

# 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  $X$

images,  
documents

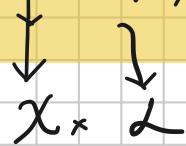
browsing profiles  
credit card history

Labels: CATS/DOGS / FOXES  $\Delta$

{0,1}

$\mathbb{R}$

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



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

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

$D$  on  $X$ :  $x_1, \dots, x_n$  are i.i.d samples from  $D$ .

$x$  is also from  $D$ .

Assumption 2: Labels cannot be arbitrary.

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

Assumptions: Prediction cannot be accurate all the time.

Accuracy of model  $\geq 90\%$

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

↓  
my prediction  
↓ "error"

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

Assumption 4: Cannot expect to succeed always!

Probability of generating 90% accurate model.  $\rightarrow$  "Probability of failure"  $\delta$ . over the dataset

PAC Model (Valiant, Vapnik, ...)

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

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

OUTPUT: Some predictor  $h$

with probability  $1-\delta$  over  $x_1, \dots, x_n$

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

PAC is a very strong requirement!

→ Often becomes intractable.

## Learning as Optimization

→ We have an underlying  $D$  on  $X$ . Labels  $d$ .

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

→ Loss function  $l: d \times d \rightarrow \mathbb{R}$

Goal: Find a hypothesis  $h$  such that  $x \rightarrow h(x) \rightarrow y$

$$\sum_{i=1}^n l(h(x_i), y_i) \text{ is small}$$

To avoid overfitting,

Parametrized hypothesis class:  $\hat{\mathcal{H}}$

Constraint the hypothesis class

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

is the predictor as specified by  
the "parameters"  $\theta$ .

# "Empirical Risk Minimization"

↓ (ERM)

We are trying loss on training data

(Ideal:

$$\mathbb{E}_{x \sim D} [l(h_\theta(x), y)].$$

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

Parametric family: Linear predictors

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

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

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

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

Least Squares Regression:

$$h_\theta(x) = \underline{\underline{\langle \theta, x \rangle}}$$

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

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

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

ERM: Find  $\theta$  to minimize  $L(\theta)$ .

Ex 2:  $l(a, b) = |a - b|$

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?  $d = 2$ , i.e.

Parametric family

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

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

↓  
Linear Threshold functions

Hypespaces

$$L(\theta) = \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{1}(\text{sign}(\langle \theta, x_i \rangle) \neq y_i)}_{\begin{array}{l} \downarrow \\ \text{sign matches } y_i \\ \text{if} \\ \text{else.} \end{array}}$$

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

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 :  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ → Parametrized family of predictors  $\mathcal{H}$ → Loss function  $l: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ 

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

(Notation: The minimizer  $\theta$  (argument)).Examples:

Predictors are linear functions

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

$$\begin{matrix} \downarrow \\ \in \mathbb{R}^d \end{matrix} \quad \begin{matrix} \searrow \\ \in \mathbb{R} \end{matrix}$$

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

(Notation for inner-product between vectors).

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

$l(a, b) = |a - b| \rightarrow$  "L<sub>1</sub>-Regression"

$l(a, b)$  = logistic loss  $\rightarrow$  Logistic Regression

$l(a, b)$  = Hinge-Loss

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



Recall LSR:

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

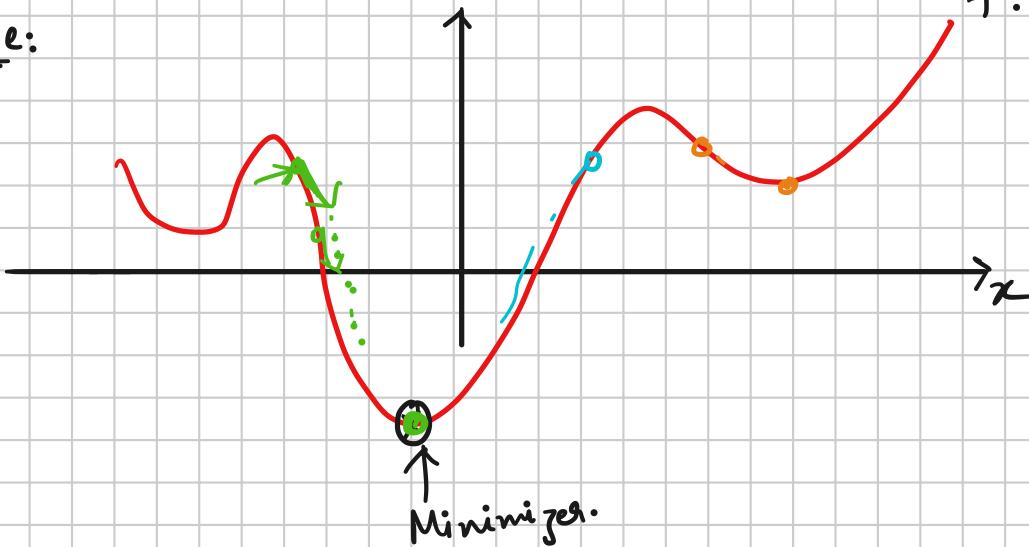
For L<sub>1</sub>-Loss:

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

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

Goal: Find  $\min f(\theta)$

Example:

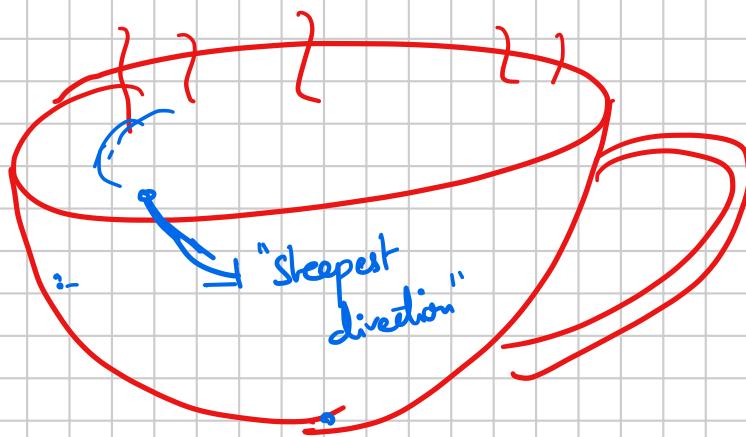


Idea: → Pick a point

→ "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  $-\nabla f(x)$ .

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

## GRADIENT DESCENT ALGORITHM (GD)

1.  $x_0$

2. For  $i = 1, \dots, T$ :

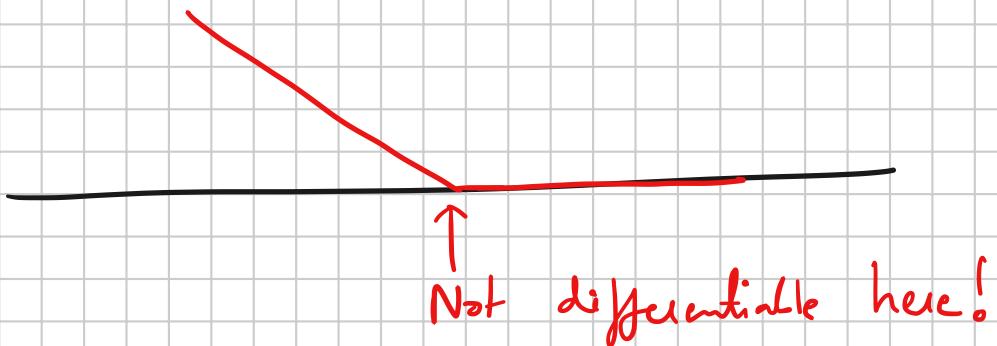
$$x_i = x_{i-1} - \eta \nabla f(x_{i-1})$$



"Step-size".

To Notebook Demo.

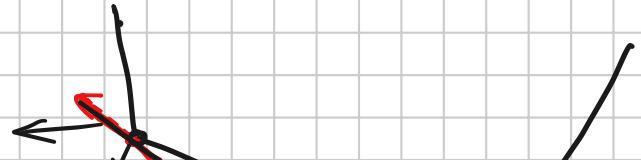
Dealing with Non-Differentiable functions.

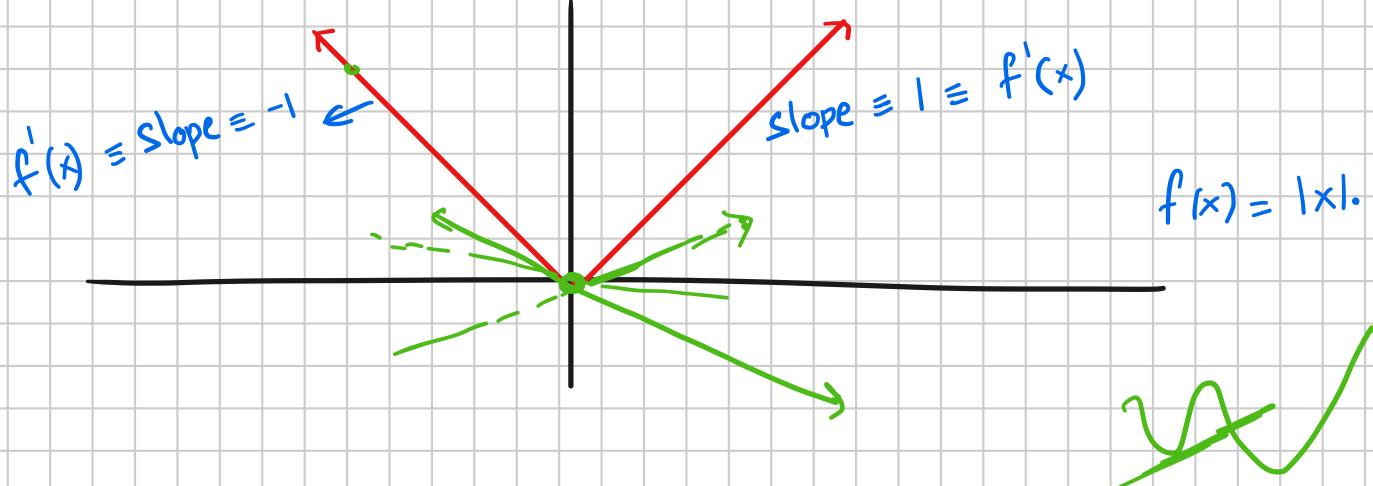
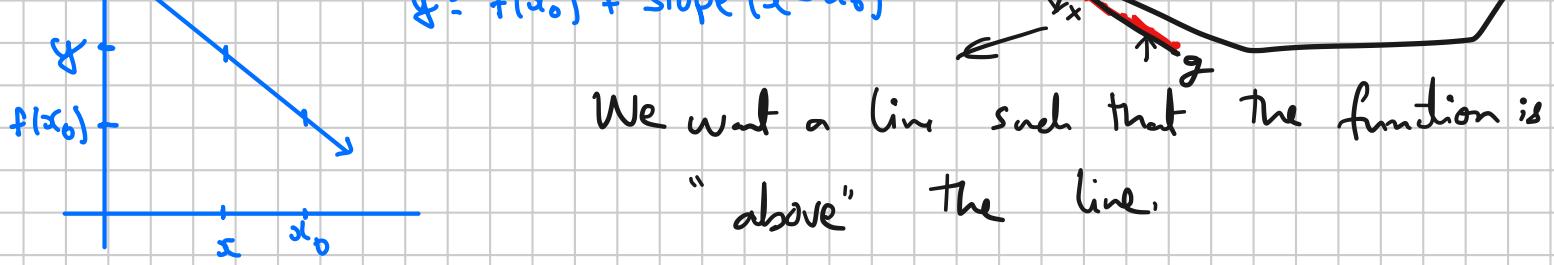


Sub-gradient: A direction  $\underline{g}$  is a sub-gradient of  $f$  at  $x_0$  if  $\forall x$   $\underline{f(x_0)} + \langle \underline{g}, x - x_0 \rangle \leq f(x)$ .

Slope of line:  $\underline{g}$

So any point  $x$   $\xrightarrow{\underline{g}}$   
 $\rightarrow f(x) \geq f(x_0) + \text{slope}(x - x_0)$





Summary: We can use subgradients as substitute for gradient.

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

$g = (\text{sign}(x_1), \text{sign}(x_2), \dots, \text{sign}(x_n))$  is a subgradient for  $f$ .

Subgradient	Descent
<ul style="list-style-type: none"> <li>→ <math>x_0</math></li> <li>→ For <math>i=1, \dots, T</math>:</li> </ul>	
	$x_i = x_{i-1} - \eta g$ <p style="text-align: right;">↑ a "Sub-gradient" of <math>f</math> at <math>x_{i-1}</math>.</p>

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

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

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

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 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  $\nabla f$ ?

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

Last class

→ Defined GD

→ Saw it 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:  $\min f(x)$

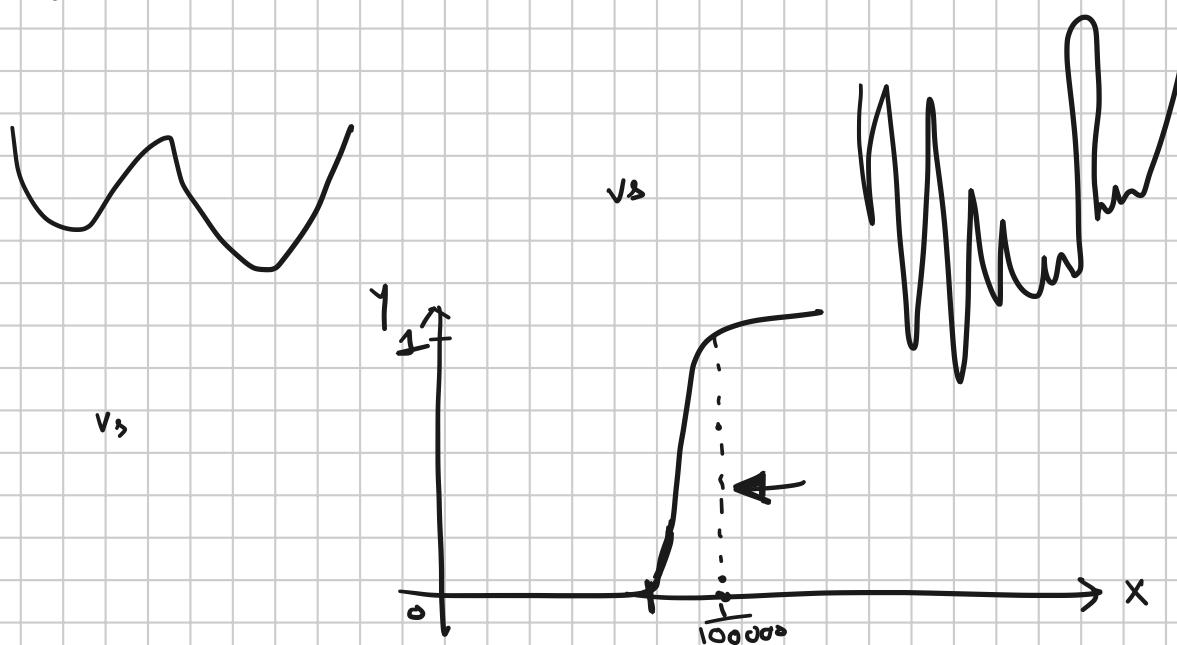
$f: \mathbb{R}^d \rightarrow \mathbb{R}$ .

GD:

→ For  $i=1, \dots, T$ :

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

Property 1: How sensitive is the function?



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

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

$\downarrow$   
 $\ell_2$ -distance between  $x$  and  $y$

2. Smoothness:  $f$  is  $\beta$ -Smooth if

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

Example:

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

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

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

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

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



$$\begin{aligned} f(x) - f(y) &= ax^2 + bx + c - (ay^2 + by + c) \\ &= a(x^2 - y^2) + b(x - y) \end{aligned}$$

$$= (x - y) \cdot \underbrace{(a(x + y) + b)}_{\downarrow}.$$

Monotonicity of GRD):

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

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

"GRD monotonically decreases the function value."

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

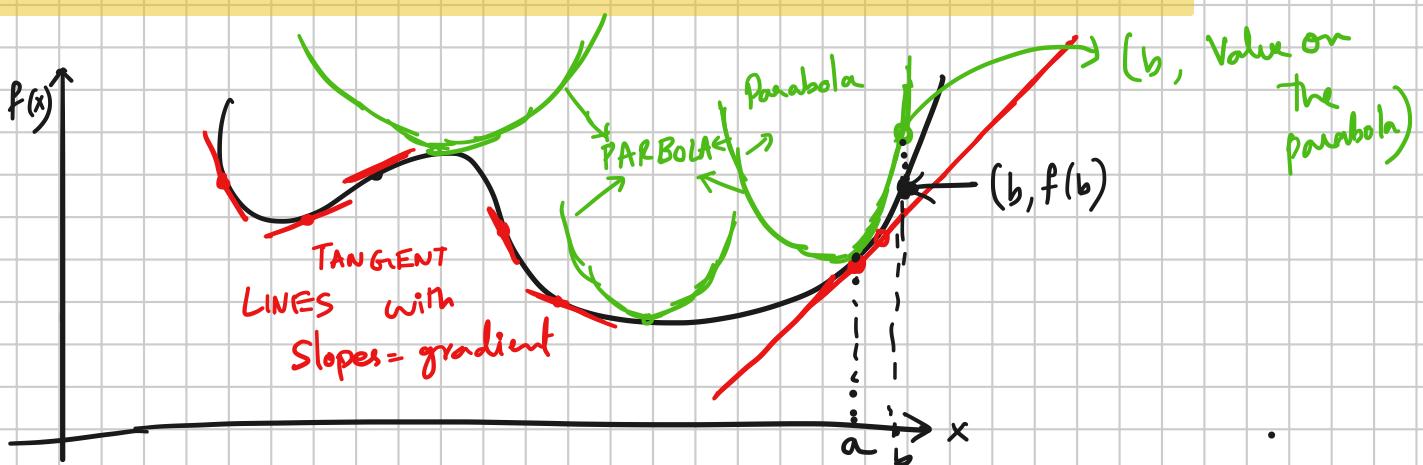
$$\|u\|_2^2 = \sum_{i=1}^d u_i^2).$$



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

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

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



Proof: (Taylor's theorem):

$$f(x+h) = f(x) + f'(x) \cdot h + \frac{1}{2} f''(x) \cdot h^2.$$

$\frac{1}{z} + \dots$

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

Prog of Monotonicity for Univariate Case:

$$\begin{aligned}
 f(x_{i+1}) &= f\left(x_i + \underbrace{\eta f'(x_i)}_{a} \right) \\
 &\leq f(x_i) + f'(x_i) \cdot \left( -\eta f'(x_i) \right) \\
 &\quad + \frac{\beta}{2} \cdot (-\eta f'(x_i))^2 \quad \left( \begin{array}{l} \text{Use Smoothness} \\ \text{upper bound} \\ x_i \end{array} \right) \\
 &= f(x_i) - \eta \cdot f'(x_i)^2 - \frac{\eta \beta}{2} f'(x_i)^2 \\
 &= f(x_i) - \eta \cdot \left( 1 - \frac{\eta \beta}{2} \right) \cdot f'(x_i)^2. \\
 &= f(x_i) - \frac{\eta}{2} \cdot f'(x_i)^2. \quad \eta \leq \frac{1}{\beta}.
 \end{aligned}$$

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

If  $f$  is  $\beta$ -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:

$$\begin{aligned} f(x_{i+1}) &= f(\underbrace{x_i - \eta \nabla f(x_i)}_y) \\ &\leq f(x_i) + \langle \underline{\nabla f(x_i)}, \underline{-\eta \nabla f(x_i)} \rangle + \frac{\beta}{2} \cdot \|(-\eta \nabla f(x_i))\|_2^2 \\ &= f(x_i) - \eta \cdot \|\nabla f(x_i)\|_2^2 + \frac{\eta^2 \beta}{2} \cdot \|\nabla f(x_i)\|_2^2 \\ &= f(x_i) - \eta \left(1 - \frac{\eta \beta}{2}\right) \cdot \|\nabla f(x_i)\|_2^2 \\ &\leq f(x_i) - \frac{\eta}{2} \cdot \|\nabla f(x_i)\|_2^2 \quad (\text{as } \eta \leq \frac{1}{\beta}). \end{aligned}$$

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

→ G1 makes progress as long as  $\eta \leq \frac{1}{\beta}$ .

(Theory to practice):

Practical tricks:

① Find largest  $\eta$  such that

$$f(x_i - \eta \nabla f(x_i)) \leq f(x_i) - \frac{\eta}{2} \cdot \|\nabla f(x_i)\|_2^2 \quad \textcircled{A}$$

(e.g.: Start with  $\eta = 1$ .)

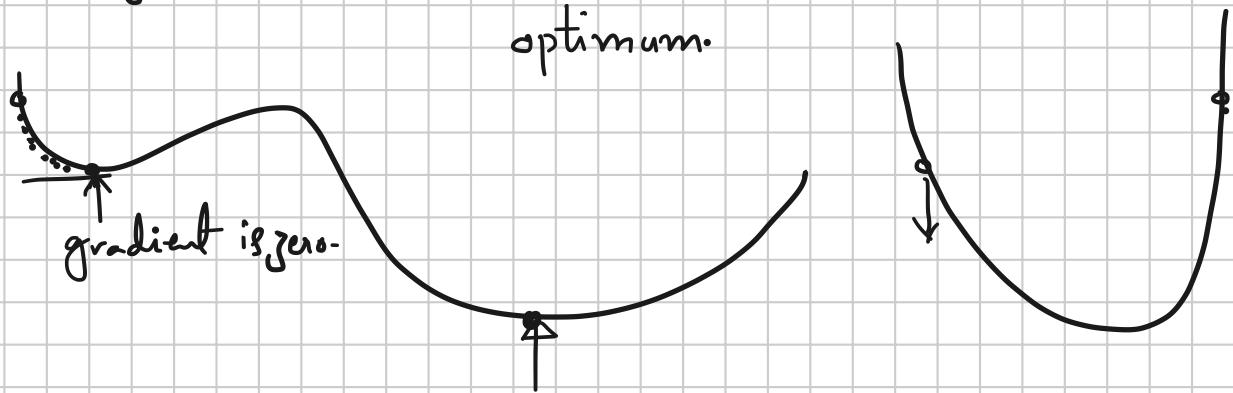
If  $\textcircled{A}$  holds, continue. Else try  $\eta = \frac{1}{2}, \dots$

② Can also do "Backtracking line search" to pick right  $\eta$ .

→ Monotonicity



We converge to the global optimum.



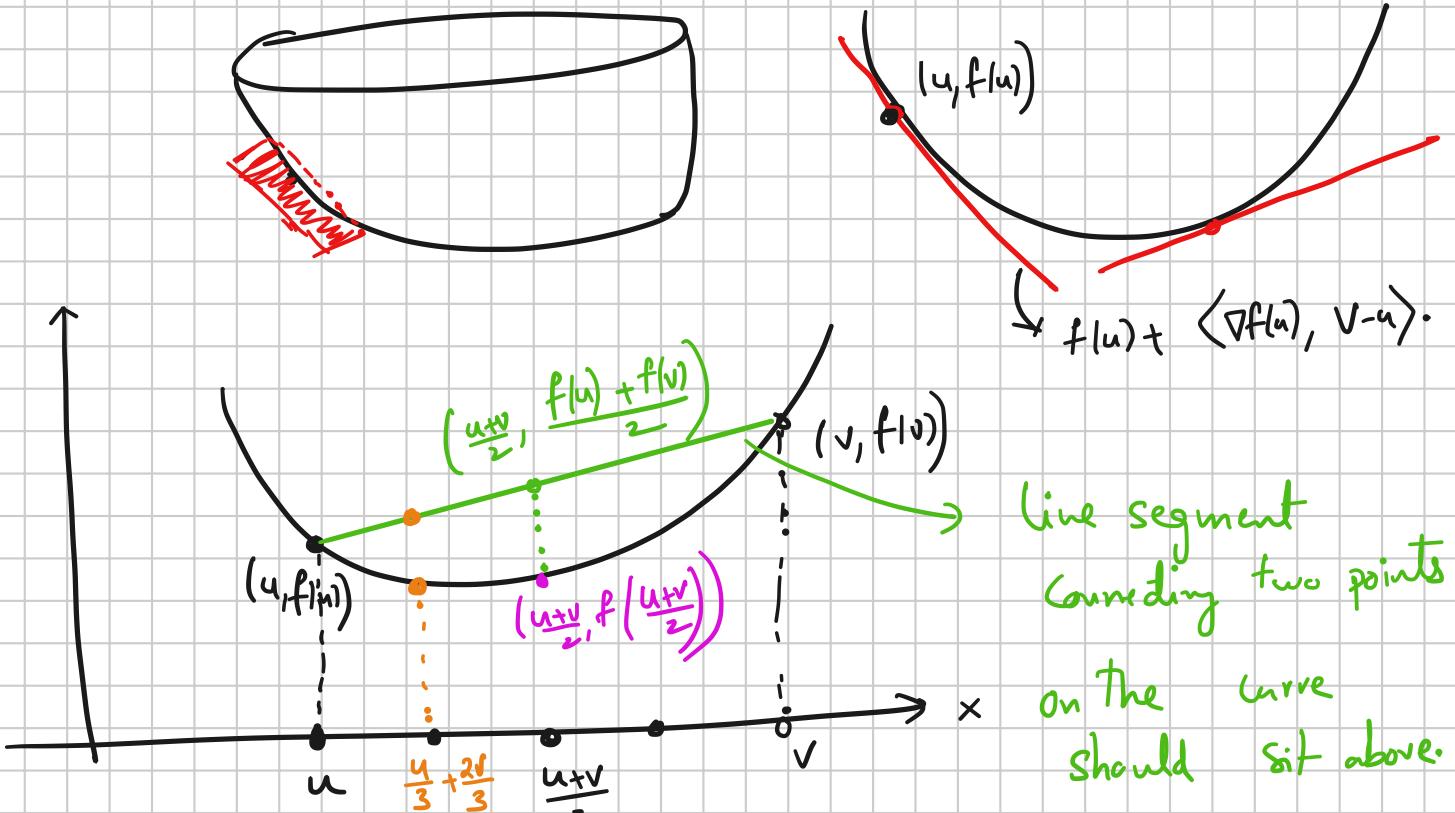
(CONVEX FUNCTIONS)

→ (MAGIC INGREDIENT IN OPTIMIZATION)

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

the tangent plane at any point is below  
the curve.

→



Equivalently:

$f: \mathbb{R}^d \rightarrow \mathbb{R}$  is convex if  
 $\forall u, v \quad f\left(\frac{u+v}{2}\right) \leq \frac{f(u) + f(v)}{2}.$

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

$\forall u, v, f(u) + \underbrace{\langle \nabla f(u), v-u \rangle}_{\downarrow} \leq f(v) \longrightarrow \textcircled{*}$   
↓  
the tangent function function.

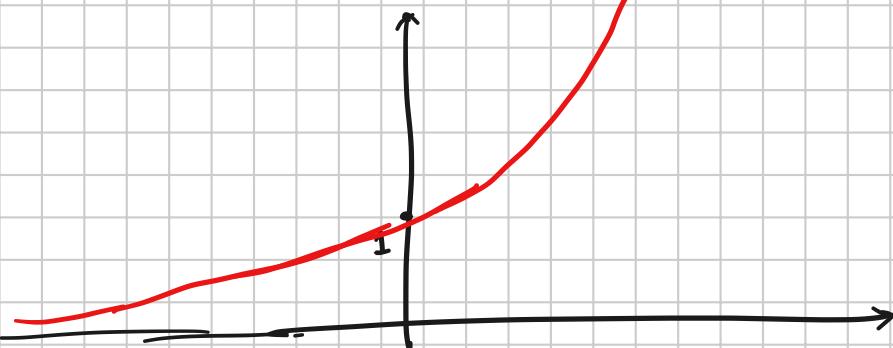
①  $f, g$  are convex  $\Rightarrow f+g$  is convex.

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

③  $\underline{g: \mathbb{R} \rightarrow \mathbb{R}}, \quad w \in \mathbb{R}^d$   
 $g_w: \mathbb{R}^d \rightarrow \mathbb{R} \quad g_w(x) = g(\langle w, x \rangle).$

$g$  is convex  $\Rightarrow g_w$  is convex.

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



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

$$g_{\omega}(x) = e^{\langle \omega, x \rangle} \text{ is Convex.}$$

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

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

$\rightarrow |x|$  is a convex function



Why Convexity:

ERM: Imagine we have parameter space  $\mathbb{H}$

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

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

$\rightarrow$  If  $l(h_{\theta}(x_i), y_i)$  is convex in  $\theta$ , then  $L$  is convex.

$\rightarrow$  Least Squares Regression:  $h_{\theta}(x_i) = \underbrace{\langle \theta, x_i \rangle}_{\text{inner-product}}$

$$l(h_{\theta}(x_i), y_i) = (\langle \theta, x_i \rangle - y_i)^2.$$

$\rightarrow$  LSR ERM:  $L(\theta) = \frac{1}{n} \sum_{i=1}^n (\langle \theta, x_i \rangle - y_i)^2$   
is a convex function in  $\theta$ .

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

$\rightarrow$  "LASSO":

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


---

$\rightarrow$  Linear Programming      }       $\rightarrow$  Minimizing a  
 $\rightarrow$  Semi-Definite Programming      }      Convex function!

CONVEX OPTIMIZATION IS EVERYWHERE

Thm: If  $f$  is  $\beta$ -Smooth and Convex, then  
 $(\text{if } n \leq \frac{1}{\beta}) \quad f(x_k) \leq f(x_\star) + \frac{\beta \cdot \|x_0 - x_\star\|^2}{2k} \xrightarrow{k \rightarrow \infty}$ .

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

② If  $f$  is  $L$ -Lipschitz, then

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

## 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.

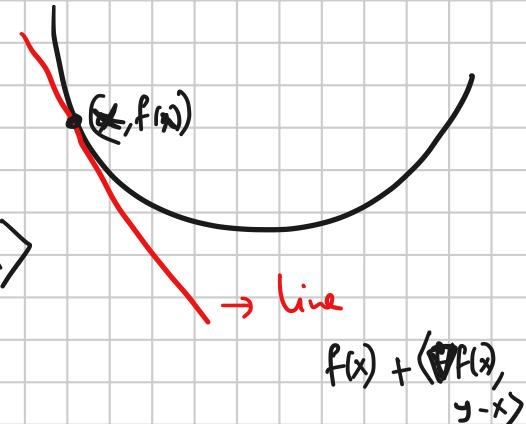
Thm:  $f$   $\beta$ -smooth and convex  $\Rightarrow$  GD works for  $n \leq \frac{1}{\beta}$ .

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

Proof:

Recall:

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

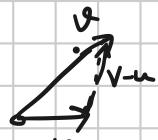


Some properties of vectors

$$\textcircled{a} \quad \|u-v\|^2 = \|u\|^2 + \|v\|^2 - 2\langle u, v \rangle.$$

$$\textcircled{a'} \quad \lambda \langle u, v \rangle = \|u\|^2 + \|v\|^2 - \|u-v\|^2. \quad \star \star$$

$$\textcircled{b} \quad |\langle u, v \rangle| \leq \|u\| \cdot \|v\| \quad (\text{Cauchy-Schwarz}).$$



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

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

(from last lecture ...)

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

(Because of Convexity!)

(1)

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

(2)

$$\text{line } f(x_k) + \langle \nabla f(x_k), y - x_k \rangle$$

Combine (1) & (2)

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

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

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

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

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

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

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

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

Therefore,

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

Add up all  $k$  inequalities.

$$\begin{aligned} f(x_1) - f(x_\star) &\leq \frac{\beta}{2} \left[ \|x_0 - x_\star\|^2 - \|x_1 - x_\star\|^2 \right] \\ f(x_2) - f(x_\star) &\leq \frac{\beta}{2} \left[ \|x_1 - x_\star\|^2 - \|x_2 - x_\star\|^2 \right] \\ &\vdots \\ &\vdots \\ f(x_{k+1}) - f(x_\star) &\leq \frac{\beta}{2} \left[ \|x_k - x_\star\|^2 - \|x_{k+1} - x_\star\|^2 \right]. \end{aligned}$$

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

Now: By the monotonicity of  $f$ ,

$$f(x_{k+1}) - f(x_\star) \leq \boxed{f(x_i) - f(x_\star)} \quad \forall i \leq k+1$$

$$\begin{aligned} \Rightarrow (k+1) \cdot (f(x_{k+1}) - f(x_\star)) &\leq \frac{\beta}{2} \left[ \|x_0 - x_\star\|^2 - \|x_{k+1} - x_\star\|^2 \right] \\ &\leq \frac{\beta}{2} \|x_0 - x_\star\|^2 \end{aligned}$$

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

Summary:

If  $f$  is convex  $\beta$ -smooth, then

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

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:

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

Remark: To get within  $\epsilon$  of the optimum  
GD takes  $\rightarrow \frac{1}{\epsilon}$  iterations

NAGD takes  $\rightarrow \frac{1}{\sqrt{\epsilon}}$  iterations!

Start with  $x_0 = y_0 = z_0$

### Momentum Updates

For  $i=0, \dots :$

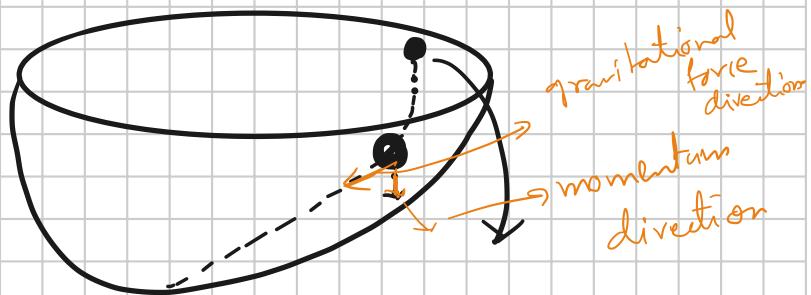
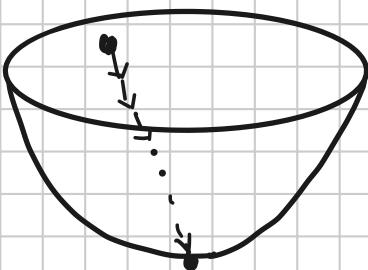
$$x_{i+1} = y_i - \eta \nabla f(y_i)$$

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

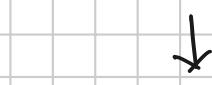
$$y_{i+1} = \alpha_i z_i + (1-\alpha_i) x_i$$

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

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



→ At every point gravitational pull  $\equiv$  along gradient.



New velocity is a combination of

current velocity and force.

Intuitively: Velocity  $\equiv$  Change in position.

" $x_{k+1} - x_k$ "  $\equiv$  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:  $(x_1, y_1), \dots, (x_n, y_n)$

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

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

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

In matrix notation:

$$L(w) = \frac{1}{n} \|$$

$$\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_n \end{array} \underbrace{\quad}_{X} \quad \underbrace{\begin{array}{c} w \\ \vdash \end{array}}_{w} - \begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \underbrace{\quad}_{y} \|^2$$

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

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

# Algorithmic ML

## Lecture 5 :

Last 2 lectures :

Gradient Descent (GD/NAGD)

Analysis of GD

This lecture :

Stochastic Gradient Descent (SGD)

GD : (parameter : "step size" or "learning rate"  $\eta$ )

1. Choose  $w_0 \in \mathbb{R}^d$

2. for  $i = 1, \dots, T$ :

$$w_i = w_{i-1} - \eta \nabla f(w_{i-1})$$

Theorem (GD convergence)  $f: \mathbb{R}^d \rightarrow \mathbb{R}$   $\beta$ -smooth & convex

$$\Rightarrow f(w_k) \leq f(w^*) + \frac{\beta \|w^* - w_0\|}{2k} \quad \text{when } \eta = \gamma_\beta .$$

↳ if  $\beta, \|x^* - x_0\| = O(1)$

then GD returns  $w$  such that  $f(w) \leq f(w^*) + \varepsilon$

in  $O(\gamma_\varepsilon)$  iterations.

Pro. : Always converges in  $O(\gamma_\varepsilon)$  iterations.

Con. : What is the time complexity per iteration?

→ Cost of computing  $\nabla f(w)$ .

Taking a step back: what  $f$  are we working with?

Recall ERM: (empirical risk minimization) Given labeled examples ("training data")  $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R} \dots$

Empirical loss function  $\underset{\text{define}}{\rightarrow} L(w) = \frac{1}{n} \sum_{i=1}^n l(h_w(x_i), y_i)$  where

Our "f" in GD.  $h_w: \mathbb{R}^d \rightarrow \mathbb{R}$  the hypothesis corresponding to  $w$  (parametrized by)

Trying to find  $w^*$ . that minimizes this loss.  $l: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  a loss function.

$$l(a, b)$$

$$\begin{aligned} \text{Gradient: } \nabla L(w) &= \frac{1}{n} \sum_{i=1}^n \nabla l(h_w(x_i), y_i) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial a} l(h_w(x_i), y_i) \frac{\partial}{\partial w} h(x_i) \end{aligned}$$

Example: LSR (least squares regression)

$$\left. \begin{array}{l} \bullet h_w(x_i) = \langle w, x_i \rangle \\ \bullet l(a, b) = (a - b)^2 \end{array} \right\} \Rightarrow L(w) = \frac{1}{n} \sum_{i=1}^n (\langle w, x_i \rangle - y_i)^2$$

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

$$\begin{aligned} \nabla L(w)_j &= \frac{1}{n} \sum_{i=1}^n 2(\langle w, x_i \rangle - y_i) \underbrace{\frac{\partial}{\partial w_j} \langle w, x_i \rangle}_{= x_{ij}} \\ &= \frac{2}{n} \sum_{i=1}^n (\langle w, x_i \rangle - y_i) x_{ij} = \frac{\partial}{\partial w_j} \sum_{i=1}^n w_i x_{ij} = x_{ij} \end{aligned}$$



Costs  $O(n \cdot d)$  time to compute.

( $n$  examples,  $d$  dimensions)

$\Rightarrow$  Time complexity of GD  $\approx O\left(\frac{n \cdot d}{\epsilon}\right)$

$n \cdot d$  very bad in many applications.

### Stochastic Gradient Descent:

Idea: efficiently estimate  $\nabla f(w)$ .

- Let  $G(w)$  be an estimator for  $\nabla f(w)$

i.e.  $E[G(w)] = \nabla f(w)$  (unbiased estimator)

$\rightarrow G(w)$  a random vector in  $\mathbb{R}^d$

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

Could use:  $G(w) = 2 \left( \langle w, x_{i^*} \rangle - y_{i^*} \right) x_{i^*}$  where  
 $i^*$  sampled u.a.r.

or . for  $k \ll n$ :

$G(w) = \frac{2}{k} \sum_{i \in S} (\langle w, x_i \rangle - y_i) x_i$  where  
 $S \subseteq \{1, \dots, n\}$ ,  $|S| = k$  random.

SGD: (using estimator  $G$ )

1) choose  $w_0$

2) for  $i=1, \dots, T$ :

let  $w_{i+1} = w_i - \eta G(w_i)$

How "good" is  $G_i$ ? Small variance is good.

$$\cdot \text{Var}(G_i) = \mathbb{E}[\|G_i(\omega) - \mathbb{E}[G_i(\omega)]\|_2^2]$$

$$\begin{aligned} & \mathbb{E}[\|G_i - \mathbb{E}[G_i]\|_2^2] \\ &= \mathbb{E}\left[\sum_{j=1}^d (G_{ij} - \mathbb{E}[G_{ij}])^2\right] = \sum_{j=1}^d \mathbb{E}[(G_{ij} - \mathbb{E}[G_{ij}])^2] \\ &= \sum_{j=1}^d \text{Var}(G_{ij}) \end{aligned}$$

Then (SGD convergence) :

$f$  convex &  $\beta$ -smooth ,  $\eta \leq 1/\beta$ ,

$$\text{Var}(G_i(\omega)) \leq \sigma^2$$

$$\Rightarrow \mathbb{E}\left[f(\bar{\omega}_k)\right] \leq f(\omega^*) + \frac{\|\omega_0 - \omega^*\|_2^2}{2\eta \cdot k} + \eta \sigma^2$$

$$\bar{\omega}_k = \frac{1}{k} (\omega_0 + \dots + \omega_k)$$

error term reduced by  
error term varying on

Remarks :

- SGD is what is most commonly used in practice (variants).
- Can get "accelerated" version of SGD (as in NAGD) just by replacing  $\nabla f(\omega_k)$  with  $G_i(\omega_k)$  in NAGD.

Proof of SGD convergence :

useful things:  $\forall u, v \in \mathbb{R}^d$ :

- convexity:  $f(u) \leq f(v) + \langle \nabla f(v), u - v \rangle$

- smoothness:  $f(u) \leq f(v) + \langle \nabla f(v), u - v \rangle + \frac{\kappa}{2} \|u - v\|_2^2$

- cosine formula:  $2 \langle u, v \rangle = \|u\|_2^2 + \|v\|_2^2 - \|u - v\|_2^2$

- SGD update:  $w_{t+1} = w_t - \eta_t G(w_t)$

- unbiased estimator:  $\mathbb{E}[G(\omega)] = \nabla f(\omega)$

$$\Rightarrow \mathbb{E}\left[\|G(\omega)\|_2^2\right] - \|\nabla f(\omega)\|_2^2 \leq \sigma^2$$

$$\begin{aligned} \sigma^2 &\geq \mathbb{E}\left[\|G(\omega) - \mathbb{E}[G(\omega)]\|_2^2\right] = \mathbb{E}\left[\|G(\omega)\|_2^2\right] - \|\mathbb{E}[G(\omega)]\|_2^2 \\ &= \mathbb{E}\left[\|G(\omega)\|_2^2\right] - \|\nabla f(\omega)\|_2^2 \end{aligned}$$

Clm 1 : (SGD function decrease ineq)  $\forall k$  (from smoothness)

$$\star \quad \mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(w_k)] - \frac{\eta}{2} \mathbb{E}\left[\|\nabla f(w_k)\|_2^2\right] + \frac{\eta\sigma^2}{2}$$

pf:

$$\begin{aligned} (\text{smoothness}) \quad f(w_{k+1}) &\leq f(w_k) + \langle \nabla f(w_k), w_{k+1} - w_k \rangle + \frac{\beta}{2} \|w_{k+1} - w_k\|_2^2 \\ &= f(w_k) + \langle \nabla f(w_k), -\eta G(w_k) \rangle + \frac{\beta\eta^2}{2} \|G(w)\|_2^2 \end{aligned}$$

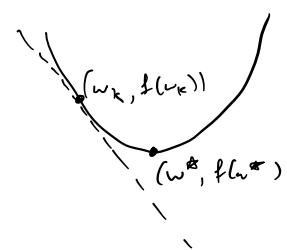
$$\begin{aligned} \sigma^2 &\geq \mathbb{E}\left[\|G(w) - \mathbb{E}[G(w)]\|_2^2\right] = \mathbb{E}\left[\|G(w)\|_2^2\right] - \|\mathbb{E}[G(w)]\|_2^2 \\ &= \mathbb{E}\left[\|G(w)\|_2^2\right] - \|\nabla f(w)\|_2^2 \end{aligned}$$

$$\begin{aligned} \mathbb{E}[f(w_{k+1})] &\leq \mathbb{E}[f(w_k)] - \eta \left( \mathbb{E}\left[\|\nabla f(w_k)\|_2^2\right] + \frac{\beta\eta^2}{2} \mathbb{E}\left[\|\nabla f(w_k)\|_2^2\right] + \frac{\beta\eta^2\sigma^2}{2} \right) \\ \left(\eta \leq \frac{1}{\beta}\right) \quad &\leq \mathbb{E}[f(w_k)] - \frac{\eta}{2} \mathbb{E}\left[\|\nabla f(w_k)\|_2^2\right] + \frac{\eta\sigma^2}{2}. \quad \square \end{aligned}$$

Now, similar to GD analysis, let's use convexity of  $f$ :

$$\star \star \quad \text{Convexity: } f(w_k) \leq f(w^*) + \langle \nabla f(w_k), w_k - w^* \rangle$$

$$(- \langle \nabla f(w_k), w_k - w^* \rangle \leq f(w^*) - f(w_k))$$



$$\begin{aligned} \star \star \quad \frac{1}{2\eta} \mathbb{E}\left[\|w_{k+1} - w^*\|_2^2 - \|w_k - w^*\|_2^2\right] &\quad (\text{Relates point distance to the gradient}) \\ \text{Twist ineq:} \quad &= \frac{\eta}{2} \mathbb{E}\left[\|\nabla f(w_k)\|_2^2\right] - \mathbb{E}\left[\langle \nabla f(w_k), w_k - w^* \rangle\right] + \frac{\eta\sigma^2}{2} \end{aligned}$$

$$\text{Pf : } \|\omega_{k+1} - \omega^*\|_2^2 = \underbrace{\|\omega_k - \omega^* - \eta g(\omega_k)\|_2^2}_{\text{in}} = \|\omega_k - \omega^*\|_2^2 + \eta^2 \|g(\omega_k)\|_2^2 - 2\eta \langle \omega_k - \omega^*, g(\omega_k) \rangle$$

$$\Rightarrow \mathbb{E} \left[ \|\omega_{k+1} - \omega^*\|_2^2 - \|\omega_k - \omega^*\|_2^2 \right] + \eta^2 \mathbb{E} \left[ \|\nabla f(\omega_k)\|_2^2 + \eta^2 \sigma^2 \right] - 2\eta \mathbb{E} \left[ \langle \nabla f(\omega_k), \omega_k - \omega^* \rangle \right]$$

multiply by  $\frac{1}{2\eta} \dots \square$

Main proof continued...

$$(\text{---} \mathbb{E}[\langle \nabla f(\omega_k), \omega_k - \omega^* \rangle]) \leq f(\omega^*) - \mathbb{E}[f(\omega_k)]$$

$\star\star\star + \star\star \Rightarrow$

$$\begin{aligned} \frac{1}{2\eta} \mathbb{E} \left[ \|\omega_{k+1} - \omega^*\|_2^2 - \|\omega_k - \omega^*\|_2^2 \right] &\leq \frac{\eta}{2} \mathbb{E} \left[ \|\nabla f(\omega_k)\|_2^2 \right] + f(\omega^*) - \mathbb{E}[f(\omega_k)] + \frac{\eta\sigma^2}{2} \\ &\leq f(\omega^*) - \mathbb{E}[f(\omega_{k+1})] + \eta\sigma^2 \quad (\text{by } \star) \end{aligned}$$

$$\Rightarrow \forall i=1, \dots, k : \quad (\text{Telescoping sum trick})$$

$$\mathbb{E}[f(\omega_i)] \leq f(\omega^*) - \frac{1}{2\eta} \mathbb{E} \left[ \|\omega_i - \omega^*\|_2^2 - \|\omega_{i-1} - \omega^*\|_2^2 \right] + \eta\sigma^2$$

$$\begin{aligned} \Rightarrow \sum_{i=1}^k \mathbb{E}[f(\omega_i)] &\leq k \cdot f(\omega^*) - \frac{1}{2\eta} \left( \|\omega_k - \omega^*\|_2^2 - \|\omega_0 - \omega^*\|_2^2 \right) + k\eta\sigma^2 \\ &\leq k \cdot f(\omega^*) + \frac{\|\omega_0 - \omega^*\|_2^2}{2\eta} + k\eta\sigma^2 \end{aligned}$$

$$\Rightarrow \mathbb{E} \left[ \frac{1}{k} \sum_{i=1}^k f(\omega_i) \right] \leq f(\omega^*) + \frac{\|\omega_0 - \omega^*\|_2^2}{2\eta} + \eta\sigma^2$$

$$\text{Now, } f \text{ convex} \Rightarrow f \left( \frac{1}{k} \sum_{i=1}^k \omega_i \right) \leq \frac{1}{k} \sum_{i=1}^k f(\omega_i)$$

and recall that SGD output is  $\bar{\omega}_k = \frac{1}{k} \sum_{i=1}^k \omega_i$ .  $\square$

## Concluding Remarks :

Then (SGD convergence) :

$f$  convex &  $\beta$ -smooth ,  $\eta \leq \frac{1}{\beta}$ ,

$$\text{Var}(G(\omega)) \leq \sigma^2$$

$$\Rightarrow \mathbb{E} \left[ f(\bar{\omega}_k) \right] \leq f(\omega^*) + \frac{\|\omega_0 - \omega^*\|_2^2}{2\eta \cdot k} + \eta \sigma^2$$

Choosing step-size :

in standard GD :

just want to pick  $\eta$  as large as possible (i.e.  $\eta = \frac{1}{\beta}$ )

large  $\eta$  is good:  
allows us to converge faster  
(as long as  $\eta \leq \frac{1}{\beta}$ )

small  $\eta$  is good:  
mitigates error of estimator

in SGD : have to balance the 2 error terms

general guideline : choose  $\eta \approx \frac{1}{\sqrt{k}}$

$$\eta = \frac{1}{\sqrt{k}} \Rightarrow \mathbb{E} \left[ f(\bar{\omega}_k) \right] \leq f(\omega^*) + \frac{1}{\sqrt{k}} \left( \frac{\|\omega_0 - \omega^*\|_2^2}{2} + \sigma^2 \right)$$

$$\eta = \frac{\|\omega_0 - \omega^*\|_2}{\sigma \sqrt{2k}} \Rightarrow \mathbb{E} \left[ f(\bar{\omega}_k) \right] \leq f(\omega^*) + \frac{\frac{1}{2} \cdot \sigma \|\omega_0 - \omega^*\|_2}{\sqrt{k}}$$

~ in practice, step-size choice can be more nuanced depending on the specific application.

... or use  $\eta_t = \frac{1}{\sqrt{t}}$  (step-size decreases over time)

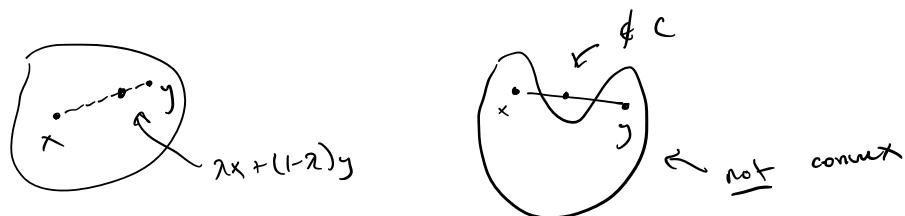
Constrained Optimization: So far we have only considered optimizing over all of  $\mathbb{R}^d$ .

i.e. 
$$\boxed{\underset{w \in \mathbb{R}^d}{\operatorname{argmin}} L(w)}$$

In many situations we want  $w$  to satisfy some constraints.  
i.e. want  $w$  to lie in some bounded region of  $\mathbb{R}^d$ .

It turns out we can adapt GD / SGD / NAGD to this situation  
as long as the constraint region is a convex set.

Defn:  $C \subseteq \mathbb{R}^d$  is convex set if  
 $\forall x, y \in C, \frac{x+y}{2} \in C$  (midpoint also in  $C$ )  
equiv:  $\forall x, y \in C, \forall \lambda \in [0, 1], \lambda x + (1-\lambda)y \in C$



Goal: compute 
$$\boxed{\underset{w \in C}{\operatorname{argmin}} L(w)}$$

Projected Gradient Descent:

Projection:  $\text{Proj}_C(y) = \underset{x \in C}{\operatorname{argmin}} \|x-y\|_2$  (closest point in  $C$  to  $y$ )

PGD update :  $w_{k+1} = \text{Proj}_C(w_k - \alpha \nabla f(w_k))$   
 or  $G(w_k)$  in case of SGD

Remark : Everything known for GDO/SGO/NAGD

extends to PGD as long as

$C$  is a convex set.

---

