COMP 0299

— Intermediate Maths for CS.

Week 1 - Sequences of Functions

Function: $f : I \mapsto \mathbb{R}$ $(I \subseteq \mathbb{R})$

function takes an input and returns a real numbr where input is a real number (subset)

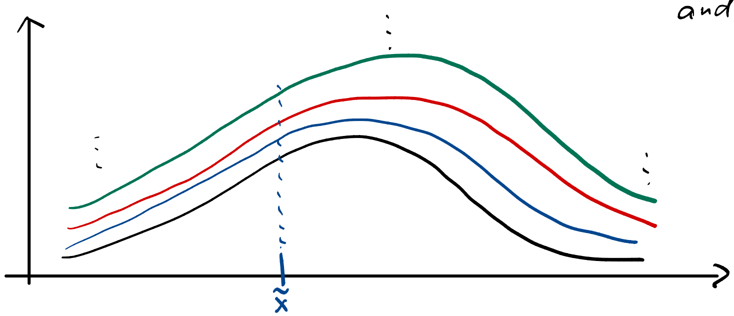
Continuous function: $f : \mathbb{R} \mapsto \mathbb{R}$

Function that is connected

Bounded function: $f : I \to \mathbb{R}$ is bounded iff
$\{f(x) \mid x \in I\} = f[I]$ is bounded set in $\mathbb{R}$
↳ If range of function is bound.
$(\iff \sup_{x \in I} |f(x)| < \infty)$

Sequence of functions:

Assuming that $\forall n \in \mathbb{N}$, function $f_n : A \to \mathbb{R}$
$(f_n) \to$ sequence of functions
$[(f_n)]$
where sequence is $(f_1, f_2, f_3, \dots f_n)$ where $n \geq 1$
and sequence members
$f_1 : I \mapsto \mathbb{R}$
$f_2 : I \mapsto \mathbb{R}$
$f_3 : I \mapsto \mathbb{R}$
$\vdots$
$f_n : I \mapsto \mathbb{R}$



The functions must have the same domain

For any fixed $\tilde{x} \in I$, $(f_1(\tilde{x}), f_2(\tilde{x}), f_3(\tilde{x}), ..)$, we will get a sequence of real numbers

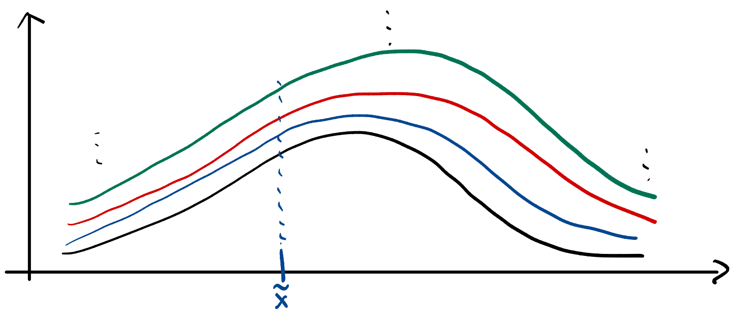
Pointwise convergence:

→ defined as (f_n)

$(f_1, f_2, f_3, f_4 \ldots)$ is pointwise convergent to a function $f: I \mapsto \mathbb{R}$ if for all $\tilde{x} \in I$ st $(f_1(\tilde{x}), f_2(\tilde{x}), f_3(\tilde{x}) \ldots)$ is convergent to $f(\tilde{x})$ where $f(\tilde{x})$ is a limit function

Mathematical definition:

$$[\forall x \in I, \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall n \geq N : |f_n(x) - \underbrace{f(x)}_{\text{(pointwise) limit function}}| < \varepsilon]$$



All the codomain values (vertical slices) are convergent sequences (they converge to a limit) → distance to limit gets progressively smaller (denoted by ε), but doesn't touch limit (tends to 0).

Formal definition [from lecture notes]:

- By definition, for a fixed value $x = a$, we have a sequence $(f_n(a))_{n \in \mathbb{N}}$
- We recall that the f_n are defined on the interval A
- If for all $a \in A$ there exists a finite value $\lim_{n\to\infty}(f_n(a))$

  then we can define the limit function: $f : x \mapsto \lim_{n\to\infty}(f_n(x))$
- We say that the sequence $(f_n(a))_{n \in \mathbb{N}}$ converges *pointwise* towards f.

New notion of convergence

Note that limit function doesn't need to be continuous (uniform convergence ensures/proves continuity)

Proving pointwise convergence

Consider $f_n = x^n$ on $[0,1]$. Prove (f_n) converges pointwise.

POINTWISE definition:

$\forall \varepsilon > 0,\ \forall x \in [0,1], \exists N \in \mathbb{N}$ st $\forall n \geq N,\ |f_n(x) - f(x)| < \varepsilon$

If $x=0$, $f_n(0) = 0^n = 0 \to 0$, $|f_n(0) - 0| = |0-0| = 0 < \varepsilon$ ✓

$x=1$, $f_n(1) = 1^n = 1 \to 1$, $|f_n(1) - 1| = |1-1| = 0 < \varepsilon$ ✓

$0<x<1$, $f_n(x) = x^n \to 0$, $|f_n(x) - 0| = |x^n - 0| = |x^n|$

→ we need to find an N where $\exists N \in \mathbb{N}$:

$= x^n < \varepsilon$

$= \ln x^n < \ln \varepsilon$

$= n \ln x < \ln \varepsilon$

$= n > \frac{\ln \varepsilon}{\ln x}$

Choose $N > \frac{\ln \varepsilon}{\ln x}$

Proof: let $\varepsilon > 0$ and $x \in [0,1]$

If $x=0$, $f_n(0) = 0$, so take any $N \in \mathbb{Z}^*$ then for any $n > N$, $|f_n(0) - 0|$
$= |0-0| = 0 < \varepsilon$

If $x=1$, $f_n(1) = 1^n = 1$, so take any $N \in \mathbb{Z}^*$ then for any $n > N$, $|f_n(1) - 1|$
$= |1-1| = 0 < \varepsilon$

If $0<x<1$, choose $N > \frac{\ln \varepsilon}{\ln x}$. Then for any $n > N$,
$|f_n(x) - 0| = |x^n - 0| = |x^n| = x^n$

Since $n > N > \frac{\ln \varepsilon}{\ln x}$, we have $n > \frac{\ln \varepsilon}{\ln x}$, so $n \ln x < \ln \varepsilon$, so
$\ln x < \frac{\ln \varepsilon}{n}$, so $x < e^{\frac{\ln \varepsilon}{n}}$

Thus $|f_n(x) - 0| = x^n < \left(e^{\frac{\ln \varepsilon}{n}}\right)^n = e^{\ln \varepsilon} = \varepsilon$

$\therefore x^n < \varepsilon$ ∎

$\therefore f_n(x) = x^n,\ x \in [0,1],\ f(x) = \begin{cases} 0, & 0 \leq x < 1 \\ 1 & x = 1 \end{cases}$

Cauchy's Criterion:

↳ Method to show if the terms of sequence gets closer to each other → it will show convergence for a sequence of numbers. (as n gets bigger)

↳ used to prove convergence without finding limit of sequence.

A sequence $(u_0, u_1, \ldots, u_n)$ is Cauchy if for all $\varepsilon > 0$, there exists some N st $|u_m - u_n| < \varepsilon$ for all $m, n > N$.

$$\ulcorner \quad \forall \varepsilon > 0, \exists N \in \mathbb{N}, \forall m, n \geq N, \ |u_n - u_m| < \varepsilon \quad \lrcorner$$

↳ shows that the terms gets arbitrarily closer together. (i.e. distance between terms gets progressively smaller (distance denoted by ε)

IMPORTANT: for a sequence of real numbers.

Cauchy sequence $\Longleftrightarrow$ Convergent sequence

Proving Cauchy sequence

Prove the sequence $u_n = \left(\frac{1}{n}\right)$ is Cauchy

Cauchy: $\forall \varepsilon > 0, \exists N \in \mathbb{N}$ st $\forall m, n \geq N : |u_m - u_n| < \varepsilon$

natural number

Proof

Let $\varepsilon > 0$. Choose $N > 2/\varepsilon$ [there exists a number larger than a nat num]

$\forall n, m > N, \left|\frac{1}{n} - \frac{1}{m}\right| \leq \frac{1}{n} + \frac{1}{m}$

$< \frac{1}{N} + \frac{1}{N}$

$< \varepsilon/2 + \varepsilon/2$

$= \varepsilon$ ∎

As $n > N$, $\frac{1}{n} < \frac{1}{N}$

$m > N$, $\frac{1}{m} < \frac{1}{N}$

$\therefore$

① Sketch ← do that first

$\varepsilon > 0$, $|u_m - u_n| < \varepsilon$

$\left|\frac{1}{m} + \left(-\frac{1}{n}\right)\right| \leq \left|\frac{1}{m}\right| + \left|-\frac{1}{n}\right|$

$= \frac{1}{m} + \frac{1}{n} < \varepsilon$

$\frac{1}{n} < \varepsilon/2$

$2 < n\varepsilon$

$n > 2/\varepsilon$

Choose $N > 2/\varepsilon$

↳ Archmedian property

Prove that sequence $\frac{1}{3^n}$ is Cauchy

Proof:

Let $\varepsilon > 0$, $\left|\frac{1}{3^n} - \frac{1}{3^m}\right| \leq \left|\frac{1}{3^n}\right| + \left|\frac{1}{3^m}\right|$

and $m, n > N$

$= \frac{1}{3^n} + \frac{1}{3^m}$

$< \frac{1}{3^N} + \frac{1}{3^N}$

$< \varepsilon/2 + \varepsilon/2$

$= \varepsilon$ ∎

Cauchy

$m > N \Rightarrow \frac{1}{m} < \frac{1}{N}$

$\frac{1}{3^m} < \frac{1}{3^N}$

$n > N \Rightarrow \frac{1}{3^n} < \frac{1}{3^N}$

As: $n > N > \frac{\log 2/\varepsilon}{\log 3}$

$n > \frac{\log 2/\varepsilon}{\log 3}$

$n \log 3 > \log 2/\varepsilon$

$3^n > 2/\varepsilon$

$\boxed{\frac{1}{3^n} < \varepsilon/2}$

Sketch:

$\forall \varepsilon > 0, |u_n - u_m| < \varepsilon$

$= \left|\frac{1}{3^n} - \frac{1}{3^m}\right| \leq \left|\frac{1}{3^n}\right| + \left|-\frac{1}{3^m}\right|$

$= \frac{1}{3^n} + \frac{1}{3^m}$

$< \frac{1}{3^n} + \frac{1}{3^n} < \varepsilon$

$\frac{2}{3^n} < \varepsilon$

$\frac{2}{3^m} < \varepsilon$

$3^n > 2/\varepsilon$

$\log 3^n > \log 2/\varepsilon$

$n > \frac{\log 2/\varepsilon}{\log 3}$. Choose $N > \frac{\log 2/\varepsilon}{\log 3}$

Uniform convergence

Sequence of function (f_n) is uniform convergent to $f: A \to \mathbb{R}$ where A is the interval

from which only more in the interval

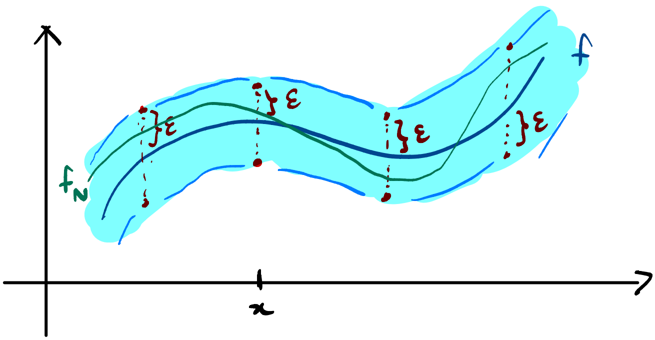
$$\forall \varepsilon > 0, \exists N \in \mathbb{N} \text{ st } \forall n \geq N, \forall x \in A, |f_n(x) - f(x)| < \varepsilon$$

for any area size

works uniformly for all x (there exist a rank)

all terms have their graph within the area

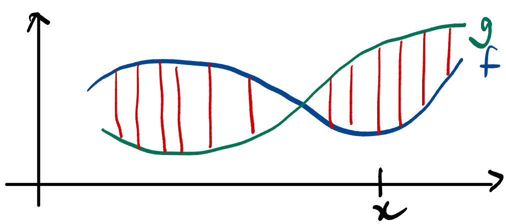
(note: pointwise has $\forall x \in A$ first where we pick N depending on x, but in uniform, N works uniformly for all x in interval at the same time)



region of length 2ε is along all points (epsilon tube around graph of function f)
inequality holds for all x in interval A.

Graph of function f_N needs to lie inside the epsilon tube, along with all the graphs in the sequence that come after N
→ shows it is possible to measure distance between 2 functions.

Distance of functions



$f: I \to \mathbb{R}$
$g: I \to \mathbb{R}$

using absolute value: $|f(x) - g(x)|$

largest distance $\Leftrightarrow$ maximum absolute value obtained

BUT we don't know if max exist so we use supremum: $\sup_{x \in I} |f(x) - g(x)|$

(largest possible distance between numbers)

Supremum: least upper bound of a set of numbers that exists for any bounded set of real numbers even when a maximum doesn't exist

Supremum norm: $\|f-g\|_\infty = \sup_{x\in I} |f(x) - g(x)|$
of $f-g$

Uniform convergence means: $\|f_n - f\|_\infty \xrightarrow{n\to\infty} 0$

↳ largest possible distance between sequence of functions and limit of function must approach 0 as n approaches infinity

NOTE:

Pointwise convergence $\not\Rightarrow$ Uniform convergence

but $\Leftarrow$

If all functions in a sequence are continuous, is the limit continuous?

YES - if sequence converges uniformly

NO - if sequence ONLY converges pointwise.

- Uniform convergence preserves continuity

If sequence $(f_n(a))_{n\in\mathbb{N}}$ converges uniformly to f on $[a,b]$ then:

$$\lim_{n\to\infty} \int_a^b f_n(x)\,dx = \lim_{n\to\infty} \int_a^b f(x)\,dx$$

If sequence (u_n) consist of differentiable functions in $[a,b]$

If $((f_n(a))_n$ converges to $f(a)$, and if (f'_n) converges uniformly on $[a,b]$,

then

- Convergence of (f'_n) is Uniform
- $\lim_{n\to\infty} f'_n(x) = f'(x)$

↳ For differentiation, condition of uniform convergence is on sequence of derivatives

Proving uniform convergence

6. By using uniform convergence, show that $\lim_{n\to\infty}(\int_1^2 e^{-nx}dx) = 0$.

1. (x^n) converges uniformly on $[0, \frac{1}{2}]$ → using supremum

Uniform convergence:

$\forall \varepsilon > 0, \exists N \in \mathbb{N}$ st $\forall n > N, \forall x \in [0, \frac{1}{2}]$, $|f_n(x) - f(x)| < \varepsilon$ ← supremum

$x \le \frac{1}{2} \Rightarrow x^n \le (\frac{1}{2})^n = \frac{1}{2^n}$

Scratchwork (Goal:)

$|x^n - 0| < \varepsilon$

$|x^n| = x^n < \varepsilon$

$x^n < \varepsilon$

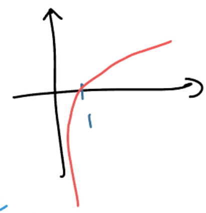
$x^n \le (\frac{1}{2})^n < \varepsilon$

$\log (\frac{1}{2})^n < \log \varepsilon$

$n \log(\frac{1}{2}) < \log \varepsilon$

$n > \frac{\log \varepsilon}{\log \frac{1}{2}}$

$N > \frac{\log \varepsilon}{\log \frac{1}{2}}$



Proof:

Let $\underbrace{\varepsilon > 0}_{\text{start}}$. Choose $\underbrace{N > \frac{\log \varepsilon}{\log x}}$, then $\forall n > N, \forall x \in [0, \frac{1}{2}]$.

Find natural number N st $\forall n > N$ and $\forall x \in$ domain, distance between $f_n(x)$ and $f(x)$ is less than ε.

$|x^n - 0| = |x^n| = x^n \le (\frac{1}{2})^n$

Since $n > N > \frac{\log \varepsilon}{\log \frac{1}{2}}$ so

$n > \frac{\log \varepsilon}{\log \frac{1}{2}}$, so $n \log \frac{1}{2} < \log \varepsilon$

$(\frac{1}{2})^n < \varepsilon$

Hence $|x^n - 0| \le (\frac{1}{2})^n < \varepsilon$ ~~m~~ uniform convergent.

Series of Functions

$\forall n \in \mathbb{N}, \; f_n : A \to \mathbb{R}$

If (f_n) is a sequence of functions, then the series $\sum f_n$ is the series of functions where

$$\sum f_n = f_1 + f_2 + f_3 + \dots + f_n + \dots\dots = \text{'infinite sum'}$$

of sequence of function

$\sum f_n$ converges pointwise to f if $S_n = f_1 + \dots + f_n$ converges pointwise to f. [sequence of partial sums]

$\sum f_n$ converges uniformly to f if $S_n = f_1 + \dots + f_n$ converges uniformly to f.

where $(S_n)_n = \sum_{n=1}^{n} f_n$ = sequence of partial sums

Weierstrass M-test

- Comparison test if series of functions converges uniformly
  - Can use any series convergence test

- We have a sequence of functions $f_n : A \to \mathbb{R}$
  and a sequence of non-negative numbers M_n.
- If we have $\forall n \in \mathbb{N}, \forall x \in A, |f_n(x)| \leq M_n$
  and if the series of numbers $\sum M_n$ converges,

  Then the series of functions $\sum f_n$ converges uniformly on A.

compare the functions to the terms of a convergent series [not depending on x]

For a sequence of functions (f_n), let $M_n > 0$ be s.t. if $\forall n \in \mathbb{N}, \forall x \in A, |f_n(x)| \leq M_n$.

If $\sum M_n$ converges uniformly then $\sum f_n$ converges uniformly on A.

To test if $\sum M_n$ converges uniformly, there are many ways to test series convergence e.g. Cauchy's Criterion

$$\iff \forall \varepsilon > 0, \exists N \in \mathbb{N} \text{ s.t. } \forall n,m > N, |M_n - M_m| < \varepsilon$$

Alternatively can use p-series, comparison test, ratio test, root-test, test for divergence, direct comparison test, limit comparison test

BUT Weierstrass Test doesn't tell us much about the limit itself.

Power Series:

- series of functions in the form $\sum_{n \geq 0} a_n (x-b)^n$ where (a_n) is a sequence of numbers
- it is centred in b. Power series centred at 0 is $\sum_{n \geq 0} a_n x^n$
- used to represent functions with infinite polynomials

$$= \sum_{n \geq 0}^{\infty} a_n (x-b)^n$$

centre of power series

co-efficient

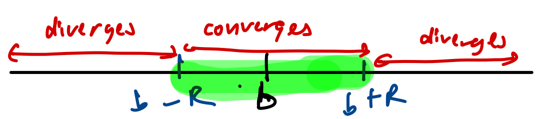
↳ can be found by Taylor Series

Convergence of Power Series:

- There exists a number $r \geq 0$ = radius of convergence st:
- On interval $(b-r, b+r)$, series converges uniformly
- when $x > b+r$, or $x < b-r$, series diverges

$|x-b| < R \to$ converges

$|x-b| > R \to$ diverges



Finding ratio of convergence :

↳ By ratio test

$|x-b| < L$ [radius of convergence]

For power series $\sum a_n (x-b)^n$, expressed as $\sum c_n$

$\sum c_n$ converges if $\lim_{n\to\infty} \left| \frac{c_{n+1}}{c_n} \right| < 1$

We must have $\lim_{n\to\infty} \left| \frac{a_{n+1} (x-b)^{\cancel{n}+1}}{a_n \cancel{(x-b)^n}} \right| < 1$

so $\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| |x-b| < 1$

we need $|x-a| < \underbrace{\lim_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right|}_{L \text{ [radius]}}$

If power series is raised to power k i.e. $\sum a_n (x-b)^{kn}$

$$|x-a| < \underbrace{\sqrt[k]{\lim_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right|}}_{L \text{ [radius]}}$$

Summary:

If
- $\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = L \neq 0 \Rightarrow$ radius of convergence $= 1/L$
- $\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = 0 \Rightarrow$ radius of convergence $= +\infty$
  → converges uniformly everywhere
- $\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = \infty \Rightarrow$ radius of convergence $= 0$

Taylor Series

- Converge to well-known functions [↑ degree, closer to real value]
- Approximation of series of functions expressed in polynomials

$$f(x) \approx \sum_{n=0}^{\infty} a_n (x-b)^n$$

- Differentiation used to find coefficients
- A type of power series

↳ infinite sum (series) of polynomials to approx a function e.g. e^x, $\sin(x)$, etc.

$$f(x) \approx a_0 + a_1(x-b) + a_2(x-b)^2 + \dots$$

$$f^{(1)}(x) \approx a_1 + 2a_2(x-b) + 3a_3(x-b)^2 + \dots$$

$$f^{(2)}(x) \approx 2a_2 + 3\cdot 2 a_3 (x-b) + \dots$$

$$f^{(3)}(x) \approx 3\cdot 2\cdot 1\, a_3 + \dots$$

$f'(x) \approx P_n^{(1)}(x)$

$f'(0) = 1/2$, $P_n^{(1)}(0) = 1/2$

using derivatives as coefficients for approx.

$$f^{(3)}(b) \approx 3!\, a_3$$

$$a_3 = \frac{f^{(3)}(b)}{3!}, \quad a_4 = \frac{f^{(4)}(b)}{4!}, \quad a_n = \frac{f^{(n)}(b)}{n!}$$

← to find coefficients

Taylor Series of f centred at b:

when $|x-a| < R$

R: radius of convergence

$$\sum_{n \geq 0} \frac{f^{(n)}(b)}{n!}(x-b)^n$$

$$= f(b) + f^{(1)}(b)(x-b) + \frac{f^{(2)}(b)}{2!}(x-b)^2 + \dots + \frac{f^{(n)}(b)}{n!}(x-b)^n$$

[where $f^{(0)}(b) = f(b)$]

when $b=0$, it is known as Maclaurin Series.

Taylor's Remainder

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

as f(x) provides approximation but remainder used to get exact value [error]

$R_n(x,b) \Rightarrow$ error

$= O((x-b)^{n+1})$ ← grows at most quickly as $(x-b)^{n+1}$ [UB]

$= o((x-b)^n)$ ← grows slower than $(x-b)^n$ [LB]

Asymptotic comparison between ↓

where $R_n(x,b) = \dfrac{f^{(n+1)}(c)}{(n+1)!}(x-b)^{n+1}$

E.g.

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots = x + R_n(x,b)$$

$$= x + o(x) = x + O(x^2)$$

$$\therefore \frac{\sin(x)}{x} = 1 + \frac{o(x)}{x}$$

$$\therefore \lim_{x\to 0} \frac{\sin(x)}{x} = \lim_{x\to 0} 1 + \lim_{x\to 0} \frac{o(x)}{x}$$

$$= \lim_{x\to 0} 1 + o(x) \lim_{x\to 0} \frac{1}{x}$$

$$= 1 + 0$$

$$= 1$$

[$O(x)$ = UB

$o(x)$ = LB]

Linear Algebra

recap vector space:

V

$\vec{a} \in V$ where $\vec{a}$ is an element in vector space V

elements abide by properties of vectors. , i.e. they are vectors

e.g. $\vec{a}+\vec{b} = \vec{b}+\vec{a}$, $\vec{a}+(\vec{b}+\vec{c}) = (\vec{a}+\vec{b})+\vec{c}$

$\vec{a}+(-\vec{a}) = \vec{0}$, $\vec{a}+\vec{0} = \vec{a}$, $c(\vec{a}+\vec{b}) = c\vec{a}+c\vec{b}$

$(c+d)\vec{a} = c\vec{a}+d\vec{a}$, $cd(\vec{a}) = cd\vec{a}$, $1\vec{a} = \vec{a}$

$= c(d\vec{a})$

vector space V = non-empty set equipped with two maps:

1. Vector addition map: $V \times V \to V$, $(u, v) \to u+v$
2. Scalar multiplication map: $V \times \mathbb{R} \to V$, $(v, \lambda) \to \lambda v$

[where $v = (v_1, v_2, v_3 \ldots v_n)$]

VECTOR SPACE AXIOMS:

1. $u+v = v+u$
2. $(u+v)+w = u+(v+w)$
3. $v+0 = 0+v = v$
4. $v+(-v) = (-v)+v = 0$
5. $\lambda(u+v) = \lambda u+\lambda v$
6. $(\lambda+\mu)v = \lambda v+\mu v$
7. $(\lambda\mu)v = \lambda(\mu v)$
8. $1.v = v$

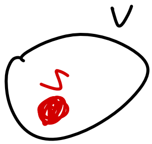
Example: set of real numbers $\mathbb{R}$

1. $5 \in \mathbb{R}$, $c = 2$, $2(5) = 10$ $10 \in \mathbb{R}$ ✓
2. $5 \in \mathbb{R}$, $3 \in \mathbb{R}$, then $5+3 = 8 \in \mathbb{R}$ ✓

Set of real numbers is a vector space as it satisfies all 2 closure properties

- Vector spaces can be made of functions

Subspaces:



$S \subseteq V$

every element in subspace S is also an element in vector space V.

S is also a vector space → has to satisfy closure

$S \subseteq V$ is a subspace, provided:

1. S is non-empty [$\exists u \in U$]
2. S closed under addition [For $u_1, u_2 \in U \Rightarrow u_1 + u_2 \in U$]
3. S closed under scalar multiplication [For $\lambda \in \mathbb{R}, u \in U \Rightarrow \lambda u \in U$]

To prove if a subspace, ensure all 3 properties are satisfied

e.g. $S = \{0\}$, $0 + 0 = \underline{0} \in S$

$1\underline{0} = \underline{0} \in S$

Subspace Test

$S \subseteq V$ for V a vector space, is a subspace provided:

(I) $\underline{0} \in S$

(II) $\lambda u_1 + u_2 \in U$

[for $\lambda \in \mathbb{R}$, $u_1, u_2 \in S$]

– show the test upholds the definition of subspace and vice versa

Span:

for a vector space V

$\vec{v}_1, \vec{v}_2, \vec{v}_3, \dots \vec{v}_n \in V$

$a_1\vec{v}_1 + a_2\vec{v}_2 + \dots + a_n\vec{v}_n$ ⇒ linear combination

Set of all linear combinations = span

e.g. $\text{span}(\vec{v}_1, \vec{v}_2, \vec{v}_3) = a\vec{v}_1 + b\vec{v}_2 + c\vec{v}_3$

span of any number of elements of vector space V is also a subspace of V

→ also smallest subspace of V that contains set of elements

You use span to describe a vector space

A span is linearly independent if you cannot express elements as linear combination of the other element [set $\{v_1, \dots v_n\}$ linearly independent ⟺ the only solution to $\alpha_1 v_1 + \dots + \alpha_n v_n = 0$ is $\alpha_1 = \alpha_2 = \alpha_n = 0$]

Linear Transformations

- sort of functions in vector spaces
- map one vector space to another [or back to its own vector space) $L: V \to V$

$$V \xrightarrow{L} W$$

$$L: V \to W \qquad [\text{e.g. } L: \mathbb{R}^2 \to \mathbb{R}\ ]$$

take vectors of length 2 and return scalars

Linear Transformation Properties

1. $T(v_1 + v_2) = T(v_1) + T(v_2)$, $\forall v_1, v_2 \in V$
2. $T(\lambda v) = \lambda T(v)$, $\forall v \in V$ and scalars $\lambda \in \mathbb{F}$

$\iff T(\alpha v_1 + \beta v_2) = \alpha T(v_1) + \beta T(v_2)$ - used to show if linear map.

You represent linear transformation with matrices

$T(\vec{v}) = A\vec{v}$ e.g. if $T: \mathbb{R}^n \to \mathbb{R}^m$

where A is $m \times n$ matrix

that can be obtained by transforming standard basis of $\mathbb{R}^n$

e.g. $T: \mathbb{R}^2 \to \mathbb{R}^3$, $T(\vec{v}) = \begin{bmatrix} v_2 \\ v_1 + v_2 \\ v_1 - v_2 \end{bmatrix}$

standard basis: $\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

$$T\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad T\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \quad \therefore A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$Av = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0v_1 + v_2 \\ v_1 + v_2 \\ v_1 + (-v_2) \end{bmatrix} = \begin{bmatrix} v_2 \\ v_1 + v_2 \\ v_1 - v_2 \end{bmatrix}$$

$T[\underline{0}]$ must be $[\underline{0}]$ otherwise not a linear map.
- can use that to show if linear map!

Kernel / Image

$$V \xrightarrow{L} W$$

Kernel: Set of vectors in space V that map to zero vector in W
- $\ker(L)$ is subspace in V

Image: S is subspace in V $[S \subseteq V]$
- $L(S)$ = image of S
- $L(S)$ is a subspace in W

Linear Independence $\to$ If none of vectors can be written in terms of the others

Dimension $\to$ number of elements in basis [provided vector space has basis]

Y2 content:

Direct Sum of Subspaces

For two subspaces U, V in vector space W
subspace $X = U \oplus V$ iff:

1) $X = U + V$, then for $w \in W$,
$w = u + v$ for $u \in U, v \in V$

2) $U \cap V = \{\underline{0}\}$

Direct sum is when two subspaces do not overlap

Example:
Suppose $V = \mathbb{R}^3$ [3-d vector space]
$X = \{(x, y, z) : x + 2y - z = 0\}$
(1) $Y = \{(-y, y, z) : y, z \in \mathbb{R}\}$ $\Rightarrow$ Not a Direct sum
Is it true that $V = X \oplus Y$ → check (ii) [$X \cap Y$] first then (i) [$V = X + Y$]

Consider $X \cap Y$: $x + 2y = z$, $x = -y$ $\Leftrightarrow y = z$, $y = -x$ $\Leftrightarrow y = z = -x$

e.g. $(1, -1, -1) \in X \cap Y$ ✓
$\neq (0, 0, 0)$ $\therefore$ NOT a direct sum

(2) $Y = \{(y, y, y) : y \in \mathbb{R}\}$ $\Rightarrow$ Is a direct sum

1. Consider $X \cap Y$: $x + 2y = z$, $x = y = z$ $\Leftrightarrow 3y = z$ $\Leftrightarrow 3z = z$ $\Leftrightarrow z = 0$

$\therefore x = 0, y = 0$
$(0, 0, 0) \in X \cap Y$
$= (0, 0, 0)$ $\therefore$ satisfies 2^{nd} condition

2. Consider $v = (x, y, z) \in \mathbb{R}^3$
= something from X + something in Y
Y = basis = $(1, 1, 1)$
X = basis = $\{(1, 0, 1), (0, 1, 2)\}$ [$(-2, 1, 0)$ can be made from the basis]

$$V = \alpha(1,0,1) + \beta(0,1,2) + \gamma(1,1,1)$$

1st: $x = \alpha + \gamma \rightarrow \alpha = x - \gamma$

2nd: $y = \beta + \gamma \rightarrow \beta = y - \gamma$

3rd: $z = \alpha + 2\beta + \gamma \rightarrow z = x - \gamma + 2(y - \gamma) + \gamma$

$= x + 2y - 2\gamma$

$\gamma = \frac{x+2y-z}{2}$

$\beta = \frac{z-x}{2}$, $\alpha = \frac{x-2y+z}{2}$ $\quad \therefore V = X \oplus Y$

(iii). $Y = \{(x,0,x) : x \in \mathbb{R}\}$ $\Rightarrow$ Fail condition 2

Consider $X \cap Y \rightarrow x + 2y - z = 0 \Rightarrow x = z$

$\boxed{\begin{matrix} x = z \\ y = 0 \end{matrix}}$

e.g. $(1,0,1) \in X \cap Y$

$\neq (0,0,0)$

$\therefore$ NOT a direct sum

<u>Direct Sum</u>

Consider two subspaces U, V of a vector space W.

We can always define their sum $U + V = \{u + v, u \in U, v \in V\}$.

If we also have $U \cap V = \{0\}$, then we say that U, V are in <u>direct sum.</u>

We denote by $U \oplus V$ this property, and the sum set $U + V$.

<u>Example:</u> $U = span(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}); V = span(\begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix})$ $\qquad \mathbb{R}^3 = U \oplus V$

To prove that $W = U \oplus V$, we need to show that:

$W = U + V$ Each vector of W can be written as u+v

$U \cap V = \{0\}$ The only vector u and v have in common in 0

<u>Theorem</u>

If $(u_1, \ldots, u_n)$ is a basis of U and $(v_1, \ldots, v_m)$ a basis of V, then $U \oplus V$ if and only if the family $(u_1, \ldots, u_n, v_1, \ldots, v_m)$ is linearly independent.

This becomes a basis for $U \oplus V$

The bases can be merged together in the case of a direct sum!

<u>Theorem:</u>

$W = U \oplus V$ if and only if For every $w \in W$ there exist unique vectors $u \in U$ and $v \in V$ such that $u + v = w$.

Similar definition for n subspaces:

$V = V_1 \oplus V_2 \oplus \cdots \oplus V_n$ $\qquad$ For every $v \in V$ there exist unique vectors $v_i \in V_i$ such that $v = v_1 + \ldots v_n$.

("if and only if")

Linear Maps

$f: V \to W$, $g: U \to V$

can be composed into $f \circ g: U \to W$

Linear Maps have properties of:

- Associativity $\quad f \circ (g \circ h) = (f \circ g) \circ h$
- Identity $\quad f \circ id = id \circ f = f$
- Distributivity $\quad f \circ (g+h) = f \circ g + f \circ h$

$$(g+h) \circ f = g \circ f + h \circ f$$

Rank Nullity Theorem

$T: V \to W$ is a linear transformation

$$\mathrm{Ker}(T) = \{v \in V \mid T(v) = 0\}$$ - vectors in V that map to $\underline{0}$ under the linear transformation

$\ulcorner \mathrm{Ker}(T) \subseteq V \lrcorner$

$$\mathrm{Im}(T) = \{T(v) \mid v \in V\}$$

$\ulcorner \mathrm{Im}(T) \subseteq W \lrcorner$

- can use subspace test to show that Ker / Im are subspaces

rank/nullity

$$\dim(\mathrm{Ker}(T)) := \mathrm{NULLITY}(T)$$

Ker(T) has a basis [as its a valid subspace]
where number of vectors in basis will get the dimension

$$\dim(\mathrm{Im}(T)) := \mathrm{RANK}(T)$$

Rank Nullity Theorem

Let V, W be finite dimensional vector spaces. Let $T: V \to W$ be a linear transformation:

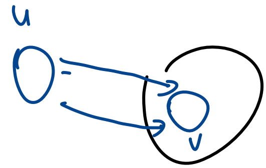
$$\dim(V) = \mathrm{rank}(T) + \mathrm{nullity}(T)$$

Properties of Linear Maps

Injective:

A map ρ is injective if:

$$\forall x, y \in U, \quad \rho(x) = \rho(y) \Rightarrow x = y$$



$$\rho : U \to V$$

A map is injective $\iff$ $\ker(\rho) = \{\underline{0}\}$

Proof: Injective $\Rightarrow \ker(\rho) = \{0\}$

Assume ρ injective. $\exists x \in \ker(\rho)$

$$\rho(x) = \underline{0}, \quad \rho(\underline{0}) = \underline{0}$$

$$x = \underline{0} \in \ker(\rho)$$

$\ker(\rho) = \{0\} \Rightarrow$ Injective

Assume $\ker(\rho) = \{0\}$. $\exists x, y \in U$ st $\rho(x) = \rho(y)$

$$\rho(x) - \rho(y) = \rho(x - y) = \underline{0}$$, hence $x - y \in U$

As $x - y = 0 \Rightarrow x = y \Rightarrow \rho$ is injective QED

If ρ is injective, then $\dim(U) \leq \dim(V)$

↳ use rank-nullity theorem

$$\dim(U) = \dim(\operatorname{Im}(\rho)) + \dim(\ker(\rho))$$
$$= \dim(\operatorname{Im}(\rho)) + \dim(\{0\}) = \dim(\operatorname{Im}(\rho)) \leq \dim(V)$$

Given a map ρ and matrix M representing it in basis B

ρ is injective $\iff$ columns of M are linearly independent

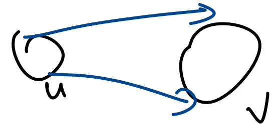
↳ if column with only 0s → no injectivity!

Surjective maps:

$p: U \to V$

Map p is surjective if: [entire space V can be reached by p]

$\forall v \in V, \exists u \in U$ st $p(u) = v$



p is surjective $\iff \operatorname{Im}(p) = V$

$\quad \quad \iff \dim(U) \geq \dim(V)$

- **Example:** Are the following maps surjective?

$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \mapsto \begin{bmatrix} y \\ 2x \end{bmatrix}$ Yes. Every vector can be obtained as an image: $\begin{bmatrix} \frac{1}{2}y \\ x \\ 0 \end{bmatrix} \mapsto \begin{bmatrix} x \\ y \end{bmatrix}$

$\begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} x \\ x \end{bmatrix}$ No. $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ is not part of the image

$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \mapsto \begin{bmatrix} y \\ 0 \\ x \end{bmatrix}$ No. $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ is not part of the image

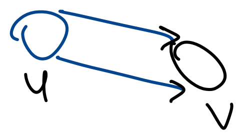
- Any map can be made surjective by restricting its codomain to its image

Bijective Maps

Map p is bijective $\iff$ p is injective and surjective

$\iff \dim(U) = \dim(V)$

$\iff$ there exists a map $p^{-1}: V \to U$ $[p \circ p^{-1} = p^{-1} \circ p = id]$



Bijective map is also known as isomorphism

[$\forall x, y \in U \Rightarrow x = y$

$\forall v \in V, \exists u \in U$ st $p(u) = v$]

- We must cover each element of V exactly once

Bijection = One-to-one correspondence

$\forall v \in V$, there exists a unique $u \in U$ st $p(u) = v$

- **Example:** Are the following maps bijective?

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \mapsto \begin{bmatrix} y \\ 2x \end{bmatrix} \qquad \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} 2y \\ x \end{bmatrix} \qquad \begin{bmatrix} x \\ y \\ z \end{bmatrix} \mapsto \begin{bmatrix} x \\ x+y \\ x+y+z \end{bmatrix}$$

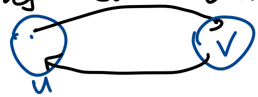
No	Yes	Yes
Dimensions not matching	Injective and surjective	Harder to see why
(it is not injective)		

Isomorphism

$\dim(U) = \dim(V) \iff$ there exists an isomorphism between them = they are isomorphic

$\iff$ equal spaces → same behaviour

Bijective maps are isomorphic



Bijective maps can be represented by SQUARE matrices (must be square) i.e. same dimension

Matrix is invertible iff its corresponding map is bijective.

Endomorphisms

$g: U \to U$, assuming U is finite-dimensional

g is then:
- Injective
- Surjective
- Bijective

} at same time.

Can prove g is bijective by showing $\ker(g) = \{0\}$

g injective $\iff \ker(g) = \{0\}$

$\operatorname{Im}(g) = U$, shows g is surjective

Ker / Im subspaces of same space

g bijective $\iff$ maps a basis to basis

Injective → preserve linear independence

Surjective → cover all vectors in space

- <u>Change of bases:</u>
  We consider an endomorphism g of the space U, represented by a matrix M in the basis B_1. How to represent it in another basis B_2?

- Construct the matrix P whose columns are the coordinates of the vectors of B_2 in the basis B_1. Invert it and calculate P^{-1}.

- Finally, $\boxed{P^{-1}M\,P}$ is the matrix representing g in the basis B_2.

- P is called the transition matrix, or change of basis matrix.

Change of bases:

basis B_1 → basis B_2
P: in our basis B_1
$P^{-1}MP\vec{v}$ will get the matrix representation in basis B_2

basis 1 → basis 2

change of basis matrix

$$\underbrace{P^{-1}\, M}_{\text{transformed matrix in her language}}\, P\, \vec{v}$$

inverse change of basis matrix

transformation matrix in our language [basis B_1] aka endomorphism

→ 3B1B way of explaining change of bases as diff languages

- To get the transformation matrix for g in the basis B_2
- A way of representing the same vector, where the vector is a linear combination of basis B_1 vectors, as vector in terms of basis B_2.

M has columns made of the new basis vectors

- Orders matters in matrix multiplication

- <u>Remark:</u>
  If we have $P^{-1}M\,P = N$, then $PNP^{-1} = M$.

  In other words, if P is the transition matrix from B_1 to B_2, then P^{-1} is the transition matrix from B_2 to B_1.

$$\begin{tikzcd} B_2 \arrow[r, bend left, "P"] & B_1 \arrow[l, bend left, "P^{-1}"] \end{tikzcd}$$

[Inversing matrix]

Step 1: replace every entry by its minor

Given an entry in a 3 by 3 matrix, cross out its entire row and column, and take the determinant of the 2 by 2 matrix that remains (this is called the **minor**).

In our example, this gives us

$$\begin{pmatrix} (-4)\times 1-(-2)\times 4 & 1\times 1-(-2)\times(-3) & 1\times 4-(-4)\times(-3) \\ (-3)\times 1-(-2)\times 4 & 0\times 1-(-2)\times(-3) & 0\times 4-(-3)\times(-3) \\ (-3)\times(-2)-(-2)\times(-4) & 0\times(-2)-(-2)\times 1 & 0\times(-4)-(-3)\times 1 \end{pmatrix} = \begin{pmatrix} 4 & -5 & -8 \\ 5 & -6 & -9 \\ -2 & 2 & 3 \end{pmatrix}.$$

Step 2: change some of the signs

We now change the signs of some of the minors, according to the pattern

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

thus creating what's called the **matrix of cofactors**. In our case, this is

$$\begin{pmatrix} 4 & 5 & -8 \\ -5 & -6 & 9 \\ -2 & -2 & 3 \end{pmatrix}.$$

Step 3: transpose

We now transpose the matrix of cofactors. In our case, we get

$$\begin{pmatrix} 4 & -5 & -2 \\ 5 & -6 & -2 \\ -8 & 9 & 3 \end{pmatrix}.$$

Step 4: divide by the determinant

Finally, we divide by the determinant of the original matrix. In our case, the determinant is

$$\det\begin{pmatrix} 0 & -3 & -2 \\ 1 & -4 & -2 \\ -3 & 4 & 1 \end{pmatrix} = 0\times\det\begin{pmatrix} -4 & -2 \\ 4 & 1 \end{pmatrix} + 3\times\det\begin{pmatrix} 1 & -2 \\ -3 & 1 \end{pmatrix} - 2\times\det\begin{pmatrix} 1 & -4 \\ -3 & 4 \end{pmatrix} = 1,$$

so the inverse is simply

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

- <u>Some remarkable endomorphisms:</u>

- In the vector space $\mathbb{R}^n$ with the basis $(u_1, \ldots, u_n)$
  the following matrix represents a projection on the subspace $span(u_1, u_2, \ldots, u_k)$

- It can be seen as the identity matrix on the desired subspace, and 0 everywhere else.
- It is not injective

$$\begin{array}{c} u_1, u_2, \ldots, u_k \\ \begin{bmatrix} 1 & 0 & & \cdots & 0 \\ 0 & \ddots & & & \\ \vdots & & 1 & & \vdots \\ & & & & \\ 0 & & \cdots & & 0 \end{bmatrix} \end{array}$$

- The vectors $u_1, u_2, \ldots, u_k$ don't have to be the first ones in the basis, but we can always change the basis to obtain this form.

- In the vector space $\mathbb{R}^n$ with the basis $(u_1, \ldots, u_n)$
  the following matrix represents a scaling or homothety with factor a.

- It is equal to $a \times I_n$
- It represents a mapping $v \mapsto av$
- It is invertible if $a \neq 0$

$$\begin{bmatrix} a & 0 & \cdots & 0 \\ 0 & \ddots & & 0 \\ \vdots & & & \vdots \\ 0 & & \cdots & a \end{bmatrix}$$

- It has the same form in every basis!

$$P^{-1}(aI_n)P = a \times P^{-1}I_nP = a \times P^{-1}P = aI_n$$

(Counter clockwise rotation matrix is always seen as change of basis)

Matrix Reductions

Eigenvector:

- Example of an invariant subspace

$Mr = \lambda r \Rightarrow r$ is eigenvector, when $r \neq 0$
for eigenvalue λ and matrix M

- When endomorphism acts like a scaling

$$\left[\begin{array}{l} Mr = \lambda r \\ Mr - \lambda r = 0 \\ Mr - \lambda I_n r = 0 \\ (M - \lambda I_n) r = 0, \ v \neq 0 \\ \det(M - \lambda I_n) = 0 \end{array}\right.$$

λ is eigenvalue $\Longleftrightarrow$ $(M - \lambda I_n)$ is not injective or surjective or bijective

$\Longleftrightarrow \det|M - \lambda I| = 0 \rightarrow$

$Ker(M - \lambda I_n)$ = set of eigenvectors associated with λ and vector $\underline{0}$.
= invariant subspace [eigenspace]
↳ subspace of vector space on which matrix M acts.
↳ closed under addition / scalar mult.
↳ use Ker to find eigenvectors

- **Theorem:**

Suppose $\lambda_1, \ldots, \lambda_m$ are distinct eigenvalues of f and $v_1, \ldots, v_m$ are corresponding eigenvectors. Then $v_1, \ldots, v_m$ are linearly independent.

- **Corollary:**
An endomorphism $f : U \to U$ cannot have more than $dim(U)$ eigenvectors.

Eigenspace

→ All eigenvectors that correspond to some eigenvalue λ

$$E_\lambda = N(M - \lambda I_n)$$

If rows of a matrix are not independent, they will have eigenvalue 0, thus not invertible matrix.

If sum of rows are same, an obvious eigenvector for matrix is $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

For a matrix M, λ is an eigenvalue $\Leftrightarrow$ $(M-\lambda I_n)$ NOT bijective
not invertible

$\det(M-\lambda I_n)=0$, whenever λ is an eigenvalue
↳ obtain characteristic polynomial + find eigenvalues

- Roots of characteristic polynomial χ_M = eigenvalues of M
  ↳ splitting polynomial becomes crucial

χ^M = degree n.

- **Examples:**
  Let's find some eigenvectors of $M = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$
- $M-(-1)I_2 = \begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix}$

As expected, this matrix is not injective, ie -1 is an eigenvalue

- $Ker(M-(-1)I_2) = \{ \begin{bmatrix} y \\ x \end{bmatrix}, x+y=0 \text{ and } -2x-2y=0\} = span(\begin{bmatrix} 1 \\ -1 \end{bmatrix})$

Eigenvector associated with -1

- For a linear map T we can define the following objects:

$$T^0 = I, \qquad T\circ T = T^2, \qquad T\circ T\circ\cdots\circ T = T^n$$

- The same applies to matrices.
- Hence if we have a polynomial $P = a_0 + a_1X + \cdots + a_nX^n$
  it makes sense to define $P(T) = a_0I + a_1T + \cdots + a_nT^n$
- We can even define $\chi_M(M)$

Cayley-Hamilton Theorem

- **Cayley-Hamilton Theorem**
  For a square matrix M, we have:
  $$\chi_M(M) = 0$$

- The combination defined by the characteristic polynomial cancels M.
- **Example:** $M = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \qquad \chi_M = X^2 + 3X + 2 = (X+1)(X+2)$

$$\chi_M(M) = M^2 + 3M + 2I_2 = \begin{bmatrix} -2 & -3 \\ 6 & 7 \end{bmatrix} + 3\begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} + 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

- As eigenvectors are linearly independent, we can use them to build a new basis of eigenvectors

- We will use our results to reduce our matrix to a simpler form. First, recall the standard decomposition of a polynomial. If $\mathbb{F} = \mathbb{R}$, we know that we can write:

$$\boxed{\chi_M = a \times (P_1)^{h_1} \times \cdots \times (P_n)^{h_n} \times (X - \lambda_1)^{k_1} \times \cdots \times (X - \lambda_m)^{k_m}}$$

$P_1, \ldots, P_n$ distinct irreducible polynomials of degree 2 and $\lambda_1, \ldots, \lambda_n$ distinct eigenvalues

- <u>Theorem:</u> the vector space can be decomposed as:

$$\boxed{U = Ker(P_1^{h_1}(M)) \oplus \cdots \oplus Ker(P_n^{h_n}(M)) \oplus Ker((M - \lambda_1 I)^{k_1}) \oplus \cdots \oplus Ker((M - \lambda_m I)^{k_m})}$$

- If $\mathbb{F} = \mathbb{C}$ it is easier because every polynomial splits:

$$\boxed{\chi_M = a \times (X - \lambda_1)^{k_1} \times \cdots \times (X - \lambda_m)^{k_m}}$$

$\lambda_1, \ldots, \lambda_n$ distinct eigenvalues

- Theorem: the vector space can be decomposed as:

$$\boxed{U = Ker((M - \lambda_1 I)^{k_1}) \oplus \cdots \oplus Ker((M - \lambda_m I)^{k_m})}$$

$Ker(M - \lambda I) \neq Ker(M - \lambda I)^k$

$Mx = \lambda x, \qquad (M - \lambda I)^k x = 0$

- multiplicity 1 for all roots

Matrix Diagonalization - depends on root/multiplicity

<u>Theorem:</u>
A matrix of dimension n with n distinct eigenvalues can be expressed as a diagonal matrix in the basis of eigenvectors.

$$\begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & \lambda_n \end{bmatrix}$$

New basis:
$e_1 \in Ker(M - \lambda_1 I)$
$e_2 \in Ker(M - \lambda_2 I)$
$\vdots$
$e_n \in Ker(M - \lambda_n I)$

Each eigenspace is stable, and the mapping behaves like a scaling

- Method to change to basis of eigenvects
- Take a matrix and write as product of matrices

- <u>Proof & Method:</u>
  If we have n distinct eigenvalues, then we know: $\chi_M = a \times (X - \lambda_1) \times \cdots \times (X - \lambda_n)$
  Hence $U = Ker(M - \lambda_1 I) \oplus \cdots \oplus Ker(M - \lambda_n I)$
- We have n distinct eigenspaces, the corresponding eigenvectors form a basis of U.

- Only works for matrices with unique eigenvalues

$D = P^{-1} M P$ $\left[D = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \right]$

- Diagonal matrix D made up of eigenvalues of matrix M
- Matrix P made up of eigenvectors of M
- Matrix M = $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$

M can be obtained by $M = PDP^{-1}$

- Example: $M = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$

- We start by calculating the characteristic polynomial:

$$\chi_M = \det \begin{bmatrix} 3-X & 2 \\ 2 & 3-X \end{bmatrix} = X^2 - 6X + 5$$

- With usual methods we obtain:

$$\chi_M = (X-5)(X-1)$$

Therefore M has 2 distinct eigenvalues, and can be diagonalised.

- Now we study the eigenspaces to find a basis of eigenvectors:

$$v = \begin{bmatrix} x \\ y \end{bmatrix} \in Ker(M - 5I_2) \iff Mv = 5v$$

$$\iff 3x + 2y = 5x \text{ and } 2x + 3y = 5y$$

$$\iff x = y$$

$$\iff v \in span \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

- As expected, we find a 1-dimensional eigenspace for the eigenvalue 5.
- Same thing for the eigenvalue 1:

$$v = \begin{bmatrix} x \\ y \end{bmatrix} \in Ker(M - I_2) \iff Mv = v$$

$$\iff 3x + 2y = x \text{ and } 2x + 3y = y$$

$$\iff x = -y$$

$$\iff v \in span \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

- Now we have both eigenvectors and they form a basis.

- Example: $M = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$

eigenvalue 5

eigenvectors

← eigenvalue 1

order of eigenvalues / eigenvectors MATTERS

- Matrix for change of basis: $P = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$

- We express our endomorphism in this new basis:

$$D = P^{-1} M P = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix}$$

Diagonalised!

2^{nd} case:

If eigenvalues not distinct [multiplicity in χ_M]

We cannot always diagonalise. But:

Theorem:
If the characteristic polynomial has the following form:

$$\chi_M = (X - \lambda_1)^{k_1} \times \cdots \times (X - \lambda_m)^{k_m}$$

Then there exists a basis in which our matrix is triangular.

$$\begin{bmatrix} \lambda_1 & a_{12} & \cdots & a_{1n} \\ 0 & \lambda_1 & & a_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & & & \lambda_n \end{bmatrix}$$

- In this triangular form, the eigenvalues are on the diagonal.

- Note that (as we have seen) this is always possible if $\mathbb{F} = \mathbb{C}$

- <u>Second best case</u>: what if eigenvalues are not distinct? (multiplicities in χ_M)

  To find the basis in which the matrix is triangular, we have to find a basis for each space $Ker(M - \lambda I)^k$, since we have the following:

$$U = Ker((M - \lambda_1 I)^{k_1}) \oplus \cdots \oplus Ker((M - \lambda_m I)^{k_m})$$

- We haven't said it is impossible to diagonalise! Just not guaranteed in general.
  Example: $\begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ Non distinct eigenvalues, but diagonalised

Worst case: Irreducible Factors - can't find enough eigenvalues

- Here, it is possible that our matrix cannot be expressed as a triangular or diagonal matrix.
- <u>Example</u>: $M = \begin{bmatrix} 0 & 1 & 1 \\ -2 & 0 & 5 \\ 0 & 0 & 3 \end{bmatrix}$ $\qquad \chi_M = (3 - X)(X^2 + 2)$

One eigenvalue found, with multiplicity 1

Irreducible part (no root)

In $\mathbb{R}$,we cannot make M diagonal, or triangular.

How to find eigenvalues: [obvious ways than solving kernel]
- Use properties of determinant
- If matrix in diagonal form, eigenvalues are the diagonal
- If coefficients on each row add up to k, then k is eigenvalue associated with vector $\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

e.g. $\begin{array}{c} 3 \\ 3 \\ 3 \end{array} \begin{bmatrix} 1 & 3 & -1 \\ 2 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

Differential Calculus / Multivariate Functions

LDEs

$$a_n y^{(n)} + a_{n-1} y^{(n-1)} + \dots + a_1 y' + a_0 y = 0$$

Theorem:
If two functions are solutions of a homogeneous linear differential equation, then every linear combination of them is also a solution.

Proof idea: by linearity of the equation.

In particular, the sum of solutions is a solution.

- Consider a simple case first:
$$ay^{(2)}(x) + by^{(1)}(x) + cy(x) = 0$$
- We remember that for first order equations, the solution involves exponentials.
- Let us try $e^{\lambda x}$. The equation becomes: $a\lambda^2 e^{\lambda x} + b\lambda e^{\lambda x} + ce^{\lambda x} = 0$
- In particular for $x = 0$, we obtain $\boxed{a\lambda^2 + b\lambda + c = 0}$
- If λ is solution of this polynomial equation, then $e^{\lambda x}$ is a solution of the DE.

$$\left[\begin{aligned} y &= e^{3t} \\ y' &= 3e^{3t} = 3y \\ y'' &= 9e^{3t} = 9y \end{aligned}\right]$$

E.g. $ay'' + by' + cy = 0$ $\qquad y =$

Suggest $y = e^{\lambda x}$
$y' = \lambda e^{\lambda x}$
$y'' = \lambda^2 e^{\lambda x}$
$\Rightarrow a\lambda^2 e^{\lambda x} + b\lambda e^{\lambda x} + ce^{\lambda x} = 0$
$e^{\lambda x}(a\lambda^2 + b\lambda + c) = 0$

never zero, so $a\lambda^2 + b\lambda + c = 0$

↳ AUXILIARY EQUATION

- Get auxiliary solution = polynominal equation $\mathcal{X}[x] = 0$ and find roots
  If roots distinct: 2 solutions of λ
  repeated roots: 1 solution $[y = ce^{\lambda x} + cxe^{\lambda x}]$

Higher Order

General Method: $a_n y^{(n)} + a_{n-1} y^{(n-1)} + \dots + a_1 y' + a_0 y = 0$

- How to find roots? → use auxiliary equation
- If single multiplicity:
$\mathcal{X}[x] = Q(x)(x - \lambda)$, $\lambda \in \mathbb{R}$, λ not root of Q
→ factor theorem
$x \mapsto Ae^{\lambda x}$, $A \in \mathbb{R}$ [use: $e^{\lambda x}$]

You add up the roots at the end as general solution

- If multiplicity k (real root)

$$\chi(x) = (x-\lambda)^k Q(x), \quad \lambda \in \mathbb{R}, \ \lambda \text{ not root}$$

- repeated roots k times

use: $A_0 e^{\lambda x}, \lambda e^{\lambda t}, \ldots, \lambda^{k-1} e^{\lambda t}$

- If complex roots: $[\lambda = a \pm ib \in \mathbb{C}]$

use:
$e^{ax}\cos(bx)$
$e^{ax}\sin(bx)$

$\lambda = a+ib$

$$e^{\lambda x} = e^{(a+ib)x}$$
$$= e^{ax} e^{ibx}$$
$$= e^{ax}[\cos(bx) + i\sin(bx)]$$
$$= e^{ax}\cos(bx) + ie^{ax}\sin(bx)$$

Euler's Formula $[e^{i\theta} = \cos(\theta) + i\sin(\theta)]$

$$Re(e^{\lambda x}) = e^{ax}\cos(bx), \quad Im(e^{\lambda x}) = e^{ax}\sin(bx)$$

If $\lambda = a - ib$,

$$e^{\lambda x} = e^{ax}\cos(-bx) + i\sin(-bx)$$
$$= e^{ax}\cos(bx) - i\sin(bx)$$

General solution
- Contain sum of all solutions obtained from roots
  - as equation is linear so sum of solutions is also a solution

e.g. $y^{(4)} - 4y^{(3)} + 5y^{(2)} - 4y^{(1)} + 4y = 0$

TURN TO AUXILIARY: $X^4 - 4X^3 + 5X^2 - 4X + 4 = 0$

$$= (X-2)^2(X^2+1) = 0$$

$X = 2$ with mult 2

$= Ae^{2x} + Bxe^{2x}$

$X = \pm i$ with mult 1

$X = i$

$= C\cos x + D\sin x$

$X = -i$

$= E\cos x - F\sin x$

$[E\cos(-x) + F\sin(-x) = E\cos x - F\sin x$

$\therefore GE = Ae^{2x} + Bxe^{2x} + C\cos x + D\sin x$

- Particular solution can be found if initial conditions specified

Multivariate Functions

- $f(x,y) = z$

  $\mathbb{R}^2 \to \mathbb{R}$

- Represented as a surface defined by points $(x, y, f(x,y))$ in 3D plane

Integrating 2 variables → double integral [measure volume under surface]

$$\int_X \int_Y f(x,y)\,dy\,dx = \int_X \left(\int_Y f(x,y)\,dy \right) dx$$

- Order of integration in general matters but not for well-behaved functions

- <u>Theorem: Fubini's Integration</u>

  If the integral $\int_X \int_Y |f(x,y)|dydx$ is well-defined, then we have:

  Must have a finite value

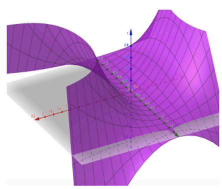
$$\int_X (\int_Y f(x,y)dy)dx = \int_Y (\int_X f(x,y)dx)dy = \iint_{X\times Y} f(x,y)dxdy$$

Either order works

- Example:

  $X = [0,1] \quad Y = [0, \frac{\pi}{2}]$

$$\begin{aligned}\iint_{X\times Y} x^2\cos(y)dxdy &= \int_{x=0}^{1}\left(\int_{y=0}^{\frac{\pi}{2}} x^2\cos(y)dy\right)dx \\ &= \int_{x=0}^{1}\left[x^2\sin(y)\right]_{y=0}^{\frac{\pi}{2}}dx \\ &= \int_{x=0}^{1}(x^2\times 1)dx \\ &= \left[\frac{1}{3}x^3\right]_{x=0}^{1} dx \\ &= \frac{1}{3}\end{aligned}$$



- We can use level curves to represent functions with 2 variables [sets of point with fixed altitude]
  - curve is like a section of surface, taken at regular intervals of altitude

- <u>Example</u>

  What are the level curves of $f(x,y) = x^2 - y$

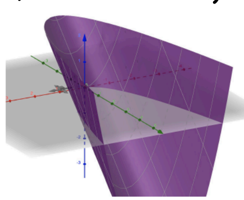
- Fixed altitude means:

$$f(x,y) = x^2 - y = b \text{ for } b \in \mathbb{R}$$

  What does this equation define?

- In this case we can write it as: $y = x^2 - b$

  Hence the level curves are <u>parabolas</u>.



Let's see it in real time

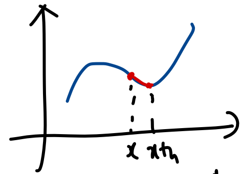
Differentiating multivariate functions [partial differentiation – PDE]

↳ differentiate independently for each variable : partial derivatives

$$f(x,y)$$

$$\frac{\partial f}{\partial x}(x,y) \qquad \frac{\partial f}{\partial y}(x,y)$$

- when differentiating for x, y acts as constant



approx gradient as h gets closer to x

example:

1 variable diff:

$$\frac{df}{dx} = \lim_{h\to 0}\left(\frac{f(x+h)-f(x)}{h}\right)$$

how f changes in x direction

2 variable diff: f depends on x and y

$z = f(x,y)$

$$\frac{\partial f}{\partial x} = \lim_{h\to 0}\left(\frac{f(x+h,y)-f(x,y)}{h}\right)$$ ⇒ add h to x [change x only] value of y is constant

notation for PD

$$\frac{\partial f}{\partial y} = \lim_{h\to 0}\left(\frac{f(x,y+h)-f(x,y)}{h}\right)$$ ⇒ how f acts when y changes only value of x is constant

when doing partial derivative of x, treat y as constant, and vice versa

Example: $f(x,y) = xy^2 + yx^3$

$$\frac{\partial f}{\partial x} = y^2 + 3yx^2 \quad , \quad \frac{\partial f}{\partial y} = 2xy + x^3$$

$$\frac{\partial^2 f}{\partial x} = 0 + 6yx \quad , \quad \frac{\partial^2 f}{\partial y^2} = 2x$$

Quotient / Product / Chain Rule apply in Partial Differentiation

$$f(x,y) = (x^2+2x)\sin(x^2+y) + e^{y-2x}$$

$$\frac{\partial f}{\partial x} = (2x+2)\sin(x^2+y^2) + (x^2+2x)\cos(x^2+y)\cdot 2x - 2e^{y-2x}$$

$$\frac{\partial f}{\partial y} = (x^2+2x)\cos(x^2+y) + e^{y-2x}$$

$f(x,y)$

$f_x = 0 \Rightarrow f = g(y)$

$f_{yx} = 0 \Rightarrow f = G(y) + h(x)$

$\frac{\partial f}{\partial x} = f_x = 0$

$f = y, y^2, \sin y, e^{y+y^2+y^4}$ [f doesn't depend on x]

[Any function with y only will get $\frac{\partial f}{\partial x}$ be 0]

$= g(y)$ where g can be any function taking in y and return anything with y

$f_{yx} = 0$

$f_y = g(y)$ ↙ integration constant replaced as integration function [function of x]

$f = \int g(y)\, dy + h(x)$

$f = G(y) + h(x)$ [$f_{yx} = 0$]

$u(x,y)$, $u_{xx} - u = 0$

$u = a(y)e^x + b(y)e^{-x}$

$u_x = a(y)e^x - b(y)e^{-x}$

$u_{xx} = a(y)e^x + b(y)e^{-x}$

$\frac{d^2u}{dx^2} - u = 0$

$u \sim e^{nx} \Rightarrow n^2 - 1 = 0$, $n = \pm 1$

$u = \underbrace{a(y)}_{\text{Integration function}} e^x + b(y)e^{-x}$

[Similar to IC in ODE] → can pretend to solve ODE and convert to PDE

Gradient of multivariate function

$f(x,y)$ will get you a surface in 3D plane [not line]

↳ to find gradient, perform partial derivatives on both variables [partial diff definition enables other variable to be constant]

- We can gather both partial derivatives and form the *gradient vector*

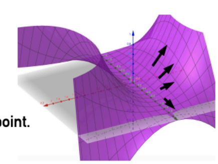
$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x}(x,y) \\ \frac{\partial f}{\partial y}(x,y) \end{bmatrix}$$

- This vector takes different values for each (x,y). It is the 2-dimensional equivalent of a derivative.

- Example: $f(x,y) = x^2\cos(y)$

$$\frac{\partial f}{\partial x}(x,y) = 2x\cos(y) \qquad \frac{\partial f}{\partial y}(x,y) = -x^2\sin(y)$$

$$\nabla f = \begin{bmatrix} 2x\cos(y) \\ -x^2\sin(y) \end{bmatrix}$$



- As with the 1-dimensional derivative, the gradient shows the variation of the function at the considered point.

Finding Critical Point (Stationary)

- Gradient for both partial derivatives have to be 0.
- Critical points consist of stationary points, maxima, minima

e.g. $f(x,y) = x^2 + 2xy - y^2 + y^3$

$$\frac{\partial f}{\partial x} = 2x + 2y = 0$$
$$\frac{\partial f}{\partial y} = 2x - 2y + 3y^2 = 0$$

$y = -x$

$-4y + 3y^2 = 0$, $y[3y - 4] = 0$, $y = 0, 4/3$

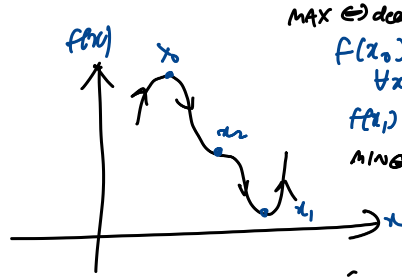
$x = 0, -4/3$

critical points: $(0,0)$, $(-4/3, 4/3)$

Classifying critical points

3 critical points:

- MAXIMUM e.g. $z = -x^2 - y^2$
- MINIMUM e.g. $z = x^2 + y^2$
- SADDLE e.g. $z = x^2 - y^2$
  [like a pringles]



MAX ⟺ decreasing [$f''(x) > 0$]

$f(x_0) > f(x)$
$\forall x$ 'nearby' $\neq x_0$

$f(x_1) < f(x)$
$\forall x$ 'nearby' $\neq x_1$

MIN ⟺ increasing
$f''(x) > 0$

2^{nd} derivative = tells about the change in gradient

$0 \to +$ number $\to$ MIN [$f''(x) > 0$]

$0 \to -$ $\to$ MAX [$f''(x) < 0$]

2^{nd} order partial derivatives

↳ For MAX: all need to be < 0 . $f_{xx} < 0$, $f_{yy} < 0$, $f_{xy} = f_{yx} < 0$

↳ MIN: all need to be > 0 , $f_{xx} > 0$, $f_{yy} > 0$, $f_{xy} = f_{yx} > 0$

- **Theorem:**
  If the partial derivatives of order 2 are all continuous near (x,y), then we have:
$$\frac{\partial^2 f}{\partial xy}(x,y) = \frac{\partial^2 f}{\partial yx}(x,y)$$

$$f_{xy} = \frac{\partial^2 f}{\partial xy} = f_{yx} = \frac{\partial^2 f}{\partial yx}$$

Taylor's Expansion of $f(x,y)$ =

$$f(x,y) \approx f(x_0,y_0) + (x-x_0) f_x(x_0,y_0) + (y-y_0) f_y(x_0,y_0) + \dots$$
$$+ \frac{(x-x_0)^2}{2} f_{xx}(x_0,y_0) + \frac{(y-y_0)^2}{2} f_{yy}(x_0,y_0)$$
$$+ \frac{2(x-x_0)(y-y_0)}{2} f_{xy}(x_0,y_0) + \dots$$

all nearby points

$$\underbrace{f(x,y) - f(x_0,y_0)}_{LHS} \approx \frac{(x-x_0)^2}{2} f_{xx}(x_0,y_0) + \frac{(y-y_0)^2}{2} f_{yy}(x_0,y_0) + (x-x_0)(y-y_0) f_{xy}(x_0,y_0)$$

$$\approx \frac{1}{2f_{xx}}\left[\left(\underbrace{(x-x_0)f_{xx} + (y-y_0)f_{xy}}_{\geq 0}\right)^2 + \left(\underbrace{f_{xx}f_{yy} - f_{xy}^2}_{\text{DISCRIMINANT}}\right)\underbrace{(y-y_0)^2}_{\geq 0}\right]$$

for $f_{xx} \neq 0 \Rightarrow$ MAX : $LHS < 0 \Leftrightarrow f_{xx} < 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$

MIN : $LHS > 0 \Leftrightarrow f_{xx} > 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$

SADDLE : $f_{xx}f_{yy} - f_{xy}^2 < 0$

① calculate discriminant $[f_{xx}f_{yy} - f_{xy}^2]$ [if < 0, SADDLE; > 0, check f_{xx} → if < 0 : MAX; > 0 : MIN]

Example:

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

$\frac{\partial f}{\partial x} = 2x + 2y$, $\frac{\partial f}{\partial y} = 2x - 2y + 3y^2$

Critical points: $(0,0)$, $(-4/3, 4/3)$

$f_{xx} = 2$

$f_{yy} = -2 + 6y$

f_{xy} [f_x differentiate with respect to y] $= 2$

$f_{yx} = 2$ $\therefore f_{yx} = f_{xy}$

Discriminant:

At $(0,0)$: $f_{xx}f_{yy} - f_{xy}^2$

$= 2(-2+6(0)) - 2^2$

$= -8$ $\therefore$ Saddle

At $(-4/3, 4/3)$ $= 2(-2+6(4/3)) - 2^2$

$= 8$ $\therefore$ Minimum

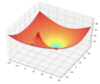
$f_{xx} > 0$ → min

- **Classification of Stationary Points:**

If $\det(\mathcal{H}_f(x,y)) > 0$

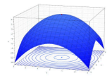
if $\frac{\partial^2 f}{\partial xx}(x,y) > 0$

Local Minimum



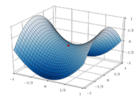
if $\frac{\partial^2 f}{\partial xx}(x,y) < 0$

Local Maximum



If $\det(\mathcal{H}_f(x,y)) < 0$

Saddle Point



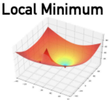
- **Classification of Stationary Points:**

If $\det(\mathcal{H}_f(x,y)) > 0$

$f_{xx} > 0$ Same rules as for 1 variable! $f_{xx} < 0$

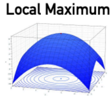
if $\frac{\partial^2 f}{\partial xx}(x,y) > 0$

Local Minimum



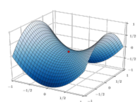
if $\frac{\partial^2 f}{\partial xx}(x,y) < 0$

Local Maximum



If $\det(\mathcal{H}_f(x,y)) < 0$

Saddle Point



How to understand and remember these rules:

- The Hessian is a symmetrical matrix (in continuous cases).
  We will see later that it means it can be diagonalised!
- Picture the diagonal form of the Hessian at a stationary point (a,b) : $\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$
- Diagonalising means that, if we look at it the right way (in the basis of eigenvectors), it's almost as if:

$$\frac{\partial^2 f}{\partial xx}(a,b) = \lambda_1 \qquad \frac{\partial^2 f}{\partial xy}(a,b) = \frac{\partial^2 f}{\partial yx}(a,b) = 0 \qquad \frac{\partial^2 f}{\partial yy}(a,b) = \lambda_2$$

It's not strictly true to write it like this. For it to be true we'd have to change x,y according to the eigenvectors.

How to understand and remember these rules:

- Finally, note that $\det \mathcal{H}_f(a,b) = \lambda_1 \times \lambda_2$
- Therefore the sign of the determinant informs us on whether the eigenvalues have the same sign or opposite signs.
- You don't have to diagonalise the Hessian! But that explains the method.

For function $\mathbb{R}^n \to \mathbb{R}^m$

V. Higher Dimensions

UCL

- A quick look at functions $\mathbb{R}^n \to \mathbb{R}^m$
- We denote such a function by $f : (v) \mapsto (f_1(v), \dots f_m(v))$
  with $v = (x_1, \dots, x_n)$

The equivalent of first order derivative is:

The Jacobian Matrix

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Each row contains the gradient of a component of the function

Euclidean Spaces

↳ vector spaces with an inner product

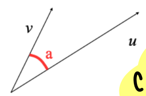
Inner Product

↳ simple case: dot product

- The simplest case of inner product is the dot product seen in $\mathbb{R}^2$
- It is defined as: $u \cdot v = \begin{bmatrix} x \\ y \end{bmatrix} \cdot \begin{bmatrix} z \\ t \end{bmatrix} = x \times z + y \times t$
- Or equivalently $u \cdot v = \|u\| \times \|v\| \times \cos(a)$
- Some important remarks:

$$u \cdot v = v \cdot u \qquad |u \cdot v| \leq \|u\| \times \|v\|$$

"length" of the vector



$$\cos(a) = \frac{u \cdot v}{\|u\|\|v\|}$$

$\|u\|$ = length of vector [pythagoras]

Inner Product

- Consider a vector space U,

Inner product: $U \times U \to \mathbb{R}$ [billinear form]

To test if billinear:

(i) Linearity in 1st position → [need to check if inner product]

$$\langle \underbrace{\alpha_1 v_1 + \alpha_2 v_2}_{\in U}, \underbrace{v_3}_{\in U} \rangle = \alpha_1 \langle v_1, v_3 \rangle + \alpha_2 \langle v_2, v_3 \rangle$$

[$\forall \alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$
$\forall v_1, v_2, v_3 \in V$]

(ii) Linearity in 2nd position:

$$\langle v_1, \alpha_2 v_2 + \alpha_3 v_3 \rangle = \alpha_2 \langle v_1, v_2 \rangle + \alpha_3 \langle v_1, v_3 \rangle$$

Example of billinear form: dot product in $\mathbb{R}^n$ vector space [Euclidean Space]
[vector space over real numbers]

$$\langle \underline{u}, \underline{v} \rangle = (u_1, u_2 \dots u_n) \cdot (v_1, v_2 \dots v_n)$$
$$= u_1 v_1 + u_2 v_2 + \dots u_n v_n$$

Consider: $\langle \alpha_1 \underline{u} + \alpha_2 \underline{v}, \underline{w} \rangle = (\alpha_1 \underline{u} + \alpha_2 \underline{v}) \cdot \underline{w}$

$$= (\alpha_1 u_1 + \alpha_2 v_1, \dots, \alpha_1 u_n + \alpha_2 v_n) \cdot (w_1 \dots w_n)$$
$$= \alpha_1 u_1 w_1 + \alpha_2 v_1 w_1 + \dots + \alpha_1 u_n w_n + \alpha_2 v_n w_n$$
$$= \alpha_1 (u_1 w_1 + \dots + u_n w_n) + \alpha_2 (v_1 w_1 + \dots + v_n w_n)$$

Inner product defined as dot product $= \alpha_1 \langle \underline{u}, \underline{w} \rangle + \alpha_2 \langle \underline{v}, \underline{w} \rangle$

$$= \alpha_1 \langle \underline{w}, \underline{u} \rangle + \alpha_2 \langle \underline{w}, \underline{v} \rangle \quad \therefore \text{(i) satisfied}$$

proving that satisfies (ii)

Consider $= \langle \underline{u}, \alpha_2 \underline{v} + \alpha_3 \underline{w} \rangle = u \cdot (\alpha_2 \underline{v} + \alpha_3 \underline{w})$

$= (u_1 \dots u_n) \cdot (\alpha_2 v_1 + \alpha_3 w_1 \dots \alpha_2 v_n + \alpha_3 w_n)$

$= \alpha_2 v_1 u_1 + \alpha_3 w_1 u_1 \cdot \quad \alpha_2 v_n u_n + \alpha_3 w_n u_n$

$= \alpha_2 (v_1 u_1 + \dots v_n u_n) + \alpha_3 (w_1 u_1 + \dots + w_n u_n)$

$= \alpha_2 \langle \underline{u}, \underline{v} \rangle + \alpha_3 \langle \underline{u}, \underline{w} \rangle$

$= \alpha_2 \langle \underline{v}, \underline{u} \rangle + \alpha_3 \langle \underline{w}, \underline{u} \rangle \therefore$ (ii) satisfied

If a billinear form satisfies $\langle \underline{u}, \underline{v} \rangle = \langle \underline{v}, \underline{u} \rangle$ then it is symmetric

Billinear form is positive definite $\Leftrightarrow \langle \underline{v}, \underline{v} \rangle \geq 0, \forall \underline{v} \in V$
and $= 0 \Leftrightarrow \underline{v} = \underline{0}$

$\langle \underline{v}, \underline{v} \rangle = (v_1 \dots v_n) \cdot (v_1 \dots v_n)$

$= v_1^2 + \dots + v_n^2 \geq 0$

↳ only can be zero if v is zero vector

Inner Product Space [Euclidean Space)

Real vector space equipped with inner product :

[Map $\langle -, - \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$]

i) BILINEAR FORM : $\langle \alpha_1 v_1 + \alpha_2 v_2, v_3 \rangle = \alpha_1 \langle \underline{v_1}, \underline{v_3} \rangle + \alpha_2 \langle \underline{v_2}, \underline{v_3} \rangle$

ii) SYMMETRIC : $\langle \underline{u}, \underline{v} \rangle = \langle \underline{v}, \underline{u} \rangle$

iii) POSITIVE DEFINITE : $\langle \underline{v}, \underline{v} \rangle \geq 0$ and $= 0 \Leftrightarrow \underline{v} = \underline{0}$
[positive except for zero vector]

If $\langle \underline{u}, \underline{v} \rangle = 0$ then u, v are orthogonal

A real vector space with an inner product is known as Euclidean Space (over real numbers) → dot product is inner product in Euclidean space

- <u>In finite dimension:</u>

Consider the case $U = \mathbb{R}^n$. The canonical inner product is:

$$\boxed{\langle u, v \rangle = \langle \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}, \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \rangle = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n}$$

Example: $\langle \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ -1 \\ 5 \end{bmatrix} \rangle = 1 \times 3 + 0 \times (-1) + 2 \times 5 = 13$

- This inner product works for any vector space isomorphic to $U = \mathbb{R}^n$.
- For instance recall that polynomials can be represented as vectors:

$$ax^2 + bx + c \longleftrightarrow \begin{bmatrix} c \\ b \\ a \end{bmatrix}$$

We see the coefficients as coordinates of the vector

- Therefore $\mathbb{R}^2$ is isomorphic to $\mathbb{R}_{n-1}[X]$
- We can then calculate an inner product of polynomials as we did for vectors:

$$\langle a_2x^2 + a_1x + a_0, b_2x^2 + b_1x + b_0 \rangle = a_2b_2 + a_1b_1 + a_0b_0$$

Inner Products can be represented by matrices

Theorem:

- Every inner product can be represented by a matrix
- Every matrix that is symmetric and positive definite represents an inner product.

- Example:

  The matrix $M = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$ represents an inner product, which we denote by $\langle \_, \_ \rangle_M$

- M is symmetric.
- Proving that M is positive definite is not easy - see end of lecture.
- Example of calculation:

$$\langle \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ 4 \end{bmatrix} \rangle_M = [1,2]\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \end{bmatrix} = [1(2)+2(1), 1(1)+2(3)]\begin{bmatrix} 3 \\ 4 \end{bmatrix} = [4\ 7]\begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

$$= 40$$

In infinite dimension:

Inner product on S [space of continuous functions on interval $[0,1]$]

↳ vector space of infinite dimension

$$\langle f, g \rangle = \int_0^1 f(x)\,g(x)\,dx \qquad [f, g \in \mathbb{P}_n[x]]$$

Example of proof:

(ii) consider $\langle g, f \rangle = \int_0^1 g(x)\,f(x)\,dx = \int_0^1 f(x)g(x)\,dx = \langle f, g \rangle$ commutative

(i) $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \int_0^1 (\alpha_1 f_1 + \alpha_2 f_2) g\,dx$

$= \alpha_1 \int_0^1 f_1 g\,dx + \alpha_2 \int_0^1 f_2 g\,dx$

$= \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$

linearity

(iii) $\langle f, f \rangle = \int_0^1 f^2 \, dx \geq 0$ [always be ≥ 0) positive definite as $f^2 \geq 0$ [$=0$ if $f=0$]

$\therefore$ $\langle f, g \rangle = \int_0^1 f(x) g(x) \, dx$ is an INNER PRODUCT

e.g. $\langle \sqrt{x}, x^2 \rangle = \int_0^1 \sqrt{x} \cdot x^2 \, dx$

$$= \int_0^1 x^{5/2} \, dx$$

$$= \left[\tfrac{2}{7} x^{7/2} \right]_0^1$$

$$= 2/7$$

NOTE: If <

Norms - measuring vectors

For $u \in \mathbb{R}^n$, we define:

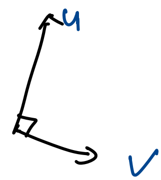
$$||u|| = \sqrt{\langle u, u \rangle} = \sqrt{u_1^2 + u_2^2 + \dots + u_n^2}$$

[Standard norm)

"Euclidean

Orthogonal when $\langle \underline{u}, \underline{v} \rangle = 0$

[right angle)



- The idea of a norm is to “measure” vectors, i.e. generalise the idea of length.

Definition:
In a vector space with an inner product, $\langle -, - \rangle$, the Euclidean norm of a vector v is:

$$\|v\| = \sqrt{\langle v, v \rangle}$$

- Because of positive definiteness, we immediately have:

$$\|v\| \geq 0$$
$$\|v\| = 0 \Rightarrow v = 0$$

This allows to prove that a vector is equal to 0

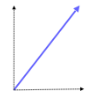
- **Examples:**

- In $\mathbb{R}$, the canonical inner product is the multiplication.
  The corresponding norm is the absolute value:

$$\langle x, y\rangle = x \times y \qquad \|x\| = \sqrt{\langle x, x\rangle} = \sqrt{x^2} = |x|$$

- In $\mathbb{R}^2$, we have:

$$\left\|\begin{bmatrix} x \\ y \end{bmatrix}\right\| = \sqrt{\left\langle \begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} x \\ y \end{bmatrix} \right\rangle} = \sqrt{x^2+y^2}$$



In continuous functions on $[0,1]$:

$$\|f\| = \sqrt{\int_0^1 f^2\, dx}$$

e.g. $\|x+1\| = \sqrt{\int_0^1 (x+1)^2\, dx} = \sqrt{\int_0^1 x^2+2x+1\, dx}$

$= \sqrt{\left[\frac{1}{3}x^3+x^2+x\right]_0^1}$

$= \sqrt{7/3}$

- Norm provides size of functions

Cauchy - Schwarz Inequality

Theorem: Cauchy-Schwarz Inequality
For two vectors in a Euclidean Space, we have:

$$\boxed{|\langle u, v\rangle| \leq \|u\| \times \|v\|}$$

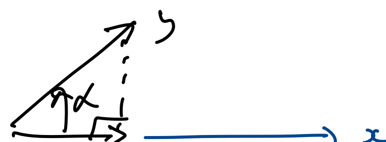
- Makes sense with the usual dot product in the plane $u \cdot v = \|u\| \times \|v\| \times \cos(a)$
- The inner product cannot be larger than the product of the lengths.

$|\cos(a)| \leq 1$

Let $(X, \langle -, - \rangle)$ be inner product space and $\|x\| = \sqrt{\langle x, x\rangle}$

Then for all $x, y \in X$

$$|\langle x, y\rangle| \leq \|x\| \cdot \|y\|$$



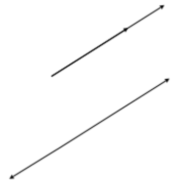
$|\langle x, y\rangle| = \|x\|\,\|y\| \iff x, y$ linearly independent

- It is easy to show that the equality is obtained when vectors are colinear:

$$u \in span(v) \Rightarrow |\langle u, v \rangle| = \|u\| \times \|v\|$$

We will prove it in problem class

- In the plane, it corresponds to the case $cos(a) = 1$ or -1.



Distance

- **Distance:**
For a space U, a distance or metric is a function $\quad U \times U \to \mathbb{R}$

  which verifies: $\quad (u, v) \mapsto d(u, v)$

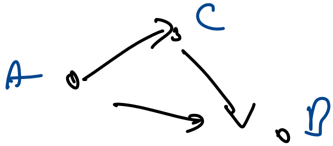
  - Symmetry $\quad \forall u, v, \quad d(u,v) = d(v,u)$
  - Positive definiteness $\quad \forall u, v, \quad d(u,v) \geq 0 \quad$ and $\quad d(u,v) = 0 \Leftrightarrow u = v$
  - Triangle Inequality $\quad \forall u, v, w, \quad d(u,v) \leq d(u,w) + d(w,v)$

- A space with a distance is called metric space.

Doesn't have to be a vector space!

Triangle Inequality:

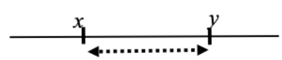
$$\forall u, v, w, \quad d(u,v) \leq d(u,w) + d(w,v)$$



Can't go faster via C if going from $A \to B$

- **Examples:**

- In $\mathbb{R}$, an obvious distance is $\quad d(x,y) = |x - y|$



- In any space, we can define a trivial metric:

$$d(x,y) = 1 \text{ if } x \neq y, \quad \text{and } d(x,y) = 0 \text{ if } x = y$$

- In the space of words of length n, we can define the Hamming distance:

$$d(\texttt{what}, \texttt{that}) = 1 \qquad d(\texttt{that}, \texttt{they}) = 2$$

Proving Triangle Inequality:

- **Proof:**
  Symmetry and positive definiteness are easy. Let's prove the triangle inequality.

$$
\begin{aligned}
\|u-v\|^2 = \|u-v+w-w\|^2 &= \|(u-w)+(w-v)\|^2 \\
&= \langle (u-w)+(w-v), (u-w)+(w-v) \rangle \\
&= \langle u-w, u-w \rangle + \langle w-v, w-v \rangle + 2\langle u-w, w-v \rangle \\
&= \|u-w\|^2 + \|w-v\|^2 + 2\langle u-w, w-v \rangle \\
&\leq \|u-w\|^2 + \|w-v\|^2 + 2\|u-w\| \times \|w-v\| \\
&\leq (\|u-w\| + \|w-v\|)^2
\end{aligned}
$$

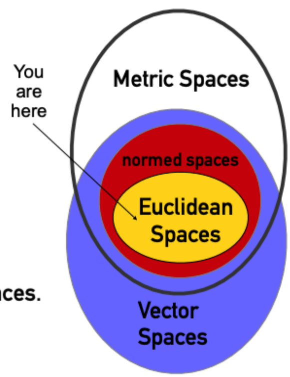
Finally since the square root function is continuous and increasing, it preserves the inequality and we obtain:

$$
\begin{aligned}
\sqrt{\|u-v\|^2} &\leq \sqrt{(\|u-w\| + \|w-v\|)^2} \\
\|u-v\| &\leq \|u-w\| + \|w-v\| \\
d(u,v) &\leq d(u,w) + d(w,v)
\end{aligned}
$$

The story so far:

- Some vector spaces can be provided with an inner product
- This always gives them a metric
- Not all vector spaces are Euclidean

  (But counterexamples are very complicated)
- Not all metric spaces are Euclidean, or Vector spaces.



Orthogonality

2 vectors v, u are orthogonal if $\langle u, v \rangle = 0$

— Depends on inner product [have to be 0]

Pythagorean Theorem:
If two vectors u, v of a Euclidean space U are orthogonal, then we have:

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2$$

← Example case → Pythagoras Theorem

Proof:

$$
\begin{aligned}
\|u+v\|^2 &= \langle u+v, u+v \rangle \\
&= \langle u,u \rangle + \langle v,v \rangle + \overset{=0}{\langle u,v \rangle} + \overset{=0}{\langle v,u \rangle} \\
&= \|u\|^2 + \|v\|^2
\end{aligned}
$$

[$\langle u, v \rangle = 0$]
↑ to satisfy orthogonality!

Orthonormal

Orthonormal Set: $\{v_1, v_2, \ldots, v_n\}$ is inner product space V st:

$$\langle v_i, v_j \rangle = \delta_{ij} = \begin{cases} 1 & , i=j \\ 0 & , i \neq 0 \end{cases} \rightarrow \|v_i\| = 1$$

v_i has [length 1] $\forall i$

$[= \langle v_i, v_i \rangle]$

Take inner product between any 2 vectors is 1 if same vector, 0 otherwise

\# ELEMENTS = dim of $V \Rightarrow$ BASIS

Lemma: $\{v_1, \ldots, v_n\}$ orthonormal set, then $\{v_1, v_2, \ldots v_n\}$ is linearly independent

Proof:

Consider $\alpha_1 v_1 + \cdots \alpha_n v_n = 0$

$$\begin{aligned} 0 = \langle v_i, 0 \rangle &= \langle v_i, \alpha_1 v_1 + \ldots + \alpha_n v_n \rangle \\ &= \alpha \langle v_i, v_1 \rangle + \ldots + \alpha_n \langle v_i, v_n \rangle \\ &= \alpha_i \underbrace{\langle v_i, v_i \rangle}_{=1} \\ &= \alpha_i \end{aligned}$$

$[\langle v_i, v_n \rangle = 0$ if $i \neq n]$

$\therefore \alpha_i = 0, \forall i \Rightarrow$ linear independence

If we have linearly independent set that has same dimension as vector space spanning therefore set has a basis

Orthonormal basis

A family of vectors $(u_1, \ldots, u_n)$ is orthonormal if they are pairwise orthogonal and if they all verify $\| u_i \| = 1$. ← length 1 [unit vectors → can normalize to get unit vectors]

If they form a basis, it is called an orthonormal basis.

Examples:

The canonical basis of $\mathbb{R}$ is orthonormal w.r.t the canonical inner product.

A less canonical orthonormal family is $\begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

Corollary:
Every orthonormal family of vectors is linearly independent.

- An orthonormal basis is a convenient tool:

Theorem:
If $(u_1, u_2, \ldots, u_k)$ is an orthonormal basis then we have for every vector v:

$$v = \langle v, u_1\rangle u_1 + \langle v, u_2\rangle u_2 + \cdots + \langle v, u_k\rangle u_k$$

$$\|v\| = \langle v, u_1\rangle^2 + \langle v, u_2\rangle^2 + \cdots + \langle v, u_k\rangle^2$$

The coordinates in an orthonormal basis are obtained with inner products

In Euclidean space U, orthogonal complement of subspace V:

$$V^\perp = \{ u \in U \mid \forall v \in V, \langle u, v\rangle = 0 \}$$

- Set of all vectors orthogonal to all vectors of V.
- $V^\perp$ is subspace and $U^\perp = \{0\}$, $\{0\}^\perp = U$

$$V \oplus V^\perp = U$$

- A vector space can always be subdivided into orthogonal subspaces

- **Example**
  What is the orthogonal complement of $V = span(\overset{v_1}{\begin{bmatrix}1\\1\\0\end{bmatrix}}, \overset{v_2}{\begin{bmatrix}-1\\0\\1\end{bmatrix}})$ in $\mathbb{R}^3$?

- First, note that for any vector we have $u \in V^\perp \Leftrightarrow \langle u, v_1\rangle = 0 \text{ and } \langle u, v_2\rangle = 0$

verify this

$$\langle u, v_1\rangle = 0 \Leftrightarrow \langle \begin{bmatrix}x\\y\\z\end{bmatrix}, \begin{bmatrix}1\\1\\0\end{bmatrix}\rangle = 0 \Leftrightarrow x + y = 0$$

$$\langle u, v_1\rangle = 0 \Leftrightarrow \langle \begin{bmatrix}x\\y\\z\end{bmatrix}, \begin{bmatrix}-1\\0\\1\end{bmatrix}\rangle = 0 \Leftrightarrow -x + z = 0$$

$x+y=0$, $x=-y$,
$-x+z=0$, $x=z$ $\therefore u \in \begin{bmatrix}1\\-1\\1\end{bmatrix}$

$u = \begin{bmatrix}c\\-c\\c\end{bmatrix}$, $\forall c \in \mathbb{R}$

Orthogonal Projection

As $V \oplus V^{\perp} = U$, we can rewrite each vector of U as:

$$u = v + v', \quad v \in V, \; v' \in V^{\perp}$$

Transformation mapping u to v the orthogonal projection on V $\Rightarrow$ denoted by P_V

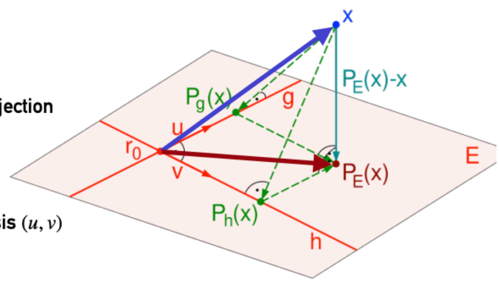
Theorem:
If $(v_1, \dots v_n)$ is an orthonormal basis of V, then we have:

$$P_V(u) = \langle u, v_1 \rangle v_1 + \langle u, v_2 \rangle v_2 + \cdots + \langle u, v_k \rangle v_k$$

One can easily show that $u - P_V(u)$ is orthogonal to all vectors of V.

By direct sum

- Illustration
- Here we have $E = span(u, v)$.
- We calculate the orthogonal projection of the vector x on E.
- This projection is a vector of E.
- It can be decomposed in the basis (u, v) by using the previous formula:
- $P_E(x) = \langle x, u \rangle u + \langle x, v \rangle v$



The projection is like the "shadow" of the vector

- Example: calculate the projection of $u = \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix}$ on $V = span(\underset{v1}{\begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}}, \underset{v2}{\begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}})$

$$P_V = \langle u, v_1 \rangle v_1 + \langle u, v_2 \rangle v_2$$

$$= (3 \cdot 0 + 1 \cdot -1 + 4 \cdot 0) v_1 + (3 \cdot 1/\sqrt{2} + 1 \cdot 0 + 4 \cdot \tfrac{1}{\sqrt{2}}) v_2$$

$$= -v_1 + 7/\sqrt{2}\, v_2$$

$$= -\begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} + 7/\sqrt{2} \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ 1/\sqrt{2} \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 7/2 \\ 0 \\ 7/\sqrt{2} \end{bmatrix}$$

$$= \begin{bmatrix} 7/2 \\ 1 \\ 7/2 \end{bmatrix}$$

Spectral Theorem

Square matrix M is orthogonal if its columns form an orthonormal basis [orthonormal set of vectors]

$$MM^T = M^T M = I \quad \text{, therefore } M^T = M^{-1}$$

- Symmetric matrices in $\mathbb{R}$ can be diagonalised in orthonormal basis

e.g. $D = P^T M P^T$

$M^{-1} = M^T$ if M is orthogonal [easy to find inverse then]

Consequences of the Spectral Theorem:

- Eigenvectors of a symmetric matrix are orthogonal.
- Matrices representing inner products have eigenvalues and can be represented as a diagonal matrix in an orthonormal basis.
- We can now easily characterise such matrices.

- Consider a matrix M representing an inner product.
- We diagonalise it in an orthonormal basis $(e_1, \ldots, e_n)$ and obtain $D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & \lambda_n \end{bmatrix}$
- If we write the coordinates of our vectors in this new basis, we now have:

$$\langle u, v \rangle = u^T D v = \lambda_1 u_1 v_1 + \ldots + \lambda_n u_n v_n$$

Almost like our canonical inner product!

A square matrix represents an inner product iff:

1) SYMMETRIC
2) only has STRICTLY POSITIVE eigenvalues

positive definite

e.g. $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ represents inner product because:

- Symmetric
- Eigenvalues are 1, 3

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

$\lambda = 1, 3$

Numerical Methods

Newton's method for zeros [Newton Raphson method]

- Find estimate of x when $f(x) = 0$.
- Assuming f is differentiable

Algorithm

- Define a value d for the desired precision.
- Make a first guess x_0
- While $f(x_n) > d$, iterate by using the formula $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$
- Return x_n

If

Theorem:
If the sequence defined by $x_{n+1} = x_n - \dfrac{f(x_n)}{f'(x_n)}$ has a limit l, then it verifies:

$$f(l) = 0$$

Proof idea:
Any zero of f is a fixed point of the function $x \mapsto x - \dfrac{f(x)}{f'(x)}$ which defines the recursive sequence.

might not be correct

Probabilities

Random variables

↳ Map outcomes of random processes to numbers

↳ quantify outcomes

- Consider a random variable X.
  The *target space* *T* of X is the set of values that X can take.

- When T is continuous, X is a continuous random variable.

i.e. $X = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases}$

Examples:

Discrete → distinct / separate value

Result of a die — target space: $\{1, 2, 3, 4, 5, 6\}$

Continuous → any value in interval

Lifespan of a lightbulb — target space: $[0, +\infty[$

- For a continuous random variable, we are interested in calculating the probability of intervals rather than individual values.

$$Pr(X \geq a) \qquad Pr(a \leq X \leq b) \qquad Pr(X < b)$$

- We will consider the target space of continuous random variables to be $T \subseteq \mathbb{R}^n$

- There are conditions from the theory of *measurable spaces*, but we will only consider well-behaved cases.

- For more information, see Measure Theory for Probabilities

You calculate probability of intervals when dealing with DISCRETE random variables.

Probability Density Function

- A function is called a *probability density function* if it verifies:
  1. $f : T \to \mathbb{R}$
  2. $\forall x \in T, f(x) \geq 0$
  3. $\int_T f(x)dx$ exists and $\int_T f(x)dx = 1$

Can be a multi-dimensional integral!

We will start by studying it in dimension 1, then higher.

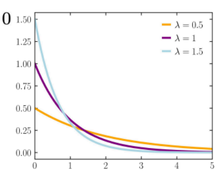
must be true when verifying if prob. density function

$$\boxed{\int_{-\infty}^{\infty} f(x)\,dx = 1}$$

or if range $a \leq x \leq b$

$$\int_a^b f(x)dx = 1$$

- Example: $T = [0, +\infty[\quad f : x \mapsto \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$



- We verify the conditions:

$\forall x \in T, f(x) \geq 0$ ✓

$$\int_T f(x)dx = \int_0^{+\infty} \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_0^{+\infty} = e^0 = 1$$ ✓

↳ known as exponential density function.

- Is the condition $\int_T f(x)dx = 1$ so important?

- Yes, because it is significant for probabilities.

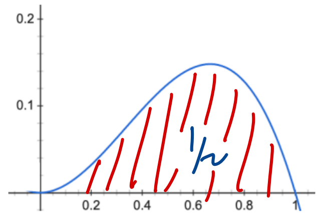
- But any integrable function can be scaled / normalised to verify this condition!

$$e^{-\lambda x} \longrightarrow \lambda e^{-\lambda x}$$

Does not verify the condition

verifies the condition

- **Example:**
Is the following a probability density function? $T = [0,1]$ $\quad f(x) = x^2(1-x)$



$$\int_0^1 x^2(1-x)\,dx$$
$$= \int_0^1 x^2 - x^3\,dx$$
$$= \left[\tfrac{1}{3}x^3 - \tfrac{1}{4}x^4\right]_0^1$$
$$= \left[\tfrac{1}{3} - \tfrac{1}{4} - 0\right]$$
$$= \left[\tfrac{1}{12}\right] \neq 1$$

$\therefore$ NOT a probability density function

But can normalise and define:

$$g(x) = 12 f(x)$$

to get a probability density function

For random variable X associated with probability density function f:

$$Pr(a \le x \le b) = \int_a^b f(x)\,dx$$

$\Rightarrow$ Find probability between interval $[a,b]$

For single value e.g. $Pr(x=a) = \int_a^a f(x)\,dx = 0$

$\int_T f(x)\,dx = 1 \Rightarrow$ total of probabilities MUST equal to 1.

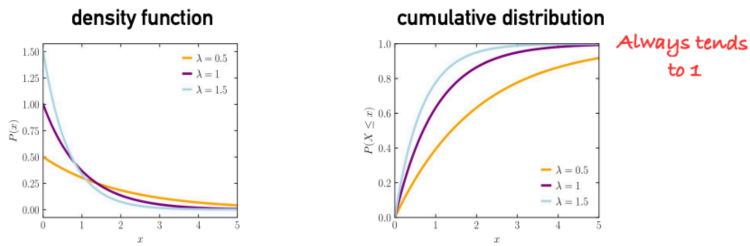
Cumulative Distribution Function

For random variable X, cumulative distribution function is:

$$F_X(x) = Pr(X \leq x)$$
$$= \int_{-\infty}^{x} f(x)\ dx$$

↳ If X is associated with a probability density function

- **Example:**
For the exponential distribution we have:



Expected Value

For random variable X and probability density function f, expected value:

$$E(X) = \int_T x f(x)\ dx$$

For discrete random variables, $E(X) = \sum_T x Pr(X = x)$

- Expected value ⇒ represents mean / average value
- Sometimes written as μ.

Covariance:

Covariance of 2 variables X, Y:

$$Cov(X,Y) = E(X \times Y) - E(X)E(Y)$$ → How 2 variables are dependent on one another

Variance of 1 variable = covariance of itself: → How variable is spread around average

$$Var(X) = E(X^2) - [E(X)]^2$$

Standard Deviation

= Square root of variance $\left[\sqrt{Var(X)} = \sigma(X) \right]$

$X \times Y \Rightarrow$ new variable, whose values is products of values of X and Y with corresponding probabilities

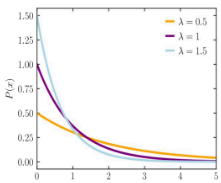
How to calculate variance:

$$Var(x) = \int_T x^2 f(x)\,dx - \mu^2 = \int_T (x-\mu)^2 f(x)\,dx$$

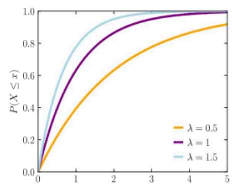
<u>Some Classic Density Functions</u>

- Exponential Distribution

$$T = [0, +\infty[\quad f : x \mapsto \begin{cases} \lambda e^{-\lambda x} \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases}$$



Density



Cumulative Distribution

$$E(X) = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$

$$E(X) = \int_0^\infty \lambda x e^{-\lambda x}\,dx$$

$$\left.\begin{array}{ll} u = x & v' = \lambda e^{-\lambda x} \\ u' = 1 & v = -e^{-\lambda x} \end{array}\right\} \text{Integration by parts}$$

$$= \left[-xe^{-\lambda x}\right]_0^\infty - \int_0^\infty -e^{-\lambda x}\,dx$$

$$= \left[\lim_{x\to\infty}\left[-xe^{-\lambda x}\right] - 0e^{-\lambda 0}\right] + \left[-\tfrac{1}{\lambda} e^{-\lambda x}\right]_0^\infty$$

$$= 0 + \left[\lim_{x\to\infty}\left[-\tfrac{1}{\lambda}e^{-\lambda x}\right] - \left[-\tfrac{1}{\lambda}e^{0}\right]\right]$$

$$= 0 + \left[0 + \tfrac{1}{\lambda}\right]$$

$$= \tfrac{1}{\lambda}$$

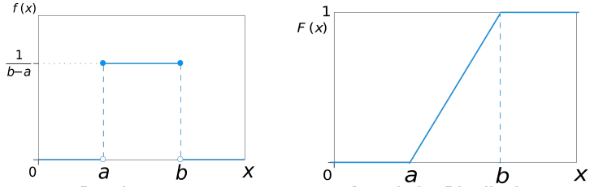
Uniform Distribution

Some Classic Density Functions

- **Uniform Distribution**

$$T = [a, b] \qquad f : x \mapsto \begin{cases} \frac{1}{b-a} \text{ if } a \le x \le b \\ 0 \text{ otherwise} \end{cases}$$

(constant)



Density   Cumulative Distribution

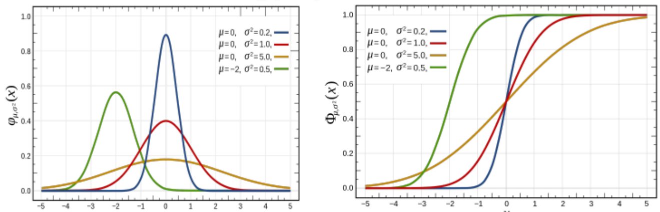
Expected value
And variance
in problem
class

Normal Distribution

Some Classic Density Functions

$$T = \mathbb{R} \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- **Normal Distribution** $\mathcal{N}(\mu, \sigma^2)$ aka Gaussian



Density   Cumulative Distribution

$$E(X) = \mu$$
$$Var(X) = \sigma^2$$

Given as
parameters

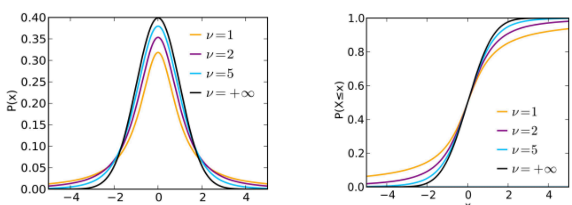
- Need to use calculator / table to find probability as there is no explicit formula for antiderivative for f

T-Distribution

Some Classic Density Functions

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

- **Student's t-Distribution**



Density   Cumulative Distribution

This is the 'Gamma'
function

Similar to a normal
distribution, but not
exactly the same

- Used to construct confidence intervals for means
- Derived from sum of normal distributions

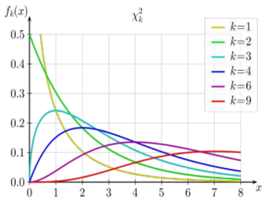
χ^2 Distribution

- Distribution of sum of normally distributed random variables

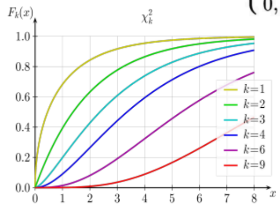
Some Classic Density Functions

- **Chi-squared Distribution**

$$f(x;\,k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\,\Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$



Density



Cumulative Distribution

- used to construct confidence intervals

Joint Probabilities

→ 2 random variables → joint probabilities:

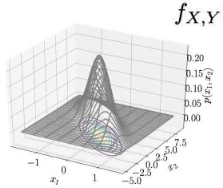
$$Pr(X = a,\ Y = b) = Pr(X = a \cap Y = b)$$

Target space of joint variable = product of both target spaces
→ Can define multi-dimensional distributions
In discrete, table used; but in continuous, prob distribution used.

- Joint distribution of 2 continuous variables:

Instead of a table, we get a 2-variable probability distribution

$f_{X,Y}(x,y)$



We will focus on continuous variables

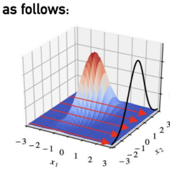
$$Pr(a \le X \le b, c \le Y \le d) = \int_a^b \int_c^d f_{X,Y}(x,y)dydx$$

↳ ah yea, double integral. fun

- From the joint probability distribution $f_{X,Y}(x,y)$, we can find the probability distribution of the individual variables X,Y.
- These are called *marginal distributions* and obtained as follows:

$$\boxed{\begin{aligned} f_X(x) &= \int f_{X,Y}(x,y)\,dy \\ f_Y(y) &= \int f_{X,Y}(x,y)\,dx \end{aligned}}$$

Integrate along the other variable

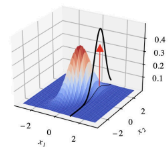


You can obtain conditional probabilities from joint distribution

$$\boxed{Pr(X \leq x | Y = y) = \frac{\int_{-\infty}^{x} f_{X,Y}(t, y)dt}{f_Y(y)}}$$

Integrate on the corresponding 'slice'

normalise



Law of Large Numbers

Consider a sequence of *independent* and *identically distributed* random variables $X_1, X_2, \ldots, X_n$. Assume that $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$.
We define the sequence of averages $\overline{X_n} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$. Then we have:

$$\boxed{\overline{X_n} \to \mu}$$

The average converges to the expected value.

- $Pr((\lim \overline{X_n}) = \mu) = 1$ when $n \to \infty$

$X_n = X_1 + X_2 + \ldots + X_n$

$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots X_n) \to \mu$

$E(X_1) = E(X_2) \ldots = E(X_n) = \mu$

- Probability distribution of $\overline{X}_n$ = Normal distribution spread out around the average μ → hence the CLT

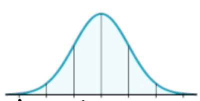
Central Limit Theorem

Consider a sequence of *independent* and *identically distributed* random variables $X_1, X_2, \ldots, X_n$.

$\forall i,\ X_i \sim N(\mu, \sigma^2)$

- Assume that $E(X_1) = E(X_2) = \cdots = E(X_n) = \mu$
- Assume $Var(X_1) = Var(X_2) = \cdots = Var(X_n) = \sigma^2$.
- Consider $\overline{X_n} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$.

Then for n large enough, the distribution of $\overline{X_n}$ is arbitrarily close to a Normal Distribution $\boxed{\mathcal{N}(\mu, \frac{\sigma^2}{n})}$



- When repeating a sampling/random process, we expect a normal distribution

[Distribution of $\sqrt{n}(\overline{X_n} - \mu)$ converges to $N(0, \sigma^2)$]

- Used to approx a distribution [valid for $n \geq 30$]
  ↳ CLT used to approx average or sum

EXAMPLE: Approximating Binomial Law

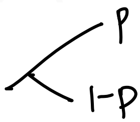
- Application: approximate the Binomial Law

We start by applying the CLT to a single Bernoulli variable

Random Variable X_i

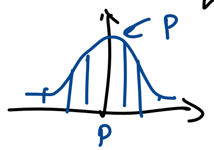
Expected value: p

Variance: $p(1-p)$



Approximation of $\frac{1}{n}(X_1 + X_2 + \cdots + X_n)$

CLT: $\sim N\left(p, \frac{p(1-p)}{n}\right)$



- Application: approximate the Binomial Law

Now for the full Binomial variable:

Random Variable X

$$X = X_1 + X_2 + \cdots + X_n$$

Expected value: np

Variance: $np(1-p)$

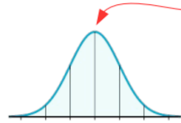
$$\mathcal{B}(n, p)$$

Allows approximation of Binomial → Normal

Approximation of $X = X_1 + X_2 + \cdots + X_n$

$$\mathcal{N}(np, np(1-p))$$

np



EXAMPLE: $X \sim B(100, 0.5) \approx Y \sim N(100(0.5), 100(0.5)(1-0.5))$
$\sim N(50, 25)$

Visualising the approximation:

- Comparison with binomial law:

$\mathcal{B}(100, 0.5)$	$\mathcal{N}(50, 25)$
$Pr(X \leq 40) = \sum_{k=0}^{40} \binom{100}{k} 0.5^k \times 0.5^{100-k}$	$Pr(X \leq 40) = \int_0^{40} \frac{1}{5\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-50}{5}\right)^2} dx$
$\boxed{= 0.02844}$	$\boxed{= 0.02872}$

$\sqrt{25} = 5 = 5$

Statistics

[Bayesian Statistics]

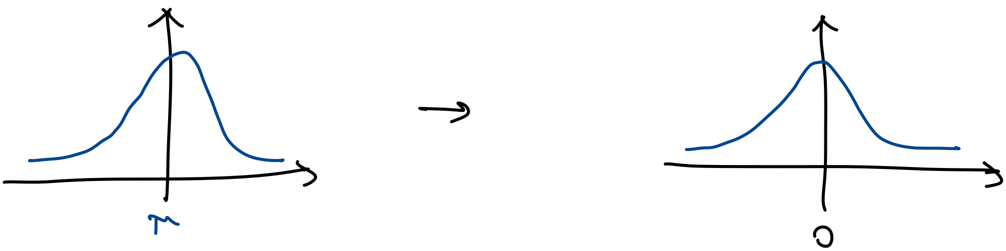
1. Interpret probabilities as degree of belief

Normal Distribution

Standardisation

$$X \sim N(\mu, \sigma^2) \longrightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

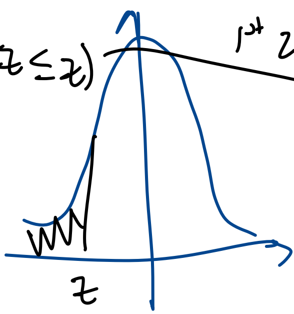
Transform normally distributed X to Z with mean 0 / variance 1



$Pr(Z \leq z)$

← z-scores [known values on tables]

– signed distance to mean, expressed as standard deviation.



1st 2 digits of z-score

next 2 digits of z score

1st 2 digits →

z	.00	.01	...
-3.4			
-3.3			
⋮			
-2.9		$\boxed{0.018}$	

$$Pr(Z \leq -2.91) = 0.018$$

Example Question

Example:
Consider the exam scores of a cohort of students. We assume it is normally distributed with mean 75 and standard deviation of 15.
What percentage of students have scored less than 50?

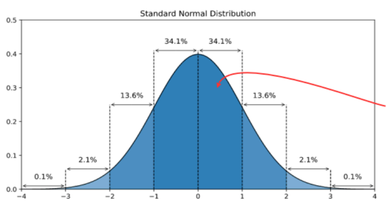
$$X \sim N(75, 15^2)$$

① NORMALISE

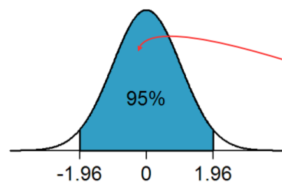
$$Z = \frac{X-75}{15} \sim N(0,1)$$

$$\begin{aligned} Pr(X \le 50) &= Pr\left(Z \le \frac{50-75}{15}\right) \\ &= Pr\left(Z \le -\frac{25}{15}\right) \\ &= Pr(Z \le -1.66) \\ &= \underbrace{0.0485}_{\text{From table}} = 4.85\% \end{aligned}$$

Known values of Normal Distribution:

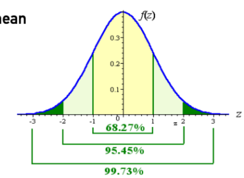


34.1% probability to be between 0 and 1

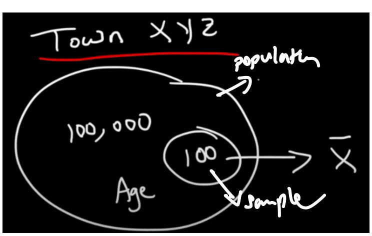


95% probability to be Between -1.96 and 1.96

- 99% of values are within a distance of 3 to the mean
- For a non-standardised variable, that means a distance of $3 \times \sigma$ to the mean.



Estimating Parameters



$\bar{X}$ - sample mean (statistic)
μ = population mean [Parameter]

Statistic		Parameter
Sample		population
$\bar{X}$	Mean	μ
s	stdev.	σ
s^2	variance	σ^2
$\hat{p}$	proportion	p
n	size	N

Sample: subset of population
Population: everything in study

Statistic - characteristic that describe a sample
parameter - characteristics that describe a parameter

Given data $X_1, X_2 \dots X_n$ = statistic
$X_1 \dots X_n \to n$ independent observations from population of mean μ, variance σ^2
Estimator - Statistic used to estimate a parameter
- [no condition - some estimators better than others]

Estimate - value of estimator

An estimator is UNBIASED if expected value equal to the parameter it is estimating.

<u>Estimating the mean</u>

- The sample average $\overline{X_n} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ is a statistic.
- We can use it as an estimator of the mean.

The law of large numbers suggests it's a good estimator

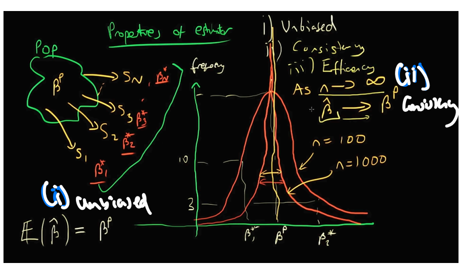
- It is unbiased: $E(\overline{X_n}) = \frac{1}{n}(E(X_1) + \dots E(X_n)) = \mu$
- <u>Example</u>
  We consider the set of UCL students and we are interested in their average height.
  We take a sample $X_1, \dots, X_n$ of student sizes and calculate the sample average.

How close are we from average μ of distribution?
↳ we approximation of CLT and assume sample average follows normal distribution.

$$\overline{X_n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

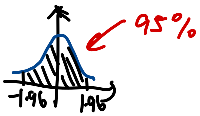
→ Standardise

$$\overline{Y_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0, 1)$$



Confidence Intervals

95% $\Rightarrow$



$$-1.96 \le y_n \le 1.96$$

$$-1.96 \le \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \le 1.96$$

$$-1.96\sigma \le \sqrt{n}(\bar{X}_n - \mu) \le 1.96\sigma$$

$$-1.96\frac{\sigma}{\sqrt{n}} \le \bar{X}_n - \mu \le 1.96\frac{\sigma}{\sqrt{n}}$$

$$\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}$$

→ CI with 95% Probability

$$\left[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

need real value of variance / std deviation

→ 95% sure average over whole population lies inside this confidence interval.

→ It depends on sample size

Estimating Variance

Sample variance $\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

↳ estimator of variance

↳ NOT unbiased $\left[E(\hat{s}^2) = \frac{n-1}{n}\sigma^2\right]$

↳ Correct the estimator:

$$s^2 = \frac{n}{n-1} \times \hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

s^2 is unbiased estimator of variance

$\left[\frac{n}{n-1}\hat{s}^2\right] \Rightarrow \frac{n}{n-1} \times$ sample variance

[dividing by n makes it biased estimator]

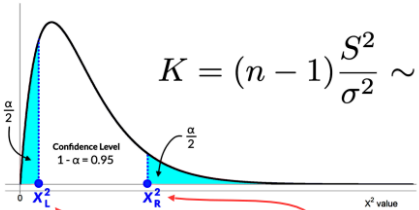
Standardisation:
of estimating variance

$$s^2 \longrightarrow \quad K = (n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$$

↳ get CI

Chi-squared distribution of (n-1) degrees of freedom

- It is then possible to obtain confidence intervals for the variance.



$$K = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

- For the level of confidence desired, we must find the 2 values above in a table.

Estimating the variance

- We denote by $\chi^2_{n-1}[c]$ the number such that $Pr(X \leq \chi^2_{n-1}[c]) = c$ for X a random variable following this chi-squared distribution.

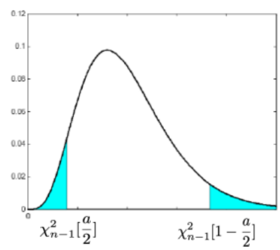
- Example: $c = 0.05, \qquad \chi^2_5[c] = 1.145$

by definition:

- (1-a)% of values are between $\chi^2_{n-1}[\frac{a}{2}]$ and $\chi^2_{n-1}[1-\frac{a}{2}]$

For instance:

- 95% of values are between $\chi^2_{n-1}[0.025]$ and $\chi^2_{n-1}[0.975]$



These values depend on n!

$$\chi^2_{n-1}[a/2] \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{n-1}[1-a/2]$$

$$\frac{(n-1)s^2}{\chi^2_{n-1}[1-a/2]} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n-1}[a/2]}$$

$$CI = \left[\frac{(n-1)s^2}{\chi^2_{n-1}[1-a/2]}, \frac{(n-1)s^2}{\chi^2_{n-1}[a/2]}\right]$$

A few subtleties:

- In practice, we calculate confidence intervals for the average by using the estimated variance instead of the real value.

$$Y_n = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\underbrace{\sigma}_{\text{real}}} \qquad Z_n = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\underbrace{S}_{\text{estimated}}}$$

- The distribution of Y_n is standard normal, but the distribution of Z_n is actually not.
- Z_n follows a t-distribution with (n-1) degrees of freedom (similar to normal).
- Taking this into account, the resulting confidence interval is less precise.

Hypothesis Testing

T-test – Test our new hypothesis

H_0: default situation

H_1: Our hypothesis

example:

- We define a theoretical value μ_0 of the average.
- We want to test a hypothesis about the true average μ. For instance:

$$H_1 : \mu \neq \mu_0$$
$$H_1 : \mu < \mu_0$$
$$H_1 : \mu > \mu_0$$

We have no value for the real average.
We just compare it to a hypothesised value.

- We also set the Null Hypothesis as being the default situation: $H_0 : \mu = \mu_0$

- **Example:**

  Consider knee surgery patients. A researcher thinks that if patients go to physical therapy 3 times a week instead of 2, their recovery period will be shorter. Average recovery times for knee surgery patients is 8 weeks.

- **We must start by formulating a hypothesis about the average:**

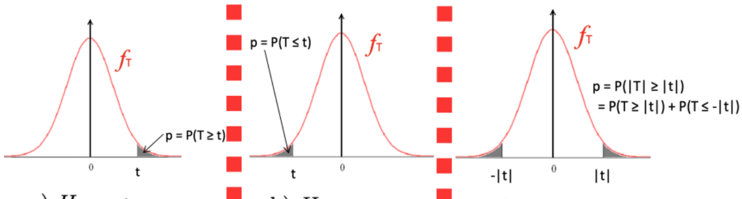
  If patients go to physical therapy 3 times a week, then their recovery times will be shorter than 8 weeks.

H_0: average recovery time = 8 weeks
H_1: average recovery time < 8 weeks

How do we test?

2-tailed

- Several cases, depending on the hypothesis:



a) $H_1 : \mu > \mu_0$ b) $H_1 : \mu < \mu_0$ c) $H_1 : \mu \neq \mu_0$

- We calculate these probabilities by using a table for t-distributions / a calculator.

Find the p-value [probability] → based on hypothesis we made and compare to threshold of significance α. [often 5% used]

$p < \alpha \implies$ reject H_0 in favour of H_1

$p > \alpha \implies$ do not reject H_0 – cannot conclude

Patients have gone to physical therapy 3 times a week. We study a sample of 12 patients. The average recovery time is 7.2 weeks, with a standard deviation of 1.5.

Does that mean our hypothesis was correct?

Statistic: $\frac{7.2-8}{1.5}\sqrt{12} \approx -1.848$

$Pr(t < -1.848) = P(t > 1.848)$
$= 1 - P(t < 1.848)$

Table of t-distribution for $n = 11 \implies 0.9503$

$\therefore P(t < -1.848) = 1 - 0.9503$
$= 0.0497$ $[< 0.05]$

$\therefore$ reject H_0, accept H_1 as smaller than 5% → data is significant to confirm our hypothesis

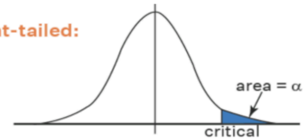
We can use critical values to compare values

- Instead of p-values, another equivalent method relies on comparing the value of t with known values for the distribution: the critical values.

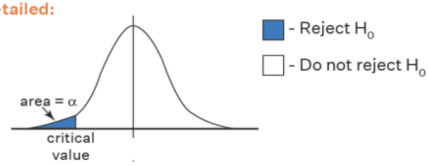
- These values can be found in tables.

- Depending where t is, we may reject H_0.

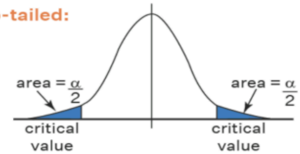
right-tailed:



left-tailed:



two-tailed:



Bootstrapping:

method of resampling a sample with NO distribution to form a new sample of size n and calculate mean.

You repeat resampling until normal distribution is obtained. [via CLT]

Bootstrapping method of resampling:

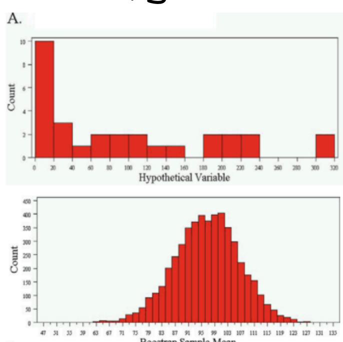
- Assume we have a sample of size n, with no obvious distribution.

- We resample the sample, with replacement, to form a new sample of size n.

We calculate the mean of this sample.

- We repeat the resampling and consider the distribution of our means: it is normal!

- Use if for confidence intervals for the mean.



To find CI → find where 95% of bootstrapped sample mean are

→ order all values in increasing order [find top 2.5%, bottom 2.5%]

→ we do CI [Confident 95% of population mean (real average) is in interval]

- Bootstrapping allows to extract a lot of information from a single sample
- Resampling with replacement allows to represent the distribution of the sample (most frequent elements are overrepresented)
- Works for a small sample (too small for CLT to apply)
- Works without assumptions on the distribution

- Statistics rely on comparisons with mathematical models (distributions)
- We standardise and compare with classic distributions
- Estimating parameters of a distribution: mean, variance
- Precision of estimation is given by confidence intervals
- Other approach: bootstrapping