V. REGELSKIS

7PAM2003-0901

DATA SCIENCE CORE SKILLS: MATHEMATICS

SEMESTER B, 2021-22

DEPARTMENT OF PHYSICS, ASTRONOMY AND MATHEMATICS

Contents

1
Calculus

Lecture 1

Additional Reading:

- ***J. Feldman, A. Rechnitzer***, *Differential Calculus Notes*
- ***S. K. Chung***, *Understanding Basic Calculus*

1.1 Functions

1.1.1 Numbers, sets, and intervals

We start with a very basic concept in mathematics that you all must be familiar with:

> *A **set** is a collection of distinct objects. The objects are referred to as **elements** or **members** of the set. They can be anything: numbers, people, letters of the alphabet, other sets, and so on.*

Sets are also present in Python, e.g.

```
myset = {"a", "b", "c"}
```

It is common to use capital letters to denote sets, A, B, C, X, Y, etc., and to use lower case letters to denote their elements, a, b, c, x, y, etc. Below we list some important sets of numbers that we will dealing with in this course:

- ***Natural numbers*** are the "whole numbers" $1, 2, 3, \ldots$ that we learn first at about the same time as we learn the alphabet. We will denote this set of numbers by the symbol "$\mathbb{N}$".

  In Python you will meet `Integer` and `Float` numeric data types that are used to store Integer and Real numbers.

  The $\mathbb{N}$ is closed under addition and multiplication. This means that if you take any two natural numbers and add them you get another natural number. Similarly if you take any two natural numbers and multiply them you get another natural number. However $\mathbb{N}$ is not closed under subtraction or division; we need negative numbers and fractions to make collections of numbers closed under subtraction and division.

- ***Integers*** are all positive and negative numbers together with the zero. We denote the set of all integers by the symbol "$\mathbb{Z}$".

  The $\mathbb{Z}$ stands for the German Zahlen meaning numbers.

  The $\mathbb{Z}$ is closed under addition, subtraction and multiplication, but not division.

- ***Rational numbers*** are all numbers that can be written as the ratio of two integers. That is, any rational number r can be written as p/q where p, q are integers. We denote this set by the symbol "$\mathbb{Q}$".

  The Q stands for Italian quoziente meaning quotient or ratio.

  Now we finally have a set of numbers which is closed under addition, subtraction, multiplication and division (of course you still need to be careful not to divide by zero).

- ***Real numbers*** – generally we think of these numbers as numbers that can be written as decimal expansions and we denote it by "$\mathbb{R}$". It is beyond the scope of these lectures to go into the details of how to give a precise definition of real numbers, and the notion that a real number can be written as a decimal expansion will be sufficient.

  Infinity is not a real number because it is not compatible with the usual rules of arithmetic; for example

  $$\infty + 1 = \infty \implies 0 = 1$$

  $$2 \times \infty = \infty \implies 2 = 1$$

  which do not make sense.

Now that we have defined sets, we can ask "if an object is in the set?"; the answer can be "yes" or "no". We write this as $a \in A$ or $a \notin A$. For example, $4 \in \mathbb{N}$ and $-2 \notin \mathbb{N}$. Here the symbol $\in$ is a mathematical shorthand for "is an element of", and $\notin$ is a shorthand for "is not an element of".

We have already seen a few important sets – $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. However, arguably the most important set in mathematics is the empty set.

Inspection in Python:

```
A = {1, 2, 3}
# Returns True
1 in A
# Returns False
5 in A
```

> *The **empty set** (or null set or void set) is the set which contains no elements. It is denoted $\emptyset$. For any object x, we always have $x \notin \emptyset$; hence $\emptyset = \{\}$.*

Note that the empty set is not nothing; think of it as an empty bag.

The empty set in Python:

```
emptyset = { }
```

When a set does not contain too many elements it is fine to specify it by listing out its elements. But for infinite sets or even just big sets we can't do this and instead we have to give the defining rule. For example the set of all perfect square numbers we write as

$$S = \{x : x = k^2,\ k \in \mathbb{Z}\}$$

Here ":" (also written as "|" or "*s.t.*") stands for "such that".

Example 1.1: Sets

- Some finite sets: $A = \{2,3,5,7,11,13,17,19\}$ and $B = \{a \in A : a < 8\} = \{2,3,5,7\}$
- Natural numbers greater than 10: $\mathbb{N}_{>10} = \{n \in \mathbb{N} : n > 10\} = \{11,12,13,\ldots\}$
- Even integers: $E = \{n : n = 2k,\ k \in \mathbb{Z}\} = \{2n : n \in \mathbb{Z}\}$
- Odd integers: $O = \{n : n = 2k+1,\ k \in \mathbb{Z}\} = \{2n+1 : n \in \mathbb{Z}\}$

The sets A and B in the above example illustrate an important point. Every element in B is an element in A, and so we say that B is a subset of A.

Subsets in Python:

```
A = {1, 2, 3}
B = {1, 2, 3, 4, 5}
# Returns True
A.issubset(B)
# Returns False
B.issubset(A)
```

> *Let A and B be sets. We say **A is a subset of B** if every element of A is also an element of B. We denote this $A \subseteq B$ (or $B \supseteq A$).*
>
> *If A is a subset of B and A and B are not the same, so that there is some $b \in B$ such that $b \notin A$, then we say that **A is a proper subset of B**. We denote this by $A \subset B$ (or $B \supset A$).*

Two important things to note about subsets are:

- if A is a set, then $\emptyset \subset A$, and
- if a set A is not a subset of B, we write $A \not\subseteq B$ meaning that there is some $a \in A$ such that $a \notin B$.

In much of our work with functions later in the course we will need to work with subsets of real numbers, particularly segments of the "real line". A convenient and standard way of representing such subsets is with interval notation.

Let $a, b \in \mathbb{R}$ such that $a < b$. We name the subset of all numbers between a and b in different ways depending on whether or not the ends of the interval (a and b) are elements of the subset.

- *The **closed interval** $[a,b] = \{x \in \mathbb{R} : a \leq x \leq b\}$ – both end points are included.*
- *The **open interval** $(a,b) = \{x \in \mathbb{R} : a < x < b\}$ – neither end point is included.*

*We also define **half-open intervals** which contain one end point but not the other:*

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

*We sometimes also need **unbounded intervals***

$$[a,\infty) = \{x \in \mathbb{R} : a \leq x\} \qquad (a,\infty) = \{x \in \mathbb{R} : a < x\}$$

$$(-\infty,b] = \{x \in \mathbb{R} : x \leq b\} \qquad (-\infty,b) = \{x \in \mathbb{R} : x < b\}$$

These unbounded intervals do not include ∞, so that the end of an interval is always open.

We now know how to say that one set is contained within another. Next we will define some other operations on sets.

*Let A and B be sets. We define the **union of A and B**, denoted $A \cup B$, to be the set of all elements that are in at least one of A or B,*

$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$

*We define the **intersection of A and B**, denoted $A \cap B$, to be the set of elements that belong to both A and B,*

$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$

Here we are using the word "or" in a mathematical sense. We mean that x belongs to A or x belongs to B or both. Whereas in normal everyday English "or" is often used to be "exclusive or" – A or B but not both.

Union and intersection in Python:

```
A = {1, 2, 3}
B = {3, 4, 5}

# Returns {1, 2, 3, 4, 5}
A.union(B)

# Returns {3}
A.intersection(B)
```

Example 1.2: Subsets, unions, intersections

- Let $S = \{1,2\}$. What are all the subsets of S? Each element of S can either be in the subset or not (independent of the other elements of the set). So we have $2^2 = 4$ possibilities:

  $$\emptyset, \{1\}, \{2\}, \{1,2\} \subseteq S$$

  This can be generalised to show that a set that contains exactly n elements has exactly 2^n subsets. It is called the *power set* of S, and is usually denoted by $P(S)$.

- Let $A = \{1,2,3,4\}$, $B = \{p : p \text{ is prime}\}$, $C = \{5,7,9\}$ and $D = \{\text{even positive integers}\}$. Then

  $$A \cap B = \{2,3\} \qquad B \cap D = \{2\}$$
  $$A \cup C = \{1,2,3,4,5,7,9\} \qquad A \cap C = \emptyset$$

  In this last case we see that the two sets have no elements in common – they are said to be *disjoint*.

1.1.2 *Functions, formulae, and domains*

Functions encapsulate mathematical transformations. We put some kind of mathematical object into the function and another one comes out. For example, there is a function called sin; if we put $\pi/2$ in, 1 comes out. We write this as $\sin(\pi/2) = 1$.

A formula describes a sequence of operations or rules that are to be applied to some mathematical object; for example, x^2 is a formula.

Formulae and functions are closely related, but are not the same thing! To turn a formula into a function, we have to add two extra pieces of information. The first thing we have do is to specify what the variable (x, in the x^2 example) is allowed to be. Are we thinking about squaring integers, real numbers, matrices? We do this by specifying a set called the ***domain*** of the function, which contains every object we might want to feed into the function. If we want to think about squaring real numbers, we set the domain be $\mathbb{R}$. If we want to think about squaring 3×3 real matrices, we set the domain to be the set of all real 3×3 matrices, $\mathbb{R}^{3\times 3}$.

We will talk about matrices in the second part of this course.

The second (and, for our purposes, less important) piece of a function is called the ***codomain***. It specifies what kind of objects come out of the function. Unlike the domain, it is not an exact description of every value that actually comes out of the function; rather, it is a container large enough to hold every possible value. For example, if our formula is x^2 and our domain is $\mathbb{R}$, we could let the codomain be $\mathbb{R}$, or $[0, \infty)$ - but we could not let it be $\mathbb{Z}$, as squaring a real number does not necessarily give an integer. These three pieces of information together give us a true function.

The ***range*** or ***image*** of the function is the exact set of values that it takes on: the set of all $f(x)$ as x varies through the domain of f. We can write

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = x^2$$

or

$$f : \mathbb{R} \to \mathbb{R}, \quad f : x \mapsto x^2$$

to denote the function mapping $\mathbb{R}$ to $\mathbb{R}$ by squaring. We could also write

$$f : \mathbb{R} \to [0, \infty), \quad f(x) = x^2$$

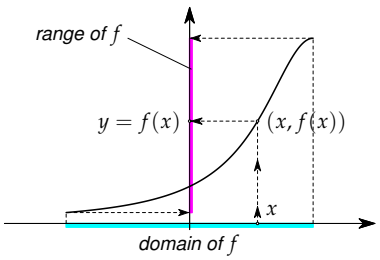
which is a slightly different function. It has the same and domain and range, but a different codomain. Once we have created a function, the letters we use do not matter:

$$f : \mathbb{R} \to \mathbb{R}, \quad f : x \mapsto x^2$$

and

$$f : \mathbb{R} \to \mathbb{R}, \quad f : y \mapsto y^2$$

are exactly the same. A formula like x^2 is about squaring something called x. The function we have defined is about squaring a real number.



In Computer Science, this dissociation of the function from the variable is known as λ-calculus.

Functions also go beyond formulae: all we need to define a function is an unambiguous rule which, given an element of the domain, describes ***exactly one*** element of the codomain. This does not have to be a simple algebraic formula; for example

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

with domain and codomain $\mathbb{R}$ describes the real absolute value function. Similarly,

$$H(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

with domain and codomain $\mathbb{R}$ is perfectly valid; this function is known as the Heaviside function (after Oliver Heaviside, who realised how useful it could be).

To get from a function to a formula, we need to specify a domain and a codomain. In Calculus, we often do this in the simplest way we can, as follows:

- Let the codomain be $\mathbb{R}$.
- Let the domain be the set of all real numbers for which the formula makes sense. This is called the ***maximal domain*** or ***natural domain*** (within $\mathbb{R}$) associated with the formula.

To find the maximal domain, look out for such things as:

- division by zero,
- square roots of negative numbers,
- logarithms of non-positive numbers.

This is very important in programming: you must make sure you always call functions with the right arguments to avoid abnormal behaviour.

If you have not seen logarithms before, we will discuss them later in this lecture.

Example 1.3: Functions, maximal domains, and ranges

Below we list a few functions together with their maximal domains and ranges. The codomain in each case can be taken to be the whole real line, $\mathbb{R}$. To find the range we need the concept of a limit which will be introduced later.

$f_1(x) = x^2 + 1$	$\mathbb{R}$	$[1, \infty)$
$f_2(y) = \dfrac{1}{y^2+1}$	$\mathbb{R}$	$(0, \infty)$
$f_3(z) = \dfrac{1}{z}$	$\mathbb{R} \setminus \{0\} = (-\infty, 0) \cup (0, \infty)$	$\mathbb{R} \setminus \{0\}$
$f_4(u) = \dfrac{u}{u^2-1}$	$\mathbb{R} \setminus \{-1, 1\} = (-\infty, -1) \cup (-1, 1) \cup (1, \infty)$	$\mathbb{R}$
$f_5(v) = \log(1+v)$	$(-1, \infty)$	$\mathbb{R}$
$f_6(w) = \sqrt{1-w}$	$(-\infty, 1]$	$[0, \infty)$

The logarithm in f_5 will be introduced properly later: it means the real-valued natural logarithm defined on $(0, \infty)$.

1.1.3 *Elementary functions*

Power function. Let $n, m \in \mathbb{N}$ and $x \in \mathbb{R}$.The power function is

$$f(x) = x^n = \underbrace{x \cdot x \cdots x}_{n \text{ times}}$$

Note that $x^0 = 1$. The power function satisfies

$$x^n x^m = x^{n+m} \qquad x^n / x^m = x^{n-m} \qquad (x^n)^m = x^{n \cdot m}$$

It also makes sense to have negative and rational powers,

$$x^{-n} = \frac{1}{x^n} \qquad x^{n/m} = (\sqrt[m]{x})^n$$

where $\sqrt[m]{x}$ is the number $y \in \mathbb{R}_{>0}$, such that $y^m = x$. It is called the m-th root of x.

Polynomial function. Let $n \in \mathbb{N}$ and $a_0, \ldots a_n \in \mathbb{R}$ such that $a_n \neq 0$. A polynomial of degree n is the function

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$$

Rational function. Let P and Q be polynomials. Their quotient is called a rational function,

$$R(x) = \frac{P(x)}{Q(x)}$$

Exponential function. Let $x \in \mathbb{R}$. The exponential function is

$$f(x) = \exp(x) = e^x$$

where $e = 2.71828182\ldots$ is a mathematical (Euler's) constant. It can be defined as the sum of the infinite series

$$e = 1 + \frac{1}{1} + \frac{1}{1 \cdot 2} + \frac{1}{1 \cdot 2 \cdot 3} + \ldots$$

Logarithmic function. Let $a, x \in \mathbb{R}_{>0}$ and $a \neq 1$. The logarithmic function to the base a, denoted $\log_a(x)$, is the function that gives the power to which a must be raised to get x, that is

$$a^{\log_a(x)} = x \tag{1.1}$$

Replacing x with a^x gives $a^{\log_a(a^x)} = a^x$, that is

$$\log_a(a^x) = x \tag{1.2}$$

The following identities can be derived from the ones above:

$$\log_a(x) = \frac{\log_b(x)}{\log_b(a)} \qquad \log_a(xy) = \log_a(x) + \log_a(y)$$
$$\log_a(x^y) = y \log_a(x) \qquad \log_a(x/y) = \log_a(x) - \log_a(y)$$

The natural logarithmic function. Let $x \in \mathbb{R}_{>0}$. The natural logarithmic function is defined as the logarithmic function to the base e,

$$\ln(x) = \log_e(x)$$

It is often also written as $\log(x)$. The identities (1.1) and (1.2) become

$$e^{\ln(x)} = x \qquad \ln(e^x) = x$$

Elementary functions in Python:

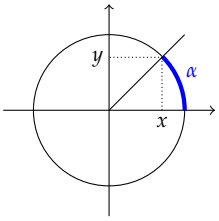
```
import math

# x to the y power
x**y

# exp and natural log function
math.exp(x)
math.log(x)

# trigonometric functions
math.sin(x)
math.cos(x)
math.tan(x)

# hyperbolic functions
math.sinh(x)
math.cosh(x)
math.tanh(x)
```

If $x \in \mathbb{Q}$ it makes sense to think of e^x as of a power function. If $x \in \mathbb{R}$ is not a rational number, then one can find the value of e^x by summing the infinite series

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{2 \cdot 3} + \frac{x^4}{2 \cdot 3 \cdot 4} + \ldots$$

We will revisit the exp function in Lecture 3.

Trigonometric functions. Consider a unit circle around the origin in a plane. Choose a point on the circle with coordinates (x, y). Then consider a half-line with its one end-point in the origin. Choose it initially to lie along the positive x-axis, and rotate it anti-clockwise around the origin until it reaches the (x, y) point. Then define the "sine", "cosine" and "tangent" functions by requiring



$$\sin(\alpha) = y \qquad \cos(\alpha) = x \qquad \tan(\alpha) = \frac{\sin(\alpha)}{\cos(\alpha)}$$

These are periodic functions, e.g. $\sin(\alpha + 2\pi n) = \sin(\alpha)$ for $n \in \mathbb{Z}$. Moreover, they satisfy a number of "useful" identities, for instance,

2π is the circumference of the unit circle. Thus adding $2\pi n$ corresponds to a full n-times rotation anti-clockwise.

$$\cos^2(\alpha) + \sin^2(\alpha) = 1$$
$$\sin(2\alpha) = 2\sin(\alpha)\cos(\alpha) \qquad \cos(2\alpha) = \cos^2(\alpha) - \sin^2(\alpha)$$

Search Google or Wikipedia for more "trigonometric identities"

1.1.4 Inverse functions

The last thing that we should review before diving into the main material of the course is the inverse functions. As we have seen above functions are really just rules for taking an input (almost always a number), processing it somehow (usually by a formula) and then returning an output (again, almost always a number):

$$\text{input number } x \;\rightarrow\; f \text{ does "stuff" to } x \;\rightarrow\; \text{return number } f(x)$$

In many situations it will turn out to be very useful if we can undo whatever it is that our function has done, that is

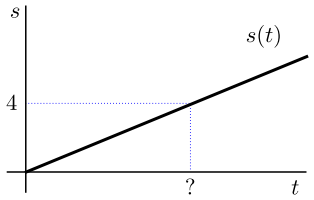
$$\text{take the } f(x) \;\rightarrow\; \text{do "stuff" to } f(x) \;\rightarrow\; \text{return the original } x$$

When it exists, the function "which undoes" the function $f(x)$ is found by solving $y = f(x)$ for x as a function of y and is called the ***inverse function*** of f, and is denoted by f^{-1}. It turns out that it is not always possible to solve $y = f(x)$ for x as a function of y. Even when it is possible, it can be really hard to do.

You should be careful not to confuse $f^{-1}(x)$ with $\dfrac{1}{f(x)}$

For instance, suppose that a particle's position, s, at time t is given by the formula $s(t) = 7t$. Given any particular time t, you can quickly work out the corresponding position s. However, if you are asked the question "When does the particle reach $s = 4$?" then you need to "undo" $s(t) = 7t = 4$ to find t in terms of s. This is easy to do:

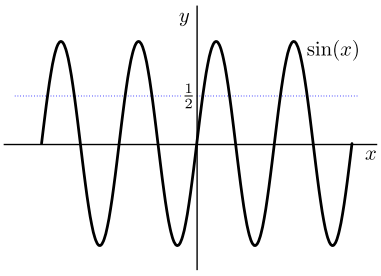
A sketch of $s(t) = 7t$:



$$s(t) = 7t \quad \Longrightarrow \quad t(s) = \frac{s}{7}$$

and so $t(4) = \frac{4}{7}$. However, this question is not always so easy. Consider the function $f(x) = sin(x)$. When is $f(x) = \frac{1}{2}$? In other words, we want to know at which values of x does the curve $sin(x)$ cross the horizontal line $y = \frac{1}{2}$? We can see that there are infinitely many x-values that give $sin(x) = \frac{1}{2}$. This means that $f(x) = \sin(x)$ is not a ***one-to-one*** function. Because of this it can not be "undone". The "one-to-one-ness" is crucial, hence we will state this property precisely.

A sketch of $y = \sin(x)$:



A function f is one-to-one (injective) when it never takes the same value more than once. That is

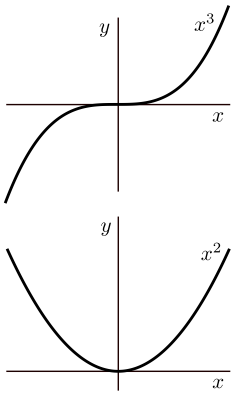
$$\text{if } x_1 \neq x_2 \text{ then } f(x_1) \neq f(x_2)$$

There is an easy way to test if a function is one-to-one if you have a plot of it. This test is called ***the horizontal line test***.

A function is one-to-one if and only if no horizontal line $y = c$ intersects the graph $y = f(x)$ more than once.

Therefore every horizontal line intersects the graph either zero or one times. Never twice or more. This test tell us that, for instance, $y = x^3$ is one-to-one, but $y = x^2$ is not. However note that if we restrict the domain of $y = x^2$ to $x \geq 0$ then the horizontal line test is passed. This is one of the reasons we have to be careful to consider the domain of the function.

A sketch of $y = x^3$ and $y = x^2$:

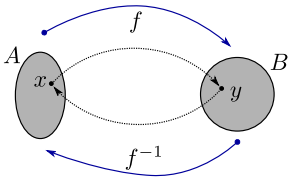


When a function is one-to-one then it has an inverse function.

Let f be a one-to-one function with domain A and range B. Then its inverse function, f^{-1}, has domain B and range A. It is defined by

$$f^{-1}(x) = y \text{ whenever } f(x) = y$$

A one-to-one function and its inverse:



As a result, we have

$$f^{-1}(f(x)) = x \qquad \text{for any } x \in A$$
$$f(f^{-1}(y)) = y \qquad \text{for any } y \in B$$

Example 1.4: Inverse functions

Let $f(x) = x^5 + 3$ on the domain $\mathbb{R}$. To find its inverse we first write $y = f(x)$, that is $y = x^5 + 3$. Solving this equality in terms of x gives

$$y = x^5 + 3 \quad \Longrightarrow \quad x^5 = y - 3 \quad \Longrightarrow \quad x = (y-3)^{1/5}$$

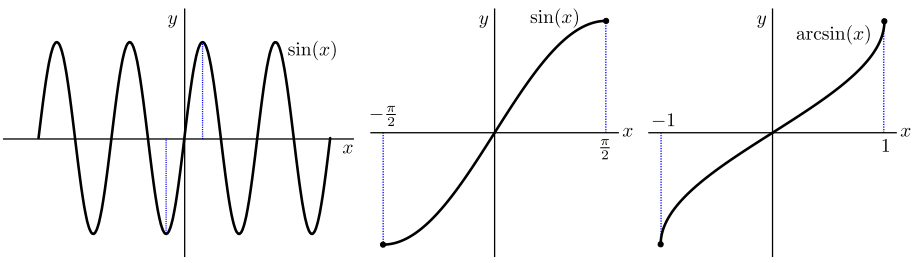
and so $f^{-1}(y) = (y-3)^{1/5}$. The variable "y" is a dummy variable. We can replace it with any other variable, the favourite being "x". Therefore the inverse of $f(x) = x^5 + 3$ is $f^{-1}(x) = (x-3)^{1/5}$.

Let $g(x) = \sqrt{x-1}$ on the domain $x \geq 1$. To find the inverse we repeat the steps as above,

$$y = \sqrt{x-1} \quad \Longrightarrow \quad y^2 = x - 1 \quad \Longrightarrow \quad x = y^2 + 1$$

Therefore the inverse of $g(x) = \sqrt{x-1}$ is $g^{-1}(x) = x^2 + 1$.

Finally, let $h(x) = sin(x)$ on the domain $\mathbb{R}$. Notice that as x runs from $-\frac{\pi}{2} \leq x \leq \frac{\pi}{2}$, $\sin(x)$ increases from -1 to $+1$. Thus if we restrict the domain of $\sin(x)$ to $-\frac{\pi}{2} \leq x \leq \frac{\pi}{2}$, it does have an inverse function, traditionally called $\arcsin(x)$.



Lecture 2

1.2 Limits

Before we step into differential calculus, we need to understand the concept of the limit. The definition of the limit is not very enlightening when people see it for the first time, hence we will start with two informal problems.

1.2.1 Two problems

Suppose an object moves along the x-axis and its displacement s (in meters) at time t (in seconds) is given by

$$s(t) = t^2 \qquad t \geq 0$$

We want to consider its velocity at a certain time instant, say at $t = 2$. Velocity (or speed) is defined by

$$\text{velocity} = \frac{\text{distance travelled}}{\text{time elapsed}} \tag{1.3}$$

It can only be applied to find average velocities over time intervals. We (still) don't have a definition for velocity at $t = 2$. To define the velocity at $t = 2$, we consider short time intervals about $t = 2$, say from $t = 2$ to $t = 2 + 1/2^n$. Using (1.3) we can compute the average velocity v_n over the time intervals with $n = 1, 2, 3, 4, \ldots$

n	Time interval	Velocity
1	[2, 2.5]	4.5 m/s
2	[2, 2.25]	4.25 m/s
3	[2, 2.125]	4.125 m/s
4	[2, 2.0625]	4.0625 m/s
⋮	⋮	⋮

In general, the velocity v_n over the time interval $[2,\, 2 + 1/2^n]$ is

$$v_n = \frac{(2 + 1/2^n)^2 - 2^2}{1/2^n} = \frac{4 + 2 \cdot 2 \cdot 1/2^n + 1/2^{2n} - 4}{1/2^n} = 4 + 1/2^n$$

It is clear that if n is very large (that is, if the time interval is very short), v_n is very close to 4 m/s. The velocity, called the instantaneous velocity, at $t = 2$ is (defined to be) 4 m/s.

Next, suppose we want to find the area of the region that lies under the curve $y = x^2$ and above the x-axis for x between 0 and 1. We divide the interval $[0,\, 1]$ into finitely many subintervals of equal lengths,

$$\left[0, \frac{1}{n}\right], \quad \left[\frac{1}{n}, \frac{2}{n}\right], \quad \left[\frac{2}{n}, \frac{3}{n}\right], \quad \ldots, \quad \left[\frac{n-1}{n}, 1\right]$$

For each subinterval $[\frac{i-1}{n}, \frac{i}{n}]$ we consider the rectangular region with base on the subinterval and height $(\frac{i-1}{n})^2$ (the largest region that lies under the curve). If we add the area of these rectangular regions, the sum is smaller than that of the required region. However, if n is very large, the error is very small and we get a good approximation for the required area.

In general, if there are n subintervals, the sum S_n of the areas of the rectangular regions is

$$\begin{aligned} S_n &= \tfrac{1}{n}\cdot 0^2 + \tfrac{1}{n}\cdot\left(\tfrac{1}{n}\right)^2 + \tfrac{1}{n}\cdot\left(\tfrac{2}{n}\right)^2 + \ldots + \tfrac{1}{n}\cdot\left(\tfrac{n-1}{n}\right)^2 \\ &= \frac{1^2+2^2+\ldots+(n-1)^2}{n^3} \\ &= \frac{n(n-1)(2n-1)}{6n^3} \\ &= \frac{2n^3-3n^2+n}{6n^3} \\ &= \frac{1}{3}-\frac{1}{2n}+\frac{1}{6n^2} \end{aligned}$$

In this example we need to use the sum of squares formula:

$$1^2+2^2+\ldots+n^2 = \frac{n(n+1)(2n+1)}{6}$$

It is clear that if n is very large, S_n is very close the region under the curve, and is very close to $1/3$. Hence the area of the region is $1/3$.

1.2.2 *Limits of sequences*

In the last subsection, we obtained formulas for quantities v_n and S_n. Each of these formulas gives a sequence (which is a special type of function):

- *A **sequence** is a function whose domain is* $\mathbb{N}$.
- *A **sequence of real numbers** is a sequence whose codomain is* $\mathbb{R}$.

Let $f : \mathbb{N} \to \mathbb{R}$ be a sequence (of real numbers). For each positive integer n, the value $f(n)$ is called the nth term of the sequence and is usually denoted by a small letter together with n in the subscript, for example a_n. The sequence is also denoted by $(a_n)_{n=1}^{\infty}$ because if we know all the a_n's, then we know the sequence.

Sometimes, we represent a sequence $(a_n)_{n=1}^{\infty}$ by listing the first few terms of it:

$$a_1,\ a_2,\ a_3,\ a_4,\ \ldots$$

For instance, if $a_n = 4 + 1/2^n$, then $(a_n)_{n=1}^{\infty}$ can be represented by

$$\frac{9}{2},\ \frac{17}{4},\ \frac{33}{8},\ \frac{65}{16},\ \cdots$$

However, it is not always a good way to describe a sequence by listing a few terms in the sequence, since it may not be easy to find a formula for the nth term. Moreover, different people may obtain different formulas. Thus it is better to describe a sequence by writing down a formula for the nth term explicitly.

In this course our focus will be on the convergent sequences.

- *A sequence $(a_n)_{n=1}^{\infty}$ is said to be **convergent** if there exists a real number L such that a_n is arbitrarily close to L if n is sufficiently large.*
- *We then say that L is the limit of $(a_n)_{n=1}^{\infty}$ and write*

$$\lim_{n\to\infty} a_n = L$$

Here the "arbitrarily close" means that we can make $|a_n - L|$ as small as we want by taking n large enough. For the sequence $(a_n)_{n=1}^{\infty}$ where $a_n = 1/2^n$, we can make a_n arbitrarily close to $L = 0$ by taking n large enough. For example, if we want $|1/2^n - 0| < 0.01$, we can take any $n > 7$; if we want $|1/2^n - 0| < 0.001$, we can take any $n > 10$, etc.

A sequence that does not converge is called ***divergent***. An obvious example of a divergent sequence is

$$(n)_{n=1}^{\infty} = 1, 2, 3, 4, \ldots$$

In this case the limit does not exists and we write $\lim_{n\to\infty} a_n = \infty$. Another example of a divergent sequence is

$$((-1)^n)_{n=1}^{\infty} = -1, +1, -1, +1, \ldots$$

In this case we only say that the limit does not exist.

In general, in order to compute the limit of a sequence, we need to use the following rules:

These rules can be proved using the "ϵ–δ" method, however this is beyond the scope of these lectures.

(L1) $\lim\limits_{n\to\infty} k = k$ where k is a constant

(L2) $\lim\limits_{n\to\infty} \dfrac{1}{n^p} = 0$ where p is a positive constant

(L3) $\lim\limits_{n\to\infty} \dfrac{1}{b^n} = 0$ where b is a positive constant greater than 1

(L4) $\lim\limits_{n\to\infty} (a_n + b_n) = \lim\limits_{n\to\infty} a_n + \lim\limits_{n\to\infty} b_n$

(L5) $\lim\limits_{n\to\infty} a_n b_n = \lim\limits_{n\to\infty} a_n \cdot \lim\limits_{n\to\infty} b_n$

(L6) $\lim\limits_{n\to\infty} \dfrac{a_n}{b_n} = \dfrac{\lim\limits_{n\to\infty} a_n}{\lim\limits_{n\to\infty} b_n}$ provided $\lim\limits_{n\to\infty} b_n \neq 0$

- The meaning of (L1) is that if $a_n = k$ for all n, then the sequence $(a_n)_{n=1}^{\infty}$ is convergent and its limit is k.
- The meaning of (L4) is that if both $(a_n)_{n=1}^{\infty}$ and $(b_n)_{n=1}^{\infty}$ are convergent and their limits are L and M respectively, then $(a_n + b_n)_{n=1}^{\infty}$ is also convergent and its limit is $L + M$.
- Applying (L1) with $a_n = k$ to (L5) gives a special case of it,

(L5′) $\lim\limits_{n\to\infty} kb_n = k \lim\limits_{n\to\infty} b_n$

- Using (L5′) and (L4) we get

(L4′) $\lim\limits_{n\to\infty} (a_n - b_n) = \lim\limits_{n\to\infty} a_n - \lim\limits_{n\to\infty} b_n$

The rules (L4) and (L4′) are valid for any sum and any difference of finitely many sequences. Similarly, the rule (L5) is valid for any product of finitely many sequences.

You may ask if we can simply write $\lim a_n = L$, omitting $n \to \infty$? For sequences, this will not cause ambiguity. However, for functions in general, limits at infinity as well as limits at a point a (where $a \in \mathbb{R}$) the notations $\lim_{x\to\infty} f(x)$ and $\lim_{x\to a} f(x)$ have different meanings.

Example 1.5: Limits of sequences

(i) Find $\lim_{n\to\infty}(4+1/2^n)$, if it exists:

$$\begin{aligned}\lim_{n\to\infty}(4+1/2^n) &= \lim_{n\to\infty}4+\lim_{n\to\infty}1/2^n && \text{by (L4)}\\ &= 4+0 && \text{by (L1) and (L3)}\\ &= 4\end{aligned}$$

(ii) Find $\lim_{n\to\infty}\dfrac{2n^3-3n^2+n}{6n^3}$, if it exists:

$$\begin{aligned}\lim_{n\to\infty}\frac{2n^3-3n^2+n}{6n^3} &= \lim_{n\to\infty}\left(\frac{1}{3}-\frac{1}{2n}+\frac{1}{6n^2}\right) && \text{by splitting the fraction}\\ &= \lim_{n\to\infty}\frac{1}{3}-\lim_{n\to\infty}\left(\frac{1}{2}\cdot\frac{1}{n}\right)+\lim_{n\to\infty}\left(\frac{1}{6}\cdot\frac{1}{n^2}\right) && \text{by (L4)}\\ &= \frac{1}{3}-\frac{1}{2}\cdot\lim_{n\to\infty}\frac{1}{n}+\frac{1}{6}\cdot\lim_{n\to\infty}\frac{1}{n^2} && \text{by (L1) and (L5')}\\ &= \frac{1}{3}-\frac{1}{2}\cdot 0+\frac{1}{6}\cdot 9 && \text{by (L2)}\\ &= \frac{1}{3}\end{aligned}$$

(iii) Find $\lim_{n\to\infty}(1+2n)$, if it exists:

Limit does not exist. This is because we can't find any real number L satisfying the condition that $2n+1$ is close to L if n is large. If we apply rules for limits, we get

$$\begin{aligned}\lim_{n\to\infty}(2n+1) &= \lim_{n\to\infty}2n+\lim_{n\to\infty}1 && \text{by (L4)}\\ &= 2\lim_{n\to\infty}n+1 && \text{by (L1) and (L5')}\end{aligned}$$

However, we can't proceed because $\lim_{n\to\infty}n$ does not exist. From this, we see that the given limit does not exist.

(iv) Find $\lim_{n\to\infty}\dfrac{n+1}{2n+1}$, if it exists:

We can't use (L6) because limits of the numerator and the denominator do not exist. However, we can't conclude from this that the given limit does not exist. To find the limit, we use a trick: divide the numerator and the denominator by n,

$$\begin{aligned}\lim_{n\to\infty}\frac{n+1}{2n+1} &= \lim_{n\to\infty}\frac{\frac{n+1}{n}}{\frac{2n+1}{n}} && \text{divide numerator and denominator by } n\\ &= \frac{\lim_{n\to\infty}\frac{n+1}{n}}{\lim_{n\to\infty}\frac{2n+1}{n}} && \text{by (L6)}\\ &= \frac{\lim_{n\to\infty}1+\lim_{n\to\infty}\frac{1}{n}}{\lim_{n\to\infty}2+\lim_{n\to\infty}\frac{1}{n}} && \text{by (L4)}\\ &= \frac{1+0}{2+0} && \text{by (L1) and (L2)}\\ &= \frac{1}{2}\end{aligned}$$

1.2.3 *Limits of functions at infinity*

Recall that a sequence is a function from $\mathbb{N}$ to $\mathbb{R}$. When we consider limit of a sequence $(a_n)_{n=1}^{\infty}$, we let n approach ∞ through the discrete points $n = 1, 2, 3, \ldots$

In many cases, we will consider functions f from a subset of $\mathbb{R}$ to $\mathbb{R}$ such that $f(x)$ is defined when x is large. For such functions, we can let x approach ∞ continuously (through large real numbers) and consider the behaviour of $f(x)$.

*Let f be a function such that $f(x)$ is defined for sufficiently large x. Suppose L is a real number such that $f(x)$ is arbitrarily close to L if x is sufficiently large. Then we say that L is the **limit** of f **at infinity** and write*

$$\lim_{x\to\infty} f(x) = L$$

Here the condition "$f(x)$ is defined for sufficiently large x" means that there is a real number r such that $f(x)$ is defined for all $x > r$.

In general, $f(x)$ does not need to approach a fixed number L as x approached infinity.

Let f be a function such that $f(x)$ is defined for sufficiently large x. Suppose that $f(x)$ is positive or negative and arbitrarily large if x is sufficiently large. Then we say that the limit of f at infinity does not exists and write

$$\lim_{x\to\infty} f(x) = +\infty \quad \textit{or} \quad \lim_{x\to\infty} f(x) = -\infty$$

respectively.

Finding limits of functions at infinity is very similar to that of sequences. We need to use the following rules, where f and g are functions such that $f(x)$ and $g(x)$ are defined for sufficiently large x:

(L1) $\lim_{x\to\infty} k = k$ where k is a constant

(L2) $\lim_{x\to\infty} \frac{1}{x^p} = 0$ where p is a positive constant

(L3) $\lim_{x\to\infty} \frac{1}{b^x} = 0$ where b is a positive constand greater than 1

(L4) $\lim_{x\to\infty} (f(x) \pm g(x)) = \lim_{x\to\infty} f(x) \pm \lim_{x\to\infty} g(x)$

(L5) $\lim_{x\to\infty} f(x) \cdot g(x) = \lim_{x\to\infty} f(x) \cdot \lim_{x\to\infty} g(x)$

(L5') $\lim_{x\to\infty} k \cdot g(x) = k \cdot \lim_{x\to\infty} g(x)$

(L6) $\lim_{x\to\infty} \frac{f(x)}{g(x)} = \frac{\lim_{x\to\infty} f(x)}{\lim_{x\to\infty} g(x)}$ provided $\lim_{x\to\infty} g(x) \neq 0$

To consider limits of functions at infinity, we should first check the domains of the functions. For example, if $f(x) = \sqrt{1-x}$ the domain of f is $x \leq 1$; it is meaningless to talk about the limit of f at infinity.

You need to be careful when calculations involve infinities:

$r + \infty = \infty$ for any $r \in \mathbb{R}$
$r \cdot \infty = \infty$ for any $r \in \mathbb{R}_{>0}$

However

$\infty - \infty$ is undefined
$0 \cdot \infty$ is undefined
$0/0$ is undefined
∞/∞ is undefined

Example 1.6: Limits of functions at infinity

(i) Find $\lim_{x\to\infty}(1-2/x^3)$, if it exists:

$$\begin{aligned}\lim_{x\to\infty}(1-2/x^3) &= \lim_{x\to\infty}1-2\lim_{x\to\infty}1/x^3 && \text{by (L4) and (L5')}\\ &= 1-2\cdot 0 && \text{by (L1)}\\ &= 1\end{aligned}$$

(ii) Find $\lim_{x\to\infty}(1+x^2)$, if it exists:

$$\begin{aligned}\lim_{x\to\infty}(1+x^2) &= \lim_{x\to\infty}1+\lim_{x\to\infty}x^2 && \text{by (L4)}\\ &= 1+\infty && \text{by (L1)}\\ &= \infty && \text{the limit does not exist}\end{aligned}$$

(iii) Find $\lim_{x\to\infty}\dfrac{x^2+1}{3x^3-4x+5}$, if it exists:

$$\begin{aligned}\lim_{x\to\infty}\frac{x^2+1}{3x^3-4x+5} &= \lim_{x\to\infty}\frac{\frac{x^2+1}{x^3}}{\frac{3x^3-4x+5}{x^3}} && \text{divide numerator and denominator by } x^3\\ &= \frac{\lim_{x\to\infty}\left(\frac{1}{x}+\frac{1}{x^3}\right)}{\lim_{n\to\infty}\left(3-\frac{4}{x^2}+\frac{5}{x^3}\right)} && \text{by (L6)}\\ &= \frac{\lim_{x\to\infty}\frac{1}{x}+\lim_{x\to\infty}\frac{1}{x^3}}{\lim_{n\to\infty}3-\lim_{x\to\infty}\frac{4}{x^2}+\lim_{x\to\infty}\frac{5}{x^3}} && \text{by (L4)}\\ &= \frac{0+0}{3-0+0} && \text{by (L1), (L2) and (L5')}\\ &= 0\end{aligned}$$

Leading term rule. Let $f(x)$ and $g(x)$ be polynomials of degree n and m respectively, then

$$\lim_{x\to\infty}\frac{f(x)}{g(x)}=\lim_{x\to\infty}\frac{a_nx^n+a_{n-1}x^{n-1}+\ldots+a_1x+a_0}{b_mx^m+b_{m-1}x^{m-1}+\ldots+b_1x+b_0}=\lim_{x\to\infty}\frac{a_nx^n}{b_mx^m}$$

This is because we can write the fraction $f(x)/g(x)$ as

$$\frac{f(x)}{g(x)}=\frac{a_nx^n}{b_mx^m}\cdot\underbrace{\frac{1+\frac{a_{n-1}}{a_n}\cdot\frac{1}{x}+\frac{a_{n-2}}{a_n}\cdot\frac{1}{x^2}+\ldots+\frac{a_1}{a_n}\cdot\frac{1}{x^{n-1}}+\frac{a_0}{a_n}\cdot\frac{1}{x^n}}{1+\frac{b_{m-1}}{b_m}\cdot\frac{1}{x}+\frac{b_{m-2}}{b_m}\cdot\frac{1}{x^2}+\ldots+\frac{b_1}{b_m}\cdot\frac{1}{x^{m-1}}+\frac{b_0}{b_m}\cdot\frac{1}{x^m}}}_{\text{this goes to } 1/1 \text{ as } x \text{ goes to } \infty}$$

Thus,

$$\lim_{x\to\infty}\frac{f(x)}{g(x)}=\begin{cases}\infty & \text{if } n>m\\ a_n/b_n & \text{if } n=m\\ 0 & \text{if } n<m\end{cases}$$

For example,

$$\lim_{x\to\infty}\frac{x^3+1}{3x^3-4x+5}=\lim_{x\to\infty}\frac{x^3}{3x^3}=\frac{1}{3}$$

Limits at negative infinity. Similar to limits at infinity, we may consider limits at negative infinity provided that $f(x)$ is defined for x sufficiently large negative. The calculations works in the same way, for example,

$$\lim_{x\to-\infty} 1/x = 0$$
$$\lim_{x\to-\infty} x^3 = -\infty$$
$$\lim_{x\to-\infty} (1-x^3) = +\infty$$

1.2.4 *Limits of functions at a point*

We now consider limits of a function at a point $a \in \mathbb{R}$. Note that x can approach a from the left-side or from the right-side, hence we must have left-side and right-side limits. They are called ***one-sided limits***.

> *Let $a \in \mathbb{R}$ and let f be a function such that $f(x)$ is defined for x sufficiently close to and greater than a. Suppose L is a real number such that $f(x)$ is arbitrarily close to L if x is sufficiently close to and greater than a. Then we say that L is the **right-side limit** of f at a and write*
>
> $$\lim_{x\to a+} f(x) = L$$

Here the condition "$f(x)$ is defined for x sufficiently close to and greater than a" means that there is a positive real number δ such that $f(x)$ is defined for all $x \in (a, a+\delta)$ but not necessarily at $x = a$. Even if $f(a)$ is defined, it has no effect on the existence and the value of $\lim_{x\to a+} f(x)$. For simplicity, we will often say "f is defined on the right-side of a".

The left-side limit of f at a is defined similarly.

> *Let $a \in \mathbb{R}$ and let f be a function such that $f(x)$ is defined for x sufficiently close to and less than a. Suppose L is a real number such that $f(x)$ is arbitrarily close to L if x is sufficiently close to and less than a. Then we say that L is the **left-side limit** of f at a and write*
>
> $$\lim_{x\to a-} f(x) = L$$

We are now at a position to define the two-sided limit of a function at a point.

> *Let $a \in \mathbb{R}$ and let f be a function defined on the left-side and right-side of a. Suppose that both $\lim_{x\to a-} f(x)$ and $\lim_{x\to a+} f(x)$ exist and are equal (with the common limit denoted by L which is a real number). Then the **two-sided limit**, or more simply, the limit of f at a is defined to be L, and is written as*
>
> $$\lim_{x\to a} f(x) = L$$

Here the condition "f be a function defined on the left-side and the right-side of a" means that there is a positive real number δ such that

$f(x)$ is defined for all $x \in (a, a-\delta) \cup (a, a+\delta)$ but not necessarily at $x = a$.

If the left-side and right-side limits of f at a are not the same, or do not exist, we say that the limit of f at a does not exist.

For example, consider the function $f(x) = |x|/x$. Its domain is $\mathbb{R}\backslash\{0\}$, thus it makes sense to talk about the limit of $f(x)$ at $x = 0$. It is easy to see that the one-sided limits are

$$\lim_{x\to 0-} |x|/x = -1 \qquad \lim_{x\to 0+} |x|/x = +1$$

Hence the limit of $f(x) = |x|/x$ at $x = 0$ does not exist.

As a second example, consider the function $f(x) = 1/x^2$. Its domain is also $\mathbb{R}\backslash\{0\}$, and the one-sided limits are

$$\lim_{x\to 0-} 1/x^2 = \infty \qquad \lim_{x\to 0+} 1/x^2 = \infty$$

Hence

$$\lim_{x\to 0} 1/x^2 = \infty$$

To find limits at a point of complicated functions we need to use the following rules:

(La1) $\lim_{x\to a} k = k$ where k is a constant

(La2) $\lim_{x\to a} x^n = a^n$ where n is a positive integer

(La2′) $\lim_{x\to a} \sqrt[n]{x} = \sqrt[n]{a}$ where n is an odd positive integer

(La2″) $\lim_{x\to a} \sqrt[n]{x} = \sqrt[n]{a}$ where $a > 0$ and n is an even positive integer

(La3) $\lim_{x\to a} b^x = b^a$ where b is a positive real number

(La4) $\lim_{x\to a}(f(x) \pm g(x)) = \lim_{x\to a} f(x) \pm \lim_{x\to a} g(x)$

(La5) $\lim_{x\to a} f(x) \cdot g(x) = \lim_{x\to a} f(x) \cdot \lim_{x\to a} g(x)$

(La5′) $\lim_{x\to a} k \cdot g(x) = k \cdot \lim_{x\to a} g(x)$ where k is a constant

(La6) $\lim_{x\to a} \dfrac{f(x)}{g(x)} = \dfrac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)}$ provided $\lim_{x\to a} g(x) \neq 0$

The rule (La4) is valid for any sum and any difference of finitely many functions; the rule (La5) is valid for any product of finitely many functions.

Before moving to complicated examples, we want to state the limit property of polynomials:

> *Let $P(x)$ be a polynomial and let $a \in \mathbb{R}$. Then*
>
> $$\lim_{x\to a} P(x) = P(a)$$

Example 1.7: Limits of functions at a point

(i) Find $\lim_{x\to -1} x\sqrt{x^2+1}$, if it exists:

$$\begin{aligned}
\lim_{x\to -1} x\sqrt{x^2+1} &= \lim_{x\to -1} x \cdot \lim_{x\to -1} \sqrt{x^2+1} && \text{by (La5)}\\
&= (-1)\cdot\sqrt{(-1)^2+1} && \text{by the limit property of polynomials and (La2")}\\
&= -\sqrt{2}
\end{aligned}$$

(ii) Find $\lim_{x\to 2} \dfrac{x-1}{x^2+x-2}$, if it exists:

$$\begin{aligned}
\lim_{x\to 2} \frac{x-1}{x^2+x-2} &= \frac{\lim_{n\to 2}(x-1)}{\lim_{n\to 2}(x^2+x-2)} && \text{by (La6)}\\
&= \frac{2-1}{2^2+2-2} && \text{by the limit property of polynomials}\\
&= \frac{1}{4}
\end{aligned}$$

(iii) Find $\lim_{n\to 1} \dfrac{x-1}{x^2+x-2}$, if it exists:
Notice that repeating the same steps as in (i) gives $0/0$ which is undefined, thus we need to proceed in a slightly different manner

$$\begin{aligned}
\lim_{x\to 1} \frac{x-1}{x^2+x-2} &= \lim_{n\to 1} \frac{x-1}{(x-1)(x+2)} && \text{by factorising denominator}\\
&= \lim_{x\to 1} \frac{1}{x+2} && \text{by simplifying the fraction}\\
&= \frac{\lim_{x\to 2} 1}{\lim_{n\to 2}(x+2)} && \text{by (La6)}\\
&= \frac{1}{1+2} && \text{by the limit property of polynomials}\\
&= \frac{1}{3}
\end{aligned}$$

(iv) Find $\lim_{h\to 0} \dfrac{f(x+h)-f(x)}{h}$ for $f(x)=x^2+3$, if it exists:
The expression $(f(x+h)-f(x))/h$ is called a difference quotient and involves two variables, x and h. The questions asks for the limit at $h=0$, thus the x needs to be considered as a constant. Hence

$$\begin{aligned}
\lim_{h\to 0} \frac{f(x+h)-f(x)}{h} &= \lim_{h\to 0} \frac{((x+h)^2+3)-(x^2+3)}{h}\\
&= \lim_{h\to 0} \frac{(x^2+2xh+h^2+3)-(x^2+3)}{h}\\
&= \lim_{h\to 0} \frac{2xh+h^2}{h}\\
&= \lim_{h\to 0} (2x+h)\\
&= 2x && \text{by (La4)}
\end{aligned}$$

1.2.5 Continuous functions

We saw above that for "nice" functions $\lim_{x\to a} f(x)$ equals $f(a)$. Functions with this property are called continuous functions.

> *Let $a \in \mathbb{R}$ and let f be a function such that $f(x)$ is defined for x sufficiently close to a (including a). Then f is **continuous** at a if*
>
> $$\lim_{x\to a} f(x) = f(a)$$
>
> *If $\lim_{x\to a} f(x)$ does not exist or if $\lim_{x\to a} f(x)$ exists but does not equal $f(a)$, then f is **discontinuous** at a.*

Roughly speaking, the condition above means that if x changes from a to $a+\delta$, where δ is a small real number, the corresponding change of $f(x)$ is also small.

We can extend the notion of continuity from a point to an open interval.

> *Let $I \subset \mathbb{R}$ be an open interval and let f be a function defined on I. We say that f is continuous on I if it is continuous at every $a \in I$.*

Geometrically, a function f is continuous on an open interval I means that the graph of f on I has no "break"; if we use a pen to draw the graph on paper, we can draw it continuously without raising the pen above the paper. For example, $f(x) = |x|$ is continuous everywhere, but $f(x) = |x|/x$ is discontinuous at $x = 0$; at this point it "jumps" from -1 to 1.

Strictly speaking, $f(x) = |x|/x$ is not defined at $x = 0$, but this can be "fixed" in the following way:

$$f(x) = \begin{cases} |x|/x & \text{if } x \neq 1 \\ 0 & \text{if } x = 0 \end{cases}$$

The following two results give two more examples of continuous functions:

> - *Every polynomial function is continuous on $\mathbb{R}$.*
> - *Every rational function is continuous on every open interval contained in its domain.*

Next, we introduction the notion of a one-sided continuity.

> *Let $a \in \mathbb{R}$ and let f be a function defined on the right-side of a as well as at a. Then f is **right-continuous** at a if*
>
> $$\lim_{x\to a+} f(x) = f(a)$$
>
> *Let $b \in \mathbb{R}$ and let f be a function defined on the left-side of b as well as at b. Then f is **left-continuous** at b if*
>
> $$\lim_{x\to b-} f(x) = f(b)$$

We can now combine the different notions of continuity.

> *Let $[a,b] \subset \mathbb{R}$ and let f be a function defined on $[a,b]$. We say that f is continuous on $[a,b]$ if it is continuous at every $x \in (a,b)$, right-continuous at a, and left-continuous at b.*

For example, $f(x) = \sqrt{x}$ is right-continuous at $x = 0$. It is both left- and right-continuous at any $x > 0$. In fact, it is continuous on any interval $[a, b]$ such that $0 \leq a < b$.

The following statement is an important property of continuous functions on closed intervals that plays a crucial role in Data Science, which will be explained later in these lectures.

> *Let f be a function that is defined and continuous on an interval $[a, b]$. Then f attains its maximum and minimum in $[a, b]$, that is, there exist $x_1, x_2 \in [a, b]$ such that*
>
> $$f(x_1) \leq f(x) \leq f(x_2) \quad \text{for all } x \in [a, b]$$

Here the interval $[a, b]$ is closed since the minimum or the maximum could be attained at the end of the interval. For example, it makes no sense to talk about the minimum of $f(x) = \sqrt{x}$ on (a, b) with $0 < a < b$.

1.3 Differentiation

Lecture 3

In this lecture we define the derivative as the slope of a tangent line, using a particular limit. This will allow us to find the equation of the tangent line to a function at any point. However, the main importance of derivatives does not come from this application. Instead, (arguably) it comes from the interpretation of the derivative as the instantaneous rate of change of a quantity and can be used to find the (local) extremum points: the minimum, the maximum and the inflection points. In Data Science, differentiation helps us to find the critical values of the unknown parameters in linear regression models and to find the values of parameters that minimise the loss function in the deep learning algorithms.

1.3.1 Defining the derivative

> *Suppose $D \subseteq \mathbb{R}$ and $f : D \to \mathbb{R}$. Let $a \in D$. The **derivative** of f at a is*
>
> $$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$
>
> *provided the limit exists. When it does exist, the function is said to be **differentiable** at a. If the derivative $f'(x)$ exists for all $x \in (a, b) \subseteq D$ we say that f is differentiable on (a, b).*

The following notations are all used for "the derivative of f with respect to x":

$$f'(x) \quad \frac{df}{dx} \quad \frac{d}{dx}f(x) \quad D_x f(x)$$

The primed notation is due to I. Newton. The d/dx notation is due to G. W. Leibniz. In fact, you can think of the derivative as a "generalised" function that maps differentiable functions into functions:

$$D_x : \mathcal{C}^1 \to \mathcal{C}^0$$

where $\mathcal{C}^1$ is set of all at least once differentiable functions and $\mathcal{C}^0$ is the set of all functions.

The derivative $f'(a)$ at a specific point $x = a$ is the ***slope*** of the line ***tangent*** to f at a. We will explore this property in a bit later.

When we allow a to be any number within the domain, we are thinking of the derivative as an operation we can do on a function. We then simply write

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

or equivalently

$$f'(x) = \lim_{a \to x} \frac{f(x) - f(a)}{x - a}$$

Let us now compute the derivatives of some simple functions. This will help us to build a toolbox for computing derivatives of complicated functions.

Example 1.8: A few simple derivatives

Suppose $a, b, c \in \mathbb{R}$ and $n \in \mathbb{N}$. Then we have the following derivatives:

$$\begin{aligned}
f(x) &= a & f'(x) &= 0 \\
f(x) &= ax + b & f'(x) &= a \\
f(x) &= ax^2 + bx + c & f'(x) &= 2ax + b \\
f(x) &= x^n & f'(x) &= nx^{n-1} \\
f(x) &= 1/x & f'(x) &= -1/x^2 \qquad (x \neq 0)
\end{aligned}$$

Constants: if $f(x) = a$ ($x \in \mathbb{R}$), then

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

Linear functions: if $f(x) = ax + b$ ($x \in \mathbb{R}$), then

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{a(x+h) + b - ax - b}{h} = a$$

Quadratics: if $f(x) = ax^2 + bx + c$ ($x \in \mathbb{R}$), then

$$\begin{aligned}
f'(x) &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \\
&= \lim_{h \to 0} \frac{a(x+h)^2 + b(x+h) + c - ax^2 - bx - c}{h} \\
&= \lim_{h \to 0} (2ax + b + h) \\
&= 2ax + b
\end{aligned}$$

Powers: if $n \in \mathbb{N}$ and $f(x) = x^n$ ($x \in \mathbb{R}$), then

$$\begin{aligned}
f'(x) &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \\
&= \lim_{h \to 0} \frac{1}{h} \left[x^n + nhx^{n-1} + n(n-1)h^2 x^{n-2}/2 + \cdots + h^n - x^n \right] \\
&= \lim_{h \to 0} \left[nx^{n-1} + n(n-1)hx^{n-2}/2 + \cdots + h^{n-1} \right]
\end{aligned}$$

All terms apart from the first are of the form $h^k \times$ (something not depending on h) for some $k \in \mathbb{N}$. All such terms tend to zero, so f is differentiable at x and $f'(x) = nx^{n-1}$.

Reciprocal: if $f(x) = 1/x$ ($x \in \mathbb{R}$, $x \neq 0$), then

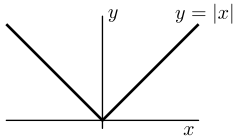
$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \to 0} \frac{1}{h} \left[\frac{1}{x+h} - \frac{1}{x} \right] = \lim_{h \to 0} \frac{-1}{x(x+h)} = -\frac{1}{x^2}$$

According to definition, the derivative $f'(a)$ exists precisely when the limit $\lim_{h\to 0}(f(x+h)-f(x))/h$ exists. This limit does not exist if the function is not defined at $x=a$ or it does not have a tangent line at $x=a$. We have met the first scenario in the example above. The function $f(x)=1/x$ "blows up" (i.e. becomes infinite) at $x=0$. It does not have a tangent line at $x=0$ and its derivative does not exist at $x=0$. The second scenario is demonstrated in the example below.

Example 1.9: A function not differentiable at a point

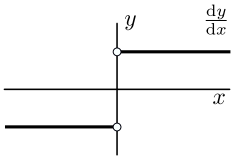
Define $f:\mathbb{R}\to\mathbb{R}$ by $f(x)=|x|$, that is

$$f(x)=\begin{cases}-x & \text{if } x\le 0\\ x & \text{if } x>0\end{cases}$$



Then f is differentiable everywhere except at 0. Its derivative is the function:

$$f'(x)=\begin{cases}-1 & \text{if } x<0\\ \text{undefined} & \text{if } x=0\\ 1 & \text{if } x>0\end{cases}$$



The calculations go as follows. Assume $x<0$. Then

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

Now assume $x>0$. Then

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

Finally, when $x=0$ we need to compute left and right limits:

$$\lim_{h\to 0-}\frac{f(0+h)-f(0)}{h}=-1,\qquad \lim_{h\to 0+}\frac{f(0+h)-f(0)}{h}=1$$

Since the one-sided limits differ, the limit as $h\to 0$ does not exist. And thus the derivative $f'(x)$ does not exist at $x=0$. We can not draw a tangent line to f at this point.

If f is differentiable at a, then $\lim_{x\to a-}f(x)=\lim_{x\to a+}f(x)$ meaning that f is continuous at a. In short, every differentiable function is continuous. The converse is not true, e.g.

$$f(x)=\begin{cases}0 & \text{if } x\le 0\\ x & \text{if } x>0\end{cases}$$

is continuous but not differentiable. Most easily-described continuous functions are, like this one, non-differentiable only at a few points. But, it is possible to construct continuous functions which are nowhere differentiable. Surprisingly, there is a technical sense in which "almost all" continuous functions are nowhere differentiable.

An example of a continuous nowhere differentiable function is the ***Weierstrass function***. You may look it up on Wikipedia.

1.3.2 *Equation of the tangent line*

We want to find an equation of the tangent line to f at a. Consider the secant through $(a, f(a))$ and $(a+h, f(a+h))$. Squeezing $h \to 0$ the secant approaches the tangent line and the slope of the secant becomes the slope of the tangent,

$$\lim_{h\to 0} \frac{f(a+h)-f(a)}{h} = [\text{the slope of the tangent to } f \text{ at } a]$$

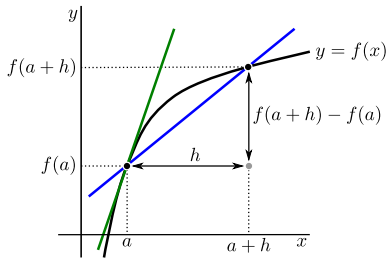
We know that the tangent line has slope $f'(a)$ and passes through the point $(a, f(a))$. If (x, y) is any other point on the tangent line, then

$$\frac{y-f(a)}{x-a} = f'(a)$$

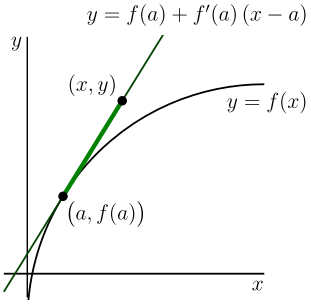
Cross multiplying gives us the equation of the tangent line:

> The ***tangent line*** to f at a is given by the equation
>
> $$y = f(a) + f'(a)\,(x-a)$$
>
> provided the derivative $f'(a)$ exists.

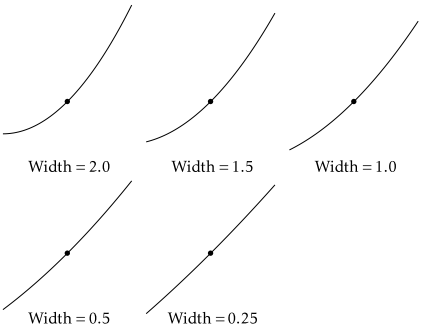
The tangent line is also known at the ***linearisation*** of f about a or the ***local linear approximation*** to f about a. This is because a differentiable function in the neighbourhood of any point looks very much like a straight line.



The blue line is the secant through $(a, f(a))$ and $(a+h, f(a+h))$, and the green line is the tangent line to f at a. When $h \to 0$, the secant becomes the tangent.



$f(x) = x^2$ under a microscope at $x = 1$:



1.3.3 *Higher-order derivatives*

If $f : D \to \mathbb{R}$ is differentiable then its derivative f' is also a function $f' : D \to \mathbb{R}$. If this is also differentiable, we denote its derivative by f'':

$$f'' = (f')'$$

and call this the ***second derivative***. We can similarly consider the ***third derivative*** f'''. This notation soon becomes unwieldy!

We also use the notation

$$f^{(0)} = f \qquad f^{(n)} = \left(f^{(n-1)}\right)'$$

so $f^{(n)}$ is the nth derivative of the function f (assuming $f^{(0)}, \ldots, f^{(n-1)}$ are all differentiable on the domain in question. Alternatively, we can write

$$\frac{df}{dx} \qquad \frac{d^2f}{dx^2} \qquad \frac{d^3f}{dx^3} \qquad \cdots \qquad \frac{d^nf}{dx^n}$$

The term ***order*** is often used: ***second-order derivative***, ***higher-order derivative***, etc.

1.3.4 *The rules of differentiation - a toolbox*

We will now build a collection of tools that help reduce the problem of computing the derivative of a complicated function to that of computing the derivatives of a number of simple functions. Suppose f, g are differentiable at x.

Additivity: $f+g$ is differentiable at x and

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

This corresponds to the fact that if we add two linear functions, then their slopes add: $(ah+b)+(ch+d) = (a+c)h+(b+d)$.

The product rule: fg is differentiable at x and

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

This is less obvious: if we multiply two linear functions then we have $(ah+b)(ch+d) = (ad+cb)h + bd + ach^2$. Provided h is small, this is well approximated by the linear part, with gradient $ad+cb$.

The chain rule: if f is differentiable at $g(x)$ then $f \circ g$ is differentiable at x and

$$(f \circ g)'(x) = f'(g(x))\,g'(x)$$

Here $f \circ g$ denotes the composition of functions, that is $(f \circ g)(x) = f(g(x))$.

This corresponds to the fact that if we compose linear functions then their gradients multiply: $a(cx+d)+b = (ac)x + ad + b$.

The quotient rule: if $g'(x) \neq 0$ then $(f/g)'(x)$ is differentiable at x and

$$(f/g)'(x) = \frac{f'(x)\,g(x) - f(x)\,g'(x)}{g(x)^2}$$

It is most easily derived from the product rule and the chain rule,

$$\begin{aligned}(f/g)'(x) &= f'(x)/g(x) + f(x)\,(1/g)'(x) \\ &= f'(x)\,g(x)/g^2(x) + f(x)(-1/g^2(x))\,g'(x)\end{aligned}$$

In Leibniz notation these rules read as:

Additivity: $\frac{d}{dx}(f+g) = \frac{df}{dx} + \frac{dg}{dx}$

The product rule: $\frac{d}{dx}fg = f\frac{dg}{dx} + g\frac{df}{dx}$

The chain rule: $\frac{d}{dx}f \circ g = \frac{df}{dg}\frac{dg}{dx}$

The quotient rule: $\frac{d}{dx}\frac{f}{g} = \frac{g\frac{df}{dx} - f\frac{dg}{dx}}{g^2}$

In the last step we have used that if $f(x) = 1/x$ then $f'(x) = -1/x^2$, which we have shown in Example 1.8.

Below we state two useful generalisations of the additivity and the product rules. Let $f_1, f_2, \ldots, f_n$ are differentiable at x.

Linearity: if $a_1, a_2, \ldots, a_n$ are any constants then $a_1 f_1(x) + \ldots + a_n f_n(x)$ is differentiable at x and

$$(a_1 f_1(x) + \ldots + a_n f_n(x))' = a_1 f_1'(x) + \ldots + a_n f_n'(x)$$

This is easily derived from the additivity and the product rule; we leave this as an exercise to the reader.

The power rule: if n is a natural number then $f^n(x)$ is differentiable at x and

$$(f^n(x))' = n f^{n-1}(x)\,f'(x)$$

When $f(x) = x$ the power rule gives

$$(x^n)' = n\,x^{n-1}$$

This follows directly from the product rule,

$$\begin{aligned}(f_1(x)\,f_2(x)\cdots f_n(x))' = {} & f_1'(x)\,f_2(x)\cdots f_n(x) \\ & + f_1(x)\,f_2'(x)\,f_3(x)\cdots f_n(x) \\ & \cdots \\ & + f_1(x)\cdots f_{n-1}(x)\,f_n'(x)\end{aligned}$$

Setting $f_1(x) = f_2(x) = \ldots = f_n(x) = f(x)$ gives the wanted result. In fact, the power rule is valid for any rational power. We will not provide a proof of this fact.

l'Hôpital's rule: Suppose f and g are differentiable at x_0, $f(x_0) = g(x_0) = 0$ and $g'(x_0) \neq 0$. Then

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \frac{f'(x_0)}{g'(x_0)}$$

This is a basic form of l'Hôpital's rule, a useful device for calculating limits of the "0/0" variety.

Example 1.10: Complicated derivatives

We want to find derivatives of the following functions:

$$f_1(x) = 5x - 4 \qquad f_2(x) = x \cdot (5x - 4) \qquad f_3(x) = 2x^3 + 4x^5$$
$$f_4(x) = \frac{x}{3x + 2} \qquad f_5(x) = \sqrt{x^2 - 1} \qquad f_6(x) = \frac{x}{2x + \frac{1}{3x+1}}$$

The derivatives of f_1, f_2 and f_3 are found using linearity, product and power rules:

$$f_1'(x) = (5x - 4)' = (5x)' - (4)' = 5 - 0 = 5$$
$$f_2'(x) = (x \cdot (5x - 4))' = x' \cdot (5x - 4) + x \cdot (5x - 4)' = 1 \cdot (5x - 4) + x \cdot 5 = 10x - 4$$
$$f_3'(x) = (2x^3 + 4x^5)' = 2 \cdot (x^3)' + 4 \cdot (x^5)' = 2 \cdot 3x^{3-1} + 4 \cdot 5x^{5-1} = 6x^2 + 20x^4$$

To find the derivative of f_4 we need to use the quotient rule together with linearity:

$$f_4'(x) = \left(\frac{x}{3x + 2}\right)' = \frac{x' \cdot (3x + 2) - x \cdot (3x + 2)'}{(3x + 2)^2} = \frac{3x + 2 - x \cdot 3}{(3x + 2)^2} = \frac{2}{(3x + 2)^2}$$

Function f_5 involves a square root. Taking the square root is the same as taking the $1/2$ power. Using the power rule and the chain we find:

$$f_5'(x) = ((x^2 - 1)^{1/2})' = 1/2 \cdot (x^2 - 1)^{1/2-1} \cdot (x^2 - 1)' = 1/2 \cdot (x^2 - 1)^{-1/2} \cdot 2x = \frac{x}{\sqrt{x^2 - 1}}$$

To find the derivative of f_6 we denote its denominator by $g(x) = 2x + \frac{1}{3x+1}$ giving

$$f_6'(x) = \left(\frac{x}{g(x)}\right)' = \frac{x' \cdot g(x) - x \cdot g'(x)}{g^2(x)}$$

where

$$g'(x) = \left(2x + (3x + 1)^{-1}\right)' = 2 + (-1) \cdot (3x + 1)^{-2} \cdot (3x + 1)' = 2 - \frac{3}{(3x + 1)^2}$$

giving

$$x' \cdot g(x) - x \cdot g'(x) = \left(2x + \frac{1}{3x + 1}\right) - x \cdot \left(2 - \frac{3}{(3x + 1)^2}\right) = \frac{1}{3x + 1} + \frac{3x}{(3x + 1)^2} = \frac{6x + 1}{(3x + 1)^2}$$

and

$$g^2(x) = \left(2x + \frac{1}{3x + 1}\right)^2 = \left(\frac{2x(3x + 1) + 1}{3x + 1}\right)^2 = \frac{(6x^2 + 2x + 1)^2}{(3x + 1)^2}$$

Therefore

$$f_6'(x) = \frac{\frac{6x+1}{(3x+1)^2}}{\frac{(6x^2+2x+1)^2}{(3x+1)^2}} = \frac{6x + 1}{(6x^2 + 2x + 1)^2}$$

1.4 Application of differentiation

Lecture 4

1.4.1 Approximating functions near a specified point

Suppose that you are interested in the values of some function $f(x)$ for x near some fixed point a. When the function is a polynomial or a rational function we can use some arithmetic (and maybe some hard work) to write down the answer. For example, if

$$f(x) = \frac{x^2 - 3}{x^2 - 2x + 4}$$

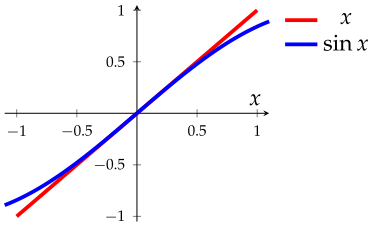
then $f(1/5)$ is

$$f(1/5) = \frac{1/25 - 3}{1/25 - 2/5 + 4} = \frac{(1 - 75)/25}{(1 - 10 + 100)/75} = -\frac{74}{91}$$

Tedious, but we can do it. On the other hand if you are asked to compute $\sin(1/10)$ then what can we do? We know that a calculator can work it out

$$\sin(1/10) = 0.09983341\ldots$$

But how does the calculator do this? How did people compute this before calculators? A hint comes from the following sketch of $\sin(x)$ for x around 0. The figure shows that the curves $y = x$ and $y = \sin(x)$ are almost the same when x is close to 0. Hence if we want the value of $\sin(1/10)$ we could just use this approximation $y = x$ to get



$$\sin(1/10) \approx 0.1$$

Of course, in this case we simply observed that one function was a good approximation of the other. We need to know how to find such approximations more systematically. More precisely, say we are given a function $f(x)$ that we wish to approximate close to some point $x = a$. Then

> *We need to find function $F(x)$ that*
> - *is simple and easy to compute,*
> - *is a good approximation to $f(x)$ for x values close to a.*

Further, we would like to understand how good our approximation is. Namely we need to be able to estimate the error $|f(x) - F(x)|$. There are many different ways to approximate a function and we will discuss one family of approximations: ***Taylor polynomials***. This is an infinite family of ever improving approximations, and our starting point is the very simplest.

1.4.2 Zeroth approximation – the constant approximation

The simplest functions are those that are constants. And our zeroth approximation will be by a constant function. That is, the approximating function will have the form

$$F(x) = A$$

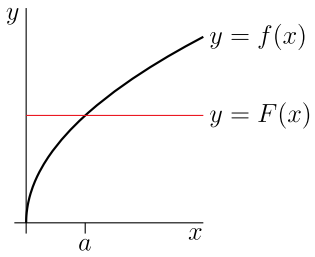
for some constant A. Notice that this function is a polynomial of degree zero. To ensure that $F(x)$ is a good approximation for x close to a, we choose A so that $f(x)$ and $F(x)$ take exactly the same value when $x = a$, that is $A = f(a)$. Therefore

The zeroth approximation of $f(x)$ is

$$f(x) \approx f(a)$$

An important point to note is that we need to know $f(a)$ – if we cannot compute that easily then we are not going to be able to proceed. We will often have to choose a (the point around which we are approximating $f(x)$) with some care to ensure that we can compute $f(a)$.

The figure on the right shows the graphs of a typical $f(x)$ and approximating function $F(x) = f(a)$. For x very near a, the values of $f(x)$ and $F(x)$ remain close together. But the quality of the approximation deteriorates fairly quickly as x moves away from a. Clearly we could do better with a straight line that follows the slope of the curve. That is our next approximation.



1.4.3 *First approximation – the linear approximation*

Our first approximation improves on our zeroth approximation by allowing the approximating function to be a linear function of x rather than just a constant function. That is,

$$F(x) = A + Bx$$

for some constants A and B. To ensure that $F(x)$ is a good approximation for x close to a, we still require that $f(x)$ and $F(x)$ have the same value at $x = a$ (that was our zeroth approximation). Our additional requirement is that their tangent lines at $x = a$ have the same slope – that the derivatives of $f(x)$ and $F(x)$ are the same at $x = a$. Hence

$$\begin{aligned} F(x) = A + Bx \quad &\Longrightarrow \quad F(a) = A + Ba = f(a) \\ F'(x) = B \quad &\Longrightarrow \quad F'(a) = \quad B = f'(a) \end{aligned}$$

Hence we must have $B = f'(a)$ and $A = f(a) - Ba = f(a) - f'(a) \cdot a$. Therefore

$$F(x) = (f(a) - af'(a)) + f'(a) \cdot x = f(a) + f'(a) \cdot (x - a)$$

We write it in this form because we can now clearly see that our first approximation is just an extension of our zeroth approximation.

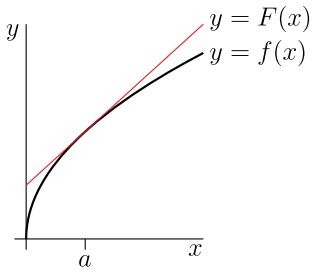
The linear approximation of $f(x)$ is

$$f(x) \approx f(a) + f'(a) \cdot (x - a)$$

This is the local linear approximation of $f(x)$ at $x = a$ we met in Lecture 3.

We should again stress that in order to form this approximation we need to know $f(a)$ and $f'(a)$ – if we cannot compute them easily then we are not going to be able to proceed.

Recall from Lecture 3 that $y = f(a) + f'(a) \cdot (x - a)$ is exactly the equation of the tangent line to the curve $y = f(x)$ at a. The figure on the right shows the graphs of a typical $f(x)$ and the approximating function $F(x) = f(a) + f'(a) \cdot (x - a)$. Observe that the graph of $F(x)$ remains close to the graph of $f(x)$ for a much larger range of x than did the graph of our constant approximation. One can also see that we can improve this approximation if we can use a function that curves down rather than being perfectly straight. That is our next approximation.



1.4.4 *Second approximation – the quadratic approximation*

We next develop a still better approximation by now allowing the approximating function be to a quadratic function of x. That is,

$$F(x) = A + Bx + Cx^2$$

for some constants A, B and C. To ensure that $F(x)$ is a good approximation for x close to a, we choose A, B and C so that

- $f(a) = F(a)$ – just as in our zeroth approximation,
- $f'(a) = F'(a)$ – just as in our first approximation,
- $f''(a) = F''(a)$ – this is a new condition.

These conditions give us the following equations

$$\begin{aligned} F(x) = A + Bx + Cx^2 &\implies F(a) = A + Ba + Ca^2 = f(a) \\ F'(x) = B + 2Cx &\implies F'(a) = B + 2Ca = f'(a) \\ F''(x) = 2C &\implies F''(a) = 2C = f''(a) \end{aligned}$$

We need to solve these for C first, then B and finally A:

$$\begin{aligned} C &= \tfrac{1}{2} f''(a) \\ B &= f'(a) - 2Ca = f'(a) - f'(a) \cdot a \\ A &= f(a) - Ba - Ca^2 = f(a) - \big(f'(a) - f'(a) \cdot a\big) \cdot a - \tfrac{1}{2} f''(a) \cdot a^2 \end{aligned}$$

Then put things back together to build up $F(x)$:

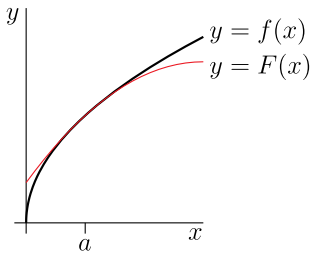
$$\begin{aligned} F(x) &= \underbrace{f(a) - f'(a) \cdot a - f'(a) \cdot a^2 - \tfrac{1}{2} f''(a) \cdot a^2}_{=A} + \underbrace{f'(a) \cdot x - f'(a) \cdot ax}_{=Bx} + \underbrace{\tfrac{1}{2} f''(a) x^2}_{Cx^2} \\ &= f(a) + f'(a) \cdot (x - a) + \tfrac{1}{2} f''(a) \cdot (x - a)^2 \end{aligned}$$

We again write it in this form because we can now clearly see that our second approximation is just an extension of our first approximation. Our second approximation is called the quadratic approximation:

> *The quadratic approximation of $f(x)$ is*
>
> $$f(x) \approx f(a) + f'(a) \cdot (x - a) + \tfrac{1}{2} f''(a) \cdot (x - a)^2$$

We stress once again that we need to know $f(a)$, $f'(a)$ and $f''(a)$ – if we cannot compute them easily then we are not going to be able to proceed.

The figure on the right shows graphs of a typical $f(x)$ and the approximating function $F(x) = f(a) + f'(a) \cdot (x-a) + \frac{1}{2} f''(a) \cdot (x-a)^2$. This new approximation looks better than both the first and second. It also gives an idea for a still better approximation.



1.4.5 *Still better approximations – Taylor polynomials*

We will use the same strategy to generate still better approximations by polynomials of any degree we like. As was the case with the approximations above, we determine the coefficients of the polynomial by requiring, that at the point $x = a$, the approximation and its first n derivatives agree with those of the original function.

Rather than simply moving to a cubic polynomial, let us try to write things in a more general way. We will consider approximating the function $f(x)$ using a polynomial, $T_n(x)$, of degree n,

$$\begin{aligned} T_n(x) &= c_0 + c_1(x-a) + c_2(x-a)^2 + \ldots + c_n(x-a)^n \\ &= \sum_{k=0}^{n} c_k (x-a)^k \end{aligned}$$

The above form makes it very easy to evaluate this polynomial and its derivatives at $x = a$:

$$\begin{aligned} T_n(x) &= c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + \ldots + c_n(x-a)^n \\ T_n'(x) &= c_1 + 2c_2(x-a) + 3c_3(x-a)^2 + \ldots + nc_n(x-a)^{n-1} \\ T_n''(x) &= 2c_2 + 6c_3(x-a) + \ldots + (n-1)nc_n(x-a)^{n-2} \\ T_n'''(x) &= 6c_3 + \ldots + (n-2)(n-1)nc_n(x-a)^{n-3} \\ &\vdots \\ T_n^{(n)}(x) &= n!\, c_n \end{aligned}$$

where $n! = 1 \cdot 2 \cdots (n-2) \cdot (n-1) \cdot n$ is the factorial of n. (Note that $0! = 1$.) Now notice that when we substitute $x = a$ into the above expressions only the constant terms survive and we get

$$\begin{aligned} T_n(a) &= c_0 &&= 0! \cdot c_0 & \qquad T_n'''(a) &= 6c_0 = 3! \cdot c_3 \\ T_n'(a) &= c_1 &&= 1! \cdot c_1 & \qquad &\vdots \\ T_n''(a) &= 2c_2 &&= 2! \cdot c_2 & \qquad T_n^{(n)}(a) &= n! \cdot c_n \end{aligned}$$

So now if we want to set the coefficients of $T_n(x)$ so that it agrees with $f(x)$ at $x = a$ then we need

$$T_n(a) = c_0 = f(a) \quad \Longrightarrow \quad c_0 = f(a) = \frac{1}{0!} f(a)$$

We also want the first n derivatives of $T_n(x)$ to agree with the derivatives of $f(x)$ at $x = a$, so

$$T_n^{(k)}(a) = k! \cdot c_k = f^{(k)}(a) \quad \Longrightarrow \quad c_k = \frac{1}{k!} f^{(k)}(a)$$

for $k = 1, 2, \ldots, n$. Putting this all together we have

The n-th order approximation of $f(x)$ is

$$f(x) \approx T_n(x)$$

where $T_n(x)$ is the n-th degree ***Taylor polynomial*** *for $f(x)$ about $x = a$,*

$$T_n(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(a) \cdot (x-a)^k$$

The special case $a = 0$ is called the ***Maclaurin polynomial***.

Before we proceed with examples, a couple of remarks are in order:

- While we can compute a Taylor polynomial about any a-value (providing the derivatives exist), in order to be a useful approximation, we must be able to compute $f^{(k)}(a)$ for $k = 1 \dots n$ easily. This means we must choose the point a with care. Indeed for many functions the choice $a = 0$ is very natural – hence the prominence of Maclaurin polynomials.
- If we have computed the approximation $T_n(x)$, then we can readily extend this to the next Taylor polynomial $T_{n+1}(x)$ since

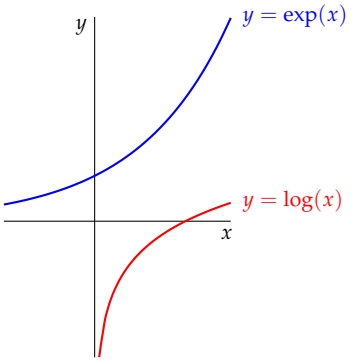
  $$T_{n+1}(x) = T_n(x) + \frac{1}{(n+1)!} f^{(n+1)}(a) \cdot (x-a)^{n+1}$$

  This is very useful if we discover that $T_n(x)$ is an insufficient approximation, because then we can produce $T_{n+1}(x)$ without having to start again from scratch.

1.4.6 *Approximation of some elementary functions*

The exponential function and the logarithm.

There is a unique function $\exp : \mathbb{R} \to \mathbb{R}$ *such that* $\exp' = \exp$ *and* $\exp(0) = 1$. *Any function $f : \mathbb{R} \to \mathbb{R}$ such that $f' = f$ is a constant multiple of* exp.



Consequences:

- $\exp(x+y) = \exp(x)\exp(y)$ for any $x, y \in \mathbb{R}$.
- $\exp(x) > 0$ for all $x \in \mathbb{R}$ and exp is strictly increasing, tends to 0 at $-\infty$ and tends to ∞ at $+\infty$.
- If we define $e = \exp(1)$ then we have $\exp(q) = e^q$ for any $q \in \mathbb{Q}$.
- There is an inverse function $\log : (0, \infty) \to \mathbb{R}$ such that $\log(\exp(x)) = x$ and $\exp(\log(y)) = y$ for all $x \in \mathbb{R}$ and $y \in (0, \infty)$.
- log is differentiable on $(0, \infty)$ and $\log'(y) = 1/y$.
- $\log(1) = 0$ and $\log(ab) = \log(a) + \log(b)$ for $a, b \in (0, \infty)$.

Taylor polynomial approximation:

- $\exp(x) \approx \sum_{k=0}^{n} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \dots$
- $\log(1+x) \approx \sum_{k=0}^{n} \frac{(-1)^k}{k+1} x^{k+1} = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots$

In the limit $n \to \infty$ the approximation becomes precise for $x \in \mathbb{R}$ for exp, and for $|x| < 1$ for log.

Arbitrary powers. This agrees with, and extends, the earlier definition of x^q for $q \in \mathbb{Q}$:

> *For $x \in (0, \infty)$ and $y \in \mathbb{R}$, we define*
>
> $$x^y = \exp(y \log x)$$
>
> *We also adopt the conventions that $0^y = 0$ for $y > 0$ and $0^0 = 1$.*

Consequences:

- $\log(x^y) = y \log(x)$ for $x > 0$.
- $(x^y)^z = x^{yz}$ and $x^{y+z} = x^y x^z$.
- The derivatives w.r.t. x and y of x^y are

$$\frac{d}{dx} x^y = y x^{y-1} \qquad \frac{d}{dy} x^y = x^y \log(x)$$

Trigonometric functions.

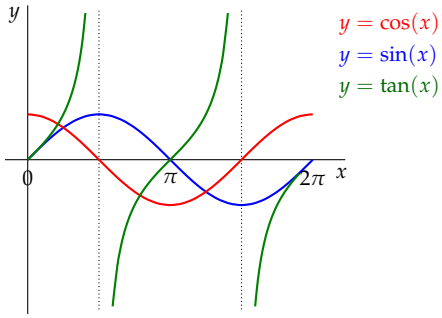
> *There is a unique function* $\sin : \mathbb{R} \to \mathbb{R}$ *such that* $\sin'' = -\sin$, $\sin(0) = 0$ *and* $\sin'(0) = 1$.

Consequences:

- Define $\cos = \sin'$, then $\cos' = -\sin$.
- Define $\tan = \sin / \cos$, then $\tan' = 1 + \tan^2$.
- sin is an odd function and cos is an even function, that is

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

- Define π to be the smallest number such that $\pi > 0$ and $\sin(\pi) = 0$. Then sin and cos are 2π-periodic (and 2π is the minimal period).
- All the known trigonometric identities can be derived starting from this definition of sin and cos.



Some special values useful to remember:

θ	$\sin(\theta)$	$\cos(\theta)$	$\tan(\theta)$
0	0	1	0
$\frac{\pi}{6}$	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{3}}{3}$
$\frac{\pi}{4}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{2}}{2}$	1
$\frac{\pi}{3}$	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$	$\sqrt{3}$
$\frac{\pi}{2}$	1	0	undefined

Taylor polynomial approximation:

- $\sin(x) \approx \sum_{k=0}^{n} \frac{(-1)^k}{(2k+1)!} x^{2k+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots$
- $\cos(x) \approx \sum_{k=0}^{n} \frac{(-1)^k}{(2k)!} x^{2k} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots$

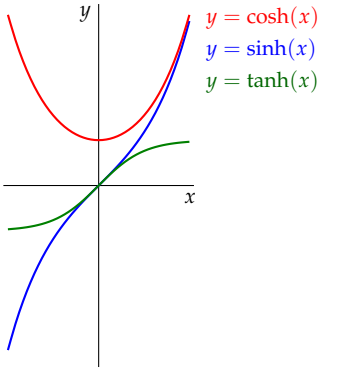
In the limit $n \to \infty$ the approximation becomes precise for $x \in \mathbb{R}$.

Hyperbolic functions.

> $$\sinh(x) = \frac{e^x - e^{-x}}{2} \qquad \cosh(x) = \frac{e^x + e^{-x}}{2}$$

Consequences:

- $\sinh' = \cosh$ and $\cosh' = \sinh$.
- Define $\tanh = \sinh / \cosh$, then $\tanh' = 1 - \tanh^2$.



- Hyperbolic functions satisfy a number of useful identities, e.g.

$$\cosh^2(x) - \sinh^2(x) = 1$$
$$\sinh(2x) = 2\sinh(x)\cosh(x) \quad \cosh(2x) = \cosh^2(x) + \sinh^2(x)$$

Taylor polynomial approximation:

- $\sinh(x) \approx \sum_{k=0}^{n} \frac{x^{2k+1}}{(2k+1)!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \dots$
- $\cosh(x) \approx \sum_{k=0}^{n} \frac{x^{2k}}{(2k)!} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \dots$

In the limit $n \to \infty$ the approximation becomes precise for $x \in \mathbb{R}$.

1.4.7 The error in the Taylor polynomial approximations

Any time you make an approximation, it is desirable to have some idea of the size of the error you introduced. That is, we would like to know the difference (remainder) $R(x)$ between the original function $f(x)$ and our approximation $F(x)$:

$$R(x) = f(x) - F(x)$$

Of course if we know $R(x)$ exactly, then we could recover $f(x) = F(x) + R(x)$ – so this is an unrealistic hope. In practice we would simply like to bound $R(x)$:

$$|R(x)| = |f(x) - F(x)| \leq M$$

where (hopefully) M is some small number. It is worth stressing that we do not need the tightest possible value of M, we just need a relatively easily computed M that isn't too far off the true value of $|f(x) - F(x)|$.

We will now develop a formula for the error introduced by the constant approximation,

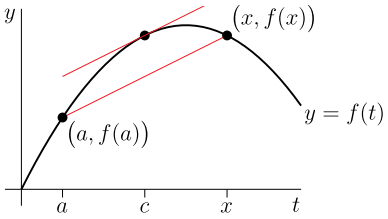
$$f(x) \approx f(a) = T_0(x)$$

The resulting formula can be used to get an upper bound on the size of the error $|R(x)|$. The main ingredient we will need is the ***Mean-Value Theorem*** (MVT):

> *Let f be a function that is defined and continuous on an interval $[a, b]$. Suppose that f is differentiable on (a, b). Then there is a number $c \in (a, b)$ such that*
>
> $$f'(c) = \frac{f(b) - f(a)}{b - a}$$

Consider the following obvious statement:

$$\begin{aligned} f(x) &= \underbrace{f(a)}_{=T_0(x)} + (f(x) - f(a)) \cdot \underbrace{\frac{x-a}{x-a}}_{=1} \\ &= T_0(x) + \frac{f(x) - f(a)}{x - a} \cdot (x - a) \end{aligned} \tag{1.4}$$



The coefficient $\frac{f(x)-f(a)}{x-a}$ of $(x-a)$ is the average slope of $f(t)$ as t moves from $t=a$ to $t=x$. We can picture this as the slope of the secant joining the points $(a, f(a))$ and $(x, f(x))$ in the sketch on the right. As t moves a to x, the instantaneous slope $f'(t)$ keeps changing. Sometimes $f'(t)$ might be larger than the average slope $\frac{f(x)-f(a)}{x-a}$, and sometimes $f'(t)$ might be smaller than $\frac{f(x)-f(a)}{x-a}$. However, by the MVT, there must be some number c strictly between a and x, for which $f'(c) = \frac{f(x)-f(a)}{x-a}$ exactly. Substituting this into formula (1.4) gives

$$f(x) = T_0(x) + \underbrace{f'(c)\cdot(x-a)}_{R_0(x)} \tag{1.5}$$

for some c strictly between a and x. The MVT does not tell us the value of c, however we do know that it lies strictly between x and a. So if we can get a good bound on $f'(c)$ on this interval then we can get a good bound on the error.

There are formulae similar to equation (1.5), that can be used to bound the error in higher order approximations; all are based on generalisations of the MVT. The next one – for the linear approximation – is

$$f(x) = T_1(x) + \underbrace{\tfrac{1}{2}f''(c)\cdot(x-a)^2}_{R_1(x)} \tag{1.6}$$

for some c strictly between a and x. More generally,

$$f(x) = T_n(x) + \underbrace{\frac{1}{(n+1)!}f^{(n+1)}(c)\cdot(x-a)^{n+1}}_{R_n(x)} \tag{1.7}$$

for some c strictly between a and x. The $R_n(x)$ stated above is called the ***Lagrange form of the remainder***. We will make use of the formula (1.7) in the next lecture.

Example 1.11: Estimating e

We want to estimate the value of e. Consider the linear approximation of $f(x) = e^x$ about $a = 0$:

$$f(x) \approx T_1(x) = f(0) + f'(0)\cdot x = 1 + x$$

So at $x = 1$ we have $e \approx T_1(1) = 2$. The error term in this approximation is

$$R_1(x) = \tfrac{1}{2}f''(c)\cdot x^2 = \tfrac{1}{2}\,e^c\cdot x^2 \quad\Longrightarrow\quad R_1(1) = \tfrac{1}{2}\,e^c$$

for some $c \in (0,1)$. Since e^x is an increasing function, it follows that $e^c < e$. Therefore

$$e = T_1(1) + R_1(1) = 2 + \tfrac{1}{2}\,e^c < 2 + \tfrac{1}{2}\,e \quad\Longrightarrow\quad e < 4$$

This isn't as tight as we would like – so now do the same with the quadratic approximation about $a = 0$:

$$f(x) \approx T_2(x) = f(0) + f'(0)\cdot x + \tfrac{1}{2}f''(0)\cdot x^2 = 1 + x + \tfrac{1}{2}x^2$$

So at $x = 1$ we have $e \approx T_2(1) = 1 + 1 + 1/2 = 5/2$. The error term in this approximation is

$$R_2(x) = \tfrac{1}{6}f'''(c)\cdot x^3 = \tfrac{1}{6}\,e^c\cdot x^3 \quad\Longrightarrow\quad R_2(1) = \tfrac{1}{6}\,e^c$$

for some $c \in (0,1)$. By the same arguments as before,

$$e = T_2(1) + R_2(1) = \tfrac{5}{2} + \tfrac{1}{6}\,e^c < \tfrac{5}{2} + \tfrac{1}{6}\,e \quad \Longrightarrow \quad e < 3$$

which is fairly close to the actual value, $e = 2.71828\ldots$

Example 1.12: Estimating sin

We want to estimate the value of $\sin 46^\circ$ using Taylor polynomials about $a = 45^\circ$, and estimate the resulting error. We start by rewriting a in terms of radians: $45^0 = 45/180 \cdot \pi = \pi/4$. Then

$$\begin{aligned} f(x) &= \sin(x) &\Longrightarrow \quad f(a) &= 1/\sqrt{2} \\ f'(x) &= \cos(x) &\Longrightarrow \quad f'(a) &= 1/\sqrt{2} \\ f''(x) &= -\sin(x) &\Longrightarrow \quad f''(a) &= -1/\sqrt{2} \\ f'''(x) &= -\cos(x) &\Longrightarrow \quad f'''(a) &= -1/\sqrt{2} \end{aligned}$$

The constant, linear and quadratic Taylor approximations for $\sin(x)$ about $\pi/4$ are

$$\begin{aligned} T_0(x) &= f(a) &&= 1/\sqrt{2} \\ T_1(x) &= T_0(x) + f'(a)\cdot(x-a) &&= 1/\sqrt{2} + 1/\sqrt{2}\cdot(x-\pi/4) \\ T_2(x) &= T_1(x) + \tfrac{1}{2}f''(a)\cdot(x-a)^2 &&= 1/\sqrt{2} + 1/\sqrt{2}\cdot(x-\pi/4) - 1/(2\sqrt{2})\cdot(x-\pi/4)^2 \end{aligned}$$

The approximations for $\sin(46^\circ)$ are

$$\begin{aligned} \sin(46^\circ) &= T_0(46\pi/180) = 1/\sqrt{2} &&= 0.70710678 \\ \sin(46^\circ) &= T_1(46\pi/180) = 1/\sqrt{2} + 1/\sqrt{2}\cdot(46\pi/180 - \pi/4) &&= 0.71944812 \\ \sin(46^\circ) &= T_2(46\pi/180) = 1/\sqrt{2} + 1/\sqrt{2}\cdot(x-\pi/4) - 1/(2\sqrt{2})\cdot(x-\pi/4)^2 &&= 0.71934042 \end{aligned}$$

The errors in those approximations are (respectively)

$$\begin{aligned} R_0(a) &= f'(c)\cdot(x-a) &&= \cos(c)\cdot\pi/180 \\ R_1(a) &= \tfrac{1}{2}f''(c)\cdot(x-a)^2 &&= -\tfrac{1}{2}\sin(c)\cdot(\pi/180)^2 \\ R_2(a) &= \tfrac{1}{6}f'''(c)\cdot(x-a)^3 &&= -\tfrac{1}{6}\cos(c)\cdot(\pi/180)^3 \end{aligned}$$

In each of these three cases c must lie somewhere between 45° and 46°. However, independently of the value of c, $\sin(c) \le 1$ and $\cos(c) \le 1$. Therefore

$$\begin{aligned} R_0(a) &\le \pi/180 &&< 0.018 \\ R_1(a) &\le -\tfrac{1}{2}\cdot(\pi/180)^2 &&< 0.00015 \\ R_2(a) &\le -\tfrac{1}{6}\cdot(\pi/180)^3 &&< 0.0000009 \end{aligned}$$

1.5 Optimisation

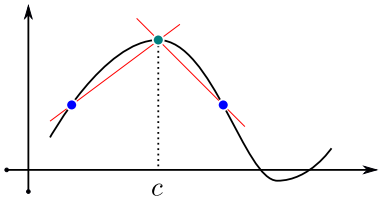
Lecture 5

An important application of differential calculus is finding the extremal value of some function: the minimum cost, the maximum probability, and so on. This is an often scenario in Data Science.

1.5.1 Local and global maxima and minima

Suppose that the maximum value of $f(x)$ is $f(c)$. Then for all "nearby" points, the function should be smaller. Consider the derivative $f'(c)$:

$$f'(c) = \lim_{h \to 0} \frac{f(c+h) - f(c)}{h}$$



Split the above limit into the left and right limits:

- Consider points to the right of $x = c$, then for all $h > 0$,

$$f(c+h) \leq f(c) \quad \Longrightarrow \quad \frac{f(c+h) - f(c)}{h} \leq 0$$

  Squeezing $h \to 0$ we get (provided the limit exists)

$$\lim_{h \to 0+} \frac{f(c+h) - f(c)}{h} \leq 0$$

- Consider points to the left of $x = c$, then for all $h < 0$,

$$f(c+h) \leq f(c) \quad \Longrightarrow \quad \frac{f(c+h) - f(c)}{h} \geq 0$$

  Squeezing $h \to 0$ we get (provided the limit exists)

$$\lim_{h \to 0-} \frac{f(c+h) - f(c)}{h} \geq 0$$

The "$\leq$" became "$\geq$" because we divided by h which is negative, i.e. dividing $1 \leq 2$ by -1 gives $-1 \geq -2$

Hence if the derivative $f'(c)$ exists, then the above right- and left-hand limits must agree, forcing $f'(c) = 0$.

Using similar reasoning one can also show that $f'(c) = 0$ if $f(c)$ is the minimum value of $f(x)$.

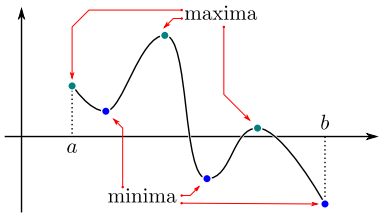
> *If $f(c)$ is the minimum or the maximum value of $f(x)$ and $f'(c)$ exists, then $f'(c) = 0$.*

Geometrically, the tangent line at the minimum or the maximum point of the curve must be horizontal, i.e. its slope is zero.

Notice two things about the above reasoning:

- Firstly, in order for the argument to work we only need that $f(x) < f(c)$ for x close to c – it does not matter what happens for x values far from c.
- Secondly, in the above argument we had to consider $f(x)$ for x both to the left of and to the right of c. If the function $f(x)$ is defined on a closed interval $[a, b]$, then the above arguments only apply when $a < c < b$ – not when c is either of the endpoints a and b.

Consider the function on the right. This function has only 1 maximum value (the middle green point in the graph) and 1 minimum value (the rightmost blue point), however it has 4 points at which the derivative is zero. In the small intervals around those points where the derivative is zero, we can see that function is ***locally*** a maximum or minimum, even if it is not the ***global*** maximum or minimum. We clearly need to be more careful distinguishing between these cases.

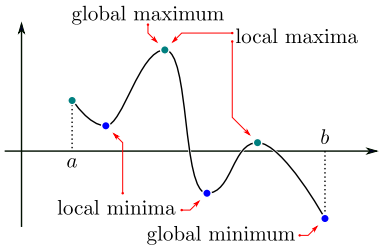


Let $a \leq b$ and let the function $f(x)$ be defined for all $x \in [a, b]$. Assume that $a \leq c \leq b$. Then

- *We say that $f(x)$ has a **global** (or absolute) **minimum** at $x = c$ if $f(x) \geq f(c)$ for all $a \leq x \leq b$.*
- *Similarly, we say that $f(x)$ has a **global** (or absolute) **maximum** at $x = c$ if $f(x) \leq f(c)$ for all $a \leq x \leq b$.*

Now assume that $a < c < b$. Then

- *We say that $f(x)$ has a **local minimum** at $x = c$ if there are a' and b' obeying $a \leq a' < c < b' \leq b$ such that $f(x) \geq f(c)$ for all x obeying $a' < x < b'$.*
- *Similarly, we say that $f(x)$ has a **local maximum** at $x = c$ if there are a' and b' obeying $a \leq a' < c < b' \leq b$ such that $f(x) \leq f(c)$ for all x obeying $a' < x < b'$.*



Consider again the function we showed on the right. It has 2 local maxima and 2 local minima. The global maximum occurs at the middle green point (which is also a local maximum), while the global minimum occurs at the rightmost blue point (which is not a local minimum).

We can summarise what we have learned so far:

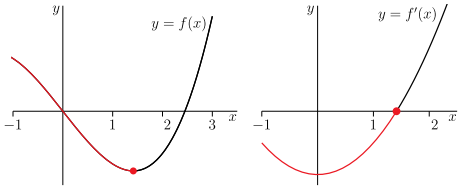
If a function $f(x)$ has a local minimum or maximum at $x = c$ and if $f'(c)$ exists, then $f'(c) = 0$.

It is often the case that, when $f(x)$ has a local maximum at $x = c$, the function $f(x)$ increases as x approaches c from the left and decreases as x leaves c to the right. That is, $f'(x) > 0$ for x just to the left of c and $f'(x) < 0$ for x just to the right of c. Similarly, it is often the case that, when $f(x)$ has a local minimum at $x = c$, $f'(x) < 0$ for x just to the left of c and $f'(x) > 0$ for x just to the right of c. Hence when $f(x)$ has a local maximum or minimum at $x = c$, there are two possibilities:

- the derivative $f'(c) = 0$, and
- the derivative $f'(c)$ does not exist.

This demonstrates that the points at which the derivative is zero or does not exist are very important. We will thus give names to these points.

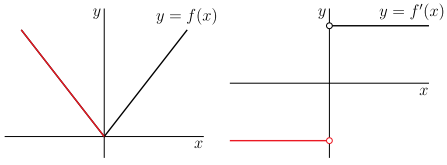
Possibility 1: e.g. $f'(x)$ changes continuously from negative to positive at the local minimum, taking the value zero at the local minimum (the red dot).



Let $f(x)$ be a function and let c be a point in its domain. Then

- *if $f'(c)$ exists and is zero we call $x = c$ a **critical point** of the function, and*
- *if $f'(c)$ does not exist then we call $x = c$ a **singular point** of the function.*

Possibility 2: e.g. $f'(x)$ changes discontinuously from negative to positive at the local minimum ($x = 0$) and $f'(0)$ does not exist.

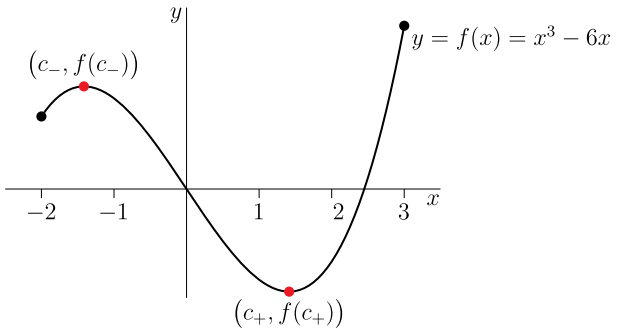


We'll now look at a few simple examples involving local maxima and minima, critical points and singular points. Then we will move on to global maxima and minima.

Example 1.13: Local minima and maxima

We want to find local and global minima and maxima of the function $f(x) = x^3 - 6x$ on $[-2, 3]$.

- First, we compute the derivative:
$$f'(x) = 3x^2 - 6$$
Since this is a polynomial it is defined everywhere on the domain and so there will be no singular points. So we now look for critical points.
- We rewrite the derivative as
$$f'(x) = 3x^2 - 6 = 3(x^2 - 2) = 3(x - \sqrt{2})(x + \sqrt{2})$$
Hence there two are critical points, $x = c_- = -\sqrt{2}$ and $x = c_+ = \sqrt{2}$.
- Finally, we need to determine the type of the critical points. For this we need to look at the values of $f(x)$ nearby the critical points. Thus let us sketch $f(x)$ on $[-2, 3]$:



From the sketch we see that:
  - $f(x)$ has a local minimum at $x = c_+$, i.e. $f(x) \geq f(c_+)$ for x close to c_+,
  - $f(x)$ has a local maximum at $x = c_-$, i.e. $f(x) \leq f(c_+)$ for x close to c_-,
  - the global minimum of $f(x)$ on $[-2, 3]$ is at $x = c_+$, i.e. $f(x) \geq f(c_+)$ for $x \in [-2, 3]$,
  - the global maximum of $f(x)$ on $[-2, 3]$ is at $x = 3$, i.e. $f(x) \leq f(3)$ for $x \in [-2, 3]$.
- Note that we have carefully constructed this example to illustrate that the global maximum (or minimum) of a function on an interval may or may not also be a local maximum (or minimum).
- This example also illustrates why we are looking for minima and maxima on closed intervals. The function $f(x)$ does have a global maximum on the open interval $(-2, 3)$.

Next, we want to find local and global minima and maxima of the function $f(x) = |x|$ on $[-1, 1]$.

- The derivative is
$$f'(x) = \begin{cases} -1 & \text{if } x < 0 \\ \text{undefined} & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$
- This derivative never takes the value 0, so the function does not have any critical points. However the derivative does not exist at the point $x = 0$, so that point is a singular point.
- Since $|x| \geq 0$ we know that $x = 0$ is the global minimum. Moreover, the global maximum is attained at $x = -1$ and $x = 1$ points.
- This example illustrates that we need to consider both critical points and singular points when we look for maxima and minima. It also illustrates that there does not need to be a unique global minimum or maximum.

The example above illustrates the following statement:

If $f(x)$ has a global minimum or maximum for $a \leq x \leq b$, at $x = c$, then there are 3 possibilities. Either

- $f'(c) = 0$, *or*
- $f'(c)$ *does not exist, or*
- $c = a$ *or* $c = b$.

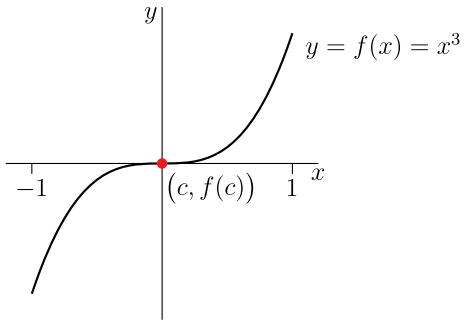
That is, a global maximum or minimum must occur either at a critical point, a singular point or at the endpoints of the interval.

We are left with one more scenario, illustrated in the next example.

Example 1.14: Inflection point

We want to find local and global minima and maxima of the function $f(x) = x^3$ on $[-1, 1]$.

- First, we compute the derivative: $f'(x) = 3x^2$. It is defined everywhere on the domain and so there will be no singular points but there may be critical points.
- The derivative is zero only when $x = 0$, so $x = c = 0$ is the only critical point.
- We need to determine the type of the critical point. We sketch the function:



  From the sketch we see that:
  - $f(x)$ has no local minimum or maximum at $x = c$ despite the fact that $f'(c) = 0$ – we have $f(x) < f(c) = 0$ for all $x < c = 0$ and $f(x) > f(c) = 0$ for all $x > c = 0$,
  - the global minimum of $f(x)$ on $[-1, 1]$ is at $x = -1$, i.e. $f(x) \geq f(-1)$ for $x \in [-1, 1]$,
  - the global maximum of $f(x)$ on $[-1, 1]$ is at $x = 1$, i.e. $f(x) \leq f(1)$ for $x \in [-1, 1]$.
- This example illustrates that a critical point need not be a local maximum or minimum for the function. The point $x = c = 0$ in this case is called an ***inflection point***.

To summarise, the strategy of finding the global minimum and maximum is thus as follows:

Let $f(x)$ be a continuous function on $[a, b]$. Then to find the global maximum and minimum of the function:

- *Compute the function at all the critical points, singular points, and endpoints.*
- *Evaluate $f(c)$ for each c in that list. The largest (or smallest) of those values is the largest (or smallest) value of $f(x)$ on $[a, b]$.*

1.5.2 *Second derivative test*

The second derivative $f''(x)$ tells us the rate at which the derivative changes. It can be used to determine the type of a critical point under certain conditions.

> *Suppose f is a twice differentiable function defined on $[a,b] \subset \mathbb{R}$ and let $x_0 \in (a,b)$ be such that $f'(x_0) = 0$. Then:*
> - *if $f''(x_0) > 0$ then x_0 is a local minimum,*
> - *if $f''(x_0) < 0$ then x_0 is a local maximum,*
> - *if $f''(x_0) = 0$ then the test is inconclusive.*

To see why this works, write down the Taylor expansion of f about x_0 to first order with the Lagrange form of the remainder:

$$f(x) = f(x_0) + f'(x_0) \cdot (x - x_0) + \tfrac{1}{2} f''(c) \cdot (x - x_0)^2$$

for some c between x and x_0. Since $f'(x_0) = 0$, this reduces to

$$f(x) = f(x_0) + \tfrac{1}{2} f''(c) \cdot (x - x_0)^2$$

Now, if x is close to, but not equal to, x_0 then $f''(c)$ is close to $f''(x_0)$. It follows that if $f''(x_0) < 0$ then $f''(c) < 0$ and if $f''(x_0) > 0$ then $f''(c) > 0$. Because $(x - x_0)^2 > 0$, we have $f(x) > f(x_0)$ if $f''(x_0) > 0$ and $f(x) < f(x_0)$ if $f''(x_0) < 0$. That is, x_0 is a local minimum if $f''(x_0) > 0$ and a local maximum if $f''(x_0) < 0$. But if $f''(x_0) = 0$, then any behaviour is possible.

Example 1.15: Second derivative test

We want to find local minima and maxima of the function $f(x) = \exp(-x) \cdot (x^3 + 2x + 2)$ on $\mathbb{R}$.

- We start by computing the first derivative (using the product rule):

$$\begin{aligned} f'(x) &= (\exp(-x))' \cdot (x^3 + 2x + 2) + \exp(-x) \cdot (x^3 + 2x + 2)' \\ &= -\exp(-x) \cdot (x^3 + 2x + 2) + \exp(-x) \cdot (3x^2 + 2) \\ &= -\exp(-x) \cdot (x^3 - 3x^2 + 2x) \end{aligned}$$

  Since $e^{-x} > 0$ and $x^3 - 3x^2 + 2x = x(x-1)(x-2)$, the critical points are $x_1 = 0$, $x_2 = 1$, $x_3 = 2$.

- We then compute the second derivative:

$$\begin{aligned} f''(x) &= -(\exp(-x))' \cdot (x^3 - 3x^2 + 2x) - \exp(-x) \cdot (x^3 - 3x^2 + 2x)' \\ &= \exp(-x) \cdot (x^3 - 3x^2 + 2x) - \exp(-x) \cdot (3x^2 - 6x + 2) \\ &= \exp(-x) \cdot (x^3 - 6x^2 + 8x - 2) \end{aligned}$$

  It remains to subsitute the critical values

$$\begin{aligned} f''(0) &= e^0 \cdot (-2) = -2 < 0 && \implies && x = 0 \text{ is a local maximum} \\ f''(1) &= e^{-1} \cdot (1 - 6 + 8 - 2) = e^{-1} > 0 && \implies && x = 1 \text{ is a local minimum} \\ f''(2) &= e^{-2} \cdot (8 - 24 + 16 - 2) = -2e^{-2} < 0 && \implies && x = 2 \text{ is a local maximum} \end{aligned}$$

- This example illustrates that the local minima and maxima always alternate; there can only be inflection points in between.

1.5.3 *Multivariable calculus*

So far we were concerned with functions that map the real numbers to real number, $f : \mathbb{R} \to \mathbb{R}$, sometimes called "real functions of one (or single) variable", meaning the "input" is a single real number and the "output" is likewise a single real number.

In this last part on this lecture we turn to real functions of multiple variables,

$$f : D \to \mathbb{R} \quad \text{where} \quad D \subseteq \mathbb{R}^n$$

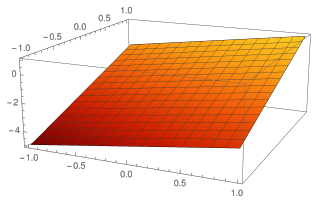
Here $\mathbb{R}^n$ denotes n copies of $\mathbb{R}$. We will deal primarily with $n = 2$ and to a lesser extent $n = 3$; in fact many of the techniques we discuss can be applied to larger values of n as well.

When $n = 2$, a function $f : \mathbb{R}^2 \to \mathbb{R}$ maps a pair of values (x, y) to single real number. Such function can be used to define surfaces in the three dimensional space. For example

$$z = f(x, y) = x + 2y - 2$$

is an incline plane, and

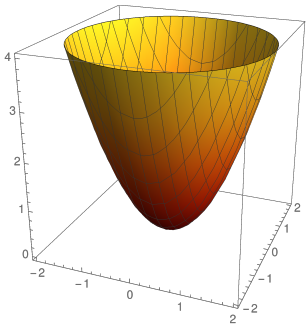
Incline plane $z = x + 2y - 2$



$$z = f(x, y) = x^2 + y^2$$

is a paraboloid.

Paraboloid $z = x^2 + y^2$



When $n \geq 3$, a function $f : \mathbb{R}^n \to \mathbb{R}$ maps an n-tuple of values $(x_1, x_2, \ldots, x_n)$ to a single real number. The principal difficulty with such functions is visualizing them, as they do not "fit" in the three dimensions we are familiar with. For three variables there are various ways to interpret functions that make them easier to understand. For example, $f(x, y, z)$ could represent the temperature at the point (x, y, z), or the pressure, or the strength of a magnetic field. It is often useful to consider those points at which $f(x, y, z) = k$, where k is some fixed value. If $f(x, y, z)$ is temperature, the set of points $(x, y, z) \in \mathbb{R}^3$ such that $f(x, y, z) = k$ is the collection of points in space with temperature k; in general this is called a ***level set***.

Our goal is to develop a method for finding local minimum and maximum of multivariable functions. We start by introducing the notion of a partial derivative.

> *Suppose $D \subseteq \mathbb{R}^n$ and $f : D \to \mathbb{R}$. Let $(a_1, a_2, \ldots, a_n) \in D$. The* ***partial derivative*** *of f at $(a_1, a_2, \ldots, a_n)$ with respect to its i-th variable is*
>
> $$\frac{\partial f}{\partial x_i}(a_1, \ldots, a_n) = \lim_{h \to 0} \frac{f(a_1, \ldots, a_i + h, \ldots, a_n) - f(a_1, \ldots, a_n)}{h}$$
>
> *provided the limit exists.*

The following notations are all used for "the partial derivative of f with respect to x":

$$\frac{\partial f}{\partial x} \quad \frac{\partial}{\partial x} f \quad \partial_x f \quad f'_x$$

The symbol "∂" has replaced the "d" used in the single variable calculus to remind us that there are more variables than x_i, but that only x_i is being treated as a variable. For example, the partial derivative $\partial/\partial y$ applied to $f(x, y, z)$ treats it as a function of a single variable y only.

The symbol "∂" is also very handy for denoting the higher-order partial derivatives. For example, the second-order derivatives of $f(x,y)$ are

$$\frac{\partial^2 f}{\partial x^2} \qquad \frac{\partial^2 f}{\partial x \partial y} \qquad \frac{\partial^2 f}{\partial y^2}$$

or equivalently

$$\partial_x^2 f \qquad \partial_x \partial_y f \qquad \partial_y^2 f$$

It is important to note that the order of differentiation does not matter. That's why $\partial y \partial x f$ is not present in the list above.

Example 1.16: Partial derivatives

Compute all first- and second-order partial derivatives of $f(x,y) = y^2 + 3xy + \sin(xy)$.

- The first-order partial derivatives are

$$\begin{aligned}\partial_x f(x,y) &= \partial_x(y^2) + \partial_x(3xy) + \partial_x(\sin(xy)) \\ &= 0 + 3y + \cos(xy) \cdot \partial_x(xy) \\ &= 3y + y\cos(xy)\end{aligned}$$

$$\begin{aligned}\partial_y f(x,y) &= \partial_y(y^2) + \partial_y(3xy) + \partial_y(\sin(xy)) \\ &= 2y + 3x + \cos(xy) \cdot \partial_y(xy) \\ &= 2y + 3x + x\cos(xy)\end{aligned}$$

- The second-order partial derivatives are

$$\begin{aligned}\partial_x^2 f(x,y) &= \partial_x(3y + y\cos(xy)) \\ &= 0 - y\sin(xy) \cdot \partial_x(xy) \\ &= -y^2\sin(xy)\end{aligned}$$

$$\begin{aligned}\partial_x \partial_y f(x,y) &= \partial_x(2y + 3x + x\cos(xy)) \\ &= 3 + (\partial_x x) \cdot \cos(xy) + x \cdot \partial_x \cos(xy) \\ &= 3 + \cos(xy) - xy\sin(xy)\end{aligned}$$

$$\begin{aligned}\partial_y^2 f(x,y) &= \partial_y(2y + 3x + x\cos(xy)) \\ &= 2 + 0 - x\sin(xy) \cdot \partial_y(xy) \\ &= 2 - x^2\sin(xy)\end{aligned}$$

Compute the first-order partial derivatives of $g(x,y) = xy\log(xy) + c$ where c is a fixed number.

- Using $\log(xy) = \log(x) + \log(y)$ we find

$$\begin{aligned}\partial_x g(x,y) &= \partial_x(xy\log(x) + xy\log(y)) + \partial_x c \\ &= (\partial_x xy) \cdot \log(x) + xy \cdot \partial_x \log(x) + y\log(y) + 0 \\ &= y\log(x) + \frac{xy}{x} + y\log(y) \\ &= y + y\log(xy)\end{aligned}$$

- Since $g(x,y)$ is symmetric with respect to x and y, we can immediately deduce that

$$\partial_y g(x,y) = x + x\log(xy)$$

Suppose a surface given by $z = f(x, y)$ has a local maximum at (x_0, y_0). Geometrically, this point on the surface looks like the top of a hill. If we look at the cross-section in the plane $y = y_0$, we will see a local maximum on the curve at $f(x, y_0)$ at $x = x_0$, and we know from the single-variable calculus that $\partial_x f = 0$ at this point. Likewise, in the plane $x = x_0$, we must have $\partial_y f = 0$ at $y = y_0$. So if there is a local maximum at (x_0, y_0). Both partial derivatives must be zero at the same point, and likewise for a local minimum. Thus, to find local maximum and minimum points, we need only consider those points at which both partial derivatives are zero.

However, as in the single-variable case, it is possible for the derivatives to be zero at a point that is neither a maximum or a minimum, so we need to test these points further. For this we use a two-variable analogue of the second derivative test.

Suppose f is a twice differentiable function defined on $D \subset \mathbb{R}^2$ and let (x_0, y_0) be a point inside D such that $\partial_x f(x_0, y_0) = \partial_y f(x_0, y_0) = 0$. Denote by Δ the discriminant

$$\Delta = (\partial_x^2 f(x_0, y_0)) \cdot (\partial_y^2 f(x_0, y_0)) - (\partial_x \partial_y f(x_0, y_0))^2$$

Then:

- *if $\Delta > 0$ and $\partial_x^2 f(x_0, y_0) > 0$ then (x_0, y_0) is a local minimum,*
- *if $\Delta > 0$ and $\partial_x^2 f(x_0, y_0) < 0$ then (x_0, y_0) is a local maximum,*
- *if $\Delta < 0$ then (x_0, y_0) is neither a local minimum nor a maximum,*
- *if $\Delta = 0$ then the test is inconclusive.*

Recall that when we did single-variable global maximum and minimum problems, we restricted analysis to a closed interval, for then we simply had to check all critical values and the endpoints. We will make a similar assumption for two-variable functions.

If $f(x, y)$ is continuous on a closed and bounded subset of $\mathbb{R}^2$, then it has both a global minimum and maximum.

As in the case of single variable functions, this means that the maximum and minimum values must occur at a critical point or on the boundary; in the two variable case, however, the boundary is a curve, not merely two endpoints.

Example 1.17: Local minima and maxima of a surface

Find all local maxima and minima for $f(x, y) = x^2 + y^2$.

- First, we compute all the necessary derivatives:

$$\partial_x f(x, y) = 2x \qquad \partial_y f(x, y) = 2y \qquad \partial_x^2 f(x, y) = 2 \qquad \partial_y^2 f(x, y) = 2 \qquad \partial_x \partial_y f(x, y) = 0$$

- Equating $\partial_x f(x, y) = \partial_y f(x, y) = 0$ we find that there is a single critical point, at $(0, 0)$.
- Applying the second derivative test we find

$$\Delta = (\partial_x^2 f(0, 0)) \cdot (\partial_y^2 f(0, 0)) - (\partial_x \partial_y f(0, 0))^2 = 2 \cdot 2 - 0^2 = 4 > 0$$

Since $\partial_x^2 f(0, 0)) = 2 > 0$ the point $(0, 0)$ is a local minimum.

Example 1.18: Local minima and maxima of a surface

Find all local maxima and minima for $g(x,y) = x^2 - y^2$.

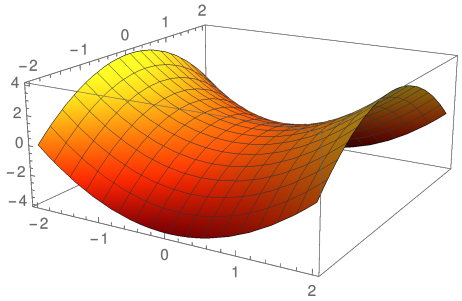
- First, we compute all the necessary derivatives:

$$\partial_x g(x,y) = 2x \qquad \partial_y g(x,y) = -2y \qquad \partial_x^2 g(x,y) = 2 \qquad \partial_y^2 g(x,y) = -2 \qquad \partial_x \partial_y g(x,y) = 0$$

- Again, there is a single critical point, at $(0,0)$.
- Applying the second derivative test we find

$$\Delta = (\partial_x^2 g(0,0)) \cdot (\partial_y^2 g(0,0)) - (\partial_x \partial_y g(0,0))^2 = 2 \cdot (-2) - 0^2 = -4 < 0$$

Thus the point $(0,0)$ is neither a minimum nor a maximum. It is called a ***saddle point***.



Find all local maxima and minima for $h(x,y) = x^3 + y^3$.

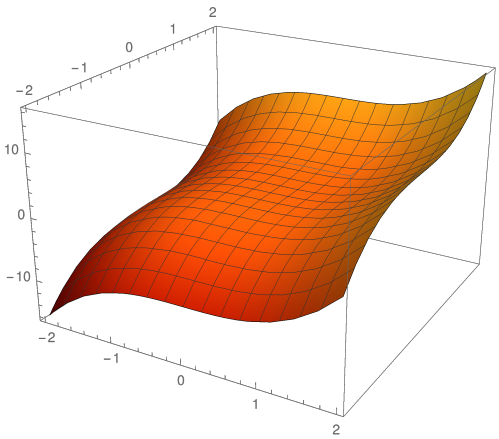
- Again, we compute all the necessary derivatives:

$$\partial_x h(x,y) = 3x^2 \qquad \partial_y h(x,y) = 3y^2 \qquad \partial_x^2 h(x,y) = 6x \qquad \partial_y^2 h(x,y) = 6y \qquad \partial_x \partial_y h(x,y) = 0$$

- There is a single critical point, at $(0,0)$.
- Applying the second derivative test we find

$$\Delta = (\partial_x^2 h(0,0)) \cdot (\partial_y^2 h(0,0)) - (\partial_x \partial_y h(0,0))^2 = 0 \cdot 0 - 0^2 = 0$$

Hence the test is inconclusive. This can be seen from inspecting the cross-sections $f(x,0) = x^3$ and $f(0,y) = y^3$ which we already know that do not have either a minimum or a maximum.



2
Linear Algebra

Lecture 6

Additional Reading:

- ***J. Brownlee***, *Basics of Linear Algebra for Machine Learning*
- ***O. Mulita***, *Lecture Notes on Linear Algebra*
- ***C. Wendlandt***, *Linear Algebra*

2.1 Systems of linear equations

In this lecture we will introduce a class of equations called linear equations. The study of these equations is the main motivation behind the development of the theory of linear algebra and matrices.

2.1.1 Linear equations

*A **linear equation** in n unknowns $x_1, x_2, \ldots, x_n$ is an equation which can be put in the form*

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n = b \tag{2.1}$$

where $b, a_1, a_2, \ldots, a_n$ are fixed real numbers.

*A **solution** to (2.1) is a sequence $s_1, \ldots, s_n \in \mathbb{R}$, often written as an n-tuple $(s_1, \ldots, s_n)$, satisfying*

$$a_1s_1 + a_2s_2 + \ldots + a_ns_n = b$$

The terminology "linear" can be explained by the fact that such equations generalize the equation of a line – we will come back to this below. Linear equations are among the simplest and most well understood equations in mathematics. Let's begin with a few examples:

- The equation $0x_1 + \ldots + 0x_n = 0$ is linear. It is called the ***trivial*** linear equation. Every sequence $s_1, \ldots, s_n \in \mathbb{R}$ is a solution of it.
- For any $a, b \in \mathbb{R}$ the equation $y = ax + b$ is linear. It is just the equation of a line with slope a and intercept b.
- $x_1 + 2x_2 + \ldots + 100x_{100} = 5050$ is linear, with a solution given by $x_1 = x_2 = \ldots = x_{100} = 1$. Note that there are more solutions.
- $x^3 + 10x = 0$ and $x_1x_2 + x_3 + x_4 = 0$ are ***not*** linear equations.

Our focus will be not just on individual linear equations, but rather on systems of linear equations.

*An $m \times n$ **system of linear equations** is a set of m linear equations in n unknowns $x_1, x_2, \ldots, x_n$:*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &+ \ldots + a_{1n}x_n &&= b_1 \\ a_{21}x_1 + a_{22}x_2 &+ \ldots + a_{2n}x_n &&= b_2 \\ \vdots \quad &\quad \vdots \quad &\vdots \quad &\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 &+ \ldots + a_{mn}x_n &&= b_n \end{aligned} \tag{2.2}$$

where each a_{ij} and each b_i is a fixed real number.

*A **solution** to the system (2.2) is a sequence $s_1, \ldots, s_n \in \mathbb{R}$, often written as $(s_1, \ldots, s_n)$, which is a solution to each of the m equations in the system.*

*If the system (2.2) admits at least one solution, it is said to be **consistent**. Otherwise, it is said to be **inconsistent.***

Let's illustrate this definition with a few examples:

- The linear equation (2.1) is itself a $1 \times n$ linear system. It is inconsistent exactly when $(a_1, \ldots, a_n, b) = (0, \ldots, 0, b)$ with $b \neq 0$.
- The following is a 2×3 linear system:

$$\begin{aligned} x + 2y + z &= 1 \\ x + z &= 1 \end{aligned}$$

  Every solution of this system takes the form

$$x = s \qquad y = 0 \qquad z = 1 - s$$

  where s in an ***independent*** (of ***free***) variable, which can take any value. The above is called the ***general solution*** to the system, as it describes all of its solutions.
- An example of a 4×4 linear system is

$$\begin{aligned} x_1 + x_2 + 6x_4 &= 0 \\ 2x_2 + x_3 + x_4 &= 0 \\ x_1 + 2x_2 + 2x_3 + 5x_4 &= 0 \\ x_1 + 2x_2 + x_3 + 6x_4 &= 0 \end{aligned}$$

  It has a unique solution $0 = x_1 = x_2 = x_3 = x_4$. However, at this point it is unclear how the general solution to a larger system like this can be found (and $n = m = 4$ is still very small).

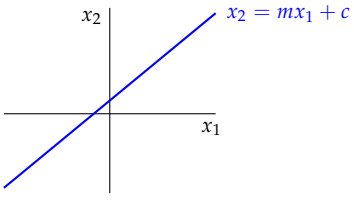
Before continuing with our study of linear systems in general, let's focus in on the special case where $n = 2$. In this setting, the corresponding systems and their solutions can be analysed geometrically on the plane $\mathbb{R}^2$. Indeed, each equation in such a system is either inconsistent, trivial, or of the form

$$a_1 x_1 + a_2 x_2 = b$$

with a_1 and a_2 not both zero. If $a_2 \neq 0$ the above is equivalent to

$$x_2 = m x_1 + c \quad \text{where} \quad m = -\frac{a_1}{a_2} \quad c = \frac{b}{a_2}$$

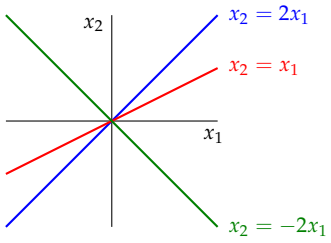
This is the equation of a line with slope m and intercept c. If instead $a_2 = 0$, the above equation is the vertical line $x_1 = b/a_1$.



Next, consider an $m \times 2$ system of consistent equations. A solution to such a system system is precisely a point (x_1, x_2) on the plane where the associated lines intersect. In particular, given a graph of the lines, one may solve the system geometrically by identifying the points of intersection. There can be several different scenarios that we will illustrate with simple examples:

- The 3×2 system

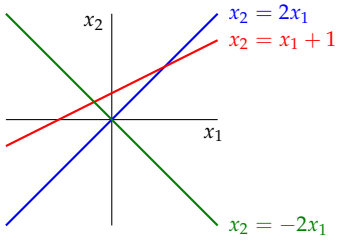
$$\begin{aligned} x_1 - x_2 &= 0 \\ 2x_1 + x_2 &= 0 \\ -2x_1 + x_2 &= 0 \end{aligned}$$



corresponds to the three lines $x_2 = x_1$, $x_2 = -2x_1$ and $x_2 = 2x_1$. It is clear from the graph on the right that there is a *unique solution*, which occurs at $(x_1, x_2) = (0, 0)$.

- The 3×2 system

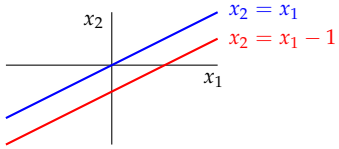
$$\begin{aligned} x_1 - x_2 &= -1 \\ 2x_1 + x_2 &= 0 \\ -2x_1 + x_2 &= 0 \end{aligned}$$



corresponds to the three lines $x_2 = x_1 + 1$, $x_2 = -2x_1$ and $x_2 = 2x_1$. Since there is no mutual intersection, the system has *no solution* and thus it is inconsistent.

- The 2×2 system

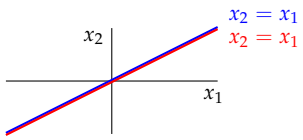
$$\begin{aligned} x_1 - x_2 &= 0 \\ 2x_1 - 2x_2 &= 2 \end{aligned}$$



corresponds to the two lines $x_2 = x_1$ and $x_2 = x_1 - 1$. These lines are parallel and the system has *no solution* – it is inconsistent.

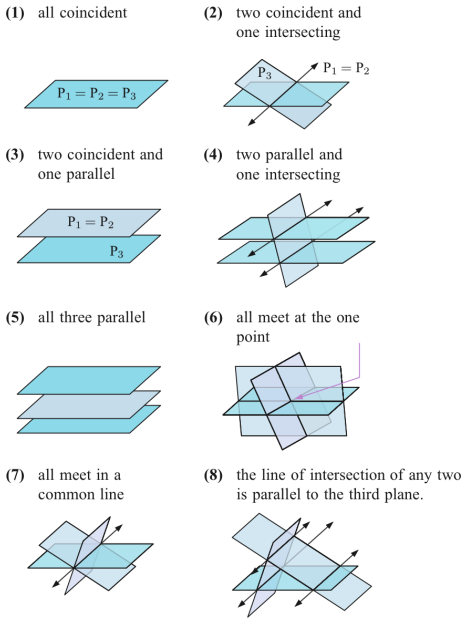
- The 2×2 system

$$\begin{aligned} x_1 - x_2 &= 0 \\ 3x_1 - 3x_2 &= 0 \end{aligned}$$



corresponds to a single line $x_2 = x_1$. Thus there are infinitely many points of intersection – every point $(x_1, x_2) = (s, s)$ with $s \in \mathbb{R}$ is on both lines, and hence a solution to both equations. Therefore, the system has *infinitely many solutions*.

When $n = 3$, lines in $\mathbb{R}^2$ are replaced with planes in $\mathbb{R}^3$. For example, solutions to a 3×3 system could have any of the eight arrangements shown on the right. When $n > 3$ it is typically near impossible to try to solve an $m \times n$ linear system geometrically. There are two standard high school methods for solving systems of linear equations algebraically:



- the ***substitution*** method, where you express variables in terms of the other variables and substitute the result in the remaining equations, and
- the ***elimination*** method, also called the Gauss-Jordan method.

The substitution method if efficient only for very small m and n. The elimination method is much more effective, thus we would like to briefly recall it.

- Consider a 2×2 system

$$2x_1 + x_2 = 1 \tag{E1}$$
$$4x_1 + 2x_2 = 1 \tag{E2}$$

  Replacing (E2) with (E2) $-\ 2 \times$ (E1) gives $0 = -1$. This means that the system is inconsistent.

- Consider a 2×2 system

$$2x_1 + x_2 = 1 \tag{E1$'$}$$
$$4x_1 + x_2 = 1 \tag{E2$'$}$$

  Replacing (E2$'$) with (E2$'$) $-$ (E1$'$) gives $2x_1 = 0$, and so

$$x_1 = 0 \tag{E3$'$}$$

  Replacing (E1$'$) with (E1$'$) $- 2 \times$ (E3$'$) gives $x_2 = 1$. Thus $(0, 1)$ is the unique solution.

- Consider a 2×2 system

$$2x_1 + x_2 = 1 \tag{E1$''$}$$
$$4x_1 + 2x_2 = 2 \tag{E2$''$}$$

  Replacing (E2$''$) with (E2$''$) $-$ (E1$''$) gives $0 = 0$, so (E2$''$) is redundant, meaning (E1$''$) is the only linear equation we need to solve. Put $x_2 = s$ where s is an independent variable. Then $x_1 = (1 - s)/2$ and thus $(x_1, x_2) = ((1 - s)/2, s)$ is the general solution.

We would like to make the elimination operations more mechanical by forgetting about the variable names $x_1, x_2, \ldots$, and replacing the whole process with a certain matrix computation.

2.1.2 *The augmented matrix*

Consider an $m \times n$ system of linear equations (2.2). Collect the fixed real numbers a_{ij} and b_i into an $m \times (n+1)$ ***augmented matrix*** $(A|\boldsymbol{b})$, that is,

The peculiar notation $(A|\boldsymbol{b})$ will become clear in Lecture 7.

$$(A|\boldsymbol{b}) = \left[\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} & b_n \end{array}\right]$$

The vertical line in the matrix is put there just to remind us that the rightmost column is different from the others, and arises from the constants on the right hand side of the equations.

Label the rows of $(A|\boldsymbol{b})$ by $\mathrm{r}_1, \mathrm{r}_2, \ldots, \mathrm{r}_m$. Motivated by the examples in the previous subsection we define three types of ***elementary row operations*** on $(A|\boldsymbol{b})$:

- Add a multiple of row r_j to row r_i. For example,

$$\left[\begin{array}{ccc|c} 3 & 1 & 9 & 0 \\ 4 & 6 & 7 & 2 \\ 2 & 5 & 8 & 11 \end{array}\right] \xrightarrow{\mathrm{r}_3 \to \mathrm{r}_3 - 3\mathrm{r}_1} \left[\begin{array}{ccc|c} 3 & 1 & 9 & 0 \\ 4 & 6 & 7 & 2 \\ -7 & 2 & -19 & 11 \end{array}\right]$$

- Interchange rows r_i and r_j. For example,

$$\left[\begin{array}{ccc|c} 3 & 1 & 9 & 0 \\ 4 & 6 & 7 & 2 \\ 2 & 5 & 8 & 11 \end{array}\right] \xrightarrow{\mathrm{r}_1 \leftrightarrow \mathrm{r}_3} \left[\begin{array}{ccc|c} 2 & 5 & 8 & 11 \\ 4 & 6 & 7 & 2 \\ 3 & 1 & 9 & 0 \end{array}\right]$$

- Multiply row r_i by a non-zero scalar. For example,

$$\left[\begin{array}{ccc|c} 3 & 1 & 9 & 0 \\ 4 & 6 & 7 & 2 \\ 2 & 5 & 8 & 11 \end{array}\right] \xrightarrow{\mathrm{r}_2 \to 4\mathrm{r}_2} \left[\begin{array}{ccc|c} 3 & 1 & 9 & 0 \\ 16 & 24 & 28 & 8 \\ 2 & 5 & 8 & 11 \end{array}\right]$$

Example 2.1: Elimination method

Suppose we want to solve the following system of linear equations:

$$\begin{aligned} 2x_1 - x_2 + 4x_3 - x_4 &= 1 \\ x_1 + 2x_2 + x_3 + x_4 &= 2 \\ x_1 - 3x_2 + 3x_3 - 2x_4 &= -1 \\ -3x_1 - x_2 - 5x_3 &= -3 \end{aligned}$$

We will apply the elementary row operations onto the augmented matrix of this system:

$$\left[\begin{array}{cccc|c} 2 & -1 & 4 & -1 & 1 \\ 1 & 2 & 1 & 1 & 2 \\ 1 & -3 & 3 & -2 & -1 \\ -3 & -1 & -5 & 0 & -3 \end{array}\right] \xrightarrow{\mathrm{r}_1 \to \frac{1}{2}\mathrm{r}_1} \left[\begin{array}{cccc|c} 1 & -1/2 & 2 & -1/2 & 1/2 \\ 1 & 2 & 1 & 1 & 2 \\ 1 & -3 & 3 & -2 & -1 \\ -3 & -1 & -5 & 0 & -3 \end{array}\right]$$

$$\xrightarrow[\mathrm{r}_4 \to \mathrm{r}_4 + 3\mathrm{r}_1]{\substack{\mathrm{r}_2 \to \mathrm{r}_2 - \mathrm{r}_1 \\ \mathrm{r}_3 \to \mathrm{r}_3 - \mathrm{r}_1}} \left[\begin{array}{cccc|c} 1 & -1/2 & 2 & -1/2 & 1/2 \\ 0 & 5/2 & -1 & 3/2 & 3/2 \\ 0 & -5/2 & 1 & -3/2 & -3/2 \\ 0 & -5/2 & 1 & -3/2 & -3/2 \end{array}\right] \xrightarrow[\mathrm{r}_4 \to \mathrm{r}_4 + \mathrm{r}_2]{\mathrm{r}_3 \to \mathrm{r}_3 + \mathrm{r}_2} \left[\begin{array}{cccc|c} 1 & -1/2 & 2 & -1/2 & 1/2 \\ 0 & 5/2 & -1 & 3/2 & 3/2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

$$\xrightarrow{\mathrm{r}_2 \to \frac{2}{5}\mathrm{r}_2} \left[\begin{array}{cccc|c} 1 & -1/2 & 2 & -1/2 & 1/2 \\ 0 & 1 & -2/5 & 3/5 & 3/5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \xrightarrow{\mathrm{r}_1 \to \mathrm{r}_1 + \frac{1}{2}\mathrm{r}_2} \left[\begin{array}{cccc|c} 1 & 0 & 9/5 & -1/5 & 4/5 \\ 0 & 1 & -2/5 & 3/5 & 3/5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

The original system has been transformed to the following equivalent system, that is, both systems have the same solutions:

$$\begin{aligned} x_1 + \tfrac{9}{5}x_3 - \tfrac{1}{5}x_4 &= \tfrac{4}{5} \\ x_2 - \tfrac{2}{5}x_3 + \tfrac{3}{5}x_4 &= \tfrac{3}{5} \end{aligned}$$

In a solution to the latter system, the variables x_3 and x_4 can take arbitrary values, say s and t. Then these equations tell us that

$$\begin{aligned} x_1 &= -\tfrac{9}{5}s + \tfrac{1}{5}t + \tfrac{4}{5} \\ x_2 &= \tfrac{2}{5}s - \tfrac{3}{5}t + \tfrac{3}{5} \end{aligned}$$

Hence the general solution is

$$(x_1, x_2, x_3, x_4) = \left(-\tfrac{9}{5}s + \tfrac{1}{5}t + \tfrac{4}{5},\ \tfrac{2}{5}s - \tfrac{3}{5}t + \tfrac{3}{5},\ s,\ t\right)$$

2.1.3 *Row reduction*

Let $A = (a_{ij})$ be an $m \times n$ matrix. For the ith row, let $c(i)$ denote the position of the first (leftmost) non-zero entry in that row. In other words, $a_{i,c(i)} \neq 0$ while $a_{ij} = 0$ for all $j < c(i)$. It will make things a little easier to write if we use the convention that $c(i) = \infty$ if the ith row is entirely zero.

We will describe a procedure, analogous to solving systems of linear equations by elimination, which starts with a matrix, performs certain row operations, and finishes with a new matrix in a special form. After applying this procedure, the resulting matrix will have the following properties:

(i) All zero rows are below all non-zero rows.

(ii) Each non-zero row has 1 as its *first* non-zero entry: $a_{i,c(i)} = 1$ for all $1 \leq i \leq s$, where s is the number of non-zero rows.

(iii) The first non-zero entry of each row is strictly to the right of the first non-zero entry of the row above, that is, $c(1) < \cdots < c(s)$.

Note that this implies that if row i is non-zero, then all entries below the first non-zero entry of row i are zero: $a_{k,c(i)} = 0$ for $k > i$.

> *A matrix satisfying properties* (i)–(iii) *above is said to be in a* ***row echelon form****.*

We can also add the following property:

(iv) If row i is non-zero, then all entries both above and below the first non-zero entry of row i are zero: $a_{k,c(i)} = 0$ for all $k \neq i$.

Note that this implies that if row i is non-zero, then all entries below the first non-zero entry of row i are zero: $a_{k,c(i)} = 0$ for $k > i$.

> *A matrix satisfying properties* (i)–(iv) *above is said to be in the* ***reduced row echelon form*** *or simply in the* ***row reduced form****.*

A row echelon form of a matrix will be used later to calculate the rank of a matrix. The row reduced form (the use of the definite article is intended: this form is, indeed, unique, though we shall not prove this here) is used to solve systems of linear equations. In this light, the following statement says that every system of linear equations can be solved by the Gauss-Jordan (Elimination) method.

> *Every matrix can be brought to the row reduced form by elementary row transformations.*

We will now describe an algorithm to achieve this. We need to show that:

- the algorithm must terminate after finitely many steps, and
- the resulting matrix must be in the row reduced form.

This will be clear from the nature of the algorithm. At any stage in the procedure we will be looking at the entry a_{ij} in a particular position (i, j) of the matrix. We will call (i, j) the ***pivot*** position, and

The following matrices are in a *row echelon form*:

$$\begin{bmatrix} 1 & 3 & 0 & 5 \\ 0 & 1 & 2 & 7 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 4 & 3 & 2 & 0 \\ 0 & 0 & 1 & 0 & 7 \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The following matrices are in the *row reduced form*:

$$\begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 4 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 7 \\ 0 & 0 & 0 & 1 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

In Python, the `sympy` (symbolic python) package provides a row reduction tool:

```
import sympy as sp
A = [[1,2,3],[2,3,4]]
sp.Matrix(A).rref()
```

It returns two elements:
– the reduced row form of A, and
– a tuple of indices of the pivot entries.

a_{ij} – the ***pivot entry***. We start with $(i, j) = (1, 1)$ and proceed as follows:

(1) If a_{ij} and all entries below it in its columns are zero, that is, if $a_{kj} = 0$ for all $k \geq i$, then move the pivot one place to the right, to $(i, j+1)$ and repeat Step 1, or terminate if $j = n$.

(2) If $a_{ij} = 0$ but $a_{kj} \neq 0$ for some $k > i$ then interchange rows r_i and r_k, i.e., apply $\mathrm{r}_i \leftrightarrow \mathrm{r}_k$.

(3) At this stage $a_{ij} \neq 0$. If $a_{ij} \neq 1$, then multiply row r_i by a_{ij}^{-1}, i.e., apply $\mathrm{r}_i \to a_{ij}^{-1}\mathrm{r}_i$.

(4) At this stage $a_{ij} = 1$. If, for any $k \neq i$, $a_{kj} \neq 0$, then subtract row r_i multiplied by a_{kj} from row r_k, i.e., apply $\mathrm{r}_k \to \mathrm{r}_k - a_{kj}\mathrm{r}_i$.

(5) At this stage, $a_{kj} = 0$ for all $k \neq i$. If $i = m$ or $j = n$ then terminate. Otherwise, move the pivot diagonally down to the right to $(i+1, j+1)$, and go back to Step 1.

If one needs only a row echelon form, this can done faster by replacing Steps 4 and 5 with weaker and faster steps as follows:

Though this algorithm always works, in practice there is often more efficient routes to obtaining the reduced row echelon form of a matrix, that involve skipping or combining several steps.

(4′) At this stage $a_{ij} = 1$. If, for any $k > i$, $a_{kj} \neq 0$, then subtract row r_i multiplied by a_{kj} from row r_k, i.e., apply $\mathrm{r}_k \to \mathrm{r}_k - a_{kj}\mathrm{r}_i$.

(5′) At this stage, $a_{kj} = 0$ for all $k > i$. If $i = m$ or $j = n$ then terminate. Otherwise, move the pivot diagonally down to the right to $(i+1, j+1)$, and go back to Step 1.

In the example below, we find a row echelon form of a matrix by applying the faster algorithm.

Example 2.2: Row echelon form

Matrix	Pivot	Step	Operation
$\begin{bmatrix} 2 & 4 & 2 & -4 & 2 \\ 0 & 0 & 1 & 2 & 1 \\ 3 & 6 & 3 & -6 & 3 \\ 1 & 2 & 3 & 3 & 3 \end{bmatrix}$	$(1,1)$	3	$\mathrm{r}_1 \to \frac{1}{2}\mathrm{r}_1$
$\begin{bmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 3 & 6 & 3 & -6 & 3 \\ 1 & 2 & 3 & 3 & 3 \end{bmatrix}$	$(1,1)$	4′	$\mathrm{r}_3 \to \mathrm{r}_3 - 3\mathrm{r}_1$ $\mathrm{r}_4 \to \mathrm{r}_4 - \mathrm{r}_1$
$\begin{bmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 5 & 2 \end{bmatrix}$	$(1,1) \to (2,2) \to (2,3)$	5′, 1, 4′	$\mathrm{r}_4 \to \mathrm{r}_4 - 2\mathrm{r}_2$
$\begin{bmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$	$(2,3) \to (3,4)$	5′, 2	$\mathrm{r}_3 \leftrightarrow \mathrm{r}_4$
$\begin{bmatrix} 1 & 2 & 1 & -2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$(3,4) \to (4,5)$	5′, 1	Terminate

The analysis above provides a proof of the following theorem:

Let $(A|\boldsymbol{b})$ be an $m \times (n+1)$ augmented matrix in the reduced row echelon form, with r nonzero rows. Then the associated $m \times n$ linear system is inconsistent if and only if the last nonzero row of B is

$$\left[\, 0 \;\cdots\; 0 \mid 1 \,\right]$$

Otherwise, the system is consistent with $r \leq n$ and it has

- *a unique solution if $r = n$, or*
- *infinitely many solutions with $(n - r)$ free variables if $r < n$.*

In these two cases, the system admits a general solution in which the r dependent variables correspond to the r columns with leading entries.

The theorem above implies the following useful corollary.

Suppose that $m < n$. Then any $m \times n$ consistent linear system has infinitely many solutions.

For example, a 9×10 linear system cannot possibly have a unique solution. Neither can a single linear equation in $n > 1$ unknowns. However, it is certainly possible for a 10×9 system to have a unique solution.

To summarise, to solve a given $m \times n$ system of linear equations, we need to:

- find the augmented matrix $(A|\boldsymbol{b})$ of the system;
- find the reduced row echelon form of $(A|\boldsymbol{b})$;
- write down the corresponding reduced linear system, and either conclude the system is inconsistent or find its general solution.

If at any point in the row reduction process one encounters a row of the form

$$[\,0 \;\cdots\; 0 \,|\, b\,] \quad \text{with} \quad b \neq 0$$

then one can stop the process and conclude the corresponding system is inconsistent.

2.1.4 *Homogeneous systems*

There is a special class of linear systems which are ubiquitous in mathematics and are always consistent. These are the so-called homogeneous systems.

*An $m \times n$ linear system is called **homogeneous** if it is of the form*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 \;+ \ldots + a_{1n}x_n \;&= 0 \\ a_{21}x_1 + a_{22}x_2 \;+ \ldots + a_{2n}x_n \;&= 0 \\ \vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \qquad\quad &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \;&= 0 \end{aligned} \tag{2.3}$$

*A homogeneus system always admits a **trivial solution** given by*

$$(x_1, \ldots, x_n) = (0, \ldots, 0)$$

*Any other solution is said to be **nontrivial**.*

If a homogeneous system admits a single nontrivial solution, then it automatically has infinitely many solutions.

Suppose that $m < n$. Then any $m \times n$ homogeneous linear system has infinitely many solutions.

2.1.5 *Column reduction*

In analogy to elementary row operations, one can introduce elementary column operations. Let A be an $m \times n$ matrix. Label the columns of A by $c_1, c_2, \ldots, c_n$. There are three types of ***elementary column operations*** on A:

- Add a multiple of column c_j to column c_i. For example,

$$\begin{bmatrix} 3 & 1 & 9 & 0 \\ 4 & 6 & 7 & 2 \end{bmatrix} \xrightarrow{c_3 \to c_3 - 3c_1} \begin{bmatrix} 3 & 1 & 0 & 0 \\ 4 & 6 & -5 & 2 \end{bmatrix}$$

- Interchange columns c_i and c_j. For example,

$$\begin{bmatrix} 3 & 1 & 9 & 0 \\ 4 & 6 & 7 & 2 \end{bmatrix} \xrightarrow{c_1 \leftrightarrow c_3} \begin{bmatrix} 9 & 1 & 3 & 0 \\ 7 & 6 & 4 & 2 \end{bmatrix}$$

- Multiply column c_i by a non-zero scalar. For example,

$$\begin{bmatrix} 3 & 1 & 9 & 0 \\ 4 & 6 & 7 & 2 \end{bmatrix} \xrightarrow{c_2 \to 4c_2} \begin{bmatrix} 3 & 1 & 36 & 0 \\ 4 & 6 & 28 & 2 \end{bmatrix}$$

Elementary column operations change a linear system and ***cannot*** be applied to solve a system of linear equations. However, they are useful for reducing a matrix to a very simple form.

By applying elementary row and column operations, an $m \times n$ matrix can be brought into the block form

$$\begin{bmatrix} I_s & 0_{s,n-s} \\ 0_{m-s,s} & 0_{m-s,n-s} \end{bmatrix} \tag{2.4}$$

where I_s is an $s \times s$ matrix with zeros everywhere except with 1's on the main diagonal, and $0_{k,l}$ denotes the $k \times l$ zero matrix.

An example of such a matrix is

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

Indeed, use elementary row operations to reduce A to row reduced form. Then all $a_{i,c(i)} = 1$. We can use these leading entries in each row to make all the other entries zero: for each $a_{ij} \neq 0$ with $j \neq c(i)$, replace c_j with $c_j - a_{ij}c_{c(i)}$. Finally the only nonzero entries of our matrix are $a_{i,c(i)} = 1$. Now for each number i starting from $i = 1$, exchange c_i and $c_{c(i)}$, putting all the zero columns at the right-hand side.

*A matrix in the form (2.4) is said to be in **row and column reduced form**. This is sometimes called the **Smith normal form**.*

The number of non-zero entries in the row and column reduced form of a matrix does not depend on the particular order that we apply elementary row and column operations. We will see later that this number represents an important property of a matrix.

Example 2.3: Row and column reduced form

Assume that a matrix is already in a row reduced form. The column reduction then gives:

$$\begin{bmatrix} 1 & 2 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix} \xrightarrow[c_5 \to c_5 - c_1]{c_2 \to c_2 - 2c_1} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{bmatrix} \xrightarrow[c_5 \to c_5 - 3c_4]{c_2 \leftrightarrow c_3} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \xrightarrow[c_5 \to c_5 - 2c_2]{c_3 \leftrightarrow c_4} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

2.2 *Matrices and vectors*

Lecture 7

So far, we have thought of matrices as merely a convenient way to encode systems of linear equations. Technically, we could have done everything up until this point without a matrix at all (perhaps a bit uncomfortably). We now begin the process of changing this viewpoint, and shifting our focus from linear systems to matrices themselves. We start by introducing basic operations on matrices and vectors.

2.2.1 *The basics*

For any $m, n \in \mathbb{N}$ we will use the following notation:

- $A \in \mathbb{R}^{m\times n}$ will denote a ***matrix*** with m rows and n columns, and with entries being real numbers. We always use upper case letters to denote matrices.
- The (i, j)-th entry of a matrix A will be denoted by $(A)_{ij}$ or a_{ij}.
- $\boldsymbol{x} \in \mathbb{R}^n$ will denote a ***vector*** with n entries, the entries being real numbers.
  - By convention, a vector with n entries is thought as a ***column vector*** – a matrix with n rows and 1 column.
  - If we want to explicitly represent a ***row vector*** – a matrix with 1 row and n columns – we will write $\boldsymbol{x}^T$, where T denotes the transpose operation, which we will explain a bit later.

  We will always use bold lower case letters to denote vectors.
- The i-th element of a vector $\boldsymbol{x}$ will be denoted by $(\boldsymbol{x})_i$ or x_i.

An $m \times n$ matrix:

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

An n-dimensional column vector

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

An n-dimensional row vector

$$\boldsymbol{x}^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

Other common notations for vectors are

$$\vec{x} \quad \underline{x} \quad \vec{x}^T \quad \underline{x}^T$$

A matrix $A \in \mathbb{R}^{m\times n}$ is called a ***square*** matrix if $m = n$. Otherwise, it is called a ***rectangular*** matrix. There are two types of square matrices that will be of a special interest to us:

- We will typically write $D = \text{diag}(d_1, d_2, \ldots, d_n) \in \mathbb{R}^{n\times n}$ to denote a ***diagonal matrix*** with d_i's on the diagonal and zeros everywhere else, that is, $(D)_{ij} = \delta_{ij} d_i$.
- We will write $I = \text{diag}(1, \ldots, 1) \in \mathbb{R}^{n\times n}$ to denote the ***identity matrix***, a diagonal matrix with ones on the diagonal and zeros everywhere else, that is, $(I)_{ij} = \delta_{ij}$.

The symbol δ_{ij} is called the "Kronecker delta". It is defined by

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

The basic ***binary operations*** on matrices and vectors are:

A *binary operation* is a rule that combines two elements to obtain another element.

- ***Matrix sum and difference***. Given $A, B \in \mathbb{R}^{m\times n}$ their sum is the matrix $C = A + B \in \mathbb{R}^{m\times n}$ with entries given by

$$c_{ij} = a_{ij} + b_{ij}$$

  Their difference is the matrix $C = A - B \in \mathbb{R}^{m\times n}$ with entries given by

$$c_{ij} = a_{ij} - b_{ij}$$

  For example,

$$\begin{bmatrix} 5 & 3 \\ 4 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 6 & 5 \\ 7 & 6 \end{bmatrix} \qquad \begin{bmatrix} 5 & 3 \\ 4 & 2 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 1 & -2 \end{bmatrix}$$

- ***Vector sum and difference***. Given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ their sum is the vector $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y} \in \mathbb{R}^n$ with entries given by

$$z_i = x_i + y_i$$

Their difference is the vector $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{y} \in \mathbb{R}^n$ with entries given by

$$z_i = x_i - y_i$$

For example,

$$\begin{bmatrix} 5 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 7 \end{bmatrix} \qquad \begin{bmatrix} 5 \\ 4 \end{bmatrix} - \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

Note that in order for the matrix sum or difference to exist, the matrices must be of the same size. Also note that subtracting a matrix from itself gives a zero matrix – a matrix with zero entries. The same is true for vectors. We will write $A - A = 0$ and $\boldsymbol{x} - \boldsymbol{x} = \boldsymbol{0}$. It is clear that $A + 0 = 0 + A = A$ and $\boldsymbol{x} + \boldsymbol{0} = \boldsymbol{0} + \boldsymbol{x} = \boldsymbol{x}$.

- ***Matrix multiplication***. Given $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ their product is the matrix $C = AB \in \mathbb{R}^{m \times p}$ with entries given by

$$c_{ij} = (AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

This sum is called the inner product of the i-th row of A and j-th column of B. For example,

$$\begin{bmatrix} 5 & 3 \\ 4 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 5 \cdot 1 + 3 \cdot 3 & 5 \cdot 2 + 3 \cdot 4 \\ 4 \cdot 1 + 2 \cdot 3 & 4 \cdot 2 + 2 \cdot 4 \end{bmatrix} = \begin{bmatrix} 14 & 22 \\ 10 & 16 \end{bmatrix}$$

Note that in order for the matrix product to exist, the number of columns in A must equal the number of rows in B.

- ***Matrix-vector multiplication***. Given $A \in \mathbb{R}^{m \times n}$ and $\boldsymbol{x} \in \mathbb{R}^n$, their product is a vector $\boldsymbol{y} = A\boldsymbol{x} \in \mathbb{R}^m$ with entries given by

$$y_i = (A\boldsymbol{x})_i = \sum_{j=1}^{n} A_{ij} x_j$$

For example,

$$\begin{bmatrix} 5 & 3 \\ 4 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 5 \cdot 1 + 3 \cdot 3 \\ 4 \cdot 1 + 2 \cdot 3 \end{bmatrix} = \begin{bmatrix} 14 \\ 10 \end{bmatrix}$$

- ***Vector-matrix multiplication***. Given $\boldsymbol{x} \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$, their product is a row vector $\boldsymbol{y}^T = \boldsymbol{x}^T A \in \mathbb{R}^m$ with entries given by

$$y_i = (\boldsymbol{x}^T A)_i = \sum_{i=1}^{n} x_j A_{ji}$$

For example,

$$\begin{bmatrix} 5 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 5 \cdot 1 + 3 \cdot 3 & 5 \cdot 2 + 3 \cdot 4 \end{bmatrix} = \begin{bmatrix} 14 & 22 \end{bmatrix}$$

Note that writing $\boldsymbol{x} \in \mathbb{R}^{n \times 1}$ and $\boldsymbol{x}^T \in \mathbb{R}^{1 \times n}$ the matrix-vector and the vector-matrix multiplications are just special cases of the matrix multiplication.

In Python, the `numpy` (numeric python) package provides matrix algebra tools:

```
import numpy as np

# Matrices:
A = np.array([[5,3],[4,2]])
B = np.array([[1,2],[3,4]])

# Column vector:
x = np.array([[1],[3]])

# Row vector:
y = np.array([[5,3]])

# Matrix sum and difference:
A + B
A - B

# Matrix multiplication:
A.dot(B)
A @ B

# Matrix-vector multiplication:
A.dot(x)
A @ x

# Vector-matrix multiplication:
y.dot(A)
y @ A

# Scalar multiplication:
3 * A

# Hadamard multiplication:
A * B

# Inner product:
y.dot(x)
y @ x

# Outer product:
x.dot(y)
x @ y

# Transpose:
np.transpose(A)

# Trace:
np.trace(A)

# Norm:
np.linalg.norm(x)
```

Vectors can also be defined as lists:

```
z = np.array([1,2,3,4])
w = np.array([5,6,7,8])
```

Such vectors do not have column and row representations and are treated "universally":

```
# Matrix-vector multiplication
z @ A

# Vector-matrix multiplication
A @ z

# Inner product:
z @ w
np.inner(z,w)

# Outer product:
np.outer(z,w)
```

- ***Scalar multiplication***. Given $A \in \mathbb{R}^{m\times n}$ and $\alpha \in \mathbb{R}$ the scalar multiplication of A by α gives the matrix $C = \alpha A \in \mathbb{R}^{m\times n}$ with entries given by

$$c_{ij} = \alpha a_{ij}$$

  In other words, we multiply each entry of A by α. For example,

$$7 \cdot \begin{bmatrix} 5 & 3 \\ 4 & 2 \end{bmatrix} = \begin{bmatrix} 7\cdot 5 & 7\cdot 3 \\ 7\cdot 4 & 7\cdot 2 \end{bmatrix} = \begin{bmatrix} 35 & 21 \\ 28 & 14 \end{bmatrix}$$

- ***Hadamard product***. Given $A, B \in \mathbb{R}^{m\times n}$ their entry-wise product is the matrix $C = A \odot B \in \mathbb{R}^{m\times n}$ with entries $c_{ij} = a_{ij}b_{ij}$. For example,

The Hadamard product is very seldom used in Mathematics, but is very common in Data Science.

$$\begin{bmatrix} 5 & 3 \\ 4 & 2 \end{bmatrix} \odot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 5\cdot 1 & 3\cdot 2 \\ 4\cdot 3 & 2\cdot 4 \end{bmatrix} = \begin{bmatrix} 5 & 6 \\ 12 & 8 \end{bmatrix}$$

  Note that the entry-wise product is commutative, $A \odot B = B \odot A$.

- ***Inner product***. Given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, their inner product is the real number $\boldsymbol{x}^T\boldsymbol{y} \in \mathbb{R}$ given by

$$\boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

  For example,

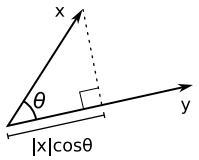
$$\begin{bmatrix} 5 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 3 \end{bmatrix} = 5\cdot 1 + 3\cdot 3 = 14$$

  Note that this is a special case of the matrix multiplication. It is also a generalisation of the usual inner (dot) product of vectors in $\mathbb{R}^2$; we say that vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are ***orthogonal*** if $\boldsymbol{x}^T\boldsymbol{y} = 0$.

Inner (dot) product of vectors in $\mathbb{R}^2$:

$$\boldsymbol{x} \cdot \boldsymbol{y} = |\boldsymbol{x}||\boldsymbol{y}| \cos\theta$$

where $|\cdot|$ denotes the length of a vector.



- ***Outer product***. Given $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$ their outer product is the matrix $\boldsymbol{x}\boldsymbol{y}^T \in \mathbb{R}^{m\times n}$ with entries given by $x_i y_j$. For example,

$$\begin{bmatrix} 5 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \end{bmatrix} = \begin{bmatrix} 5\cdot 1 & 5\cdot 2 \\ 4\cdot 1 & 4\cdot 2 \end{bmatrix} = \begin{bmatrix} 5 & 10 \\ 4 & 8 \end{bmatrix}$$

  This is also a special case of the matrix multiplication.

Matrix multiplication satisfies the following ***fundamental laws*** whenever the sums and products involved are defined:

- Matrix multiplication is not commutative in general, that is

$$AB \neq BA$$

  For example,

$$\begin{bmatrix} 5 & 3 \\ 4 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 14 & 22 \\ 10 & 16 \end{bmatrix} \qquad \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 5 & 3 \\ 4 & 2 \end{bmatrix} = \begin{bmatrix} 13 & 7 \\ 31 & 17 \end{bmatrix}$$

- Matrix multiplication is associative, that is,

$$(AB)C = A(BC)$$

which we will simply write as ABC. For example,

$$\left(\begin{bmatrix}5 & 3\\ 4 & 2\end{bmatrix}\cdot\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\right)\cdot\begin{bmatrix}-3 & 2\\ 2 & -1\end{bmatrix}=\begin{bmatrix}14 & 22\\ 10 & 16\end{bmatrix}\cdot\begin{bmatrix}-3 & 2\\ 2 & -1\end{bmatrix}=\begin{bmatrix}2 & 6\\ 2 & 4\end{bmatrix}$$

$$\begin{bmatrix}5 & 3\\ 4 & 2\end{bmatrix}\cdot\left(\begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}\cdot\begin{bmatrix}-3 & 2\\ 2 & -1\end{bmatrix}\right)=\begin{bmatrix}5 & 3\\ 4 & 2\end{bmatrix}\cdot\begin{bmatrix}1 & 0\\ -1 & 2\end{bmatrix}=\begin{bmatrix}2 & 6\\ 2 & 4\end{bmatrix}$$

- Matrix multiplication is distributive, that is,

$$A(B+C)=AB+AC \qquad (A+B)C=AC+BC$$

- Multiplication by a scalar satisfies

$$A(\alpha B)=\alpha(AB)$$

The basic ***unary operations*** on matrices and vectors are:

A *unary operation* is a rule that takes a single element and gives a single output.

- ***The transpose*** of a matrix $A\in\mathbb{R}^{n\times m}$, is the matrix $A^T\in\mathbb{R}^{m\times n}$ with its rows and columns transposed: $(A^T)_{ij}=A_{ji}$. For example,

$$\begin{bmatrix}5 & 3 & 2\\ 4 & 2 & 7\\ 1 & 0 & 8\end{bmatrix}^T=\begin{bmatrix}5 & 4 & 1\\ 3 & 2 & 0\\ 2 & 7 & 8\end{bmatrix}$$

  The transpose satisfies

$$(A^T)^T=A \qquad (AB)^T=B^TA^T \qquad (A+B)^T=A^T+B^T$$

- ***The trace*** of a square matrix $A\in\mathbb{R}^{n\times n}$, denoted by $\operatorname{tr}(A)$ or simply by $\operatorname{tr}A$, is the sum of its diagonal elements,

$$\operatorname{tr}A=\sum_{i=1}^{n}a_{ii}$$

  For example,

$$\operatorname{tr}\begin{bmatrix}5 & 3 & 2\\ 4 & 2 & 7\\ 1 & 0 & 8\end{bmatrix}=5+2+8=15$$

  The trace satisfies

$$\operatorname{tr}A^T=\operatorname{tr}A \qquad \operatorname{tr}(A+B)=\operatorname{tr}A+\operatorname{tr}B \qquad \operatorname{tr}(AB)=\operatorname{tr}(BA)$$

- ***A norm*** of a vector $\boldsymbol{x}\in\mathbb{R}^n$ is the real number $\|\boldsymbol{x}\|_2\in\mathbb{R}$ given by

$$\|\boldsymbol{x}\|_2=\sqrt{\boldsymbol{x}^T\boldsymbol{x}}=\sqrt{\sum_{i=1}^{n}x_i^2}$$

  This is called the Euclidean or ℓ_2 norm. Other examples of norms are the ℓ_1 and the ℓ_∞ norms,

$$\|\boldsymbol{x}\|_1=\sum_{i=1}^{n}|x_i| \qquad \|\boldsymbol{x}\|_\infty=\max_i|x_i|$$

  All three norms are often used in Data Science.

  For example,

$$\left\|\begin{bmatrix}2\\ 1\\ 4\end{bmatrix}\right\|_2=\sqrt{4+1+16}=\sqrt{21}$$

  Informally, the norm is the measure of the "length" of a vector. It can be viewed as a function $\|\cdot\|:\mathbb{R}^n\to\mathbb{R}$ satisfying

  - $\|\boldsymbol{x}\|\geq 0$ for all $\boldsymbol{x}\in\mathbb{R}^n$ (non-negativity)
  - $\|\boldsymbol{x}\|=0$ if and only if $\boldsymbol{x}=0$ (definiteness)
  - $\|a\,\boldsymbol{x}\|=|a|\cdot\|\boldsymbol{x}\|$ for all $a\in\mathbb{R}$ (homogeneity)
  - $\|\boldsymbol{x}+\boldsymbol{y}\|\leq\|\boldsymbol{x}\|+\|\boldsymbol{y}\|$ for all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^n$ (triangle inequality)

2.2.2 *The matrix form of linear systems*

An important application of matrices and vectors is that they provide a very elegant way of encoding an $m \times n$ system of linear equations as a ***matrix equation***. Indeed, suppose were are given an $m \times n$ system with an augmented matrix $(A|\boldsymbol{b})$. Denote by $\boldsymbol{x}$ the n-dimensional column vector with entries $x_1, x_2, \ldots, x_n$. Then

$$A\boldsymbol{x} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + \ldots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \ldots + a_{mn}x_n \end{bmatrix}$$

It follows that the underlying system is equivalent to the matrix equation

$$A\boldsymbol{x} = \boldsymbol{b} \quad \text{where} \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

and $\boldsymbol{x}$ is treated as a vector of n-unknowns. For example, the 3×4 linear system

$$\begin{aligned} 3x_1 + 2x_2 - 4x_4 &= 5 \\ -x_2 + 3x_3 + 7x_4 &= 0 \\ x_1 + 5x_2 + x_3 + 2x_4 &= 7 \end{aligned}$$

is equivalent to following the matrix equation

$$\begin{bmatrix} 3 & 2 & 0 & -4 \\ 0 & -1 & 3 & 7 \\ 1 & 5 & 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \\ 7 \end{bmatrix}$$

Example 2.4: Matrix form

We want to find all values of $x_1, x_2, x_3 \in \mathbb{R}$ which solve the vector equation

$$x_1 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

We start by rewriting this equation in a matrix form,

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

Hence, we are back to the usual problem of solving a system of linear equations, which we are well equipped to handle. We row reduce the associated augmented matrix:

$$\left[\begin{array}{ccc|c} 1 & 2 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{array}\right] \xrightarrow{r_3 \to r_3 - r_1} \left[\begin{array}{ccc|c} 1 & 2 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{array}\right] \xrightarrow[r_2 \to r_2 - r_3]{r_1 \to r_1 + r_3} \left[\begin{array}{ccc|c} 1 & 2 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{array}\right] \xrightarrow{r_1 \to r_1 - 2r_2} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{array}\right]$$

Therefore, there is a unique solution $(x_1, x_2, x_3) = (3, -1, 1)$.

2.2.3 *Linear dependence and indepedence*

*A vector $\boldsymbol{b} \in \mathbb{R}^m$ is a **linear combination** of $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n \in \mathbb{R}^m$ if it can be expressed in the form*

$$\boldsymbol{b} = x_1\boldsymbol{v}_1 + x_2\boldsymbol{v}_2 + \ldots + x_n\boldsymbol{v}_n \tag{2.5}$$

for some $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

Let's begin with some basic examples:

- A vector $\boldsymbol{v} \in \mathbb{R}^n$ is a linear combination of a single vector $\boldsymbol{w} \in \mathbb{R}^n$ if and only if it is a scalar multiple of it

  $$\boldsymbol{v} = k\boldsymbol{w}$$

  for some $k \in \mathbb{R}$.
- Every vector $\boldsymbol{x} \in \mathbb{R}^2$ can be written as

  $$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

  and is thus a linear combination of the vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ defined by

  $$\boldsymbol{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \boldsymbol{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

- More generally, every vector $\boldsymbol{x} \in \mathbb{R}^n$ can be written as

  $$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \ldots + x_n \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

  and thus is linear combination of vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ defined by

  $$\boldsymbol{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \boldsymbol{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \qquad \ldots \qquad \boldsymbol{e}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

  These are called the standard ***unit vectors*** in $\mathbb{R}^n$.

This is not an unfamiliar concept to us. Consider an $m \times n$ linear system, $A\boldsymbol{x} = \boldsymbol{b}$. Write the matrix A as

$$A = [\boldsymbol{v}_1 \; \boldsymbol{v}_2 \; \cdots \; \boldsymbol{v}_n] \quad \text{where} \quad \boldsymbol{v}_i = \begin{bmatrix} a_{1i} \\ \vdots \\ a_{mi} \end{bmatrix}$$

Then vector $\boldsymbol{b} \in \mathbb{R}^n$ is a linear combination of $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ precisely if the linear system has a solution $\boldsymbol{x} \in \mathbb{R}^n$. We are thus back to the problem of determining when an $m \times n$ system of linear equations is consistent.

Example 2.5: Linear combination

We want to write $\boldsymbol{b}$ as a linear combination of $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$, where

$$\boldsymbol{b} = \begin{bmatrix} 27 \\ 17 \\ 22 \end{bmatrix} \qquad \boldsymbol{a}_1 = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \qquad \boldsymbol{a}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \qquad \boldsymbol{a}_3 = \begin{bmatrix} 8 \\ 5 \\ 7 \end{bmatrix}$$

This is equivalent to solving the matrix equation

$$\begin{bmatrix} 2 & 1 & 8 \\ 1 & 1 & 5 \\ 2 & 0 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 27 \\ 17 \\ 22 \end{bmatrix}$$

which is nothing but a 3×3 system of linear equations. To solve it, we row reduce the associated augmented matrix, as usual:

$$\left[\begin{array}{ccc|c} 2 & 1 & 8 & 27 \\ 1 & 1 & 5 & 17 \\ 2 & 0 & 7 & 22 \end{array}\right] \xrightarrow{r_1 \to r_1 - r_2} \left[\begin{array}{ccc|c} 1 & 0 & 3 & 10 \\ 1 & 1 & 5 & 17 \\ 2 & 0 & 7 & 22 \end{array}\right] \begin{array}{c} r_2 \to r_2 - r_1 \\ r_3 \to r_3 - 2r_1 \end{array} \left[\begin{array}{ccc|c} 1 & 0 & 3 & 10 \\ 0 & 1 & 2 & 7 \\ 0 & 0 & 1 & 2 \end{array}\right] \begin{array}{c} r_1 \to r_1 - 3r_3 \\ r_2 \to r_2 - 2r_3 \end{array} \left[\begin{array}{ccc|c} 1 & 0 & 0 & 4 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{array}\right]$$

Therefore, we can write $\boldsymbol{b}$ as the linear combination $\boldsymbol{b} = 4\boldsymbol{a}_1 + 3\boldsymbol{a}_2 + 3\boldsymbol{a}_3$.

Let $\mathbf{0}$ denote the zero vector – a vector with all entries equal 0.

A set of m-dimensional vectors $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\} \subset \mathbb{R}^m$ is said to be ***linearly independent*** *if the vector equation*

$$x_1\boldsymbol{v}_1 + x_2\boldsymbol{v}_2 + \ldots + x_n\boldsymbol{v}_n = \mathbf{0}$$

has only the trivial solution, $x_1 = x_2 = \ldots = x_n = 0$.

If a set of vector is not linearly independent, then they are is said to be ***linearly dependent****.*

For example, the unit vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n \in \mathbb{R}^n$ form a linearly independent set. Indeed, we have

$$x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 + \ldots + x_n\boldsymbol{e}_n = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

which is equal to $\mathbf{0}$ if and only if each x_i is zero.

The problem of determining when $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\} \subset \mathbb{R}^m$ is linearly independent is not a new problem for us either. It is equivalent to determining when the matrix equation

$$A\boldsymbol{x} = \mathbf{0} \quad \text{where} \quad A = [\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n]$$

has a *unique solution*. This amounts to solving a homogeneous $m \times n$ system of linear equations. We will illustrate this with two examples.

Example 2.6: Linear independence

We want to verify if the following vectors are linearly independent:

$$v_1 = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \quad v_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \quad v_3 = \begin{bmatrix} 8 \\ 5 \\ 7 \end{bmatrix}$$

This amounts to solving the matrix equation

$$\begin{bmatrix} 2 & 1 & 8 \\ 1 & 1 & 5 \\ 2 & 0 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

To solve this, we apply Gauss-Jordan elimination to $(A|\mathbf{0})$, or just to A as the last column consisting of all zeroes will not change. This has already been done in Example 2.5: A has row reduced form

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

which represents a homogeneous system with a unique solution, the trivial solution, thus the set $\{v_1, v_2, v_3\}$ is *linearly independent*.

Example 2.7: Linear dependence

We want to verify if the following vectors are linearly dependent:

$$v_1 = \begin{bmatrix} 5 \\ 1 \\ 2 \\ 0 \end{bmatrix} \quad v_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ -2 \end{bmatrix} \quad v_3 = \begin{bmatrix} 3 \\ 3 \\ 1 \\ 5 \end{bmatrix} \quad v_4 = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 4 \end{bmatrix}$$

Again, this amounts to solving the matrix equation $A\boldsymbol{x} = \mathbf{0}$ with $A = [v_1\ v_2\ v_3\ v_4]$ – we need to find the row reduced form of A:

$$\begin{bmatrix} 5 & 1 & 3 & 2 \\ 1 & 1 & 3 & 1 \\ 2 & 0 & 1 & 1 \\ 0 & -2 & 5 & 4 \end{bmatrix} \xrightarrow[\mathrm{r}_4 \to -1/2\,\mathrm{r}_4]{\mathrm{r}_3 \leftrightarrow \mathrm{r}_1} \begin{bmatrix} 2 & 0 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 5 & 1 & 3 & 2 \\ 0 & 1 & -5/2 & -2 \end{bmatrix} \xrightarrow[\mathrm{r}_2 \to \mathrm{r}_2 - 1/2\,\mathrm{r}_1]{\mathrm{r}_1 \to 1/2\,\mathrm{r}_1} \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 5/2 & 1/2 \\ 0 & 0 & -2 & -1 \\ 0 & 1 & -5/2 & -2 \end{bmatrix}$$

$$\xrightarrow[\mathrm{r}_3 \to -1/2\,\mathrm{r}_3]{\mathrm{r}_4 \to \mathrm{r}_4 - \mathrm{r}_2} \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 5/2 & 1/2 \\ 0 & 0 & 1 & 1/2 \\ 0 & 0 & -5 & -5/2 \end{bmatrix} \xrightarrow[\mathrm{r}_4 \to \mathrm{r}_4 + 5\mathrm{r}_3]{\substack{\mathrm{r}_1 \to \mathrm{r}_1 - 1/2\,\mathrm{r}_3 \\ \mathrm{r}_2 \to \mathrm{r}_2 - 5/2\,\mathrm{r}_3}} \begin{bmatrix} 1 & 0 & 0 & 1/4 \\ 0 & 1 & 0 & -3/4 \\ 0 & 0 & 1 & 1/2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

This corresponds to the homogeneous system

$$\begin{aligned} x_1 + 1/4\, x_4 &= 0 \\ x_2 - 3/4\, x_4 &= 0 \\ x_3 + 1/2\, x_4 &= 0 \end{aligned}$$

Hence $x_4 = 4s$ with $s \in \mathbb{R}$ is a free variable, and the general solution is given by

$$(x_1, x_2, x_3, x_4) = (-s, -3s, -2s, 4s)$$

We thus conclude that the given vectors are *linearly dependent*.

Finally, let us state two more important facts.

- *Suppose $\boldsymbol{v}_1 \in \mathbb{R}^m$ can be written as a linear combination of vectors $\boldsymbol{v}_2, \ldots, \boldsymbol{v}_n \in \mathbb{R}^m$. Then the set $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ is linearly dependent.*
- *Suppose that $n > m$. Then any set of vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ in $\mathbb{R}^m$ is linearly dependent.*

The first statement is true since by the initial assumption,

$$\boldsymbol{v}_1 = x_2\boldsymbol{v}_2 + \ldots + x_n\boldsymbol{v}_n$$

for some $x_2, \ldots, x_n \in \mathbb{R}$. But then $(y_1, y_2, \ldots, y_n) = (-1, x_2, \ldots, x_n)$ is a nontrivial solution of

$$y_1\boldsymbol{v}_1 + y_2\boldsymbol{v}_2 + \ldots + y_n\boldsymbol{v}_n = \boldsymbol{0}$$

and so $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ is linearly dependent. For example, let

$$\boldsymbol{v}_1 = 2\underbrace{\begin{bmatrix}1\\0\end{bmatrix}}_{\boldsymbol{v}_2} + 3\underbrace{\begin{bmatrix}0\\1\end{bmatrix}}_{\boldsymbol{v}_3} = \begin{bmatrix}2\\3\end{bmatrix}$$

Then

$$-1\underbrace{\begin{bmatrix}2\\3\end{bmatrix}}_{\boldsymbol{v}_1} + 2\underbrace{\begin{bmatrix}1\\0\end{bmatrix}}_{\boldsymbol{v}_2} + 3\underbrace{\begin{bmatrix}0\\1\end{bmatrix}}_{\boldsymbol{v}_3} = \underbrace{\begin{bmatrix}0\\0\end{bmatrix}}_{\boldsymbol{0}}$$

The second statement is also true since the equation

$$x_1\boldsymbol{v}_1 + \ldots + x_n\boldsymbol{v}_n = \boldsymbol{0}$$

is equivalent to an $m \times n$ homogeneous system of equations. Since $m < n$, this system has infinitely many solutions and hence $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is linearly dependent. For example, let $m = 2$ and $n = 3$. Suppose

$$\boldsymbol{v}_1 = \begin{bmatrix}1\\0\end{bmatrix} \quad \boldsymbol{v}_2 = \begin{bmatrix}3\\2\end{bmatrix} \quad \boldsymbol{v}_3 = \begin{bmatrix}8\\6\end{bmatrix}$$

Then the relation

$$x_1\boldsymbol{v}_1 + x_2\boldsymbol{v}_2 + x_3\boldsymbol{v}_3 = \boldsymbol{0}$$

is equivalent to the 2×3 system

$$\begin{bmatrix}1 & 3 & 8\\0 & 2 & 6\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$$

Since

$$\begin{bmatrix}1 & 3 & 8\\0 & 2 & 6\end{bmatrix} \xrightarrow{\mathrm{r}_2 \to 1/2\,\mathrm{r}_2} \begin{bmatrix}1 & 3 & 8\\0 & 1 & 3\end{bmatrix} \xrightarrow{\mathrm{r}_1 \to \mathrm{r}_1 - 3\mathrm{r}_1} \begin{bmatrix}1 & 0 & -1\\0 & 1 & 3\end{bmatrix}$$

the initial system is equivalent to

$$\begin{aligned} x_1 - x_3 &= 0\\ x_2 + 3x_3 &= 0 \end{aligned}$$

Thus the general solution is $(x_1, x_2, x_3) = (s, -3s, s)$ for any $s \in \mathbb{R}$. In other words,

$$s(\boldsymbol{v}_1 - 3\boldsymbol{v}_2 + \boldsymbol{v}_3) = \boldsymbol{0}$$

2.3 Matrix rank, inverse and determinant

Lecture 8

In this lecture will mostly narrow our focus to square matrices in $\mathbb{R}^{n\times n}$ and sets of n vectors in $\mathbb{R}^n$. Our goal is to develop a method for finding the inverse of a square matrix, if it exists.

2.3.1 Singular and nonsingular matrices

> *A matrix $A \in \mathbb{R}^{n\times n}$ is called **singular** if the equation $A\boldsymbol{x} = \mathbf{0}$ has a nontrivial solution $\boldsymbol{x} \neq \mathbf{0} \in \mathbb{R}^n$. A matrix that is not singular is called **nonsingular**.*

Recall that an $n \times n$ homogeneous linear system $A\boldsymbol{x} = \mathbf{0}$ has either a unique solution, the trivial solution, or infinitely many solutions.

The definition says that a matrix A is *nonsingular* exactly when the matrix equation

$$A\boldsymbol{x} = \mathbf{0}$$

has a unique solution, the trivial solution $\boldsymbol{x} = \mathbf{0}$. We can rephrase this in terms of linear independence as follows. Let $\boldsymbol{a}_j$ denote the jth column of A. Then

$$A\boldsymbol{x} = \begin{bmatrix} \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = x_1\boldsymbol{a}_1 + \ldots + x_n\boldsymbol{a}_n$$

Thus, the matrix A is nonsingular precisely when its set of columns $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n\}$ is a linearly independent set in $\mathbb{R}^n$.

> *A matrix $A \in \mathbb{R}^{n\times n}$ is nonsingular if and only if its columns, viewed as vectors, form a linearly independent set in $\mathbb{R}^n$.*

A homogeneous equation $A\boldsymbol{x} = \mathbf{0}$ has a unique solution exactly when the augmented matrix $(A|\mathbf{0})$ has the reduced echelon form $(C|\mathbf{0})$, where C in an $n \times n$ reduced echelon matrix with $r = n$ nonzero rows. However, there is only one such matrix, that is

$$C = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I$$

We have thus shown the following.

> *If $A \in \mathbb{R}^{n\times n}$ is nonsingular, then it is row reduction equivalent to the identity matrix.*

We state the theorem below without an immediate proof. We will explain why this statement is indeed true in Subsection 2.3.4.

> *Let $A \in \mathbb{R}^{n\times n}$. Then the matrix equation $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution for every $\boldsymbol{b} \in \mathbb{R}^n$ if and only if A is nonsingular.*

2.3.2 *Matrix rank*

The ***column rank*** of a matrix $A \in \mathbb{R}^{m \times n}$ is the size of the largest subset of columns of A that constitute a linearly independent set. With some abuse of terminology, this is often referred to simply as the number of linearly independent columns of A. In the same way, the ***row rank*** is the largest number of rows of A that constitute a linearly independent set. For example, consider the matrix

$$M = \begin{bmatrix} 1 & 4 & 2 \\ 2 & 1 & -3 \\ 3 & 5 & -1 \end{bmatrix}$$

Its column rank is 2 because we the third column can be written as a linear combinations of the first two columns. Its row rank is also 2 since the third row can be written as a sum of the first two rows.

For any matrix $A \in \mathbb{R}^{m \times n}$, it turns out that the column rank of A is equal to the row rank of A (though we will not prove this), and so both quantities are referred to collectively as the ***rank*** of A, denoted as $\operatorname{rank}(A)$ or simply $\operatorname{rank} A$. For example, we found above that

$$\operatorname{rank} M = 2$$

In Python, the `numpy` package provides a matrix rank function:

```
import numpy as np
A = [[1,4,2],[2,1,-3],[3,5,-1]]
np.linalg.matrix_rank(A)
```

The following are some basic properties of the rank:

- for $A \in \mathbb{R}^{m \times n}$, $\operatorname{rank} A \leq \min(m, n)$. If $\operatorname{rank} A = \min(m, n)$, then A is said to be ***full rank***;
- for $A \in \mathbb{R}^{m \times n}$, $\operatorname{rank} A = (\operatorname{rank} A^T)$;
- for $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $\operatorname{rank}(AB) \leq \min(\operatorname{rank} A, \operatorname{rank} B)$;
- for $A, B \in \mathbb{R}^{m \times n}$, $\operatorname{rank}(A + B) \leq \operatorname{rank} A + \operatorname{rank} B$.

And most crucially:

> *A square matrix $A \in \mathbb{R}^{n \times n}$ is nonsingular if and only if it is **full rank**, that is* $\operatorname{rank} A = n$.

2.3.3 *Matrix inverse*

If a is a nonzero real number, then there exists a unique $b \in \mathbb{R}$ satisfying

$$ab = 1 = ba$$

The number b is called the ***multiplicative inverse*** of a, and is nothing but the reciprocal $b = 1/a$.

For a square matrix $A \in \mathbb{R}^{n \times n}$, the analogue of the above equation is

$$AB = I = BA$$

where I is the $n \times n$ identity matrix and B is called the ***inverse matrix*** to A, and is denoted by A^{-1}.

Not all matrices have inverses. Non-square matrices, for example, do not have inverses by definition. However, even for some square matrices A, it may still be the case that A^{-1} may not exist.

> *We say that a square matrix $A \in \mathbb{R}^{n \times n}$ is **invertible** if A^{-1} exists, and it is **non-invertible** if A^{-1} does not exist.*

Here are some simple examples of matrix inverses:

- The $n \times n$ identity matrix $I = \operatorname{diag}(1, \ldots, 1)$ has inverse $I^{-1} = I$.
- Any $n \times n$ diagonal matrix

$$D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$$

is invertible exactly when all the d_i's are nonzero. In this case the inverse matrix is obtained by taking the reciprocal of each d_i:

$$D^{-1} = \operatorname{diag}(d_1^{-1}, d_2^{-1}, \ldots, d_n^{-1})$$

The general situation is much more complicated; the inverse of an arbitrary invertible matrix A is not obtained by just inverting each entry of A.

Our current goal is to determine when A^{-1} exists, while simultaneously uncovering an algorithm for computing it. With this goal in mind, consider equation $AB = I$. It can be written equivalently as

$$AB = [A\boldsymbol{b}_1 \; A\boldsymbol{b}_2 \; \cdots A\boldsymbol{b}_n] = [\boldsymbol{e}_1 \; \boldsymbol{e}_2 \; \cdots \; \boldsymbol{e}_n]$$

where $\boldsymbol{b}_i$'s denote columns of the matrix B and $\boldsymbol{e}_i$'s are the unit vectors in $\mathbb{R}^n$. This means that solving $AB = I$ for $B \in \mathbb{R}^{n \times n}$ is equivalent to solving the n linear equations

$$A\boldsymbol{b}_1 = \boldsymbol{e}_1 \quad A\boldsymbol{b}_2 = \boldsymbol{e}_2 \quad \ldots \quad A\boldsymbol{b}_n = \boldsymbol{e}_n$$

for $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n \in \mathbb{R}^n$. In particular, the matrix A is only invertible if all these equations can be solved. However, this doesn't tell us that a solution B of $AB = I$ is unique. Moreover, this doesn't say anything about the second wanted identity, $BA = I$. The three statements below clarify these points.

- *Suppose matrices $A, B \in \mathbb{R}^{n \times n}$ satisfy $AB = I$. Then both matrices, A and B, are nonsingular.*
- *If a matrix $A \in \mathbb{R}^{n \times n}$ is nonsingular, then there is a unique matrix $B \in \mathbb{R}^{n \times n}$ such that $AB = I$.*
- *Suppose matrices $A, B \in \mathbb{R}^{n \times n}$ satisfy $AB = I$. Then A is invertible with the inverse $A^{-1} = B$, that is $AB = I = BA$.*

We illustrate these statements with a simple example. Consider the 2×2 matrix

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix}$$

It is a nonsingular matrix and has a unique inverse, which is also nonsingular,

$$A^{-1} = \begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix}$$

Indeed,

$$AA^{-1} = \begin{bmatrix} 1 \cdot (-5) + 2 \cdot 3 & 1 \cdot 2 + 2 \cdot (-1) \\ 3 \cdot (-5) + 5 \cdot 3 & 3 \cdot 2 + 5 \cdot (-1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

$$A^{-1}A = \begin{bmatrix} (-5) \cdot 1 + 2 \cdot 3 & (-5) \cdot 2 + 2 \cdot 5 \\ 3 \cdot 1 + (-1) \cdot 3 & 3 \cdot 2 + (-1) \cdot 5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

We now know that a matrix $A \in \mathbb{R}^{n\times n}$ is invertible precisely if it is nonsingular. Suppose that a matrix $B \in \mathbb{R}^{n\times n}$ is also nonsingular. The matrices A and B then satisfy the following useful identities:

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

The last ingredient that we need is the precise method of determining the inverse. Let $A \in \mathbb{R}^{n\times n}$. The following algorithm gives us a concrete method for both determining when A^{-1} exists and computing it:

(1) Solve the equation $A\boldsymbol{x} = \boldsymbol{e}_j$ for each $1 \leq j \leq n$. This is done as usual, by row reducing $(A|\boldsymbol{e}_j)$ to the reduced row echelon form.

(2) If any of these equations are inconsistent, A is not invertible. Otherwise, A is invertible and has the reduced row echelon form I (because it is nonsingular), so the above step will take us to a matrix of the form $(I|\boldsymbol{b}_j)$.

(3) The inverse A^{-1} is then equal to the matrix $A^{-1} = [\boldsymbol{b}_1 \ \cdots \ \boldsymbol{b}_n]$.

Note that we can solve all the equations $A\boldsymbol{x} = \boldsymbol{e}_j$ at the same time by forming the $n \times 2n$ matrix

$$(A|\boldsymbol{e}_1 \ \cdots \ \boldsymbol{e}_n) = (A|I)$$

and applying Gauss-Jordan elimination until we obtain a matrix of the form

$$(C|B)$$

with C the unique reduced row echelon form matrix, row equivalent to A. If $C = I$, A is invertible with $A^{-1} = B$. Otherwise, A is not invertible.

In Python, the `numpy` package provides a matrix inverse function:

```
import numpy as np
A = np.array([[1,2],[2,3]])
np.linalg.inv(A)
```

Example 2.8: Matrix inverse

We want to find A^{-1} if it exists, where

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

We need to row reduce $(A|I)$:

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \xrightarrow{r_2 \to 1/2\, r_2} \left[\begin{array}{ccc|ccc} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \xrightarrow[r_2 \to r_2 - r_3]{r_1 \to r_1 - r_3} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & 0 & 1/2 & -1 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right]$$

This shows that A is invertible with

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

Example 2.9: Matrix inverse

We want to find A^{-1} if it exists, where

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 2 & 2 \\ 2 & 4 & 6 \end{bmatrix}$$

Again, we need to row reduce $(A|I)$:

$$\left[\begin{array}{ccc|ccc} 1 & 1 & 2 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 \\ 2 & 4 & 6 & 0 & 0 & 1 \end{array}\right] \xrightarrow{r_3 \to r_3 - 2r_1} \left[\begin{array}{ccc|ccc} 1 & 1 & 2 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 \\ 0 & 2 & 2 & -2 & 0 & 1 \end{array}\right] \xrightarrow{r_3 \to r_3 - 2r_2} \left[\begin{array}{ccc|ccc} 1 & 1 & 2 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & -2 & -1 & 1 \end{array}\right]$$

At this point there is no need to continue. The last row represents the three inconsistent equations

$$0 = -2 \qquad 0 = -1 \qquad 0 = 1$$

Therefore, A is not invertible.

2.3.4 *Solving linear systems using an inverse*

Suppose we are given an $n \times n$ system of linear equations with augmented matrix $(A|\boldsymbol{b})$. The system is then equivalent to the matrix equation

$$A\boldsymbol{x} = \boldsymbol{b}$$

Suppose we know that A is nonsingular. Then A is invertible. Left multiplying both sides of the equality with A^{-1} we find that

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

If we already know A^{-1}, this immediately gives as the solution to the underlying system.

Example 2.10: Solving systems using inverses

We want to solve the 2×2 system

$$2x_1 + 3x_3 = 4$$
$$6x_1 + 7x_2 = 0$$

It is equivalent to the matrix equation

$$A\boldsymbol{x} = \begin{bmatrix} 2 & 3 \\ 6 & 7 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

We have already computed the inverse A^{-1} in an example above. Hence the above 2×2 system has a unique solution

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} -7 & 3 \\ 6 & -2 \end{bmatrix} \begin{bmatrix} 4 \\ 0 \end{bmatrix} = \begin{bmatrix} -7 & 3 \\ 6 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -7 \\ 6 \end{bmatrix}$$

2.3.5 *Determinant*

There are other ways of finding the inverse of a matrix. In particular, there is a very simple formula for determining if a 2×2 matrix is nonsingular and computing its inverse.

Let $A \in \mathbb{R}^{2\times 2}$. Then A is nonsingular precisely if the quantity

$$\det A = a_{11}a_{22} - a_{12}a_{21}$$

*called the **determinant** of A, is non-zero. In this case the inverse is given by*

$$A^{-1} = \frac{1}{\det A}\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$$

For example, suppose that

$$A = \begin{bmatrix} 2 & 3 \\ 6 & 7 \end{bmatrix}$$

Then

$$A^{-1} = \frac{1}{2\cdot 7 - 3\cdot 6}\begin{bmatrix} 7 & -3 \\ -6 & 2 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} -7 & 3 \\ 6 & -2 \end{bmatrix}$$

The key object in this construction is the *determinant*. It is a scalar value that can be computed from the elements of a square matrix and encodes certain properties of the the matrix.

Geometrically, the determinant can be viewed as the volume scaling factor of the linear transformation described by the matrix; we will touch this topic in the next lecture.

*Given a square matrix $A \in \mathbb{R}^{3\times 3}$ its determinant can be computed using the **expansion by the first row rule***

$$\det(A) = a_{11}\det(A_{11}) - a_{12}\det(A_{12}) + a_{13}\det(A_{13})$$

where A_{ij} represents the matrix obtained by eliminating from A the ith row and the jth column.

For example, let

$$A = \begin{bmatrix} 3 & 0 & 1 \\ 1 & 2 & 5 \\ -1 & 4 & 2 \end{bmatrix}$$

Then

$$\det(A) = 3\cdot\begin{vmatrix} 2 & 5 \\ 4 & 2 \end{vmatrix} - 0\cdot\begin{vmatrix} 1 & 5 \\ -1 & 2 \end{vmatrix} + 1\cdot\begin{vmatrix} 1 & 2 \\ -1 & 4 \end{vmatrix} = 3\cdot(-16) + 1\cdot 6 = -42$$

In fact, we can expand $\det(A)$ by any row or column following the pattern of the signs

$$\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix}$$

For example, using the expansion by the second column we find

$$\det(A) = -0\cdot\begin{vmatrix} 1 & 5 \\ -1 & 2 \end{vmatrix} + 2\cdot\begin{vmatrix} 3 & 1 \\ -1 & 2 \end{vmatrix} - 4\cdot\begin{vmatrix} 3 & 1 \\ 1 & 5 \end{vmatrix} = 2\cdot 7 - 4\cdot 14 = -42$$

The expansion rules above are particular cases of the generic ***Laplace rule*** for the expansion of the determinant. This rule allows us to reduce the computation of the determinant of an $n \times n$ matrix to that of n determinants of certain $(n-1) \times (n-1)$ matrices.

In Python, the `numpy` package provides a matrix determinant function:

```
import numpy as np
A = [[1,2],[2,3]]
np.linalg.det(A)
```

Given a square matrix $A \in \mathbb{R}^{n \times n}$ with $n \geq 2$ its determinant can be computed using the ***expansion by the ith row rule***

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij})$$

or using the ***expansion by the jth column rule***

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij})$$

where A_{ij} represents the matrix obtained by eliminating from A the ith row and the jth column.

- The scalar $\det(A_{ij})$ is called the ***complementary minor*** associated with the entry a_{ij}.
- The scalar $c_{ij} = (-1)^{i+j} \det(A_{ij})$ is called the ***cofactor*** of the entry a_{ij}.
- The matrix of cofactors is called the ***cofactor matrix*** of A, and is denoted by $\mathrm{Cof}(A) = (c_{ij})$.
- The transpose of cofactor matrix is called the ***adjoint matrix*** of A, and is denoted by $\mathrm{Adj}(A) = \mathrm{Cof}(A^T)$.

The determinant combined with the adjoint matrix can be used to find the inverse of a matrix.

Let $A \in \mathbb{R}^{n \times n}$ with $n \geq 2$. Then

$$\mathrm{Adj}(A)A = \det(A)I$$

In particular, if $\det(A) \neq 0$, then

$$A^{-1} = \frac{1}{\det(A)} \mathrm{Adj}(A) \tag{2.6}$$

Computing determinants and matrix inverses using the Laplace rule can be a tedious exercise and thus is seldom done by hand in its "original form". Formula (2.6) is only efficient for small matrices and sparse matrices only, since for general matrices this requires to compute an exponential number of determinants, even if care is taken to compute each minor only once.

The determinant satisfies the following basic properties:

- $\det(A^T) = \det(A)$;
- $\det(A^{-1}) = 1/\det(A)$;
- $\det(cA) = c^n \det(A)$ for any scalar c;
- $\det(AB) = \det(A)\det(B)$.

The task of computing the determinant can be simplified significantly with the help of the following additional properties:

- if two columns or rows are interchanged, the determinant changes by a sign;
- if the matrix B is obtained by multiplying row i of A by a scalar c, then $\det(B) = c\det(A)$;
- if the matrix B is obtained by adding c times row i of A to row j of A, then $\det(B) = \det(A)$;
- if the matrix B is obtained by adding c times column i of A to column j of A, then $\det(B) = \det(A)$.

We say that a matrix $A \in \mathbb{R}^{n\times n}$ is ***upper triangular*** if $a_{ij} = 0$ for all $i > j$, that is,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2n} \\ 0 & 0 & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

By transposing an upper triangular matrix, i.e., A^T, we obtain a ***lower triangular*** matrix. Triangular matrices have the following important property.

> *Let $A \in \mathbb{R}^{n\times n}$ with $n \geq 2$ be an upper or lower triangular matrix. Then*
>
> $$\det(A) = a_{11}a_{22}\cdots a_{nn}$$
>
> *In particular,* $\det(I) = 1$.

We can thus use row reduction to simplify computations of determinants.

Example 2.11: Determinant

Consider the 4×4 matrix

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

Using row reduction, we find

$$\det(A) \overset{\substack{r_2\to r_2+2r_1 \\ r_3\to r_3+r_1 \\ r_4\to r_4-r_1}}{=\!=\!=} \begin{vmatrix} -1 & 0 & 1 & 1 \\ 0 & -1 & 2 & 4 \\ 0 & 2 & 2 & 0 \\ 0 & -1 & 0 & -1 \end{vmatrix} \overset{\substack{r_3\to r_3+2r_2 \\ r_4\to r_4-r_2}}{=\!=\!=} \begin{vmatrix} -1 & 0 & 1 & 1 \\ 0 & -1 & 2 & 4 \\ 0 & 0 & 6 & 8 \\ 0 & 0 & -2 & -5 \end{vmatrix}$$

$$\overset{r_3\to r_3+3r_4}{=\!=\!=} \begin{vmatrix} -1 & 0 & 1 & 1 \\ 0 & -1 & 2 & 4 \\ 0 & 0 & 0 & -7 \\ 0 & 0 & -2 & -5 \end{vmatrix} \overset{r_3\leftrightarrow r_4}{=\!=} - \begin{vmatrix} -1 & 0 & 1 & 1 \\ 0 & -1 & 2 & 4 \\ 0 & 0 & -2 & -5 \\ 0 & 0 & 0 & -7 \end{vmatrix} = -(-1)(-1)(-2)(-7) = -14$$

2.4 Vector spaces and linear mappings

Lecture 9

In this lecture we shift our focus to more abstract mathematical concepts that will encapsulate our knowledge of vectors of matrices.

2.4.1 Vector spaces and subspaces

*A real **vector space** V is an algebraic structure consisting of a non-empty set of elements, called **vectors**, in which two operations are defined:*

- *addition "$+$": for all $\boldsymbol{v}_1, \boldsymbol{v}_2 \in V$ we have $\boldsymbol{v}_1 + \boldsymbol{v}_2 \in V$*
- *multiplication by scalars "$\cdot$": for all $c \in \mathbb{R}$ and $\boldsymbol{v} \in V$ we have $c \cdot \boldsymbol{v} \in V$, often written simply as $c\boldsymbol{v}$.*

The two operations, "$+$" and "$\cdot$", are required to satisfy eight properties, called axioms of a vector space, that encode our "usual experience" of adding and multiplying real numbers. Using only these axioms, one can show that the *zero vector* $\mathbf{0}$ is unique, and the *opposite vector* $(-\boldsymbol{v})$ is also unique for each $\boldsymbol{v} \in V$. We shall say that V is ***closed*** under vector addition and multiplication by scalars.

In these lectures we only consider real vector spaces, i.e., vector spaces defined over real numbers. There exist more general vector spaces, however we will not encounter them (in fact, we will discuss complex vector spaces very briefly). Basic examples of real vector spaces are:

- $V = \{\mathbf{0}\}$ – the set containing the null vector only; we have $\mathbf{0} + \mathbf{0} = \mathbf{0}$ and $c \cdot \mathbf{0} = \mathbf{0}$ for all $c \in \mathbb{R}$.
- $V = \mathbb{R}$ – the set of real numbers equipped with the usual addition and multiplication rules.
- $V = \mathbb{R}^n$ – the set of n-tuples of real numbers with the usual addition and multiplication rules is the fundamental (prototypical) example of a real vector space.
- Let $D \subseteq \mathbb{R}$. Then the set of functions $f : D \to \mathbb{R}$ with the usual addition and multiplication rules is a real vector space. The null vector is the function $f(x) = 0$ for all $x \in D$, the opposite of f is the function $(-f)$ defined by $(-f)(x) = -f(x)$.
- The set of polynomials $P_n(x) : \mathbb{R} \to \mathbb{R}$ with the usual addition and multiplication rules is also a real vector space. The null vector is the constant polynomial $P_0(x) = a_0$ with $a_0 = 0$.

*Let V be a real vector space. A non-empty subset $W \subseteq V$ is called a vector **subspace** if and only if it is a real vector space itself.*

Given a subset $W \subseteq V$, in order to ensure that W is a subspace we only need to verify that

- $\boldsymbol{w}_1 + \boldsymbol{w}_2 \in W$ for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in W$,
- $c \cdot \boldsymbol{w} \in W$ for all $c \in \mathbb{R}$ and $\boldsymbol{w} \in W$, and
- $\mathbf{0} \in W$.

Vector addition and multiplication by scalar operations are required to satisfy the following eight properties

(V1) $\boldsymbol{v}_1 + \boldsymbol{v}_2 = \boldsymbol{v}_2 + \boldsymbol{v}_1$ for all $\boldsymbol{v}_1, \boldsymbol{v}_2 \in V$

(V2) $(\boldsymbol{v}_1 + \boldsymbol{v}_2) + \boldsymbol{v}_3 = \boldsymbol{v}_1 + (\boldsymbol{v}_2 + \boldsymbol{v}_3)$ for all $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3 \in V$

(V3) there exists an element $\mathbf{0} \in V$, called ***zero*** of the ***null vector***, such that $\boldsymbol{v} + \mathbf{0} = \boldsymbol{v}$ for all $\boldsymbol{v} \in V$

(V4) for all $\boldsymbol{v} \in V$ there exists an ***opposite*** $(-\boldsymbol{v}) \in V$ such that $\boldsymbol{v} + (-\boldsymbol{v}) = \mathbf{0}$

(V5) $1 \cdot \boldsymbol{v} = \boldsymbol{v}$ for all $\boldsymbol{v} \in V$

(V6) $a(\boldsymbol{v}_1 + \boldsymbol{v}_2) = a\boldsymbol{v}_1 + a\boldsymbol{v}_2$ for all $a \in \mathbb{R}$ and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in V$

(V7) $(a + b)\boldsymbol{v} = a\boldsymbol{v} + b\boldsymbol{v}$ for all $a, b \in \mathbb{R}$ and $\boldsymbol{v} \in V$

(V8) $(ab)\boldsymbol{v} = a(b\boldsymbol{v})$ for all $a, b \in \mathbb{R}$ and $\boldsymbol{v} \in V$

The remaining properties (from V1–V8) are automatic.

Here are some basic examples of vector spaces together with their subspaces:

- The set consisting of the zero vector only, $\{\mathbf{0}\}$, is a subspace of any vector space V; it is called the ***zero*** or ***trivial subspace***.
- The set of continuous differentiable functions $\mathcal{C}^1(D)$ on a domain $D \subseteq \mathbb{R}$ is a subspace of the vector space of continuous functions $\mathcal{C}(D)$.
- The vector space $\mathbb{R}^2$ is *not* a subspace of $\mathbb{R}^3$ because $\mathbb{R}^2$ is not even a subset of $\mathbb{R}^3$. Vectors in $\mathbb{R}^3$ have three components, whereas the vectors in $\mathbb{R}^2$ have only two. We can define a subset $H \subset \mathbb{R}^3$ that "looks" and "acts" like $\mathbb{R}^2$

$$H = \left\{ \begin{bmatrix} a \\ b \\ 0 \end{bmatrix} \in \mathbb{R}^3 : a, b \in \mathbb{R} \right\}$$

  The subset H is a subspace of $\mathbb{R}^3$. Geometrically, it is a horizontal plane.
- Let V be a vector space and let $U \subseteq V$ and $W \subseteq V$ be subspaces. We denote by $U \cap W$ the intersection of U and W, i.e., the set of elements which lie both in U and W,

$$U \cap W = \{\boldsymbol{v} \in V : \boldsymbol{v} \in U \text{ and } \boldsymbol{v} \in W\}$$

  The intersection $U \cap W$ is a subspace of V. For instance, if $V = \mathbb{R}^3$ and U, W are two planes in $\mathbb{R}^3$ passing through the origin, then their intersection is a straight line passing through the origin.

Example 2.12: Vector subspaces

We want to determine if $V \subset \mathbb{R}^3$ is a subspace, where

$$V = \left\{ \begin{bmatrix} v_1 \\ v_2 \\ 3v_1 + v_2 \end{bmatrix} : v_1, v_2 \in \mathbb{R} \right\}$$

We need to verify that $\mathbf{0} \in V$ and that V is closed under addition and multiplication by a scalar:

- the zero vector is obtained by choosing $v_1 = v_2 = 0$, thus $\mathbf{0} \in V$;
- suppose $\boldsymbol{v}, \boldsymbol{w} \in V$, then

$$\boldsymbol{v} + \boldsymbol{w} = \begin{bmatrix} v_1 \\ v_2 \\ 3v_1 + v_2 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \\ 3w_1 + w_2 \end{bmatrix} = \begin{bmatrix} v_1 + w_1 \\ v_2 + w_2 \\ 3(v_1 + w_1) + (v_2 + w_2) \end{bmatrix} \in V$$

- suppose $k \in \mathbb{R}$ and $\boldsymbol{v} \in V$, then

$$k\boldsymbol{v} = k \begin{bmatrix} v_1 \\ v_2 \\ 3v_1 + v_2 \end{bmatrix} = \begin{bmatrix} kv_1 \\ kv_2 \\ 3kv_1 + kv_2 \end{bmatrix} \in V$$

Example 2.13: Vector subspaces

We want to determine if $V \subset \mathbb{R}^2$ is a subspace, where

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

Clearly, $\mathbf{0} \in V$, however V is not closed under addition. Indeed, if $\boldsymbol{v}, \boldsymbol{w} \in V$, then

$$\boldsymbol{v} + \boldsymbol{w} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} v_1 + w_1 \\ v_2 + w_2 \end{bmatrix}$$

where

$$(v_1 + w_1)(v_2 + w_2) = v_1 v_2 + v_1 w_1 + v_2 w_2 + w_1 w_2 = v_1 w_1 + v_2 w_2 \neq 0$$

Hence $\boldsymbol{v} + \boldsymbol{w} \notin V$, and so V is not a subspace of $\mathbb{R}^2$.

2.4.2 *Generators*

*Let V be a vector space and $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ a set of n vectors in V. We call the subspace **generated** (or **spanned**) by $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ the set of all linear combinations of $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$:*

$$\operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\} = \{c_1 \boldsymbol{v}_1 + \ldots + c_n \boldsymbol{v}_n : c_1, \ldots, c_n \in \mathbb{R}\}$$

For instance, in Example 2.12, we showed that $V \subset \mathbb{R}^3$ defined by

$$V = \left\{ \begin{bmatrix} v_1 \\ v_2 \\ 3v_1 + v_2 \end{bmatrix} : v_1, v_2 \in \mathbb{R} \right\}$$

is a vector subspace of $\mathbb{R}^3$. Every vector in V can be written as

$$\begin{bmatrix} v_1 \\ v_2 \\ 3v_1 + v_2 \end{bmatrix} = v_1 \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix} + v_2 \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

Thus a spanning set for V is

$$V = \operatorname{span} \left\{ \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix}, \ \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right\}$$

Let V be a vector space and let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\} \subseteq V$ be any subset. It can be verified that $\operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a vector subspace of V.

We say that the set $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ generates V if for every $\boldsymbol{v} \in V$, there exist coefficients $c_1, \ldots, c_n \in \mathbb{R}$ such that

$$\boldsymbol{v} = c_1 \boldsymbol{v}_1 + \ldots + c_n \boldsymbol{v}_n.$$

In other words,

$$V = \operatorname{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$$

*and $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is called a **system of generators** for V.*

Example 2.14: Generators

Let $V \subset \mathbb{R}^3$ denote the subspace generated by the set $\{v_1, v_2\}$, where

$$v_1 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} \qquad v_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{2.7}$$

We want to find sufficient and necessary conditions for a vector $\boldsymbol{b} \in \mathbb{R}^3$ to belong to $V = \text{span}\{v_1, v_2\}$. By definition, $\boldsymbol{b} \in V$ if and only if there exist constants $x_1, x_2 \in \mathbb{R}$ such that

$$x_1 v_1 + x_2 v_2 = \boldsymbol{b}$$

Hence we are back to the problem of determining when a system is consistent, namely the 3×2 system with the matrix form

$$A\boldsymbol{x} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \boldsymbol{b}$$

We row reduce the corresponding augmented matrix:

$$\left[\begin{array}{cc|c} 1 & 1 & b_1 \\ 2 & 1 & b_2 \\ 0 & 1 & b_3 \end{array}\right] \xrightarrow{r_2 \to r_2 - 2r_1} \left[\begin{array}{cc|c} 1 & 1 & b_1 \\ 0 & -1 & b_2 - 2b_1 \\ 0 & 1 & b_3 \end{array}\right] \xrightarrow{r_3 \to r_3 + r_2} \left[\begin{array}{cc|c} 1 & 1 & b_1 \\ 0 & -1 & b_2 - 2b_1 \\ 0 & 0 & b_3 + b_2 - 2b_1 \end{array}\right]$$

Therefore, the system is consistent precisely when $b_3 + b_2 - 2b_1 = 0$. This allows us to conclude that $\boldsymbol{b} \in \mathbb{R}^3$ belongs to V is an only if $b_3 = 2b_1 - b_2$. Equivalently,

$$V = \text{span}\{v_1, v_2\} = \left\{ \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} : b_3 = 2b_1 - b_2, \;\; b_1, b_2 \in \mathbb{R} \right\} \tag{2.8}$$

2.4.3 *Bases and dimensions of vector spaces*

> *We call a **basis** of V any set of linearly independent vectors $\{v_1, \ldots v_n\}$ that generates V.*

Since $\{v_1, \ldots v_n\}$ is a system of generators of V, it follows that any vector $v \in V$ can be written as

$$v = c_1 v_1 + \ldots + c_n v_n \tag{2.9}$$

for some $c_1, \ldots, c_n \in \mathbb{R}$. The scalars $c_1, \ldots, c_n$ are called the ***coefficients*** (or ***components*** or ***coordinates***) of v with respect to the basis $\{v_1, \ldots v_n\}$. The decomposition (2.9) is called the decomposition of v with respect to the basis $\{v_1, \ldots v_n\}$.

An important difference between a spanning set (or a system of generators) and a basis is that the former do not need to be linearly independent. For example, consider the subspaces $V = \text{span}\{v_1, v_2\} \subset \mathbb{R}^3$ and $W = \text{span}\{v_1, v_2, v_3\} \subset \mathbb{R}^3$, where v_1 and v_1 are given by (2.7), and $v_3 = v_1 + v_2$. Repeating the same steps as in Example 2.14 we would arrive to the conclusion that W is also given by the formula (2.8) implying that $V = W$, and so they must have the same basis.

The next theorem gives us useful information about the number of elements of any basis of a vector space. The main result is that any two bases of a vector space have the same number of elements. Moreover, it gives us a useful criterion to determine when a set of elements of a vector space is a basis.

Let V be a real vector space and let $\{\boldsymbol{v}_1, \dots \boldsymbol{v}_n\}$ be a basis of V. Then, the following statements hold true:

- *any set of linearly independent vectors in V has at most n elements;*
- *any other basis has exactly n elements;*
- *any set of n linearly independent vectors in V must generate V, and hence, form a basis.*

Here are some simple examples:

- Continuing from the previous example, any two linearly independent vectors $\boldsymbol{v}, \boldsymbol{w} \in V$ form a basis of the space $V = \text{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2\}$; for instance

$$\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2 = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \qquad \boldsymbol{w} = \boldsymbol{v}_1 - \boldsymbol{v}_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

- The most commonly used basis for $\mathbb{R}^n$ is the canonical basis, or the natural basis, which is given by the set $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ where

$$\boldsymbol{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \boldsymbol{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \qquad \dots \qquad \boldsymbol{e}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

  The basis $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ is an ***orthonormal*** basis, since vectors $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ with $i \neq j$ are orthogonal, that is $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ if $i \neq j$, and are normalised to 1, that is $\|\boldsymbol{e}_i\|_2 = 1$.

- The set $\{\boldsymbol{j}_1, \boldsymbol{j}_2, \dots, \boldsymbol{j}_n\}$, where

$$\boldsymbol{j}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \boldsymbol{j}_2 = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ 0 \end{bmatrix} \qquad \dots \qquad \boldsymbol{j}_n = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ n \end{bmatrix}$$

  is also a basis of $\mathbb{R}^n$: every vector $v \in \mathbb{R}^n$ has a unique decomposition with respect to the basis $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$.

The theorem and the examples above suggest that there exists a number that is uniquely associated to any vector space V – the number of elements of any basis of V. This number is an intrinsic property of V that does not depend on a particular choice of the basis.

*Let V be a vector space having a basis consisting of n elements. We call n the **dimension** of V and we write*

$$\dim(V) = n$$

If V consists of $\mathbf{0}$ alone, then V does not have a basis, and we shall say that V has dimension 0. A vector space which has a basis consisting of a finite number of elements, or the zero vector space, is called ***finite dimensional***. Otherwise, if for any n, there exist n linearly independent vectors of V, then V is called ***infinite dimensional*** and we write $\dim(V) = \infty$. However, in these lectures we will restrict to finite dimensional vector spaces only.

Let V be a vector space having a basis of n elements. Let $W \subseteq V$ be a vector subspace. Then

$$\dim(W) \leq \dim(V)$$

Below we have listed dimensions of some commonly used vector spaces:

- Viewing $\mathbb{R}$ as a vector space its dimension is $\dim(\mathbb{R}) = 1$. Indeed, the element 1 forms a basis of $\mathbb{R}$ over $\mathbb{R}$, since any element $c \in \mathbb{R}$ has a unique expression $c = c \cdot 1$.
- The dimension of $\mathbb{R}^n$ is $\dim(\mathbb{R}^n) = n$.
- The dimension of the space of continuous function on $D \subseteq \mathbb{R}$ is $\dim(\mathcal{C}(D)) = \infty$.

2.4.4 *Linear mappings*

We have learned that a subspace of $\mathbb{R}^n$ is a subset $V \subseteq \mathbb{R}^n$ with some additional structure which allows one to do vector operations inside of V. Linear mappings are special kinds of functions from one subspace to another which "keep track" of information about these underlying vector operations. Before formalizing this into a definition, we briefly recall some basic notions related to functions.

Let V and W be sets. A function f from V to W, written as $f : V \to W$, is a rule that assigns to each $x \in V$ a point $f(x) \in W$. The set V is called the domain of f, and W is called the codomain of f. Here are some simple examples:

- $f_1 : \mathbb{R} \to \mathbb{R}$, given by $f_1(x) = \exp(x)$;
- $f_2 : \mathbb{R}^2 \to \mathbb{R}$, given by $f_2\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = x_1 + x_2$;
- $f_3 : \mathbb{R}^2 \to \mathbb{R}$, given by $f_3\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = x_1 x_2$;
- $f_4 : \mathbb{R}^2 \to \mathbb{R}^3$, given by $f_4\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1 \\ x_1 + 2x_2 \\ x_2 \end{bmatrix}$;
- $f_5 : \mathbb{R}^n \to \mathbb{R}$, given by $f_5(\boldsymbol{x}) = \det(\boldsymbol{x})$;
- $f_6 : \mathbb{R}^n \to \mathbb{R}^n$, given by $f_6(\boldsymbol{x}) = \boldsymbol{x}^T$.

As indicated above, linear transformations are special kinds of functions between subspaces which preserve certain information about vector operations. Let's make this precise.

Given two real vector spaces, V and W, a mapping $L : V \to W$ is called ***linear*** *if it preserves the operations of vector addition and multiplication by scalar, that is,*

(i) $L(\boldsymbol{v}_1 + \boldsymbol{v}_2) = L(\boldsymbol{v}_1) + L(\boldsymbol{v}_2)$ *for all* $\boldsymbol{v}_1, \boldsymbol{v}_2 \in V$, *and*

(ii) $L(c\boldsymbol{v}) = cL(\boldsymbol{v})$ *for all* $c \in \mathbb{R}$ *and all* $\boldsymbol{v} \in V$.

For example, functions f_2, f_4, f_5 (when $n = 1$) and f_6 are linear, and functions f_1, f_3 and f_5 (when $n \geq 2$) are not linear (are nonlinear).

The condition *(ii)* implies that

$$L(\mathbf{0}) = \mathbf{0} \tag{2.10}$$

Moreover, a repeated application of both conditions produces the generalization

$$L(c_1\boldsymbol{v}_1 + \ldots + c_n\boldsymbol{v}_n) = c_1L(\boldsymbol{v}_1) + \ldots + c_nL(\boldsymbol{v}_n) \tag{2.11}$$

for all $c_1, \ldots, c_n \in \mathbb{R}$ and all $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in V$.

Suppose that we are given a basis of V and we want to know the value of the linear mapping for a generic vector $\boldsymbol{x} \in V$. The crucial consequence of linearity is that if we know the value of the mapping for each vector in a basis of V, then we know its value for each vector in the entire space. Indeed, suppose that $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\} \subset V$ is a basis so that each $\boldsymbol{v}$ in V can be uniquely written as

$$\boldsymbol{v} = c_1\boldsymbol{v}_1 + \ldots + c_n\boldsymbol{v}_n$$

for unique scalars $c_1, \ldots, c_n \in \mathbb{R}$. The condition (2.11) then implies that

$$L(\boldsymbol{v}) = c_1L(\boldsymbol{v}_1) + \ldots + c_nL(\boldsymbol{v}_n)$$

The only freedom for a linear mapping is its value on the basis vectors. Once we settle the mapping of the basis vectors of V, the mapping of any vector in V is also settled.

Linear mappings are often called ***linear transformations***. A mapping will also be called a map, for the sake of brevity. A linear map from V into itself is called a linear ***operator*** on V. Like for classic terminology for functions, the vector space V is called the ***domain*** of L, and W is called the ***codomain*** of L. For $\boldsymbol{v} \in V$, the vector $L(\boldsymbol{v})$ is called the ***image*** of $\boldsymbol{v}$ under the action of L.

2.4.5 *The space of linear mappings*

Let V and W be two real vector spaces. We denote the space of all linear maps from V to W by

$$\mathcal{L}(V, W) = \{L : V \to W \text{ linear mapping}\}$$

We shall define the operations of addition multiplication by scalars in $\mathcal{L}(V, W)$ in such a way as to make it into a vector space:

- Given $L, F \in \mathcal{L}(V, W)$ their addition $L + F$ is the map $L + F : V \to W$, such that, for all $\boldsymbol{v} \in V$,

$$(L + F)(\boldsymbol{v}) = L(\boldsymbol{v}) + F(\boldsymbol{v})$$

It can be easily verified that the map $L + F$ satisfies the two conditions *(i)* and *(ii)* which define a linear map.

- Given $c \in \mathbb{R}$ and $L \in \mathcal{L}(V, W)$ their multiplication cL is the map $(cL) : V \to W$ such that, for all $\boldsymbol{v} \in V$,

$$(cL)(\boldsymbol{v}) = cL(\boldsymbol{v})$$

  Again, it can be easily verified that cL is a linear map.

- Given $L \in \mathcal{L}(V, W)$, we define the opposite map to be the map $(-1)L$.
- The zero map is $L(\mathbf{0}) = \mathbf{0}$.

We can now verify that the properties (V1)–(V8) defining a vector space are satisfied and conclude that the set of linear maps between two vector spaces is itself a vector space. Its dimension is

$$\dim(\mathcal{L}(V, W)) = \dim(V) \cdot \dim(W) \tag{2.12}$$

The reason why this identity is true will be explained at the end of this lecture.

2.4.6 *Image and kernel of linear mappings*

> *Let V and W be real vector spaces, and let $L : V \to W$ be a linear map. The **kernel** of L is the set of all elements in V, whose image in W is the zero vector,*
>
> $$\ker(L) = \{\boldsymbol{v} \in V : L(\boldsymbol{v}) = \mathbf{0}\} \subseteq V$$

A kernel is always a non-empty set, since $\mathbf{0} \in V$ is always an element of the kernel. It can be proved that the kernel of a linear map L is a vector subspace of V.

The kernel of a linear map is useful to determine when the map is injective.

> *A linear mapping $L : V \to W$ is **injective**, or **one-to-one**, if different vectors in V are mapped into different vectors in W, that is, if $\boldsymbol{v}_1, \boldsymbol{v}_2$ are elements of V such that $L(\boldsymbol{v}_1) = L(\boldsymbol{v}_2)$, then $\boldsymbol{v}_1 = \boldsymbol{v}_2$.*

Let $L : V \to W$ be a linear map. Then the following two conditions are equivalent:

- $\ker(L) = \{\mathbf{0}\}$;
- L is injective.

For example, consider linear mappings $f, g : \mathbb{R}^3 \to \mathbb{R}^3$ given by

$$f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \begin{bmatrix} x_1 + x_2 \\ x_2 + x_3 \\ x_3 + x_1 \end{bmatrix} \qquad g\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \begin{bmatrix} x_1 - x_2 \\ x_2 - x_3 \\ x_3 - x_1 \end{bmatrix} \tag{2.13}$$

The mapping f is injective since $\ker(f) = \mathbf{0}$, but g is not injective since

$$\ker(g) = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} : x_1 = x_2 = x_3,\ x_1, x_2, x_3 \in \mathbb{R} \right\}$$

An important property of injective mappings is that they respect linear independence of vectors.

> *Let $L : V \to W$ be an injective linear map. If $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are linearly independent elements of V, then $L(\boldsymbol{v}_1), \ldots, L(\boldsymbol{v}_n)$ are also linearly independent elements of W.*

For example, the mapping f given by (2.13) maps the natural basis vectors of $\mathbb{R}^3$ to vectors

$$f\left(\begin{bmatrix}1\\0\\0\end{bmatrix}\right) = \begin{bmatrix}1\\0\\1\end{bmatrix} \qquad f\left(\begin{bmatrix}0\\1\\0\end{bmatrix}\right) = \begin{bmatrix}1\\1\\0\end{bmatrix} \qquad f\left(\begin{bmatrix}0\\0\\1\end{bmatrix}\right) = \begin{bmatrix}0\\1\\1\end{bmatrix}$$

that are also linearly independent in $\mathbb{R}^3$.

We define the image of L to be the set of all elements in W that are the image through L of at least one element in V,

$$\text{im}(L) = \{\boldsymbol{w} \in W : \exists \boldsymbol{v} \in V \; s.t. \; L(\boldsymbol{v}) = \boldsymbol{w}\} \subseteq W$$

The image of L is a vector subspace of W. In particular, it is always non-empty since the zero vector in V is always mapped to the zero vector in W.

> *A linear map $L : V \to W$ is **surjective**, or **onto**, if the domain V is mapped through L to the entire codomain W, that is $\text{im}(V) = W$.*

For example, the image of the mapping f defined by (2.13) is $\text{im}(f) = \mathbb{R}^3$ and thus f is surjective, but

$$\text{im}(g) = \left\{ \begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} : x_1 + x_2 + x_3 = 0, \; x_1, x_2, x_3 \in \mathbb{R} \right\}$$

and thus g is not surjective. We will explain how this was obtained in the next subsection.

> *Let $L : V \to W$ be a linear map. Then*
>
> $$\dim(V) = \dim(\ker(L)) + \dim(\text{im}(L))$$

For example, we have $\dim(\ker(f)) = 0$ and $\dim(\text{im}(f)) = 3$, and $\dim(\ker(g)) = 1$ and $\dim(\text{im}(g)) = 2$. In both cases their sum is 3 which match $\dim(\mathbb{R}^3) = 3$.

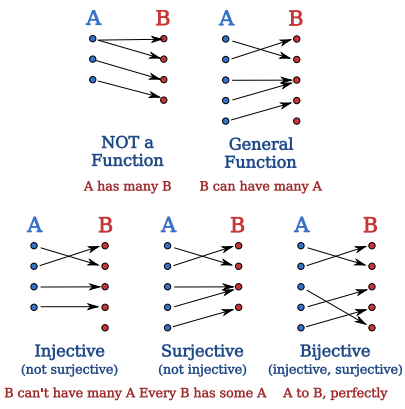
> *A linear map is called **bijective** if it is both injective and surjective.*

In other words, a map is bijective if each element of the codomain is mapped to by exactly one element of the domain.

> *Let V and W be two vector spaces for which $\dim(V) = \dim(W)$. Let $L : V \to W$ be a linear map. If L is injective, or if L is surjective, then L is bijective.*

For example, the mapping f is bijective, but g is not.

Possible mappings between sets A and B:



2.4.7 *Linear mappings and matrices*

Given an $A \in \mathbb{R}^{m\times n}$, we can always associate with A a linear map

$$L_A : \mathbb{R}^n \to \mathbb{R}^m$$

by defining

$$L_A(\boldsymbol{x}) = A\boldsymbol{x}$$

for any column vector $\boldsymbol{x} \in \mathbb{R}^n$. Both A and $\boldsymbol{x}$ are compatible for multiplication and the result of the product will be a column vector in $\mathbb{R}^m$. Linearity is an immediate consequence of properties of matrix multiplication. Indeed, we have that

$$L_A(\boldsymbol{x}_1 + \boldsymbol{x}_2) = A(\boldsymbol{x}_1 + \boldsymbol{x}_2) = A\boldsymbol{x}_1 + A\boldsymbol{x}_2 = L_A(\boldsymbol{x}_1) + L_A(\boldsymbol{x}_2)$$

for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$, and

$$L_A(c\boldsymbol{x}) = A(c\boldsymbol{x}) = cA\boldsymbol{x} = cL_A(\boldsymbol{x})$$

for all $c \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n$.

The L_A is called the linear map ***associated*** with the matrix A. It can be easily shown that different matrices give rise to different associated maps. In other words, if two matrices give rise to the same linear map, then they are necessarily equal.

We have seen above that any matrix in $\mathbb{R}^{m\times n}$ leads immediately to a linear map from $\mathbb{R}^n$ into $\mathbb{R}^m$. The opposite direction holds too.

Let $L : \mathbb{R}^n \to \mathbb{R}^m$ be a linear map. Then there exists a unique matrix A such that $L = L_A$. The column vectors of A are the values of the linear map in the vectors of the canonical basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ of $\mathbb{R}^n$. Precisely, for $j = 1, \ldots, n$, the jth column of A is found by applying L to the jth standard basis vector,

$$A = [L(\boldsymbol{e}_1) \ \ldots \ L(\boldsymbol{e}_n)] \in \mathbb{R}^{m\times n}$$

For example, the mappings f and g defined by (2.13) are associated with the matrices

$$F = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \qquad G = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix} \tag{2.14}$$

*Let $A \in \mathbb{R}^{m\times n}$. Then kernel of the associated linear mapping L_A is the **null space** of A, denoted by $\mathcal{N}(A)$. It is the subset of $\mathbb{R}^n$ given by*

$$A\boldsymbol{x} = \boldsymbol{0}$$

*The image of L_A is the **range** of A, denoted by $\mathcal{R}(A)$. It is the subset of $\mathbb{R}^m$ given by*

$$\mathcal{R}(A) = \{\boldsymbol{b} \in \mathbb{R}^m : A\boldsymbol{x} = \boldsymbol{b} \text{ for some } \boldsymbol{x} \in \mathbb{R}^n\}$$

Example 2.15: Kernel and image

We want to determine the kernel and the image of linear mappings L_F and L_G associated with matrices F and G in (2.14).

- To detremine $\ker(L_F)$ we need to solve the associated homogeneous system, $F\boldsymbol{x} = \boldsymbol{0}$. We start by row reducing the matrix F:

$$F = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \xrightarrow{\mathrm{r}_3 \to \mathrm{r}_3 + \mathrm{r}_2 - \mathrm{r}_1} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix} \xrightarrow{\mathrm{r}_3 \to 1/2\,\mathrm{r}_3} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \xrightarrow[\mathrm{r}_2 \to \mathrm{r}_2 - \mathrm{r}_3]{\mathrm{r}_1 \to \mathrm{r}_1 - \mathrm{r}_2 + \mathrm{r}_3} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

  Hence the equation $F\boldsymbol{x} = \boldsymbol{0}$ has a trivial solution only, and so

$$\ker(L_F) = \boldsymbol{0}$$

  Repating the same steps for L_G we find

$$G = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix} \xrightarrow{\mathrm{r}_3 \to \mathrm{r}_3 + \mathrm{r}_2 + \mathrm{r}_1} \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

  and so $G\boldsymbol{x} = \boldsymbol{0}$ has infinitely many solutions given by $x_1 = x_2 = x_3$, thus

$$\ker(L_G) = \{\boldsymbol{x} \in \mathbb{R}^3 : x_1 = x_2 = x_3\}$$

- To determine $\mathrm{im}(L_F)$ we need to find all $\boldsymbol{b}$ satisfying $F\boldsymbol{x} = \boldsymbol{b}$ for some $\boldsymbol{x} \in \mathbb{R}^3$. Applying the row reduction procedure to $(F|\boldsymbol{b})$ we find

$$(F|\boldsymbol{b}) = \left[\begin{array}{ccc|c} 1 & 1 & 0 & b_1 \\ 0 & 1 & 1 & b_2 \\ 1 & 0 & 1 & b_3 \end{array}\right] \longrightarrow \left[\begin{array}{ccc|c} 1 & 0 & 0 & 1/2\,(b_1 - b_2 + b_3) \\ 0 & 1 & 0 & 1/2\,(b_2 - b_3 + b_1) \\ 0 & 0 & 1 & 1/2\,(b_3 - b_1 + b_2) \end{array}\right]$$

  giving

$$\begin{aligned} x_1 &= 1/2\,(b_1 - b_2 + b_3) \\ x_2 &= 1/2\,(b_2 - b_3 + b_1) \\ x_3 &= 1/2\,(b_3 - b_1 + b_2) \end{aligned}$$

  and so for each $\boldsymbol{b} \in \mathbb{R}^3$ there exists $\boldsymbol{x} \in \mathbb{R}^3$ satisfying $F\boldsymbol{x} = \boldsymbol{b}$ implying that

$$\mathrm{im}(F) = \mathbb{R}^3$$

  Repeating the same steps for L_G we find

$$(G|\boldsymbol{b}) = \left[\begin{array}{ccc|c} 1 & -1 & 0 & b_1 \\ 0 & 1 & -1 & b_2 \\ -1 & 0 & 1 & b_3 \end{array}\right] \longrightarrow \left[\begin{array}{ccc|c} 1 & -1 & 0 & b_1 \\ 0 & 1 & -1 & b_2 \\ 0 & 0 & 0 & b_1 + b_2 + b_3 \end{array}\right]$$

  giving

$$\begin{aligned} x_1 - x_2 &= b_1 \\ x_2 - x_3 &= b_2 \\ 0 &= b_1 + b_2 + b_3 \end{aligned}$$

  and so

$$\mathrm{im}(L_G) = \left\{ \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \in \mathbb{R}^3 : b_1 + b_2 + b_3 = 0,\ b_1, b_2, b_3 \in \mathbb{R} \right\}$$

We have investigated so far the relation between matrices and linear maps of $\mathbb{R}^n$ into $\mathbb{R}^m$. A question arises naturally: do linear maps between arbitrary vector spaces over $\mathbb{R}$ have matrix representation? The answer, provided that the vector spaces are finite-dimensional, is yes. We will find out that, once the basis for each space are settled, the matrix representation of the map is unique.

Consider a finite-dimensional vector space V and let n be its dimension. Given a basis $\{v_1, \ldots, v_n\}$ of V, any vector $\boldsymbol{v} \in V$ admits the unique representation

$$\boldsymbol{v} = c_1\boldsymbol{v}_1 + \ldots + c_n\boldsymbol{v}_n$$

for suitable scalars $c_1, \ldots, c_n \in \mathbb{R}$. Then we can consider the bijective map

$$\Phi : \mathbb{R}^n \to V$$

defined by

$$(c_1, \ldots, c_n) \mapsto c_1\boldsymbol{v}_1 + \ldots + c_n\boldsymbol{v}_n$$

that maps the unique coordinates of any vector to the representation of the vector as a linear combination of the coordinates. We say that V is ***isomorphic*** to $\mathbb{R}^n$ under the map Φ and we write

$$V \cong \mathbb{R}^n$$

Assume we are given two finite-dimensional real vector spaces V and W and a linear map $L : V \to W$. Assume, moreover, that $\dim(V) = n$ and $\dim(W) = m$ and let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ and $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m\}$ be bases of V and W respectively. Using the isomorphism Φ, we can identify $V \cong \mathbb{R}^n$ and $W \cong \mathbb{R}^m$ and interpret L as linear map of $\mathbb{R}^n$ into $\mathbb{R}^m$, and thus we can associate a matrix with L. This matrix strongly depends on the choice of the basis. Precisely, it is the unique matrix A having the property that if we denote by $[\boldsymbol{x}]_{\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n\}}$ the (column) coordinate vector of a vector $\boldsymbol{v} \in V$, relative to the basis $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$, then $A[\boldsymbol{x}]_{\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n\}}$ is the (column) coordinate vector of $L(\boldsymbol{x})$, relative to the basis $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m\}$, that is

$$L(\boldsymbol{x})_{\{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_m\}} = A[\boldsymbol{x}]_{\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n\}}$$

The following notation

$$A^{\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n\}}_{\{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_m\}}(L)$$

is often used to denote that A is the matrix associated to the linear map L and to indicate the respective choices of the basis for V and W.

In short, the matrix A carries all the essential information. If the basis is known, and the matrix is known, then the transformation of every vector is known. The coding of the information is simple. To transform a space to itself, one basis is enough. A transformation from one space to another requires a basis for each.

2.5 Eigenanalysis

Lecture 10

Eigenanalysis play a central role in the principal component analysis (PCA) and dimensional reduction of multidimensional data that help to reduce the amount of data one has to handle and to minimize the information loss.

Principal components are new variables that are constructed as linear combinations of the initial variables. These combinations are constructed in such a way that the new variables are uncorrelated and most of the information within the initial variables is stored in the first components. An n-dimensional data has n principal components, but PCA tries to put maximum possible information into the first ones, so that, if one wants to reduce the dataset's dimensionality, one can focus their analysis on the first few components without suffering a great penalty in terms of information loss.

2.5.1 Eigenvalues and eigenvectors

*Let V be a real vector space and let $L : V \to V$ be a linear map. A nonzero vector $\boldsymbol{v} \in V$ is called an **eigenvector**, or an **eigenfunction** of L if there exists a scalar $\lambda \in \mathbb{R}$ such that*

$$L\boldsymbol{v} = \lambda \boldsymbol{v} \tag{2.15}$$

*The scalar λ is called an **eigenvalue** of L corresponding to the eigenvector $\boldsymbol{v}$. In other words, L acts on eigenvectors as multiplication by scalars (the corresponding eigenvalues).*

The prefix eigen- is adopted from the German word "eigen", which means "owned by" or "peculiar", "specific", "characteristic" to. It was the mathematician David Hilbert who introduced the terms Eigenwert and Eigenfunktion. Eigenvalues and eigenvectors are equivalently called characteristic values and characteristic vectors, proper values and proper vectors, or latent values and latent vectors.

Note that the definition of eigenvector requires that it is a nonzero vector. In fact, if $\boldsymbol{v} = \boldsymbol{0}$ were allowed, then any scalar λ would be an eigenvalue since $L\boldsymbol{0} = \lambda\boldsymbol{0}$ holds for any $\lambda \in \mathbb{R}$. On the other hand, we can have $\lambda = 0$ and $\boldsymbol{v} \neq \boldsymbol{0}$.

Geometrically, (2.15) means that eigenvectors are those vectors that experience only changes in magnitude or sign under the action of L. Specifically, the eigenvalue λ is simply the amount of "stretch" or "shrink" to which the eigenvector $\boldsymbol{x}$ is subjected when transformed by L. For instance, $L\boldsymbol{x} = 2\boldsymbol{x}$ says that $\boldsymbol{x}$ is stretched by factor 2 under transformation by L. Similarly, $L\boldsymbol{x} = (-1/3)\boldsymbol{x}$ says that L reverses $\boldsymbol{x}$ and shrinks it by a factor $1/3$.

Let $A \in \mathbb{R}^{n\times n}$. An eigenvector of A is an eigenvector of the associated linear map $L_A : \mathbb{R}^n \to \mathbb{R}^n$, defined by $L_A\boldsymbol{x} = A\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^n$. Thus an eigenvector of A is a nonzero vector $\boldsymbol{x} \in \mathbb{R}^n$ for which there exists $\lambda \in \mathbb{R}$ such that $A\boldsymbol{x} = \lambda\boldsymbol{x}$.

Eigenvalues and eigenvectors in Python:

```
import numpy as np
from numpy import linalg as la
A = np.array([[1,2,3],
              [2,1,0],
              [3,0,2]])
# Eigenvalues and eigenvectors
values, vectors = la.eig(A)
# Verify the result
A @ vectors - values * vectors
```

Here `vectors` is a matrix eigenvectors and `values` is an array of eigenvalues

Here are some basic examples:

- Let $A \in \mathbb{R}^{n\times n}$ be given by

$$A = \begin{bmatrix}\lambda_1\boldsymbol{e}_1 & \dots & \lambda_n\boldsymbol{e}_n\end{bmatrix} = \begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix}$$

  Then each λ_i is an eigenvalue of A with eigenvector $\boldsymbol{e}_i$:

$$A\boldsymbol{e}_i = \lambda_i\boldsymbol{e}_i \quad \text{for} \quad 1 \leq i \leq n$$

- Let $A \in \mathbb{R}^{2\times 2}$ and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^2$ be given by

$$A = \begin{bmatrix}4 & 2\\ 2 & 4\end{bmatrix} \qquad \boldsymbol{v}_1 = \begin{bmatrix}1\\ 1\end{bmatrix} \qquad \boldsymbol{v}_2 = \begin{bmatrix}1\\ -1\end{bmatrix} \tag{2.16}$$

Then $\lambda_1 = 6$ and $\lambda_2 = 2$ are eigenvalues of A with eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, respectively, that is

$$A\boldsymbol{v}_1 = 6\boldsymbol{v}_1 \qquad A\boldsymbol{v}_2 = 2\boldsymbol{v}_2 \tag{2.17}$$

Eigenvectors are never unique. If $\boldsymbol{v}$ is an eigenvector for L with associated eigenvalue λ, then so is $c\boldsymbol{v}$, for any non-zero $c \in \mathbb{R}$. Indeed,

$$L(c\boldsymbol{v}) = cL(\boldsymbol{v}) = c(\lambda\boldsymbol{v}) = \lambda(c\boldsymbol{v})$$

More generally, let

$$E_\lambda = \{\boldsymbol{v} \in V : L\boldsymbol{v} = \lambda\boldsymbol{v}\}$$

that is, E_λ is the set of all eigenvectors having eigenvalue λ and the zero vector $\mathbf{0}$. The set E_λ is called the ***eigenspace*** of L corresponding to λ. Observe that $L\boldsymbol{v} = \lambda\boldsymbol{v}$ can be rewritten as $(L - \lambda I)\boldsymbol{v} = \mathbf{0}$. This means that

$$E_\lambda = \ker(L - \lambda I)$$

> *Let V be a finite dimensional vector space and let $L : V \to V$ be a linear map. Then $\lambda \in \mathbb{R}$ is an eigenvalue of L*
>
> - *if and only if $E_\lambda \neq \{\mathbf{0}\}$, or equivalently*
> - *if and only if $\det(L - \lambda I)$ is not invertible.*

The following theorem is an important result that states that distinct eigenvalues correspond to linearly independent eigenvectors.

> *Let V be a finite dimensional vector space and let $L : V \to V$ be a linear map. Let $\lambda_1, \ldots, \lambda_n$ be eigenvalues of L, and let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be associated eigenvectors. If $\lambda_i \neq \lambda_j$ for any $i \neq j$, then the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are linearly independent.*

For example, the matrix A in (2.16) has two distinct eigenvalues, $\lambda_1 = 6$ and $\lambda_2 = 2$. The associated eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are indeed linearly independent: the vector equation $c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 = \mathbf{0}$ has a trivial solution only.

> *Let V be a finite dimensional vector space and let $L : V \to V$ be a linear map having n eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ whose eigenvalues $\lambda_1, \ldots, \lambda_n$ are distinct. Then $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ is a basis of V.*

An important consequence of the statement above is that any $A \in \mathbb{R}^{n\times n}$ can have at most n distinct eigenvalues. Indeed, if A had $m > n$ distinct eigenvalues, then the associated eigenvectors would be a set of m linearly independent vectors of $\mathbb{R}^n$, which is impossible.

2.5.2 *Computation of eigenpairs*

We address the following questions. If V is a real vector space, and $L : V \to V$ is a linear map, how can we find out the set of eigenvalues of L? Once the latest are at our disposal, how can we find out the associated eigenvectors (or eigenspaces)?

*Given $A \in \mathbb{R}^{n\times n}$, the **characteristic polynomial** associated with A is defined by*

$$p_A(\lambda) = \det(A - \lambda I)$$

By expanding $\det(A - \lambda I)$ using Laplace rule according to the first column, it is possible to show that

$$\begin{aligned} p_A(\lambda) &= \det(A - \lambda I) \\ &= (a_{11} - \lambda)\cdots(a_{nn} - \lambda) + \ldots \\ &= (-1)^n\lambda^n + \text{lower order terms in } \lambda \end{aligned}$$

Thus we deduce that $p_A(\lambda)$ is a polynomial of degree n in the variable λ, whose leading term is $(-1)^n\lambda^n$.

The scalar λ is an eigenvalue for any given matrix $A \in \mathbb{R}^{n\times n}$ if and only if $p_A(\lambda) = 0$. The equation $p_A(\lambda) = 0$ is called the ***characteristic equation*** of the matrix A. Determinants give us a direct way of computing eigenvalues of a matrix, since the latest are the solutions of the characteristic equation or, equivalently, the roots of the characteristic polynomial. Clearly, this is possible whenever we can determine explicitly the roots of $p_A(\lambda)$. Sometimes this is an easy task, but in the majority of cases it is a formidable task.

For example, let

$$A = \begin{bmatrix} 1 & 9 \\ 1 & 1 \end{bmatrix}$$

Then

$$p_A(\lambda) = \det\begin{bmatrix} 1-\lambda & 9 \\ 1 & 1-\lambda \end{bmatrix} = (\lambda - 1)^2 - 3^2 = (\lambda - 4)(\lambda + 2)$$

Use the formula:
$a^2 - b^2 = (a - b)(a + b)$

Hence there are two eigenvalues: $\lambda_1 = 4$, and $\lambda_2 = -2$.

Once the eigenvalues are known, it is straightforward to compute the corresponding eigenvectors and eigenspaces. Indeed, in order to find the eigenvectors corresponding to λ_1 and λ_2, we only need to find a nontrivial solution to the corresponding homogeneous equation

$$(A - \lambda_i I)\boldsymbol{v}_i = \mathbf{0}$$

Continuing from the previous example, let $\boldsymbol{v}_1 \in \mathbb{R}^2$ be eigenvector with eigenvalue $\lambda_1 = 4$. Then

$$\begin{bmatrix} 1-\lambda_1 & 9 \\ 1 & 1-\lambda_1 \end{bmatrix}\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} -3 & 9 \\ 1 & -3 \end{bmatrix}\begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This leads to a system of linear equations,

$$\begin{aligned} -3v_{11} + 9v_{12} &= 0 \\ v_{11} - 3v_{12} &= 0 \end{aligned}$$

Notice that the equations are multiples of each other, so it is sufficient to solve one of them, say $v_{11} - 3v_{12} = 0$, giving $(v_{11}, v_{12}) = (s, 3s)$ for any non-zero $s \in \mathbb{R}$. Hence the set of all eigenvectors is a line with the origin missing, and the eigenspace for $\lambda_1 = 4$ is

$$E_{\lambda_1} = E_4 = \left\{ s\begin{bmatrix} 1 \\ 3 \end{bmatrix} : s \in \mathbb{R} \right\}$$

The value $s = 0$ is also included since the eigenspace contains the zero vector, $\mathbf{0}$, by definition.

As a second example, consider the matrix

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

In this case we find

$$p_A(\lambda) = \det \begin{bmatrix} -\lambda & 1 \\ -1 & -\lambda \end{bmatrix} = \lambda^2 + 1$$

The characteristic equation $\lambda^2 + 1 = 0$ has no solutions in real numbers. However it does have solutions in complex numbers. Thus considering A as an element in $\mathbb{C}^{2\times 2}$ we have that $p_A(\lambda) = (\lambda - i)(\lambda + 1)$, and so A has two complex eigenvalues, $\lambda_1 = i$ and $\lambda_2 = -i$. The corresponding eigenvectors are scalar multiples of the vectors

$$v_1 = \begin{bmatrix} 1 \\ i \end{bmatrix} \quad \text{and} \quad v_2 = \begin{bmatrix} 1 \\ -i \end{bmatrix}$$

respectively.

The set of complex numbers, $\mathbb{C}$, is the set of all numbers expressed in the form $a + ib$, where a and b are real numbers, and the symbol i is an "imaginary" number satisfying $i^2 = -1$, that is

$$\mathbb{C} = \{a + ib : a, b \in \mathbb{R},\ i^2 = -1\}$$

You can think of $\mathbb{C}$ as the set of all points on a plane.

Note that in Python, the symbol i is replaced with j.

Let's complete the discussion by recalling a powerful result given by the ***fundamental theorem of algebra***, which states that every polynomial of degree n in the domain of complex numbers has exactly n roots. Some of these roots may be complex numbers (even if all the coefficients in the polynomial are real), and some roots may be repeated. Therefore, altogether any $n \times n$ matrix A over $\mathbb{C}$ has exactly n eigenvalues. Some eigenvalues might lie in $\mathbb{C}$, while others might be repeated. In particular, if the entries of A lie in $\mathbb{R} \subset \mathbb{C}$, complex eigenvalues occur in conjugate pairs, i.e., if $\lambda \in \sigma(A)$, then $\overline{\lambda} \in \sigma(A)$. This is a consequence of the fact that the roots of a polynomial with real coefficients occur in conjugate pairs.

Each complex number $c = a + ib$ has its complex conjugate given by $\bar{c} = a - ib$. Let a be the x-coordinate, and let b be the y-coordinate. Then the numbers c and $\bar{c}$ are related to each other via the reflection in the x-axis.

Let's see now some examples of computations of particularly favorable matrices.

Diagonal matrix. The eigenvalues of a diagonal matrix are its diagonal entries. Indeed, consider the 2×2 diagonal matrix

$$D = \begin{bmatrix} -2 & 0 \\ 0 & 3 \end{bmatrix}$$

It is easy to verify that

$$D \begin{bmatrix} 1 \\ 0 \end{bmatrix} = -2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{and} \quad D \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 3 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and so D has eigenvalues $\lambda_1 = -2$ and $\lambda_2 = 3$ with the corresponding eigenspaces

$$E_{-2} = \left\{ c \begin{bmatrix} 1 \\ 0 \end{bmatrix} : c \in \mathbb{R} \right\} \qquad E_3 = \left\{ c \begin{bmatrix} 0 \\ 1 \end{bmatrix} : c \in \mathbb{R} \right\}$$

Projection matrix. The only possible eigenvalues for any projection matrix are 1 and 0. Indeed, consider the 2×2 projection matrix

$$\Pi = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Then

$$\Pi \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \Pi \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 0 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and so Π has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 0$ with the corresponding eigenspaces

$$E_1 = \left\{ c \begin{bmatrix} 1 \\ 1 \end{bmatrix} : c \in \mathbb{R} \right\} \qquad E_0 = \left\{ c \begin{bmatrix} 1 \\ -1 \end{bmatrix} : c \in \mathbb{R} \right\}$$

Everytime $\lambda = 1$, the associated eigenvector is projected into itself, and everytime $\lambda = 0$, the associated eigenvector is projected to the zero vector. Like every other scalar, zero might or might be not an eigenvalue and there is nothing exceptional about that. There is an information that we can deduce if zero is an eigenvalue and it is that the matrix is singular (not invertible), i.e., its determinant is zero. Invertible matrices must have all eigenvalues different from zero.

A matrix $\Pi \in \mathbb{R}^{n \times n}$ is called a *projection matrix* if

$$\Pi^2 = \Pi$$

For example, if

$$\Pi = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

then

$$\Pi^2 = \frac{1}{4}\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} = \Pi$$

Triangular matrix. Every triangular matrix as has the eigenvalues sitting along the main diagonal. Indeed, consider the 3×3 matrix

$$A = \begin{bmatrix} 1 & 9 & 3 \\ 0 & 3 & -1 \\ 0 & 0 & -4 \end{bmatrix}$$

Since the determinant of any triangular matrix is the product of the diagonal entries, we have that

$$p_A(\lambda) = \det \begin{bmatrix} 1-\lambda & 9 & 3 \\ 0 & 3-\lambda & -1 \\ 0 & 0 & -4-\lambda \end{bmatrix} = (1-\lambda)(3-\lambda)(-4-\lambda)$$

giving $\lambda_1 = 1$, $\lambda_2 = 3$ and $\lambda_3 = -4$, which are the diagonal entries of A.

2.5.3 *Diagonalisation of a matrix*

> *Two matrices $A, B \in \mathbb{R}^{n \times n}$ are called **similar** if there exists an invertible matrix P such that $B = P^{-1}AP$. Going from one to the other is called a **similarity transformation**.*

Similar matrices have the same characteristic equation, hence the same eigenvalues. Indeed, by exploiting the properties of the determinant one can easily deduce that

$$\det(B) = \det(P^{-1}AP) = \det(A)$$

In particular, we recall the different matrices corresponding to the same linear map are all similar, therefore they all have the same characteristic equation. This implies that the eigenvalues of a linear map are independent on the particular choice of the basis.

*A matrix $A \in \mathbb{R}^{n\times n}$ is called **diagonalizable** if it is similar to a diagonal matrix.*

There is a strong connection between diagonalizable matrices and eigenvectors. More specifically, suppose that A has n linearly independent eigenvectors. Let S be the square matrix whose columns are given by the eigenvectors of A. Then, it turns out that the matrix $S^{-1}AS$ is a diagonal matrix D whose diagonal entries are the eigenvalues of A:

$$S^{-1}AS = D = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

We will refer to S as the ***eigenvector matrix*** and D the ***eigenvalue matrix***. The eigenvector matrix S converts the matrix A into a diagonal matrix (its eigenvalue matrix D).

For example, consider a matrix $A \in \mathbb{R}^{2\times 2}$ and vectors $v_1, v_2 \in \mathbb{R}^2$ given by

$$A = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix} \qquad v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad v_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

We know that (cf. (2.17)) $Av_1 = 6v_1$ and $Av_2 = 2v_2$. Hence A must be similar to the diagonal matrix $D = \operatorname{diag}(6, 2)$. Indeed,

$$\frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 6 & 2 \\ 6 & -2 \end{bmatrix} = \begin{bmatrix} 6 & 0 \\ 0 & 2 \end{bmatrix}$$

We have used the formula

$$S^{-1} = \frac{1}{\det S}\begin{bmatrix} s_{22} & -s_{12} \\ -s_{21} & s_{11} \end{bmatrix}$$

to compute the matrix inverse.

If the matrix A has n distinct eigenvalues, then the corresponding n eigenvectors are automatically linearly independent. Consequently, any $n \times n$ matrix with n distinct eigenvalues can by diagonalized. The converse, however, is not true. Matrices that have repeated eigenvalues may be a case of not diagonalizable matrices, but in general it is not always so. Diagonalisability only depends on whether the corresponding n eigenvectors are linearly independent or not. An obvious counter-example is given by the identity matrix of order n, which has only a single eigenvalue, the scalar 1, repeated n times, but it is diagonal already.

The only possible matrices S that diagonalize a given matrix A are those whose columns are eigenvectors of A. Other choices of S will not produce a diagonal matrix from $S^{-1}AS$.

The eigenvector matrix S is not unique. Indeed, each eigenvector can be multiplied by a nonzero scalar and it remains an eigenvector. If we pre-multiply the columns of S by any nonzero constant, we will get a new valid S for which $S^{-1}AS = D$.

Not all matrices are diagonalizable since not all matrices posses n distinct eigenvectors and we cannot construct S in those cases. These are called ***defective matrices***.

2.5.4 *Diagonalisation of a symmetric matrix*

Symmetric real matrices enjoy some nicer properties. Their eigenvalues are all real, eigenvectors corresponding to distinct eigenvalues are orthogonal, and those matrices can be diagonalized by real orthogonal matrices.

> *A matrix $Q \in \mathbb{R}^{n\times n}$ is called **orthogonal** if $Q^TQ = I$. That is, if Q is invertible with $Q^{-1} = Q^T$.*

An example of an orthogonal matrix is

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

Indeed,

$$Q^TQ = \frac{1}{4}\begin{bmatrix} \sqrt{3} & -1 \\ 1 & \sqrt{3} \end{bmatrix}\begin{bmatrix} \sqrt{3} & 1 \\ -1 & \sqrt{3} \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 3+1 & \sqrt{3}-\sqrt{3} \\ \sqrt{3}-\sqrt{3} & 1+3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

To summarise:

> *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then:*
> - *there exists a real orthogonal matrix Q such that $Q^TAQ = D$;*
> - *it has n real eigenvalues;*
> - *its eigenvectors corresponding to distinct eigenvalues are orthogonal.*

For example, consider a symmetric matrix $A \in \mathbb{R}^{2\times 2}$ and vectors $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^2$ given by

$$A = \frac{1}{4}\begin{bmatrix} 5 & 3\sqrt{3} \\ 3\sqrt{3} & -1 \end{bmatrix} \qquad \boldsymbol{v}_1 = \begin{bmatrix} \sqrt{3} \\ 1 \end{bmatrix} \qquad \boldsymbol{v}_2 = \begin{bmatrix} -1 \\ \sqrt{3} \end{bmatrix}$$

Then it is easy to verify that

$$A\boldsymbol{v}_1 = 2\boldsymbol{v}_1, \qquad A\boldsymbol{v}_2 = -\boldsymbol{v}_2$$

and

$$Q^TAQ = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}$$

Example 2.16: Diagonalisation of a symmetric matrix

We want to diagonalise the symmetric matrix

$$A = \begin{bmatrix} 5 & 2 & -4 \\ 2 & 5 & -4 \\ -4 & -4 & 11 \end{bmatrix}$$

- We start by computing the characteristic polynomial $p_A(\lambda) = \det(A - \lambda I)$:

$$p_A(\lambda) = \det \begin{bmatrix} 5-\lambda & 2 & -4 \\ 2 & 5-\lambda & -4 \\ -4 & -4 & 11-\lambda \end{bmatrix} \xrightarrow{r_3 \to r_3 + 2r_2} \det \begin{bmatrix} 5-\lambda & 2 & -4 \\ 2 & 5-\lambda & -4 \\ 0 & 6-2\lambda & 3-\lambda \end{bmatrix}$$

$$\xrightarrow{c_2 \to c_2 - 2c_3} \det \begin{bmatrix} 5-\lambda & 10 & -4 \\ 2 & 13-\lambda & -4 \\ 0 & 0 & 3-\lambda \end{bmatrix} \xrightarrow{c_1 \to c_1 - \frac{2}{13-\lambda} c_2} \det \begin{bmatrix} 5-\lambda+\frac{20}{13-\lambda} & 10 & -4 \\ 0 & 13-\lambda & -4 \\ 0 & 0 & 3-\lambda \end{bmatrix}$$

$$= ((5-\lambda)(13-\lambda)+20)(3-\lambda) = (15-\lambda)(3-\lambda)^2$$

  Thus the eigenvalues of A are $\lambda_1 = 15$ and $\lambda_2 = \lambda_3 = 3$.

- Next, we need to find the associated eigenvectors. For this we need to solve the homogeneous equation $(A - \lambda I)\boldsymbol{x} = \mathbf{0}$. Let $\lambda = 15$. Then

$$A - 15I = \begin{bmatrix} -10 & 2 & -4 \\ 2 & -10 & -4 \\ -4 & -4 & -4 \end{bmatrix} \xrightarrow[r_3 \to r_3 + 2r_2]{r_1 \to r_1 + 5r_2} \begin{bmatrix} 0 & -48 & -24 \\ 2 & -10 & -4 \\ 0 & -24 & -12 \end{bmatrix}$$

$$\xrightarrow[r_1 \to -1/48\, r_1]{\substack{r_3 \to r_3 - 1/2\, r_1 \\ r_2 \to 1/2\, r_2}} \begin{bmatrix} 0 & 1 & 1/2 \\ 1 & -5 & -2 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{r_2 \to r_2 + 5r_1} \begin{bmatrix} 0 & 1 & 1/2 \\ 1 & 0 & 1/2 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{r_1 \leftrightarrow r_2} \begin{bmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 0 & 0 & 0 \end{bmatrix}$$

  Hence the general solution to $(A - 15I)\boldsymbol{x} = \mathbf{0}$ is $(x_1, x_2, x_3) = (-1/2s, -1/2s, s)$ with $s \in \mathbb{R}$; its vector form is

$$\boldsymbol{x} = s \begin{bmatrix} -1/2 \\ -1/2 \\ 1 \end{bmatrix} \quad \text{with} \quad s \in \mathbb{R}$$

- Now let $\lambda = 3$. Then

$$A - 3I = \begin{bmatrix} 2 & 2 & -4 \\ 2 & 2 & -4 \\ -4 & -4 & 8 \end{bmatrix} \xrightarrow[r_3 \to r_3 + 2r_1]{r_2 \to r_2 - r_1} \begin{bmatrix} 2 & 2 & -4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{r_1 \to 1/2\, r_1} \begin{bmatrix} 1 & 1 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

  Hence the general solution to $(A - 3I)\boldsymbol{x} = \mathbf{0}$ is $(x_1, x_2, x_3) = (2u - t, t, u)$ with $t, u \in \mathbb{R}$; its vector form is

$$\boldsymbol{x} = t \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + u \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \quad \text{with} \quad t, u \in \mathbb{R}$$

- We can choose the three linearly independent eigenvectors of A to be

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \qquad \boldsymbol{v}_2 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \qquad \boldsymbol{v}_3 = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$$

Then set

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

To find the inverse S^{-1} we row reduce the matrix

$$(S|I) = \left[\begin{array}{ccc|ccc} 1 & -1 & 2 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ -2 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \xrightarrow[\mathrm{r}_1 \to \mathrm{r}_1 + \mathrm{r}_2]{\mathrm{r}_3 \to \mathrm{r}_3 + \mathrm{r}_1 + \mathrm{r}_2} \left[\begin{array}{ccc|ccc} 2 & 0 & 2 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 3 & 1 & 1 & 1 \end{array}\right]$$

$$\xrightarrow[\mathrm{r}_3 \to 1/3\,\mathrm{r}_3]{\mathrm{r}_1 \to 1/2\,\mathrm{r}_1} \left[\begin{array}{ccc|ccc} 1 & 0 & 1 & 1/2 & 1/2 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1/3 & 1/3 & 1/3 \end{array}\right] \xrightarrow[\mathrm{r}_2 \to \mathrm{r}_2 - \mathrm{r}_1 + \mathrm{r}_3]{\mathrm{r}_1 \to \mathrm{r}_1 - \mathrm{r}_3} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1/6 & 1/6 & -1/3 \\ 0 & 1 & 0 & -1/6 & 5/6 & 1/3 \\ 0 & 0 & 1 & 1/3 & 1/3 & 1/3 \end{array}\right]$$

giving

$$S^{-1} = \frac{1}{6} \begin{bmatrix} 1 & 1 & -2 \\ -1 & 5 & 2 \\ 2 & 2 & 2 \end{bmatrix}$$

Therefore

$$S^{-1}AS = \frac{1}{6} \begin{bmatrix} 1 & 1 & -2 \\ -1 & 5 & 2 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} 5 & 2 & -4 \\ 2 & 5 & -4 \\ -4 & -4 & 11 \end{bmatrix} \begin{bmatrix} 1 & -1 & 2 \\ 1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & -2 \\ -1 & 5 & 2 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} 5 & -1 & 2 \\ 5 & 1 & 0 \\ -10 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 15 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

- We could have chosen the three linearly independent eigenvectors of A to be

$$v_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \qquad v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \qquad v_3 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{2.18}$$

Then

$$Q = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 & -\sqrt{3} & \sqrt{2} \\ 1 & \sqrt{3} & \sqrt{2} \\ -2 & 0 & \sqrt{2} \end{bmatrix}$$

is an orthogonal matrix, $Q^TQ = I$, and $Q^TAQ = \mathrm{diag}(15, 3, 3)$. The eigenvectors (2.18) form an orthogonal basis of $\mathbb{R}^3$ and are obtained by the Gram-Schmidt procedure; this procedure is beyond the scope of these lectures.