Sets and functions

A set is an unordered collection of distinct elements
(A tuple is ordered)

$$S = \{2, 4, 6, 8\} = \{8, 4, 2, 6\}$$

$$2 \in S$$

$$3 \notin S$$

If $n \in S$, then n is even

$$S = \{n \text{ is an integer} : 1 \leq n < 10,\ n \text{ is even}\}$$

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

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

$A \cup B = A + B - A \cap B$ (for non-infinite sets)

---

72 English, 43 French, 100 total, How many only speak English? French?

$|E| = 72 \quad |F| = 43$

$|E \cup F| = 100$

$|E \cup F| = |E| + |F| - |E \cap F|$

$100 = 72 + 43 - |E \cap F|$

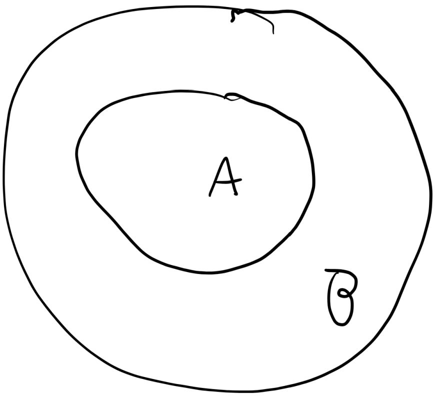
$|E \cap F| = 15$

$72 - 15 = \boxed{57}$

$43 - 15 = \boxed{28}$

Subsets

A is a subset of B ($A \subseteq B$) if every element of A is also a member of B.



(If $x \in A$, $x \in B$

A is a proper subset of B if there is an element x such that $x \in B$ and $x \notin A$

($A \subset B$)

Empty Set $\emptyset$

Natural Numbers $\mathbb{N}$ $\{0, 1, 2, \ldots\}$

Integers $\mathbb{Z}$ $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$

Rational numbers $\mathbb{Q} = \{m/n : m, n \in \mathbb{Z}, n \neq 0\}$

Real numbers $\mathbb{R}$ - all numbers including infinite decimals

Complex numbers $\mathbb{C} = \{a+bi : a, b \in \mathbb{R}\}$

Real intervals

Closed interval

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

Open interval

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

Half open interval

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

$$[5,\infty) := \{x \in \mathbb{R} : 5 \leq x\}$$

Set Minus

$$A \setminus B = \{a : a \in A, a \notin B\}$$

Powerset of A – set of all subsets

$$\wp(A) := \{A' : A' \subseteq A\}$$

Cartesian Product

The set of ordered pairs where the first element is from A and the second element is from B

$A \times B$:

Functions

$f: A \to B$ is an assignment of a unique element of B to every element of A.

More generally f could be defined as the output of a computer program

for a function $f: A \to B$

A is the domain (source)

B is the codomain (target)

If $f(a) = b$, b is the image of a

The pre-image of b is the set $\{a : f(a) = b\}$

The image or range of f is the set of images of elements of A.

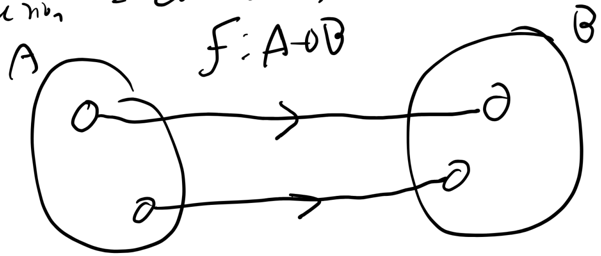
Composition of Functions

Suppose $f : A \to B$ and $g : B \to C$

The composition of g and f = $g \circ f : A \to C$
(or $gf : A \to C$)

where $g \circ f(a) = g(f(a))$

Injection - each element in set A has a unique image in set B



$(|\text{Set } B| \geq |\text{Set } A|)$

Surjection - Codomain equal to the range ($|Set\ A| \geq |Set\ B|$)

Bijection - f is both injective and Surjective.
(Set A and Set B the same size)

$f: \mathbb{N} \to \mathbb{N}$ where $f(x) = x+1$

$\Downarrow$

Injective but not Surjective as nothing maps to 0.

$f: \mathbb{Z} \to \mathbb{Z}$ where $f(x) = x+1$

$\Downarrow$

Bijective

$f: \mathbb{Z} \to \mathbb{N}$, $f(x) = x^2+1$

$\Downarrow$

Neither Injective or Surjective
($f(-1) = f(1)$ so not unique image, $f(x)$ never equals 0)

$f: \mathbb{R} \to \mathbb{R}, \ f(x) = x^3 + 1$

$\Downarrow$

Bijective

$f: \mathbb{R} \to \mathbb{R}, \ f(x) = x^3 + x^2$

$\Downarrow$

Surjective but not injective

$\left(f(0) = f(-1) = 0 \right)$

Inverse functions

Consider $f: A \to B$

The inverse function, if exists: $f^{-1}: B \to A$

If $f(a) = b$ then $f^{-1}(b) = a$

When does f^{-1} exist?

- f must be injective
  - If $f(a_1) = f(a_2) = b$ then $f^{-1}(b)$ cannot be defined to be both a_1 and a_2

- f must be surjective
  - All elements in set B must map back to an element in set A.

$\therefore f$ must be bijective

To find the inverse function set $x = f(y)$ and rearrange for y

e.g. $f(x) = 5x + 3$

$$5y + 3 = x$$

$$y = \frac{x-3}{5}$$

$$\therefore f^{-1} = \frac{x-3}{5}$$

Cardinality

For a finite set, the cardinality is the number of distinct elements. Two sets have the same size if there is a bijection between them.

If the number of elements is unbounded, cardinality is infinite

- If bijection with set $\mathbb{N}$ → countably infinite
- If no bijection, uncountably infinite.

$A = \{a, b, e, t\}$, $|A| = 4$

$f(a) = 1$

$f(b) = 3$

$f(e) = 4$

$f(t) = 5$

$B = \{1, 3, 4, 5\}$, $|B| = 4$

$f: \mathbb{N} \to \mathbb{Z}$

$$f(n) = \begin{cases} n/2, & n \text{ is even} \\ -\frac{(n+1)}{2}, & n \text{ is odd} \end{cases}$$

$f(0) = 0$, $f(1) = -1$, $f(2) = 1$, $f(3) = -2$, $f(4) = 2$

A graph of a function $f: A \to B$ is the locus of points $(a, f(a))$

Proof

Theorem $\to$ If hypotheses then conclusion

e.g. If n is even, n^2 is divisible by 4

Proof: $n = 2k$

$n^2 = (2k) \cdot (2k) = 4k^2$

$\therefore n^2/4 = k^2$

For 2 positive real numbers, x, y, Prove

$$\frac{x+y}{2} > \sqrt{xy}$$

$$\frac{x+y}{2} > \sqrt{xy}$$

$$\left(\frac{x+y}{2}\right)^2 > xy$$

$$(x+y)^2 > 4xy$$

$$x^2 + 2xy + y^2 > 4xy$$

$$x^2 - 2xy + y^2 > 0$$

$$\underline{(x-y)^2 > 0}$$

Prove $2 = 1$

Assume $2 = 1$ $\Downarrow$

So $1 = 2$

Add $\Downarrow$

$3 = 3$

$\therefore$ not true since does not work in reverse

Proof types
- Direct
- Contraposition
- Exhaustion
- Contradiction
- Induction

Proof by Contradiction

Theorem: p

Proof: Assume not p

Show that not $p \Rightarrow q$, where q is known to be false

Since not $p \Rightarrow$ false, p must be true

e.g. Theorem: $\sqrt{2} \notin \mathbb{Q}$

Proof: Assume $\sqrt{2} \in \mathbb{Q}$

i.e. $\sqrt{2} = \frac{m}{n}$, $\frac{m}{n}$ simplified fraction

$\Downarrow$

$\frac{m^2}{n^2} = 2$ So $m^2 = 2n^2$ and

So must be even

Let $m = 2k$, then $4k^2 = 2n^2$ and

$2k^2 = n^2$

Hence $\frac{m}{n} = \frac{4k^2}{2k^2}$ is not a simplified fraction, so the assumption is false

$\therefore \sqrt{2} \notin \mathbb{Q}$

Proof by induction

Theorem: Statement $S(n)$ holds for all integers $n \geq 1$

Proof: Check that $S(1)$ is true. If $S(1)$ not true, Theorem false.

Prove that if $S(k)$ holds for a fixed value $k \geq 1$, then $S(k+1)$ also holds

These two steps together then imply $S(n)$

holds for all $n \geq 1$

Example: prove for $n \in \mathbb{N}$ that $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$

Base Case $n=1$

$\sum_{i=1}^{1} i = 1$, $\frac{1 \times 2}{2} = 1$ So holds

Induction

Assume $\sum_{i=1}^{k} i = \frac{k(k+1)}{2}$

Let $n = k+1$

$$\sum_{i=1}^{k+1} i = \sum_{i=1}^{k} i + (k+1) = \frac{k(k+1)}{2} + \frac{2(k+1)}{2}$$

$$= \frac{(k+1)(k+2)}{2}$$

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

So if Statement holds for $n=k$, it also holds for $n=k+1$

$\therefore$ By induction, Statement holds for all $n \in \mathbb{N}$

Example: $S(n)$: $11^n - 6$ is divisible by 5 for all $n \geq 1$

Base case: $S(1) = 11 - 6 = 5$

Inductive Step: $S(n) \to S(n+1)$

Assume $S(k) = 11^k - 6$ holds for $k \geq 1$

$11^k - 6 = 5m$

$n = k+1$

$$S(k+1) : 11^{k+1} - 6 = 11 \cdot 11^k - 6$$
$$= 11(5m+6) - 6$$
$$= 55m + 60$$
$$= 5(11m+12)$$

Proof for $n \leq 2^n$ for $n \geq 0$ on lecture slides (wk3)

Sometimes have $S(0)$, $S(1)$, $S(2)$ as base case

Then prove $S(n+3)$ given $S(n)$ true, and other variants.

Differentiation

The derivative of a function $f(x)$ at a point $x = x_0$ is the instantaneous rate of change of f at the $x_0 \to$ gradient.

f is differentiable at $x = x_0$ only if $\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$ exists

with $h = x - x_0$
the derivative $= \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$

The derivative at x_0 is denoted $f'(x)$ or $\frac{df}{dx}(x_0)$

If f is differentiable at all points in the interval (a, b) then the derivative is a function $f' : (a, b) \to \mathbb{R}$
= the function that maps any point on f in (a, b) to its derivative.

If $f(x) = c$ for some constant c, then $f'(x) = 0 \ \forall x$

If $f(x) = x^a$ for some $a \in \mathbb{R}$, then $f'(x) = a x^{a-1}$

If $f(x) = \sin \alpha x$ for some $\alpha \in \mathbb{R}$, then $f'(x) = \alpha \cos \alpha x$

If $f(x) = \cos \alpha x$ for some $\alpha \in \mathbb{R}$, then $f'(x) = -\alpha \sin \alpha x$

If $f(x) = g(x) + h(x)$ then $f'(x) = g'(x) + h'(x)$

Example: $f(x) = \begin{cases} 2 - x, & x < 1 \\ x^2, & x \geq 1 \end{cases}$

Assuming $h > 0$
so $f(1+h)$ uses x^2

$$\frac{f(1+h) - f(1)}{h} = \frac{(1+h)^2 - 1}{h} = 2 + h, \quad h > 0$$

Assuming $h < 0$
so $f(1+h)$ uses $2 - x$

$$\frac{f(1+h)-f(1)}{h} = \frac{(2-(1+h))-1}{h} = -1 \qquad h<0$$

We write $\lim_{h\to 0^+}$ for the limit restricted to positive h: ‘from above’ or ‘right hand limit’.

We write $\lim_{h\to 0^-}$ for the limit restricted to negative h: ‘from below’ or ‘left hand limit’.

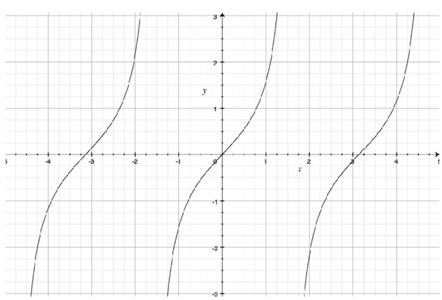
Here: $\lim_{h\to 0^+} \frac{f(1+h)-f(1)}{h} = 2$, but $\lim_{h\to 0^-} \frac{f(1+h)-f(1)}{h} = -1$ so the limit (unrestricted) and therefore the derivative do not exist.

A function $f:\mathbb{R}\to\mathbb{R}$ is said to be continuous at $x=a$ if
$\lim_{x\to a^-} f(x) = \lim_{x\to a^+} f(x) = L$ and $f(a) = L$

A function $f:\mathbb{R}\to\mathbb{R}$ is said to be differentiable at $x=a$ if
$\lim_{x\to a} \frac{f(x)-f(a)}{f(x-a)}$ exists, meaning f must be continuous at $x=a$

If you can keep differentiating a function and always get a differentiable result then it is called smooth

Not continuous



Consider $\tan x = \frac{\sin x}{\cos x}$

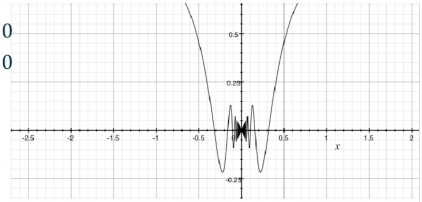
$\lim_{x\to\frac{\pi}{2}} \sin(x) = 1$ and $\lim_{x\to\frac{\pi}{2}} \cos(x) = 0$

I.e. as we approach $\frac{\pi}{2}$ from below, $\cos x$ is +ve, so $\tan x$ gets arbitrarily large +ve.

As we approach $\frac{\pi}{2}$ from above, $\cos x$ is –ve, $\tan x$ gets arbitrarily large -ve.

Still a useful function, and smooth on $\left(-\frac{\pi}{2},\frac{\pi}{2}\right)$.

Define a function $f(x) = \begin{cases} x \sin\frac{1}{x} & x \neq 0 \\ 0 & x = 0 \end{cases}$



For $x \neq 0$ we can write $f(x) = \frac{\sin\frac{1}{x}}{\frac{1}{x}}$.

So when x tends to 0, the numerator is bounded between -1 and $+1$, and the denominator is unbounded, which gives $\lim_{x\to 0} f(x) = 0$.

Hence f is continuous on $(-\infty, \infty)$.

For $x_0 = 0$, the derivative is $\lim_{x\to 0}\frac{f(x)-f(0)}{x-0} = \lim_{x\to 0}\frac{x\sin\frac{1}{x}}{x} = \lim_{x\to 0}\sin\frac{1}{x}$.

Since this oscillates between -1 and $+1$ for arbitrarily small x, the limit does not exist. Hence f is not differentiable at 0.

$\nexists$ as $x \to 0$, $\sin \frac{1}{x}$ does not converge.

$\therefore$ derivative does not exist.

Differentiable but not twice differentiable

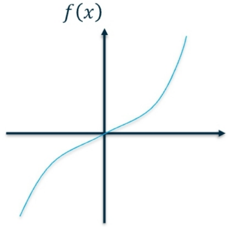
Define a function $f(x) = \begin{cases} -\frac{x^2}{2} & for\, x < 0 \\ \frac{x^2}{2} & for\, x \geq 0 \end{cases}$

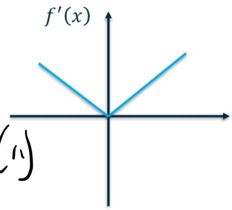
Now $f'(x) = \begin{cases} -x & for\, x < 0 \\ x & for\, x \geq 0 \end{cases} = |x|$

So f' is not differentiable at 0.

Could be 1 or -1 depending on direction of approach

$\therefore \lim_{x\to a^+} f(x) \neq \lim_{x \to a^-} f(x)$





Durham University

If $f(x), g(x)$ differentiable at x_0:

Product rule $\rightarrow$

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

Quotient rule → $\dfrac{d\left(\frac{f}{g}\right)}{dx} = \dfrac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$

Chain rule → $\dfrac{d\, f \circ g(x)}{dx} = f'\big(g(x_0)\big)g'(x_0)$

Proof for Product Rule

$f(x)g(x)$ $\quad \dfrac{d\, f(x)g(x)}{dx} = f'(x)g(x) + f(x)g'(x)$

$$\frac{d\, f(x)g(x)}{dx} = \lim_{x \to x_0} \frac{f(x)g(x) - f(x_0)g(x_0)}{x - x_0}$$

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

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

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

Proof for Chain rule

$$\frac{d\big(f \circ (g(x))\big)}{dx} = f'(g(x_0))\, g'(x_0)$$

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

$$= \lim_{x \to x_0} \frac{f(g(x)) - f(g(x_0))}{g(x) - g(x_0)} \cdot \underbrace{\frac{g(x) - g(x_0)}{x - x_0}}_{= g'(x_0)}$$

$x \to x_0$, $g(x) \to g(x_0)$

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

$$= f'(g(x_0))$$

$$\therefore \frac{d(f \circ g(x))}{dx} = f'(g(x_0))g'(x_0)$$

for a Composition of functions $f \circ g(x) = f(g(x))$ often better to set

$u = g(x)$

so $\frac{df}{dx} = \frac{df}{du} \frac{du}{dx}$

e.g. Differentiate $\sin(x^2+3)$

Let $u = x^2+3$ and $f(u) = \sin(u)$

$f'(u) = \cos u$, $u'(x) = 2x$

$$\frac{df}{dx} = \frac{df}{du} \times \frac{du}{dx} = \cos u \cdot 2x = 2x \cos(x^2+3)$$

$f(x) = \sin\left(\sqrt{x^2+1}\right)$

$$\frac{d}{dx}\left(x^2+1\right)^{\frac{1}{2}} = \frac{1}{2} \times 2x \left(x^2+1\right)^{-\frac{1}{2}}$$

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

$$\therefore f'(x) = x\left(x^2+1\right)^{-\frac{1}{2}} \sin\left(\left(x^2+1\right)^{\frac{1}{2}}\right)$$

Proof of Quotient rule

Proof:

$$\frac{d\left(\frac{f}{g}\right)}{dx} = f'(x)\left(\frac{1}{g(x)}\right) + f(x)\frac{d\left(\frac{1}{g}\right)}{dx}$$

By the chain rule, setting $u = g(x)$,

$$\frac{d\left(\frac{1}{g}\right)}{dx} = \frac{d\left(\frac{1}{u}\right)}{du}\frac{du}{dx} = -\frac{1}{u^2}g'(x) = -\frac{g'(x)}{g(x)^2}.$$

Putting these together give the result.

Extrema

Extrema

Let $f(x)$ be a function defined on an interval $[a, b]$.

A point $x_0 \in [a, b]$ is:

- an absolute maximum if $f(x_0) \geq f(x) \ \forall\, x \in [a, b]$
- an absolute minimum if $f(x_0) \leq f(x) \ \forall\, x \in [a, b]$
- a local maximum if $\exists \delta > 0$: $f(x_0) \geq f(x_0 + h) \ \forall\, |h| < \delta$
- a local minimum if $\exists \delta > 0$: $f(x_0) \leq f(x_0 + h) \ \forall\, |h| < \delta$

functions of multiple real variables

A function $f: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ can be viewed as a 2D surface

e.g. $f(x, y) = \sin x + \cos y + 1.5$

We may be interested in
- gradient in a specific direction
- Limits of f as approaching a given (x, y)
- The direction of steepest ascent from a given point
- A linear approximation to the surface at a given point

Example: $f(x,y) = 2x^2y^3 - xy + 2x - 3y + 5$

What does this look like if:
- we fix $y=0$? $f(x,0) = 2x+5$ - a line
- we fix $y=1$? $f(x,1) = 2x^2 + x + 2$ - a +ve parabola
- we fix $y=-1$? $f(x,-1) = -2x^2 + 3x + 8$ - a -ve parabola
- we fix $x=0$? $f(0,y) = -3y+5$ - a line
- we fix $x=1$? $f(1,y) = 2y^3 - 4y + 7$ - cubic
- we fix $x=-1$? $f(-1,y) = 2y^3 - 2y^2 + 3$ - cubic

For a fixed y we can work out the gradient in the x direction at any point.

- If we fix $y=0$, $f(x,0) = 2x+5$
- $\frac{d(2x+5)}{dx} = 2$, so at a point with $y=0$, moving in the x direction, the gradient is 2

Example: $f(x,y) = 2x^2y^3 - xy + 2x - 3y + 5$

If we think of $y = y_0$ as fixed but unknown, f is a function of just x and we can differentiate it at $x = x_0$. This is the **partial derivative of f with respect to x**.

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{x \to x_0} \frac{f(x, y_0) - f(x_0, y_0)}{x - x_0}$$

As with single-variable derivatives, $\frac{\partial f}{\partial x}$ is itself a function $\frac{\partial f}{\partial x}: A \to \mathbb{R}$, where $A \subseteq \mathbb{R}^2$ is the domain on which the limit exists.

Here:

$$\frac{\partial f}{\partial x} = \frac{\partial(2y^3x^2 - xy + 2x - 3y + 5)}{\partial x} = 4y^3x - y + 2$$

E.g. at point (1,3) the gradient moving in the x-direction is $108 - 3 + 2 = 107$.

The partial derivative wrt y is where you treat x as fixed but unknown, e.g. $\frac{\partial f}{\partial y} = 6xy^2 - x - 3$ using above ex.

(At point $(0,3)$ the gradient in the y direction is -3.)

$\frac{\partial f}{\partial x}$ can be written as f_x

$\frac{\partial f}{\partial y}$ " " " " f_y

Continuity & partial derivatives

Consider $f(x,y) = \begin{cases} 1 & xy = 0 \\ 0 & xy \neq 0 \end{cases}$

This has partial derivatives at $(0,0)$ but f is not continuous at $(0,0)$ since $\left.\begin{array}{l} f(0,\epsilon) = 1 \\ f(\epsilon,\epsilon) = 0 \end{array}\right\}$ for arbitrarily small ϵ

As a result, the existence of partial derivatives does not imply continuity, or that the function is 'nice' ☹

Directional derivatives

Consider a point $x_0 = (x_0, y_0)$ and a direction given by a unit vector $\underline{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix}$:

If not unit vector then uh-oh!

The **directional derivative** of $f(x,y)$ in the direction $\boldsymbol{v}$ is defined to be:

$$\nabla_v f(x_0, y_0) = \lim_{h\to 0} \frac{f(x_0 + h.v_x, y_0 + h.v_y) - f(x_0, y_0)}{h} = \lim_{h \to 0} \frac{f(\boldsymbol{x_0} + h\boldsymbol{v}) - f(\boldsymbol{x_0})}{h}$$

I.e. this gives the gradient of the slope of f if you were to move in the $\boldsymbol{v}$ direction through point $\boldsymbol{x_0} = (x_0, y_0)$.

this $\underline{x}_0$ is the point (x_0, y_0) not just the x coordinate x_0 (shite notation)

E.g.

Example: $f(x,y) = 2x^2y^3 - xy + 2x - 3y + 5$

What about moving in the direction $\begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}$ through $(1,3)$?

$$\lim_{h\to 0} \frac{f\left(1 + h.\frac{1}{\sqrt{5}}, 3 + h.\frac{2}{\sqrt{5}}\right) - f(1,3)}{h}$$

$$= \lim_{h\to 0} \frac{2\left(1+h.\frac{1}{\sqrt{5}}\right)^2\left(3+h.\frac{2}{\sqrt{5}}\right)^3 - \left(1+h.\frac{1}{\sqrt{5}}\right)\left(3+h.\frac{2}{\sqrt{5}}\right) + 2\left(1+h.\frac{1}{\sqrt{5}}\right) - 3\left(3+h.\frac{2}{\sqrt{5}}\right) + 5 - 49}{h}$$

$$= \lim_{h\to 0} \frac{54-3+2-9+5-49+h\left(\frac{2.2.27}{\sqrt{5}} + \frac{2.27.2}{\sqrt{5}} - \frac{3}{\sqrt{5}} - \frac{2}{\sqrt{5}} + \frac{2}{\sqrt{5}} - \frac{3.2}{\sqrt{5}}\right) + h^2(?) + \cdots}{h}$$

$$= \lim_{h\to 0} \frac{h\left(\frac{207}{\sqrt{5}}\right) + h^2(?) + \cdots}{h} = \lim_{h\to 0} \frac{207}{\sqrt{5}} + h(?) = \frac{207}{\sqrt{5}} \approx 93.$$

irrelevant as h fucks off

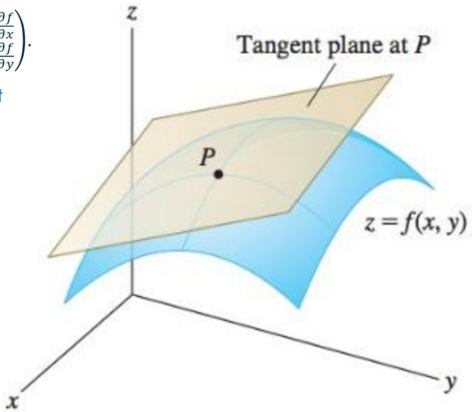
An easier way?

If the surface is well behaved, i.e. approximately flat at a small scale, then we can determine the slope in any direction from the slope in the x and y directions:

The slope in direction $\boldsymbol{v} = \binom{v_x}{v_y}$ will be $v_x f_x + v_y f_y$. ← just use this ⇓

I.e. $\nabla_{\boldsymbol{v}} f = \boldsymbol{v}.\binom{f_x}{f_y} = \boldsymbol{v}.\begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}$.

bro spawned in dot product out of nowhere



Example: $f(x,y) = 2x^2y^3 - xy + 2x - 3y + 5$

Same ex. as before much quicker method ⇓

- At point $(1,3)$ we saw that $f_x(1,3) = 107 \quad f_y(1,3) = 50$.

What about moving in the direction $\begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}$ through $(1,3)$?

$$= \frac{1}{\sqrt{5}}.107 + \frac{2}{\sqrt{5}}.50 = \frac{207}{\sqrt{5}} \approx 93.$$

We say f is differentiable if for all unit vectors $\underline{v}$:

$$\nabla_v f = v_x \frac{\partial f}{\partial x} + v_y \frac{\partial f}{\partial y} = \underline{v}.\begin{pmatrix} f_x \\ f_y \end{pmatrix}$$

⇓

this means that close to point (x,y), f is like a flat plane

The vector of partial derivatives $\binom{f_x}{f_y}$ is denoted ∇f.

This symbol '∇' is a 'nabla'.

In this case it is pronounced del when it is an operator, or grad when it is the result, similar to "derivative of f" and "gradient of f" when f is a function of a single variable.

∇f is a vector pointing in the direction of the greatest rate of increase of f and having magnitude the rate of increase of f in that direction.

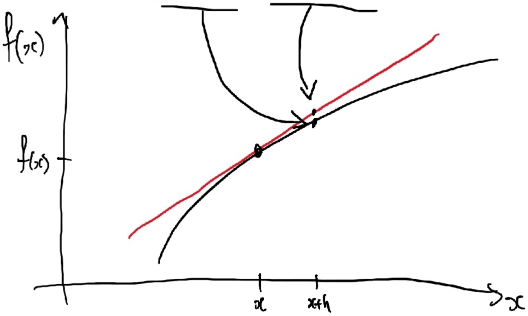
Linear approximations

For a univariate function the derivative gives details of the tangent line

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

So for small h, $hf'(x) \approx f(x+h) - f(x)$

$$f(x+h) \approx f(x) + hf'(x)$$



Durham

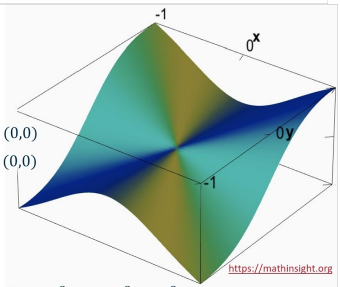
For a surface we have seen that $\boldsymbol{v}.\nabla f(\boldsymbol{x_0})$ is the slope in the direction $\boldsymbol{v}$.

$$f(\boldsymbol{x_0} + h.\boldsymbol{v}) \approx f(\boldsymbol{x_0}) + h\boldsymbol{v}.\nabla f(\boldsymbol{x_0})$$

Which is

$$f\left(x_0 + h.v_x, y_0 + h.v_y\right) \approx f(x_0, y_0) + h.v_x \frac{\partial f}{\partial x} + h.v_y \frac{\partial f}{\partial y}$$

Example



Consider $f(x,y) = \begin{cases} \frac{x^2 y}{x^2+y^2} & (x,y) \neq (0,0) \\ 0 & (x,y) = (0,0) \end{cases}$

Now f is continuous at all points.

It has partial derivatives:

$$\frac{\partial f}{\partial x} = \frac{2xy(x^2+y^2) - x^2y(2x)}{(x^2+y^2)^2} = \frac{2x^3y + 2xy^3 - 2x^3y}{(x^2+y^2)^2}$$

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

Even at $(0,0)$ these exist and are both 0.

But the directional derivatives at (0,0) do not agree with partial derivatives: no plane approximation.

$\therefore$ Cannot use easy method !!

Multivariate calculus also works with 3 or more dimensions

e.g. $f : \mathbb{R}^n \to \mathbb{R}$, $f(x_1, x_2, \dots, x_n)$

Let $\boldsymbol{e}_i$ denote the unit vector which is 1 in position i and 0 elsewhere.

E.g. $f(x_1, x_2, \dots, x_n): \mathbb{R}^n \to \mathbb{R}$.

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}) = \lim_{h \to 0} \frac{f(\boldsymbol{x} + h.\boldsymbol{e}_i) - f(\boldsymbol{x})}{h}.$$

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n}\right)^T = \begin{pmatrix} f_{x_1} \\ \vdots \\ f_{x_n} \end{pmatrix}$$

$\underline{x}$ is just a vector $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

wtf is T?

f is **differentiable** at $\boldsymbol{x_0}$ if for all unit vectors $\boldsymbol{v}$ we have

$$\nabla_v f(\boldsymbol{x_0}) = \boldsymbol{v}.\nabla f$$

$f \circ g(x)$: the chain rule can be written $\frac{df}{dx} = \frac{df}{dg}\frac{dg}{dx}$

If I wiggle g by a small amount ϵ, how much does f vary?

If I wiggle x a bit, how much does f vary?

If $f(x, y): \mathbb{R}^2 \to \mathbb{R}$, where x, y are functions of s, t, what is $\frac{\partial f}{\partial t}$?

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t}$$

Wiggling t causes x to vary by $\frac{\partial x}{\partial t}$ and y to vary by $\frac{\partial y}{\partial t}$,

and varying x and y by these amounts causes f to vary by $\frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t}$

because the directional derivative in direction $\boldsymbol{v} = \left(\frac{\partial x}{\partial t}, \frac{\partial y}{\partial t}\right)^T$ is

$$\nabla_v f(\boldsymbol{x_0}) = \boldsymbol{v}.\nabla f$$

Extrema:

Extrema for univariate functions

Let $f(x)$ be a function defined on an interval $[a, b]$ and differentiable at a point $x_0 \in [a, b]$.

If x_0 is a maximum or minimum of f, **then** $f'(x_0) = 0$.

not always true in reverse

- not all stationary points are extrema (points of inflexion)

Example: $f(x) = \frac{x^3}{3} + 2x^2 + 3x + 1$

f continuous + differentiable for all x, so stationary points when $f'(x) = 0$

$f'(x) = x^2 + 4x + 3 = (x+1)(x+3)$

So stationary points when $x = -1$, $x = -3$

Consider f' close to -1, i.e. $x = -1+h$

$$f'(-1+h) = h(h+2)$$

for small h this is positive for $h>0$ and negative for $h<0$

(i.e. f is sloping downwards to the left of $x=-1$, and upwards to the right of $x=-1$)

$\therefore$ minimum at $x=-1$

Near -3, $f'(-3+h) = (-2+h)h$ which is +ve for $h<0$ and −ve for $h>0$

$\therefore$ maximum at $x=-3$

We can determine the same thing by looking at $f''(x)$ (sometimes called the curvature)

If $f'(x_0)=0$ and $f''(x_0)>0$ then we must have $f'(x)<0$ for x just less than x_0 and $f'(x)>0$ for x just greater than x_0, i.e. $f(x_0)$ is a minimum.

$f''(x)=0 \rightarrow$ inflection point

$f''(x)<0 \rightarrow$ concave (down)

$f''(x)>0 \rightarrow$ convex / concave up

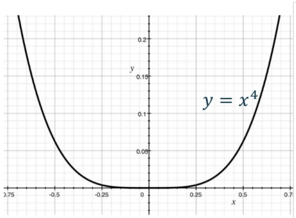
Univariate extrema summary

For a twice differentiable univariate function $f(x)$ with continuous derivatives:

- If $f'(x) = 0$ we have a **stationary point**.
- If $f'(x) = 0$ and $f''(x) < 0$ we have a **maximum**.
- If $f'(x) = 0$ and $f''(x) > 0$ we have a **minimum**.
- If $f'(x) = 0$ and $f''(x) = 0$ we **may** have a **stationary inflection point**.
- If $f'(x) \neq 0$ and $f''(x) = 0$ we **may** have a **non-stationary inflection point**.

Why only *may*? Consider $f(x) = x^4$.
At $x = 0$ we have $f(x) = f'(x) = f''(x) = 0$.

In general, you need to look at the first non-zero derivative. If it is the k^{th} derivative and k is odd it is an inflection. If k is even, it is an extremum.



Quadratic approximation

Recall that the gradient gives the best linear approximation to f.

$$f(x_0 + x) \approx f(x_0) + f'(x_0)x$$

The curvature gives the best 2nd order (or "quadratic") approximation:

$$f(x_0 + x) \approx f(x_0) + f'(x_0)x + \frac{f''(x_0)}{2}x^2$$

At a stationary point the second term is zero, so

$$f(x_0 + x) \approx f(x_0) + \frac{f''(x_0)}{2}x^2$$

Which is a parabola open up when $f''(x_0) > 0$, parabola open down when $f''(x_0) < 0$, and we are not sure if $f''(x_0) = 0$.

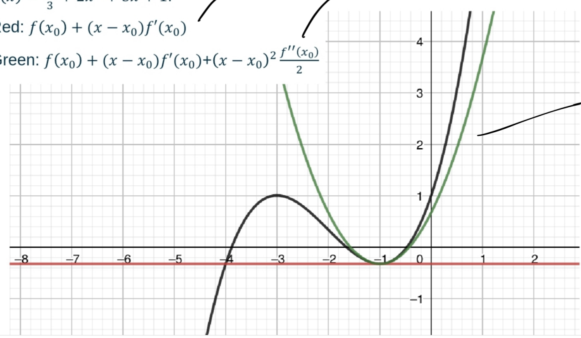
Example

linear quadratic

$f(x) = \frac{x^3}{3} + 2x^2 + 3x + 1.$

Red: $f(x_0) + (x - x_0)f'(x_0)$

Green: $f(x_0) + (x - x_0)f'(x_0) + (x - x_0)^2 \frac{f''(x_0)}{2}$



Over a small interval approximations 'accurate', but quadratic is better.

At points of inflexion, linear = quadratic since $f''(-1) = 0$

Brackets important → f^2 is squared

For a function f, the kth derivative $f^{(k)}$ is the function obtained by differentiating k times.

A function is in the class C^0 is continuous

A continuous function with a continuous derivative is class C^1, etc.

A function that is continuous in class C^∞ is smooth.

Smoothness for multivariate functions

For a function $f(x_1, x_2, \dots x_n)$, a k^{th} partial derivative is a function obtained by differentiating k times with respect to members of $(x_1, x_2, \dots x_n)$.

$f(x_1, \dots, x_8) = x_7 x_2^2 x_1 + x_7^4 + x_8^3 x_1^7 + 5x_2$

E.g. a third derivative is $\frac{\partial}{\partial x_1}\left[\frac{\partial}{\partial x_7}\left[\frac{\partial f}{\partial x_2}\right]\right] = \frac{\partial^3 f}{\partial x_1 \partial x_7 \partial x_2} = f_{x_2 x_7 x_1}$.

partial der'tive wrt x_2 | Part. diff. wrt x_7 of f_{x_2} | p.d. of $f_{x_2x_7}$ wrt x_1

$f_{x_2} = 2x_7 x_2 x_1 + 5,\quad f_{x_2 x_7} = 2x_2 x_1,\quad f_{x_2 x_7 x_1} = 2x_2$

A function f such that partial derivatives of all orders exist and are continuous is called **smooth**.

Multivariate polynomials are smooth functions.

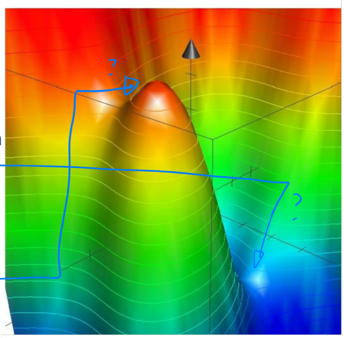
Bivariate extrema

Example: $f(x, y) = x^3 + 3y^3 - 3x - y$

$\frac{\partial f}{\partial x} = 3x^2 - 3$, zero at $x = \pm 1$

$\frac{\partial f}{\partial y} = 9y^2 - 1$, zero at $y = \pm \frac{1}{3}$

Stationary points at: $\left(1, \frac{1}{3}\right), \left(-1, \frac{1}{3}\right), \left(1, -\frac{1}{3}\right), \left(-1, -\frac{1}{3}\right)$.



$\frac{\partial^2 f}{\partial x^2} = 6x$, and $\frac{\partial^2 f}{\partial y^2} = 18y$

So at $\left(1, \frac{1}{3}\right)$ both $\frac{\partial^2 f}{\partial x^2}$ and $\frac{\partial^2 f}{\partial y^2}$ are positive and so the curvature is "concave up" in both directions: a minimum.

At $\left(-1, -\frac{1}{3}\right)$ both $\frac{\partial^2 f}{\partial x^2}$ and $\frac{\partial^2 f}{\partial y^2}$ are negative and so the curvature is "concave down" in both directions: a maximum (red peak).

At $\left(-1, \frac{1}{3}\right)$, $\frac{\partial^2 f}{\partial x^2}$ is negative, $\frac{\partial^2 f}{\partial y^2}$ is positive.

This gives a "saddle point".

At $\left(1, -\frac{1}{3}\right)$, $\frac{\partial^2 f}{\partial x^2}$ is positive, $\frac{\partial^2 f}{\partial y^2}$ is negative.

Again a saddle point.

It is not enough to check that $\frac{d^2 f}{dy^2}$ and $\frac{d^2 f}{dx^2}$ are both +ve or -ve to identify a minimum or maximum.

No: Consider $f(x, y) = x^2 + y^2 + axy$.

$\frac{\partial f}{\partial x} = 2x + ay,\quad \frac{\partial^2 f}{\partial x^2} = 2$, and

$\frac{\partial f}{\partial y} = 2y + ax,\quad \frac{\partial^2 f}{\partial y} = 2$.

So positive curvature (concave up) in both the x and y directions.

But as a varies the shape of the surface changes from a minimum to a saddle.

To identify a minimum or maximum we must also consider $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}\left[\frac{\partial f}{\partial y}\right] = (f_y)_x = f_{yx}$ and $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial x}\right] = f_{xy}$.

Note: For most C^2 functions $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$, but not in general.

We gather all the 2nd order partial derivatives into a matrix called the Hessian matrix:

$$\boldsymbol{H}_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} f_{xx} & f_{yx} \\ f_{xy} & f_{yy} \end{bmatrix}$$

This contains information about curvature in all directions.

Also $\boldsymbol{H}_f$ is the Jacobian of ∇f !!! [We will talk about Jacobians later]

A well-known test for determining the form of bivariate extrema.

Theorem:

Suppose $f(x,y) \in C^2$ and $f_x(x_0, y_0) = 0 = f_y(x_0, y_0)$, then

- (x_0, y_0) is a local maximum if $f_{xx}f_{yy} - f_{xy}^2 > 0$ and $f_{xx} < 0$ at (x_0, y_0)
- (x_0, y_0) is a local minimum if $f_{xx}f_{yy} - f_{xy}^2 > 0$ and $f_{xx} > 0$ at (x_0, y_0)
- (x_0, y_0) is a saddle point if $f_{xx}f_{yy} - f_{xy}^2 < 0$ at (x_0, y_0)
- If $f_{xx}f_{yy} - f_{xy}^2 = 0$ then the test is inconclusive and higher order derivatives must be analysed.

Note: $f_{xx}f_{yy} - f_{xy}^2 = f_{xx}f_{yy} - f_{xy}f_{yx} = \det(H_f)$ for smooth functions.

Oro Started Waffling

LEARN THIS

2nd derivative test and eigenvalues

Recall that for univariate functions:

$$f(x_0 + x) \approx f(x_0) + f'(x_0)x + \frac{1}{2}f''(x_0)x^2$$

For multivariate the linear approximation is:

$$f: \mathbb{R}^n \to \mathbb{R}, \qquad f(\boldsymbol{x_0} + \boldsymbol{v}) \approx f(\boldsymbol{x_0}) + \nabla f(\boldsymbol{x_0}).\boldsymbol{v}$$

The 2nd order partial derivatives again give a quadratic approximation:

$$f(\boldsymbol{x_0} + \boldsymbol{v}) \approx f(\boldsymbol{x_0}) + \nabla f(\boldsymbol{x_0}).\boldsymbol{v} + \frac{1}{2}\boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{H}_{\boldsymbol{f}}(\boldsymbol{x_0})\boldsymbol{v}$$

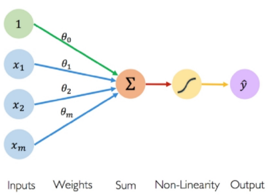
At a stationary point the second term is zero, so:

$$f(\boldsymbol{x_0} + \boldsymbol{v}) \approx f(\boldsymbol{x_0}) + \frac{1}{2}\boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{H}_{\boldsymbol{f}}(\boldsymbol{x_0})\boldsymbol{v}$$

Automatic Differentiation

Why?

An artificial neuron

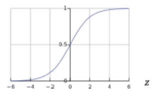


Activation Functions

$$\hat{y} = g(\theta_0 + X^T\theta)$$

- Example: sigmoid function

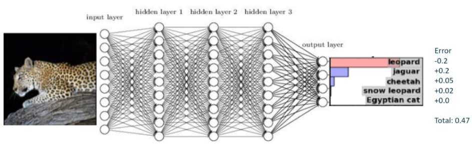
$$g(z) = \sigma(z) = \frac{1}{1+e^{-z}}$$



MIT: Alexander Amini, 2018 introtodeeplearning.com

Durham University

Learning



1. Give the network an input for which we know the correct output.
2. Compute the network's output vector.
3. Compute the error relative to the ground truth.
4. Adjust the 60-million parameters a tiny bit to reduce the error.

So we have a function $f: \mathbb{R}^{60000000} \to \mathbb{R}$ and we want to know how the error will be affected by a change in parameters and what direction to move our parameter vector in to get the greatest reduction in total error: we need to compute $v.\nabla f$ and ∇f.

Unfortunately, f is a large and complex function – but at least it is made up of the composition of many simpler functions!

Solution: use Automatic Differentiation to compute the gradient!

Computational Differentiation methods

Symbolic diff. → apply mathematical rules to generate solutions

Problem: a lot of terms.

Numerical diff. → Estimate derivative from limit formula

Problem: inaccurate for large systems

Automatic diff. → efficient and exact. :)

What is Automatic Differentiation?

It is just the chain rule

Suppose we have $e(f(g(h(u))))$

$$\frac{de}{du} = \frac{de}{df}\frac{df}{du} = \frac{de}{df}\frac{df}{dg}\frac{dg}{du} = \frac{de}{df}\frac{df}{dg}\frac{dg}{dh}\frac{dh}{du}$$

$$\frac{dh}{du} = \frac{dh}{du}$$

$$\frac{dg}{du} = \frac{dg}{dh}\frac{dh}{du}$$

[literally just chain rule]

$$\frac{df}{du} = \frac{df}{dg}\frac{dg}{du}$$

$$\frac{de}{du} = \frac{de}{df}\cdot\frac{df}{du}$$

example of chain rule

Example

$f(x) = \sin(\cos((\sin(x^3 - 2x)^5))$. Compute $f'(2)$

$g(x) = x^3 - 2x$	$\frac{dg}{dx} = 10$	$\frac{dg}{dx} = 10$
$h(g) = g^5$	$\frac{dh}{dg} = 1280$	$\frac{dh}{dx} = \frac{dh}{dg}\frac{dg}{dx} = 1280 \times 10 = 12800$
$i(h) = \sin(h)$	$\frac{di}{dh} = 0.98$	$\frac{di}{dx} = \frac{di}{dh}\frac{dh}{dx} = 0.98 \times 12800 = 12544$
$j(i) = \cos(i)$	$\frac{dj}{di} = 0.16$	$\frac{dj}{dx} = \frac{dj}{di}\frac{di}{dx} = 0.16 \times 12544 = 2007$
$k(j) = \sin(j)$	$\frac{dk}{dj} = 0.55$	$\frac{dk}{dx} = \frac{dk}{dj}\frac{dj}{dx} = 0.55 \times 2007 = 1103.9$

$k(j(i(h(g(x))))) = \sin(\cos((\sin(x^3 - 2x)^5))$

$$\frac{dk}{dx}(2) = \frac{dk}{dj}\frac{dj}{di}\frac{di}{dh}\frac{dh}{dg}\frac{dg}{dx}(2) = \cos(j)\,(-\sin(i))\cos(h)\,5g(3x^2 - 2)$$
$$= 10 \times 1280 \times 0.98 \times 0.16 \times 0.55 = 1103.9$$

tivariate chain rule

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial t}$$

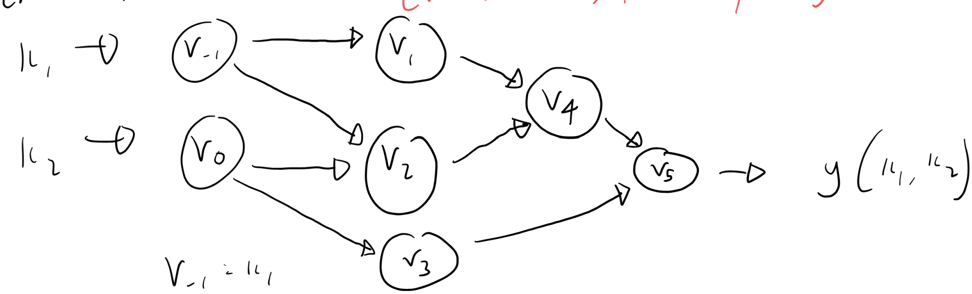
(Changing t by a small amount changes x by $\frac{\partial x}{\partial t}$

y by $\frac{\partial y}{\partial t}$

Example (of what?)

$$y(x_1, x_2) = \ln(x_1) + x_1 x_2 - \sin(x_2)$$

1. Create computation graph of atomic operations (where did this graph come from?)



$v_{-1} = x_1$

$v_0 = x_2$

$v_1 = \ln(v_{-1})$

$v_2 = v_{-1} \cdot v_0$

$v_3 = \sin(v_0)$

$v_4 = v_1 + v_2$

$v_5 = v_4 - v_3$

② Evaluate the function

Forward Primal Trace

v_{-1}	$= x_1$	$= 2$
v_0	$= x_2$	$= 5$
v_1	$= \ln v_{-1}$	$= \ln 2$
v_2	$= v_{-1} \times v_0$	$= 2 \times 5$
v_3	$= \sin v_0$	$= \sin 5$
v_4	$= v_1 + v_2$	$= 0.693 + 10$
v_5	$= v_4 - v_3$	$= 10.693 + 0.959$
y	$= v_5$	$= 11.652$

$$\frac{\partial y}{\partial x_1} = \frac{\partial v_5}{\partial x_1} = \frac{\partial v_5}{\partial v_4}\frac{\partial v_4}{\partial x_1} + \frac{\partial v_5}{\partial v_3}\frac{\partial v_3}{\partial x_1}$$

v_0 and x_1 unrelated so $= 0$

$$\frac{\partial v_3}{\partial x_1} = \frac{\partial v_3}{\partial v_0} \cdot \frac{\partial v_0}{\partial x_1} = 0$$

$$\frac{\partial v_4}{\partial x_1} = \frac{\partial v_4}{\partial v_1} \cdot \frac{\partial v_1}{\partial x_1} + \frac{\partial v_4}{\partial v_2} \cdot \frac{\partial v_2}{\partial x_1}$$

Substituting in gives us:

$$\frac{\partial y}{\partial x_1} = \frac{\partial v_5}{\partial x_1} = \frac{\partial v_5}{\partial v_4}\frac{\partial v_4}{\partial v_1}\frac{\partial v_1}{\partial x_1} + \frac{\partial v_5}{\partial v_4}\frac{\partial v_4}{\partial v_2}\frac{\partial v_2}{\partial x_1}$$

$$\frac{\partial v_1}{\partial x_1} = \frac{\partial v_1}{\partial v_{-1}}\frac{\partial v_{-1}}{\partial x_1} = \frac{\partial v_1}{\partial v_{-1}}.1$$

$$\frac{\partial v_2}{\partial x_1} = \frac{\partial v_2}{\partial v_0}\frac{\partial v_0}{\partial x_1} + \frac{\partial v_2}{\partial v_{-1}}\frac{\partial v_{-1}}{\partial x_1} = 0 + \frac{\partial v_2}{\partial v_{-1}}.1$$

But as before, rather than working everything out separately and doing a large multiplication, instead we multiply at each stage and keep a running total.
I.e. work out in turn:
$\frac{\partial v_i}{\partial x_1}$ for i from -1 to 5. evaluated at (2,5)

③ Compute partial derivatives of atomic functions

$\dot{v}_j = \frac{\partial v_j}{\partial x}$

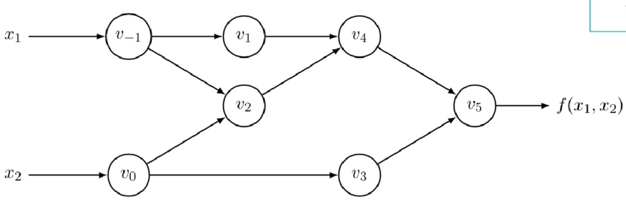
Forward Primal Trace			Forward Tangent (Derivative)	
v_{-1}	$= x_1$	$= 2$	$\dot{v}_{-1}$	$= \dot{x}_1$
v_0	$= x_2$	$= 5$	$\dot{v}_0$	$= \dot{x}_2$
v_1	$= \ln v_{-1}$	$= \ln 2$	$\dot{v}_1$	$= \dot{v}_{-1}/v_{-1}$
v_2	$= v_{-1} \times v_0$	$= 2 \times 5$	$\dot{v}_2$	$= \dot{v}_{-1} \times v_0 + \dot{v}_0 \times v_{-1}$
v_3	$= \sin v_0$	$= \sin 5$	$\dot{v}_3$	$= \dot{v}_0 \times \cos v_0$
v_4	$= v_1 + v_2$	$= 0.693 + 10$	$\dot{v}_4$	$= \dot{v}_1 + \dot{v}_2$
v_5	$= v_4 - v_3$	$= 10.693 + 0.959$	$\dot{v}_5$	$= \dot{v}_4 - \dot{v}_3$
y	$= v_5$	$= 11.652$	$\dot{\boldsymbol{y}}$	$= \dot{\boldsymbol{v}}_5$

④

Step 4a: Evaluate derivative $\frac{\partial y}{\partial x_i}$ by setting $\dot{x}_i = 1$ others $= 0$.

Example: $y(x_1, x_2) = \ln(x_1) + x_1 x_2 - \sin(x_2)$ at point (2,5)

Notation:
$\dot{v}_j = \frac{\partial v_j}{\partial \boldsymbol{x}}$

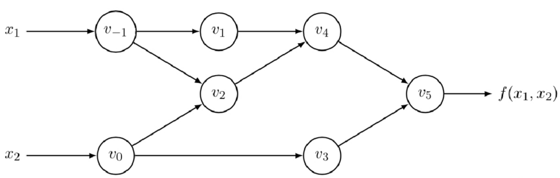


Forward Primal Trace			Forward Tangent (Derivative) Trace		
v_{-1}	$= x_1$	$= 2$	$\dot{v}_{-1}$	$= \dot{x}_1$	$= 1$
v_0	$= x_2$	$= 5$	$\dot{v}_0$	$= \dot{x}_2$	$= 0$
v_1	$= \ln v_{-1}$	$= \ln 2$	$\dot{v}_1$	$= \dot{v}_{-1}/v_{-1}$	$= 1/2$
v_2	$= v_{-1} \times v_0$	$= 2 \times 5$	$\dot{v}_2$	$= \dot{v}_{-1} \times v_0 + \dot{v}_0 \times v_{-1}$	$= 1 \times 5 + 0 \times 2$
v_3	$= \sin v_0$	$= \sin 5$	$\dot{v}_3$	$= \dot{v}_0 \times \cos v_0$	$= 0 \times \cos 5$
v_4	$= v_1 + v_2$	$= 0.693 + 10$	$\dot{v}_4$	$= \dot{v}_1 + \dot{v}_2$	$= 0.5 + 5$
v_5	$= v_4 - v_3$	$= 10.693 + 0.959$	$\dot{v}_5$	$= \dot{v}_4 - \dot{v}_3$	$= 5.5 - 0$
y	$= v_5$	$= 11.652$	$\dot{\boldsymbol{y}}$	$= \dot{\boldsymbol{v}}_5$	$= \mathbf{5.5}$

$\frac{\partial y}{\partial x_1}\mathbf{(2, 5)}$

Step 4b: Evaluate directional derivative in direction (2,1)

Example: $y(x_1, x_2) = \ln(x_1) + x_1 x_2 - \sin(x_2)$ at point (2,5)



Forward Tangent (Derivative)	
$\dot{v}_{-1}$	$= \dot{x}_1$
$\dot{v}_0$	$= \dot{x}_2$
$\dot{v}_1$	$= \dot{v}_{-1}/v_{-1}$
$\dot{v}_2$	$= \dot{v}_{-1} \times v_0 + \dot{v}_0 \times v_{-1}$
$\dot{v}_3$	$= \dot{v}_0 \times \cos v_0$
$\dot{v}_4$	$= \dot{v}_1 + \dot{v}_2$
$\dot{v}_5$	$= \dot{v}_4 - \dot{v}_3$
$\dot{\boldsymbol{y}}$	$= \dot{\boldsymbol{v}}_5$

Unit vector in direction (2,1) is $\left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right)$.

Set $\dot{x}_1 = \frac{2}{\sqrt{5}}, \dot{x}_2 = \frac{1}{\sqrt{5}}$.

$\dot{v}_1 = \frac{2}{\sqrt{5}} . \frac{1}{2} = \frac{1}{\sqrt{5}}$

$\dot{v}_2 = \frac{2}{\sqrt{5}} . 5 + \frac{1}{\sqrt{5}} . 2 = \frac{12}{\sqrt{5}}$

$\dot{v}_3 = \frac{1}{\sqrt{5}} . \cos 5$

$\dot{v}_4 = \frac{1}{\sqrt{5}} + \frac{12}{\sqrt{5}} = \frac{13}{\sqrt{5}}$

$\dot{v}_5 = \frac{13}{\sqrt{5}} - \frac{\cos 5}{\sqrt{5}} \approx 5.69$

Forward mode: when $f: \mathbb{R}^n \to \mathbb{R}$ one pass can compute the directional derivative.
But forward mode requires n passes to calculate the full gradient ∇f.

To compute ∇f in one pass, use reverse mode

Suppose we have a function $e(f(g(h(x))))$

Chain rule: $\frac{de}{dx} = \frac{de}{df}\frac{df}{dg}\frac{dg}{dh}\frac{dh}{dx}$

Working from the inside:

$$\frac{dh}{dx} = \frac{dh}{dx}$$
$$\frac{dg}{dx} = \frac{dg}{dh}\frac{dh}{dx}$$
$$\frac{df}{dx} = \frac{df}{dg}\frac{dg}{dx}$$
$$\frac{de}{dx} = \frac{de}{df}\frac{df}{dx}$$

Durham University

forward mode

(all derivatives calculated are in terms of x, so for every value of x, all need to be recalculated)

Now: Work from the outside!

$$\frac{de}{de} = 1$$
$$\frac{de}{df} = \frac{de}{de}\frac{de}{df}$$
$$\frac{de}{dg} = \frac{de}{df}\frac{df}{dg}$$
$$\frac{de}{dh} = \frac{de}{dg}\frac{dg}{dh}$$
$$\frac{de}{dx} = \frac{de}{dh}\frac{dh}{dx}$$

only this calculation in terms of x so changing x only results in this needing to be recalculated

reverse mode

Notation $\bar{v}_i = \frac{\partial y}{\partial v_i}$

Computation:

- Start with $v_5 = y$, so $\overline{v_5} = \frac{\partial y}{\partial v_5} = 1$.
- Work backwards through the previous computation adding in the contribution of each variable as it occurs using the chain rule.
- $\overline{v_4} = \frac{\partial y}{\partial v_4} = \frac{\partial y}{\partial v_5}\frac{\partial v_5}{\partial v_4}$ since v_4 only contributes to v_5.
- $$\overline{v_0} = \bar{v}_2\frac{\partial v_2}{\partial v_0} + \bar{v}_3\frac{\partial v_3}{\partial v_0}$$ since v_0 contributes to v_2 and v_3.

Forward Primal Trace

$v_{-1} = x_1$	$= 2$	
$v_0 = x_2$	$= 5$	
$v_1 = \ln v_{-1}$	$= \ln 2$	
$v_2 = v_{-1} \times v_0$	$= 2 \times 5$	
$v_3 = \sin v_0$	$= \sin 5$	
$v_4 = v_1 + v_2$	$= 0.693 + 10$	
$v_5 = v_4 - v_3$	$= 10.693 + 0.959$	
$y = v_5$	$= 11.652$	

Reverse Adjoint (Derivative) Trace $\nabla y(2,5)$

$\bar{x}_1 = \bar{v}_{-1}$		$\mathbf{= 5.5}$
$\bar{x}_2 = \bar{v}_0$		$\mathbf{= 1.716}$
$\bar{v}_{-1} = \bar{v}_{-1} + \bar{v}_1 \frac{\partial v_1}{\partial v_{-1}}$	$= \bar{v}_{-1} + \bar{v}_1 / v_{-1}$	$= 5.5$
$\bar{v}_0 = \bar{v}_0 + \bar{v}_2 \frac{\partial v_2}{\partial v_0}$	$= \bar{v}_0 + \bar{v}_2 \times v_{-1}$	$= 1.716$
$\bar{v}_{-1} = \bar{v}_2 \frac{\partial v_2}{\partial v_{-1}}$	$= \bar{v}_2 \times v_0$	$= 5$
$\bar{v}_0 = \bar{v}_3 \frac{\partial v_3}{\partial v_0}$	$= \bar{v}_3 \times \cos v_0$	$= -0.284$
$\bar{v}_2 = \bar{v}_4 \frac{\partial v_4}{\partial v_2}$	$= \bar{v}_4 \times 1$	$= 1$
$\bar{v}_1 = \bar{v}_4 \frac{\partial v_4}{\partial v_1}$	$= \bar{v}_4 \times 1$	$= 1$
$\bar{v}_3 = \bar{v}_5 \frac{\partial v_5}{\partial v_3}$	$= \bar{v}_5 \times (-1)$	$= -1$
$\bar{v}_4 = \bar{v}_5 \frac{\partial v_5}{\partial v_4}$	$= \bar{v}_5 \times 1$	$= 1$
$\bar{v}_5 = \bar{y}$	$= 1$	

Vector - valued Functions

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)^T$$

f is differentiable at x_0 if for all unit vectors $\underline{v}$ we have $\nabla_v f(\underline{x}_0) = \underline{v} \cdot \nabla f(\underline{x}_0)$

So far we have considered a vector input and real valued output.

E.g. $f: \mathbb{R}^n \to \mathbb{R}$ given by $f(x_1, x_2, \dots, x_n) = \sum_j \frac{x_j}{n}$ might give the average brightness of a digital camera sensor.

What if we instead wanted to adjust the brightness?

We could map each pixel i with a function $f_i(x_1, x_2, \dots, x_n) = x_i - \sum_j \frac{x_j}{n} + 128$ to make the average brightness 128.

But rather than writing out separate functions for every pixel, we can think of the output as a vector $(f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \dots, f_n(\boldsymbol{x}))$.

Now we have a **vector valued** function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^n$.

The derivative of the i^{th} component of the output with respect to the j^{th} input is:

$$\frac{\partial f_i}{\partial x_j}.$$

We have $\underline{f}: \mathbb{R}^n \to \mathbb{R}^m$ where $\underline{f} = (f_1(\underline{x}), \dots, f_m(\underline{x}))$

Set $\underline{J}_{ij} = \frac{\partial f_i}{\partial x_j}$

The resulting matrix of partial derivatives is called the Jacobian Matrix

$$\underline{J} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

So the ith row of this matrix is the overall gradient, ∇f_i

jth column is $\frac{\partial f}{\partial x_j}$

Examples:

$$f : \mathbb{R}^3 \to \mathbb{R}^2 = (\underbrace{2x_1 + x_2 - x_3}_{f_1}, \underbrace{x_1 + 3x_2 - 4x_3}_{f_2})$$

Inputs: x_1, x_2, x_3
Outputs: f_1, f_2

$$g : \mathbb{R}^2 \to \mathbb{R}^2 = (\sin(x_1 + x_2), \cos(x_1 - x_2))$$

Inputs: x_1, x_2
Outputs: g_1, g_2

$$\frac{\partial f_1}{\partial x_1} = \frac{\partial(2x_1 + x_2 - x_3)}{\partial x_1} = 2$$

$$\frac{\partial f_1}{\partial x_2}$$

$$J_f = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 3 & -4 \end{pmatrix}$$

$$\frac{\partial f_2}{\partial x_1}$$

etc.

$$J_g = \begin{pmatrix} \cos(x_1 + x_2) & \cos(x_1 + x_2) \\ -\sin(x_1 - x_2) & \sin(x_1 - x_2) \end{pmatrix}$$

e.g. $J_g\left(\frac{3\pi}{4}, \frac{\pi}{4}\right) = \begin{pmatrix} \cos\pi & \cos\pi \\ -\sin\frac{\pi}{2} & \sin\frac{\pi}{2} \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}$

What about $g \circ f : \mathbb{R}^3 \to \mathbb{R}^2$?

$$g \circ f = (\sin(3x_1 + 4x_2 - 5x_3), \cos(x_1 - 2x_2 + 3x_3))$$

replace x_1 with f_1
x_2 with f_2

Since g takes f as inputs

Since $g_1 = \sin(x_1 + x_2)$

$$(g \circ f)_1 = \sin(f_1 + f_2)$$
$$= \sin(2x_1 + x_2 - x_3 + x_1 + 3x_2 - 4x_3)$$
$$=$$

$$J_{g\circ f} = \begin{array}{ccc} \frac{\partial (g\circ f)_1}{\partial x_1} & \frac{\partial (g\circ f)_1}{\partial x_2} & \\ \begin{pmatrix} 3\cos(3x_1+4x_2-5x_3) & 4\cos(3x_1+4x_2-5x_3) & -5\cos(3x_1+4x_2-5x_3) \\ -\sin(x_1-2x_2+3x_3) & 2\sin(x_1-2x_2+3x_3) & -3\sin(x_1-2x_2+3x_3) \end{pmatrix} \\ \frac{\partial (g\circ f)_2}{\partial x_1} & \text{etc} & \end{array}$$

e.g.

$$J_{g\circ f}\left(\frac{8\pi}{20}, -\frac{\pi}{20}, 0\right) = \begin{pmatrix} 3\cos(\pi) & 4\cos(\pi) & -5\cos(\pi) \\ -\sin(\pi/2) & 2\sin(\pi/2) & -3\sin(\pi/2) \end{pmatrix} = \begin{pmatrix} -3 & -4 & 5 \\ -1 & 2 & -3 \end{pmatrix}$$

Chain rule: $\underline{J}_{g\circ f} = \underline{J}_g \underline{J}_f$

$$\frac{\partial\, g\circ f_i}{\partial x_j} = \frac{\partial g_i}{\partial f_1}\frac{\partial f_1}{\partial x_j} + \frac{\partial g_i}{\partial f_2}\frac{\partial f_2}{\partial x_j} = \nabla g_i \frac{\partial \underline{f}}{\partial x_j}$$

Linear approximations

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

So $f(x+h) \approx f(x) + hf'(x)$

For a surface we have seen that $\boldsymbol{v}.\nabla f(\boldsymbol{x_0})$ is the slope in the direction $\boldsymbol{v}$.

$$f(\boldsymbol{x_0} + h.\boldsymbol{v}) \approx f(\boldsymbol{x_0}) + h\boldsymbol{v}.\nabla f(\boldsymbol{x_0})$$

Which is

$$f(x_0 + h.v_x, y_0 + h.v_y) \approx f(x_0, y_0) + h.v_x\frac{\partial f}{\partial x} + h.v_y\frac{\partial f}{\partial y}$$

For a multivariate function: $\mathbb{R}^n \to \mathbb{R}^m$ the Jacobian matrix gives the best linear approximation to $\underline{f}$ at $\underline{x}$.

$$f(\underline{x}+h\underline{v}) \simeq \underline{f}(\underline{x}) + J_f(\underline{x})(h\underline{v})$$

where $\underline{x}$, $\underline{v}$, $h\underline{v}$ are n-dimensional vectors

J_f is a $m \times n$ matrix

$\underline{f}(\underline{x})$, $J_f(\underline{x})(h\underline{v})$ are m dimensional vectors.

Example:

$\boldsymbol{g}: \mathbb{R}^2 \to \mathbb{R}^2 = (\sin(x_1+x_2), \cos(x_1-x_2))$

$$\boldsymbol{J_g} = \begin{pmatrix} \cos(x_1+x_2) & \cos(x_1+x_2) \\ -\sin(x_1-x_2) & \sin(x_1-x_2) \end{pmatrix}$$

$$\boldsymbol{J_g}\left(\frac{3\pi}{4}, \frac{\pi}{4}\right) = \begin{pmatrix} \dfrac{\partial g_1}{\partial x_1} & \dfrac{\partial g_1}{\partial x_2} \\ \dfrac{\partial g_2}{\partial x_1} & \dfrac{\partial g_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} \cos(\pi) & \cos(\pi) \\ -\sin(\pi/2) & \sin(\pi/2) \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}$$

So near $\left(\frac{3\pi}{4}, \frac{\pi}{4}\right)$ we can approximate $\boldsymbol{g}$ by:

$$\boldsymbol{g}\left(\frac{3\pi}{4}+\delta_x, \frac{\pi}{4}+\delta_y\right) = \underbrace{\boldsymbol{g}\left(\frac{3\pi}{4}, \frac{\pi}{4}\right)}_{f(\underline{x})} + \underbrace{\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}\overbrace{\begin{pmatrix} \delta_x \\ \delta_y \end{pmatrix}}^{h\underline{v}}}_{J_f(\underline{x})(h\underline{v})} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} -\delta_x - \delta_y \\ -\delta_x + \delta_y \end{pmatrix} = \begin{pmatrix} -\delta_x - \delta_y \\ -\delta_x + \delta_y \end{pmatrix}$$

Extrema

To identify a minimum or maximum we must also consider $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial}{\partial x}\left[\frac{\partial f}{\partial y}\right] = f_{yx}$ and $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left[\frac{\partial f}{\partial x}\right] = f_{xy}$.

Note: For most C^2 functions $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$, but not in general.

We gather all the 2nd order partial derivatives into a matrix called the Hessian matrix:

$$\boldsymbol{H}_f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x^2} & \dfrac{\partial^2 f}{\partial x \partial y} \\ \dfrac{\partial^2 f}{\partial y \partial x} & \dfrac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} f_{xx} & f_{yx} \\ f_{xy} & f_{yy} \end{bmatrix}$$

This contains information about curvature in all directions.

For C^2 this is symmetric. Also $\boldsymbol{H}_f$ is the Jacobian of ∇f !!!

Exponential, Logarithmic and Hyperbolic functions

Euler's number, $e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = \lim_{n \to \infty} \left(1 + n \cdot \frac{1}{n} + \binom{n}{2}\left(\frac{1}{n}\right)^2 + \dots + \binom{n}{n} \cdot \left(\frac{1}{n}\right)^n\right)$

Suppose we are interested in a function $f(x) = a^x$ for some constant $a \geq 1$, such that the graph of f at $x = 0$ has gradient 1. I.e. $f'(0) = 1$.

By definition

$$f'(0) = \lim_{h \to 0}\left(\frac{a^{0+h} - a^0}{h}\right) = \lim_{h \to 0}\left(\frac{a^h - 1}{h}\right).$$

If we take $h = 1/n$ then as $n \to \infty$, $h \to 0$.

So a is being chosen such that

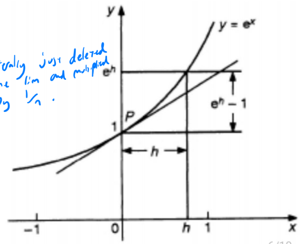
$$f'(0) = \lim_{n \to \infty}\left(\frac{a^{1/n} - 1}{1/n}\right) = 1.$$

literally just deleted the lim and multiplied by $1/n$.

I.e. for large n we have $a^{1/n} - 1 \approx 1/n$, or

$$a \approx \left(1 + \frac{1}{n}\right)^n$$

And in the limit $a = \lim_{n \to \infty}\left(1 + \frac{1}{n}\right)^n = e$.



Consider $f(x) = e^x$

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

By chain rule: $\frac{d}{dx} e^{g(x)} = g'(x) e^{g(x)}$

e^x grows more rapidly than x^p for any positive p

Using

$$e = \sum_{r=0}^{\infty} \frac{1}{r!} = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \frac{1}{720} + \cdots$$

We can approximate $e = 2.71821828459$ ...

Also (no proof) $e^x = \sum_{r=0}^{\infty} \frac{x^r}{r!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots$

Theorem: $\lim_{x \to \infty} \frac{e^x}{x^p} \to \infty$ for all $p > 0$.

Proof:

$$\frac{e^x}{x^p} = \frac{1 + x + \frac{x^2}{2!} + \cdots + \frac{x^p}{p!} + \boxed{\frac{x^{p+1}}{(p+1)!}} + \cdots}{\boxed{x^p}} > \frac{x}{(p+1)!}$$

this is the same as this. As $x \to \infty$, LHS has additional positive terms so LHS > RHS

So

$$\lim_{x \to \infty} \frac{e^x}{x^p} > \lim_{x \to \infty} \frac{x}{(p+1)!} \to \infty.$$

Corollary: $\lim_{x \to \infty} \frac{x^p}{e^x} = 0$ for all $p > 0$.

If $f(x) = e^x$, $f: \mathbb{R} \to \mathbb{R}^{>0}$ is bijective so an inverse exists.

$f^{-1}(x) = \ln x = \log_e x$.

If $y = \ln x$, $e^y = x$ → $\frac{d}{dx}$ both sides (implicit for LHS)

$$\frac{dy}{dx} e^y = 1$$

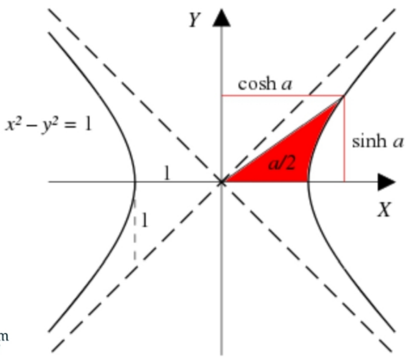
$$\frac{dy}{dx} = \frac{1}{e^y} = \frac{1}{x}$$

$\ln(x)$ grows more slowly than x^p for any $p>0$

Hyperbolic functions

Given the **unit hyperbola** $x^2 - y^2 = 1$, the hyperbolic functions are defined by the coordinates of a point at which area $\alpha/2$ has been swept out:

Defined to be $(\cosh\alpha, \sinh\alpha)$. "hyperbolic sin/cos" or "sinsh" and "cosh"



With a bit more work, we can derive the formulae:

$$\sinh(\alpha) = \frac{e^\alpha - e^{-\alpha}}{2}$$

and

$$\cosh(\alpha) = \frac{e^\alpha + e^{-\alpha}}{2}$$

Certainly, the formulae give the set of points on the hyperbola.

Durham University

Hyperbolic functions

As for regular trigonometry, there are the following variants:

$$\tan\theta = \frac{\sin\theta}{\cos\theta} \qquad \tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\cot\theta = \frac{1}{\tan\theta} = \frac{\cos\theta}{\sin\theta} \qquad \coth x = \frac{1}{\tanh x} = \frac{e^x + e^{-x}}{e^x - e^{-x}}$$

$$\sec\theta = \frac{1}{\cos\theta} \qquad \operatorname{sech} x = \frac{2}{e^x + e^{-x}}$$

$$\operatorname{cosec}\theta = \frac{1}{\sin\theta} \qquad \operatorname{cosech} x = \frac{2}{e^x - e^{-x}}$$

Derivatives of hyperbolic functions

$$\frac{d}{dx}[\sinh x] = \frac{d}{dx}\left[\frac{e^x - e^{-x}}{2}\right] = \frac{e^x + e^{-x}}{2} = \cosh x$$

$$\frac{d}{dx}[\cosh x] = \frac{d}{dx}\left[\frac{e^x + e^{-x}}{2}\right] = \frac{e^x - e^{-x}}{2} = \sinh x$$

$$\frac{d}{dx}[\tanh x] = \frac{d}{dx}\left[\frac{e^x - e^{-x}}{e^x + e^{-x}}\right] = \frac{(e^x+e^{-x})(e^x+e^{-x})-(e^x-e^{-x})(e^x-e^{-x})}{(e^x+e^{-x})^2} = \frac{\cosh^2 x-\sinh^2 x}{\cosh^2 x} = \operatorname{sech}^2 x$$

$$\frac{d}{dx}[\sigma(x)] = \frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right] = \frac{-(-e^{-x})}{(1+e^{-x})^2} = \frac{1}{(1+e^{-x})}\frac{((1+e^{-x})-1)}{(1+e^{-x})} = \sigma(x)[1-\sigma(x)]$$