CS260B: Algorithmic Machine Learning

→ What is Learning?

Examples ... → Reinforcement Learning
→ Memorization
→ Classification (image/document/fraud detection)
→ Image Generation
→ Regression
→ Advertising
→ Recommendation Systems (Movies, Friends, Tweets)

Some Categories:
→ Supervised Learning
→ Unsupervised Learning
→ Semi-supervised Learning
→ Active Learning
→ Reinforcement Learning

How to model Supervised Learning?

INPUT: Some domain $\mathcal{X}$

images, documents

browsing profiles

credit card history

Labels: CATS / DOGS / FOXES $\mathcal{L}$

$\{0, 1\}$

$\mathbb{R}$

DATASET: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$x \in \mathcal{X}$, $y \in \mathcal{L}$

Goal: Given a new $x \in \mathcal{X}$, want to predict its label!

Assumption 1: There is an underlying ground truth distribution on training & test examples.

D on $\mathcal{X}$: $x_1, \ldots, x_n$ are i.i.d samples from D.

x is also from D.

Assumption 2: Labels cannot be arbitrary.

There is a class of functions $\mathcal{H}: X \to \mathcal{L}$

Assumptions: Prediction cannot be accurate all the time.

$$\to \Pr_{x \sim D}\left[Y \neq f(x) \right] \leq \varepsilon$$

$f(x)$: my prediction; Y: "error"

We need to relax "$\neq$" depending on application.

Assumption 4: Cannot expect to succeed always!

→ "Probability of failure" δ.

PAC Model (Valiant, Vapnik, ...)

An algorithm (ε, δ)-PAC learns a hypothesis class $\mathcal{H}$ with sample complexity $n(\varepsilon, \delta, \mathcal{H})$ if

INPUT: $(x_1, f^*(x_1)), \ldots, (x_n, f^*(x_n))$ where $x_i \leftarrow D$, $f^* \in \mathcal{H}$.

OUTPUT: Some predictor h with probability $1-\delta$ over $x_1, \ldots, x_n$

$$\Pr_{x \leftarrow D}\left[h(x) = f^*(x) \right] \geq 1 - \varepsilon$$

PAC is a very strong requirement!

→ Often becomes intractable.

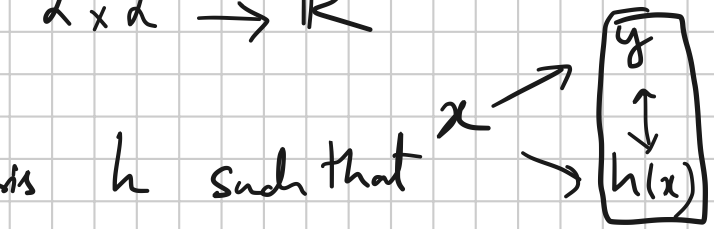
Learning as Optimization

→ We have an underlying $\mathcal{D}$ on X. Labels $\mathcal{L}$.

→ Dataset: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

→ Loss function $\ell : \mathcal{L} \times \mathcal{L} \to \mathbb{R}$

Goal: Find a hypothesis h such that



$$\sum_{i=1}^{n} \ell\left(h(x_i), y_i\right) \quad \text{is small}$$

Parametrized hypothesis class: $\mathcal{H}$

$$\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_\theta(x_i), y_i\right)$$

$h_\theta(x_i)$ is the predictor as specified by the "parameters" θ.

"Empirical Risk Minimization"

(ERM)

We are using loss on training data

(Ideal: $\mathbb{E}_{x \sim D}\left[\ell\left[h_\theta(x), y\right]\right]$).

Example 1 (ERM): $\mathcal{X} \equiv \mathbb{R}^d$; $\mathcal{A} \equiv \mathbb{R}$

Parametric family: Linear predictors

$\left(\Theta \equiv \mathbb{R}^d\right)$

$$h_\theta(x) = \sum_{i=1}^{d} \theta_i x_i$$

$$= \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d .$$

(More generally, $h_{\theta,t}(x) = \theta_1 x_1 + \cdots + \theta_d x_d + t$).

Least Squares Regression: $h_\theta(x) = \langle \theta, x \rangle$

$\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ $\quad \ell(a,b) = (a-b)^2$.

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

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\langle \theta, x_i \rangle - y_i\right)^2 .$$

ERM: Find θ to minimize $L(\theta)$.

Ex 2: $\ell(a,b) = |a-b|$.

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

Ex 3: What if labels are discrete? $\mathcal{Y} \equiv \{0,1\}$

Parametric family $h_\theta(x) = \begin{cases} 1 & \text{if } \langle \theta, x \rangle > 0 \\ 0 & \text{if } \langle \theta, x \rangle < 0 \end{cases}$

$\ell(a,b) = \begin{cases} 1 & \text{if } a \neq b \\ 0 & \text{else} \end{cases}$

↓

Linear Threshold functions

Halfspaces

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left(\underbrace{\text{sign}(\langle \theta, x_i \rangle)}_{\downarrow \text{ sign matches } y_i} \neq y_i\right)$$

$\begin{cases} 0 & \text{if} \\ 1 & \text{else.} \end{cases}$

LECTURE 2 03/30/22

Summary from last class

→ PAC Model

→ Learning as optimization: ERM

Today

→ Examples of ERM

→ How to solve ERM?

→ GD: A meta algorithm...

Last class:

ERM: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

→ Parametrized family of predictors $\mathcal{H}$

→ Loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$$\arg\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_\theta(x_i), y_i\right)$$

(Notation: The minimizer θ (argument)).

Examples:

Predictors are linear functions

→ $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

→ $h_\theta(x) = \sum_{j=1}^{d} \theta_j x_j \equiv \langle \theta, x \rangle$

(Notation for inner-product between vectors).

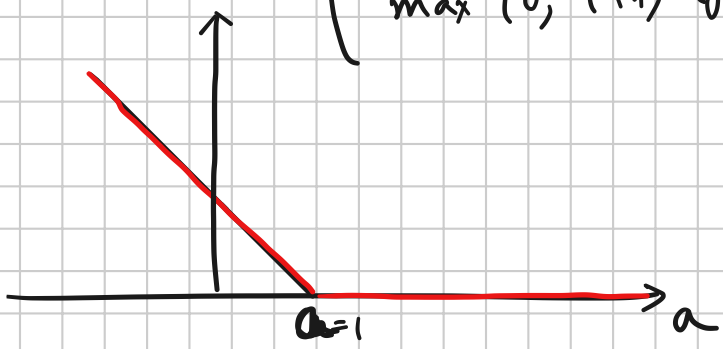
Examples: $\ell(a,b) = (a-b)^2 \rightarrow$ Least Squares Regression

$\ell(a,b) = |a-b| \rightarrow$ "L_1 - Regression"

$\ell(a,b) \equiv$ logistic loss $\rightarrow$ Logistic Regression

$\ell(a,b) \equiv$ Hinge-Loss

$$\ell(a,b) = \begin{cases} \max(0, 1-a) & \text{if } b>0 \\ \max(0, 1+a) & \text{if } b<0 \end{cases}$$



Recall LSR:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\langle \theta, x_i \rangle - y_i \right)^2 .$$

For L_1- Loss:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left| \langle \theta, x_i \rangle - y_i \right|$$

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

Goal: Find $\min_{\theta} f(\theta)$

$\theta \in \mathbb{R}^d$

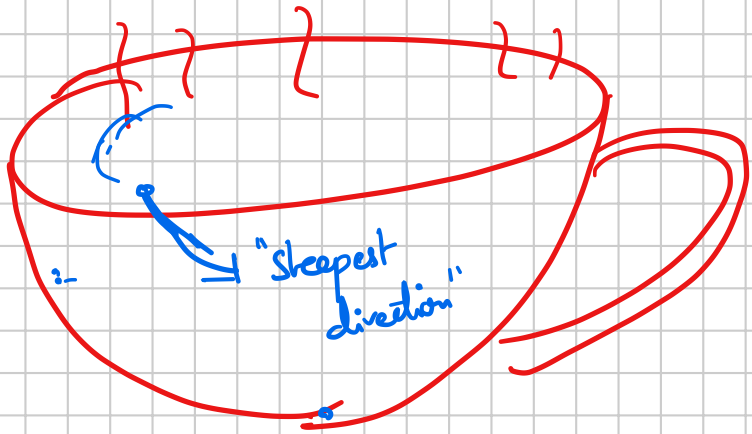
Example:

Idea: → Pick a point

→ "Move in a direction that reduces function value."

→ Pick a point to start

→ Move in the direction of steepest descent.



Claim: Steepest descent direction is $-\nabla f(x)$.

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d} \right)$$

$\nabla f(x) \equiv \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right)$.

GRADIENT DESCENT ALGORITHM (GD)

1. X_0
2. For $i = 1, \ldots, T$:

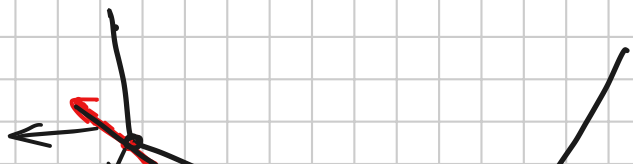
$$X_i = X_{i-1} - \eta \nabla f(x_{i-1})$$

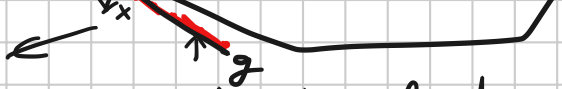
η: "Step-Size".

To Notebook Demo.

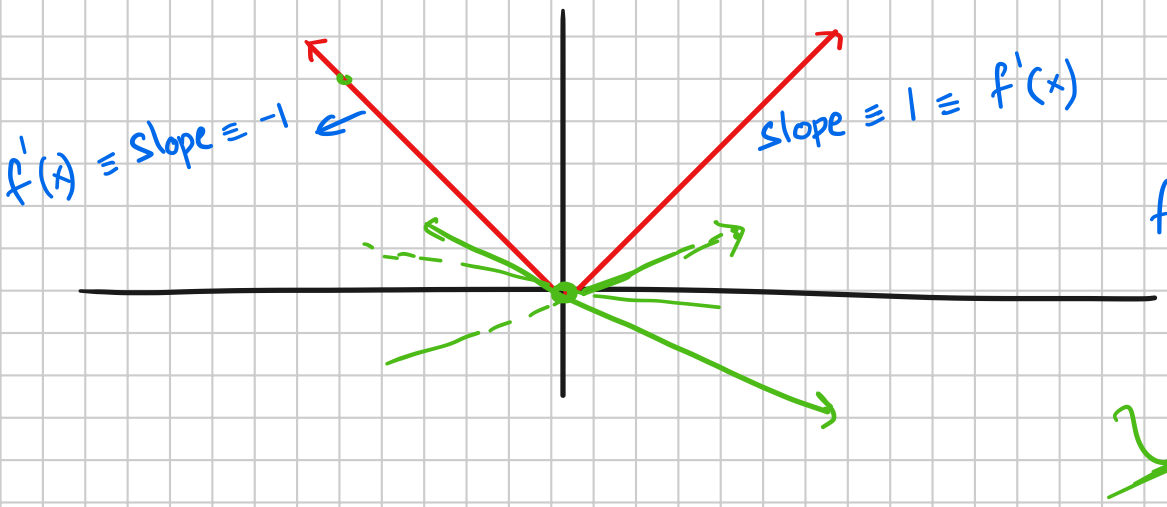
Dealing with Non-Differentiable functions.

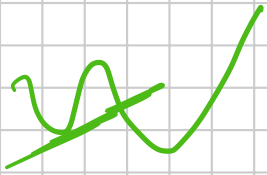
Sub-gradient: A direction g is a sub-gradient of f at x_0 if $\forall x \quad f(x_0) + \langle g, x - x_0 \rangle \leq f(x)$.





We want a line such that the function is "above" the line.





Sumarize: We can use sub-gradients as substitute for gradient.

$f(x) = |x_1| + |x_2| + \cdots + |x_n|$.

$g \equiv (\text{sign}(x_1), \text{sign}(x_2), \ldots, \text{sign}(x_n))$ is a sub-gradient for f.

Subgradient Descent

- x_0
- For $i = 1 \ldots T$:

$$x_i = x_{i-1} - \eta g$$

g: a "sub-gradient" of f at x_{i-1}.

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

- x_0
- For $i = 1, \ldots, T$:

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

What would we like to say/know about GD on f?

1. Does GD get me to the minimum?
2. How many steps would GD to take to get to minimum?
3. How to pick the starting point?
4. How to choose the step-size?
5. When do I stop?
6. How do I compute ∇f?

LECTURE 3: 01/04/22: What does GD do

Last class
- → Defined GD
- → Sawit in action

Today
- → What can we say about GD?
- → Convexity.

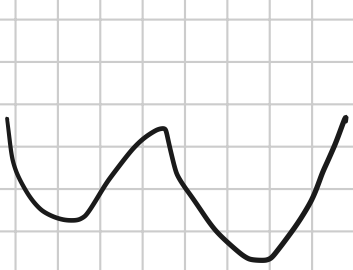
ANNOUNCEMENTS
- → HW1 out 04/06 (wednesday)
- → Special instructor: 04/11/22 (HADLEY BLACK)

Recall: Our goal: $\min f(x)$ $\quad f: \mathbb{R}^d \to \mathbb{R}$.

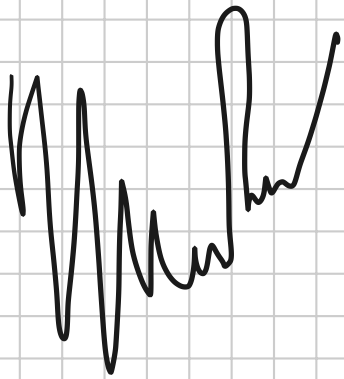
GD:
→ For $i = 1, \ldots, T$:

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

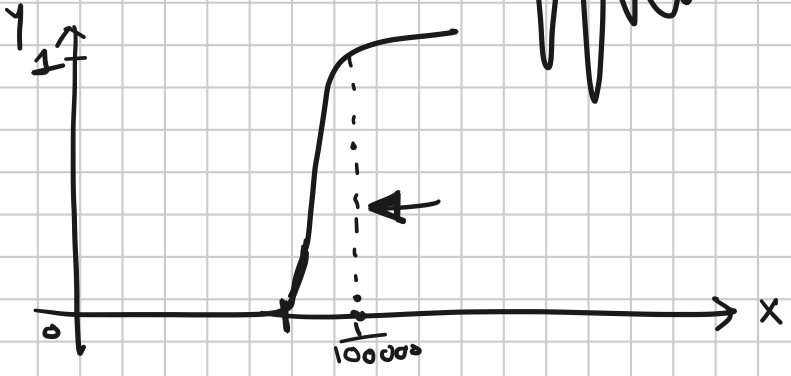
Property 1: How sensitive is the function?



vs



vs



1. Lipschitzness: f is L-Lipschitz if $(f: \mathbb{R}^d \to \mathbb{R})$

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

$\|x-y\|_2$ — ℓ_2-distance between x and y

2. Smoothness: f is β-smooth if

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

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

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

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

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

$\Rightarrow$ f is $(2a)$-Smooth.

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

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

$$= (x-y) \cdot (a(x+y) + b) .$$

(Step-size in GD)

Thm 1 (Monotonicity of GD):

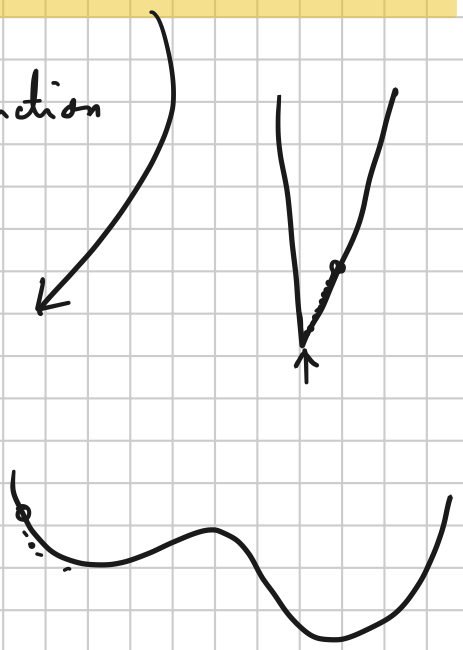
f is a β-smooth function, if $\boxed{\eta \leq \frac{1}{\beta}}$. Then,

$$f(x_{i+1}) \leq f(x_i) - \frac{\eta}{2} \|\nabla f(x_i)\|_2^2$$

"GD monotonically decreases the function value."

(Recall: for any vector $u \in \mathbb{R}^d$,

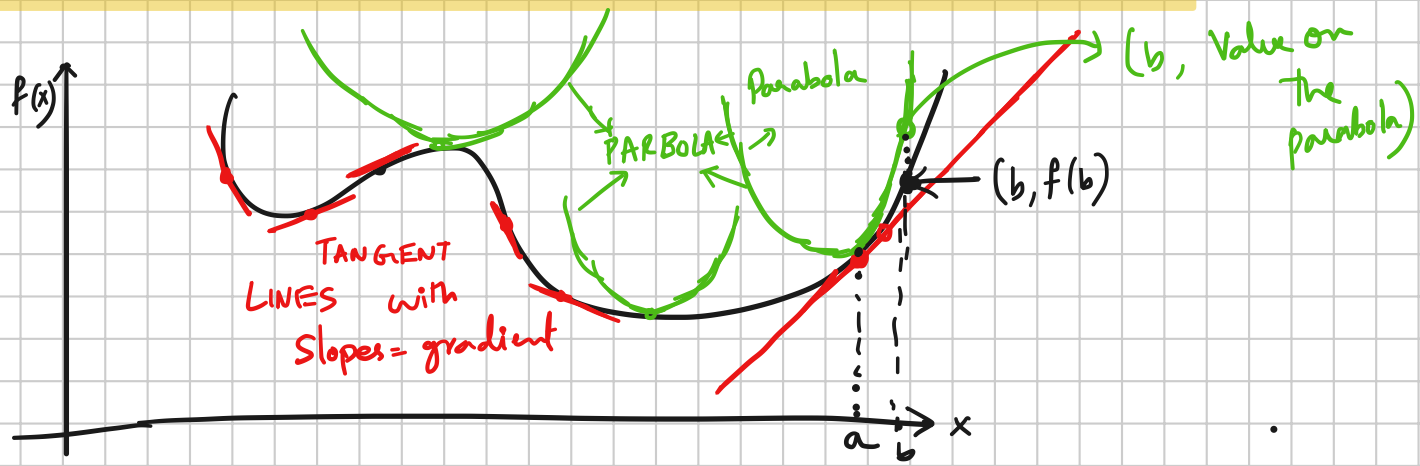
$$\|u\|_2^2 = \sum_{i=1}^{d} u_i^2 \Big).$$



Proof of Monotonicity: $f: \mathbb{R} \to \mathbb{R}$

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

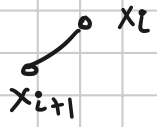
$$\forall a, b \quad f(b) \leq f(a) + f'(a)\cdot(b-a) + \frac{\beta}{2}(b-a)^2. \quad \rightarrow$$



Proof: (Taylor's theorem): $f(x+h) = f(x) + f'(x)\cdot h + f''(x)\cdot h^2$

(Taylor's theorem with a remainder term) $f(x+h) = f(x) + f'(x)\cdot h + \int_0^1 \left(f'(x+th) - f'(x)\right) t h \, dt.$

Proof of Monotonicity for Univariate Case:



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

Use Smoothness upper bound (x_i → a)

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

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

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

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

$\eta \leq \frac{1}{\beta}$.

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

Smoothness upper bound for multi-variate functions: $f: \mathbb{R}^d \to \mathbb{R}$

If f is β-smooth, then

$\forall x, y \quad f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\beta}{2}\cdot\|y-x\|_2^2.$

↓ (inner-product)

Proof of Monotonicity for all functions:

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

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

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

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

$$\leq f(x_i) - \frac{\eta}{2} \cdot \|\nabla f(x_i)\|_2^2 \qquad (\text{as } \eta \leq \tfrac{1}{\beta}).$$

(Claim: $\langle u, u \rangle = \|u\|_2^2$.)

→ GD makes progress as long as $\eta \leq \frac{1}{\beta}$.

(Theory to practice):

Practical tricks:

① Find largest η such that

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

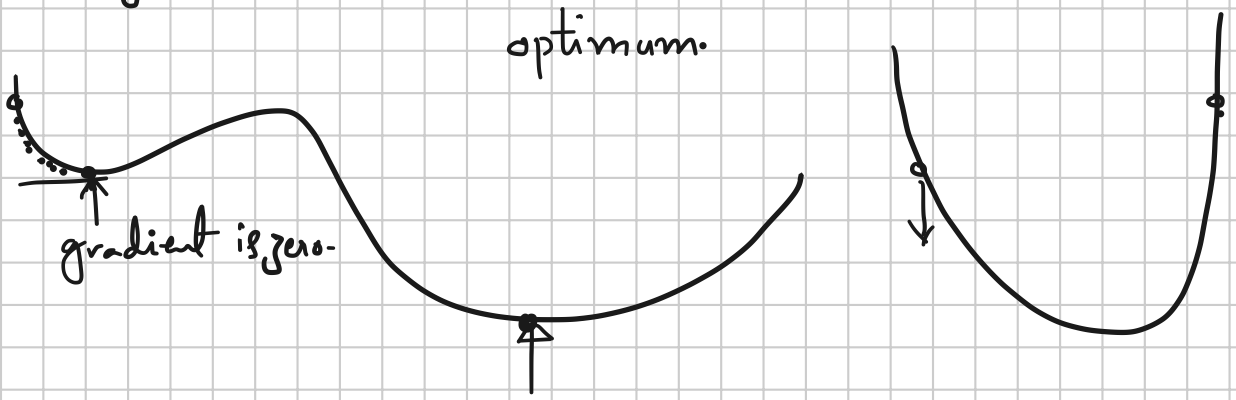
(e.g.: start with $\eta = 1$.

If (*) holds, continue. Else try $\eta = \frac{1}{2}, \ldots$

② Can also do "Backtracking line search"

to pick right η.

→ Monotonicity $\not\Rightarrow$ We converge to the global optimum.

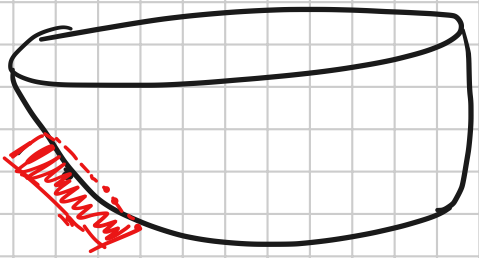


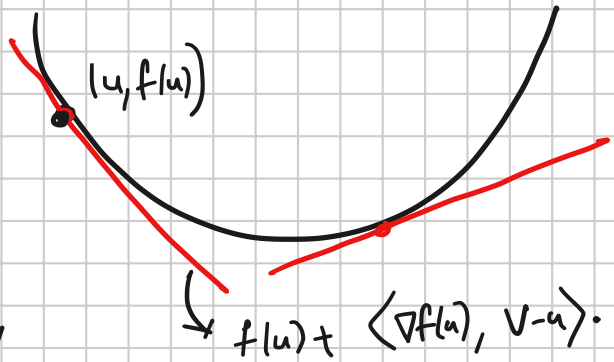
(CONVEX FUNCTIONS)

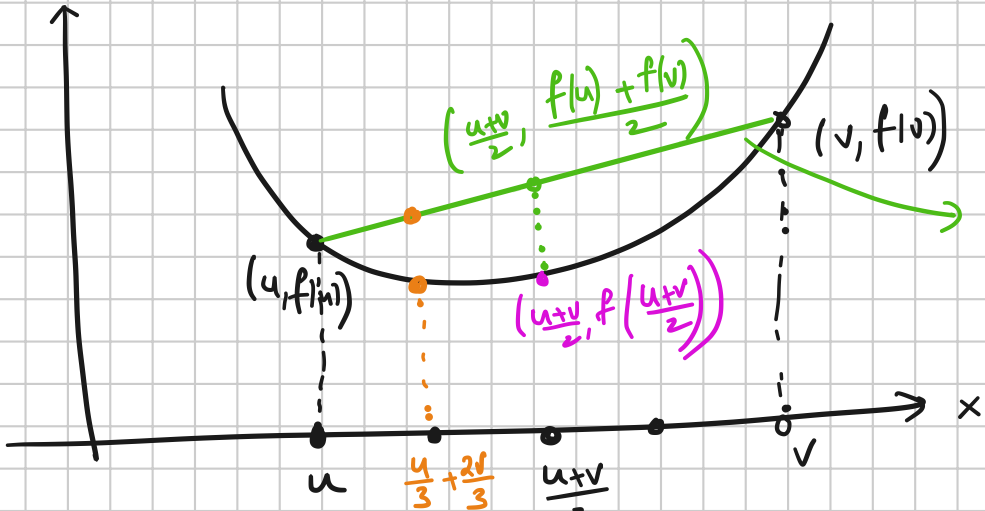
→ (MAGIC INGREDIENT IN OPTIMIZATION)

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

→







Line segment connecting two points on the curve should sit above.

Equivalently:

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

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

$\to \forall\ u, v,\ \lambda \in (0,1) \qquad f(\lambda u + (1-\lambda)v) \leq \lambda \cdot f(u) + (1-\lambda) f(v)$

$\to \forall\ u, v, \qquad f(u) + \langle \nabla f(u), v-u \rangle \leq f(v) \longrightarrow (\star)$

$\langle \nabla f(u), v-u \rangle$: the tangent function

$f(v)$: function.

---

① f, g are Convex $\Rightarrow$ $f+g$ is Convex.

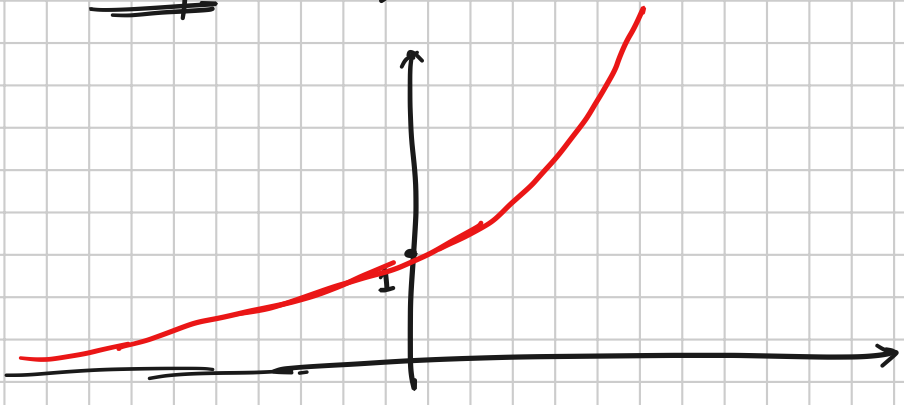
② f is convex $\Rightarrow$ $a \cdot f$ is Convex for $a > 0$.

③ $g: \mathbb{R} \to \mathbb{R}$, $w \in \mathbb{R}^d$

$g_w: \mathbb{R}^d \to \mathbb{R} \qquad g_w(x) = g(\langle w, x \rangle).$

g is convex $\Rightarrow$ g_w is convex.

Example: $\to e^x$ is a convex function.



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

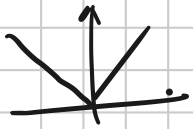
$g_w(x) = e^{\langle w, x \rangle}$ is Convex.

→ x^2 is a Convex function. ⇒ $f(x) = \langle w, x \rangle^2$ is a Convex function.

→ $(x-a)^2$ is a Convex function

⇒ $f(x) = (\langle w, x \rangle - a)^2$ is a Convex Function ...

→ $|x|$ 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 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

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

→ Least Squares Regression: $h_\theta(x_i) = \langle \theta, x_i \rangle$ (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 : $L_1(\theta) = \frac{1}{n} \sum_{i=1}^{n} |\langle \theta, x_i \rangle - y_i|$

→ "LASSO":

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

→ Linear Programming
→ Semi-Definite Programming
} → Minimizing a Convex function!

CONVEX OPTIMIZATION IS EVERYWHERE

Thm: If f is β-Smooth and Convex, then (if $\eta \leq \frac{1}{\beta}$)

$$f(x_k) \leq f(x_*) + \frac{\beta \cdot \|x_0 - x_*\|^2}{2k} \to (\text{Fix}).$$

($f(x_*)$: global optimum)

(Remark: ① Minimizing a Convex function is "Easy".

② If f is L-Lipschitz, then

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

04/06 LECTURE 4: GD Analysis, Variants

Last class
- → GD makes progress for smooth functions
- → Convex functions
- → Why Convexity

Today
- → GD solves convex optimization
- → Can you do better than GD?
- → NAGD, SGD
- → Demo.

Thm: f β-Smooth and Convex $\Rightarrow$ GD works for $\eta \le \frac{1}{\beta}$.

$$f(x_k) \le f(x_*) + \frac{2\beta\|x_* - x_0\|}{k} \quad (\text{if } \eta = \tfrac{1}{\beta}).$$

Proof:

Recall:

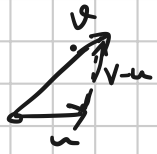
$$\forall\ x, y \quad f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle$$

Some properties of vectors

(a) $\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2\langle u, v \rangle$.

(a)' $2\langle u, v \rangle = \|u\|^2 + \|v\|^2 - \|u - v\|^2$. — (★★)

(b) $|\langle u, v \rangle| \le \|u\| \cdot \|v\|$ (Cauchy–Schwarz).



$\eta = \frac{1}{\beta}$

$$f(x_{k+1}) \le f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2 \quad -(1)$$

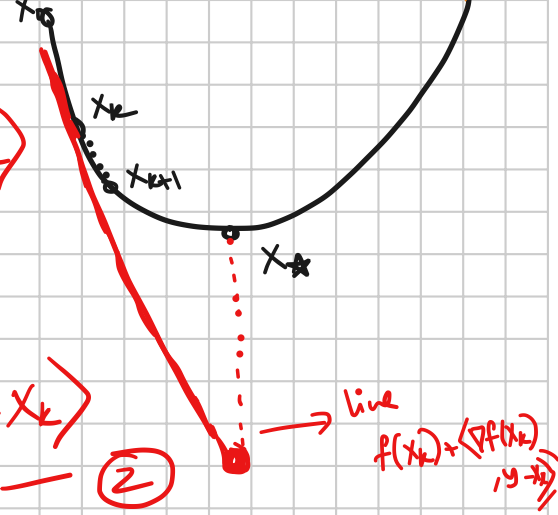
(From last lecture ...)

$$f(x_*) \geq f(x_k) + \langle \nabla f(x_k), x_* - x_k \rangle$$

(Because of Convexity!)

$$\Downarrow (1)$$

$$f(x_k) \leq f(x_*) - \langle \nabla f(x_k), x_* - x_k \rangle \quad - (2)$$



Combine (1) & (2)

$$f(x_{k+1}) \leq f(x_*) - \underbrace{\langle \nabla f(x_k), x_* - x_k \rangle} - \frac{1}{2\beta} \|\nabla f(x_k)\|^2.$$

$$f(x_{k+1}) - f(x_*) \leq \frac{1}{2\beta} \left[2\beta \cdot \langle \nabla f(x_k), x_k - x_* \rangle - \|\nabla f(x_k)\|^2 \right]$$

Recall: $x_{k+1} = x_k - \frac{1}{\beta} \nabla f(x_k)$

$$\Rightarrow \nabla f(x_k) = \beta \cdot (x_k - x_{k+1}).$$

$$f(x_{k+1}) - f(x_*) \leq \frac{1}{2\beta} \left[2\beta \langle \beta \cdot (x_k - x_{k+1}), x_k - x_* \rangle - \beta^2 \cdot \|x_k - x_{k+1}\|^2 \right]$$

$$= \frac{\beta}{2} \left[2 \cdot \langle \underbrace{x_k - x_{k+1}}_{u}, \underbrace{x_k - x_*}_{v} \rangle - \|x_k - x_{k+1}\|^2 \right]$$

$$\left(2\langle u, v \rangle = \|u\|^2 + \|v\|^2 - \|u - v\|^2. \right)$$

$$= \frac{\beta}{2} \left[\|x_k - x_{k+1}\|^2 + \|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2 \right].$$

$$- \|x_k - x_{k+1}\|^2$$

Therefore,

$$f(x_{k+1}) - f(x_*) \le \frac{\beta}{2}\left[\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2\right].$$

Add up all k inequalities.

$$f(x_1) - f(x_*) \le \frac{\beta}{2}\left[\|x_0 - x_*\|^2 - \|x_1 - x_*\|^2\right]$$

$$f(x_2) - f(x_*) \le \frac{\beta}{2}\left[\|x_1 - x_*\|^2 - \|x_2 - x_*\|^2\right]$$

$$\vdots$$

$$f(x_{k+1}) - f(x_*) \le \frac{\beta}{2}\left[\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2\right].$$

$$\sum_{i=1}^{k+1}\left(f(x_i) - f(x_*)\right) \le \frac{\beta}{2}\left[\|x_0 - x_*\|^2 - \|x_{k+1} - x_*\|^2\right].$$

Now: By the monotonicity of GD,

$$f(x_{k+1}) - f(x_*) \le f(x_i) - f(x_*) \quad \forall\ i \le k+1$$

$$\Rightarrow (k+1)\cdot\left(f(x_{k+1}) - f(x_*)\right) \le \frac{\beta}{2}\left[\|x_0 - x_*\|^2 - \|x_{k+1} - x_*\|^2\right]$$

$$\le \frac{\beta}{2}\|x_0 - x_*\|^2$$

$$\Rightarrow \boxed{f(x_{k+1}) - f(x_*) \le \frac{\beta}{2(k+1)}\|x_0 - x_*\|^2}.$$

Summary: If f is Convex β-Smooth, then

$$f(x_{k+1}) \leq f(x_*) + \frac{\beta}{2(k+1)} \|x_0 - x_*\|^2$$

What do you really need to run GD?

→ The only ability required is to compute gradients.

First-Order Methods of Optimization:
We have a subroutine that computes $\nabla f(x)$ at any point x.

→ What is the best you can do with First-Order Methods?

Nesterov's Accelerated Gradient Descent (NAGD) 1983:

$$f(x_k) \leq f(x_*) + \frac{\|x_0 - x_*\|^2}{\beta \cdot k^2}.$$

Remark: To get within ε of the optimum

GD takes → $\frac{1}{\varepsilon}$ iterations

NAGD takes → $\frac{1}{\sqrt{\varepsilon}}$ iterations!

NAGD

Momentum Updates

Start with $x_0 = y_0 = z_0$

For $k = 0, \ldots$:

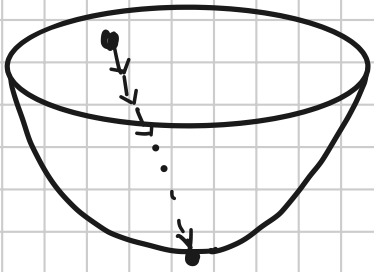
$$x_{i+1} = y_i - \eta \nabla f(y_i)$$

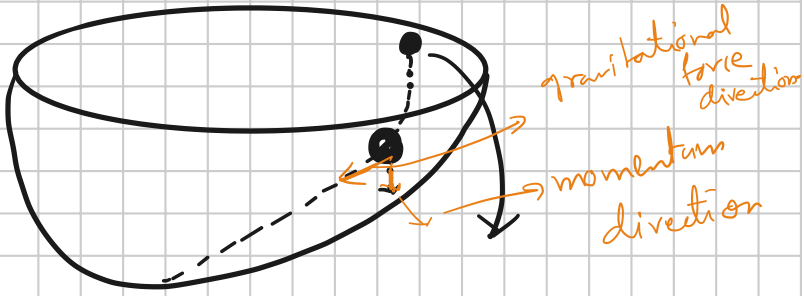
$$z_{i+1} = z_i - \eta_i \nabla f(y_i)$$

$$y_{i+1} = \alpha_i z_i + (1-\alpha_i) x_i$$

Thm: f is convex and smooth. $\eta \le \frac{1}{\beta}$, $\eta_i = \frac{(i+1)\eta}{2}$, $\alpha_i = \frac{2}{i+3}$.

$$f(x_k) \le f(x_*) + \frac{2\beta \|x_0 - x_*\|^2}{k^2}.$$





→ At every point gravitational pull ≡ along gradient.

↓

New velocity is a combination of

Current velocity and Force.

Intuitively: Velocity ≡ Change in position.

" $x_{k+1} - x_k$ " ≡ Some combination of " $x_k - x_{k-1}$ " and " $-\nabla f(x_k)$ "

NAGD is the best you can do among all First-Order Methods!

Demo:

Least Squares Regression: $(x_1, y_1), \ldots, (x_n, y_n)$

Parameter family: $h_w(x) = \langle w, x \rangle$

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

$$\nabla L(w) = \frac{1}{n} \sum_{i=1}^{n} 2(\langle w, x_i \rangle - y_i)\, x_i$$

In matrix notation: $L(w) = \frac{1}{n} \left\| \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} w - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right\|^2$

X $\quad$ $\tilde{y}$

$$= \frac{1}{n} \| Xw - y \|^2.$$

$$\nabla L(w) = \left(\frac{2}{n}\right) X^T (Xw - y).$$