Classnotes - MA1102

Series and Matrices

Arindama Singh
Department of Mathematics
Indian Institute of Technology Madras

This write-up is bare-minimum, lacking motivation.
So, it is not a substitute for attending classes.
It will help you if you have missed a class.
Have suggestions? Contact asingh@iitm.ac.in.

Contents

Part I

Series

Chapter 1

Series of Numbers

1.1 Preliminaries

We use the following notation:

$\varnothing$ = the empty set.

$\mathbb{N} = \{1, 2, 3, \ldots\}$, the set of natural numbers.

$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$, the set of integers.

$\mathbb{Q} = \{\frac{p}{q} : p \in \mathbb{Z},\ q \in \mathbb{N}\}$, the set of rational numbers.

$\mathbb{R}$ = the set of real numbers.

$\mathbb{R}_+$ = the set of all positive real numbers.

$\mathbb{N} \subsetneq \mathbb{Z} \subsetneq \mathbb{Q} \subsetneq \mathbb{R}$. The numbers in $\mathbb{R} - \mathbb{Q}$ is the set of irrational numbers.

Examples are $\sqrt{2}$, $3.10110111011110\cdots$ etc.

Along with the usual laws of $+,\ \cdot, <$, $\mathbb{R}$ satisfies the **Archimedian property**:

If $a > 0$ and $b > 0$, then there exists an $n \in \mathbb{N}$ such that $na \geq b$.

Also $\mathbb{R}$ satisfies the **completeness property**:

Every nonempty subset of $\mathbb{R}$ having an upper bound has a least upper bound (**lub**) in $\mathbb{R}$.

Explanation: Let A be a nonempty subset of $\mathbb{R}$. A real number u is called an upper bound of A iff each element of A is less than or equal to u. An upper bound ℓ of A is called a least upper bound iff all upper bounds of A are greater than or equal to ℓ.

Notice that $\mathbb{Q}$ does not satisfy the completeness property. For example, the nonempty set $A = \{x \in \mathbb{Q} : x^2 < 2\}$ has an upper bound, say, 2. But its least upper bound is $\sqrt{2}$, which is not in $\mathbb{Q}$.

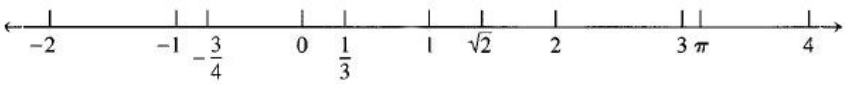
Let A be a nonempty subset of $\mathbb{R}$. A real number v is called a lower bound of A iff each element of A is greater than or equal to v. A lower bound m of A is called a greatest lower bound iff all lower bounds of A are less than or equal to m. The completeness property of $\mathbb{R}$ implies that

Every nonempty subset of $\mathbb{R}$ having a lower bound has a greatest lower bound (**glb**) in $\mathbb{R}$.

When the $\text{lub}(A) \in A$, this lub is defined as the **maximum of** A and is denoted as $\max(A)$. Similarly, if the $\text{glb}(A) \in A$, this glb is defined as the **minimum of** A and is denoted by $\min(A)$.

Moreover, both $\mathbb{Q}$ and $\mathbb{R} - \mathbb{Q}$ are **dense** in $\mathbb{R}$. That is, if $x < y$ are real numbers then there exist a rational number a and an irrational number b such that $x < a < y$ and $x < b < y$.

These properties allow $\mathbb{R}$ to be visualized as a number line: $\mathbb{R}$ is a straight line made of expansible rubber of no thickness!



From the archemedian property it follows that the **greatest interger function** is well defined. That is, for each $x \in \mathbb{R}$, there corresponds, the number $[x]$, which is the greatest integer less than or equal to x. Moreover, the correspondence $x \mapsto [x]$ is a function.

Let $a, b \in \mathbb{R},\ a < b$.

$[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$, the closed interval $[a, b]$.

$(a, b] = \{x \in \mathbb{R} : a < x \leq b\}$, the semi-open interval $(a, b]$.

$[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$, the semi-open interval $[a, b)$.

$(a, b) = \{x \in \mathbb{R} : a < x < b\}$, the open interval (a, b).

$(-\infty, b] = \{x \in \mathbb{R} : x \leq b\}$, the closed infinite interval $(-\infty, b]$.

$(-\infty, b) = \{x \in \mathbb{R} : x < b\}$, the open infinite interval $(-\infty, b)$.

$[a, \infty) = \{x \in \mathbb{R} : x \geq a\}$, the closed infinite interval $[a, \infty)$.

$(a, \infty) = \{x \in \mathbb{R} : x \leq b\}$, the open infinite interval (a, ∞).

$(-\infty, \infty) = \mathbb{R}$.

We also write $\mathbb{R}_+$ for $(0, \infty)$ and $\mathbb{R}_-$ for $(-\infty, 0)$. These are, respectively, the set of all positive real numbers, and the set of all negative real numbers.

A **neighborhood** of a point c is an open interval $(c - \delta, c + \delta)$ for some $\delta > 0$.

The **absolute value** of $x \in \mathbb{R}$ is defined as $|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0. \end{cases}$

Thus $|x| = \sqrt{x^2}$. And $|-a| = a$ or $a \geq 0$; $|x - y|$ is the distance between real numbers x and y. Moreover, if $a, b \in \mathbb{R}$, then

$$|-a| = |a|, \quad |ab| = |a|\,|b|, \quad \left|\frac{a}{b}\right| = \frac{|a|}{|b|} \text{ if } b \neq 0, \quad |a + b| \leq |a| + |b|, \quad |\,|a| - |b|\,| \leq |a - b|.$$

Let $x \in \mathbb{R}$ and let $a > 0$. The following are true:

1. $|x| = a$ iff $x = \pm a$.
2. $|x| < a$ iff $-a < x < a$ iff $x \in (-a, a)$.
3. $|x| \leq a$ iff $-a \leq x \leq a$ iff $x \in [-a, a]$.
4. $|x| > a$ iff $-a < x$ or $x > a$ iff $x \in (-\infty, -a) \cup (a, \infty)$ iff $x \in \mathbb{R} \setminus [-a, a]$.
5. $|x| \geq a$ iff $-a \leq x$ or $x \geq a$ iff $x \in (-\infty, -a] \cup [a, \infty)$ iff $x \in \mathbb{R} \setminus (-a, a)$.
6. For $a \in \mathbb{R}, \delta > 0, \quad |x - a| < \delta$ iff $a - \delta < x < a + \delta$.

The following statements are useful in proving equalities from inequalities:

Let $a, b \in \mathbb{R}$.

1. If for each $\epsilon > 0,\ |a| < \epsilon$, then $a = 0$.
2. If for each $\epsilon > 0,\ 0 \leq a < \epsilon$, then $a = 0$.
3. If for each $\epsilon > 0,\ a < b + \epsilon$, then $a \leq b$.

1.2 Sequences

A **sequence of real numbers** is a function $f : \mathbb{N} \to \mathbb{R}$. The values of the function are $f(1),\ f(2),\ f(3),\ \ldots$ These are called the **terms of the sequence**. With $f(n) = x_n$, the nth term of the sequence, we write the sequence in many ways such as

$$(x_n) = (x_n)_{n=1}^{\infty} = \{x_n\}_{n=1}^{\infty} = \{x_n\} = (x_1, x_2, x_3, \ldots)$$

showing explicitly its terms. For example, $x_n = n$ defines the sequence

$$f : \mathbb{N} \to \mathbb{R} \ \text{ with } f(n) = n,$$

that is, the sequence is $(1, 2, 3, 4, \ldots)$, the sequence of natural numbers. Informally, we say "the sequence $x_n = n$."

The sequence $x_n = 1/n$ is the sequence $(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots)$; formally, $\{1/n\}$ or $(1/n)$.

The sequence $x_n = 1/n^2$ is the sequence $(1/n^2)$, or $\{1/n^2\}$, or $(1, \frac{1}{4}, \frac{1}{9}, \frac{1}{16}, \ldots)$.

The **constant sequence** $\{c\}$ for a given real number c is the constant function $f : \mathbb{N} \to \mathbb{R}$, where $f(n) = c$ for each $n \in \mathbb{N}$. It is $(c, c, c, \ldots)$.

A sequence is an infinite list of real numbers; it is ordered like natural numbers, and unlike a set of numbers.

Let $\{x_n\}$ be a sequence. Let $a \in \mathbb{R}$. We say that $\{x_n\}$ **converges to** a iff for each $\epsilon > 0$, there exists an $m \in \mathbb{N}$ such that if $n \geq m$ is any natural number, then $|x_n - a| < \epsilon$.

Example 1.1. Show that the sequence $\{1/n\}$ converges to 0.

Let $\epsilon > 0$. Take $m = [1/\epsilon] + 1$. That is, m is the natural number such that $m - 1 \leq 1/\epsilon < m$. Then $1/m < \epsilon$. Moreover, if $n > m$, then $1/n < 1/m < \epsilon$. That is, for any such given $\epsilon > 0$, there exists an m, (we have defined it here) such that for every $n \geq m$, we see that $|1/n - 0| < \epsilon$. Therefore, $\{1/n\}$ converges to 0.

Notice that in Example **??**, we could have resorted to the Archimedian principle and chosen any natural number $m > 1/\epsilon$.

Now that $\{1/n\}$ converges to 0, the sequence whose first 1000 terms are like (n) and 1001st term onward, it is like $(1/n)$ also converges to 0. Because, for any given $\epsilon > 0$, we choose our m as $[1/\epsilon] + 1001$. Moreover, the sequence whose first 1000 terms are like $\{n\}$ and then onwards it is $1, 1/2, 1/3, \ldots$ converges to 0 for the same reason. That is, convergence behavior of a sequence does not change if first finite number of terms are changed.

For a constant sequence $x_n = c$, suppose $\epsilon > 0$ is given. We see that for each $n \in \mathbb{N}$, $|x_n - c| = 0 < \epsilon$. Therefore, the constant sequence $\{c\}$ converges to c.

Sometimes, it is easier to use the condition $|x_n - a| < \epsilon$ as $a - \epsilon < x_n < a + \epsilon$ or as $x \in (a - \epsilon, a + \epsilon)$.

A sequence thus converges to a implies the following:

1. Each neighborhood of a contains a tail of the sequence.
2. Every tail of the sequence contains numbers arbitrarily close to a.

We say that a sequence $\{x_n\}$ **converges** iff it converges to some a. A sequence **diverges** iff it does not converge to any real number.

There are two special cases of divergence.

Let $\{x_n\}$ be a sequence. We say that $\{x_n\}$ **diverges to** ∞ iff for every $r > 0$, there exists an $m \in \mathbb{N}$ such that if $n > m$ is any natural number, then $x_n > r$.

We call an open interval (r, ∞) a neighborhood of ∞. A sequence thus diverges to ∞ implies the following:

1. Each neighborhood of ∞ contains a tail of the sequence.
2. Every tail of the sequence contains arbitrarily large positive numbers.

In this case, we write $\lim_{n\to\infty} x_n = \infty$; we also write it as "$x_n \to \infty$ as $n \to \infty$" or as $x_n \to \infty$.

We say that $\{x_n\}$ **diverges to** $-\infty$ iff for every $r \in \mathbb{R}$, there exists an $m \in \mathbb{N}$ such that if $n > m$ is any natural number, then $x_n < r$.

Calling an open interval $(-\infty, s)$ a neighborhood of $-\infty$, we see that a sequence diverges to $-\infty$ implies the following:

1. Each neighborhood of $-\infty$ contains a tail of the sequence.

2. Every tail of the sequence contains arbitrarily small negative numbers.

In this case, we write $\lim_{n\to\infty} x_n = -\infty$; we also write it as "$x_n \to -\infty$ as $n \to \infty$" or as $x_n \to -\infty$.

We use a unified notation for convergence to a real number and divergence to $\pm\infty$.

For $\ell \in \mathbb{R} \cup \{-\infty, \infty\}$, the notations

$$\lim_{n\to\infty} x_n = \ell, \quad \lim x_n = \ell, \quad x_n \to \ell \text{ as } n \to \infty, \quad x_n \to \ell$$

all stand for the phrase **limit of $\{x_n\}$ is ℓ**. When $\ell \in \mathbb{R}$, the limit of $\{x_n\}$ is ℓ means that $\{x_n\}$ converges to ℓ; and when $\ell = \pm\infty$, the limit of $\{x_n\}$ is ℓ means that $\{x_n\}$ diverges to $\pm\infty$.

Example 1.2. Show that (a) $\lim \sqrt{n} = \infty$; (b) $\lim \ln(1/n) = -\infty$.

(a) Let $r > 0$. Choose an $m > r^2$. Let $n > m$. Then $\sqrt{n} > \sqrt{m} > r$. Therefore, $\lim \sqrt{n} = \infty$.

(b) Let $r > 0$. Choose a natural number $m > e^r$. Let $n > m$. Then $1/n < 1/m < e^{-r}$. Consequently, $\ln(1/n) < \ln e^{-r} = -r$. Therefore, $\ln(1/n) \to -\infty$.

We state a result connecting the limit notion of a function and limit of a sequence. We use the idea of a constant sequence.

Theorem 1.1. (Sandwich Theorem)*: Let the terms of the sequences $\{x_n\}$, $\{y_n\}$, and $\{z_n\}$ satisfy $x_n \le y_n \le z_n$ for all n greater than some m. If $x_n \to \ell$ and $z_n \to \ell$, then $y_n \to \ell$.*

Theorem 1.2. (Limits of sequences to limits of functions)*: Let $a < c < b$. Let $f : D \to \mathbb{R}$ be a function where D contains $(a, c) \cup (c, b)$. Let $\ell \in \mathbb{R}$. Then $\lim_{x\to c} f(x) = \ell$ iff for each non-constant sequence $\{x_n\}$ converging to c, the sequence of functional values $\{f(x_n)\}$ converges to ℓ.*

Theorem 1.3. (Limits of functions to limits of sequences)*: Let $k \in \mathbb{N}$. Let $f(x)$ be a function defined for all $x \ge k$. Let $\{a_n\}$ be a sequence of real numbers such that $a_n = f(n)$ for all $n \ge k$. If $\lim_{x\to\infty} f(x) = \ell$, then $\lim_{n\to\infty} a_n = \ell$.*

For instance, the function $\ln x$ is defined on $[1, \infty)$. Using L' Hospital's rule, we have

$$\lim_{x\to\infty} \frac{\ln x}{x} = \lim_{x\to\infty} \frac{1}{x} = 0.$$

Therefore, $\lim_{n\to\infty} \dfrac{\ln n}{n} = \lim_{x\to\infty} \dfrac{\ln x}{x} = 0.$

1.3 Series

A series is an infinite sum of numbers. We say that the series $\sum x_n$ is convergent iff the sequence $\{s_n\}$ is convergent, where the nth **partial sum** s_n is given by $s_n = \sum_{k=1}^{n} x_n$.

Thus we may define convergence of a series as follows:

We say that the series $\sum x_n$ **converges to** $\ell \in \mathbb{R}$ iff for each $\epsilon > 0$, there exists an $m \in \mathbb{N}$ such that for each $n \geq m$, $|\sum_{k=1}^{n} x_k - \ell| < \epsilon$. In this case, we write $\sum x_n = \ell$.

Further, we say that the series $\sum x_n$ **converges** iff the series converges to some $\ell \in \mathbb{R}$.

The series is said to be **divergent** or to **diverge** iff it is not convergent.

That is, the series $\sum x_n$ **diverges to** ∞ iff for each $r > 0$, there exists $m \in \mathbb{N}$ such that for each $n \geq m$, $\sum_{k=1}^{n} x_k > r$. We write it as $\sum x_n = \infty$.

Similarly, the series $\sum x_n$ **diverges to** $-\infty$ iff for each $r \in \mathbb{R}$, there exists $m \in \mathbb{N}$ such that for each $n \geq m$, $\sum_{k=1}^{n} x_k < r$. We write it as $\sum x_n = -\infty$.

In the unified notation, we say that a series $\sum a_n$ **sums to** $\ell \in \mathbb{R} \cup \{\infty, -\infty\}$, when either the series converges to the real number ℓ or it diverges to $\pm\infty$. In all these cases we write $\sum a_n = \ell$.

There can be series which diverge but neither to ∞ nor to $-\infty$. For example, the series

$$\sum_{n=0}^{\infty} (-1)^n = 1 - 1 + 1 - 1 + 1 - 1 + \cdots$$

neither diverges to ∞ nor to $-\infty$. But it is a divergent series. Can you see why?

Example 1.3.

(a) The series $\displaystyle\sum_{n=1}^{\infty} \frac{1}{2^n}$ sums to 1. Because, if $\{s_n\}$ is the sequence of partial sums, then

$$s_n = \sum_{k=1}^{n} \frac{1}{2^k} = \frac{1}{2} \cdot \frac{1 - (1/2)^n}{1 - 1/2} = 1 - \frac{1}{2^n} \to 1.$$

(b) The series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$ sums to ∞. To see this, let $s_n = \sum_{k=1}^{n} \frac{1}{k}$ be the partial sum up to n terms. Let m be the natural number such that $2^m \leq n < 2^{m+1}$. Then

$$\begin{aligned}
s_n &= \sum_{k=1}^{n} \frac{1}{k} \geq 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{2^m - 1} \\
&= 1 + \Big(\frac{1}{2} + \frac{1}{3}\Big) + \Big(\frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7}\Big) + \cdots + \Big(\sum_{k=2^{m-1}}^{2^m - 1} \frac{1}{k}\Big) \\
&> 1 + \Big(\frac{1}{4} + \frac{1}{4}\Big) + \Big(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\Big) + \cdots + \Big(\sum_{k=2^{m-1}}^{2^m - 1} \frac{1}{2^m}\Big) \\
&= 1 + \frac{1}{2} + \frac{1}{2} + \cdots + \frac{1}{2} = 1 + \frac{m-1}{2}.
\end{aligned}$$

As $n \to \infty$, we see that $m \to \infty$. Consequently, $s_n \to \infty$. That is, the series diverges to (sums to) ∞. This is called the **harmonic series**.

(c) The series $-1 - 2 - 3 - 4 - \cdots - n - \cdots$ sums to $-\infty$.

(d) The series $1 - 1 + 1 - 1 + \cdots$ diverges. It neither diverges to ∞ nor to $-\infty$. Because, the sequence of partial sums here is $1, 0, 1, 0, 1, 0, 1, \ldots$ which does not converge.

Example 1.4. Let $a \neq 0$. Consider the **geometric series**

$$\sum_{n=1}^{\infty} ar^{n-1} = a + ar + ar^2 + ar^3 + \cdots .$$

The nth partial sum of the geometric series is

$$s_n = a + ar + ar^2 + ar^3 + \cdots ar^{n-1} = \frac{a(1-r^n)}{1-r}.$$

(a) If $|r| < 1$, then $r^n \to 0$. The geometric series sums to $\lim_{n\to\infty} s_n = \dfrac{a}{1-r}$.

Therefore, $\displaystyle\sum_{n=0}^{\infty} ar^n = \sum_{n=1}^{\infty} ar^{n-1} = \frac{a}{1-r}$.

(b) If $|r| \geq 1$, then r^n diverges. The geometric series $\sum ar^{n-1}$ diverges.

1.4 Some results on convergence

Theorem 1.4. *If a series $\sum a_n$ sums to $\ell \in \mathbb{R} \cup \{\infty, -\infty\}$, then ℓ is unique.*

Proof. Suppose the series $\sum a_n$ sums to ℓ and also to s, where both $\ell, s \in \mathbb{R} \cup \{\infty, -\infty\}$. Suppose that $\ell \neq s$. We consider the following exhaustive cases.

Case 1: $\ell, s \in \mathbb{R}$. Then $|s - \ell| > 0$. Choose $\epsilon = |s-\ell|/2$. We have natural numbers k and m such that for every $n \geq k$ and $n \geq m$,

$$\Big|\sum_{j=1}^{n} a_j - \ell\Big| < \epsilon \quad \text{and} \quad \Big|\sum_{j=1}^{n} a_j - s\Big| < \epsilon.$$

Fix one such n, say $M > \max\{k, m\}$. Both the above inequalities hold for $n = M$. Then

$$|s-\ell| = \Big|s - \sum_{j=1}^{M} a_j + \sum_{j=1}^{M} a_j - \ell\Big| \leq \Big|\sum_{j=1}^{M} a_j - s\Big| + \Big|\sum_{j=1}^{M} a_j - \ell\Big| < 2\epsilon = |s-\ell|.$$

This is a contradiction.

Case 2: $\ell \in \mathbb{R}$ and $s = \infty$. Then there exists a natural number k such that for every $n \geq k$, we have

$$\Big|\sum_{j=1}^{n} a_j - \ell\Big| < 1.$$

Since the series sums to ∞, we have $m \in \mathbb{N}$ such that for every $n \geq m$,

$$\sum_{j=1}^{n} a_j > \ell + 1.$$

Now, fix an $M > \max\{k, m\}$. Then both of the above hold for this $n = M$. So,

$$\sum_{j=1}^{M} a_j < \ell + 1 \quad \text{and} \quad \sum_{j=1}^{M} a_j > \ell + 1.$$

This is a contradiction.

Case 3: $\ell \in \mathbb{R}$ and $s = -\infty$. It is similar to Case 2. Choose "less than $\ell - 1$".

Case 4: $\ell = \infty,\ s = -\infty$. Again choose an M so that $\sum_{j=1}^{n} a_n$ is both greater than 1 and also less than -1 leading to a contradiction. □

Theorem 1.5. (1) (Cauchy Criterion) *A series $\sum a_n$ converges iff for each $\epsilon > 0$, there exists a $k \in \mathbb{N}$ such that $|\sum_{j=m}^{n} a_j| < \epsilon$ for all $n \geq m \geq k$.*

(2) (Weirstrass Criterion) *Let $\sum a_n$ be a series of non-negative terms. Suppose there exists $c \in \mathbb{R}$ such that each partial sum of the sries is less than c, i.e., for each n, $\sum_{j=1}^{n} a_j < c$. Then $\sum a_n$ is convergent.*

Theorem 1.6. *If a series $\sum a_n$ converges, then the sequence $\{a_n\}$ converges to 0.*

Proof: Let s_n denote the partial sum $\sum_{k=1}^{n} a_k$. Then $a_n = s_n - s_{n-1}$. If the series converges, say, to ℓ, then $\lim s_n = \ell = \lim s_{n-1}$. It follows that $\lim a_n = 0$. □

It says that if $\lim a_n$ does not exist, or if $\lim a_n$ exists but is not equal to 0, then the series $\sum a_n$ diverges.

The series $\displaystyle\sum_{n=1}^{\infty} \frac{-n}{3n+1}$ diverges because $\displaystyle\lim_{n\to\infty} \frac{-n}{3n+1} = -\frac{1}{3} \neq 0$.

The series $\sum(-1)^n$ diverges because $\lim(-1)^n$ does not exist.

Notice what Theorem **??** does not say. The harmonic series diverges even though $\lim \frac{1}{n} = 0$. Recall that for a real number ℓ our notation says that $\ell + \infty = \infty$ and $\ell - \infty = -\infty$. Further, if $\ell > 0$, then $\ell \cdot (\infty) = \infty$ and $\ell \cdot (-\infty) = -\infty$; and if $\ell < 0$, then $\ell \cdot (\infty) = -\infty$ and $\ell \cdot (-\infty) = \infty$. Similarly, we accept the convention that $\infty + \infty = \infty$ and $-\infty - \infty = -\infty$.

Theorem 1.7. (1) *If $\sum a_n$ sums to a and $\sum b_n$ sums to b, then the series $\sum(a_n + b_n)$ sums to $a + b$; $\sum(a_n - b_n)$ sums to $a - b$; and $\sum kb_n$ sums to kb; where k is any real number.*

(2) *If $\sum a_n$ converges and $\sum b_n$ diverges, then $\sum(a_n + b_n)$ and $\sum(a_n - b_n)$ diverge.*

(3) *If $\sum a_n$ diverges and $k \neq 0$, then $\sum ka_n$ diverges.*

Notice that sum of two divergent series can converge. For example, both $\sum(1/n)$ and $\sum(-1/n)$ diverge but their sum $\sum 0$ converges.

Since deleting a finite number of terms of a sequence does not alter its convergence, omitting a finite number of terms or adding a finite number of terms to a convergent (divergent) series implies the convergence (divergence) of the new series. Of course, the sum of the convergent series will be affected. For example,

$$\sum_{n=3}^{\infty} \left(\frac{1}{2^n}\right) = \sum_{n=1}^{\infty} \left(\frac{1}{2^n}\right) - \frac{1}{2} - \frac{1}{4}.$$

However,

$$\sum_{n=3}^{\infty}\Big(\frac{1}{2^{n-2}}\Big)=\sum_{n=1}^{\infty}\Big(\frac{1}{2^{n}}\Big).$$

This is called **re-indexing** the series. As long as we preserve the order of the terms of the series, we can re-index without affecting its convergence and sum.

1.5 Comparison tests

Theorem 1.8. (Comparison Test) *Let $\sum a_n$ and $\sum b_n$ be series of non-negative terms. Suppose there exists $k>0$ such that $0\le a_n\le kb_n$ for each n greater than some natural number m.*

1. *If $\sum b_n$ converges, then $\sum a_n$ converges.*
2. *If $\sum a_n$ diverges to ∞, then $\sum b_n$ diverges to ∞.*

Proof: (1) Consider all partial sums of the series having more than m terms. We see that

$$a_1+\cdots+a_m+a_{m+1}+\cdots+a_n\le a_1+\cdots+a_m+k\sum_{j=m+1}^{n}b_j.$$

Since $\sum b_n$ converges, so does $\sum_{j=m+1}^{n}b_j$. By Weirstrass criterion, $\sum a_n$ converges.

(2) Similar to (1). □

Caution: The comparison test holds for series of non-negative terms.

Theorem 1.9. (Ratio Comparison Test) *Let $\sum a_n$ and $\sum b_n$ be series of non-negative terms. Suppose there exists $m\in\mathbb{N}$ such that for each $n>m$, $a_n>0$, $b_n>0$, and $\dfrac{a_{n+1}}{a_n}\le\dfrac{b_{n+1}}{b_n}$.*

1. *If $\sum b_n$ converges, then $\sum a_n$ converges.*
2. *If $\sum a_n$ diverges to ∞, then $\sum b_n$ diverges to ∞.*

Proof: For $n>m$,

$$a_n=\frac{a_n}{a_{n-1}}\,\frac{a_{n-1}}{a_{n-2}}\,\cdots\frac{a_{m+1}}{a_m}\,a_m\le\frac{b_n}{b_{n-1}}\,\frac{b_{n-1}}{b_{n-2}}\,\cdots\frac{b_{m+1}}{b_m}\,b_m=\frac{a_m}{b_m}\,b_n.$$

By the Comparison test, if $\sum b_n$ converges, then $\sum a_n$ converges. This proves (1). And, (2) follows from (1) by contradiction. □

Theorem 1.10. (Limit Comparison Test) *Let $\sum a_n$ and $\sum b_n$ be series of non-negative terms. Suppose that there exists $m\in\mathbb{N}$ such that for each $n>m$, $a_n>0$, $b_n>0$, and $\lim\limits_{n\to\infty}\dfrac{a_n}{b_n}=k$.*

1. *If* $k > 0$ *then* $\sum b_n$ *and* $\sum a_n$ *either both converge, or both diverge to* ∞.

2. *If* $k = 0$ *and* $\sum b_n$ *converges, then* $\sum a_n$ *converges.*

3. *If* $k = \infty$ *and* $\sum b_n$ *diverges to* ∞ *then* $\sum a_n$ *diverges to* ∞.

Proof: (1) Let $\epsilon = k/2 > 0$. The limit condition implies that there exists $m \in \mathbb{N}$ such that

$$\frac{k}{2} < \frac{a_n}{b_n} < \frac{3k}{2} \quad \text{for each} \quad n > m.$$

By the Comparison test, the conclusion is obtained.

(2) Let $\epsilon = 1$. The limit condition implies that there exists $m \in \mathbb{N}$ such that

$$-1 < \frac{a_n}{b_n} < 1 \quad \text{for each} \quad n > m.$$

Using the right hand inequality and the Comparison test we conclude that convergence of $\sum b_n$ implies the convergence of $\sum a_n$.

(3) If $k > 0$, $\lim(b_n/a_n) = 1/k$. Use (1). If $k = \infty$, $\lim(b_n/a_n) = 0$. Use (2). □

Example 1.5. For each $n \in \mathbb{N}$, $n! \geq 2^{n-1}$. That is, $\dfrac{1}{n!} \leq \dfrac{1}{2^{n-1}}$.

Since $\displaystyle\sum_{n=1}^{\infty} \frac{1}{2^{n-1}}$ is convergent, $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n!}$ is convergent. Therefore, adding 1 to it, the series

$$1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} + \cdots$$

is convergent. In fact, this series converges to e. To see this, consider

$$s_n = 1 + 1 + \frac{1}{2!} + \cdots + \frac{1}{n!}, \quad t_n = \Big(1 + \frac{1}{n}\Big)^n.$$

By the Binomial theorem,

$$t_n = 1 + 1 + \frac{1}{2!}\Big(1 - \frac{1}{n}\Big) + \cdots + \frac{1}{n!}\Big[\big(1 - \frac{1}{n}\big)\big(1 - \frac{2}{n}\big)\cdots\big(1 - \frac{n-1}{n}\big)\Big] \leq s_n.$$

Thus taking limit as $n \to \infty$, we have

$$e = \lim_{n\to\infty} t_n \leq \lim_{n\to\infty} s_n.$$

Also, for $n > m$, where m is any fixed natural number,

$$t_n \geq 1 + 1 + \frac{1}{2!}\Big(1 - \frac{1}{n}\Big) + \cdots + \frac{1}{m!}\Big[(1 - 1/n)(1 - 2/n)\cdots\big(1 - \frac{m-1}{n}\big)\Big]$$

Taking limit as $n \to \infty$ we have

$$e = \lim_{n\to\infty} t_n \geq s_m.$$

Since m is arbitrary, taking the limit as $m \to \infty$, we have

$$e \geq \lim_{m\to\infty} s_m.$$

Therefore, $\displaystyle\lim_{m\to\infty} s_m = e$. That is, the series $\displaystyle\sum_{n=0}^{\infty} \frac{1}{n!} = e$.

Example 1.6. Determine when the series $\displaystyle\sum_{n=1}^{\infty} \frac{n+7}{n(n+3)\sqrt{n+5}}$ converges.

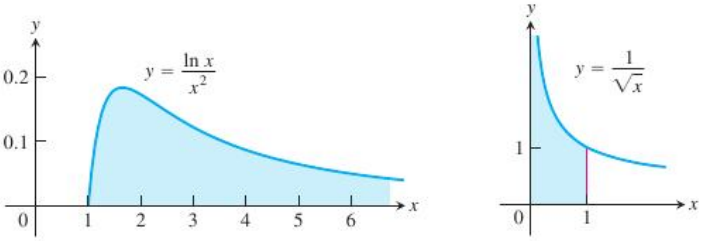
Let $a_n = \dfrac{n+7}{n(n+3)\sqrt{n+5}}$ and $b_n = \dfrac{1}{n^{3/2}}$. Then

$$\frac{a_n}{b_n} = \frac{\sqrt{n}(n+7)}{(n+3)\sqrt{n+5}} \to 1 \text{ as } n \to \infty.$$

Since $\dfrac{1}{n^{3/2}}$ is convergent, Limit comparison test says that the given series is convergent.

1.6 Improper integrals

In the definite integral $\int_a^b f(x)dx$ we required that both a, b are finite and also the range of $f(x)$ is a subset of some finite interval. However, there are functions which violate one or both of these requirements, and yet, the area under the curves and above the x-axis remain bounded.



Such integrals are called **Improper Integrals**. Suppose $f(x)$ is continuous on $[0, \infty)$. It makes sense to write

$$\int_0^{\infty} f(x)dx = \lim_{b\to\infty} \int_0^b f(x)dx$$

provided that the limit exists. In such a case, we say that the improper integral $\int_0^{\infty} f(x)dx$ **converges** and its *value* is given by the limit. We say that the improper integral **diverges** iff it is not convergent.

Here are the possible types of improper integrals.

1. If $f(x)$ is continuous on $[a, \infty)$, then $\displaystyle\int_a^{\infty} f(x)\,dx = \lim_{b\to\infty} \int_a^b f(x)\,dx.$

2. If $f(x)$ is continuous on $(-\infty, b]$, then $\displaystyle\int_{-\infty}^{b} f(x)\,dx = \lim_{a\to-\infty} \int_a^b f(x)\,dx.$

3. If $f(x)$ is continuous on $(-\infty, \infty)$, then
$\displaystyle\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{c} f(x)\,dx + \int_c^{\infty} f(x)\,dx$, for any $c \in \mathbb{R}$.

4. If $f(x)$ is continuous on $(a, b]$ and discontinuous at $x = a$, then
$\displaystyle\int_a^b f(x)\,dx = \lim_{t\to a+} \int_t^b f(x)\,dx.$

5. If $f(x)$ is continuous on $[a, b)$ and discontinuous at $x = b$, then
$$\int_a^b f(x)\,dx = \lim_{t\to b-} \int_a^t f(x)\,dx.$$
6. If $f(x)$ is continuous on $[a, c) \cup (c, b]$ and discontinuous at $x = c$, then
$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

In each case, if the right hand side (along with the limits of the concerned integrals) is finite, then we say that the improper integral on the left **converges**, else, the improper integral **diverges**; the finite value as obtained from the right hand side is the *value* of the improper integral. A convergent improper integral converges to its value.

Two important sub-cases of divergent improper integrals are when the limit of the concerned integral is ∞ or $-\infty$. In these cases, we say that the improper integral **diverges to ∞ or to $-\infty$** as is the case.

Example 1.7. For what values of $p \in \mathbb{R}$, the improper integral $\displaystyle\int_1^\infty \frac{dx}{x^p}$ converges? What is its value, when it converges?

Case 1: $p = 1$.
$$\int_1^b \frac{dx}{x^p} = \int_1^b \frac{dx}{x} = \ln b - \ln 1 = \ln b.$$
Since $\lim\limits_{b\to\infty} \ln b = \infty$, the improper integral diverges to ∞.

Case 2: $p < 1$.
$$\int_1^b \frac{dx}{x^p} = \frac{-x^{-p+1}}{-p+1}\Big|_1^b = \frac{1}{1-p}(b^{1-p} - 1).$$
Since $\lim\limits_{b\to\infty} b^{1-p} = \infty$, the improper integral diverges to ∞.

Case 3: $p > 1$.
$$\int_1^b \frac{dx}{x^p} = \frac{1}{1-p}(b^{1-p} - 1) = \frac{1}{1-p}\Big(\frac{1}{b^{p-1}} - 1\Big).$$
Since $\lim\limits_{b\to\infty} \dfrac{1}{b^{p-1}} = 0$, we have
$$\int_1^\infty \frac{dx}{x^p} = \lim_{b\to\infty} \int_1^b \frac{dx}{x^p} = \lim_{b\to\infty} \frac{1}{1-p}\Big(\frac{1}{b^{p-1}} - 1\Big) = \frac{1}{p-1}.$$
Hence, the improper integral $\displaystyle\int_1^\infty \frac{dx}{x^p}$ converges to $\dfrac{1}{p-1}$ for $p > 1$ and diverges to ∞ for $p \le 1$.

Example 1.8. For what values of $p \in \mathbb{R}$, the improper integral $\displaystyle\int_0^1 \frac{dx}{x^p}$ converges?

Case 1: $p = 1$.

$$\int_0^1 \frac{dx}{x^p} = \lim_{a\to 0+} \int_a^1 \frac{dx}{x} = \lim_{a\to 0+} [\ln 1 - \ln a] = \infty.$$

Therefore, the improper integral diverges to ∞.

Case 2: $p < 1$.

$$\int_0^1 \frac{dx}{x^p} = \lim_{a\to 0+} \int_a^1 \frac{dx}{x^p} = \lim_{a\to 0+} \frac{1 - a^{1-p}}{1-p} = \frac{1}{1-p}.$$

Therefore, the improper integral converges to $1/(1-p)$.

Case 3: $p > 1$.

$$\int_0^1 \frac{dx}{x^p} = \lim_{a\to 0+} \frac{1 - a^{1-p}}{1-p} = \lim_{a\to 0+} \frac{1}{p-1}\Big(\frac{1}{a^{p-1}} - 1\Big) = \infty.$$

Hence the improper integral diverges to ∞.

The improper integral $\displaystyle\int_0^1 \frac{dx}{x^p}$ converges to $\displaystyle\frac{1}{1-p}$ for $p < 1$ and diverges to ∞ for $p \geq 1$.

1.7 Convergence tests for improper integrals

Theorem 1.11. (Comparison Test) *Let $f(x)$ and $g(x)$ be real valued continuous functions on $[a, \infty)$. Suppose that $0 \leq f(x) \leq g(x)$ for all $x \geq a$.*

1. *If $\int_a^\infty g(x)\,dx$ converges, then $\int_a^\infty f(x)\,dx$ converges.*
2. *If $\int_a^\infty f(x)\,dx$ diverges to ∞, then $\int_a^\infty g(x)\,dx$ diverges to ∞.*

Proof: Since $0 \leq f(x) \leq g(x)$ for all $x \geq a$,

$$\int_a^b f(x)dx \leq \int_a^b g(x)dx.$$

As $\lim_{b\to\infty} \int_a^b g(x)dx = \ell$ for some $\ell \in \mathbb{R}$, $\lim_{b\to\infty} \int_a^b f(x)dx$ exists and the limit is less than or equal to ℓ. This proves (1). Proof of (2) is similar to that of (1). □

Theorem 1.12. (Limit Comparison Test) *Let $f(x)$ and $g(x)$ be continuous functions on $[a, \infty)$ with $f(x) \geq 0$ and $g(x) \geq 0$. If $\lim_{x\to\infty} \frac{f(x)}{g(x)} = L$, where $0 < L < \infty$, then $\int_a^\infty f(x)dx$ and $\int_a^\infty g(x)dx$ either both converge, or both diverge.*

Theorem 1.13. *Let $f(x)$ be a real valued continuous function on $[a, b)$, for $b \in \mathbb{R} \cup \{\infty\}$. If the improper integral $\int_a^b |f(x)|\,dx$ converges, then the improper integral $\int_a^b f(x)\,dx$ also converges.*

Example 1.9.

(a) $\displaystyle\int_1^\infty \frac{\sin^2 x}{x^2}\,dx$ converges because $\displaystyle\frac{\sin^2 x}{x^2} \le \frac{1}{x^2}$ for all $x \ge 1$ and $\displaystyle\int_1^\infty \frac{dx}{x^2}$ converges.

(b) $\displaystyle\int_2^\infty \frac{dx}{\sqrt{x^2-1}}$ diverges to ∞ because (Recall: $\displaystyle\lim_{x\to\infty} \ln x = \infty$.)

$\displaystyle\frac{1}{\sqrt{x^2-1}} \ge \frac{1}{x}$ for all $x \ge 2$ and $\displaystyle\int_2^\infty \frac{dx}{x}$ diverges to ∞.

(c) $\displaystyle\int_1^\infty \frac{dx}{1+x^2}$ converges or diverges?

Since $\displaystyle\lim_{x\to\infty}\left[\frac{1}{1+x^2}\Big/\frac{1}{x^2}\right] = \lim_{x\to\infty}\frac{x^2}{1+x^2} = 1,$ the limit comparison test says that the given improper integral and $\displaystyle\int_1^\infty \frac{dx}{x^2}$ both converge or diverge together. The latter converges, so does the former. However, they may converge to different values.

$$\int_1^\infty \frac{dx}{1+x^2} = \lim_{b\to\infty}[\tan^{-1} b - \tan^{-1} 1] = \frac{\pi}{2} - \frac{\pi}{4} = \frac{\pi}{4}.$$

$$\int_1^\infty \frac{dx}{x^2} = \lim_{b\to\infty}\left(\frac{-1}{b} - \frac{-1}{1}\right) = 1.$$

(d) Does the improper integral $\displaystyle\int_1^\infty \frac{10^{10}dx}{e^x+1}$ converge?

$$\lim_{x\to\infty}\frac{10^{10}}{e^x+1}\Big/\frac{1}{e^x} = \lim_{x\to\infty}\frac{10^{10}e^x}{e^x+1} = 10^{10}.$$

Also, $e \ge 2$ implies that for all $x \ge 1,\ e^x \ge x^2$. So, $e^{-x} \le x^{-2}$. Since $\displaystyle\int_1^\infty \frac{dx}{x^2}$ converges, $\displaystyle\int_1^\infty \frac{dx}{e^x}$ also converges. By limit comparison test, the given improper integral converges.

Example 1.10. Show that $\Gamma(x) = \displaystyle\int_0^\infty e^{-t}t^{x-1}\,dt$ converges for each $x > 0$.

Fix $x > 0$. Since $\displaystyle\lim_{t\to\infty} e^{-t}t^{x+1} = 0$, there exists $t_0 \ge 1$ such that $0 < e^{-t}t^{x+1} < 1$ for $t > t_0$. That is,

$$0 < e^{-t}t^{x-1} < t^{-2} \quad \text{for} \quad t > t_0.$$

Since $\int_1^\infty t^{-2}\,dt$ is convergent, $\int_{t_0}^\infty t^{-2}\,dt$ is also convergent. By the comparison test,

$$\int_{t_0}^\infty e^{-t}t^{x-1}\,dt \text{ is convergent.}$$

The integral $\int_1^{t_0} e^{-t}t^{x-1}\,dt$ exists and is not an improper integral.

Next, we consider the improper integral $\int_0^1 e^{-t}t^{x-1}\,dt$. Let $0 < a < 1$.

For $a \le t \le 1$, we have $0 < e^{-t}t^{x-1} < t^{x-1}$. So,

$$\int_a^1 e^{-t}t^{x-1}\,dt < \int_a^1 t^{x-1}\,dt = \frac{1-a^x}{x} < \frac{1}{x}.$$

Taking the limit as $a \to 0+$, we see that the

$$\int_0^1 e^{-t}t^{x-1}\,dt \quad \text{is convergent,}$$

and its value is less than or equal to $1/x$. Therefore,

$$\int_0^\infty e^{-t}t^{x-1}\,dt = \int_0^1 e^{-t}t^{x-1}\,dt + \int_1^{t_0} e^{-t}t^{x-1}\,dt + \int_{t_0}^\infty e^{-t}t^{x-1}\,dt$$

is convergent.

The function $\Gamma(x)$ is defined on $(0,\infty)$. For $x > 0$, using integration by parts,

$$\Gamma(x+1) = \int_0^\infty t^x e^{-t}\,dt = \Big[-t^x - e^{-t}\Big]_0^\infty - \int_0^\infty xt^{x-1}(-e^{-t})\,dt = x\Gamma(x).$$

It thus follows that $\Gamma(n+1) = n!$ for any non-negative integer n. We take $0! = 1$.

The Gamma function takes other forms by substitution of the variable of integration. Substituting t by rt we have

$$\Gamma(x) = r^x \int_0^\infty e^{-rt}t^{x-1}\,dt \quad \text{for } \ 0 < r,\ 0 < x.$$

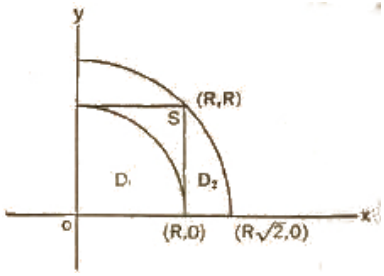
Substituting t by t^2, we have

$$\Gamma(x) = 2\int_0^\infty e^{-t^2}t^{2x-1}\,dt \quad \text{for } \ 0 < x.$$

Example 1.11. Show that $\Gamma(1/2) = 2\int_0^\infty e^{-t^2}\,dt = \sqrt{\pi}$.

$$\Gamma\Big(\frac{1}{2}\Big) = \int_0^\infty e^{-x}x^-1/2\,dx = 2\int_0^\infty e^{-t^2}\,dt \qquad (x = t^2)$$

To evaluate this integral, consider the double integral of $e^{-x^2-y^2}$ over two circular sectors D_1 and D_2, and the square S as indicated below.



Since the integrand is positive, we have $\iint_{D_1} < \iint_S < \iint_{D_2}$.
Now, evaluate these integrals by converting them to iterated integrals as follows:

$$\int e^{-r^2}r\,dr \int_0^{\pi/2} d\theta < \int_0^R e^{-x^2}\,dx \int_0^R e^{-y^2}\,dy < \int_0^{R\sqrt{2}} e^{-r^2}r\,dr \int_0^{\pi/2} d\theta$$

$$\frac{\pi}{4}(1 - e^{-R^2}) < \Big(\int_0^R e^{-x^2}\,dx\Big)^2 < \frac{\pi}{4}(1 - e^{-2R^2})$$

Take the limit as $R \to \infty$ to obtain

$$\Big(\int_0^\infty e^{-x^2}\, dx \Big)^2 = \frac{\pi}{4}$$

From this, the result follows.

Example 1.12. Test the convergence of $\int_{-\infty}^{\infty} e^{-t^2}\, dt$.

Since e^{-t^2} is continuous on $[-1, 1]$, $\int_{-1}^{1} e^{-t^2}\, dt$ exists.

For $t > 1$, we have $t < t^2$. So, $0 < e^{-t^2} < e^{-t}$. Since $\int_1^\infty e^{-t}\, dt$ is convergent, by the comparison test, $\int_1^\infty e^{-t^2}\, dt$ is convergent.

Now, $\int_{-a}^{-1} e^{-t^2}\, dt = \int_a^1 e^{-t^2} d(-t) = \int_1^a e^{-t^2} dt$. Taking limit as $a \to \infty$, we see that $\int_{-\infty}^{1} e^{-t^2}\, dt$ is convergent and its value is equal to $\int_1^\infty e^{-t^2}\, dt$.

Combining the three integrals above, we conclude that $\int_{-\infty}^{\infty} e^{-t^2}\, dt$ is convergent.

Example 1.13. Prove: $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, dt$ converges for $x > 0, y > 0$.

We write the integral as a sum of two integrals:

$$B(x, y) = \int_0^{1/2} t^{x-1}(1-t)^{y-1}\, dt + \int_{1/2}^1 t^{x-1}(1-t)^{y-1}\, dt$$

Setting $u = 1 - t$, the second integral looks like

$$\int_{1/2}^1 t^{x-1}(1-t)^{y-1}\, dt = \int_0^{1/2} u^{y-1}(1-u)^{x-1}\, dt$$

Therefore, it is enough to show that the first integral converges. Notice that here, $0 < t \leq 1/2$.

Case 1: $x \geq 1$.

For $0 < t < 1/2$, $1 - t > 0$. Therefore, for all $y > 0$, the function $(1-t)^{y-1}$ is well defined, continuous, and bounded on $(0, 1/2]$. So is the function t^{x-1}. Therefore, the integral $\int_0^{1/2} t^{x-1}(1-t)^{y-1}\, dt$ exists and is not an improper integral.

Case 2: $0 < x < 1$.

Here, the function t^{x-1} is well defined and continuous on $(0, 1/2]$. By Example **??**, the integral $\int_0^{1/2} t^{x-1}\, dt$ converges. for $0 < t \leq 1/2$, $t^{x-1}(1-t)^{y-1} \leq t^{x-1}$. So, $\int_0^{1/2} t^{x-1}(1-t)^{y-1}\, dt$ converges.

By setting t as $1 - t$, we see that $B(x, y) = B(y, x)$.

Like the Gamma function, the Beta function takes various forms.

By substituting t with $\sin^2 t$, the Beta function can be written as

$$B(x, y) = 2 \int_0^{\pi/2} (\sin t)^{2x-1} (\cos t)^{2y-1}\, dt, \quad \text{for } \ x > 0,\ y > 0.$$

Changing the variable t to $t/(1+t)$, the Beta function can be written as

$$B(x,y) = \int_0^\infty \frac{t^{x+1}}{(1+t)^{x+y}}\, dt \quad \text{for} \;\; x>0,\; y>0.$$

Again, using multiple integrals it can be shown that

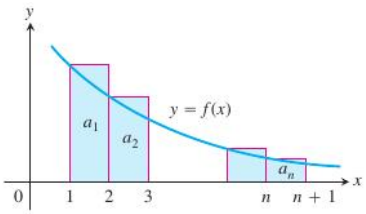
$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \quad \text{for} \;\; x>0,\; y>0.$$

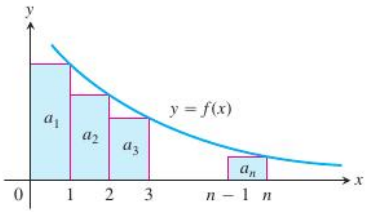
1.8 Tests of convergence for series

Theorem 1.14. (Integral Test) *Let $\sum a_n$ be a series of positive terms. Let $f:[1,\infty)\to\mathbb{R}$ be a continuous, positive and non-increasing function such that $a_n = f(n)$ for each $n\in\mathbb{N}$.*

1. *If $\int_1^\infty f(t)dt$ is convergent, then $\sum a_n$ is convergent.*
2. *If $\int_1^\infty f(t)dt$ diverges to ∞, then $\sum a_n$ diverges to ∞.*

Proof: Since f is a positive and non-increasing, the integrals and the partial sums have a certain relation.





$$\int_1^{n+1} f(t)\,dt \;\le\; a_1 + a_2 + \cdots + a_n \;\le\; a_1 + \int_1^n f(t)\,dt.$$

If $\int_1^n f(t)\,dt$ is finite, then the right hand inequality shows that $\sum a_n$ is convergent.
If $\int_1^n f(t)\,dt = \infty$, then the left hand inequality shows that $\sum a_n$ diverges to ∞. □

Notice that when the series converges, the value of the integral can be different from the sum of the series. Moreover, Integral test assumes implicitly that $\{a_n\}$ is a monotonically decreasing sequence. Further, the integral test is also applicable when the interval of integration is $[m,\infty)$ instead of $[1,\infty)$.

Example 1.14. Show that $\displaystyle\sum_{n=1}^{\infty}\frac{1}{n^p}$ converges for $p>1$ and diverges for $p\le 1$.

For $p=1$, the series is the harmonic series; and it diverges. Suppose $p\neq 1$. Consider the function $f(t)=1/t^p$ from $[1,\infty)$ to $\mathbb{R}$. This is a continuous, positive and decreasing function.

$$\int_1^\infty \frac{1}{t^p}\,dt = \lim_{b\to\infty}\frac{t^{-p+1}}{-p+1}\Big|_1^b = \frac{1}{1-p}\lim_{b\to\infty}\left(\frac{1}{b^{p-1}}-1\right) = \begin{cases} \frac{1}{p-1} & \text{if } p>1 \\ \infty & \text{if } p<1. \end{cases}$$

Then the Integral test proves the statement. We note that for $p>1$, the sum of the series $\sum n^{-p}$ need not be equal to $(p-1)^{-1}$.

Theorem 1.15. (D' Alembert Ratio Test)
Let $\sum a_n$ be a series of positive terms. Suppose $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = \ell$.

1. *If $\ell < 1$, then $\sum a_n$ converges.*
2. *If $\ell > 1$ or $\ell = \infty$, then $\sum a_n$ diverges to ∞.*
3. *If $\ell = 1$, then no conclusion is obtained.*

Proof: (1) Given that $\lim(a_{n+1}/a_n) = \ell < 1$. Choose δ such that $\ell < \delta < 1$. There exists $m \in \mathbb{N}$ such that for each $n > m$, $a_{n+1}/a_n < \delta$. Then

$$\frac{a_n}{a_{m+1}} = \frac{a_n}{a_{n-1}} \frac{a_{n-1}}{a_{n-2}} \cdots \frac{a_{m+2}}{a_{m+1}} < \delta^{n-m}.$$

Thus, $a_n < \delta^{n-m} a_{m+1}$. Consequently,

$$a_{m+1} + a_{m+2} + \cdots + a_n < a_{m+1}(1 + \delta + \delta^2 + \cdots \delta^{n-m}).$$

Since $\delta < 1$, this approaches a limit as $n \to \infty$. Therefore, the series

$$a_{m+1} + a_{m+2} + \cdots a_n + \cdots$$

converges. In that case, the series $\sum a_n = (a_1 + \cdots + a_m) + a_{m+1} + a_{m+2} + \cdots$ converges.

(2) Given that $\lim(a_{n+1}/a_n) = \ell > 1$. Then there exists $m \in \mathbb{N}$ such that for each $n > m$, $a_{n+1} > a_n$. Then

$$a_{m+1} + a_{m+2} + \cdots + a_n > a_{m+1}(n - m).$$

Since $a_{m+1} > 0$, this approaches ∞ as $n \to \infty$. Therefore, the series

$$a_{m+1} + a_{m+2} + \cdots a_n + \cdots$$

diverges to ∞. In that case, the series $\sum a_n = (a_1 + \cdots + a_m) + a_{m+1} + a_{m+2} + \cdots$ diverges to ∞. The other case of $\ell = \infty$ is similar.

(3) The series $\sum(1/n)$ diverges to ∞. But $\lim(a_{n+1}/a_n) = \lim(n/(n+1)) = 1$.
But the series $\sum(1/n^2)$ is convergent although $\lim(a_{n+1}/a_n) = 1$. □

Example 1.15. Does the series $\displaystyle\sum_{n=1}^{\infty} \frac{n!}{n^n}$ converge?

Write $a_n = n!/(n^n)$. Then

$$\frac{a_{n+1}}{a_n} = \frac{(n+1)!n^n}{(n+1)^{n+1}(n!)} = \Big(\frac{n}{n+1}\Big)^n \to \frac{1}{e} < 1 \text{ as } n \to \infty.$$

By D' Alembert's ratio test, the series converges.

Remark: It follows that the sequence $\{n!/n^n\}$ converges to 0.

Theorem 1.16. (Cauchy Root Test)
Let $\sum a_n$ be a series of positive terms. Suppose $\lim_{n\to\infty}(a_n)^{1/n} = \ell$.

1. *If $\ell < 1$, then $\sum a_n$ converges.*
2. *If $\ell > 1$ or $\ell = \infty$, then $\sum a_n$ diverges to ∞.*
3. *If $\ell = 1$, then no conclusion is obtained.*

Proof: (1) Suppose $\ell < 1$. Choose δ such that $\ell < \delta < 1$. Due to the limit condition, there exists an $m \in \mathbb{N}$ such that for each $n > m$, $(a_n)^{1/n} < \delta$. That is, $a_n < \delta^n$. Since $0 < \delta < 1$, $\sum \delta^n$ converges. By Comparison test, $\sum a_n$ converges.
(2) Given that $\ell > 1$ or $\ell = \infty$, we see that $(a_n)^{1/n} > 1$ for infinitely many values of n. That is, the sequence $\{(a_n)^{1/n}\}$ does not converge to 0. Therefore, $\sum a_n$ is divergent. It diverges to ∞ since it is a series of positive terms.
(3) Once again, for both the series $\sum(1/n)$ and $\sum(1/n^2)$, we see that $(a_n)^{1/n}$ has the limit 1. But one is divergent, the other is convergent. □

Remark: In fact, for a sequence $\{a_n\}$ of positive terms if $\lim_{n\to\infty}\dfrac{a_{n+1}}{a_n}$ exists, then $\lim_{n\to\infty}(a_n)^{1/n}$ exists and the two limits are equal.

To see this, suppose $\lim_{n\to\infty}\dfrac{a_{n+1}}{a_n} = \ell$. Let $\epsilon > 0$. Then we have an $m \in \mathbb{N}$ such that for all $n > m$, $\ell - \epsilon < \dfrac{a_{n+1}}{a_n} < \ell + \epsilon$. Use the right side inequality first. For all such n, $a_n < (\ell+\epsilon)^{n-m}a_m$. Then

$$(a_n)^{1/n} < (\ell+\epsilon)((\ell+\epsilon)^{-m}a_m)^{1/n} \to \ell + \epsilon \text{ as } n \to \infty.$$

Therefore, $\lim(a_n)^{1/n} \le \ell + \epsilon$ for every $\epsilon > 0$. That is, $\lim(a_n)^{1/n} \le \ell$.
Similarly, the left side inequality gives $\lim(a_n)^{1/n} \ge \ell$.

Notice that this gives an alternative proof of Theorem **??**.

Example 1.16. Does the series $\displaystyle\sum_{n=1}^{\infty} 2^{(-1)^n - n} = 2 + \frac{1}{4} + \frac{1}{2} + \frac{1}{16} + \cdots$ converge?

Let $a_n = 2^{(-1)^n - n}$. Then

$$\frac{a_{n+1}}{a_n} = \begin{cases} 1/8 & \text{if } n \text{ even} \\ 2 & \text{if } n \text{ odd.} \end{cases}$$

Clearly, its limit does not exist. But

$$(a_n)^{1/n} = \begin{cases} 2^{1/n-1} & \text{if } n \text{ even} \\ 2^{-1/n-1} & \text{if } n \text{ odd} \end{cases}$$

This has limit $1/2 < 1$. Therefore, by Cauchy root test, the series converges.

1.9 Alternating series

Theorem 1.17. (Leibniz Alternating Series Test)
Let $\{a_n\}$ be a sequence of positive terms decreasing to 0; that is, for each n, $a_n \geq a_{n+1} > 0$, and $\lim_{n\to\infty} a_n = 0$. Then the series $\sum_{n=1}^{\infty}(-1)^n a_n$ converges, and its sum lies between $a_1 - a_2$ and a_1.

Proof: The partial sum upto $2n$ terms is

$$s_{2n} = (a_1 - a_2) + (a_3 - a_4) + \cdots + (a_{2n-1} - a_{2n}) = a_1 - \big[(a_2 - a_3) + \cdots + (a_{2n-2} - a_{2n-1}) - a_{2n}.$$

It is a sum of n positive terms bounded above by a_1 and below by $a_1 - a_2$. Hence

$$s_{2n} \text{ converges to some } s \text{ such that } a_1 - a_2 \leq s \leq a_1.$$

The partial sum upto $2n+1$ terms is $s_{2n+1} = s_{2n} + a_{2n+1}$. It converges to s as $\lim a_{2n+1} = 0$. Hence the series converges to some s with $a_1 - a_2 \leq s \leq a_1$. □

The bounds for s can be sharpened by taking $s_{2n} \leq s \leq s_{2n-1}$ for each $n > 1$.

Leibniz test now implies that the series $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} + \cdots$ is convergent to some s with $1/2 \leq s \leq 1$. By taking more terms, we can have different bounds such as

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} = \frac{7}{12} \leq s \leq 1 - \frac{1}{2} + \frac{1}{3} = \frac{10}{12}$$

In contrast, the harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots$ diverges to ∞.

We say that the series $\sum a_n$ is **absolutely convergent** iff the series $\sum |a_n|$ is convergent.

An alternating series $\sum a_n$ is said to be **conditionally convergent** iff it is convergent but it is not absolutely convergent.

Theorem 1.18. *An absolutely convergent series is convergent.*

Proof: Let $\sum a_n$ be an absolutely convergent series. Then $\sum |a_n|$ is convergent. Let $\epsilon > 0$. By Cauchy criterion, there exists an $n_0 \in \mathbb{N}$ such that for all $n > m > n_0$, we have

$$|a_m| + |a_{m+1}| + \cdots + |a_n| < \epsilon.$$

Now,

$$|a_m + a_{m+1} + \cdots + a_n| \leq |a_m| + |a_{m+1}| + \cdots + |a_n| < \epsilon.$$

Again, by Cauchy criterion, the series $\sum a_n$ is convergent. □

The series $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} + \cdots$ is a conditionally convergent series.

An absolutely convergent series can be rearranged in any way we like, but the sum remains the same. Whereas a rearrangement of the terms of a conditionally convergent series may lead to divergence or convergence to any other number. In fact, a conditionally convergent series can always be rearranged in a way so that the rearranged series converges to any desired number; we will not prove this fact.

Example 1.17. Do the series (a) $\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{2^n}$ (b) $\sum_{n=1}^{\infty}\frac{\cos n}{n^2}$ converge?

(a) $\sum(2)^{-n}$ converges. Therefore, the given series converges absolutely; hence it converges.

(b) $\left|\frac{\cos n}{n^2}\right| \leq \frac{1}{n^2}$ and $\sum(n^{-2})$ converges. By comparison test, the given series converges absolutely; and hence it converges.

Example 1.18. Discuss the convergence of the series $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^p}$.

For $p > 1$, the series $\sum n^{-p}$ converges. Therefore, the given series converges absolutely for $p > 1$.

For $0 < p \leq 1$, by Leibniz test, the series converges. But $\sum n^{-p}$ does not converge. Therefore, the given series converges conditionally for $0 < p \leq 1$.

For $p \leq 0$, $\lim \frac{(-1)^{n+1}}{n^p} \neq 0$. Therefore, the given series diverges in this case.

Chapter 2

Series Representation of Functions

2.1 Power series

Let $a \in \mathbb{R}$. A **power series about** $x = a$ is a series of the form

$$\sum_{n=0}^{\infty} c_n(x-a)^n = c_0 + c_1(x-a) + c_2(x-a)^2 + \cdots$$

The point a is called the **center** of the power series and the real numbers $c_0, c_1, \cdots, c_n, \cdots$ are its **coefficients**.

For example, the geometric series

$$1 + x + x^2 + \cdots + x^n + \cdots$$

is a power series about $x = 0$ with each coefficient as 1. We know that its sum is $\dfrac{1}{1-x}$ for $-1 < x < 1$. And we know that for $|x| \geq 1$, the geometric series does not converge. That is, the series defines a function from $(-1, 1)$ to $\mathbb{R}$ and it is not meaningful for other values of x.

Example 2.1. Show that the following power series converges for $0 < x < 4$.

$$1 - \frac{1}{2}(x-2) + \frac{1}{4}(x-2)^2 + \cdots + \frac{(-1)^n}{2^n}(x-2)^n + \cdots$$

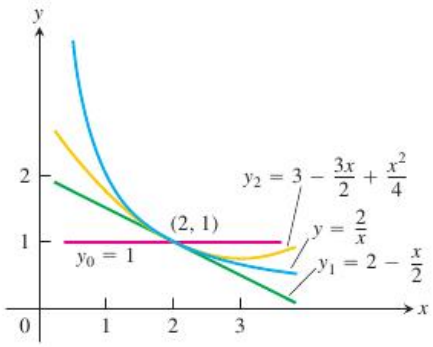
It is a geometric series with the ratio as $r = (-1/2)(x-2)$. Thus it converges for $|(-1/2)(x-2)| < 1$. Simplifying we get the constraint as $0 < x < 4$.

Notice that the power series sums to

$$\frac{1}{1-r} = \frac{1}{1-\frac{-1}{2(x-2)}} = \frac{2}{x}.$$

Thus, the power series gives a series expansion of the function $\dfrac{2}{x}$ for $0 < x < 4$.

Truncating the series to n terms give us polynomial approximations of the function $\dfrac{2}{x}$.



Theorem 2.1. (Convergence Theorem for Power Series) *Suppose that the power series* $\sum_{n=0}^{\infty} a_n x^n$ *is convergent for* $x = c$ *and divergent for* $x = d$ *for some* $c > 0,\ d > 0$. *Then the power series converges absolutely for all* x *with* $|x| < c$; *and it diverges for all* x *with* $|x| > d$.

Proof: The power series converges for $x = c$ means that $\sum a_n c^n$ converges. Thus $\lim_{n\to\infty} a_n c^n = 0$. Then we have an $M \in \mathbb{N}$ such that for all $n > M$, $|a_n c^n| < 1$.
Let $x \in \mathbb{R}$ be such that $|x| < c$. Write $t = |\frac{x}{c}|$. For each $n > M$, we have

$$|a_n|\,|x|^n = |a_n x^n| = |a_n c^n|\,|\tfrac{x}{c}|^n < |\tfrac{x}{c}|^n = t^n.$$

As $0 \le t < 1$, the geometric series $\sum_{n=M+1}^{\infty} t^n$ converges. By comparison test, for any x with $|x| < c$, the series $\sum_{n=M+1}^{\infty} |a_n x|^n$ converges. However, $\sum_{n=0}^{M} |a_n x^n|$ is finite. Therefore, the power series $\sum_{n=0}^{\infty} a_n x^n$ converges absolutely for all x with $|x| < c$.

For the divergence part of the theorem, suppose, on the contrary that the power series converges for some $\alpha > d$. By the convergence part, the series must converge for $x = d$, a contradiction. □

Notice that if the power series is about a point $x = a$, then we take $t = x - a$ and apply Theorem **??**. Also, for $x = 0$, the power series $\sum a_n x^n$ always converges.

Consider the power series $\sum_{n=0}^{\infty} a_n (x-a)^n$. The real number

$$R = \text{lub}\{c \ge 0 :\ \text{the power series converges for all } x \text{ with } |x-a| < c\}$$

is called the **radius of convergence** of the power series.

That is, R is such non-negative number that the power series converges for all x with $|x-a| < R$ and it diverges for all x with $|x-a| > R$.

If the radius of convergence of the power series $\sum a_n (x-a)^n$ is R, then the **interval of convergence** of the power series is

$(a-R, a+R)$ if it diverges at both $x = a-R$ and $x = a+R$.

$[a-R, a+R)$ if it converges at $x = a-R$ and diverges at $x = a+R$.

$(a-R, a+R]$ if it diverges at $x = a-R$ and converges at $x = a+R$.

Theorem **??** guarantees that the power series converges everywhere inside the interval of convergence, it converges absolutely inside the open interval $(a-R, a+R)$, and it diverges everywhere outside the interval of convergence.

Also, see that when $R=\infty$, the power series converges for all $x \in \mathbb{R}$, and when $R=0$, the power series converges only at the point $x=a$, whence its sum is a_0.

To determine the interval of convergence, you may find the radius of convergence R, and then test for its convergence separately for the end-points $x=a-R$ and $x=a+R$.

2.2 Determining radius of convergence

Theorem 2.2. *The radius of convergence of the power series* $\sum_{n=0}^{\infty} a_n(x-a)^n$ *is given by* $\lim_{n\to\infty} |a_n|^{-1/n}$ *provided that this limit is either a real number or equal to* ∞.

Proof: Let R be the radius of convergence of the power series $\sum_{n=0}^{\infty} a_n(x-a)^n$. Let $\lim_{n\to\infty} |a_n|^{1/n} = r$. We consider three cases and show that

$$(1)\ \ r>0 \Rightarrow R=\frac{1}{r}, \quad (2)\ \ r=\infty \Rightarrow R=0, \quad (3)\ \ r=0 \Rightarrow R=\infty.$$

(1) Let $r>0$. By the root test, the series is absolutely convergent whenever

$$\lim_{n\to\infty} |a_n(x-a)^n|^{1/n} < 1 \ \text{ i.e., } \ |x-a| \lim_{n\to\infty} |a_n|^{1/n} < 1 \ \text{ i.e., } \ |x-a| < \frac{1}{r}.$$

It also follows from the root test that the series is divergent when $|x-a| > 1/r$. Hence $R = 1/r$.

(2) Let $r=\infty$. Then for any $x \neq a$, $\lim |a_n(x-a)^n| = \lim |x-a||a_n|^{1/n} = \infty$. By the root test, $\sum a_n(x-a)^n$ diverges for each $x \neq a$. Thus, $R=0$.

(3) Let $r=0$. Then for any $x \in \mathbb{R}$, $\lim |a_n(x-a)^n|^{1/n} = |x-a| \lim |a_n|^{1/n} = 0$. By the root test, the series converges for each $x \in \mathbb{R}$. So, $R=\infty$. □

Theorem 2.3. *The radius of convergence of the power series* $\sum_{n=0}^{\infty} a_n(x-a)^n$ *is given by* $\lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right|$, *provided that this limit is either a real number or equal to* ∞.

Example 2.2. For what values of x, do the following power series converge?

$$\text{(a)} \quad \sum_{n=0}^{\infty} n!x^n \qquad \text{(b)} \quad \sum_{n=0}^{\infty} \frac{x^n}{n!} \qquad \text{(c)} \quad \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}$$

(a) $a_n = n!$. Thus $\lim |a_n/a_{n+1}| = \lim 1/(n+1) = 0$. Hence $R=0$. That is, the series is convergent only at $x=0$.

(b) $a_n = 1/n!$. Thus $\lim |a_n/a_{n+1}| = \lim(n+1) = \infty$. Hence $R = \infty$. That is, the series is convergent for all $x \in \mathbb{R}$.

(c) Here, the power series is not in the form $\sum b_n x^n$. The series can be thought of as

$$x\Big(1 - \frac{x^2}{3} + \frac{x^4}{5} + \cdots\Big) = x\sum_{n=0}^{\infty}(-1)^n \frac{t^n}{2n+1} \quad \text{for } \ t = x^2$$

Now, for the power series $\sum(-1)^n \dfrac{t^n}{2n+1}$, $a_n = (-1)^n/(2n+1)$.
Thus $\lim |a_n/a_{n+1}| = \lim \frac{2n+3}{2n+1} = 1$. Hence $R = 1$. That is, for $|t| = x^2 < 1$, the series converges and for $|t| = x^2 > 1$, the series diverges.

Alternatively, you can use the geometric series. That is, for any $x \in \mathbb{R}$, consider the series

$$x\Big(1 - \frac{x^2}{3} + \frac{x^4}{5} + \cdots\Big).$$

By the ratio test, the series converges if

$$\lim_{n\to\infty}\Big|\frac{u_n}{u_{n+1}}\Big| = \lim_{n\to\infty}\frac{2n+3}{2n+1}|x^2| = x^2 < 1.$$

That is, the power series converges for $-1 < x < 1$. Also, by the ratio test, the series diverges for $|x| > 1$.

What happens for $|x| = 1$?
For $x = -1$, the original power series is an alternating series; it converges due to Liebniz. Similarly, for $x = 1$, the alternating series also converges.
Hence the interval of convergence for the original power series (in x) is $[-1, 1]$.

If R is the radius of convergence of a power series $\sum a_n(x-a)^n$, then the series defines a function $f(x)$ from the open interval $(a-R, a+R)$ to $\mathbb{R}$ by

$$f(x) = a_0 + a_1(x-a) + a_2(x-a)^2 + \cdots = \sum_{n=0}^{\infty} a_n(x-a)^n \quad \text{for } \ x \in (a-R, a+R).$$

Theorem 2.4. *Let the power series $\sum_{n=0}^{\infty} a_n(x-a)^n$ have radius of convergence $R > 0$. Then the power series defines a function $f : (a-R, a+R) \to \mathbb{R}$. Further, $f'(x)$ and $\int f(x)dx$ exist as functions from $(a-R, a+R)$ to $\mathbb{R}$ and these are given by*

$$f(x) = \sum_{n=0}^{\infty} a_n(x-a)^n, \quad f'(x) = \sum_{n=1}^{\infty} n a_n(x-a)^{n-1}, \quad \int f(x)dx = \sum_{n=0}^{\infty} a_n \frac{(x-a)^{n+1}}{n+1} + C,$$

where all the three power series converge for all $x \in (a-R, a+R)$.

Caution: Term by term differentiation may not work for series, which are not power series. For example, $\displaystyle\sum_{n=1}^{\infty} \frac{\sin(n!x)}{n^2}$ is convergent for all x. The series obtained by term-by-term differentiation is $\displaystyle\sum_{n=1}^{\infty} \frac{n!\cos(n!x)}{n^2}$; it diverges for all x.

Theorem 2.5. *Let the power series $\sum a_n x^n$ and $\sum b_n x^n$ have the same radius of convergence $R > 0$. Then their multiplication has the same radius of convergence R. Moreover, the functions they define satisfy the following:*

$$\text{If } f(x) = \sum a_n x^n,\ g(x) = \sum b_n x^n \text{ then } f(x)g(x) = \sum c_n x^n \text{ for } a - R < x < a + R$$

where $c_n = \sum_{k=0}^{n} a_k b_{n-k} = a_0 b_n + a_1 b_{n-1} + \cdots + a_{n-1} b_1 + a_n b_0$.

Example 2.3. Determine power series expansions of the functions (a) $\dfrac{2}{(x-1)^3}$ (b) $\tan^{-1} x$.

(a) For $-1 < x < 1$, $\dfrac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots$.

Differentiating term by term, we have

$$\frac{1}{(1-x)^2} = 1 + 2x + 3x^2 + 4x^3 + \cdots$$

Differentiating once more, we get

$$\frac{2}{(1-x)^3} = 2 + 6x + 12x^2 + \cdots = \sum_{n=2}^{\infty} n(n-1)x^{n-2} \quad \text{for} \quad -1 < x < 1.$$

(b) $\dfrac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + x^8 - \cdots$ for $|x^2| < 1$.

Integrating term by term we have

$$\tan^{-1} x + C = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots \quad \text{for} \quad -1 < x < 1.$$

Evaluating at $x = 0$, we see that $C = 0$. Hence the power series for $\tan^{-1} x$.

2.3 Taylor's formulas

Theorem 2.6. (Taylor's Formula in Differential Form) *Let $n \in \mathbb{N}$. Suppose that $f^{(n)}(x)$ is continuous on $[a, b]$ and is differentiable on (a, b). Then there exists a point $c \in (a, b)$ such that*

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}.$$

Proof: For $x = a$, the formula holds. So, let $x \in (a, b]$. For any $t \in [a, x]$, let

$$p(t) = f(a) + f'(a)(t-a) + \frac{f''(a)}{2!}(t-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(t-a)^n.$$

Here, we treat x as a certain point, not a variable; and t as a variable. Write

$$g(t) = f(t) - p(t) - \frac{f(x) - p(x)}{(x-a)^{n+1}}\,(t-a)^{n+1}.$$

We see that $g(a) = 0,\ g'(a) = 0,\ g''(a) = 0,\ \ldots, g^{(n)}(a) = 0,$ and $g(x) = 0.$

By Rolle's theorem, there exists $c_1 \in (a, x)$ such that $g'(c_1) = 0$. Since $g(a) = 0$, apply Rolle's theorem once more to get a $c_2 \in (a, c_1)$ such that $g''(c_2) = 0$.

Continuing this way, we get a $c_{n+1} \in (a, c_n)$ such that $g^{(n+1)}(c_{n+1}) = 0$.

Since $p(t)$ is a polynomial of degree at most n, $p^{(n+1)}(t) = 0$. Then

$$g^{(n+1)}(t) = f^{(n+1)}(t) - \frac{f(x) - p(x)}{(x-a)^{n+1}}\,(n+1)!.$$

Evaluating at $t = c_{n+1}$ we have $f^{(n+1)}(c_{n+1}) - \dfrac{f(x) - p(x)}{(x-a)^{n+1}}\,(n+1)! = 0$. That is,

$$\frac{f(x) - p(x)}{(x-a)^{n+1}} = \frac{f^{(n+1)}(c_{n+1})}{(n+1)!}.$$

Consequently, $g(t) = f(t) - p(t) - \dfrac{f^{(n+1)}(c_{n+1})}{(n+1)!}\,(t-a)^{n+1}$.

Evaluating it at $t = x$ and using the fact that $g(x) = 0$, we get

$$f(x) = p(x) + \frac{f^{(n+1)}(c_{n+1})}{(n+1)!}\,(x-a)^{n+1}.$$

Since x is an arbitrary point in $(a, b]$, this completes the proof. □

The polynomial

$$p(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n$$

in Taylor's formula is called as **Taylor's polynomial of order** n. Notice that the degree of the Taylor's polynomial may be less than or equal to n. The expression given for $f(x)$ there is called **Taylor's formula** for $f(x)$. Taylor's polynomial is an approximation to $f(x)$ with **the error**

$$R_n(x) = \frac{f^{(n+1)}(c_{n+1})}{(n+1)!}\,(x-a)^{n+1}.$$

How good $f(x)$ is approximated by $p(x)$ depends on the smallness of the error $R_n(x)$.

For example, if we use $p(x)$ of order 5 for approximating $\sin x$ at $x = 0$, then we get

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} + R_6(x), \quad \text{where} \quad R_6(x) = \frac{\sin\theta}{6!}x^6.$$

Here, θ lies between 0 and x. The absolute error is bounded above by $|x|^6/6!$. However, if we take the Taylor's polynomial of order 6, then $p(x)$ is the same as in the above, but the absolute error is now $|x|^7/7!$. If x is near 0, this is smaller than the earlier bound.

Notice that if $f(x)$ is a polynomial of degree n, then Taylor's polynomial of order n is equal to the original polynomial.

Theorem 2.7. (Taylor's Formula in Integral Form) *Let $f(x)$ be an $(n+1)$-times continuously differentiable function on an open interval I containing a. Let $x \in I$. Then*

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n(x),$$

where $R_n(x) = \displaystyle\int_a^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)\, dt$. *An estimate for $R_n(x)$ is given by*

$$\frac{m\, x^{n+1}}{(n+1)!} \le R_n(x) \le \frac{M\, x^{n+1}}{(n+1)!}$$

where $m \le f^{n+1}(x) \le M$ for $x \in I$.

Proof: We prove it by induction on n. For $n = 0$, we should show that

$$f(x) = f(a) + R_0(x) = f(a) + \int_a^x f'(t)\, dt.$$

But this follows from the Fundamental theorem of calculus. Now, suppose that Taylor's formula holds for $n = m$. That is, we have

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(m)}(a)}{m!}(x-a)^m + R_m(x),$$

where $R_m(x) = \displaystyle\int_a^x \frac{(x-t)^m}{m!} f^{(m+1)}(t)\, dt$. We evaluate $R_m(x)$ using integration by parts with the first function as $f^{(m+1)}(t)$ and the second function as $(x-t)^m/m!$. Remember that the variable of integration is t and x is a fixed number. Then

$$\begin{aligned} R_m(x) &= \Big[- f^{(m+1)}(t) \frac{(x-t)^{m+1}}{(m+1)!} \Big]_a^x + \int_a^x f^{(m+2)}(t) \frac{(x-t)^{m+1}}{(m+1)!}\, dt \\ &= f^{(m+1)}(a) \frac{(x-a)^{m+1}}{(m+1)!} + \int_a^x f^{(m+2)}(t) \frac{(x-t)^{m+1}}{(m+1)!}\, dt \\ &= \frac{f^{(m+1)}(a)}{(m+1)!}(x-a)^{m+1} + R_{m+1}(x). \end{aligned}$$

This completes the proof of Taylor's formula. For the estimate of $R_n(x)$, Observe that

$$\begin{aligned} R_n(x) &= \int_a^x \frac{(x-t)^n}{n!}(t)\, dt = (-1)^n \int_a^x \frac{(t-x)^n}{n!}(t)\, dt \\ &= (-1)^n \Big[\frac{(t-x)^{n+1}}{(n+1)!} \Big]_a^x = -(-1)^n (-1)^{n+1} \frac{(x-a)^{n+1}}{(n+1)!} = \frac{(x-a)^{n+1}}{(n+1)!}. \end{aligned}$$

Now, the estimate for $R_n(x)$ follows. □

2.4 Taylor series

Taylor's formulas (Theorem **??** and Theorem **??**) say that under suitable hypotheses a function can be written in the following forms:

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}.$$

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \int_a^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)\, dt.$$

It is thus clear that whenever one (form) of the remainder term

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1} \quad \text{OR} \quad R_n(x) = \int_a^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)\, dt$$

converges to 0 for all x in an interval around the point $x = a$, the series on the right hand side would converge and then the function can be written in the form of a series. That is, under the conditions that $f(x)$ has derivatives of all order, and $R_n(x) \to 0$ for all x in an interval around $x = a$, the function $f(x)$ has a power series representation

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \cdots$$

Such a series is called the **Taylor series** expansion of the function $f(x)$. When $a = 0$, the Taylor series is called the **Maclaurin series**.

Conversely, if a function $f(x)$ has a power series expansion about $x = a$, then by repeated differentiation and evaluation at $x = a$ shows that the coefficients of the power series are precisely of the form $\dfrac{f^{(n)}(a)}{n!}$ as in the Taylor series.

Example 2.4. Find the Taylor series expansion of the function $f(x) = 1/x$ at $x = 2$. In which interval around $x = 2$, the series converges?

We see that $f(x) = x^{-1},\ f(2) = \dfrac{1}{2};\quad \cdots ; f^{(n)}(x) = (-1)^n n! x^{-(n+1)},\ f^{(n)}(2) = (-1)^n n! 2^{-(n+1)}$.

Hence the Taylor series for $f(x) = 1/x$ is

$$\frac{1}{2} - \frac{x-2}{2^2} + \frac{(x-2)^2}{2^3} - \cdots + (-1)^n \frac{(x-2)^n}{2^{n+1}} + \cdots$$

A direct calculation can be done looking at the Taylor series so obtained. Here, the series is a geometric series with ratio $r = -(x-2)/2$. Hence it converges absolutely whenever

$$|r| < 1, \quad \text{i.e.,} \quad |x-2| < 2 \quad \text{i.e.,} \quad 0 < x < 4.$$

Does this convergent series converge to the given function? We now require the remainder term in the Taylor expansion. The absolute value of the remainder term in the differential form is (for any c, x in an interval around $x = 2$)

$$|R_n| = \left| \frac{f^{(n+1)}(c)}{(n+1)!}(x-2)^{n+1} \right| = \left| \frac{(x-2)^{n+1}}{c^{n+2}} \right|$$

Here, c lies between x and 2. Clearly, if x is near 2, $|R_n| \to 0$. Hence the Taylor series represents the function near $x = 2$.

Example 2.5. Consider the function $f(x) = e^x$. For its Maclaurin series, we find that

$$f(0) = 1,\ f'(0) = 1, \cdots, f^{(n)}(0) = 1, \cdots$$

Hence

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots$$

By the ratio test, this power series has the radius of convergence

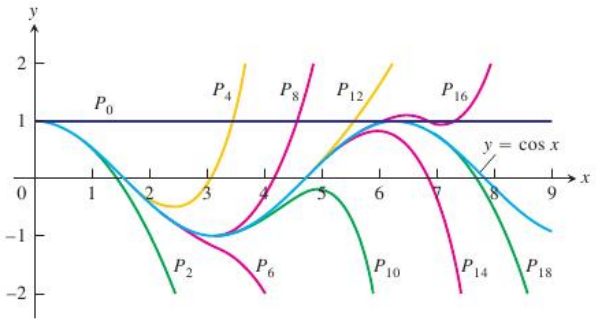
$$R = \lim_{n\to\infty} \frac{a_n}{a_{n+1}} = \lim_{n\to\infty} \frac{(n+1)!}{n!} = \infty.$$

Therefore, for every $x \in \mathbb{R}$ the above series converges. Using the integral form of the remainder,

$$|R_n(x)| = \Big| \int_a^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)\, dt \Big| = \Big| \int_0^x \frac{(x-t)^n}{n!} e^t\, dt \Big| \to 0 \quad \text{as} \quad n \to \infty.$$

Hence, e^x has the above power series expansion for each $x \in \mathbb{R}$.

Similarly, you can show that $\cos x = \displaystyle\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}$. The Taylor polynomials approximating $\cos x$ are therefore $P_{2n}(x) = \displaystyle\sum_{k=n}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!}$. The following picture shows how these polynomials approximate $\cos x$ for $0 \le x \le 9$.



In the above Maclaurin series expansion of $\cos x$, we have the absolute value of the remainder in the differential form as

$$|R_{2n}(x)| = \frac{|x|^{2n+1}}{(2n+1)!} \to 0 \quad \text{as} \quad n \to \infty$$

for any $x \in \mathbb{R}$. Hence the series represents $\cos x$ for each $x \in \mathbb{R}$.

Example 2.6. Let $m \in \mathbb{R}$. Show that, for $-1 < x < 1$,

$$(1+x)^m = 1 + \sum_{n=1}^{\infty} \binom{m}{n} x^n, \quad \text{where} \quad \binom{m}{n} = \frac{m(m-1)\cdots(m-n+1)}{n!}.$$

To see this, find the derivatives of the given function:

$$f(x) = (1+x)^m, \quad f^{(n)}(x) = m(m-1)\cdots(m-n+1)x^{m-n}.$$

Then the Maclaurin series for $f(x)$ is the given series. You must show that the series converges for $-1 < x < 1$ and then the remainder term in the Maclaurin series expansion goes to 0 as $n \to \infty$ for all such x. The series so obtained is called a **binomial series** expansion of $(1+x)^m$. Substituting values of m, we get series for different functions. For example, with $m = 1/2$, we have

$$(1+x)^{1/2} = 1 + \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16} - \cdots \quad \text{for} \quad -1 < x < 1.$$

Notice that when $m \in \mathbb{N}$, the binomial series terminates to give a polynomial and it represents $(1+x)^m$ for each $x \in \mathbb{R}$.

2.5 Fourier series

A **trigonometric series** is of the form

$$\frac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx)$$

Since both cosine and sine functions are periodic of period 2π, if the trigonometric series converges to a function $f(x)$, then necessarily $f(x)$ is also periodic of period 2π. Thus,

$$f(0) = f(2\pi) = f(4\pi) = f(6\pi) = \cdots \quad \text{and} \quad f(-\pi) = f(\pi), \quad \text{etc.}$$

Moreover, if $f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx)$, say, for all $x \in [-\pi, \pi]$, then the coefficients can be determined from $f(x)$. Towards this, multiply $f(t)$ by $\cos mt$ and integrate to obtain:

$$\begin{aligned}\int_{-\pi}^{\pi} f(t)\cos mt\,dt &= \frac{1}{2}a_0\int_{-\pi}^{\pi}\cos mt\,dt + \sum_{n=1}^{\infty}a_n\int_{-\pi}^{\pi}\cos nt\cos mt\,dt \\ &\quad + \sum_{n=1}^{\infty}b_n\int_{-\pi}^{\pi}\sin nt\cos mt\,dt.\end{aligned}$$

For $m, n = 0, 1, 2, 3, \ldots,$

$$\int_{-\pi}^{\pi}\cos nt\cos mt = \begin{cases} 0 & \text{if } n \neq m \\ \pi & \text{if } n = m > 0 \\ 2\pi & \text{if } n = m = 0 \end{cases} \quad \text{and} \quad \int_{-\pi}^{\pi}\sin nt\cos mt = 0.$$

Thus, we obtain $\displaystyle\int_{-\pi}^{\pi} f(t)\cos mt = \pi a_m, \quad \text{for all } \ m = 0, 1, 2, 3, \cdots$

Similarly, by multiplying $f(t)$ by $\sin mt$ and integrating, and using the fact that

$$\int_{-\pi}^{\pi} \sin nt \sin mt = \begin{cases} 0 & \text{if } n \neq m \\ \pi & \text{if } n = m > 0 \\ 0 & \text{if } n = m = 0 \end{cases}$$

we obtain $\displaystyle\int_{-\pi}^{\pi} f(t) \sin mt = \pi b_m, \quad \text{for all } m = 1, 2, 3, \cdots.$

Assuming that $f(x)$ has period 2π, we then give the following definition.

Let $f : [-\pi, \pi] \to \mathbb{R}$ be an integrable function extended to $\mathbb{R}$ by periodicity of period 2π, i.e., $f : \mathbb{R} \to \mathbb{R}$ satisfies

$$f(x + 2\pi) = f(x) \quad \text{for all } x \in \mathbb{R}.$$

Let $a_n = \dfrac{1}{\pi}\displaystyle\int_{-\pi}^{\pi} f(t) \cos nt\, dt$ for $n = 0, 1, 2, 3, , \ldots,$ and $b_n = \dfrac{1}{\pi}\displaystyle\int_{-\pi}^{\pi} f(t) \sin nt\, dt$ for $n = 1, 2, 3, \ldots.$ Then the trigonometric series $\dfrac{1}{2}a_0 + \displaystyle\sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx)$ is called the **Fourier series** of $f(x)$.

Recall that a function is called piecewise continuous on an interval iff all points in that interval, where the function is discontinuous, are finite in number; and at such points c, the left and right sided limits $f(c-)$ and $f(c+)$ exist.

Theorem 2.8. (Convergence of Fourier Series) *Let $f : [-\pi, \pi] \to \mathbb{R}$ be a function extended to $\mathbb{R}$ by periodicity of period 2π. Let $f(x)$ satisfy* at least one of *the following conditions:*

1. *$f(x)$ is a bounded and monotonic function.*
2. *$f(x)$ is piecewise continuous, and $f(x)$ has both left hand derivative and right hand derivative at each $x \in (-\pi, \pi)$.*

Then $f(x)$ is equal to its Fourier series at all points where $f(x)$ is continuous; and at a point c, where $f(x)$ is discontinuous, the Fourier series converges to $\frac{1}{2}[f(c+) + f(c-)]$.

In particular, if $f(x)$ and $f'(x)$ are continuous on $[-\pi, \pi]$ with period 2π, then

$$f(x) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx) \quad \text{for all } x \in \mathbb{R}.$$

Further, if $f(x)$ is an odd function, i.e., $f(-x) = f(x)$, then for $n = 0, 1, 2, 3, \ldots,$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \cos nt\, dt = 0$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \sin nt\, dt = \frac{2}{\pi}\int_{0}^{\pi} f(t) \sin nt\, dt.$$

In this case,

$$f(x) = \sum_{n=1}^{\infty} b_n \sin nx \quad \text{for all} \quad x \in \mathbb{R}.$$

Similarly, if $f(x)$ and $f'(x)$ are continuous on $[-\pi, \pi]$ with period 2π and if $f(x)$ is an even function, i.e., $f(-x) = f(x)$, then

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nx \quad \text{for all} \quad x \in \mathbb{R},$$

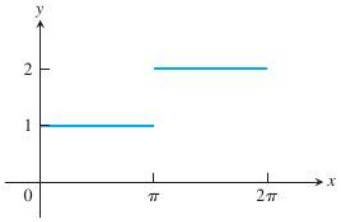
where

$$a_n = \frac{2}{\pi} \int_0^{\pi} f(t) \cos nt \, dt \quad \text{for} \quad n = 0, 1, 2, 3, \ldots$$

Fourier series can represent functions which cannot be represented by a Taylor series, or a conventional power series; for example, a step function.

Example 2.7. Find the Fourier series of the function $f(x)$ given by the following which is extended to $\mathbb{R}$ with the periodicity 2π:

$$f(x) = \begin{cases} 1 & \text{if} \quad 0 \le x < \pi \\ 2 & \text{if} \quad \pi \le x < 2\pi \end{cases}$$



Due to periodic extension, we can rewrite the function $f(x)$ on $[-\pi, \pi]$ as

$$f(x) = \begin{cases} 2 & \text{if} \quad -\pi \le x < 0 \\ 1 & \text{if} \quad 0 \le x < \pi \end{cases}$$

Then the coefficients of the Fourier series are computed as follows:

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{0} f(t) \, dt + \frac{1}{\pi} \int_0^{\pi} f(t) \, dt = 3.$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{0} \cos nt \, dt + \frac{1}{\pi} \int_0^{\pi} 2 \cos nt \, dt = 0.$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{0} \sin nt \, dt + \frac{1}{\pi} \int_0^{\pi} 2 \sin nt \, dt = \frac{(-1)^n - 1}{n\pi}.$$

Notice that $b_1 = -\frac{2}{\pi}$, $b_2 = 0$, $b_3 = -\frac{2}{3\pi}$, $b_4 = 0, \ldots$. Therefore,

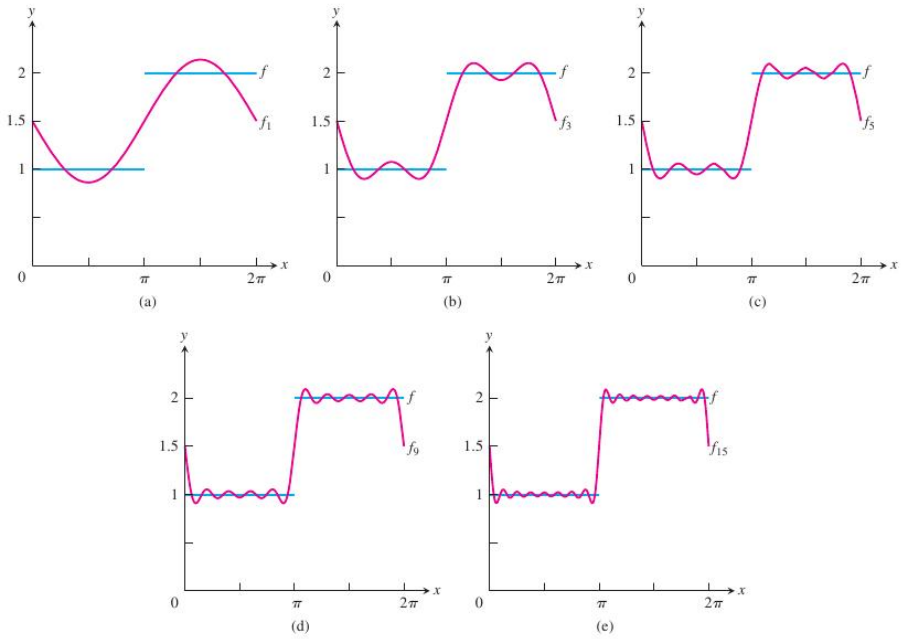
$$f(x) = \frac{3}{2} - \frac{2}{\pi}\Big(\sin x + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \cdots \Big).$$

Here, the last expression for $f(x)$ holds for all $x \in \mathbb{R}$; however, the function here has been extended to $\mathbb{R}$ by using its periodicity as 2π. If we do not extend but find the Fourier series for the function as given on $[-\pi, \pi)$, then also for all $x \in [-\pi, \pi)$, the same expression holds.

Once we have a series representation of a function, we should see how the **partial sums** of the series approximate the function. In the above example, let us write

$$f_m(x) = \frac{1}{2}a_0 + \sum_{n=1}^{m}(a_n \cos nx + b_n \sin nx).$$

The approximations $f_1(x)$, $f_3(x)$, $f_5(x)$, $f_9(x)$ and $f_{15}(x)$ to $f(x)$ are shown in the figure below.



Example 2.8. Show that $x^2 = \dfrac{\pi^2}{3} + 4\sum_{n=1}^{\infty}(-1)^n \dfrac{\cos nx}{n^2}$ for all $x \in [-\pi, \pi]$.

The extension of $f(x) = x^2$ to $\mathbb{R}$ is not the function x^2. For instance, in the interval $[\pi, 3\pi]$, its extension looks like $f(x) = (x - 2\pi)^2$. Remember that the extension has period 2π. Also, notice that $f(\pi) = f(-\pi)$; thus we have no problem at the point π in extending the function continuously. With this understanding, we go for the Fourier series expansion of $f(x) = x^2$ in the interval $[-\pi, \pi]$. We also see that $f(x)$ is an even function. Its Fourier series is a cosine series. The coefficients of the series are as follows:

$$a_0 = \frac{2}{\pi}\int_0^{\pi} t^2\, dt = \frac{2}{3}\pi^2.$$

$$a_n = \frac{2}{\pi}\int_0^{\pi} t^2 \cos nt\, dt = \frac{4}{n^2}(-1)^n.$$

Therefore,

$$f(x) = x^2 = \frac{\pi^2}{3} + 4\sum_{n=1}^{\infty}(-1)^n \frac{\cos nx}{n^2} \quad \text{for all} \quad x \in [-\pi, \pi].$$

In particular, by taking $x = 0$ and $x = \pi$, we have

$$\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^2} = \frac{\pi^2}{12}, \qquad \sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}.$$

Due to the periodic extension of $f(x)$ to $\mathbb{R}$, we see that

$$(x-2\pi)^2 = \frac{\pi^2}{3} + 4\sum_{n=1}^{\infty}(-1)^n \frac{\cos nx}{n^2} \quad \text{for all} \;\; x \in [\pi, 3\pi].$$

It also follows that the same sum (of the series) is equal to $(x-4\pi)^2$ for $x \in [3\pi, 5\pi]$, etc.

Example 2.9. Show that the Fourier series for $f(x) = x^2$ defined on $(0, 2\pi)$ is given by

$$\frac{4\pi^2}{6} + \sum_{n=1}^{\infty}\Big(\frac{4}{\pi^2}\cos nx - \frac{4\pi}{n}\sin nx\Big).$$

Extend $f(x)$ to $\mathbb{R}$ by periodicity 2π and by taking $f(0) = f(2\pi)$. Then

$$f(-\pi) = f(-\pi+2\pi) = f(\pi) = \pi^2, \; f(-\pi/2) = f(-\pi/2+2\pi) = f(3\pi/2) = (3\pi/2)^2, \; f(0) = f(2\pi).$$

Thus the function $f(x)$ on $[-\pi, \pi]$ is defined by

$$f(x) = \begin{cases} (x+2\pi)^2 & \text{if} \quad -\pi \le x < 0 \\ x^2 & \text{if} \;\; 0 \le x \le \pi. \end{cases}$$

Notice that $f(x)$ is neither odd nor even. The coefficients of the Fourier series for $f(x)$ are

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\,dt = \frac{1}{\pi}\int_0^{2\pi} t^2\,dt = \frac{8\pi^2}{3}.$$

$$a_n = \frac{1}{\pi}\int_0^{2\pi} t^2 \cos nt\,dt = \frac{4}{n^2}.$$

$$b_n = \frac{1}{\pi}\int_0^{2\pi} t^2 \sin nt\,dt = -\frac{4\pi}{n}.$$

Hence the Fourier series for $f(x)$ is as claimed.

As per the extension of $f(x)$ to $\mathbb{R}$, we see that in the interval $(2k\pi, 2(k+1)\pi)$, the function is defined by $f(x) = (x-2k\pi)^2$. Thus it has discontinuities at the points $x = 0, \; \pm 2\pi, \; \pm 4\pi, \; \ldots$ At such a point $x = 2k\pi$, the series converges to the average value of the left and right side limits, i.e., the series when evaluated at $2k\pi$ yields the value

$$\frac{1}{2}\Big[\lim_{x\to 2k\pi-} f(x) + \lim_{x\to 2k\pi+} f(x)\Big] = \frac{1}{2}\Big[\lim_{x\to 2k\pi-}(x-2k\pi)^2 + \lim_{x\to 2k\pi+}(x-2(k+1)\pi)^2\Big] = 2\pi^2.$$

Notice that since $f(x)$ is extended by periodicity, whether we take the basic interval as $[-\pi, \pi]$ or as $[0, 2\pi]$ does not matter in the calculation of coefficients. We will follow this suggestion elsewhere instead of always redefining $f(x)$ on $[-\pi, \pi]$. However, the odd or even classification of $f(x)$ may break down.

Example 2.10. Show that for $0 < x < 2\pi$, $\quad \frac{1}{2}(\pi - x) = \sum_{n=1}^{\infty} \frac{\sin nx}{n}$.

Let $f(x) = x$ for $0 < x < 2\pi$. Extend $f(x)$ to $\mathbb{R}$ by taking the periodicity as 2π and with the condition that $f(0) = f(2\pi)$. As in Example **??**, $f(x)$ is not an odd function. For instance, $f(-\pi/2) = f(3\pi/2) = 3\pi/2 \neq f(\pi/2) = \pi/2$.

The coefficients of the Fourier series for $f(x)$ are as follows:

$$a_0 = \frac{1}{\pi}\int_0^{2\pi} t\,dt = 2\pi, \quad a_n = \frac{1}{\pi}\int_0^{2\pi} t\cos nt\,dt = 0.$$

$$b_n = \frac{1}{\pi}\int_0^{2\pi} t\sin nt\,dt = \frac{1}{\pi}\Big[\frac{-n\cos nt}{n}\Big]_0^{2\pi} + \frac{1}{n\pi}\int_0^{2\pi}\cos nt\,dt = -\frac{2}{\pi}.$$

By the convergence theorem, $x = \pi - 2\sum_{n=1}^{\infty}\frac{\sin nx}{n}$ for $0 < x < 2\pi$, which yields the required result.

Example 2.11. Find the Fourier series expansion of $f(x) = \begin{cases} x & \text{if } 0 \le x \le \pi/2 \\ \pi - x & \text{if } \pi/2 \le x \le \pi. \end{cases}$

Notice that $f(x)$ has the domain as an interval of length π and not 2π. Thus, there are many ways of extending it to $\mathbb{R}$ with periodicity 2π.

1. Odd Extension:

First, extend $f(x)$ to $[-\pi, \pi]$ by requiring that $f(x)$ is an odd function. This requirement forces $f(-x) = -f(x)$ for each $x \in [-\pi, \pi]$. Next, we extend this $f(x)$ which has been now defined on $[-\pi, \pi]$ to $\mathbb{R}$ with periodicity 2π.

The Fourier series expansion of this extended $f(x)$ is a sine series, whose coefficients are given by

$$b_n = \frac{2}{\pi}\int_0^{\pi} f(t)\sin nt\,dt = \frac{2}{\pi}\int_0^{\pi/2} t\sin nt\,dt + \frac{2}{\pi}\int_{\pi/2}^{\pi}(\pi - t)\sin nt\,dt = (-1)^{(n-1)/2}\frac{\pi}{4n^2}.$$

Thus $f(x) = \frac{\pi}{4}\Big(\frac{\sin x}{1^2} - \frac{\sin 3x}{3^2} + \frac{\sin 5x}{5^2} - \cdots\Big)$.
In this case, we say that the fourier series is a **sine series expansion** of f(x).

2. Even Extension:

First, extend $f(x)$ to $[-\pi, \pi]$ by requiring that $f(x)$ is an even function. This requirement forces $f(-x) = f(x)$ for each $x \in [-\pi, \pi]$. Next, we extend this $f(x)$ which has been now defined on $[-\pi, \pi]$ to $\mathbb{R}$ with periodicity 2π.

The Fourier series expansion of this extended $f(x)$ is a cosine series, whose coefficients are

$$a_n = \frac{2}{\pi}\int_0^{\pi} f(t)\cos nt\,dt = \frac{2}{\pi}\int_0^{\pi/2} t\cos nt\,dt + \frac{2}{\pi}\int_{\pi/2}^{\pi}(\pi - t)\cos nt\,dt = -\frac{2}{n^2\pi} \quad \text{for } 4 \nmid n.$$

And $a_0 = \pi/4,\ a_{4k} = 0$. Thus $f(x) = \frac{\pi}{4} - \frac{2}{\pi}\Big(\frac{\cos 2x}{1^2} + \frac{\cos 6x}{3^2} + \frac{\sin 10x}{5^2} + \cdots\Big)$.
In this case, we say that the fourier series is a **cosine series expansion** of f(x).

3. Scaling to length 2π:

We define a bijection $g : [-\pi, \pi] \to [0, \pi]$. Then consider the composition $h = (f \circ g) : [-\pi, \pi] \to \mathbb{R}$. We find the Fourier series for $h(y)$ and then resubstitute $y = g^{-1}(x)$ for obtaining Fourier series for $f(x)$. Notice that in computing the Fourier series for $h(y)$, we must extend $h(y)$ to $\mathbb{R}$ using periodicity of period 2π and $h(-\pi) = h(\pi)$.

In this approach, we consider

$$x = g(y) = \frac{1}{2}(y + \pi), \quad h(y) = f\Big(\frac{y+\pi}{2}\Big) = \begin{cases} \frac{1}{2}(y+\pi) & \text{if } -\pi \le y \le 0 \\ \frac{1}{2}(3\pi - y) & \text{if } 0 \le y \le \pi. \end{cases}$$

The Fourier coefficients are as follows:

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{0} \frac{t+\pi}{2}\, dt + \frac{1}{\pi}\int_{0}^{\pi} \frac{3\pi - t}{2}\, dt = 2\pi.$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{0} \frac{t+\pi}{2}\cos nt\, dt + \frac{1}{\pi}\int_{0}^{\pi} \frac{3\pi - t}{2}\cos nt\, dt = \pi n^2(1-(-1)^n) = \begin{cases} \frac{2}{\pi n^2} & \text{if } n \text{ odd} \\ 0 & \text{if } n \text{ even.} \end{cases}$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{0} \frac{t+\pi}{2}\sin nt\, dt + \frac{1}{\pi}\int_{0}^{\pi} \frac{3\pi - t}{2}\sin nt\, dt = \frac{1}{2n}[2(-1)^n - 1 + 3 - 2(-1)^n] = \frac{1}{n}.$$

Then the Fourier series for $h(y)$ is given by

$$\pi + \sum_{n=1}^{\infty} \Big(a_n \cos ny + \frac{1}{n}\sin ny\Big).$$

Using $y = g^{-1}(x) = 2x - \pi$, we have the Fourier series for $f(x)$ as

$$\pi + \sum_{n=1}^{\infty} \Big(a_n \cos n(2x-\pi) + \frac{1}{n}\sin n(2x-\pi)\Big).$$

Notice that this is neither a cosine series nor a sine series.

This example suggests three ways of construction of Fourier series for a function $f(x)$, which might have been defined on any arbitrary interval $[a, b]$.

The **first approach** says that we define a function $g(y) : [0, \pi] \to [a, b]$ and consider the composition $f \circ g$. Now, $f \circ g : [0, \pi] \to \mathbb{R}$. Next, we take an odd extension of $f \circ g$ with periodicity 2π; and call this extended function as h. We then construct the Fourier series for $h(y)$. Finally, substitute $y = g^{-1}(x)$. This will give a sine series.

In the **second approach**, we define $g(y)$ as in the first approach and take an even extension of $f \circ g$ with periodicity 2π; and call this extended function as h. We then construct the Fourier series for $h(y)$. Finally, substitute $y = g^{-1}(x)$. This gives a cosine series.

These two approaches lead to the so-called **half range Fourier series** expansions.

In the **third approach**, we define a function $g(y) : [-\pi, \pi] \to [a, b]$ and consider the composition $f \circ g$. Now, $f \circ g : [-\pi, \pi] \to \mathbb{R}$. Next, we extend $f \circ g$ with periodicity 2π;

and call this extended function as h. We then construct the Fourier series for $h(y)$. Finally, substitute $y = g^{-1}(x)$. This may give a general Fourier series involving both sine and cosine terms.

In particular, a function $f : [-\ell, \ell] \to \mathbb{R}$ which is known to have period 2ℓ can easily be expanded in a Fourier series by considering the new function $g(x) = f(\ell x/\pi)$. Now, $g : [-\pi, \pi] \to \mathbb{R}$ has period 2π. We construct a Fourier series of $g(x)$ and then substitute x with $\pi x/\ell$ to obtain a Fourier series of $f(x)$. This is the reason the third method above is called **scaling**.

In this case, the Fourier coefficients are given by

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f\Big(\frac{\ell}{\pi}t\Big)\cos nt\,dt, \quad b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f\Big(\frac{\ell}{\pi}t\Big)\cos nt\,dt$$

Substituting $s = \frac{\ell}{\pi}t, \quad dt = \frac{\pi}{\ell}ds$, we have

$$a_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(s)\cos ns\,ds, \quad b_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(s)\cos ns\,ds$$

And the Fourier series for $f(x)$ is then, with the original variable x,

$$\frac{a_0}{2} + \sum_{n=1}^{\infty}\Big(a_n\cos\frac{n\pi}{\ell}x + b_n\sin\frac{n\pi}{\ell}x\Big).$$

Example 2.12. Construct the Fourier series for $f(x) = |x|$ which is defined on $[-\ell, \ell]$ for some $\ell > 0$, and then extended to $\mathbb{R}$ with period 2ℓ.

Notice that the function $f : \mathbb{R} \to \mathbb{R}$ is not $|x|$; it is $|x|$ on $[-\ell, \ell]$. Due to its period as 2ℓ, it is $|x - 2\ell|$ on $[\ell, 3\ell]$ etc. However, it is an even function; so its Fourier series is a cosine function.

The Fourier coefficients are

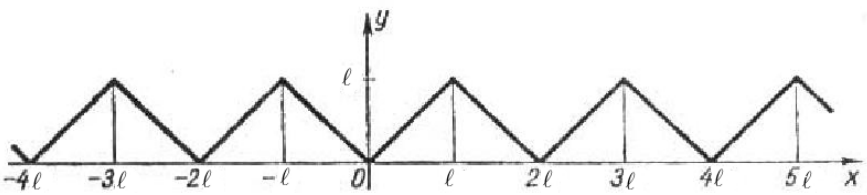
$$b_n = 0, \quad a_0 = \frac{1}{\ell}\int_{-\ell}^{\ell}|s|\,ds = \frac{2}{\ell}\int_0^{\ell} s\,ds = \ell,$$

$$a_n = \frac{2}{\ell}\int_0^{\ell} s\cos\Big(\frac{n\pi s}{\ell}\Big)\,ds = \begin{cases} 0 & \text{if } n \text{ even} \\ -\frac{4\ell}{n^2\pi^2} & \text{if } n \text{ odd} \end{cases}$$

Therefore the Fourier series for $f(x)$ shows that in $[-\ell, \ell]$,

$$|x| = \frac{\ell}{2} - \frac{4\ell}{\pi^2}\left[\frac{\cos(\pi/\ell)x}{1} + \frac{\cos(3\pi/\ell)x}{3^2} + \cdots + \frac{\cos((2n+1)\pi/\ell)x}{(2n+1)^2} + \cdots\right].$$

As our extension of $f(x)$ to $\mathbb{R}$ shows, the above Fourier series represents the function given in the following figure:



A Fun Problem: Show that the nth partial sum of the Fourier series for $f(x)$ can be written as the following integral:

$$s_n(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x+t)\frac{\sin(2n+1)t/2}{2\sin t/2}\,dt.$$

We know that $s_n(x) = \dfrac{a_0}{2} + \displaystyle\sum_{k=1}^{n}(a_k\cos kx + b_k \sin kx)$, where

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos kt\,dt, \quad b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin kt\,dt$$

Substituting these values in the expression for $s_n(x)$, we have

$$\begin{aligned}
s_n(x) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\,dt + \frac{1}{\pi}\sum_{k=1}^{n}\Big[\int_{-\pi}^{\pi} f(t)\cos kx\,\cos kt\,dt + \int_{-\pi}^{\pi} f(t)\sin kx\,\sin kt\,dt\Big] \\
&= \frac{1}{\pi}\int_{-\pi}^{\pi}\Big[\frac{f(t)}{2} + \sum_{k=1}^{n}\{f(t)\cos kx\,\cos kt + f(t)\sin kx\,\sin kt\}\Big]\,dt \\
&= \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\Big[\frac{1}{2} + \sum_{k=1}^{n}\cos k(t-x)\Big]\,dt := \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sigma_n(t-x)\,dt.
\end{aligned}$$

The expression $\sigma_n(z)$ for $z = t - x$ can be re-written as follows:

$$\sigma_n(z) = \frac{1}{2} + \cos z + \cos 2z + \cdots + \cos nz$$

Thus

$$\begin{aligned}
2\sigma_n(z)\cos z &= \cos z + 2\cos z\cos z + 2\cos z\cos 2z + \cdots + 2\cos z\cos nz \\
&= \cos z + [1+\cos 2z] + [\cos z + \cos 3z] + \cdots + [\cos(n-1)z + \cos(n+1)z] \\
&= 1 + 2\cos z + 2\cos 2z + \cdots + 2\cos(n-1)z + 2\cos nz + 2\cos(n+1)z \\
&= 2\sigma_n(z) - \cos nz + \cos(n+1)z
\end{aligned}$$

This gives

$$\sigma_n(z) = \frac{\cos nz - \cos(n+1)z}{2(1-\cos z)} = \frac{\sin(2n+1)z/2}{2\sin z/2}$$

Therefore, substituting $\sigma_n(z)$ with $z = t - x$, we have

$$s_n(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\frac{\sin(2n+1)(t-x)/2}{2\sin(t-x)/2}\,dt$$

Since the integrand is periodic of period 2π, the value of the integral remains same on any interval of length 2π. Thus

$$s_n(x) = \frac{1}{\pi}\int_{x-\pi}^{x+\pi} f(t)\frac{\sin(2n+1)(t-x)/2}{2\sin(t-x)/2}\,dt$$

Introduce a new variable $y = t - x$, i.e., $t = x + y$. And then write the integral in terms of t instead of y to obtain

$$s_n(x) = \int_{-\pi}^{\pi} f(x+y) \frac{\sin(2n+1)y/2}{2\sin y/2}\, dy = \int_{-\pi}^{\pi} f(x+t) \frac{\sin(2n+1)t/2}{2\sin t/2}\, dt$$

This integral is called the **Dirichlet Integral**. In particular, taking $f(x) = 1$, we see that $a_0 = 2$, $a_k = 0$ and $b_k = 0$ for $k \in \mathbb{N}$; and then we get the identity

$$\frac{1}{\pi} \int_{-\pi}^{\pi} \frac{\sin\big((2n+1)t/2\big)}{2\sin\big(t/2\big)}\, dt = 1 \quad \text{for each} \quad n \in \mathbb{N}.$$

Part II

Matrices

Chapter 3

Matrix Operations

3.1 Preliminary matrix operations

A **matrix** is a rectangular array of symbols. For us these symbols are real numbers or, in general, complex numbers. The individual numbers in the array are called the **entries** of the matrix. The number of rows and the number of columns in any matrix are necessarily positive integers. A matrix with m rows and n columns is called an $m \times n$ matrix and it may be written as

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix},$$

or as $A = [a_{ij}]$ for short with $a_{ij} \in \mathbb{F}$ for $i = 1, \ldots, m] \, j = 1, \ldots, n$. The number a_{ij} which occurs at the entry in ith row and jth column is referred to as the (ij)th entry (sometimes as (i, j)-th entry) of the matrix $[a_{ij}]$.

As usual, $\mathbb{R}$ denotes the set of all real numbers and $\mathbb{C}$ denotes the set of all complex numbers. We will write $\mathbb{F}$ for either $\mathbb{R}$ or $\mathbb{C}$. The numbers in $\mathbb{F}$ will also be referred to as **scalars**. Thus each entry of a matrix is a scalar.

Any matrix with m rows and n columns will be referred as an $m \times n$ matrix. The set of all $m \times n$ matrices with entries from $\mathbb{F}$ will be denoted by $\mathbb{F}^{m \times n}$.

A **row vector** of size n is a matrix in $\mathbb{F}^{1 \times n}$. A typical row vector is written as $[a_1 \quad \cdots \quad a_n]$. Similarly, a **column vector** of size n is a matrix in $\mathbb{F}^{n \times 1}$. A typical column vector is written as

$$\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \quad \text{or as} \quad [a_1 \quad \cdots \quad a_n]^T$$

for saving vertical space. We will write both $\mathbb{F}^{1 \times n}$ and $\mathbb{F}^{n \times 1}$ as $\mathbb{F}^n$. The vectors in $\mathbb{F}^n$ will be written as

$$(a_1, \ldots, a_n).$$

Sometimes, we will write a row vector as $(a_1, \ldots, a_n)$ and a column vector as $(a_1, \ldots, a_m)^T$ also.

Any matrix in $\mathbb{F}^{m\times n}$ is said to have its **size** as $m \times n$. If $m = n$, the rectangular array becomes a square array with n rows and n columns; and the matrix is called a square matrix of **order** n.

Two matrices of the same size are considered **equal** when their corresponding entries coincide, i.e., if $A = [a_{ij}]$ and $B = [b_{ij}]$ are in $\in \mathbb{F}^{m\times n}$, then

$$A = B \quad \text{iff} \quad a_{ij} = b_{ij}$$

for each $i \in \{1, \ldots, m\}$ and for each $j \in \{1, \ldots, n\}$. Thus matrices of different sizes are unequal.

Sum of two matrices of the same size is a matrix whose entries are obtained by adding the corresponding entries in the given two matrices. That is, if $A = [a_{ij}]$ and $B = [b_{ij}]$ are in $\in \mathbb{F}^{m\times n}$, then

$$A + B = [a_{ij} + b_{ij}] \in \mathbb{F}^{m\times n}.$$

For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{bmatrix} + \begin{bmatrix} 3 & 1 & 2 \\ 2 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 3 & 5 \\ 4 & 4 & 4 \end{bmatrix}.$$

We informally say that matrices are added entry-wise. Matrices of different sizes can never be added.

It then follows that

$$A + B = B + A.$$

Similarly, matrices can be **multiplied by a scalar** entry-wise. If $A = [a_{ij}] \in \mathbb{F}^{m\times n}$, and $\alpha \in \mathbb{F}$, then

$$\alpha\, A = [\alpha\, a_{ij}] \in \mathbb{F}^{m\times n}.$$

We write the **zero matrix** in $\mathbb{F}^{m\times n}$, all entries of which are 0, as 0. Thus,

$$A + 0 = 0 + A = A$$

for all matrices $A \in \mathbb{F}^{m\times n}$, with an implicit understanding that $0 \in \mathbb{F}^{m\times n}$. For $A = [a_{ij}]$, the matrix $-A \in \mathbb{F}^{m\times n}$ is taken as one whose (ij)th entry is $-a_{ij}$. Thus

$$-A = (-1)A \quad \text{and} \quad A + (-A) = -A + A = 0.$$

We also abbreviate $A + (-B)$ to $A - B$, as usual.

For example,

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

The addition and scalar multiplication as defined above satisfy the following properties:
Let $A, B, C \in \mathbb{F}^{m\times n}$. Let $\alpha, \beta \in \mathbb{F}$.

1. $A + B = B + A$.
2. $(A + B) + C = A + (B + C)$.
3. $A + 0 = 0 + A = A$.
4. $A + (-A) = (-A) + A = 0$.
5. $\alpha(\beta A) = (\alpha\beta)A$.
6. $\alpha(A + B) = \alpha A + \alpha B$.
7. $(\alpha + \beta)A = \alpha A + \beta A$.
8. $1\,A = A$.

Notice that whatever we discuss here for matrices apply to row vectors and column vectors, in particular. But remember that a row vector cannot be added to a column vector unless both are of size 1×1.

Another operation that we have on matrices is **multiplication of matrices**, which is a bit involved. Let $A = [a_{ik}] \in \mathbb{F}^{m\times n}$ and $B = [b_{kj}] \in \mathbb{F}^{n\times r}$. Then their **product** AB is a matrix $[c_{ij}] \in \mathbb{F}^{m\times r}$, where the entries are

$$c_{ij} = a_{i1}b_{1j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj}.$$

Notice that the matrix product AB is defined only when the number of columns in A is equal to the number of rows in B. Now, look back at the compact way of writing a linear system. Does it make sense?

A particular case might be helpful. Suppose A is a row vector in $\mathbb{F}^{1\times n}$ and B is a column vector in $\mathbb{F}^{n\times 1}$. Then their product $AB \in \mathbb{F}^{1\times 1}$; it is a matrix of size 1×1. Often we will identify such matrices with numbers. The product now looks like:

$$\begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} a_1b_1 + \cdots + a_nb_n \end{bmatrix}$$

This is helpful in visualizing the general case, which looks like

$$\begin{bmatrix} a_{11} & & a_{1k} & & a_{1n} \\ a_{i1} & \cdots & a_{ik} & \cdots & a_{in} \\ a_{m1} & & a_{mk} & & a_{mn} \end{bmatrix} \begin{bmatrix} b_{11} & b_{1j} & b_{1r} \\ & \vdots & \\ b_{\ell 1} & b_{\ell j} & b_{\ell r} \\ & \vdots & \\ b_{n1} & b_{nj} & b_{nr} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{1j} & c_{1r} \\ c_{i1} & c_{ij} & c_{ir} \\ c_{m1} & c_{mj} & c_{mr} \end{bmatrix}$$

The ith row of A multiplied with the jth column of B gives the (ij)th entry in AB. Thus to get AB, you have to multiply all m rows of A with all r columns of B.

For example,

$$\begin{bmatrix} 3 & 5 & -1 \\ 4 & 0 & 2 \\ -6 & -3 & 2 \end{bmatrix} \begin{bmatrix} 2 & -2 & 3 & 1 \\ 5 & 0 & 7 & 8 \\ 9 & -4 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 22 & -2 & 43 & 42 \\ 26 & -16 & 14 & 6 \\ -9 & 4 & -37 & -28 \end{bmatrix}.$$

If $u \in \mathbb{F}^{1\times n}$ and $v \in \mathbb{F}^{n\times 1}$, then $uv \in \mathbb{F}^{1\times 1}$, which we identify with a scalar; but $vu \in \mathbb{F}^{n\times n}$.

$$\begin{bmatrix} 3 & 6 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 19 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} \begin{bmatrix} 3 & 6 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 1 \\ 6 & 12 & 2 \\ 12 & 24 & 4 \end{bmatrix}.$$

It shows clearly that matrix multiplication is not commutative. Commutativity can break down due to various reasons. First of all when AB is defined, BA may not be defined. Secondly, even when both AB and BA are defined, they may not be of the same size; and thirdly, even when they are of the same size, they need not be equal. For example,

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 7 \\ 6 & 11 \end{bmatrix} \quad \text{but} \quad \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 8 & 13 \end{bmatrix}.$$

It does not mean that AB is never equal to BA. There can be some particular matrices A and B both in $\mathbb{F}^{n\times n}$ such that $AB = BA$. An extreme case is $A\,I = I\,A$, where I is the **identity matrix** defined by $I = [\delta_{ij}]$, where Kronecker's delta is defined as follows:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \qquad \text{for } i, j \in \mathbb{N}.$$

In fact, I serves as the identity of multiplication. I looks like

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ & & \vdots & & \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}.$$

We often do not write the zero entries for better visibility of some pattern.

Unlike numbers, product of two nonzero matrices can be a zero matrix. For example,

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

It is easy to verify the following properties of matrix multiplication:

1. If $A \in \mathbb{F}^{m\times n}$, $B \in \mathbb{F}^{n\times r}$ and $C \in \mathbb{F}^{r\times p}$, then $(AB)C = A(BC)$.
2. If $A, B \in \mathbb{F}^{m\times n}$ and $C \in \mathbb{F}^{n\times r}$, then $(A + B)C = AB + AC$.
3. If $A \in \mathbb{F}^{m\times n}$ and $B, C \in \mathbb{F}^{n\times r}$, then $A(B + C) = AB + AC$.

4. If $\alpha \in \mathbb{F},\ A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{n\times r}$, then $\alpha(AB) = (\alpha A)B = A(\alpha B)$.

You can see matrix multiplication in a block form. Suppose $A \in \mathbb{F}^{m\times n}$. Write its ith row as $A_{i\star}$ Also, write its kth column as $A_{\star k}$. Then we can write A as a row of columns and also as a column of rows in the following manner:

$$A = [a_{ik}] = \begin{bmatrix} A_{\star 1} & \cdots & A_{\star n} \end{bmatrix} = \begin{bmatrix} A_{1\star} \\ \vdots \\ A_{m\star} \end{bmatrix}.$$

Write $B \in \mathbb{F}^{n\times r}$ similarly as

$$B = [b_{kj}] = \begin{bmatrix} B_{\star 1} & \cdots & B_{\star r} \end{bmatrix} = \begin{bmatrix} B_{1\star} \\ \vdots \\ B_{\star n} \end{bmatrix}.$$

Then their product AB can now be written as

$$AB = \begin{bmatrix} AB_{\star 1} & \cdots & AB_{\star r} \end{bmatrix} = \begin{bmatrix} A_{1\star}B \\ \vdots \\ A_{m\star}B \end{bmatrix}.$$

When writing this way, we ignore the extra brackets [and].

Powers of square matrices can be defined inductively by taking

$$A^0 = I \quad \text{and} \quad A^n = AA^{n-1} \ \text{ for } \ n \in \mathbb{N}.$$

A square matrix A of order m is called **invertible** iff there exists a matrix B of order m such that

$$AB = I = BA.$$

Such a matrix B is called an **inverse** of A. If C is another inverse of A, then

$$C = CI = C(AB) = (CA)B = IB = B.$$

Therefore, an inverse of a matrix is unique and is denoted by A^{-1}. We talk of invertibility of square matrices only; and all square matrices are not invertible. For example, I is invertible but 0 is not. If $AB = 0$ for square matrices A and B, then neither A nor B is invertible.

It is easy to verify that if $A, B \in \mathbb{F}^{n\times n}$ are invertible matrices, then $(AB)^{-1} = B^{-1}A^{-1}$.

3.2 Transpose and adjoint

Given a matrix $A \in \mathbb{F}^{m\times n}$, its **transpose** is a matrix in $\mathbb{F}^{n\times m}$, which is denoted by A^T, and is defined by

$$\text{the } (ij)\text{th entry of } A^T = \text{the } (ji)\text{th entry of} A.$$

That is, the ith column of A^T is the column vector $(a_{i1}, \cdots, a_{in})^T$. The rows of A are the columns of A^T and the columns of A become the rows of A^T. In particular, if $u = [a_1 \ \cdots \ a_m]$ is a row vector, then its transpose is

$$u^T = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix},$$

which is a column vector. Similarly, the transpose of a column vector is a row vector. Notice that the transpose notation goes well with our style of writing a column vector as the transpose of a row vector. If you write A as a row of column vectors, then you can express A^T as a column of row vectors, as in the following:

$$A = \begin{bmatrix} A_{\star 1} & \cdots & A_{\star n} \end{bmatrix} \Rightarrow A^T = \begin{bmatrix} A_{\star 1}^T \\ \vdots \\ A_{\star n}^T \end{bmatrix}.$$

$$A = \begin{bmatrix} A_{1\star} \\ \vdots \\ A_{m\star} \end{bmatrix} \Rightarrow A^T = \begin{bmatrix} A_{1\star}^T & \cdots & A_{m\star}^T \end{bmatrix}.$$

For example,

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{bmatrix} \Rightarrow A^T = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 1 \end{bmatrix}.$$

The following are some of the properties of this operation of transpose.

1. $(A^T)^T = A$.
2. $(A + B)^T = A^T + B^T$.
3. $(\alpha A)^T = \alpha A^T$.
4. $(AB)^T = B^T A^T$.
5. If A is invertible, then A^T is invertible, and $(A^T)^{-1} = (A^{-1})^T$.

In the above properties, we assume that the operations are allowed, that is, in (2), A and B must be of the same size. Similarly, in (4), the number of columns in A must be equal to the number of rows in B; and in (5), A must be a square matrix.

It is easy to see all the above properties, except perhaps the fourth one. For this, let $A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{n\times r}$. Now, the (ji)th entry in $(AB)^T$ is the (ij)th entry in AB; and it is given by

$$a_{i1}b_{j1} + \cdots + a_{in}b_{jn}.$$

On the other side, the (ji)th entry in B^TA^T is obtained by multiplying the jth row of B^T with the ith column of A^T. This is same as multiplying the entries in the jth column of B with the corresponding entries in the ith row of A, and then taking the sum. Thus it is

$$b_{j1}a_{i1} + \cdots + b_{jn}a_{in}.$$

This is the same as computed earlier.

Close to the operations of transpose of a matrix is the adjoint. Let $A = [a_{ij}] \in \mathbb{F}^{m\times n}$. The **adjoint** of A is denoted as A^*, and is defined by

$$\text{the } (ij)\text{th entry of } A^* = \text{the complex conjugate of } (ji)\text{th entry of} A.$$

We write $\overline{\alpha}$ for the **complex conjugate** of a scalar α. That is, $\overline{\alpha + i\beta} = \alpha - i\beta$. Thus, if $a_{ij} \in \mathbb{R}$, then $\overline{a}_{ij} = a_{ij}$. When A has only real entries, $A^* = A^T$. The ith column of A^* is the column vector $(\overline{a}_{i1}, \cdots, \overline{a}_{in})^T$. For example,

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{bmatrix} \Rightarrow A^* = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 1 \end{bmatrix}.$$

$$A = \begin{bmatrix} 1+i & 2 & 3 \\ 2 & 3 & 1-i \end{bmatrix} \Rightarrow A^* = \begin{bmatrix} 1-i & 2 \\ 2 & 3 \\ 3 & 1+i \end{bmatrix}.$$

Similar to the transpose, the adjoint satisfies the following properties:

1. $(A^*)^* = A$.
2. $(A + B)^* = A^* + B^*$.
3. $(\alpha A)^* = \overline{\alpha}A^*$.
4. $(AB)^* = B^*A^*$.
5. If A is invertible, then A^* is invertible, and $(A^*)^{-1} = (A^{-1})^*$.

Here also, in (2), the matrices A and B must be of the same size, and in (4), the number of columns in A must be equal to the number of rows in B. The adjoint of A is also called the **conjugate transpose** of A.

Further, the familiar dot product in $\mathbb{R}^{1\times 3}$ can be generalized to $\mathbb{F}^{1\times n}$ or to $\mathbb{F}^{n\times 1}$. The generalization can be given via matrix product. For vectors $u, v \in \mathbb{F}^{1\times n}$, we define their **inner product** as

$$\langle u, v\rangle = uv^*.$$

For example,

$$u = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}, \quad v = \begin{bmatrix} 2 & 1 & 3 \end{bmatrix} \quad \Rightarrow \quad \langle u, v\rangle = 1 \times 2 + 2 \times 1 + 3 \times 3 = 13.$$

Similarly, for $x, y \in \mathbb{F}^{n\times 1}$, we define their inner product as

$$\langle x, y\rangle = y^* x.$$

In case, $\mathbb{F} = \mathbb{R}$, the x^* becomes x^T. The inner product satisfies the following properties:

For $x, y, z \in \mathbb{F}^n$ and $\alpha, \beta \in \mathbb{F}$,

1. $\langle x, x\rangle \geq 0$.
2. $\langle x, x\rangle = 0$ iff $x = 0$.
3. $\langle x, y\rangle = \overline{\langle y, x\rangle}$.
4. $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$.
5. $\langle z, x + y\rangle = \langle z, x\rangle + \langle z, y\rangle$.
6. $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$.
7. $\langle x, \beta y\rangle = \overline{\beta}\langle x, y\rangle$.

The inner product gives rise to the length of a vector as in the familiar case of $\mathbb{R}^{1\times 3}$. We now call the generalized version of length as the *norm*. If $u \in \mathbb{F}^n$, we define its **norm**, denoted by $\|u\|$ as the nonnegative square root of $\langle u, u\rangle$. That is,

$$\|u\| = \sqrt{\langle u, u\rangle}.$$

The norm satisfies the following properties:

For $x, y \in \mathbb{F}^n$ and $\alpha \in \mathbb{F}$,

1. $\|x\| \geq 0$.
2. $\|x\| = 0$ iff $x = 0$.
3. $\|\alpha x\| = |\alpha|\, \|x\|$.
4. $|\langle x, y\rangle| \leq \|x\|\, \|y\|$. (*Cauchy-Schwartz inequality*)
5. $\|x + y\| = \|x\| + \|y\|$. (*Triangle inequality*)

Using these properties, the acute (non-obtuse) angle between any to vectors can be defined. Let $x, y \in \mathbb{F}^n$. The acute **angle** θ **between** x and y, denoted by $\theta(x, y)$ is defined by

$$\cos\theta(x, y) = \frac{|\langle x, y\rangle|}{\|x\|\, \|y\|}.$$

In particular, when $\theta(x, y) = \pi/2$, we say that the vectors x and y are **orthogonal**, and we write this as $x \perp y$. That is,

$$x \perp y \quad \text{iff} \quad \langle x, y\rangle = 0.$$

It follows that if $x \perp y$, then $\|x\|^2 + \|y\|^2 = \|x + y\|^2$. This is referred to as **Pythagoras law**. The converse of Pythagoras law holds when $\mathbb{F} = \mathbb{R}$. For $\mathbb{F} = \mathbb{C}$, it does not hold, in general.

Adjoints of matrices behave in a very predictable way with the inner product.

Theorem 3.1. *Let $A \in \mathbb{F}^{m\times n}$; $x \in \mathbb{F}^{n\times 1}$; $y \in \mathbb{F}^{m\times 1}$. Then*

$$\langle Ax, y\rangle = \langle x, A^*y\rangle \quad \textit{and} \quad \langle A^*y, x\rangle = \langle y, Ax\rangle.$$

Proof: Recall that in $\mathbb{F}^{r\times 1}$, $\langle u, v\rangle = v^*u$. Further, $Ax \in \mathbb{F}^{m\times 1}$ and $A^*y \in \mathbb{F}^{n\times 1}$. We are using the same notation for both the inner products in $\mathbb{F}^{m\times 1}$ and in $\mathbb{F}^{n\times 1}$. We then have

$$\langle Ax, y\rangle = y^*Ax = (A^*y)^*x = \langle x, A^*y\rangle.$$

The second equality follows from the first. □

Often the definition of an adjoint is taken using this identity:

$$\langle Ax, y\rangle = \langle x, A^*y\rangle.$$

3.3 Special types of matrices

Recall that the *zero matrix* is a matrix each entry of which is 0. We write 0 for all zero matrices of all sizes. The size is to be understood from the context.

Let $A = [a_{ij}] \in \mathbb{F}^{n\times n}$ be a square matrix of order n. The entries a_{ii} are called as the **diagonal entries** of A. The first diagonal diagonal entry is a_{11}, and the last diagonal entry is a_{nn}. The entries of A, which are not the diagonal entries, are called as **off-diagonal entries** of A; they are a_{ij} for $i \neq j$. In the following matrix, the diagonal entries are shown in red:

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

Here, 1 is the first diagonal entry, 3 is the second diagonal entry and 0 is the third and the last diagonal entry.

If all off-diagonal entries of A are 0, then A is said to be a **diagonal matrix**. Only a square matrix can be a diagonal matrix. There is a way to generalize this notion to any matrix, but we do not require it. Notice that the diagonal entries in a diagonal matrix need not all be nonzero. For example, the zero matrix of order n is also a diagonal matrix. The following is a diagonal matrix. We follow the convention of not showing the off-diagonal entries in a diagonal matrix.

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

We also write a diagonal matrix with diagonal entries $d_1, \dots, d_n$ as $\operatorname{diag}(d_1, \dots, d_n)$. Thus the above diagonal matrix is also written as

$$\operatorname{diag}(1, 3, 0).$$

The *identity matrix* is a square matrix of which each diagonal entry is 1 and each off-diagonal entry is 0. Obviously,

$$I^T = I^* = I^{-1} = \operatorname{diag}(1, \ldots, 1) = I.$$

When identity matrices of different orders are used in a context, we will use the notation I_m for the identity matrix of order m. If $A \in \mathbb{F}^{m\times n}$, then $AI_n = A$ and $I_m A = A$.

We write e_i for a column vector whose ith component is 1 and all other components 0. That is, the jth component of e_i is δ_{ij}. In $\mathbb{F}^{n\times 1}$, there are then n distinct column vectors

$$e_1, \ldots, e_n.$$

The e_is are referred to as the **standard basis vectors**. These are the columns of the identity matrix of order n, in that order; that is, e_i is the ith column of I. The transposes of these e_is are the rows of I.
That is, the ith row of I is e_i^T. Thus

$$I = \begin{bmatrix} e_1 & \cdots & e_n \end{bmatrix} = \begin{bmatrix} e_1^T \\ \vdots \\ e_n^T \end{bmatrix}.$$

A **scalar matrix** is a matrix of which each diagonal entry is a scalar, the same scalar, and each off-diagonal entry is 0. That is, a scalar matrix is of the form αI, for some scalar α. The following is a scalar matrix:

$$\begin{bmatrix} 3 & & & \\ & 3 & & \\ & & 3 & \\ & & & 3 \end{bmatrix}.$$

It is also written as $\operatorname{diag}(3, 3, 3, 3)$. If $A, B \in \mathbb{F}^{m\times m}$ and A is a scalar matrix, then $AB = BA$. Conversely, if $A \in \mathbb{F}^{m\times m}$ is such that $AB = BA$ for all $B \in \mathbb{F}^{m\times m}$, then A must be a scalar matrix. This fact is not obvious, and you should try to prove it.

A matrix $A \in \mathbb{F}^{m\times n}$ is said to be **upper triangular** iff all entries below the diagonal are zero. That is, $A = [a_{ij}]$ is upper triangular when $a_{ij} = 0$ for $i > j$. In writing such a matrix, we simply do not show the zero entries below the diagonal. Similarly, a matrix is called **lower triangular** iff all its entries above the diagonal are zero. Both upper triangular and lower triangular matrices are referred to as **triangular** matrices. A diagonal matrix is both upper triangular and lower triangular. Transpose of a lower triangular matrix is an upper triangular matrix and vice versa. In the following, L is a lower triangular matrix, and U is an upper triangular matrix, both of order 3.

$$L = \begin{bmatrix} 1 & & \\ 2 & 3 & \\ 3 & 4 & 5 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 2 & 3 \\ & 3 & 4 \\ & & 5 \end{bmatrix}.$$

A square matrix A is called **hermitian**, iff $A^* = A$. And A is called **skew hermitian** iff $A^* = -A$. A hermitian matrix with real entries satisfies $A^T = A$; and accordingly, such a

matrix is called a **real symmetric** matrix. In general, A is called a **symmetric** matrix iff $A^T = A$. We also say that a matrix is **skew symmetric** iff $A^T = -A$. In the following, B is symmetric, C is skew-symmetric, D is hermitian, and E is skew-hermitian. B is also hermitian and C is also skew-hermitian.

$$B = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 2 & -3 \\ -2 & 0 & 4 \\ 3 & -4 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} i & 2i & 3 \\ 2i & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix}, \quad E = \begin{bmatrix} 0 & 2+i & 3 \\ 2-i & i & 4i \\ 3 & -4i & 0 \end{bmatrix}$$

Notice that a skew-symmetric matrix must have a zero diagonal, and the diagonal entries of a skew-hermitian matrix must be 0 or purely imaginary. Reason:

$$a_{ii} = -\overline{a}_{ii} \Rightarrow 2\mathrm{Re}(a_{ii}) = 0.$$

Let A be a square matrix. Since $A + A^T$ is symmetric and $A - A^T$ is skew symmetric, every square matrix can be written as a sum of a symmetric matrix and a skew symmetric matrix:

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

Similar rewriting is possible with hermitian and skew hermitian matrices:

$$A = \frac{1}{2}(A + A^*) + \frac{1}{2}(A - A^*).$$

A square matrix A is called **unitary** iff $A^*A = I = AA^*$. In addition, if A is real, then it is called an orthogonal matrix. That is, an **orthogonal matrix** is a matrix with real entries satisfying $A^TA = I = AA^T$. Notice that a square matrix is unitary iff it is invertible and its inverse is equal to its adjoint. Similarly, a real matrix is orthogonal iff it is invertible and its inverse is its transpose. In the following, B is a unitary matrix of order 2, and C is an orthogonal matrix (also unitary) of order 3:

$$B = \frac{1}{2}\begin{bmatrix} 1+i & 1-i \\ 1-i & 1+i \end{bmatrix}, \quad C = \frac{1}{3}\begin{bmatrix} 2 & 1 & 2 \\ -2 & 2 & 1 \\ 1 & 2 & -2 \end{bmatrix}.$$

The following are examples of orthogonal 2×2 matrices:

$$O_1 := \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \quad O_2 := \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}.$$

O_1 is said to be a *rotation by an angle* θ and O_2 is called a *reflection by an angle* $\theta/2$ along the x-axis. Can you say why are they so called?

A square matrix A is called **normal** iff $A^*A = AA^*$. All hermitian matrices and all real symmetric matrices are normal matrices. For example,

$$\begin{bmatrix} 1+i & 1+i \\ -1-i & 1+i \end{bmatrix}$$

is a normal matrix; verify this. Also see that this matrix is neither hermitian nor skew-hermitian. In fact, a matrix is normal iff it is in the form $B + iC$, where B, C are hermitian and $BC = CB$. Can you prove this fact?

3.4 Elementary row operations

There are three kinds of **Elementary Row Operations** for a matrix $A \in \mathbb{F}^{m\times n}$:

$ER1$. Exchange of two rows.

$ER2$. Multiplication of a row by a nonzero constant.

$ER3$. Replacing a row by sum of that row with a nonzero constant multiple of another row.

Example 3.1. You must find out how exactly the operations have been applied in the following computation:

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 3 & 3 \end{bmatrix} \xrightarrow{ER3} \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{ER3} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We define the **elementary matrices** as in the following, which will help us seeing the elementary row operations as matrix products.

1. $E[i,j]$ is the matrix obtained from I by exchanging the ith and jth rows.
2. $E_\alpha[i]$ is the matrix obtained from I by replacing its ith row with α times the ith row.
3. $E_\alpha[i,j]$ is the matrix obtained from I by replacing its ith row by the ith row plus α times the jth row.

Let $A \in \mathbb{F}^{m\times n}$. Consider $E[i,j],\ E_\alpha[i],\ E_\alpha[i,j] \in \mathbb{F}^{m\times m}$. The following may be verified:

1. $E[i,j]A$ is the matrix obtained from A by exchanging the ith and the jth rows. It corresponds to an elementary row operation $ER1$.
2. $E_\alpha[i]A$ is the matrix obtained from A by replacing the ith row with α times the ith row. It corresponds to an elementary row operation $ER2$.
3. $E_\alpha[i,j]A$ is the matrix obtained from A by replacing the ith row with the ith row plus α times the jth row. It corresponds to an elementary row operation $ER3$.

The computation with elementary row operations in the above example can now be written as

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 3 & 3 \end{bmatrix} \xrightarrow{E_{-3}[3,1]} \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{E_{-2}[2,1]} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Often we will apply elementary operations in a sequence. In this way, the above operations could be shown in one step as $E_{-3}[3,1],\ E_{-2}[2,1]$.

3.5 Row reduced echelon form

The first nonzero entry (from left) in a nonzero row of a matrix is called a **pivot**. We denote a pivot in a row by putting a box around it. A column where a pivot occurs is called a **pivotal column**.

A matrix $A \in \mathbb{F}^{m \times n}$ is said to be in **row reduced echelon form (RREF)** iff the following conditions are satisfied:

(1) Each pivot is equal to 1.

(2) The column index of the pivot in any nonzero row R is smaller than the column index of the pivot in any row below R.

(3) In a pivotal column, all entries other than the pivot, are zero.

(4) All zero rows are at the bottom.

Example 3.2. The matrix $\begin{bmatrix} \boxed{1} & 2 & 0 & 0 \\ 0 & 0 & \boxed{1} & 0 \\ 0 & 0 & 0 & \boxed{1} \end{bmatrix}$ is in row reduced echelon form whereas the matrices

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

are not in row reduced echelon form.

Reduction to RREF

1. Set the work region R as the whole matrix A.
2. If all entries in R are 0, then stop.
3. If there are nonzero entries in R, then find the leftmost nonzero column. Mark it as the pivotal column.
4. Find the topmost nonzero entry in the pivotal column. Suppose it is α. Box it; it is a pivot.
5. If the pivot is not on the top row of R, then exchange the row of A which contains the top row of R with the row where the pivot is.
6. If $\alpha \neq 1$, then replace the top row of R in A by $1/\alpha$ times that row.
7. Make all entries, except the pivot, in the pivotal column as zero by replacing each row above and below the top row of R using elementary row operations in A with that row and the top row of R.

8. Find the sub-matrix to the right and below the pivot. If no such sub-matrix exists, then stop. Else, reset the work region R to this sub-matrix, and go to Step 2.

We will refer to the output of the above reduction algorithm as *the row reduced echelon form* or, the RREF of a given matrix.

Example 3.3.

$$A = \begin{bmatrix} \boxed{1} & 1 & 2 & 0 \\ 3 & 5 & 7 & 1 \\ 1 & 5 & 4 & 5 \\ 2 & 8 & 7 & 9 \end{bmatrix} \xrightarrow{R1} \begin{bmatrix} \boxed{1} & 1 & 2 & 0 \\ 0 & 2 & 1 & 1 \\ 0 & 4 & 2 & 5 \\ 0 & 6 & 3 & 9 \end{bmatrix} \xrightarrow{E_{1/2}[2]} \begin{bmatrix} \boxed{1} & 1 & 2 & 0 \\ 0 & \boxed{1} & \frac{1}{2} & \frac{1}{2} \\ 0 & 4 & 2 & 5 \\ 0 & 6 & 3 & 9 \end{bmatrix}$$

$$\xrightarrow{R2} \begin{bmatrix} \boxed{1} & 0 & \frac{3}{2} & -\frac{1}{2} \\ 0 & \boxed{1} & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 6 \end{bmatrix} \xrightarrow{E_{1/3}[3]} \begin{bmatrix} \boxed{1} & 0 & \frac{3}{2} & -\frac{1}{2} \\ 0 & \boxed{1} & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \boxed{1} \\ 0 & 0 & 0 & 6 \end{bmatrix} \xrightarrow{R3} \begin{bmatrix} \boxed{1} & 0 & \frac{3}{2} & 0 \\ 0 & \boxed{1} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \boxed{1} \\ 0 & 0 & 0 & 0 \end{bmatrix} = B$$

Here, $R1 = E_{-3}[2,1]$, $E_{-1}[3,1]$, $E_{-2}[4,1]$; $R2 = E_{-1}[2,1]$, $E_{-4}[3,2]$, $E_{-6}[4,2]$; and $R3 = E_{1/2}[1,3]$, $E_{-1/2}[2,3]$, $E_{-6}[4,3]$. Notice that

$$B = E_{-6}[4,3]\, E_{-1/2}[2,3]\, E_{1/2}[1,3]\, E_{-1/3}[3]\, E_{-6}[4,2]\, E_{-4}[3,2]\, E_{-1}[2,1] E_{-1/2}[2] \\ E_{-2}[4,1]\, E_{-1}[3,1]\, E_{-3}[2,1]\, A.$$

The products are in reverse order.

Theorem 3.2. *A square matrix is invertible iff it is a product of elementary matrices.*

Proof: $E[i,j]$ is its own inverse, $E_{1/\alpha}[i]$ is the inverse of $E_\alpha[i]$, and $E_{-\alpha}[i,j]$ is the inverse of $E_\alpha[i,j]$. So, product of elementary matrices is invertible.

Conversely, suppose that A is invertible. Let EA^{-1} be the RREF of A^{-1}. If EA^{-1} has a zero row, then $EA^{-1}A$ also has a zero row. That is, E has a zero row, which is impossible. So, EA^{-1} does not have a zero row. Then each row in the square matrix EA^{-1} has a pivot. But the only square matrix in RREF having a pivot at each row is the identity matrix. Therefore, $EA^{-1} = I$. That is, $A = E$, a product of elementary matrices. □

Theorem 3.3. *Let $A \in \mathbb{F}^{m\times n}$. There exists a unique matrix in $\mathbb{F}^{m\times n}$ in row reduced echelon form obtained from A by elementary row operations.*

Proof: Suppose $B, C \in \mathbb{F}^{m\times n}$ are matrices in RREF such that each has been obtained from A by elementary row operations. Observe that elementary matrices are invertible and their inverses are also elementary matrices. Then $B = E_1 A$ and $C = E_2 A$ for some invertible matrices $E_1, E_2 \in \mathbb{F}^{m\times m}$. Now, $B = E_1 A = E_1 (E_2)^{-1} C$. Write $E = E_1 (E_2)^{-1}$ to have $B = EC$, where E is invertible.

Assume, on the contrary, that $B \neq C$. Then there exists a column index, say $k \geq 1$, such that the first $k-1$ columns of B coincide with the first $k-1$ columns of C, respectively;

and the kth column of B is not equal to the kth column of C. Let u be the kth column of B, and let v be the kth column of C. We have $u = Ev$ and $u \neq v$.

Suppose the pivotal columns that appear within the first $k-1$ columns in C are $e_1, \ldots, e_j$. Then $e_1, \ldots, e_j$ are also the pivotal columns in B that appear within the first $k-1$ columns. Since $B = EC$, we have $C = E^{-1}B$; and consequently,

$$e_1 = Ee_1 = E^{-1}e_1, \ \ldots, \ e_j = Ee_j = E^{-1}e_j.$$

Since C is in RREF, either $u = e_{j+1}$ or $u = \alpha_1 e_1 + \cdots + \alpha_j e_j$ for some scalars $\alpha_1, \ \ldots, \ \alpha_j$. (See it.) The latter case includes the possibility that $u = 0$. Similarly, either $v = e_{j+1}$ or $v = \beta_1 e_1 + \cdots + \beta_j e_j$ for some scalars $\beta_1, \ \ldots, \ \beta_j$. We consider the following exhaustive cases.

If $u = e_{j+1}$ and $v = e_{j+1}$, then $u = v$.

If $u = e_{j+1}$ and $v = \beta_1 e_1 + \cdots + \beta_j e_j$, then

$$u = Ev = \beta_1 Ee_1 + \cdots + \beta_j Ee_j = \beta_1 e_1 + \cdots + \beta_j e_j = v.$$

If $u = \alpha_1 e_1 + \cdots + \alpha_j u_j$ (and whether $v = e_{j+1}$ or $v = \beta_1 e_1 + \cdots + \beta_j e_j$), then

$$v = E^{-1}u = \alpha_1 E^{-1}e_1 + \cdots + \alpha_j E^{-1}e_j = \alpha_1 e_1 + \cdots + \alpha_j e_j = u.$$

In either case, $u = v$; and this is a contradiction. Therefore, $B = C$. □

3.6 Determinant

The sum of all diagonal entries of a square matrix is called the **trace** of the matrix. That is, if $A = [a_{ij}] \in \mathbb{F}^{m\times m}$, then

$$\mathrm{tr}(A) = a_{11} + \cdots + a_{nn} = \sum_{k=1}^{n} a_{kk}.$$

In addition to $\mathrm{tr}(I_m) = m, \ \mathrm{tr}(0) = 0$, the trace satisfies the following properties:

Let $A \in \mathbb{F}^{n\times n}$. Let $\beta \in \mathbb{F}$.

1. $\mathrm{tr}(\beta A) = \beta \mathrm{tr}(A)$.
2. $\mathrm{tr}(A^T) = \mathrm{tr}(A)$ and $\mathrm{tr}(A^*) = \overline{\mathrm{tr}(A)}$.
3. $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.
4. $\mathrm{tr}(A^*A) = 0$ iff $\mathrm{tr}(AA^*) = 0$ iff $A = 0$.

(4) follows from the observation that $\mathrm{tr}(A^*A) = \sum_{i=1}^{m}\sum_{j=1}^{m} |a_{ij}|^2 = \mathrm{tr}(AA^*)$. The second quantity, called the **determinant** of a square matrix $A = [a_{ij}] \in \mathbb{F}^{n\times n}$, written as $\det(A)$, is defined inductively as follows:

If $n = 1$, then $\det(A) = a_{11}$.

If $n > 1$, then $\det(A) = \sum_{j=1}^{n} (-1)^{1+j}\, a_{1j} \det(A_{1j})$

where the matrix $A_{1j} \in \mathbb{F}^{(n-1)\times(n-1)}$ is obtained from A by deleting the first row and the jth column of A.

When $A = [a_{ij}]$ is written showing all its entries, we also write $\det(A)$ by replacing the two big closing brackets [and] by two vertical bars | and |. For a 2×2 matrix, its determinant is seen as follows:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = (-1)^{1+1} a_{11} \det[a_{22}] + (-1)^{1+2} a_{12} \det[a_{21}] = a_{11}a_{22} - a_{12}a_{21}.$$

Similarly, for a 3×3 matrix, we need to compute three 2×2 determinants. For example,

$$\begin{aligned} \det \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix} &= \begin{vmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{vmatrix} \\ &= (-1)^{1+1} \times 1 \times \begin{vmatrix} 3 & 1 \\ 1 & 2 \end{vmatrix} + (-1)^{1+2} \times 2 \times \begin{vmatrix} 2 & 1 \\ 3 & 2 \end{vmatrix} + (-1)^{1+3} \times 3 \times \begin{vmatrix} 2 & 3 \\ 3 & 1 \end{vmatrix} \\ &= 1 \times \begin{vmatrix} 3 & 1 \\ 1 & 2 \end{vmatrix} - 2 \times \begin{vmatrix} 2 & 1 \\ 3 & 2 \end{vmatrix} + 3 \times \begin{vmatrix} 2 & 3 \\ 3 & 1 \end{vmatrix} \\ &= (3 \times 2 - 1 \times 1) - 2 \times (2 \times 2 - 1 \times 3) + 3 \times (2 \times 1 - 3 \times 3) \\ &= 5 - 2 \times 1 + 3 \times (-7) = -18. \end{aligned}$$

The determinant of any triangular matrix (upper or lower), is the product of its diagonal entries. In particular, the determinant of a diagonal matrix is also the product of its diagonal entries. Thus, if I is the identity matrix of order n, then $\det(I) = 1$ and $\det(-I) = (-1)^n$.

Let $A \in \mathbb{F}^{n\times n}$. The sub-matrix of A obtained by deleting the ith row and the jth column is called the (ij)th **minor** of A, and is denoted by A_{ij}. The (ij)th **co-factor** of A is $(-1)^{i+j}\det(A_{ij})$; it is denoted by $C_{ij}(A)$. Sometimes, when the matrix A is fixed in a context, we write $C_{ij}(A)$ as C_{ij}. The **adjugate** of A is the $n \times n$ matrix obtained by taking transpose of the matrix whose (ij)th entry is $C_{ij}(A)$; it is denoted by $\operatorname{adj}(A)$. That is, $\operatorname{adj}(A) \in \mathbb{F}^{n\times n}$ is the matrix whose (ij)th entry is the (ji)th co-factor $C_{ji}(A)$. Denote by $A_j(x)$ the matrix obtained from A by replacing the jth column of A by the (new) column vector $x \in \mathbb{F}^{n\times 1}$. Some important facts about the determinant are listed below without proof.

Theorem 3.4. *Let $A \in \mathbb{F}^{n\times n}$. Let $i, j, k \in \{1, \ldots, n\}$. Then the following statements are true.*

1. $\det(A) = \sum_i a_{ij}(-1)^{i+j} \det(A_{ij}) = \sum_i a_{ij} C_{ij}(A)$ *for any fixed j.*
2. *For any $j \in \{1, \ldots, n\}$,* $\det(\, A_j(x + y)\,) = \det(\, A_j(x)\,) + \det(\, A_j(y)\,)$.
3. *For any $\alpha \in \mathbb{F}$,* $\det(\, A_j(\alpha x)\,) = \alpha \det(\, A_j(x)\,)$.
4. *For $A \in \mathbb{F}^{n\times n}$, let $B \in \mathbb{F}^{n\times n}$ be the matrix obtained from A by interchanging the jth and the kth columns, where $j \neq k$. Then* $\det(B) = -\det(A)$.

5. *If some column of A is the zero vector, then* $\det(A) = 0$.

6. *If one column of A is a scalar multiple of another column, then* $\det(A) = 0$.

7. *If a column of A is replaced by that column plus a scalar multiple of another column, then determinant does not change.*

8. *If A is a triangular matrix, then* $\det(A)$ *is equal to the product of the diagonal entries of A.*

9. $\det(AB) = \det(A)\det(B)$ *for any matrix* $B \in \mathbb{F}^{n\times n}$.

10. *If A is invertible, then* $\det(A) \neq 0$ *and* $\det(A^{-1}) = (\det(A))^{-1}$.

11. *Columns of A are linearly dependent iff* $\det(A) = 0$.

12. *If A and B are similar matrices, then* $\det(A) = \det(B)$.

13. $\det(A) = \sum_j a_{ij}(-1)^{i+j}A_{ij}$ *for any fixed i.*

14. $\operatorname{rank}(A) = n$ *iff* $\det(A) \neq 0$.

15. *All of* (2)-(7) *and* (11) *are true for rows instead of columns.*

16. $\det(A^T) = \det(A)$.

17. $A\operatorname{adj}(A) = \operatorname{adj}(A)A = \det(A)\, I$. *In particular, if* $\det(A) \neq 0$, *then A is invertible.*

The statements (10) and the particular case in (17), in the above theorem, say that a square matrix is invertible iff its determinant is nonzero.

Example 3.4.

$$\begin{vmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{vmatrix} \overset{R1}{=} \begin{vmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & -1 & 1 & 2 \\ 0 & -1 & -1 & 2 \end{vmatrix} \overset{R2}{=} \begin{vmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & -1 & 4 \end{vmatrix} \overset{R3}{=} \begin{vmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{vmatrix} = 8.$$

Here, $R1 = E_1[2,1];\ E_1[3,1];\ E_1[4,1]$, $R2 = E_1[3,2];\ E_1[4,2]$, and $R3 = E_1[4,3]$.

Finally, the upper triangular matrix has the required determinant.

Recall that a hermitian matrix is one for which its adjoint is the same as itself; and a unitary matrix is one for which its adjoint coincides with its inverse. The determinant of a hermitian matrix is a real number since $A = A^*$ implies

$$\det(A) = \det(A^*) = \det(\overline{A}) = \overline{\det(A)}.$$

Here, $\overline{A}$ denotes the matrix whose (ij)th entry is the complex conjugate of the (ij)th entry of A. For a unitary matrix, taking determinant, we see that

$$1 = \det(I) = \det(A^*A) = \det(\overline{A})\det(A) = \overline{\det(A)}\det(A) = |\det(A)|^2.$$

That is, the determinant of a unitary matrix has absolute value 1. Since orthogonal matrices are real unitary matrices, the determinant of an orthogonal matrix is also ± 1.

3.7 Computing inverse of a matrix

Look at the sequence of elementary matrices corresponding to the elementary operations used in this row reduction of A. The product of these elementary matrices is A^{-1}, since this product times A is I, which is the row reduced form of A. Now, if we use the same elementary operations on I, then the result will be $A^{-1}I = A^{-1}$. Thus we obtain a procedure to compute the inverse of a matrix A provided it is invertible.

If $A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{m\times k}$, then the matrix $[A|B] \in \mathbb{F}^{m\times(n+k)}$ obtained from A and B by writing first all the columns of A and then the columns of B, in that order, is called an **augmented matrix**.

For computing the inverse of a matrix, start with the augmented matrix $[A|I]$. Then we reduce $[A|I]$ to its RREF. If the A-portion in the RREF is I, then the I-portion in the RREF gives A^{-1}. If the A-portion in the RREF contains a zero row, then A is not invertible.

Example 3.5. For illustration, consider the following square matrices:

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

We want to find the inverses of the matrices, if at all they are invertible.

Augment A with an identity matrix to get

$$\left[\begin{array}{cccc|cccc} \boxed{1} & -1 & 2 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 2 & 0 & 1 & 0 & 0 \\ 2 & 1 & -1 & -2 & 0 & 0 & 1 & 0 \\ 1 & -2 & 4 & 2 & 0 & 0 & 0 & 1 \end{array}\right].$$

Use elementary row operations. Since $a_{11} = 1$, we leave $row(1)$ untouched. To zero-out the other entries in the first column, we use the sequence of elementary row operations $E_1[2,1]$, $E_{-2}[3,1]$, $E_{-1}[4,1]$ to obtain

$$\left[\begin{array}{cccc|cccc} \boxed{1} & -1 & 2 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 2 & 2 & 1 & 1 & 0 & 0 \\ 0 & 3 & -5 & -2 & -2 & 0 & 1 & 0 \\ 0 & -1 & 2 & 2 & -1 & 0 & 0 & 1 \end{array}\right].$$

The pivot is -1 in $(2,2)$ position. Use $E_{-1}[2]$ to make the pivot 1.

$$\left[\begin{array}{cccc|cccc} 1 & -1 & 2 & 0 & 1 & 0 & 0 & 0 \\ 0 & \boxed{1} & -2 & -2 & -1 & -1 & 0 & 0 \\ 0 & 3 & -5 & -2 & -2 & 0 & 1 & 0 \\ 0 & -1 & 2 & 2 & -1 & 0 & 0 & 1 \end{array}\right].$$

Use $E_1[1,2]$; $E_{-3}[3,2]$; $E_1[4,2]$ to zero-out all non-pivot entries in the pivotal column to 0:

$$\left[\begin{array}{cccc|cccc} \boxed{1} & 0 & 0 & -2 & 0 & -1 & 0 & 0 \\ 0 & \boxed{1} & -2 & -2 & -1 & -1 & 0 & 0 \\ 0 & 0 & \boxed{1} & 4 & 1 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & -2 & -1 & 0 & 1 \end{array}\right].$$

Since a zero row has appeared in the A portion, A is not invertible. And $\text{rank}(A) = 3$, which is less than the order of A. The second portion of the augmented matrix has no meaning now. However, it records the elementary row operations which were carried out in the reduction process. Verify that this matrix is equal to

$$E_1[4,2]\, E_{-3}[3,2]\, E_1[1,2]\, E_{-1}[2]\, E_{-1}[4,1]\, E_{-2}[3,1]\, E_1[2,1]$$

and that the first portion is equal to this matrix times A.

For B, we proceed similarly. The augmented matrix $[B|I]$ with the first pivot looks like:

$$\left[\begin{array}{cccc|cccc} \boxed{1} & -1 & 2 & 0 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 2 & 0 & 1 & 0 & 0 \\ 2 & 1 & -1 & -2 & 0 & 0 & 1 & 0 \\ 0 & -2 & 0 & 2 & 0 & 0 & 0 & 1 \end{array}\right].$$

The sequence of elementary row operations $E_1[2,1]$; $E_{-2}[3,1]$ yields

$$\left[\begin{array}{cccc|cccc} \boxed{1} & -1 & 2 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 2 & 2 & 1 & 1 & 0 & 0 \\ 0 & 3 & -5 & -2 & -2 & 0 & 1 & 0 \\ 0 & -2 & 0 & 2 & 0 & 0 & 0 & 1 \end{array}\right].$$

Next, the pivot is -1 in $(2,2)$ position. Use $E_{-1}[2]$ to get the pivot as 1.

$$\left[\begin{array}{cccc|cccc} \boxed{1} & -1 & 2 & 0 & 1 & 0 & 0 & 0 \\ 0 & \boxed{1} & -2 & -2 & -1 & -1 & 0 & 0 \\ 0 & 3 & -5 & -2 & -2 & 0 & 1 & 0 \\ 0 & -2 & 0 & 2 & 0 & 0 & 0 & 1 \end{array}\right].$$

And then $E_1[1,2]$; $E_{-3}[3,2]$; $E_2[4,2]$ gives

$$\left[\begin{array}{cccc|cccc} \boxed{1} & 0 & 0 & -2 & 0 & -1 & 0 & 0 \\ 0 & \boxed{1} & -2 & -2 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 4 & 1 & 3 & 1 & 0 \\ 0 & 0 & -4 & -2 & -2 & -2 & 0 & 1 \end{array}\right].$$

Next pivot is 1 in $(3,3)$ position. Now, $E_2[2,3]$; $E_4[4,3]$ produces

$$\left[\begin{array}{cccc|cccc} \boxed{1} & 0 & 0 & -2 & 0 & -1 & 0 & 0 \\ 0 & \boxed{1} & 0 & 6 & 1 & 5 & 2 & 0 \\ 0 & 0 & \boxed{1} & 4 & 1 & 3 & 1 & 0 \\ 0 & 0 & 0 & 14 & 2 & 10 & 4 & 1 \end{array}\right].$$

Next pivot is 14 in $(4,4)$ position. Use $[4; 1/14]$ to get the pivot as 1:

$$\left[\begin{array}{cccc|cccc} \boxed{1} & 0 & 0 & -2 & 0 & -1 & 0 & 0 \\ 0 & \boxed{1} & 0 & 6 & 1 & 5 & 2 & 0 \\ 0 & 0 & \boxed{1} & 4 & 1 & 3 & 1 & 0 \\ 0 & 0 & 0 & \boxed{1} & 1/7 & 5/7 & 2/7 & 1/14 \end{array}\right].$$

Use $E_2[1,4]$; $E_{-6}[2,4]$; $E_{-4}[3,4]$ to zero-out the entries in the pivotal column:

$$\left[\begin{array}{cccc|cccc} \boxed{1} & 0 & 0 & 0 & 2/7 & 3/7 & 4/7 & 1/7 \\ 0 & \boxed{1} & 0 & 0 & 1/7 & 5/7 & 2/7 & -3/7 \\ 0 & 0 & \boxed{1} & 0 & 3/7 & 1/7 & -1/7 & -2/7 \\ 0 & 0 & 0 & \boxed{1} & 1/7 & 5/7 & 2/7 & 1/14 \end{array}\right].$$

Thus $B^{-1} = \frac{1}{7}\begin{bmatrix} 2 & 3 & 4 & 1 \\ 1 & 5 & 2 & -3 \\ 3 & 1 & -1 & -2 \\ 1 & 5 & 2 & \frac{1}{2} \end{bmatrix}$. Verify that $B^{-1}B = BB^{-1} = I$.

Chapter 4

Linear Systems

4.1 Linear independence

If $v = (a_1, \ldots, a_n)^T \in \mathbb{F}^{n\times 1}$, then we can express v in terms of the special vectors $e_1, \ldots, e_n$ as $v = a_1e_1 + \cdots + a_ne_n$. We now generalize the notions involved.

If $v_1, \ldots, v_m \in \mathbb{F}^{n\times 1}$, then the vector

$$\alpha_1 v_1 + \cdots \alpha_m v_m$$

is called a **linear combination** of these vectors, where $\alpha_1, \ldots, \alpha_m \in \mathbb{F}$ are some scalars.

For example, in $\mathbb{F}^{2\times 1}$, one linear combination of $v_1 = (1, 1)^T$ and $v_2 = (1, -1)^T$ is as follows:

$$2\begin{bmatrix}1\\1\end{bmatrix} + 1\begin{bmatrix}1\\-1\end{bmatrix} = \begin{bmatrix}3\\1\end{bmatrix}.$$

Is $(4, -2)^T$ a linear combination of v_1 and v_2? Yes, since

$$\begin{bmatrix}4\\-1\end{bmatrix} = 1\begin{bmatrix}1\\1\end{bmatrix} + 3\begin{bmatrix}1\\-1\end{bmatrix}.$$

In fact, every vector in $\mathbb{F}^{2\times 1}$ is a linear combination of v_1 and v_2. Reason:

$$\begin{bmatrix}a\\b\end{bmatrix} = \frac{a+b}{2}\begin{bmatrix}1\\1\end{bmatrix} + \frac{a-b}{2}\begin{bmatrix}1\\-1\end{bmatrix}.$$

However, every vector in $\mathbb{F}^{2\times 1}$ is not a linear combination of $(1, 1)^T$ and $(2, 2)^T$. Reason? Any linear combination of these two vectors is a multiple of $(1, 1)^T$. Then $(1, 0)^T$ is not a linear combination of these two vectors.

The vectors $v_1, \ldots, v_m$ in $\mathbb{F}^{n\times 1}$ are called **linearly dependent** iff at least one of them is a linear combination of others. The vectors are called **linearly independent** iff none of them is a linear combination of others.

For example, $(1, 1)^T$, $(1, -1)^T$, $(4, -1)^T$ are linearly dependent vectors whereas $(1, 1)^T$, $(1, -1)^T$ are linearly independent vectors.

Theorem 4.1. *The vectors $v_1, \dots, v_m \in \mathbb{F}^n$ are linearly independent iff for all $\alpha_1, \dots, \alpha_m \in \mathbb{F}$,*

$$\alpha_1 v_1 + \cdots \alpha_m v_m = 0 \text{ implies that } \alpha_1 = \cdots = \alpha_m = 0.$$

Notice that if $\alpha_1 = \cdots = \alpha_m = 0$, then obviously, the linear combination $\alpha_1 v_1 + \cdots + \alpha_m v_m$ evaluates to 0. But the above characterization demands its converse. We prove the following statement:

$v_1, \dots, v_m$ are linearly dependent iff $\alpha_1 v_1 + \cdots + \alpha_m v_m = 0$ for scalars $\alpha_1, \dots, \alpha_m$ not all zero.

Proof: If the vectors $v_1, \dots, v_m$ are linearly dependent then one of them is a linear combination of others. That is, we have an $i \in \{1, \dots, m\}$ such that

$$v_i = \alpha_1 v_1 + \cdots + \alpha_{i-1} v_{i-1} + \alpha_{i+1} v_{i+1} + \cdots + \alpha_m v_m.$$

Then

$$\alpha_1 v_1 + \cdots + \alpha_{i-1} v_{i-1} + (-1) v_i + \alpha_{i+1} v_{i+1} + \cdots + \alpha_m v_m = 0.$$

Here, we see that a linear combination becomes zero, where at least one of the coefficients, that is, the ith one is nonzero.
Conversely, suppose we have scalars $\alpha_1, \dots, \alpha_m$ not all zero such that

$$\alpha_1 v_1 + \cdots + \alpha_m v_m = 0.$$

Suppose $\alpha_j \neq 0$. Then

$$v_j = -\frac{\alpha_1}{\alpha_j} v_1 - \cdots - \frac{\alpha_{j-1}}{\alpha_j} v_{j-1} - \frac{\alpha_{j+1}}{\alpha_j} v_{j+1} - \cdots - \frac{\alpha_m}{\alpha_j} v_m.$$

That is, $v_1, \dots, v_n$ are linearly dependent. □

Example 4.1. Are the vectors $(1,1,1),\ (2,1,1),\ (3,1,0)$ linearly independent?

Let

$$a(1,1,1) + b(2,1,1) + c(3,1,0) = (0,0,0).$$

Comparing the components, we have

$$a + 2b + 3c = 0,\ a + b + c = 0,\ a + b = 0.$$

The last two equations imply that $c = 0$. Substituting in the first, we see that

$$a + 2b = 0.$$

This and the equation $a + b = 0$ give $b = 0$. Then it follows that $a = 0$.

We conclude that the given vectors are linearly independent.

For solving linear systems, it is of primary importance to find out which equations linearly depend on others. Once determined, such equations can be thrown away, and the rest can be solved.

Suppose that you are given with m number of vectors from $\mathbb{F}^{1\times n}$, say,

$$u_1 = (u_{11}, \dots, u_{1n}), \dots, u_m = (u_{m1}, \dots, u_{mn}).$$

We form the matrix A with rows as $u_1, \dots, u_m$. We then reduce A to its RREF, say, B. If there are r number of nonzero rows in B, then the rows corresponding to those rows in A are linearly independent, and the other rows (which have become the zero rows in B) are linear combinations of those r rows.

Example 4.2. From among the vectors $(1, 2, 2, 1),\ (2, 1, 0, -1),\ (4, 5, 4, 1),\ (5, 4, 2, -1)$, find linearly independent vectors; and point out which are the linear combinations of these independent ones.

We form a matrix with the given vectors as rows and then bring the matrix to its RREF.

$$\begin{bmatrix} \boxed{1} & 2 & 2 & 1 \\ 2 & 1 & 0 & -1 \\ 4 & 5 & 4 & 1 \\ 5 & 4 & 2 & -1 \end{bmatrix} \xrightarrow{R1} \begin{bmatrix} \boxed{1} & 2 & 2 & 1 \\ 0 & \boxed{-3} & -4 & -3 \\ 0 & -3 & -4 & -3 \\ 0 & -6 & -8 & -6 \end{bmatrix} \xrightarrow{R2} \begin{bmatrix} \boxed{1} & 0 & -\frac{2}{3} & -1 \\ 0 & \boxed{1} & \frac{4}{3} & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Here, $R1 = E_{-2}[2,1]$, $E_{-4}[3,1]$, $E_{-5}[4,1]$ and $R2 = E_{-3}[2]$, $E_{-2}[1,2]$, $E_3[3,2]$, $E_6[4,2]$.
No row exchanges have been applied in this reduction, and the nonzero rows are the first and the second rows. Therefore, the linearly independent vectors are $(1, 2, 2, 1),\ (0, 1, 4/3, 1)$; and the third and the fourth are linear combinations of these.

The same method can be used in $\mathbb{F}^{n\times 1}$. Just use the transpose of the columns, form a matrix, and continue with row reductions. Finally, take the transposes of the nonzero rows in the RREF.

A question: can you find four linearly independent vectors in $\mathbb{R}^{1\times 3}$?

4.2 Gram-Schmidt orthogonalization

There is another method for determining linear independence of a finite list of vectors using the inner product.

Let $v_1, \dots, v_n \in \mathbb{F}^n$. We say that these vectors are **orthogonal** iff $\langle v_i, v_j\rangle = 0$ for all pairs of indices i, j with $i \neq j$. When two vectors u and v are orthogonal, we write $u \perp v$.

Orthogonality is stronger than independence.

Theorem 4.2. *Any orthogonal list of nonzero vectors in $\mathbb{F}^n$ is linearly independent.*

Proof: Let $v_1, \dots, v_n \in \mathbb{F}^n$ be nonzero vectors. For scalars $a_1, \dots, a_n$, let $a_1v_1 + \cdots + a_nv_n = 0$. Take inner product of both the sides with v_1. Since $\langle v_i, v_1\rangle = 0$ for each $i \neq 1$, we obtain

$\langle a_1 v_1, v_1\rangle = 0$. But $\langle v_1, v_1\rangle \neq 0$. Therefore, $a_1 = 0$. Similarly, it follows that each $a_i = 0$. $\square$

It will be convenient to use the following terminology. We denote the set of all linear combinations of vectors $v_1, \ldots, v_n$ by $\text{span}(v_1, \ldots, v_n)$; and read it as the **span of** the vectors $v_1, \ldots, v_n$.

Our procedure, called **Gram-Schmidt orthogonalization**, constructs orthogonal vectors $v_1, \ldots, v_m$ from the given vectors $u_1, \ldots, u_n$ so that

$$\text{span}(v_1, \ldots, v_m) = \text{span}(u_1, \ldots, u_n).$$

It is described in Theorem **??** below.

Theorem 4.3. *Let $u_1, u_2, \ldots, u_m \in \mathbb{F}^n$. Define*

$$\begin{aligned} v_1 &= u_1 \\ v_2 &= u_2 - \frac{\langle u_2, v_1\rangle}{\langle v_1, v_1\rangle} v_1 \\ &\vdots \\ v_m &= u_m - \frac{\langle u_m, v_1\rangle}{\langle v_1, v_1\rangle} v_1 - \cdots - \frac{\langle u_m, v_{m-1}\rangle}{\langle v_{m-1}, v_{m-1}\rangle} v_{m-1} \end{aligned}$$

In the above process, if $v_i = 0$, then it is ignored for the rest of the steps. After ignoring such v_is suppose we obtain the vectors as $v_{j1}, \ldots, v_{jk}$. Then $v_{j1}, \ldots, v_{jk}$ are orthogonal and $\text{span}(v_{j1}, \ldots, v_{jk}) = \text{span}\{u_1, u_2, \ldots, u_m\}$.

Proof: $\langle v_2, v_1\rangle = \langle u_2 - \frac{\langle u_2, v_1\rangle}{\langle v_1, v_1\rangle} v_1, v_1\rangle = \langle u_2, v_1\rangle - \frac{\langle u_2, v_1\rangle}{\langle v_1, v_1\rangle}\langle v_1, v_1\rangle = 0.$

Notice that if $v_2 = 0$, then u_2 is a scalar multiple of u_1. That is, u_1, u_2 are linearly dependent. On the other hand, if $v_2 \neq 0$, then u_1, u_2 are linearly independent.

For the general case, use induction. $\square$

Example 4.3. Consider the vectors $u_1 = (1, 0, 0)$, $u_2 = (1, 1, 0)$ and $u_3 = (1, 1, 1)$. Apply Gram-Schmidt Orthogonalization.

$v_1 = (1, 0, 0)$.

$$v_2 = u_2 - \frac{\langle u_2, v_1\rangle}{\langle v_1, v_1\rangle} v_1 = (1, 1, 0) - \frac{(1, 1, 0) \cdot (1, 0, 0)}{(1, 0, 0) \cdot (1, 0, 0)}(1, 0, 0) = (1, 1, 0) - 1\,(1, 0, 0) = (0, 1, 0).$$

$$\begin{aligned} v_3 &= u_3 - \frac{\langle u_3, v_1\rangle}{\langle v_1, v_1\rangle} v_1 - \frac{\langle u_3, v_2\rangle}{\langle v_2, v_2\rangle} v_2 \\ &= (1, 1, 1) - (1, 1, 1) \cdot (1, 0, 0)(1, 0, 0) - (1, 1, 1) \cdot (0, 1, 0)(0, 1, 0) \\ &= (1, 1, 1) - (1, 0, 0) - (0, 1, 0) = (0, 0, 1). \end{aligned}$$

The set $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ is orthogonal; and span of the new vectors is the same as span of the old ones, which is $\mathbb{R}^3$.

If u_i is a linear combination of earlier us, then the corresponding v_i will be the zero vector.

Example 4.4. Use Gram-Schmidt orthogonalization on the vectors $u_1 = (1, 1, 0, 1)$, $u_2 = (0, 1, 1, -1)$ and $u_3 = (1, 3, 2, -1)$.

$v_1 = (1, 1, 0, 1)$.

$$v_2 = u_2 - \frac{\langle u_2, v_1\rangle}{\langle v_1, v_1\rangle} v_1 = (0, 1, 1, -1) - \frac{\langle (0, 1, 1, -1), (1, 1, 0, 1)\rangle}{\langle (1, 1, 0, 1), (1, 1, 0, 1)\rangle} (1, 1, 0, 1) = (0, 1, 1, -1).$$

$$\begin{aligned} v_3 &= u_3 - \frac{\langle u_3, v_1\rangle}{\langle v_1, v_1\rangle} v_1 - \frac{\langle u_3, v_2\rangle}{\langle v_2, v_2\rangle} v_2 \\ &= (1, 3, 2, -1) - \frac{\langle (1, 3, 2, -1), (1, 1, 0, 1)\rangle}{\langle (1, 1, 0, 1), (1, 1, 0, 1)\rangle} (1, 1, 0, 1) - \frac{\langle (1, 3, 2, -1), (0, 1, 1, -1)\rangle}{\langle (0, 1, 1, -1), (0, 1, 1, -1)\rangle} (0, 1, 1, -1) \\ &= (1, 3, 2, -1) - (1, 1, 0, 1) - 2(0, 1, 1, -1) = (0, 0, 0, 0). \end{aligned}$$

Notice that since u_1, u_2 are already orthogonal, Gram-Schmidt process returned $v_2 = u_2$. Next, the process also revealed the fact that $u_3 = u_1 - 2u_2$.

Example 4.5. We redo Example **??** using Gram-Schmidt orthogonalization. There, we had

$$u_1 = (1, 2, 2, 1),\ u_2 = (2, 1, 0, -1),\ u_3 = (4, 5, 4, 1),\ u_4 = (5, 4, 2, -1).$$

Then

$v_1 = (1, 2, 2, 1)$.

$$v_2 = (2, 1, 0, -1) - \frac{\langle (2, 1, 0, -1), (1, 2, 2, 1)\rangle}{\langle (1, 2, 2, 1), (1, 2, 2, 1)\rangle} (1, 2, 2, -1) = (\tfrac{17}{10}, \tfrac{2}{5}, -\tfrac{3}{5}, -\tfrac{13}{10}).$$

$$\begin{aligned} v_3 = (4, 5, 4, 1) &- \frac{\langle (4, 5, 4, 1), (1, 2, 2, 1)\rangle}{\langle (1, 2, 2, 1), (1, 2, 2, 1)\rangle} (1, 2, 2, 1) \\ &- \frac{\langle (4, 5, 4, 1), (\frac{3}{2}, 0, -1, \frac{1}{2})\rangle}{\langle (\frac{17}{10}, \frac{2}{5}, -\frac{3}{5}, -\frac{13}{10}), (\frac{17}{10}, \frac{2}{5}, -\frac{3}{5}, -\frac{13}{10})\rangle} (\tfrac{17}{10}, \tfrac{2}{5}, -\tfrac{3}{5}, -\tfrac{13}{10}) = (0, 0, 0, 0). \end{aligned}$$

So, we ignore v_3; and mark that u_3 is a linear combination of u_1 and u_2. Next, we compute

$$v_4 = u_4 - \frac{\langle u_4, v_1\rangle}{\langle v_1, v_1\rangle} v_1 - \frac{\langle u_4, v_2\rangle}{\langle v_2, v_2\rangle} v_2 = 0.$$

Therefore, we conclude that u_4 is a linear combination of u_1 and u_2. Finally, we obtain the orthogonal vectors v_1, v_2 such that $\operatorname{span}(u_1, u_2, u_3, u_4) = \operatorname{span}(v_1, v_2)$.

Though Gram-Schmidt process is more difficult than the row reduction, it provides more information. Further, we do not have to compute the orthogonal vectors manually; softwares are available.

4.3 Rank of a matrix

The number of pivots in the RREF of A is called the **rank** of A, and is denoted by $\operatorname{rank}(A)$.

If $\operatorname{rank}(A) = r$, then there are r number of linearly independent columns in A and other columns are linear combinations of these r columns. The linearly independent columns

correspond to the pivotal columns in the RREF of A. Also, there exist r number of linearly independent rows of A such that other rows are linear combinations of these r rows. The linearly independent rows correspond to the nonzero rows in the RREF of A.

Example 4.6. Consider $A = \begin{bmatrix} 1 & 1 & 1 & 2 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 3 & 5 & 3 & 4 & 3 \\ -1 & 0 & -1 & -3 & -1 \end{bmatrix}$.

Here, we see that $row(3) = row(1) + 2row(2), \quad row(4) = row(2) - 2row(1)$.
But $row(2)$ is not a scalar multiple of $row(1)$, that is, $row(1),\ row(2)$ are linearly independent; and all other rows are linear combinations of $row(1)$ and $row(2)$.

Also, we see that

$$col(3) = col(5) = col(1),\ col(4) = 3col(1) - col(2).$$

And columns one and two are linearly independent. You can compute its RREF to find that $\text{rank}(A) = 2$.

It raises a question. Suppose for a matrix A, we find r number of linearly independent rows such that other rows are linear combinations of these r rows. Can it happen that there are also k rows which are linearly independent and other rows are linear combinations of these k rows, and that $k \neq r$?

Theorem 4.4. *Let $A \in \mathbb{F}^{m\times n}$. Let $P \in \mathbb{F}^{m\times m}$ and $Q \in \mathbb{F}^{n\times n}$ be invertible. Suppose there exist r linearly independent rows and r linearly independent columns in A such that other rows are linear combinations of these r rows; and the other columns are linear combinations of these r columns. Then there exist r linearly independent rows and r linearly independent columns in PAQ such that other rows are linear combinations of these r rows, and other columns are linear combinations of these r columns.*

Proof: Let $u_1, \ldots, u_r, u \in \mathbb{F}^{m\times 1}$. Let $a_1, \ldots, a_r \in \mathbb{F}$. Observe that

$$u = a_1u_1 + \cdots + a_ru_r \quad \text{iff} \quad Pu = a_1Pu_1 + \cdots a_rPu_r.$$

Taking $u = 0$, we see that the vectors $u_1, \ldots, u_r \in \mathbb{F}^{m\times 1}$ are linearly independent iff $Pu_1, \ldots, Pu_r$ are linearly independent. Further, the above equation implies that if there exist r number of columns in A which are linearly independent and other columns are linear combinations of these r columns, then the same is true for the matrix PA.

We now show a similar fact about the rows of A. For this purpose, let $v_1, \ldots, v_r, v \in \mathbb{F}^{1\times n}$. Let $a_1, \ldots, a_r \in \mathbb{F}$. Then

$$v = a_1v_1 + \cdots + a_rv_r \quad \text{iff} \quad vQ = a_1v_1Q + \cdots a_rv_rQ.$$

Analogously, it follows that there exist r number of linearly independent rows in A and other rows are linear combinations of these rows iff the same happens in AQ.

This completes the proof. □

Each elementary matrix is invertible. Thus, Theorem **??** implies that if we apply a sequence of elementary row operations on A, then this number of linearly independent rows (or columns) such that other rows (columns) are linear combinations of those linearly independent ones, does not change. From the RREF of a matrix we thus conclude that this number of linearly independent rows is equal to the number of such linearly independent columns, and this is equal to the rank of the matrix. That is,

$$\begin{aligned}\text{rank}(A) &= \text{the maximum number of linearly independent rows in } A \\ &= \text{the maximum number of linearly independent columns in } A.\end{aligned}$$

Therefore, $\text{rank}(A) = \text{rank}(A^T)$.

Linear combinations have some thing to do with the RREF of a matrix. Suppose A has been reduced to its RREF. Let $R_{i1}, \ldots, R_{ir}$ be the rows of A which have become the nonzero rows in the RREF, and other rows have become the zero rows. Also, suppose $C_{j1}, \ldots, C_{jr}$ for $j1 < \cdots < jr$, be the columns of A which have become the pivotal columns in the RREF, other columns being non-pivotal. Using Theorem **??**, we see that the following are true:

1. All rows of A other than $R_{i1}, \ldots, R_{ir}$ are linear combinations of $R_{i1}, \ldots, R_{ir}$.
2. All columns of A other than $C_{j1}, \ldots, C_{jr}$ are linear combinations of $C_{j1}, \ldots, C_{jr}$.
3. The columns $C_{j1}, \ldots, C_{jr}$ have respectively become $e_1, \ldots, e_r$ in the RREF.
4. If $e_1, \ldots, e_k$ are all the pivotal columns in the RREF that occur to the left of a non-pivotal column, then the non-pivotal column is in the form $(a_1, \ldots, a_k, 0, \ldots, 0)^T$. Further, if a column C in A has become this non-pivotal column in the RREF, then $C = a_1 C_{j1} + \cdots + a_k C_{jk}$.

We say that A and B are **equivalent** iff there exist invertible matrices $P \in \mathbb{F}^{m \times m}$ and $Q \in \mathbb{F}^{n \times n}$ such that $B = PAQ$

From Theorem **??** it follows that equivalent matrices have the same rank. On the other hand, if $\text{rank}(A) = r$, then using elementary row and column operations, it can be reduced to a matrix of the following form:

$$\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}.$$

Here, the first r columns are $e_1, \ldots, e_r$; and other columns are zero columns. This proves the *Rank Theorem*, which states the following:

Two matrices of the same size are equivalent iff they have the same rank.

You have seen earlier that there do not exist three linearly independent vectors in $\mathbb{R}^2$. With the help of rank, now we can see why does it happen.

Theorem 4.5. *Let $u_1, \ldots, u_k$ and $v_1, \ldots, v_m$ be vectors in $\mathbb{F}^n$. Suppose $u_1, \ldots, u_k$ are linearly independent, $v_1, \ldots, v_m \in \text{span}(u_1, \ldots, u_k)$ and $m > k$. Then $v_1, \ldots, v_m$ are linearly dependent.*

Proof: Consider all vectors as row vectors. Form the matrix A by taking its rows as $u_1, \ldots, u_k$, $v_1, \ldots, v_m$ in that order. Now, $\text{rank}(A) = m$. Similarly, construct the matrix B by taking its rows as $v_1, \ldots, v_m, u_1, \ldots, u_k$, in that order. Now, A and B are equivalent. Therefore, $\text{rank}(A) = \text{rank}(B)$. Since $m > k = \text{rank}(B)$, out of $v_1, \ldots, v_m$ at most k vectors can be linearly independent. It follows that $v_1, \ldots, v_m$ are linearly dependent. □

Closely connected with this notion is that of similarity of two matrices. We say that two matrices $A, B \in \mathbb{F}^{n \times n}$ are **similar** iff there exists an invertible matrix P such that $B = P^{-1}AP$. Though equivalence is easy to characterize through the rank, similarity is a much more complex notion.

4.4 Linear equations

A system of linear equations, also called a **linear system** with m equations in n unknowns looks like:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots a_{mn}x_n &= b_m \end{aligned}$$

As you know, using the abbreviation $x = (x_1, \ldots, x_n)^T$, $b = (b_1, \ldots, b_m)^T$ and $A = [a_{ij}]$, the system can be written in the compact form:

$$Ax = b.$$

Here, $A \in \mathbb{F}^{m \times n}$, $x \in \mathbb{F}^{n \times 1}$ and $b \in \mathbb{F}^{m \times 1}$ so that m is the *number of equations* and n is the *number of unknowns* in the system. Notice that for linear systems, we deviate from our symbolism and write b as a column vector and x_i are unknown scalars. The system $Ax = b$ is **solvable**, also said *to have a solution*, iff there exists a vector $u \in \mathbb{F}^{n \times 1}$ such that $Au = b$.

Thus, the system $Ax = b$ is solvable iff b is a linear combination of columns of A. Also, $Ax = b$ has a unique solution iff b is a linear combination of columns of A and the columns of A are linearly independent. These issues are better tackled with the help of the corresponding **homogeneous system**

$$Ax = 0.$$

The homogeneous system always has a solution, namely, $x = 0$. It has infinitely many solutions iff it has a nonzero solution. For, if u is a solution, so is αu for any scalar α.

To study the non-homogeneous system, we use the augmented matrix $[A|b] \in \mathbb{F}^{m \times (n+1)}$ which has its first n columns as those of A in the same order, and the $(n+1)$th column is b.

Theorem 4.6. *Let $A \in \mathbb{F}^{m\times n}$ and let $b \in \mathbb{F}^{m\times 1}$. Then the following statements are true:*

1. *If $[A' \,|\, b']$ is obtained from $[A \,|\, b]$ by applying a finite sequence of elementary row operations, then each solution of $Ax = b$ is a solution of $A'x = b'$, and vice versa.*
2. **(Consistency)** *$Ax = b$ has a solution iff* $\operatorname{rank}([A \,|\, b]) = \operatorname{rank}(A)$.
3. *If u is a (particular) solution of $Ax = b$, then each solution of $Ax = b$ is given by $u + y$, where y is a solution of the homogeneous system $Ax = 0$.*
4. *If $r = \operatorname{rank}([A \,|\, b]) = \operatorname{rank}(A) < n$, then there are $n - r$ unknowns which can take arbitrary values; and other r unknowns can be determined from the values of these $n-r$ unknowns.*
5. *If $m < n$, then the homogeneous system has infinitely many solutions.*
6. *$Ax = b$ has a unique solution iff* $\operatorname{rank}([A \,|\, b]) = \operatorname{rank}(A) = n$.
7. *If $m = n$, then $Ax = b$ has a unique solution iff* $\det(A) \neq 0$.
8. **(Cramer's Rule)** *If $m = n$ and* $\det(A) \neq 0$, *then the solution of $Ax = b$ is given by $x_j = \det(\, A_j(b)\,)/\det(A)$ for each $j \in \{1, \ldots, n\}$.*

Proof: (1) If $[A' \,|\, b']$ has been obtained from $[A \,|\, b]$ by a finite sequence of elementary row operations, then $A' = EA$ and $b' = Eb$, where E is the product of corresponding elementary matrices. The matrix E is invertible. Now, $A'x = b'$ iff $EAx = Eb$ iff $Ax = E^{-1}Eb = b$.

(2) Due to (1), we assume that $[A \,|\, b]$ is in RREF. Suppose $Ax = b$ has a solution. If there is a zero row in A, then the corresponding entry in b is also 0. Therefore, there is no pivot in b. Hence $\operatorname{rank}([A \,|\, b]) = \operatorname{rank}(A)$.

Conversely, suppose that $\operatorname{rank}([A \,|\, b]) = \operatorname{rank}(A) = r$. Then there is no pivot in b. That is, b is a non-pivotal column in $[A \,|\, b]$. Thus, b is a linear combination of pivotal columns, which are some columns of A. Therefore, $Ax = b$ has a solution.

(3) Let u be a solution of $Ax = b$. Then $Au = b$. Now, z is a solution of $Ax = b$ iff $Az = b$ iff $Az = Au$ iff $A(z - u) = 0$ iff $z - u$ is a solution of $Ax = 0$. That is, each solution z of $Ax = b$ is expressed in the form $z = u + y$ for a solution y of the homogeneous system $Ax = 0$.

(4) Let $\operatorname{rank}([A \,|\, b]) = \operatorname{rank}(A) = r < n$. By (2), there exists a solution. Due to (3), we consider solving the corresponding homogeneous system. Due to (1), assume that A is in RREF. There are r number of pivots in A and $m - r$ number of zero rows. Omit all the zero rows; it does not affect the solutions. Write the system as linear equations. Rewrite the equations by keeping the unknowns corresponding to pivots on the left hand side, and taking every other term to the right hand side. The unknowns corresponding to pivots are now expressed in terms of the other $n - r$ unknowns. For obtaining a solution, we may arbitrarily assign any values to these $n - r$ unknowns, and the unknowns corresponding to the pivots get evaluated by the equations.

(5) Let $m < n$. Then $r = \text{rank}(A) \le m < n$. Consider the homogeneous system $Ax = 0$. By (4), there are $n - r \ge 1$ number of unknowns which can take arbitrary values, and other r unknowns are determined accordingly. Each such assignment of values to the $n-r$ unknowns gives rise to a distinct solution resulting in infinite number of solutions of $Ax = 0$.

(6) It follows from (3) and (4). The unique solution is given by $x = A^{-1}b$.

(7) If $A \in \mathbb{F}^{n\times n}$, then it is invertible iff $\text{rank}(A) = n$ iff $\det(A) \neq 0$. Then use (6).

(8) Recall that $A_j(b)$ is the matrix obtained from A by replacing the jth column of A with the vector b. Since $\det(A) \neq 0$, by (6), $Ax = b$ has a unique solution, say $y \in \mathbb{F}^{n\times 1}$. Write the identity $Ay = b$ in the form:

$$y_1 \begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix} + \cdots + y_j \begin{bmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{bmatrix} + \cdots + y_n \begin{bmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}.$$

This gives

$$y_1 \begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix} + \cdots + \begin{bmatrix} (y_j a_{1j} - b_1) \\ \vdots \\ (y_j a_{nj} - b_n) \end{bmatrix} + \cdots + y_n \begin{bmatrix} a_{1n} \\ \cdots \\ a_{nn} \end{bmatrix} = 0.$$

In this sum, the jth vector is a linear combination of other vectors, where $-y_j$s are the coefficients. Therefore,

$$\begin{vmatrix} a_{11} & \cdots & (y_j a_{1j} - b_1) & \cdots & a_{1n} \\ & & \vdots & & \\ a_{n1} & \cdots & (y_j a_{nj} - b_n) & \cdots & a_{nn} \end{vmatrix} = 0.$$

From Property (6) of the determinant, it follows that

$$y_j \begin{vmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ & & \vdots & & \\ a_{n1} & \cdots & a_{nj} & \cdots & a_{nn} \end{vmatrix} - \begin{vmatrix} a_{11} & \cdots & b_1 & \cdots & a_{1n} \\ & & \vdots & & \\ a_{n1} & \cdots & b_n & \cdots & a_{nn} \end{vmatrix} = 0.$$

Therefore, $y_j = \det(\, A_j(b)\,)/\det(A)$. □

A system of linear equations $Ax = b$ is said to be **consistent** iff $\text{rank}([A|b]) = \text{rank}(A)$.

Theorem **??**(1) says that only consistent systems have solutions. Theorem **??**(2) asserts that all solutions of the non-homogeneous system can be obtained by adding a particular solution to solutions of the corresponding homogeneous system.

4.5 Gauss-Jordan elimination

To determine whether a system of linear equations is consistent or not, it is enough to convert the augmented matrix $[A|b]$ to its RREF and then check whether an entry in the b portion

of the augmented matrix has become a pivot or not. In fact, the pivot check shows that corresponding to the zero rows in the portion of A in the RREF of $[A|b]$, all the entries in b must be zero. Thus an entry in the b portion has become a pivot guarantees that the system is inconsistent, else the system is consistent.

Example 4.7. Is the following system of linear equations consistent?

$$\begin{aligned} 5x_1 + 2x_2 - 3x_3 + x_4 &= 7 \\ x_1 - 3x_2 + 2x_3 - 2x_4 &= 11 \\ 3x_1 + 8x_2 - 7x_3 + 5x_4 &= 8 \end{aligned}$$

We take the augmented matrix and reduce it to its RREF by elementary row operations.

$$\left[\begin{array}{cccc|c} 5 & 2 & -3 & 1 & 7 \\ 1 & -3 & 2 & -2 & 11 \\ 3 & 8 & -7 & 5 & 8 \end{array}\right] \xrightarrow{R1} \left[\begin{array}{cccc|c} \boxed{1} & 2/5 & -3/5 & 1/5 & 7/5 \\ 0 & -17/5 & 13/5 & -11/5 & 48/5 \\ 0 & 34/5 & -26/5 & 22/5 & -19/5 \end{array}\right]$$

$$\xrightarrow{R2} \left[\begin{array}{cccc|c} \boxed{1} & 0 & -5/17 & -1/17 & 43/17 \\ 0 & \boxed{1} & -13/17 & 11/17 & -48/17 \\ 0 & 0 & 0 & 0 & \boxed{77/5} \end{array}\right]$$

Here, $R_1 = E_{1/5}[1]$, $E_{-1}[2,1]$, $E_{-3}[3,1]$ and $R2 = E_{-5/17}[2]$, $E_{-2/5}[1,2]$, $E_{-34/5}[3,2]$. Since an entry in the b portion has become a pivot, the system is inconsistent. In fact, you can verify that the third row in A is simply first row minus twice the second row, whereas the third entry in b is not the first entry minus twice the second entry. Therefore, the system is inconsistent.

Gauss-Jordan elimination is an application of converting the augmented matrix to its RREF for solving linear systems.

Example 4.8. We change the last equation in the previous example to make it consistent. The system now looks like:

$$\begin{aligned} 5x_1 + 2x_2 - 3x_3 + x_4 &= 7 \\ x_1 - 3x_2 + 2x_3 - 2x_4 &= 11 \\ 3x_1 + 8x_2 - 7x_3 + 5x_4 &= -15 \end{aligned}$$

The reduction to echelon form will change that entry as follows: take the augmented matrix and reduce it to its echelon form by elementary row operations.

$$\left[\begin{array}{cccc|c} 5 & 2 & -3 & 1 & 7 \\ 1 & -3 & 2 & -2 & 11 \\ 3 & 8 & -7 & 5 & -15 \end{array}\right] \xrightarrow{R1} \left[\begin{array}{cccc|c} \boxed{1} & 2/5 & -3/5 & 1/5 & 7/5 \\ 0 & -17/5 & 13/5 & -11/5 & 48/5 \\ 0 & 34/5 & -26/5 & 22/5 & -96/5 \end{array}\right]$$

$$\xrightarrow{R2} \left[\begin{array}{cccc|c} \boxed{1} & 0 & -5/17 & -1/17 & 43/17 \\ 0 & \boxed{1} & -13/17 & 11/17 & -48/17 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

with $R_1 = E_{1/5}[1]$, $E_{-1}[2,1]$, $E_{-3}[3,1]$ and $R2 = E_{-5/17}[2]$, $E_{-2/5}[1,2]$, $E_{-34/5}[3,2]$. This expresses the fact that the third equation is redundant. Now, solving the new system in RREF is easier. Writing as linear equations, we have

$$\begin{array}{rrrrrr} \boxed{1}\, x_1 & & -\frac{5}{17}x_3 & -\frac{1}{17}x_4 & = & \frac{43}{17} \\ & \boxed{1}\, x_2 & -\frac{13}{17}x_3 & +\frac{11}{17}x_4 & = & -\frac{48}{17} \end{array}$$

The unknowns corresponding to the pivots are called the **basic variables** and the other unknowns are called the **free variable**. By assigning the free variables to any arbitrary values, the basic variables can be evaluated. So, we assign a free variable x_i an arbitrary number, say α_i, and express the basic variables in terms of the free variables to get a solution of the equations.

In the above reduced system, the basic variables are x_1 and x_2; and the unknowns x_3, x_4 are free variables. We assign x_3 to α_3 and x_4 to α_4. The solution is written as follows:

$$x_1 = \frac{43}{17} + \frac{5}{17}\alpha_3 + \frac{1}{17}\alpha_4, \quad x_2 = -\frac{48}{17} + \frac{13}{17}\alpha_3 - \frac{11}{17}\alpha_4, \quad x_3 = \alpha_3, \quad x_4 = \alpha_4.$$

Chapter 5

Matrix Eigenvalue Problem

5.1 Eigenvalues and eigenvectors

Let $A \in \mathbb{F}^{n\times n}$. A scalar $\lambda \in \mathbb{F}$ is called an **eigenvalue** of A iff there exists a non-zero vector $v \in \mathbb{F}^{n\times 1}$ such that $Av = \lambda v$. Such a vector v is called an **eigenvector of A for** (or, associated with, or, corresponding to) the eigenvalue λ.

Example 5.1. Consider the matrix $A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$. It has an eigenvector $(0, 0, 1)^T$ associated with the eigenvalue 1. Is $(0, 0, c)^T$ also an eigenvector associated with the same eigenvalue 1?

In fact, corresponding to an eigenvalue, there are infinitely many eigenvectors.

Theorem 5.1. *Let $A \in \mathbb{F}^{n\times n}$. Let $v \in \mathbb{F}^{n\times 1}$, $v \neq 0$. Then, v is an eigenvector of A for the eigenvalue $\lambda \in \mathbb{F}$ iff v is a nonzero solution of the homogeneous system $(A - \lambda I)x = 0$ iff* $\det(A - \lambda I) = 0$.

Proof: The scalar λ is an eigenvalue of A iff we have a nonzero vector $v \in \mathbb{F}^{n\times 1}$ such that $Av = \lambda v$ iff $(A - \lambda I)v = 0$ and $v \neq 0$ iff $A - \lambda I$ is not invertible iff $\det(A - \lambda I) = 0$. □

5.2 Characteristic polynomial

The polynomial $\det(A - tI)$ is called the **characteristic polynomial** of the matrix A. A zero of the characteristic polynomial of A is called a **characteristic root** of the matrix A. Recall that if $p(t)$ is a polynomial, its zero is a scalar α such that $p(\alpha) = 0$.

The above theorem says that a scalar is an eigenvalue of A iff it is a characteristic root. That is, if $A \in \mathbb{R}^{n\times n}$, then the scalars are supposed to be real numbers, and then any real characteristic root of A is an eigenvalue of A. However, there can be characteristic roots of A which are not real. Such characteristic roots are not eigenvalues of A.

However, each real number is a complex number. Thus, each matrix in $\mathbb{R}^{n\times n}$ is also in $\mathbb{C}^{n\times n}$. For convenience, we will make it a *convention that all our matrices are complex matrices.*

In this sense, a characteristic root of a matrix will be called a *complex eigenvalue* of the matrix.

Since the characteristic polynomial of a matrix A of order n is a polynomial of degree n in t, it has exactly n, not necessarily distinct, complex zeros. And these are the eigenvalues of A. Notice that, here, we are using the fundamental theorem of algebra which says that each polynomial of degree n with complex coefficients can be factored into exactly n linear factors.

Caution: When λ is a complex eigenvalue of $A \in \mathbb{F}^{n\times n}$, a corresponding eigenvector x is, in general, a vector in $\mathbb{C}^{n\times 1}$.

Example 5.2. Find the eigenvalues and corresponding eigenvectors of the matrix

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

The characteristic polynomial is

$$\det(A - tI) = \begin{vmatrix} 1-t & 0 & 0 \\ 1 & 1-t & 0 \\ 1 & 1 & 1-t \end{vmatrix} = (1-t)^3.$$

The characteristic roots are $1, 1, 1$. These are the eigenvalues of A.

To get an eigenvector, we solve $A(a, b, c)^T = (a, b, c)^T$ or that

$$a = a,\ a + b = b,\ a + b + c = c.$$

It gives $b = c = 0$ and $a \in \mathbb{F}$ can be arbitrary. Since an eigenvector is nonzero, all the eigenvectors are given by $(a, 0, 0)^T$, for any $a \neq 0$.

The eigenvalue λ being a characteristic root has certain multiplicity. That is, the maximum k such that $(t-\lambda)^k$ divides the characteristic polynomial is called the **algebraic multiplicity** of the eigenvalue λ.

In Example **??**, the algebraic multiplicity of the eigenvalue 1 is 3.

Example 5.3. For $A \in \mathbb{R}^{2\times 2}$, given by

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

the characteristic polynomial is $t^2+1 = 0$. It has no real roots. However, with our convention at work, i and $-i$ are its (complex) eigenvalues. That is, the same matrix $A \in \mathbb{C}^{2\times 2}$ has eigenvalues as i and $-i$. The corresponding eigenvectors are obtained by solving

$$A(a, b)^T = i(a, b)^T \text{ and } A(a, b) = -i(a, b)^T.$$

For $\lambda = i$, we have $b = ia, -a = ib$. Thus, $(a, ia)^T$ is an eigenvector for $a \neq 0$.
For the eigenvalue $-i$, the eigenvectors are $(a, -ia)$ for $a \neq 0$.
Here, algebraic multiplicity of each eigenvalue is 1.

Theorem 5.2.

1. *A matrix and its transpose have the same eigenvalues.*
2. *Similar matrices have the same eigenvalues.*
3. *The diagonal entries of any triangular matrix are precisely its eigenvalues.*

Proof: (1) $\det(A^T - tI) = \det((A - tI)^T) = \det(A - tI)$.
(2) $\det(P^{-1}AP - tI) = \det(P^{-1}(A - tI)P) = \det(P^{-1})\det(A - tI)\det(P) = \det(A - tI)$.
(3) In all these cases, $\det(A - tI) = (a_{11} - t) \cdots (a_{nn} - t)$. □

Theorem 5.3. $\det(A)$ *equals the product and* $\mathrm{tr}(A)$ *equals the sum of all eigenvalues of* A.

Proof: Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A, not necessarily distinct. Now,

$$\det(A - tI) = (\lambda_1 - t) \cdots (\lambda_n - t).$$

Put $t = 0$. It gives $\det(A) = \lambda_1 \cdots \lambda_n$.
Expand $\det(A - tI)$ and equate the coefficients of t^{n-1} to get

$$\begin{aligned}&\text{Coeff of } t^{n-1} \text{ in } \det(A - tI) = \text{Coeff of } t^{n-1} \text{ in } (a_{11} - t) \cdot A_{11}\\ &= \cdots = \text{Coeff of } t^{n-1} \text{ in } (a_{11} - t) \cdot (a_{22} - t) \cdots (a_{nn} - t) = (-1)^{n-1}\mathrm{tr}(A).\end{aligned}$$

But Coeff of t^{n-1} in $\det(A - tI) = (-1)^{n-1} \cdot \sum \lambda_i$. □

Theorem 5.4. (Caley-Hamilton) *Any square matrix satisfies its characteristic polynomial.*

Proof: Let $A \in \mathbb{F}^{n \times n}$. Its characteristic polynomial is

$$p(t) = (-1)^n \det(A - t\, I).$$

We show that $p(A) = 0$, the zero matrix. Theorem **??**(14) with the matrix $A - t\, I$ says that

$$p(t)\, I = (-1)^n \det(A - tI)\, I = (-1)^n (A - tI)\, \mathrm{adj}\,(A - tI).$$

The entries in $\mathrm{adj}\,(A - tI)$ are polynomials in t of degree at most $n - 1$. Write

$$\mathrm{adj}\,(A - tI) := B_0 + tB_1 + \cdots + t^{n-1}B_{n-1},$$

where $B_0, \ldots, B_{n-1} \in \mathbb{F}^{n \times n}$. Then

$$p(t)I = (-1)^n (A - t\, I)(B_0 + tB_1 + \cdots t^{n-1}B_{n-1}).$$

Notice that this is an identity in polynomials, where the coefficients of t^j are matrices. Substituting t by any matrix of the same order will satisfy the equation. In particular, substituting A for t we obtain $p(A) = 0$. □

5.3 Hermitian and unitary matrices

Theorem 5.5. *Let $A \in \mathbb{F}^{n\times n}$. Let λ be any complex eigenvalue of A.*

1. *If A is hermitian or real symmetric, then $\lambda \in \mathbb{R}$.*
2. *If A is skew-hermitian or skew-symmetric, then λ is purely imaginary or zero.*
3. *If A is unitary or orthogonal, then $|\lambda| = 1$.*

Proof: Let $\lambda \in \mathbb{C}$ be an eigenvalue of A with an eigenvector $v \in \mathbb{C}^{n\times 1}$. Now, $Av = \lambda v$. Pre-multiplying with v^*, we have $v^*Av = \lambda v^*v \in \mathbb{C}$.

(1) Let A be hermitian, i.e., $A = A^*$. Now,

$$(v^*Av)^* = v^*A^*v = v^*Av \quad \text{and} \quad (v^*v)^* = v^*v.$$

So, both v^*Av and v^*v are real. Therefore, in $v^*Av = \lambda v^*v$, λ is also real.

(2) When A is skew-hermitian, $(v^*Av)^* = -v^*Av$. Then $v^*Av = \lambda v^*v$ implies that

$$(\lambda v^*v)^* = -\lambda(v^*v).$$

Since $v \neq 0,\ v^*v \neq 0$. Therefore, $\lambda^* = \bar{\lambda} = -\lambda$. That is, $2\mathrm{Re}(\lambda) = 0$. This shows that λ is purely imaginary or zero.

(3) Let A be unitary, i.e., $A^*A = I$. Now, $Av = \lambda v$ implies $v^*A^* = (\lambda v)^* = \bar{\lambda}v^*$. Then

$$v^*v = v^*Iv = v^*A^*Av = \bar{\lambda}\lambda v^*v = |\lambda|^2 v^*v.$$

Since $v^*v \neq 0$, $|\lambda| = 1$. □

Not only each eigenvalue of a real symmetric matrix is real, but also a corresponding real eigenvector can be chosen. To see this, let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Let $\lambda \in \mathbb{R}$ be an eigenvalue of A. If $v = x + iy \in \mathbb{C}^{n\times 1}$ is a corresponding eigenvector with $x, y \in \mathbb{R}^{n\times 1}$, then

$$A(x + iy) = \lambda(x + iy).$$

Comparing the real and imaginary parts, we have

$$Ax = \lambda x, \quad Ay = \lambda y.$$

Since $x + iy \neq 0$, at least one of x or y is nonzero. Choose one nonzero vector out of x and y. That is a real eigenvector corresponding to the eigenvalue λ of A.

Thus, a real symmetric matrix of order n has n real eigenvalues counting multiplicities.

Theorem 5.6. *Let $A \in \mathbb{F}^{n\times n}$ be a unitary or an orthogonal matrix.*

1. *For each pair of vectors x, y, $\langle Ax, Ay\rangle = \langle x, y\rangle$. In particular, $\|Ax\| = \|x\|$ for any x.*
2. *The columns of A are orthogonal and each is of norm* 1.
3. *The rows of A are orthogonal, and each is of norm* 1.
4. $|\det(A)| = 1$.

Proof: (1) $\langle Ax, Ay\rangle = \langle x, A^*Ay\rangle = \langle x, y\rangle$. Take $x = y$ for the second equality.

(2) Since $A^*A = I$, the ith row of A^* multiplied with the jth column of A gives δ_{ij}. However, this product is simply the inner product of the jth column of A with the ith column of A.

(3) It follows from (2). Also, considering $AA^* = I$, we get this result.

(4) Since $\det(A)$ is the product of the eigenvalues of A, and each eigenvalue has absolute value 1, the result follows. □

It thus follows that the determinant of an orthogonal matrix is either 1 or -1.

5.4 Diagonalization

Theorem 5.7. *Eigenvectors associated with distinct eigenvalues of an $n \times n$ matrix are linearly independent.*

Proof: Let $\lambda_1, \ldots, \lambda_m$ be all the distinct eigenvalues of $A \in \mathbb{F}^{n\times n}$. Let $v_1, \ldots, v_m$ be corresponding eigenvectors. We use induction on $i \in \{1, \ldots, m\}$.
For $i = 1$, since $v_1 \neq 0$, $\{v_1\}$ is linearly independent,
Induction Hypothesis: for $i = k$ suppose $\{v_1, \ldots, v_k\}$ is linearly independent. We use the characterization of linear independence as proved in Theorem **??**.
Our induction hypothesis implies that if we equate any linear combination of $v_1, \ldots, v_k$ to 0, then the coefficients in the linear combination must all be 0. Now, for $i = k + 1$, we want to show that $v_1, \ldots, v_k, v_{k+1}$ are linearly independent. So, we start equating an arbitrary linear combination of these vectors to 0. Our aim is to derive that each scalar coefficient in such a linear combination must be 0. Towards this, assume that

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k + \alpha_{k+1} v_{k+1} = 0. \tag{5.1}$$

Then, $A(\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k + \alpha_{k+1} v_{k+1}) = 0$. Since $Av_j = \lambda_j v_j$, we have

$$\alpha_1\lambda_1 v_1 + \alpha_2\lambda_2 v_2 + \cdots + \alpha_k\lambda_k v_k + \alpha_{k+1}\lambda_{k+1} v_{k+1} = 0. \tag{5.2}$$

Multiply (**??**) with λ_{m+1}. Subtract from (**??**) to get:

$$\alpha_1(\lambda_1 - \lambda_{m+1})v_1 + \cdots + \alpha_k(\lambda_k - \lambda_{k+1})v_k = 0.$$

By the Induction Hypothesis, $\alpha_j(\lambda_j - \lambda_{k+1}) = 0$ for each $j = 1, \ldots, k$. Since $\lambda_1, \ldots, \lambda_{k+1}$ are distinct, we conclude that $\alpha_1 = \cdots = \alpha_k = 0$. Then, from (**??**), it follows that $\alpha_{k+1} v_{k+1} = 0$. As $v_{k+1} \neq 0$, we have $\alpha_{k+1} = 0$. □

Suppose an $n \times n$ matrix A has n distinct eigenvalues. Let v_i be an eigenvector corresponding to λ_i for $i = 1, \ldots, n$. The list of vectors $v_1, \ldots, v_n$ is linearly independent and it contains n vectors. We find that

$$Av_1 = \lambda_1 v_1, \quad \ldots, Av_n = \lambda_n v_n.$$

Construct the matrix $P \in \mathbb{F}^{n \times n}$ by taking its columns as the eigenvectors $v_1, \ldots, v_n$. That is, let

$$P = \begin{bmatrix} v_1 & v_2 & \cdots & v_{m-1} & v_m \end{bmatrix}.$$

Also, construct the diagonal matrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. That is,

$$D = \begin{bmatrix} \lambda_1 & \cdots & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}.$$

Then the above product of A with the v_is can be written as a single equation

$$AP = PD.$$

Now, $\operatorname{rank}(P) = n$. So, P is an invertible matrix. Then the above equation shows that

$$P^{-1}AP = D.$$

That is, A is similar to a diagonal matrix.

We give a definition and then summarize our result as in the following.

Let $A \in \mathbb{F}^{n \times n}$. We call A to be **diagonalizable** iff A is similar to a diagonal matrix. We also say that A is **diagonalizable by the matrix** P iff $P^{-1}AP = D$.

Theorem 5.8. *Let $A \in \mathbb{F}^{n \times n}$ have n distinct eigenvalues. Then A is similar to the diagonal matrix, whose diagonal entries are the eigenvalues of A.*

To **diagonalize** a matrix A means that we determine an invertible matrix P and a diagonal matrix D such that $P^{-1}AP = D$. Notice that only square matrices can possibly be diagonalized.

In general, diagonalization starts with determining eigenvalues and corresponding eigenvectors of A. We then construct the diagonal matrix D by taking the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A. Next, we construct P by putting the corresponding eigenvectors $v_1, \ldots, v_n$ as columns of P in that order. Then $P^{-1}AP = D$. This work succeeds provided that the list of eigenvectors $v_1, \ldots, v_n$ in $\mathbb{F}^{n \times 1}$ are linearly independent.

Theorem 5.9. *An $n \times n$ matrix is diagonalizable iff it has n linearly independent eigenvectors.*

Proof: In fact, we have already proved the theorem. Let us repeat it.

Suppose $A \in \mathbb{F}^{n\times n}$ is diagonalizable. Let $P = [v_1, \ \cdots \ v_n]$ be an invertible matrix and let $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ such that $P^{-1}AP = D$. Then $AP = PD$. Then $Av_i = \lambda_i v_i$ for $1 \leq i \leq n$. That is, each v_i is an eigenvector of A. Moreover, P is invertible implies that $v_1, \ldots, v_n$ are linearly independent.

Conversely, suppose $u_1, \ldots, u_n$ are linearly independent eigenvectors of $A \in \mathbb{F}^{n\times n}$ with corresponding eigenvalues $d_1, \ldots, d_n$. Then $Q = [u_1 \ \cdots \ u_n]$ is invertible. And, $Au_j = d_j u_j$ implies that $AQ = Q\Delta$, where $\Delta = \text{diag}(d_1, \ldots, d_n)$. Therefore, A is diagonalizable. □

Theorem 5.10. (Spectral Theorem) *Each normal matrix is diagonalizable by a unitary matrix. In particular, each hermitian matrix is diagonalizable by a unitary matrix, and each real symmetric matrix is diagonalizable by an orthogonal matrix.*

It is easy to check that if a matrix A is diagonalizable by a unitary matrix, then A must be a normal matrix.

Once we know that a matrix is A diagonalizable, we can give a procedure to diagonalize it. All we have to do is determine the eigenvalues and corresponding eigenvectors so that the eigenvectors are linearly independent and their number is equal to the order of A. Then, put the eigenvectors as columns to construct the matrix P. Then $P^{-1}AP$ is a diagonal matrix.

Example 5.4. $A = \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}$ is real symmetric.

It has eigenvalues $-1, \ 2, \ 2$. To find the associated eigenvectors, we must solve the linear systems of the form $Ax = \lambda x$.

For the eigenvalue -1, the system $Ax = -x$ gives

$$x_1 - x_2 - x_3 = -x_1, \ -x_1 + x_2 - x_3 = -x_2, \ -x_1 - x_2 + x_3 = -x_3.$$

It gives $x_1 = x_2 = x_3$. One eigenvector is $(1, 1, 1)^T$.

For the eigenvalue 2, we have the equations as

$$x_1 - x_2 - x_3 = 2x_1, \ -x_1 + x_2 - x_3 = 2x_2, \ -x_1 - x_2 + x_3 = 2x_3.$$

Which gives $x_1 + x_2 + x_3 = 0$. We can have two linearly independent eigenvectors such as $(-1, 1, 0)^T$ and $(-1, -1, 2)^T$.

The three eigenvectors are orthogonal to each other. To orthonormalize, we deivide each by its norm. We end up at the following orthonormal eigenvectors:

$$\begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix}, \quad \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}, \quad \begin{bmatrix} -1/\sqrt{6} \\ -1/\sqrt{6} \\ 2/\sqrt{6} \end{bmatrix}.$$

They are orthogonal vectors in $\mathbb{R}^{3\times 1}$, each of norm 1. Taking

$$P = \begin{bmatrix} 1/\sqrt{3} & -1/\sqrt{2} & -1/\sqrt{6} \\ 1/\sqrt{3} & 1/\sqrt{2} & -1/\sqrt{6} \\ 1/\sqrt{3} & 0 & 2/\sqrt{6} \end{bmatrix},$$

we see that $P^{-1} = P^T$ and

$$P^{-1}AP = P^TAP = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Bibliography

[1] *Advanced Engineering Mathematics, 10th Ed.,* E. Kreyszig, John Willey & Sons, 2010.

[2] *Introduction to Linear Algebra, 2nd Ed.,* S. Lang, Springer-Verlag, 1986.

[3] *Calculus of One Variable,* M.T. Nair, Ane Books, 2014.

[4] *Differential and Integral Calculus Vol. 1-2,* N. Piskunov, Mir Publishers, 1974.

[5] *Introduction to Matrix Theory*, A. Singh, Ane Books, 2018.

[6] *Linear Algebra and its Applications,* G. Strang, Cengage Learning, 4th Ed., 2006.

[7] *Thomas Calculus*, G.B. Thomas, Jr, M.D. Weir, J.R. Hass, Pearson, 2009.

Index