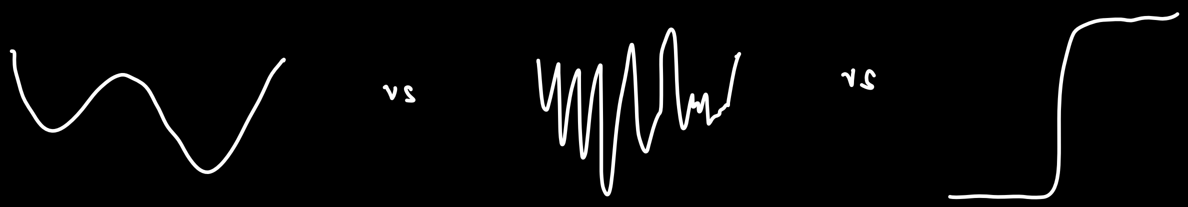
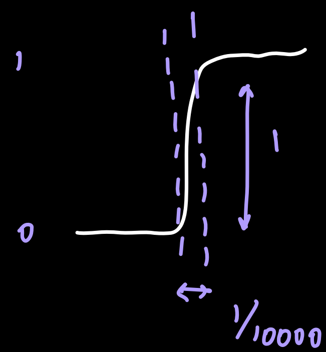
GOAL: Which functions are easier to optimize?

PROPERTY 1: How sensitive is the function?



Sharp bump is not good for GD.

A small step in x → Large change in $f(x)$.

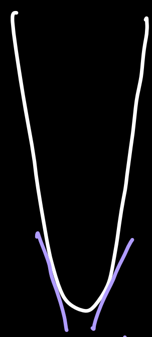


1. LIPSCHITZNESS: [Function doesn't change much for a step]

f is L-lipschitz if $(f: \mathbb{R}^d \to \mathbb{R})$

$$\forall x, y \quad |f(x) - f(y)| \leq L \underbrace{\|x - y\|_2}_{\text{L2 distance between } x \text{ and } y.}$$

What if function is:



large change in gradients.

2. SMOOTHNESS: [Gradient should also not change quickly]

f is β-smooth if

$$\forall\, x, y \quad \|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \cdot \|x - y\|_2$$

SMOOTHNESS IS STRICTER THAN LIPSCHITZNESS

if f is β-smooth $\Rightarrow$ f is L-Lipschitz.

only if the input values

are bounded!

Example :

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

$$f(x) = ax^2 + bx + c$$

$$f'(x) = 2ax + b$$

$$|f'(x) - f'(y)| = 2a|x - y|$$

$\Rightarrow$ f is $(2a)$ - smooth

$$f(x) - f(y) = ax^2 + bx + c - (ay^2 + by + c)$$

$$= a(x^2 - y^2) + b(x - y)$$

$$= (x - y)\left[a(x + y) + b\right]$$

So if $(x-y)$ is bounded, we can sort of say it is Lipschitz.

But cannot be proven explicitly for $\forall x, y$.

Theorem 1: Monotonicity of GD

f is a β-smooth function, if $\eta \leq 1/\beta$. Then,

$$f(x_{i+1}) \leq f(x_i) - \frac{\eta}{2} \|\nabla f(x_i)\|^2$$

"GD monotonically decreases the function value".

Recall : For any vector $u \in \mathbb{R}^d$

$$\|u\|_2^2 = \sum_{i=1}^{d} u_i^2$$

$$x_i = x_{i-1} - \eta \nabla f(x_{i-1})$$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \cdot \|x - y\|_2$$

Proof of Monotonicity:

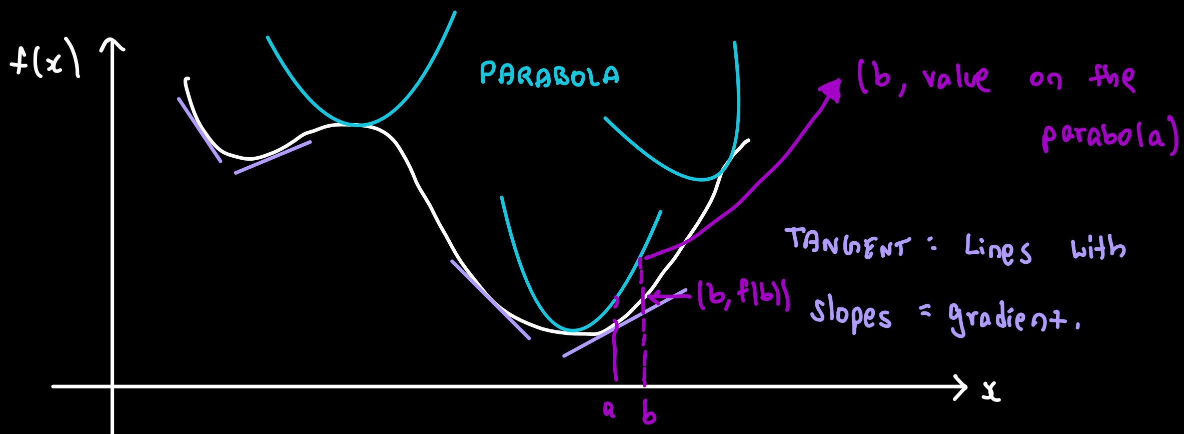
Assume univariate function $f: \mathbb{R} \to \mathbb{R}$

Smoothness upper bound: f is β-smooth $(f: \mathbb{R} \to \mathbb{R})$

$$\forall\, a, b \qquad f(b) \leq f(a) + f'(a) \cdot (b-a) + \frac{\beta}{2} (b-a)^2$$

$\Rightarrow$ So we can create a parabola as a function of β, such that parabola is above the function.

[β ensures that the parabola is above but as close as possible]



PROOF : Based on Taylor's Theorem

$$f(x+h) = f(x) + f'(x) \cdot h + f''(x) \cdot \frac{h^2}{2} + \dots$$

Taylor's theorem with a remainder term : $f(x+h) = f(x) + f'(x) \cdot h + \int_0^1 (f'(x+th) - f'(x)) \cdot th \, dt \cdot$

PROOF OF MONOTONICITY FOR UNIVARIATE CASE:

$$f(x_{i+1}) = f(x_i - \eta f'(x_i))$$

Use smoothness upper bound :

$$f(b) \leq f(a) + f'(a) \cdot (b-a) + \frac{\beta}{2} (b-a)^2$$

$$b = x_i - \eta f'(x_i)$$

$$a = x_i$$

$$f(x_{i+1}) = f(x_i - \eta f'(x_i)) \leq f(x_i) + f'(x_i) \cdot (-\eta f'(x_i)) + \frac{\beta}{2}(-\eta f'(x_i))^2$$

$$= f(x_i) - \eta f'(x_i)^2 + \frac{\beta}{2} \eta^2 f'(x_i)^2$$

$$= f(x_i) - \eta \left(1 - \frac{\eta \beta}{2}\right) f'(x_i)^2$$

$\eta \leq 1/\beta$

$$= f(x_i) - \frac{\eta}{2} f'(x_i)^2$$

Smoothness upper bound for multivariate functions:

$$f : \mathbb{R}^d \to \mathbb{R}$$

If f is β-smooth, then

$$\forall x, y \quad f(y) \leq f(x) + \underbrace{\langle \nabla f(x), y - x \rangle}_{\text{(inner-product)}} + \frac{\beta}{2} \|y - x\|_2^2$$

PROOF OF MONOTONICITY FOR ALL FUNCTIONS:

$$f(x_{i+1}) = f(\underbrace{\underset{x}{x_i} - \eta \nabla f(x_i)}_{y})$$

$$\leq f(x_i) + \langle \nabla f(x_i), -\eta \nabla f(x_i) \rangle + \frac{\beta}{2} \| (-\eta \nabla f(x_i)) \|_2^2$$

$$= f(x_i) - \eta \| \nabla f(x_i) \|_2^2 + \frac{\eta^2 \beta}{2} \| \nabla f(x_i) \|_2^2$$

$$= f(x_i) - \eta \left(1 - \frac{\eta \beta}{2}\right) \| \nabla f(x_i) \|_2^2$$

$$\leq f(x_i) - \frac{\eta}{2} \| \nabla f(x_i) \|_2^2 \qquad \eta \leq 1/\beta$$

Summary:

→ GD makes progress as long as $\eta \leq 1/\beta$

(Theory to practice):

$\boxed{\text{Practical tricks:}}$

1. Find largest η such that

$$f(x_i - \eta \nabla f(x_i)) \leq f(x_i) - \frac{\eta}{2} \| \nabla f(x_i) \|^2 \quad (*)$$

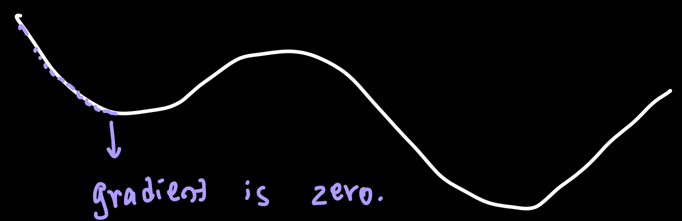
(eg: start with $\eta = 1$

if $(*)$ holds, continue. else try $\eta = 1/2, \ldots$

2. Can also do "Backtracking line search"

to pick right η.

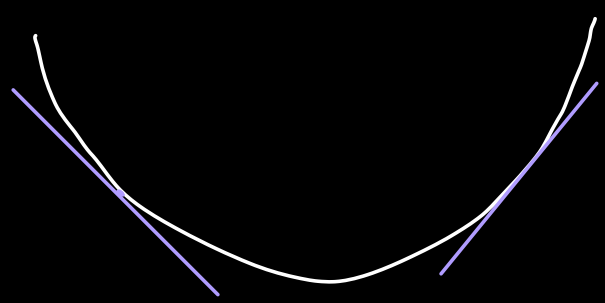
→ Monotonicity $\not\Rightarrow$ we converge to the global minimum.

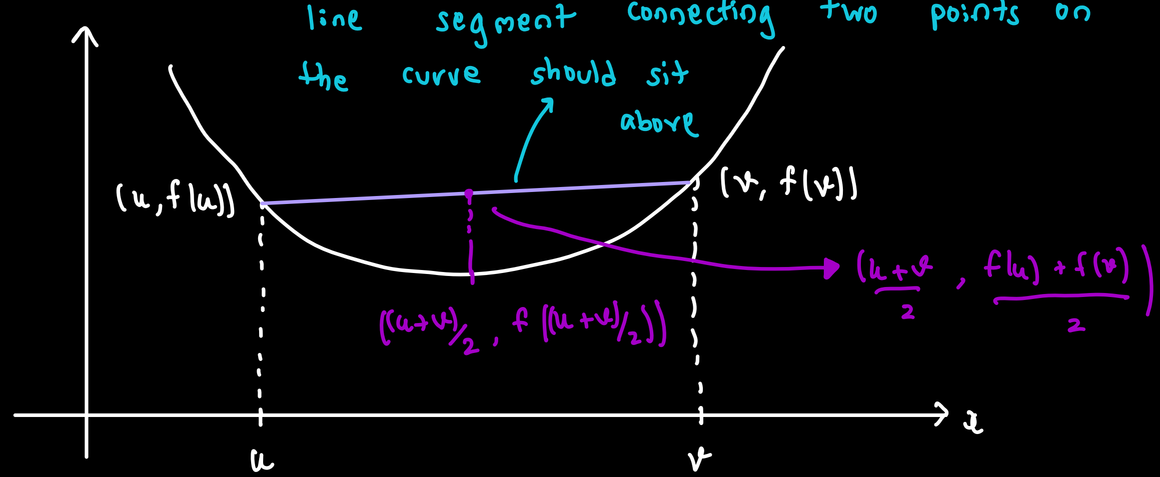


CONVEX FUNCTIONS

[Magic Ingredient in optimization]

Convex: $f : \mathbb{R}^d \to \mathbb{R}$ is convex if the tangent plane at any point is below the curve.





Equivalently :

$f : \mathbb{R}^d \to \mathbb{R}$ is convex if

$\to \forall u, v \quad f\left(\frac{u+v}{2}\right) \leq \frac{f(u) + f(v)}{2}$

$\to \forall u, v, \lambda \in [0,1) \quad f(\lambda u + (1-\lambda)v) \leq \lambda . f(u) + (1-\lambda) f(v)$

$\to \forall u, v, \; \underbrace{f(u) + \langle \nabla f(u), v-u \rangle}_{\text{the tangent function}} \leq \underset{\text{function}}{\underset{\downarrow}{f(v)}} \quad (\bigstar)$

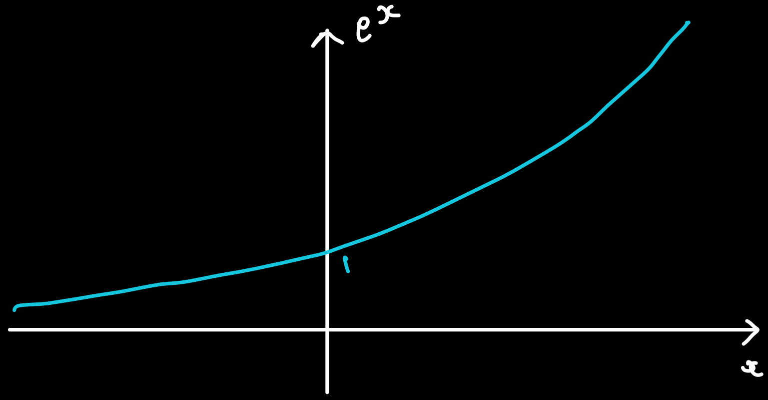
* f, g are convex $\Rightarrow$ $f + g$ is convex.
* f is convex $\Rightarrow$ $a . f$ is convex for $a > 0$.
* $g : \mathbb{R} \to \mathbb{R}$, $w \in \mathbb{R}^d$

  $g_w : \mathbb{R}^d \to \mathbb{R}$

  $g_w(x) = g(\langle w, x \rangle)$

  g is convex $\Rightarrow$ g_w is convex.

Example: e^x is a convex function.



$\Rightarrow \forall w \quad g_w : \mathbb{R}^d \to \mathbb{R}$ as

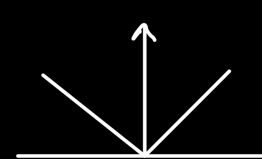
$g_w(x) = e^{\langle w, x \rangle}$ is convex.

$\Rightarrow x^2$ is a convex function $\Rightarrow f(x) = \langle w, x \rangle^2$ is a convex function.

$\Rightarrow (x-a)^2$ is a convex function

$\Rightarrow f(x) = (\langle w, x \rangle - a)^2$ is a convex function.

$\Rightarrow |x|$ is a convex function



Why Convexity :

ERM : Imagine we have parameter space Θ

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(h_\theta(x_i), y_i)$$

Dataset $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$

$\rightarrow$ If $\ell(h_\theta(x_i), y_i)$ is convex in θ, then L is convex.

Least Squares Regression :

$$h_\theta(x_i) = \underbrace{\langle \theta, x_i \rangle}_{\text{inner-product}}$$

$$\ell(h_\theta(x_i), y_i) = (\langle \theta, x_i \rangle - y_i)^2$$

LSR ERM : $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} (\langle \theta, x_i \rangle - y_i)^2$

is a convex function in θ.

L_1 ERM : $L_1(\theta) = \frac{1}{n}\sum_{i=1}^{n} |\langle \theta, x_i \rangle - y_i|$

is a convex function in θ.

"LASSO" : $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} (\langle \theta, x_i \rangle - y_i)^2 + \lambda(|\theta_1| + |\theta_2| + \dots + |\theta_n|)$

is a convex function in θ.

→ Linear programming

→ Semi-definite programming

} → Minimizing a convex function.

CONVEX OPTIMIZATION IS EVERYWHERE!

THEOREM: If f is β-Smooth and convex, then

(if $\eta \le 1/\beta$) $$f(x_k) \le f(x_*) + \frac{2\beta \cdot \| x_0 - x_*)}{k}$$

$f(x_*)$ ↓ global optimum

k ↳ number of iterations

(Remark: Minimizing a convex function is "easy")

for a given accuracy and we know β, we know the number of iterations needed, to reach within that accuracy of the global optimum.

(Remark: If f is L-Lipschitz, then

$$f(x_k) \le f(x_*) + \frac{L \cdot \|x_0 - x_*\|}{\sqrt{k}}.$$