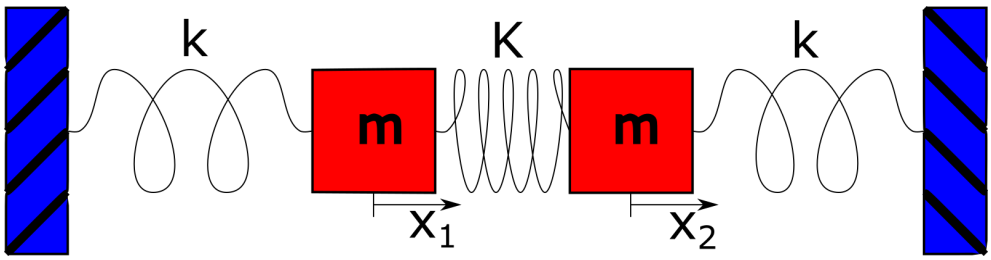
Applied Linear Algebra and Differential Equations

Lecture notes for MATH 2350

Jeffrey R. Chasnov



THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

The Hong Kong University of Science and Technology
Department of Mathematics
Clear Water Bay, Kowloon
Hong Kong

Copyright © 2017-2019 by Jeffrey Robert Chasnov

This work is licensed under the Creative Commons Attribution 3.0 Hong Kong License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/hk/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

Preface

What follows are my lecture notes for a mathematics course offered to second-year engineering students at the the Hong Kong University of Science and Technology. Material from our usual courses on linear algebra and differential equations have been combined into a single course (essentially, two half-semester courses) at the request of our Engineering School. I have tried my best to select the most essential and interesting topics from both courses, and to show how knowledge of linear algebra can improve students' understanding of differential equations.

All web surfers are welcome to download these notes and to use the notes and videos freely for teaching and learning.

I also have some online courses on Coursera. You can click on the links below to explore these courses.

If you want to learn differential equations, have a look at

Differential Equations for Engineers

If your interests are matrices and elementary linear algebra, try

Matrix Algebra for Engineers

If you want to learn vector calculus (also known as multivariable calculus, or calculus three), you can sign up for

Vector Calculus for Engineers

And if your interest is numerical methods, have a go at

Numerical Methods for Engineers

Jeffrey R. Chasnov
Hong Kong
January 2020

Contents

0 A short mathematical review **1**
- 0.1 The trigonometric functions . . . 1
- 0.2 The exponential function and the natural logarithm . . . 1
- 0.3 Definition of the derivative . . . 2
- 0.4 Differentiating a combination of functions . . . 2
  - 0.4.1 The sum or difference rule . . . 2
  - 0.4.2 The product rule . . . 2
  - 0.4.3 The quotient rule . . . 2
  - 0.4.4 The chain rule . . . 2
- 0.5 Differentiating elementary functions . . . 3
  - 0.5.1 The power rule . . . 3
  - 0.5.2 Trigonometric functions . . . 3
  - 0.5.3 Exponential and natural logarithm functions . . . 3
- 0.6 Definition of the integral . . . 3
- 0.7 The fundamental theorem of calculus . . . 4
- 0.8 Definite and indefinite integrals . . . 5
- 0.9 Indefinite integrals of elementary functions . . . 5
- 0.10 Substitution . . . 6
- 0.11 Integration by parts . . . 6
- 0.12 Taylor series . . . 6
- 0.13 Functions of several variables . . . 7
- 0.14 Complex numbers . . . 8

I Linear algebra **13**

1 Matrices **17**
- 1.1 Definition of a matrix . . . 17
- 1.2 Addition and multiplication of matrices . . . 17
- 1.3 The identity matrix and the zero matrix . . . 19
- 1.4 General notation, transposes, and inverses . . . 19
- 1.5 Rotation matrices and orthogonal matrices . . . 23
- 1.6 Matrix representation of complex numbers . . . 25
- 1.7 Permutation matrices . . . 25
- 1.8 Projection matrices . . . 26

2 Systems of linear equations **27**
- 2.1 Gaussian Elimination . . . 27
- 2.2 When there is no unique solution . . . 28
- 2.3 Reduced row echelon form . . . 29
- 2.4 Computing inverses . . . 30
- 2.5 LU decomposition . . . 31

3 Vector spaces **35**
3.1 Vector spaces . . . 35
3.2 Linear independence . . . 37
3.3 Span, basis and dimension . . . 38
3.4 Inner product spaces . . . 39
3.5 Vector spaces of a matrix . . . 41
3.5.1 Null space . . . 41
3.5.2 Application of the null space . . . 42
3.5.3 Column space . . . 43
3.5.4 Row space, left null space and rank . . . 44
3.6 Gram-Schmidt process . . . 44
3.7 Orthogonal projections . . . 46
3.8 QR factorization . . . 47
3.9 The least-squares problem . . . 48
3.10 Solution of the least-squares problem . . . 49

4 Determinants **53**
4.1 Two-by-two and three-by-three determinants . . . 53
4.2 Laplace expansion and Leibniz formula . . . 54
4.3 Properties of the determinant . . . 55
4.4 Cramer's rule . . . 60
4.5 Calculating the inverse matrix using determinants . . . 62
4.6 Use of determinants in Vector Calculus . . . 64

5 Eigenvalues and eigenvectors **67**
5.1 The eigenvalue problem . . . 67
5.2 Matrix diagonalization . . . 71
5.3 Symmetric and Hermitian matrices . . . 72

II Differential equations **75**

6 Introduction to odes **79**
6.1 The simplest type of differential equation . . . 79

7 First-order odes **81**
7.1 The Euler method . . . 81
7.2 Separable equations . . . 82
7.3 Linear equations . . . 86
7.4 Applications . . . 88
7.4.1 Compound interest . . . 88
7.4.2 Chemical reactions . . . 89
7.4.3 Terminal velocity . . . 91
7.4.4 Escape velocity . . . 92
7.4.5 RC circuit . . . 93
7.4.6 The logistic equation . . . 95

8 Second-order odes, constant coefficients **97**
8.1 The Euler method . . . 97
8.2 The principle of superposition . . . 98
8.3 The Wronskian . . . 98
8.4 Homogeneous odes . . . 99

8.4.1 Distinct real roots . . . 100
8.4.2 Distinct complex-conjugate roots . . . 102
8.4.3 Degenerate roots . . . 104
8.5 Difference equations . . . 105
8.6 Inhomogeneous odes . . . 106
8.7 Resonance . . . 110
8.8 Applications . . . 113
8.8.1 RLC circuit . . . 113
8.8.2 Mass on a spring . . . 115
8.8.3 Pendulum . . . 116
8.9 Damped resonance . . . 117

9 Series solutions **119**
9.1 Ordinary points . . . 119

10 Systems of linear differential equations **125**
10.1 Distinct real eigenvalues . . . 125
10.2 Solution by diagonalization . . . 127
10.3 Solution by the matrix exponential . . . 128
10.4 Distinct complex-conjugate eigenvalues . . . 129
10.5 Repeated eigenvalues with one eigenvector . . . 131
10.6 Normal modes . . . 133

11 Nonlinear differential equations **137**
11.1 Fixed points and stability . . . 137
11.1.1 One dimension . . . 137
11.1.2 Two dimensions . . . 138
11.2 Bifurcation theory . . . 140
11.2.1 Saddle-node bifurcation . . . 141
11.2.2 Transcritical bifurcation . . . 142
11.2.3 Supercritical pitchfork bifurcation . . . 143
11.2.4 Subcritical pitchfork bifurcation . . . 143
11.2.5 Application: a mathematical model of a fishery . . . 146

Chapter 0

A short mathematical review

A basic understanding of pre-calculus, calculus, and complex numbers is required for this course. This zero chapter presents a concise review.

0.1 The trigonometric functions

The Pythagorean trigonometric identity is

$$\sin^2 x + \cos^2 x = 1,$$

and the addition theorems are

$$\begin{aligned}\sin(x+y) &= \sin(x)\cos(y) + \cos(x)\sin(y),\\ \cos(x+y) &= \cos(x)\cos(y) - \sin(x)\sin(y).\end{aligned}$$

Also, the values of $\sin x$ in the first quadrant can be remembered by the rule of quarters, with $0^\circ = 0$, $30^\circ = \pi/6$, $45^\circ = \pi/4$, $60^\circ = \pi/3$, $90^\circ = \pi/2$:

$$\sin 0^\circ = \sqrt{\frac{0}{4}}, \qquad \sin 30^\circ = \sqrt{\frac{1}{4}}, \qquad \sin 45^\circ = \sqrt{\frac{2}{4}},$$
$$\sin 60^\circ = \sqrt{\frac{3}{4}}, \qquad \sin 90^\circ = \sqrt{\frac{4}{4}}.$$

The following symmetry properties are also useful:

$$\sin(\pi/2 - x) = \cos x, \quad \cos(\pi/2 - x) = \sin x;$$

and

$$\sin(-x) = -\sin(x), \quad \cos(-x) = \cos(x).$$

0.2 The exponential function and the natural logarithm

The transcendental number e, approximately 2.71828, is defined as

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n.$$

The exponential function $\exp(x) = e^x$ and natural logarithm $\ln x$ are inverse functions satisfying

$$e^{\ln x} = x, \quad \ln e^x = x.$$

The usual rules of exponents apply:

$$e^x e^y = e^{x+y}, \quad e^x / e^y = e^{x-y}, \quad (e^x)^p = e^{px}.$$

The corresponding rules for the logarithmic function are

$$\ln(xy) = \ln x + \ln y, \quad \ln(x/y) = \ln x - \ln y, \quad \ln x^p = p \ln x.$$

0.3 Definition of the derivative

The derivative of the function $y = f(x)$, denoted as $f'(x)$ or dy/dx, is defined as the slope of the tangent line to the curve $y = f(x)$ at the point (x, y). This slope is obtained by a limit, and is defined as

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}. \tag{1}$$

0.4 Differentiating a combination of functions

0.4.1 The sum or difference rule

The derivative of the sum of $f(x)$ and $g(x)$ is

$$(f+g)' = f' + g'.$$

Similarly, the derivative of the difference is

$$(f-g)' = f' - g'.$$

0.4.2 The product rule

The derivative of the product of $f(x)$ and $g(x)$ is

$$(fg)' = f'g + fg',$$

and should be memorized as "the derivative of the first times the second plus the first times the derivative of the second."

0.4.3 The quotient rule

The derivative of the quotient of $f(x)$ and $g(x)$ is

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2},$$

and should be memorized as "the derivative of the top times the bottom minus the top times the derivative of the bottom over the bottom squared."

0.4.4 The chain rule

The derivative of the composition of $f(x)$ and $g(x)$ is

$$\Big(f\big(g(x)\big)\Big)' = f'\big(g(x)\big) \cdot g'(x),$$

and should be memorized as "the derivative of the outside times the derivative of the inside."

0.5 Differentiating elementary functions

0.5.1 The power rule

The derivative of a power of x is given by

$$\frac{d}{dx}x^p = px^{p-1}.$$

0.5.2 Trigonometric functions

The derivatives of $\sin x$ and $\cos x$ are

$$(\sin x)' = \cos x, \quad (\cos x)' = -\sin x.$$

We thus say that "the derivative of sine is cosine," and "the derivative of cosine is minus sine." Notice that the second derivatives satisfy

$$(\sin x)'' = -\sin x, \quad (\cos x)'' = -\cos x.$$

0.5.3 Exponential and natural logarithm functions

The derivative of e^x and $\ln x$ are

$$(e^x)' = e^x, \quad (\ln x)' = \frac{1}{x}.$$

0.6 Definition of the integral

The definite integral of a function $f(x) > 0$ from $x = a$ to b ($b > a$) is defined as the area bounded by the vertical lines $x = a$, $x = b$, the x-axis and the curve $y = f(x)$. This "area under the curve" is obtained by a limit. First, the area is approximated by a sum of rectangle areas. Second, the integral is defined to be the limit of the rectangle areas as the width of each individual rectangle goes to zero and the number of rectangles goes to infinity. This resulting infinite sum is called a *Riemann Sum*, and we define

$$\int_a^b f(x)dx = \lim_{h\to 0}\sum_{n=1}^{N} f\big(a+(n-1)h\big)\cdot h, \tag{2}$$

where $N = (b-a)/h$ is the number of terms in the sum. The symbols on the left-hand-side of (2) are read as "the integral from a to b of f of x dee x." The Riemann Sum definition is extended to all values of a and b and for all values of $f(x)$ (positive and negative). Accordingly,

$$\int_b^a f(x)dx = -\int_a^b f(x)dx \quad \text{and} \quad \int_a^b (-f(x))dx = -\int_a^b f(x)dx.$$

Also,

$$\int_a^c f(x)dx = \int_a^b f(x)dx + \int_b^c f(x)dx,$$

which states when $f(x) > 0$ and $a < b < c$ that the total area is equal to the sum of its parts.

0.7 The fundamental theorem of calculus

View tutorial on YouTube

Using the definition of the derivative, we differentiate the following integral:

$$\begin{aligned}\frac{d}{dx}\int_a^x f(s)ds &= \lim_{h\to 0}\frac{\int_a^{x+h} f(s)ds - \int_a^x f(s)ds}{h}\\ &= \lim_{h\to 0}\frac{\int_x^{x+h} f(s)ds}{h}\\ &= \lim_{h\to 0}\frac{hf(x)}{h}\\ &= f(x).\end{aligned}$$

This result is called the fundamental theorem of calculus, and provides a connection between differentiation and integration.

The fundamental theorem teaches us how to integrate functions. Let $F(x)$ be a function such that $F'(x) = f(x)$. We say that $F(x)$ is an antiderivative of $f(x)$. Then from the fundamental theorem and the fact that the derivative of a constant equals zero,

$$F(x) = \int_a^x f(s)ds + c.$$

Now, $F(a) = c$ and $F(b) = \int_a^b f(s)ds + F(a)$. Therefore, the fundamental theorem shows us how to integrate a function $f(x)$ provided we can find its antiderivative:

$$\int_a^b f(s)ds = F(b) - F(a). \tag{3}$$

Unfortunately, finding antiderivatives is much harder than finding derivatives, and indeed, most complicated functions cannot be integrated analytically.

We can also derive the very important result (3) directly from the definition of the derivative (1) and the definite integral (2). We will see it is convenient to choose the same h in both limits. With $F'(x) = f(x)$, we have

$$\begin{aligned}\int_a^b f(s)ds &= \int_a^b F'(s)ds\\ &= \lim_{h\to 0}\sum_{n=1}^{N} F'\big(a+(n-1)h\big)\cdot h\\ &= \lim_{h\to 0}\sum_{n=1}^{N}\frac{F(a+nh) - F\big(a+(n-1)h\big)}{h}\cdot h\\ &= \lim_{h\to 0}\sum_{n=1}^{N} F(a+nh) - F\big(a+(n-1)h\big).\end{aligned}$$

The last expression has an interesting structure. All the values of $F(x)$ evaluated at the points lying between the endpoints a and b cancel each other in consecutive terms. Only the value $-F(a)$ survives when $n = 1$, and the value $+F(b)$ when $n = N$, yielding again (3).

0.8 Definite and indefinite integrals

The Riemann sum definition of an integral is called a *definite integral*. It is convenient to also define an indefinite integral by

$$\int f(x)dx = F(x),$$

where F(x) is the antiderivative of $f(x)$.

0.9 Indefinite integrals of elementary functions

From our known derivatives of elementary functions, we can determine some simple indefinite integrals. The power rule gives us

$$\int x^n dx = \frac{x^{n+1}}{n+1} + c, \quad n \neq -1.$$

When $n = -1$, and x is positive, we have

$$\int \frac{1}{x} dx = \ln x + c.$$

If x is negative, using the chain rule we have

$$\frac{d}{dx} \ln(-x) = \frac{1}{x}.$$

Therefore, since

$$|x| = \begin{cases} -x & \text{if } x < 0; \\ x & \text{if } x > 0, \end{cases}$$

we can generalize our indefinite integral to strictly positive or strictly negative x:

$$\int \frac{1}{x} dx = \ln |x| + c.$$

Trigonometric functions can also be integrated:

$$\int \cos x dx = \sin x + c, \quad \int \sin x dx = -\cos x + c.$$

Easily proved identities are an addition rule:

$$\int \big(f(x) + g(x)\big)dx = \int f(x)dx + \int g(x)dx;$$

and multiplication by a constant:

$$\int Af(x)dx = A \int f(x)dx.$$

This permits integration of functions such as

$$\int (x^2 + 7x + 2)dx = \frac{x^3}{3} + \frac{7x^2}{2} + 2x + c,$$

and

$$\int (5\cos x + \sin x)dx = 5\sin x - \cos x + c.$$

0.10 Substitution

More complicated functions can be integrated using the chain rule. Since

$$\frac{d}{dx}f\big(g(x)\big) = f'\big(g(x)\big)\cdot g'(x),$$

we have

$$\int f'\big(g(x)\big)\cdot g'(x)dx = f\big(g(x)\big) + c.$$

This integration formula is usually implemented by letting $y = g(x)$. Then one writes $dy = g'(x)dx$ to obtain

$$\begin{aligned}\int f'\big(g(x)\big)g'(x)dx &= \int f'(y)dy \\ &= f(y) + c \\ &= f\big(g(x)\big) + c.\end{aligned}$$

0.11 Integration by parts

Another integration technique makes use of the product rule for differentiation. Since

$$(fg)' = f'g + fg',$$

we have

$$f'g = (fg)' - fg'.$$

Therefore,

$$\int f'(x)g(x)dx = f(x)g(x) - \int f(x)g'(x)dx.$$

Commonly, the above integral is done by writing

$$\begin{aligned} u &= g(x) & dv &= f'(x)dx \\ du &= g'(x)dx & v &= f(x). \end{aligned}$$

Then, the formula to be memorized is

$$\int udv = uv - \int vdu.$$

0.12 Taylor series

A Taylor series of a function $f(x)$ about a point $x = a$ is a power series representation of $f(x)$ developed so that all the derivatives of $f(x)$ at a match all the derivatives of the power series. Without worrying about convergence here, we have

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \ldots.$$

Notice that the first term in the power series (the zeroth derivative of $f(x)$ at a) matches $f(a)$, all other terms vanishing, the second term matches $f'(a)$, all other terms vanishing, etc. Commonly, the Taylor series is developed with $a = 0$. We will

also make use of the Taylor series in a slightly different form, with $x = x_* + \epsilon$ and $a = x_*$:

$$f(x_* + \epsilon) = f(x_*) + f'(x_*)\epsilon + \frac{f''(x_*)}{2!}\epsilon^2 + \frac{f'''(x_*)}{3!}\epsilon^3 + \ldots.$$

Another way to view this series is that of $g(\epsilon) = f(x_* + \epsilon)$, expanded about $\epsilon = 0$.

Taylor series that are commonly used include

$$\begin{aligned}
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots, \\
\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots, \\
\cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots, \\
\frac{1}{1+x} &= 1 - x + x^2 - \ldots, \quad \text{for } |x| < 1, \\
\ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots, \quad \text{for } |x| < 1.
\end{aligned}$$

0.13 Functions of several variables

For simplicity, we consider a function $f = f(x, y)$ of two variables, though the results are easily generalized. The partial derivative of f with respect to x is defined as

$$\frac{\partial f}{\partial x} = \lim_{h \to 0} \frac{f(x+h, y) - f(x, y)}{h},$$

and similarly for the partial derivative of f with respect to y. To take the partial derivative of f with respect to x, say, take the derivative of f with respect to x holding y fixed. As an example, consider

$$f(x, y) = 2x^3y^2 + y^3.$$

We have

$$\frac{\partial f}{\partial x} = 6x^2y^2, \quad \frac{\partial f}{\partial y} = 4x^3y + 3y^2.$$

Second derivatives are defined as the derivatives of the first derivatives, so we have

$$\frac{\partial^2 f}{\partial x^2} = 12xy^2, \quad \frac{\partial^2 f}{\partial y^2} = 4x^3 + 6y;$$

and the mixed second partial derivatives are

$$\frac{\partial^2 f}{\partial x \partial y} = 12x^2y, \quad \frac{\partial^2 f}{\partial y \partial x} = 12x^2y.$$

In general, mixed partial derivatives are independent of the order in which the derivatives are taken.

Partial derivatives are necessary for applying the chain rule. Consider

$$df = f(x + dx, y + dy) - f(x, y).$$

We can write df as

$$\begin{aligned} df &= [f(x+dx, y+dy) - f(x, y+dy)] + [f(x, y+dy) - f(x,y)] \\ &= \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy. \end{aligned}$$

If one has $f = f(x(t), y(t))$, say, then

$$\frac{df}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}.$$

And if one has $f = f(x(r,\theta), y(r,\theta))$, say, then

$$\frac{\partial f}{\partial r} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial r}, \quad \frac{\partial f}{\partial \theta} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial \theta} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \theta}.$$

A Taylor series of a function of several variables can also be developed. Here, all partial derivatives of $f(x,y)$ at (a,b) match all the partial derivatives of the power series. With the notation

$$f_x = \frac{\partial f}{\partial x}, \quad f_y = \frac{\partial f}{\partial y}, \quad f_{xx} = \frac{\partial^2 f}{\partial x^2}, \quad f_{xy} = \frac{\partial^2 f}{\partial x \partial y}, \quad f_{yy} = \frac{\partial^2 f}{\partial y^2}, \quad \text{etc.},$$

we have

$$\begin{aligned} f(x,y) &= f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b) \\ &\quad + \frac{1}{2!}\left(f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b) + f_{yy}(a,b)(y-b)^2\right) + \ldots \end{aligned}$$

0.14 Complex numbers

View tutorial on YouTube: Complex Numbers
View tutorial on YouTube: Complex Exponential Function

We define the imaginary number i to be one of the two numbers that satisfies the rule $(i)^2 = -1$, the other number being $-i$. Formally, we write $i = \sqrt{-1}$. A complex number z is written as

$$z = x + iy,$$

where x and y are real numbers. We call x the real part of z and y the imaginary part and write

$$x = \operatorname{Re} z, \quad y = \operatorname{Im} z.$$

Two complex numbers are equal if and only if their real and imaginary parts are equal.

The complex conjugate of $z = x + iy$, denoted as $\bar{z}$, is defined as

$$\bar{z} = x - iy.$$

Using z and $\bar{z}$, we have

$$\operatorname{Re} z = \frac{1}{2}(z + \bar{z}), \quad \operatorname{Im} z = \frac{1}{2i}(z - \bar{z}). \tag{4}$$

Furthermore,

$$\begin{aligned} z\bar{z} &= (x+iy)(x-iy) \\ &= x^2 - i^2y^2 \\ &= x^2 + y^2; \end{aligned}$$

and we define the absolute value of z, also called the modulus of z, by

$$\begin{aligned} |z| &= (z\bar{z})^{1/2} \\ &= \sqrt{x^2+y^2}. \end{aligned}$$

We can add, subtract, multiply and divide complex numbers to get new complex numbers. With $z = x + iy$ and $w = s + it$, and x, y, s, t real numbers, we have

$$z + w = (x+s) + i(y+t); \quad z - w = (x-s) + i(y-t);$$

$$\begin{aligned} zw &= (x+iy)(s+it) \\ &= (xs - yt) + i(xt + ys); \end{aligned}$$

$$\begin{aligned} \frac{z}{w} &= \frac{z\bar{w}}{w\bar{w}} \\ &= \frac{(x+iy)(s-it)}{s^2+t^2} \\ &= \frac{(xs+yt)}{s^2+t^2} + i\frac{(ys-xt)}{s^2+t^2}. \end{aligned}$$

Furthermore,

$$\begin{aligned} |zw| &= \sqrt{(xs-yt)^2 + (xt+ys)^2} \\ &= \sqrt{(x^2+y^2)(s^2+t^2)} \\ &= |z||w|; \end{aligned}$$

and

$$\begin{aligned} \overline{zw} &= (xs-yt) - i(xt+ys) \\ &= (x-iy)(s-it) \\ &= \bar{z}\bar{w}. \end{aligned}$$

Similarly

$$\left|\frac{z}{w}\right| = \frac{|z|}{|w|}, \quad \overline{(\frac{z}{w})} = \frac{\bar{z}}{\bar{w}}.$$

Also, $\overline{z+w} = \bar{z} + \bar{w}$. However, $|z+w| \leq |z| + |w|$, a theorem known as the triangle inequality.

It is especially interesting and useful to consider the exponential function of an imaginary argument. Using the Taylor series expansion of an exponential function, we have

$$\begin{aligned} e^{i\theta} &= 1 + (i\theta) + \frac{(i\theta)^2}{2!} + \frac{(i\theta)^3}{3!} + \frac{(i\theta)^4}{4!} + \frac{(i\theta)^5}{5!} \cdots \\ &= \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \ldots\right) + i\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} + \ldots\right) \\ &= \cos\theta + i\sin\theta. \end{aligned}$$

Since we have determined that

$$\cos\theta = \mathrm{Re}\, e^{i\theta}, \quad \sin\theta = \mathrm{Im}\, e^{i\theta}, \tag{5}$$

we also have using (4) and (5), the frequently used expressions

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2}, \quad \sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}.$$

The much celebrated Euler's identity derives from $e^{i\theta} = \cos\theta + i\sin\theta$ by setting $\theta = \pi$, and using $\cos\pi = -1$ and $\sin\pi = 0$:

$$e^{i\pi} + 1 = 0,$$

and this identity links the five fundamental numbers—0, 1, i, e and π—using three basic mathematical operations—addition, multiplication and exponentiation—only once.

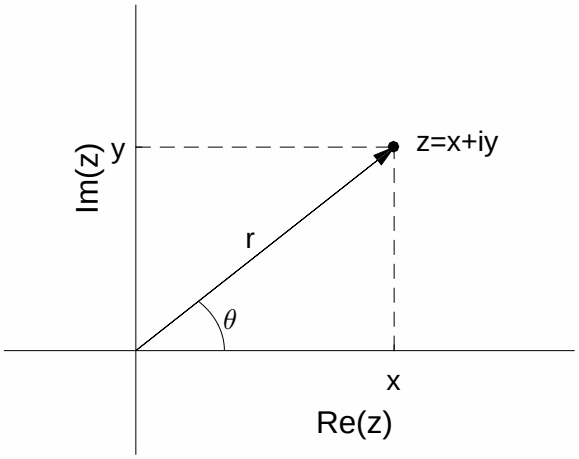


Figure 1: *The complex plane*

The complex number z can be represented in the complex plane with $\mathrm{Re}(z)$ as the x-axis and $\mathrm{Im}(z)$ as the y-axis (see Fig. 1). Using the definition of cosine and sine, we have $x = r\cos\theta$ and $y = r\sin\theta$, so that $z = r(\cos\theta + i\sin\theta)$. This leads to the polar representation

$$z = re^{i\theta},$$

where $r = |z|$ and $\tan\theta = y/x$. We define $\arg z = \theta$. Note that θ is not unique, though it is conventional to choose the value such that $-\pi < \theta \leq \pi$, and $\theta = 0$ when $r = 0$.

The polar form of a complex number can be useful when multiplying numbers. For example, if $z_1 = r_1 e^{i\theta_1}$ and $z_2 = r_2 e^{i\theta_2}$, then $z_1 z_2 = r_1 r_2 e^{i(\theta_1+\theta_2)}$. In particular, if $r_2 = 1$, then multiplication of z_1 by z_2 spins the representation of z_1 in the complex plane an angle θ_2 counterclockwise.

Useful trigonometric relations can be derived using $e^{i\theta}$ and properties of the exponential function. The addition law can be derived from

$$e^{i(x+y)} = e^{ix}e^{iy}.$$

We have

$$\begin{aligned}\cos(x+y)+i\sin(x+y) &= (\cos x + i\sin x)(\cos y + i\sin y)\\ &= (\cos x\cos y - \sin x\sin y) + i(\sin x\cos y + \cos x\sin y);\end{aligned}$$

yielding

$$\cos(x+y) = \cos x\cos y - \sin x\sin y, \quad \sin(x+y) = \sin x\cos y + \cos x\sin y.$$

De Moivre's Theorem derives from $e^{in\theta} = (e^{i\theta})^n$, yielding the identity

$$\cos(n\theta) + i\sin(n\theta) = (\cos\theta + i\sin\theta)^n.$$

For example, if $n = 2$, we derive

$$\begin{aligned}\cos 2\theta + i\sin 2\theta &= (\cos\theta + i\sin\theta)^2\\ &= (\cos^2\theta - \sin^2\theta) + 2i\cos\theta\sin\theta.\end{aligned}$$

Therefore,

$$\cos 2\theta = \cos^2\theta - \sin^2\theta, \quad \sin 2\theta = 2\cos\theta\sin\theta.$$

Example: Write $\sqrt{i}$ as a standard complex number

To solve this example, we first need to define what is meant by the square root of a complex number. The meaning of $\sqrt{z}$ is the complex number whose square is z. There will always be two such numbers, because $(\sqrt{z})^2 = (-\sqrt{z})^2 = z$. One can not define the positive square root because complex numbers are not defined as positive or negative.

We will show two methods to solve this problem. The first most straightforward method writes

$$\sqrt{i} = x + iy.$$

Squaring both sides, we obtain

$$i = x^2 - y^2 + 2xyi;$$

and equating the real and imaginary parts of this equation yields the two real equations

$$x^2 - y^2 = 0, \qquad 2xy = 1.$$

The first equation yields $y = \pm x$. With $y = x$, the second equation yields $2x^2 = 1$ with two solutions $x = \pm\sqrt{2}/2$. With $y = -x$, the second equation yields $-2x^2 = 1$, which has no solution for real x. We have therefore found that

$$\sqrt{i} = \pm\left(\frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}\right).$$

The second solution method makes use of the polar form of complex numbers. The algebra required for this method is somewhat simpler, especially for finding cube roots, fourth roots, etc. We know that $i = e^{i\pi/2}$, but more generally because of the periodic nature of the polar angle, we can write

$$i = e^{i(\frac{\pi}{2}+2\pi k)},$$

where k is an integer. We then have

$$\sqrt{i} = i^{1/2} = e^{i(\frac{\pi}{4}+\pi k)} = e^{i\pi k} e^{i\pi/4} = \pm e^{i\pi/4},$$

where we have made use of the usual properties of the exponential function, and $e^{i\pi k} = \pm 1$ for k even or odd. Converting back to standard form, we have

$$\sqrt{i} = \pm\left(\cos \pi/4 + i \sin \pi/4\right) = \pm\left(\frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}\right).$$

The fundamental theorem of algebra states that every polynomial equation of degree n has exactly n complex roots, counted with multiplicity. Two familiar examples would be $x^2 - 1 = (x+1)(x-1) = 0$, with two roots $x_1 = -1$ and $x_2 = 1$; and $x^2 - 2x + 1 = (x-1)^2 = 0$, with one root $x_1 = 1$ with multiplicity two.

The problem of finding the nth roots of unity is to solve the polynomial equation

$$z^n = 1$$

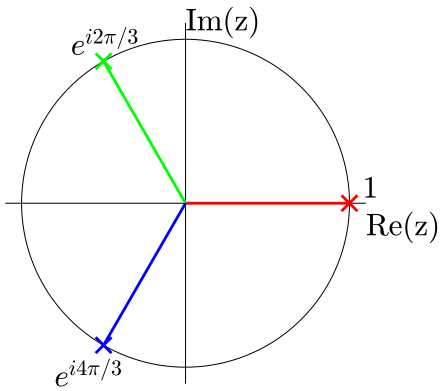
for the n complex values of z. We have $z_1 = 1$ for $n = 1$; and $z_1 = 1$, $z_2 = -1$ for $n = 2$. Beyond $n = 2$, some of the roots are complex and here we find the cube roots of unity, that is, the three values of z that satisfy $z^3 = 1$. Writing $1 = e^{i2\pi k}$, where k is an integer, we have

$$z = (1)^{1/3} = \left(e^{i2\pi k}\right)^{1/3} = e^{i2\pi k/3} = \begin{cases} 1; \\ e^{i2\pi/3}; \\ e^{i4\pi/3}. \end{cases}$$

Using $\cos(2\pi/3) = -1/2$, $\sin(2\pi/3) = \sqrt{3}/2$, $\cos(4\pi/3) = -1/2$, $\sin(4\pi/3) = -\sqrt{3}/2$, the three cube roots of unity are given by

$$z_1 = 1, \qquad z_2 = -\frac{1}{2} + i\frac{\sqrt{3}}{2}, \qquad z_3 = -\frac{1}{2} - i\frac{\sqrt{3}}{2}.$$

These three roots are evenly spaced around the unit circle in the complex plane, as shown in the figure below.



Part I

Linear algebra

The first part of this course is on linear algebra. We begin by introducing matrices and matrix algebra. Next, the important algorithms of Gaussian elimination and the LU-decomposition are presented and used to solve a system of linear equations and invert a matrix. We then discuss the abstract concept of vector and inner product spaces, and show how these concepts are related to matrices. Finally, a thorough presentation of determinants is given and the determinant is then used to solve the very important eigenvalue problem.

Chapter 1

Matrices

1.1 Definition of a matrix

View *Definition of a Matrix* on YouTube

An m-by-n matrix is a rectangular array of numbers (or other mathematical objects) with m rows and n columns. For example, a two-by-two matrix A, with two rows and two columns, looks like

$$\mathrm{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

(Sometimes brackets are used instead of parentheses.) The first row has elements a and b, the second row has elements c and d. The first column has elements a and c; the second column has elements b and d. As further examples, 2-by-3 and 3-by-2 matrices look like

$$\mathrm{B} = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}, \quad \mathrm{C} = \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix}.$$

Of special importance are the so-called row matrices and column matrices. These matrices are also called row vectors and column vectors. The row vector is in general 1-by-n and the column vector is n-by-1. For example, when $n = 3$, we would write

$$\mathrm{v} = \begin{pmatrix} a & b & c \end{pmatrix}$$

as a row vector, and

$$\mathrm{v} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

as a column vector.

1.2 Addition and multiplication of matrices

View *Addition & Multiplication of Matrices* on YouTube

Matrices can be added and multiplied. Matrices can be added only if they have the same dimension, and addition proceeds element by element. For example,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix}.$$

Multiplication of a matrix by a scalar is also simple. The rule is to just multiply every element of the matrix by the scalar. The 2-by-2 case is illustrated as

$$k \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} ka & kb \\ kc & kd \end{pmatrix}.$$

Matrix multiplication, however, is more complicated. Matrices can be multiplied only if the number of columns of the left matrix equals the number of rows of the right matrix. In other words, an m-by-n matrix on the left can be multiplied by an n-by-k matrix on the right. The result will be an m-by-k matrix. Evidently, matrix multiplication cannot commute for rectangular matrices. And in general, matrix multiplication doesn't commute for square matrices either.

We can illustrate matrix multiplication using two 2-by-2 matrices, writing

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}.$$

The standard way to multiply matrices is as follows. The first row of the left matrix is multiplied against and summed with the first column of the right matrix to obtain the element in the first row and first column of the product matrix. Next, the first row is multiplied against and summed with the second column; then the second row is multiplied against and summed with the first column; and finally the second row is multiplied against and summed with the second column.

In general, a particular element in the resulting product matrix, say in row k and column l, is obtained by multiplying and summing the elements in row k of the left matrix with the elements in column l of the right matrix.

Example: Consider the Fibonacci Q-matrix given by

$$\mathrm{Q} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

Determine Q^n in terms of the Fibonacci numbers.

The famous Fibonacci sequence is 1, 1, 2, 3, 5, 8, 13, ..., where each number in the sequence is the sum of the preceeding two numbers, and the first two numbers are set equal to one. With F_n the nth Fibonacci number, the mathematical definition is

$$F_{n+1} = F_n + F_{n-1}, \quad F_1 = F_2 = 1,$$

and we may define $F_0 = 0$ so that $F_0 + F_1 = F_2$.

Notice what happens when a matrix is multiplied by Q on the left:

$$\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a+c & b+d \\ a & b \end{pmatrix}.$$

The first row is replaced by the sum of the first and second rows, and the second row is replaced by the first row. Using the Fibonacci numbers, we can cleverly write the Fibonacci Q-matrix as

$$\mathrm{Q} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} F_2 & F_1 \\ F_1 & F_0 \end{pmatrix};$$

and then using the Fibonacci recursion relation we have

$$\mathrm{Q}^2 = \begin{pmatrix} F_3 & F_2 \\ F_2 & F_1 \end{pmatrix}, \quad \mathrm{Q}^3 = \begin{pmatrix} F_4 & F_3 \\ F_3 & F_2 \end{pmatrix}.$$

More generally, for $n \geq 1$,

$$\mathrm{Q}^n = \begin{pmatrix} F_{n+1} & F_n \\ F_n & F_{n-1} \end{pmatrix}.$$

1.3 The identity matrix and the zero matrix

View *Special Matrices* on YouTube

Two special matrices are the identity matrix, denoted by I, and the zero matrix, denoted simply by 0. The zero matrix can be m-by-n and is a matrix consisting of all zero elements. The identity matrix is a square matrix. If A and I are of the same size, then the identity matrix satisfies

$$\mathrm{AI} = \mathrm{IA} = \mathrm{A},$$

and plays the role of the number one in matrix multiplication. The identity matrix consists of ones along the diagonal (from top left to bottom right, sometimes called the main diagonal) and zeros elsewhere. For example, the 3-by-3 zero and identity matrices are given by

$$0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathrm{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and it is easy to check that

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}.$$

Although strictly speaking, the symbols 0 and I represent different matrices depending on their size, we will just use these symbols and leave their exact size to be inferred.

1.4 General notation, transposes, and inverses

View *Transpose Matrix* on YouTube
View *Inner and Outer Products* on YouTube
View *Inverse Matrix* on YouTube

A useful notation for writing a general m-by-n matrix A is

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}. \tag{1.1}$$

Here, the matrix element of A in the ith row and the jth column is denoted as a_{ij}.

Matrix multiplication can be written in terms of the matrix elements. Let A be an m-by-n matrix and let B be an n-by-p matrix. Then $C = \mathrm{AB}$ is an m-by-p matrix, and its ij element can be written as

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}. \tag{1.2}$$

Notice that the second index of a and the first index of b are summed over.

We can define the *transpose* of the matrix A, denoted by A^{T} and spoken as A-transpose, as the matrix for which the rows become the columns and the columns become the rows. Here, using (1.1),

$$\mathrm{A}^{\mathrm{T}} = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{pmatrix},$$

where we would write

$$a_{ij}^{\mathrm{T}} = a_{ji}.$$

Evidently, if A is m-by-n then A^{T} is n-by-m. As a simple example, view the following pair:

$$\mathrm{A} = \begin{pmatrix} a & d \\ b & e \\ c & f \end{pmatrix}, \quad \mathrm{A}^{\mathrm{T}} = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}. \tag{1.3}$$

A square matrix that satisfies $\mathrm{A}^{\mathrm{T}} = \mathrm{A}$ is called *symmetric*. For example the 3-by-3 matrix

$$\mathrm{A} = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}$$

is symmetric. A matrix that satisfies $\mathrm{A}^{\mathrm{T}} = -\mathrm{A}$ is called *skew symmetric*. For example,

$$\mathrm{A} = \begin{pmatrix} 0 & b & c \\ -b & 0 & e \\ -c & -e & 0 \end{pmatrix}$$

is skew symmetric. Notice that the diagonal elements must be zero. A sometimes useful fact is that every square matrix can be written as the sum of a symmetric and a skew-symmetric matrix using

$$\mathrm{A} = \frac{1}{2}\left(\mathrm{A} + \mathrm{A}^{\mathrm{T}}\right) + \frac{1}{2}\left(\mathrm{A} - \mathrm{A}^{\mathrm{T}}\right).$$

This is just like the fact that every function can be written as the sum of an even and an odd function.

How do we write the transpose of the product of two matrices? Let $[\mathrm{X}]_{ij}$ denote the element in row i and column j of the matrix X. Again, let A be an m-by-n matrix and B be an n-by-p matrix. Then

$$\left[(\mathrm{AB})^{\mathrm{T}}\right]_{ij} = [\mathrm{AB}]_{ji} = \sum_{k=1}^{n} a_{jk} b_{ki} = \sum_{k=1}^{n} b_{ik}^{\mathrm{T}} a_{kj}^{\mathrm{T}} = \left[\mathrm{B}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}\right]_{ij}.$$

Therefore,

$$(\mathrm{AB})^{\mathrm{T}} = \mathrm{B}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}.$$

In words, the transpose of the product of matrices is equal to the product of the transposes with the order of multiplication reversed.

The transpose of a column vector is a row vector. The inner product (or dot product) between two vectors is obtained by the product of a row vector and a

column vector, and is treated as a scalar (not a one-by-one matrix). With column vectors

$$\mathrm{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \quad \mathrm{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

the inner product between these two vectors becomes

$$\mathrm{u}^{\mathrm{T}}\mathrm{v} = \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = u_1v_1 + u_2v_2 + u_3v_3.$$

The norm-squared of a vector, or its magnitude squared, is defined as

$$\mathrm{u}^{\mathrm{T}}\mathrm{u} = \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = u_1^2 + u_2^2 + u_3^2,$$

whereas the norm of a vector, or its magnitude, is the positive square root of this quantity.

We say that two column vectors are *orthogonal* if their inner product is zero. We say that a column vector is *normalized* if it has a norm of one. A set of column vectors that are normalized and mutually orthogonal are said to be *orthonormal*.

When the vectors are complex, the inner product needs to be defined differently. Instead of a transpose of a matrix, one defines the conjugate transpose as the transpose together with taking the complex conjugate of every element of the matrix. The symbol used is that of a dagger, so that continuing the example from above,

$$\mathrm{u}^{\dagger} = \begin{pmatrix} \overline{u}_1 & \overline{u}_2 & \overline{u}_3 \end{pmatrix}.$$

Then

$$\mathrm{u}^{\dagger}\mathrm{u} = \begin{pmatrix} \overline{u}_1 & \overline{u}_2 & \overline{u}_3 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = |u_1|^2 + |u_2|^2 + |u_3|^2.$$

As noted above, when a real matrix is equal to its transpose we say that it is symmetric. When a complex matrix is equal to its conjugate transpose, we say that it is *Hermitian*. Hermitian matrices play a fundamental role in quantum physics.

An outer product is also defined, and is used in some applications. The outer product between u and v is given by

$$\mathrm{u}\mathrm{v}^{\mathrm{T}} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} = \begin{pmatrix} u_1v_1 & u_1v_2 & u_1v_3 \\ u_2v_1 & u_2v_2 & u_2v_3 \\ u_3v_1 & u_3v_2 & u_3v_3 \end{pmatrix}.$$

Notice that every column is a multiple of the single vector u, and every row is a multiple of the single vector v^{T}.

The transpose operation can also be used to make square matrices. If A is an m-by-n matrix, then A^{T} is n-by-m and $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ is an n-by-n matrix. For example, using (1.3), we have

$$\mathrm{A}^{\mathrm{T}}\mathrm{A} = \begin{pmatrix} a^2 + b^2 + c^2 & ad + be + cf \\ ad + be + cf & d^2 + e^2 + f^2 \end{pmatrix}$$

Notice that $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ is symmetric because

$$(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{\mathrm{T}} = \mathrm{A}^{\mathrm{T}}\mathrm{A}.$$

The trace of a square matrix A, denoted as Tr A, is the sum of the diagonal elements of A. So if A is an n-by-n matrix, then

$$\mathrm{Tr}\,\mathrm{A} = \sum_{i=1}^{n} a_{ii}.$$

Example: Let A **be an** m**-by-**n **matrix. Prove that** $\mathrm{Tr}(\mathrm{A}^{\mathrm{T}}\mathrm{A})$ **is the sum of the squares of all the elements of** A**.**

Note that $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ is an n-by-n matrix. We have

$$\begin{aligned}
\mathrm{Tr}(\mathrm{A}^{\mathrm{T}}\mathrm{A}) &= \sum_{i=1}^{n} (\mathrm{A}^{\mathrm{T}}\mathrm{A})_{ii} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^{\mathrm{T}} a_{ji} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ji} a_{ji} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^{2}.
\end{aligned}$$

Square matrices may also have inverses. Later, we will see that for a matrix to have an inverse its determinant, which we will define in general, must be nonzero. Here, if an n-by-n matrix A has an inverse, denoted as A^{-1}, then

$$\mathrm{A}\mathrm{A}^{-1} = \mathrm{A}^{-1}\mathrm{A} = \mathrm{I}.$$

If both the n-by-n matrices A and B have inverses then we can ask what is the inverse of the product of these two matrices? From the definition of an inverse,

$$(\mathrm{AB})^{-1}(\mathrm{AB}) = \mathrm{I}, \quad (\mathrm{AB})(\mathrm{AB})^{-1} = \mathrm{I}.$$

Either multiply the first equation on the right by B^{-1}, and then by A^{-1}, or multiply the second equation on the left by A^{-1}, and then by B^{-1}, to obtain

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

Again in words, the inverse of the product of matrices is equal to the product of the inverses with the order of multiplication reversed. Be careful here: this rule applies only if both matrices in the product are invertible.

Example: Assume that A **is an invertible matrix. Prove that** $(\mathrm{A}^{-1})^{\mathrm{T}} = (\mathrm{A}^{\mathrm{T}})^{-1}$**. In words: the transpose of the inverse matrix is the inverse of the transpose matrix.**

We know that

$$\mathrm{A}\mathrm{A}^{-1} = \mathrm{I} \quad \text{and} \quad \mathrm{A}^{-1}\mathrm{A} = \mathrm{I}.$$

Taking the transpose of these equations, and using $(\mathrm{AB})^{\mathrm{T}} = \mathrm{B}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}$ and $\mathrm{I}^{\mathrm{T}} = \mathrm{I}$, we obtain

$$(\mathrm{A}^{-1})^{\mathrm{T}}\mathrm{A}^{\mathrm{T}} = \mathrm{I} \quad \text{and} \quad \mathrm{A}^{\mathrm{T}}(\mathrm{A}^{-1})^{\mathrm{T}} = \mathrm{I}.$$

We can therefore conclude that $(\mathrm{A}^{-1})^{\mathrm{T}} = (\mathrm{A}^{\mathrm{T}})^{-1}$.

It is illuminating to derive the inverse of a two-by-two matrix. To find the inverse of A given by

$$\mathrm{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

the most direct approach would be to write

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and solve for x_1, x_2, y_1, and y_2. There are two inhomogeneous and two homogeneous equations given by

$$\begin{aligned} ax_1 + by_1 &= 1, \qquad & cx_1 + dy_1 &= 0, \\ cx_2 + dy_2 &= 1, \qquad & ax_2 + by_2 &= 0. \end{aligned}$$

To solve, we can eliminate y_1 and y_2 using the two homogeneous equations, and then solve for x_1 and x_2 using the two inhomogeneous equations. Finally, we use the two homogeneous equations to solve for y_1 and y_2. The solution for A^{-1} is found to be

$$\mathrm{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{1.4}$$

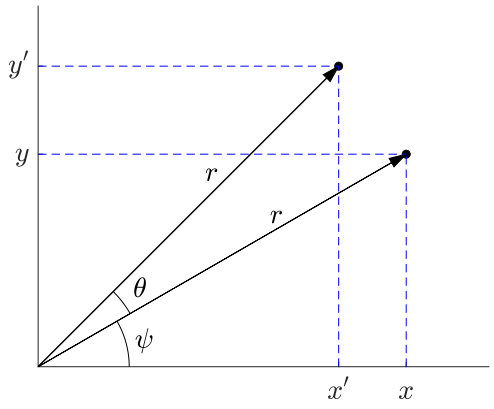
The factor in front of the matrix is the definition of the determinant for our two-by-two matrix A:

$$\det \mathrm{A} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

The determinant of a two-by-two matrix is the product of the diagonals minus the product of the off-diagonals. Evidently, A is invertible only if $\det \mathrm{A} \neq 0$. Notice that the inverse of a two-by-two matrix, in words, is found by switching the diagonal elements of the matrix, negating the off-diagonal elements, and dividing by the determinant. It can be useful in a linear algebra course to remember this formula.

1.5 Rotation matrices and orthogonal matrices

View *Rotation Matrix* on YouTube
View *Orthogonal Matrices* on YouTube



Rotating a vector in the x-y plane.

Consider the two-by-two rotation matrix that rotates a vector counterclockwise through an angle θ in the x-y plane, shown above. Trigonometry and the addi-

tion formula for cosine and sine results in

$$\begin{aligned} x' &= r\cos(\theta+\psi) \\ &= r(\cos\theta\cos\psi - \sin\theta\sin\psi) \\ &= x\cos\theta - y\sin\theta \end{aligned} \qquad \begin{aligned} y' &= r\sin(\theta+\psi) \\ &= r(\sin\theta\cos\psi + \cos\theta\sin\psi) \\ &= x\sin\theta + y\cos\theta. \end{aligned}$$

Writing the equations for x' and y' in matrix form, we have

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

The above two-by-two matrix is called a rotation matrix and is given by

$$\mathrm{R}_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Example: Find the inverse of the rotation matrix R_θ.

The inverse of R_θ rotates a vector clockwise by θ. To find R_θ^{-1}, we need only change $\theta \to -\theta$:

$$\mathrm{R}_\theta^{-1} = \mathrm{R}_{-\theta} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}.$$

This result agrees with (1.4) since $\det \mathrm{R}_\theta = 1$.

Notice that $\mathrm{R}_\theta^{-1} = \mathrm{R}_\theta^{\mathrm{T}}$. In general, a square n-by-n matrix Q with real entries that satisfies

$$\mathrm{Q}^{-1} = \mathrm{Q}^{\mathrm{T}}$$

is called an *orthogonal matrix*. Since $\mathrm{QQ}^{\mathrm{T}} = \mathrm{I}$ and $\mathrm{Q}^{\mathrm{T}}\mathrm{Q} = \mathrm{I}$, and since QQ^{T} multiplies the rows of Q against themselves (and summing the products), and $\mathrm{Q}^{\mathrm{T}}\mathrm{Q}$ multiplies the columns of Q against themselves, both the rows of Q and the columns of Q must form an orthonormal set of vectors (normalized and mutually orthogonal). For example, the column vectors of R, given by

$$\begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \quad \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix},$$

are orthonormal.

It is clear that rotating a vector around the origin doesn't change its length. More generally, orthogonal matrices preserve inner products. To prove, let Q be an orthogonal matrix and x a column vector. Then

$$(\mathrm{Qx})^{\mathrm{T}}(\mathrm{Qx}) = \mathrm{x}^{\mathrm{T}}\mathrm{Q}^{\mathrm{T}}\mathrm{Qx} = \mathrm{x}^{\mathrm{T}}\mathrm{x}.$$

The complex matrix analogue of an orthogonal matrix is a *unitary matrix* U. Here, the relationship is

$$\mathrm{U}^{-1} = \mathrm{U}^{\dagger}.$$

Like Hermitian matrices, unitary matrices also play a fundamental role in quantum physics.

1.6 Matrix representation of complex numbers

In our studies of complex numbers, we noted that multiplication of a complex number by $e^{i\theta}$ rotates that complex number an angle θ in the complex plane (counterclockwise if $\theta > 0$ and clockwise if $\theta < 0$). This leads to the idea that we might be able to represent complex numbers as matrices with $e^{i\theta}$ as the rotation matrix.

Accordingly, we begin by representing $e^{i\theta}$ as the rotation matrix, that is,

$$\begin{aligned} e^{i\theta} &= \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \\ &= \cos\theta \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \sin\theta \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \end{aligned}$$

Since $e^{i\theta} = \cos\theta + i\sin\theta$, we are led to the matrix representations of the unit numbers as

$$1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad i = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

A general complex number $z = x + iy$ is then represented as

$$z = \begin{pmatrix} x & -y \\ y & x \end{pmatrix}.$$

The complex conjugate operation, where $i \to -i$, is seen to be just the matrix transpose.

Example: Show that $i^2 = -1$ in the matrix representation.

We have

$$i^2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = -1.$$

Example: Show that $z\bar{z} = x^2 + y^2$ in the matrix representation.

We have

$$z\bar{z} = \begin{pmatrix} x & -y \\ y & x \end{pmatrix} \begin{pmatrix} x & y \\ -y & x \end{pmatrix} = \begin{pmatrix} x^2+y^2 & 0 \\ 0 & x^2+y^2 \end{pmatrix} = (x^2+y^2) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = (x^2+y^2).$$

We can now see that there is a one-to-one correspondence between the set of complex numbers and the set of all two-by-two matrices with equal diagonal elements and opposite signed off-diagonal elements. If you do not like the idea of $\sqrt{-1}$, then just imagine the arithmetic of these two-by-two matrices!

1.7 Permutation matrices

View *Permutation Matrices* on YouTube

A permutation matrix is another type of orthogonal matrix. When multiplied on the left, an n-by-n permutation matrix reorders the rows of an n-by-n matrix, and when multiplied on the right, reorders the columns. For example, let the string 12 represent the order of the rows (columns) of a two-by-two matrix. Then the two possible permutations of the rows (columns) are given by 12 and 21. The first permutation

is no permutation at all, and the corresponding permutation matrix is simply the identity matrix. The second permutation of the rows (columns) is achieved by

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} c & d \\ a & b \end{pmatrix}, \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} b & a \\ d & c \end{pmatrix}.$$

The rows (columns) of a 3-by-3 matrix has $3! = 6$ possible permutations, namely 123, 132, 213, 231, 312, 321. For example, the row permutation 312 is obtained by

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = \begin{pmatrix} g & h & i \\ a & b & c \\ d & e & f \end{pmatrix}.$$

Evidently, the permutation matrix is obtained by permuting the corresponding rows of the identity matrix, as seen by the identity $\mathrm{P} = \mathrm{PI}$. Because the columns and rows of the identity matrix are orthonormal, the permutation matrix is an orthogonal matrix.

1.8 Projection matrices

The two-by-two projection matrix projects a vector onto a specified vector in the x-y plane. Let $\mathbf{u}$ be a unit vector in $\mathbb{R}^2$. The projection of an arbitrary vector $\mathbf{x} = \langle x_1, x_2 \rangle$ onto the vector $\mathbf{u} = \langle u_1, u_2 \rangle$ is determined from

$$\mathrm{Proj}_{\mathbf{u}}(\mathbf{x}) = (\mathbf{x} \cdot \mathbf{u})\mathbf{u} = (x_1 u_1 + x_2 u_2)\langle u_1, u_2 \rangle.$$

In matrix form, this becomes

$$\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} u_1^2 & u_1 u_2 \\ u_1 u_2 & u_2^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

The projection matrix $\mathrm{P}_{\mathbf{u}}$, then, can be defined as

$$\begin{aligned} \mathrm{P}_{\mathbf{u}} &= \begin{pmatrix} u_1^2 & u_1 u_2 \\ u_1 u_2 & u_2^2 \end{pmatrix} \\ &= \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \begin{pmatrix} u_1 & u_2 \end{pmatrix} \\ &= \mathrm{u}\mathrm{u}^{\mathrm{T}}, \end{aligned}$$

which is an outer product. Notice that $\mathrm{P}_{\mathbf{u}}$ is symmetric.

Example: Show that $\mathrm{P}_{\mathbf{u}}^2 = \mathrm{P}_{\mathbf{u}}$.

It should be obvious that two projections is the same as one. To prove, we have

$$\begin{aligned} \mathrm{P}_{\mathbf{u}}^2 &= (\mathrm{u}\mathrm{u}^{\mathrm{T}})(\mathrm{u}\mathrm{u}^{\mathrm{T}}) \\ &= \mathrm{u}(\mathrm{u}^{\mathrm{T}}\mathrm{u})\mathrm{u}^{\mathrm{T}} && \text{(associative law)} \\ &= \mathrm{u}\mathrm{u}^{\mathrm{T}} && (\mathbf{u} \text{ is a unit vector}) \\ &= \mathrm{P}_{\mathbf{u}}. \end{aligned}$$

Chapter 2

Systems of linear equations

Consider the system of n linear equations and n unknowns, given by

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ \vdots \qquad\qquad & \qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n. \end{aligned}$$

We can write this system as the matrix equation

$$\mathrm{A}\mathrm{x} = \mathrm{b}, \tag{2.1}$$

with

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad \mathrm{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathrm{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}.$$

This chapter details the standard algorithm to solve (2.1) for the unknown vector x.

2.1 Gaussian Elimination

View *Gaussian Elimination* on YouTube

The standard algorithm to solve a system of linear equations is called Gaussian elimination. It is easiest to illustrate this algorithm by example.

Consider the linear system of equations given by

$$\begin{aligned} -3x_1 + 2x_2 - x_3 &= -1, \\ 6x_1 - 6x_2 + 7x_3 &= -7, \\ 3x_1 - 4x_2 + 4x_3 &= -6, \end{aligned} \tag{2.2}$$

which can be rewritten in matrix form as

$$\begin{pmatrix} -3 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -4 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 \\ -7 \\ -6 \end{pmatrix}.$$

To perform Gaussian elimination, we form what is called an *augmented matrix* by combining the matrix A with the column vector b:

$$\left(\begin{array}{ccc|c} -3 & 2 & -1 & -1 \\ 6 & -6 & 7 & -7 \\ 3 & -4 & 4 & -6 \end{array}\right).$$

Row reduction is then performed on this matrix. Allowed operations are (1) multiply any row by a nonzero constant, (2) add a multiple of one row to another row, (3)

interchange the order of any rows. It is easy to confirm that these operations do not change the solution of the original equations. The goal here is to convert the matrix A into a matrix with all zeros below the diagonal. This is called an *upper-triangular matrix*, from which one can quickly solve for the unknowns x.

We start with the first row of the matrix and work our way down as follows. The key element is called the pivot, which is the diagonal element that we use to zero all the elements below it. The pivot in the first row is the diagonal entry -3. To zero the 6 in the second row below the pivot, we multiply the first row by 2 and add it to the second row. To zero the 3 in the third row below the pivot, we add the first row to the third row:

$$\begin{pmatrix} -3 & 2 & -1 & -1 \\ 6 & -6 & 7 & -7 \\ 3 & -4 & 4 & -6 \end{pmatrix} \rightarrow \begin{pmatrix} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 0 & -2 & 3 & -7 \end{pmatrix}.$$

We then go to the second row. The new pivot is the number -2 in the diagonal of the second row. To zero the -2 below the pivot, we multiply the second row by -1 and add it to the third row:

$$\begin{pmatrix} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 0 & -2 & 3 & -7 \end{pmatrix} \rightarrow \begin{pmatrix} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 0 & 0 & -2 & 2 \end{pmatrix}.$$

The original matrix A is now upper triangular, and the transformed equations can be determined from the augmented matrix as

$$\begin{aligned} -3x_1 + 2x_2 - x_3 &= -1, \\ -2x_2 + 5x_3 &= -9, \\ -2x_3 &= 2. \end{aligned}$$

These equations can be solved by back substitution, starting from the last equation and working backwards. We have

$$\begin{aligned} x_3 &= -\frac{1}{2}(2) = -1 \\ x_2 &= -\frac{1}{2}(-9 - 5x_3) = 2, \\ x_1 &= -\frac{1}{3}(-1 - 2x_2 + x_3) = 2. \end{aligned}$$

Therefore,

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix}.$$

2.2 When there is no unique solution

Given n equations and n unknowns, one usually expects a unique solution. But two other possibilities exist: there could be no solution, or an infinite number of solutions. We will illustrate what happens during Gaussian elimination in these two cases. Consider

$$\begin{aligned} -3x_1 + 2x_2 - x_3 &= -1, \\ 6x_1 - 6x_2 + 7x_3 &= -7, \\ 3x_1 - 4x_2 + 6x_3 &= b. \end{aligned}$$

Note that the first two equations are the same as in (2.2), but the left-hand-side of the third equation has been replaced by the sum of the left-hand-sides of the first two equations, and the right-hand-side has been replaced by the parameter b. If $b = -8$, then the third equation is just the sum of the first two equations and adds no new information to the system. In this case, the equations should admit an infinite number of solutions. However, if $b \neq -8$, then the third equation is inconsistent with the first two equations and there should be no solution.

We solve by Gaussian elimination to see how it plays out. Writing the augmented matrix and doing row elimination, we have

$$\begin{pmatrix} -3 & 2 & -1 & -1 \\ 6 & -6 & 7 & -7 \\ 3 & -4 & 6 & b \end{pmatrix} \rightarrow \begin{pmatrix} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 0 & -2 & 5 & b-1 \end{pmatrix} \rightarrow \begin{pmatrix} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 0 & 0 & 0 & b+8 \end{pmatrix}.$$

Evidently, Gaussian elimination has reduced the last row of the matrix A to zeros, and the last equation becomes

$$0 = b + 8.$$

If $b \neq -8$, there will be no solution, and if $b = -8$, the under-determined systems of equations becomes

$$\begin{aligned} -3x_1 + 2x_2 - x_3 &= -1 \\ -2x_2 + 5x_3 &= -9. \end{aligned}$$

The unknowns x_1 and x_2 can be solved in terms of x_3 as

$$x_1 = \frac{10}{3} + \frac{4}{3}x_3, \quad x_2 = \frac{9}{2} + \frac{5}{2}x_3,$$

indicating an infinite family of solutions dependent on the free choice of x_3.

To be clear, for a linear system represented by $\mathrm{Ax} = \mathrm{b}$, if there is a unique solution then A is invertible and the solution is given formally by

$$\mathrm{x} = \mathrm{A}^{-1}\mathrm{b}.$$

If there is not a unique solution, then A is not invertible. We then say that the matrix A is singular. Whether or not an n-by-n matrix A is singular can be determined by row reduction on A. After row reduction, if the last row of A is all zeros, then A is a singular matrix; if not, then A is an invertible matrix. We have already shown in the two-by-two case, that A is invertible if and only if $\det \mathrm{A} \neq 0$, and we will later show that this is also true for n-by-n matrices.

2.3 Reduced row echelon form

View *Reduced Row Echelon Form* on YouTube

If we continue the row elimination procedure so that all the pivots are one, and all the entries in the columns above and below the pivots are zero, then the resulting matrix is in the so-called *reduced row echelon form*. We write the reduced row echelon form of a matrix A as $\mathrm{rref}(\mathrm{A})$. If A is an invertible square matrix, then $\mathrm{rref}(\mathrm{A}) = \mathrm{I}$.

Instead of Gaussian elimination and back substitution, a system of equations can be solved by bringing a matrix to reduced row echelon form. We can illustrate this

by solving again our first example. Beginning with the same augmented matrix, we have

$$\begin{pmatrix} -3 & 2 & -1 & -1 \\ 6 & -6 & 7 & -7 \\ 3 & -4 & 4 & -6 \end{pmatrix} \to \begin{pmatrix} -3 & 2 & -1 & -1 \\ 0 & -2 & 5 & -9 \\ 0 & -2 & 3 & -7 \end{pmatrix} \to \begin{pmatrix} -3 & 0 & 4 & -10 \\ 0 & -2 & 5 & -9 \\ 0 & 0 & -2 & 2 \end{pmatrix}$$
$$\to \begin{pmatrix} -3 & 0 & 4 & -10 \\ 0 & -2 & 5 & -9 \\ 0 & 0 & 1 & -1 \end{pmatrix} \to \begin{pmatrix} -3 & 0 & 0 & -6 \\ 0 & -2 & 0 & -4 \\ 0 & 0 & 1 & -1 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Once A has been transformed into the identity matrix, the resulting system of equations is just the solution, that is, $x_1 = 2$, $x_2 = 2$ and $x_3 = -1$.

2.4 Computing inverses

View *Computing Inverses* on YouTube

Calculating the reduced row echelon form of an n-by-n invertible matrix A can be used to compute the inverse matrix A^{-1}.

For example, recall how we found the general inverse of a two-by-two matrix by writing $\mathrm{AA}^{-1} = \mathrm{I}$, that is,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This single matrix equation is equivalent to two sets of two equations and two unknowns, namely

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

We can solve these two equations by bringing A to reduced row echelon form. There is no point in doing this twice, so instead we form a doubly augmented matrix and go to work on that:

$$\begin{pmatrix} a & b & 1 & 0 \\ c & d & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & b/a & 1/a & 0 \\ c & d & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & b/a & 1/a & 0 \\ 0 & \frac{ad-bc}{a} & -c/a & 1 \end{pmatrix} \to$$
$$\begin{pmatrix} 1 & b/a & 1/a & 0 \\ 0 & 1 & -\frac{c}{ad-bc} & \frac{a}{ad-bc} \end{pmatrix} \to \begin{pmatrix} 1 & 0 & \frac{d}{ad-bc} & -\frac{b}{ad-bc} \\ 0 & 1 & -\frac{c}{ad-bc} & \frac{a}{ad-bc} \end{pmatrix}.$$

The third column of the reduced matrix corresponds to the first column of the inverse matrix, and the fourth column of the reduced matrix correponds to the second column of the inverse matrix. Therefore, we have rederived

$$\mathrm{A}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

In other words, by moving A to reduced row echelon form while simultaneously performing the same operations on the identity matrix I, we achieve the following transformation:

$$\begin{pmatrix} \mathrm{A} & \mathrm{I} \end{pmatrix} \to \begin{pmatrix} \mathrm{I} & \mathrm{A}^{-1} \end{pmatrix}.$$

To illustrate this algorithm further, we find the inverse of the three-by-three matrix used in our first example. We have

$$\begin{pmatrix} -3 & 2 & -1 & 1 & 0 & 0 \\ 6 & -6 & 7 & 0 & 1 & 0 \\ 3 & -4 & 4 & 0 & 0 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} -3 & 2 & -1 & 1 & 0 & 0 \\ 0 & -2 & 5 & 2 & 1 & 0 \\ 0 & -2 & 3 & 1 & 0 & 1 \end{pmatrix} \rightarrow$$
$$\begin{pmatrix} -3 & 0 & 4 & 3 & 1 & 0 \\ 0 & -2 & 5 & 2 & 1 & 0 \\ 0 & 0 & -2 & -1 & -1 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} -3 & 0 & 0 & 1 & -1 & 2 \\ 0 & -2 & 0 & -1/2 & -3/2 & 5/2 \\ 0 & 0 & -2 & -1 & -1 & 1 \end{pmatrix} \rightarrow$$
$$\begin{pmatrix} 1 & 0 & 0 & -1/3 & 1/3 & -2/3 \\ 0 & 1 & 0 & 1/4 & 3/4 & -5/4 \\ 0 & 0 & 1 & 1/2 & 1/2 & -1/2 \end{pmatrix};$$

and one can check that

$$\begin{pmatrix} -3 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -4 & 4 \end{pmatrix} \begin{pmatrix} -1/3 & 1/3 & -2/3 \\ 1/4 & 3/4 & -5/4 \\ 1/2 & 1/2 & -1/2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It is also interesting to check that the solution to the equation $Ax = b$ is $x = A^{-1}b$. Using the b from our first example, we have

$$x = \begin{pmatrix} -1/3 & 1/3 & -2/3 \\ 1/4 & 3/4 & -5/4 \\ 1/2 & 1/2 & -1/2 \end{pmatrix} \begin{pmatrix} -1 \\ -7 \\ -6 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix},$$

as obtained previously.

2.5 LU decomposition

View *LU Decomposition* on YouTube
View *Solving LUx = b* on YouTube

The process of Gaussian elimination also results in the factoring of the matrix A to

$$A = LU,$$

where L is a lower triangular matrix and U is an upper triangular matrix. Using the same matrix A as in the last section, we show how this factorization is realized. We have

$$\begin{pmatrix} -3 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -4 & 4 \end{pmatrix} \rightarrow \begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 3 & -4 & 4 \end{pmatrix} = M_1A,$$

where

$$M_1A = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -3 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -4 & 4 \end{pmatrix} = \begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 3 & -4 & 4 \end{pmatrix}.$$

Note that the matrix M_1 performs row elimination on the second row using the first row. Two times the first row is added to the second row.

The next step is

$$\begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 3 & -4 & 4 \end{pmatrix} \rightarrow \begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 0 & -2 & 3 \end{pmatrix} = M_2M_1A,$$

where

$$M_2M_1A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 3 & -4 & 4 \end{pmatrix} = \begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 0 & -2 & 3 \end{pmatrix}.$$

Note that the matrix M_2 performs row elimination on the third row using the first row. One times the first row is added to the third row.

The last step is

$$\begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 0 & -2 & 3 \end{pmatrix} \to \begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 0 & 0 & -2 \end{pmatrix} = M_3M_2M_1A,$$

where

$$M_3M_2M_1A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 0 & -2 & 3 \end{pmatrix} = \begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 0 & 0 & -2 \end{pmatrix}.$$

Here, M_3 performs row elimination on the third row using the second row. Minus one times the second row is added to the third row. We now have

$$M_3M_2M_1A = U$$

or

$$A = M_1^{-1}M_2^{-1}M_3^{-1}U.$$

The inverse matrices are easy to find. The matrix M_1 multiples the first row by 2 and adds it to the second row. To invert this operation, we simply need to multiply the first row by -2 and add it to the second row, so that

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad M_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

To check that

$$M_1M_1^{-1} = I,$$

we multiply

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Similarly,

$$M_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad M_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix},$$

and

$$M_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}, \quad M_3^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

Therefore,

$$L = M_1^{-1}M_2^{-1}M_3^{-1}$$

is given by

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix},$$

which is lower triangular. Notice that the off-diagonal elements of M_1^{-1}, M_2^{-1}, and M_3^{-1} are simply combined to form L. Without actually multiplying matrices, one could obtain this result by considering how an elementary matrix performs row reduction on another elementary matrix. Our LU decomposition is therefore

$$\begin{pmatrix} -3 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -4 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 0 & 0 & -2 \end{pmatrix}.$$

Another nice feature of the LU decomposition, if done by computer, is that A can be overwritten, therefore saving memory if the matrix A is very large.

The LU decomposition is useful when one needs to solve $Ax = b$ for x when A is fixed and there are many different b's. First one determines L and U using Gaussian elimination. Then one writes

$$(LU)x = L(Ux) = b.$$

We let

$$y = Ux,$$

and first solve

$$Ly = b$$

for y by forward substitution, starting from the first equation and working forward to complete the solution. We then solve

$$Ux = y$$

for x by back substitution. If we count operations, we can show that solving $(LU)x = b$ is a factor of n faster once L and U are in hand than solving $Ax = b$ directly by Gaussian elimination.

We now illustrate the solution of $LUx = b$ using our previous example, where

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 0 & 0 & -2 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ -7 \\ -6 \end{pmatrix}.$$

With $y = Ux$, we first solve $Ly = b$, that is

$$\begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -1 \\ -7 \\ -6 \end{pmatrix}.$$

Using forward substitution

$$\begin{aligned} y_1 &= -1, \\ y_2 &= -7 + 2y_1 = -9, \\ y_3 &= -6 + y_1 - y_2 = 2. \end{aligned}$$

We now solve Ux = y, that is

$$\begin{pmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 0 & 0 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 \\ -9 \\ 2 \end{pmatrix}.$$

Using back substitution,

$$\begin{aligned} x_3 &= -\frac{1}{2}(2) = -1, \\ x_2 &= -\frac{1}{2}(-9 - 5x_3) = 2, \\ x_1 &= -\frac{1}{3}(-1 - 2x_2 + x_3) = 2, \end{aligned}$$

and we have once again determined

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix}.$$

When performing Gaussian elimination, recall that the diagonal element that one uses during the elimination procedure is called the pivot. To obtain the correct multiple, one uses the pivot as the divisor to the elements below the pivot. Gaussian elimination in this form will fail if the pivot is zero. In this case, a row interchange must be performed.

Even if the pivot is not identically zero, a small value can result in an unstable numerical computation. For large matrices solved by a computer, one can easily lose all accuracy in the solution. To avoid these round-off errors arising from small pivots, row interchanges are made, and the numerical technique is called partial pivoting. This method of LU decomposition with partial pivoting is the one usually taught in a standard numerical analysis course.

Chapter 3

Vector spaces

Linear algebra abstracts the vector concept, introducing new vocabulary and definitions that are widely used by scientists and engineers. Vector spaces, subspaces, inner product spaces, linear combinations, linear independence, linear dependence, span, basis, dimension, norm, unit vectors, orthogonal, orthonormal: this is the vocabulary that you need to know.

3.1 Vector spaces

View *Vector Spaces* on YouTube

In multivariable, or vector calculus, a vector is defined to be a mathematical construct that has both direction and magnitude. In linear algebra, vectors are defined more abstractly. Vectors are mathematical constructs that can be added and multiplied by scalars under the usual rules of arithmetic. Vector addition is commutative and associative, and scalar multiplication is distributive and associative. Let u, v, and w be vectors, and let a, b, and c be scalars. Then the rules of arithmetic say that

$$\mathrm{u}+\mathrm{v}=\mathrm{v}+\mathrm{u}, \qquad \mathrm{u}+(\mathrm{v}+\mathrm{w})=(\mathrm{u}+\mathrm{v})+\mathrm{w};$$

and

$$a(\mathrm{u}+\mathrm{v})=a\mathrm{u}+a\mathrm{v}, \qquad a(b\mathrm{u})=(ab)\mathrm{u}.$$

A *vector space* consists of a set of vectors and a set of scalars that is closed under vector addition and scalar multiplication. That is, when you multiply any two vectors in a vector space by scalars and add them, the resulting vector is still in the vector space.

We can give some examples of vector spaces. Let the scalars be the set of real numbers and let the vectors be column matrices of a specified type. One example of a vector space is the set of all three-by-one column matrices. If we let

$$\mathrm{u}=\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \quad \mathrm{v}=\begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

then

$$\mathrm{w}=a\mathrm{u}+b\mathrm{v}=\begin{pmatrix} au_1+bv_1 \\ au_2+bv_2 \\ au_3+bv_3 \end{pmatrix}$$

is evidently a three-by-one matrix, so that the set of all three-by-one matrices (together with the set of real numbers) forms a vector space. This vector space is usually called $\mathbb{R}^3$, which maps one-to-one with the three-dimensional vectors of Vector Calculus.

A vector subspace is a vector space that is a subset of another vector space. For example, a vector subspace of $\mathbb{R}^3$ could be the set of all three-by-one matrices with

zero in the third row. If we let

$$\mathrm{u} = \begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix}, \quad \mathrm{v} = \begin{pmatrix} v_1 \\ v_2 \\ 0 \end{pmatrix},$$

then

$$\mathrm{w} = a\mathrm{u} + b\mathrm{v} = \begin{pmatrix} au_1 + bv_1 \\ au_2 + bv_2 \\ 0 \end{pmatrix}$$

is evidently also a three-by-one matrix with zero in the third row. This subspace of $\mathbb{R}^3$ is closed under scalar multiplication and vector addition and is therefore a vector space. This vector space is usually called $\mathbb{R}^2$.

Another example of a vector subspace of $\mathbb{R}^3$ would be the set of all three-by-one matrices where the first row is equal to the third row. Two vectors in this subspace could be

$$\mathrm{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_1 \end{pmatrix}, \quad \mathrm{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_1 \end{pmatrix}.$$

Then

$$\mathrm{w} = a\mathrm{u} + b\mathrm{v} = \begin{pmatrix} au_1 + bv_1 \\ au_2 + bv_2 \\ au_1 + bv_1, \end{pmatrix}$$

is also a three-by-one matrix with its first row equal to its third row, so that this subspace is also closed under scalar multiplication and vector addition.

Of course, not all subsets of $\mathbb{R}^3$ form a vector space. A simple example would be the set of all three-by-one matrices where the row elements sum to one. If, say, $\mathrm{u} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^{\mathrm{T}}$, then au is a vector whose rows sum to a, which can be different than one.

The zero vector must be a member of every vector space. If u is in the vector space, then so is 0u which is just the zero vector. Another argument would be that if u is in the vector space, then so is $(-1)\mathrm{u} = -\mathrm{u}$, and $\mathrm{u} - \mathrm{u}$ is again equal to the zero vector.

The concept of vector spaces is more general than a set of column matrices. Here are some examples where the vectors are functions.

Example: Consider vectors consisting of all real polynomials in x of degree less than or equal to n. Show that this set of vectors (together with the set of real numbers) form a vector space.

Consider the polynomials of degree less than or equal to n given by

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n, \qquad q(x) = b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n,$$

where $a_0, a_1, \ldots, a_n$ and $b_0, b_1, \ldots, b_n$ are real numbers. Clearly, multiplying these polynmials by real numbers still results in a polynomial of degree less than or equal to n. Adding these polynomials results in

$$p(x) + q(x) = (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2 + \cdots + (a_n + b_n)x^n,$$

which is another polynomial of degree less than or equal to n. Since this set of polynomials is closed under scalar multiplication and vector addition, it forms a vector space. This vector space is designated as $\mathbb{P}_n$.

Example: Consider a function $y = y(x)$ and the differential equation $d^3y/dx^3 = 0$. Find the vector space associated with the general solution of this differential equation.

From Calculus, we know that the function whose third derivative is zero is a polynomial of degree less than or equal to two. That is, the general solution to the differential equation is

$$y(x) = a_0 + a_1 x + a_2 x^2,$$

which is just all possible vectors in the vector space $\mathbb{P}_2$.

Example: Consider a function $y = y(x)$ and the differential equation $d^2y/dx^2 + y = 0$. Find the vector space associated with the general solution of this differential equation.

Again from Calculus, we know that the trigonometric functions $\cos x$ and $\sin x$ have second derivatives that are the negative of themselves. The general solution to the differential equation consists of all vectors of the form

$$y(x) = a\cos x + b\sin x,$$

which is just all possible vectors in the vector space consisting of a linear combination of $\cos x$ and $\sin x$.

3.2 Linear independence

View *Linear Independence* on YouTube

A set of vectors, $\{\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_n\}$, is said to be *linearly independent* if for any scalars c_1, c_2, $\ldots$, c_n, the equation

$$c_1\mathrm{u}_1 + c_2\mathrm{u}_2 + \cdots + c_n\mathrm{u}_n = 0$$

has only the solution $c_1 = c_2 = \cdots = c_n = 0$. That is, a set of vectors is linearly independent if one is unable to write any of the vectors $\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_n$ as a linear combination of any of the other vectors. For instance, if there was a solution to the above equation with $c_1 \neq 0$, then we could solve that equation for u_1 in terms of the other vectors with nonzero coefficients.

As an example consider whether the following three three-by-one column vectors are linearly independent:

$$\mathrm{u} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathrm{v} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathrm{w} = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix}.$$

Indeed, they are not linearly independent, that is, they are *linearly dependent*, because w can be written in terms of u and v. In fact, $\mathrm{w} = 2\mathrm{u} + 3\mathrm{v}$. Now consider the three three-by-one column vectors given by

$$\mathrm{u} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathrm{v} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathrm{w} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

These three vectors are linearly independent because you cannot write any one of these vectors as a linear combination of the other two. If we go back to our definition of linear independence, we can see that the equation

$$a\mathrm{u} + b\mathrm{v} + c\mathrm{w} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

has as its only solution $a = b = c = 0$.

3.3 Span, basis and dimension

View *Span, Basis and Dimension* on YouTube

Given a set of vectors, one can generate a vector space by forming all linear combinations of that set of vectors. The *span* of the set of vectors $\{\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_n\}$ is the vector space consisting of all linear combinations of $\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_n$. We say that a set of vectors spans a vector space.

For example, the set of three-by-one column matrices given by

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} \right\}$$

spans the vector space of all three-by-one matrices with zero in the third row. This vector space is a *vector subspace* of all three-by-one matrices.

One doesn't need all three of these vectors to span this vector subspace because any one of these vectors is linearly dependent on the other two. The smallest set of vectors needed to span a vector space forms a *basis* for that vector space. Here, given the set of vectors above, we can construct a basis for the vector subspace of all three-by-one matrices with zero in the third row by simply choosing two out of three vectors from the above spanning set. Three possible bases are given by the sets

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}, \quad \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} \right\}, \quad \left\{ \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix} \right\}.$$

Although all three combinations form a basis for the vector subspace, the first combination is usually preferred because this is an orthonormal basis. The vectors in this basis are mutually orthogonal and of unit norm.

The number of vectors in a basis gives the dimension of the vector space. Here, the dimension of the vector space of all three-by-one matrices with zero in the third row is two.

Example: Find an orthonormal basis for the set of all three-by-one matrices where the first row is equal to the third row.

There are many different solutions to this example, but a rather simple orthonormal basis is given by the set

$$\left\{ \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \frac{\sqrt{2}}{2} \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

Any other three-by-one matrix with first row equal to third row can be written as a linear combination of these two basis vectors, and the dimension of this vector

space is also two.

Example: Determine a basis for $\mathbb{P}_2$, the vector space consisting of all polynomials of degree less than or equal to two. Again, there are many possible choices for a basis, but perhaps the simplest one is given by the set

$$\left\{1, \quad x, \quad x^2\right\}.$$

Clearly, any polynomial of degree less than or equal to two can be written as a linear combination of these basis vectors. The dimension of $\mathbb{P}_2$ is three.

Example: Determine a basis for the vector space given by the general solution of the differential equation $d^2y/dx^2 + y = 0$. The general solution is given by

$$y(x) = a\cos x + b\sin x,$$

and a basis for this vector space are just the set of functions

$$\{\cos x, \quad \sin x\}.$$

The dimension of the vector space given by the general solution of the differential equation is two. This dimension is equal to the order of the highest derivative in the differential equation.

3.4 Inner product spaces

We have discussed the inner product (or dot product) between two column matrices. Recall that the inner product between, say, two three-by-one column matrices

$$\mathrm{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \qquad \mathrm{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

is given by

$$\mathrm{u}^{\mathrm{T}}\mathrm{v} = u_1v_1 + u_2v_2 + u_3v_3.$$

We now generalize the inner product so that it is applicable to any vector space, including those containing functions.

We will denote the inner product between any two vectors u and v as (u, v), and require the inner product to satisfy the same arithmetic rules that are satisfied by the dot product. With u, v, w vectors and c a scalar, these rules can be written as

$$(\mathrm{u}, \mathrm{v}) = (\mathrm{v}, \mathrm{u}), \quad (\mathrm{u} + \mathrm{v}, \mathrm{w}) = (\mathrm{u}, \mathrm{w}) + (\mathrm{v}, \mathrm{w}), \quad (\mathrm{cu}, \mathrm{v}) = \mathrm{c}(\mathrm{u}, \mathrm{v}) = (\mathrm{u}, \mathrm{cv});$$

and $(\mathrm{u}, \mathrm{u}) \geq 0$, where the equality holds if and only if $\mathrm{u} = 0$.

Generalizing our definitions for column matrices, the *norm* of a vector u is defined as

$$||\mathrm{u}|| = (\mathrm{u}, \mathrm{u})^{1/2}.$$

A *unit vector* is a vector whose norm is one. Unit vectors are said to be *normalized to unity*, though sometimes we just say that they are *normalized*. We say two vectors are *orthogonal* if their inner product is zero. We also say that a basis is *orthonormal* (as in

an orthonormal basis) if all the vectors are mutually orthogonal and are normalized to unity. For an orthonormal basis consisting of the vectors $v_1, v_2, \ldots, v_n$, we write

$$(v_i, v_j) = \delta_{ij},$$

where δ_{ij} is called the Kronecker delta, defined as

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{if } i \neq j. \end{cases}$$

Oftentimes, basis vectors are used that are orthogonal but are normalized to other values besides unity.

Example: Define an inner product for $\mathbb{P}_n$.

Let $p(x)$ and $q(x)$ be two polynomials in $\mathbb{P}_n$. One possible definition of an inner product is given by

$$(p, q) = \int_{-1}^{1} p(x)q(x)dx.$$

You can check that all the conditions of an inner product are satisfied.

Example: Show that the first four Legendre polynomials form an orthogonal basis for $\mathbb{P}_3$ using the inner product defined above.

The first four Legendre polynomials are given by

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2 - 1), \quad P_3(x) = \frac{1}{2}(5x^3 - 3x),$$

and these four polynomials form a basis for $\mathbb{P}_3$. With an inner product defined on $\mathbb{P}_n$ as

$$(p, q) = \int_{-1}^{1} p(x)q(x)dx,$$

it can be shown by explicit integration that

$$(P_m, P_n) = \frac{2}{2n+1}\delta_{m,n},$$

so that the first four Legendre polynomials are mutually orthogonal. They are normalized so that $P_n(1) = 1$.

Example: Define an inner product on $\mathbb{P}_n$ such that the Hermite polynomials are orthogonal.

For instance, the first four Hermite polynomials are given by

$$H_0(x) = 1, \quad H_1(x) = 2x, \quad H_2(x) = 4x^2 - 2, \quad H_3(x) = 8x^3 - 12x,$$

which also form a basis for $\mathbb{P}_3$. Here, define an inner product on $\mathbb{P}_n$ as

$$(p, q) = \int_{-\infty}^{\infty} p(x)q(x)e^{-x^2}dx.$$

It can be shown that

$$(H_m, H_n) = 2^n \pi^{1/2} n! \delta_{m,n},$$

so that the Hermite polynomials are orthogonal with this definition of the inner product. These Hermite polynomials are normalized so that the leading coefficient of H_n is given by 2^n.

3.5 Vector spaces of a matrix

3.5.1 Null space

View *Null Space* on YouTube

The null space of a matrix A is the vector space spanned by all vectors x that satisfy the matrix equation

$$\mathrm{A}\mathrm{x} = 0.$$

If the matrix A is m-by-n, then the column vector x is n-by-one and the null space of A is a subspace of $\mathbb{R}^n$. If A is a square invertible matrix, then the null space consists of just the zero vector.

To find a basis for the null space of a noninvertible matrix, we bring A to row reduced echelon form. We demonstrate by example. Consider the three-by-five matrix given by

$$\mathrm{A} = \begin{pmatrix} -3 & 6 & -1 & 1 & -7 \\ 1 & -2 & 2 & 3 & -1 \\ 2 & -4 & 5 & 8 & -4 \end{pmatrix}.$$

By judiciously permuting rows to simplify the arithmetic, one pathway to construct rref(A) is

$$\begin{pmatrix} -3 & 6 & -1 & 1 & -7 \\ 1 & -2 & 2 & 3 & -1 \\ 2 & -4 & 5 & 8 & -4 \end{pmatrix} \to \begin{pmatrix} 1 & -2 & 2 & 3 & -1 \\ -3 & 6 & -1 & 1 & -7 \\ 2 & -4 & 5 & 8 & -4 \end{pmatrix} \to$$
$$\begin{pmatrix} 1 & -2 & 2 & 3 & -1 \\ 0 & 0 & 5 & 10 & -10 \\ 0 & 0 & 1 & 2 & -2 \end{pmatrix} \to \begin{pmatrix} 1 & -2 & 2 & 3 & -1 \\ 0 & 0 & 1 & 2 & -2 \\ 0 & 0 & 5 & 10 & -10 \end{pmatrix} \to$$
$$\begin{pmatrix} 1 & -2 & 0 & -1 & 3 \\ 0 & 0 & 1 & 2 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We can now write the matrix equation $\mathrm{A}\mathrm{x} = 0$ for the null space using rref(A). Writing the variable associated with the pivot columns on the left-hand-side of the equations, we have from the first and second rows

$$x_1 = 2x_2 + x_4 - 3x_5,$$
$$x_3 = -2x_4 + 2x_5.$$

Eliminating x_1 and x_3, we now write the general solution for vectors in the null space as

$$\begin{pmatrix} 2x_2 + x_4 - 3x_5 \\ x_2 \\ -2x_4 + 2x_5 \\ x_4 \\ x_5 \end{pmatrix} = x_2 \begin{pmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 1 \\ 0 \\ -2 \\ 1 \\ 0 \end{pmatrix} + x_5 \begin{pmatrix} -3 \\ 0 \\ 2 \\ 0 \\ 1 \end{pmatrix},$$

where x_2, x_4, and x_5 are called free variables, and can take any values.

The vector multiplying the free variable x_2 has a one in the second row and all the other vectors have a zero in this row. Similarly, the vector multiplying x_4 has a one in the fourth row and all the other vectors have a zero in this row. And the vector multiplying x_5 has a one in the fifth row and all the other vectors have a zero

in this row. Therefore, these three vectors must be linearly independent and they form a basis for the null space. The basis is given by the set

$$\left\{ \begin{pmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ -2 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} -3 \\ 0 \\ 2 \\ 0 \\ 1 \end{pmatrix} \right\}.$$

The null space is seen to be a three-dimensional subspace of $\mathbb{R}^5$, and its dimension is equal to the number of free variables of rref(A). The number of free variables is, of course, equal to the number of columns minus the number of pivot columns.

3.5.2 Application of the null space

View *Application of the Null Space* on YouTube

An underdetermined system of linear equations $\mathrm{Ax} = \mathrm{b}$ with more unknowns than equations may not have a unique solution. If u is the general form of a vector in the null space of A, and v is any vector that satisfies $\mathrm{Av} = \mathrm{b}$, then $x = \mathrm{u} + \mathrm{v}$ satisfies $\mathrm{Ax} = \mathrm{A}(\mathrm{u} + \mathrm{v}) = \mathrm{Au} + \mathrm{Av} = 0 + b = b$. The general solution of $\mathrm{Ax} = \mathrm{b}$ can therefore be written as the sum of a general vector in $\mathrm{Null}(\mathrm{A})$ and a particular vector that satisfies the underdetermined system.

As an example, suppose we want to find the general solution to the linear system of two equations and three unknowns given by

$$\begin{aligned} 2x_1 + 2x_2 + x_3 &= 0, \\ 2x_1 - 2x_2 - x_3 &= 1, \end{aligned}$$

which in matrix form is given by

$$\begin{pmatrix} 2 & 2 & 1 \\ 2 & -2 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

We first bring the augmented matrix to reduced row echelon form:

$$\begin{pmatrix} 2 & 2 & 1 & 0 \\ 2 & -2 & -1 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 & 1/4 \\ 0 & 1 & 1/2 & -1/4 \end{pmatrix}.$$

The null space is determined from $x_1 = 0$ and $x_2 = -x_3/2$, and taking $x_3 = 2$, we can write

$$\mathrm{Null}(\mathrm{A}) = \mathrm{span} \left\{ \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix} \right\}.$$

A particular solution for the inhomogeneous system is found by solving $x_1 = 1/4$ and $x_2 + x_3/2 = -1/4$. Here, we simply take the free variable x_3 to be zero, and we find $x_1 = 1/4$ and $x_2 = -1/4$. The general solution to the original underdetermined linear system is the sum of the null space and the particular solution and is given by

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = a \begin{pmatrix} 0 \\ -1 \\ 2 \end{pmatrix} + \frac{1}{4} \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}.$$

3.5.3 Column space

View *Column Space* on YouTube

The column space of a matrix is the vector space spanned by the columns of the matrix. When a matrix is multiplied by a column vector, the resulting vector is in the column space of the matrix, as can be seen from the example

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix} = x \begin{pmatrix} a \\ c \end{pmatrix} + y \begin{pmatrix} b \\ d \end{pmatrix},$$

where the right-hand side is seen to be a linear combination of the columns of A. In general, Ax is a linear combination of the columns of A, and the equation $Ax = 0$ expresses the linear dependence of the columns of A. If the columns of A are linearly independent, then the null space of A is the zero vector. If the columns of a square matrix A are linearly independent, then A is an invertible matrix.

Given an m-by-n matrix A, what is the dimension of the column space of A, and how do we find a basis? Note that since A has m rows, the column space of A is a subspace of $\mathbb{R}^m$.

Fortunately, a basis for the column space of A can be found from rref(A). Consider the example of §3.5.1, where

$$A = \begin{pmatrix} -3 & 6 & -1 & 1 & -7 \\ 1 & -2 & 2 & 3 & -1 \\ 2 & -4 & 5 & 8 & -4 \end{pmatrix},$$

and

$$\mathrm{rref}(A) = \begin{pmatrix} 1 & -2 & 0 & -1 & 3 \\ 0 & 0 & 1 & 2 & -2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The matrix equation $Ax = 0$ is equivalent to $\mathrm{rref}(A)x = 0$, and the latter equation can be expressed as

$$x_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} -2 \\ 0 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -1 \\ 2 \\ 0 \end{pmatrix} + x_5 \begin{pmatrix} 3 \\ -2 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Only the pivot columns of $\mathrm{rref}(A)$, here the first and third columns, are linearly independent. For example, the second column is -2 times the first column; and whatever linear dependence relations hold true for the columns of $\mathrm{rref}(A)$ also hold true for the original matrix A. (You can try and check this fact.) The dimension of the column space of A is therefore equal to the number of pivot columns of A, and here it is two. A basis for the column space is given by the first and third columns of A (not $\mathrm{rref}(A)$), and is

$$\left\{ \begin{pmatrix} -3 \\ 1 \\ 2 \end{pmatrix}, \quad \begin{pmatrix} -1 \\ 2 \\ 5 \end{pmatrix} \right\}.$$

Recall that the dimension of the null space is the number of non-pivot columns, so that the sum of the dimensions of the null space and the column space is equal to the total number of columns. A statement of this theorem is as follows. Let A be an m-by-n matrix. Then

$$\dim(\mathrm{Col}(A)) + \dim(\mathrm{Null}(A)) = n.$$

3.5.4 Row space, left null space and rank

View *Row Space, Left Null Space and Rank* on YouTube

In addition to the column space and the null space, a matrix A has two more vector spaces associated with it, namely the column space and null space of A^T, which are called the row space and the left null space of A.

If A is an m-by-n matrix, then the row space and the null space are subspaces of $\mathbb{R}^n$, and the column space and the left null space are subspaces of $\mathbb{R}^m$.

The null space consists of all vectors x such that $Ax = 0$, that is, the null space is the set of all vectors that are orthogonal to the row space of A. We say that these two vector spaces are orthogonal.

A basis for the row space of a matrix can be found from computing $\text{rref}(A)$, and is found to be rows of $\text{rref}(A)$ (written as column vectors) with pivot columns. The dimension of the row space of A is therefore equal to the number of pivot columns, while the dimension of the null space of A is equal to the number of nonpivot columns. The union of these two subspaces make up the vector space of all n-by-one matrices and we say that these subspaces are *orthogonal complements* of each other.

Furthermore, the dimension of the column space of A is also equal to the number of pivot columns, so that the dimensions of the column space and the row space of a matrix are equal. We have

$$\dim(\text{Col}(A)) = \dim(\text{Row}(A)).$$

We call this dimension the rank of the matrix A. This is an amazing result since the column space and row space are subspaces of two different vector spaces. In general, we must have $\text{rank}(A) \leq \min(m, n)$. When the equality holds, we say that the matrix is of full rank. And when A is a square matrix and of full rank, then the dimension of the null space is zero and A is invertible.

We summarize our results in the table below. The null space of A^T is also called the left null space of A and the column space of A^T is also called the row space of A. The null space of A and the row space of A are orthogonal complements as is the left null space of A and the column space of A. The dimension of the column space of A is equal to the dimension of the row space of A and this dimension is called the rank of A.

Table 3.1: The four fundamental subspaces of an m-by-n matrix

vector space	subspace of	dimension
$\text{Null}(A)$	$\mathbb{R}^n$	$n -$ # of pivot columns
$\text{Col}(A)$	$\mathbb{R}^m$	# of pivot columns
$\text{Null}(A^T)$	$\mathbb{R}^m$	$m -$ # of pivot columns
$\text{Col}(A^T)$	$\mathbb{R}^n$	# of pivot columns

3.6 Gram-Schmidt process

View *Gram-Schmidt Process* on YouTube
View *Gram-Schmidt Process Example* on YouTube

Given any basis for a vector space, we can use an algorithm called the Gram-Schmidt process to construct an orthonormal basis for that space. Let the vectors $v_1, v_2, \ldots, v_n$ be a basis for some n-dimensional vector space. We will assume here that these vectors are column matrices, but this process also applies more generally.

We will construct an orthogonal basis $u_1, u_2, \ldots, u_n$, and then normalize each vector to obtain an orthonormal basis. First, define $u_1 = v_1$. To find the next orthogonal basis vector, define

$$u_2 = v_2 - \frac{(u_1^T v_2)u_1}{u_1^T u_1}.$$

Observe that u_2 is equal to v_2 minus the component of v_2 that is parallel to u_1. By multiplying both sides of this equation with u_1^T, it is easy to see that $u_1^T u_2 = 0$ so that these two vectors are orthogonal.

The next orthogonal vector in the new basis can be found from

$$u_3 = v_3 - \frac{(u_1^T v_3)u_1}{u_1^T u_1} - \frac{(u_2^T v_3)u_2}{u_2^T u_2}.$$

Here, u_3 is equal to v_3 minus the components of v_3 that are parallel to u_1 and u_2. We can continue in this fashion to construct n orthogonal basis vectors. These vectors can then be normalized via

$$\widehat{u}_1 = \frac{u_1}{(u_1^T u_1)^{1/2}}, \quad \text{etc.}$$

Since u_k is a linear combination of $v_1, v_2, \ldots, v_k$, the vector subspace spanned by the first k basis vectors of the original vector space is the same as the subspace spanned by the first k orthonormal vectors generated through the Gram-Schmidt process. We can write this result as

$$\text{span}\{u_1, u_2, \ldots, u_k\} = \text{span}\{v_1, v_2, \ldots, v_k\}.$$

To give an example of the Gram-Schmidt process, consider a subspace of $\mathbb{R}^4$ with the following basis:

$$W = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \right\} = \{v_1, v_2, v_3\}.$$

We use the Gram-Schmidt process to construct an orthonormal basis for this subspace. Let $u_1 = v_1$. Then u_2 is found from

$$\begin{aligned} u_2 &= v_2 - \frac{(u_1^T v_2)u_1}{u_1^T u_1} \\ &= \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \frac{3}{4}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{4}\begin{pmatrix} -3 \\ 1 \\ 1 \\ 1 \end{pmatrix}. \end{aligned}$$

Finally, we compute u_3:

$$u_3 = v_3 - \frac{(u_1^T v_3)u_1}{u_1^T u_1} - \frac{(u_2^T v_3)u_2}{u_2^T u_2}$$
$$= \begin{pmatrix}0\\0\\1\\1\end{pmatrix} - \frac{1}{2}\begin{pmatrix}1\\1\\1\\1\end{pmatrix} - \frac{1}{6}\begin{pmatrix}-3\\1\\1\\1\end{pmatrix} = \frac{1}{3}\begin{pmatrix}0\\-2\\1\\1\end{pmatrix}.$$

Normalizing the three vectors, we obtain the orthonormal basis

$$W' = \left\{\frac{1}{2}\begin{pmatrix}1\\1\\1\\1\end{pmatrix}, \quad \frac{1}{2\sqrt{3}}\begin{pmatrix}-3\\1\\1\\1\end{pmatrix}, \quad \frac{1}{\sqrt{6}}\begin{pmatrix}0\\-2\\1\\1\end{pmatrix}\right\}.$$

3.7 Orthogonal projections

View *Orthogonal Projections* on YouTube

Suppose that V is an n-dimensional vector space and W is a p-dimensional subspace of V. Let $\{s_1, s_2, \ldots, s_p\}$ be an orthonormal basis for W. Extending the basis for W, let $\{s_1, s_2, \ldots, s_p, t_1, t_2, \ldots, t_{n-p}\}$ be an orthonormal basis for V.

Any vector v in V can be written in terms of the basis for V as

$$v = a_1 s_1 + a_2 s_2 + \cdots + a_p s_p + b_1 t_1 + b_2 t_2 + b_{n-p} t_{n-p}.$$

The orthogonal projection of v onto W is then defined as

$$v_{\text{proj}_W} = a_1 s_1 + a_2 s_2 + \cdots + a_p s_p,$$

that is, the part of v that lies in W.

If you only know the vector v and the orthonormal basis for W, then the orthogonal projection of v onto W can be computed from

$$v_{\text{proj}_W} = (v^T s_1)s_1 + (v^T s_2)s_2 + \cdots + (v^T s_p)s_p,$$

that is, $a_1 = v^T s_1$, $a_2 = v^T s_2$, etc.

We can prove that the vector v_{proj_W} is the vector in W that is closest to v. Let w be any vector in W different than v_{proj_W}, and expand w in terms of the basis vectors for W:

$$w = c_1 s_1 + c_2 s_2 + \cdots + c_p s_p.$$

The distance between v and w is given by the norm $||v - w||$, and we have

$$||v - w||^2 = (a_1 - c_1)^2 + (a_2 - c_2)^2 + \cdots + (a_p - c_p)^2 + b_1^2 + b_2^2 + \cdots + b_{n-p}^2$$
$$\geq b_1^2 + b_2^2 + \cdots + b_{n-p}^2 = ||v - v_{\text{proj}_W}||^2,$$

or $||v - v_{\text{proj}_W}|| \leq ||v - w||$, a result that will be used later in the problem of least squares.

3.8 QR factorization

The Gram-Schmidt process naturally leads to a matrix factorization. Let A be an m-by-n matrix with n linearly-independent columns given by $\{x_1, x_2, \ldots, x_n\}$. Following the Gram-Schmidt process, it is always possible to construct an orthornormal basis for the column space of A, denoted by $\{q_1, q_2, \ldots, q_n\}$. An important feature of this orthonormal basis is that the first k basis vectors from the orthonormal set span the same vector subspace as the first k columns of the matrix A. For some coefficients r_{ij}, we can therefore write

$$\begin{aligned}
x_1 &= r_{11}q_1, \\
x_2 &= r_{12}q_1 + r_{22}q_2, \\
x_3 &= r_{13}q_1 + r_{23}q_2 + r_{33}q_3, \\
&\vdots \qquad\qquad \vdots \\
x_n &= r_{1n}q_1 + r_{2n}q_2 + \cdots + r_{nn}q_n;
\end{aligned}$$

and these equations can be written in matrix form as

$$\begin{pmatrix} | & | & | & & | \\ x_1 & x_2 & x_3 & \cdots & x_n \\ | & | & | & & | \end{pmatrix} = \begin{pmatrix} | & | & | & & | \\ q_1 & q_2 & q_3 & \cdots & q_n \\ | & | & | & & | \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{pmatrix}.$$

This form represents the matrix factorization called the QR factorization, and is usually written as

$$A = QR,$$

where Q is an orthogonal matrix and R is an upper triangular matrix. The diagonal elements of R can also be made non-negative by suitably adjusting the signs of the orthonormal basis vectors.

As a concrete example, we will find the the QR factorization of the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} | & | \\ a_1 & a_2 \\ | & | \end{pmatrix}.$$

Applying the Gram-Schmidt process to the column vectors of A, we have for the unnormalized orthogonal vectors

$$\begin{aligned}
q_1 &= a_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \\
q_2 &= a_2 - \frac{(q_1^T a_2)q_1}{q_1^T q_1} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} - \frac{4}{5}\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 6/5 \\ -3/5 \end{pmatrix} = \frac{3}{5}\begin{pmatrix} 2 \\ -1 \end{pmatrix},
\end{aligned}$$

and normalizing, we obtain

$$q_1 = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad q_2 = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ -1 \end{pmatrix}.$$

The projection of the columns of A onto the set of orthonormal vectors is given by

$$a_1 = (a_1^T q_1)q_1, \quad a_2 = (a_2^T q_1)q_1 + (a_2^T q_2)q_2,$$

and with $r_{ij} = a_j^T q_i$, we compute

$$r_{11} = a_1^T q_1 = \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \frac{1}{\sqrt{5}} = \sqrt{5},$$

$$r_{12} = a_2^T q_1 = \begin{pmatrix} 2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \frac{1}{\sqrt{5}} = \frac{4\sqrt{5}}{5},$$

$$r_{22} = a_2^T q_2 = \begin{pmatrix} 2 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ -1 \end{pmatrix} \frac{1}{\sqrt{5}} = \frac{3\sqrt{5}}{5}.$$

The QR factorization of A is therefore given by

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ 2/\sqrt{5} & -1/\sqrt{5} \end{pmatrix} \begin{pmatrix} \sqrt{5} & 4\sqrt{5}/5 \\ 0 & 3\sqrt{5}/5 \end{pmatrix}.$$

3.9 The least-squares problem

View *The Least-Squares Problem Using Matrices* on YouTube

Suppose there is some experimental data that is suspected to satisfy a functional

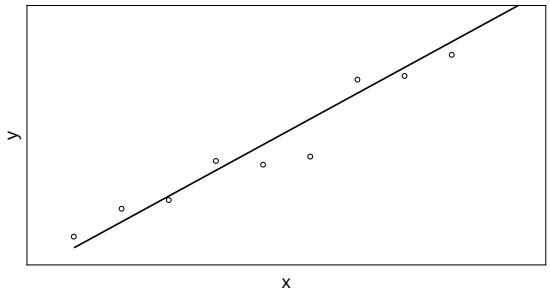


Figure 3.1: *Linear regression.*

relationship. The simplest such relationship is linear, and suppose one wants to fit a straight line to the data. An example of such a linear regression problem is shown in Fig. 3.1.

In general, let the data consist of a set of n points given by (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Here, the x values are exact, and the y values are noisy. We assume that a line of the form

$$y = \beta_0 + \beta_1 x$$

is the best fit to the data. Although we know that the line will not go through all of the data points, we can still write down the resulting equations. We have

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1, \\ y_2 &= \beta_0 + \beta_1 x_2, \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_n. \end{aligned}$$

These equations are a system of n equations in the two unknowns β_0 and β_1. The corresponding matrix equation is given by

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

This is an overdetermined system of equations that obviously has no solution. The problem of least-squares is to find the best solution of these equations for β_0 and β_1.

We can generalize this problem as follows. Suppose we are given the matrix equation

$$\mathrm{Ax} = \mathrm{b}$$

that has no solution because b is not in the column space of A. Instead of exactly solving this matrix equation, we want to solve another approximate equation that minimizes the error between Ax and b. The error can be defined as the norm of $\mathrm{Ax} - \mathrm{b}$, and the square of the error is just the sum of the squares of the components. Our search is for the least-squares solution.

3.10 Solution of the least-squares problem

View *Solution of the Least-Squares Problem by the Normal Equations* on YouTube

The problem of least-squares can be cast as the problem of solving an overdetermined matrix equation $\mathrm{Ax} = \mathrm{b}$ when b is not in the column space of A. By replacing b by its orthogonal projection onto the column space of A, the solution minimizes the norm $||\mathrm{Ax} - \mathrm{b}||$.

Now $\mathrm{b} = \mathrm{b}_{\mathrm{proj}_{\mathrm{Col(A)}}} + (\mathrm{b} - \mathrm{b}_{\mathrm{proj}_{\mathrm{Col(A)}}})$, where $\mathrm{b}_{\mathrm{proj}_{\mathrm{Col(A)}}}$ is the projection of b onto the column space of A. Since $(\mathrm{b} - \mathrm{b}_{\mathrm{proj}_{\mathrm{Col(A)}}})$ is orthogonal to the column space of A, it is in the nullspace of A^{T}. Therefore, $\mathrm{A}^{\mathrm{T}}(\mathrm{b} - \mathrm{b}_{\mathrm{proj}_{\mathrm{Col(A)}}}) = 0$, and it pays to multiply $\mathrm{Ax} = \mathrm{b}$ by A^{T} to obtain

$$\mathrm{A}^{\mathrm{T}}\mathrm{Ax} = \mathrm{A}^{\mathrm{T}}\mathrm{b}.$$

These equations, called the normal equations for $\mathrm{Ax} = \mathrm{b}$, determine the least-squares solution for x, which can be found by Gaussian elimination. When A is an m-by-n matrix, then $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ is an n-by-n matrix, and it can be shown that $\mathrm{A}^{\mathrm{T}}\mathrm{A}$ is invertible when the columns of A are linearly independent. When this is the case, one can rewrite the normal equations by multiplying both sides by $\mathrm{A}(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}$ to obtain

$$\mathrm{Ax} = \mathrm{A}(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}\mathrm{A}^{\mathrm{T}}\mathrm{b},$$

where the matrix

$$\mathrm{P} = \mathrm{A}(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}\mathrm{A}^{\mathrm{T}}$$

projects a vector onto the column space of A. It is easy to prove that $\mathrm{P}^2 = \mathrm{P}$, which

states that two projections is the same as one. We have

$$\begin{aligned} \mathrm{P}^2 &= (\mathrm{A}(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}\mathrm{A}^{\mathrm{T}})(\mathrm{A}(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}\mathrm{A}^{\mathrm{T}}) \\ &= \mathrm{A}\left[(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}(\mathrm{A}^{\mathrm{T}}\mathrm{A})\right](\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}\mathrm{A}^{\mathrm{T}} \\ &= \mathrm{A}(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}\mathrm{A}^{\mathrm{T}} = P. \end{aligned}$$

As an example, consider the toy least-squares problem of fitting a line through the three data points $(1,1)$, $(2,3)$ and $(3,2)$. With the line given by $y = \beta_0 + \beta_1 x$, the overdetermined system of equations is given by

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}.$$

The least-squares solution is determined by solving

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix},$$

or

$$\begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 6 \\ 13 \end{pmatrix}.$$

We can solve either by directly inverting the two-by-two matrix or by using Gaussian elimination. Inverting the two-by-two matrix, we have

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix} \begin{pmatrix} 6 \\ 13 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \end{pmatrix},$$

so that the least-squares line is given by

$$y = 1 + \frac{1}{2}x.$$

The graph of the data and the line is shown below.

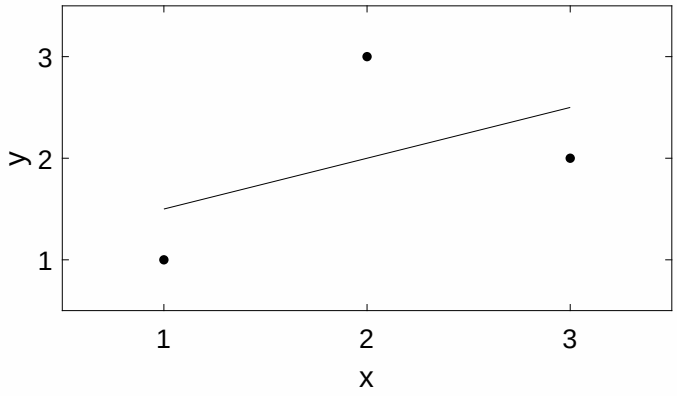


Figure 3.2: *Solution of the toy least-squares problem.*

Chapter 4

Determinants

4.1 Two-by-two and three-by-three determinants

View *Two-by-Two and Three-by-Three Determinants* on YouTube

Our first introduction to determinants was the definition for the general two-by-two matrix

$$\mathrm{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \quad \det \mathrm{A} = ad - bc.$$

Other widely used notations for the determinant include

$$\det \mathrm{A} = \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = |\mathrm{A}| = \begin{vmatrix} a & b \\ c & d \end{vmatrix}.$$

By explicit construction, we have seen that a two-by-two matrix A is invertible if and only if $\det \mathrm{A} \neq 0$. If a square matrix A is invertible, then the equation $\mathrm{Ax} = \mathrm{b}$ has the unique solution $\mathrm{x} = \mathrm{A}^{-1}\mathrm{b}$. But if A is not invertible, then $\mathrm{Ax} = \mathrm{b}$ may have no solution or an infinite number of solutions. When $\det \mathrm{A} = 0$, we say that A is a singular matrix.

Here, we would like to extend the definition of the determinant to an n-by-n matrix. Before we do so, let us display the determinant for a three-by-three matrix. We consider the system of equations $\mathrm{Ax} = 0$ and find the condition for which $\mathrm{x} = 0$ is the only solution. This condition must be equivalent to $\det \mathrm{A} \neq 0$. With

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = 0,$$

one can do the messy algebra of elimination to solve for x_1, x_2, and x_3. One finds that $x_1 = x_2 = x_3 = 0$ is the only solution when $\det \mathrm{A} \neq 0$, where the definition is given by

$$\det \mathrm{A} = aei + bfg + cdh - ceg - bdi - afh. \tag{4.1}$$

A way to remember this result for the three-by-three matrix is by the following picture:

$$\begin{pmatrix} a & b & c & a & b \\ & \searrow & \searrow & \searrow & \\ d & e & f & d & e \\ & \searrow & \searrow & \searrow & \\ g & h & i & g & h \end{pmatrix} - \begin{pmatrix} a & b & c & a & b \\ & \swarrow & \swarrow & \swarrow & \\ d & e & f & d & e \\ & \swarrow & \swarrow & \swarrow & \\ g & h & i & g & h \end{pmatrix}.$$

The matrix A is periodically extended two columns to the right, drawn explicitly here but usually only imagined. Then the six terms comprising the determinant are made evident, with the lines slanting down towards the right getting the plus signs and the lines slanting down towards the left getting the minus signs. Unfortunately, this mnemonic is only valid for three-by-three matrices.

4.2 Laplace expansion and Leibniz formula

View *Laplace Expansion for Computing Determinants* on YouTube
View *Leibniz Formula for Computing Determinants* on YouTube

There are two ways to view the three-by-three determinant that do in fact generalize to n-by-n matrices. The first way writes

$$\begin{aligned}\det \mathrm{A} &= aei + bfg + cdh - ceg - bdi - afh \\ &= a(ei - fh) - b(di - fg) + c(dh - eg) \\ &= a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}.\end{aligned}$$

The three-by-three determinant is found from lower-order two-by-two determinants, and a recursive definition of the determinant is possible. This method of computing a determinant is called a Laplace expansion, or cofactor expansion, or expansion by minors. The minors refer to the lower-order determinants, and the cofactor refers to the combination of the minor with the appropriate plus or minus sign. The rule here is that one goes across the first row of the matrix, multiplying each element in the first row by the determinant of the matrix obtained by crossing out the element's row and column. The sign of the terms alternate as we go across the row.

Instead of going across the first row, we could have gone down the first column using the same method to obtain

$$\det \mathrm{A} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - d\begin{vmatrix} b & c \\ h & i \end{vmatrix} + g\begin{vmatrix} b & c \\ e & f \end{vmatrix},$$

also equivalent to (4.1). In fact, this expansion by minors can be done across any row or down any column. The sign of each term in the expansion is given by $(-1)^{i+j}$ when the number multiplying each minor is drawn from the ith-row and j-th column. An easy way to remember the signs is to form a checkerboard pattern, exhibited here for the three-by-three matrix:

$$\begin{pmatrix} + & - & + \\ - & + & - \\ + & - & + \end{pmatrix}.$$

The second way to generalize the determinant is called the Leibniz formula, or more descriptively, the big formula. One notices that each term in (4.1) has only a single element from each row and from each column. As we can choose one of three elements from the first row, then one of two elements from the second row, and only one element from the third row, there are $3! = 6$ terms in the expansion. For a general n-by-n matrix there are $n!$ terms.

The sign of each term depends on whether it derives from an even or odd permutation of the columns numbered $\{1, 2, 3, \ldots, n\}$, with even permutations getting a plus sign and odd permutations getting a minus sign. An even permutation is one that can be obtained by switching pairs of numbers in the sequence $\{1, 2, 3, \ldots, n\}$ an even number of times, and an odd permutation corresponds to an odd number of times. As examples from the three-by-three case, the terms aei, bfg, and cdh correspond to the column numberings $\{1, 2, 3\}$, $\{2, 3, 1\}$, and $\{3, 1, 2\}$, which can

be seen to be even permutations of $\{1,2,3\}$, and the terms ceg, bdi, and afh correspond to the column numberings $\{3,2,1\}$, $\{2,1,3\}$, and $\{1,3,2\}$, which are odd permutations.

Either the Laplace expansion or the Leibniz formula can be used to define the determinant of an n-by-n matrix. It will, however, be more illuminating to define the determinant from three of its fundamental properties. These properties will lead us to the determinant's most important practical application: $\det \mathrm{A} \neq 0$ for an invertible matrix. But we will also elucidate many other useful properties.

4.3 Properties of the determinant

View *Properties of the Determinant* on YouTube

The determinant, as we know, is a function that maps an n-by-n matrix to a scalar. We now define this determinant function by the following three properties.

Property 1: The determinant of the identity matrix is one, i.e.,

$$\det \mathrm{I} = 1.$$

This property essentially normalizes the determinant. The two-by-two illustration is

$$\begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1 \times 1 - 0 \times 0 = 1.$$

Property 2: The determinant changes sign under row exchange. The two-by-two illustration is

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc = -(cb - da) = -\begin{vmatrix} c & d \\ a & b \end{vmatrix}.$$

Property 3: The determinant is a linear function of the first row, holding all other rows fixed. The two-by-two illustration is

$$\begin{vmatrix} ka & kb \\ c & d \end{vmatrix} = kad - kbc = k(ad - bc) = k\begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

and

$$\begin{vmatrix} a + a' & b + b' \\ c & d \end{vmatrix} = (a + a')d - (b + b')c = (ad - bc) + (a'd - b'c) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} + \begin{vmatrix} a' & b' \\ c & d \end{vmatrix}.$$

Remarkably, Properties 1-3 are all we need to uniquely define the determinant function. It is easy to show explicitly that these three properties hold for the determinants of two-by-two and three-by-three matrices. And not too hard to show that they hold for our definitions of the Laplace expansion and the Leibniz formula for the determinant of an n-by-n matrix.

We now discuss further properties that follow from Properties 1-3. We will continue to illustrate these properties using a two-by-two matrix.

Property 4: The determinant is a linear function of all the rows, e.g.,

$$\begin{aligned}\begin{vmatrix} a & b \\ kc & kd \end{vmatrix} &= -\begin{vmatrix} kc & kd \\ a & b \end{vmatrix} && \text{(Property 2)}\\ &= -k\begin{vmatrix} c & d \\ a & b \end{vmatrix} && \text{(Property 3)}\\ &= \quad k\begin{vmatrix} a & b \\ c & d \end{vmatrix}, && \text{(Property 2)}\end{aligned}$$

and similarly for the second linearity condition.

Property 5: If a matrix has two equal rows, then the determinant is zero, e.g.,

$$\begin{aligned}\begin{vmatrix} a & b \\ a & b \end{vmatrix} &= -\begin{vmatrix} a & b \\ a & b \end{vmatrix} && \text{(Property 2)}\\ &= 0,\end{aligned}$$

since zero is the only number equal to its negative.

Property 6: If we add k times row-i to row-j the determinant doesn't change, e.g.,

$$\begin{aligned}\begin{vmatrix} a & b \\ c+ka & d+kb \end{vmatrix} &= \begin{vmatrix} a & b \\ c & d \end{vmatrix} + k\begin{vmatrix} a & b \\ a & b \end{vmatrix} && \text{(Property 4)}\\ &= \begin{vmatrix} a & b \\ c & d \end{vmatrix}. && \text{(Property 5)}\end{aligned}$$

This property together with Property 2 and 3 allows us to perform row reduction on a matrix to simplify the calculation of a determinant.

Property 7: The determinant of a matrix with a row of zeros is zero, e.g.,

$$\begin{aligned}\begin{vmatrix} a & b \\ 0 & 0 \end{vmatrix} &= 0\begin{vmatrix} a & b \\ 0 & 0 \end{vmatrix} && \text{(Property 4)}\\ &= 0.\end{aligned}$$

Property 8: The determinant of a diagonal matrix is just the product of the diagonal elements, e.g.,

$$\begin{aligned}\begin{vmatrix} a & 0 \\ 0 & d \end{vmatrix} &= ad\begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} && \text{(Property 4)}\\ &= ad. && \text{(Property 1)}\end{aligned}$$

Property 9: The determinant of an upper or lower triangular matrix is just the product of the diagonal elements, e.g.,

$$\begin{aligned}\begin{vmatrix} a & b \\ 0 & d \end{vmatrix} &= \begin{vmatrix} a & 0 \\ 0 & d \end{vmatrix} && \text{(Property 6)}\\ &= ad. && \text{(Property 8)}\end{aligned}$$

In the above calculation, Property 6 is applied by multiplying the second row by $-b/d$ and adding it to the first row.

Property 10: A matrix with a nonzero determinant is invertible. A matrix with a zero determinant is not invertible. Row reduction (Property 6), row exchange (Property 2), and multiplication of a row by a nonzero scalar (Property 4) can bring a square matrix to its reduced row echelon form. If $\operatorname{rref}(\mathrm{A}) = \mathrm{I}$, then the determinant is nonzero and the matrix is invertible. If $\operatorname{rref}(\mathrm{A}) \neq \mathrm{I}$, then the last row is all zeros, the determinant is zero, and the matrix is not invertible.

Property 11: The determinant of the product is equal to the product of the determinants, i.e.,

$$\det \mathrm{AB} = \det \mathrm{A} \det \mathrm{B}.$$

This identity turns out to be very useful, but its proof for a general n-by-n matrix is difficult. The proof for a two-by-two matrix can be done directly. Let

$$\mathrm{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \mathrm{B} = \begin{pmatrix} e & f \\ g & h \end{pmatrix}.$$

Then

$$\mathrm{AB} = \begin{pmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{pmatrix},$$

and

$$\begin{aligned}
\det \mathrm{AB} &= \begin{vmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{vmatrix} \\
&= \begin{vmatrix} ae & af \\ ce + dg & cf + dh \end{vmatrix} + \begin{vmatrix} bg & bh \\ ce + dg & cf + dh \end{vmatrix} \\
&= \begin{vmatrix} ae & af \\ ce & cf \end{vmatrix} + \begin{vmatrix} ae & af \\ dg & dh \end{vmatrix} + \begin{vmatrix} bg & bh \\ ce & cf \end{vmatrix} + \begin{vmatrix} bg & bh \\ dg & dh \end{vmatrix} \\
&= ac \begin{vmatrix} e & f \\ e & f \end{vmatrix} + ad \begin{vmatrix} e & f \\ g & h \end{vmatrix} + bc \begin{vmatrix} g & h \\ e & f \end{vmatrix} + bd \begin{vmatrix} g & h \\ g & h \end{vmatrix} \\
&= (ad - bc) \begin{vmatrix} e & f \\ g & h \end{vmatrix} \\
&= \det \mathrm{A} \det \mathrm{B}.
\end{aligned}$$

Property 12: Commuting two matrices doesn't change the value of the determinant, i.e., $\det \mathrm{AB} = \det \mathrm{BA}$. The proof is simply

$$\begin{aligned}
\det \mathrm{AB} &= \det \mathrm{A} \det \mathrm{B} && \text{(Property 11)} \\
&= \det \mathrm{B} \det \mathrm{A} \\
&= \det \mathrm{BA}. && \text{(Property 11)}
\end{aligned}$$

Property 13: The determinant of the inverse is the inverse of the determinant, i.e., if A is invertible, then $\det(\mathrm{A}^{-1}) = 1/\det \mathrm{A}$. The proof is

$$\begin{aligned}
1 &= \det \mathrm{I} && \text{(Property 1)} \\
&= \det(\mathrm{AA}^{-1}) \\
&= \det \mathrm{A} \det \mathrm{A}^{-1}. && \text{(Property 11)}
\end{aligned}$$

Therefore,

$$\det \mathrm{A}^{-1} = \frac{1}{\det \mathrm{A}}.$$

Property 14: The determinant of a matrix raised to an integer power is equal to the determinant of that matrix, raised to the integer power. Note that $\mathrm{A}^2 = \mathrm{AA}$, $\mathrm{A}^3 = \mathrm{AAA}$, etc. This property in equation form is given by

$$\det(\mathrm{A}^p) = (\det \mathrm{A})^p,$$

where p is an integer. This result follows from the successive application of Property 11.

Property 15: If A is an n-by-n matrix, then

$$\det k\mathrm{A} = k^n \det \mathrm{A}.$$

Note that kA multiplies every element of A by the scalar k. This property follows simply from Property 4 applied n times.

Property 16: The determinant of the transposed matrix is equal to the determinant of the matrix, i.e.

$$\det \mathrm{A}^{\mathrm{T}} = \det \mathrm{A}.$$

When $\mathrm{A} = \mathrm{LU}$ without any row exchanges, we have $\mathrm{A}^{\mathrm{T}} = \mathrm{U}^{\mathrm{T}}\mathrm{L}^{\mathrm{T}}$ and

$$\begin{aligned}
\det \mathrm{A}^{\mathrm{T}} &= \det \mathrm{U}^{\mathrm{T}}\mathrm{L}^{\mathrm{T}} \\
&= \det \mathrm{U}^{\mathrm{T}} \det \mathrm{L}^{\mathrm{T}} && \text{(Property 11)} \\
&= \det \mathrm{U} \det \mathrm{L} && \text{(Property 9)} \\
&= \det \mathrm{LU} && \text{(Property 11)} \\
&= \det \mathrm{A}.
\end{aligned}$$

The same result can be shown to hold even if row interchanges are needed. The implication of Property 16 is that any statement about the determinant and the rows of A also apply to the columns of A. To compute the determinant, one can do either row reduction or column reduction!

It is time for some examples. We start with a simple three-by-three matrix and illustrate some approaches to a hand calculation of the determinant.

Example: Compute the determinant of

$$\mathrm{A} = \begin{pmatrix} 1 & 5 & 0 \\ 2 & 4 & -1 \\ 0 & -2 & 0 \end{pmatrix}.$$

We show computations using the Leibniz formula and the Laplace expansion.

Method 1 (Leibniz formula): We compute the six terms directly by periodically extending the matrix and remembering that diagonals slanting down towards the right get plus signs and diagonals slanting down towards the left get minus signs. We have

$$\det \mathrm{A} = 1 \cdot 4 \cdot 0 + 5 \cdot (-1) \cdot 0 + 0 \cdot 2 \cdot (-2) - 0 \cdot 4 \cdot 0 - 5 \cdot 2 \cdot 0 - 1 \cdot (-1) \cdot (-2) = -2.$$

Method 2 (Laplace expansion): We expand using minors. We should choose an expansion across the row or down the column that has the most zeros. Here, the obvious

choices are either the third row or the third column, and we can show both. Across the third row, we have

$$\det \mathrm{A} = -(-2) \cdot \begin{vmatrix} 1 & 0 \\ 2 & -1 \end{vmatrix} = -2,$$

and down the third column, we have

$$\det \mathrm{A} = -(-1) \cdot \begin{vmatrix} 1 & 5 \\ 0 & -2 \end{vmatrix} = -2.$$

Example: Compute the determinant of

$$\mathrm{A} = \begin{pmatrix} 6 & 3 & 2 & 4 & 0 \\ 9 & 0 & -4 & 1 & 0 \\ 8 & -5 & 6 & 7 & 1 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 2 & 3 & 2 & 0 \end{pmatrix}.$$

We first expand in minors across the fourth row:

$$\begin{vmatrix} 6 & 3 & 2 & 4 & 0 \\ 9 & 0 & -4 & 1 & 0 \\ 8 & -5 & 6 & 7 & 1 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 2 & 3 & 2 & 0 \end{vmatrix} = -3 \begin{vmatrix} 3 & 2 & 4 & 0 \\ 0 & -4 & 1 & 0 \\ -5 & 6 & 7 & 1 \\ 2 & 3 & 2 & 0 \end{vmatrix}.$$

We then expand in minors down the fourth column:

$$-3 \begin{vmatrix} 3 & 2 & 4 & 0 \\ 0 & -4 & 1 & 0 \\ -5 & 6 & 7 & 1 \\ 2 & 3 & 2 & 0 \end{vmatrix} = 3 \begin{vmatrix} 3 & 2 & 4 \\ 0 & -4 & 1 \\ 2 & 3 & 2 \end{vmatrix}.$$

We can then multiply the third column by 4 and add it to the second column:

$$= 3 \begin{vmatrix} 3 & 2 & 4 \\ 0 & -4 & 1 \\ 2 & 3 & 2 \end{vmatrix} = 3 \begin{vmatrix} 3 & 18 & 4 \\ 0 & 0 & 1 \\ 2 & 11 & 2 \end{vmatrix},$$

and finally expand in minors across the second row:

$$3 \begin{vmatrix} 3 & 18 & 4 \\ 0 & 0 & 1 \\ 2 & 11 & 2 \end{vmatrix} = -3 \begin{vmatrix} 3 & 18 \\ 2 & 11 \end{vmatrix} = -3(33 - 36) = 9.$$

The technique here is to try and zero out all the elements in a row or a column except one before proceeding to expand by minors across that row or column.

Example: Recall the Fibonacci Q-matrix, which satisfies

$$\mathrm{Q} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathrm{Q}^n = \begin{pmatrix} F_{n+1} & F_n \\ F_n & F_{n-1} \end{pmatrix},$$

where F_n is the nth Fibonacci number. Prove Cassini's identity

$$F_{n+1}F_{n-1} - F_n^2 = (-1)^n.$$

Repeated us of Property 10 yields $\det(\mathrm{Q}^n) = (\det \mathrm{Q})^n$. Applying this identity to the Fibonacci Q-matrix results in Cassini's identity. For example, with $F_5 = 5$, $F_6 = 8$, $F_7 = 13$, we have $13 \cdot 5 - 8^2 = 1$.

Example: Consider the tridiagonal matrix with ones on the main diagonal, ones on the first diagonal below the main, and negative ones on the first diagonal above the main. The matrix denoted by T_n is the n-by-n version of this matrix. For example, the first four matrices are given by

$$T_1 = (1)\,, \quad T_2 = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad T_3 = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix} \quad T_4 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & -1 & 0 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Show that $|T_n| = F_{n+1}$.

Let's compute the first three determinants. We have $|T_1| = 1 = F_2$ and $|T_2| = 2 = F_3$. We compute $|T_3|$ going across the first row using minors:

$$|T_3| = 1 \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} + 1 \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 2 + 1 = 3 = F_4.$$

To prove that $|T_n| = F_{n+1}$, we need only prove that $|T_{n+1}| = |T_n| + |T_{n-1}|$. We expand $|T_{n+1}|$ in minors across the first row. Using $|T_4|$ as an example, it is easy to see that

$$|T_{n+1}| = |T_n| + \begin{vmatrix} 1 & -1 & 0 & 0 & 0 & \dots \\ 0 & 1 & -1 & 0 & 0 & \dots \\ 0 & 1 & 1 & -1 & 0 & \dots \\ 0 & 0 & 1 & 1 & -1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots \end{vmatrix}.$$

The remaining determinant can be expanded down the first column to obtain $|T_{n-1}|$ so that $|T_{n+1}| = |T_n| + |T_{n-1}|$.This Fibonacci recursion relation together with $|T_1| = 1$ and $|T_2| = 2$ results in $|T_n| = F_{n+1}$.

4.4 Cramer's rule

Cramer's rule, first published in the year 1750, is a formula that uses determinants to find the solution of a linear system of equations. It is useful only for small systems. We will illustrate the derivation of Cramer's rule for three equation and three unknowns.

Consider the linear system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3. \end{aligned}$$

As usual, we write this as the matrix equation Ax = b, where x is the unknown column vector and b is the right-hand side. The coefficient matrix A is given by

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

Using the properties of a determinant (Properties 4, 6, and 16), we can write the following equalities:

$$\begin{aligned} x_1 \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= \begin{vmatrix} a_{11}x_1 & a_{12} & a_{13} \\ a_{21}x_1 & a_{22} & a_{23} \\ a_{31}x_1 & a_{32} & a_{33} \end{vmatrix} \\ &= \begin{vmatrix} a_{11}x_1 + a_{12}x_2 & a_{12} & a_{13} \\ a_{21}x_1 + a_{22}x_2 & a_{22} & a_{23} \\ a_{31}x_1 + a_{32}x_2 & a_{32} & a_{33} \end{vmatrix} \\ &= \begin{vmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 & a_{12} & a_{13} \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 & a_{22} & a_{23} \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 & a_{32} & a_{33} \end{vmatrix}. \end{aligned}$$

Then replacing the first column of the right-hand side determinant by the right-hand side of the system of equations, we get

$$x_1 \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}.$$

Solving for the first unknown x_1, we find

$$x_1 = \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix} \Bigg/ \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}.$$

Two similar calculations yield

$$x_2 = \begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix} \Bigg/ \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix},$$

$$x_3 = \begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix} \Bigg/ \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}.$$

Example: Using Cramer's rule, solve the system of linear equations given by

$$\begin{aligned} -3x_1 + 2x_2 - x_3 &= -1, \\ 6x_1 - 6x_2 + 7x_3 &= -7, \\ 3x_1 - 4x_2 + 4x_3 &= -6. \end{aligned}$$

We can find the required determinants using row reduction and the Laplace expansion. We have for the determinant of the A matrix,

$$\begin{vmatrix} -3 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -4 & 4 \end{vmatrix} = \begin{vmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 0 & -2 & 3 \end{vmatrix} = -3 \begin{vmatrix} -2 & 5 \\ -2 & 3 \end{vmatrix} = -3(-6+10) = -12.$$

Then replacing the first column of the A matrix by the right-hand side of the equation, we have

$$\begin{vmatrix} -1 & 2 & -1 \\ -7 & -6 & 7 \\ -6 & -4 & 4 \end{vmatrix} = \begin{vmatrix} -1 & 2 & -1 \\ 0 & -20 & 14 \\ 0 & -16 & 10 \end{vmatrix} = -4 \begin{vmatrix} -10 & 7 \\ -8 & 5 \end{vmatrix} = -4(-50+56) = -24,$$

and

$$x_1 = -24/(-12) = 2.$$

Replacing the second column, we have

$$\begin{vmatrix} -3 & -1 & -1 \\ 6 & -7 & 7 \\ 3 & -6 & 4 \end{vmatrix} = \begin{vmatrix} -3 & -1 & -1 \\ 0 & -9 & 5 \\ 0 & -7 & 3 \end{vmatrix} = -3 \begin{vmatrix} -9 & 5 \\ -7 & 3 \end{vmatrix} = -3(-27+35) = -24,$$

and

$$x_2 = -24/(-12) = 2.$$

Finally, replacing the third column, we have

$$\begin{vmatrix} -3 & 2 & -1 \\ 6 & -6 & -7 \\ 3 & -4 & -6 \end{vmatrix} = \begin{vmatrix} -3 & 2 & -1 \\ 0 & -2 & -9 \\ 0 & -2 & -7 \end{vmatrix} = -3 \begin{vmatrix} -2 & -9 \\ -2 & -7 \end{vmatrix} = -3(14-18) = 12,$$

and

$$x_3 = 12/(-12) = -1,$$

and we have solved our three equations for three unknowns.

4.5 Calculating the inverse matrix using determinants

If A is an invertible n-by-n matrix, then Cramer's rule can be applied to solve the equation

$$\mathrm{A}\mathrm{A}^{-1} = \mathrm{I}. \tag{4.2}$$

Each column of A^{-1} can be found using the corresponding column of I on the right-hand side. We will derive the method using a general three-by-three matrix, while introducing some commonly used terminology.

Let A be a three-by-three matrix given by

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

with $\det \mathrm{A} \neq 0$. The matrix equation we want to solve is

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{4.3}$$

where the unknown inverse matrix is given by

$$\mathrm{A}^{-1} = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix}.$$

The first column of the inverse matrix can be found by solving the system of equations given by

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Applying Cramer's rule and a Laplace expansion, we solve for x_1 times the determinant of A:

$$x_1 \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} 1 & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}.$$

Similarly, we solve for y_1 times the determinant of A:

$$y_1 \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{11} & 1 & a_{13} \\ a_{21} & 0 & a_{23} \\ a_{31} & 0 & a_{33} \end{vmatrix} = - \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}.$$

Note the minus sign in front of the two-by-two determinant because the one from the identity matrix is located in row one and column two of the three-by-three matrix. Finally, we solve for z_1 times the determinant of A:

$$z_1 \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} & 1 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & 0 \end{vmatrix} = \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}.$$

We can continue in this manner to determine the second column of the inverse matrix, and then finally the third column.

At this point, textbook writers usually introduce some additional terminology. The minor of the element a_{ij} of matrix A is defined to be the determinant of the submatrix formed by deleting the ith row and jth column of the matrix A. We denote the value of this determinant by M_{ij}. For example, the minor M_{23} of element a_{23} from a three-by-three matrix is computed by

$$M_{23} = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix} = a_{11}a_{32} - a_{12}a_{31}.$$

The cofactor of the element a_{ij} is the minor M_{ij} multiplied by $(-1)^{i+j}$, that is, the cofactor equals the minor if the row plus column number is even, and equals the negative of the minor if the row plus column number is odd. We denote the cofactor by C_{ij}, so that

$$C_{ij} = (-1)^{i+j} M_{ij}.$$

We now see that the first column of the inverse matrix can be written as

$$x_1 \det \mathrm{A} = C_{11}, \quad y_1 \det \mathrm{A} = C_{12}, \quad z_1 \det \mathrm{A} = C_{13}.$$

Continuing to compute the second and third columns of the inverse matrix, we obtain

$$\mathrm{A}^{-1} \det \mathrm{A} = \begin{pmatrix} C_{11} & C_{21} & C_{31} \\ C_{12} & C_{22} & C_{32} \\ C_{13} & C_{23} & C_{33} \end{pmatrix}.$$

Observe that the indexing of the cofactors here is column-row and not the usual row-column. If we define a cofactor matrix in the more standard way, with

$$\mathrm{C} = \begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{21} \\ C_{31} & C_{32} & C_{33} \end{pmatrix},$$

then we have determined that

$$A^{-1} = C^T / \det A,$$

where C^T is the transpose of C. The transpose of the cofactor matrix of A used to be called the adjoint matrix of A but is now called the adjugate matrix of A, and is denoted by adj(A). So the formula for the inverse matrix is often very neatly written as

$$A^{-1} = \mathrm{adj}(A) / \det A.$$

Example: Using the adjugate matrix of A, compute the inverse of $A = \begin{pmatrix} -3 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -4 & 4 \end{pmatrix}$.

We will need to compute one three-by-three determinant and nine two-by-two determinants. The three-by-three determinant was already found using

$$\begin{vmatrix} -3 & 2 & -1 \\ 6 & -6 & 7 \\ 3 & -4 & 4 \end{vmatrix} = \begin{vmatrix} -3 & 2 & -1 \\ 0 & -2 & 5 \\ 0 & -2 & 3 \end{vmatrix} = -3(-6+10) = -12.$$

The minor matrix is obtained by systematically deleting the rows and columns to compute all nine two-by-two determinants. I let the reader do this to find

$$M = \begin{pmatrix} 4 & 3 & -6 \\ 4 & -9 & 6 \\ 8 & -15 & 6 \end{pmatrix}.$$

The cofactor matrix is determined from the minor matrix after multiplying each element by either plus or minus one, so that

$$C = \begin{pmatrix} 4 & -3 & -6 \\ -4 & -9 & -6 \\ 8 & 15 & 6 \end{pmatrix}.$$

And the adjugate matrix is found from the transpose of the cofactor matrix:

$$\mathrm{adj}(A) = \begin{pmatrix} 4 & -4 & 8 \\ -3 & -9 & 15 \\ -6 & -6 & 6 \end{pmatrix}.$$

Then, dividing the adjugate matrix by $\det(A) = -12$, we find

$$A^{-1} = \mathrm{adj}(A)/\det A = \begin{pmatrix} -1/3 & 1/3 & -2/3 \\ 1/4 & 3/4 & -5/4 \\ 1/2 & 1/2 & -1/2 \end{pmatrix}.$$

4.6 Use of determinants in Vector Calculus

Consider two three-dimensional vectors $\mathbf{u} = u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k}$ and $\mathbf{v} = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$, written as you would find them in Calculus rather than as column matrices in Linear Algebra. The dot product of the two vectors is defined in Calculus as

$$\mathbf{u} \cdot \mathbf{v} = u_1v_1 + u_2v_2 + u_3v_3,$$

and the cross product is defined as

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} = \mathbf{i}(u_2v_3 - u_3v_2) - \mathbf{j}(u_1v_3 - u_3v_1) + \mathbf{k}(u_1v_2 - u_2v_1),$$

where the determinant from Linear Algebra is used as a mnemonic to remember the definition. If the angle between the two vectors is given by θ, then trigonometry can be used to show that

$$\mathbf{u} \cdot \mathbf{v} = |\mathbf{u}||\mathbf{v}| \cos\theta, \quad |\mathbf{u} \times \mathbf{v}| = |\mathbf{u}||\mathbf{v}| \sin\theta.$$

Now, if $\mathbf{u}$ and $\mathbf{v}$ lie in the x-y plane, then the area of the parallelogram formed from these two vectors, determined from base times height, is given by

$$\begin{aligned} \text{area} &= |\mathbf{u} \times \mathbf{v}| \\ &= |u_1v_2 - u_2v_1| \\ &= \left| \det \begin{pmatrix} u_1 & u_2 \\ v_1 & v_2 \end{pmatrix} \right|. \end{aligned}$$

This result also generalizes to three dimensions. The volume of a parallelopiped formed by the three vectors $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ is given by

$$\begin{aligned} \text{volume} &= |\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})| \\ &= \left| \det \begin{pmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{pmatrix} \right|. \end{aligned}$$

An important application of this result is the change-of-variable formula for multi-dimensional integration. Consider the double integral

$$I = \int \int_A \ldots dxdy$$

over some unspecified function of x and y and over some designated area A in the x-y plane. Suppose we make a linear transformation from the x-y coordinate system to some u-v coordinate system. That is, let

$$u = ax + by, \quad v = cx + dy,$$

or in matrix notation,

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Observe that the orthonormal basis vectors $\mathbf{i}$ and $\mathbf{j}$ transform into the vectors $a\mathbf{i} + c\mathbf{j}$ and $b\mathbf{i} + d\mathbf{j}$ so that a rectangle in the x-y coordinate system transforms into a parallelogram in the u-v coordinate system. The area A of the parallelogram in the u-v coordinate system is given by

$$A = \left| \det \begin{pmatrix} a & c \\ b & d \end{pmatrix} \right| = \left| \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} \right|.$$

Notice that because this was a linear transformation, we could have also written the area as

$$A = \left| \det \begin{pmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{pmatrix} \right|,$$

which is called the Jacobian determinant, or just the Jacobian. This result also applies to infinitesimal areas where a linear approximation can be made, and with

$$u = u(x, y), \quad v = v(x, y),$$

the change of variables formula becomes

$$dudv = \left| \det \frac{\partial(u, v)}{\partial(x, y)} \right| dxdy,$$

where in general, the Jacobian matrix is defined as

$$\frac{\partial(u, v)}{\partial(x, y)} = \begin{pmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{pmatrix}.$$

Note that sometimes we define the change of coordinates as

$$x = x(u, v), \quad y = y(u, v),$$

and the change of variables formula will be

$$dxdy = \left| \det \frac{\partial(x, y)}{\partial(u, v)} \right| dudv.$$

We can give two very important examples. The first in two dimensions is the change of variables from rectangular to polar coordinates. We have

$$x = r\cos\theta, \quad y = r\sin\theta,$$

and the Jacobian of the transformation is

$$\left| \det \frac{\partial(x, y)}{\partial(r, \theta)} \right| = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r.$$

So to find the area of a circle of radius R, with formula $x^2 + y^2 = R^2$, we have

$$\int_{-R}^{R} \int_{-\sqrt{R^2-y^2}}^{\sqrt{R^2-y^2}} dx\, dy = \int_0^{2\pi} \int_0^R r\, dr\, d\theta = \pi R^2.$$

The second example in three dimensions is from cartesian to spherical coordinates. Here,

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\theta.$$

The Jacobian is

$$\left| \det \frac{\partial(x, y, z)}{\partial(r, \theta, \phi)} \right| = \begin{vmatrix} \sin\theta\cos\phi & r\cos\theta\cos\phi & -r\sin\theta\sin\phi \\ \sin\theta\sin\phi & r\cos\theta\sin\phi & r\sin\theta\cos\phi \\ \cos\theta & -r\sin\theta & 0 \end{vmatrix} = r^2\sin\theta.$$

So to find the area of a sphere of radius R, with formula $x^2 + y^2 + z^2 = R^2$, we have

$$\int_{-R}^{R} \int_{-\sqrt{R^2-z^2}}^{\sqrt{R^2-z^2}} \int_{-\sqrt{R^2-y^2-z^2}}^{\sqrt{R^2-y^2-z^2}} dx\, dy\, dz = \int_0^{2\pi} \int_0^{\pi} \int_0^R r^2\sin\theta\, dr\, d\theta\, d\phi = \frac{4}{3}\pi R^3.$$

Chapter 5

Eigenvalues and eigenvectors

5.1 The eigenvalue problem

View *The Eigenvalue Problem* on YouTube
View *Finding Eigenvalues and Eigenvectors (Part 1)* on YouTube
View *Finding Eigenvalues and Eigenvectors (Part 2)* on YouTube

Let A be an n-by-n matrix, x an n-by-1 column vector, and λ a scalar. The eigenvalue problem for a given matrix A solves

$$\mathrm{A}\mathrm{x} = \lambda \mathrm{x} \tag{5.1}$$

for n eigenvalues λ_i with corresponding eigenvectors x_i. Since $\mathrm{I}\mathrm{x} = \mathrm{x}$, where I is the n-by-n identity matrix, we can rewrite the eigenvalue equation (5.1) in homogeneous form as

$$(\mathrm{A} - \lambda \mathrm{I})\mathrm{x} = 0. \tag{5.2}$$

The trivial solution to this equation is $\mathrm{x} = 0$, and for nontrivial solutions to exist, the n-by-n matrix $\mathrm{A} - \lambda \mathrm{I}$, which is the matrix A with λ subtracted from its main diagonal, must be singular. Hence, to determine the nontrivial solutions, we require that

$$\det(\mathrm{A} - \lambda \mathrm{I}) = 0. \tag{5.3}$$

Using the Leibniz formula for the determinant, we see that (5.3) is an n-th order polynomial equation in λ, called the *characteristic equation* of A. The characteristic equation can be solved for the eigenvalues, and for each eigenvalue, a corresponding eigenvector can be determined directly from (5.2).

We can demonstrate how to find the eigenvalues of a general 2-by-2 matrix given by

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

We have

$$\begin{aligned} 0 &= \det(\mathrm{A} - \lambda \mathrm{I}) \\ &= \begin{vmatrix} a - \lambda & b \\ c & d - \lambda \end{vmatrix} \\ &= (a - \lambda)(d - \lambda) - bc \\ &= \lambda^2 - (a + d)\lambda + (ad - bc), \end{aligned}$$

which can be more generally written as

$$\lambda^2 - \mathrm{Tr}\,\mathrm{A}\ \lambda + \det \mathrm{A} = 0, \tag{5.4}$$

where Tr A is the trace, or sum of the diagonal elements, of the matrix A.

Example: The Cayley-Hamilton theorem states that every square matrix satisfies its own characteristic equation. By explicit calculation, prove the Cayley-Hamilton theorem for a two-by-two matrix.

The Cayley-Hamilton theorem for a two-by-two matrix A states that

$$A^2 - (\mathrm{Tr}\, A)A + (\det A)I = 0.$$

Notice that we need to replace the one in the third term of the characteristic polynomial by the identity matrix so that each term in the polynomial becomes a two-by-two matrix. Let

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

We compute:

$$A^2 - (\mathrm{Tr}\, A)A + (\det A)I = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} - (a+d)\begin{pmatrix} a & b \\ c & d \end{pmatrix} + (ad-bc)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} a^2+bc & ab+bd \\ ac+cd & bc+d^2 \end{pmatrix} - \begin{pmatrix} a^2+ad & ab+bd \\ ac+cd & ad+d^2 \end{pmatrix} + \begin{pmatrix} ad-bc & 0 \\ 0 & ad-bc \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

thereby proving the Cayley-Hamilton theorem for two-by-two matrices.

Example: Use the Cayley-Hamilton theorem to find the inverse of a two-by-two matrix.
Again, the Cayley-Hamilton theorem for a two-by-two matrix is

$$A^2 - (\mathrm{Tr}\, A)A + (\det A)I = 0.$$

We multiply this equation by A^{-1} to obtain

$$A - (\mathrm{Tr}\, A)I + (\det A)A^{-1} = 0.$$

Assuming $\det A \neq 0$, we can solve for A^{-1} to obtain

$$A^{-1} = \frac{1}{\det A}\big((\mathrm{Tr}\, A)I - A\big).$$

With

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

we have

$$\big((\mathrm{Tr}\, A)I - A\big) = \begin{pmatrix} a+d & 0 \\ 0 & a+d \end{pmatrix} - \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

so that

$$A^{-1} = \frac{1}{\det A}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Since the characteristic equation of a two-by-two matrix is a quadratic equation, it can have either (i) two distinct real roots; (ii) two distinct complex conjugate roots; or (iii) one degenerate real root. That is, eigenvalues and eigenvectors can be real or complex, and that for certain defective matrices, there may be less than n distinct eigenvalues and eigenvectors.

If λ_1 is an eigenvalue of our 2-by-2 matrix A, then the corresponding eigenvector x_1 may be found by solving

$$\begin{pmatrix} a-\lambda_1 & b \\ c & d-\lambda_1 \end{pmatrix}\begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix} = 0, \tag{5.5}$$

where the equation of the second row will always be a multiple of the equation of the first row because the determinant of the matrix on the left-hand-side is zero. The eigenvector x_1 can be multiplied by any nonzero constant and still be an eigenvector. We could normalize x_1, for instance, by taking $x_{11} = 1$ or $|\mathrm{x}_1| = 1$, or whatever, depending on our needs.

The equation from the first row of (5.5) is

$$(a - \lambda_1)x_{11} + bx_{21} = 0,$$

and we could take $x_{11} = 1$ to find $x_{21} = (\lambda_1 - a)/b$. These results are usually derived as needed when given specific matrices.

Example: Find the eigenvalues and eigenvectors of the following matrices:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}.$$

For

$$\mathrm{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

the characteristic equation is

$$\lambda^2 - 1 = 0,$$

with solutions $\lambda_1 = 1$ and $\lambda_2 = -1$. The first eigenvector is found by solving $(\mathrm{A} - \lambda_1 \mathrm{I})\mathrm{x}_1 = 0$, or

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix} = 0,$$

so that $x_{21} = x_{11}$. The second eigenvector is found by solving $(\mathrm{A} - \lambda_2 \mathrm{I})\mathrm{x}_2 = 0$, or

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix} = 0,$$

so that $x_{22} = -x_{12}$. The eigenvalues and eigenvectors are therefore given by

$$\lambda_1 = 1, \;\; \mathrm{x}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}; \quad \lambda_2 = -1, \;\; \mathrm{x}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

To find the eigenvalues and eigenvectors of the second matrix we can follow this same procedure. Or better yet, we can take a shortcut. Let

$$\mathrm{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \mathrm{B} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}.$$

We know the eigenvalues and eigenvectors of A and that $\mathrm{B} = \mathrm{A} + 3\mathrm{I}$. Therefore, with λ_B representing the eigenvalues of B, and λ_A representing the eigenvalues of A, we have

$$0 = \det(\mathrm{B} - \lambda_\mathrm{B}\mathrm{I}) = \det(\mathrm{A} + 3\mathrm{I} - \lambda_\mathrm{B}\mathrm{I}) = \det(\mathrm{A} - (\lambda_\mathrm{B} - 3)\mathrm{I}) = \det(\mathrm{A} - \lambda_\mathrm{A}\mathrm{I}).$$

Therefore, $\lambda_\mathrm{B} = \lambda_\mathrm{A} + 3$ and the eigenvalues of B are 4 and 2. The eigenvectors remain the same.

It is useful to notice that, for a two-by-two matrix, $\lambda_1 + \lambda_2 = \mathrm{Tr}\,\mathrm{A}$ and that $\lambda_1\lambda_2 = \det \mathrm{A}$. The analogous result for n-by-n matrices is also true and worthwhile to remember. In particular, summing the eigenvalues and comparing to the trace of the matrix provides a rapid check on your algebra.

Example: Find the eigenvalues and eigenvectors of the following matrices

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

For

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

the characteristic equation is

$$\lambda^2 + 1 = 0,$$

with solutions i and $-i$. Notice that if the matrix A is real, then the complex conjugate of the eigenvalue equation $\mathrm{Ax} = \lambda \mathrm{x}$ is $\mathrm{A}\bar{\mathrm{x}} = \bar{\lambda}\bar{\mathrm{x}}$. So if λ and x are an eigenvalue and eigenvector of a real matrix A, then so are the complex conjugates $\bar{\lambda}$ and $\bar{\mathrm{x}}$. Eigenvalues and eigenvectors of a real matrix appear as complex conjugate pairs.

The eigenvector associated with $\lambda = i$ is determined from $(\mathrm{A} - i\mathrm{I})\mathrm{x} = 0$, or

$$\begin{pmatrix} -i & -1 \\ 1 & -i \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0,$$

or $x_1 = ix_2$. The eigenvectors and eigenvectors of A are therefore given by

$$\lambda = i, \;\; \mathrm{x} = \begin{pmatrix} i \\ 1 \end{pmatrix}; \quad \bar{\lambda} = -i, \;\; \bar{\mathrm{x}} = \begin{pmatrix} -i \\ 1 \end{pmatrix}.$$

For

$$\mathrm{B} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

the characteristic equation is

$$\lambda^2 = 0,$$

so that there is a degenerate eigenvalue of zero. The eigenvector associated with the zero eigenvalue is found from $\mathrm{Bx} = 0$ and has zero second component. Therefore, this matrix is defective and has only one eigenvalue and eigenvector given by

$$\lambda = 0, \;\; \mathrm{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Example: Find the eigenvalues and eigenvectors of the rotation matrix

$$\mathrm{R} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

The characteristic equation is given by

$$\lambda^2 - 2\cos\theta\lambda + 1 = 0,$$

with solution

$$\lambda_\pm = \cos\theta \pm \sqrt{\cos^2\theta - 1} = \cos\theta \pm i\sin\theta = e^{\pm i\theta}.$$

The eigenvector corresponding to $\lambda = e^{i\theta}$ is found from

$$-i\sin\theta x_1 - \sin\theta x_2 = 0,$$

or $x_2 = -ix_1$. Therefore, the eigenvalues and eigenvectors are

$$\lambda = e^{i\theta}, \;\; \mathrm{x} = \begin{pmatrix} 1 \\ -i \end{pmatrix}$$

and their complex conjugates.

5.2 Matrix diagonalization

View *Matrix Diagonalization* on YouTube
View *Powers of a Matrix* on YouTube
View *Powers of a Matrix Example* on YouTube

For concreteness, consider a 2-by-2 matrix A with eigenvalues and eigenvectors given by

$$\lambda_1, \ \mathrm{x}_1 = \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}; \quad \lambda_2, \ \mathrm{x}_2 = \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix}.$$

Now, consider the matrix product and factorization

$$\mathrm{A}\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} = \begin{pmatrix} \lambda_1 x_{11} & \lambda_2 x_{12} \\ \lambda_1 x_{21} & \lambda_2 x_{22} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

We define S to be the matrix whose columns are the eigenvectors of A, and Λ to be the diagonal eigenvalue matrix. Then generalizing to any square matrix with a complete set of eigenvectors, we have

$$\mathrm{AS} = \mathrm{S}\Lambda.$$

Multiplying both sides on the right or the left by S^{-1}, we have found

$$\mathrm{A} = \mathrm{S}\Lambda\mathrm{S}^{-1} \quad \text{and} \quad \Lambda = \mathrm{S}^{-1}\mathrm{AS}.$$

To memorize the order of the S matrices in these formulas, just remember that A should be multiplied on the right by S.

Diagonalizing a matrix facilitates finding powers of that matrix. For instance,

$$\mathrm{A}^2 = (\mathrm{S}\Lambda\mathrm{S}^{-1})(\mathrm{S}\Lambda\mathrm{S}^{-1}) = \mathrm{S}\Lambda^2\mathrm{S}^{-1},$$

where in the 2-by-2 example, Λ^2 is simply

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{pmatrix}.$$

In general, Λ^2 has the eigenvalues squared down the diagonal. More generally, for p a positive integer,

$$\mathrm{A}^p = \mathrm{S}\Lambda^p\mathrm{S}^{-1}.$$

Example: Recall the Fibonacci Q-matrix, which satisfies

$$\mathrm{Q} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathrm{Q}^n = \begin{pmatrix} F_{n+1} & F_n \\ F_n & F_{n-1} \end{pmatrix}.$$

Using Q and Q^n, derive Binet's formula for F_n.

The characteristic equation of Q is given by

$$\lambda^2 - \lambda - 1 = 0,$$

with solutions

$$\lambda_1 = \frac{1+\sqrt{5}}{2} = \Phi, \quad \lambda_2 = \frac{1-\sqrt{5}}{2} = -\phi.$$

The irrational number Φ is called the golden ratio, and the irrational number ϕ is called the golden ratio conjugate. The numerical values are approximately $\Phi = 1.618\ldots$ and $\phi = 0.618\ldots$. Useful identities are

$$\Phi = 1 + \phi, \quad \Phi = 1/\phi, \quad \text{and} \quad \Phi + \phi = \sqrt{5}.$$

The eigenvector corresponding to Φ can be found from

$$x_1 - \Phi x_2 = 0,$$

and the eigenvector corresponding to $-\phi$ can be found from

$$x_1 + \phi x_2 = 0.$$

Therefore, the eigenvalues and eigenvectors can be written as

$$\lambda_1 = \Phi, \;\; \mathrm{x}_1 = \begin{pmatrix} \Phi \\ 1 \end{pmatrix}; \quad \lambda_2 = -\phi, \;\; \mathrm{x}_2 = \begin{pmatrix} -\phi \\ 1 \end{pmatrix}.$$

The eigenvector matrix S becomes

$$\mathrm{S} = \begin{pmatrix} \Phi & -\phi \\ 1 & 1 \end{pmatrix};$$

and the inverse of this 2-by-2 matrix is given by

$$\mathrm{S}^{-1} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & \phi \\ -1 & \Phi \end{pmatrix}.$$

Our diagonalization is therefore

$$\mathrm{Q} = \frac{1}{\sqrt{5}} \begin{pmatrix} \Phi & -\phi \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \Phi & 0 \\ 0 & -\phi \end{pmatrix} \begin{pmatrix} 1 & \phi \\ -1 & \Phi \end{pmatrix}.$$

Raising to the nth power, we have

$$\begin{aligned} \mathrm{Q}^n &= \frac{1}{\sqrt{5}} \begin{pmatrix} \Phi & -\phi \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \Phi^n & 0 \\ 0 & (-\phi)^n \end{pmatrix} \begin{pmatrix} 1 & \phi \\ -1 & \Phi \end{pmatrix} \\ &= \frac{1}{\sqrt{5}} \begin{pmatrix} \Phi & -\phi \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \Phi^n & \Phi^{n-1} \\ -(-\phi)^n & -(-\phi)^{n-1} \end{pmatrix} \\ &= \frac{1}{\sqrt{5}} \begin{pmatrix} \Phi^{n+1} - (-\phi)^{n+1} & \Phi^n - (-\phi)^n \\ \Phi^n - (-\phi)^n & \Phi^{n-1} - (-\phi)^{n-1} \end{pmatrix}. \end{aligned}$$

Using Q^n written in terms of the Fibonacci numbers, we have derived Binet's formula

$$F_n = \frac{\Phi^n - (-\phi)^n}{\sqrt{5}}.$$

5.3 Symmetric and Hermitian matrices

When a real matrix A is equal to its transpose, $\mathrm{A}^{\mathrm{T}} = \mathrm{A}$, we say that the matrix is symmetric. When a complex matrix A is equal to its conjugate transpose, $\mathrm{A}^{\dagger} = \mathrm{A}$, we say that the matrix is Hermitian.

One of the reasons symmetric and Hermitian matrices are important is because their eigenvalues are real and their eigenvectors are orthogonal. Let λ_i and λ_j be eigenvalues and x_i and x_j eigenvectors of the possibly complex matrix A. We have

$$\mathrm{A}x_i = \lambda_i x_i, \quad \mathrm{A}x_j = \lambda_j x_j.$$

Multiplying the first equation on the left by $x_j^\dagger$, and taking the conjugate transpose of the second equation and multiplying on the right by x_i, we obtain

$$x_j^\dagger \mathrm{A}x_i = \lambda_i x_j^\dagger x_i, \quad x_j^\dagger \mathrm{A}^\dagger x_i = \bar{\lambda}_j x_j^\dagger x_i.$$

If A is Hermitian, then $\mathrm{A}^\dagger = \mathrm{A}$, and subtracting the second equation from the first yields

$$(\lambda_i - \bar{\lambda}_j) x_j^\dagger x_i = 0.$$

If $i = j$, then since $x_i^\dagger x_i > 0$, we have $\bar{\lambda}_i = \lambda_i$: all eigenvalues are real. If $i \neq j$ and $\lambda_i \neq \lambda_j$, then $x_j^\dagger x_i = 0$: eigenvectors with distinct eigenvalues are orthogonal. Usually, the eigenvectors are made orthonormal, and diagonalization makes use of real orthogonal or complex unitary matrices.

Example: Diagonalize the symmetric matrix

$$\mathrm{A} = \begin{pmatrix} a & b \\ b & a \end{pmatrix}.$$

The characteristic equation of A is given by

$$(a - \lambda)^2 = b^2,$$

with real eigenvalues $\lambda_1 = a + b$ and $\lambda_2 = a - b$. The eigenvector with eigenvalue λ_1 satisfies $-x_1 + x_2 = 0$, and the eigenvector with eigenvalue λ_2 satisfies $x_1 + x_2 = 0$. Normalizing the eigenvectors, we have

$$\lambda_1 = a + b,\ \mathrm{X}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}; \qquad \lambda_2 = a - b,\ \mathrm{X}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Evidently, the eigenvectors are orthonormal. The diagonalization using $\mathrm{A} = \mathrm{Q}\Lambda\mathrm{Q}^{-1}$ is given by

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} a+b & 0 \\ 0 & a-b \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

which can be verified directly by matrix multiplication. The matrix Q is a symmetric orthogonal matrix so that $\mathrm{Q}^{-1} = \mathrm{Q}$.

Part II

Differential equations

The second part of this course is on differential equations. We begin with first-order odes and explain how to solve separable and linear equations. A range of applications are given. We then discuss the important case of second-order odes with constant coefficients. Homogeneous and inhomogeneous equations are solved, and the phenomena of resonance is discussed. When the coefficients are not constant, a series solution is often required and we discuss this important technique. We next study a system of linear differential equations and show how some of our knowledge of linear algebra can aid in their solution. We finish by considering nonlinear equations and the ideas of fixed points and linear stability analysis.

Chapter 6

Introduction to odes

A differential equation is an equation for a function that relates the values of the function to the values of its derivatives. An ordinary differential equation (ode) is a differential equation for a function of a single variable, e.g., $x(t)$, while a partial differential equation (pde) is a differential equation for a function of several variables, e.g., $v(x, y, z, t)$. An ode contains ordinary derivatives and a pde contains partial derivatives. Typically, pde's are much harder to solve than ode's.

6.1 The simplest type of differential equation

View tutorial on YouTube

The simplest ordinary differential equations can be integrated directly by finding antiderivatives. These simplest odes have the form

$$\frac{d^n x}{dt^n} = G(t),$$

where the derivative of $x = x(t)$ can be of any order, and the right-hand-side may depend only on the independent variable t. As an example, consider a mass falling under the influence of constant gravity, such as approximately found on the Earth's surface. Newton's law, $F = ma$, results in the equation

$$m\frac{d^2 x}{dt^2} = -mg,$$

where x is the height of the object above the ground, m is the mass of the object, and $g = 9.8$ meter/sec^2 is the constant gravitational acceleration. As Galileo suggested, the mass cancels from the equation, and

$$\frac{d^2 x}{dt^2} = -g.$$

Here, the right-hand-side of the ode is a constant. The first integration, obtained by antidifferentiation, yields

$$\frac{dx}{dt} = A - gt,$$

with A the first constant of integration; and the second integration yields

$$x = B + At - \frac{1}{2}gt^2,$$

with B the second constant of integration. The two constants of integration A and B can then be determined from the initial conditions. If we know that the initial height of the mass is x_0, and the initial velocity is v_0, then the initial conditions are

$$x(0) = x_0, \quad \frac{dx}{dt}(0) = v_0.$$

Substitution of these initial conditions into the equations for dx/dt and x allows us to solve for A and B. The unique solution that satisfies both the ode and the initial conditions is given by

$$x(t) = x_0 + v_0 t - \frac{1}{2} g t^2. \tag{6.1}$$

For example, suppose we drop a ball off the top of a 50 meter building. How long will it take the ball to hit the ground? This question requires solution of (6.1) for the time T it takes for $x(T) = 0$, given $x_0 = 50$ meter and $v_0 = 0$. Solving for T,

$$\begin{aligned} T &= \sqrt{\frac{2x_0}{g}} \\ &= \sqrt{\frac{2 \cdot 50}{9.8}} \text{sec} \\ &\approx 3.2 \text{sec}. \end{aligned}$$

Chapter 7

First-order differential equations

Reference: Boyce and DiPrima, Chapter 2

The general first-order differential equation for the function $y = y(x)$ is written as

$$\frac{dy}{dx} = f(x, y), \tag{7.1}$$

where $f(x, y)$ can be any function of the independent variable x and the dependent variable y. We first show how to determine a numerical solution of this equation, and then learn techniques for solving analytically some special forms of (7.1), namely, *separable* and *linear* first-order equations.

7.1 The Euler method

View tutorial on YouTube

Although it is not always possible to find an analytical solution of (7.1) for $y = y(x)$, it is always possible to determine a unique numerical solution given an initial value $y(x_0) = y_0$, and provided $f(x, y)$ is a well-behaved function. The differential equation (7.1) gives us the slope $f(x_0, y_0)$ of the tangent line to the solution curve $y = y(x)$ at the point (x_0, y_0). With a small step size $\Delta x = x_1 - x_0$, the initial condition (x_0, y_0) can be marched forward to (x_1, y_1) along the tangent line using Euler's method (see Fig. 7.1)

$$y_1 = y_0 + \Delta x f(x_0, y_0).$$

This solution (x_1, y_1) then becomes the new initial condition and is marched forward to (x_2, y_2) along a newly determined tangent line with slope given by $f(x_1, y_1)$. For small enough Δx, the numerical solution converges to the exact solution.

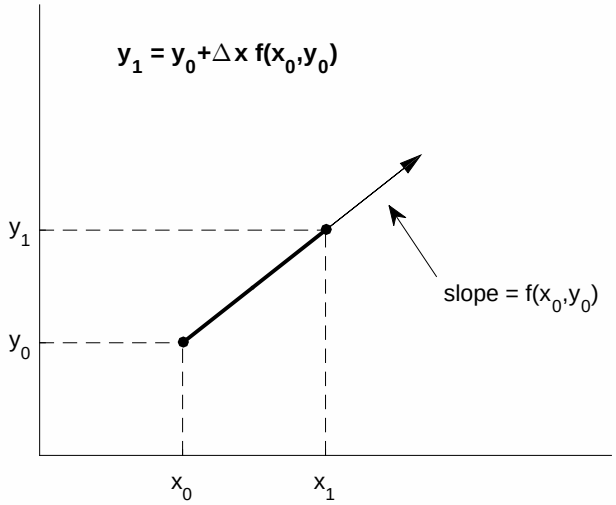


Figure 7.1: *The differential equation* $dy/dx = f(x, y)$, $y(x_0) = y_0$, *is integrated to* $x = x_1$ *using the Euler method* $y_1 = y_0 + \Delta x f(x_0, y_0)$, *with* $\Delta x = x_1 - x_0$.

7.2 Separable equations

View tutorial on YouTube

A first-order ode is separable if it can be written in the form

$$g(y)\frac{dy}{dx} = f(x), \quad y(x_0) = y_0, \tag{7.2}$$

where the function $g(y)$ is independent of x and $f(x)$ is independent of y. Integration from x_0 to x results in

$$\int_{x_0}^{x} g(y(x))y'(x)dx = \int_{x_0}^{x} f(x)dx.$$

Noticing that $dy = y'(x)dx$, using $y(x_0) = y_0$ and denoting $y(x) = y$, we have upon changing variables

$$\int_{y_0}^{y} g(y)dy = \int_{x_0}^{x} f(x)dx. \tag{7.3}$$

A simpler procedure that also yields (7.3) is to treat dy/dx in (7.2) like a fraction. Multiplying (7.2) by dx results in

$$g(y)dy = f(x)dx,$$

which is a separated equation with all the dependent variables on the left-side, and all the independent variables on the right-side. Equation (7.3) then results directly upon integration.

Example: Solve $\frac{dy}{dx} + \frac{1}{2}y = \frac{3}{2}$**, with** $y(0) = 2$**.**

We first manipulate the differential equation to the form

$$\frac{dy}{dx} = \frac{1}{2}(3 - y), \tag{7.4}$$

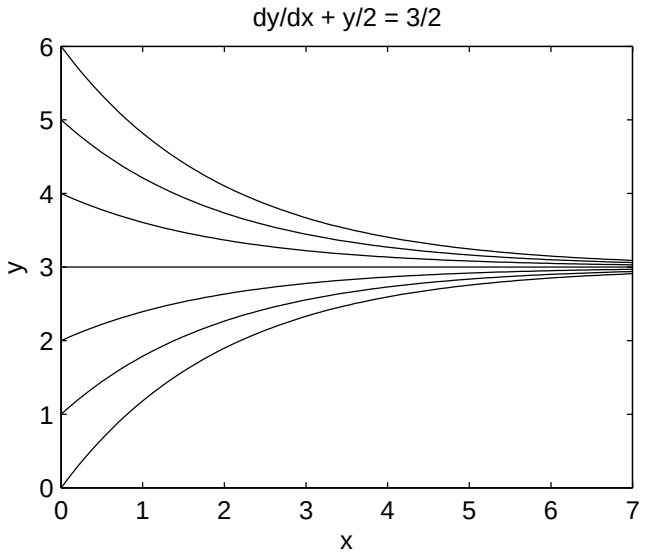


Figure 7.2: *Solution of the following ode:* $\frac{dy}{dx} + \frac{1}{2}y = \frac{3}{2}$.

and then treat dy/dx as if it was a fraction to separate variables:

$$\frac{dy}{3-y} = \frac{1}{2}dx.$$

We integrate the right-side from the initial condition $x = 0$ to x and the left-side from the initial condition $y(0) = 2$ to y. Accordingly,

$$\int_2^y \frac{dy}{3-y} = \frac{1}{2}\int_0^x dx. \tag{7.5}$$

The integrals in (7.5) need to be done. Note that $y(x) < 3$ for finite x or the integral on the left-side diverges. Therefore, $3 - y > 0$ and integration yields

$$\begin{aligned} -\ln(3-y)\big]_2^y &= \frac{1}{2}x\big]_0^x, \\ \ln(3-y) &= -\frac{1}{2}x, \\ 3-y &= e^{-x/2}, \\ y &= 3 - e^{-x/2}. \end{aligned}$$

Since this is our first nontrivial analytical solution, it is prudent to check our result. We do this by differentiating our solution:

$$\begin{aligned} \frac{dy}{dx} &= \frac{1}{2}e^{-x/2} \\ &= \frac{1}{2}(3-y); \end{aligned}$$

and checking the initial condition, $y(0) = 3 - e^0 = 2$. Therefore, our solution satisfies both the original ode and the initial condition.

Example: Solve $\frac{dy}{dx} + \frac{1}{2}y = \frac{3}{2}$, with $y(0) = 4$.

This is the identical differential equation as before, but with different initial conditions. We will jump directly to the integration step:

$$\int_4^y \frac{dy}{3-y} = \frac{1}{2}\int_0^x dx.$$

Now $y(x) > 3$, so that $y - 3 > 0$ and integration yields

$$\begin{aligned} -\ln(y-3)\big]_4^y &= \frac{1}{2}x\big]_0^x, \\ \ln(y-3) &= -\frac{1}{2}x, \\ y-3 &= e^{-x/2}, \\ y &= 3 + e^{-x/2}. \end{aligned}$$

The solution curves for a range of initial conditions are presented in Fig. 7.2. All solutions have a horizontal asymptote at $y = 3$ at which $dy/dx = 0$. For $y(0) = y_0$, the general solution can be shown to be $y(x) = 3 + (y_0 - 3)\exp(-x/2)$.

Example: Solve $\frac{dy}{dx} = \frac{2\cos 2x}{3+2y}$, with $y(0) = -1$. (i) For what values of $x > 0$ does the solution exist? (ii) For what value of $x > 0$ is $y(x)$ maximum?

Notice that the derivative of y diverges when $y = -3/2$, and that this may cause some problems with a solution.

We solve the ode by separating variables and integrating from initial conditions:

$$\begin{aligned} (3+2y)dy &= 2\cos 2x\, dx \\ \int_{-1}^y (3+2y)dy &= 2\int_0^x \cos 2x\, dx \\ 3y + y^2\big]_{-1}^y &= \sin 2x\big]_0^x \\ y^2 + 3y + 2 - \sin 2x &= 0 \\ y_\pm &= \frac{1}{2}[-3 \pm \sqrt{1 + 4\sin 2x}]. \end{aligned}$$

Solving the quadratic equation for y has introduced a spurious solution that does not satisfy the initial conditions. We test:

$$y_\pm(0) = \frac{1}{2}[-3 \pm 1] = \begin{cases} \text{-1;} \\ \text{-2.} \end{cases}$$

Only the $+$ root satisfies the initial condition, so that the unique solution to the ode and initial condition is

$$y = \frac{1}{2}[-3 + \sqrt{1 + 4\sin 2x}]. \tag{7.6}$$

To determine (i) the values of $x > 0$ for which the solution exists, we require

$$1 + 4\sin 2x \geq 0,$$

or

$$\sin 2x \geq -\frac{1}{4}. \tag{7.7}$$

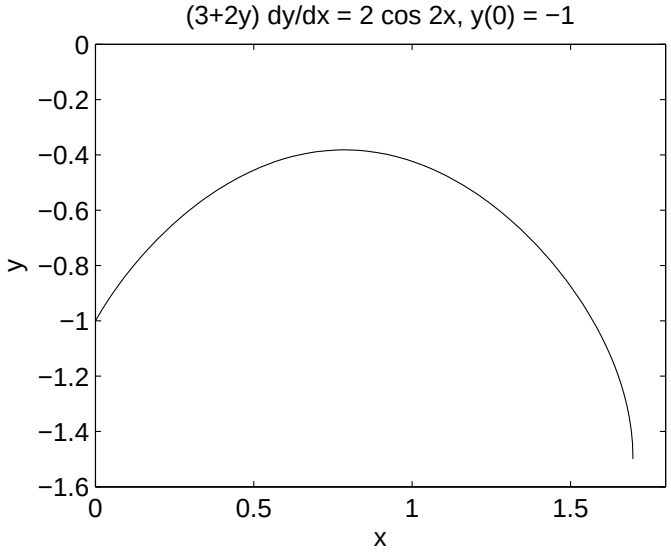


Figure 7.3: *Solution of the following ode:* $(3+2y)y' = 2\cos 2x$, $y(0) = -1$.

Notice that at $x = 0$, we have $\sin 2x = 0$; at $x = \pi/4$, we have $\sin 2x = 1$; at $x = \pi/2$, we have $\sin 2x = 0$; and at $x = 3\pi/4$, we have $\sin 2x = -1$ We therefore need to determine the value of x such that $\sin 2x = -1/4$, with x in the range $\pi/2 < x < 3\pi/4$. The solution to the ode will then exist for all x between zero and this value.

To solve $\sin 2x = -1/4$ for x in the interval $\pi/2 < x < 3\pi/4$, one needs to recall the definition of arcsin, or $\sin^{-1}$, as found on a typical scientific calculator. The inverse of the function

$$f(x) = \sin x, \quad -\pi/2 \leq x \leq \pi/2$$

is denoted by arcsin. The first solution with $x > 0$ of the equation $\sin 2x = -1/4$ places $2x$ in the interval $(\pi, 3\pi/2)$, so to invert this equation using the arcsine we need to apply the identity $\sin(\pi - x) = \sin x$, and rewrite $\sin 2x = -1/4$ as $\sin(\pi - 2x) = -1/4$. The solution of this equation may then be found by taking the arcsine, and is

$$\pi - 2x = \arcsin(-1/4),$$

or

$$x = \frac{1}{2}\left(\pi + \arcsin\frac{1}{4}\right).$$

Therefore the solution exists for $0 \leq x \leq (\pi + \arcsin(1/4))/2 = 1.6971\ldots$, where we have used a calculator value (computing in radians) to find $\arcsin(0.25) = 0.2527\ldots$. At the value $(x, y) = (1.6971\ldots, -3/2)$, the solution curve ends and dy/dx becomes infinite.

To determine (ii) the value of x at which $y = y(x)$ is maximum, we examine (7.6) directly. The value of y will be maximum when $\sin 2x$ takes its maximum value over the interval where the solution exists. This will be when $2x = \pi/2$, or $x = \pi/4 = 0.7854\ldots$.

The graph of $y = y(x)$ is shown in Fig. 7.3.

7.3 Linear equations

View tutorial on YouTube

The linear first-order differential equation (linear in y and its derivative) can be written in the form

$$\frac{dy}{dx} + p(x)y = g(x), \tag{7.8}$$

with the initial condition $y(x_0) = y_0$. Linear first-order equations can be integrated using an integrating factor $\mu(x)$. We multiply (7.8) by $\mu(x)$,

$$\mu(x)\left[\frac{dy}{dx} + p(x)y\right] = \mu(x)g(x), \tag{7.9}$$

and try to determine $\mu(x)$ so that

$$\mu(x)\left[\frac{dy}{dx} + p(x)y\right] = \frac{d}{dx}[\mu(x)y]. \tag{7.10}$$

Equation (7.9) then becomes

$$\frac{d}{dx}[\mu(x)y] = \mu(x)g(x). \tag{7.11}$$

Equation (7.11) is easily integrated using $\mu(x_0) = \mu_0$ and $y(x_0) = y_0$:

$$\mu(x)y - \mu_0 y_0 = \int_{x_0}^{x} \mu(x)g(x)dx,$$

or

$$y = \frac{1}{\mu(x)}\left(\mu_0 y_0 + \int_{x_0}^{x} \mu(x)g(x)dx\right). \tag{7.12}$$

It remains to determine $\mu(x)$ from (7.10). Differentiating and expanding (7.10) yields

$$\mu\frac{dy}{dx} + p\mu y = \frac{d\mu}{dx}y + \mu\frac{dy}{dx};$$

and upon simplifying,

$$\frac{d\mu}{dx} = p\mu. \tag{7.13}$$

Equation (7.13) is separable and can be integrated:

$$\int_{\mu_0}^{\mu} \frac{d\mu}{\mu} = \int_{x_0}^{x} p(x)dx,$$
$$\ln\frac{\mu}{\mu_0} = \int_{x_0}^{x} p(x)dx,$$
$$\mu(x) = \mu_0 \exp\left(\int_{x_0}^{x} p(x)dx\right).$$

Notice that since μ_0 cancels out of (7.12), it is customary to assign $\mu_0 = 1$. The solution to (7.8) satisfying the initial condition $y(x_0) = y_0$ is then commonly written as

$$y = \frac{1}{\mu(x)}\left(y_0 + \int_{x_0}^{x} \mu(x)g(x)dx\right),$$

with

$$\mu(x) = \exp\left(\int_{x_0}^{x} p(x)dx\right)$$

the integrating factor. This important result finds frequent use in applied mathematics.

Example: Solve $\frac{dy}{dx} + 2y = e^{-x}$, with $y(0) = 3/4$.

Note that this equation is not separable. With $p(x) = 2$ and $g(x) = e^{-x}$, we have

$$\begin{aligned}\mu(x) &= \exp\left(\int_0^x 2dx\right)\\ &= e^{2x},\end{aligned}$$

and

$$\begin{aligned}y &= e^{-2x}\left(\frac{3}{4} + \int_0^x e^{2x}e^{-x}dx\right)\\ &= e^{-2x}\left(\frac{3}{4} + \int_0^x e^{x}dx\right)\\ &= e^{-2x}\left(\frac{3}{4} + (e^x - 1)\right)\\ &= e^{-2x}\left(e^x - \frac{1}{4}\right)\\ &= e^{-x}\left(1 - \frac{1}{4}e^{-x}\right).\end{aligned}$$

Example: Solve $\frac{dy}{dx} - 2xy = x$, with $y(0) = 0$.

This equation is separable, and we solve it in two ways. First, using an integrating factor with $p(x) = -2x$ and $g(x) = x$:

$$\begin{aligned}\mu(x) &= \exp\left(-2\int_0^x xdx\right)\\ &= e^{-x^2},\end{aligned}$$

and

$$y = e^{x^2}\int_0^x xe^{-x^2}dx.$$

The integral can be done by substitution with $u = x^2$, $du = 2xdx$:

$$\begin{aligned}\int_0^x xe^{-x^2}dx &= \frac{1}{2}\int_0^{x^2} e^{-u}du\\ &= -\frac{1}{2}e^{-u}\Big]_0^{x^2}\\ &= \frac{1}{2}\left(1 - e^{-x^2}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}y &= \frac{1}{2}e^{x^2}\left(1 - e^{-x^2}\right)\\ &= \frac{1}{2}\left(e^{x^2} - 1\right).\end{aligned}$$

Second, we integrate by separating variables:

$$\begin{aligned}
\frac{dy}{dx} - 2xy &= x, \\
\frac{dy}{dx} &= x(1+2y), \\
\int_0^y \frac{dy}{1+2y} &= \int_0^x x dx, \\
\frac{1}{2}\ln(1+2y) &= \frac{1}{2}x^2, \\
1+2y &= e^{x^2}, \\
y &= \frac{1}{2}\left(e^{x^2}-1\right).
\end{aligned}$$

The results from the two different solution methods are the same, and the choice of method is a personal preference.

7.4 Applications

7.4.1 Compound interest

View tutorial on YouTube

The equation for the growth of an investment with continuous compounding of interest is a first-order differential equation. Let $S(t)$ be the value of the investment at time t, and let r be the annual interest rate compounded after every time interval Δt. We can also include deposits (or withdrawals). Let k be the annual deposit amount, and suppose that an installment is deposited after every time interval Δt. The value of the investment at the time $t + \Delta t$ is then given by

$$S(t+\Delta t) = S(t) + (r\Delta t)S(t) + k\Delta t, \tag{7.14}$$

where at the end of the time interval Δt, $r\Delta t S(t)$ is the amount of interest credited and $k\Delta t$ is the amount of money deposited ($k > 0$) or withdrawn ($k < 0$). As a numerical example, if the account held \$10,000 at time t, and $r = 6\%$ per year and $k = \$12{,}000$ per year, say, and the compounding and deposit period is $\Delta t = 1$ month $= 1/12$ year, then the interest awarded after one month is $r\Delta t S = (0.06/12) \times \$10{,}000 = \$50$, and the amount deposited is $k\Delta t = \$1000$.

Rearranging the terms of (7.14) to exhibit what will soon become a derivative, we have

$$\frac{S(t+\Delta t) - S(t)}{\Delta t} = rS(t) + k.$$

The equation for continuous compounding of interest and continuous deposits is obtained by taking the limit $\Delta t \to 0$. The resulting differential equation is

$$\frac{dS}{dt} = rS + k, \tag{7.15}$$

which can solved with the initial condition $S(0) = S_0$, where S_0 is the initial capital. We can solve either by separating variables or by using an integrating factor; I solve

here by separating variables. Integrating from $t = 0$ to a final time t,

$$\begin{aligned}
\int_{S_0}^{S} \frac{dS}{rS+k} &= \int_0^t dt, \\
\frac{1}{r} \ln\left(\frac{rS+k}{rS_0+k}\right) &= t, \\
rS + k &= (rS_0 + k)e^{rt}, \\
S &= \frac{rS_0 e^{rt} + ke^{rt} - k}{r}, \\
S &= S_0 e^{rt} + \frac{k}{r} e^{rt}\left(1 - e^{-rt}\right), \qquad (7.16)
\end{aligned}$$

where the first term on the right-hand side of (7.16) comes from the initial invested capital, and the second term comes from the deposits (or withdrawals). Evidently, compounding results in the exponential growth of an investment.

As a practical example, we can analyze a simple retirement plan. It is easiest to assume that all amounts and returns are in real dollars (adjusted for inflation). Suppose a 25 year-old plans to set aside a fixed amount every year of his/her working life, invests at a real return of 6%, and retires at age 65. How much must he/she invest each year to have HK$8,000,000 at retirement? (Note: 1 US$≈ 8 HK$.) We need to solve (7.16) for k using $t = 40$ years, $S(t) = \$8{,}000{,}000$, $S_0 = 0$, and $r = 0.06$ per year. We have

$$\begin{aligned}
k &= \frac{rS(t)}{e^{rt} - 1}, \\
k &= \frac{0.06 \times 8{,}000{,}000}{e^{0.06 \times 40} - 1}, \\
&= \$47{,}889\,\text{year}^{-1}.
\end{aligned}$$

To have saved approximately one million US$ at retirement, the worker would need to save about HK$50,000 per year over his/her working life. Note that the amount saved over the worker's life is approximately $40 \times \$50{,}000 = \$2{,}000{,}000$, while the amount earned on the investment (at the assumed 6% real return) is approximately $\$8{,}000{,}000 - \$2{,}000{,}000 = \$6{,}000{,}000$. The amount earned from the investment is about 3× the amount saved, even with the modest real return of 6%. Sound investment planning is well worth the effort.

7.4.2 Chemical reactions

Suppose that two chemicals A and B react to form a product C, which we write as

$$A + B \xrightarrow{k} C,$$

where k is called the rate constant of the reaction. For simplicity, we will use the same symbol C, say, to refer to both the chemical C and its concentration. The law of mass action says that dC/dt is proportional to the product of the concentrations A and B, with proportionality constant k; that is,

$$\frac{dC}{dt} = kAB. \qquad (7.17)$$

Similarly, the law of mass action enables us to write equations for the time-derivatives of the reactant concentrations A and B:

$$\frac{dA}{dt} = -kAB, \quad \frac{dB}{dt} = -kAB. \tag{7.18}$$

The ode given by (7.17) can be solved analytically using conservation laws. We assume that A_0 and B_0 are the initial concentrations of the reactants, and that no product is initially present. From (7.17) and (7.18),

$$\begin{aligned} \frac{d}{dt}(A + C) = 0 \qquad &\Longrightarrow \quad A + C = A_0, \\ \frac{d}{dt}(B + C) = 0 \qquad &\Longrightarrow \quad B + C = B_0. \end{aligned}$$

Using these conservation laws, (7.17) becomes

$$\frac{dC}{dt} = k(A_0 - C)(B_0 - C), \quad C(0) = 0,$$

which is a nonlinear equation that may be integrated by separating variables. Separating and integrating, we obtain

$$\begin{aligned} \int_0^C \frac{dC}{(A_0 - C)(B_0 - C)} &= k \int_0^t dt \\ &= kt. \end{aligned} \tag{7.19}$$

The remaining integral can be done using the method of partial fractions. We write

$$\frac{1}{(A_0 - C)(B_0 - C)} = \frac{a}{A_0 - C} + \frac{b}{B_0 - C}. \tag{7.20}$$

The cover-up method is the simplest method to determine the unknown coefficients a and b. To determine a, we multiply both sides of (7.20) by $A_0 - C$ and set $C = A_0$ to find

$$a = \frac{1}{B_0 - A_0}.$$

Similarly, to determine b, we multiply both sides of (7.20) by $B_0 - C$ and set $C = B_0$ to find

$$b = \frac{1}{A_0 - B_0}.$$

Therefore,

$$\frac{1}{(A_0 - C)(B_0 - C)} = \frac{1}{B_0 - A_0}\left(\frac{1}{A_0 - C} - \frac{1}{B_0 - C}\right),$$

and the remaining integral of (7.19) becomes (using $C < A_0, B_0$)

$$\begin{aligned} \int_0^C \frac{dC}{(A_0 - C)(B_0 - C)} &= \frac{1}{B_0 - A_0}\left(\int_0^C \frac{dC}{A_0 - C} - \int_0^C \frac{dC}{B_0 - C}\right) \\ &= \frac{1}{B_0 - A_0}\left(-\ln\left(\frac{A_0 - C}{A_0}\right) + \ln\left(\frac{B_0 - C}{B_0}\right)\right) \\ &= \frac{1}{B_0 - A_0}\ln\left(\frac{A_0(B_0 - C)}{B_0(A_0 - C)}\right). \end{aligned}$$

Using this integral in (7.19), multiplying by $(B_0 - A_0)$ and exponentiating, we obtain

$$\frac{A_0(B_0 - C)}{B_0(A_0 - C)} = e^{(B_0 - A_0)kt}.$$

Solving for C, we finally obtain

$$C(t) = A_0 B_0 \frac{e^{(B_0 - A_0)kt} - 1}{B_0 e^{(B_0 - A_0)kt} - A_0},$$

which appears to be a complicated expression, but has the simple limits

$$\begin{aligned}\lim_{t\to\infty} C(t) &= \begin{cases} A_0, & \text{if } A_0 < B_0, \\ B_0, & \text{if } B_0 < A_0 \end{cases} \\ &= \min(A_0, B_0).\end{aligned}$$

As one would expect, the reaction stops after one of the reactants is depleted; and the final concentration of product is equal to the initial concentration of the depleted reactant.

7.4.3 Terminal velocity

View tutorial on YouTube

Using Newton's law, we model a mass m free falling under gravity but with air resistance. We assume that the force of air resistance is proportional to the speed of the mass and opposes the direction of motion. We define the x-axis to point in the upward direction, opposite the force of gravity. Near the surface of the Earth, the force of gravity is approximately constant and is given by $-mg$, with $g = 9.8\,\mathrm{m/s^2}$ the usual gravitational acceleration. The force of air resistance is modeled by $-kv$, where v is the vertical velocity of the mass and k is a positive constant. When the mass is falling, $v < 0$ and the force of air resistance is positive, pointing upward and opposing the motion. The total force on the mass is therefore given by $F = -mg - kv$. With $F = ma$ and $a = dv/dt$, we obtain the differential equation

$$m\frac{dv}{dt} = -mg - kv. \tag{7.21}$$

The terminal velocity v_∞ of the mass is defined as the asymptotic velocity after air resistance balances the gravitational force. When the mass is at terminal velocity, $dv/dt = 0$ so that

$$v_\infty = -\frac{mg}{k}. \tag{7.22}$$

The approach to the terminal velocity of a mass initially at rest is obtained by solving (7.21) with initial condition $v(0) = 0$. The equation is both linear and separable, and I solve by separating variables:

$$\begin{aligned} m\int_0^v \frac{dv}{mg + kv} &= -\int_0^t dt, \\ \frac{m}{k}\ln\left(\frac{mg + kv}{mg}\right) &= -t, \\ 1 + \frac{kv}{mg} &= e^{-kt/m}, \\ v &= -\frac{mg}{k}\left(1 - e^{-kt/m}\right). \end{aligned}$$

Therefore, $v = v_\infty \left(1 - e^{-kt/m}\right)$, and v approaches v_∞ as the exponential term decays to zero.

As an example, a skydiver of mass $m = 100\,\text{kg}$ with his parachute closed may have a terminal velocity of 200 km/hr. With

$$g = (9.8\,\text{m/s}^2)(10^{-3}\,\text{km/m})(60\,\text{s/min})^2(60\,\text{min/hr})^2 = 127,008\,\text{km/hr}^2,$$

one obtains from (7.22), $k = 63,504\,\text{kg/hr}$. One-half of the terminal velocity for free-fall (100 km/hr) is therefore attained when $(1 - e^{-kt/m}) = 1/2$, or $t = m \ln 2/k \approx 4\,\text{sec}$. Approximately 95% of the terminal velocity (190 km/hr) is attained after 17 sec.

7.4.4 Escape velocity

View tutorial on YouTube

An interesting physical problem is to find the smallest initial velocity for a mass on the Earth's surface to escape from the Earth's gravitational field, the so-called escape velocity. Newton's law of universal gravitation asserts that the gravitational force between two massive bodies is proportional to the product of the two masses and inversely proportional to the square of the distance between them. For a mass m a position x above the surface of the Earth, the force on the mass is given by

$$F = -G\frac{Mm}{(R+x)^2},$$

where M and R are the mass and radius of the Earth and G is the gravitational constant. The minus sign means the force on the mass m points in the direction of decreasing x. The approximately constant acceleration g on the Earth's surface corresponds to the absolute value of F/m when $x = 0$:

$$g = \frac{GM}{R^2},$$

and $g \approx 9.8\ \text{m/s}^2$. Newton's law $F = ma$ for the mass m is thus given by

$$\begin{aligned} \frac{d^2x}{dt^2} &= -\frac{GM}{(R+x)^2} \\ &= -\frac{g}{(1+x/R)^2}, \end{aligned} \tag{7.23}$$

where the radius of the Earth is known to be $R \approx 6350$ km.

A useful trick allows us to solve this second-order differential equation as a first-order equation. First, note that $d^2x/dt^2 = dv/dt$. If we write $v(t) = v(x(t))$—considering the velocity of the mass m to be a function of its distance above the Earth—we have using the chain rule

$$\begin{aligned} \frac{dv}{dt} &= \frac{dv}{dx}\frac{dx}{dt} \\ &= v\frac{dv}{dx}, \end{aligned}$$

where we have used $v = dx/dt$. Therefore, (7.23) becomes the first-order ode

$$v\frac{dv}{dx} = -\frac{g}{(1+x/R)^2},$$

which may be solved assuming an initial velocity $v(x=0)=v_0$ when the mass is shot vertically from the Earth's surface. Separating variables and integrating, we obtain

$$\int_{v_0}^{v} v dv = -g\int_0^x \frac{dx}{(1+x/R)^2}.$$

The left integral is $\frac{1}{2}(v^2-v_0^2)$, and the right integral can be performed using the substitution $u=1+x/R$, $du=dx/R$:

$$\begin{aligned}\int_0^x \frac{dx}{(1+x/R)^2} &= R\int_1^{1+x/R}\frac{du}{u^2}\\ &= -\frac{R}{u}\Big]_1^{1+x/R}\\ &= R-\frac{R^2}{x+R}\\ &= \frac{Rx}{x+R}.\end{aligned}$$

Therefore,

$$\frac{1}{2}(v^2-v_0^2) = -\frac{gRx}{x+R},$$

which when multiplied by m is an expression of the conservation of energy (the change of the kinetic energy of the mass is equal to the change in the potential energy). Solving for v^2,

$$v^2 = v_0^2 - \frac{2gRx}{x+R}.$$

The escape velocity is defined as the minimum initial velocity v_0 such that the mass can *escape* to infinity. Therefore, $v_0 = v_{\text{escape}}$ when $v\to 0$ as $x\to\infty$. Taking this limit, we have

$$\begin{aligned}v_{\text{escape}}^2 &= \lim_{x\to\infty}\frac{2gRx}{x+R}\\ &= 2gR.\end{aligned}$$

With $R\approx 6350$ km and $g=127\,008\,\text{km/hr}^2$, we determine $v_{\text{escape}}=\sqrt{2gR}\approx 40\,000$ km/hr. In comparison, the muzzle velocity of a modern high-performance rifle is 4300 km/hr, almost an order of magnitude too slow for a bullet, shot into the sky, to escape the Earth's gravity.

7.4.5 RC circuit

View tutorial on YouTube

Consider a resister R and a capacitor C connected in series as shown in Fig. 7.4. A battery providing an electromotive force, or emf $\mathcal{E}$, connects to this circuit by a switch. Initially, there is no charge on the capacitor. When the switch is thrown to a, the battery connects and the capacitor charges. When the switch is thrown to b, the battery disconnects and the capacitor discharges, with energy dissipated in the resister. Here, we determine the voltage drop across the capacitor during charging and discharging.

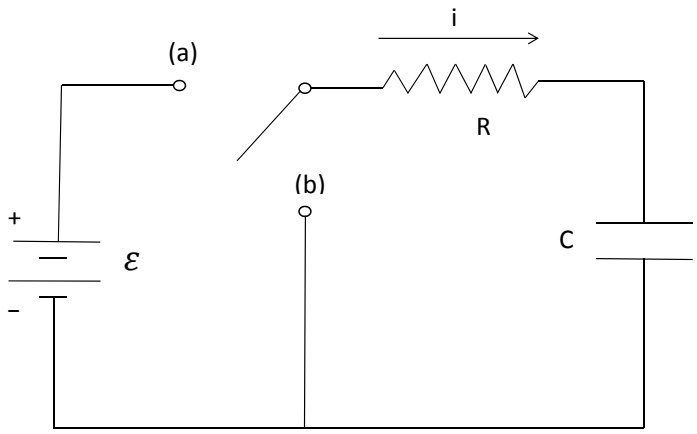


Figure 7.4: RC circuit diagram.

The equations for the voltage drops across a capacitor and a resister are given by

$$V_C = q/C, \quad V_R = iR, \tag{7.24}$$

where C is the capacitance and R is the resistance. The charge q and the current i are related by

$$i = \frac{dq}{dt}. \tag{7.25}$$

Kirchhoff's voltage law states that the emf $\mathcal{E}$ in any closed loop is equal to the sum of the voltage drops in that loop. Applying Kirchhoff's voltage law when the switch is thrown to a results in

$$V_R + V_C = \mathcal{E}. \tag{7.26}$$

Using (7.24) and (7.25), the voltage drop across the resister can be written in terms of the voltage drop across the capacitor as

$$V_R = RC\frac{dV_C}{dt},$$

and (7.26) can be rewritten to yield the linear first-order differential equation for V_C given by

$$\frac{dV_C}{dt} + V_C/RC = \mathcal{E}/RC, \tag{7.27}$$

with initial condition $V_C(0) = 0$.

The integrating factor for this equation is

$$\mu(t) = e^{t/RC},$$

and (7.27) integrates to

$$V_C(t) = e^{-t/RC}\int_0^t (\mathcal{E}/RC)e^{t/RC}dt,$$

with solution

$$V_C(t) = \mathcal{E}\left(1 - e^{-t/RC}\right).$$

The voltage starts at zero and rises exponentially to $\mathcal{E}$, with characteristic time scale given by RC.

When the switch is thrown to b, application of Kirchhoff's voltage law results in

$$V_R + V_C = 0,$$

with corresponding differential equation

$$\frac{dV_C}{dt} + V_C/RC = 0.$$

Here, we assume that the capacitance is initially fully charged so that $V_C(0) = \mathcal{E}$. The solution, then, during the discharge phase is given by

$$V_C(t) = \mathcal{E}e^{-t/RC}.$$

The voltage starts at $\mathcal{E}$ and decays exponentially to zero, again with characteristic time scale given by RC.

7.4.6 The logistic equation

View tutorial on YouTube

Let $N(t)$ be the number of individuals in a population at time t, and let b and d be the average per capita birth rate and death rate, respectively. In a short time Δt, the number of births in the population is $b\Delta tN$, and the number of deaths is $d\Delta tN$. An equation for N at time $t + \Delta t$ is then determined to be

$$N(t + \Delta t) = N(t) + b\Delta tN(t) - d\Delta tN(t),$$

which can be rearranged to

$$\frac{N(t + \Delta t) - N(t)}{\Delta t} = (b - d)N(t);$$

and as $\Delta t \to 0$, and with $r = b - d$, we have

$$\frac{dN}{dt} = rN.$$

This is the Malthusian growth model (Thomas Malthus, 1766-1834), and is the same equation as our compound interest model.

Under a Malthusian growth model, the population size grows exponentially like

$$N(t) = N_0e^{rt},$$

where N_0 is the initial population size. However, when the population growth is constrained by limited resources, a heuristic modification to the Malthusian growth model results in the Verhulst equation,

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right), \tag{7.28}$$

where K is called the carrying capacity of the environment. Making (7.28) dimensionless using $\tau = rt$ and $x = N/K$ leads to the logistic equation,

$$\frac{dx}{d\tau} = x(1 - x),$$

where we may assume the initial condition $x(0) = x_0 > 0$. Separating variables and integrating

$$\int_{x_0}^{x} \frac{dx}{x(1-x)} = \int_0^{\tau} d\tau.$$

The integral on the left-hand-side can be done using the method of partial fractions:

$$\frac{1}{x(1-x)} = \frac{a}{x} + \frac{b}{1-x},$$

and the cover-up method yields $a = b = 1$. Therefore,

$$\begin{aligned}\int_{x_0}^{x} \frac{dx}{x(1-x)} &= \int_{x_0}^{x} \frac{dx}{x} + \int_{x_0}^{x} \frac{dx}{(1-x)} \\ &= \ln \frac{x}{x_0} - \ln \frac{1-x}{1-x_0} \\ &= \ln \frac{x(1-x_0)}{x_0(1-x)} \\ &= \tau.\end{aligned}$$

Solving for x, we first exponentiate both sides and then isolate x:

$$\begin{aligned}\frac{x(1-x_0)}{x_0(1-x)} &= e^{\tau}, \\ x(1-x_0) &= x_0 e^{\tau} - x x_0 e^{\tau}, \\ x(1-x_0+x_0e^{\tau}) &= x_0 e^{\tau}, \\ x &= \frac{x_0}{x_0 + (1-x_0)e^{-\tau}}.\end{aligned} \tag{7.29}$$

We observe that for $x_0 > 0$, we have $\lim_{\tau\to\infty} x(\tau) = 1$, corresponding to

$$\lim_{t\to\infty} N(t) = K.$$

The population, therefore, grows in size until it reaches the carrying capacity of its environment.

Chapter 8

Linear second-order differential equations with constant coefficients

Reference: Boyce and DiPrima, Chapter 3

The general linear second-order differential equation with independent variable t and dependent variable $x = x(t)$ is given by

$$\ddot{x} + p(t)\dot{x} + q(t)x = g(t), \tag{8.1}$$

where we have used the standard physics notation $\dot{x} = dx/dt$ and $\ddot{x} = d^2x/dt^2$. Herein, we assume that $p(t)$ and $q(t)$ are continuous functions on the time interval for which we solve (8.1). A unique solution of (8.1) requires initial values $x(t_0) = x_0$ and $\dot{x}(t_0) = u_0$. The equation with constant coefficients—on which we will devote considerable effort—assumes that $p(t)$ and $q(t)$ are constants, independent of time. The linear second-order ode is said to be *homogeneous* if $g(t) = 0$.

8.1 The Euler method

View tutorial on YouTube

In general, (8.1) cannot be solved analytically, and we begin by deriving an algorithm for numerical solution. Consider the general second-order ode given by

$$\ddot{x} = f(t, x, \dot{x}).$$

We can write this second-order ode as a pair of first-order odes by defining $u = \dot{x}$, and writing the first-order system as

$$\dot{x} = u, \tag{8.2}$$
$$\dot{u} = f(t, x, u). \tag{8.3}$$

The first ode, (8.2), gives the slope of the tangent line to the curve $x = x(t)$; the second ode, (8.3), gives the slope of the tangent line to the curve $u = u(t)$. Beginning at the initial values $(x, u) = (x_0, u_0)$ at the time $t = t_0$, we move along the tangent lines to determine $x_1 = x(t_0 + \Delta t)$ and $u_1 = u(t_0 + \Delta t)$:

$$x_1 = x_0 + \Delta t u_0,$$
$$u_1 = u_0 + \Delta t f(t_0, x_0, u_0).$$

The values x_1 and u_1 at the time $t_1 = t_0 + \Delta t$ are then used as new initial values to march the solution forward to time $t_2 = t_1 + \Delta t$. As long as $f(t, x, u)$ is a well-behaved function, the numerical solution converges to the unique solution of the ode as $\Delta t \to 0$.

8.2 The principle of superposition

View tutorial on YouTube

Consider the homogeneous linear second-order ode:

$$\ddot{x} + p(t)\dot{x} + q(t)x = 0; \tag{8.4}$$

and suppose that $x = X_1(t)$ and $x = X_2(t)$ are solutions to (8.4). We consider a linear combination of X_1 and X_2 by letting

$$X(t) = c_1 X_1(t) + c_2 X_2(t), \tag{8.5}$$

with c_1 and c_2 constants. The *principle of superposition* states that $x = X(t)$ is also a solution of (8.4). To prove this, we compute

$$\begin{aligned}
\ddot{X} + p\dot{X} + qX &= c_1\ddot{X}_1 + c_2\ddot{X}_2 + p\left(c_1\dot{X}_1 + c_2\dot{X}_2\right) + q\left(c_1X_1 + c_2X_2\right) \\
&= c_1\left(\ddot{X}_1 + p\dot{X}_1 + qX_1\right) + c_2\left(\ddot{X}_2 + p\dot{X}_2 + qX_2\right) \\
&= c_1 \times 0 + c_2 \times 0 \\
&= 0,
\end{aligned}$$

since X_1 and X_2 were assumed to be solutions of (8.4). We have therefore shown that any linear combination of solutions to the homogeneous linear ode is also a solution.

8.3 The Wronskian

View tutorial on YouTube

Suppose that having determined that two solutions of (8.4) are $x = X_1(t)$ and $x = X_2(t)$, we attempt to write the general solution to (8.4) as (8.5). We must then ask whether this general solution will be able to satisfy two initial conditions given by

$$x(t_0) = x_0, \quad \dot{x}(t_0) = u_0, \tag{8.6}$$

for any initial time t_0, and initial values x_0 and u_0. Applying these initial conditions to (8.5), we obtain

$$\begin{aligned}
c_1X_1(t_0) + c_2X_2(t_0) &= x_0, \\
c_1\dot{X}_1(t_0) + c_2\dot{X}_2(t_0) &= u_0,
\end{aligned} \tag{8.7}$$

which is a system of two linear equations for the two unknowns c_1 and c_2. In matrix form,

$$\begin{pmatrix} X_1(t_0) & X_2(t_0) \\ \dot{X}_1(t_0) & \dot{X}_2(t_0) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} x_0 \\ u_0 \end{pmatrix}. \tag{8.8}$$

We can solve (8.8) for any specified values of t_0, x_0 and u_0 if the 2-by-2 matrix is invertible, and that will be the case if its determinant is nonzero. The determinant is called the Wronskian and is defined by

$$W = X_1\dot{X}_2 - \dot{X}_1X_2. \tag{8.9}$$

In the language of linear algebra, when $W \neq 0$ the functions $X_1 = X_1(t)$ and $X_2 = X_2(t)$ are linearly independent and span the solution space of the second-order linear differential equation given by (8.4). The solution space of the ode satisfies the conditions of a vector space and the two solutions X_1 and X_2 act as basis vectors for this space. The dimension of this space is two, corresponding to the order of the differential equation.

Example: Show that the functions $X_1(t) = \cos\omega t$ and $X_2(t) = \sin\omega t$ have a nonzero Wronskian when $\omega \neq 0$.

We have

$$\begin{aligned} X_1(t) &= \cos\omega t, & X_2(t) &= \sin\omega t \\ \dot{X}_1(t) &= -\omega\sin\omega t, & \dot{X}_2(t) &= \omega\cos\omega t. \end{aligned}$$

The Wronskian is given by

$$W = \begin{vmatrix} \cos\omega t & \sin\omega t \\ -\omega\sin\omega t & \omega\cos\omega t \end{vmatrix} = \omega(\cos^2\omega t + \sin^2\omega t) = \omega,$$

so that $W \neq 0$ when $\omega \neq 0$.

Example: Show that the functions $X_1(t) = e^{at}$ and $X_2(t) = e^{bt}$ have a nonzero Wronskian when $a \neq b$.

We have

$$\begin{aligned} X_1(t) &= e^{at}, & X_2(t) &= e^{bt} \\ \dot{X}_1(t) &= ae^{at}, & \dot{X}_2(t) &= be^{bt}. \end{aligned}$$

The Wronskian is given by

$$W = \begin{vmatrix} e^{at} & e^{bt} \\ ae^{at} & be^{bt} \end{vmatrix} = (b-a)e^{(a+b)t},$$

so that $W \neq 0$ when $a \neq b$.

Example: Show that the functions $X_1(t) = 1$ and $X_2(t) = t$ have a nonzero Wronskian.

We have

$$\begin{aligned} X_1(t) &= 1, & X_2(t) &= t \\ \dot{X}_1(t) &= 0, & \dot{X}_2(t) &= 1. \end{aligned}$$

The Wronskian is given by

$$W = \begin{vmatrix} 1 & t \\ 0 & 1 \end{vmatrix} = 1.$$

8.4 Homogeneous linear second-order ode with constant coefficients

View tutorial on YouTube

We now study solutions of the homogeneous, constant coefficient ode, written as

$$a\ddot{x} + b\dot{x} + cx = 0, \tag{8.10}$$

with a, b, and c constants. Such an equation arises for the charge on a capacitor in an unpowered RLC electrical circuit, or for the position of a freely-oscillating frictional mass on a spring, or for a damped pendulum. Our solution method finds two linearly independent solutions to (8.10), multiplies each of these solutions by a constant, and adds them. The two free constants can then be used to satisfy two given initial conditions.

Because of the differential properties of the exponential function, a natural ansatz, or educated guess, for the form of the solution to (8.10) is $x = e^{rt}$, where r is a constant to be determined. Successive differentiation results in $\dot{x} = re^{rt}$ and $\ddot{x} = r^2 e^{rt}$, and substitution into (8.10) yields

$$ar^2 e^{rt} + bre^{rt} + ce^{rt} = 0. \tag{8.11}$$

Our choice of exponential function is now rewarded by the explicit cancelation in (8.11) of e^{rt}. The result is a quadratic equation for the unknown constant r:

$$ar^2 + br + c = 0. \tag{8.12}$$

Our ansatz has thus converted a differential equation into an algebraic equation. Equation (8.12) is called the *characteristic equation* of (8.10). (Recall that $\det(\mathrm{A} - \lambda \mathrm{I}) = 0$ was also called the characteristic equation of the matrix A. We will see later that this is not a coincidence.)

Using the quadratic formula, the two solutions of the characteristic equation (8.12) are given by

$$r_\pm = \frac{1}{2a}\left(-b \pm \sqrt{b^2 - 4ac}\right).$$

There are three cases to consider: (1) if $b^2 - 4ac > 0$, then the two roots are distinct and real; (2) if $b^2 - 4ac < 0$, then the two roots are complex conjugates (3) if $b^2 - 4ac = 0$, then the two roots are degenerate and there is only one real root. We will consider these three cases in turn.

8.4.1 Distinct real roots

When $r_+ \neq r_-$ are real roots, then the general solution to (8.10) can be written as a linear superposition of the two solutions $e^{r_+ t}$ and $e^{r_- t}$; that is,

$$x(t) = c_1 e^{r_+ t} + c_2 e^{r_- t}.$$

The unknown constants c_1 and c_2 can then be determined by the given initial conditions $x(t_0) = x_0$ and $\dot{x}(t_0) = u_0$. We now present two examples.

Example 1: Solve $\ddot{x} + 5\dot{x} + 6x = 0$ with $x(0) = 2$, $\dot{x}(0) = 3$, and find the maximum value attained by x.

View tutorial on YouTube

We take as our ansatz $x = e^{rt}$ and obtain the characteristic equation

$$r^2 + 5r + 6 = 0,$$

which factors to

$$(r + 3)(r + 2) = 0.$$

The general solution to the ode is thus

$$x(t) = c_1 e^{-2t} + c_2 e^{-3t}.$$

The solution for $\dot{x}$ obtained by differentiation is

$$\dot{x}(t) = -2c_1 e^{-2t} - 3c_2 e^{-3t}.$$

Use of the initial conditions then results in two equations for the two unknown constant c_1 and c_2:

$$\begin{aligned} c_1 + c_2 &= 2, \\ -2c_1 - 3c_2 &= 3. \end{aligned}$$

Adding three times the first equation to the second equation yields $c_1 = 9$; and the first equation then yields $c_2 = 2 - c_1 = -7$. Therefore, the unique solution that satisfies both the ode and the initial conditions is

$$\begin{aligned} x(t) &= 9e^{-2t} - 7e^{-3t} \\ &= 9e^{-2t}\left(1 - \frac{7}{9}e^{-t}\right). \end{aligned}$$

Note that although both exponential terms decay in time, their sum increases initially since $\dot{x}(0) > 0$. The maximum value of x occurs at the time t_m when $\dot{x} = 0$, or

$$t_m = \ln(7/6).$$

The maximum $x_m = x(t_m)$ is then determined to be

$$x_m = 108/49.$$

Example 2: Solve $\ddot{x} - x = 0$ with $x(0) = x_0$, $\dot{x}(0) = u_0$.

Again our ansatz is $x = e^{rt}$, and we obtain the characteristic equation

$$r^2 - 1 = 0,$$

with solution $r_\pm = \pm 1$. Therefore, the general solution for x is

$$x(t) = c_1 e^t + c_2 e^{-t},$$

and the derivative satisfies

$$\dot{x}(t) = c_1 e^t - c_2 e^{-t}.$$

Initial conditions are satisfied when

$$\begin{aligned} c_1 + c_2 &= x_0, \\ c_1 - c_2 &= u_0. \end{aligned}$$

Adding and subtracting these equations, we determine

$$c_1 = \frac{1}{2}(x_0 + u_0), \quad c_2 = \frac{1}{2}(x_0 - u_0),$$

so that after rearranging terms

$$x(t) = x_0\left(\frac{e^t + e^{-t}}{2}\right) + u_0\left(\frac{e^t - e^{-t}}{2}\right).$$

The terms in parentheses are the usual definitions of the hyperbolic cosine and sine functions; that is,

$$\cosh t = \frac{e^t + e^{-t}}{2}, \quad \sinh t = \frac{e^t - e^{-t}}{2}.$$

Our solution can therefore be rewritten as

$$x(t) = x_0 \cosh t + u_0 \sinh t.$$

Note that the relationships between the trigonometric functions and the complex exponentials were given by

$$\cos t = \frac{e^{it} + e^{-it}}{2}, \quad \sin t = \frac{e^{it} - e^{-it}}{2i},$$

so that

$$\cosh t = \cos it, \quad \sinh t = -i \sin it,$$

and

$$\cosh^2 t - \sinh^2 t = 1.$$

Also note that the hyperbolic trigonometric functions satisfy the differential equations

$$\frac{d}{dt} \sinh t = \cosh t, \quad \frac{d}{dt} \cosh t = \sinh t,$$

which though similar to the differential equations satisfied by the more commonly used trigonometric functions, is absent a minus sign.

8.4.2 Distinct complex-conjugate roots

View tutorial on YouTube

We now consider a characteristic equation (8.12) with $b^2 - 4ac < 0$, so the roots occur as complex conjugate pairs. With

$$\lambda = -\frac{b}{2a}, \quad \mu = \frac{1}{2a}\sqrt{4ac - b^2},$$

the two roots of the characteristic equation are $\lambda + i\mu$ and $\lambda - i\mu$. We have thus found the following two complex exponential solutions to the differential equation:

$$Z_1(t) = e^{\lambda t} e^{i\mu t}, \quad Z_2(t) = e^{\lambda t} e^{-i\mu t}.$$

Applying the principle of superposition, any linear combination of Z_1 and Z_2 is also a solution to the second-order ode.

Recall that if $z = x + iy$, then $\operatorname{Re} z = (z + \bar{z})/2$ and $\operatorname{Im} z = (z - \bar{z})/2i$. We can therefore form two different linear combinations of $Z_1(t)$ and $Z_2(t)$ that are real, namely $X_1(t) = \operatorname{Re} Z_1(t)$ and $X_2(t) = \operatorname{Im} Z_1(t)$. We have

$$X_1(t) = e^{\lambda t} \cos \mu t, \quad X_2(t) = e^{\lambda t} \sin \mu t.$$

Having found these two real solutions, $X_1(t)$ and $X_2(t)$, we can then apply the principle of superposition a second time to determine the general solution for $x(t)$:

$$x(t) = e^{\lambda t} \left(A \cos \mu t + B \sin \mu t \right). \tag{8.13}$$

It is best to memorize this result. The real part of the roots of the characteristic equation goes into the exponential function; the imaginary part goes into the argument of cosine and sine.

Example 1: Solve $\ddot{x} + x = 0$ with $x(0) = x_0$ and $\dot{x}(0) = u_0$.

View tutorial on YouTube

The characteristic equation is

$$r^2 + 1 = 0,$$

with roots

$$r_\pm = \pm i.$$

The general solution of the ode is therefore

$$x(t) = A\cos t + B\sin t.$$

The derivative is

$$\dot{x}(t) = -A\sin t + B\cos t.$$

Applying the initial conditions:

$$x(0) = A = x_0, \quad \dot{x}(0) = B = u_0;$$

so that the final solution is

$$x(t) = x_0\cos t + u_0\sin t.$$

Recall that we wrote the analogous solution to the ode $\ddot{x} - x = 0$ as $x(t) = x_0\cosh t + u_0\sinh t$.

Example 2: Solve $\ddot{x} + \dot{x} + x = 0$ with $x(0) = 1$ and $\dot{x}(0) = 0$.

The characteristic equation is

$$r^2 + r + 1 = 0,$$

with roots

$$r_\pm = -\frac{1}{2} \pm i\frac{\sqrt{3}}{2}.$$

The general solution of the ode is therefore

$$x(t) = e^{-\frac{1}{2}t}\left(A\cos\frac{\sqrt{3}}{2}t + B\sin\frac{\sqrt{3}}{2}t\right).$$

The derivative is

$$\dot{x}(t) = -\frac{1}{2}e^{-\frac{1}{2}t}\left(A\cos\frac{\sqrt{3}}{2}t + B\sin\frac{\sqrt{3}}{2}t\right) + \frac{\sqrt{3}}{2}e^{-\frac{1}{2}t}\left(-A\sin\frac{\sqrt{3}}{2}t + B\cos\frac{\sqrt{3}}{2}t\right).$$

Applying the initial conditions $x(0) = 1$ and $\dot{x}(0) = 0$:

$$A = 1,$$
$$-\frac{1}{2}A + \frac{\sqrt{3}}{2}B = 0;$$

or

$$A = 1, \quad B = \frac{\sqrt{3}}{3}.$$

Therefore,

$$x(t) = e^{-\frac{1}{2}t}\left(\cos\frac{\sqrt{3}}{2}t + \frac{\sqrt{3}}{3}\sin\frac{\sqrt{3}}{2}t\right).$$

8.4.3 Degenerate roots

View tutorial on YouTube

Finally, we consider the characteristic equation,

$$ar^2 + br + c = 0,$$

with $b^2 - 4ac = 0$. The degenerate root is then given by

$$r = -\frac{b}{2a},$$

yielding only a single solution to the ode:

$$x_1(t) = \exp\left(-\frac{bt}{2a}\right). \tag{8.14}$$

To satisfy two initial conditions, a second independent solution must be found with nonzero Wronskian, and apparently this second solution is not of the form of our ansatz $x = \exp(rt)$.

One method to determine this missing second solution is to try the ansatz

$$x(t) = y(t)x_1(t), \tag{8.15}$$

where $y(t)$ is an unknown function that satisfies the differential equation obtained by substituting (8.15) into (8.10). This standard technique is called the reduction of order method and enables one to find a second solution of a homogeneous linear differential equation if one solution is known. Upon substitution of (8.15), one obtains a differential equation for y that contains only y' and y''. By letting $w = y'$, one then obtains a linear first-order equation for w that we already know how to solve.

Here, however, I choose to determine this missing second solution through a limiting process. We will start with the solution obtained for complex roots of the characteristic equation, and then arrive at the solution obtained for degenerate roots by taking the limit $\mu \to 0$.

Now, the general solution for complex roots was given by (8.13), and to properly limit this solution as $\mu \to 0$ requires first satisfying the specific initial conditions $x(0) = x_0$ and $\dot{x}(0) = u_0$. Solving for A and B, the general solution given by (8.13) becomes the specific solution

$$x(t;\mu) = e^{\lambda t}\left(x_0\cos\mu t + \frac{u_0 - \lambda x_0}{\mu}\sin\mu t\right).$$

Here, we have written $x = x(t;\mu)$ to show explicitly that x depends on μ.

Taking the limit as $\mu \to 0$, and using $\lim_{\mu\to 0} \mu^{-1} \sin \mu t = t$, we have

$$\lim_{\mu\to 0} x(t;\mu) = e^{\lambda t}\big(x_0 + (u_0 - \lambda x_0)t\big).$$

The second solution is observed to be a constant, $u_0 - \lambda x_0$, times t times the first solution, $e^{\lambda t}$. Our general solution to the ode (8.10) when $b^2 - 4ac = 0$ can therefore be written in the form

$$x(t) = (c_1 + c_2 t)e^{rt},$$

where r is the repeated root of the characteristic equation. The main result to be remembered is that for the case of repeated roots, the second solution is t times the first solution.

Example: Solve $\ddot{x} + 2\dot{x} + x = 0$ with $x(0) = 1$ and $\dot{x}(0) = 0$.

The characteristic equation is

$$\begin{aligned} r^2 + 2r + 1 &= (r+1)^2 \\ &= 0, \end{aligned}$$

which has a repeated root given by $r = -1$. Therefore, the general solution to the ode is

$$x(t) = c_1 e^{-t} + c_2 t e^{-t},$$

with derivative

$$\dot{x}(t) = -c_1 e^{-t} + c_2 e^{-t} - c_2 t e^{-t}.$$

Applying the initial conditions, we have

$$\begin{aligned} c_1 &= 1, \\ -c_1 + c_2 &= 0, \end{aligned}$$

so that $c_1 = c_2 = 1$. Therefore, the solution is

$$x(t) = (1 + t)e^{-t}.$$

8.5 Homogeneous linear second-order difference equations with constant coefficients

The solution of linear difference equations is similar to that of linear differential equations. Here, we solve one interesting example.

Example: Find an explicit formula for the nth Fibonacci number F_n, where

$$F_{n+1} = F_n + F_{n-1}, \quad F_1 = F_2 = 1.$$

This will be our second derivation of Binet's formula (see §5.2 for the first derivation). We consider the relevant difference equation

$$x_{n+1} - x_n - x_{n-1} = 0, \tag{8.16}$$

and try to solve it using a method similar to the solution of a second-order differential equation. An appropriate ansatz here is

$$x_n = \lambda^n, \tag{8.17}$$

where λ is an unknown constant. Substitution of (8.17) into (8.16) results in

$$\lambda^{n+1} - \lambda^n - \lambda^{n-1} = 0,$$

or upon division by λ^{n-1},

$$\lambda^2 - \lambda - 1 = 0.$$

Use of the quadratic formula yields two roots. We have

$$\lambda_1 = \frac{1+\sqrt{5}}{2} = \Phi, \quad \lambda_2 = \frac{1-\sqrt{5}}{2} = -\phi,$$

where Φ is the golden ratio and ϕ is the golden ratio conjugate.

We have thus found two independent solutions to (8.16) of the form (8.17), and we can now use these two solutions to determine a formula for F_n. Multiplying the solutions by constants and adding them, we obtain

$$F_n = c_1\Phi^n + c_2(-\phi)^n, \tag{8.18}$$

which must satisfy the initial values $F_1 = F_2 = 1$. The algebra for finding the unknown constants can be made simpler, however, if instead of F_2, we use the value $F_0 = F_2 - F_1 = 0$.

Application of the values for F_0 and F_1 results in the system of equations given by

$$\begin{aligned} c_1 + c_2 &= 0, \\ c_1\Phi - c_2\phi &= 1. \end{aligned}$$

We use the first equation to write $c_2 = -c_1$, and substitute into the second equation to get

$$c_1(\Phi + \phi) = 1.$$

Since $\Phi + \phi = \sqrt{5}$, we can solve for c_1 and c_2 to obtain

$$c_1 = 1/\sqrt{5}, \quad c_2 = -1/\sqrt{5}. \tag{8.19}$$

Using (8.19) in (8.18) then derives Binet's formula

$$F_n = \frac{\Phi^n - (-\phi)^n}{\sqrt{5}}. \tag{8.20}$$

8.6 Inhomogeneous linear second-order ode

We now consider the general inhomogeneous linear second-order ode (8.1):

$$\ddot{x} + p(t)\dot{x} + q(t)x = g(t), \tag{8.21}$$

with initial conditions $x(t_0) = x_0$ and $\dot{x}(t_0) = u_0$. There is a three-step solution method when the inhomogeneous term $g(t) \neq 0$. (i) Find the general solution of the homogeneous equation

$$\ddot{x} + p(t)\dot{x} + q(t)x = 0. \tag{8.22}$$

Let us denote the homogeneous solution by

$$x_h(t) = c_1X_1(t) + c_2X_2(t),$$

where X_1 and X_2 are linearly independent solutions of (8.22), and c_1 and c_2 are as yet undetermined constants. (ii) Find a *particular* solution x_p of the inhomogeneous equation (8.21). A particular solution is readily found when $p(t)$ and $q(t)$ are constants, and when $g(t)$ is a combination of polynomials, exponentials, sines and cosines. (iii) Write the general solution of (8.21) as the sum of the homogeneous and particular solutions,

$$x(t) = x_h(t) + x_p(t), \tag{8.23}$$

and apply the initial conditions to determine the constants c_1 and c_2. Note that because of the linearity of (8.21),

$$\begin{aligned}\ddot{x} + p\dot{x} + qx &= \frac{d^2}{dt^2}(x_h + x_p) + p\frac{d}{dt}(x_h + x_p) + q(x_h + x_p) \\ &= (\ddot{x}_h + p\dot{x}_h + qx_h) + (\ddot{x}_p + p\dot{x}_p + qx_p) \\ &= 0 + g \\ &= g,\end{aligned}$$

so that (8.23) solves (8.21), and the two free constants in x_h can be used to satisfy the initial conditions.

We will consider here only the constant coefficient case. We now illustrate the solution method by an example.

Example: Solve $\ddot{x} - 3\dot{x} - 4x = 3e^{2t}$ with $x(0) = 0$ and $\dot{x}(0) = 0$.

View tutorial on YouTube

First, we solve the homogeneous equation. The characteristic equation is

$$\begin{aligned}r^2 - 3r - 4 &= (r-4)(r+1) \\ &= 0,\end{aligned}$$

so that

$$x_h(t) = c_1 e^{4t} + c_2 e^{-t}.$$

Second, we find a particular solution of the inhomogeneous equation. The form of the particular solution is chosen such that the exponential will cancel out of both sides of the ode. The ansatz we choose is

$$x(t) = Ae^{2t}, \tag{8.24}$$

where A is a yet undetermined coefficient. Upon substituting x into the ode, differentiating using the chain rule, and canceling the exponential, we obtain

$$4A - 6A - 4A = 3,$$

from which we determine $A = -1/2$. Obtaining a solution for A independent of t justifies the ansatz (8.24). Third, we write the general solution to the ode as the sum of the homogeneous and particular solutions, and determine c_1 and c_2 that satisfy the initial conditions. We have

$$x(t) = c_1 e^{4t} + c_2 e^{-t} - \frac{1}{2}e^{2t};$$

and taking the derivative,

$$\dot{x}(t) = 4c_1 e^{4t} - c_2 e^{-t} - e^{2t}.$$

Applying the initial conditions,

$$c_1 + c_2 - \frac{1}{2} = 0,$$
$$4c_1 - c_2 - 1 = 0;$$

or

$$c_1 + c_2 = \frac{1}{2},$$
$$4c_1 - c_2 = 1.$$

This system of linear equations can be solved for c_1 by adding the equations to obtain $c_1 = 3/10$, after which $c_2 = 1/5$ can be determined from the first equation. Therefore, the solution for $x(t)$ that satisfies both the ode and the initial conditions is given by

$$\begin{aligned} x(t) &= \frac{3}{10}e^{4t} - \frac{1}{2}e^{2t} + \frac{1}{5}e^{-t} \\ &= \frac{3}{10}e^{4t}\left(1 - \frac{5}{3}e^{-2t} + \frac{2}{3}e^{-5t}\right), \end{aligned}$$

where we have grouped the terms in the solution to better display the asymptotic behavior for large t.

We now find particular solutions for some relatively simple inhomogeneous terms using this method of undetermined coefficients.

Example: Find a particular solution of $\ddot{x} - 3\dot{x} - 4x = 2\sin t$.

View tutorial on YouTube

We show two methods for finding a particular solution. The first more direct method tries the ansatz

$$x(t) = A\cos t + B\sin t,$$

where the argument of cosine and sine must agree with the argument of sine in the inhomogeneous term. The cosine term is required because the derivative of sine is cosine. Upon substitution into the differential equation, we obtain

$$(-A\cos t - B\sin t) - 3(-A\sin t + B\cos t) - 4(A\cos t + B\sin t) = 2\sin t,$$

or regrouping terms,

$$-(5A + 3B)\cos t + (3A - 5B)\sin t = 2\sin t.$$

This equation is valid for all t, and in particular for $t = 0$ and $\pi/2$, for which the sine and cosine functions vanish, respectively. For these two values of t, we find

$$5A + 3B = 0, \quad 3A - 5B = 2;$$

and solving for A and B, we obtain

$$A = \frac{3}{17}, \quad B = -\frac{5}{17}.$$

The particular solution is therefore given by

$$x_p = \frac{1}{17}\left(3\cos t - 5\sin t\right).$$

The second solution method makes use of the relation $e^{it} = \cos t + i\sin t$ to convert the sine inhomogeneous term to an exponential function. We introduce the complex function $z(t)$ by letting

$$z(t) = x(t) + iy(t),$$

and rewrite the differential equation in complex form. We can rewrite the equation in one of two ways. On the one hand, if we use $\sin t = \mathrm{Re}\{-ie^{it}\}$, then the differential equation is written as

$$\ddot{z} - 3\dot{z} - 4z = -2ie^{it}; \tag{8.25}$$

and by equating the real and imaginary parts, this equation becomes the two real differential equations

$$\ddot{x} - 3\dot{x} - 4x = 2\sin t, \quad \ddot{y} - 3\dot{y} - 4y = -2\cos t.$$

The solution we are looking for, then, is $x_p(t) = \mathrm{Re}\{z_p(t)\}$.

On the other hand, if we write $\sin t = \mathrm{Im}\{e^{it}\}$, then the complex differential equation becomes

$$\ddot{z} - 3\dot{z} - 4z = 2e^{it}, \tag{8.26}$$

which becomes the two real differential equations

$$\ddot{x} - 3\dot{x} - 4x = 2\cos t, \quad \ddot{y} - 3\dot{y} - 4y = 2\sin t.$$

Here, the solution we are looking for is $x_p(t) = \mathrm{Im}\{z_p(t)\}$.

We will proceed here by solving (8.26). As we now have an exponential function as the inhomogeneous term, we can make the ansatz

$$z(t) = Ce^{it},$$

where we now expect C to be a complex constant. Upon substitution into the ode (8.26) and using $i^2 = -1$:

$$-C - 3iC - 4C = 2;$$

or solving for C:

$$\begin{aligned} C &= \frac{-2}{5+3i} \\ &= \frac{-2(5-3i)}{(5+3i)(5-3i)} \\ &= \frac{-10+6i}{34} \\ &= \frac{-5+3i}{17}. \end{aligned}$$

Therefore,

$$\begin{aligned} x_p &= \mathrm{Im}\{z_p\} \\ &= \mathrm{Im}\left\{\frac{1}{17}(-5+3i)(\cos t + i\sin t)\right\} \\ &= \frac{1}{17}(3\cos t - 5\sin t). \end{aligned}$$

Example: Find a particular solution of $\ddot{x} + \dot{x} - 2x = t^2$.

View tutorial on YouTube

The correct ansatz here is the polynomial

$$x(t) = At^2 + Bt + C.$$

Upon substitution into the ode, we have

$$2A + 2At + B - 2At^2 - 2Bt - 2C = t^2,$$

or

$$-2At^2 + 2(A - B)t + (2A + B - 2C)t^0 = t^2.$$

Equating powers of t,

$$-2A = 1, \quad 2(A - B) = 0, \quad 2A + B - 2C = 0;$$

and solving,

$$A = -\frac{1}{2}, \quad B = -\frac{1}{2}, \quad C = -\frac{3}{4}.$$

The particular solution is therefore

$$x_p(t) = -\frac{1}{2}t^2 - \frac{1}{2}t - \frac{3}{4}.$$

8.7 Resonance

View tutorial on YouTube

Resonance occurs when the frequency of the inhomogeneous term matches the frequency of the homogeneous solution. To illustrate resonance in its simplest embodiment, we consider the second-order linear inhomogeneous ode

$$\ddot{x} + \omega_0^2 x = f\cos\omega t, \quad x(0) = 0,\ \dot{x}(0) = 0. \tag{8.27}$$

Our main goal is to determine what happens to the solution in the limit $\omega \to \omega_0$.

The homogeneous equation has characteristic equation

$$r^2 + \omega_0^2 = 0,$$

so that $r_\pm = \pm i\omega_0$. Therefore,

$$x_h(t) = c_1\cos\omega_0 t + c_2\sin\omega_0 t. \tag{8.28}$$

To find a particular solution, we note the absence of a first-derivative term, and simply try

$$x(t) = A\cos\omega t.$$

Upon substitution into the ode, we obtain

$$-\omega^2 A + \omega_0^2 A = f,$$

or

$$A = \frac{f}{\omega_0^2 - \omega^2}.$$

Therefore,

$$x_p(t) = \frac{f}{\omega_0^2 - \omega^2} \cos \omega t.$$

Our general solution is thus

$$x(t) = c_1 \cos \omega_0 t + c_2 \sin \omega_0 t + \frac{f}{\omega_0^2 - \omega^2} \cos \omega t,$$

with derivative

$$\dot{x}(t) = \omega_0 (c_2 \cos \omega_0 t - c_1 \sin \omega_0 t) - \frac{f\omega}{\omega_0^2 - \omega^2} \sin \omega t.$$

Initial conditions are satisfied when

$$0 = c_1 + \frac{f}{\omega_0^2 - \omega^2}, \qquad 0 = c_2 \omega_0,$$

so that

$$c_1 = \frac{f}{\omega^2 - \omega_0^2}, \qquad c_2 = 0.$$

Therefore, the solution to the ode that satisfies the initial conditions is

$$\begin{aligned} x(t) &= \frac{f}{\omega^2 - \omega_0^2} \cos \omega_0 t - \frac{f}{\omega^2 - \omega_0^2} \cos \omega t \\ &= \frac{f(\cos \omega_0 t - \cos \omega t)}{\omega^2 - \omega_0^2}. \end{aligned}$$

Resonance occurs in the limit $\omega \to \omega_0$; that is, the frequency of the inhomogeneous term (the external force) matches the frequency of the homogeneous solution (the free oscillation). Using L'Hospital's rule, we can determine the indeterminate $0/0$ limit by differentiating the numerator and denominator with respect to ω:

$$\begin{aligned} \lim_{\omega \to \omega_0} \frac{f(\cos \omega_0 t - \cos \omega t)}{\omega^2 - \omega_0^2} &= \lim_{\omega \to \omega_0} \frac{ft \sin \omega t}{2\omega} \\ &= \frac{ft \sin \omega_0 t}{2\omega_0}. \end{aligned} \tag{8.29}$$

At resonance, the term proportional to the amplitude f of the inhomogeneous term increases linearly with t, resulting in larger-and-larger amplitudes of oscillation for $x(t)$. In general, if the inhomogeneous term in the differential equation is a solution of the corresponding homogeneous differential equation, then the correct ansatz for the particular solution is a constant times the inhomogeneous term times t.

To illustrate this same example further, we return to the original ode, now assumed to be exactly at resonance,

$$\ddot{x} + \omega_0^2 x = f \cos \omega_0 t,$$

and find a particular solution directly. The particular solution is the real part of the particular solution of

$$\ddot{z} + \omega_0^2 z = f e^{i\omega_0 t}.$$

If we try $z_p = Ce^{i\omega_0 t}$, we obtain $0 = f$, showing that the particular solution is not of this form. Because the inhomogeneous term is a solution of the homogeneous equation, we should take as our ansatz

$$z_p = Ate^{i\omega_0 t}.$$

We have

$$\dot{z}_p = Ae^{i\omega_0 t}\left(1 + i\omega_0 t\right), \quad \ddot{z}_p = Ae^{i\omega_0 t}\left(2i\omega_0 - \omega_0^2 t\right);$$

and upon substitution into the ode

$$\begin{aligned}\ddot{z}_p + \omega_0^2 z_p &= Ae^{i\omega_0 t}\left(2i\omega_0 - \omega_0^2 t\right) + \omega_0^2 Ate^{i\omega_0 t} \\ &= 2i\omega_0 Ae^{i\omega_0 t} \\ &= fe^{i\omega_0 t}.\end{aligned}$$

Therefore,

$$A = \frac{f}{2i\omega_0},$$

and

$$\begin{aligned}x_p &= \mathrm{Re}\{\frac{ft}{2i\omega_0}e^{i\omega_0 t}\} \\ &= \frac{ft\sin\omega_0 t}{2\omega_0},\end{aligned}$$

the same result as (8.29).

Example: Find a particular solution of $\ddot{x} - 3\dot{x} - 4x = 5e^{-t}$.

View tutorial on YouTube

If we naively try the ansatz

$$x = Ae^{-t},$$

and substitute this into the inhomogeneous differential equation, we obtain

$$A + 3A - 4A = 5,$$

or $0 = 5$, which is clearly nonsense. Our ansatz therefore fails to find a solution. The cause of this failure is that the corresponding homogeneous equation has solution

$$x_h = c_1 e^{4t} + c_2 e^{-t},$$

so that the inhomogeneous term $5e^{-t}$ is one of the solutions of the homogeneous equation. To find a particular solution, we should therefore take as our ansatz

$$x = Ate^{-t},$$

with first- and second-derivatives given by

$$\dot{x} = Ae^{-t}(1 - t), \quad \ddot{x} = Ae^{-t}(-2 + t).$$

Substitution into the differential equation yields

$$Ae^{-t}(-2+t) - 3Ae^{-t}(1-t) - 4Ate^{-t} = 5e^{-t}.$$

The terms containing t cancel out of this equation, resulting in $-5A = 5$, or $A = -1$. Therefore, the particular solution is

$$x_p = -te^{-t}.$$

8.8 Applications

View *Nondimensionalization* on YouTube

8.8.1 RLC circuit

View *RLC circuit* on YouTube

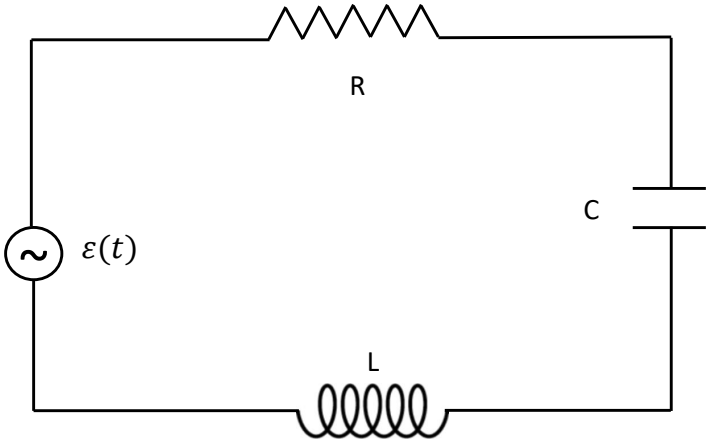


Figure 8.1: RLC circuit diagram.

Consider a resister R, an inductor L, and a capacitor C connected in series as shown in Fig. 8.1. An AC generator provides a time-varying electromotive force (emf), $\mathcal{E}(t)$, to the circuit. Here, we determine the differential equation satisfied by the charge on the capacitor.

The constitutive equations for the voltage drops across a capacitor, a resister, and an inductor are given by

$$V_C = q/C, \quad V_R = iR, \quad V_L = \frac{di}{dt}L, \tag{8.30}$$

where C is the capacitance, R is the resistance, and L is the inductance. The charge q and the current i are related by

$$i = \frac{dq}{dt}. \tag{8.31}$$

Kirchhoff's voltage law states that the emf $\mathcal{E}$ applied to any closed loop is equal to the sum of the voltage drops in that loop. Applying Kirchhoff's voltage law to the RLC ciruit results in

$$V_L + V_R + V_C = \mathcal{E}(t); \tag{8.32}$$

or using (8.30) and (8.31),

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{1}{C}q = \mathcal{E}(t).$$

The equation for the RLC circuit is a second-order linear inhomogeneous differential equation with constant coefficients.

The AC voltage can be written as $\mathcal{E}(t) = \mathcal{E}_0 \cos \omega t$, and the governing equation for $q = q(t)$ can be written as

$$\frac{d^2q}{dt^2} + \frac{R}{L}\frac{dq}{dt} + \frac{1}{LC}q = \frac{\mathcal{E}_0}{L}\cos \omega t. \tag{8.33}$$

Nondimensionalization of this equation will be shown to reduce the number of free parameters.

To construct dimensionless variables, we first define the natural frequency of oscillation of a system to be the frequency of oscillation in the absence of any driving or damping forces. The iconic example is the simple harmonic oscillator, with equation given by

$$\ddot{x} + \omega_0^2 x = 0,$$

and general solution given by $x(t) = A\cos \omega_0 t + B \sin \omega_0 t$. Here, the natural frequency of oscillation is ω_0, and the period of oscillation is $T = 2\pi/\omega_0$. For the RLC circuit, the natural frequency of oscillation is found from the coefficient of the q term, and is given by

$$\omega_0 = \frac{1}{\sqrt{LC}}.$$

Making use of ω_0, with units of one over time, we can define a dimensionless time τ and a dimensionless charge Q by

$$\tau = \omega_0 t, \quad Q = \frac{\omega_0^2 L}{\mathcal{E}_0} q.$$

The resulting dimensionless equation for the RLC circuit is then given by

$$\frac{d^2Q}{d\tau^2} + \alpha\frac{dQ}{d\tau} + Q = \cos \beta\tau, \tag{8.34}$$

where α and β are dimensionless parameters given by

$$\alpha = \frac{R}{L\omega_0}, \quad \beta = \frac{\omega}{\omega_0}.$$

Notice that the original five parameters in (8.33), namely R, L, C, $\mathcal{E}_0$ and ω, have been reduced to the two dimensionless parameters α and β. We will return later to solve (8.34) after visiting two more applications that will be shown to be governed by the same dimensionless equation.

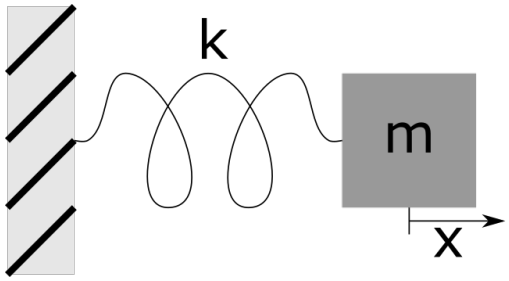


Figure 8.2: Mass-spring system (top view).

8.8.2 Mass on a spring

View *Mass on a Spring* on YouTube

Consider a mass lying on a flat surface connected by a spring to a wall, with top view shown in Fig. 8.2. The spring force is modeled by Hooke's law, $F_s = -kx$, and sliding friction is modeled as $F_f = -cdx/dt$. An external force is applied and is assumed to be sinusoidal, with $F_e = F_0 \cos \omega t$. Newton's equation, $F = ma$, results in

$$m\frac{d^2x}{dt^2} = -kx - c\frac{dx}{dt} + F_0 \cos \omega t.$$

Rearranging terms, we obtain

$$\frac{d^2x}{dt^2} + \frac{c}{m}\frac{dx}{dt} + \frac{k}{m}x = \frac{F_0}{m}\cos \omega t.$$

Here, the natural frequency of oscillation is given by

$$\omega_0 = \sqrt{\frac{k}{m}},$$

and we define a dimensionless time τ and a dimensionless position X by

$$\tau = \omega_0 t, \quad X = \frac{m\omega_0^2}{F_0}x.$$

The resulting dimensionless equation is given by

$$\frac{d^2X}{d\tau^2} + \alpha\frac{dX}{d\tau} + X = \cos \beta\tau, \tag{8.35}$$

where here, α and β are dimensionless parameters given by

$$\alpha = \frac{c}{m\omega_0}, \quad \beta = \frac{\omega}{\omega_0}.$$

Notice that even though the physical problem is different, and the dimensionless variables are defined differently, the resulting dimensionless equation (8.35) for the mass-spring system is the same as that for the *RLC* circuit (8.34).

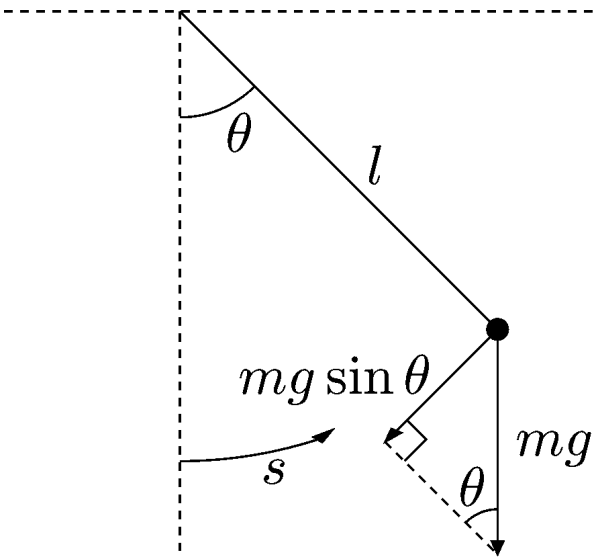


Figure 8.3: The pendulum.

8.8.3 Pendulum

View *Pendulum* on YouTube

Here, we consider a mass that is attached to a massless rigid rod and is constrained to move along an arc of a circle centered at the pivot point (see Fig. 8.3). Suppose l is the fixed length of the connecting rod, and θ is the angle it makes with the vertical.

We can apply Newton's equation, $F = ma$, to the mass with origin at the bottom and axis along the arc with positive direction to the right. The position s of the mass along the arc is given by $s = l\theta$. The relevant gravitational force on the pendulum is the component along the arc, and from Fig. 8.3 is observed to be

$$F_g = -mg \sin\theta.$$

We model friction to be proportional to the velocity of the pendulum along the arc, that is

$$F_f = -c\dot{s} = -cl\dot{\theta}.$$

With a sinusoidal external force, $F_e = F_0 \cos\omega t$, Newton's equation $m\ddot{s} = F_g + F_f + F_e$ results in

$$ml\ddot{\theta} = -mg \sin\theta - cl\dot{\theta} + F_0 \cos\omega t.$$

Rewriting, we have

$$\ddot{\theta} + \frac{c}{m}\dot{\theta} + \frac{g}{l}\sin\theta = \frac{F_0}{ml}\cos\omega t.$$

At small amplitudes of oscillation, we can approximate $\sin\theta \approx \theta$, and the natural frequency of oscillation ω_0 of the mass is given by

$$\omega_0 = \sqrt{\frac{g}{l}}.$$

Nondimensionalizing time as $\tau = \omega_0 t$, the dimensionless pendulum equation becomes

$$\frac{d^2\theta}{d\tau^2} + \alpha\frac{d\theta}{d\tau} + \sin\theta = \gamma\cos\beta\tau,$$

where α, β, and γ are dimensionless parameters given by

$$\alpha = \frac{c}{m\omega_0}, \quad \beta = \frac{\omega}{\omega_0}, \quad \gamma = \frac{F_0}{ml\omega_0^2}.$$

The nonlinearity of the pendulum equation, with the term $\sin\theta$, results in the additional dimensionless parameter γ. For small amplitude of oscillation, however, we can scale θ by $\theta = \gamma\Theta$, and the small amplitude dimensionless equation becomes

$$\frac{d^2\Theta}{d\tau^2} + \alpha\frac{d\Theta}{d\tau} + \Theta = \cos\beta\tau, \tag{8.36}$$

the same equation as (8.34) and (8.35).

8.9 Damped resonance

View *Damped Resonance* on YouTube

We now solve the dimensionless equations given by (8.34), (8.35) and (8.36), written here as

$$\ddot{x} + \alpha\dot{x} + x = \cos\beta t, \tag{8.37}$$

where the physical constraints of our three applications requires that $\alpha > 0$. The homogeneous equation has characteristic equation

$$r^2 + \alpha r + 1 = 0,$$

so that the solutions are

$$r_\pm = -\frac{\alpha}{2} \pm \frac{1}{2}\sqrt{\alpha^2 - 4}.$$

When $\alpha^2 - 4 < 0$, the motion of the unforced oscillator is said to be underdamped; when $\alpha^2 - 4 > 0$, overdamped; and when $\alpha^2 - 4 = 0$, critically damped. For all three types of damping, the roots of the characteristic equation satisfy $\mathrm{Re}(r_\pm) < 0$. Therefore, both linearly independent homogeneous solutions decay exponentially to zero, and the long-time asymptotic solution of (8.37) reduces to the non-decaying particular solution. Since the initial conditions are satisfied by the free constants multiplying the decaying homogeneous solutions, the long-time asymptotic solution is independent of the initial conditions.

If we are only interested in the long-time asymptotic solution of (8.37), we can proceed directly to the determination of a particular solution. As before, we consider the complex ode

$$\ddot{z} + \alpha\dot{z} + z = e^{i\beta t},$$

with $x_p = \mathrm{Re}(z_p)$. With the ansatz $z_p = Ae^{i\beta t}$, we have

$$-\beta^2 A + i\alpha\beta A + A = 1,$$

or

$$\begin{aligned} A &= \frac{1}{(1-\beta^2)+i\alpha\beta} \\ &= \left(\frac{1}{(1-\beta^2)^2+\alpha^2\beta^2}\right)\Big((1-\beta^2)-i\alpha\beta\Big). \end{aligned} \tag{8.38}$$

To determine x_p, we utilize the polar form of a complex number. The complex number $z = x + iy$ can be written in polar form as $z = re^{i\phi}$, where $r = \sqrt{x^2+y^2}$ and $\tan\phi = y/x$. We therefore write

$$(1-\beta^2) - i\alpha\beta = re^{i\phi},$$

with

$$r = \sqrt{(1-\beta^2)^2+\alpha^2\beta^2}, \quad \tan\phi = -\frac{\alpha\beta}{1-\beta^2}.$$

Using the polar form, A in (8.38) becomes

$$A = \left(\frac{1}{\sqrt{(1-\beta^2)^2+\alpha^2\beta^2}}\right)e^{i\phi},$$

and $x_p = \mathrm{Re}(Ae^{i\beta t})$ becomes

$$\begin{aligned} x_p &= \left(\frac{1}{\sqrt{(1-\beta^2)^2+\alpha^2\beta^2}}\right)\mathrm{Re}\left\{e^{i(\beta t+\phi)}\right\} \\ &= \left(\frac{1}{\sqrt{(1-\beta^2)^2+\alpha^2\beta^2}}\right)\cos(\beta t+\phi). \end{aligned} \tag{8.39}$$

We conclude with a couple of observations about (8.39). First, if the forcing frequency ω is equal to the natural frequency ω_0 of the undamped oscillator, then $\beta = 1$ and $A = 1/i\alpha$, and $x_p = (1/\alpha)\sin t$. The oscillator position is observed to be $\pi/2$ out of phase with the external force, or in other words, the velocity of the oscillator, not the position, is in phase with the force. Second, the value of β that maximizes the amplitude of oscillation is the value of β that minimizes the denominator of (8.39). To determine β_m we thus minimize the function $g(\beta^2)$ with respect to β^2, where

$$g(\beta^2) = (1-\beta^2)^2 + \alpha^2\beta^2.$$

Taking the derivative of g with respect to β^2 and setting this to zero to determine β_m yields

$$-2(1-\beta_m^2) + \alpha^2 = 0,$$

or

$$\beta_m = \sqrt{1-\frac{\alpha^2}{2}} \approx 1 - \frac{\alpha^2}{4},$$

the last approximation valid if $\alpha << 1$ and the dimensionless damping coefficient is small. We can interpret this result by saying that small damping slightly lowers the "resonance" frequency of the undamped oscillator.

Chapter 9

Series solutions of homogeneous linear second-order differential equations

Reference: Boyce and DiPrima, Chapter 5

We consider the homogeneous linear second-order differential equation for $y = y(x)$:

$$P(x)y'' + Q(x)y' + R(x)y = 0, \tag{9.1}$$

where $P(x)$, $Q(x)$ and $R(x)$ are polynomials or convergent power series around $x = x_0$, with no common polynomial factors that could be divided out. The value $x = x_0$ is called an *ordinary point* of (9.1) if $P(x_0) \neq 0$, and is called a singular point if $P(x_0) = 0$. Singular points can be further classified as *regular singular points* and *irregular singular points*. Here, we will only consider series expansions about ordinary points. Our goal is to find two independent solutions of (9.1).

9.1 Ordinary points

If x_0 is an ordinary point of (9.1), then it is possible to determine two power series (i.e., Taylor series) solutions for $y = y(x)$ centered at $x = x_0$. We illustrate the method of solution by solving two examples, with $x_0 = 0$.

Example: Find the general solution of $y'' + y = 0$.

View tutorial on YouTube

By now, you should know that the general solution is $y(x) = a_0 \cos x + a_1 \sin x$, with a_0 and a_1 constants. To find a power series solution about the point $x_0 = 0$, we write

$$y(x) = \sum_{n=0}^{\infty} a_n x^n;$$

and upon differentiating term-by-term

$$y'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1},$$

and

$$y''(x) = \sum_{n=2}^{\infty} n(n-1) a_n x^{n-2}.$$

Substituting the power series for y and its derivatives into the differential equation to be solved, we obtain

$$\sum_{n=2}^{\infty} n(n-1)a_n x^{n-2} + \sum_{n=0}^{\infty} a_n x^n = 0. \tag{9.2}$$

The power-series solution method requires combining the two sums on the left-hand-side of (9.2) into a single power series in x. To shift the exponent of x^{n-2} in the first sum upward by two to obtain x^n, we need to shift the summation index downward by two; that is,

$$\sum_{n=2}^{\infty} n(n-1)a_n x^{n-2} = \sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2} x^n.$$

We can then combine the two sums in (9.2) to obtain

$$\sum_{n=0}^{\infty} \Big((n+2)(n+1)a_{n+2} + a_n \Big) x^n = 0. \tag{9.3}$$

For (9.3) to be satisfied, the coefficient of each power of x must vanish separately. (This can be proved by setting $x = 0$ after successive differentiation.) We therefore obtain the *recurrence relation*

$$a_{n+2} = -\frac{a_n}{(n+2)(n+1)}, \quad n = 0, 1, 2, \ldots.$$

We observe that even and odd coefficients decouple. We thus obtain two independent sequences starting with first term a_0 or a_1. Developing these sequences, we have for the sequence beginning with a_0:

$$\begin{aligned}
&a_0, \\
&a_2 = -\frac{1}{2} a_0, \\
&a_4 = -\frac{1}{4 \cdot 3} a_2 = \frac{1}{4 \cdot 3 \cdot 2} a_0, \\
&a_6 = -\frac{1}{6 \cdot 5} a_4 = -\frac{1}{6!} a_0;
\end{aligned}$$

and the general coefficient in this sequence for $n = 0, 1, 2, \ldots$ is

$$a_{2n} = \frac{(-1)^n}{(2n)!} a_0.$$

Also, for the sequence beginning with a_1:

$$\begin{aligned}
&a_1, \\
&a_3 = -\frac{1}{3 \cdot 2} a_1, \\
&a_5 = -\frac{1}{5 \cdot 4} a_3 = \frac{1}{5 \cdot 4 \cdot 3 \cdot 2} a_1, \\
&a_7 = -\frac{1}{7 \cdot 6} a_5 = -\frac{1}{7!} a_1;
\end{aligned}$$

and the general coefficient in this sequence for $n = 0, 1, 2, \ldots$ is

$$a_{2n+1} = \frac{(-1)^n}{(2n+1)!} a_1.$$

Using the principle of superposition, the general solution is therefore

$$\begin{aligned} y(x) &= a_0 \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} + a_1 \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} \\ &= a_0 \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots\right) + a_1 \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots\right) \\ &= a_0 \cos x + a_1 \sin x, \end{aligned}$$

as expected.

In our next example, we will solve the *Airy's Equation*. This differential equation arises in the study of optics, fluid mechanics, and quantum mechanics.

Example: Find the general solution of $y'' - xy = 0$.

View tutorial on YouTube

With

$$y(x) = \sum_{n=0}^{\infty} a_n x^n,$$

the differential equation becomes

$$\sum_{n=2}^{\infty} n(n-1) a_n x^{n-2} - \sum_{n=0}^{\infty} a_n x^{n+1} = 0. \tag{9.4}$$

We shift the first sum to x^{n+1} by shifting the exponent up by three, i.e.,

$$\sum_{n=2}^{\infty} n(n-1) a_n x^{n-2} = \sum_{n=-1}^{\infty} (n+3)(n+2) a_{n+3} x^{n+1}.$$

When combining the two sums in (9.4), we separate out the extra $n = -1$ term in the first sum given by the constant term $2a_2$. Therefore, (9.4) becomes

$$2a_2 + \sum_{n=0}^{\infty} \Big((n+3)(n+2) a_{n+3} - a_n \Big) x^{n+1} = 0. \tag{9.5}$$

Setting coefficients of powers of x to zero, we first find $a_2 = 0$, and then obtain the recursion relation

$$a_{n+3} = \frac{1}{(n+3)(n+2)} a_n. \tag{9.6}$$

Three sequences of coefficients—those starting with either a_0, a_1 or a_2—decouple. In particular the three sequences are

$$\begin{aligned} &a_0, a_3, a_6, a_9, \ldots; \\ &a_1, a_4, a_7, a_{10}, \ldots; \\ &a_2, a_5, a_8, a_{11} \ldots. \end{aligned}$$

Since $a_2 = 0$, we find immediately for the last sequence

$$a_2 = a_5 = a_8 = a_{11} = \cdots = 0.$$

We compute the first four nonzero terms in the power series with coefficients corresponding to the first two sequences. Starting with a_0, we have

$$\begin{aligned} &a_0, \\ &a_3 = \frac{1}{3 \cdot 2} a_0, \\ &a_6 = \frac{1}{6 \cdot 5 \cdot 3 \cdot 2} a_0, \\ &a_9 = \frac{1}{9 \cdot 8 \cdot 6 \cdot 5 \cdot 3 \cdot 2} a_0; \end{aligned}$$

and starting with a_1,

$$\begin{aligned} &a_1, \\ &a_4 = \frac{1}{4 \cdot 3} a_1, \\ &a_7 = \frac{1}{7 \cdot 6 \cdot 4 \cdot 3} a_1, \\ &a_{10} = \frac{1}{10 \cdot 9 \cdot 7 \cdot 6 \cdot 4 \cdot 3} a_1. \end{aligned}$$

The general solution for $y = y(x)$, can therefore be written as

$$\begin{aligned} y(x) &= a_0 \left(1 + \frac{x^3}{6} + \frac{x^6}{180} + \frac{x^9}{12960} + \dots\right) + a_1 \left(x + \frac{x^4}{12} + \frac{x^7}{504} + \frac{x^{10}}{45360} + \dots\right) \\ &= a_0 y_0(x) + a_1 y_1(x). \end{aligned}$$

Suppose we would like to graph the solutions $y = y_0(x)$ and $y = y_1(x)$ versus x by solving the differential equation $y'' - xy = 0$ numerically. What initial conditions should we use? Clearly, $y = y_0(x)$ solves the ode with initial values $y(0) = 1$ and $y'(0) = 0$, while $y = y_1(x)$ solves the ode with initial values $y(0) = 0$ and $y'(0) = 1$.

The numerical solutions, obtained using MATLAB, are shown in Fig. 9.1. Note that the solutions oscillate for negative x and grow exponentially for positive x. This can be understood by recalling that $y'' + y = 0$ has oscillatory sine and cosine solutions and $y'' - y = 0$ has exponential hyperbolic sine and cosine solutions.

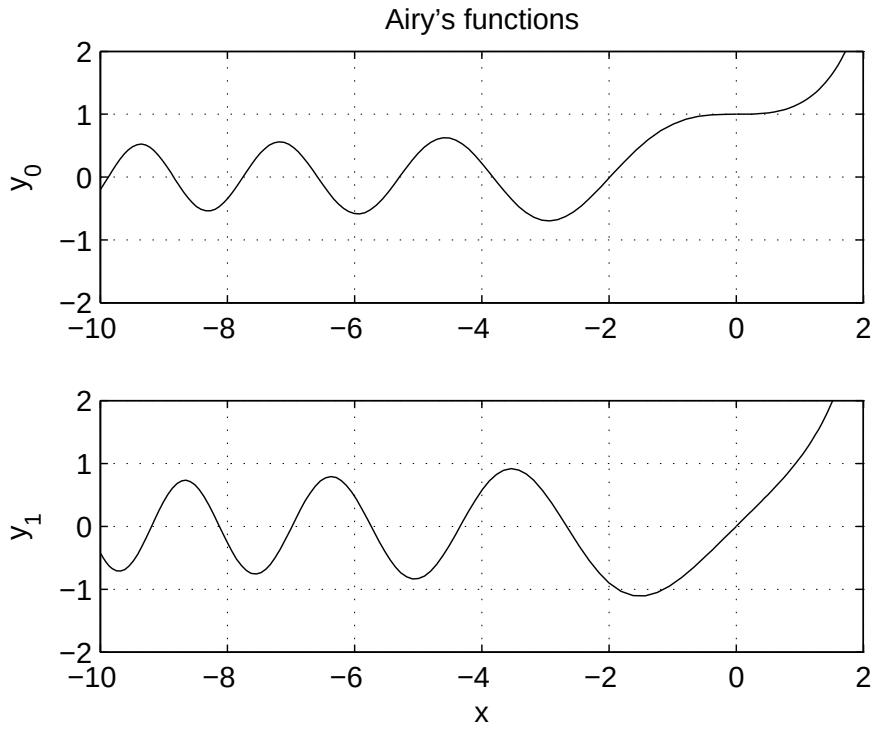


Figure 9.1: *Numerical solution of Airy's equation.*

Chapter 10

Systems of linear differential equations

Reference: Boyce and DiPrima, Chapter 7

Here, we consider the simplest case of a system of two coupled homogeneous linear first-order equations with constant coefficients. The general system is given by

$$\dot{x}_1 = ax_1 + bx_2, \quad \dot{x}_2 = cx_1 + dx_2, \tag{10.1}$$

or in matrix form as

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

The short-hand notation will be

$$\dot{\mathrm{x}} = \mathrm{A}\mathrm{x}. \tag{10.2}$$

Although we can write these two first-order equations as a single second-order equation, we will instead make use of our newly learned techniques in matrix algebra. We will also introduce the important concept of the phase space, and the physical problem of coupled oscillators.

10.1 Distinct real eigenvalues

We illustrate the solution method by example.

Example: Find the general solution of $\dot{x}_1 = x_1 + x_2$, $\dot{x}_2 = 4x_1 + x_2$.

View tutorial on YouTube

The equation to be solved may be rewritten in matrix form as

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 4 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{10.3}$$

or in short hand as (10.2).

We take as our ansatz $\mathrm{x}(t) = \mathrm{v}e^{\lambda t}$, where v is a vector and λ is a scalar, and both are independent of t. Upon substitution into (10.2), we obtain

$$\lambda \mathrm{v}e^{\lambda t} = \mathrm{A}\mathrm{v}e^{\lambda t};$$

and upon cancellation of the exponential, we obtain the eigenvalue problem

$$\mathrm{A}\mathrm{v} = \lambda \mathrm{v}. \tag{10.4}$$

Finding the characteristic equation using (5.4), we have

$$\begin{aligned} 0 &= \det(\mathrm{A} - \lambda \mathrm{I}) \\ &= \lambda^2 - 2\lambda - 3 \\ &= (\lambda - 3)(\lambda + 1). \end{aligned}$$

Therefore, the two eigenvalues are $\lambda_1 = 3$ and $\lambda_2 = -1$.

To determine the corresponding eigenvectors, we substitute the eigenvalues successively into

$$(\mathrm{A} - \lambda \mathrm{I})\mathrm{v} = 0. \tag{10.5}$$

We will write the corresponding eigenvectors v_1 and v_2 using the matrix notation

$$\begin{pmatrix} \mathrm{v}_1 & \mathrm{v}_2 \end{pmatrix} = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix},$$

where the components of v_1 and v_2 are written with subscripts corresponding to the first and second columns of a 2-by-2 matrix.

For $\lambda_1 = 3$, and unknown eigenvector v_1, we have from (10.5)

$$\begin{aligned} -2v_{11} + v_{21} &= 0, \\ 4v_{11} - 2v_{21} &= 0. \end{aligned}$$

Clearly, the second equation is just the first equation multiplied by -2, so only one equation is linearly independent. This will always be true, so for the 2-by-2 case we need only consider the first row of the matrix. The first eigenvector therefore satisfies $v_{21} = 2v_{11}$. Recall that an eigenvector is only unique up to multiplication by a constant: we may therefore take $v_{11} = 1$ for convenience.

For $\lambda_2 = -1$, and eigenvector $\mathrm{v}_2 = (v_{12}, v_{22})^T$, we have from (10.5)

$$2v_{12} + v_{22} = 0,$$

so that $v_{22} = -2v_{12}$. Here, we take $v_{12} = 1$.

Therefore, our eigenvalues and eigenvectors are given by

$$\lambda_1 = 3,\ \mathrm{v}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}; \qquad \lambda_2 = -1,\ \mathrm{v}_2 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

Using the principle of superposition, the general solution to the ode is therefore

$$\mathrm{X}(t) = c_1 \mathrm{v}_1 e^{\lambda_1 t} + c_2 \mathrm{v}_2 e^{\lambda_2 t},$$

or explicitly writing out the components,

$$\begin{aligned} x_1(t) &= c_1 e^{3t} + c_2 e^{-t}, \\ x_2(t) &= 2c_1 e^{3t} - 2c_2 e^{-t}. \end{aligned} \tag{10.6}$$

We can obtain a new perspective on the solution by drawing a phase portrait, shown in Fig. 10.1, with "x-axis" x_1 and "y-axis" x_2. Each curve corresponds to a different initial condition, and represents the trajectory of a particle with velocity given by the differential equation. The dark lines represent trajectories along the direction of the eigenvectors. If $c_2 = 0$, the motion is along the eigenvector v_1 with $x_2 = 2x_1$ and the motion with increasing time is away from the origin (arrows pointing out) since the eigenvalue $\lambda_1 = 3 > 0$. If $c_1 = 0$, the motion is along the eigenvector v_2 with $x_2 = -2x_1$ and motion is towards the origin (arrows pointing in) since the eigenvalue $\lambda_2 = -1 < 0$. When the eigenvalues are real and of opposite signs, the origin is called a *saddle point*. Almost all trajectories (with the exception of those with initial conditions exactly satisfying $x_2(0) = -2x_1(0)$) eventually move away from the origin as t increases. When the eigenvalues are real and of the same sign, the origin is called a *node*. A node can be *stable* (negative eigenvalues) or *unstable* (positive eigenvalues).

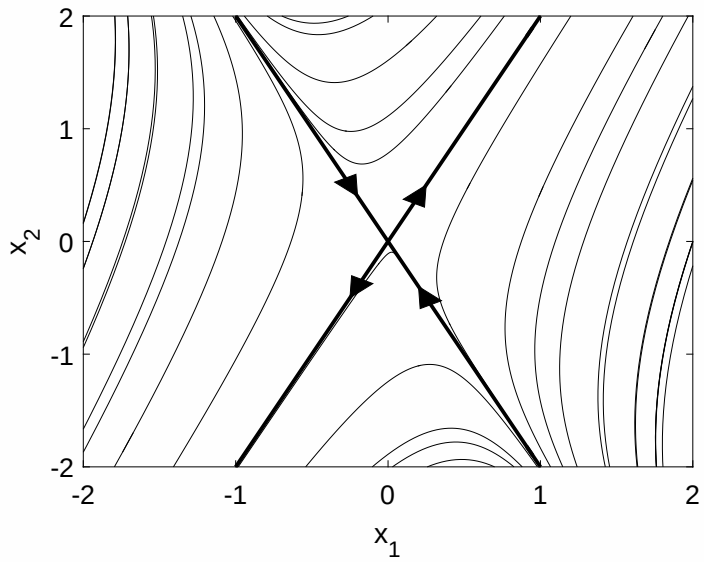


Figure 10.1: *Phase portrait for example with two real eigenvalues of opposite sign.*

10.2 Solution by diagonalization

Another way to view the problem of coupled first-order linear odes is from the perspective of matrix diagonalization. With

$$\dot{x} = Ax, \tag{10.7}$$

we suppose A can be diagonalized using

$$S^{-1}AS = \Lambda, \tag{10.8}$$

where Λ is the diagonal eigenvalue matrix, and the columns of S hold the eigenvectors. We can change variables in (10.7) using

$$x = Sy \tag{10.9}$$

and obtain

$$S\dot{y} = ASy.$$

Multiplication on the left by S^{-1} and using (10.8) results in

$$\begin{aligned}\dot{y} &= S^{-1}ASy \\ &= \Lambda y.\end{aligned}$$

The first-order differential equations in the y-variables are now uncoupled and can be immediately solved, and the x-variables can be recovered using (10.9).

Example: Solve the previous example

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 4 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

by the diagonalization method.

The eigenvalues and eigenvectors are known, and we have

$$\Lambda = \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}, \quad S = \begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix},$$

and the uncoupled y-equations are given by

$$\dot{y}_1 = 3y_1, \quad \dot{y}_2 = -y_2,$$

with solution

$$y_1(t) = c_1 e^{3t}, \quad y_2 = c_2 e^{-t}.$$

Transforming back to the x-variables using (10.9), we have

$$\begin{aligned} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= \begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} c_1 e^{3t} \\ c_2 e^{-t} \end{pmatrix} \\ &= \begin{pmatrix} c_1 e^{3t} + c_2 e^{-t} \\ 2c_1 e^{3t} - 2c_2 e^{-t} \end{pmatrix}, \end{aligned}$$

which agrees with solution (10.6).

10.3 Solution by the matrix exponential

Another interesting approach to this problem makes use of the matrix exponential. Let A be a square matrix, tA the matrix A multiplied by the scalar t, and A^n the matrix A multiplied by itself n times. We define the matrix exponential function e^{tA} similar to the way the exponential function may be defined using its Taylor series. The corresponding definition is

$$e^{tA} = I + tA + \frac{t^2A^2}{2!} + \frac{t^3A^3}{3!} + \frac{t^4A^4}{4!} + \dots.$$

We can differentiate e^{tA} with respect to t and obtain

$$\begin{aligned} \frac{d}{dt}e^{tA} &= A + tA^2 + \frac{t^2A^3}{2!} + \frac{t^3A^4}{3!} + \dots \\ &= A\left(I + tA + \frac{t^2A^2}{2!} + \frac{t^3A^3}{3!} + \dots\right) \\ &= Ae^{tA}, \end{aligned}$$

as one would expect from differentiating the exponential function. We can therefore formally write the solution of

$$\dot{x} = Ax$$

as

$$x(t) = e^{tA}x(0).$$

If the matrix A is diagonalizable such that $\mathrm{A} = \mathrm{S}\Lambda\mathrm{S}^{-1}$, then observe that

$$\begin{aligned} e^{t\mathrm{A}} &= e^{t\mathrm{S}\Lambda\mathrm{S}^{-1}} \\ &= \mathrm{I} + t\mathrm{S}\Lambda\mathrm{S}^{-1} + \frac{t^2(\mathrm{S}\Lambda\mathrm{S}^{-1})^2}{2!} + \frac{t^3(\mathrm{S}\Lambda\mathrm{S}^{-1})^3}{3!} + \dots \\ &= \mathrm{I} + t\mathrm{S}\Lambda\mathrm{S}^{-1} + \frac{t^2\mathrm{S}\Lambda^2\mathrm{S}^{-1}}{2!} + \frac{t^3\mathrm{S}\Lambda^3\mathrm{S}^{-1}}{3!} + \dots \\ &= \mathrm{S}\left(\mathrm{I} + t\Lambda + \frac{t^2\Lambda^2}{2!} + \frac{t^3\Lambda^3}{3!} + \dots\right)\mathrm{S}^{-1} \\ &= \mathrm{S}e^{t\Lambda}\mathrm{S}^{-1}. \end{aligned}$$

If Λ is a diagonal matrix with diagonal elements λ_1, λ_2, etc., then the matrix exponential $e^{t\Lambda}$ is also a diagonal matrix with diagonal elements given by $e^{\lambda_1 t}$, $e^{\lambda_2 t}$, etc. We can now use the matrix exponential to solve a system of linear differential equations.

Example: Solve the previous example

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 4 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

by matrix exponentiation.

We know that

$$\Lambda = \begin{pmatrix} 3 & 0 \\ 0 & -1 \end{pmatrix}, \quad \mathrm{S} = \begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix}, \quad \mathrm{S}^{-1} = -\frac{1}{4}\begin{pmatrix} -2 & -1 \\ -2 & 1 \end{pmatrix}.$$

The solution to the system of differential equations is then given by

$$\begin{aligned} \mathrm{x}(t) &= e^{t\mathrm{A}}\mathrm{x}(0) \\ &= \mathrm{S}e^{t\Lambda}\mathrm{S}^{-1}\mathrm{x}(0) \\ &= -\frac{1}{4}\begin{pmatrix} 1 & 1 \\ 2 & -2 \end{pmatrix}\begin{pmatrix} e^{3t} & 0 \\ 0 & e^{-t} \end{pmatrix}\begin{pmatrix} -2 & -1 \\ -2 & 1 \end{pmatrix}\mathrm{x}(0). \end{aligned}$$

Successive matrix multiplication results in

$$\begin{aligned} x_1(t) &= \frac{1}{4}(2x_1(0) + x_2(0))e^{3t} + \frac{1}{4}(2x_1(0) - x_2(0))e^{-t}, \\ x_2(t) &= \frac{1}{2}(2x_1(0) + x_2(0))e^{3t} - \frac{1}{2}(2x_1(0) - x_2(0))e^{-t}, \end{aligned}$$

which is the same solution as previously found, but here the c_1 and c_2 free coefficients are replaced by the initial values of $\mathrm{x}(0)$.

10.4 Distinct complex-conjugate eigenvalues

Example: Find the general solution of $\dot{x}_1 = -\frac{1}{2}x_1 + x_2$, $\dot{x}_2 = -x_1 - \frac{1}{2}x_2$.

View tutorial on YouTube

The equations in matrix form are

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -1/2 & 1 \\ -1 & -1/2 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

The ansatz $\mathrm{x} = \mathrm{v}e^{\lambda t}$ leads to the equation

$$\begin{aligned} 0 &= \det(\mathrm{A} - \lambda \mathrm{I}) \\ &= \lambda^2 + \lambda + \frac{5}{4}. \end{aligned}$$

Therefore, $\lambda = -1/2 \pm i$; and we observe that the eigenvalues occur as a complex conjugate pair. We will denote the two eigenvalues as

$$\lambda = -\frac{1}{2} + i \quad \text{and} \quad \bar{\lambda} = -\frac{1}{2} - i.$$

Now, if A a real matrix, then $\mathrm{Av} = \lambda \mathrm{v}$ implies $\mathrm{A}\overline{\mathrm{v}} = \bar{\lambda}\overline{\mathrm{v}}$, so the eigenvectors also occur as a complex conjugate pair. The eigenvector v associated with eigenvalue λ satisfies $-iv_1 + v_2 = 0$, and normalizing with $v_1 = 1$, we have

$$\mathrm{v} = \begin{pmatrix} 1 \\ i \end{pmatrix}.$$

We have therefore determined two independent complex solutions to the ode, that is,

$$\mathrm{v}e^{\lambda t} \quad \text{and} \quad \overline{\mathrm{v}}e^{\bar{\lambda} t},$$

and we can form a linear combination of these two complex solutions to construct two independent real solutions. Namely, if the complex functions $z(t)$ and $\bar{z}(t)$ are written as

$$z(t) = \mathrm{Re}\{z(t)\} + i\mathrm{Im}\{z(t)\}, \quad \bar{z}(t) = \mathrm{Re}\{z(t)\} - i\mathrm{Im}\{z(t)\},$$

then two real functions can be constructed from the following linear combinations of z and $\bar{z}$:

$$\frac{z + \bar{z}}{2} = \mathrm{Re}\{z(t)\} \quad \text{and} \quad \frac{z - \bar{z}}{2i} = \mathrm{Im}\{z(t)\}.$$

Thus the two real vector functions that can be constructed from our two complex vector functions are

$$\begin{aligned} \mathrm{Re}\{\mathrm{v}e^{\lambda t}\} &= \mathrm{Re}\left\{ \begin{pmatrix} 1 \\ i \end{pmatrix} e^{(-\frac{1}{2}+i)t} \right\} \\ &= e^{-\frac{1}{2}t}\mathrm{Re}\left\{ \begin{pmatrix} 1 \\ i \end{pmatrix} (\cos t + i \sin t) \right\} \\ &= e^{-\frac{1}{2}t} \begin{pmatrix} \cos t \\ -\sin t \end{pmatrix}; \end{aligned}$$

and

$$\begin{aligned} \mathrm{Im}\{\mathrm{v}e^{\lambda t}\} &= e^{-\frac{1}{2}t}\mathrm{Im}\left\{ \begin{pmatrix} 1 \\ i \end{pmatrix} (\cos t + i \sin t) \right\} \\ &= e^{-\frac{1}{2}t} \begin{pmatrix} \sin t \\ \cos t \end{pmatrix}. \end{aligned}$$

Taking a linear superposition of these two real solutions yields the general solution to the ode, given by

$$\mathrm{x} = e^{-\frac{1}{2}t} \left[A \begin{pmatrix} \cos t \\ -\sin t \end{pmatrix} + B \begin{pmatrix} \sin t \\ \cos t \end{pmatrix} \right].$$

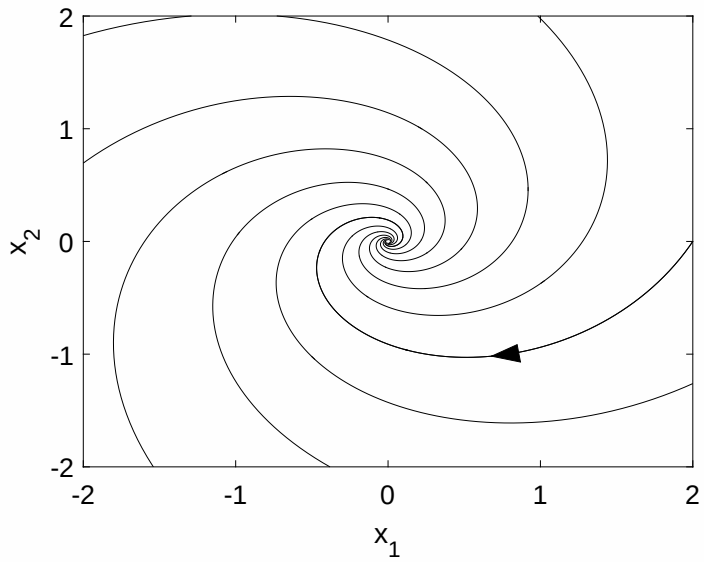


Figure 10.2: *Phase portrait for example with complex conjugate eigenvalues.*

The corresponding phase portrait is shown in Fig. 10.2. We say the origin is a *spiral point*. If the real part of the complex eigenvalue is negative, as is the case here, then solutions spiral into the origin. If the real part of the eigenvalue is positive, then solutions spiral out of the origin.

The direction of the spiral—here, it is clockwise—can be determined easily. Reconsider the original differential equations given by

$$\dot{x}_1 = -\frac{1}{2}x_1 + x_2, \qquad \dot{x}_2 = -x_1 - \frac{1}{2}x_2.$$

If we examine these equations with $x_1 = 0$ and $x_2 = 1$, we see that $\dot{x}_1 = 1$ and $\dot{x}_2 = -1/2$. The trajectory at the point $(0, 1)$ is moving to the right and downward, and this is possible only if the spiral is clockwise. A counterclockwise trajectory would be moving to the left and downward.

10.5 Repeated eigenvalues with one eigenvector

Example: Find the general solution of $\dot{x}_1 = x_1 - x_2$, $\dot{x}_2 = x_1 + 3x_2$.

View tutorial on YouTube

The equations in matrix form are

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{10.10}$$

The ansatz $\mathrm{x} = \mathrm{v}e^{\lambda t}$ leads to the characteristic equation

$$\begin{aligned} 0 &= \det(\mathbf{A} - \lambda\mathbf{I}) \\ &= \lambda^2 - 4\lambda + 4 \\ &= (\lambda - 2)^2. \end{aligned}$$

Therefore, $\lambda = 2$ is a repeated eigenvalue. The associated eigenvector is found from $-v_1 - v_2 = 0$, or $v_2 = -v_1$; and normalizing with $v_1 = 1$, we have

$$\lambda = 2, \quad \mathrm{v} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

We have thus found a single solution to the ode, given by

$$\mathrm{x}_1(t) = c_1 \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{2t},$$

and we need to find the missing second solution to be able to satisfy the initial conditions. An ansatz of t times the first solution is tempting, but will fail. Here, we will cheat and find the missing second solution by solving the equivalent second-order, homogeneous, constant-coefficient differential equation.

We already know that this second-order differential equation for $x_1(t)$ has a characteristic equation with a degenerate eigenvalue given by $\lambda = 2$. Therefore, the general solution for x_1 is given by

$$x_1(t) = (c_1 + tc_2)e^{2t}.$$

Since from the first differential equation, $x_2 = x_1 - \dot{x}_1$, we compute

$$\dot{x}_1 = \big(2c_1 + (1+2t)c_2\big)e^{2t},$$

so that

$$\begin{aligned} x_2 &= x_1 - \dot{x}_1 \\ &= (c_1 + tc_2)e^{2t} - \big(2c_1 + (1+2t)c_2\big)e^{2t} \\ &= -c_1 e^{2t} + c_2(-1-t)e^{2t}. \end{aligned}$$

Combining our results for x_1 and x_2, we have therefore found

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{2t} + c_2 \left[\begin{pmatrix} 0 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \end{pmatrix} t \right] e^{2t}.$$

Our missing linearly independent solution is thus determined to be

$$\mathrm{x}_2(t) = c_2 \left[\begin{pmatrix} 0 \\ -1 \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \end{pmatrix} t \right] e^{2t}. \tag{10.11}$$

The second term of (10.11) is just t times the first solution; however, this is not sufficient. Indeed, the correct ansatz to find the second solution directly is given by

$$\mathrm{x} = (\mathrm{w} + t\mathrm{v})\, e^{\lambda t}, \tag{10.12}$$

where λ and v are the eigenvalue and eigenvector of the first solution, and w is an unknown vector to be determined. To illustrate this direct method, we substitute (10.12) into $\dot{\mathrm{x}} = \mathrm{A}\mathrm{x}$, assuming $\mathrm{A}\mathrm{v} = \lambda \mathrm{v}$. Canceling the exponential, we obtain

$$\mathrm{v} + \lambda\,(\mathrm{w} + t\mathrm{v}) = \mathrm{A}\mathrm{w} + \lambda t \mathrm{v}.$$

Further canceling the common term $\lambda t \mathrm{v}$ and rewriting yields

$$(\mathrm{A} - \lambda \mathrm{I})\, \mathrm{w} = \mathrm{v}. \tag{10.13}$$

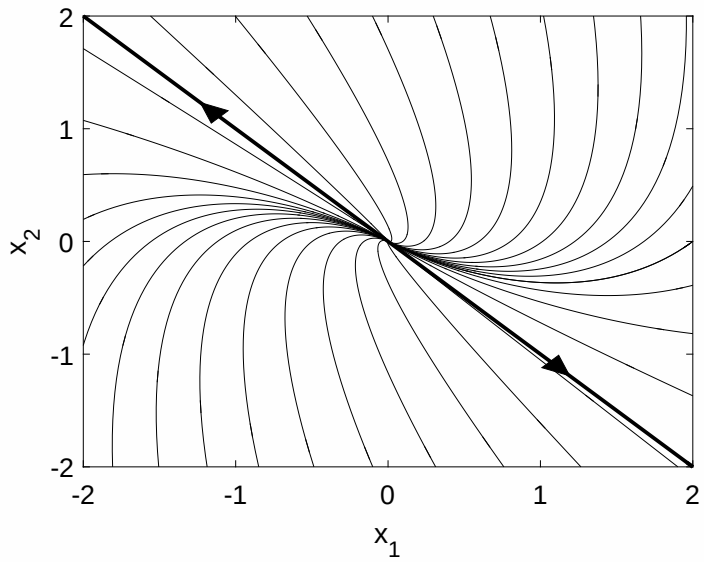


Figure 10.3: *Phase portrait for example with only one eigenvector.*

If A has only a single linearly independent eigenvector v, then (10.13) can be solved for w (otherwise, it cannot). Using A, λ and v of our present example, (10.13) is the system of equations given by

$$\begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

The first and second equation are the same, so that $w_2 = -(w_1 + 1)$. Therefore,

$$\begin{aligned} \mathrm{w} &= \begin{pmatrix} w_1 \\ -(w_1+1) \end{pmatrix} \\ &= w_1 \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} 0 \\ -1 \end{pmatrix}. \end{aligned}$$

Notice that the first term repeats the first found solution, i.e., a constant times the eigenvector, and the second term is new. We therefore take $w_1 = 0$ and obtain

$$\mathrm{w} = \begin{pmatrix} 0 \\ -1 \end{pmatrix},$$

as before.

The phase portrait for this ode is shown in Fig. 10.3. The dark line is the single eigenvector **v** of the matrix **A**. When there is only a single eigenvector, the origin is called an *improper node*.

10.6 Normal modes

View tutorials on YouTube: Part 1 Part 2

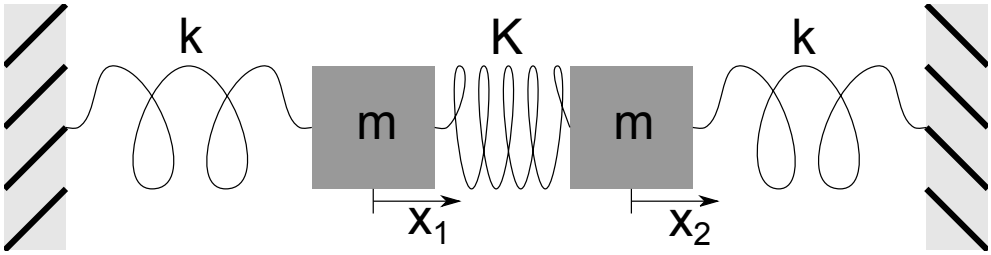


Figure 10.4: *Top view of a double mass, triple spring system.*

We now consider an application of the eigenvector analysis to the coupled mass-spring system shown in Fig. 10.4. The position variables x_1 and x_2 are measured from the equilibrium positions of the masses. Hooke's law states that the spring force is linearly proportional to the extension length of the spring, measured from equilibrium. By considering the extension of the spring and the sign of the force, we write Newton's law $F = ma$ separately for each mass:

$$\begin{aligned} m\ddot{x}_1 &= -kx_1 - K(x_1 - x_2), \\ m\ddot{x}_2 &= -kx_2 - K(x_2 - x_1). \end{aligned}$$

Further rewriting by collecting terms proportional to x_1 and x_2 yields

$$\begin{aligned} m\ddot{x}_1 &= -(k+K)x_1 + Kx_2, \\ m\ddot{x}_2 &= Kx_1 - (k+K)x_2. \end{aligned}$$

The equations for the coupled mass-spring system form a system of two second-order linear homogeneous odes. In matrix form, $m\ddot{\mathrm{x}} = \mathrm{Ax}$, or explicitly,

$$m\frac{d^2}{dt^2}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -(k+K) & K \\ K & -(k+K) \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{10.14}$$

In analogy to a system of first-order equations, we try the ansatz $\mathrm{x} = \mathrm{v}e^{rt}$, and upon substitution into (10.14) we obtain the eigenvalue problem $\mathrm{Av} = \lambda\mathrm{v}$, with $\lambda = mr^2$. The eigenvalues are determined by solving the characteristic equation

$$\begin{aligned} 0 &= \det(\mathrm{A} - \lambda\mathrm{I}) \\ &= \begin{vmatrix} -(k+K) - \lambda & K \\ K & -(k+K) - \lambda \end{vmatrix} \\ &= (\lambda + k + K)^2 - K^2. \end{aligned}$$

The solution for λ is

$$\lambda = -k - K \pm K,$$

and the two eigenvalues are

$$\lambda_1 = -k, \quad \lambda_2 = -(k+2K).$$

The corresponding values of r in our ansatz $\mathrm{x} = \mathrm{v}e^{rt}$, with $r = \pm\sqrt{\lambda/m}$, are

$$r_1 = i\sqrt{k/m}, \quad \bar{r}_1, \quad r_2 = i\sqrt{(k+2K)/m}, \quad \bar{r}_2.$$

Since the values of r are pure imaginary, we know that $x_1(t)$ and $x_2(t)$ will oscillate with angular frequencies $\omega_1 = \mathrm{Im}\{r_1\}$ and $\omega_2 = \mathrm{Im}\{r_2\}$, that is,

$$\omega_1 = \sqrt{k/m}, \quad \omega_2 = \sqrt{(k+2K)/m}.$$

The positions of the oscillating masses in general contain time dependencies of the form $\sin\omega_1 t$, $\cos\omega_1 t$, and $\sin\omega_2 t$, $\cos\omega_2 t$.

It is of further interest to determine the eigenvectors, or so-called *normal modes* of oscillation, associated with the two distinct angular frequencies. With specific initial conditions proportional to an eigenvector, the mass will oscillate with a single frequency. The eigenvector with eigenvalue λ_1 satisfies

$$-Kv_{11} + Kv_{12} = 0,$$

so that $v_{11} = v_{12}$. The normal mode with frequency $\omega_1 = \sqrt{k/m}$ thus follows a motion where $x_1 = x_2$. Referring to Fig. 10.4, during this motion the center spring length does not change, which is why the frequency of oscillation is independent of K.

Next, we determine the eigenvector with eigenvalue λ_2:

$$Kv_{21} + Kv_{22} = 0,$$

so that $v_{21} = -v_{22}$. The normal mode with frequency $\omega_2 = \sqrt{(k+2K)/m}$ thus follows a motion where $x_1 = -x_2$. Again referring to Fig. 10.4, during this motion the two equal masses symmetrically push or pull against each side of the middle spring. This two-sided push and pull results in the contribution of $2K$ to the frequency.

A general solution for $\mathbf{x}(t)$ can be constructed from the eigenvalues and eigenvectors. Our ansatz was $\mathrm{x} = \mathrm{v}e^{rt}$, and for each of two eigenvectors v, we have a pair of complex conjugate values for r. Accordingly, we first apply the principle of superposition to obtain four real solutions, and then apply the principle again to obtain the general solution. With $\omega_1 = \sqrt{k/m}$ and $\omega_2 = \sqrt{(k+2K)/m}$, the general solution is given by

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} (A\cos\omega_1 t + B\sin\omega_1 t) + \begin{pmatrix} 1 \\ -1 \end{pmatrix} (C\cos\omega_2 t + D\sin\omega_2 t),$$

where the now real constants A, B, C, and D can be determined from the four independent initial conditions, $x_1(0)$, $x_2(0)$, $\dot{x}_1(0)$, and $\dot{x}_2(0)$.

Chapter 11

Nonlinear differential equations

Reference: Strogatz, Sections 2.2, 2.4, 3.1, 3.2, 3.4, 6.3, 6.4, 8.2

We now turn our attention to nonlinear differential equations. In particular, we study how small changes in the parameters of a system can result in qualitative changes in the dynamics. These qualitative changes in the dynamics are called bifurcations. To understand bifurcations, we first need to understand the concepts of fixed points and stability.

11.1 Fixed points and stability

11.1.1 One dimension

View tutorial on YouTube

Consider the one-dimensional differential equation for $x = x(t)$ given by

$$\dot{x} = f(x). \tag{11.1}$$

We say that x_* is a *fixed point*, or *equilibrium point*, of (11.1) if $f(x_*) = 0$. At the fixed point, $\dot{x} = 0$. The terminology *fixed point* is used since the solution to (11.1) with initial condition $x(0) = x_*$ is $x(t) = x_*$ for all time t.

A fixed point, however, can be stable or unstable. A fixed point is said to be *stable* if a small perturbation of the solution from the fixed point decays in time; it is said to be *unstable* if a small perturbation grows in time. We can determine stability by a *linear analysis*. Let $x = x_* + \epsilon(t)$, where ϵ represents a small perturbation of the solution from the fixed point x_*. Because x_* is a constant, $\dot{x} = \dot{\epsilon}$; and because x_* is a fixed point, $f(x_*) = 0$. Taylor series expanding about $\epsilon = 0$, we have

$$\begin{aligned}\dot{\epsilon} &= f(x_* + \epsilon) \\ &= f(x_*) + \epsilon f'(x_*) + \dots \\ &= \epsilon f'(x_*) + \dots .\end{aligned}$$

The omitted terms in the Taylor series expansion are proportional to ϵ^2, and can be made negligible over a short time interval with respect to the kept term, proportional to ϵ, by taking $\epsilon(0)$ sufficiently small. Therefore, at least over short times, the differential equation to be considered, $\dot{\epsilon} = f'(x_*)\epsilon$, is linear and has by now the familiar solution

$$\epsilon(t) = \epsilon(0)e^{f'(x_*)t}.$$

The perturbation of the fixed point solution $x(t) = x_*$ thus decays exponentially if $f'(x_*) < 0$, and we say the fixed point is stable. If $f'(x_*) > 0$, the perturbation grows exponentially and we say the fixed point is unstable. If $f'(x_*) = 0$, we say the fixed point is marginally stable, and the next higher-order term in the Taylor series expansion must be considered.

Example: Find all the fixed points of the logistic equation $\dot{x} = x(1-x)$ and determine their stability.

There are two fixed points at which $\dot{x} = 0$, given by $x_* = 0$ and $x_* = 1$. Stability of these equilibrium points may be determined by considering the derivative of $f(x) = x(1-x)$. We have $f'(x) = 1 - 2x$. Therefore, $f'(0) = 1 > 0$ so that $x_* = 0$ is an unstable fixed point, and $f'(1) = -1 < 0$ so that $x_* = 1$ is a stable fixed point. Indeed, we have previously found that all solutions approach the stable fixed point asymptotically.

11.1.2 Two dimensions

View tutorial on YouTube

The idea of fixed points and stability can be extended to higher-order systems of odes. Here, we consider a two-dimensional system and will need to make use of the two-dimensional Taylor series expansion of a function $F(x,y)$ about the origin. In general, the Taylor series of $F(x,y)$ is given by

$$F(x,y) = F + x\frac{\partial F}{\partial x} + y\frac{\partial F}{\partial y} + \frac{1}{2}\left(x^2\frac{\partial^2 F}{\partial x^2} + 2xy\frac{\partial^2 F}{\partial x \partial y} + y^2\frac{\partial^2 F}{\partial y^2}\right) + \ldots,$$

where the function F and all of its partial derivatives on the right-hand-side are evaluated at the origin. Note that the Taylor series is constructed so that all partial derivatives of the left-hand-side match those of the right-hand-side at the origin.

We now consider the two-dimensional system given by

$$\dot{x} = f(x,y), \quad \dot{y} = g(x,y). \tag{11.2}$$

The point (x_*, y_*) is said to be a fixed point of (11.2) if $f(x_*, y_*) = 0$ and $g(x_*, y_*) = 0$. Again, the local stability of a fixed point can be determined by a linear analysis. We let $x(t) = x_* + \epsilon(t)$ and $y(t) = y_* + \delta(t)$, where ϵ and δ are small independent perturbations from the fixed point. Making use of the two dimensional Taylor series of $f(x,y)$ and $g(x,y)$ about the fixed point, or equivalently about $(\epsilon, \delta) = (0,0)$, we have

$$\begin{aligned}
\dot{\epsilon} &= f(x_* + \epsilon, y_* + \delta) \\
&= f + \epsilon\frac{\partial f}{\partial x} + \delta\frac{\partial f}{\partial y} + \ldots \\
&= \epsilon\frac{\partial f}{\partial x} + \delta\frac{\partial f}{\partial y} + \ldots. \\
\dot{\delta} &= g(x_* + \epsilon, y_* + \delta) \\
&= g + \epsilon\frac{\partial g}{\partial x} + \delta\frac{\partial g}{\partial y} + \ldots \\
&= \epsilon\frac{\partial g}{\partial x} + \delta\frac{\partial g}{\partial y} + \ldots,
\end{aligned}$$

where in the Taylor series f, g and all their partial derivatives are evaluated at the fixed point (x_*, y_*). Neglecting higher-order terms in the Taylor series, we thus have a system of odes for the perturbation, given in matrix form as

$$\frac{d}{dt}\begin{pmatrix} \epsilon \\ \delta \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{pmatrix}\begin{pmatrix} \epsilon \\ \delta \end{pmatrix}. \tag{11.3}$$

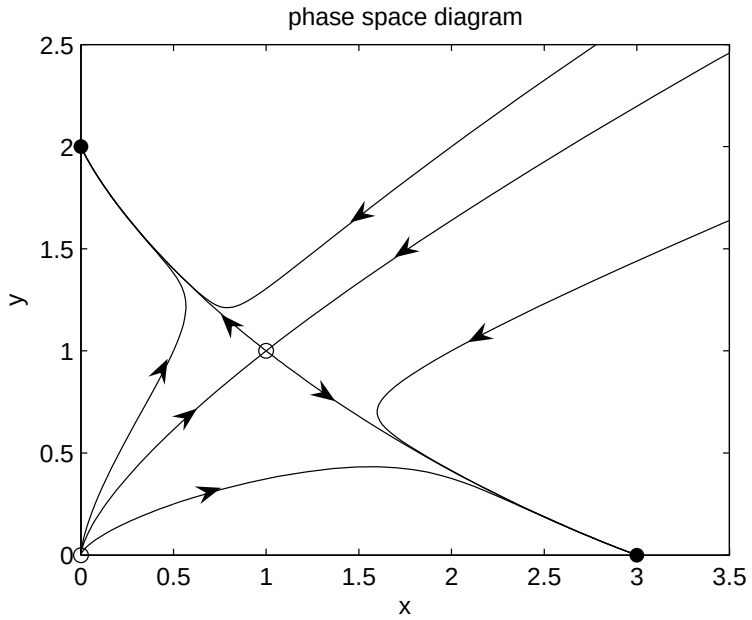


Figure 11.1: *Phase space plot for two-dimensional nonlinear system.*

The two-by-two matrix in (11.3) is called the Jacobian matrix at the fixed point. An eigenvalue analysis of the Jacobian matrix will typically yield two eigenvalues λ_1 and λ_2. These eigenvalues may be real and distinct, complex conjugate pairs, or repeated. The fixed point is stable (all perturbations decay exponentially) if both eigenvalues have negative real parts. The fixed point is unstable (some perturbations grow exponentially) if at least one of the eigenvalues has a positive real part. Fixed points can be further classified as stable or unstable nodes, unstable saddle points, stable or unstable spiral points, or stable or unstable improper nodes.

Example: Find all the fixed points of the nonlinear system $\dot{x} = x(3 - x - 2y)$, $\dot{y} = y(2 - x - y)$, and determine their stability.

View tutorial on YouTube

The fixed points are determined by solving

$$f(x,y) = x(3 - x - 2y) = 0, \quad g(x,y) = y(2 - x - y) = 0.$$

Evidently, $(x,y) = (0,0)$ is a fixed point. On the one hand, if only $x = 0$, then the equation $g(x,y) = 0$ yields $y = 2$. On the other hand, if only $y = 0$, then the equation $f(x,y) = 0$ yields $x = 3$. If both x and y are nonzero, then we must solve the linear system

$$x + 2y = 3, \quad x + y = 2,$$

and the solution is easily found to be $(x,y) = (1,1)$. Hence, we have determined the four fixed points $(x_*, y_*) = (0,0)$, $(0,2)$, $(3,0)$, $(1,1)$. The Jacobian matrix is given by

$$\begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{pmatrix} = \begin{pmatrix} 3 - 2x - 2y & -2x \\ -y & 2 - x - 2y \end{pmatrix}.$$

Stability of the fixed points may be considered in turn. With J_* the Jacobian matrix evaluated at the fixed point, we have

$$(x_*, y_*) = (0,0): \quad \mathrm{J}_* = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}.$$

The eigenvalues of J_* are $\lambda = 3, 2$ so that the fixed point $(0,0)$ is an unstable node. Next,

$$(x_*, y_*) = (0,2): \quad \mathrm{J}_* = \begin{pmatrix} -1 & 0 \\ -2 & -2 \end{pmatrix}.$$

The eigenvalues of J_* are $\lambda = -1, -2$ so that the fixed point $(0,2)$ is a stable node. Next,

$$(x_*, y_*) = (3,0): \quad \mathrm{J}_* = \begin{pmatrix} -3 & -6 \\ 0 & -1 \end{pmatrix}.$$

The eigenvalues of J_* are $\lambda = -3, -1$ so that the fixed point $(3,0)$ is also a stable node. Finally,

$$(x_*, y_*) = (1,1): \quad \mathrm{J}_* = \begin{pmatrix} -1 & -2 \\ -1 & -1 \end{pmatrix}.$$

The characteristic equation of J_* is given by $(-1-\lambda)^2 - 2 = 0$, so that $\lambda = -1 \pm \sqrt{2}$. Since one eigenvalue is negative and the other positive the fixed point $(1,1)$ is an unstable saddle point. From our analysis of the fixed points, one can expect that all solutions will asymptote to one of the stable fixed points $(0,2)$ or $(3,0)$, depending on the initial conditions.

It is of interest to sketch the phase space diagram for this nonlinear system. The eigenvectors associated with the unstable saddle point $(1,1)$ determine the directions of the flow into and away from this fixed point. The eigenvector associated with the positive eigenvalue $\lambda_1 = -1 + \sqrt{2}$ can be determined from the first equation of $(\mathrm{J}_* - \lambda_1 \mathrm{I})\mathrm{V}_1 = 0$, or

$$-\sqrt{2}v_{11} - 2v_{12} = 0,$$

so that $v_{12} = -(\sqrt{2}/2)v_{11}$. The eigenvector associated with the negative eigenvalue $\lambda_1 = -1 - \sqrt{2}$ satisfies $v_{22} = (\sqrt{2}/2)v_{21}$. The eigenvectors give the slope of the lines with origin at the fixed point for incoming (negative eigenvalue) and outgoing (positive eigenvalue) trajectories. The outgoing trajectories have negative slope $-\sqrt{2}/2$ and the incoming trajectories have positive slope $\sqrt{2}/2$. A rough sketch of the phase space diagram can be made by hand (as demonstrated in class). Here, a computer generated plot obtained from numerical solution of the nonlinear coupled odes is presented in Fig. 11.1. The curve starting from the origin and at infinity, and terminating at the unstable saddle point is called the separatrix. This curve separates the phase space into two regions: initial conditions for which the solution asymptotes to the fixed point $(0,2)$, and initial conditions for which the solution asymptotes to the fixed point $(3,0)$.

11.2 Bifurcation theory

A bifurcation occurs in a nonlinear differential equation when a small change in a parameter results in a qualitative change in the long-time solution. Examples of bifurcations are when fixed points are created or destroyed, or change their stability.

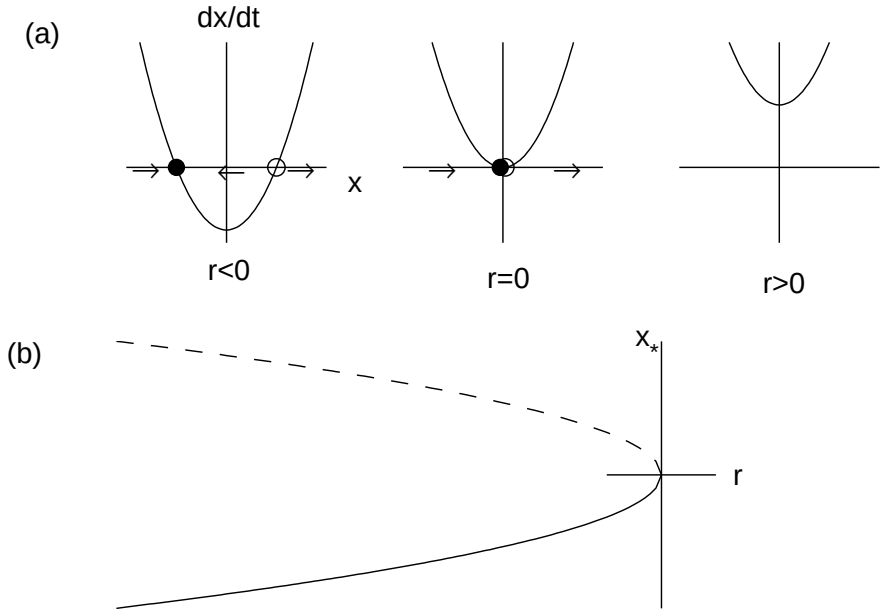


Figure 11.2: *Saddlenode bifurcation. (a) $\dot{x}$ versus x; (b) bifurcation diagram.*

We now consider four classic bifurcations of one-dimensional nonlinear differential equations: saddle-node bifurcation, transcritical bifurcation, supercritical pitchfork bifurcation, and subcritical pitchfork bifurcation. The corresponding differential equation will be written as

$$\dot{x} = f_r(x),$$

where the subscript r represents a parameter that results in a bifurcation when varied across zero. The simplest differential equations that exhibit these bifurcations are called the *normal forms*, and correspond to a local analysis (i.e., Taylor series expansion) of more general differential equations around the fixed point, together with a possible rescaling of x.

11.2.1 Saddle-node bifurcation

View tutorial on YouTube

The saddle-node bifurcation results in fixed points being created or destroyed. The normal form for a saddle-node bifurcation is given by

$$\dot{x} = r + x^2.$$

The fixed points are $x_* = \pm\sqrt{-r}$. Clearly, two real fixed points exist when $r < 0$ and no real fixed points exist when $r > 0$. The stability of the fixed points when $r < 0$ are determined by the derivative of $f(x) = r + x^2$, given by $f'(x) = 2x$. Therefore, the negative fixed point is stable and the positive fixed point is unstable.

Graphically, we can illustrate this bifurcation in two ways. First, in Fig. 11.2(a), we plot $\dot{x}$ versus x for the three parameter values corresponding to $r < 0$, $r = 0$ and

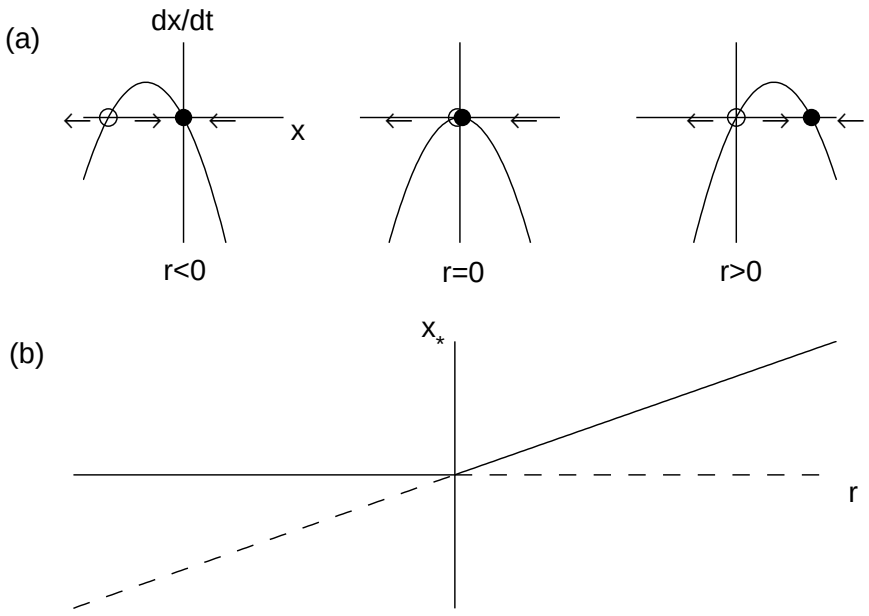


Figure 11.3: *Transcritical bifurcation. (a) $\dot{x}$ versus x; (b) bifurcation diagram.*

$r > 0$. The values at which $\dot{x} = 0$ correspond to the fixed points, and arrows are drawn indicating how the solution $x(t)$ evolves (to the right if $\dot{x} > 0$ and to the left if $\dot{x} < 0$). The stable fixed point is indicated by a filled circle and the unstable fixed point by an open circle. Note that when $r = 0$, solutions converge to the origin from the left, but diverge from the origin on the right. Second, in Fig. 11.2(b), we plot a bifurcation diagram illustrating the fixed point x_* versus the bifurcation parameter r. The stable fixed point is denoted by a solid line and the unstable fixed point by a dashed line. Note that the two fixed points collide and annihilate at $r = 0$, and there are no fixed points for $r > 0$.

11.2.2 Transcritical bifurcation

View tutorial on YouTube

A transcritical bifurcation occurs when there is an exchange of stabilities between two fixed points. The normal form for a transcritical bifurcation is given by

$$\dot{x} = rx - x^2.$$

The fixed points are $x_* = 0$ and $x_* = r$. The derivative of the right-hand-side is $f'(x) = r - 2x$, so that $f'(0) = r$ and $f'(r) = -r$. Therefore, for $r < 0$, $x_* = 0$ is stable and $x_* = r$ is unstable, while for $r > 0$, $x_* = r$ is stable and $x_* = 0$ is unstable. The two fixed points thus exchange stability as r passes through zero. The transcritical bifurcation is illustrated in Fig. 11.3.

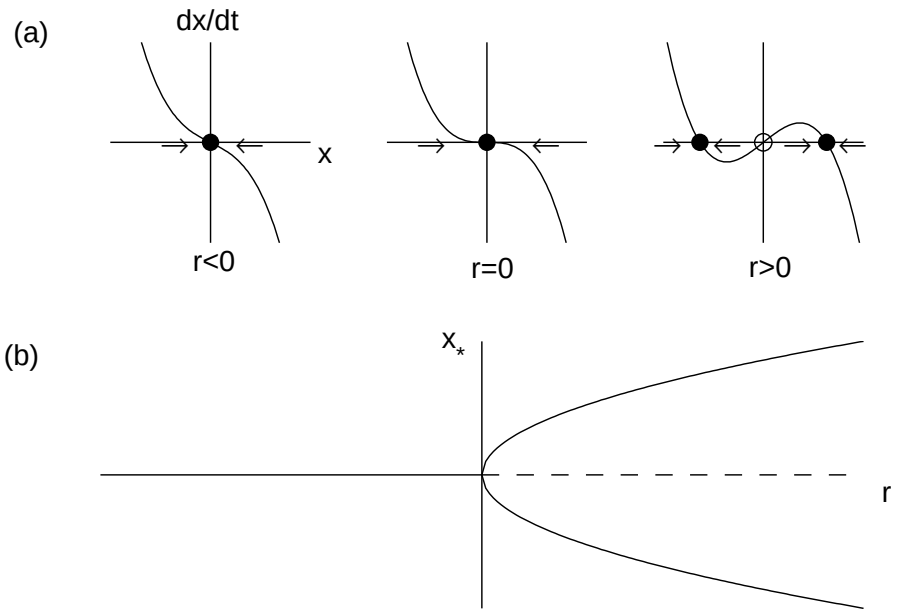


Figure 11.4: *Supercritical pitchfork bifurcation. (a) $\dot{x}$ versus x; (b) bifurcation diagram.*

11.2.3 Supercritical pitchfork bifurcation

View tutorial on YouTube

The pitchfork bifurcations occur in physical models where fixed points appear and disappear in pairs due to some intrinsic symmetry of the problem. Pitchfork bifurcations can come in one of two types. In the supercritical bifurcation, a pair of stable fixed points are created at the bifurcation (or critical) point and exist after (super) the bifurcation. In the subcritical bifurcation, a pair of unstable fixed points are created at the bifurcation point and exist before (sub) the bifurcation.

The normal form for the supercritical pitchfork bifurcation is given by

$$\dot{x} = rx - x^3.$$

Note that the linear term results in exponential growth when $r > 0$ and the nonlinear term stabilizes this growth. The fixed points are $x_* = 0$ and $x_* = \pm\sqrt{r}$, the latter fixed points existing only when $r > 0$. The derivative of f is $f'(x) = r - 3x^2$ so that $f'(0) = r$ and $f'(\pm\sqrt{r}) = -2r$. Therefore, the fixed point $x_* = 0$ is stable for $r < 0$ and unstable for $r > 0$ while the fixed points $x = \pm\sqrt{r}$ exist and are stable for $r > 0$. Notice that the fixed point $x_* = 0$ becomes unstable as r crosses zero and two new stable fixed points $x_* = \pm\sqrt{r}$ are born. The supercritical pitchfork bifurcation is illustrated in Fig. 11.4.

11.2.4 Subcritical pitchfork bifurcation

View tutorial on YouTube

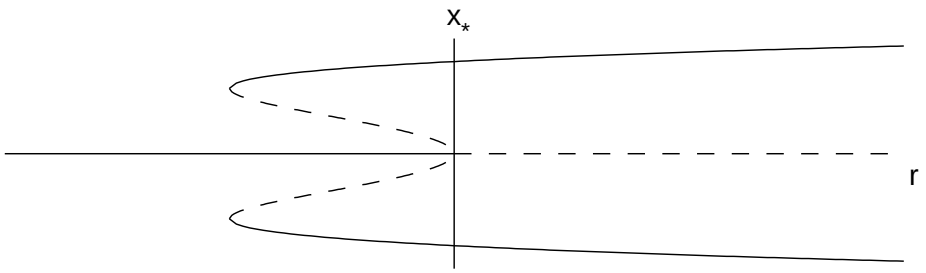


Figure 11.5: *Subcritical pitchfork bifurcation.*

In the subcritical case, the cubic term is destabilizing. The normal form (to order x^3) is

$$\dot{x} = rx + x^3.$$

The fixed points are $x_* = 0$ and $x_* = \pm\sqrt{-r}$, the latter fixed points existing only when $r \leq 0$. The derivative of the right-hand-side is $f'(x) = r + 3x^2$ so that $f'(0) = r$ and $f'(\pm\sqrt{-r}) = -2r$. Therefore, the fixed point $x_* = 0$ is stable for $r < 0$ and unstable for $r > 0$ while the fixed points $x = \pm\sqrt{-r}$ exist and are unstable for $r < 0$. There are no stable fixed points when $r > 0$.

The absence of stable fixed points for $r > 0$ indicates that the neglect of terms of higher-order in x than x^3 in the normal form may be unwarranted. Keeping to the intrinsic symmetry of the equations (only odd powers of x) we can add a stabilizing nonlinear term proportional to x^5. The extended normal form (to order x^5) is

$$\dot{x} = rx + x^3 - x^5,$$

and is somewhat more difficult to analyze. The fixed points are solutions of

$$x(r + x^2 - x^4) = 0.$$

The fixed point $x_* = 0$ exists for all r, and four additional fixed points can be found from the solutions of the quadratic equation in x^2:

$$x_* = \pm\sqrt{\frac{1}{2}(1 \pm \sqrt{1 + 4r})}.$$

These fixed points exist only if x_* is real. Clearly, for the inner square-root to be real, $r \geq -1/4$. Also observe that $1 - \sqrt{1 + 4r}$ becomes negative for $r > 0$. We thus have three intervals in r to consider, and these regions and their fixed points are

$$\begin{aligned}
r < -\frac{1}{4}: &\quad x_* = 0 \quad \text{(one fixed point)};\\
-\frac{1}{4} < r < 0: &\quad x_* = 0, \quad x_* = \pm\sqrt{\frac{1}{2}(1 \pm \sqrt{1 + 4r})} \quad \text{(five fixed points)};\\
r > 0: &\quad x_* = 0, \quad x_* = \pm\sqrt{\frac{1}{2}(1 + \sqrt{1 + 4r})} \quad \text{(three fixed points)}.
\end{aligned}$$

Stability is determined from $f'(x) = r + 3x^2 - 5x^4$. Now, $f'(0) = r$ so $x_* = 0$ is stable for $r < 0$ and unstable for $r > 0$. The calculation for the other four roots can

be simplified by noting that x_* satisfies $r + x_*^2 - x_*^4 = 0$, or $x_*^4 = r + x_*^2$. Therefore,

$$\begin{aligned} f'(x_*) &= r + 3x_*^2 - 5x_*^4 \\ &= r + 3x_*^2 - 5(r + x_*^2) \\ &= -4r - 2x_*^2 \\ &= -2(2r + x_*^2). \end{aligned}$$

With $x_*^2 = \frac{1}{2}(1 \pm \sqrt{1+4r})$, we have

$$\begin{aligned} f'(x_*) &= -2\left(2r + \frac{1}{2}(1 \pm \sqrt{1+4r})\right) \\ &= -\left((1+4r) \pm \sqrt{1+4r}\right) \\ &= -\sqrt{1+4r}\left(\sqrt{1+4r} \pm 1\right). \end{aligned}$$

Clearly, the plus root is always stable since $f'(x_*) < 0$. The minus root exists only for $-\frac{1}{4} < r < 0$ and is unstable since $f'(x_*) > 0$. We summarize the stability of the various fixed points:

$$\begin{aligned} r < -\frac{1}{4}: &\quad x_* = 0 \quad \text{(stable)}; \\ -\frac{1}{4} < r < 0: &\quad x_* = 0, \quad \text{(stable)} \\ &\quad x_* = \pm\sqrt{\frac{1}{2}(1+\sqrt{1+4r})} \quad \text{(stable)}; \\ &\quad x_* = \pm\sqrt{\frac{1}{2}(1-\sqrt{1+4r})} \quad \text{(unstable)}; \\ r > 0: &\quad x_* = 0 \quad \text{(unstable)} \\ &\quad x_* = \pm\sqrt{\frac{1}{2}(1+\sqrt{1+4r})} \qquad \text{(stable)}. \end{aligned}$$

The bifurcation diagram is shown in Fig. 11.5, and consists of a subcritical pitchfork bifurcation at $r = 0$ and two saddle-node bifurcations at $r = -1/4$. We can imagine what happens to x as r increases from negative values, supposing there is some small noise in the system so that $x = x(t)$ will diverge from unstable fixed points. For $r < -1/4$, the equilibrium value of x is $x_* = 0$. As r increases into the range $-1/4 < r < 0$, x will remain at $x_* = 0$. However, a catastrophe occurs as soon as $r > 0$. The $x_* = 0$ fixed point becomes unstable and the solution will jump up (or down) to the only remaining stable fixed point. Such behavior is called a jump bifurcation. A similar catastrophe can happen as r decreases from positive values. In this case, the jump occurs as soon as $r < -1/4$.

Since the stable equilibrium value of x depends on whether we are increasing or decreasing r, we say that the system exhibits *hysteresis*. The existence of a subcritical pitchfork bifurcation can be very dangerous in engineering applications since a small change in a problem's parameters can result in a large change in the equilibrium state. Physically, this can correspond to a collapse of a structure, or the failure of a component.

11.2.5 Application: a mathematical model of a fishery

View tutorial on YouTube

We illustrate the utility of bifurcation theory by analyzing a simple model of a fishery. We utilize the logistic equation (see §7.4.6) to model a fish population in the absence of fishing. To model fishing, we assume that the government has established fishing quotas so that at most a total of C fish per year may be caught. We assume that when the fish population is at the carrying capacity of the environment, fisherman can catch nearly their full quota. When the fish population drops to lower values, fish may be harder to find and the catch rate may fall below C, eventually going to zero as the fish population diminishes. Combining the logistic equation together with a simple model of fishing, we propose the mathematical model

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right) - \frac{CN}{A+N}, \tag{11.4}$$

where N is the fish population size, t is time, r is the maximum potential growth rate of the fish population, K is the carrying capacity of the environment, C is the maximum rate at which fish can be caught, and A is a constant satisfying $A < K$ that is used to model the idea that fish become harder to catch when scarce.

We nondimensionalize (11.4) using $x = N/K$, $\tau = rt$, $c = C/rK$, $\alpha = A/K$:

$$\frac{dx}{d\tau} = x(1-x) - \frac{cx}{\alpha + x}. \tag{11.5}$$

Note that $0 \leq x \leq 1$, $c > 0$ and $0 < \alpha < 1$. We wish to qualitatively describe the equilibrium solutions of (11.5) and the bifurcations that may occur as the nondimensional catch rate c increases from zero. Practically, a government would like to issue each year as large a catch quota as possible without adversely affecting the number of fish that may be caught in subsequent years.

The fixed points of (11.5) are $x_* = 0$, valid for all c, and the solutions to $x^2 - (1-\alpha)x + (c-\alpha) = 0$, or

$$x_* = \frac{1}{2}\left[(1-\alpha) \pm \sqrt{(1+\alpha)^2 - 4c}\right]. \tag{11.6}$$

The fixed points given by (11.6) are real only if $c < \frac{1}{4}(1+\alpha)^2$. Furthermore, the minus root is greater than zero only if $c > \alpha$. We therefore need to consider three intervals over which the following fixed points exist:

$$\begin{aligned}
0 \leq c \leq \alpha: &\quad x_* = 0, \quad x_* = \frac{1}{2}\left[(1-\alpha) + \sqrt{(1+\alpha)^2 - 4c}\right];\\
\alpha < c < \frac{1}{4}(1+\alpha)^2: &\quad x_* = 0, \quad x_* = \frac{1}{2}\left[(1-\alpha) \pm \sqrt{(1+\alpha)^2 - 4c}\right];\\
c > \frac{1}{4}(1+\alpha)^2: &\quad x_* = 0.
\end{aligned}$$

The stability of the fixed points can be determined with rigor analytically or graphically. Here, we simply apply biological intuition together with knowledge of the types of one dimensional bifurcations. An intuitive argument is made simpler if we consider c decreasing from large values. When c is large, that is $c > \frac{1}{4}(1+\alpha)^2$, too many fish are being caught and our intuition suggests that the fish population goes extinct. Therefore, in this interval, the single fixed point $x_* = 0$ must be stable. As

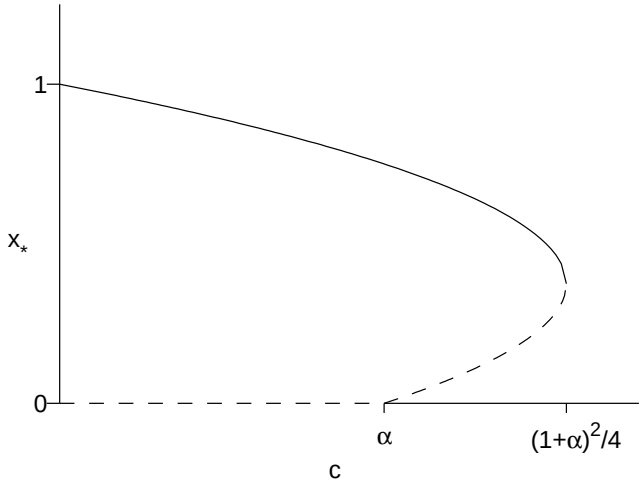


Figure 11.6: *Fishery bifurcation diagram.*

c decreases, a bifurcation occurs at $c = \frac{1}{4}(1+\alpha)^2$ introducing two additional fixed points at $x_* = (1-\alpha)/2$. The type of one dimensional bifurcation in which two fixed points are created as a square root becomes real is a saddlenode bifurcation, and one of the fixed points will be stable and the other unstable. Following these fixed points as $c \to 0$, we observe that the plus root goes to one, which is the appropriate stable fixed point when there is no fishing. We therefore conclude that the plus root is stable and the minus root is unstable. As c decreases further from this bifurcation, the minus root collides with the fixed point $x_* = 0$ at $c = \alpha$. This appears to be a transcritical bifurcation and assuming an exchange of stability occurs, we must have the fixed point $x_* = 0$ becoming unstable for $c < \alpha$. The resulting bifurcation diagram is shown in Fig. 11.6.

The purpose of simple mathematical models applied to complex ecological problems is to offer some insight. Here, we have learned that overfishing (in the model $c > \frac{1}{4}(1+\alpha)^2$) during one year can potentially result in a sudden collapse of the fish catch in subsequent years, so that governments need to be particularly cautious when contemplating increases in fishing quotas.