}{2},2}R_{-3,1\to 2}A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix},$$

and now add –2 times Row 2 onto Row 1:

$$R_{-2,2\to 1}R_{-\frac{1}{2},2}R_{-3,1\to 2}A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We now have the inverse: the product of the row-operation matrices:

$$\left(R_{-2,2\to 1}R_{-\frac{1}{2},2}R_{-3,1\to 2}\right)A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

so $A^{-1} = R_{-2,2\to 1}R_{-\frac{1}{2},2}R_{-3,1\to 2} = \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix}$.

Consider the 3 × 3 matrix

$$B = \begin{pmatrix} 9 & 3 & 1 \\ 4 & 2 & 1 \\ 1 & -1 & 1 \end{pmatrix}.$$

In order to show that this is row equivalent to the identity matrix, we must start by demonstrating the steps necessary to convert the matrix to row-echelon form. The row operations include

1. swapping Rows 1 and 3 (included for simplicity),
2. adding –4 times Row 1 to Row 2,
3. adding –9 times Row 1 onto Row 3, and
4. adding –2 times Row 2 onto Row 3,

yielding

$$\begin{pmatrix} 9 & 3 & 1 \\ 4 & 2 & 1 \\ 1 & -1 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & -1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & -1 & 1 \\ 0 & 6 & -3 \\ 0 & 12 & -8 \end{pmatrix} \sim \begin{pmatrix} 1 & -1 & 1 \\ 0 & 6 & -3 \\ 0 & 0 & -2 \end{pmatrix}.$$

This is followed by

5. scaling Row 2 by 1/6 and
6. scaling Row 3 by –1/2,

yielding

$$\begin{pmatrix} 9 & 3 & 1 \\ 4 & 2 & 1 \\ 1 & -1 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & -1 & 1 \\ 0 & 6 & -3 \\ 0 & 0 & -2 \end{pmatrix} \sim \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -\frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}$$

and now we may proceed to eliminate the strictly upper triangular entries by

7. adding 1/2 times Row 3 onto Row 2,
8. adding –1 times Row 3 onto Row 1, and
9. adding Row 2 onto Row 1,

yielding

$$\begin{pmatrix} 9 & 3 & 1 \\ 4 & 2 & 1 \\ 1 & -1 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -\frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

If we multiply these operations together, we get

$$B^{-1} = R_{1,2\to 1} R_{-1,3\to 1} R_{\frac{1}{2},3\to 2} R_{-\frac{1}{2},3} R_{\frac{1}{6},2} R_{-2,2\to 3} R_{-9,1\to 3} R_{-4,1\to 2} R_{1\leftrightarrow 3}$$

Multiplying these out explicitly gives us that

$$B^{-1}=\begin{pmatrix}1&1&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&-1\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&\frac{1}{2}\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\0&0&-\frac{1}{2}\end{pmatrix}\begin{pmatrix}1&0&0\\0&\frac{1}{6}&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\0&-2&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\-9&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\-4&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}$$

or

$$B^{-1}=\begin{pmatrix}\frac{1}{4}&-\frac{1}{3}&\frac{1}{12}\\-\frac{1}{4}&\frac{2}{3}&-\frac{5}{12}\\-\frac{1}{2}&1&\frac{1}{2}\end{pmatrix}.$$

Now, recall that we began by swapping Rows 1 and 3. If we did not, the operations would have changed, but the end result would still be the identity matrix, and consequently, the inverse will also be the same.

In order to find the inverse of a matrix, however, recording these steps is rather painful, and instead, we'd like to multiply out the row operations as we apply them to the matrix. Fortunately, we can do this quite easily: create an adjunct matrix of the matrix we are inverting and the identity matrix. Let us do a different example, say, find the inverse of

$$C=\begin{pmatrix}2&-1&1\\-4&2&-1\\1&2&-3\end{pmatrix}.$$

Now we create the adjunct matrix

$$\left(\begin{array}{ccc:ccc}2&-1&1&1&0&0\\-4&2&-1&0&1&0\\1&2&-3&0&0&1\end{array}\right).$$

Now, when we perform a row operation on this matrix, we are simultaneously calculating the product of the row operations on the right-hand side:

$$\left(\begin{array}{ccc:ccc}2&-1&1&1&0&0\\-4&2&-1&0&1&0\\1&2&-3&0&0&1\end{array}\right)\sim\left(\begin{array}{ccc:ccc}2&-1&1&1&0&0\\0&0&1&2&1&0\\1&2&-3&0&0&1\end{array}\right)\sim\left(\begin{array}{ccc:ccc}2&-1&1&1&0&0\\0&0&1&2&1&0\\0&\frac{5}{2}&-\frac{7}{2}&-\frac{1}{2}&0&1\end{array}\right)$$

now we swap Rows 2 and 3, and continue

$$\left(\begin{array}{ccc:ccc}2&-1&1&1&0&0\\-4&2&-1&0&1&0\\1&2&-3&0&0&1\end{array}\right)\sim\left(\begin{array}{ccc:ccc}2&-1&1&1&0&0\\0&\frac{5}{2}&-\frac{7}{2}&-\frac{1}{2}&0&1\\0&0&1&2&1&0\end{array}\right).$$

Scaling Rows 1 and 2 by $\frac{1}{2}$ and $\frac{2}{5}$, respectively, we get,

$$\left(\begin{array}{ccc:ccc}2&-1&1&1&0&0\\-4&2&-1&0&1&0\\1&2&-3&0&0&1\end{array}\right)\sim\left(\begin{array}{ccc:ccc}1&-\frac{1}{2}&\frac{1}{2}&\frac{1}{2}&0&0\\0&1&-\frac{7}{5}&-\frac{1}{5}&0&\frac{2}{5}\\0&0&1&2&1&0\end{array}\right).$$

Finally, eliminating the strictly upper-triangular component, we must now

1. add $\frac{7}{5}$ times Row 3 onto Row 2,
2. add $-\frac{1}{2}$ times Row 3 onto Row 1, and
3. add $\frac{1}{2}$ times Row 2 onto Row 1,

yielding the sequence

$$\left(\begin{array}{ccc|ccc} 2 & -1 & 1 & 1 & 0 & 0 \\ -4 & 2 & -1 & 0 & 1 & 0 \\ 1 & 2 & -3 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & \frac{13}{5} & \frac{7}{5} & \frac{2}{5} \\ 0 & 0 & 1 & 2 & 1 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & -\frac{1}{2} & 0 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 1 & 0 & \frac{13}{5} & \frac{7}{5} & \frac{2}{5} \\ 0 & 0 & 1 & 2 & 1 & 0 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{4}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & 1 & 0 & \frac{13}{5} & \frac{7}{5} & \frac{2}{5} \\ 0 & 0 & 1 & 2 & 1 & 0 \end{array}\right).$$

Therefore, the inverse is the commutation of the row operations on the right-hand matrix, or

$$C^{-1} = \begin{pmatrix} \frac{4}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{13}{5} & \frac{7}{5} & \frac{2}{5} \\ 2 & 1 & 0 \end{pmatrix}.$$

We can confirm this by multiply $CC^{-1} = C^{-1}C = Id_3$.

This can be used for finding the inverse, but as you can see, there is a lot of work that must be done:

1. For each off diagonal entry, we must add a multiple of a row onto another. This requires $n(n-1)\cdot 2n$ multiplications and additions, and
2. For each row, we will likely have to scale the row, requiring $n \cdot 2n$ multiplications.

Therefore, the total work required is approximately n^3 multiplications—inverting a 1000 × 1000 matrix would require two billion multiplications, and even inverting a 5 × 5 matrix would require 250 multiplications. Not only that, we will later see that matrix inversion is said to be *numerically unstable*, meaning that if you try to do this in a computer, it is very likely to have significant numerical error.

If you are trying to solve $A\mathbf{x} = \mathbf{b}$, use the other techniques we have described in this course. Under no circumstances should you ever calculate A^{-1} and then attempt to calculate $A^{-1}\mathbf{b}$. There are cases where you will need to explicitly calculate the inverse, but this will most likely be related to your second-year calculus courses, in which case, you can use the technique described above.

10.3.1.3 Dealing with error

Recall that we use partial pivoting in Gaussian elimination to ensure that we never magnify an error. Let us try to find the inverse of the matrix

$$A=\begin{pmatrix}1&4&7\\2&5&8\\3&6&12\end{pmatrix}$$

Performing the previous operations,

$$\left(\begin{array}{ccc:ccc}1&4&7&1&0&0\\2&5&8&0&1&0\\3&6&12&0&0&1\end{array}\right)\sim\left(\begin{array}{ccc:ccc}3&6&12&0&0&1\\2&5&8&0&1&0\\1&4&7&1&0&0\end{array}\right)$$
$$\sim\left(\begin{array}{ccc:ccc}3&6&12&0&0&1\\0&1&0&0&1&-\frac{2}{3}\\1&4&7&1&0&0\end{array}\right)$$
$$\sim\left(\begin{array}{ccc:ccc}3&6&12&0&0&1\\0&1&0&0&1&-\frac{2}{3}\\0&2&3&1&0&-\frac{1}{3}\end{array}\right)$$
$$\sim\left(\begin{array}{ccc:ccc}3&6&12&0&0&1\\0&2&3&1&0&-\frac{1}{3}\\0&1&0&0&1&-\frac{2}{3}\end{array}\right)$$
$$\sim\left(\begin{array}{ccc:ccc}3&6&12&0&0&1\\0&2&3&1&0&-\frac{1}{3}\\0&0&-\frac{3}{2}&-\frac{1}{2}&1&-\frac{1}{2}\end{array}\right)$$
$$\sim\left(\begin{array}{ccc:ccc}3&6&12&0&0&1\\0&2&3&1&0&-\frac{1}{3}\\0&0&-\frac{3}{2}&-\frac{1}{2}&1&-\frac{1}{2}\end{array}\right)$$

At this point, we run into a problem: we can no longer pivot to bring the largest entry onto the diagonal, and to change 6 to a zero, we must add –3 times Row 2 onto Row 1. Any error in the values of Row 2 have now been magnified by a factor of three, and thus, we see that calculating the inverse cannot be done so as to minimize the error. Consequently, this is an operation you should never perform except when you are certain that there is no error in the coefficients, such as when you are performing a change of coordinates—an operation you will see your vector calculus course.

Regardless, proceeding forward, it is now easiest to make all the entries on the diagonal equal to one by dividing each of the rows by 3, 2 and $-\frac{3}{2}$:

$$\left(\begin{array}{ccc:ccc}1&4&7&1&0&0\\2&5&8&0&1&0\\3&6&12&0&0&1\end{array}\right)\sim\left(\begin{array}{ccc:ccc}3&6&12&0&0&1\\0&2&3&1&0&-\frac{1}{3}\\0&0&-\frac{3}{2}&-\frac{1}{2}&1&-\frac{1}{2}\end{array}\right)$$
$$\sim\left(\begin{array}{ccc:ccc}1&2&4&0&0&\frac{1}{3}\\0&1&\frac{3}{2}&\frac{1}{2}&0&-\frac{1}{6}\\0&0&1&\frac{1}{3}&-\frac{2}{3}&\frac{1}{3}\end{array}\right)$$

and now we continue to perform row operations to eliminate the upper triangular component of the left-hand matrix:

$$\left(\begin{array}{ccc:ccc} 1 & 4 & 7 & 1 & 0 & 0 \\ 2 & 5 & 8 & 0 & 1 & 0 \\ 3 & 6 & 12 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc:ccc} 1 & 2 & 4 & 0 & 0 & \frac{1}{3} \\ 0 & 1 & \frac{3}{2} & \frac{1}{2} & 0 & -\frac{1}{6} \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{2}{3} & \frac{1}{3} \end{array}\right)$$

$$\sim \left(\begin{array}{ccc:ccc} 1 & 0 & 1 & -1 & 0 & \frac{2}{3} \\ 0 & 1 & \frac{3}{2} & \frac{1}{2} & 0 & -\frac{1}{6} \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{2}{3} & \frac{1}{3} \end{array}\right)$$

$$\sim \left(\begin{array}{ccc:ccc} 1 & 0 & 1 & -1 & 0 & \frac{2}{3} \\ 0 & 1 & 0 & 0 & 1 & -\frac{2}{3} \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{2}{3} & \frac{1}{3} \end{array}\right)$$

$$\sim \left(\begin{array}{ccc:ccc} 1 & 0 & 0 & -\frac{4}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & 1 & 0 & 0 & 1 & -\frac{2}{3} \\ 0 & 0 & 1 & \frac{1}{3} & -\frac{2}{3} & \frac{1}{3} \end{array}\right)$$

We have now found a sequence of row operations that reduced the matrix to the identity matrix, and therefore the product of those row operations is preserved in the right-hand matrix.

$$A^{-1} = \begin{pmatrix} -\frac{4}{3} & \frac{2}{3} & \frac{1}{3} \\ 0 & 1 & -\frac{2}{3} \\ \frac{1}{3} & -\frac{2}{3} & \frac{1}{3} \end{pmatrix},$$

and you will note that $A^{-1}A = AA^{-1} = Id_3$.

Now, let us see what happens in the process if we attempt to do the same sequence of operations with a matrix that is not one-to-one:

$$\left(\begin{array}{ccc:ccc} 1 & 4 & 7 & 1 & 0 & 0 \\ 2 & 5 & 8 & 0 & 1 & 0 \\ 3 & 6 & 9 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc:ccc} 3 & 6 & 9 & 0 & 0 & 1 \\ 2 & 5 & 8 & 0 & 1 & 0 \\ 1 & 4 & 7 & 1 & 0 & 0 \end{array}\right)$$

$$\sim \left(\begin{array}{ccc:ccc} 3 & 6 & 9 & 0 & 0 & 1 \\ 0 & 1 & 2 & 0 & 1 & -\frac{2}{3} \\ 1 & 4 & 7 & 1 & 0 & 0 \end{array}\right)$$

$$\sim \left(\begin{array}{ccc:ccc} 3 & 6 & 9 & 0 & 0 & 1 \\ 0 & 1 & 2 & 0 & 1 & -\frac{2}{3} \\ 0 & 2 & 3 & 1 & 0 & -\frac{1}{3} \end{array}\right)$$

$$\sim \left(\begin{array}{ccc:ccc} 3 & 6 & 9 & 0 & 0 & 1 \\ 0 & 2 & 4 & 1 & 0 & -\frac{1}{3} \\ 0 & 1 & 2 & 0 & 1 & -\frac{2}{3} \end{array}\right)$$

$$\sim \left(\begin{array}{ccc:ccc} 3 & 6 & 9 & 0 & 0 & 1 \\ 0 & 2 & 4 & 1 & 0 & -\frac{1}{3} \\ 0 & 0 & 0 & -\frac{1}{2} & 1 & -\frac{1}{2} \end{array}\right)$$

We note that it is no longer possible to multiply Row 3 so as to get 1 in location (3,3). The matrix is not invertible.

In Matlab, you can find the inverse

Previously, we used Gaussian elimination on the augmented matrix to simply zero out all entries below the diagonal using pivoting to always bring the largest entry in absolute value onto the diagonal before zeroing out all entries below it. This pivoting (swapping of rows) ensures numerical stability.

To find the inverse,

$$\left(\begin{array}{ccc|ccc} 2 & 2 & 3 & 1 & 0 & 0 \\ 2 & -3 & 1 & 0 & 1 & 0 \\ 1 & 1 & 2 & 0 & 0 & 1 \end{array}\right) \to \left(\begin{array}{ccc|ccc} 1 & 1 & \frac{3}{2} & \frac{1}{2} & 0 & 0 \\ 2 & -3 & 1 & 0 & 1 & 0 \\ 1 & 1 & 2 & 0 & 0 & 1 \end{array}\right)$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 1 & \frac{3}{2} & \frac{1}{2} & 0 & 0 \\ 0 & -5 & 2 & -1 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & 1 \end{array}\right)$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 0 & \frac{11}{10} & \frac{3}{10} & \frac{1}{5} & 0 \\ 0 & 1 & \frac{2}{5} & \frac{1}{5} & -\frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 0 & 1 \end{array}\right)$$

$$\to \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{7}{5} & \frac{1}{5} & -\frac{11}{5} \\ 0 & 1 & 0 & \frac{3}{5} & -\frac{1}{5} & -\frac{4}{5} \\ 0 & 0 & 1 & -1 & 0 & 2 \end{array}\right)$$

Finding the inverse in Matlab is quite straight-forward: raise the matrix to the exponent –1:

```
>> M = [2 2 3; 2 -3 1; 1 1 2];    % This is the matrix in the previous example
>> M^-1
  ans =
      1.4000    0.2000   -2.2000
      0.6000   -0.2000   -0.8000
     -1.0000         0    2.0000
```

The most important rule in engineering:

Do not calculate the inverse, and if you must, don't. If you have a 2 × 2 matrix, where

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \begin{pmatrix} \frac{d}{\Delta} & -\frac{b}{\Delta} \\ -\frac{c}{\Delta} & \frac{a}{\Delta} \end{pmatrix}^{-1} \quad \text{where } \Delta = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc \,.$$

Recall that matrix multiplication is not commutative, that is, for two $n \times n$ matrices **A** and **B**, in general, $\mathbf{AB} \neq \mathbf{BA}$. Thus, if we have determined that $\mathbf{AA}^{-1} = \mathbf{Id}_n$, does it follow that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{Id}_n$?

Recall that the determinant is the ratio of the change of volume with a negative determinant indicating a change in orientation. Thus, one may deduce that for two invertible $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$, it follows that $\left|\mathbf{AB}\right| = \left|\mathbf{A}\right|\left|\mathbf{B}\right|$. Since the identity matrix does not change any volume, it also follows that $\left|\mathbf{Id}_n\right| = 1$. Therefore, as

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{Id}_n,$$

it follows that

$$\left|\mathbf{A}\mathbf{A}^{-1}\right| = \left|\mathbf{A}^{-1}\mathbf{A}\right| = \mathbf{Id}_n.$$

If

$$\left(\alpha\mathbf{A}\right)\mathbf{A}^{-1} = \alpha\left(\mathbf{A}\mathbf{A}^{-1}\right) = \alpha\mathbf{Id}_n,$$

and therefore

$$\mathbf{Id}_n = \frac{1}{\alpha}\left(\alpha\mathbf{A}\right)\mathbf{A}^{-1} = \left(\alpha\mathbf{A}\right)\left(\frac{1}{\alpha}\mathbf{A}^{-1}\right),$$

Thus, the inverse of $\alpha\mathbf{A}$ is $\frac{1}{\alpha}\mathbf{A}^{-1}$.

10.4 Finding the inverse

Suppose you

Problems

1. Find the inverse of the matrix

$$\begin{pmatrix} 3 & 2 & -2 \\ 0 & 3 & 1 \\ -3 & 2 & -3 \end{pmatrix}$$

11 The adjoint of a linear operator (transpose and conjugate transpose)

Given two inner product spaces $\mathcal{U}$ and $\mathcal{V}$ (a vector space together with an inner product), the *adjoint* of a linear operator $A:U\to V$ is that operator $A^*:V\to U$ such that

$$\langle A\mathbf{u},\mathbf{v}\rangle = \langle \mathbf{u},A^*\mathbf{v}\rangle$$

for all $\mathbf{u}\in U$ and $\mathbf{v}\in V$. Note that the left-hand side uses the inner product in *V* and the right-hand side uses the inner product in *U*. Usually, however, are most significant interest will be when *A* maps a vector space onto itself; that is, when $A:V\to V$.

This is very significant in the application of linear operators, and we will look specifically at those linear operators for which the adjoint equals the linear operator itself (that is, we will look at those linear operators where $A=A^*$) and those operators where the adjoint equals the inverse (so where $A^{-1}=A^*$). We will begin by:

1. looking at the properties of the adjoint,
2. finding the adjoint of a real finite-dimensional linear operator (that is, the adjoint of a matrix),
3. finding the adjoint of a complex finite-dimensional linear operator,
4. defining and considering self-adjoint and skew-adjoint linear operators and their properties,
5.
6. defining unitary and orthogonal linear operators,

11.1 Properties of the adjoint

We will look at various properties of the adjoint of a linear operator.

Theorem

The adjoint of the adjoint of a linear operator *A* is *A* itself; that is, $\left(A^*\right)^* = A$

Proof:

Let $A:U\to V$ and let $\mathbf{u}\in U$ and $\mathbf{v}\in V$ be arbitrary vectors within the appropriate vector spaces, then recalling the property that for the inner product, $\langle \mathbf{u},\mathbf{v}\rangle = \langle \mathbf{v},\mathbf{u}\rangle^*$, we have that

$$\langle A\mathbf{u},\mathbf{v}\rangle = \langle \mathbf{u},A^*\mathbf{v}\rangle = \langle A^*\mathbf{v},\mathbf{u}\rangle^* = \left\langle \mathbf{v},\left(A^*\right)^*\mathbf{u}\right\rangle^* = \left\langle \left(A^*\right)^*\mathbf{u},\mathbf{v}\right\rangle .$$

As this is true for all vectors, $\left(A^*\right)^* = A$. ∎

Theorem
The adjoint of a scalar multiple of a linear mapping is the scalar multiple of the complex conjugate of the scalar and the mapping.

Proof:
Let $A:\mathcal{U}\to\mathcal{V}$ and let $\mathbf{u}\in\mathcal{U}$ and $\mathbf{v}\in\mathcal{V}$, then

$$\langle(\alpha A)\mathbf{u},\mathbf{v}\rangle=\langle\mathbf{u},(\alpha A)^*\mathbf{v}\rangle,$$

but

$$\begin{aligned}\langle(\alpha A)\mathbf{u},\mathbf{v}\rangle&=\langle\alpha(A\mathbf{u}),\mathbf{v}\rangle\\&=\alpha^*\langle A\mathbf{u},\mathbf{v}\rangle\\&=\alpha^*\langle\mathbf{u},A^*\mathbf{v}\rangle\\&=\langle\mathbf{u},\alpha^*(A^*\mathbf{v})\rangle\\&=\langle\mathbf{u},(\alpha^*A^*)\mathbf{v}\rangle\end{aligned}$$

and therefore, as $\langle\mathbf{u},(\alpha A)^*\mathbf{v}\rangle=\langle\mathbf{u},(\alpha^*A^*)\mathbf{v}\rangle$ for all **u** and **v**, it follows that $(\alpha A)^*=\alpha^*A^*$. ∎

Theorem
In a real vector space, the operation of the adjoint is linear.

Proof:
We have that, by definition, $\langle(\alpha A+\beta B)\mathbf{u},\mathbf{v}\rangle=\langle(\alpha A+\beta B)^*\mathbf{u},\mathbf{v}\rangle$, but

$$\begin{aligned}\langle(\alpha A+\beta B)\mathbf{u},\mathbf{v}\rangle&=\langle\alpha_1A_1\mathbf{u}+\alpha_2A_2\mathbf{u},\mathbf{v}\rangle\\&=\langle\alpha_1A_1\mathbf{u},\mathbf{v}\rangle+\langle\alpha_2A_2\mathbf{u},\mathbf{v}\rangle\ \text{ but for a real vector space, }\langle\alpha\mathbf{u},\mathbf{v}\rangle=\alpha\langle\mathbf{u},\mathbf{v}\rangle\\&=\alpha_1\langle A_1\mathbf{u},\mathbf{v}\rangle+\alpha_2\langle A_2\mathbf{u},\mathbf{v}\rangle\\&=\alpha_1\langle\mathbf{u},A_1^*\mathbf{v}\rangle+\alpha_2\langle\mathbf{u},A_2^*\mathbf{v}\rangle\\&=\langle\mathbf{u},\alpha_1A_1^*\mathbf{v}\rangle+\langle\mathbf{u},\alpha_2A_2^*\mathbf{v}\rangle\\&=\langle\mathbf{u},\alpha_1A_1^*\mathbf{v}+\alpha_2A_2^*\mathbf{v}\rangle\\&=\langle\mathbf{u},(\alpha_1A_1^*+\alpha_2A_2^*)\mathbf{v}\rangle\end{aligned}$$

and therefore $(\alpha_1A_1+\alpha_2A_2)^*=\alpha_1A_1^*+\alpha_2A_2^*$. ∎

Theorem
The adjoint of a sum of linear operators is the sum of the adjoints.

Proof:
Let $A_1, A_2 : U \to V$ and let $\mathbf{u} \in U$ and $\mathbf{v} \in V$, then

$$\langle (A_1 + A_2)\mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, (A_1 + A_2)^* \mathbf{v} \rangle,$$

but

$$\begin{aligned} \langle (A_1 + A_2)\mathbf{u}, \mathbf{v} \rangle &= \langle A_1\mathbf{u} + A_2\mathbf{u}, \mathbf{v} \rangle \\ &= \langle A_1\mathbf{u}, \mathbf{v} \rangle + \langle A_1\mathbf{u}, \mathbf{v} \rangle \\ &= \langle \mathbf{u}, A_1^*\mathbf{v} \rangle + \langle \mathbf{u}, A_2^*\mathbf{v} \rangle \\ &= \langle \mathbf{u}, A_1^*\mathbf{v} + A_2^*\mathbf{v} \rangle \\ &= \langle \mathbf{u}, (A_1^* + A_2^*)\mathbf{v} \rangle \end{aligned}$$

and therefore, as $\langle \mathbf{u}, (A_1 + A_2)^* \mathbf{v} \rangle = \langle \mathbf{u}, (A_1^* + A_2^*)\mathbf{v} \rangle$ for all **u** and **v**, it follows that $(A_1 + A_2)^* = A_1^* + A_2^*$. ∎

Theorem
The adjoint of a composition of linear operators is the composition of the adjoints in reverse order.

Proof:
Assume that $A: U \to V$ and $B: V \to W$ are linear operators, $\mathbf{u} \in U$ and $\mathbf{w} \in W$, using the property of associativity, we have

$$\langle (BA)\mathbf{u}, \mathbf{w} \rangle = \langle \mathbf{u}, (BA)^* \mathbf{w} \rangle$$

but

$$\begin{aligned} \langle (BA)\mathbf{u}, \mathbf{w} \rangle &= \langle B(A\mathbf{u}), \mathbf{w} \rangle \\ &= \langle A\mathbf{u}, B^*\mathbf{w} \rangle \\ &= \langle \mathbf{u}, A^*(B^*\mathbf{w}) \rangle \\ &= \langle \mathbf{u}, (A^*B^*)\mathbf{w} \rangle \end{aligned}$$

and therefore, as $\langle \mathbf{u}, (BA)^* \mathbf{w} \rangle = \langle \mathbf{u}, (A^*B^*)\mathbf{w} \rangle$ for all **u** and **w**, it follows that $(BA)^* = A^*B^*$. ∎

Note that the inner product $\langle BA\mathbf{u}, \mathbf{w} \rangle$ is the inner product in *W*, $\langle A\mathbf{u}, B^*\mathbf{w} \rangle$ is an inner product in V, and $\langle \mathbf{u}, (A^*B^*)\mathbf{w} \rangle$ is an inner product in *U*.

Note that this strongly relationship between the complex conjugate and the adjoint, and thus, they both use the same symbol. The adjoint has many of the same, or similar, properties of the complex conjugate, as we will see. You simply have to be careful to see what you are applying the superscript star to: for α^*, α is a scalar, and so it refers to the complex conjugate, while for the operator A, A^* is the adjoint.

Theorem
If a linear operator is invertible, the inverse of the adjoint of a matrix is the adjoint of the inverse of the matrix.

Proof:
Using the definitions,

$$\begin{aligned}\langle \mathbf{u}, \mathbf{v} \rangle &= \left\langle \left(AA^{-1} \right)\mathbf{u}, \mathbf{v} \right\rangle \\ &= \left\langle A\left(A^{-1}\mathbf{u} \right), \mathbf{v} \right\rangle \\ &= \left\langle A^{-1}\mathbf{u}, A^*\mathbf{v} \right\rangle \\ &= \left\langle \mathbf{u}, \left(A^{-1} \right)^* \left(A^*\mathbf{v} \right) \right\rangle \\ &= \left\langle \mathbf{u}, \left(\left(A^{-1} \right)^* A^* \right)\mathbf{v} \right\rangle \\ &= \left\langle \mathbf{u}, \left(AA^{-1} \right)^* \mathbf{v} \right\rangle \\ &= \langle \mathbf{u}, \mathbf{v} \rangle\end{aligned}$$

Therefore, $\left(A^{-1} \right)^* = \left(A^* \right)^{-1}$. ∎

Note that many of the properties of the complex conjugate are shared by the adjoint:

	Complex conjugate	**Adjoint**
Self-inverse	$\left(\alpha^* \right)^* = \alpha$	$\left(A^* \right)^* = A$
Distributes across addition	$\left(\alpha + \beta \right)^* = \alpha^* + \beta^*$	$\left(A + B \right)^* = A^* + B^*$
Distributes across subtraction	$\left(\alpha - \beta \right)^* = \alpha^* - \beta^*$	$\left(A - B \right)^* = A^* - B^*$
Distributes across scalar multiplication	$\left(\alpha\beta \right)^* = \alpha^*\beta^*$	$\left(\alpha A \right)^* = \alpha^* A^*$
Commutes with matrix powers	$\left(\alpha^n \right)^* = \left(\alpha^* \right)^n$	$\left(A^n \right)^* = \left(A^* \right)^n$
Commutes with the inverse	$\left(\alpha^{-1} \right)^* = \left(\alpha^* \right)^{-1}$ if $\alpha \neq 0$	$\left(A^{-1} \right)^* = \left(A^* \right)^{-1}$ if A is invertible

The adjoint differs slightly from complex conjugation when it comes to distributing across multiplication:

	Complex conjugate	**Adjoint**
Distributes across multiplication	$\left(\alpha\beta \right)^* = \alpha^*\beta^*$	$\left(AB \right)^* = B^*A^*$

Now that we've looked at the properties of the adjoint, let's look at how the adjoint manifests itself in real and complex finite-dimensional vector spaces.

Problems

1. What is the adjoint of CBC where A, B and C are linear mappings?

2. What is the adjoint of $A + B + C$, where A, B and C are linear mappings?

3. What is the relationship between the domains and co-domains of the mappings in Question 1?

4. What is the relationship between the domains and co-domains of the mappings in Question 2?

5. What is the adjoint of AA^*A?

6. If $A: \mathcal{U} \to \mathcal{V}$, what are the domains and co-domains of AA^*, AA^*, A^*AA^* and AA^*A?

7. What is the adjoint of $3A_1 - 4A_2$?

8. What is the adjoint of $3jA_1 - jA_2$?

Solutions

1. $(CBA)^* = A^*B^*C^*$

3. If $A: \mathcal{U} \to \mathcal{V}, B: \mathcal{V} \to \mathcal{W}, C: \mathcal{W} \to \mathcal{X}$, $CBA: \mathcal{U} \to \mathcal{X}$ and $A^*B^*C^*: \mathcal{X} \to \mathcal{U}$.

5. A^*AA^*

7. $3A_1^* - 4A_2^*$

11.2 The adjoint for real finite-dimensional vector spaces

If $A:\mathbf{R}^n \to \mathbf{R}^m$ is the matrix representation of a linear operator in a finite-dimensional real vector space and $\mathbf{u} \in \mathbf{R}^n$ and $\mathbf{v} \in \mathbf{R}^m$, then it must be true that

$$\langle A\mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, A^*\mathbf{v} \rangle$$

and we know that the jth entry of $A\mathbf{u}$ is $(A\mathbf{u})_i = \sum_{j=1}^{n} a_{i,j}u_j$, and thus

$$\begin{aligned}
\langle A\mathbf{u}, \mathbf{v} \rangle &= \sum_{i=1}^{m} (A\mathbf{u})_i \, v_i \\
&= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{i,j}u_j \right) v_i \\
&= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{i,j}u_j v_i \right) \\
&= \sum_{j=1}^{n} \left(\sum_{i=1}^{m} a_{i,j}u_j v_i \right) \\
&= \sum_{j=1}^{n} u_j \left(\sum_{i=1}^{m} a_{i,j} v_i \right)
\end{aligned}$$

Now, the inner sum $\sum_{i=1}^{m} a_{i,j}v_i$ looks like a matrix-vector product, but see that it must be an $m \times n$ matrix, where the $(i, j)^{\text{th}}$ entry is actually $a_{j,i}$. For finite-dimensional real matrices, we have a special name this adjoint matrix, the *transpose*, and it is customary to write the transpose of a matrix A as A^{T}. For example, here $A:\mathbf{R}^4 \to \mathbf{R}^3$,

$$A = \begin{pmatrix} -3 & -1 & 1 & 0 \\ 0 & 1 & -2 & 3 \\ -1 & 2 & 4 & 2 \end{pmatrix},$$

so if $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$, we see that $A\mathbf{u} = \begin{pmatrix} -2 \\ 8 \\ 23 \end{pmatrix}$, and so we see that $\langle A\mathbf{u}, \mathbf{v} \rangle = -2 \cdot 2 + 8 \cdot 3 + 23 \cdot 4 = 112$.

Similarly, we have that

$$A^{\mathrm{T}} = \begin{pmatrix} -3 & 0 & -1 \\ -1 & 1 & 2 \\ 1 & -2 & 4 \\ 0 & 3 & 2 \end{pmatrix}$$

and so $A^{\mathrm{T}}\mathbf{v} = \begin{pmatrix} -10 \\ 9 \\ 12 \\ 17 \end{pmatrix}$ and thus $\left\langle \mathbf{u}, A^{\mathrm{T}}\mathbf{v} \right\rangle = 1\cdot(-10) + 2\cdot 9 + 3\cdot 12 + 4\cdot 17 = 112$.

Theorem

If $A: \mathbf{R}^n \to \mathbf{R}^m$ and $A\mathbf{u}_1, \ldots, A\mathbf{u}_k$ is a linearly independent set, then $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is also linearly independent.

Proof:

Suppose that $A: \mathbf{R}^n \to \mathbf{R}^m$ and $A\mathbf{u}_1, \ldots, A\mathbf{u}_k$ is a linearly independent set, but that $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are not linearly dependent. Therefore, by definition, there must be a collection of scalars $\alpha_1, \ldots, \alpha_k$ not all zero such that

$$a_1\mathbf{u}_1 + \cdots + a_k\mathbf{u}_k = \mathbf{0}_n,$$

but in this case,

$$\begin{aligned} & A(\alpha_1\mathbf{u}_1 + \cdots + \alpha_k\mathbf{u}_k) = \mathbf{0}_m \\ \therefore \quad & A(\alpha_1\mathbf{u}_1) + \cdots + A(\alpha_k\mathbf{u}_k) = \mathbf{0}_m \\ \therefore \quad & \alpha_1 A\mathbf{u}_1 + \cdots + \alpha_k A\mathbf{u}_k = \mathbf{0}_m \end{aligned}$$

This, however, implies that $A\mathbf{u}_1, \ldots, A\mathbf{u}_k$ is linearly dependent, which contradicts our assumption. Therefore the set $\mathbf{u}_1, \ldots, \mathbf{u}_k$ is also linearly independent. ∎

Theorem

If $A: \mathbf{F}^n \to \mathbf{F}^m$, then the rank of A equals the rank of A^*.

Proof:

If the rank of A is rank(A), then there must exists rank(A) vectors in $\mathbf{F}^m$ in such that

$$\mathbf{v}_1, \ldots, \mathbf{v}_{\text{rank}(A)}$$

form a basis for range(A). Because all of these images are in the range of A, this means that there must exist rank(A) vectors in $\mathbf{F}^n$ such that

$$\mathbf{v}_1 = A\mathbf{u}_1, \ldots, \mathbf{v}_{\text{rank}(A)} = A\mathbf{u}_{\text{rank}(A)}.$$

Now, from the previous theorem, the vectors

$$\mathbf{u}_1, \ldots, \mathbf{u}_{\text{rank}(A)}$$

must be linearly independent. Now, the adjoint (transpose for real matrices, conjugate transpose for complex matrices) maps $A^*: \mathbf{F}^m \to \mathbf{F}^n$ and thus the composition $A^*A: \mathbf{F}^n \to \mathbf{F}^n$. Note that both A and A^*A have the same domain: $\mathbf{F}^m$.

Now, suppose that $\mathbf{u}$ is in the null space of A; that is, $A\mathbf{u} = \mathbf{0}_m$. In this case, $A^*A\mathbf{u} = A^*\mathbf{0}_m = \mathbf{0}_n$, so $\mathbf{u}$ is also in the null space of A^*A. Suppose that $A\mathbf{u} \neq \mathbf{0}_m$. In this case, $\|A\mathbf{u}\|_2 = \langle A\mathbf{u}, A\mathbf{u} \rangle > 0$ so $\left\langle A^*A\mathbf{u}, \mathbf{u} \right\rangle > 0$. Thus, $A^*A\mathbf{u} \neq \mathbf{0}_m$.

Therefore, the null space of A and A^*A are identical, and thus dim(null(A)) = dim(null(A^*A)). However, for any finite-dimensional linear mapping, the dimension of the null space plus the dimension of the range equals the dimension of the domain; that is, in this case,

$$\dim(\text{null}(A)) + \dim(\text{range}(A)) = n \text{ and } \dim(\text{null}(A^*A)) + \dim(\text{range}(A^*A)) = n.$$

Therefore, because two of the three of each of these equations are equal, it follows that

$$\text{rank}(A) = \dim(\text{range}(A)) = \dim(\text{range}(A^*A)) = \text{rank}(A^*A).$$

Now, given that dim(range(A^*A)), there exists a basis for that range

$$A^*A\mathbf{u}_1, \ldots, A^*A\mathbf{u}_{\text{rank}(A)}.$$

Consequently, there are rank(A) linearly independent vectors that are images of vectors in $\mathbf{F}^m$:

$$A^*\mathbf{v}_1, \ldots, A^*\mathbf{v}_{\text{rank}(A)}$$

Because these are linearly independent, each contributes one towards the dimension of the range of A^*. Thus, the rank of A^* is at least equal to the rank of A, as the span of these vectors may only form a subspace of the range.

$$\text{rank}\left(A^*\right) = \dim\left(\text{range}\left(A^*\right)\right) \geq \text{rank}\left(A\right).$$

Now, repeat this entire process, but start with A^*: recall that $A^* : \mathbf{F}^m \to \mathbf{F}^n$. If the rank of A^* is rank(A^*), then there must exists rank(A^*) vectors in $\mathbf{F}^n$ in such that

$$\mathbf{u}_1, \ldots, \mathbf{u}_{\text{rank}(A^*)}$$

form a basis for range(A^*). Because all of these images are in the range of A^*, this means that there must exist rank(A^*) vectors in $\mathbf{F}^m$ such that

$$\mathbf{u}_1 = A\mathbf{v}_1, \ldots, \mathbf{u}_{\text{rank}(A^*)} = A\mathbf{v}_{\text{rank}(A^*)}.$$

Now, from the previous theorem, the vectors

$$\mathbf{v}_1, \ldots, \mathbf{v}_{\text{rank}(A^*)}$$

must be linearly independent. Now, the adjoint of the adjoint (transpose of the transpose for real matrices, conjugate transpose of the conjugate transpose for complex matrices) maps $\left(A^*\right)^* : \mathbf{F}^n \to \mathbf{F}^m$ and $\left(A^*\right)^* = A$, and thus the composition $AA^* : \mathbf{F}^m \to \mathbf{F}^m$. Note that both A and AA^* have the same domain: $\mathbf{F}^n$.

Now, suppose that $\mathbf{v}$ is in the null space of A^*; that is, $A^*\mathbf{v} = \mathbf{0}_n$. In this case, $AA^*\mathbf{v} = A^*\mathbf{0}_n = \mathbf{0}_m$, so $\mathbf{v}$ is also in the null space of AA^*. Suppose that $A^*\mathbf{v} \neq \mathbf{0}_n$. In this case, $\left\|A^*\mathbf{v}\right\|_2 = \left\langle A^*\mathbf{v}, A^*\mathbf{v}\right\rangle > 0$ so $\left\langle \left(A^*\right)^* A^*\mathbf{v}, \mathbf{v}\right\rangle = \left\langle AA^*\mathbf{v}, \mathbf{v}\right\rangle > 0$. Thus, $AA^*\mathbf{v} \neq \mathbf{0}_m$.

Therefore, the null space of A^* and AA^* are identical, and thus dim(null(A^*)) = dim(null(AA^*)). However, for any finite-dimensional linear mapping, the dimension of the null space plus the dimension of the range equals the dimension of the domain; that is, in this case,

$$\dim(\text{null}(A^*)) + \dim(\text{range}(A^*)) = m \text{ and } \dim(\text{null}(AA^*)) + \dim(\text{range}(AA^*)) = m.$$

Therefore, because two of the three of each of these equations are equal, it follows that

$$\text{rank}(A^*) = \dim(\text{range}(A^*)) = \dim(\text{range}(AA^*)) = \text{rank}(AA^*).$$

Now, given that dim(range(AA^*)), there exists a basis for that range

$$AA^*\mathbf{v}_1, \ldots, AA^*v_{\text{rank}(A^*)}.$$

Consequently, there are rank(A^*) linearly independent vectors that are images of vectors in $\mathbf{F}^n$:

$$A\mathbf{u}_1, \ldots, A\mathbf{u}_{\text{rank}(A^*)}.$$

Because these are linearly independent, each contributes one towards the dimension of the range of *A*. Thus, the rank of *A* is at least equal to the rank of A^*, as the span of these vectors may only form a subspace of the range.

$$\text{rank}(A) = \dim(\text{range}(A)) \geq \text{rank}(A^*).$$

Therefore, as both $\text{rank}(A^*) \geq \text{rank}(A)$ and $\text{rank}(A) \geq \text{rank}(A^*)$ are true, it follows that $\text{rank}(A) = \text{rank}(A^*)$. ∎

In Matlab, the transpose of a matrix is found using the apostrophe operator:

```
>> A = rand( 3, 3 )
    A =
        0.8147    0.9134    0.2785
        0.9058    0.6324    0.5469
        0.1270    0.0975    0.9575
>> A'
    ans =
        0.8147    0.9058    0.1270
        0.9134    0.6324    0.0975
        0.2785    0.5469    0.9575

>> B = rand( 5, 2 )
    B =
        0.9595    0.6787
        0.6557    0.7577
        0.0357    0.7431
        0.8491    0.3922
        0.9340    0.6555

>> B'
    ans =
        0.9595    0.6557    0.0357    0.8491    0.9340
        0.6787    0.7577    0.7431    0.3922    0.6555
>> u = [1 2]';
>> v = [3 2 1 0 -1]';
>> (B*u)'*v
    ans =
       10.5704

>> u'*(B'*v)
    ans =
       10.5704
```

Problems

1. If $A: \mathbf{R}^3 \to \mathbf{R}^5$, describe the shape of A and A^{T} and the domain and codomain of A^{T}.

2. If $A^{\mathrm{T}}: \mathbf{R}^4 \to \mathbf{R}^2$, describe the shape of A and A^{T} and domain and codomain of A.

3. Find the transpose of the following three matrices:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix}.$$

4. Find the transpose of the following three matrices:

$$\begin{pmatrix} 1.2 & -2.1 & 0.4 & 0.1 \\ -0.7 & 1.7 & 0.8 & -0.9 \end{pmatrix}, \begin{pmatrix} 3.7 & -0.8 \\ 0.9 & 4.5 \end{pmatrix}, \begin{pmatrix} -0.5 & 0.1 \\ 0.4 & 1.7 \\ 0.8 & 0.3 \\ 0.6 & 0.2 \\ 0.3 & 0.1 \\ -0.1 & 0 \end{pmatrix}.$$

5. Demonstrate that the definition of the adjoint holds for the transpose with the matrix $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$ by calculating both $\langle A\mathbf{u}, \mathbf{v} \rangle$ and $\langle \mathbf{u}, A^{\mathrm{T}}\mathbf{v} \rangle$ first with $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and a second time with $\mathbf{u} = \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$.

6. Demonstrate that the definition of the adjoint holds for the transpose with the matrix $A = \begin{pmatrix} 4 & -2 & 2 \\ -2 & 1 & -1 \\ 0 & 1 & 3 \\ -2 & 2 & 2 \end{pmatrix}$ by calculating both $\langle A\mathbf{u}, \mathbf{v} \rangle$ and $\langle \mathbf{u}, A^{\mathrm{T}}\mathbf{v} \rangle$ first with $\mathbf{u} = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 1 \end{pmatrix}$, a second time with $\mathbf{u} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ 1 \\ -1 \\ 1 \end{pmatrix}$, and a third time with $\mathbf{u} = \begin{pmatrix} -2 \\ -3 \\ 1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 23 \\ 46 \\ 123 \\ 32 \end{pmatrix}$

Solutions

1. A must be a 5 × 3 matrix, and thus A^{T} must be a 3 × 5 matrix, and $A^{\mathrm{T}} : \mathbf{R}^3 \to \mathbf{R}^5$.

3. Always remember to ensure the diagonal entries ($a_{i,i}$) remain unchanged and the first row becomes the first column, and so on and so forth:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} 1 & 6 & 11 \\ 2 & 7 & 12 \\ 3 & 8 & 13 \\ 4 & 9 & 14 \\ 5 & 10 & 15 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{pmatrix}.$$

5. $A\mathbf{u} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$, so if $\mathbf{v} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$, $\langle A\mathbf{u}, \mathbf{v} \rangle = 2 + 2 = 4$. Also, $A^{\mathrm{T}}\mathbf{v} = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$ and thus $\langle \mathbf{u}, A^{\mathrm{T}}\mathbf{v} \rangle = 2 + 2 + 0 = 4$.

Similarly, $A\mathbf{u} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} -1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 6 \\ 12 \end{pmatrix}$, so if $\mathbf{v} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$, $\langle A\mathbf{u}, \mathbf{v} \rangle = -12 + 12 = 0$, so $A\mathbf{u}$ and $\mathbf{v}$ are orthogonal.

Also, $A^{\mathrm{T}}\mathbf{v} = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$ and thus $\langle \mathbf{u}, A^{\mathrm{T}}\mathbf{v} \rangle = -2 + 2 + 0 = 0$, so $\mathbf{u}$ and $A^{\mathrm{T}}\mathbf{v}$ are orthogonal.

11.3 The adjoint for complex finite-dimensional vector spaces

If A is the matrix representation of a linear operator in a finite-dimensional complex vector space, because $\langle A\mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, A^*\mathbf{v} \rangle$, it follows that

$$\begin{aligned} \langle A\mathbf{u}, \mathbf{v} \rangle &= \sum_{i=1}^{n} (A\mathbf{u})_i^* v_i \\ &= \sum_{i=1}^{n} \left(\sum_{j=1}^{n} (a_{i,j})^* u_j^* \right) v_i \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \left(v_i (a_{i,j})^* u_j \right) \end{aligned}$$

but

$$\begin{aligned} \langle \mathbf{u}, A^*\mathbf{v} \rangle &= \sum_{i=1}^{n} u_i (A^*\mathbf{v})_i \\ &= \sum_{i=1}^{n} u_i^* \left(\sum_{j=1}^{n} a_{i,j}^* v_j \right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \left(u_i^* a_{i,j}^* v_j \right) \end{aligned}$$

and thus it follows that, with a similar rearrangement, $\sum_{i=1}^{n} \sum_{j=1}^{n} \left(v_i (a_{i,j})^* u_j^* \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(v_i a_{j,i}^* u_j^* \right)$, and so therefore, $a_{j,i}^* = (a_{i,j})^*$. Like the case of the real finite-dimensional matrices, this matrix A^* is called the *conjugate transpose*.[27] This course will always use A^*. For example, if

$$\mathbf{u} = \begin{pmatrix} 2 + j \\ 3 \\ -1 - j \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} 1 - j \\ -4j \\ 2 \end{pmatrix} \text{ and } A = \begin{pmatrix} 1 + 2j & 2 + j & 3 \\ 4 - j & 5 & 6j \\ 7 & 8 - j & 9 \end{pmatrix},$$

then

[27] The conjugate transpose is also called the *Hermitian transpose*, and can be represented by either A^{H} or $A^{\dagger}$ instead of A^*. In this course, we will always use the terminology of conjugate transpose and A^*.

$$A\mathbf{u} = \begin{pmatrix} 3+5j \\ 30-4j \\ 29-5j \end{pmatrix} \text{ so } \langle A\mathbf{u}, \mathbf{v} \rangle = \begin{pmatrix} 3-5j & 30-4j & 29-5j \end{pmatrix} \begin{pmatrix} 1-j \\ -4j \\ 2 \end{pmatrix} = 72-118j$$

while

$$A^* = \begin{pmatrix} 1-2j & 4+j & 7 \\ 2-j & 5 & 8+j \\ 3 & -6j & 9 \end{pmatrix} \text{ so } A^*\mathbf{v} = \begin{pmatrix} 17-19j \\ 17-21j \\ -3-3j \end{pmatrix} \text{ and thus } \langle \mathbf{u}, A^*\mathbf{v} \rangle = \begin{pmatrix} 2-j & 3 & -1+j \end{pmatrix} \begin{pmatrix} 17-19j \\ 17-21j \\ -3-3j \end{pmatrix} = 72-118j \text{ .}$$

In Matlab, the apostrophe operator actually does compute the conjugate transpose if any of the entries are complex:

```
>> A = rand( 3, 3 ) + 1j*rand( 3, 3 )
    A =
       0.1712 + 0.9502i   0.2769 + 0.3816i   0.8235 + 0.1869i
       0.7060 + 0.0344i   0.0462 + 0.7655i   0.6948 + 0.4898i
       0.0318 + 0.4387i   0.0971 + 0.7952i   0.3171 + 0.4456i
>> A'
    ans =
       0.1712 - 0.9502i   0.7060 - 0.0344i   0.0318 - 0.4387i
       0.2769 - 0.3816i   0.0462 - 0.7655i   0.0971 - 0.7952i
       0.8235 - 0.1869i   0.6948 - 0.4898i   0.3171 - 0.4456i

>> B = rand( 5, 2 ) + 1j*rand( 5, 2 )
    B =
       0.6463 + 0.3404i   0.6551 + 0.5060i
       0.7094 + 0.5853i   0.1626 + 0.6991i
       0.7547 + 0.2238i   0.1190 + 0.8909i
       0.2760 + 0.7513i   0.4984 + 0.9593i
       0.6797 + 0.2551i   0.9597 + 0.5472i

>> B'
    ans =
       0.6463-0.3404i  0.7094-0.5853i  0.7547-0.2238i  0.2760-0.7513i  0.6797-0.2551i
       0.6551-0.5060i  0.1626-0.6991i  0.1190-0.8909i  0.4984-0.9593i  0.9597-0.5472i
>> u = [1; 2-1j];
>> v = [-2j; 3-4j; 2+3j; -1; 2-1j];
>> (B*u)'*v
    ans =
       9.6214 - 17.1990i

>> u'*(B'*v)
    ans =
       9.6214 - 17.1990i
```

Problems

1. If $A: \mathbf{C}^3 \to \mathbf{C}^5$, describe the shape of A and A^* and the domain and codomain of A^*.

2. If $A^*: \mathbf{C}^4 \to \mathbf{C}^2$, describe the shape of A and A^* and domain and codomain of A.

3. Find the conjugate transpose of the following three matrices:

$$\begin{pmatrix} 1-j & 2 & 3j & 4+j & -5 \\ -6j & 7-j & -8 & 9 & -10j \\ 11 & 12j & 13+2j & 14-j & 15+16j \end{pmatrix}, \begin{pmatrix} 1-j & 2j & 3+2j \\ 4-3j & 5 & 6j \\ -7 & -8j & 9j \end{pmatrix}, \begin{pmatrix} 1+j & 1 & 0 \\ 1 & 2-j & 1 \\ 0 & 1 & -3-2j \\ 0 & 0 & 1 \end{pmatrix}.$$

4. Find the conjugate transpose of the following three matrices:

$$\begin{pmatrix} 1.2-0.3j & -2.1 & 0.4+0.2j & 0.1-0.4j \\ -0.7+0.3j & 1.7-0.9j & 0.8j & -0.9-0.2j \end{pmatrix}, \begin{pmatrix} 3.7-4.7j & -0.8+0.6j \\ 0.9-0.9j & 4.5+5.7j \end{pmatrix}, \begin{pmatrix} -0.5-0.3j & 0.1+0.7j \\ 0.4+0.1j & 1.7-0.4j \\ 0.8-0.7j & 0.3+0.1j \\ 0.6+1.2j & 0.2+0.7j \\ 0.3-4.5j & 0.1-0.8j \\ -0.1 & 0 \end{pmatrix}.$$

5. Demonstrate that the definition of the adjoint holds for the conjugate transpose with the matrix $A = \begin{pmatrix} 1 & -2j & 3j \\ -4j & 5 & -6j \end{pmatrix}$ by calculating both $\langle A\mathbf{u}, \mathbf{v} \rangle$ and $\langle \mathbf{u}, A^* \mathbf{v} \rangle$ first with $\mathbf{u} = \begin{pmatrix} 1 \\ 2j \\ -2 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ -j \end{pmatrix}$, and a second time with $\mathbf{u} = \begin{pmatrix} -0.6j \\ 2j \\ 1 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1 \\ -j \end{pmatrix}$.

6. Demonstrate that the definition of the adjoint holds for the conjugate transpose with the matrix $A = \begin{pmatrix} 2+j & 1 & 0 \\ 1 & 3j & 1 \\ 0 & 1 & 1-j \\ 0 & 0 & 1 \end{pmatrix}$ by calculating both $\langle A\mathbf{u}, \mathbf{v} \rangle$ and $\langle \mathbf{u}, A^{\mathrm{T}} \mathbf{v} \rangle$ first with $\mathbf{u} = \begin{pmatrix} 0 \\ 0 \\ j \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 1+j \\ 0 \\ 0 \\ -1 \end{pmatrix}$, and a second time with $\mathbf{u} = \begin{pmatrix} 0 \\ 0 \\ j \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} 2+3j \\ 1 \\ 0 \\ -1 \end{pmatrix}$.

Solutions

1. A must be a 5 × 3 matrix, and thus A^* must be a 3 × 5 matrix, and $A^* : \mathbf{R}^3 \to \mathbf{R}^5$.

3. Always remember to ensure the diagonal entries ($a_{i,i}$) remain unchanged and the first row becomes the first column, and so on and so forth:

$$\begin{pmatrix} 1-j & 2 & 3j & 4+j & -5 \\ -6j & 7-j & -8 & 9 & -10j \\ 11 & 12j & 13+2j & 14-j & 15+16j \end{pmatrix}^* = \begin{pmatrix} 1+j & 6j & 11 \\ 2 & 7+j & -12j \\ -3j & -8 & 13-2j \\ 4-j & 9 & 14+j \\ -5 & 10j & 15-16j \end{pmatrix},$$

$$\begin{pmatrix} 1-j & 2j & 3+2j \\ 4-3j & 5 & 6j \\ -7 & -8j & 9j \end{pmatrix}^* = \begin{pmatrix} 1+j & 4+3j & -7 \\ -2j & 5 & 8j \\ 3-2j & -6j & -9j \end{pmatrix}, \quad \begin{pmatrix} 1+j & 1 & 0 \\ 1 & 2-j & 1 \\ 0 & 1 & -3-2j \\ 0 & 0 & 1 \end{pmatrix}^* = \begin{pmatrix} 1-j & 1 & 0 & 0 \\ 1 & 2+j & 1 & 0 \\ 0 & 1 & -3+2j & 1 \end{pmatrix}.$$

5. $A\mathbf{u} = \begin{pmatrix} 1 & -2j & 3j \\ -4j & 5 & -6j \end{pmatrix}\begin{pmatrix} 1 \\ 2j \\ -2 \end{pmatrix} = \begin{pmatrix} 5-6j \\ 18j \end{pmatrix}$, so if $\mathbf{v} = \begin{pmatrix} 1 \\ -j \end{pmatrix}$, $\langle A\mathbf{u}, \mathbf{v} \rangle = 5+6j-18 = -13+6j$.

Also, $A^*\mathbf{v} = \begin{pmatrix} 1 & 4j \\ 2j & 5 \\ -3j & 6j \end{pmatrix}\begin{pmatrix} 1 \\ -j \end{pmatrix} = \begin{pmatrix} 5 \\ -3j \\ 6-3j \end{pmatrix}$ and thus $\langle \mathbf{u}, A^*\mathbf{v} \rangle = 5-6-12+6j = -13+6j$.

Similarly, $A\mathbf{u} = \begin{pmatrix} 1 & -2j & 3j \\ -4j & 5 & -6j \end{pmatrix}\begin{pmatrix} -0.6 \\ 2j \\ 1 \end{pmatrix} = \begin{pmatrix} 4+2.4j \\ -2.4+4j \end{pmatrix}$, so if $\mathbf{v} = \begin{pmatrix} 1 \\ -j \end{pmatrix}$, $\langle A\mathbf{u}, \mathbf{v} \rangle = (4-2.4j)+(-4+2.4j) = 0$,

so $A\mathbf{u}$ and $\mathbf{v}$ are orthogonal. Also, $A^*\mathbf{v} = \begin{pmatrix} 1 & 4j \\ 2j & 5 \\ -3j & 6j \end{pmatrix}\begin{pmatrix} 1 \\ -j \end{pmatrix} = \begin{pmatrix} 5 \\ -3j \\ 6-3j \end{pmatrix}$ and thus $\langle \mathbf{u}, A^*\mathbf{v} \rangle = 3j-6+6-3j = 0$, so $\mathbf{u}$ and $A^*\mathbf{v}$ are orthogonal.

11.4 A new definition of the inner product

To this point, we have used the bracket notation for inner products. Suppose, however, that we write a vector in the form

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_1 \\ \vdots \\ u_n \end{pmatrix}$$

and now write the adjoint (be it the transpose or conjugate transpose) as

$$\mathbf{u}^{\mathrm{T}} = \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix} \text{ or } \mathbf{u}^{*} = \begin{pmatrix} u_1^{*} & u_2^{*} & \cdots & u_n^{*} \end{pmatrix}.$$

In this case, does not the inner product $\langle \mathbf{u}, \mathbf{v} \rangle$ closely resemble

$$\mathbf{u}^{\mathrm{T}}\mathbf{v} = \begin{pmatrix} u_1 & u_2 & \cdots & u_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_1 \\ \vdots \\ v_n \end{pmatrix} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n$$

or

$$\mathbf{u}^{*}\mathbf{v} = \begin{pmatrix} u_1^{*} & u_2^{*} & \cdots & u_n^{*} \end{pmatrix} \begin{pmatrix} v_1 \\ v_1 \\ \vdots \\ v_n \end{pmatrix} = u_1^{*} v_1 + u_2^{*} v_2 + \cdots + u_n^{*} v_n \, ?$$

This is a notation of convenience, but it is often used by engineers, and it is a notation used in Matlab, where the inner product is write as `u'*v`.

It should be important to note that **u** must be a vector. If we were to multiply the vector **v** by the adjoint of an $n \times 1$ matrix, the result would not be a scalar, but rather a 1-dimensional vector. The distinction may appear to be a triviality or subtlety, but remember that a matrix is a linear operator between vector spaces while the inner product is a scalar. Throughout this course, we will continue to use the bracket notation, but you must also be familiar with the use of the adjoint to represent the inner product.

11.5 Self-adjoint and skew-adjoint operators

A linear operator $A: \mathcal{U} \to \mathcal{U}$ is said to be *self-adjoint* if $A = A^*$, whereas it is said to be *skew-adjoint* if $A = -A^*$, or, equivalently, $-A = A^*$. Thus, for a self-adjoint linear operator,

$$\langle A\mathbf{u}_1, \mathbf{u}_2 \rangle = \langle \mathbf{u}_1, A\mathbf{u}_2 \rangle$$

for all vectors **u** and **v**, while for a skew-adjoint operator,

$$\langle A\mathbf{u}_1, \mathbf{u}_2 \rangle = \langle \mathbf{u}_1, -A\mathbf{u}_2 \rangle = -\langle \mathbf{u}_1, A\mathbf{u}_2 \rangle .$$

It is necessary that *A* be a linear operator; a linear mapping between two different vector spaces can never satisfy this requirement. We have two theorems about the properties of self-adjoint and skew-adjoint operators.

11.5.1 Symmetric and skew-symmetric matrices in real finite-dimensional vector spaces

The class of self-adjoint linear operators for finite-dimensional real vector spaces have matrix representations of the class of *symmetric* square matrices where if $A = \left(a_{i,j}\right)$, then $a_{i,j} = a_{j,i}$. An example of a symmetric matrix in $\mathbf{R}^4$ is

$$A = \begin{pmatrix} 3 & 1 & 4 & 1 \\ 1 & 6 & 3 & -2 \\ 4 & 3 & 7 & 4 \\ 1 & -2 & 4 & 9 \end{pmatrix}.$$

The class of skew-adjoint linear operators for finite-dimensional real vector spaces are the class of *skew-symmetric* matrices where if $A = \left(a_{i,j}\right)$ then $a_{i,j} = -a_{j,i}$. As a consequence of this, we first note that the diagonal entries must be zero: if $a_{j,j} = -a_{j,j}$, it follows that $a_{j,j} = 0$. An example of a skew-symmetric matrix in $\mathbf{R}^4$ is

$$A = \begin{pmatrix} 0 & 1 & 4 & 1 \\ -1 & 0 & 3 & -2 \\ -4 & -3 & 0 & 4 \\ -1 & 2 & -4 & 0 \end{pmatrix}.$$

11.5.2 Conjugate symmetric and conjugate skew-symmetric matrices in complex finite-dimensional vector spaces

If *A* is the matrix representation of a linear operator in a complex vector space, it is self-adjoint if the matrix is *conjugate symmetric*[28]; that is, if it equals its conjugate transpose, or $a_{i,j} = a^*_{j,i}$. As a consequence of this, we note that the diagonal entries must be real: if $a_{j,j} = a^*_{j,j}$, it follows that $a_{j,j}$ is real. As an example, the following matrix is conjugate symmetric:

[28] Conjugate symmetric matrices are also called *Hermitian* matrices, and the relationship may be stated as either $A = A^{\mathrm{H}}$ or $A = A^{\dagger}$. In this course, we will always refer to such matrices as being conjugate symmetric.

$$A = \begin{pmatrix} 11 & j & 4 & 1 \\ -j & 8 & 3-j & 2+3j \\ 4 & 3+j & 7 & -2j \\ 1 & 2-3j & 2j & 9 \end{pmatrix}.$$

A linear operator is skew-adjoint, *conjugate–skew-symmetric*[29] if $a_{i,j} = -a^*_{j,i}$. In this case, the diagonal entries must therefore be purely imaginary, for $a_{j,j} = -a^*_{j,j}$ implies as much. As an example, the following matrix is conjugate skew-symmetric:

$$A = \begin{pmatrix} 11j & j & 4 & -1 \\ j & 0 & 3-j & -2-3j \\ -4 & -3-j & 7j & 2j \\ 1 & 2-3j & 2j & 9j \end{pmatrix}.$$

11.5.3 Other self-adjoint operators

In your quantum mechanics course, you will use self-adjoint operators for the vector space of square integrable functions in order to extract properties from wave functions.

[29] [29] Conjugate–skew-symmetric matrices are also called *skew-Hermitian* matrices, and the relationship may be stated as either $A = -A^{\mathrm{H}}$ or $A = -A^{\dagger}$. In this course, we will always refer to such matrices as being conjugate skew-symmetric.

11.6 Results regarding self-adjoint and skew-adjoint linear operators

Theorem

For any linear mapping $A:\mathcal{U}\to\mathcal{V}$, $A^*A:\mathcal{U}\to\mathcal{U}$ and $AA^*:\mathcal{V}\to\mathcal{V}$ are self-adjoint operators.

Proof:

For the first case,

$$\begin{aligned}\left\langle\left(AA^*\right)\mathbf{u},\mathbf{v}\right\rangle &= \left\langle A\left(A^*\mathbf{u}\right),\mathbf{v}\right\rangle\\ &= \left\langle A^*\mathbf{u},A^*\mathbf{v}\right\rangle\\ &= \left\langle\mathbf{u},A\left(A^*\mathbf{v}\right)\right\rangle\\ &= \left\langle\mathbf{u},\left(AA^*\right)\mathbf{v}\right\rangle\end{aligned}$$

and therefore $\left(AA^*\right)^* = AA^*$, although, it also follows from the fact that $\left(AA^*\right)^* = \left(A^*\right)^*A^* = AA^*$. The proof that the other is self-adjoint is left to the reader. ∎

For example, if $A=\begin{pmatrix}5&1&2\\-1&6&0\\3&4&10\end{pmatrix}$ is a real matrix, then we see that $AA^{\mathrm{T}}=\begin{pmatrix}30&1&39\\1&37&21\\39&21&125\end{pmatrix}$ and $A^{\mathrm{T}}A=\begin{pmatrix}35&11&40\\11&53&42\\40&42&104\end{pmatrix}$ are both symmetric. Similarly, if $A=\begin{pmatrix}3-j&4+2j\\6&-3j\end{pmatrix}$ is a complex matrix and we see that both $AA^*=\begin{pmatrix}30&12+6j\\12-6j&45\end{pmatrix}$ and $A^*A=\begin{pmatrix}46&10-8j\\10+8j&29\end{pmatrix}$ are conjugate symmetric.

Theorem

For any linear operator $A:\mathcal{U}\to\mathcal{U}$, $A + A^*$ is self-adjoint and $A - A^*$ is skew-adjoint; that is, for

1. real finite-dimensional vector spaces, $A+A^{\mathrm{T}}$ is symmetric and $A-A^{\mathrm{T}}$ is skew-symmetric, and
2. complex finite-dimensional vector spaces, $A+A^*$ is conjugate symmetric and $A-A^*$ is conjugate skew-symmetric.

Proof:

We will prove this using the definition of the adjoint of a linear operator:

$$\begin{aligned}\left\langle\left(A+A^*\right)\mathbf{u},\mathbf{v}\right\rangle &= \left\langle A\mathbf{u},\mathbf{v}\right\rangle+\left\langle A^*\mathbf{u},\mathbf{v}\right\rangle\\ &= \left\langle\mathbf{u},A^*\mathbf{v}\right\rangle+\left\langle\mathbf{u},\left(A^*\right)^*\mathbf{v}\right\rangle\\ &= \left\langle\mathbf{u},A^*\mathbf{v}\right\rangle+\left\langle\mathbf{u},A\mathbf{v}\right\rangle\\ &= \left\langle\mathbf{u},A\mathbf{v}\right\rangle+\left\langle\mathbf{u},A^*\mathbf{v}\right\rangle\\ &= \left\langle\mathbf{u},\left(A+A^*\right)\mathbf{v}\right\rangle\end{aligned}$$

but again, it also follows from previous properties: $\left(A+A^*\right)^* = A^*+\left(A^*\right)^* = A^*+A = A+A^*$. The proof that $A-A^*$ is skew-adjoint is left to the reader. ∎

For example, if $A=\begin{pmatrix} 5 & 1 & 2 \\ -1 & 6 & 0 \\ 3 & 4 & 10 \end{pmatrix}$ is a real matrix, we see that

$$A+A^{\mathrm{T}}=\begin{pmatrix} 10 & 0 & 5 \\ 0 & 12 & 4 \\ 5 & 4 & 20 \end{pmatrix} \text{ and } A-A^{\mathrm{T}}=\begin{pmatrix} 0 & 2 & -1 \\ -2 & 0 & -4 \\ 1 & 4 & 0 \end{pmatrix}$$

are symmetric and skew symmetric, respectively.

Similarly, if $A=\begin{pmatrix} 3-j & 4+2j \\ 6 & -3j \end{pmatrix}$ is a complex matrix, we see that

$$A+A^{*}=\begin{pmatrix} 6 & 10+2j \\ 10-2j & 0 \end{pmatrix} \text{ and } A-A^{*}=\begin{pmatrix} -2j & -2+2j \\ 2+2j & -6j \end{pmatrix}$$

are conjugate symmetric and conjugate skew-symmetric, respectively.

Theorem
Every linear operator is the sum of a self-adjoint and skew-adjoint operator.

Proof:
Given a linear operator *A*, it follows that

$$\begin{aligned} A &= \frac{1}{2}A+\frac{1}{2}A+\frac{1}{2}A^{*}-\frac{1}{2}A^{*} \\ &= \frac{1}{2}\left(A+A^{*}\right)+\frac{1}{2}\left(A-A^{*}\right) \end{aligned}$$

where, from the previous results, the first is self-adjoint and the second is skew-adjoint. ∎

In our previous two examples, we see that $\begin{pmatrix} 5 & 1 & 2 \\ -1 & 6 & 0 \\ 3 & 4 & 10 \end{pmatrix}=\begin{pmatrix} 5 & 0 & 2.5 \\ 0 & 6 & 2 \\ 2.5 & 2 & 10 \end{pmatrix}+\begin{pmatrix} 0 & 1 & -0.5 \\ -1 & 0 & -2 \\ 0.5 & 2 & 0 \end{pmatrix}$ and $\begin{pmatrix} 3-j & 4+2j \\ 6 & -3j \end{pmatrix}=\begin{pmatrix} 3 & 5+j \\ 5-j & 0 \end{pmatrix}+\begin{pmatrix} -j & -1+1j \\ 1+j & -3j \end{pmatrix}$.

You may remember the definition of an even function (one where $f(-x) = f(x)$) and an odd function (one where $f(-x) = -f(-x)$). You may have then learned that every function is the sum of an even function plus an odd function, and you found those by defining two new functions $f_{\text{even}}(x)\overset{\text{def}}{=}\frac{f(x)+f(-x)}{2}$ and $f_{\text{odd}}(x)\overset{\text{def}}{=}\frac{f(x)-f(-x)}{2}$, the first of which is even, and the second of which is odd, and $f(x)=f_{\text{even}}(x)+f_{\text{odd}}(x)$. For example, given the polynomial $p(x) = x^3 - 4x^2 + 3x + 1$, we can calculate $p_{\text{even}}(x) = -4x^2 + 1$ and $p_{\text{odd}}(x) = x^3 + 3x$.

11.7 Normal operators

A linear operator *A* is *normal* if it commutes with its adjoint, that is, $AA^* = A^*A$. Clearly, $A:\mathcal{U}\to\mathcal{U}$ must map a vector space onto itself, otherwise, AA^* and A^*A would be mappings in different vector spaces. For finite dimensional linear operators, this says that they must have square matrices representations. Not all matrices are normal; for example,

$$\begin{pmatrix}1 & 2\\ 3 & 4\end{pmatrix}\begin{pmatrix}1 & 3\\ 2 & 4\end{pmatrix}=\begin{pmatrix}5 & 11\\ 11 & 25\end{pmatrix} \text{ but } \begin{pmatrix}1 & 3\\ 2 & 4\end{pmatrix}\begin{pmatrix}1 & 2\\ 3 & 4\end{pmatrix}=\begin{pmatrix}10 & 14\\ 14 & 20\end{pmatrix}.$$

We will look at some linear operators that are normal.

Theorem
Every self-adjoint linear operator is normal.

Proof:
As the linear operator is self-adjoint, $A = A^*$, and therefore $AA^* = A^2 = AA^*$. ∎

Theorem
Every skew-adjoint linear operator is normal.

Proof:
As the linear operator is skew-adjoint, $A = -A^*$, and therefore $AA^* = -A^2 = AA^*$. ∎

Not every normal matrix is either self or skew adjoint. An example, given in Wikipedia, is

$$A=\begin{pmatrix}1 & 1 & 0\\ 0 & 1 & 1\\ 1 & 0 & 1\end{pmatrix} \text{ where } AA^*=\begin{pmatrix}2 & 1 & 1\\ 1 & 2 & 1\\ 1 & 1 & 2\end{pmatrix}=A^*A.$$

For 2-dimensional real matrices, we can categorize all normal matrices as follows:

If $A:\mathbf{R}^2\to\mathbf{R}^2$ then $A=\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}$ and so $AA^{\mathrm{T}}=\begin{pmatrix}\alpha^2+\beta^2 & \alpha\gamma+\beta\delta\\ \alpha\gamma+\beta\delta & \gamma^2+\delta^2\end{pmatrix}$ and $A^{\mathrm{T}}A=\begin{pmatrix}\alpha^2+\gamma^2 & \alpha\beta+\gamma\delta\\ \alpha\beta+\gamma\delta & \beta^2+\delta^2\end{pmatrix}$ and thus it follows that for *A* to be normal, first $\beta^2+\gamma^2$ so $\beta=\pm\gamma$, and thus we have two cases to consider:

1. if $\beta=\gamma$, it follows that $\alpha\gamma+\beta\delta=\alpha\beta+\beta\delta=\alpha\beta+\gamma\delta$, so there are no restrictions on α and δ, but
2. if $\beta=-\gamma$, it follows that $ac+bd=-ab+bd$ and $ab+cd=ab-bd$, so for $-ab+bd=ab-bd$, we require that $b(d-a)=b(a-d)$, so either $b=0$ or $a=d$.

Consequently, real normal matrices are either of the form

$$A=\begin{pmatrix}\alpha & \beta\\ \beta & \delta\end{pmatrix} \text{ or } A=\begin{pmatrix}\alpha & \beta\\ -\beta & \alpha\end{pmatrix}.$$

For a complex linear operator $A:\mathbf{C}^2\to\mathbf{C}^2$, we have that $|b|^2=|c|^2$ so $|b|=|c|$, so either $b=c=0$ or $c=be^{j\theta}$ for $0\le\theta<2\pi$ and then *a* and *b* must satisfy $\left(b(a-d)^*e^{j\theta}\right)^*=b(a-d)^*$.

Problems

1. Prove that for any matrix A, AA^* and A^*A are normal.

2. Prove that for any matrix A, $A + A^*$ and $A - A^*$ are normal.

3. For what values of α is the complex matrix $\begin{pmatrix} 1 & \alpha \\ 1 & 1 \end{pmatrix}$ normal?

4. For what values of α and β is the complex matrix $\begin{pmatrix} 1 & \alpha \\ \beta & 2 \end{pmatrix}$ normal?

Solutions

1. The easy way: AA^* and A^*A are self-adjoint, and all self-adjoint matrices are normal.

The hard way: $(AA^*)(AA^*)^* = AA^*(A^*)^*A^* = AA^*AA^* = (A^*)^*A^*AA^* = (AA^*)^*(AA^*)$.

3. $\begin{pmatrix} 1 & \alpha \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & \alpha \\ 1 & 1 \end{pmatrix}^* = \begin{pmatrix} 1 & \alpha \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ \alpha & 1 \end{pmatrix} = \begin{pmatrix} 1+\alpha^2 & 1+\alpha \\ 1+\alpha & 2 \end{pmatrix}$ and $\begin{pmatrix} 1 & \alpha \\ 1 & 1 \end{pmatrix}^*\begin{pmatrix} 1 & \alpha \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \alpha & 1 \end{pmatrix}\begin{pmatrix} 1 & \alpha \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1+\alpha \\ 1+\alpha & 1+\alpha^2 \end{pmatrix}$.

For these to be equal, $1 + \alpha^2 = 2$, or $\alpha^2 = 1$, so $\alpha = 1$ or $\alpha = -1$.

11.8 Isometric operators, orthogonal matrices and unitary matrices

A linear operator $\mathcal{A}: \mathcal{U} \longrightarrow \mathcal{U}$ is said to be *isometric* ("same measure") if

$$\|A\mathbf{u}\|_2 = \|\mathbf{u}\|_2$$

for all vectors $\mathbf{u} \in \mathcal{U}$. We can make some interesting observations about such matrices. First, clearly, *A* must be non-singular, for if *A* were singular, then $A\mathbf{u} = \mathbf{0}$ for some vector $\mathbf{u} \neq \mathbf{0}$, but then $\|A\mathbf{u}\|_2 = \|\mathbf{0}\|_2 = 0 \neq \|\mathbf{u}\|_2$; however, there is a much more fundamental result:

Theorem

A is isometric if and only if $AA^* = A^*A = Id$; that is, $A^{-1} = A^*$.

Proof:

If *A* is isometric, then $\|A\mathbf{u}\|_2 = \|\mathbf{u}\|_2$ and so $\|A\mathbf{u}\|_2^2 = \|\mathbf{u}\|_2^2$; however,

$$\|A\mathbf{u}\|_2^2 = \langle A\mathbf{u}, A\mathbf{u}\rangle = \langle A^*A\mathbf{u}, \mathbf{u}\rangle = \langle \mathbf{u}, \mathbf{u}\rangle = \|\mathbf{u}\|_2^2$$

for all vectors **u**. Thus, $A^*A = Id$. Similarly, if $A^*A = Id$, we may follow the same procedure to find that $\|A\mathbf{u}\|_2 = \|\mathbf{u}\|_2$. ∎

In finite dimensions, there are some additional interesting results we may gain from isometric matrices.

Theorem

If $A: \mathbf{F}^n \to \mathbf{F}^n$ is an isometric matrix, the columns of *A* and the rows of *A* form orthogonal bases of the vector space.

Proof:

If $A = \begin{pmatrix} | & | & | & | \\ \mathbf{a}_{:,1} & \mathbf{a}_{:,2} & \cdots & \mathbf{a}_{:,n} \\ | & | & | & | \end{pmatrix}$ then $A^* = \begin{pmatrix} — & \mathbf{a}_{:,1}^* & — \\ — & \mathbf{a}_{:,2}^* & — \\ & \vdots & \\ — & \mathbf{a}_{:,n}^* & — \end{pmatrix}$, where the kth column of *A* is the kth row of A^*. As $A^*A = Id$, it follows that

$$\langle \mathbf{a}_{:,i}, \mathbf{a}_{:,j} \rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases},$$

the very definition that a set of vectors is orthogonal. As there are *n* columns, this set of *n* vectors must form an orthonormal basis for $\mathbf{R}^n$. ∎

Not only are the columns orthogonal, but so are the rows:

Theorem

If $A:\mathbf{F}^n \to \mathbf{F}^n$ is an isometric matrix, the rows of *A* are also orthonormal.

Proof:

If we consider the rows of $A = \begin{pmatrix} — & \mathbf{a}_{1,\cdots} & — \\ — & \mathbf{a}_{2,\cdots} & — \\ & \vdots & \\ — & \mathbf{a}_{m,\cdots} & — \end{pmatrix}$, then $A^* = \begin{pmatrix} | & | & | & | \\ \mathbf{a}^*_{1,\cdots} & \mathbf{a}^*_{2,\cdots} & \cdots & \mathbf{a}^*_{m,\cdots} \\ | & | & | & | \end{pmatrix}$, where the kth row of *A* is the kth column of A^*. As $A^*A = Id$, it also follows that $A^*A = AA^* = Id$, and therefore

$$\left\langle \mathbf{a}_{i,\cdots}, \mathbf{a}_{j,\cdots} \right\rangle^* = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}.$$

Thus, the rows of *A* form an orthonormal basis. ∎

Definition

If $A:\mathbf{R}^n \to \mathbf{R}^n$ is an isometric matrix, we say that *A* is an *orthogonal* matrix and its inverse is its transpose, or $A^{-1} = A^{\mathrm{T}}$.

Definition

If $A:\mathbf{C}^n \to \mathbf{C}^n$ is an isometric matrix, we say that *A* is a *unitary* matrix and its inverse is its conjugate transpose, or $A^{-1} = A^*$.

An example of an orthogonal matrix $A:\mathbf{R}^4 \to \mathbf{R}^4$ is

$$A = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 \end{pmatrix}.$$

You will note that the rows form a different orthonormal basis than do the columns:

$$\begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}, \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}, \begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix} \text{ versus } \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}, \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \\ -\frac{1}{2} \end{pmatrix}.$$

An example of a matrix $B:\mathbf{C}^4 \to \mathbf{C}^4$ that is unitary is

$$A = \tfrac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & j & -1 & -j \\ 1 & -1 & 1 & 1 \\ 1 & -j & -1 & j \end{pmatrix} = \tfrac{1}{2}\begin{pmatrix} j^0 & j^0 & j^0 & j^0 \\ j^0 & j^1 & j^2 & j^3 \\ j^0 & j^2 & j^4 & j^6 \\ j^0 & j^3 & j^6 & j^9 \end{pmatrix}.$$

This matrix forms one of many cores to the fast Fourier transform—the single most important algorithm in digital signal processing. In this case, you will note that the matrix is symmetric, but not conjugate symmetric. Another example of a useful orthogonal matrix is that used in JPEG encoding.

In summary, note that in a sense, isometric operators are similar to the complex numbers on unit circle $|z| = 1$, for the complex numbers *z* on the unit circle are the only numbers are the only complex numbers such that $|zw| = |w|$ for all complex numbers *w*, and the multiplicative inverse of *z* is z^*, so $zz^* = z^*z = 1$.

11.9 Least squares

In a previous chapter, we found that the scalar multiple of a fixed vector **v** that best approximates a given vector **v** is the projection of **v** onto **u**:

$$\text{proj}_{\mathbf{u}}(\mathbf{v}) = \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u} \ .$$

In the special case that $\hat{\mathbf{u}}$ is normalized, this simplifies to

$$\text{proj}_{\hat{\mathbf{u}}}(\mathbf{v}) = \langle \hat{\mathbf{u}}, \mathbf{v} \rangle \hat{\mathbf{u}} \ .$$

Thus, the projection is the scalar multiple of **u** or $\hat{\mathbf{u}}$ that best approximates **v**.

Suppose, however, you have a collection of ℓ vectors $\mathbf{u}_1, \ldots, \mathbf{u}_\ell$ and you would like to find which linear combination of these vectors best approximates a target vector **v**; that is, which $\alpha_1, \ldots, \alpha_\ell$ allows

$$\alpha_1 \mathbf{u}_1 + \cdots + \alpha_\ell \mathbf{u}_\ell$$

to best approximate **v**. If $\mathbf{v} \in \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_\ell\}$, then at least one solution exists, and a unique solution will exist if and only if the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_\ell$ are linearly independent. If the vectors are orthogonal, it was already argued that the best approximation is the vector **v** projected onto each of the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_\ell$,

$$\text{proj}_{\mathbf{u}_1, \ldots, \mathbf{u}_\ell}(\mathbf{v}) = \sum_{k=1}^{\ell} \text{proj}_{\mathbf{u}_k}(\mathbf{v}) = \sum_{k=1}^{\ell} \frac{\langle \mathbf{u}_k, \mathbf{v} \rangle}{\langle \mathbf{u}_k, \mathbf{u}_k \rangle} \mathbf{u}_k \ ,$$

and if, in addition, the vectors $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_\ell$ are orthonormal, this simplifies to

$$\text{proj}_{\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_\ell}(\mathbf{v}) = \sum_{k=1}^{\ell} \langle \hat{\mathbf{u}}_k, \mathbf{v} \rangle \hat{\mathbf{u}}_k \ .$$

This does not tell us, however, how to find the best approximation for a collection of vectors that are not mutually orthogonal.

Before we solve this problem, it is also useful to view this problem in terms of a system of linear equations. For example, here we have a system of five equations but only two unknowns:

$$\begin{aligned}
u_{1,1}\alpha_1 + u_{1,2}\alpha_2 &= v_1 \\
u_{2,1}\alpha_1 + u_{2,2}\alpha_2 &= v_2 \\
u_{3,1}\alpha_1 + u_{3,2}\alpha_2 &= v_3 \\
u_{4,1}\alpha_1 + u_{4,2}\alpha_2 &= v_4 \\
u_{5,1}\alpha_1 + u_{5,2}\alpha_2 &= v_5
\end{aligned}$$

The probability that a solution $\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$ exists is exceptionally small, for in almost all likelihood, for if we create the matrix $A = (\mathbf{a}_1\ \mathbf{a}_2)$ the augmented matrix $(A \mid \mathbf{v})$ (where in this case, $A: \mathbf{R}^2 \to \mathbf{R}^5$) will be row equivalent to a row-echelon form matrix of the form

$$\left(\begin{array}{cc|c} a_{1,1} & a_{1,2} & v_1 \\ a_{2,1} & a_{2,2} & v_2 \\ a_{3,1} & a_{3,2} & v_3 \\ a_{4,1} & a_{4,2} & v_4 \\ a_{5,1} & a_{5,2} & v_5 \end{array}\right) \sim \left(\begin{array}{cc|c} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right),$$

in which case, no solution exists. The next question we may ask, therefore, is:

Can we find a vector $\boldsymbol{\alpha}$ that makes $A\boldsymbol{\alpha}$ as close as possible to $\mathbf{v}$?

How do we define "close"? We could use the 2-norm:

Can we find a vector $\boldsymbol{\alpha}$ that minimizes $\|\mathbf{v} - A\boldsymbol{\alpha}\|_2$?

Before we carry on, let us consider an example:

$$\begin{aligned} 2\alpha_1 + \alpha_2 &= 2 \\ \alpha_1 - \alpha_2 &= 0 \\ \alpha_2 &= 1 \\ \alpha_1 + \alpha_2 &= 1 \end{aligned}$$

Clearly, no solution exists, for if $\alpha_2 = 1$ (by Equation 3), then $\alpha_1 - 1 = 0$ (from Equation 2) and $\alpha_1 + 1 = 1$ (from Equation 3) cannot both be simultaneously true. However, if we let $\boldsymbol{\alpha} = \begin{pmatrix} 0.6 \\ 0.7 \end{pmatrix}$, then

$$\begin{pmatrix} 2 & 1 \\ 1 & -1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0.6 \\ 0.7 \end{pmatrix} = \begin{pmatrix} 1.9 \\ -0.1 \\ 0.7 \\ 1.3 \end{pmatrix}$$

You will note that none of the entries of the matrix-vector product equals what we were hoping for, but in this case, the two are reasonably close:

$$\begin{pmatrix} 1.9 \\ -0.1 \\ 0.7 \\ 1.3 \end{pmatrix} \approx \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

The 2-norm between these two is $\sqrt{0.1^2 + 0.1^2 + 0.3^2 + 0.3^2} = \sqrt{0.2} \approx 0.447$. If you were to try any other vector $\boldsymbol{\alpha}$, the distance between $A\boldsymbol{\alpha}$ and $\mathbf{v}$ would be greater than the distance between $A\begin{pmatrix} 0.6 \\ 0.7 \end{pmatrix}$ and $\mathbf{v}$. The question is, how can we find the vector $\boldsymbol{\alpha}$ that minimizes $\|\mathbf{v} - A\boldsymbol{\alpha}\|_2$?

11.9.1 Finding the *least-squares* solution to $A\mathbf{u} - \mathbf{v}$

Recall previously that the projection $\text{proj}_{\mathbf{u}}(\mathbf{v})$ was found by finding that vector $c\mathbf{u}$ that is as close as possible to the vector $\mathbf{v}$. The difference, $\mathbf{v} - \text{proj}_{\mathbf{u}}(\mathbf{v})$ was found to be perpendicular to $\mathbf{u}$.

Thus, given $A: \mathbf{R}^n \to \mathbf{R}^m$, our goal here is to find the vector $\boldsymbol{\alpha} \in \mathbf{R}^n$ such that $A\boldsymbol{\alpha}$ is as close as possible to $\mathbf{v} \in \mathbf{R}^m$, and therefore, the difference $\mathbf{v} - A\boldsymbol{\alpha}$ should be perpendicular to all vectors $A\boldsymbol{\alpha}$; in other words, $\mathbf{v} - A\boldsymbol{\alpha}$ needs to be perpendicular to all vectors in the range of U. Thus, we need the following theorem:

Theorem

A vector is perpendicular to the range of a linear operator A if and only if it is in the null space of A^*.

Proof:

Suppose $A: \mathcal{U} \to \mathcal{V}$ and a vector $\mathbf{v}$ is perpendicular to every vector in the range of A. In this case,

$$\langle A\mathbf{u}, \mathbf{v} \rangle = 0$$

for every vector $\mathbf{u} \in \mathcal{U}$. This is true if and only if

$$\langle \mathbf{u}, A^*\mathbf{v} \rangle = 0$$

for all vectors $\mathbf{u} \in \mathcal{U}$. This is only true if $\mathbf{v}$ is in the null space of A^*, and thus $\mathbf{v} \in \text{null}(A^*)$. ∎

Therefore, if $\mathbf{v} - A\boldsymbol{\alpha}$ is perpendicular to all vectors in the range of $A = (\mathbf{a}_1 \cdots \mathbf{a}_n)$, then $A^*(\mathbf{v} - A\boldsymbol{\alpha}) = \mathbf{0}_n$. Therefore,

$$A^*\mathbf{v} - A^*A\boldsymbol{\alpha} = \mathbf{0}_n$$

or

$$A^*A\boldsymbol{\alpha} = A^*\mathbf{v}.$$

Because the solution to this equation is the vector $\boldsymbol{\alpha}$ that minimizes the 2-norm $\|\mathbf{v} - A\boldsymbol{\alpha}\|_2$, we describe $\boldsymbol{\alpha}$ as the *least-squares* solution.

Going back to our example,

$$A = \begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & -1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

Therefore,

$$A^*A = \begin{pmatrix} 2 & 1 & 0 & 1 \\ 1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & -1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 2 \\ 2 & 4 \end{pmatrix} \text{ and } A^*\mathbf{v} = \begin{pmatrix} 2 & 1 & 0 & 1 \\ 1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 4 \end{pmatrix}.$$

Solving $A^*A\boldsymbol{\alpha} = A^*\mathbf{v}$ is therefore found through

$$\left(\begin{array}{cc|c} 6 & 2 & 5 \\ 2 & 4 & 4 \end{array}\right) \sim \left(\begin{array}{cc|c} 6 & 2 & 5 \\ 0 & \frac{10}{3} & \frac{7}{3} \end{array}\right),$$

and therefore, by backward substitution, we get that $\alpha_2 = 0.7$, and substituting this back into the first equation yields that $6\alpha_1 + 2 \cdot 0.7 = 5$ or $\alpha_1 = 0.6$. Thus, the vector that makes $U\boldsymbol{\alpha}$ as close as possible to our target vector is

$$\boldsymbol{\alpha} = \begin{pmatrix} 0.6 \\ 0.7 \end{pmatrix}.$$

Notice that

$$\mathbf{v} - A\boldsymbol{\alpha} = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1.9 \\ -0.1 \\ 0.7 \\ 1.3 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.3 \\ -0.3 \end{pmatrix},$$

and we note that

$$A^*\left(A\boldsymbol{\alpha} - \mathbf{v}\right) = \begin{pmatrix} 2 & 1 & 0 & 1 \\ 1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0.1 \\ 0.1 \\ 0.3 \\ -0.3 \end{pmatrix} = \begin{pmatrix} 0.2 + 0.1 - 0.3 \\ 0.1 - 0.1 + 0.3 - 0.3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

so indeed, the difference is in the null space of A^*. Also, if we explicitly find a basis for the null space of A^*, we have that

$$\left(\begin{array}{cccc|c} 2 & 1 & 0 & 1 & 0 \\ 1 & -1 & 1 & 1 & 0 \end{array}\right) \sim \left(\begin{array}{cccc|c} 2 & 1 & 0 & 1 & 0 \\ 0 & -1.5 & 1 & 0.5 & 0 \end{array}\right),$$

so α_4 and α_3 are free variables, and thus $-1.5\alpha_2 + \alpha_3 + 0.5\alpha_4 = 0$, so $\alpha_2 = \frac{2}{3}\alpha_3 + \frac{1}{3}\alpha_4$, and $2\alpha_1 + \left(\frac{2}{3}\alpha_3 + \frac{1}{3}\alpha_4\right) + \alpha_4 = 0$, and therefore $\alpha_1 = -\frac{1}{3}\alpha_3 - \frac{2}{3}\alpha_4$. Thus, the null space is the span

$$\begin{pmatrix} -\frac{1}{3}\alpha_3 - \frac{2}{3}\alpha_4 \\ \frac{2}{3}\alpha_3 + \frac{1}{3}\alpha_4 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \alpha_3 \begin{pmatrix} -\frac{1}{3} \\ \frac{2}{3} \\ 1 \\ 0 \end{pmatrix} + \alpha_4 \begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \\ 0 \\ 1 \end{pmatrix},$$

and we note that

$$0.3\begin{pmatrix} -\frac{1}{3} \\ \frac{2}{3} \\ 1 \\ 0 \end{pmatrix} - 0.3\begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.1 \\ 0.2 \\ 0.3 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.2 \\ -0.1 \\ 0 \\ -0.3 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.3 \\ -0.3 \end{pmatrix}.$$

Problems

1. Find the least-squares approximation to

$$\begin{aligned} x + y &= 1 \\ -y &= 0 \\ x - y &= 0 \\ -2x - y &= 3 \end{aligned}$$

2. Find the least-squares approximation to

$$\begin{aligned} 2x + y &= 1 \\ -x - y &= 0 \\ y &= 2 \\ -x + y &= 3 \end{aligned}$$

Solutions

1. $x = -0.8$, $y = -0.1$

11.9.2 Linear regression for orthogonal and orthonormal vectors

Suppose that the vectors $\mathbf{u}_1,\ldots,\mathbf{u}_\ell$ we are provided are orthogonal. In this case, the matrix $U=(\mathbf{u}_1\cdots\mathbf{u}_\ell)$ has columns that are mutually orthogonal, and so

$$U^*U=\begin{pmatrix}\mathbf{u}_1^*\\ \mathbf{u}_2^*\\ \mathbf{u}_2^*\\ \vdots\\ \mathbf{u}_\ell^*\end{pmatrix}\left(\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3\cdots\mathbf{u}_\ell\right)=\begin{pmatrix}\langle\mathbf{u}_1,\mathbf{u}_1\rangle & 0 & 0 & \cdots & 0\\ 0 & \langle\mathbf{u}_2,\mathbf{u}_2\rangle & 0 & \cdots & 0\\ 0 & 0 & \langle\mathbf{u}_3,\mathbf{u}_3\rangle & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & \langle\mathbf{u}_\ell,\mathbf{u}_\ell\rangle\end{pmatrix},$$

which is a diagonal matrix, and

$$U^*\mathbf{v}=\begin{pmatrix}\mathbf{u}_1^*\\ \mathbf{u}_2^*\\ \mathbf{u}_2^*\\ \vdots\\ \mathbf{u}_\ell^*\end{pmatrix}\mathbf{v}=\begin{pmatrix}\langle\mathbf{u}_1,\mathbf{v}\rangle\\ \langle\mathbf{u}_2,\mathbf{v}\rangle\\ \langle\mathbf{u}_3,\mathbf{v}\rangle\\ \vdots\\ \langle\mathbf{u}_\ell,\mathbf{v}\rangle\end{pmatrix},$$

so solving $U^*U=U^*\mathbf{v}$ reduces to solving a diagonal system, and therefore

$$\boldsymbol{\alpha}=\begin{pmatrix}\dfrac{\langle\mathbf{u}_1,\mathbf{v}\rangle}{\langle\mathbf{u}_1,\mathbf{u}_1\rangle}\\ \dfrac{\langle\mathbf{u}_2,\mathbf{v}\rangle}{\langle\mathbf{u}_2,\mathbf{u}_2\rangle}\\ \dfrac{\langle\mathbf{u}_3,\mathbf{v}\rangle}{\langle\mathbf{u}_3,\mathbf{u}_3\rangle}\\ \vdots\\ \dfrac{\langle\mathbf{u}_\ell,\mathbf{v}\rangle}{\langle\mathbf{u}_\ell,\mathbf{u}_\ell\rangle}\end{pmatrix},$$

so the linear combination of the vectors $\mathbf{u}_1,\ldots,\mathbf{u}_\ell$ that best approximates $\mathbf{v}$ is

$$\frac{\langle\mathbf{u}_1,\mathbf{v}\rangle}{\langle\mathbf{u}_1,\mathbf{u}_1\rangle}\mathbf{u}_1+\frac{\langle\mathbf{u}_2,\mathbf{v}\rangle}{\langle\mathbf{u}_2,\mathbf{u}_2\rangle}\mathbf{u}_2+\frac{\langle\mathbf{u}_3,\mathbf{v}\rangle}{\langle\mathbf{u}_3,\mathbf{u}_3\rangle}\mathbf{u}_3+\cdots+\frac{\langle\mathbf{u}_\ell,\mathbf{v}\rangle}{\langle\mathbf{u}_\ell,\mathbf{u}_\ell\rangle}\mathbf{u}_\ell.$$

If, however, the vectors $\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_\ell$ form an orthonormal set,

$$U^*U=Id,$$

so

$$U^*U\boldsymbol{\alpha}=I\boldsymbol{\alpha}=\boldsymbol{\alpha}=U^*\mathbf{v},$$

and thus

$$\boldsymbol{\alpha} = U^* \mathbf{v} = \begin{pmatrix} \hat{\mathbf{u}}_1^* \\ \hat{\mathbf{u}}_2^* \\ \hat{\mathbf{u}}_2^* \\ \vdots \\ \hat{\mathbf{u}}_\ell^* \end{pmatrix} \mathbf{v} = \begin{pmatrix} \langle \hat{\mathbf{u}}_1, \mathbf{v} \rangle \\ \langle \hat{\mathbf{u}}_2, \mathbf{v} \rangle \\ \langle \hat{\mathbf{u}}_3, \mathbf{v} \rangle \\ \vdots \\ \langle \hat{\mathbf{u}}_\ell, \mathbf{v} \rangle \end{pmatrix},$$

so the linear combination of the vectors $\hat{\mathbf{u}}_1,\ldots,\hat{\mathbf{u}}_\ell$ that best approximates **v** is

$$\langle \hat{\mathbf{u}}_1, \mathbf{v} \rangle \hat{\mathbf{u}}_1 + \langle \hat{\mathbf{u}}_2, \mathbf{v} \rangle \hat{\mathbf{u}}_2 + \langle \hat{\mathbf{u}}_3, \mathbf{v} \rangle \hat{\mathbf{u}}_3 + \cdots + \langle \hat{\mathbf{u}}_\ell, \mathbf{v} \rangle \hat{\mathbf{u}}_\ell ,$$

which is the formula described previously, though only with a hand-waving argument.

11.9.3 Existence of a least-squares solution

Now, given a non-homogeneous system of linear equations, there are normally three possibilities:

1. there are no solutions,
2. there is a unique solution, or
3. there are infinitely many solutions.

However, given that we are attempting to find a "closest" solution, we should be able to find at least one solution.

Theorem

The is always at least one solution **u** to the system of linear equations defined by $A^* A\mathbf{u} = A^* \mathbf{y}$.

Proof:

By definition, $A^* \mathbf{v} \in \text{range}\left(A^*\right)$. Previously, we saw that if $\text{rank}\left(A^*\right) = \text{rank}\left(A^* A\right)$, and thus, $\text{range}\left(A^*\right) = \text{range}\left(A^* A\right)$, and therefore there must be a pre-image **u** such that $A^* A\mathbf{u} = A^* \mathbf{y}$. ∎

Theorem

If $A : \mathbf{R}^n \to \mathbf{R}^m$ with *m* > *n*, if the columns of *A* are linearly independent, then there is a unique solution to the system of linear equations defined by $A^* A\mathbf{u} = A^* \mathbf{y}$.

Proof:

If the colums of *A* are linearly independent, then rank(*A*) = *n*, and thus rank(A^*A) = *n* where A^*A is *n* x *n*. Thus, A^*A is invertable and thus there is a unique solution to $A^* A\mathbf{u} = A^* \mathbf{y}$. ∎

If the columns are not linearly independent, the solution may not, however, be unique. For example, if we try to find a the closest linear combination of the vectors in

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \text{ to the target } \mathbf{v} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

we have

$$\left(A^*A \;\vdots\; A^*\mathbf{v}\right) = \left(\begin{array}{ccc:c} 4 & -2 & 2 & 2 \\ -2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 2 \end{array}\right) \sim \left(\begin{array}{ccc:c} 4 & -2 & 2 & 2 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right),$$

and thus there are infinitely many solutions:

$$\mathbf{u} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix},$$

each of which yields the least-squares approximation

$$A\mathbf{u} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}.$$

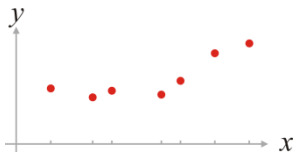
A unique solution exists if and only if rank(A) = n, in which case, rank(A^*A) = n.

11.9.4 Applications of linear regression

When we are given n points of the form $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i \neq x_k$ for $i \neq k$, we saw that we can always find an interpolating polynomial of degree less than or equal to $n-1$ that passes through each of the points. For example, we can find the affine polynomial $\alpha x + \beta$ that passes through two points (2, 5) and (4, –3) by solving the system of equations

$$\begin{aligned} 2\alpha + \beta &= \;\; 5 \\ 4\alpha + \beta &= -3 \end{aligned}$$

which yields the solution $\alpha = -4$ and $\beta = 13$. Suppose, however, that there are equations than unknowns? For example, suppose we have



Clearly, there is no linear polynomial that passes through all n points; that is, there is no coefficient vector $\mathbf{c} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ such that

$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ x_4 & 1 \\ x_5 & 1 \\ x_6 & 1 \\ x_7 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix}.$$

That is, we cannot solve $V\mathbf{c} = \mathbf{y}$ for the coefficient vector $\mathbf{c}$, and thus, regardless of what we choose, the two vectors $V\mathbf{c}$ and $\mathbf{y}$ must be different. If we could find a solution, then $V\mathbf{c} - \mathbf{y} = \mathbf{0}$, and so $\|V\mathbf{c} - \mathbf{y}\|_2 = 0$. If we cannot find a solution, perhaps the best goal would be to try to find that coefficient vector $\mathbf{c}$ that minimizes $\|V\mathbf{c} - \mathbf{y}\|_2$? We will call any polynomial that minimizes this 2-norm a *best-fitting least-squares* polynomial.

For example, if we tried to find an interpolating linear polynomial that passes through the points (2, 5), (4, –3) and (5, –1), we would have the system defined by

$$\begin{pmatrix} 2 & 1 \\ 4 & 1 \\ 5 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} -5 \\ 3 \\ 1 \end{pmatrix}.$$

As $\left(\begin{array}{cc|c} 2 & 1 & -5 \\ 4 & 1 & 3 \\ 5 & 1 & 1 \end{array}\right) \sim \left(\begin{array}{cc|c} 5 & 1 & 1 \\ 0 & 0.6 & -5.4 \\ 0 & 0 & 4 \end{array}\right)$, the system is over determined. Instead, we will define

$$\mathbf{r} = \begin{pmatrix} 2 & 1 \\ 4 & 1 \\ 5 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - \begin{pmatrix} -5 \\ 3 \\ 1 \end{pmatrix} \text{ or } \mathbf{r} = V\mathbf{c} - \mathbf{y}$$

and to find that **c** that minimizes this vector. Because Vandermonde matrix refers strictly to square matrices, we will call the above matrix *V* a *diminished* Vandermonde matrix.

The vector $\mathbf{r}$ contains the errors or *residuals* of the approximation. If it is our goal to minimize the 2-norm these residuals, this is equivalent to minimizing

$$\|\mathbf{r}\|_2^2 = \langle \mathbf{r}, \mathbf{r} \rangle = \underbrace{\langle V\mathbf{c} - \mathbf{y}, V\mathbf{c} - \mathbf{y} \rangle = \langle V\mathbf{c}, V\mathbf{c} \rangle - \langle V\mathbf{c}, \mathbf{y} \rangle - \langle \mathbf{b}, V\mathbf{c} \rangle}_{\text{dependent on } \mathbf{c}} + \underbrace{\langle \mathbf{y}, \mathbf{y} \rangle}_{\text{independent of } \mathbf{c}}.$$

We cannot change the value of $\langle \mathbf{y}, \mathbf{y} \rangle$, and thus we need only minimize

$$\langle V\mathbf{c}, V\mathbf{c} \rangle - \langle \mathbf{y}, V\mathbf{c} \rangle - \langle V\mathbf{c}, \mathbf{y} \rangle = \langle V\mathbf{c}, V\mathbf{c} \rangle - \langle V^*\mathbf{y}, \mathbf{c} \rangle - \langle \mathbf{c}, V^*\mathbf{y} \rangle.$$

In order to minimize a function *f*(*x*) in one variable, we differentiate, equate to zero and solve for *x*. If we have a function of many variables, we must differentiate with respect to each variable separately and equate each of the equations to zero. In this case, we have that

In order to derive the solution, this requires gradients, a topic of vector calculus, which is beyond the scope of this course. Thus, we will simply make a claim, and then demonstrate, that the 2-norm of the residual error is minimized when $\mathbf{c}$ is the solution to the *normal equation* $V^*V\mathbf{c} = V^*\mathbf{y}$. A solution to this equation only exists if V^*V is invertible, and this is true if and only if the columns of V are linearly independent, and for a, this means that there must be at least *n* unique *x*-values.

We will now continue with our theorem.

Theorem

The 2-norm of the residual error is minimized when $\mathbf{c}$ is a solution to $V^*V\mathbf{c} = V^*\mathbf{y}$ and this solution is unique when the columns of *A* are linearly independent.

Proof:

We would like to show that $\|V(\mathbf{c}+\mathbf{e})-\mathbf{y}\|_2^2 > \|V\mathbf{c}-\mathbf{y}\|_2^2$ whenever $\mathbf{e} \neq \mathbf{0}$ where $\mathbf{c}$ is the solution to $V^*V\mathbf{c} = V^*\mathbf{y}$. Again, using the properties of the inner product,

$$\begin{aligned}\|V(\mathbf{c}+\mathbf{e})-\mathbf{y}\|_2^2 &= \langle V(\mathbf{c}+\mathbf{e})-\mathbf{y}, V(\mathbf{c}+\mathbf{e})-\mathbf{y}\rangle \\ &= \langle V\mathbf{c}+V\mathbf{e}-\mathbf{y}, V\mathbf{c}+V\mathbf{e}-\mathbf{y}\rangle \\ &= \langle V\mathbf{c},V\mathbf{c}\rangle + \langle V\mathbf{c},V\mathbf{e}\rangle - \langle V\mathbf{c},\mathbf{y}\rangle + \langle V\mathbf{e},V\mathbf{c}\rangle + \langle V\mathbf{e},V\mathbf{e}\rangle - \langle V\mathbf{e},\mathbf{y}\rangle - \langle \mathbf{y},V\mathbf{c}\rangle - \langle \mathbf{y},V\mathbf{e}\rangle + \langle \mathbf{y},\mathbf{y}\rangle\end{aligned}$$

Next, let us group those terms that depend on**e** and those that are independent:

$$= \underbrace{\langle V\mathbf{c},V\mathbf{c}\rangle - \langle V\mathbf{c},\mathbf{y}\rangle - \langle \mathbf{y},V\mathbf{c}\rangle + \langle \mathbf{y},\mathbf{y}\rangle}_{\text{Independent of } \mathbf{e}} + \underbrace{\langle V\mathbf{c},A\mathbf{e}\rangle + \langle V\mathbf{e},A\mathbf{c}\rangle + \langle V\mathbf{e},V\mathbf{e}\rangle - \langle V\mathbf{e},\mathbf{y}\rangle - \langle \mathbf{y},V\mathbf{e}\rangle}_{\text{Dependent on } \mathbf{e}}$$

We only need to show that the second expression,

$$\langle V\mathbf{c},V\mathbf{e}\rangle + \langle V\mathbf{e},V\mathbf{c}\rangle + \langle V\mathbf{e},V\mathbf{e}\rangle - \langle V\mathbf{e},\mathbf{y}\rangle - \langle \mathbf{y},V\mathbf{e}\rangle,$$

is minimized when **e** = **0**. If **c** is the solution to $V^*V\mathbf{c} = V^*\mathbf{y}$, we may substituting this into the equation, we get

$$\begin{aligned}&\langle V\mathbf{c},M\mathbf{e}\rangle + \langle V\mathbf{e},M\mathbf{c}\rangle + \langle V\mathbf{e},M\mathbf{e}\rangle - \langle V\mathbf{e},\mathbf{y}\rangle - \langle \mathbf{y},V\mathbf{e}\rangle \\ &\quad = \langle V\mathbf{c},V\mathbf{e}\rangle + \langle V\mathbf{e},V\mathbf{c}\rangle + \langle V\mathbf{e},V\mathbf{e}\rangle - \langle \mathbf{e},V^*\mathbf{y}\rangle - \langle V^*\mathbf{y},\mathbf{e}\rangle \\ &\quad = \langle V\mathbf{c},V\mathbf{e}\rangle + \langle V\mathbf{e},V\mathbf{c}\rangle + \langle V\mathbf{e},V\mathbf{e}\rangle - \langle \mathbf{e},V^*V\mathbf{c}\rangle - \langle V^*V\mathbf{c},\mathbf{e}\rangle \\ &\quad = \cancel{\langle V\mathbf{c},V\mathbf{e}\rangle} + \cancel{\langle V\mathbf{e},V\mathbf{c}\rangle} + \langle V\mathbf{e},V\mathbf{e}\rangle - \cancel{\langle V\mathbf{e},V\mathbf{c}\rangle} - \cancel{\langle V\mathbf{c},V\mathbf{e}\rangle} \\ &\quad = \langle V\mathbf{e},V\mathbf{e}\rangle\end{aligned}$$

As the columns are linearly independent, $V\mathbf{e} = \mathbf{0}$ if and only if $\mathbf{e} = \mathbf{0}$, and therefore $\langle V\mathbf{e},V\mathbf{e}\rangle = \|V\mathbf{e}\|_2^2 > 0$ whenever $\mathbf{e} \neq \mathbf{0}$. Therefore $\|V(\mathbf{c}+\mathbf{e})-\mathbf{y}\|_2^2 > \|V\mathbf{c}-\mathbf{y}\|_2^2$ whenever $\mathbf{e} \neq \mathbf{0}$, and therefore the solution $\mathbf{c}$ to $V^*V\mathbf{c} = V^*\mathbf{y}$ minimizes the 2-norm of the residual error.[30] ∎

[30] This proof is based on a proof presented by Sheehan Olver of the School of Mathematics and Statistics at the University of Sydney in a set of lecture notes for a course in numerical complex analysis.

Theorem
The columns of a diminished $m \times n$ Vandermonde matrix where $m > n$ are linearly independent if and only if there are at least *n* unique *x* values.

Proof:
If there are greater than or equal to *n* unique *x*-values, then the rank of the $m \times n$ Vandermonde matrix equals *n*, and thus the columns are linearly independent. If there are fewer than *n* unique *x*-values, the rank equals the number of unique *x*-values, and therefore we have a greater number of columns than the rank, and therefore the columns are linearly dependent. ∎

Now that we have shown that we are looking for a solution to $V^*V\mathbf{c} = V^*\mathbf{y}$, we will consider four approaches:

1. the naïve approach,
2. Cholesky factorization, and
3. QR factorization.

For each of these, we will find the least-squares best-fitting linear and quadratic polynomials that pass through the points

$$(3, -5), (4, -7), (4, -6), (7, -8), (9, -7).$$

11.10 The naïve approach

The naïve approach is to simply calculate V^*V and $V^*\mathbf{y}$, and then solve $V^*V\mathbf{c} = V^*\mathbf{y}$. Find the least-squares solution linear polynomial, using MATLAB, we find that

```
>> x = [ 3  4  4  7  9]';
>> y = [-5 -7 -6 -8 -7]';
>> V1 =      [x.^1 x.^0];
>> V2 = [x.^2 x.^1 x.^0];
>> c1 = (V1'*V1) \ (V1'*y)
    c1 =
      -0.3095
      -4.9286

>> c2 = (V2'*V2) \ (V2'*y)
    c1 =
       0.2151
      -2.9012
       1.7093
```

Therefore, the best-fitting linear polynomial is $-0.3095x - 4.9286$ and the best-fitting least-squares quadratic polynomial is $0.2151x^2 - 2.9012x + 1.7093$.

The run time of the first is O(*mnm*) = O(m^2n), the second is O(m^2), and this produce a system of *n* linear equations in *n* unknowns, the solving of which requires O(n^3) time.

For a situation where *x*-values remain unchanged, but the *y*-values vary, we may calculate V^*V once, in which case, finding the least squares solution now requires only O($mn + n^3$) time.

11.11 Cholesky factorization

The matrix V^*V is positive definite, and therefore can be written as $A^*A = LL^*$ where L is lower triangular. Consequently, this will only become beneficial if we are in a situation where the *x*-values remain unchanged but the *y*-values may vary over time. In this case, $V^*V\mathbf{c} = V^*\mathbf{y}$ is reduced to solving $L^*L\mathbf{c} = V^*\mathbf{y}$, requiring the computation of $V^*\mathbf{y}$ requiring O(*mn*) time and the application of both backward and then forward substitution, both requiring O(n^2) time. Consequently, as *m*> *n*, the run time is O(*mn*).

11.12 QR factorization

The next matrix factorization we will look at is QR factorization (also known as QR decomposition). An $m \times n$ matrix *A* with $m \geq n$ may be written as the product of an $m \times n$ unitary matrix Q and an $n \times n$ upper triangular matrix *R*. If the matrix *M* is real, the matrix Q will be orthogonal. For example,

$$\begin{pmatrix} 1 & 2 \\ 1 & 7 \\ 3 & 9 \\ 5 & 6 \end{pmatrix} = \begin{pmatrix} \frac{1}{6} & \frac{1}{42} \\ \frac{1}{6} & \frac{31}{42} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{5}{6} & -\frac{19}{42} \end{pmatrix} \begin{pmatrix} 6 & 11 \\ 0 & 7 \end{pmatrix}$$

In general, of course, the result will not be so clean, but this example is useful to demonstrate the result. There are two techniques for finding the QR decomposition of a matrix. The first, which is intuitively easier, is also numerically less stable, but it demonstrates the transformation.

11.12.1 QR decomposition with Gram-Schmidt

The steps are straight-forward. Apply the Gram-Schmidt process to the columns of the matrix *M*, producing the matrix Q. In the above example, you will see that

$$M_{:,1} = \begin{pmatrix} 1 \\ 1 \\ 3 \\ 5 \end{pmatrix} \text{ and } \left\| M_{:,1} \right\|_2 = 6 \text{ so } Q_{:,1} = \begin{pmatrix} \frac{1}{6} \\ \frac{1}{6} \\ \frac{1}{2} \\ \frac{5}{6} \end{pmatrix}$$

$$M_{:,2} - \text{proj}_{Q_{:,1}} M_{:,2} = \begin{pmatrix} 2 \\ 7 \\ 9 \\ 6 \end{pmatrix} - \left\langle Q_{:,1}, M_{:,2} \right\rangle Q_{:,1} = \begin{pmatrix} \frac{1}{6} \\ \frac{31}{6} \\ \frac{7}{2} \\ -\frac{19}{6} \end{pmatrix} \text{ which has a 2-norm of 7, so } Q_{:,2} = \begin{pmatrix} \frac{1}{42} \\ \frac{31}{42} \\ \frac{1}{2} \\ -\frac{19}{42} \end{pmatrix}.$$

Next, we find the linear combination of the column vectors of Q that equal each of the columns of *M*:

$$M_{:,k} = \left\langle Q_{:,1}, M_{:,k} \right\rangle Q_{:,1} + \left\langle Q_{:,2}, M_{:,k} \right\rangle Q_{:,2} + \cdots + \left\langle Q_{:,n}, M_{:,k} \right\rangle Q_{:,n}$$

If we define the columns of R to be the inner products,

$$R = \begin{pmatrix} \langle Q_{:,1}, M_{:,1} \rangle & \langle Q_{:,1}, M_{:,2} \rangle & \cdots & \langle Q_{:,1}, M_{:,1} \rangle \\ \langle Q_{:,2}, M_{:,1} \rangle & \langle Q_{:,2}, M_{:,2} \rangle & \cdots & \langle Q_{:,2}, M_{:,2} \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle Q_{:,n}, M_{:,1} \rangle & \langle Q_{:,2}, M_{:,2} \rangle & \cdots & \langle Q_{:,n}, M_{:,n} \rangle \end{pmatrix},$$

we then note that this matrix is upper triangular, as $M_{:,i}$ must be perpendicular to all $Q_{:,k}$ for $k > i$. Thus, R is of the desired form, an upper triangular matrix:

$$R = \begin{pmatrix} \langle Q_{:,1}, M_{:,1} \rangle & \langle Q_{:,1}, M_{:,2} \rangle & \cdots & \langle Q_{:,1}, M_{:,1} \rangle \\ 0 & \langle Q_{:,2}, M_{:,2} \rangle & \cdots & \langle Q_{:,2}, M_{:,2} \rangle \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \langle Q_{:,n}, M_{:,n} \rangle \end{pmatrix}.$$

The process of performing the Gram Schmidt process is O(mn^2).

11.12.2 The QR factorization with Householder reflections

Another approach to a *QR* factorization is to use a *Householder reflection*. We have seen that the elementary matrix operation S^w is a reflection, and a generalization of a reflection is to take any unit vector $\mathbf{u}$ and define

$$H_{\mathbf{u}} = Id - 2\mathbf{u}\mathbf{u}^* .$$

This matrix is conjugate symmetric and unitary:

$$\begin{aligned} H_{\mathbf{u}}^* &= \left(Id - 2\mathbf{u}\mathbf{u}^*\right)^* \\ &= Id^* - 2\left(\mathbf{u}\mathbf{u}^*\right)^* \\ &= Id - 2\left(\mathbf{u}^*\right)^*\mathbf{u}^* \\ &= Id - 2\mathbf{u}\mathbf{u}^* \end{aligned} \quad \text{and} \quad \begin{aligned} H_{\mathbf{u}}^* H_{\mathbf{u}} &= \left(Id_n - 2\mathbf{u}\mathbf{u}^*\right)\left(Id_n - 2\mathbf{u}\mathbf{u}^*\right) \\ &= Id_n^2 - 4\mathbf{u}\mathbf{u}^* + 4\left(\mathbf{u}\mathbf{u}^*\right)\left(\mathbf{u}\mathbf{u}^*\right) \\ &= Id_n - 4\mathbf{u}\mathbf{u}^* + 4\mathbf{u}\left(\mathbf{u}^*\mathbf{u}\right)\mathbf{u}^* \\ &= Id_n - 4\mathbf{u}\mathbf{u}^* + 4\mathbf{u}\mathbf{u}^* = O_n \end{aligned}$$

The interpretation of $H_{\mathbf{u}}\mathbf{x}$ is a reflection of $\mathbf{x}$ through the hyperplane perpendicular to $\mathbf{u}$. Given any unit vector $\mathbf{u}$, if you reflect $\mathbf{x}$ through the vector $\dfrac{\mathbf{x} + \|\mathbf{x}\|_2 \mathbf{u}}{\left\|\mathbf{x} + \|\mathbf{x}\|_2 \mathbf{u}\right\|_2}$, this is a unit vector lying in the plane of $\mathbf{u}$ and $\mathbf{x}$ with equal angles, as shown in .

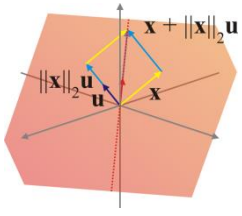


Figure 85. The vector u in black, x in yellow, the normalized sum of the scaled $\|\mathbf{x}\|_2 \mathbf{u}$ and x in red.

This results in a Householder reflection that reflects $\mathbf{x}$ onto a multiple of $\mathbf{u}$.

$$
\begin{aligned}
H_{\frac{\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}}{\|\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\|_2}}\mathbf{x} &= \left(Id_n - 2\frac{\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}}{\|\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\|_2}\left(\frac{\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}}{\|\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\|_2}\right)^*\right)\mathbf{x} \\
&= \mathbf{x} - 2\frac{\left(\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right)\left(\mathbf{x}^*+\|\mathbf{x}\|_2\mathbf{u}^*\right)}{\left\|\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right\|_2^2}\mathbf{x} \\
&= \mathbf{x} - 2\frac{\mathbf{x}\mathbf{x}^*\mathbf{x}+\|\mathbf{x}\|_2\mathbf{x}\mathbf{u}^*\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\mathbf{x}^*\mathbf{x}+\|\mathbf{x}\|_2^2\mathbf{u}\mathbf{u}^*\mathbf{x}}{\left\|\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right\|_2^2} \\
&= \mathbf{x} - 2\frac{\mathbf{x}\mathbf{x}^*\mathbf{x}+\|\mathbf{x}\|_2\mathbf{x}\mathbf{u}^*\mathbf{x}+\|\mathbf{x}\|_2^3\mathbf{u}+\|\mathbf{x}\|_2^2\mathbf{u}\mathbf{u}^*\mathbf{x}}{\left\|\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right\|_2^2} \\
&= \mathbf{x} - \frac{\mathbf{x}^*\mathbf{x}\mathbf{x}+\|\mathbf{x}\|_2\mathbf{x}^*\mathbf{u}\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}^*\mathbf{x}\mathbf{x}+\|\mathbf{x}\|_2^2\mathbf{u}^*\mathbf{u}\mathbf{x}-\|\mathbf{x}\|_2\left(\mathbf{x}^*\mathbf{x}\mathbf{u}-\|\mathbf{x}\|_2\mathbf{x}^*\mathbf{u}\mathbf{u}-\|\mathbf{x}\|_2\mathbf{u}^*\mathbf{x}\mathbf{u}-\|\mathbf{x}\|_2^2\mathbf{u}^*\mathbf{u}\mathbf{u}\right)}{\left\|\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right\|_2^2} \\
&= \mathbf{x} - \frac{\left(\mathbf{x}^*+\|\mathbf{x}\|_2\mathbf{u}^*\right)\left(\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right)\left(\mathbf{x}-\|\mathbf{x}\|_2\mathbf{u}\right)}{\left\|\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right\|_2^2} \\
&= \mathbf{x} - \frac{\left(\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right)^*\left(\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right)\left(\mathbf{x}-\|\mathbf{x}\|_2\mathbf{u}\right)}{\left\|\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right\|_2^2} \\
&\mathbf{x} - \frac{\left\|\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right\|_2^2\left(\mathbf{x}-\|\mathbf{x}\|_2\mathbf{u}\right)}{\left\|\mathbf{x}+\|\mathbf{x}\|_2\mathbf{u}\right\|_2^2} \\
&= \mathbf{x} - \mathbf{x} + \|\mathbf{x}\|_2\mathbf{u}
\end{aligned}
$$

Now, we can specifically choose $\hat{\mathbf{e}}_1$, in which case, $H_{\frac{\mathbf{x}+\|\mathbf{x}\|_2\hat{\mathbf{e}}_1}{\|\mathbf{x}+\|\mathbf{x}\|_2\hat{\mathbf{e}}_1\|_2}}\mathbf{x} = \begin{pmatrix} \|\mathbf{x}\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

11.12.3 Solving the least-squares problem with the QR factorization

Recall that we must solve $M^*M\mathbf{c} = M^*\mathbf{b}$. We now substitute

$$
\begin{aligned}
(QR)^*QR\mathbf{c} &= (QR)^*\mathbf{b} \\
R^*Q^*QR\mathbf{c} &= R^*Q^*\mathbf{b} \\
R^*R\mathbf{c} &= R^*Q^*\mathbf{b} \\
R\mathbf{c} &= Q^*\mathbf{b}
\end{aligned}
$$

Thus, if we have both *Q* and *R*, the operations are to:

1. calculate $Q^*\mathbf{b}$, which is O(m^2), and
2. solve $R\mathbf{c} = Q^*\mathbf{b}$, and because *R* is upper triangular, we need only use backward substitution, which is O(m^2).

Thus, the overall run time is O(m^2). Thus, given that the process of generating *Q* and *R* is O(m^3), the overall run time to solve a single problem is no better than solving $M^*M\mathbf{c} = M^*\mathbf{b}$; however,

11.13 Numerical error

To understand why different algorithms are necessary to achieve the same end, consider the following example: suppose we wish to find the least-squares solution of the form $\alpha x^2 + \beta x$ passing through the points

$$(0, 0), (10^{-8}, 10^{-8}), (1, 2).$$

This may occur if we are certain that the solution must pass through the origin—if there is zero voltage, there must be zero current. The correct solution is to define

$$V = \begin{pmatrix} 0 & 0 \\ 10^{-16} & 10^{-8} \\ 1 & 1 \end{pmatrix} \text{ and } \mathbf{y} = \begin{pmatrix} 0 \\ 10^{-8} \\ 2 \end{pmatrix}.$$

The correct answer to this problem is

$$\mathbf{c} = \begin{pmatrix} \frac{100000000}{99999999} \\ \frac{99999998}{99999999} \end{pmatrix} = \begin{pmatrix} 1.\overline{00000001} \\ 0.\overline{99999998} \end{pmatrix} \approx \begin{pmatrix} 1.00000001 \\ 0.99999999 \end{pmatrix}.$$

If we try the naïve approach in Matlab, however, we run into problems

```
>> x = [0 1e-8 1]';
>> y = [0 1e-8 2]';
>> V = [x.^2 x.^1];
>> V'*V
    ans =
         1     1
         1     1

>> V'*y
    ans =
         2
         2
```

This matrix is singular and the rank of the augmented matrix equals the rank of the matrix, which therefore suggests that any solution where $\alpha + \beta = 2$ is acceptable. Clearly this is false.

Next, we could try using QR decomposition with the Gram-Schmidt process:

```
>> q1 = V(:,1)/norm( V(:,1) );
>> q2 = V(:,2) - (q1'*V(:,2))*q1;
>> q2 = q2/norm( q2 );
>> Q = [q1 q2];
    Q =
                         0                    0
    0.000000000000000   1.000000000000000
    1.000000000000000                    0

>> R = [Q(:,1)'*V(:,1) Q(:,1)'*V(:,2); 0 Q(:,2)'*V(:,2)]
    R =
    1.000000000000000   1.000000000000000
                         0   0.000000010000000

>> R \ (Q'*y)
```

```
ans =
     1
     1
```

This suggests the best solution is the polynomial $x^2 + x$, which is reasonably close to the exact answer.

You may note, however, that there are two types of zero entries in the matrix Q: `0` and `0.000000000000000`. The first is an actual floating-point zero, but the second is not:

```
>> q1(2)
    ans =
         1.000000000000000e-16
```

Finally, we will look at using the QR factorization using the Householder transformations, the technique implemented in Matlab:

```
>> [Q R] = qr( V, 0 )
    Q =
                         0  -0.000000000000000
      -0.000000000000000   1.000000000000000
      -1.000000000000000  -0.000000000000000

    R =
      -1.000000000000000  -1.000000000000000
                         0   0.000000010000000
>> R\(Q'*y)
    ans =
       1.000000010000000
       0.999999990000000
```

You will see that the matrix Q has a second entry that is close to, but not exactly zero:

```
>> Q(1,2)
    ans =
        -1.000000000000000e-16

>> Q(2,1)
    ans =
        -1.000000000000000e-16
```

and consequently, this gives the best possible solution in MATLAB: the best fitting polynomial is

$$1.00000001\ x^2 + 0.99999999\ x.$$

11.14 Operator *-algebras

If we have a complex vector space U and consider all linear operators on U, the adjoint A^* for $A \in L(U,U)$ defined as defined above has three properties of interest. First, $\left(A^*\right)^* = A$, as

$$\begin{aligned}\left\langle \left(A^*\right)^* \mathbf{u}, \mathbf{v} \right\rangle &= \left\langle \mathbf{u}, A^* \mathbf{v} \right\rangle \\ &= \left\langle A\mathbf{u}, \mathbf{v} \right\rangle\end{aligned}$$

for all $\mathbf{u}, \mathbf{v} \in U$. Second, $(A+B)^* = A^* + B^*$, as

$$\begin{aligned}\left\langle (A+B)^* \mathbf{u}, \mathbf{v} \right\rangle &= \left\langle \mathbf{u}, A+B\mathbf{v} \right\rangle \\ &= \left\langle \mathbf{u}, A\mathbf{v} \right\rangle + \left\langle \mathbf{u}, B\mathbf{v} \right\rangle \\ &= \left\langle A^*\mathbf{u}, \mathbf{v} \right\rangle + \left\langle B^*\mathbf{u}, \mathbf{v} \right\rangle \\ &= \left\langle A^* + B^*\mathbf{u}, \mathbf{v} \right\rangle\end{aligned}$$

and finally, $(\alpha A)^* = \alpha^* A^*$ as

$$\begin{aligned}\left\langle (\alpha A)^* \mathbf{u}, \mathbf{v} \right\rangle &= \left\langle \mathbf{u}, \alpha A\mathbf{v} \right\rangle \\ &= \alpha^* \left\langle \mathbf{u}, A\mathbf{v} \right\rangle \\ &= \alpha^* \left\langle A^*\mathbf{u}, \mathbf{v} \right\rangle \\ &= \left\langle \alpha^* A^*\mathbf{u}, \mathbf{v} \right\rangle\end{aligned}$$

Thus, for a complex vector space U, the algebra of $L(U,U)$ is said to define a *-algebra (star-algebra). There are other examples of *-algebras; however, this is beyond the scope of this course.

11.15 Summary of adjoint operators

In summary,

12 Invariant subspaces, eigenvalues, eigenvectors and diagonalization

As we have seen, linear operators between vector spaces are reasonably complex objects. We have already seen that we may describe a matrix through its determinant, but this in itself is usually not of interest practically except when performing changes of coordinates in vector calculus. Operators themselves may or may not have inverses, they are associative, but not commutative, and they generally carry a lot of information: a mapping from $\mathbf{F}^n$ to $\mathbf{F}^m$ requires *mn* different values, and changing any one value, even only slightly, can make a non-invertible matrix invertible or vice versa. One of the goals of linear algebra is to find easier ways of describing the properties of a matrix.

Given a linear operator $A: \mathcal{U} \to \mathcal{U}$, a subspace $S \subseteq \mathcal{U}$ is said to be *invariant under A* if $AS \subseteq S$; that is,

$$\text{for all } \mathbf{u} \in S,\ A\mathbf{u} \in S,$$

or, written using logical implication operator:

$$\mathbf{u} \in S \to A\mathbf{u} \in S$$

Now, by definition, $\mathcal{U}$ is invariant under *A* for all linear operators, and so is $\{\mathbf{0}_{\mathcal{U}}\}$, but are there others?

Given any linear operator *A*, its null-space null(*A*) is invariant, for

$$\mathbf{u} \in \text{null}(A) \to A\mathbf{u} = \mathbf{0} \to A\mathbf{u} \in \text{null}(A).$$

Similarly, the range is invariant,

$$\mathbf{u} \in \text{range}(A) \to \mathbf{u} \in U \to A\mathbf{u} \in \text{range}(A)$$

We will now, however, define a special class of invariant subspaces.

12.1 1-dimensional invariant subspaces

Given a vector space $\mathcal{U}$ over a field **F** (either **R** or **C**), suppose that an invariant subspace $\mathcal{S}$ of a linear operator *A* was one-dimensional. In this case, the form of the subspace must be a line:

$$\mathcal{S} = \{\alpha \mathbf{u} : \alpha \in \mathbf{F}\}.$$

In this case, *A***u** must be a scalar multiple of **u**, and thus

$$A\mathbf{u} = \lambda \mathbf{u}$$

for some scalar λ. Let's look at some examples:

1. Given the matrix $\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$, we immediately note that $\mathcal{S}_1 = \left\{\alpha \begin{pmatrix} 1 \\ 0 \end{pmatrix} : \alpha \in \mathbf{F}\right\}$ is an invariant subspace, for $\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} \alpha \\ 0 \end{pmatrix} = \begin{pmatrix} 2\alpha \\ 0 \end{pmatrix}$. We also note that $\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} \alpha \\ 0 \end{pmatrix} = 2\begin{pmatrix} \alpha \\ 0 \end{pmatrix}$, so each vector in this subspace is stretched by a factor of two. What may be less obvious is that there is a second invariant subspace $\mathcal{S}_2 = \left\{\alpha \begin{pmatrix} 1 \\ 1 \end{pmatrix} : \alpha \in \mathbf{F}\right\}$, for $\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} \alpha \\ \alpha \end{pmatrix} = \begin{pmatrix} 3\alpha \\ 3\alpha \end{pmatrix}$, in which case, each vector in this subspace is stretched by a factor of three.
2. Let **u** be a vector in the null space of $A: \mathcal{U} \to \mathcal{U}$. Then all multiples of **u** form a 1-dimensional invariant subspace of $\mathcal{U}$, as $A\mathbf{u} = \mathbf{0}_{\mathcal{U}}$.
3. Next, let us consider the differential operator again, but now on all **F**[*t*], the space of all polynomials in *t*. Here we know that if deg(*p*) > 0, it follows that $\deg(p') = \deg(p) - 1$, and therefore, it is impossible that $\frac{d}{dt}p(t) = \lambda p(t)$. However, if deg(*q*) = 0, then *q* must be a constant polynomial, and therefore $\frac{d}{dt}q(t) = 0$. Therefore, for any constant polynomial, $\frac{d}{dt}q(t) = 0 \cdot q(t)$, and therefore the subspace of constant polynomials is an invariant 1-dimensional subspace of **F**[*t*] under differentiation.
4. If we consider the vector space of differentiable functions, then we note that $U_\lambda = \{\alpha e^{\lambda t} : \alpha \in \mathbf{C}\}$, that is, all scalar multiples of the exponentially growing $e^{\lambda t}$, then $\frac{d}{dt}\left(\alpha e^{\lambda t}\right) = (\alpha\lambda)e^{\lambda t}$, and therefore the differential operator stretches (or shrinks) the function $\alpha e^{\lambda t}$ by a factor of λ.

Definition

If $\mathcal{S} = \text{span}\{\mathbf{u}_1\}$ is a one-dimensional invariant subspace of an operator *A* where $A\mathbf{u} = \lambda\mathbf{u}$ for all $\mathbf{u} \in \mathcal{S}$, we will say that

1. λ is an *eigenvalue* of *A*,
2. any vector in $\mathcal{S}$ is an *eigenvector* of *A* corresponding to the eigenvalue λ, and
3. $\mathcal{S}$ is the *eigenspace* of *A* corresponding to the eigenvalue λ.

We only have to pick one vector in $\mathcal{S}$ to represent all vectors in $\mathcal{S}$, and it is very common to choose a normalized vector, usually using the 2-norm, but sometimes using the ∞-norm.

As an example, in the above example with the matrix $A = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$:

1. $\lambda_1 = 2$ is one eigenvalue of *A* and $\mathbf{u}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is a corresponding eigenvector, and
2. $\lambda_2 = 3$ is a second eigenvalue of *A* and $\mathbf{u}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is a corresponding eigenvector.

This matrix has two eigenvalues.

It is usual, but not necessary, to pick an eigenvector that is easily identifiable. For example, any of $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ (the previous vector 2-normalized) and $\begin{pmatrix} -1 \\ -1 \end{pmatrix}$ are eigenvectors corresponding to $\lambda_1 = 2$, but the *nicest* (visually speaking) is the first. Most programs that compute the eigenvectors will return one of the 2-normalized versions. You will never lose a mark in this class if you don't use the eigenvector we expect, so long as it is a scalar multiple thereof.

Note that with the differential operator, it depends on the space you are looking at:

1. On one hand, the differential operator on **F**[*t*] has only one eigenvalue $\lambda = 0$ with the constant polynomial $p(t) = 1$ being a corresponding eigenvector.
2. On the other, for all differentiable functions, every complex number $\lambda \in \mathbf{C}$ is an eigenvalue, and an eigenvector corresponding to each is $e^{\lambda t}$.

From this point on, we will consider only finite-dimensional vector spaces.

Looking ahead, in 2nd-year, you will learn about the *Laplace transform*. In this space, an inner product is defined as $\left\langle f, e^{-s\bullet} \right\rangle = \int_{0^-}^{\infty} f(t) e^{-st} dt$ for each $s \in \mathbf{C}$. Because the exponential functions are the *eigenvectors* (or *eigenfunctions*) of the differential operator, many operations related to differential equations will be significantly easier to manipulate.

12.1.1 Not all finite-dimensional vector spaces have eigenvalues

Do all linear operators have invariant subspaces other than these? If we have a real vector space, the answer is "no". For example, let us consider the rotation matrix in $\mathbf{R}^2$:

$$R = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

where θ is not a multiple of π. Specifically, you may think about a $\theta = 90^o$ rotation: $R = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Because every vector in $\mathbf{R}^2$ is rotated, it is impossible for a one-dimensional subspace to be mapped onto itself. Notice, however, that if we allow ourselves to use complex entries and scalars, then with a little thought, we may note that

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} j \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ j \end{pmatrix} = j\begin{pmatrix} j \\ 1 \end{pmatrix},$$

and therefore there is an invariant 1-dimensional subspace with the eigenvalue j. Similarly,

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} -j \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ -j \end{pmatrix} = -j\begin{pmatrix} -j \\ 1 \end{pmatrix},$$

and therefore there are two independent 1-dimensional invariant subspaces:

For each vector of the form $\mathbf{u} = \begin{pmatrix} j\alpha \\ \alpha \end{pmatrix}$, $R\mathbf{u} = j\mathbf{u}$, and

for each vector of the form $\mathbf{u} = \begin{pmatrix} -j\alpha \\ \alpha \end{pmatrix}$, $R\mathbf{u} = -j\mathbf{u}$.

12.1.2 Linear independence of eigenvectors corresponding to different eigenvalues

In our previous example, we saw that $A_1 = \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}$ has two eigenvalues each with a corresponding eigenvector, while $A_1 = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$ has only a single eigenvalue with one corresponding eigenvector. We will now assume that a linear operator A has m eigenvalues, and show that the collection of m corresponding eigenvectors is linearly independent.

Theorem

Given a linear operator $A: \mathcal{V} \to \mathcal{V}$ with m eigenvalues with $\lambda_1, ..., \lambda_m$, each with a corresponding eigenvector $\mathbf{u}_1, ..., \mathbf{u}_m$, respectively, then the vectors $\mathbf{u}_1, ..., \mathbf{u}_m$ are linearly independent.

Proof:

Assume the opposite: that the collection of vectors $\mathbf{u}_1, ..., \mathbf{u}_m$ is linearly dependent. In this case, there is a smallest value $2 \le k \le m$ such that $\mathbf{u}_k = \alpha_1\mathbf{u}_1 + \cdots + \alpha_{k-1}\mathbf{u}_{k-1}$ and where $\mathbf{u}_1, ..., \mathbf{u}_{k-1}$ are linearly independent. If we now multiply both sides by A, we have

$$\begin{aligned} A\mathbf{u}_k &= A\left(\alpha_1\mathbf{u}_1 + \cdots + \alpha_{k-1}\mathbf{u}_{k-1}\right) \\ &= A\left(\alpha_1\mathbf{u}_1\right) + \cdots + A\left(\alpha_{k-1}\mathbf{u}_{k-1}\right) \\ &= \alpha_1 A\mathbf{u}_1 + \cdots + \alpha_{k-1}A\mathbf{u}_{k-1} \\ &= \alpha_1\lambda_1\mathbf{u}_1 + \cdots + \alpha_{k-1}\lambda_{k-1}\mathbf{u}_{k-1} \end{aligned}$$

but, by definition, $A\mathbf{u}_k = \lambda_k\mathbf{u}_k$ and therefore $A\mathbf{u}_k = \alpha_1\lambda_k\mathbf{u}_1 + \cdots + \alpha_{k-1}\lambda_k\mathbf{u}_{k-1}$. If we equate these two, we get

$$\alpha_1\lambda_1\mathbf{u}_1 + \cdots + \alpha_{k-1}\lambda_{k-1}\mathbf{u}_{k-1} = \alpha_1\lambda_k\mathbf{u}_1 + \cdots + \alpha_{k-1}\lambda_k\mathbf{u}_{k-1}$$

and bringing all terms to the left-hand side, we have

$$\alpha_1\left(\lambda_1 - \lambda_k\right)\mathbf{u}_1 + \cdots + \alpha_{k-1}\left(\lambda_{k-1} - \lambda_k\right)\mathbf{u}_{k-1} = \mathbf{0}.$$

Because all the $\lambda_i - \lambda_k \neq 0$ for all $i = 1, \ldots, k-1$, it follows that we have a linear combination of the vectors on the left that equals the zero vector but where not all coefficients are zero. Therefore, the vectors $\mathbf{u}_1, ..., \mathbf{u}_{k-1}$ are linearly dependent, which contradicts our assumption that the vectors $\mathbf{u}_1, ..., \mathbf{u}_{k-1}$ were linearly independent and $\mathbf{u}_k$ was the first vector depending on the previous vectors. Therefore, the collection $\mathbf{u}_1, ..., \mathbf{u}_m$ could not be is linearly dependent, and therefore, $\mathbf{u}_1, ..., \mathbf{u}_m$ must be linearly independent. ∎

It follows that an *n*-dimensional linear operator *A* can have no more than *n* distinct eigenvalues.

Theorem
The eigenvalues of a self-adjoint linear operator are real.

Proof:
Let $A:\mathcal{U}\to\mathcal{U}$ be a self-adjoint linear operator with an eigenvalue λ with a corresponding eigenvector **u**, so $A\mathbf{u}=\lambda\mathbf{u}$. Now, calculate the inner product of both sides with **u**, so $\langle \mathbf{u}, A\mathbf{u}\rangle = \langle \mathbf{u}, \lambda\mathbf{u}\rangle$; however, by the properties of the inner product,

$$\begin{aligned}
\lambda\langle \mathbf{u},\mathbf{u}\rangle &= \langle \mathbf{u},\lambda\mathbf{u}\rangle && \text{but } \langle \mathbf{u},\alpha\mathbf{v}\rangle = \alpha\langle \mathbf{u},\mathbf{v}\rangle \\
&= \langle \mathbf{u}, A\mathbf{u}\rangle \\
&= \langle A^*\mathbf{u},\mathbf{u}\rangle && \text{but } A = A^* \\
&= \langle A\mathbf{u},\mathbf{u}\rangle \\
&= \langle \lambda\mathbf{u},\mathbf{u}\rangle && \text{but } \langle \alpha\mathbf{u},\mathbf{v}\rangle = \alpha^*\langle \mathbf{u},\mathbf{v}\rangle \\
&= \lambda^*\langle \mathbf{u},\mathbf{u}\rangle
\end{aligned}$$

Note, however, that $\langle \mathbf{u},\mathbf{u}\rangle = \|u\|_2^2$ is real, and therefore $\lambda=\lambda^*$, and thus λ must be real. ∎

Demonstration of this theorem

The matrix $A=\begin{pmatrix}1&1&1&1\\1&2&3&4\\1&3&2&4\\1&4&4&7\end{pmatrix}$ is self-adjoint (that is, symmetric) and its eigenvalues are –1, 0, 1, and 12.

The eigenvalues of the conjugate symmetric matrix $B=\begin{pmatrix}\alpha&\beta\\\beta^*&\gamma\end{pmatrix}$ are

$$\frac{1}{2}\left(\alpha+\gamma+\sqrt{4\beta\beta^*+\alpha^2-2\alpha\gamma+\gamma^2}\right)=\frac{1}{2}\left(\alpha+\gamma+\sqrt{4|\beta|^2+\alpha^2-2\alpha\gamma+\gamma^2}\right).$$

If $\alpha=\gamma$, $\alpha^2-2\alpha\gamma+\gamma^2=0$; while if $\alpha\neq\gamma$, $\alpha^2-2\alpha\gamma+\gamma^2>0$.

The eigenvalues of the self-adjoint (that is, conjugate symmetric) complex matrix $C=\begin{pmatrix}1&j\\-j&3\end{pmatrix}$ are $2\pm\sqrt{2}$.

You will note that this theorem applies to both real and complex vector spaces. For example, the matrix

$$\begin{pmatrix}
1.9923 & 0.0828-0.3899j & 0.8425+0.5009j & 0.2885+0.8241j & 1.5416+0.0018j \\
0.0828+0.3899j & 1.5498 & 1.0772+0.7048j & 1.1325-0.0480j & 0.6343+0.3543j \\
0.8425-0.5009j & 1.0772-0.7048j & 1.6001 & 0.5770-0.3780j & 1.0556-0.2413j \\
0.2885-0.8241j & 1.1325+0.0480j & 0.5770+0.3780j & 0.2721 & 1.7223-0.2578j \\
1.5416-0.0018j & 0.6343-0.3543j & 1.0556+0.2413j & 1.7223+0.2578j & 1.2441
\end{pmatrix}$$

has eigenvalues of –1.4915, –0.0154, 0.6851, 2.2070 and 5.2732.

Theorem
Eigenvectors corresponding to different eigenvalues of a self-adjoint linear operator are orthogonal.

Proof:
Let $A:\mathcal{U}\to\mathcal{U}$ be a self-adjoint linear operator ($A = A^*$) and let λ_1 and λ_2 be two eigenvalues where $\lambda_1 \neq \lambda_2$ with corresponding eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$. Now, we can therefore calculate

$$\begin{aligned}
\lambda_2\langle \mathbf{u}_1,\mathbf{u}_2\rangle &= \langle \mathbf{u}_1,\lambda_2\mathbf{u}_2\rangle \\
&= \langle \mathbf{u}_1, A\mathbf{u}_2\rangle \\
&= \langle A^*\mathbf{u}_1,\mathbf{u}_2\rangle \quad \text{by the definition of the adjoint} \\
&= \langle A\mathbf{u}_1,\mathbf{u}_2\rangle \quad A = A^* \\
&= \langle \lambda_1\mathbf{u}_1,\mathbf{u}_2\rangle \\
&= \lambda_1^*\langle \mathbf{u}_1,\mathbf{u}_2\rangle \\
&= \lambda_1\langle \mathbf{u}_1,\mathbf{u}_2\rangle \quad \text{as } A \text{ is self adjoint, its eigenvalues are real}
\end{aligned}$$

Therefore, $\lambda_2\langle \mathbf{u}_1,\mathbf{u}_2\rangle = \lambda_1\langle \mathbf{u}_1,\mathbf{u}_2\rangle$ and as $\lambda_1 \neq \lambda_2$, it must be true that $\langle \mathbf{u}_1,\mathbf{u}_2\rangle = 0$. ∎

Demonstration of this theorem

12.1.3 The existence of at least one complex eigenvalue in $\mathbf{C}^n$

Previous, we showed that while a linear operator $\mathbf{M}:\mathbf{R}^n$ may not have any eigenvalues, we will now demonstrate that operators

> You will recall of the *fundamental theorem of algebra*: every non-constant polynomial has at least one complex root. The immediate consequence of this is that any polynomial $\alpha_n z^n + \alpha_{n-1}z^{n-1} + \cdots + \alpha_1 z + \alpha_0$ may be written as $\beta(z-\lambda_1)(z-\lambda_2)\cdots(z-\lambda_n)$ where $\lambda_1, \lambda_2, ..., \lambda_n$ are the *n* roots of the polynomial.

Suppose that $A:\mathbf{C}^n \to \mathbf{C}^n$, in which case, choose any non-zero $\mathbf{u}\in\mathbf{C}^n$. Now, generate the collection of vectors

$$\mathbf{u}, A\mathbf{u}, A^2\mathbf{u}, \ldots, A^n\mathbf{u}$$

This collection of vectors cannot be linearly independent as there are $n + 1$ consequently, there must be a linear combination of these vectors that equals the zero vector:

$$\alpha_0\mathbf{u} + \alpha_1 A\mathbf{u} + \alpha_2 A^2\mathbf{u} + \cdots + \alpha_n A^n\mathbf{u} = \mathbf{0}$$

or

$$\left(\alpha_0 + \alpha_1 A + \alpha_2 A^2 + \cdots + \alpha_n A^n\right)\mathbf{u} = \mathbf{0}.$$

Now, in the parentheses, we have a polynomial in the matrix *A*. If we write this as a polynomial, we have the equation

$$\alpha_n z^n + \alpha_{n-1}z^{n-1} + \cdots + \alpha_1 z + \alpha_0 = 0,$$

and from the fundamental theorem of algebra, this may be written as

$$\beta(z-\lambda_1)(z-\lambda_2)\cdots(z-\lambda_n) = 0$$

where λ_1 through λ_n are the n roots of the polynomial. Consequently, we may substitute this

$$\beta(A-\lambda_1 Id)(A-\lambda_2 Id)\cdots(A-\lambda_n Id)\mathbf{u} = \mathbf{0}.$$

Now, as $\beta \neq 0$ and $\mathbf{u} \neq \mathbf{0}$, thus, one of the applications of $(A-\lambda_k Id)$ must produce the zero vector from the previous product $\mathbf{v} = (A-\lambda_{k+1} Id)\cdots(A-\lambda_n Id)\mathbf{u} \neq \mathbf{0}$. Thus, $(A-\lambda_k Id)\mathbf{v} = \mathbf{0}$, so $A\mathbf{v} - \lambda_k Id\mathbf{v} = \mathbf{0}$ and thus $A\mathbf{v} = \lambda_k Id\mathbf{v} = \lambda_k \mathbf{v}$. Therefore, **v** is an eigenvector of A with corresponding to the eigenvalue λ_k. ∎

12.1.4 Eigenvalues and invertibility

We now consider a nice theorem that states that a matrix is invertible if and only if all of its eigenvalues are non-zero.

Theorem
A matrix is invertible if and only if there are no zero eigenvalues.

Proof:
Recall that a matrix is invertible if and only if the equation $A\mathbf{v} = \mathbf{0}$ has only the trivial solution $\mathbf{v} = \mathbf{0}$. Consequently, if a matrix is not invertible, then there must exist a vector $\mathbf{v} \neq \mathbf{0}$ such that $A\mathbf{v} = \mathbf{0}$, and therefore $A\mathbf{v} = 0 \cdot \mathbf{v}$ and thus 0 is an eigenvalue of A.

On the other hand, if 0 is an eigenvalue of A, then there exists a vector **v** such that $A\mathbf{v} = \mathbf{0}$, and thus the matrix is non-invertible. ∎

12.1.5 Finding eigenvalues and eigenvectors

Finding eigenvalues is a difficult problem, and one that is not can be discussed in a first-year class. We may, however, deduce the following and use it for finding eigenvalues in either $\mathbf{F}^2$ or $\mathbf{F}^3$ either 2 × 2 or 3 × 3 matrices. First we note that if

$$A\mathbf{u} = \lambda\mathbf{u}$$

then $\lambda\mathbf{u} - A\mathbf{u} = \mathbf{0}$. Now, $\mathbf{u} = I\mathbf{u}$ and therefore $\lambda\mathbf{u} = \lambda I\mathbf{u}$, so we have that

$$\lambda I\mathbf{u} - A\mathbf{u} = (\lambda I - A)\mathbf{u} = \mathbf{0}.$$

What this says is that if **u** is an eigenvector with an eigenvalue of λ, then **u** must be in the null space of $\lambda I - A$. One means of determining if the matrix $\lambda I - A$ to be non-invertible is to check if the determinant equals zero:

$$\det(\lambda I - A) = 0.$$

If A is an $n \times n$ matrix, if you calculate this determinant, you will find that it is a polynomial in λ of degree n. This polynomial must, by the fundamental theorem of algebra, have n roots if we include multiplicity. Suppose we have found that the unique eigenvalues are $\lambda_1, \ldots, \lambda_m$ where $m \leq n$. For each unique eigenvalue λ_k, we will then find a basis for the null space by finding all solutions to solving $\lambda_k I - A = \mathbf{0}$. Remember, because $\lambda_k I - A$ is not invertable, the null space must have at least dimension one.

Note, we can, if we wish, take the basis of the null space of $\lambda_k I - A$ and create an orthonormal basis by applying Gram Schmidt; however, please note:

Important: Never apply Gram-Schmidt to eigenvectors corresponding to different eigenvalues. If you do this, the resulting orthonormal vectors may no longer be eigenvectors.

Now, the determinant of a matrix can be calculated easily for 2 × 2 or 3 × 3 matrices, so we will use that result. We will now go through six examples, finding eigenvalues of six different matrices:

$$A_1 = \begin{pmatrix} 11 & -4 \\ 4 & 1 \end{pmatrix},\ A_2 = \begin{pmatrix} 2 & -2 \\ -2 & -1 \end{pmatrix},\ A_3 = \begin{pmatrix} 2 & -4 \\ -1 & 2 \end{pmatrix},\ A_4 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},\ A_5 = \begin{pmatrix} 3 & -2 \\ 0 & 4 \end{pmatrix} \text{ and } A_6 = \begin{pmatrix} 3 & -2 \\ 0 & 3 \end{pmatrix}.$$

We will also find the eigenvalues and corresponding eigenvectors of the matrix

$$B = \begin{pmatrix} 2 & -6 & 4 \\ -2 & -2 & 4 \\ -3 & 3 & -2 \end{pmatrix}.$$

We will also find the eigenvectors of

$$C = \begin{pmatrix} -0.2 & 1.7 & -1.6 & 1.1 \\ 3.12 & -1.02 & -0.44 & -0.66 \\ 0.96 & 0.34 & -2.52 & 2.22 \\ -1.68 & 1.78 & -2.84 & 3.74 \end{pmatrix}$$

Example 1: 2 × 2

For the first example, $A_1 = \begin{pmatrix} 11 & -4 \\ 4 & 1 \end{pmatrix}$, we must calculate the determinant of

$$\lambda I - A_1 = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} - \begin{pmatrix} 11 & -4 \\ 4 & 1 \end{pmatrix} = \begin{pmatrix} \lambda - 11 & 4 \\ -4 & \lambda - 1 \end{pmatrix}$$

which is

$$(\lambda - 11)(\lambda - 1) - 4(-4) = \lambda^2 - 12\lambda + 27 = (\lambda - 3)(\lambda - 9),$$

and therefore the roots of this polynomial are the eigenvalues, which are $\lambda_1 = 3$ and $\lambda_2 = 9$. The eigenvectors corresponding to the eigenvalue $\lambda_1 = 3$ are the vectors in the null space of $3I - A_1 - = \begin{pmatrix} 3-11 & 4 \\ -4 & 3-1 \end{pmatrix} = \begin{pmatrix} -8 & 4 \\ -4 & 2 \end{pmatrix}$.

Applying Gaussian elimination, we have

$$\begin{pmatrix} -8 & 4 \\ -4 & 2 \end{pmatrix} \sim \begin{pmatrix} -8 & 4 \\ 0 & 0 \end{pmatrix},$$

and therefore, α_2 is a free variable and $8\alpha_1 - 4\alpha_2 = 0$ so $\alpha_1 = \frac{1}{2}\alpha_2$, so the null space is only one dimensional and consists of the vectors $\begin{pmatrix} \frac{1}{2}\alpha_2 \\ \alpha_2 \end{pmatrix} = \alpha_2 \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$, and therefore $\mathbf{u}_1 = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$ is an eigenvector with eigenvalue $\lambda_1 = 3$.

The eigenvectors corresponding to the eigenvalue $\lambda_2 = 9$ are the vectors in the null space of $9I - A_1 = \begin{pmatrix} 9-11 & 4 \\ -4 & 9-1 \end{pmatrix} = \begin{pmatrix} -2 & 4 \\ -4 & 8 \end{pmatrix}$. Applying Gaussian elimination, we have

$$\begin{pmatrix} -2 & 4 \\ -4 & 8 \end{pmatrix} \sim \begin{pmatrix} -4 & 8 \\ -2 & 4 \end{pmatrix} \sim \begin{pmatrix} -4 & 8 \\ 0 & 0 \end{pmatrix},$$

and therefore, α_2 is a free variable and $4\alpha_1 - 8\alpha_2 = 0$ so $\alpha_1 = 2\alpha_2$, so the null space consists of the vectors $\begin{pmatrix} 2\alpha_2 \\ \alpha_2 \end{pmatrix} = \alpha_2 \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, and therefore $\mathbf{u}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ is an eigenvector with eigenvalue $\lambda_1 = 9$. Note that

$$A_1\mathbf{u}_1 = \begin{pmatrix} 11 & -4 \\ 4 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ 3 \end{pmatrix} = 3\begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix} \text{ and } A_1\mathbf{u}_2 = \begin{pmatrix} 11 & -4 \\ 4 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 18 \\ 9 \end{pmatrix} = 9\begin{pmatrix} 2 \\ 1 \end{pmatrix},$$

so these are actually pairs of eigenvalues and eigenvectors.

Example 2: 2 × 2

For the second example, $A_2 = \begin{pmatrix} 2 & -2 \\ -2 & -1 \end{pmatrix}$, we must calculate the determinant of

$$\lambda I - A_2 = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} - \begin{pmatrix} 2 & -2 \\ -2 & -1 \end{pmatrix} = \begin{pmatrix} \lambda - 2 & 2 \\ 2 & \lambda - (-1) \end{pmatrix} = \begin{pmatrix} \lambda - 2 & 2 \\ 2 & \lambda + 1 \end{pmatrix}$$

which is

$$(\lambda - 2)(\lambda + 1) - 2 \cdot 2 = \lambda^2 - \lambda - 6 = (\lambda + 2)(\lambda - 3),$$

and therefore the eigenvalues are $\lambda_1 = -2$ and $\lambda_2 = 3$. The eigenvectors corresponding to the eigenvalue $\lambda_1 = -2$ are the vectors in the null space of $(-2)I - A_2 - = \begin{pmatrix} -2-2 & 2 \\ 2 & -2+1 \end{pmatrix} = \begin{pmatrix} -4 & 2 \\ 2 & -1 \end{pmatrix}$. Applying Gaussian elimination, we have

$$\begin{pmatrix} 4 & -2 \\ -2 & 1 \end{pmatrix} \sim \begin{pmatrix} 4 & -2 \\ 0 & 0 \end{pmatrix},$$

and therefore, α_2 is a free variable and $4\alpha_1 - 2\alpha_2 = 0$ so $\alpha_1 = \frac{1}{2}\alpha_2$, so the null space is only one dimensional and consists of the vectors $\begin{pmatrix} \frac{1}{2}\alpha_2 \\ \alpha_2 \end{pmatrix} = \alpha_2 \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$, and therefore $\mathbf{u}_1 = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}$ is an eigenvector with eigenvalue $\lambda_1 = -2$.

The eigenvectors corresponding to the eigenvalue $\lambda_2 = 3$ are the vectors in the null space of $3I - A_2 = \begin{pmatrix} 3-2 & 2 \\ 2 & 3-(-1) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$. Applying Gaussian elimination, we have

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \sim \begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix} \sim \begin{pmatrix} 2 & 4 \\ 0 & 0 \end{pmatrix},$$

and therefore, α_2 is a free variable and $-2\alpha_1 - 4\alpha_2 = 0$ so $\alpha_1 = -2\alpha_2$, so the null space consists of the vectors $\begin{pmatrix} -2\alpha_2 \\ \alpha_2 \end{pmatrix} = \alpha_2 \begin{pmatrix} -2 \\ 1 \end{pmatrix}$, and therefore $\mathbf{u}_2 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$ is an eigenvector with eigenvalue $\lambda_2 = 3$. Note that

$$A_2\mathbf{u}_1 = \begin{pmatrix} 2 & -2 \\ -2 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ -2 \end{pmatrix} = -2 \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix} \text{ and } A_2\mathbf{u}_2 = \begin{pmatrix} 2 & -2 \\ -2 & -1 \end{pmatrix} \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} -6 \\ 3 \end{pmatrix} = 3 \begin{pmatrix} -2 \\ 1 \end{pmatrix},$$

so these are actually pairs of eigenvalues and eigenvectors.

Example 3: 2 × 2

For the third example, $A_3 = \begin{pmatrix} 2 & -4 \\ -1 & 2 \end{pmatrix}$, we must calculate the determinant of

$$\lambda I - A_3 = \begin{pmatrix} \lambda - 2 & 4 \\ 1 & \lambda - 2 \end{pmatrix}$$

which is

$$(\lambda - 2)(\lambda - 2) - 1 \cdot 4 = \lambda^2 - 4\lambda + 4 - 4 = \lambda(\lambda - 4),$$

and therefore the eigenvalues are $\lambda_1 = 0$ and $\lambda_2 = 4$. The eigenvectors corresponding to the eigenvalue $\lambda_1 = 0$ are the vectors in the null space of $0 \cdot I - A_3 = \begin{pmatrix} 0-2 & 4 \\ 1 & 0-2 \end{pmatrix} = \begin{pmatrix} -2 & 4 \\ 1 & -2 \end{pmatrix}$. Applying Gaussian elimination, we have

$$\begin{pmatrix} -2 & 4 \\ 1 & -2 \end{pmatrix} \sim \begin{pmatrix} -2 & 4 \\ 0 & 0 \end{pmatrix},$$

and therefore, α_2 is a free variable and $-2\alpha_1 + 4\alpha_2 = 0$ so $\alpha_1 = 2\alpha_2$, so the null space is only one dimensional and consists of the vectors $\begin{pmatrix} 2\alpha_2 \\ \alpha_2 \end{pmatrix} = \alpha_2 \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, and therefore $\mathbf{v}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ is an eigenvector with eigenvalue $\lambda_1 = 0$.

The eigenvectors corresponding to the eigenvalue $\lambda_2 = 4$ are the vectors in the null space of $4 \cdot I - A_3 = \begin{pmatrix} 4-2 & 4 \\ 1 & 4-2 \end{pmatrix} = \begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix}$. Applying Gaussian elimination, we have

$$\begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix} \sim \begin{pmatrix} 2 & 4 \\ 0 & 0 \end{pmatrix},$$

and therefore, α_2 is a free variable and $2\alpha_1 + 4\alpha_2 = 0$ so $\alpha_1 = -2\alpha_2$, so the null space consists of the vectors $\begin{pmatrix} -2\alpha_2 \\ \alpha_2 \end{pmatrix} = \alpha_2 \begin{pmatrix} -2 \\ 1 \end{pmatrix}$, and therefore $\mathbf{v}_2 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$ is an eigenvector with eigenvalue $\lambda_2 = 4$. Note that

$$A_3\mathbf{v}_1 = \begin{pmatrix} 2 & -4 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0 \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix} \text{ and } A_3\mathbf{v}_2 = \begin{pmatrix} 2 & -4 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} -8 \\ 4 \end{pmatrix} = 4\begin{pmatrix} -2 \\ 1 \end{pmatrix},$$

so these are actually pairs of eigenvalues and eigenvectors.

Example 4: 2 × 2

For the fourth example, $A_4 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, we must calculate the determinant of

$$\lambda I - A_4 = \begin{pmatrix} \lambda - 0 & 0 \\ 0 & \lambda - 0 \end{pmatrix}$$

which is

$$(\lambda - 0)(\lambda - 0) - 0 = \lambda^2,$$

and therefore the eigenvalues are both $\lambda = 0$. The eigenvectors corresponding to the eigenvalue $\lambda = 0$ are the vectors in the null space of $0 \cdot I - A_4 = \begin{pmatrix} 0-0 & 0 \\ 0 & 0-0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$. This is already in upper-triangular form, and therefore we see that both α_2 and α_1 are free variables, so the null space is two dimensional and consists of the vectors $\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \alpha_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, and therefore $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ are both eigenvectors corresponding to the eigenvalue $\lambda = 0$.

Example 5: 2 × 2

For the fifth example, $A_5 = \begin{pmatrix} 3 & -2 \\ 0 & 4 \end{pmatrix}$, we must calculate the determinant of

$$\lambda I - A_5 = \begin{pmatrix} \lambda - 3 & 2 \\ 0 & \lambda - 4 \end{pmatrix}$$

which is

$$(\lambda - 3)(\lambda - 4) - 0 = (\lambda - 3)(\lambda - 4),$$

and therefore the eigenvalues are $\lambda_1 = 3$ and $\lambda_2 = 4$. The eigenvectors corresponding to the eigenvalue $\lambda_1 = 3$ are the vectors in the null space of $3 \cdot I - A_5 = \begin{pmatrix} 0 & 2 \\ 0 & -1 \end{pmatrix}$. Applying Gaussian elimination, we have

$$\begin{pmatrix} 0 & 2 \\ 0 & -1 \end{pmatrix} \sim \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix},$$

and therefore, $\alpha_2 = 0$ and α_1 is a free variable, so the null space is only one dimensional and consists of the vectors $\begin{pmatrix} \alpha_1 \\ 0 \end{pmatrix} = \alpha_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, and therefore $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is an eigenvector with eigenvalue $\lambda_1 = 3$.

The eigenvectors corresponding to the eigenvalue $\lambda_2 = 4$ are the vectors in the null space of $4 \cdot I - A_3 = \begin{pmatrix} 4-3 & 2 \\ 0 & 4-4 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 0 & 0 \end{pmatrix}$, and this is already in row-echelon form, and therefore α_2 is a free variable and $\alpha_1 + 2\alpha_2 = 0$ so $\alpha_1 = -2\alpha_2$, so the null space consists of the vectors $\begin{pmatrix} -2\alpha_2 \\ \alpha_2 \end{pmatrix} = \alpha_2 \begin{pmatrix} -2 \\ 1 \end{pmatrix}$, and therefore $\mathbf{v}_2 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$ is an eigenvector with eigenvalue $\lambda_2 = 4$. Note that

$$A_5 \mathbf{v}_1 = \begin{pmatrix} 3 & -2 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 3 \\ 0 \end{pmatrix} = 3 \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } A_5 \mathbf{v}_2 = \begin{pmatrix} 3 & -2 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{pmatrix} -8 \\ 4 \end{pmatrix} = 4 \begin{pmatrix} -2 \\ 1 \end{pmatrix},$$

so these are actually pairs of eigenvalues and eigenvectors. The matrix is not symmetric (self-adjoint) and therefore the eigenvectors are not orthogonal.

Example 6: 3 × 3

For the example of a 3 × 3 matrix, we must calculate the determinant of

$$\lambda I - B = \begin{pmatrix} \lambda - 2 & 6 & -4 \\ 2 & \lambda - (-2) & -4 \\ 3 & -3 & \lambda - (-2) \end{pmatrix}$$

which is

$$\begin{aligned} (\lambda-2)(\lambda+2)(\lambda+2) - 72 + 24 + 12(\lambda-2) + 12(\lambda+2) - 12(\lambda+2) &= \lambda^3 + 2\lambda^2 - 16\lambda - 32 \\ &= (\lambda+4)(\lambda+2)(\lambda-4), \end{aligned}$$

and therefore the eigenvalues are $\lambda_1 = -4$, $\lambda_2 = -2$ and $\lambda_2 = 4$.

The eigenvectors corresponding to the eigenvalue $\lambda_1 = -4$ are the vectors in the null space of $(-4)\cdot I - B = \begin{pmatrix} -4-2 & 6 & -4 \\ 2 & -4-(-2) & -4 \\ 3 & -3 & -4-(-2) \end{pmatrix} = \begin{pmatrix} -6 & 6 & -4 \\ 2 & -2 & -4 \\ 3 & -3 & -2 \end{pmatrix}$. Applying Gaussian elimination, we have

$$\begin{pmatrix} -6 & 6 & -4 \\ 2 & -2 & -4 \\ 3 & -3 & -2 \end{pmatrix} \sim \begin{pmatrix} -6 & 6 & -4 \\ 0 & 0 & -\frac{16}{3} \\ 0 & 0 & -4 \end{pmatrix} \sim \begin{pmatrix} -6 & 6 & -4 \\ 0 & 0 & -\frac{16}{3} \\ 0 & 0 & 0 \end{pmatrix},$$

and therefore, $\alpha_3 = 0$, α_2 is a free variable and $-6\alpha_1 + 6\alpha_2 = 0$ so $\alpha_1 = \alpha_2$, so the null space is only one dimensional and consists of the vectors $\begin{pmatrix} \alpha_2 \\ \alpha_2 \\ 0 \end{pmatrix} = \alpha_2 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$, and therefore $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ is an eigenvector with eigenvalue $\lambda_1 = -4$.

The eigenvectors corresponding to the eigenvalue $\lambda_2 = -2$ are the vectors in the null space of $B - (-2)\cdot Id = \begin{pmatrix} -2-2 & 6 & -4 \\ 2 & -2-(-2) & -4 \\ 3 & -3 & -2-(-2) \end{pmatrix} = \begin{pmatrix} -4 & 6 & -4 \\ 2 & 0 & -4 \\ 3 & -3 & 0 \end{pmatrix}$. Applying Gaussian elimination, we have

$$\begin{pmatrix} -4 & 6 & -4 \\ 2 & 0 & -4 \\ 3 & -3 & 0 \end{pmatrix} \sim \begin{pmatrix} -4 & 6 & -4 \\ 0 & 3 & -6 \\ 0 & \frac{3}{2} & -3 \end{pmatrix} \sim \begin{pmatrix} -4 & 6 & -4 \\ 0 & 3 & -6 \\ 0 & 0 & 0 \end{pmatrix},$$

and therefore, α_3 is a free variable and $3\alpha_2 - 6\alpha_3 = 0$ so $\alpha_2 = 2\alpha_3$, and $-4\alpha_1 + 6\alpha_2 - 4\alpha_3 = -4\alpha_1 + 6(2\alpha_3) - 4\alpha_2 = 0$, so $\alpha_1 = 2\alpha_3$, so the null space consists of the vectors $\begin{pmatrix} 2\alpha_3 \\ 2\alpha_3 \\ \alpha_3 \end{pmatrix} = \alpha_3 \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}$, and therefore $\mathbf{v}_2 = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}$ is an eigenvector with eigenvalue $\lambda_2 = -2$.

Finally, the eigenvectors corresponding to the eigenvalue $\lambda_3 = 4$ are the vectors in the null space of $2 \cdot I - B = \begin{pmatrix} 4-2 & 6 & -4 \\ 2 & 4-(-2) & -4 \\ 3 & -3 & 4-(-2) \end{pmatrix} = \begin{pmatrix} 2 & 6 & -4 \\ 2 & 6 & -4 \\ 3 & -3 & 6 \end{pmatrix}$. Applying Gaussian elimination, we have

$$\begin{pmatrix} 2 & 6 & -4 \\ 2 & 6 & -4 \\ 3 & -3 & 6 \end{pmatrix} \sim \begin{pmatrix} 3 & -3 & 6 \\ 0 & 8 & -8 \\ 0 & 8 & -8 \end{pmatrix} \sim \begin{pmatrix} 3 & -3 & 6 \\ 0 & 8 & -8 \\ 0 & 0 & 0 \end{pmatrix},$$

and therefore, α_3 is a free variable and $8\alpha_2 - 8\alpha_3 = 0$ so $\alpha_2 = \alpha_3$, and $3\alpha_1 - 3\alpha_2 + 6\alpha_3 = 3\alpha_1 - 3\alpha_3 + 6\alpha_3 = 0$, so $\alpha_1 = -\alpha_3$, so the null space consists of the vectors $\begin{pmatrix} -\alpha_3 \\ \alpha_3 \\ \alpha_3 \end{pmatrix} = \alpha_3 \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$, and therefore $\mathbf{v}_3 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$ is an eigenvector with eigenvalue $\lambda_3 = 4$.

Note that

$$B\mathbf{v}_1 = \begin{pmatrix} 2 & -6 & 4 \\ -2 & -2 & 4 \\ -3 & 3 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -4 \\ -4 \\ 0 \end{pmatrix} = -4 \cdot \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

$$B\mathbf{v}_2 = \begin{pmatrix} 2 & -6 & 4 \\ -2 & -2 & 4 \\ -3 & 3 & -2 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -4 \\ -4 \\ -2 \end{pmatrix} = -2 \cdot \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}, \text{ and}$$

$$B\mathbf{v}_3 = \begin{pmatrix} 2 & -6 & 4 \\ -2 & -2 & 4 \\ -3 & 3 & -2 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -4 \\ 4 \\ 4 \end{pmatrix} = 4 \cdot \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}.$$

so these are actually pairs of eigenvalues and eigenvectors. Also, the three eigenvalues are different, so the eigenvectors are linearly independent, but the matrix is not symmetric (self-adjoint), and therefore the eigenvectors are not orthogonal.

12.1.6 The eigenvalues of a polynomial of a linear operator

Suppose we have a linear operator *A* and a polynomial in **F**[*x*]. If we calculate

Theorem

Given a finite-dimensional linear operator *A* and a polynomial *p*, the eigenvectors of $p(A)$ are the eigenvectors of *A* and the eigenvalues of $p(A)$ are the polynomial *p* evaluated at the eigenvalues of *A*.

Proof:

Suppose we have a polynomial $p(x) = \sum_{k=0}^{n} \alpha_k x^k$. In this case, if $\mathbf{u}_i$ is an eigenvector of *A*, then

$$\begin{aligned} p(A)\mathbf{u}_i &= \left(\sum_{k=0}^{n} \alpha_k A^k\right)\mathbf{u}_i \\ &= \sum_{k=0}^{n}\left(\alpha_k A^k \mathbf{u}_i\right) \\ &= \sum_{k=0}^{n}\left(\alpha_k \lambda_i{}^k \mathbf{u}_i\right) \\ &= \left(\sum_{k=0}^{n} \alpha_k \lambda_i{}^k\right)\mathbf{u}_i \\ &= p(\lambda_i)\mathbf{u}_i \end{aligned}$$

Thus, $\mathbf{u}_i$ is an eigenvector of *p*(*A*) and the corresponding eigenvalue is $p(\lambda_i)$. ∎

Demonstration of this theorem

The linear operator $A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & -2 \\ 1 & -2 & 2 \end{pmatrix}$ has three eigenvalues, $\lambda_1 = -1$, $\lambda_2 = 2$ and $\lambda_3 = 4$, with eigenvectors $\mathbf{u}_1 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$, $\mathbf{u}_2 = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$ and $\mathbf{u}_3 = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$, respectively. The three eigenvalues are non-zero, and thus the matrix is invertible. The polynomial $p : x \mapsto x - 2$ maps the three eigenvalues onto the values –3, 0 and 2, and if you calculate *p*(*A*), you get $p(A) = A - 2 \cdot I_3 = \begin{pmatrix} -1 & 1 & 1 \\ 1 & 0 & -2 \\ 1 & -2 & 0 \end{pmatrix}$, and you will note, for example, that

$$p(A)\mathbf{u}_1 = \begin{pmatrix} -1 & 1 & 1 \\ 1 & 0 & -2 \\ 1 & -2 & 0 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ -3 \\ -3 \end{pmatrix} = p(\lambda_1)\mathbf{u}_1 \text{ and } p(A)\mathbf{u}_2 = \begin{pmatrix} -1 & 1 & 1 \\ 1 & 0 & -2 \\ 1 & -2 & 0 \end{pmatrix}\begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = p(\lambda_2)\mathbf{u}_2 .$$

You may also note that the polynomial $q(x)=(x+1)(x-2)(x-4)=x^3-5x^2+2x+8$ maps each of the eigenvalues to zero. Consequently, $q(A)$ must be the zero matrix:

$$q(A)=A^3-5A^2+2A+8I_3$$
$$=\begin{pmatrix}5 & 3 & 3\\ 3 & 33 & -31\\ 3 & -31 & 33\end{pmatrix}-5\begin{pmatrix}3 & 1 & 1\\ 1 & 9 & -7\\ 1 & -7 & 9\end{pmatrix}+2\begin{pmatrix}1 & 1 & 1\\ 1 & 2 & -2\\ 1 & -2 & 2\end{pmatrix}+8\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}=\begin{pmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}.$$

Alternatively, you can calculate

$$q(A)=(A+1)(A-2)(A-4)$$
$$=\begin{pmatrix}2 & 1 & 1\\ 1 & 3 & -2\\ 1 & -2 & 3\end{pmatrix}\begin{pmatrix}-1 & 1 & 1\\ 1 & 0 & -2\\ 1 & -2 & 0\end{pmatrix}\begin{pmatrix}-3 & 1 & 1\\ 1 & -2 & -2\\ 1 & -2 & -2\end{pmatrix}$$
$$=\begin{pmatrix}0 & 0 & 0\\ 0 & 5 & -5\\ 0 & -5 & 5\end{pmatrix}\begin{pmatrix}-3 & 1 & 1\\ 1 & -2 & -2\\ 1 & -2 & -2\end{pmatrix}=\begin{pmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}$$

You will note that A is symmetric if and only if $p(A)$ is symmetric.

12.1.7 The characteristic polynomial of a finite-dimensional linear operator

Given a finite-dimensional linear operator A, suppose we define a p_A as

$$p_A(\lambda)\overset{\text{def}}{=}\det(\lambda I-A)\ .$$

12.1.8 Vector space of discrete signals

The eigenvalues and eigenvectors of the advance3 operator are the geometric sequences: Given any complex value z, define the discrete signal x_z such that $x_z[n] = z^n$. In this case, we note that if $E\{x\}[n] = x[n + 1]$, then

$$(Ex_z)[n] = z\, x_z[n].$$

As you can see $Ex_z = z\, x_z$, and therefore the exponentially growing signals are the *eigensignals* of the advance operator and that the eigenvalue corresponding to the eigensignal x_z is z.

12.2 Vector space of functions

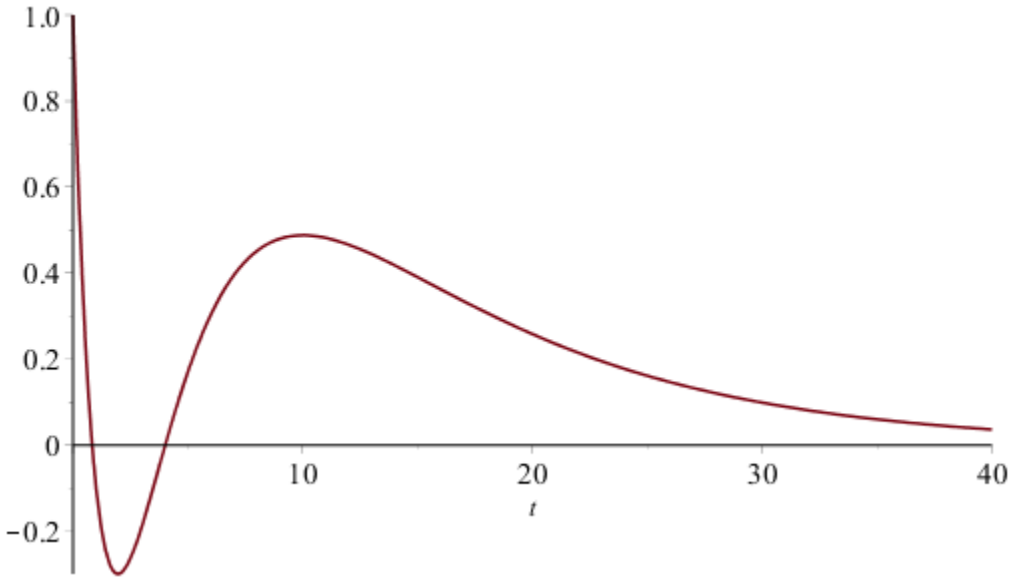
Like with the vector space of discrete signals, we can also come up with a corresponding eigenfunctions for the differential operator:

$$\frac{d}{dt}e^{\lambda t}=\lambda e^{\lambda t}\ .$$

Consequently, for every complex number λ, the exponential function $e^{\lambda t}$ is an *eigenfunction* of the differential operator, and its corresponding eigenvalue is λ. This is so trivial: indeed, it is even quite trivial to calculate higher order derivatives:

$$\frac{d^n}{dx^n}e^{\lambda x} = \lambda^n e^{\lambda x} \ .$$

If we know our solution to an initial-value problem is $y(t) = 4e^{-0.8t} - 5e^{-0.3t} + 2e^{-0.1t}$ with the graph



The derivative can therefore be trivially calculated:

$$\dot{y}(t) = -3.2e^{-0.8t} + 1.5e^{-0.3t} - 0.2e^{-0.1t} \ .$$

All we did was multiply each coefficient of each exponential by the multiplier in the exponent.

Thus, there appears to be a very special relationship between the exponential functions and the differential operator, but you will examine that phenomenon in greater detail in 2nd-year when you look at solutions to ordinary differential equations and initial-value problems. We will now, however, examine a similar phenomenon with matrices.

12.3 Diagonalization

A linear operator $A:\mathcal{U}\to\mathcal{U}$ may have the property that a collection of eigenvectors of A forms a basis for $\mathcal{U}$. What this means is that every vector in $\mathcal{U}$ may be written as a linear combination of the eigenvectors of A. If this is the case, then this is really powerful, for suppose that $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ is a basis of eigenvectors of A for the vector space $\mathcal{U}$ and that $\mathbf{v}\in\mathcal{U}$. Then $\mathbf{v}=\alpha_1\mathbf{u}_1+\cdots+\alpha_n\mathbf{u}_n$, but then it becomes trivial to calculate $A\mathbf{v}$, as

$$\begin{aligned} A\mathbf{v} &= A\left(\alpha_1\mathbf{u}_1+\cdots+\alpha_n\mathbf{u}_n\right)\\ &= \alpha_1 A\mathbf{u}_1+\cdots+\alpha_n A\mathbf{u}_n\\ &= \alpha_1\lambda_1\mathbf{u}_1+\cdots+\alpha_n\lambda_n\mathbf{u}_n \end{aligned}$$

If $A:\mathbf{F}^n\to\mathbf{F}^n$ is a matrix, no matrix multiplication was required. This also makes it relatively easy to describe the behavior of the matrix, especially if, for example, a number of the eigenvectors are zero, or if one eigenvector is large relative to the others.

Now, if a matrix has a set of vectors that forms a basis, then we can define the matrix $U=(\mathbf{u}_1\ \cdots\ \mathbf{u}_n)$ where the columns of $\mathbf{u}$ are the eigenvectors of A. To find the coefficients of $\mathbf{v}$, we must solve $U\boldsymbol{\alpha}=\mathbf{v}$. If the eigenvectors are all orthonormal, then U is a isometric matrix (either orthogonal in $\mathbf{R}^n$ or unitary in $\mathbf{C}^n$) so $\boldsymbol{\alpha}=U^*\mathbf{v}$. If the eigenvectors are not orthogonal, $\boldsymbol{\alpha}=U^{-1}\mathbf{v}$, but we are generally reduced to solving $U\boldsymbol{\alpha}=\mathbf{v}$ using a numerically stable algorithm, as calculating the inverse of U is not always advisable.

Now, once we have $A\mathbf{v}=A(U\boldsymbol{\alpha})=(AU)\boldsymbol{\alpha}$, and normally this would be foolish, as we must now calculate AU, but in this case, this is ideal, for $AU=(A\mathbf{u}_1\ \cdots\ A\mathbf{u}_n)=(\lambda_1\mathbf{u}_1\ \cdots\ \lambda_n\mathbf{u}_n)$ because, after all, the $\mathbf{u}_k$ are eigenvectors, so $A\mathbf{u}_k=\lambda_k\mathbf{u}_k$. Note, however, that this is now

$$AU=\left(\lambda_1\mathbf{u}_1\ \cdots\ \lambda_n\mathbf{u}_n\right)=U\begin{pmatrix}\lambda_1 & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \lambda_n\end{pmatrix},$$

as the diagonal matrix multiplies the first column of U (the first eigenvector) by λ_1, the second column (the second eigenvector) by λ_1, and so on. Consequently,

$$A\mathbf{v}=A(U\boldsymbol{\alpha})=(AU)\boldsymbol{\alpha}=U\begin{pmatrix}\lambda_1 & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \lambda_n\end{pmatrix}\boldsymbol{\alpha},$$

but $\boldsymbol{\alpha}=U^{-1}\mathbf{v}$, so

$$A\mathbf{v}=A(U\boldsymbol{\alpha})=(AU)\boldsymbol{\alpha}=U\begin{pmatrix}\lambda_1 & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \lambda_n\end{pmatrix}U^{-1}\mathbf{v}$$

so

$$A = U\begin{pmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n \end{pmatrix}U^{-1}.$$

We will call the middle matrix D_A, the diagonal matrix associated with the eigenvectors of A. Thus,

$$A = UD_AU^{-1}$$

and if A is real and symmetric or complex and normal,

$$A = UD_AU^{*}$$

where the adjoint is the transpose for real matrices. What this says is $A\mathbf{v}$ can be calculated by:

1. Finding the coefficients of $\mathbf{v}$ with respect to the eigenvectors of A, namely, $U^{-1}\mathbf{v}$, which results in the vector $U^{-1}\mathbf{v} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$ where $\mathbf{v} = U\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \alpha_1\mathbf{u}_1 + \cdots + \alpha_n\mathbf{u}_n$.
2. Multiplying each coefficient by the corresponding eigenvalue, namely, $D_AU^{-1}\mathbf{v}$, which results in the vector $D_AU^{-1}\mathbf{v} = \begin{pmatrix} \lambda_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_n \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \lambda_1\alpha_1 \\ \vdots \\ \lambda_n\alpha_n \end{pmatrix}$.
3. Calculating the linear combination of the eigenvectors that equals the result, namely, $UD_AU^{-1}\mathbf{v}$, which results in the vector $UD_AU^{-1}\mathbf{v} = \begin{pmatrix} \mathbf{u}_1 & \cdots & \mathbf{u}_n \end{pmatrix}\begin{pmatrix} \lambda_1\alpha_1 \\ \vdots \\ \lambda_n\alpha_n \end{pmatrix} = \lambda_1\alpha_1\mathbf{u}_1 + \cdots + \lambda_n\alpha_n\mathbf{u}_n$.

Thus, $A\mathbf{v} = \lambda_1\alpha_1\mathbf{u}_1 + \cdots + \lambda_n\alpha_n\mathbf{u}_n$.

There is an easy test to determine if a complex matrix is diagonalizable:

Theorem
In a finite-dimensional complex vector space, if a linear operator $A: \mathcal{U} \to \mathcal{U}$ is normal (that is, $A^*A = AA^*$), then any matrix representation is diagonalizable.

Unfortunately, this does not hold for real vector spaces, as a normal matrix may, never-the-less have complex eigenvalues. For example, the matrix

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

is normal, as $A^*A = AA^* = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, but the eigenvalues are $1 \pm j$ with the corresponding eigenvectors $\begin{pmatrix} \sqrt{2} \\ \pm j\sqrt{2} \end{pmatrix}$, respectively, we have that

$$\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ j\sqrt{2} & -j\sqrt{2} \end{pmatrix} \begin{pmatrix} 1+j & 0 \\ 0 & 1-j \end{pmatrix} \begin{pmatrix} \sqrt{2} & -j\sqrt{2} \\ \sqrt{2} & j\sqrt{2} \end{pmatrix}.$$

Interestingly enough, we can still make use of this in real vector spaces. We could, for example, transform a real problem into a complex problem, and then transform back. Even though the calculations were performed in

Theorem
In a finite-dimensional real vector space, if a linear operator $A: \mathcal{U} \to \mathcal{U}$ is self-adjoint (that is, symmetric or $A = A^*$), then the matrix is diagonalizable.

Sketch of a proof:
As every symmetric matrix is normal, it follows that the matrix is diagonalizable; but all the eigenvalues of a symmetric matrix are real, and therefore D_A is also real, and thus, there must be real eigenvectors, as well.

12.4 Linear operator design

We can now use diagonalization to design matrices to accomplish certain goals. For example, suppose we want to design a 4 x 4 matrix that eliminates the average value from a vector. For example, the vector $\mathbf{u} = \begin{pmatrix} 4 \\ -2 \\ -1 \\ 7 \end{pmatrix}$ has an average value of 2, and thus eliminating this would produce the vector $\mathbf{u} = \begin{pmatrix} 2 \\ -4 \\ -3 \\ 5 \end{pmatrix}$. We should observe that

$$\mathbf{u} = \begin{pmatrix} 4 \\ -2 \\ -1 \\ 7 \end{pmatrix} = \begin{pmatrix} 2 \\ -4 \\ -3 \\ 5 \end{pmatrix} + 2\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

and that the two vectors are orthogonal. Thus, what we need is a linear operator that has $\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$ as an eigenvector and the corresponding eigenvalue should be $\lambda = 0$. All other vectors should be left unchanged, recall that the following vectors fom an orthonormal basis:

$$\frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix},$$

so let us define

$$U = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix},$$

and our diagonal matrix is therefore

$$D = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Calculating $A = UDU^{\mathrm{T}}$ we get

$$A = UDU^{\mathrm{T}} = \frac{1}{4}\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}.$$

Suppose we have a vector **u** and we wish to treat its values as points $(1, u_1)$, $(2, u_1)$, …, (n, u_n), and suppose we would like to find the least-squares best approximation $(1, a + b)$, $(2, a + 2b)$, …, $(n, a + nb)$. For this, we want, for example,

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

to be left unchanged. Now, these two are not orthonormal, and therefore we should apply Gram Schmidt to get

$$\hat{\mathbf{u}}_1 = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \text{ and } \hat{\mathbf{u}}_2 = \frac{1}{2\sqrt{5}}\begin{pmatrix} -3 \\ -1 \\ 1 \\ 3 \end{pmatrix}.$$

We now need a basis for all of $\mathbf{R}^4$, so consider the other two unit vectors in the previously used basis:

$$\frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}.$$

The first is orthogonal to both vectors $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$, but the second is only orthogonal to $\hat{\mathbf{u}}_1$. To make $\hat{\mathbf{u}}_4$ orthogonal to $\hat{\mathbf{u}}_2$, once again we apply Gram Schmidt to get that

$$\hat{\mathbf{u}}_3 = \frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \text{ and } \hat{\mathbf{u}}_4 = \frac{1}{2\sqrt{5}}\begin{pmatrix} 1 \\ -3 \\ 3 \\ -1 \end{pmatrix}.$$

You will now see that we have four orthonormal vectors. Thus,

$$U = \frac{1}{2}\begin{pmatrix} 1 & -\frac{3}{\sqrt{5}} & 1 & \frac{1}{\sqrt{5}} \\ 1 & -\frac{1}{\sqrt{5}} & -1 & -\frac{3}{\sqrt{5}} \\ 1 & \frac{1}{\sqrt{5}} & -1 & \frac{3}{\sqrt{5}} \\ 1 & \frac{3}{\sqrt{5}} & 1 & -\frac{1}{\sqrt{5}} \end{pmatrix},$$

and our diagonal matrix is therefore

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

and thus

$$A = UDU^{\mathrm{T}} = \frac{1}{10}\begin{pmatrix} 7 & 4 & 1 & -2 \\ 4 & 3 & 2 & 1 \\ 1 & 2 & 3 & 4 \\ -2 & 1 & 4 & 7 \end{pmatrix}.$$

Now, if you multiply $A\mathbf{u}$ for any vector $\mathbf{u}$, you will an arithmetically increasing sequence; for example,

$$A\begin{pmatrix} 1 \\ 3 \\ 4 \\ 6 \end{pmatrix} = \begin{pmatrix} 1.1 \\ 2.7 \\ 4.3 \\ 5.9 \end{pmatrix}$$

and you will notice that the least-squares best-fitting line of the points (1, 1), (2, 3), (3, 4) and (4, 6) is $1.6x - 0.5$, and when you evaluate this at $x = 1, 2, 3$ and 4, you get the coefficients above. Thus, the linear operator A above maps a vector onto its least-squares approximation as a linear combination of $\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$.

.

12.5 Positive-definite matrices

A real matrix A is said to be *positive definite* if $\langle \mathbf{v}, A\mathbf{v} \rangle > 0$ for all vectors $\mathbf{v} \neq \mathbf{0}$. This is often written as $\mathbf{v}^{\mathrm{T}} A\mathbf{v} > 0$. Immediately, we may make some deductions:

Theorem
The eigenvalues of the matrix are all non-zero.

Proof:

Assume that A had a zero eigenvalue. In this case, $A\mathbf{v} = \mathbf{0}$. Consequently, $\langle \mathbf{v}, A\mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{0} \rangle = 0$, which contradicts our assumption that the matrix is positive definite. Therefore, $A\mathbf{v} \neq \mathbf{0}$ for all non-zero $\mathbf{v}$, and therefore A does not have a zero eigenvalue. ∎

More subtly, we may also show that:

Theorem
All the eigenvalues of a positive definite matrix are positive real numbers.

Proof:
Let λ be an eigenvalue of A. In this case, $A\mathbf{v} = \lambda\mathbf{v}$ and therefore

$$\langle \mathbf{v}, A\mathbf{v} \rangle = \langle \mathbf{v}, \lambda\mathbf{v} \rangle = \lambda \langle \mathbf{v}, \mathbf{v} \rangle = \lambda \|\mathbf{v}\|_2^2 .$$

As $\langle \mathbf{v}, A\mathbf{v} \rangle > 0$, consequently $\lambda \|\mathbf{v}\|_2^2 > 0$, and as $\|\mathbf{v}\|_2^2 > 0$, it must also be true that $\lambda > 0$. ∎

Determining if a matrix is positive definite is difficult, at best. There are, however, a few special circumstances where

A symmetric matrix is diagonally dominant if the diagonal entry is greater than the sum of the absolute values of the off-diagonal entries of the same row (or column); that is,

$$a_{i,i} > \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| a_{i,j} \right| .$$

For example, the matrix

$$\begin{pmatrix} 5.0 & 1.2 & 0.2 & 0 \\ 1.2 & 4.7 & 0.3 & -0.4 \\ 0.2 & 0.3 & 3.9 & 0.5 \\ 0 & -0.4 & 0.5 & 6.3 \end{pmatrix},$$

is diagonally dominant (and therefore positive definite) as

$$\begin{aligned} 5.0 &> 1.4 = 1.2 + 0.2 + 0.0 \\ 4.7 &> 1.9 = 1.2 + 0.3 + 0.7 \\ 3.9 &> 1.0 = 0.2 + 0.3 + 0.5 \\ 6.3 &> 0.9 = 0.0 + 0.4 + 0.5 \end{aligned}$$

Positive-definite matrices are surprisingly common in engineering, and in some cases, the matrices describing a system must be positive definite. The conductance matrix of a linear circuit consisting of resistors, inductors and capacitors is positive definite.

There are symmetric matrices that are not diagonally dominant, but are still positive definite; for example, the matrix $\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$ is diagonally dominant with eigenvalues $3 \pm 2\sqrt{2}$. Similarly, there are non-symmetric matrices that are diagonally dominant but not positive definite.

A real symmetric matrix is positive definite if and only if the *principal minors* have a positive determinant.

This condition does not apply if the matrix is not symmetric; for example, the matrix $\mathbf{M} = \begin{pmatrix} 2 & 5 \\ 1 & 3 \end{pmatrix}$ has a positive determinant and $m_{1,1} > 0$, but $\begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 2 & 5 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = -1$. Additionally, just because all the eigenvalues of a matrix are positive does not mean that the matrix itself will be positive definite. In this case, the eigenvalues of the matrix are $\frac{5 \pm \sqrt{21}}{2}$, both of which are positive.

In Matlab,

```
function [pd] = isposdef( A )
    if issymmetric( A )
        if all( 2*diag(A) > sum( abs( A ) ) )
            pd = true;
        end
    end

    pd = all( eig( A ) > 0 );
end
```

13 Singular-value decomposition (SVD)

This is not about the Soviet-era SVD sniper rifle



Figure 86. The Soviet-era SVD. Photograph by Wikipedia user Hokos.

We have seen that for a symmetric real $n \times n$ matrix A, there exists a collection of n orthogonal eigenvectors such that:

1. we can normalize these eigenvectors $\{\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_n\}$,
2. we may define a unit cube in n dimensions, the edges of which are defined by these n normalized eigenvectors, and
3. the image of the this unit cube has each edge $\hat{\mathbf{u}}_k$ stretched by λ_k.

For example, given the matrix

$$A = \begin{pmatrix} 1.2 & 0.5 & -0.7 \\ 0.5 & 0.7 & 0.2 \\ -0.7 & 0.2 & 0.3 \end{pmatrix}$$

has eigenvalues approximately equal to 1.71 , 0.76 and –0.27, and when we view a unit cube defined by three eigenvectors corresponding to these three eigenvalues, we see that the image is the original cube stretched, compressed and reflected, along the eigenvectors. You should be able to see this in Figure 87. You will also note that the ratio of the volumes of the two cubes does appear to be close to the determinant, –0.3540, and you can see that only one eigenvalue is negative, and thus there is a flip in orientation.

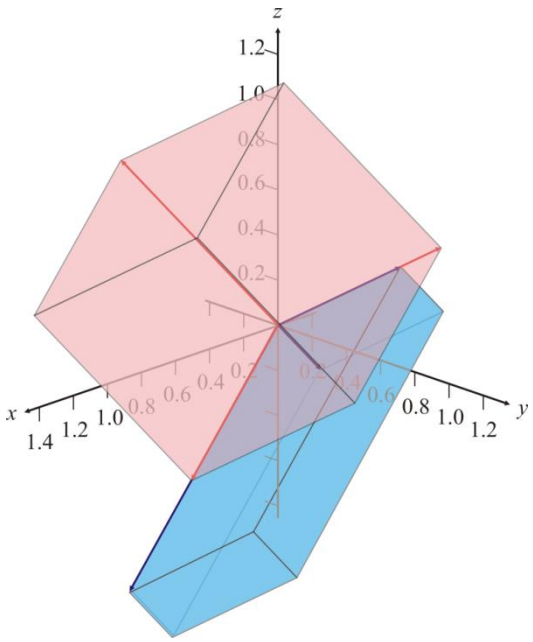


Figure 87. The cube consisting of the three eigenvectors of A and their image.

Similarly, for a normal complex $n \times n$ matrix A (that is, $AA^* = A^*A$), there exists a collection of n orthogonal eigenvectors and such the n-dimensional hypercube defined by the unit eigenvectors has an image that is an n-dimensional hyper-rectangle with each edge from the origin being a scalar multiple of the n-dimensional hypercube. Such interpretations do not exist for general finite-dimensional linear operators, as such an operator may not have a full set of n eigenvectors or the eigenvalues may only exist if the matrix is interpreted as being the mapping from $\mathbf{C}^n$ to $\mathbf{C}^n$. Additionally, even if there is a full set of n eigenvectors, they may not be orthogonal.

There is, however, a more general theorem that says:

Theorem

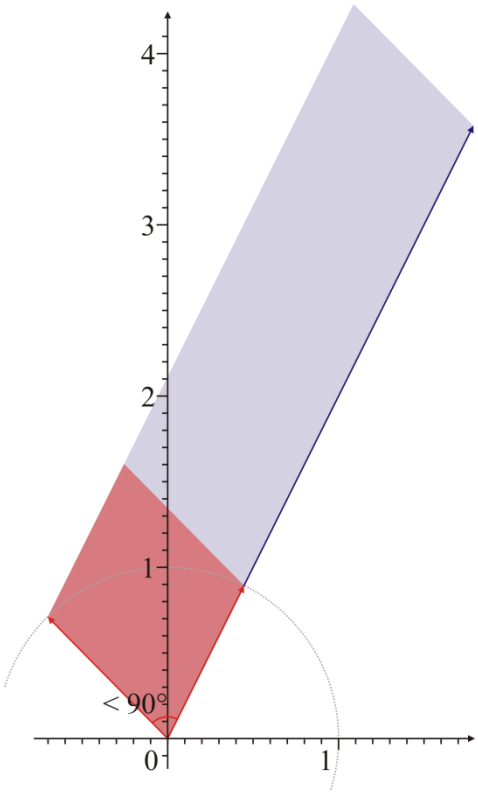
For every linear operator $A: \mathbf{F}^n \to \mathbf{F}^m$, there exists an n-dimensional unit hypercube that is mapped onto into a hyper-rectangle in $\mathbf{F}^m$.

The vectors defining the unit hypercube are called *left-singular vectors* (as they are vectors in $\mathbf{F}^n$, on the left-hand side of the mapping) and the corresponding vectors defining the m-dimensional hyper-rectangle are called *right-singular vectors*. The amounts to there being a set of n orthogonal vectors $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ in $\mathbf{F}^n$ such that $\{A\mathbf{u}_1, A\mathbf{u}_2, \ldots, A\mathbf{u}_n\}$ forms an orthogonal collection of vectors in $\mathbf{F}^m$ where some of the vectors in the image collection may be zero vectors.

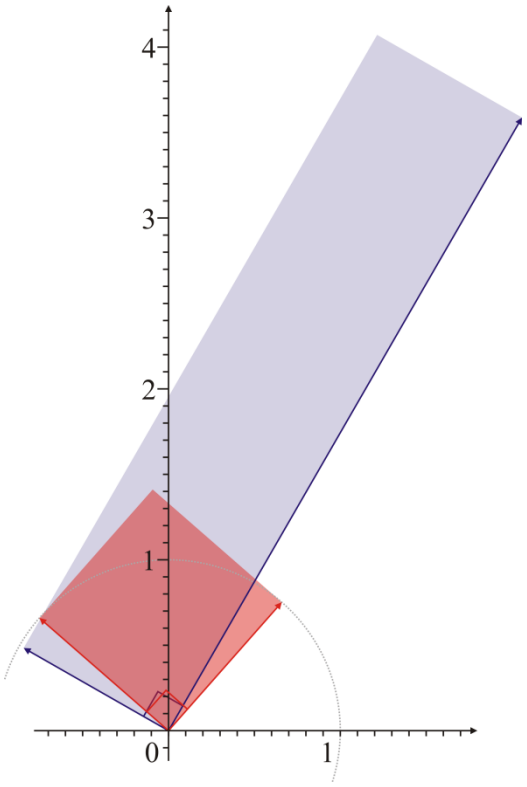
For example, consider the matrix

$$A = \begin{pmatrix} 2 & 1 \\ 2 & 3 \end{pmatrix}.$$

The eigenvalues are 1 and 4, but as the matrix is not symmetric, the corresponding eigenvectors, $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$, are not orthogonal. The image of the rhombus defined by two normalized eigenvectors is a parallelogram that has four times the area of the rhombus, as is shown in



There are two normalized vectors, however, that are orthogonal, and the image of the square defined by these two vectors is itself a rectangle, although this rectangle is rotated:



Recall that symmetric real matrices and normal complex matrices are diagonalizable with respect to an orthonormal basis of eigenvectors. Additionally, the eigenvalues of a symmetric real matrix or conjugate symmetric complex matrix are real, and that every conjugate symmetric matrix is normal.

Note that given any matrix representing a linear transformation $A:\mathbf{R}^n \to \mathbf{R}^m$, the matrices AA^* and A^*A are normal, for $\left(AA^*\right)^* = A^{**}A^* = AA^*$ and $\left(A^*A\right)^* = A^*A^{**} = A^*A$ where $A^*A:\mathbf{R}^n \to \mathbf{R}^n$ and $AA^*:\mathbf{R}^m \to \mathbf{R}^m$. Now, let $\mathbf{u}_1$ and $\mathbf{u}_2$ be eigenvectors of A^*A with corresponding real eigenvalues λ_1 and λ_2. Thus, we have two theorems:

Theorem

If the linear operator $A:\mathbf{R}^n \to \mathbf{R}^m$, all eigenvectors of A^*A are nonnegative.

Proof:

If $\mathbf{u}$ is an eigenvector of A^*A with a corresponding eigenvalue λ, then

$$\begin{aligned}
\lambda\|\mathbf{u}\|_2^2 &= \lambda\langle\mathbf{u},\mathbf{u}\rangle \\
&= \langle\mathbf{u},\lambda\mathbf{u}\rangle \\
&= \langle\mathbf{u},\left(A^*A\right)\mathbf{u}\rangle \\
&= \langle\mathbf{u},A^*\left(A\mathbf{u}\right)\rangle \\
&= \langle A\mathbf{u},A\mathbf{u}\rangle \\
&= \|A\mathbf{u}\|_2^2 \geq 0
\end{aligned}$$

It therefore follows that as $\|\mathbf{u}\|_2^2 > 0$ and $\|A\mathbf{u}\|_2^2 \geq 0$, $\lambda = \dfrac{\|A\mathbf{u}\|_2^2}{\|\mathbf{u}\|_2^2} \geq 0$. ∎

Theorem

If the linear operator $A: \mathbf{R}^n \to \mathbf{R}^m$ and $\mathbf{u}_1$ and $\mathbf{u}_2$ are eigenvectors of A^*A with corresponding (real) eigenvalues λ_1 and λ_2, then as well as $\mathbf{u}_1$ and $\mathbf{u}_2$ being orthogonal, so are $A\mathbf{u}_1$ and $A\mathbf{u}_2$.

Proof:

As A^*A is a normal complex or symmetric real matrix, $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle = 0$, and thus

$$\begin{aligned}
\langle A\mathbf{u}_1, A\mathbf{u}_2 \rangle &= \langle A^*(A\mathbf{u}_1), \mathbf{u}_2 \rangle \\
&= \langle (A^*A)\mathbf{u}_1, \mathbf{u}_2 \rangle \\
&= \langle \lambda_1 \mathbf{u}_1, \mathbf{u}_2 \rangle \\
&= \lambda_1 \langle \mathbf{u}_1, \mathbf{u}_2 \rangle \\
&= 0
\end{aligned}$$

Thus, $A\mathbf{u}_1$ and $A\mathbf{u}_2$ are orthogonal. ∎

We may now deduce the relationship that eigenvectors A^*A

Theorem

If the linear operator $A: \mathbf{R}^n \to \mathbf{R}^m$ and A^*A has an eigenvector $\mathbf{u}$ with corresponding nonnegative eigenvalue λ, then $\|A\mathbf{u}\|_2 = \sqrt{\lambda}\,\|\mathbf{u}\|_2$.

Proof:

From a previous theorem, $\lambda = \dfrac{\|A\mathbf{u}\|_2^2}{\|\mathbf{u}\|_2^2} \geq 0$, and therefore $0 \leq \sqrt{\lambda} = \sqrt{\dfrac{\|A\mathbf{u}\|_2^2}{\|\mathbf{u}\|_2^2}} = \dfrac{\|A\mathbf{u}\|_2}{\|\mathbf{u}\|_2}$. The result follows. ∎

We will now formalize the relationship between the eigenvectors and eigenvalues of A^*A and the linear operator *A*.

Definition

Given any linear operator $A: \mathbf{R}^n \to \mathbf{R}^m$, we define the *singular values* of *A* to be the square roots of the eigenvalues of A^*A with

Theorem

If $A: \mathbf{R}^n \to \mathbf{R}^n$ is a symmetric real or conjugate symmetric complex matrix, then the singular values of *A* are the eigenvalues of *A*.

Proof:

For a geometric interpretation of the singular values are as follows:

If $A:\mathbf{R}^n \to \mathbf{R}^m$ and *S* is the unit sphere in $\mathbf{R}^n$, then the image of the unit sphere is an ellipsoid. The semi-axes of the ellipsoid are those lines tangent to the surface that pass through the origin. The singular values are the lengths of those semi axes with the expected

The application of singular values is as follows:

If the linear operator $A:\mathbf{R}^n \to \mathbf{R}^m$ has the singular values $s_1, s_2, \ldots, s_n$ with $s_1 \geq s_2 \geq \cdots \geq s_n \geq 0$ with corresponding normalized singular vectors $\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_n$ where $A\hat{\mathbf{u}}_1 = s_1\hat{\mathbf{v}}_1, A\hat{\mathbf{u}}_2 = s_2\hat{\mathbf{v}}_2, \ldots, A\hat{\mathbf{u}}_n = s_n\hat{\mathbf{v}}_n$, then the action of *A* may be written as

$$A\mathbf{w} = \sum_{i=1}^{n} s_i \langle \hat{\mathbf{v}}_i, \mathbf{w} \rangle \hat{\mathbf{u}}_i ,$$

and the best approximation of the action of *A* given *k* < *n* singular values is to use the approximation

$$A\mathbf{w} \approx \sum_{i=1}^{k} s_i \langle \hat{\mathbf{v}}_i, \mathbf{w} \rangle \hat{\mathbf{u}}_i .$$

For example, the singular values of the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

are $\frac{1}{2}\sqrt{321} + \frac{1}{2}\sqrt{249}$, $\frac{1}{2}\sqrt{321} - \frac{1}{2}\sqrt{249}$ and 0, or approximately 16.85, 1.07 and 0. Consequently, we see that the most significant action of this matrix is associated with the first singular value—the other two are negligible in comparison. We may therefore approximation the action of this matrix by

$$A\mathbf{v} \approx 16.85 \left\langle \begin{pmatrix} 0.4797 \\ 0.5724 \\ 0.6651 \end{pmatrix}, \mathbf{v} \right\rangle \begin{pmatrix} 0.2148 \\ 0.5206 \\ 0.8263 \end{pmatrix}.$$

For example:

$$A\hat{\mathbf{e}}_1 = \begin{pmatrix} 1 \\ 4 \\ 7 \end{pmatrix} \text{ and } s_1 \langle \hat{\mathbf{v}}_1, \hat{\mathbf{e}}_1 \rangle \hat{\mathbf{u}}_1 = \begin{pmatrix} 0.7776 \\ 1.8843 \\ 2.9910 \end{pmatrix},$$

$$A\hat{\mathbf{e}}_2 = \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix} \text{ and } s_1 \langle \hat{\mathbf{v}}_1, \hat{\mathbf{e}}_2 \rangle \hat{\mathbf{u}}_1 \approx \begin{pmatrix} 1.8843 \\ 4.5660 \\ 7.2477 \end{pmatrix}, \text{ and}$$

$$A\hat{\mathbf{e}}_3 = \begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix} \text{ and } s_1 \langle \hat{\mathbf{v}}_1, \hat{\mathbf{e}}_3 \rangle \hat{\mathbf{u}}_1 \approx \begin{pmatrix} 2.9910 \\ 7.2477 \\ 11.5045 \end{pmatrix}.$$

Calculating $A\mathbf{w}$ in general mn multiplications and $m(n-1)$ additions, while $s_1 \langle \hat{\mathbf{v}}_1, \mathbf{w} \rangle \hat{\mathbf{u}}_1$ requires only $n + m + 1$ multiplications and $n - 1$ additions.

principle component analysis

multiplexing in MIMO communications

eigenvectors of a symmetric or normal matrix are orthogonal; however, a general matrix may not even have eigenvalues defined (for example, if $A : \mathbf{R}^3 \to \mathbf{R}^2$). However, in all cases, the matrix AA^* is symmetric

If a matrix A can be written as $A = \mathbf{u} \alpha \mathbf{v}^*$ where $\mathbf{u}$ and $\mathbf{u}$ are unit vectors, then matrix-vector multiplication can be easily reduced from an operation requiring n^2 multiplications and $n(n-1)$ additions, to one that requires only $2n + 1$ multiplications and $n - 1$ additions. The SVD can be used to determine whether or not a linear transformation can be represented

14 Matrix norms and condition

Consider the following inverse problem **Gv** = **i** with:

$$\mathbf{G} = \begin{pmatrix} 2.4 & 1.1 & 2.6 \\ 1.9 & 4.6 & 3.7 \\ 1.3 & 8.3 & 4.8 \end{pmatrix} \text{ and } \mathbf{i} = \begin{pmatrix} 0.7 \\ 1.0 \\ 1.3 \end{pmatrix}.$$

For example, the matrix may represent the conductance matrix of a circuit, and the right-hand vector **b** is a vector of the currents. We would like to find the voltages at three nodes within the circuit. This conductance matrix is invertible, and, so we can find the voltages:

$$\mathbf{v} = \begin{pmatrix} 2 \\ 1 \\ -2 \end{pmatrix}.$$

Suppose, however, we have an error in our measurement, and that error is very small:

$$\tilde{\mathbf{i}} = \begin{pmatrix} 0.7 \\ 1.001 \\ 1.3 \end{pmatrix}$$

If we now try to solve $\mathbf{G}\tilde{\mathbf{v}} = \tilde{\mathbf{i}}$, we get a completely different answer:

$$\tilde{\mathbf{v}} = \begin{pmatrix} 18.30 \\ 9.14 \\ -20.49 \end{pmatrix}$$

Consider the issue here: a minute change in the currents resulted in drastic changes in the voltages within the circuit: a 0.1 % change in one of the currents results in a 1000% change in the voltages.

This is an example of an *ill-conditioned* matrix: the matrix is invertible, but errors in the right-hand vector are significantly magnified.

If we are solving the inverse problem **Mx = b**, we will say that the matrix **M** is

1. well-conditioned if a small change in **b** results in a small change in **x**, and
2. ill conditioned if a small change in **b** results in a large change in **x**.

A *change* in a vector can be measured in terms of the norm: given a matrix **M**, if **x** is the solution to $\mathbf{Mx} = \mathbf{b}$ and $\tilde{\mathbf{x}}$ is the solution to $\mathbf{M}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ then **M** is well conditioned if $\|\mathbf{b} - \tilde{\mathbf{b}}\|_2$ is small means that $\|\mathbf{x} - \tilde{\mathbf{x}}\|_2$ is small. For example, consider the matrix

$$\mathbf{M} = \begin{pmatrix} 3 & 2 & 0 \\ -1 & 4 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

which you may already recognize as being diagonally dominant, and therefore invertible. The solution to $\mathbf{Mx} = \mathbf{b}$ when $\mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ is $\mathbf{x} = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.4 \end{pmatrix}$ and the solution to $\mathbf{M}\tilde{\mathbf{x}} = \begin{pmatrix} 1 \\ 1 \\ 1.001 \end{pmatrix}$ is $\tilde{\mathbf{x}} = \begin{pmatrix} 0.20008 \\ 0.19988 \\ 0.40056 \end{pmatrix}$, and comparing the two:

$$\left\| \mathbf{b} - \tilde{\mathbf{b}} \right\|_2 = 0.001 \text{ and } \left\| \mathbf{x} - \tilde{\mathbf{x}} \right\|_2 \approx 0.00578 .$$

The error is only magnified by a factor of approximately 6. Is this acceptable? That depends on the situation. It is unlikely, however, that an error being magnified by a factor of 10,000 is ever acceptable within an engineering solution.

14.1 Matrix 2-norms

Recall that previously we described matrices in terms of their determinant (change in volume) and their eigenvalues and eigenvectors (identifying subspaces that are invariant). We will define the *norm* of a matrix as the maximum the matrix *stretches* the length of a vector. Analytically, the definition is:

$$||| M |||_2 \overset{def}{=} \max_{\mathbf{v} \neq \mathbf{0}} \left\{ \frac{\left\| M\mathbf{v} \right\|_2}{\left\| \mathbf{v} \right\|_2} \right\} \text{ or, equivalently, } ||| M |||_2 \overset{def}{=} \max_{\|\mathbf{v}\|_2 = 1} \left\{ \left\| M\mathbf{v} \right\|_2 \right\} .$$

Note that will use three bars to represent the matrix norm, to differentiate it from the absolute value of a scalar and the norm of a vector:

scalar	absolute value	$\lvert\alpha\rvert$
vector	norm or vector norm	$\left\| \mathbf{v} \right\|$
matrix	norm or matrix norm	$\| \mathbf{M} \|$

In describing such a measure as a *norm*, we have somewhat jumped the gun, but as you may suspect, all the conditions of a norm are satisfied:

It should not be difficult to see that all the conditions of a norm are satisfied by this matrix norm:

1. For any zero matrix, $\left\| O\mathbf{v} \right\|_2 = \left\| \mathbf{0} \right\|_2 = 0$, so $||| O |||_2 = 0$, and if $M \neq O$, then $M\mathbf{v} \neq \mathbf{0}$ for some vector $\mathbf{v}$, and thus, $\left\| M \frac{\mathbf{v}}{\left\| \mathbf{v} \right\|_2} \right\|_2 > \mathbf{0}$, so $\| \mathbf{M} \|_2 > 0$.
2. From the properties of the 2-norm of vectors,

$$\begin{aligned} ||| \alpha M |||_2 &= \max_{\|\mathbf{v}\|_2 = 1} \left\{ \left\| (\alpha M)\mathbf{v} \right\|_2 \right\} \\ &= \max_{\|\mathbf{v}\|_2 = 1} \left\{ \left\| \alpha (M\mathbf{v}) \right\|_2 \right\} \\ &= \max_{\|\mathbf{v}\|_2 = 1} \left\{ |\alpha| \left\| (M\mathbf{v}) \right\|_2 \right\} \\ &= |\alpha| \max_{\|\mathbf{v}\|_2 = 1} \left\{ \left\| (M\mathbf{v}) \right\|_2 \right\} \\ &= |\alpha| ||| M |||_2 \end{aligned}$$

3. Again, based on the properties of the 2-norm,

$$
\begin{aligned}
|||A+B|||_2 &= \max_{\|\mathbf{v}\|_2=1}\left\{\left\|(A+B)\mathbf{v}\right\|_2\right\} \\
&= \max_{\|\mathbf{v}\|_2=1}\left\{\left\|A\mathbf{v}+B\mathbf{v}\right\|_2\right\} \\
&\le \max_{\|\mathbf{v}\|_2=1}\left\{\left\|A\mathbf{v}\right\|_2+\left\|B\mathbf{v}\right\|_2\right\} \\
&\le \max_{\|\mathbf{v}\|_2=1}\left\{\left\|A\mathbf{v}\right\|_2\right\}+\le \max_{\|\mathbf{v}\|_2=1}\left\{\left\|B\mathbf{v}\right\|_2\right\} \\
&= |||A|||_2+|||B|||_2
\end{aligned}
$$

Thus, the collection of matrices forms a normed vector space; however, this norm is not induced by any inner product on matrices.

As an example of how the matrix norm works, consider the matrix

$$M=\begin{pmatrix}3 & 1\\ 2 & 4\end{pmatrix}.$$

It has eigenvalues $\lambda_1=2$ and $\lambda_2=5$, and the corresponding eigenvectors are $\mathbf{v}_1=\begin{pmatrix}-1\\ 1\end{pmatrix}$ and $\mathbf{v}_2=\begin{pmatrix}1\\ 2\end{pmatrix}$, respectively.

Thus,

$$M\mathbf{v}_1=2\mathbf{v}_1 \text{ so } \frac{\left\|M\mathbf{v}_1\right\|_2}{\left\|\mathbf{v}_1\right\|_2}=2 \text{ and } M\mathbf{v}_2=5\mathbf{v}_2 \text{ so } \frac{\left\|M\mathbf{v}_2\right\|_2}{\left\|\mathbf{v}_2\right\|_2}=5.$$

You may note that the vector

$$\mathbf{v}=\begin{pmatrix}5\sqrt{5}-2\\ 11\end{pmatrix}$$

is a multiple of neither eigenvector, and yet

$$M\mathbf{v}=\begin{pmatrix}15\sqrt{5}+5\\ 10\sqrt{5}+40\end{pmatrix}$$

and that

$$\frac{\left\|M\mathbf{v}\right\|_2}{\left\|\mathbf{v}\right\|_2}=\sqrt{15+5\sqrt{5}}\approx 5.116672736,$$

and this number is greater than either eigenvalue. This is shown in , where the image of the unit circle (red) is shown in cyan, two normalized eigenvectors and their images are shown in red, and the vector that is maximally stretched is shown in blue together with its image. As you will observe, the blue vector is stretched ever so slightly more than the second eigenvector.

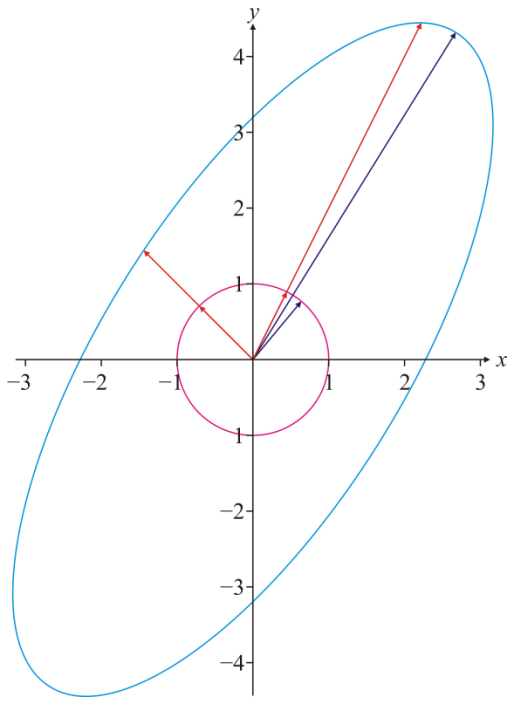


Figure 88. The image of the unit circle under the mapping $M = \begin{pmatrix} 3 & 1 \\ 2 & 4 \end{pmatrix}$ with eigenvectors and their image identified (in red) and as well as a maximally stretched vector $\mathbf{v} = \begin{pmatrix} 5\sqrt{5}-2 \\ 11 \end{pmatrix}$ (in blue).

Given the definition of the norm of a matrix,

$$\| M \|_2 = \max_{\|\mathbf{v}\|_2 = 1} \left\{ \|M\mathbf{v}\|_2 \right\},$$

it follows we need to maximize $\|M\mathbf{v}\|_2$. If a matrix is symmetric, its eigenvectors are real and its eigenvectors form an orthogonal basis, and therefore the 2-norm of a symmetric matrix is the absolute value of the maximum eigenvalue. If a matrix is not symmetric, we have a little more work to do. First, maximizing $\|M\mathbf{v}\|_2$ is equivalent to maximizing $\|M\mathbf{v}\|_2^2 = \langle M\mathbf{v}, M\mathbf{v} \rangle$, and

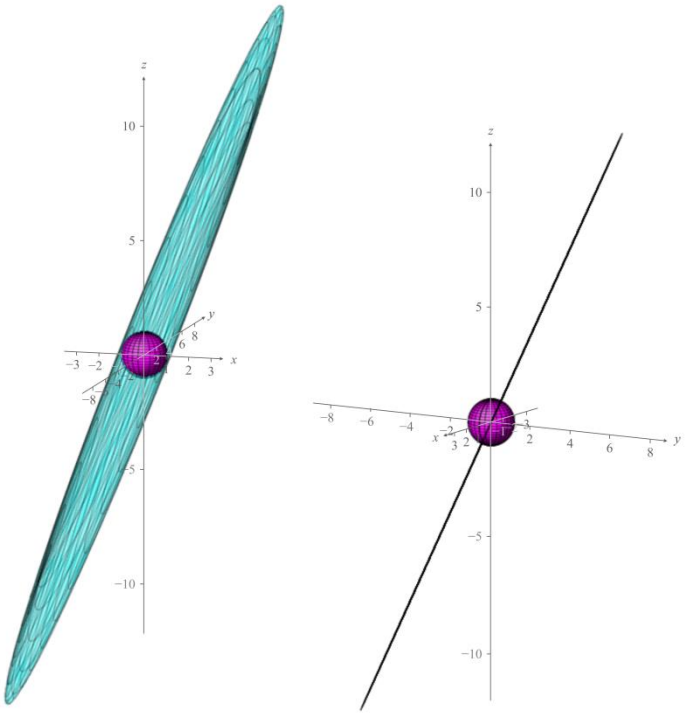
$$\begin{aligned} \|M\mathbf{v}\|_2^2 &= \langle M\mathbf{v}, M\mathbf{v} \rangle \\ &= \langle M^* M\mathbf{v}, \mathbf{v} \rangle \end{aligned}$$

Now, $M^* M$ is positive semi-definite (thus, being symmetric with all eigenvalues greater than or equal to zero), and therefore $\langle M^* M\mathbf{v}, \mathbf{v} \rangle$ is maximized when $\mathbf{v}$ is an eigenvector corresponding to the maximum eigenvalue of $M^* M$. Thus, we may deduce that

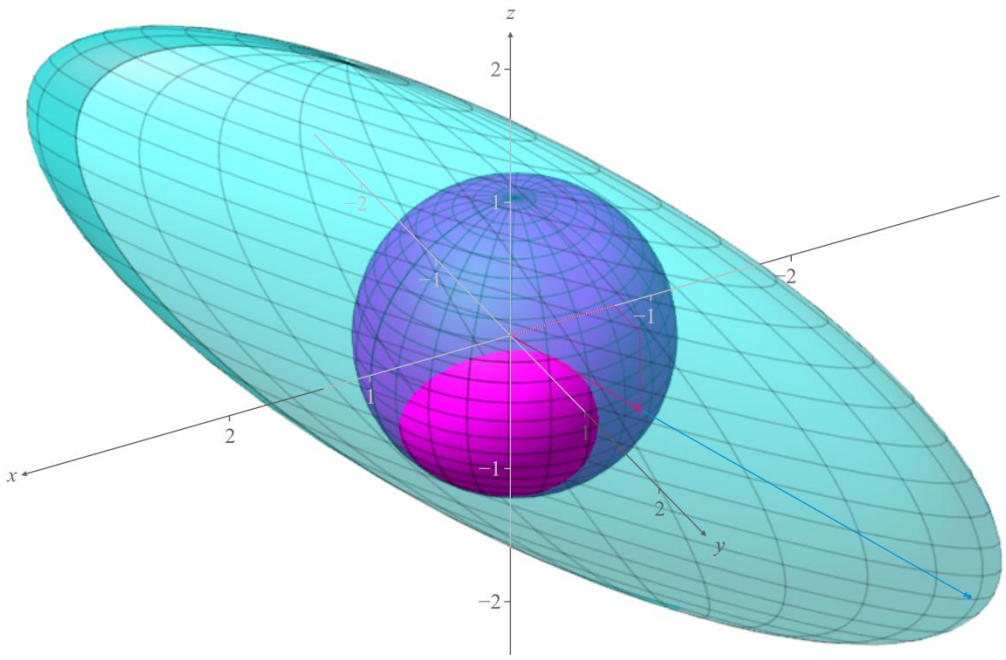
$$\| M \|_2 = \sqrt{\max \left\{ \sigma \left(M^* M \right) \right\}}$$

where, you will recall, $\sigma(M)$ is the collection of all eigenvalues of (or the *spectrum*) of the linear operator *M*. The vectors that are maximally stretched by M are those eigenvectors corresponding to the largest eigenvalue of M^*M .

As a second example, the matrix $M = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ has a 2-norm of $|||\, M \,|||_2 = \sqrt{3}\,\frac{\sqrt{107}+\sqrt{83}}{2} \approx 16.84810$ and a vector that is maximally stretched is $\mathbf{v} = \begin{pmatrix} 557794414 + 5917214\sqrt{8881} \\ 1351458100 + 14340260\sqrt{8881} \\ 2145121786 + 22763306\sqrt{8881} \end{pmatrix}$, an eigenvector of $M^*M = \begin{pmatrix} 66 & 78 & 90 \\ 78 & 93 & 108 \\ 90 & 108 & 126 \end{pmatrix}$ corresponding to the maximum eigenvalue of M^*M .



As a third example, consider the symmetric matrix $M = \begin{pmatrix} 2.4 & -0.8 & 1.0 \\ -0.8 & 1.3 & -0.3 \\ 1.0 & -0.3 & 1.2 \end{pmatrix}$. It's maximum eigenvalue is approximately 3.55166 as is its norm, and XXX shows the unit sphere (in magenta) and its image under M in cyan, together with an eigenvalue corresponding to the maximum eigenvalue.



Note also that the definition gives two suggestions:

1. by using the word max, this indicates that there must be a vector **v** such that $\|\mathbf{Mv}\|_2 = \|\mathbf{M}\|_2 \|\mathbf{v}\|_2$,
2. for all vectors, the inequality $\|\mathbf{Mv}\|_2 \le \|\mathbf{M}\|_2 \|\mathbf{v}\|_2$ must be true, and
3. the matrix **M** need not even be a square matrix.

Theorem
In a complex vector space, the 2-norm of a normal linear operator equals the absolute value of the largest eigenvalue.

Theorem
In a complex vector space, the 2-norm of a normal linear operator equals the absolute value of the largest eigenvalue.

If you pass a matrix to the norm function in Matlab, it automatically calculates the norm of the matrix:

```
>> M = [1 2 3; 4 5 6; 7 8 9];
>> norm( M )
    ans =
      16.848103352614210
```

14.2 Other induced matrix norms

Previously, we described the 2-norm of a vector as the norm *induced* by the inner product:

$$\|\mathbf{v}\|_2^2 = \langle \mathbf{v}, \mathbf{v} \rangle .$$

We have defined the 2-norm of a matrix as

$$|||\mathbf{M}|||_2 = \max_{\mathbf{v}\neq\mathbf{0}} \frac{\|\mathbf{M}\mathbf{v}\|_2}{\|\mathbf{v}\|_2} ;$$

however, we could also define matrix norms induced by the other two vector norms we have described:

$$|||\mathbf{M}|||_1 = \max_{\mathbf{v}\neq\mathbf{0}} \frac{\|\mathbf{M}\mathbf{v}\|_1}{\|\mathbf{v}\|_1} \text{ and } |||\mathbf{M}|||_\infty = \max_{\mathbf{v}\neq\mathbf{0}} \frac{\|\mathbf{M}\mathbf{v}\|_\infty}{\|\mathbf{v}\|_\infty} .$$

These are not the same.

Norm	Algorithm	Formula
1	maximum absolute column sum	$\|\|\|\mathbf{M}\|\|\|_1 = \max_{1\le j\le n}\left\{\sum_{i=1}^{m}\left\|m_{i,j}\right\|\right\}$
2	square root of the maximum eigenvalue of $\mathbf{M}\mathbf{M}^*$	$\|\|\|\mathbf{M}\|\|\|_2 = \sqrt{\max\left\{\sigma\left(\mathbf{M}\mathbf{M}^*\right)\right\}}$
infinity	maximum absolute row sum	$\|\|\|\mathbf{M}\|\|\|_\infty = \max_{1\le i\le m}\left\{\sum_{j=1}^{n}\left\|m_{i,j}\right\|\right\}$

In Matlab, we can use the above formulas to calculate the norms if we introduce the `sum` routine. If the argument to `sum` is a vector, `sum` returns the arithmetic sum of the entries of that vector. You may think that it may do the same for an $m \times n$ matrix for $n > 1$ and $m > 1$; however, here it does something else: it calculates the sum of each of the columns.

```
>> M = [1 -2 -3; 4 -5 6; -7 8 9]
    M =
         1    -2    -3
         4    -5     6
        -7     8     9

>> sum( M )
    ans =
        -2     1    12

>> norm( M, 1 )
    ans =
        18

>> max( sum( abs( M ) ) )
    ans =
        18

>> norm( M, 2 )
    ans =
       14.427729501608967

>> sqrt( max( eig( M*M' ) ) )
    ans =
       14.427729501608967

>> norm( M, Inf )
    ans =
        24

>> % the maximum absolute row sum is the max absolute column sum of the transpose
>> max( sum( abs( M ).' ) )   % recall .' means transpose and not conjugate transpose
    ans =
        24
```

There is no hard-and-fast rule as to which matrices are well conditioned, and which are ill conditioned. In this section, however, we will investigate

14.3 Condition numbers

The rules for invertibility are quite straight-forward: an $n \times n$ matrix is not invertible if any of the following conditions hold:

1. the linear operator has a zero eigenvalue,
2. the determinant of the operator is zero,
3. the dimension of the range is less than n,
4. the rank of the matrix representation is less than n.

Unfortunately, these really only work in ideal cases; where the precision is infinite. Unfortunately, when a number is stored in a computer, it is stored with only a finite amount of precision. As a double-precision floating-point number, p is stored as

$$11.001001000011111101101010100010001000010110100011000_2$$
$$3.14159265358979311599796346854418516159057617187 5_{10}$$
$$3.1415926535897932384626433832795028841971693993 7\cdots$$

For example, we know that this matrix is not invertible:

$$M = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

When you calculate the determinant, it should be zero:

```
>> M = [1 2 3; 4 5 6; 7 8 9];
>> 1*5*9 + 2*6*7 + 3*4*8 - 1*6*8 - 2*4*9 - 3*5*7
    ans =
         0

>> det( M )
    ans =
         6.661338147750939e-16
```

You may consider this an error in the implementation of the determinant in Matlab; however, this would be a purist's knee-jerk reaction, one that you need to suppress. Here is a perfectly invertible matrix that has an even smaller determinant:

$$M = \begin{pmatrix} 0.000001 & 0 & 0 \\ 0 & 0.000001 & 0 \\ 0 & 0 & 0.000001 \end{pmatrix},$$

which has the determinant 10^{-18}, and yet on

14.4 Operator C*-algebras

If we consider again the collection of all linear operators on a complex vector space U, we note that $\|A^*A\|_2 = \|A\|_2^2$, as

$$\|A^*A\|_2 = \max_{\|\mathbf{u}\|_2=1}\left\{\left\|\left(A^*A\right)\mathbf{u}\right\|_2\right\} = = \max_{\|\mathbf{u}\|_2=1}\left\{\left\langle\left(A^*A\right)\mathbf{u},\left(A^*A\right)\mathbf{u}\right\rangle\right\}$$

Consequently,

$$\begin{aligned}\left\langle (\alpha A)^* \mathbf{u}, \mathbf{v} \right\rangle &= \left\langle \mathbf{u}, \alpha A \mathbf{v} \right\rangle \\ &= \alpha^* \left\langle \mathbf{u}, A\mathbf{v} \right\rangle \\ &= \alpha^* \left\langle A^* \mathbf{u}, \mathbf{v} \right\rangle \\ &= \left\langle \alpha^* A^* \mathbf{u}, \mathbf{v} \right\rangle\end{aligned}$$

Thus, for a complex vector space U , the algebra of $L(U,U)$ is said to define a C*-algebra (see-star-algebra). Such algebras are used in quantum mechanics.

15 Iterative methods for approximating solutions

In this topic, we will now introduce iterative methods of approximation solutions to equations, and we will then look specifically at techniques for approximating solutions to systems of linear equations. The justification for using iterative techniques is that approximations to the solutions can be found much faster than using techniques such as Gaussian elimination and backward substitution. Previously, we saw that Gaussian elimination requires $\Theta(n^3)$ floating-point operations, and backward substitution requires $\Theta(n^2)$ operations.

To find such techniques, we will begin by looking at the *fixed-point theorem*.

15.1 The fixed-point theorem and applications

The fixed-point theorem says that, given an equation of the form

$$x = f(x),$$

then if we start with an arbitrary point x_0 and calculate

$$\begin{aligned} x_1 &= f(x_0) \\ x_2 &= f(x_1) \\ x_3 &= f(x_2) \\ &\vdots \end{aligned}$$

then, if this sequence converges, it will converge to a solution of the above equation. We will give this theorem without proof.

The simplest example is also likely one that you have previously tried on your calculator. If you type in any random number into your calculator and then start hitting the cos key (making sure your calculator is set to radians), you will note that very quickly it converges to approximately 0.73908513321516. For example,

1970
−0.9755896102
0.5606799346
0.8468937433
0.6623136279
0.7885715941
0.7048592701
0.7617027420
0.7236619051
0.7493860898
0.7321071960
0.7437675240
0.7359229279
0.7412115340
0.7376510942
0.7400503589
0.7384346014
0.7395231829

If we plot the values of these first forty iterations, we have the points in Figure 89.

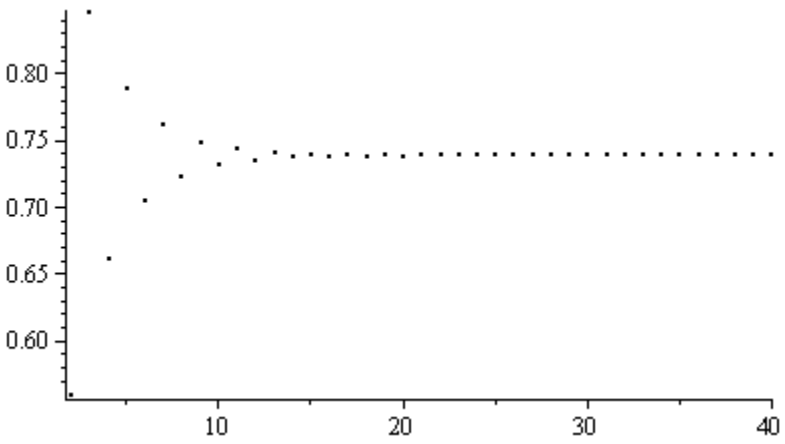


Figure 89. Forty iterations of applying $x_{n+1} = \cos(x_n)$ when $x_0 = 1970$.

On the other hand, if you perform the same iterations with the sin key, it does converge, but the convergence is very slow: even after 10 000 iterations, –0.01726512104 is not very close to the solution of $x = 0$. These first 10 000 iterations are shown in Figure 90.

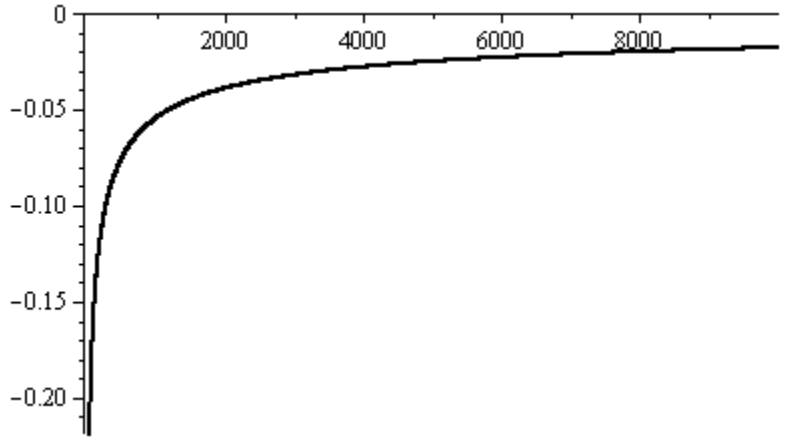


Figure 90. 10 000 iterations of applying $x_{n+1} = \sin(x_n)$ when $x_0 = 1970$.

Finally, if you perform the same iterations again with the tan key, you get a different behavior, as shown in .

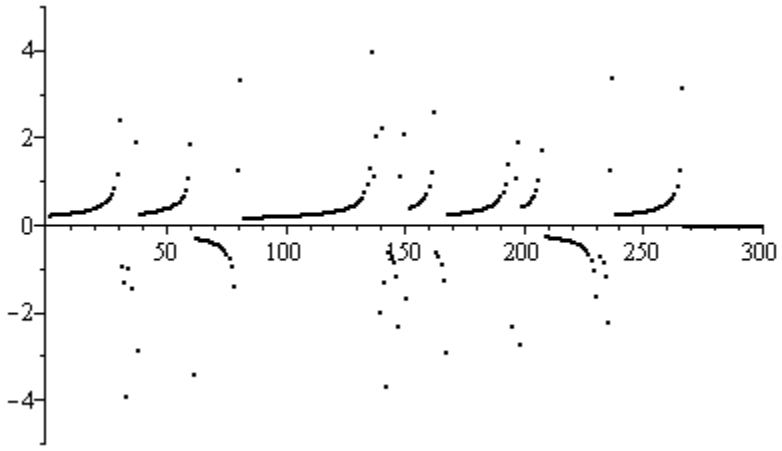


Figure 91. 300 iterations of applying $x_{n+1} = \tan(x_n)$ when $x_0 = 1970$.

It seems that after iteration 268, it seems to start converging to zero, but this is only an illusion, for if we plot all iterations from 268 to 14640, we note that again, it is diverging from the solution $x = 0$, as shown in Figure 92.

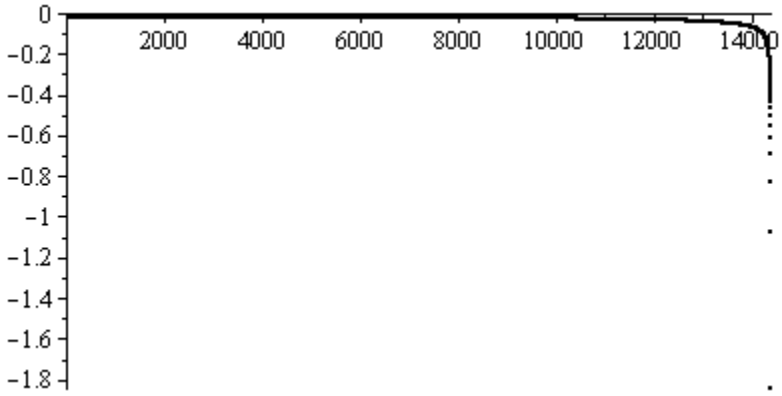


Figure 92. Iterations 268 to 14640 of applying $x_{n+1} = \tan(x_n)$ when $x_0 = 1970$.

Thus, only if the sequence $x_0, x_1, \ldots$ converges does it converge to a solution to the equation $x = f(x)$. We will look at a number of different applications of the fixed-point theorem, including the approximation of

1. the square root of two,
2. the maximum eigenvalue of a matrix,
3. the 2-norm of a matrix,
4. the minimum eigenvalue of a matrix, and
5. the condition number of a matrix.

In the next section, we will use iterative techniques to approximate solutions to systems of linear equations.

15.1.1 Approximating the square root of two

Let's look at another example, one that was known to mathematicians as long ago as the era of the Babylonian empire: suppose we want to approximate the value of $\sqrt{2}$. If we look at the equation

$$x^2 = 2,$$

we see that this is not of the form $x = f(x)$, but we could divide both sides by x to get:

$$x = \frac{2}{x}.$$

Unfortunately, this will only converge if we start with $x_0 = \sqrt{2}$ and for all other values, it will alternate between x_0 and $\frac{2}{x_0}$. We could, however, try rewriting the equation:

$$\frac{x^2}{2} = 1$$
$$\therefore x^2 - \frac{x^2}{2} = 1$$
$$\therefore x^2 = \frac{x^2}{2} + 1$$
$$\therefore x = \frac{x}{2} + \frac{1}{x}$$

Notice that the square root of two is still a solution to the last equation:

$$\frac{\sqrt{2}}{2} + \frac{1}{\sqrt{2}} = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}$$
$$= \frac{2}{\sqrt{2}} = \sqrt{2}$$

Now, if we start our iteration, we get an excellent approximation to the square root of two: with each iteration, the number of correct digits approximately doubles:

1.0
1.5
1.41666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666666667
1.41421568627450980392156862745098039215686274509803921568627450980392156862745098039215686274509803921
1.41421356237468991062629557889013491011655962211574404458490501920005437183538926835899004315764434 02
1.41421356237309504880168962350253024361498192577619742849828949862319582422892362178494183673583035 66
1.41421356237309504880168872420969807856967187537723400156101313311326525563033997853178716125071047 52
1.41421356237309504880168872420969807856967187537694807317667973799073247846210703885038753432764160 16

Using this technique to approximate the diagonal of a unit square dates back to ancient tablet, YBC 7289. This was known from the time of the Babylonian empire at approximately the era of Hammurabi (which, apart from the location, was not associated with the later neo-Babylonian empire or, more correctly named, *Chaldean* empire of

Nebuchadnezzar II and Belshazzar). This photograph was taken by Christopher Anagnostakis with permission of the Curators of the Babylonian Collection at Yale University.

Any number could be made up of up to nine 1s (or 𒁹) and up to five 10s (or 𒌋) leading to a base-60 numbering system (from which we indirectly get our 360 degrees in a circle, 60 minutes per degree, and 60 seconds per minute, although the latter are more directly inherited from the Romans). This tablet contained three numbers, as indicate in .



Here we note that the edge of the square is 30 units. The numbers in the middle, however, are more interesting. The first number could be interpreted as

$$1+\frac{24}{60}+\frac{51}{60^2}+\frac{10}{60^3}=1.41421\overline{296}$$

while the lower number is exactly 30 times this number, or

$$42+\frac{25}{60}+\frac{35}{60^2}=42.426\overline{38}.$$

Both are reasonable approximations to $\sqrt{2}\approx 1.414213562$ and $30\sqrt{2}\approx 42.42640687$ with a relative error of only 4.24×10^{-5}. You may want to recall that

$$\begin{aligned}30\cdot 1+30\cdot\frac{24}{60}+30\cdot\frac{51}{60^2}+30\cdot\frac{10}{60^3}&=30+\frac{720}{60}+\frac{1530}{60^2}+\frac{300}{60^3}\\&=30+12+\left(\frac{25}{60}+\frac{30}{60^2}\right)+\frac{5}{60^2}\\&=42+\frac{25}{60}+\frac{35}{60^2}\end{aligned}$$

The first three iterations of our formula starting with 1 in base 60 are:

$$1,1+\frac{30}{60},1+\frac{25}{60},1+\frac{24}{60}+\frac{51}{60^2}+\frac{10}{60^3}+\frac{35.\overline{2941176470588235}}{60^4}\ .$$

If this was the procedure used, the mathematician appears to have rounded down as opposed to rounding up with the last iteration.

15.1.2 Estimating the maximum eigenvalue and a corresponding eigenvector

As you are aware, finding eigenvalues and eigenvectors requires a significant amount of work; however, what if we only wish to find the maximum eigenvalue and a corresponding eigenvector?

For example, consider the matrix $A = \begin{pmatrix} 3 & 1 \\ 4 & 6 \end{pmatrix}$ which has eigenvalues $\lambda_1 = 7$ and $\lambda_2 = 2$, with corresponding eigenvectors $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, respectively. Watch what happens if we start with a completely random vector and start multiply it by *A*:

$$\mathbf{v}_0 = \begin{pmatrix} -1.0689 \\ -0.8095 \end{pmatrix}, \mathbf{v}_1 = \begin{pmatrix} -4.0161 \\ -9.1325 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} -21.1808 \\ -70.8593 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} -134.4017 \\ -509.8789 \end{pmatrix}, \mathbf{v}_4 = \begin{pmatrix} -913.0840 \\ -3596.8803 \end{pmatrix}$$

and you will see that after only four iterations, $\mathbf{v}_4 = A^4\mathbf{v}_0$ is already a close approximation to a multiple of the corresponding eigenvector, for $\frac{-913.0840}{-3596.8803} \approx 3.9393$, so the second entry is approximately four times the first, and $\frac{\|\mathbf{v}_4\|_2}{\|\mathbf{v}_3\|_2} \approx 7.0377$, which is quite close to the corresponding maximum eigenvalue.

How do we tell, however, if the sequence is not converging?

To actually find this maximum eigenvalue and its corresponding eigenvector, we must be more careful. Note that by the definition of the inner product, if $\mathbf{v}$ is an eigenvector of *A*, then $\langle \mathbf{v}, A\mathbf{v} \rangle = \langle \mathbf{v}, \lambda\mathbf{v} \rangle = \lambda \langle \mathbf{v}, \mathbf{v} \rangle$. Therefore, we will start with a vector $\hat{\mathbf{v}}_0$ and thus, given $\hat{\mathbf{v}}_k$, we will calculate $\mathbf{v}_{k+1} \leftarrow A\hat{\mathbf{v}}_k$ and let $\lambda_{k+1} = \langle \hat{\mathbf{v}}_k, \mathbf{v}_{k+1} \rangle$ and we will look at how close these two vectors are: $\|\lambda_{k+1}\hat{\mathbf{v}}_k - \mathbf{v}_{k+1}\|_2$. If this is small, we will say that the sequence has converged, and we will return λ_{k+1} and $\frac{\mathbf{v}_{k+1}}{\|\mathbf{v}_{k+1}\|_2}$.

$\hat{\mathbf{v}}_k \leftarrow \frac{\mathbf{v}_{k+1}}{\|\mathbf{v}_{k+1}\|_2}$. We will say that the process has converged if

In Matlab, the following code finds the maximum value and a corresponding eigenvector.

```
function [value, vector] = maxeig( M, epsilon, max_iter )
    if nargin ~= 3
        throw( MException( 'MATLAB:invalid_input', ...
                           sprintf( 'expecting 3 arguments, but got %d', nargin ) ) );
    end

    [m, n] = size( M );

    if m ~= n
        throw( MException( 'MATLAB:invalid_input', ...
                           'the input matrix must be square' ) );
    end

    if isreal( M ) && issymmetric( M )
        vector = rand( n, 1 );
    else
        vector = rand( n, 1 ) + 1j*rand( n, 1 );
    end

    vector = vector/norm( vector, 2 );  % Normalize the vector
    value = 1;

    for k = 1:max_iter
        vector_old = vector;
        value_old = value;

        vector = M*vector;
        value = vector_old'*vector;

        % don't check the vectors unless the eigenvalue has converged
        if abs( value - value_old ) < epsilon ...
                          && norm( vector - value*vector_old, 2 ) < epsilon
            vector = vector/norm( vector, 2 );
            return;
        end

        vector = vector/norm( vector, 2 );
    end

    throw( MException( 'MATLAB:non_convergence', ...
                       'The power method did not converge' ) );
end
```

15.1.3 Estimating the 2-norm of a matrix

Recall that the 2-condition number of a matrix A was defined as the square root of the maximum eigenvalue of A^TA. We can apply the same algorithm, but we can approach this in two manners:

1. calculate A^TA, pass it to the previous algorithm, and take the square root, or
2. write a specialized algorithm that calculates $A^T(A\mathbf{v})$ for each iteration.

The first requires the computation of A^TA, which is an $\Theta(n^3)$ operation, the second requires an extra matrix-vector multiplication, an $\Theta(n^2)$ operation, with each iteration. Thus, for sufficiently large matrices, it is more desirable to use the second approach.

In Matlab, the following code finds the maximum value and a corresponding eigenvector.

```
function [value] = norm2( M, epsilon, max_iter )
    [m, n] = size( M );

    if m ~= n
        throw( MException( 'MATLAB:invalid_input', ...
                           'the input matrix must be square' ) );
    end

    Mt = M';

    vector = rand( n, 1 );       % Create a random n-dimensional column vector
    vector = vector/norm( vector, 2 );  % Normalize the vector
    norm_old = 1.0;

    for k = 1:max_iter
        vector_old = vector;

        vector = Mt*(M*vector);
        value  = norm( vector, 2 );

        if norm( vector - vector_old, 2 ) < epsilon
            value = sqrt( value );
            return;
        else
            vector = vector/value;
        end
    end

    throw( MException( 'MATLAB:non_convergence', ...
                       'The power method did not converge' ) );
end
```

15.1.4 Estimating the minimum eigenvalue and a corresponding eigenvector

If λ is a minimum eigenvalue of the matrix A, then $\frac{1}{\lambda}$ a maximum eigenvalue of A^{-1}. Therefore, finding the minimum eigenvalue of a matrix is equivalent to finding the maximum eigenvalue of M^{-1}. In order to find the maximum eigenvalue of M^{-1}, we can use the same algorithm; however, we cannot compute the inverse of M; therefore, we will start with a random unit vector $\hat{\mathbf{v}}_0$, and then solve

$$M\mathbf{v}_{k+1} = \hat{\mathbf{v}}_k$$

and set $\hat{\mathbf{v}}_{k+1} = \dfrac{\mathbf{v}_{k+1}}{\|\mathbf{v}_{k+1}\|_2}$ for $k = 0,1,2,\ldots$.

If the matrix M is real and symmetric, all of its eigenvalues are real, and therefore the minimum eigenvalue is real. Consequently, we may start with a real vector $\hat{\mathbf{v}}_0 \in \mathbf{R}^n$. However, in any other situation, the maximum eigenvalue may indeed be complex, and if the matrix is real, for each complex eigenvalue, its complex conjugate will also be an eigenvalue. In this case, we must start with $\hat{\mathbf{v}}_0 \in \mathbf{C}^n$ and if all the entries of M are real, we must also ensure that $\hat{\mathbf{v}}_0 \in \mathbf{C}^n \setminus \mathbf{R}^n$; that is, the entries of $\hat{\mathbf{v}}_0$ cannot be all real.

For example, if we consider again the matrix $M = \begin{pmatrix} 3 & 1 \\ 4 & 6 \end{pmatrix}$ with eigenvalues $\lambda_1 = 7$ and $\lambda_2 = 2$ and corresponding eigenvectors $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$ and $\mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, respectively, and then execute the following code,

```
>> v = rand( 2, 1 )
>> for i = 1:5
       v = M \ v;
       v = v/norm( v, 2 );
end
```

we get the sequence of vectors

$$\hat{\mathbf{v}}_0 = \begin{pmatrix} 0.6584 \\ 0.7526 \end{pmatrix}, \hat{\mathbf{v}}_1 = \begin{pmatrix} 0.9932 \\ -0.1167 \end{pmatrix}, \hat{\mathbf{v}}_2 = \begin{pmatrix} 0.8148 \\ -0.5797 \end{pmatrix}, \hat{\mathbf{v}}_3 = \begin{pmatrix} 0.7381 \\ -0.6747 \end{pmatrix}, \hat{\mathbf{v}}_4 = \begin{pmatrix} 0.7160 \\ -0.6981 \end{pmatrix}, \hat{\mathbf{v}}_5 = \begin{pmatrix} 0.7096 \\ -0.7046 \end{pmatrix}, \ \cdots$$

and you will see that after only five iterations, the sequence appears to be converging to $\hat{\mathbf{v}} = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$.

Like the previous method, this method will only converge if the minimum eigenvalue is unique.

15.1.5 Estimating the condition number of a matrix

The condition number is defined as the ratio between the minimum and maximum stretch of a linear operator. Consequently, it can be defined as

$$\text{cond}(A) \overset{def}{=} \frac{|||A|||_2}{|||A^{-1}|||_2} \ .$$

15.1.6 Summary of the fixed-point theorem and applications

We have described the fixed-point theorem and looked at how it can be used to approximate the square-root of two and how it can be used to find approximations to the maximum eigenvalue and a corresponding eigenvector of a matrix. We will now look at an application of the fixed-point theorem for approximating solutions to systems of linear equations.

15.2 Iterative methods of approximating solutions to systems of linear equations

We have seen that iteration can be used to find approximations to solutions to simple problems of the form $x = f(x)$, at least under some conditions. Can we use this to approximate solutions to more complex problems?

Take for example, the inverse problem

$$A\mathbf{x} = \mathbf{b}$$

where we are trying to solve for **x**. To motivate this, we will assume we are solving this for

$$A = \begin{pmatrix} 6.2 & 1.2 & 0.5 & 0.1 \\ -2.1 & 5.9 & 1.8 & 0.4 \\ 0.9 & 1.7 & 9.3 & 2.1 \\ 0.2 & 0.8 & 1.6 & 10.5 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} 3.2 \\ 5.1 \\ -8.6 \\ -4.9 \end{pmatrix}.$$

This is an equation, but it is not of the form $\mathbf{x} = \mathbf{f}(\mathbf{x})$, and thus, there is nothing to iterate. Like before, we must somehow isolate an **x** on the left-hand side. To do this, let us first assume that all the diagonal entries of **M** are non-zero. This is generally the case in almost all engineering applications, and therefore it does not an unreasonable assumption. In this case, we can write the matrix **M** as a sum of two matrices, one consisting of the diagonal entries, and the other consisting of all the off-diagonal entries. Thus, for example, if

$$A = \begin{pmatrix} 6.2 & 1.2 & 0.5 & 0.1 \\ -2.1 & 5.9 & 1.8 & 0.4 \\ 0.9 & 1.7 & 9.3 & 2.1 \\ 0.2 & 0.8 & 1.6 & 10.5 \end{pmatrix}$$

then

$$A = \begin{pmatrix} 6.2 & 0 & 0 & 0 \\ 0 & 5.9 & 0 & 0 \\ 0 & 0 & 9.3 & 0 \\ 0 & 0 & 0 & 10.5 \end{pmatrix} + \begin{pmatrix} 0 & 1.2 & 0.5 & 0.1 \\ -2.1 & 0 & 1.8 & 0.4 \\ 0.9 & 1.7 & 0 & 2.1 \\ 0.2 & 0.8 & 1.6 & 0 \end{pmatrix}.$$

We will represent these two matrices as $A = A_{\text{diag}} + A_{\text{off}}$. In this case, our problem is now solving

$$(A_{\text{diag}} + A_{\text{off}})\mathbf{x} = \mathbf{b}$$

and thus

$$A_{\text{diag}}\,\mathbf{x} + A_{\text{off}}\,\mathbf{x} = \mathbf{b}.$$

This doesn't help us yet, but there is one real advantage to extracting the diagonal matrix—its inverse is trivial. For example, the inverse of the above diagonal matrix is

$$A_{\text{diag}}^{-1} = \begin{pmatrix} \frac{1}{6.2} & 0 & 0 & 0 \\ 0 & \frac{1}{5.2} & 0 & 0 \\ 0 & 0 & \frac{1}{9.3} & 0 \\ 0 & 0 & 0 & \frac{1}{10.5} \end{pmatrix} \approx \begin{pmatrix} 0.1613 & 0 & 0 & 0 \\ 0 & 0.1695 & 0 & 0 \\ 0 & 0 & 0.1075 & 0 \\ 0 & 0 & 0 & 0.0952 \end{pmatrix}.$$

Thus, if we bring the off-diagonal expression to the right-hand side, we have

$$A_{\text{diag}}\,\mathbf{x} = \mathbf{b} - A_{\text{off}}\,\mathbf{x},$$

and now we need only multiply both sides on the left by the inverse of A_{diag}:

$$\begin{aligned} \mathbf{x} &= \mathbf{M}_{\text{diag}}^{-1}\left(\mathbf{b} - \mathbf{M}_{\text{off}}\mathbf{x}\right) \\ &= \mathbf{M}_{\text{diag}}^{-1}\mathbf{b} - \mathbf{M}_{\text{diag}}^{-1}\left(\mathbf{M}_{\text{off}}\mathbf{x}\right) \end{aligned}.$$

Now we have an equation in the form **x** = **f**(**x**).

15.2.1 The Jacobi method

The Jacobi method uses this directly. The best initial approximation to our solution is $\mathbf{x}_0 = \mathbf{M}_{\text{diag}}^{-1}\mathbf{b}$. This is quite straight-forward to calculate, and we will use this in subsequent iterations. In our problem above,

$$\mathbf{x}_0 = \mathbf{M}_{\text{diag}}^{-1}\mathbf{b} = \begin{pmatrix} 0.5161 \\ 0.8644 \\ -0.9247 \\ -0.4667 \end{pmatrix}.$$

Now we iterate, by calculating $\mathbf{x}_1 = \mathbf{M}_{\text{diag}}^{-1}\mathbf{b} - \mathbf{M}_{\text{diag}}^{-1}\left(\mathbf{M}_{\text{off}}\mathbf{x}_0\right)$, or

$$\mathbf{x}_1 = \begin{pmatrix} 0.4309 \\ 1.3619 \\ -1.0273 \\ -0.4014 \end{pmatrix}.$$

Further computations demonstrate that it does appear to converge:

$$\mathbf{x}_2 = \begin{pmatrix} 0.3419 \\ 1.3584 \\ -1.1247 \\ -0.4221 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 0.3507 \\ 1.3578 \\ -1.1108 \\ -0.4053 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 0.3498 \\ 1.3556 \\ -1.1154 \\ -0.4075 \end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix} 0.3503 \\ 1.3567 \\ -1.1143 \\ -0.4066 \end{pmatrix}, \mathbf{x}_6 = \begin{pmatrix} 0.3500 \\ 1.3566 \\ -1.1148 \\ -0.4069 \end{pmatrix}$$

and it does appear to be converging to the solution

$$\mathbf{x} = \begin{pmatrix} 0.3500 \\ 1.3567 \\ -1.1147 \\ -0.4068 \end{pmatrix}.$$

After only six iterations, $\|\mathbf{x}_6 - \mathbf{x}\|_2 \approx 0.00011846 = 1.1846 \times 10^{-4}$.

Important: Note that calculating $\mathbf{M}_{\text{diag}}^{-1}\mathbf{b}$ is an $\Theta(n)$ problem, as we are only multiplying the diagonal entries of the matrix. Should we use associativity to calculate $\mathbf{M}_{\text{diag}}^{-1}\left(\mathbf{M}_{\text{off}}\mathbf{x}\right) = \left(\mathbf{M}_{\text{diag}}^{-1}\mathbf{M}_{\text{off}}\right)\mathbf{x}$? That is, should we find $\mathbf{M}_{\text{diag}}^{-1}\mathbf{M}_{\text{off}}$ like we found $\mathbf{M}_{\text{diag}}^{-1}\mathbf{b}$? In this case, it's actually detrimental to do so. In most case, the number of iterations will be significantly less than *n*, but finding $\mathbf{M}_{\text{diag}}^{-1}\mathbf{M}_{\text{off}}$ is an $\Theta(n^2)$ problem. Therefore, it is less work overall to calculate $\mathbf{M}_{\text{off}}\mathbf{x}_k$ with each step and then perform the $\Theta(n)$ step of multiply that result by $\mathbf{M}_{\text{diag}}^{-1}$.

In Matlab, finding these matrices is quite straight-forward:

```
M = […];
M_diag = diag( diag( M ) );
M_off = M - M_diag;
M_diag_inv = M_diag^-1;
M_diag_inv_b = M_diag_inv * b;   % matrix-vector multiplication
```

We can, however, do this more efficiently:

```
M = […];
m_diag = diag( M );
M_off = M – diag( m_diag );
m_diag_inv = m_diag.^-1;
M_diag_inv_b = m_diag_inv .* b;   % element-by-element multiplication
```

Next, we calculate

```
x = M_diag_inv_b;

for k = 1:n
    x_old = x;

    x = M_diag_inv_b - m_diag_inv .* (M_off * x);

    if norm( x - x_old ) < 1e-5
        return;
    end
end

throw …
```

15.2.2 The Gauss-Seidel method

Note that when we use the Jacobi method, if we examine the Matlab code,

```
x = M_diag_inv_b - m_diag_inv .* (M_off * x);
```

we calculate `m_diag_inv .* (M_off * x)` first, and only then do we perform the vector-vector subtraction. Visually,

$$\mathbf{x}_2 = \mathbf{M}_{\text{diag}}^{-1}\mathbf{b} - \mathbf{M}_{\text{diag}}^{-1}\left(\mathbf{M}_{\text{off}}\mathbf{x}_1\right)$$

$$\mathbf{x}_2 = \begin{pmatrix} 0.5161 \\ 0.8644 \\ -0.9247 \\ -0.4667 \end{pmatrix} - \begin{pmatrix} 0.1613 & 0 & 0 & 0 \\ 0 & 0.1695 & 0 & 0 \\ 0 & 0 & 0.1075 & 0 \\ 0 & 0 & 0 & 0.0952 \end{pmatrix} \begin{pmatrix} 0 & 1.2 & 0.5 & 0.1 \\ -2.1 & 0 & 1.8 & 0.4 \\ 0.9 & 1.7 & 0 & 2.1 \\ 0.2 & 0.8 & 1.6 & 0 \end{pmatrix} \begin{pmatrix} 0.4309 \\ 1.3619 \\ -1.0273 \\ -0.4014 \end{pmatrix}$$

$$= \begin{pmatrix} 0.5161 \\ 0.8644 \\ -0.9247 \\ -0.4667 \end{pmatrix} - \begin{pmatrix} 0.1613 & 0 & 0 & 0 \\ 0 & 0.1695 & 0 & 0 \\ 0 & 0 & 0.1075 & 0 \\ 0 & 0 & 0 & 0.0952 \end{pmatrix} \begin{pmatrix} 1.0804 \\ -2.9147 \\ 1.8600 \\ -0.4680 \end{pmatrix}$$

$$= \begin{pmatrix} 0.5161 \\ 0.8644 \\ -0.9247 \\ -0.4667 \end{pmatrix} - \begin{pmatrix} 0.1743 \\ -0.4940 \\ 0.2000 \\ 0.0446 \end{pmatrix}$$

$$= \begin{pmatrix} 0.3419 \\ 1.3584 \\ -1.1247 \\ -0.4221 \end{pmatrix}$$

We don't necessarily have to do this, however. We could, instead, calculate each entry of $\mathbf{x}_2$ serially. First, calculate the first entry of $\mathbf{x}_2$:

$$\begin{pmatrix} 0.3419 \\ ? \\ ? \\ ? \end{pmatrix} = \begin{pmatrix} 0.5161 \\ 0.8644 \\ -0.9247 \\ -0.4667 \end{pmatrix} - \begin{pmatrix} 0.1613 & 0 & 0 & 0 \\ 0 & 0.1695 & 0 & 0 \\ 0 & 0 & 0.1075 & 0 \\ 0 & 0 & 0 & 0.0952 \end{pmatrix} \begin{pmatrix} 0 & 1.2 & 0.5 & 0.1 \\ -2.1 & 0 & 1.8 & 0.4 \\ 0.9 & 1.7 & 0 & 2.1 \\ 0.2 & 0.8 & 1.6 & 0 \end{pmatrix} \begin{pmatrix} 0.4309 \\ 1.3619 \\ -1.0273 \\ -0.4014 \end{pmatrix}$$

Next, we could calculate the second entry of $\mathbf{x}_2$:

$$\begin{pmatrix} 0.3419 \\ 1.3584 \\ ? \\ ? \end{pmatrix} = \begin{pmatrix} 0.5161 \\ 0.8644 \\ -0.9247 \\ -0.4667 \end{pmatrix} - \begin{pmatrix} 0.1613 & 0 & 0 & 0 \\ 0 & 0.1695 & 0 & 0 \\ 0 & 0 & 0.1075 & 0 \\ 0 & 0 & 0 & 0.0952 \end{pmatrix} \begin{pmatrix} 0 & 1.2 & 0.5 & 0.1 \\ -2.1 & 0 & 1.8 & 0.4 \\ 0.9 & 1.7 & 0 & 2.1 \\ 0.2 & 0.8 & 1.6 & 0 \end{pmatrix} \begin{pmatrix} 0.4309 \\ 1.3619 \\ -1.0273 \\ -0.4014 \end{pmatrix}$$

Notice, however, we have already calculated a better value of the first entry: if this technique is actually converging, then 0.3419 should be a better approximation of the first value than 0.4309. Would it not be a good idea, therefore, to use the better approximation? Now, when we calculate the next value, we get 1.3267:

$$\begin{pmatrix} 0.3419 \\ 1.3267 \\ ? \\ ? \end{pmatrix} = \begin{pmatrix} 0.5161 \\ 0.8644 \\ -0.9247 \\ -0.4667 \end{pmatrix} - \begin{pmatrix} 0.1613 & 0 & 0 & 0 \\ 0 & 0.1695 & 0 & 0 \\ 0 & 0 & 0.1075 & 0 \\ 0 & 0 & 0 & 0.0952 \end{pmatrix} \begin{pmatrix} 0 & 1.2 & 0.5 & 0.1 \\ -2.1 & 0 & 1.8 & 0.4 \\ 0.9 & 1.7 & 0 & 2.1 \\ 0.2 & 0.8 & 1.6 & 0 \end{pmatrix} \begin{pmatrix} 0.3419 \\ 1.3619 \\ -1.0273 \\ -0.4014 \end{pmatrix}$$

Now we can use both these values to calculate the third entry:

$$\begin{pmatrix} 0.3419 \\ 1.3267 \\ -1.1097 \\ ? \end{pmatrix} = \begin{pmatrix} 0.5161 \\ 0.8644 \\ -0.9247 \\ -0.4667 \end{pmatrix} - \begin{pmatrix} 0.1613 & 0 & 0 & 0 \\ 0 & 0.1695 & 0 & 0 \\ 0 & 0 & 0.1075 & 0 \\ 0 & 0 & 0 & 0.0952 \end{pmatrix} \begin{pmatrix} 0 & 1.2 & 0.5 & 0.1 \\ -2.1 & 0 & 1.8 & 0.4 \\ 0.9 & 1.7 & 0 & 2.1 \\ 0.2 & 0.8 & 1.6 & 0 \end{pmatrix} \begin{pmatrix} 0.3419 \\ 1.3267 \\ -1.0273 \\ -0.4014 \end{pmatrix}$$

and finally, calculating the fourth entry:

$$\begin{pmatrix} 0.3419 \\ 1.3267 \\ -1.1097 \\ -0.4052 \end{pmatrix} = \begin{pmatrix} 0.5161 \\ 0.8644 \\ -0.9247 \\ -0.4667 \end{pmatrix} - \begin{pmatrix} 0.1613 & 0 & 0 & 0 \\ 0 & 0.1695 & 0 & 0 \\ 0 & 0 & 0.1075 & 0 \\ 0 & 0 & 0 & 0.0952 \end{pmatrix} \begin{pmatrix} 0 & 1.2 & 0.5 & 0.1 \\ -2.1 & 0 & 1.8 & 0.4 \\ 0.9 & 1.7 & 0 & 2.1 \\ 0.2 & 0.8 & 1.6 & 0 \end{pmatrix} \begin{pmatrix} 0.3419 \\ 1.3267 \\ -1.1097 \\ -0.4014 \end{pmatrix}.$$

Notice that we haven't done any more work, we have performed the exact same operations, only we have used a more precise approximation. Compare the two approximations

$$\mathbf{x}_2 = \begin{pmatrix} 0.3419 \\ 1.3584 \\ -1.1247 \\ -0.4221 \end{pmatrix} \text{ using the Jacobi method and } \mathbf{x}_2 = \begin{pmatrix} 0.3419 \\ 1.3267 \\ 1.1097 \\ -0.4052 \end{pmatrix} \text{ using the Gauss-Seidel method.}$$

With the Jacobi method $\|\mathbf{x}_2 - \mathbf{x}\|_2 \approx 0.02006$, but with the Gauss-Seidel method, $\|\mathbf{x}_2 - \mathbf{x}\|_2 \approx 0.008910$, and thus, with this simple change, the error is cut by a factor of over two. If we were to apply this at each step, including the first, we now have

$$\mathbf{x}_0 = \begin{pmatrix} 0.51612903 \\ 0.86440678 \\ -0.92473118 \\ -0.46666667 \end{pmatrix}, \mathbf{x}_1 = \begin{pmatrix} 0.43092647 \\ 1.33154718 \\ -1.10445850 \\ -0.40802757 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 0.35406055 \\ 1.35504466 \\ -1.11455618 \\ -0.40681504 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 0.35043074 \\ 1.35670724 \\ -1.11477069 \\ -0.40683754 \end{pmatrix},$$

$$\mathbf{x}_4 = \begin{pmatrix} 0.35000329 \\ 1.35666596 \\ -1.11472863 \\ -0.40683501 \end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix} 0.35000785 \\ 1.35665458 \\ -1.11472756 \\ -0.40683439 \end{pmatrix}, \mathbf{x}_6 = \begin{pmatrix} 0.35000996 \\ 1.35665496 \\ -1.11472798 \\ -0.40683440 \end{pmatrix}$$

where the actual solution is

$$\mathbf{x} = \begin{pmatrix} 0.35000989 \\ 1.35665507 \\ -1.11472799 \\ -0.40683441 \end{pmatrix}$$

so, after six iterations, $\|\mathbf{x}_6 - \mathbf{x}\|_2 \approx 0.00000012512 = 1.2612 \times 10^{-7}$. Recall that with the Jacobi method the error after six iterations was 1.1846×10^{-4}, and thus the Gauss-Seidel method has appreciably faster convergence.

16 Vector and sparse matrix data structures

Here, the "function" `MIN` is a macro that returns the minimum of the two arguments. A corresponding macro is `MAX`. The word *function* is in quotes as they are not truly functions, but rather substitutions.

```
#define MIN( a, b ) (((a) < (b)) ? (a) : (b))
#define MAX( a, b ) (((a) < (b)) ? (b) : (a))
```

16.1 Data structures for reduced accumulated error

Suppose you are trying to add three numbers, but at any step, you can only remember four significant digits of any one number. Thus, for example, you may be trying to add 93.49 + 6.560 + 0.05, which has a correct answer of 100.1. Suppose, however, we perform each addition and round each addition to four digits first:

$$\underbrace{(93.49 + 6.560)}_{100.05} + 0.05 \qquad\qquad 93.49 + \underbrace{(6.560 + 0.05000)}_{6.610}$$

$$100 \quad + 0.05 = 100.05 = 100 \qquad 93.49 + \quad 6.610 \quad = 100.1$$

Thus, in one order, we get exactly the correct answer, and in the other, we have an answer that has a relative error of 0.1%.

```
 typedef struct {
    double summand;
    double correction;
} positive_sum_t;

void positive_sum_init( positive_sum_t *sum, double initial_value ) {
    assert( initial_value >= 0.0d );

    sum->summand = initial_value;
    sum->correction = 0.0d;
}

double positive_sum_value( positive_sum_t *sum ) {
    return sum->summand;
}

void positive_sum( positive_sum_t *sum, double addend ) {
    double corrected_smaller, corrected_sum;

    assert( addend >= 0.0d );

    if ( sum->summand < addend ) {
        corrected_smaller = sum->summand - sum->correction;
        corrected_sum = addend + corrected_smaller;
        sum->correction = (corrected_sum - addend) - corrected_smaller;
    } else {
        corrected_smaller = addend - sum->correction;
        corrected_sum = sum->summand + corrected_smaller;
        sum->correction = (corrected_sum - sum->summand) - corrected_smaller;
    }

    sum->summand = corrected_sum;
}
```

16.2 Vector data structures

```
typedef struct {
    size_t dimension;
    double *a_values;
} double_t;
```

To initialize such a data structure, we require the dimension of the vector to be generated and whether or not the values should be

1. left uninitialized leaving it up to the user to assign them,
2. all the same value, or
3. copy the values from an existing vector.

These three functions are shown here. In each case, the intended dimension is saved and there is a request for a *m*emory *alloc*ation. This is a request to the operating system for an appropriate block of memory; in this case, we are requesting 8*n* bytes, as the size of a `double` is 8 bytes. In the third case, the dimensions of the two vectors may differ, in which case, only those values that can be copied are copied, and if there are extra entries, those are set to zero.

```
void vector_init( vector_t *this, size_t n ) {
    this->dimension = n;
    this->a_values = (double *)malloc( n*sizeof( double ) );
}

void vector_init_scalar( vector_t *this, size_t n, double value ) {
    size_t i;

    this->dimension = n;
    this->a_values = (double *)malloc( n*sizeof( double ) );

    for ( i = 0; i < n; ++i ) {
        this->a_values[i] = value;
    }
}

void vector_init_vector( vector_t *this, size_t n, vector_t *u ) {
    size_t i;

    this->dimension = n;
    this->a_values = (double *)malloc( n*sizeof( double ) );

    for ( i = 0; i < n; ++i ) {
        this->a_values[i] = (i < u->dimension) ? u->a_values[i] : 0.0d;
    }
}
```

The calls to `malloc` explicitly request memory from the operating system, and if we no longer require that memory, it is necessary to indicate to the operating system that we no longer need it, and that the operating system can use it again for some other purpose; that is, we must *free* the memory.

```
void vector_destroy( vector_t *this ) {
    free( this->a_values );
    this->a_values = NULL;
}
```

We now require three functions that allow the user to

1. to access the dimension of the vector,
2. get a specific entry in the vector,
3. set a specific entry to a specific value,
4. assign all entries a specific value, and
5. assign all entries those of another vector, padding with zero, if necessary.

These are implemented here:

```
size_t vector_dimension( vector_t *this ) {
    return this->dimension;
}

double vector_get( vector_t *this, size_t i ) {
    assert( i < this->dimension );
    return this->a_values[i];
}

void vector_set( vector_t *this, size_t i, double value ) {
    assert( this->dimension > i );
    this->a_values[i] = value;
}

void vector_assign_value( vector_t *this, double value ) {
    size_t i;

    for ( i = 0; i < this->dimension; ++i ) {
        this->a_values[i] = value;
    }
}

void vector_assign_vector( vector_t *this, vector_t *u ) {
    size_t i;

    for ( i = 0; i < this->dimension; ++i ) {
        this->a_values[i] = (i < u->dimension) ? u->a_values[i] : 0.0d;
    }
}
```

Finally, we get to those operations that are related to vector operations:

1. multiplying each entry in a vector by a given scalar,
2. adding to each entry of a vector a given scalar,
3. adding elementwise to each entry of a vector the entries of a second vector, and
4. adding elementwise to each entry of a vector the entries of a second vector multiplied by a scalar.

These are implemented here:

```
void vector_scalar_multiply( vector_t *this, double value ) {
    size_t i;

    for ( i = 0; i < this->dimension; ++i ) {
        this->a_values[i] *= value;
    }
}

void vector_scalar_add( vector_t *this, double value ) {
    size_t i;

    for ( i = 0; i < this->dimension; ++i ) {
        this->a_values[i] += value;
    }
}

void vector_vector_add( vector_t *this, vector_t *u ) {
    size_t i;

    assert( this->dimension == u->dimension );

    for ( i = 0; i < this->dimension; ++i ) {
        this->a_values[i] += u->a_values[i];
    }
}

void vector_vector_scalar_add( vector_t *this, vector_t *u, double value ) {
    size_t i;

    for ( i = 0; i < this->dimension; ++i ) {
        this->a_values[i] += value*u->a_values[i];
    }
}
```

ow, well beyond the scope of this course is the issue that floating-point numbers are only approximations of real numbers, and therefore floating-point operations are only approximations of actual arithmetic operations. In elementary school, you learned that you could add any three numbers in any order you wanted and you would always come up with the same answer. For example, in Matlab, you may try the following:

```
>> a = 0.00000006557406991565868;
>> b =  7922073.295595544;
>> c = -7922073.295595449;
>> (a + b) + c
       ans =
            1.601874828338623e-07
>> a + (b + c)
       ans =
            1.605689725264375e-07
```

The correct answer, to all significant digits, is $1.6057406991565868\times10^{-7}$, so comparing these three:

Calculation	Calculated value	Relative error
$(a + b) + c$	$1.601874828338623\times10^{-7}$	0.0032 %
$a + (b + c)$	$1.605689725264375\times10^{-7}$	0.2408 %

As a second example, consider adding the following 8 random numbers:

0.29140577571371085 + 0.13718742197472822 + 0.18893346303136238 + 0.95661705112294571
+ 0.29661309204132535 + 0.23720287531730688 + 0.66015752767331148 + 0.00429596790761244130

There are 8! = 40320 different permutations of these numbers, indicating as many different means of ordering them. These in turn give five different answers, with different frequencies:

Answer	Frequency
2.7724131747823025	2196
2.7724131747823030	9848
2.7724131747823035	26736
2.7724131747823040	1492
2.7724131747823045	48

But which, if any, is the correct answer? To determine this, we really need to go to binary and add the numbers using the binary representation:

$$\begin{aligned}
&0.010010101001100110010001101001001000111001001111011 01\\
&0.001000110001111010110111000000101100110111010011100 10\\
&0.001100000101110111111000110000100110101101111100001 010\\
&0.111101001110010011011010111001010101111001110111100 10\\
&0.010010111110111011010101111010011110001000000111101 11\\
&0.001111001011100101010011111000000000000101000111100 11\\
&0.101010010000000000010101011011111001000010101001100 11\\
+\ &\underline{0.000000010001100110001010011000011010101011111000111 10}\\
&10.110001011011110011011110101011001011000010000100101 10
\end{aligned}$$

The last number is exactly in-between the second and third numbers, which is hopeful, as this suggests that usually floating-point operations tend to give a reasonable answer. According to the rules of floating-point operations, this

number should be rounded to the higher of the two numbers, 2.7724131747823035, which fortunately is also the most common. Consider, however, what happens if you were to add hundreds or thousands of numbers. In such a case, the answers could vary quite widely, especially if the magnitude of the numbers are significantly different. Fortunately, William Kahan, the same person who chaired the committee that drafted the IEEE 754 floating-point standard also devised an algorithm that can add arbitrarily many positive floating-point numbers without introducing more error than would be introduced if two numbers were added. In Matlab, the Kahan algorithm for adding a vector of all positive numbers is

```
function [sum] = kahan( vector )
    assert( all( vector >= 0 ) );

    sum        = 0.0;
    correction = 0.0;

    for addend = vector
        // Ensure that addend is the smaller of the two
        //  - this will usually be the case, but just in case...
        if addend > sum
            tmp = addend;
            addend = sum;
            sum = tmp;
        end

        // Add the correction factor (initially zero) to the addend
        corrected_addend = addend - correction;

        // Calculate the new sum by adding the corrected addend to the sum
        corrected_sum = sum + corrected_addend;

        // We just calculated A = B + C, so (A – B) – C should be zero,
        // but due to floating-point operations, there might be a small error
        //  - store this error to be added to the next sum
        correction = (corrected_sum - sum) - corrected_addend;

        // Update the sum to the 'sum + corrected_addend'
        sum = corrected_sum;
    end
end
```

When this algorithm is applied to all 40320 permutations of the given eight digits, all calculate to the same and most correct value.

For our code here, we don't want to create an entirely new array of entries to be added, so the following adds $x + y$ and updates the argument `correction` to store any correcting factor:

```
double kahan_summation( double x, double y, double *correction ) {
    double corrected_addend, corrected_sum;

    if ( x < y ) {
        corrected_addend = x - *correction;
        corrected_sum = y + corrected_addend;
        *correction = (corrected_sum - y) - corrected_addend;
    } else {
        corrected_addend = y - *correction;
        corrected_sum = x + corrected_addend;
        *correction = (corrected_sum - x) - corrected_addend;
    }
```

```
    return corrected_sum;
}
```

To use this code, suppose we have an array of eight positive values for which we wish to calculate both the mean and the standard deviation.

```
double array[8] = {
    0.29140577571371085, 0.13718742197472822, 0.18893346303136238, 0.95661705112294571,
    0.29661309204132535,           0.23720287531730688,           0.66015752767331148,
0.00429596790761244130
};

double mean, standard_deviation, correction;

mean       = 0.0d;
correction = 0.0d;

for ( i = 0; i < 8; ++i ) {
    mean = kahan_summation( mean, array[i], &correction );
}

mean /= 8.0;

standard_deviation = 0.0d;
correction         = 0.0d;   // reset to zero

for ( i = 0; i < 8; ++i ) {
    delta = array[i] - mean;
    standard_deviation = kahan_summation( standard_deviation, delta*delta, &correction
);
}

standard_deviation = sqrt( standard_deviation/8.0 );
printf( "%f +/- %f\n", mean, standard_deviation );
```

We will use Kahan's summation algorithm in our implementation of the various norms, the topic which follows next.

The norms of greatest interest to engineers are the 1-, 2- and infinity-norms; however, there are many related concepts such as the mean-squared-error (MSE) and the root-mean-squared-error (RMSE). Additionally, there are many circumstances where one may need to calculate the 2-norm squared. Thus, rather than having numerous functions, we will instead create a single `norm` function that takes a second argument specifying the type of result that is desired. We could use a string, but this would be costly in terms of both memory and processing, and therefore we will opt for an *enumerated type*. Internally, the computer simply replaces these values by integers, but if the user, for example, misspells `TWO_NORM` as `TOW_NORM`, the compiler will flag an error. Strings, on the other hand, are incredibly expensive to compare, relatively speaking. In C, our enumerated type would be:

```
typedef enum {
    ONE_NORM,
    MEAN_ERROR,
    TWO_NORM,
    TWO_NORM_SQUARED,
    ROOT_MEAN_SQUARED_ERROR,
    MEAN_SQUARED_ERROR,
    INFINITY_NORM
} norm_t;
```

Next, we will implement the norm function using a switch statement, where each enumerated type is associated with one of the cases. Then, once the initial calculation is completed, we will divide the result by the number of entries or take the square root, as appropriate. You will note that there is no such concept as the *mean* infinity-norm.

```
double vector_norm( vector_t *this, norm_t n ) {
    size_t i;
    double return_value, value, correction;

    return_value = 0.0d;
    correction = 0.0d;

    switch ( n ) {
        case ONE_NORM:
        case MEAN_ERROR:
            for ( i = 0; i < this->dimension; ++i ) {
                return_value = sorted_sum( return_value, fabs( this->a_values[i] ),
                                           &correction );
            }

            if ( n == MEAN_ERROR ) {
                return_value /= (double)(this->dimension);
            }

            break;
        case TWO_NORM:
        case TWO_NORM_SQUARED:
        case ROOT_MEAN_SQUARED_ERROR:
        case MEAN_SQUARED_ERROR:
            for ( i = 0; i < this->dimension; ++i ) {
                return_value = sorted_sum( return_value,
                                            this->a_values[i] * this->a_values[i],
                                           &correction );
            }

            if ( (n == ROOT_MEAN_SQUARED_ERROR) || (n == MEAN_SQUARED_ERROR) ) {
                return_value /= (double)(this->dimension);
            }

            return_value = ((n == TWO_NORM) || (n == ROOT_MEAN_SQUARED_ERROR)) ?
                                                sqrt( return_value ) : return_value;
            break;
       case INFINITY_NORM:

            for ( i = 0; i < this->dimension; ++i ) {
                value = fabs( this->a_values[i] );

                if ( value > return_value ) {
                    return_value = value;
                }
            }

            break;
        default:
            assert( false );
    }

    return return_value;
}
```

In normalizing a vector, we will restrict the type of norm that can be used, as under no circumstances would one need to calculate, for example, $\frac{\mathbf{u}}{\|\mathbf{u}\|_2^2}$ or $\frac{\mathbf{u}}{RMSE(\mathbf{u})}$.

```
void vector_normalize( vector_t *this, norm_t n ) {
    size_t i;
    double norm = vector_norm( this, n );

    assert( n != MEAN_ERROR );
    assert( n != TWO_NORM_SQUARED );
    assert( n != MEAN_SQUARED_ERROR );
    assert( n != ROOT_MEAN_SQUARED_ERROR );

    if ( norm != 0.0d ) {
        for ( i = 0; i < this->dimension; ++i ) {
            this->a_values[i] /= norm;
        }
    }
}
```

To calculate the distance between two vectors **u** and **v**, your first thought may be to calculate the vector difference **u** – **v** and then calculate the norm of the result, but this is potentially expensive.

```
double vector_distance( vector_t *u, vector_t *v norm_t n ) {
    size_t i;
    double return_value, delta, correction;

    assert( u->dimension == v.dimension );

    return_value = 0.0d;
    correction = 0.0d;

    switch ( n ) {
        case ONE_NORM:
        case MEAN_ERROR:
            for ( i = 0; i < u->dimension; ++i ) {
                return_value = sorted_sum( return_value,
                                           fabs( u->a_values[i] – v->a_values[i] ),
                                          &correction );
            }

            if ( n == MEAN_ERROR ) {
                return_value /= (double)(u->dimension);
            }

            break;
        case TWO_NORM:
        case TWO_NORM_SQUARED:
        case ROOT_MEAN_SQUARED_ERROR:
        case MEAN_SQUARED_ERROR:
            for ( i = 0; i < u->dimension; ++i ) {
                delta = u->a_values[i] – v->a_values[i];

                return_value = sorted_sum( return_value, delta*delta, &correction );
            }

            if ( (n == ROOT_MEAN_SQUARED_ERROR) || (n == MEAN_SQUARED_ERROR) ) {
                return_value /= (double)(u->dimension);
            }

            return_value = ( (n == TWO_NORM) || (n == ROOT_MEAN_SQUARED_ERROR) ) ?
                                                    sqrt( return_value ) : return_value;
            break;
       case INFINITY_NORM:

            for ( i = 0; i < u->dimension; ++i ) {
                delta = fabs( u->a_values[i] – v->a_values[i] );

                if ( delta > return_value ) {
                    return_value = delta;
                }
            }

            break;
        default:
            assert( false );
    }

    return return_value;
}
```

Next we will implement the inner product, projections and perpendicularizations. For the first, we will implement a generic function that takes two arguments, but for the other two, we will project the current vector onto another vector. Thus, this reduces the memory requirements if all that is required is the projection of one vector onto another. Because the products of the various values in **u** and **v** may be either positive or negative, the Kahan algorithm is less likely to make a significant impact, and thus we will refrain from using it.

```
double inner_product( vector_t *u, vector_t *v ) {
    size_t i;
    mixed_sum_t sum;

    assert( u->dimension == v->dimension );

    mixed_sum_init( &sum, 0.0d );

    for ( i = 0; i < u->dimension; ++i ) {
        mixed_sum( &sum, u->a_values[i]*v->a_values[i] );
    }

    return mixed_sum_value( &sum );
}

void vector_project( vector_t *this, vector_t *u ) {
    size_t i;
    double multiplier = norm( u, TWO_NORM_SQUARED );

    assert( this->dimension == u->dimension );

    if ( multiplier == 0.0d ) {
        for ( i = 0; i < this->dimension; ++i ) {
            this->a_values[i] = 0.0d;
        }
    } else {
        multiplier = inner_product( u, this )/multiplier;

        for ( i = 0; i < this->dimension; ++i ) {
            this->a_values[i] = multiplier*u->a_values[i];
        }
    }
}

void vector_perpendicularize( vector_t *this, vector_t *u ) {
    size_t i;
    double multiplier = norm( u, TWO_NORM_SQUARED );;

    assert( this->dimension == u->dimension );

    if ( multiplier != 0.0d ) {
        multiplier = inner_product( u, this )/multiplier;

        for ( i = 0; i < this->dimension; ++i ) {
            this->a_values[i] -= multiplier*u->a_values[i];
        }
    }
}
```

We can now use these algorithms to perform the Gram-Schmidt algorithm. Assume that `a_vec` is an array of pointers to vectors. In this case, we could implement this algorithm quite easily with no significant overhead memory:

```
// Perform Gram-Schmidt on the array of n vectors
void gram_schmidt( vector_t **a_vec, size_t n ) {
    size_t i, j;

    for ( i = 0; i < n; ++i ) {
        assert( a_vec[i]->dimension == a_vec[0]->dimsions );

        for ( j = 0; j < i; ++j ) {
            vector_perpendicularize( a_vec[i], a_vec[j] );
        }

        normalize( a_vec[i], TWO );
    }
}
```

As you can see, the Gram Schmidt algorithm is much easier to implement in C than it is to execute by hand.

Finally, we will discuss a function for printing vectors. To display a vector, it makes the most sense to display it in a language already familiar to the user. We will use Maple, Matlab, C and Mathematical, and for vectors, the last two are the same. Thus, we require another enumerated type, and a function that takes three strings as an argument and prints the relevant one based on the selection `language`.

```
typedef enum {
    MAPLE,
    MATLAB,
    C,
    MATHEMATICA
} output_t;

int vector_print_language( char *maple, char *matlab, char *c_mathematica,
                           output_t language ) {
    int return_value = 0;

    switch( language ) {
        case MAPLE:
            return_value = printf( maple );
            break;
        case MATLAB:
            return_value = printf( matlab );
            break;
        case C:
        case MATHEMATICA:
            return_value = printf( c_mathematica );
            break;
        default:
            assert( false );
    }

    return return_value;
}
```

Now, in these four languages, the vector $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ would be created as:

Language	Format
Maple	`<1, 2, 3>;`

Matlab	`[1`
	`2`
	`3];`

C	`{1, 2, 3};`
Mathematica	`{1, 2, 3};`

Another issue is how are numbers to be printed. For this, we require the user to pass a formatting string. This will usually be of the form `"%f"` or `"%.17f"` for decimal floating-point, or `"%e"` or `"%.17e"` for scientific notation. The last parameter specifies the language. Now, to parallel the behavior of `printf`, we will return the total number of characters printed to the screen. This is calculated by accumulating all of the return values of the various calls to `printf`.

```
int vector_printf( vector_t *this, char *str, output_t language ) {
    int return_value;
    size_t i;

    if ( this->dimension == 0 ) {
        return_value = vector_print_language( "<NULL>:", "[];", "{};", language );
    } else {
        return_value = vector_print_language( "<", "[", "{", language );
        return_value += printf( str, this->a_values[0] );

        for ( i = 1; i < this->dimension; ++i ) {
            return_value += vector_print_language( ", ", "\n ", ", ", language );
            return_value += printf( str, this->a_values[i] );
        }

        return_value += vector_print_language( ">:", "];", "};", language );
    }

    return return_value;
}
```

Having completed our implementation of vectors, we will now proceed to look at an implementation of sparse matrices in C.

16.3 Sparse matrix data structures

To this point, we have discussed matrices, but you must also implement these matrices in programming languages like MATLAB and Python. Most of the matrices in this course have been *dense*, meaning that most of the values are non-zero. In many engineering applications, this is, however, simply not the case, and in many engineering applications, the dimensions of the matrix are, relative to examples we have used in this course, astronomically large. It is not unusual for a simulation of a circuit such as the AMD G4 to require a matrix representing a million linear equations in a million unknowns. Each value in the matrix represents the connection of one circuit element to other circuit elements, in which case, he number of connections (and therefore the number of non-zero values per row) is usually restricted to a reasonably small number like five. Similarly, if we are using finite difference methods or finite elements to simulate the behavior of an object such as a fender of a vehicle or the wing of an airplane or a graphene surface, the interactions of one element or one atom are restricted to the connections to surrounding elements or atoms and thus limiting the number of non-zero values per row to a more reasonable value such as below 10.

Now, the amount of memory required for a dense matrix that is one million by one million is 7.3 TiB, significantly beyond the memory of most computers. The memory occupied by non-zero values if seven values per row are non-zero is 54 MiB. Most matrix-matrix and matrix-vector operations, including operations involved in finding approximations to systems of linear equations, are indifferent to values that are zero, and thus, many of these operations can be reduced in run time to one dependent only on the non-zero values.

There are many possible representations of sparse matrices, and rather than discuss the benefits and drawbacks of many of the more sub-optimal designs, we will focus instead on the *new Yale* sparse matrix format. Generally, such matrices are square, and we will leave the details for storing non-square matrices as an exercise to the reader.

Programming languages such as C and Python index arrays of size m with indices from 0 to $m - 1$. There are many good reasons for doing this, however, these reasons are beyond the scope of this course. We will assume we are using such a programming language.

Given an $n \times n$ matrix A with at most N non-zero off-diagonal values, create three arrays:

1. an array `a_row_index` of unsigned integers of size $n + 1$,
2. an array `a_column_index` of unsigned integers of size N,
3. an array `a_diagonal_values` of double of size n, and
4. an array `a_off_diagonal_values` of double of size N.

In C, this could be done as

```
typedef struct {
    size_t dimension;
    size_t max_non_zero_off_diagonal_entries;
    size_t *a_row_index;
    size_t *a_column_index;
    double *a_diagonal_values;
    double *a_off_diagonal_values;
} sparse_matrix_t;
```

To initialize such a data structure, one would pass this to a function such as

```
void sparse_matrix_init( sparse_matrix_t *this, size_t n, size_t N ) {
    size_t i;

    this->dimension = n;
    this->max_non_zero_off_diagonal_entries = N;
    this->a_row_indices         = (size_t *)malloc( (n + 1)*sizeof( size_t ) );
    this->a_column_indices      = (size_t *)malloc(    N   *sizeof( size_t ) );
    this->a_diagonal_values     = (double *)malloc(    n   *sizeof( double ) );
    this->a_off_diagonal_values = (double *)malloc(    N   *sizeof( double ) );

    for ( i = 0; i < n; ++i ) {
        this->a_row_indices[i] = 0;
        this->a_digaonal_values[i] = 0.0d;
    }

    this->a_row_indices[n] = 0;
}
```

When such a data structure is no longer required, we must deallocate the memory stored internally:

```
void sparse_matrix_destroy( sparse_matrix_t *this ) {
    // Free the dynamically allocated memory of each of the four arrays
    free( this->a_row_indices );
    free( this->a_column_indices );
    free( this->a_digaonal_values );
    free( this->a_off_digaonal_values );

    // Set the pointers to zero so that they are not accidentally reused
    this->a_row_indices         = NULL;
    this->a_column_indices      = NULL;
    this->a_diagonal_values     = NULL;
    this->a_off_diagonal_values = NULL;
}
```

To use such a data structure, we would

```
sparse_matrix_t *p_matrix = (sparse_matrix_t *)malloc( sizeof( sparse_matrix_t ) );
sparse_matrix_init( p_matrix, 1000, 5000 );

// Use the matrix...

sparse_matrix_destroy( p_matrix );
free( p_matrix );
```

It is assumed that all diagonal values are non-zero, and thus, one array is always allocated to store each diagonal value. Off-diagonal values are stored as follows: The non-zero off-diagonal values are stored in row order in the array `a_off_diagonal_values`. We will demonstrate this with an example.

For example, suppose we are storing the 10 × 10 array

$$A = \begin{pmatrix} 0.0 & 0 & 0 & 0 & 0 & 0 & 0.6 & 0 & 0 & 0.9 \\ 0 & 1.1 & 0 & 1.3 & 0 & 0 & 0 & 0 & 0 & 0 \\ -2.0 & 0 & 2.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.3 & 0 & 3.5 & 0 & 0 & 0 & 3.9 \\ 0 & 0 & 0 & 0 & 4.4 & 0 & 4.6 & 0 & 0 & 0 \\ -5.0 & 0 & 0 & -5.3 & 0 & 5.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 6.6 & 0 & 6.8 & 0 \\ 0 & 0 & -7.2 & 0 & 0 & 0 & 0 & 7.7 & 0 & 0 \\ -8.0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8.8 & 0 \\ 0 & -9.1 & 0 & 0 & -9.4 & 0 & 0 & -9.7 & 0 & 9.9 \end{pmatrix}$$

and we know that at most 20 off-diagonal values are non-zero. To help you associate these non-zero values and where they appear in the corresponding arrays, note that the value is made of the row and column concatenated to create a floating-point number, and all values below the diagonal are negative. Off-diagonal values in even-numbered rows (0, 2, 4, 6 and 8) are shown in blue, while those in odd-numbered rows (1, 3, 5, 7 and 9) are shown in red. We will use the font `Consolas` for array indices and Times New Roman for matrix values. To avoid confusion, we will return to "**values**" in the matrix and "**entries**" in the four arrays.

Thus, the arrays `a_diagonal_values` and `a_off_diagonal_values` store the entries of the matrix:

`a_diagonal_values`

0	1	2	3	4	5	6	7	8	9
0.0	1.1	2.2	3.3	4.4	5.5	6.6	7.7	8.8	9.9

`a_off_diagonal_values`

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0.6	0.9	1.3	–2.0	3.5	3.9	4.6	–5.0	–5.3	6.8	–7.2	7.7	–8.0	–9.1	–9.4	–9.9				

Sixteen of the 20 possible off-diagonal entries in the second array are used and the remaining four positions are left vacant. The next array, `a_column_indices`, stores the column in which the corresponding value in `a_off_diagonal_entries` appears in the matrix.

`a_column_indices`

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
6	9	3	0	5	9	6	0	3	8	2	7	0	1	4	9				

Thus, for example, the value 1.3 appears in entry 2 of `a_off_diagonal_values` and thee column it is in is stored in entry 2 of `a_column_indices`, and in this case, `a_column_indices[2]` is 3.

What differentiates the new Yale format from others is the array `a_row_indices`. Here, each entry stores the index of the first value of each row in the arrays `a_column_indices` and `a_off_diagonal_entries`. The last entry contains the index one beyond the last non-zero entry, in this case, 16.

`a_row_indices`

0	1	2	3	4	5	6	7	8	9	10
0	**2**	**3**	**4**	**6**	**7**	**9**	**10**	**12**	**13**	**16**

For example, the first non-zero value in Row 5 appears in the entry `a_off_diagonal_entries[5]`, which is `7`, and the first non-zero value in Row 8 appears in the entry `a_off_diagonal_entries[8]`, which is `12`. All the non-zero values in Row *m* must appear in the entries from `a_off_diagonal_entries[m]` to `a_off_diagonal_entries[m + 1] - 1`.

Thus, when trying to access an entry in the matrix, there are three possibilities:

1. the entry is on the diagonal,
2. the entry is off the diagonal and non-zero, or
3. the entry is off the diagonal and zero.

First, suppose now we want to access the value at (3, 3) in the matrix. Because it is on the diagonal, we simply return the entry `a_diagonal_entries[3]`.

Second, suppose we want to access the value at (3, 5) in the matrix. It is off the diagonal, and we note that

1. `a_row_indices[3]` is assigned **`4`**, and
2. `a_row_indices[3 + 1]` is assigned **`6`**.

Therefore, there are `6 - 4 = 2` non-zero off-diagonal values in the 3rd row of the matrix, and these two values appear in entries `4` to `6 – 1` in array `a_off_diagonal_entries`. We can now check the two corresponding entries in the column index array: `a_column_indices[4]` is assigned `5`, and we are looking for the value in the 5th column, so we are done. We return `a_off_diagonal_entries[4]`.

Third, suppose we want to access the value at (9,6) in the matrix. It is off the diagonal, and we note

1. `a_row_indices[9]` is assigned **`13`**, and
2. `a_row_indices[9 + 1]` is assigned **`16`**.

Therefore, there are `16 - 13 = 3` non-zero off-diagonal values in the 9th row of the matrix, and those three values appear in entries `13` to `16 - 1` in the array `a_off_diagonal_entries`. We can now check the three corresponding entries in the column-index array: none of `a_column_indices[13]`, `a_column_indices[14]` or `a_column_indices[15]` are assigned 6, so the value at (9, 6) must be zero, which we return.

We can program this:

```
double sparse_matrix_get( sparse_matrix_t *this, size_t i, size_t j ) {
    size_t first, last, k;
    double return_value;

    assert( this->dimension > i );
    assert( this->dimension > j );

    if ( i == j ) {
        return_value = this->a_diagonal_values[i];
    } else {
        first = this->a_row_indices[i];
        last  = this->a_row_indices[i + 1];

        for ( k = first; k < last; ++k ) {
            if ( this->a_column_indices[k] >= j ) {
                return_value = ( this->a_column_indices[k] == j ) ?
                                        this->a_off_diagonal_values[k] : 0.0d;
                break;
            }
        }
    }

    return return_value;
}
```

Now, if we want to change values in the matrix, this becomes much more tedious. Whether we are setting a non-diagonal value in the matrix that was zero to a non-zero value, or setting a non-zero value in the matrix to zero, we must move or update entries in all four arrays. This could potentially result in a lot of effort. The only values that can easily be changed are those on the diagonal. We will, however, implement an *append* function that assumes that any off-diagonal value is being placed after the last existing non-zero value. This assumes that the entries of the array are being assigned in row order.

```
void sparse_matrix_append( sparse_matrix_t *this, size_t i, size_t j, double value ) {
    size_t k;

    assert( this->dimension > i );
    assert( this->dimension > j );
    assert( this->a_row_indices[this->dimension] <= this->max_non_zero_off_diagonal_values
);

    if ( i == j ) {
        this->a_diagonal_values[i] = value;
    } else {
        assert( this->a_row_indices[i + 1] == this->a_row_indices[this->dimension] );
        assert( (this->a_row_indices[i] == this->a_row_indices[i + 1])
             || (this->a_column_indices[this->a_row_indices[i + 1] - 1] < j ) );

        if ( value != 0.0d ) {
            this->     a_column_indices[this->a_row_indices[i + 1]] = j;
            this->a_off_diagonal_values[this->a_row_indices[i + 1]] = value;

            for ( k = i + 1; k <= this->dimension; ++k ) {
                ++( this->a_row_indices[k] );
            }
        }
    }
}
```

Set is much more complex:

```
void sparse_matrix_set( sparse_matrix_t *this, size_t i, size_t j, double value ) {
    size_t k, m;

    assert( this->dimension > i );
    assert( this->dimension > j );

    if ( i == j ) {
        this->a_diagonal_values[i] = value;
    } else {
        for ( k = this->a_row_indices[i];
             (k < this->a_row_indices[i + 1]) && (this->a_column_indices[k] < j);
            ++k ) {
            // do nothing
        }

        if ( k < this->a_row_indices[i + 1] && this->a_column_indices[k] == j ) {
            if ( value == 0.0d ) {
                for ( m = k + 1; m < this->a_row_indices[this->dimension]; ++m ) {
                    this->a_column_indices[m - 1]      = this->a_column_indices[m];
                    this->a_off_diagonal_values[m - 1] = this->a_off_diagonal_values[m];
                }

                for ( m = i + 1; m <= this->dimension; ++m ) {
                    --( this->a_row_indices[m] );
                }
            } else {
                this->a_off_diagonal_values[k] = value;
            }
        } else {
            if ( value != 0.0d ) {
                if ( this->a_row_indices[this->dimension] > 0 ) {
                    for ( m = this->a_row_indices[this->dimension] - 1; m > k; --m ) {
                        this->a_column_indices[m + 1]      = this->a_column_indices[m];
                        this->a_off_diagonal_values[m + 1] = this->a_off_diagonal_values[m];
                    }

                    this->a_column_indices[k + 1]      = this->a_column_indices[k];
                    this->a_off_diagonal_values[k + 1] = this->a_off_diagonal_values[k];
                } else {
                    assert( k == 0 );
                }

                for ( m = i + 1; m <= this->dimension; ++m ) {
                    ++( this->a_row_indices[m] );
                }

                this->a_off_diagonal_values[k] = value;
                this->a_column_indices[k] = j;
            }
        }
    }
}
```

Next, we will implement standard matrix operations. These will include:

1. scalar multiplication αA where if $\alpha = 0$, we eliminate all off-diagonal entries,
2. scalar addition to be interpreted as $A + \alpha Id_n$,
3. matrix-vector addition where the values of the vector are added to the corresponding diagonal entries of the matrix, and
4. matrix-matrix addition where $C = A + B$.

In the first three cases, these operations can be done in-place, but for the last, adding a matrix to an existing matrix already requires $\Theta(n)$ additional memory. Consequently, the matrix to which the sum will be assigned will be passed as a third argument.

```
void sparse_matrix_scalar_multiply( sparse_matrix_t *this, double value ) {
    size_t i;

    for ( i = 0; i < this->dimension; ++i ) {
        this->a_diagonal_values[i] *= value;
    }

    if ( value == 0.0d ) {
        for ( i = 0; i <= this->dimension; ++i ) {
            this->a_row_indices[i] = 0;
        }
    } else {
        for ( i = 0; i < this->a_row_indices[this->dimension]; ++i ) {
            this->a_off_diagonal_values[i] *= value;
        }
    }
}

void sparse_matrix_scalar_add( sparse_matrix_t *this, double value ) {
    size_t i;

    for ( i = 0; i < this->dimension; ++i ) {
        this->a_diagonal_values[i] += value;
    }
}

void sparse_matrix_vector_add( sparse_matrix_t *A, vector_t *u ) {
    size_t i;

    for ( i = 0; i < MIN( A->dimension, u->dimension ); ++i ) {
        A->a_diagonal_values([i] += u->a_values[i];
    }
}
```

```
void sparse_matrix_matrix_add( sparse_matrix_t *A, sparse_matrix_t *B, sparse_matrix_t *C )
{
    size_t i;
    size_t ka, kb, kc;
    double value;

    assert( A->dimension == B->dimension );
    assert( A->dimension == C->dimension );

    sparse_matrix_scalar_multiply( C, 0.0d );

    ka = kb = kc = 0;

    for ( i = 0; i < A->dimension; ++i ) {
        while ( ka < A->a_row_indices[i] && kb < B->a_row_indices[i] ) {
            if ( A->a_column_indices[ka] < A->a_column_indices[kb] ) {
                sparse_matrix_append( C, i, A->a_column_indices[ka],
                                      A->a_off_diagonal_values[ka] );
                ++ka;
            } else if ( A->a_column_indices[ka] == A->a_column_indices[kb] ) {
                value = A->a_off_diagonal_values[ka] + B->a_off_diagonal_values[kb];

                if ( value != 0 ) {
                    sparse_matrix_append( C, i, A->a_column_indices[ka], value );
                }

                ++ka;
                ++kb;
            } else {
                sparse_matrix_append( C, i, B->a_column_indices[kb],
                                      B->a_off_diagonal_values[kb] );
                ++kb;
            }
        }

        while ( ka < A->a_row_indices[i] ) {
            sparse_matrix_append( C, i, A->a_column_indices[ka],
                                    A->a_off_diagonal_values[ka] );
        }

        while ( kb < B->a_row_indices[i] ) {
            sparse_matrix_append( C, i, B->a_column_indices[kb],
                                    B->a_off_diagonal_values[kb] );
        }
    }
}
```

```
bool is_diagonal( sparse_matrix_t *this ) {
    return ( this->a_row_indices[this->dimension] == 0 );
}

bool is_upper_triangular( sparse_matrix_t *this ) {
    size_t i;
    bool return_value = true;

    for ( i = 1; i < this->dimension; ++i ) {
        if (
            this->a_row_indices[i] < this->a_row_indices[i + 1]
            && this->a_column_indices[this->a_row_indices[i]] < i
        ) {
            return_value = false;
            break;
        }
    }

    return return_value;
}

bool is_lower_triangular( sparse_matrix_t *this ) {
    size_t i;
    bool return_value = true;

    for ( i = 1; i < this->dimensions; ++i ) {
        if (
            this->a_row_indices[i - 1] < this->a_row_indices[i]
            this->a_column_indices[this->a_row_indices[i] - 1] > i
        ) {
            return_value = false;
            break;
        }
    }

    return return_value;
}
```

Finally, we will conclude with two printing algorithms. The first prints it currently only in Matlab format, but the second prints the internal data structure for examination.

```
int sparse_matrix_printf( sparse_matrix_t *this, char *str, output_t language ) {
    int return_value;
    size_t i, j, k;

    assert( language == MATLAB );

    if ( this->dimension == 0 ) {
        return_value = printf( "[];" );
    } else {
        k = 0;

        return_value = printf( "[\n " );
        return_value += printf( str, this->a_diagonal_values[0] );

        for ( j = 1; j < this->dimension; ++j ) {
            return_value += printf( " " );

            if ( (k < this->a_row_indices[1]) && (j == this->a_column_indices[k]) ) {
                return_value += printf( str, this->a_off_diagonal_values[k] );
                ++k;
            } else {
                return_value += printf( str, 0.0d );
            }
        }

        for ( i = 1; i < this->dimension; ++i ) {
            return_value += printf( "\n" );

            for ( j = 0; j < this->dimension; ++j ) {
                return_value += printf( " " );

                if ( i == j ) {
                    return_value += printf( str, this->a_diagonal_values[i] );
                } else {
                    if ( (k < this->a_row_indices[i + 1]) &&
                         (j == this->a_column_indices[k]) ) {
                        return_value += printf( str, this->a_off_diagonal_values[k] );
                        ++k;
                    } else {
                        return_value += printf( str, 0.0d );
                    }
                }
            }
        }

        return_value += printf( "];" );
    }

    return return_value;
}
```

```
int sparse_matrix_data_structure_printf( sparse_matrix_t *this, char *str ) {
    size_t i;

    printf( "          Dimension: %d", (int)( this->dimension ) );
    printf( "\n   Maximum non-zero: %d",
                                       (int)(  this->max_non_zero_off_diagonal_values  )
);

    printf( "\n   Diagonal entries:" );

    for ( i = 0; i < this->dimension; ++i ) {
        printf( " " );
        printf( str, this->a_diagonal_values[i] );
    }

    printf( "\n        Row indices:" );

    for ( i = 0; i <= this->dimension; ++i ) {
        printf( " %d", (int)( this->a_row_indices[i] ) );
    }

    printf( "\n     Column indices:" );

    for ( i = 0; i < this->a_row_indices[this->dimension]; ++i ) {
        printf( " %d", (int)( this->a_column_indices[i] ) );
    }

    printf( "\nOff-diagonal values:" );

    for ( i = 0; i < this->a_row_indices[this->dimension]; ++i ) {
        printf( " " );
        printf( str, this->a_off_diagonal_values[i] );
    }

    printf( "\n" );

    return 0;
}
```

Using these two output routines, the following matrix

```
[
 0.00000 0.00000 0.00000 0.41561 0.00000
 0.00000 0.00000 0.00000 0.02332 0.28551
 0.00000 0.00000 0.00000 0.00000 0.00000
 0.00000 0.20081 0.00000 0.92593 0.00000
 0.20956 0.99075 0.58670 0.43912 0.00000]
          Dimension: 5
   Maximum non-zero: 20
   Diagonal entries: 0.00000 0.00000 0.00000 0.92593 0.00000
        Row indices: 0 1 3 3 4 8
     Column indices: 3 3 4 1 0 1 2 3
Off-diagonal values: 0.41561 0.02332 0.28551 0.20081 0.20956 0.99075 0.58670 0.43912
```

16.4 Sparse-matrix–vector operations

Suppose we want to multiply a matrix and a vector. This can now be performed in $\Theta(n + N)$ time, as opposed to the significantly slower $\Theta(n^2)$ time.

```
// u <- A*u
//  - it is not possible to do this with O(1) memory,
//    an additional array of size n must be provided as an argument
void sparse_matrix_vector_multiply( vector_t *this, sparse_matrix_t *A, double *a_tmp ) {
    size_t i, j;

    assert( A->dimension == this->dimension );

    for ( i = 0; i < this->dimension; ++i ) {
        a_tmp[i] = A->a_diagonal_values[i]*this->a_values[i];

            for ( j = A->a_row_indices[i]; j < A->a_row_indices[i + 1]; ++j ) {
            a_tmp [i] +=
                A->a_off_diagonal_values[j]*this->a_values[A->a_column_indices[j]];
        }
    }

    for ( i = 0; i < this->dimension; ++i ) {
        this->a_values[i] = a_tmp[i];
    }
}
```

In general, it is not necessarily a good idea to multiply two sparse matrices. For example, the product of a matrix that has non-zero values only on the first column and a matrix that has non-zero values only on the first row results in a dense matrix.

Like matrix-vector multiplication, backward and forward substitution can be performed in $\Theta(n + N)$ time.

```
void diagonal_solve( vector_t *this, sparse_matrix_t *A ) {
    size_t i, j;
    assert( A->dimension == this->dimension );
    assert( A->a_column_indices[A->dimension] == 0 );

    for ( i = 0; i < this->dimension; ++i ) {
        this->a_values[i] /= A->a_diagonal_entries[i];
    }
}

void backward_substitute( sparse_matrix_t *A, vector_t *b, vector_t *u ) {
    size_t i, j;
    mixed_sum_t sum;
    assert( A->dimension == b->dimension );
    assert( A->dimension == u->dimension );

    j = A->a_row_indices[A->dimension] - 1;

    for ( i = A->dimension - 1; i < A->dimension; --i ) {
        assert( (A->a_row_indices[i] == A->a_row_indices[i + 1])
             || (A->a_column_indices[A->a_row_indices[i]] > i) );

        mixed_sum_init( &sum, b->a_values[i] );

        u->a_values[i] = b->a_values[i];

        for ( ; j < A->a_row_indices[A->dimension] && j >= A->a_row_indices[i]; --j ) {
            mixed_sum( &sum,
                -A->a_off_diagonal_values[j]*u->a_values[A->a_column_indices[j]] );
        }

        u->a_values[i] = mixed_sum_value( &sum )/A->a_diagonal_values[i];
    }
}

void forward_substitute( sparse_matrix_t *A, vector_t *b, vector_t *u ) {
    size_t i, j;
    mixed_sum_t sum;  // double sum;
    assert( A->dimension == b->dimension );
    assert( A->dimension == u->dimension );

    j = 0;

    for ( i = 0; i < A->dimension; ++i ) {
        assert( (A->a_row_indices[i] == A->a_row_indices[i + 1])
             || (A->a_column_indices[A->a_row_indices[i + 1] - 1] < i) );

        mixed_sum_init( &sum, b->a_values[i] );
        u->a_values[i] = b->a_values[i];


        for ( ; j < A->a_row_indices[i + 1]; ++j ) {
            mixed_sum( &sum,
                -A->a_off_diagonal_values[j]*u->a_values[A->a_column_indices[j]] );
        }

        u->a_values[i] = mixed_sum_value( &sum )/A->a_diagonal_values[i];
    }
}
```

```
bool jacobi( sparse_matrix_t *A, vector_t *b, vector_t *u,
             vector_t *tmp, size_t max_iterations, double epsilon_abs ) {
    bool is_approximation_found = false;
    size_t i, j, k;
    mixed_sum_t sum;

    assert( A->dimension ==   b->dimension );
    assert( A->dimension ==   u->dimension );
    assert( A->dimension == tmp->dimension );

    for ( k = 0; k < max_iterations; ++k ) {
        for ( i = 0; i < A->dimension; ++i ) {
            mixed_sum_init( &sum, b->a_values[i] );

            for ( j = A->a_row_indices[i]; j < A->a_row_indices[i + 1]; ++j ) {
                mixed_sum( &sum,
                    -A->a_off_diagonal_values[j]*u->a_values[A->a_column_indices[j]] );
            }

            assert( A->a_diagonal_values[i] != 0.0d );
            tmp->a_values[i] = mixed_sum_value( &sum )/A->a_diagonal_values[i];
        }

        if ( vector_distance( u, tmp, TWO_NORM ) < epsilon_abs ) {
            is_approximation_found = true;
            break;
        } else {
            vector_vector_assign( u, tmp );
        }
    }

    return is_approximation_found;
}
```

```
bool gauss_seidel( sparse_matrix_t *A, vector_t *b, vector_t *u,
             vector_t *tmp, size_t max_iterations, double epsilon_abs ) {
    bool is_approximation_found = false;
    size_t i, j, k;
    mixed_sum_t sum;

    assert( A->dimension ==   b->dimension );
    assert( A->dimension ==   u->dimension );
    assert( A->dimension == tmp->dimension );

    for ( k = 0; k < max_iterations; ++k ) {
        vector_vector_assign( tmp, u );

        for ( i = 0; i < A->dimension; ++i ) {
            mixed_sum_init( &sum, b->a_values[i] );

            for ( j = A->a_row_indices[i]; j < A->a_row_indices[i + 1]; ++j ) {
                mixed_sum( &sum,
                    -A->a_off_diagonal_values[j]*u->a_values[A->a_column_indices[j]] );
            }

            assert( A->a_diagonal_values[i] != 0.0d );
            u->a_values[i] = mixed_sum_value( &sum )/A->a_diagonal_values[i];
        }

        if ( vector_distance( u, tmp, TWO_NORM ) < epsilon_abs ) {
            is_approximation_found = true;
            break;
        }
    }

    return is_approximation_found;
}
```

One of the more difficult algorithms to author is the LU decomposition. In general, there are some observations that must be made:

1. Just because a matrix *A* is sparse, does not guarantee that its LU decomposition is sparse.
2. If the permutation matrix is not the identity matrix (for example, the matrix is not diagonally dominant), then the algorithm can become significantly more difficult.

This implementation of the *LU* decomposition that uses additional memory beyond local variables and constructs the matrices *L* and *U* in row-centric order, thus, allowing the algorithm to use the append function and not set. If the matrix is *band diagonal*, meaning all entries are within *m* entries to the diagonal, then the matrices *L* and *U* will also be band diagonal.

```
void lu( sparse_matrix_t *A, sparse_matrix_t *L, sparse_matrix_t *U ) {
    bool found, diagonal_used;
    size_t i, j;
    size_t *a_first_column_indices;
    size_t min_column, row;
    double value;
    mixed_sum_t sum;

    assert( A->dimension == L->dimension );
    assert( A->dimension == U->dimension );
    assert( sizeof( size_t ) <= sizeof( double ) );

    a_first_column_indices = (size_t *)L->a_diagonal_values;

    sparse_matrix_scalar_multiply( L, 0.0d );
    sparse_matrix_scalar_multiply( U, 0.0d );

    for ( i = 0; i < A->dimension; ++i ) {
        for ( j = 0; j < i; ++j ) {
            a_first_column_indices[j] = U->a_row_indices[j];
        }

        a_first_column_indices[i] = A->a_row_indices[i];

        diagonal_used = false;

        do {
            found = false;

            if ( a_first_column_indices[i] < A->a_row_indices[i + 1] ) {
                min_column = A->a_column_indices[a_first_column_indices[i]];
                found = true;
            }

            for ( j = L->a_row_indices[i]; j < L->a_row_indices[i + 1]; ++j ) {
                row = L->a_column_indices[j];

                if ( a_first_column_indices[row] < U->a_row_indices[row + 1] ) {
                    if ( !found
                      || (U->a_column_indices[a_first_column_indices[row]] < min_column) )
{
                        min_column = U->a_column_indices[a_first_column_indices[row]];
                        found = true;
                    }
                }
            }
```

```
if ( !diagonal_used && (A->a_diagonal_values[i] != 0.0d) ) {
    if ( !found || (found && (i < min_column)) ) {
        min_column = i;
        found = true;
    }
}
```

```
            if ( found ) {
                if ( min_column == i ) {
                    mixed_sum_init( &sum, A->a_diagonal_values[min_column] );
                    diagonal_used = true;
                } else {
                    if ( (a_first_column_indices[i] < A->a_row_indices[i + 1])
                      && (A->a_column_indices[a_first_column_indices[i]] == min_column) ) {

                        mixed_sum_init( &sum,
                                        A-
>a_off_diagonal_values[a_first_column_indices[i]] );
                        ++( a_first_column_indices[i] );
                    } else {
                        mixed_sum_init( &sum, 0.0d );
                    }
                }

                for ( j = L->a_row_indices[i]; j < L->a_row_indices[i + 1]; ++j ) {
                    row = L->a_column_indices[j];

                    if ( (a_first_column_indices[row] < U->a_row_indices[row + 1])
                      && (U->a_column_indices[a_first_column_indices[row]] == min_column) )
{
                        mixed_sum( &sum,
                                   -U->a_off_diagonal_values[a_first_column_indices[row]]
                                   *L->a_off_diagonal_values[j] );
                        ++( a_first_column_indices[row] );
                    }
                }

                value = mixed_sum_value( &sum );

                if ( value != 0.0d ) {
                    // If we are below the diagonal, update L, otherwise, update U

                    if ( min_column < i ) {
                        sparse_matrix_append( L, i, min_column,
                                              value/U->a_diagonal_values[min_column] );
                    } else {
                        sparse_matrix_append( U, i, min_column, value );
                    }
                }
            }
        } while ( !found );
    }

    sparse_matrix_scalar_add( L, 1.0d );
}
```

16.5 Solving systems of linear equations

Linear equations are equations in n variables (or *unknowns*) where all the terms in the equation are either constants or scalar multiples of the n variables; for example, the following are all linear equations

$$3x + 4y + 2 = 0$$

$$2x + 4y - z + 1 = x - 3y + 5z - 2$$

$$y = 4.532x + 0.987$$

$$y + 5 = z - \sin(4)$$

Any equation that is not a linear equation is said to be a *non-linear equation*, and these include

$$3x + 4xy + 2 = 0$$

$$2x + 4y - z + 1 = x^2 - 3y + 5z - 2$$

$$\sin(y) = 4.532x + 0.987$$

$$\frac{1}{y} + 5 = z - \sin(4)$$

On occasion, although seldom, a linear equation may disguise itself to appear to be non-linear, such as

$$2 + 3\frac{y}{x} = \frac{5}{x},$$

however, for the most part, we will always write our linear equations in a canonical form, where all terms on the left-hand side of the equality are scalar multiples of the variables, and all constants are on the right-hand side. For example, our four linear equations above written in canonical form would be

$$3x + 4y = -2$$

$$x + 7y - 6z = 1$$

$$y - 4.532x = 0.987$$

$$y - z = -5 - \sin(4)$$

As the actual variable names often mean nothing, after all, the equations

$$3x + 4y = -2$$

$$3x + 4z = -2$$

$$3a + 4b = -2$$

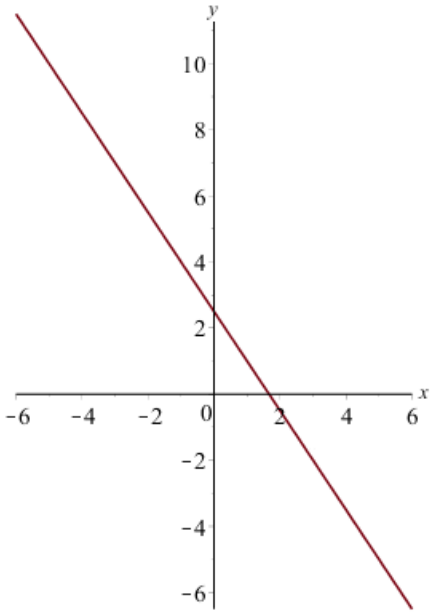
$$3x_1 + 4x_2 = -2$$

all contain the same information. Consequently, we will usually defer to the last formulation, where the *n* variables are listed as $x_1, x_2, x_3, \ldots, x_n$; although we will in many cases chose a different variable name to index, although usually we will restrict ourselves to letters in the alphabet after and including *u*.

For the most simple linear equation, any equation of the form $ax = b$, there is only one solution: $x = \frac{b}{a}$. If, however, we have two variables, it may not have just one solution; for example,

$$3x + 2y = 5.$$

Here, given any value of *x*, then if we let $y = \frac{5 - 3x}{2}$, this pair will satisfy our system of equations. For example, the pair $x = 1$ and $y = 1$ satisfies it, but so does $x = 0$ and $y = 2.5$ or $x = 2$ and $y = -0.5$. In addition, any linear equation in two variables also defines a line on the plane. For example, $3x + 2y = 5$ defines all the points the line in



However, if we have a system of two linear equations in two unknowns, each of these defines a line, so there are three possibilities: the two lines

1. intersect at a point,
2. are parallel but different, or
3. are identical.

For example, the lines defined by

$$x + y = 1 \text{ and } x - 2y = 3$$

intersect at a point. The lines defined by

$$x + y = 1 \text{ and } 2x + 2y = 3$$

are parallel, while the lines

$$x + y = 1 \text{ and } 2x + 2y = 2$$

are the same.

Similarly, every three

16.5.1 Steps in solving systems of linear equations

You will recall from secondary school that there are steps that we can take when attempting to solve a system of linear equations. For example, suppose we want to solve

$$\begin{aligned} x + 2y - z &= -6 \\ 3x - 2y + 2z &= 3 \\ -2x + 4z &= 6 \end{aligned}$$

First, the order of the equations does not matter, and therefore we can swap any two equations, so

$$\begin{aligned} -2x + 4z &= 6 \\ 3x - 2y + 2z &= 3 \\ x + 2y - z &= -6 \end{aligned}$$

represents the same three constraints.

Next we note is that we can multiply any equation by a non-zero constant, and we continue to have the same constraints, so

$$\begin{aligned} -3x - 6y + 3z &= 18 \\ 3x - 2y + 2z &= 3 \\ -2x + 4z &= 6 \end{aligned}$$

continues to have the same solution.

Finally, we note that adding a multiple of one equation onto another does not fundamentally change the constraints, and therefore if we add –3 times the first equation onto the second, and 2 times the first equation onto the third, we get

$$\begin{aligned} x + 2y - z &= -6 \\ -8y + 5z &= 21 \\ 4y + 2z &= -6 \end{aligned}$$

Now, to simplify life, we might swap the second and third equations

$$\begin{aligned} x + 2y - z &= -6 \\ 4y + 2z &= -6 \\ -8y + 5z &= 21 \end{aligned}$$

We can now add twice the second equation onto the third equation to get

$$\begin{aligned} x+2y- \ z &= -6 \\ 4x+2z &= -6 \\ 9z &= \ 9 \end{aligned}$$

We may now deduce that $z = 1$, and substitute this into the second equation to get that

$$4y+2\cdot 1=-6$$

or $y = -2$, and finally, we may substitute both of these into the first equation to get that

$$x+2\cdot(-2)-(-1)=-6$$

or $x = -1$.

Note that if we did not swap the second and third equations, we could have simplified this

$$\begin{aligned} x+2y- \ z &= -6 \\ -8y+5z &= \ 21 \\ 4y+2z &= -6 \end{aligned}$$

by adding half of the second equation onto the third to get

$$\begin{aligned} x+2y- \ \ z &= -6 \\ -8y+5 \ z &= \ 21 \\ 4.5z &= \ \ 4.5 \end{aligned}$$

which would have yielded the same solution.

16.5.2 Interpreting linear equations as constraints on vectors in a vector space

Consider the linear equation

$$\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_{n-1} x_{n-1} + \alpha_n x_n = \beta\,.$$

Another interpretation is to consider all n-dimensional vectors of the form $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}$, in which case, the equation

$$\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_{n-1} x_{n-1} + \alpha_n x_n = \beta$$

restricts the possible vectors that satisfy this condition. The set of vectors that satisfies a linear equation is a subspace if and only if $\beta = 0$; after all, the zero vector $\mathbf{0} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, when substituted into the left-hand side equals zero.

Given the linear equation $\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_{n-1} x_{n-1} + \alpha_n x_n = \beta$, a vector that is orthogonal to the plane is the vector

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_n \end{pmatrix},$$

for two vectors **u** and **v** satisfy the constraint, then

$$\alpha_1 u_1 + \alpha_2 u_2 + \alpha_3 u_3 + \cdots + \alpha_{n-1} u_{n-1} + \alpha_n u_n = \beta \text{ and } \alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 v_3 + \cdots + \alpha_{n-1} v_{n-1} + \alpha_n v_n = \beta.$$

and therefore

$$\begin{aligned} \langle \boldsymbol{\alpha}, \mathbf{u} - \mathbf{v} \rangle &= \sum_{k=1}^{n} a_k \left(u_k - v_k \right) \\ &= \sum_{k=1}^{n} \left(\alpha_k u_k - \alpha_k v_k \right) \\ &= \left(\sum_{k=1}^{n} \alpha_k u_k \right) - \left(\sum_{k=1}^{n} \alpha_k v_k \right) \\ &= \beta - \beta = 0 \end{aligned}$$

Specifically, if the linear equation defines a subspace; that is,

$$\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_{n-1} x_{n-1} + \alpha_n x_n = 0,$$

then every vector in that subspace is orthogonal to $\boldsymbol{\alpha}$ as $\langle \boldsymbol{\alpha}, \mathbf{u} \rangle = \sum_{k=1}^{n} a_k u_k = 0$.

This also suggests a different observation: given a vector $\boldsymbol{\alpha}$, the set of all vectors that satisfies $\langle \boldsymbol{\alpha}, \mathbf{u} \rangle = \beta$ is an (n – 1)-dimensional manifold.

You will note that, in general, a linear equation in n variables defines a subspace of dimension $n - 1$, as the equation makes one restriction on the scope of the variables: given values for $n - 1$ variables, the last variable is given. Thus, if our linear equation is

$$\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \cdots + \alpha_{n-1} x_{n-1} + \alpha_n x_n = \beta$$

then given values for x_1 through x_{n-1}, the value of x_n is

$$x_n = \frac{\beta - \alpha_1 x_1 - \alpha_2 x_2 - \alpha_3 x_3 - \cdots - \alpha_{n-1} x_{n-1}}{\alpha_n}.$$

16.5.3 Elementary row operations

The three elementary row operations are

We will represent these using

Operation	Representation	
Swap Rows j and k	$\mathbf{R}_{j \leftrightarrow k}$	reflection
Multiply Row j by the scalar α.	$\mathbf{R}_{\alpha; j}$	scaling
Add α times Row j on Row k.	$R_{\alpha; j \to k}$	shear

Every row operation may be represented by a matrix:

$$\mathbf{R}_{\alpha; j} = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \alpha & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ j \\ \\ \\ \end{matrix}$$

$$\mathbf{R}_{j\leftrightarrow k} = \begin{array}{c} \begin{matrix} & & & j & & & & k & & & \\ \end{matrix} \\ \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array} \begin{matrix} \\ \\ \\ j \\ \\ \\ \\ k \\ \\ \\ \end{matrix}$$

$$\mathbf{R}_{\alpha;j\rightarrow k} = \begin{array}{c} \begin{matrix} & & & j & & & & k & & & \\ \end{matrix} \\ \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \alpha & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \end{array} \begin{matrix} \\ \\ \\ j \\ \\ \\ \\ k \\ \\ \\ \end{matrix}$$

Note that when we apply matrix operations, they mirror the operations on equations:

Solve	Solve
$2x + 2y - z = -1$ $3x - 5y + 3z = -4$ $5x + 2y + z = 2$	$\left(\begin{array}{ccc\|c} 2 & 2 & -1 & -1 \\ 3 & -5 & 3 & -4 \\ 5 & 2 & 1 & 2 \end{array}\right)$
Add –1.5 times the first equation onto the second $2x + 2y - z = -1$ $-8y + 4.5z = -2.5$ $5x + 2y + z = 2$	Add –1.5 times the first row onto the second $\left(\begin{array}{ccc\|c} 2 & 2 & -1 & -1 \\ 0 & -8 & 4.5 & -2.5 \\ 5 & 2 & 1 & 2 \end{array}\right)$
Add –2.5 times the first equation onto the third $2x + 2y - z = -1$ $-8y + 4.5z = -2.5$ $-3y + 3.5z = 4.5$	Add –2.5 times the first row onto the third $\left(\begin{array}{ccc\|c} 2 & 2 & -1 & -1 \\ 0 & -8 & 4.5 & -2.5 \\ 0 & -3 & 3.5 & 4.5 \end{array}\right)$
Add –0.375 times the second equation onto the third $2x + 2y - z = -1$ $-8y + 4.5z = -2.5$ $1.8125z = 5.4375$	Add –0.375 times the second row onto the third $\left(\begin{array}{ccc\|c} 2 & 2 & -1 & -1 \\ 0 & -8 & 4.5 & -2.5 \\ 0 & 0 & 1.8125 & 5.4375 \end{array}\right)$
The last equation gives us that $z = 3$, so substitute this into the first two equations: $2x + 2y - 3 = -1$ $-8y + 13.5 = -2.5$ $z = 3$ so $2x + 2y = 2$ $-8y = -16$ $z = 3$	Divide the last row by 1.8125 $\left(\begin{array}{ccc\|c} 2 & 2 & -1 & -1 \\ 0 & -8 & 4.5 & -2.5 \\ 0 & 0 & 1 & 3 \end{array}\right)$ and substitute into the next row $\left(\begin{array}{ccc\|c} 2 & 2 & 0 & 2 \\ 0 & -8 & 0 & -16 \\ 0 & 0 & 1 & 3 \end{array}\right)$
The second equation gives us that $y = 2$, so substitute this into the first equation $2x + 4 - 3 = -1$ $y = 2$ $z = 3$ so $2x = -2$ $y = 2$ $z = 3$	Divide the second row by –8 $\left(\begin{array}{ccc\|c} 2 & 2 & 0 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right)$ and substitute into the next row $\left(\begin{array}{ccc\|c} 2 & 0 & 0 & -2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right)$
Finally, the first equation gives us that $x = -1$, so we have $x = -1$ $y = 2$ $z = 3$	Finally, dividing the first row by 2 gives us that $\left(\begin{array}{ccc\|c} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{array}\right)$

In Matlab, you can define a matrix in many different ways. All of the matrices produced in the next five examples product the same 3×5 matrix

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 \\ 11 & 12 & 13 & 14 & 15 \end{pmatrix}.$$

First, we may list the rows, hitting enter at the end of each row:

```
>> M = [ 1  2  3  4  5
         6  7  8  9 10
        11 12 13 14 15];
```

Second, we may indicate the different rows using semicolons:

```
>> M = [1 2 3 4 5; 6 7 8 9 10; 11 12 13 14 15];
```

Third, you could define five column vectors

```
>> v1 = [1  6 11]';
>> v2 = [2  7 12]';
>> v3 = [3  8 13]';
>> v4 = [4  9 14]';
>> v5 = [5 10 15]';
```

and now create a matrix of these column vectors:

```
>> M = [v1 v2 v3 v4 v5];
```

Fourth, you can define three row vectors

```
>> r1 = [ 1  2  3  4  5];
>> r2 = [ 6  7  8  9 10];
>> r3 = [11 12 13 14 15];
```

and now create a matrix of these row vectors:

```
>> M = [r1; r2; r3];
```

or

```
>> M = [r1
        r2
        r3];
```

In all of these examples, the added white space, except for the required one space between vector entries is unnecessary and added only for the clarity of the presentation. You could just as easily enter

```
>> M = [1 2 3 4 5
6 7 8 9 10
11 12 13 14 15];
```

or

```
>> M = [r1
r2
r3];
```

Just like we can join row or column vectors to create a matrix, we can also create the augmented matrix

$$\left(\begin{array}{ccc:c} 5 & 2 & 0 & 6 \\ 3 & 6 & -1 & -7 \\ 1 & 3 & 7 & 3 \end{array}\right)$$

as follows:

```
>> M = [5 2 0; 3 6 -1; 1 3 7];
>> v = [6 -7 3]';
>> Maug = [M v];
```

Now that we can define matrices, we can also solve systems of linear equations. For example, given the system of linear equations

$$5x + 2y = 6$$
$$3x + 6y - z = -7$$
$$x + 3y + 7z = 3$$

we can rewrite this as the following inverse problem, to solve

$$\begin{pmatrix} 5 & 2 & 0 \\ 3 & 6 & -1 \\ 1 & 3 & 7 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 6 \\ -7 \\ 3 \end{pmatrix}$$

Now we enter the known matrix and vector into Matlab:

```
>> M = [5 2 0; 3 6 -1; 1 3 7];
>> b = [6 -7 3]';
```

To solve this system of equations, we use the notation:

```
>> M \ b
  ans =
     2
    -2
     1
```

In reality, matrices are seldom as *clean* as in this example. Normally, to solve a system of linear equations such as

$$5.2925x + 1.8914y - 0.0052z = 5.9417$$
$$3.2702x + 5.8103y - 1.2053z = -7.0350$$
$$1.0359x + 3.1584y + 7.2783z = 3.2943$$

```
>> M = [5.2925 1.8914 -0.0052
        3.2702 5.8103 -1.2053
        1.0359 3.1584  7.2783];
>> b = [5.9417 -7.0350 3.2943]';
>> format long
>> M \ b
  ans =
     1.848988522999504
    -2.029452818534221
     1.070134038317093
```

Aside: You may have noticed that the second example is very similar to the first, only the numbers are *shaken* up, so-to-speak. We say that the second system of linear equations is a *perturbation* of the first. Notice that the answer is also close: 2 versus 1.8490…, –2 versus –2.0294… and 1 versus 1.0701…. In some cases, however, small changes to the coefficients can lead to significant changes in the answer.

```
>> M1 = [-2.0391 0.9928 0.0153; 1.0093 -2.9815 2.0359; 3.0952 0.9193 -1.9354]
  M1 =
    -2.0391   0.9928   0.0153
     1.0093  -2.9815   2.0359
     3.0952   0.9193  -1.9354
>> M2 = [-2.0425 1.0047 -0.0009; 1.0123 -3.0085 2.0146; 3.0523 0.9204 -1.9847]
  M2 =
    -2.0425   1.0047  -0.0009
     1.0123  -3.0085   2.0146
     3.0523   0.9204  -1.9847
```

Now compare these two solutions:

```
>> M1 \ [2 1 -4]'
  ans =
     3.656487807105599
     9.334224766709340
    12.348100593463442

>> M2 \ [2 1 -4]'
  ans =
    -9.692762749739293
   -17.733125656871273
   -21.114923462293326
```

Now, in any physical system, there will be errors in any readings (at least some component of which is called *noise*). If this is the case, then even the smallest error in our readings may lead to a completely different solution. Later, you will see numerical algorithms to determine when the solution of a system of linear equations can be trusted, and when its answer, no matter how precise your measurements, will always be suspect.

17 Moore-Penrose pseudo-inverse

Given a linear mapping $A: \mathcal{U} \to \mathcal{V}$ where both vector spaces have inner products (and thus, both vector spaces have a 2-norm induced by that inner product) and given a target vector $\mathbf{v} \in \mathcal{V}$, let us consider the inverse problem defined by $A\mathbf{u} = \mathbf{v}$. First, if there is a solution to $A\mathbf{u} = \mathbf{v}$, there may also be infinitely many solutions. Second, if there is no solution to $A\mathbf{u} = \mathbf{v}$, there is at least one point in range(A) that maps closest to $\mathbf{v}$, but there too may be infinitely many points that map to this same point. Thus, we define $A^+\mathbf{v}$ to be that vector $\mathbf{u} \in \mathcal{U}$ that is both smallest in 2-norm while simultaneously having an image in the range that is closest to $\mathbf{v}$ (that is, $\left\| AA^+\mathbf{v} - \mathbf{v} \right\|_2$ is minimized).

It can be shown that such a mapping exists and is linear, and thus must be representable by an $n \times m$ matrix. Such a pseudo-inverse has applications in statistics, grpahics and robotics, to name a few. A recent article in Mathematical Problems in Engineering include the use of this pseudo-inverse for digital image restoration. At this point, however, we will go on to:

1. find this pseudo-inverse in three special cases,
2. describe how to find this pseudo-inverse when the matrix is a rectangular diagaonal matrix,
3. describe how to find this pseudo-inverse for a general matrix,
4. list the alternative criteria for this pseudo-inverse as is normally given in the literature, and
5. looked at an application when the normal least-squares techniques fail.

We will start by looking at finding the Moore-Penrose pseudo-inverse when A is one-to-one and onto, one-to-one but not onto, and onto but not one-to-one.

17.1 Finding the Moore-Penrose pseudo-inverse in special cases

In three out of four cases above, it is reasonably straight-forward to find the Moore-Penrose Pseudo-inverse. We will look at these three cases, and then carry on by describing how we can find the pseudo-inverse in general. We will describe these cases when $A: \mathbf{R}^n \to \mathbf{R}^m$.

17.1.1 One-to-one and onto

If A is one-to-one and onto (only possible if $m = n$), then there is a unique $\mathbf{u}$ such that $A\mathbf{u} = \mathbf{v}$, so that unique $\mathbf{u}$ must be the Moore-Penrose pseudo-inverse, so $A^+\mathbf{v} = \mathbf{u}$, and thus $A^+ = A^{-1}$.

17.1.2 One-to-one but not onto

If A is one-to-one but not onto, then all the columns of A are linearly independent but $m > n$ so that range(A) is a subspace of $\mathbf{R}^m$ and each point in that subspace has a unique pre-image. If $\mathbf{v}$ is in range(A), then let $A^+\mathbf{v}$ be that unique vector $\mathbf{u}$ that maps to $\mathbf{v}$. If, however, it happens that no vector in the domain maps to $\mathbf{v}$, there must be a vector in range(A) that is closest to $\mathbf{v}$. This is the least-squares problem where there is a unique solution to $A^*A\mathbf{u} = A^*\mathbf{v}$ and this is the point in range(A) that is closest to $\mathbf{v}$. Thus, in this case, $A^+\mathbf{v}$ equals this unique solution, and thus $A^+ = (A^*A)^{-1}A^*$.

17.1.3 Onto but not one-to-one

If A is onto but not one-to-one (only possible if $n > m$), there are multiple vectors that map onto $\mathbf{v}$. Define $A^+\mathbf{v}$ to be that vector with smallest norm that maps onto $\mathbf{v}$. That is, $\|A^+\mathbf{v}\|_2 \le \|\mathbf{u}\|_2$ for all $\mathbf{u}$ in the domain such that $A\mathbf{u} = \mathbf{v}$ and where $AA^+\mathbf{v} = \mathbf{v}$. This is the least-norm problem. Recall that all vectors that map onto $\mathbf{v}$ may be found by finding one particular vector $\mathbf{u}_p$ that maps onto $\mathbf{v}$ and then adding $\mathbf{u}_p$ + null(A), that is adding to $\mathbf{u}_p$ all vectors in the null space, thus exactly one vector must be closest to the origin $\mathbf{0}_n$. Specifically, we want to find that unique vector that is orthogonal to every vector in the null space. Thus, $\left\langle A^+\mathbf{v}, \mathbf{u}_0 \right\rangle = 0$ for every vector $\mathbf{u}_0 \in \text{null}(A)$

Because A is onto, $A^*: \mathbf{R}^m \to \mathbf{R}^n$ is one-to-one, $AA^*\mathbf{v} = A\mathbf{u}$

17.1.4 Summary on three special cases

We have looked at three special cases where it is possible to describe A^+. Each of these, however, requires one to calculate the inverse, a numerically unstable algorithm unless performed correctly. In Section 17.3, we will see how to find the Moore-Penrose pseudo inverse, but we will first start with a special case: when the matrix is diagonal.

17.2 The Moore-Penrose pseudo-inverse of a rectangular diagonal matrix

A rectangular diagonal matrix $\Sigma : \mathbf{R}^n \to \mathbf{R}^m$ is any matrix where all off-diagonal entries are zero; that is, $\sigma_{i,j} = 0$ whenever $i \neq j$. The Moore-Penrose inverse Σ^+ is thus the conjugugate transpose of Σ with the reciprocal taken of all non-zero entries on the diagonal. We will argue that this is the case by looking at two examples: when Σ:$\mathbf{R}^3 \to \mathbf{R}^4$ and when Σ:$\mathbf{R}^4 \to \mathbf{R}^3$. Note that if at least one diagonal entry is zero, then the mapping is neither one-to-one nor onto, and yet we will see that this is indeed an appropriate formulation of this inverse.

17.2.1 Example 1: Σ:$\mathbf{R}^3 \to \mathbf{R}^4$

Thus, if $\Sigma = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -4 \\ 0 & 0 & 0 \end{pmatrix}$, then $\Sigma^+ = \begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -0.25 & 0 \end{pmatrix}$. Note that a basis for the range is $\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$ and if the vector $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix}$, then $\Sigma^+\mathbf{v} = \begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -0.25 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = \begin{pmatrix} 0.5v_1 \\ 0 \\ -0.25v_3 \end{pmatrix}$ and $\Sigma\Sigma^+\mathbf{v} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -4 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.5v_1 \\ 0 \\ -0.25v_3 \end{pmatrix} = \begin{pmatrix} v_1 \\ 0 \\ v_3 \\ 0 \end{pmatrix}$, which is the closest vector in the range to $\mathbf{v}$. Also note that the null space of Σ has a basis $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ and thus all solutions to $S\mathbf{u} = \mathbf{v}$ are $\mathbf{u} = \begin{pmatrix} 0.5v_1 \\ u_2 \\ -0.25v_3 \end{pmatrix}$ where u_2 is a free variable, and therefore we note that that vector with the smallest 2-norm is when $u_2 = 0$.

Note that for numerical stability, in general it is assumed that if one of the diagonal entries is essentially zero, we treat it as if it were zero as opposed to calculating its reciprocal.

17.2.2 Example 2: $\Sigma: R^4 \to R^3$

Thus, if $\Sigma = \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$, then $\Sigma^+ = \begin{pmatrix} -0.5 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. Note that a basis for the range is $\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ and if the vector $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$, then $\Sigma^+\mathbf{v} = \begin{pmatrix} -0.5 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} -0.5v_1 \\ 0.2v_2 \\ 0 \\ 0 \end{pmatrix}$ and $\Sigma\Sigma^+\mathbf{v} = \begin{pmatrix} -2 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} -0.5v_1 \\ 0.2v_2 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ 0 \end{pmatrix}$, which is the closest vector in the range to **v**. Also note that the null space of Σ has a basis $\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$ and thus all solutions to S**u** = **v** are $\mathbf{u} = \begin{pmatrix} -0.5v_1 \\ 0.2v_2 \\ u_3 \\ u_4 \end{pmatrix}$ where u_3 and u_4 are free variables, and therefore we note that that vector with the smallest 2-norm is when $u_3 = u_4 = 0$.

17.3 Finding the Moore-Penrose pseuodo-inverse in general

Previously, we have demonstrated that we can find the Moore-Penrose pseudo-inverse if the mapping is either one-to-one or onto (or both), however, the techniques described require the computation of the inverse, and thus we now describe a more numerically stable inverse.

For $A: \mathbf{F}^n \to \mathbf{F}^m$, the most accurate means of finding the Moore-Penrose pseudo-inverse is to find the singular-value decomposition (SVD) of the matrix $A = U\Sigma V^*$ and to then calculate $A^+ = U\Sigma^+V^*$ where Σ^+ is the Moore-Penrose pseudo-inverse of the diagonal rectangular matrix Σ. This is the approach used by J. Demmel and W. Kahan in their article "Accurate singular values of bidiagonal matrices" found in SIAM Journal on Scientific and Statistical Computing, vol. 11, no. 5, pp. 873–912, 1990.

17.4 Alternative requirements for the Moore-Penrose pseudo-inverse

The stated requirements for the Moore-Pemrose pseudo-inverse are:

1. $AA^+A = A$. This says that $AA^+\mathbf{a}_k = \mathbf{a}_k$ for each column vector of $A = (\mathbf{a}_1 \cdots \mathbf{a}_n)$.
2. $A^+AA^+ = A^+$. This says that $A^+A\mathbf{a}_k^+ = \mathbf{a}_k^+$ for each column vector of $A^+ = (\mathbf{a}_1^+ \cdots \mathbf{a}_m^+)$.
3. AA^+ is conjugate symmetric, so $\langle AA^+\mathbf{v}_1, \mathbf{v}_2 \rangle = \langle \mathbf{v}_1, AA^+\mathbf{v}_2 \rangle$.
4. A^+A is conjugate symmetric, so $\langle A^+A\mathbf{u}_1, \mathbf{u}_2 \rangle = \langle \mathbf{u}_1, A^+A\mathbf{u}_2 \rangle$.

From these four properties, it is possible to deduce the properties above; however, that is beyond the scope of this course.

17.5 Applications

In finding a least-squares solution to $A\mathbf{u} = \mathbf{v}$, meaning finding that $\mathbf{u}$ that minimizes $\|A\mathbf{u} - \mathbf{v}\|_2$, if there are infinitely many solutions to this minimization problem, $\mathbf{u} = A^+\mathbf{v}$ finds that $\mathbf{u}$ with minimal 2-norm. For example, find that linear combination of the vectors

$$\begin{pmatrix} 1 \\ -1 \\ 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 3 \\ 2 \\ 0 \end{pmatrix}$$

that best approximates the target vector $\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$. Now, there is no solution, as the target vector is not in the span of these three vectors. Thus, if you let $A = \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & 3 \\ 0 & 1 & 2 \\ 2 & 1 & 0 \end{pmatrix}$, then we note that $A^*A = A^*\mathbf{v}$, which is

$$\left(\begin{array}{ccc|c} 6 & 2 & -2 & 7 \\ 2 & 4 & 6 & 10 \\ -2 & 6 & 14 & 13 \end{array}\right)$$

has infinitely many solutions, namely $\begin{pmatrix} 0.4 \\ 2.3 \\ 0.0 \end{pmatrix} + \alpha \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$. It may be difficult to determine which of these has the least 2-norm, however, we see that

$$A^+ = \frac{1}{6}\begin{pmatrix} 0.9 & -0.7 & 0.1 & 1.7 \\ 0.6 & 0.2 & 0.4 & 0.8 \\ 0.3 & 1.1 & 0.7 & -0.1 \end{pmatrix}$$

and thus

$$A^+\mathbf{v} = \frac{1}{6}\begin{pmatrix} 0.9 & -0.7 & 0.1 & 1.7 \\ 0.6 & 0.2 & 0.4 & 0.8 \\ 0.3 & 1.1 & 0.7 & -0.1 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 1.1 \\ 0.9 \\ 0.7 \end{pmatrix}$$

where $\left\|A^+\mathbf{v}\right\|_2 \approx 1.5843$.

Thus, that linear combination of the given three vectors that comes closest to the target vector is

$$1.1\begin{pmatrix}1\\-1\\0\\2\end{pmatrix}+0.9\begin{pmatrix}1\\1\\1\\1\end{pmatrix}+0.7\begin{pmatrix}1\\3\\2\\0\end{pmatrix}=\begin{pmatrix}2.7\\1.9\\2.3\\3.1\end{pmatrix}$$

and the sum of the squares of these coefficients is a minimum.

You may note that the third vector is a linear combination of the first two vectors, and the first two vectors are linealry independent, so the solution $\begin{pmatrix}0.4\\2.3\\0.0\end{pmatrix}$ (when $\alpha = 0$) is the unique solution to the least-squares problem of finding the best approximation of a linear combination of the first two vectrs:

$$\begin{pmatrix}1\\-1\\0\\2\end{pmatrix} \text{ and } \begin{pmatrix}1\\1\\1\\1\end{pmatrix},$$

namely

$$0.4\begin{pmatrix}1\\-1\\0\\2\end{pmatrix}+2.3\begin{pmatrix}1\\1\\1\\1\end{pmatrix}=\begin{pmatrix}2.7\\1.9\\2.3\\3.1\end{pmatrix},$$

but we note that $0.4^2 + 2.3^2 = 5.45 > 2.51 = 1.1^2 + 0.9^2 + 0.7^2$.

17.6 Summary of the Moore-Penrose pseudo-inverse

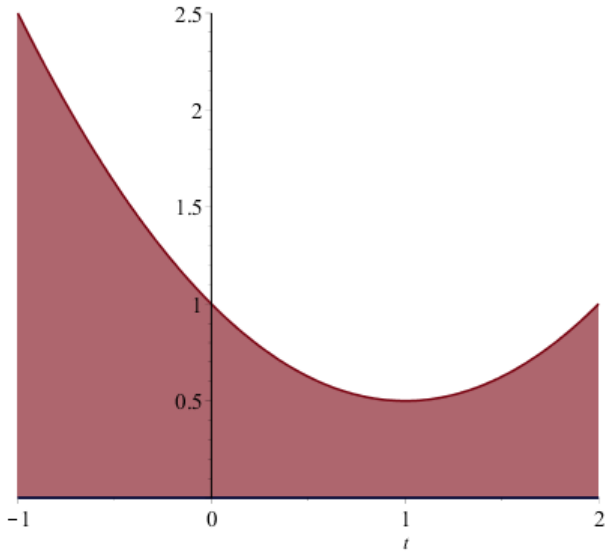
In this chapter, we described the Moore-Penrose pseudo-inverse. This finds the least-squares best approximation of $A\mathbf{u} = \mathbf{v}$, and when there are infinitely many solutions, it chooses that $\mathbf{u}$ that has the smallest 2-norm. This pseudo-inverse can be easily described for diagonal rectangular matrices, and for an arbitrary matrix, it can be found relatively easily using the SVD.

Glossary

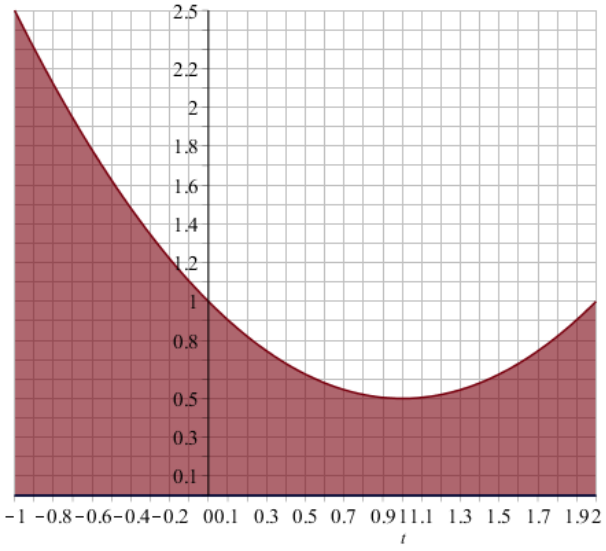
co-linear	Two vectors are *co-linear* if the vectors are scalar multiples of each other.
null space	The null space of a linear mapping $A:\mathcal{U}\to\mathcal{V}$ is the collection of all vectors in the domain that map onto the zero vector $\mathbf{0}_{\mathcal{V}}$ in the co-domain. For finite-dimensional linear operators $A:\mathbf{F}^n\to\mathbf{F}^m$, the dimension of the null space is equal to the number of free variables when solving $A\mathbf{u}=\mathbf{0}_m$.
co-domain	The complementary domain, given a mapping $A:\mathcal{U}\to\mathcal{V}$, the co-domain is the vector space into vectors are being mapped onto.
domain	The
onto	A linear mapping $A:\mathcal{U}\to\mathcal{V}$ is onto if for every vector in the co-domain $\mathbf{v}\in\mathcal{V}$, there is at least one vector in the domain such that $A\mathbf{u}=\mathbf{v}$. For finite-dimensional linear operators, $A:\mathbf{F}^n\to\mathbf{F}^m$, the linear operator is onto if rank(A) = m (or equivalently, dim(range(A)) = m).
one-to-one	A linear mapping $A:\mathcal{U}\to\mathcal{V}$ is one-to-one if for every vector $\mathbf{v}$ in the range of A, there is one and only one vector $\mathbf{u}\in\mathcal{U}$ such that $A\mathbf{u}=\mathbf{v}$. For finite-dimensional linear operators, $A:\mathbf{F}^n\to\mathbf{F}^m$, the linear operator is one-to-one if null(A) = $\{\mathbf{0}_n\}$ (or equivalently, dim(null(A)) = 0); that is, only the zero vector $\mathbf{0}_n$ is mapped onto $\mathbf{0}_m$.
many-to-one	A linear mapping $A:\mathcal{U}\to\mathcal{V}$ is many-to-one if for every vector $\mathbf{v}$ in the range of A, there is one and only one vector $\mathbf{u}\in\mathcal{U}$ such that $A\mathbf{u}=\mathbf{v}$. For finite-dimensional linear operators, $A:\mathbf{F}^n\to\mathbf{F}^m$, the linear operator is one-to-one if null(A) = $\{\mathbf{0}_n\}$ (or equivalently, dim(null(A)) = 0); that is, only the zero vector $\mathbf{0}_n$ is mapped onto $\mathbf{0}_m$.

Appendix A: A quick introduction to integration

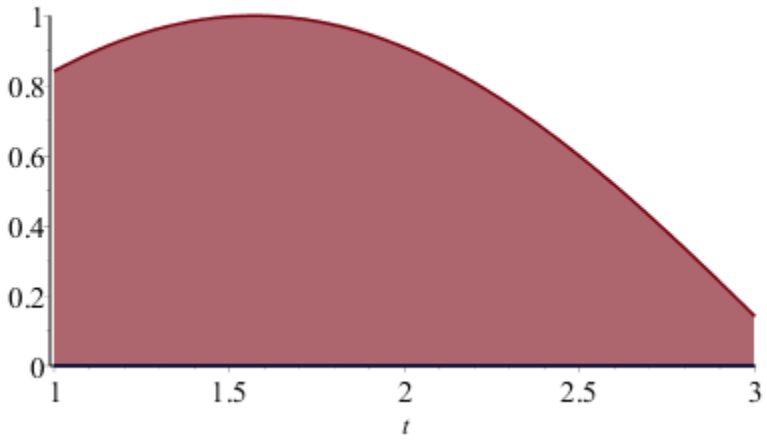
Integration is an idea: given a function $y = f(t)$, what is the area between the line $y = 0$ and $y = f(t)$? For example, given the polynomial $y(t) = 0.5t^2 - t + 1$, as shown in , you may ask what is the area shown in red? In this case, we are asking, what is the area under the function between $t = -1$ and $t = 2$.



The algorithms for determining the integrals is something you will cover in an introductory course in calculus, if you have not already learned about it. In this case, the area under the curve is 3, although usually the number is not that "clean". If you look at, each square has an area of 0.01, and if you could estimate the area, if you wish.



If you try to find the area under the sine function between $t = 1$ and $t = 3$, you would have to determine the area in .



In this case, the area is approximately equal to 1.530294802, and in a course in calculus, you will learn that the exact area is exactly equal to cos(1) – cos(3).

We will use the notation

$$\int_{1}^{3} \sin(t)\,dt$$

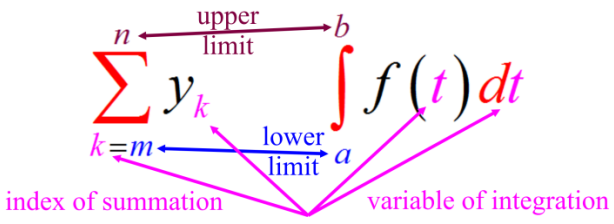
to say we want to calculate the area under the sine function from $t = 1$ to $t = 3$. The dt indicates that we are calculating the area (integrating) with respect to the variable t, just like

$$\frac{d}{dt}\sin(t)$$

indicates that we are taking the derivative of sin(t) with respect to t. The squiggly line to the left is an old way of writing the letter "s", for example, in the American Declaration of Independence, the word "measures" appears as is shown in .



Essentially, we are saying that we are "summing up the area". This is similar to the sigma notation for sums, where the Greek letter Σ was chosen for the same reason, it is a sum.



For the polynomial above, we would write

$$\int_1^3 \left(\frac{1}{2}t^2 - t + 1\right) dt \,,$$

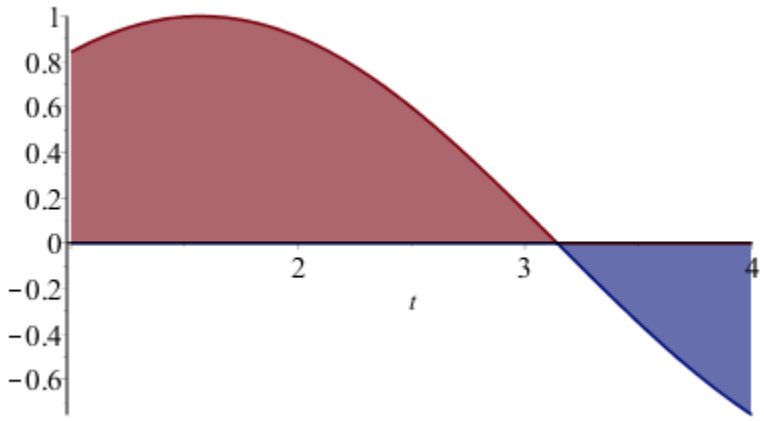
although some times the parentheses are left off with the assumption that we are integrating everything be the long s and the *dt*.

Now, to understand the significance to engineering, suppose that the function is our speed. In this case, the derivative of speed is acceleration, but the integral is the distance travelled. Thus, if $v(t)$ gives our velocity at any moment in time (perhaps accelerating at times, perhaps decelerating, or maintaining a constant speed, then

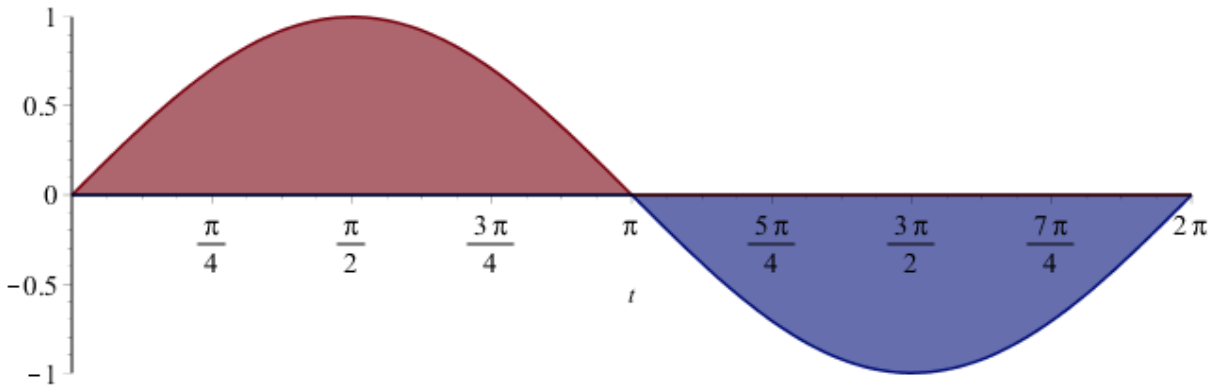
$$\int_{t_{\text{initial}}}^{t_{\text{final}}} v(t)\, dt$$

is the distance travelled between the time t_{initial} and t_{final}.

You may ask, what happens if you back up? That is, what if you start travelling forward, but at some point you go into reverse, in which case, your velocity becomes negative? In this case, we define the area underneath the line $y = 0$ as being negative, and thus, if we wanted to find the area between the function $y(t) = \sin(t)$ and the line $y = 0$, we would have to calculate the area of region where $\sin(t)$ is greater than zero, and subtract from that the area for the region where $\sin(t)$ is less than zero.



You should therefore conclude that $\int_0^{2\pi} \sin(t)\, dt = 0$, as the area above the curve (between $0 \le t \le \pi$) equals the area underneath the curve (between $\pi \le t \le 2\pi$).



Theorem

If $f(t) \geq 0$ for all values of $a \leq t \leq b$, then $\int_a^b f(t)\,dt \geq 0$.

Note that in this course, we will not require an understanding of how to find the area under a curve, but rather, we will simply use the idea.