Calculus Refresher

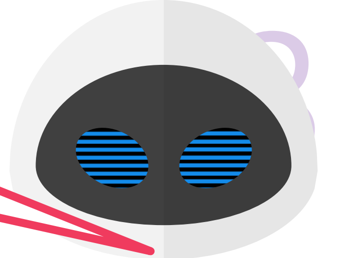
CS771: Introduction to Machine Learning

Purushottam Kar

Topics to be Covered

- Calculus basics: extrema, saddle points, gradient, Hessian,
- Dealing with non-differentiable functions
- Convex sets and convex functions

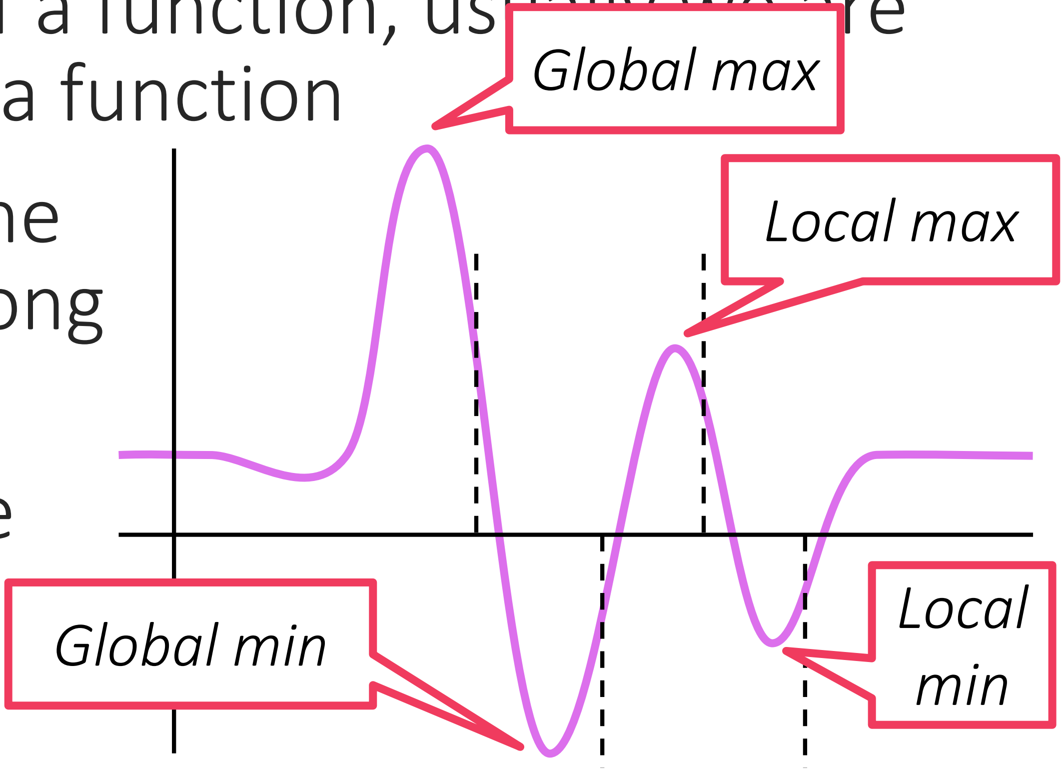
Extrema



Since we always seek the "best" values of a function, usually we are looking for the maxima or the minima of a function

Global extrema: a point which achieves the best value of the function (max/min) among all the possible points

Local extrema: a point which achieves the best value of the function only in a small region surrounding that point



Most machine learning algorithms love to find the global extrema

- E.g. we saw that CSVM wanted to find the model with max margin

Sometimes it is difficult so we settle for local extrema (e.g. deepnets)

Derivat

Derivatives only tell us how f will behave close to the point at which the derivative was calculated. If you move too much in direction of derivative, f may start decreasing. Similarly, if you move too much opposite to derivate, f may start increasing

…oint tells us
…r we should move left or right on the number line to *increase* f.
…*is positive,* …

Corollary of Taylor's Theorem

$$f(x + \Delta x) \approx f(x) + \Delta x \cdot f'(x)$$

if Δx is "small"

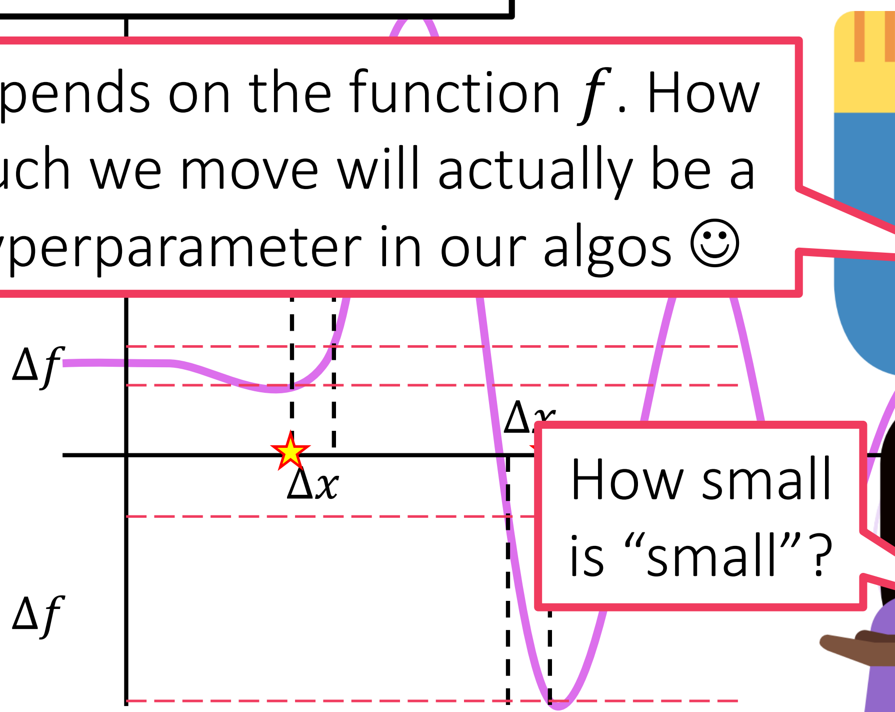
Magnitude of the d… …ld f increase if we moved a teeny tiny…

If we move a little bit opposite to the direction of derivative, then f would *decrease*

Depends on the function f. How much we move will actually be a hyperparameter in our algos ☺

What if I moved in the opposite direction of the derivative?

Why do you keep saying "little bit"? What if I move a lot?



How small is "small"?

Stationary Points

If $f''(x) < 0$ and $f'(x) = 0$ then derivative moves from +ve to -ve around this point – local/global max!

These are places where the derivative vanishes i.e. is 0

If $f''(x) = 0$ and $f'(x) = 0$ then this may be extrema/saddle – higher derivatives e.g. $f'''(x)$ needed

These can be local

The derivative bei

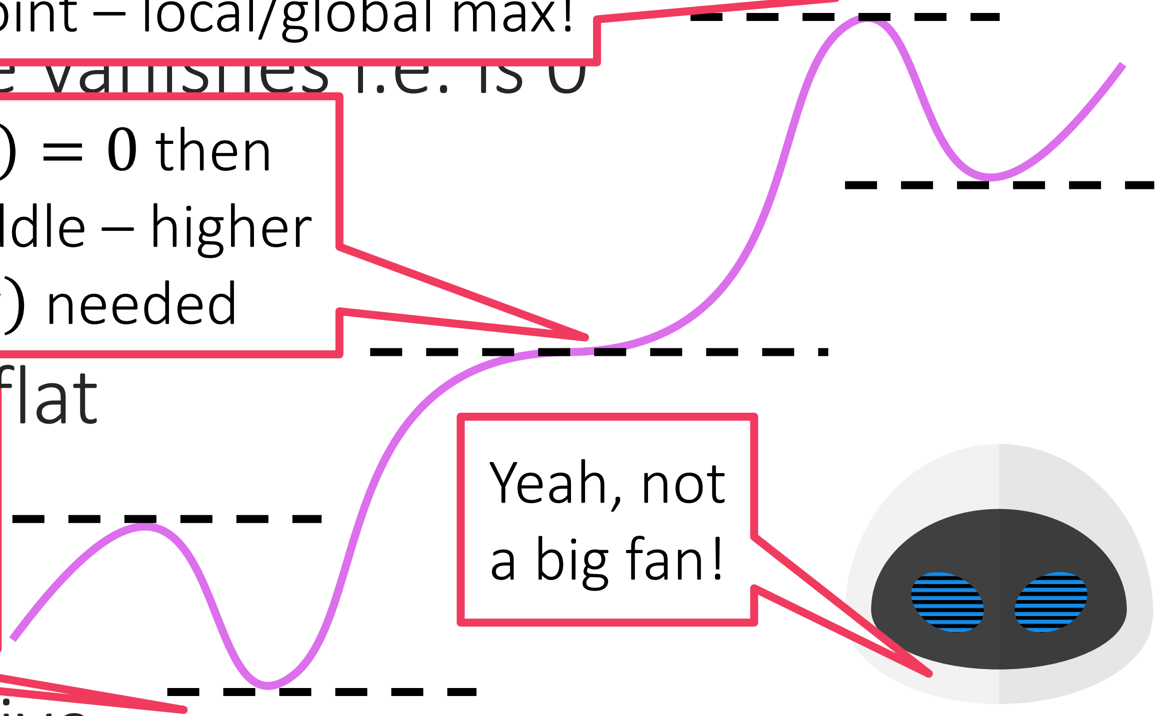
that at t ... flat

If $f''(x) > 0$ and $f'(x) = 0$ then derivative moves from -ve to +ve around this point – local/global min!

Saddle p

We can

is saddle or extrema using 2nd derivative



Just as sign of the derivative tells us if the function is increasing or decreasing if we move left a tiny bit, the 2nd derivative tells us if the derivative is increasing or decreasing if we move left a tiny bit

Rules of derivatives

Sum Rule: $\left(f(x) + g(x)\right)' = f'(x) + g'(x)$

Scaling Rule: $\left(a \cdot f(x)\right)' = a \cdot f'(x)$ if a is not a function of x

Product Rule: $\left(f(x) \cdot g(x)\right)' = f'(x) \cdot g(x) + g'(x) \cdot f(x)$

Quotient Rule: $\left(f(x)/g(x)\right)' = \left(f'(x) \cdot g(x) - g'(x)f(x)\right)/\left(g(x)\right)^2$

Chain Rule: $\left(f\left(g(x)\right)\right)' \stackrel{\text{def}}{=} (f \circ g)'(x) = f'\left(g(x)\right) \cdot g'(x)$

Most common use f is a function of t but $t = g(x)$, calculate df/dx

Multivariate Functio

Gradient

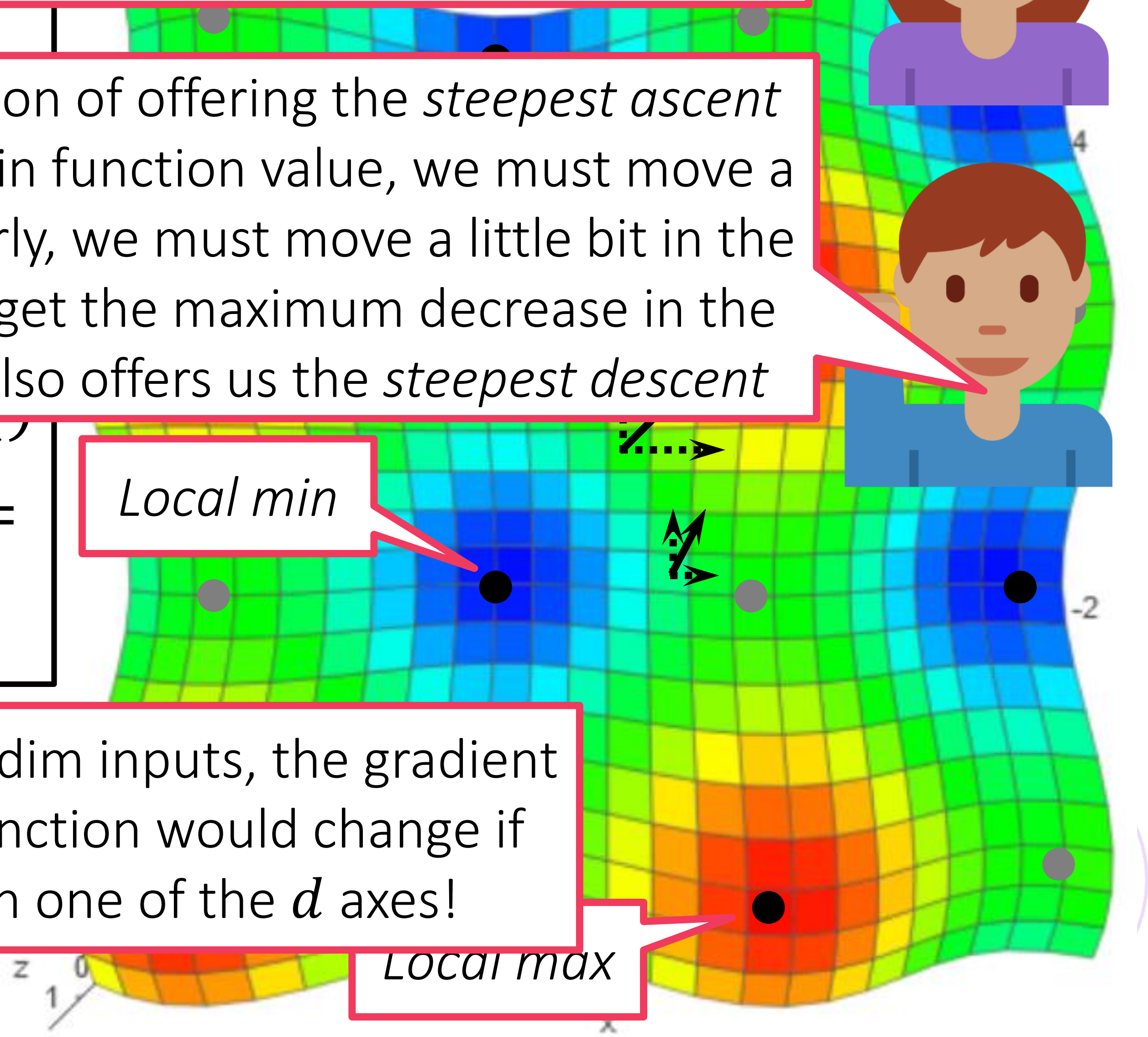
∇f

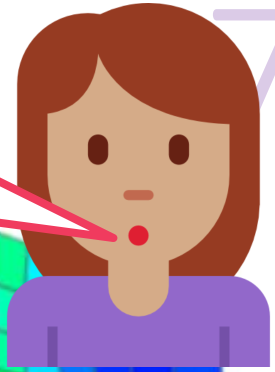
Taylor's

If we move

then $f(\mathbf{x}+\mathbf{t}) \approx f(\mathbf{x}) + \sum_{i=1}^{d} t_i \cdot \frac{\partial f(\mathbf{x})}{\partial x_i} =$

$f(\mathbf{x}) + \mathbf{t}^\top \nabla f(\mathbf{x})$ if $\mathbf{t}$ is "small"





This looks just like the 1D case except that we are summing up contributions from all d dimensions

The gradient also has the distinction of offering the *steepest ascent* i.e. if we want maximum increase in function value, we must move a little bit along the gradient. Similarly, we must move a little bit in the direction opposite to gradient to get the maximum decrease in the function value, i.e. the gradient also offers us the *steepest descent*

For multivariate functions with d-dim inputs, the gradient simply records how much the function would change if we move a little bit along each one of the d axes!

PLOTS COURTESY ACADEMO.ORG

Higher derivatives in higher dimensions

2nd derivative of $f: \mathbb{R}^d \to \mathbb{R}$ is a $d \times d$ matrix called the *Hessian*

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_d} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d x_1} & \frac{\partial^2 f}{\partial x_d x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$$

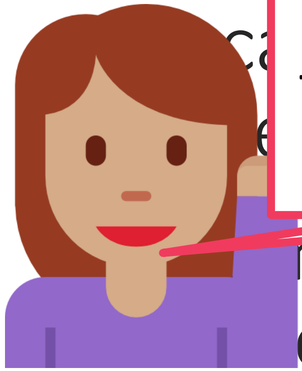
May get difficult to *visualize* higher derivatives – just go with the math

3rd and higher derivatives must be expressed as *tensors*

All rules of derivatives (chain, product etc) apply here as well

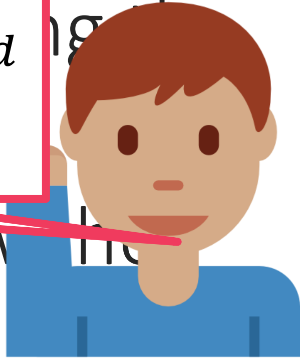
Stationary Points in d-dimensions

These are places where the gradient vanishes i.e. is a zero vector!



If a matrix satisfies $\mathbf{x}^\top A\mathbf{x} < 0$ for all $\mathbf{x} \in \mathbb{R}^d$ then it is called *negative definite (ND)*

Recall that if a square $d \times d$ symmetric matrix A satisfies $\mathbf{x}^\top A\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^d$ then it is *positive definite (PD)*

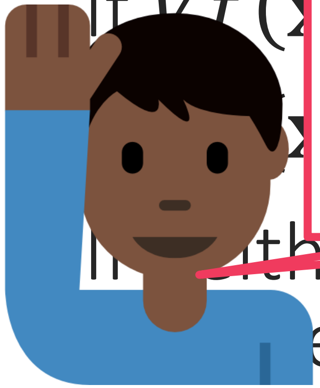


...ca... ...ona... ...ng... ...e... ...D

...nore complicated to visualize, but the Hessian tells us how ...he...

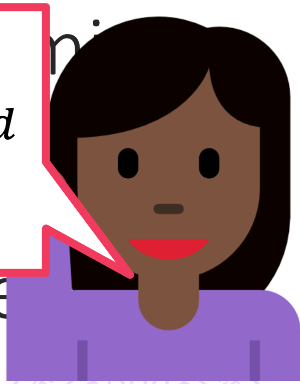
...e of the function is curved at a point

If $\nabla f(\mathbf{x}$ ... PD ... mi...



If a matrix satisfies $\mathbf{x}^\top A\mathbf{x} \leq 0$ for all $\mathbf{x} \in \mathbb{R}^d$ then it is called *negative semidefinite (NSD)*

Recall that if a square $d \times d$ symmetric matrix A satisfies $\mathbf{x}^\top A\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^d$ then it is *positive semidefinite (PSD)*



...N...

If ...ither of these are true, then either $\mathbf{x}$ is a saddle point or the...

...eed higher order derivatives to verify

Whether point is saddle or test has failed depends on ***eigenvalues*** *of* $\nabla f(\mathbf{x})$

We will learn about eigenvalues in a few weeks when we refresh linear algebra

A Toy Example – Function Values

	0	1	2	3	4	5	6	7	8
6	3	3	3	3	3	3	3	3	3
5	2	2	2	3	4	3	3	2	1
4	1	1	1	3	3	3	1	1	1
3	1	0	1	1	2	1	1	0	1
2	1	1	1	3	3	3	1	1	1
1	1	2	3	3	4	3	2	2	2
0	3	3	3	3	3	3	3	3	3

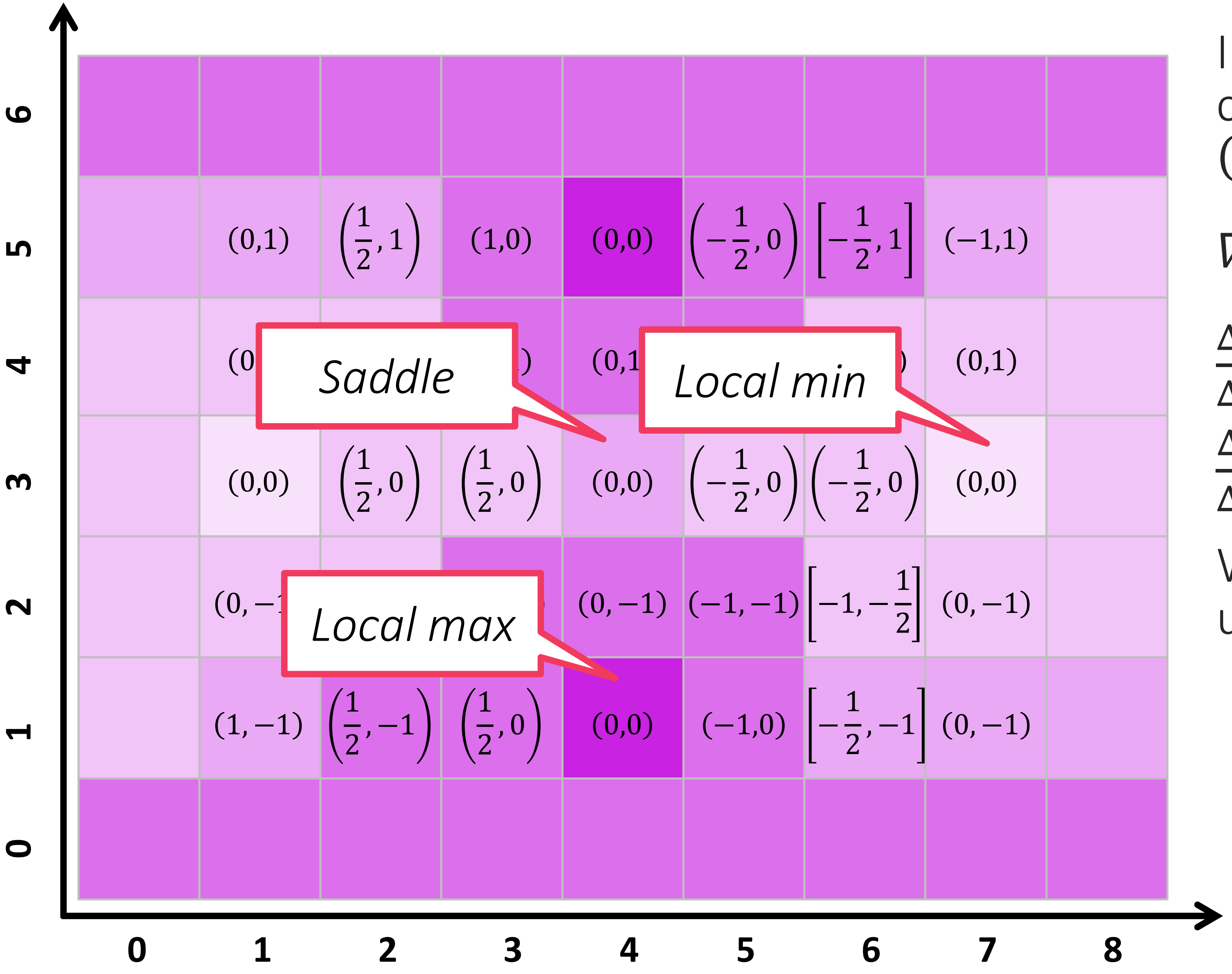
In this discrete toy example, we can calculate gradient at a point (x_0, y_0) as

$$\nabla f(x_0, y_0) = \left(\frac{\Delta f}{\Delta x}, \frac{\Delta f}{\Delta y}\right) \text{ where}$$

$$\frac{\Delta f}{\Delta x} = \frac{f(x_0+1, y_0) - f(x_0-1, y_0)}{2}$$

$$\frac{\Delta f}{\Delta y} = \frac{f(x_0, y_0+1) - f(x_0, y_0-1)}{2}$$

A Toy Example – Gradients



In this discrete toy example, we can calculate gradient at a point (x_0, y_0) as

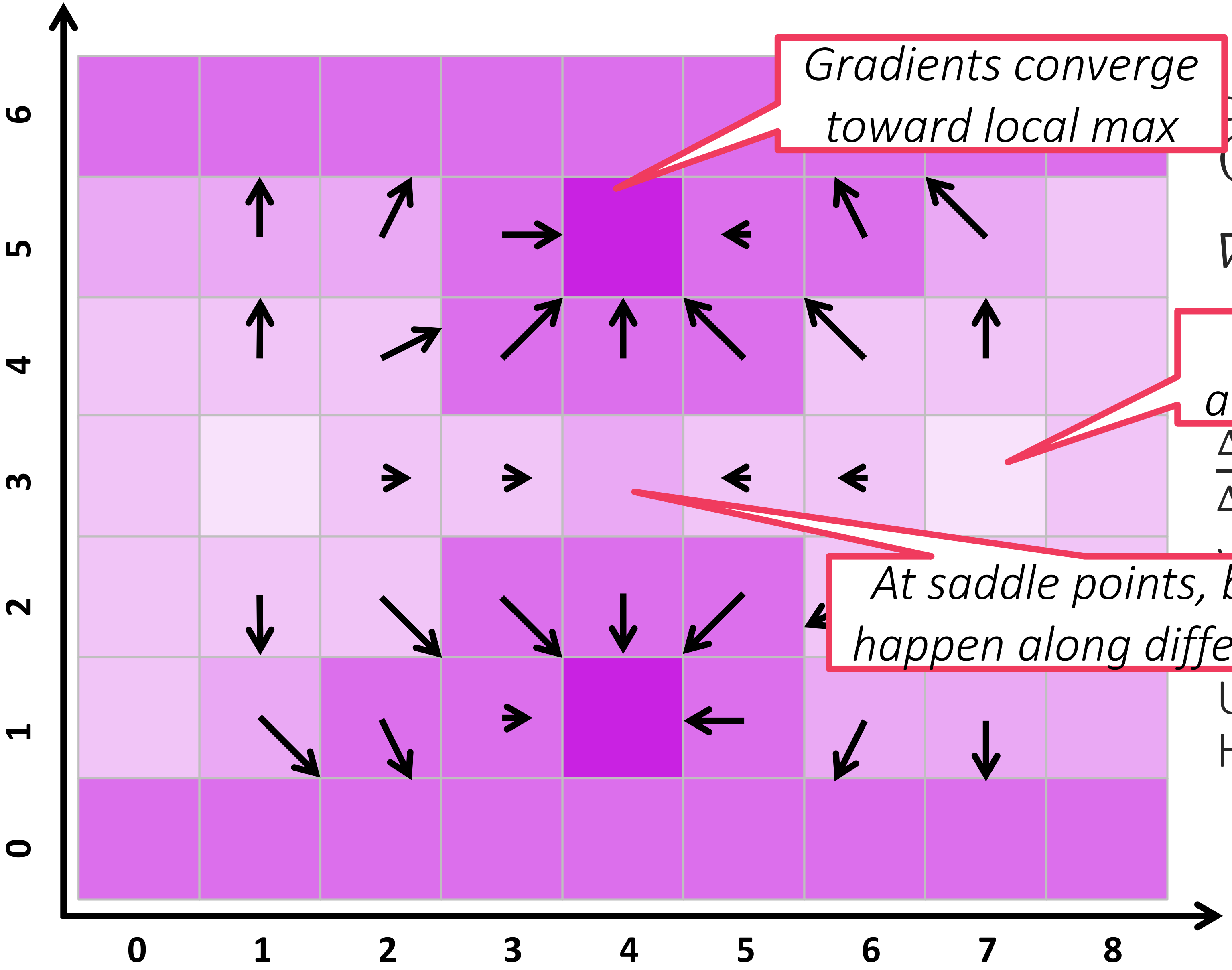
$$\nabla f(x_0, y_0) = \left(\frac{\Delta f}{\Delta x}, \frac{\Delta f}{\Delta y}\right) \text{ where}$$

$$\frac{\Delta f}{\Delta x} = \frac{f(x_0+1, y_0) - f(x_0-1, y_0)}{2}$$

$$\frac{\Delta f}{\Delta y} = \frac{f(x_0, y_0+1) - f(x_0, y_0-1)}{2}$$

We can visualize these gradients using simple arrows as well

A Toy Example – Gradients



In this discrete toy example, we can calculate gradient at a point (x_0, y_0) as

$$\nabla f(x_0, y_0) = \left(\frac{\Delta f}{\Delta x}, \frac{\Delta f}{\Delta y}\right) \text{ where}$$

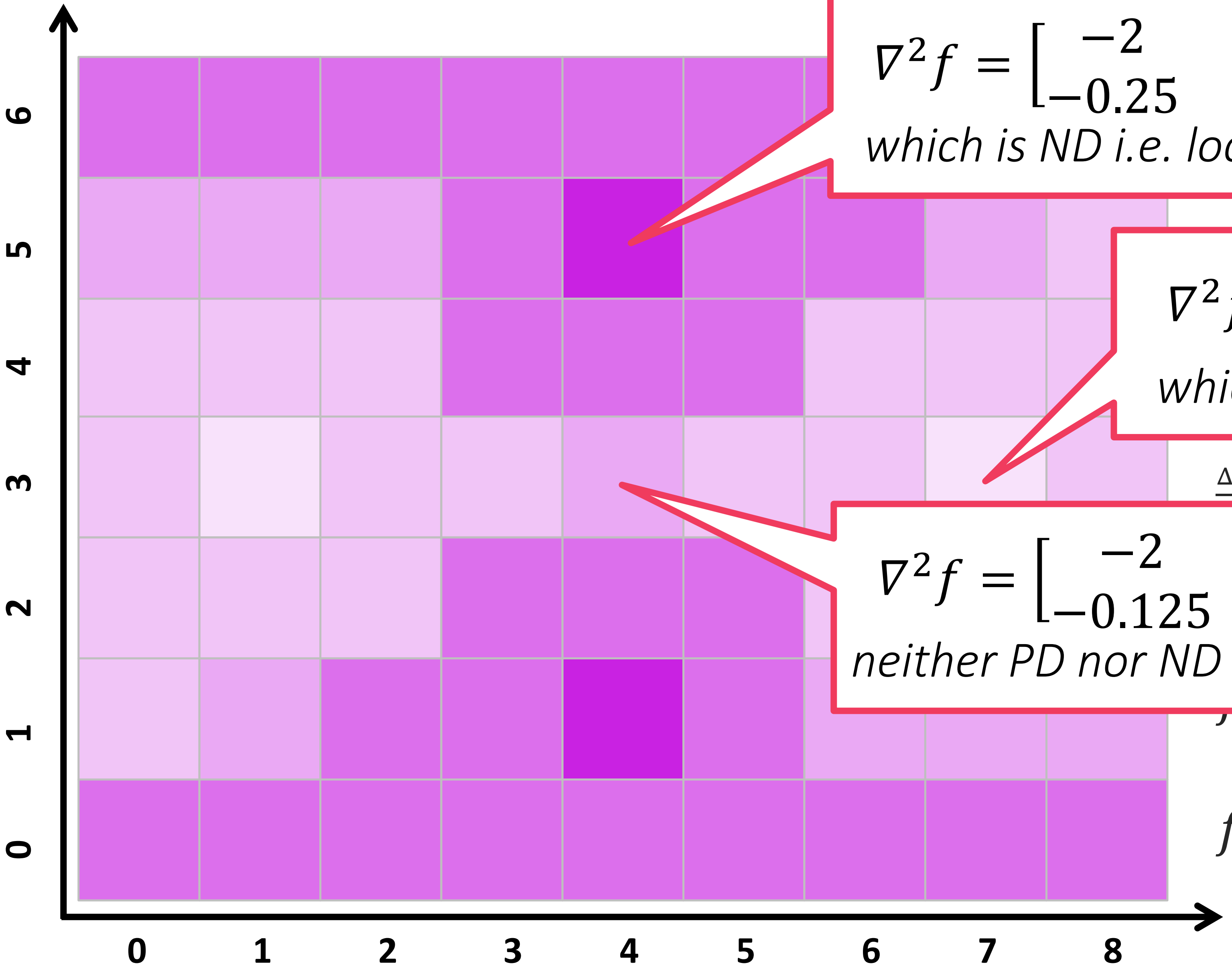
$-1, y_0)$

$$\frac{\Delta f}{\Delta y} = \frac{f(x_0, y_0+1) - f(x_0, y_0-1)}{2}$$

ualize these gradients e arrows as well

Using a similar method, the Hessian can be calculated as well!

A Toy Example – Hessians



crete toy example, we can Hessian at (x_0, y_0) as

$$\begin{bmatrix} \frac{\Delta^2 f}{\Delta x^2} & \frac{\Delta^2 f}{\Delta x \Delta y} \\ & \frac{\Delta^2 f}{\Delta y^2} \end{bmatrix}$$ where

$) - 2f(x_0, y_0)$

$\frac{\Delta^2 f}{} = f(x_0, y_0 + 1) + f(x_0, y_0 - 1) - 2f(x_0, y_0)$

$$f_{yx} = \frac{\frac{\Delta f}{\Delta y}(x_0+1, y_0) - \frac{\Delta f}{\Delta y}(x_0-1, y_0)}{2}$$

$x_0, y_0 - 1)$ ... 2

$\nabla^2 f = \begin{bmatrix} -2 & -0.25 \\ -0.25 & -2 \end{bmatrix}$

which is ND i.e. local max

$\nabla^2 f = \begin{bmatrix} 2 & 0.125 \\ 0.125 & 2 \end{bmatrix}$

which is PD i.e. local min

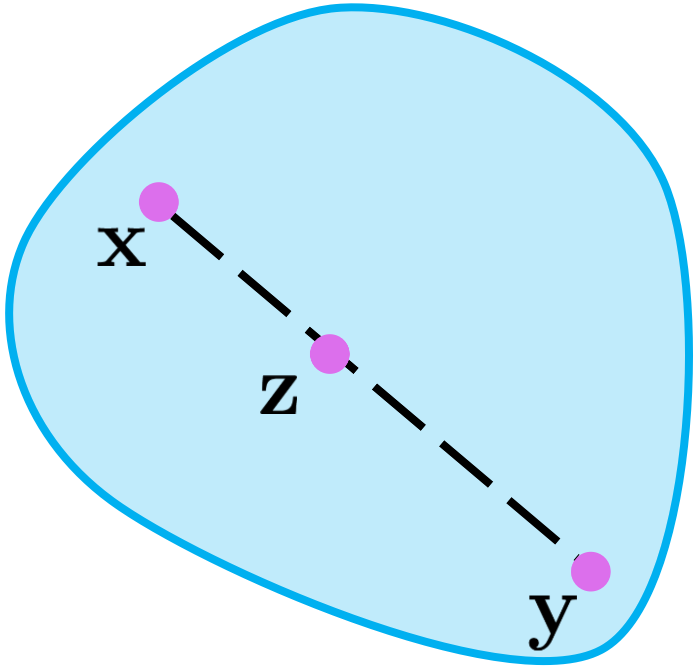
$\nabla^2 f = \begin{bmatrix} -2 & -0.125 \\ -0.125 & 2 \end{bmatrix}$ *which is neither PD nor ND (it is a saddle point)*

Convex Sets

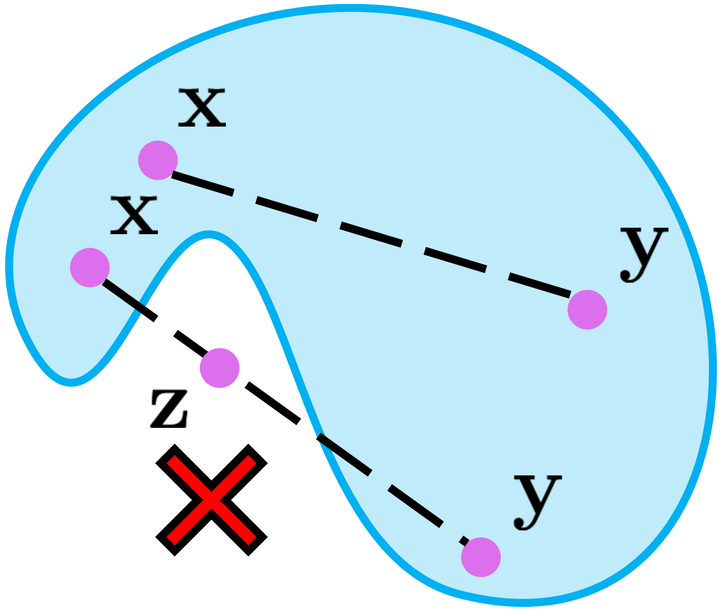
$$\mathcal{C} \subseteq \mathbb{R}^d$$

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}$$

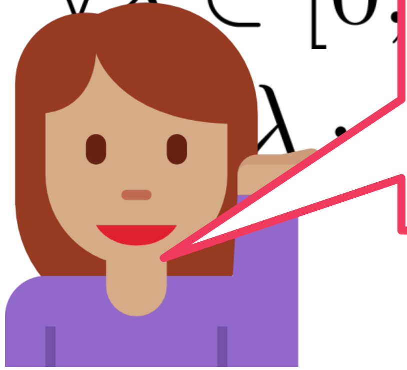
$$\forall \lambda \in [0,$$



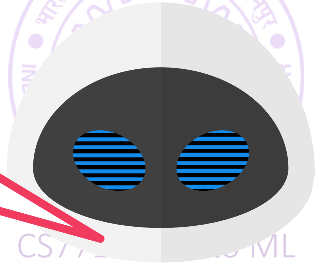
CONVEX SET



NON-CONVEX SET

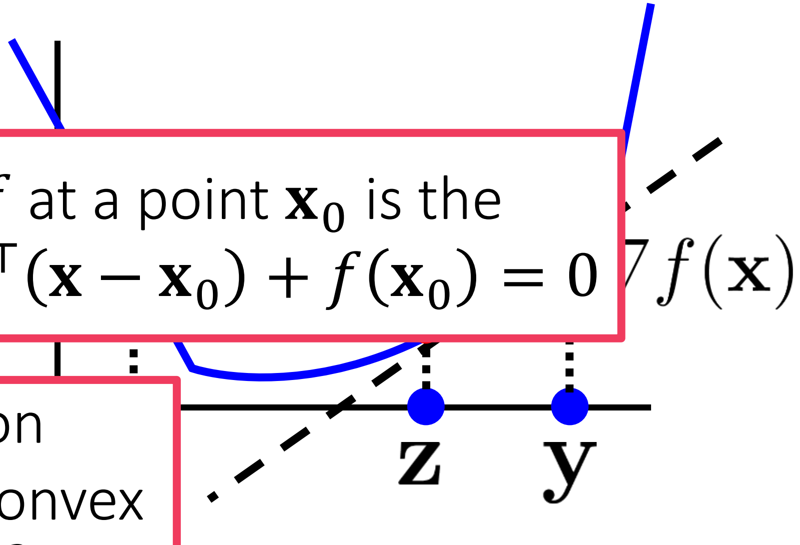


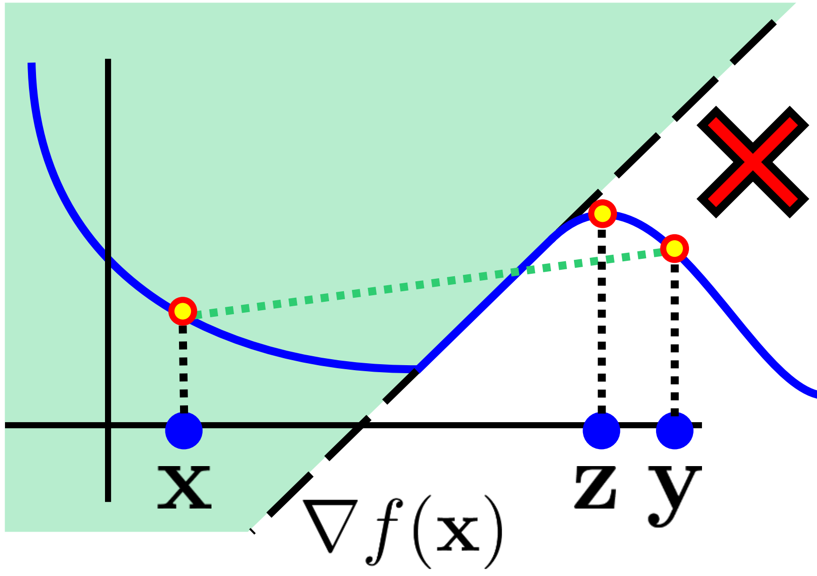
Think about which common shapes/objects are convex and which are not – balls, cuboids, stars, rectangles?



The intersection of two convex sets is always convex. The union may or may not be convex!

Convex Functions





NON-CONVEX

The tangent to f at a point $\mathbf{x}_0$ is the hyperplane $\nabla f(\mathbf{x}_0)^\top(\mathbf{x}-\mathbf{x}_0) + f(\mathbf{x}_0) = 0$

Think of common functions that are convex
1D examples: x^2
d-dim example: $\|\mathbf{x}\|_2^2$

The sum of two convex functions is always convex. The difference may or may not be convex

A differentiable convex function must lie above all its *tangents*

In fact a third definition exists for twice differentiable convex functions: their Hessian $\nabla^2 f(\mathbf{x})$ must be PSD everywhere

Checking for Convexity

All constant functions $f(\mathbf{x}) = c$ are convex

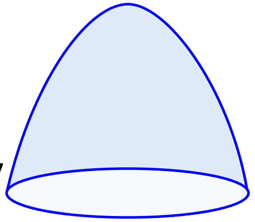
ar fu … vex

Many popular functions are concave e.g. $\log x, \sqrt{x}$. The negative of a concave function is always convex

… multiples of convex functions $c \cdot f(\mathbf{x}), c \geq 0$ are convex

$\mathbb{R}^d \to$ … convex and non-decreasing

The negative of a convex function $-f(x)$ is called a *concave* function and they look like inverted cups



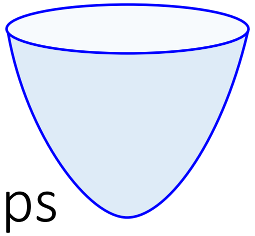
… $\mathbb{R}^d \to \mathbb{R}$ is convex

I also love concave functions since all local maxima are global maxima for a concave function

… clidean distance is convex $f(\mathbf{x}) =$ …

$\mathbb{R} \to$ … $g(\mathbf{x}) = f(\mathbf{a}^\top \mathbf{x} + b)$ is also co…

Convex functions look like cups



I love convex functions since all local minima are global minima for a convex function

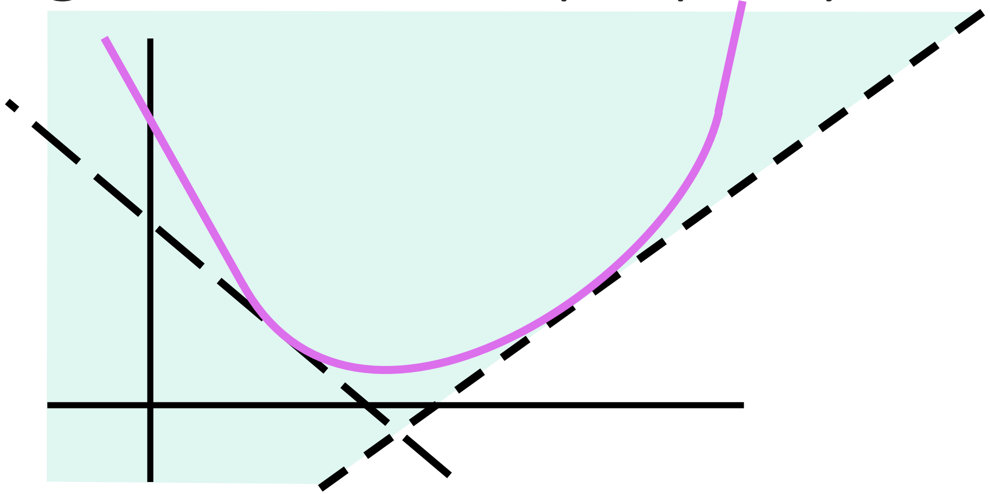
Non-differentiable Functions

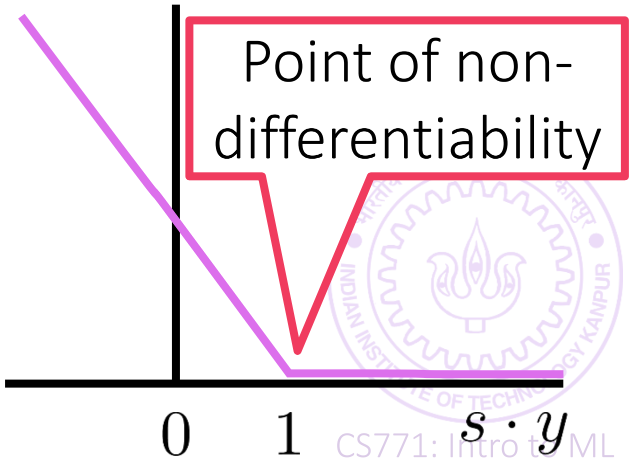
The hinge loss function is not differentiable everywhere ☹

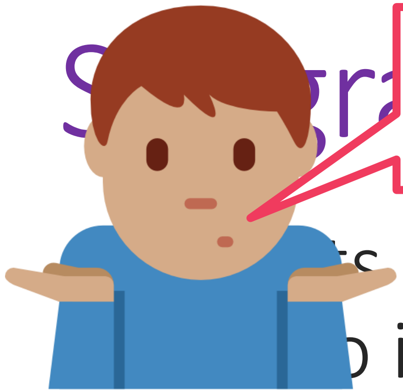
- Can we define some form of gradient for non-diff functions as well?
- Yes, if a function is convex, then no matter if it is non-differentiable, a notion of gradient called *subgradient* can always be defined for it

Recall that for differentiable functions, the gradient defines a *tangent* hyperplane at every point and the function must lie above this plane

Subgradients exploit and generalize this property ☺

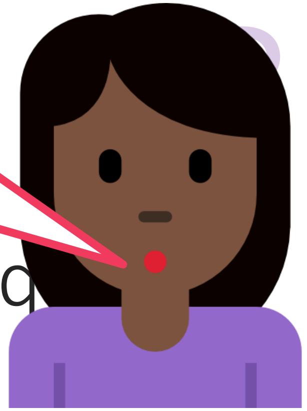






How can I find out the subgradients of a function?

S gra e sub

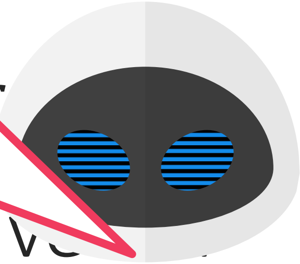


Wait! Does this mean a function can have more than one subgradient at a point $\mathbf{x}^0$

ts of a convex differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ are uniq
o its gradients

The tangent at $\mathbf{x}^0$ is the hyperp

Convex functions lie above all ta

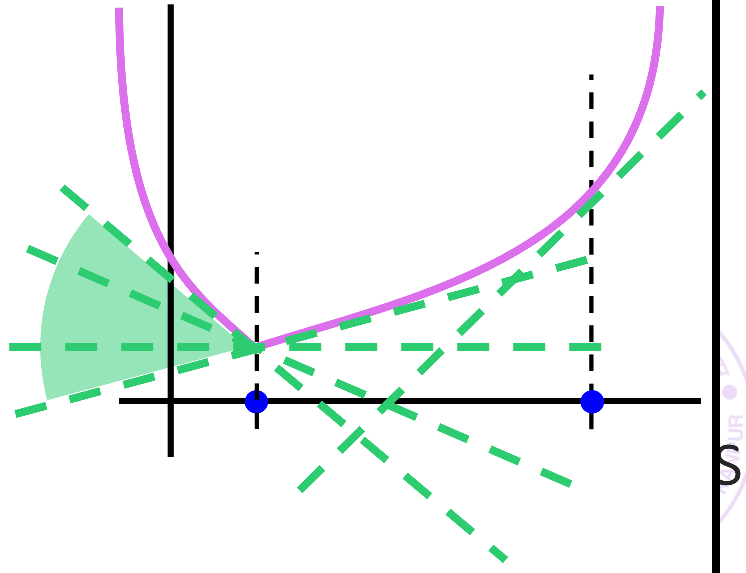


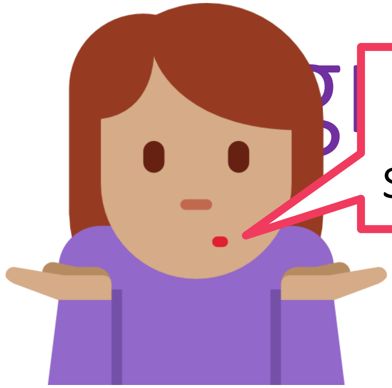
If f is non-differentiable at $\mathbf{x}^0$ then it can indeed have multiple subgradients at $\mathbf{x}^0$. However, if f is differentiable at $\mathbf{x}^0$, then it can have only one subgradient at $\mathbf{x}^0$, and that is the gradient $\nabla f(\mathbf{x}^0)$ itself ☺

Trick: turn the definition arour a ve

$\mathbf{g}$ so that the hyperplane $\mathbf{g}^\top(\mathbf{x} - \mathbf{x}^0) + f(\mathbf{x}^0) = 0$ is tangent to f at $\mathbf{x}^0$

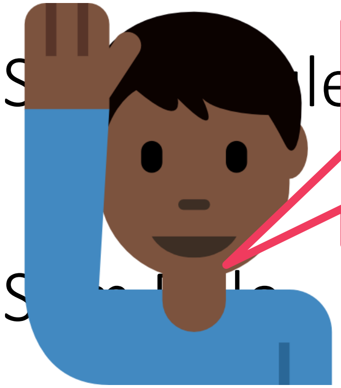
$$\partial f(\mathbf{x}^0) \triangleq \{\mathbf{g} : f(\mathbf{x}) \geq \mathbf{g}^\top(\mathbf{x} - \mathbf{x}^0) + f(\mathbf{x}^0) \;\; \forall \mathbf{x}\}$$





Gradien... ...alus

$$\mathbf{x} \in \mathbb{R}^d, \mathbf{a} \in \mathbb{R}^d, b, c \in \mathbb{R}$$



$$\partial(c \cdot f)(\mathbf{x}) = c \cdot \partial f(\mathbf{x}) = \{c \cdot \mathbf{v} : \mathbf{v} \in \partial f(\mathbf{x})\}$$

$$\nabla(f+g)(\mathbf{x}) = \nabla f(\mathbf{x}) + \nabla g(\mathbf{x}) \qquad \partial(f+g)(\mathbf{x}) = \partial f(\mathbf{x}) + \partial g(\mathbf{x}) = \{\mathbf{u} + \mathbf{v} : \mathbf{u} \in \partial f(\mathbf{x}), \mathbf{v} \in \partial g(\mathbf{x})\}$$

Chain Rule $\nabla f(\mathbf{a}^\top \mathbf{x} + b) = f'(\mathbf{a}^\top \mathbf{x} + b) \cdot \mathbf{a}$ $\quad \partial f(\mathbf{a}^\top \mathbf{x} + b) = \partial f(\mathbf{a}^\top \mathbf{x} + b) \cdot \mathbf{a} = \{c \cdot \mathbf{a} : c \in \partial f(\mathbf{a}^\top \mathbf{x} + b)\}$

Max Rule **No counterpart in general** $\quad h(\mathbf{x}) = \max\{f(\mathbf{x}), g(\mathbf{x})\}$

If $f(\mathbf{x}^0) > g(\mathbf{x}^0), \partial h(\mathbf{x}^0) = \partial f(\mathbf{x}^0)$. If $g(\mathbf{x}^0) > f(\mathbf{x}^0), \partial h(\mathbf{x}^0) = \partial g(\mathbf{x}^0)$

If $f(\mathbf{x}^0) = g(\mathbf{x}^0), \partial h(\mathbf{x}^0) = \{\lambda \mathbf{u} + (1-\lambda)\mathbf{v} : \mathbf{u} \in \partial f(\mathbf{x}^0), \mathbf{v} \in \partial g(\mathbf{x}^0), \lambda \in [0,1]\}$

Example: subgradient for hinge loss

$\ell_{\text{hinge}}(x) = \max\{1 - x, 0\} = \max\{f(x), g(x)\}$

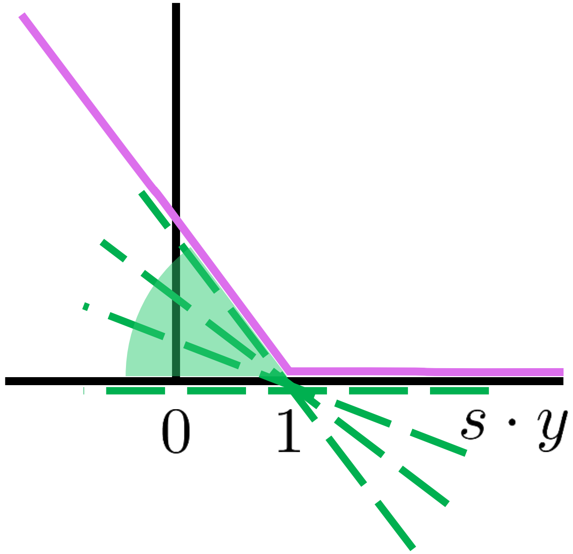
ℓ_{hinge} is differentiable at all points except $x = 1$

Thus, $\partial\ell_{\text{hinge}}(x) = \ell'_{\text{hinge}}(x)$ *if* $x \neq 1$

$) = -1$

0

$\in [0,1]\} = [-1,0]$



Applying subgradient chain rule gives us

$$\ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle) = [1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle]_+$$

Need $\mathbf{v}^i \in \partial\ell_{\text{hinge}}(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle)$

$$\mathbf{v}^i = \begin{cases} \mathbf{0} & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle > 1 \\ -y^i \cdot \mathbf{x}^i & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle < 1 \\ c \cdot y^i \cdot \mathbf{x}^i & \text{if } y^i \langle \mathbf{w}, \mathbf{x}^i \rangle = 1 \\ c \in [-1, 0] & \end{cases}$$