LEARNING:

→ Reinforcement , Classification etc.

MODULE - 1

LEARNING AS OPTIMIZATION:

Input : $x_1, x_2 \dots x_n \in D$

↳ Distribution

$y_1, y_2 \dots y_n$ → Labels such that

$$\mathcal{H} : x \to y$$

(There is a hypothesis function)

Loss FUNCTION:

$$l : \text{Label} \times \text{Label} \to \mathbb{R}$$

Output : h such that loss is least but to avoid overfitting we will constrain it by parameter θ : Parameterized Hypothesis Class. (H)

GOAL: EMPIRICAL RISK MINIMIZATION (ERM)

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

GD :

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

Functions that are Good/Bad for GD:

Lipschitzness:

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

f is L-Lipschitz if

$$\forall x, y \qquad \boxed{|f(x) - f(y)| \leq L \|x - y\|_2}$$

L_2 norm $\sqrt{\sum x_i^2}$

Smoothness:

f is β-Smooth if

$$\forall x, y \qquad \boxed{\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2}$$

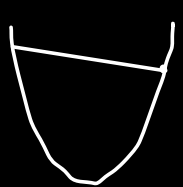
[Smoothness is a stronger constrain]

Monotonicity of GD:

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

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

[Idea: So β gives an idea of how much to jump without gradient being too different]



avoids such jumps.

So monotonic!

note: * If β is small, we say that gradient change is small → better for GD.

So η can be large. Learn fast.

* Stops when gradient = 0.

SMOOTHNESS UPPERBOUND:

* f is β-smooth $\forall a, b$ $\quad [f: \mathbb{R} \to \mathbb{R}]$

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

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

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

TAYLOR'S THEOREM:

$$f(x+h) = f(x) + f'(x) \cdot h + \underbrace{\int_0^1 (f'(x+th) - f'(x))\, th\, dt}$$

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

MONOTONICITY PROOF:

UNIVARIATE:

Given:

- $\| \nabla f(x) - \nabla f(y) \|_2 \leq \beta \| x - y \|_2$ [Smoothness]
- $x_{i+1} = x_i - \eta f'(x_i)$ [GD]
- $f(b) \leq f(a) + f'(a) \cdot (b-a) + \frac{\beta}{2}(b-a)^2$ [Smoothness upper bound]

PROOF:

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

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

$$a = x_i$$

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

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

$$\eta \leq 1/\beta \qquad \beta \geq \frac{1}{\eta}$$

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

MULTIVARIATE:

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

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

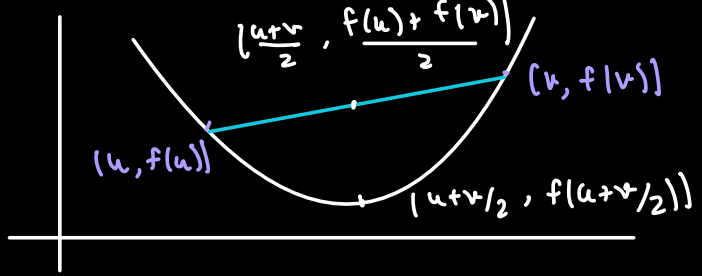
$$\leq f(x_i) - \eta \| \nabla f(x_i) \|_2^2 + \frac{\eta^2 \beta}{2} \| \nabla f(x_i) \|_2^2$$

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

note: $\boxed{\langle x, x \rangle = \|x\|_2^2}$

CONVEX FUNCTIONS: [Only one minima]

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



$\forall\ u, v$

$$f\left(\frac{u+v}{2}\right) \leq \frac{f(u)+f(v)}{2}$$

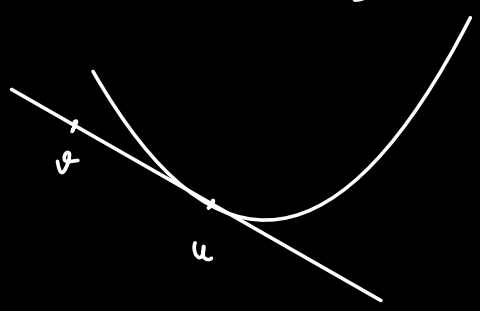
$\lambda \in (0,1)$

$$f(\lambda u + (1-\lambda) v) \leq \lambda f(u) + (1-\lambda) f(v)$$

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

↓

value at v along the tangent line



value at v = $f(u)$ + slope $(v-u)$

f, g : convex → $f+g$, $a \cdot f$ } Convex

$a > 0$

$f - g$ - need not be Convex

$\Rightarrow$ $g : \mathbb{R} \to \mathbb{R}$: convex

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

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

↳ Convex!

eg: e^x, x^2, $(x-a)^2$, $|x|$

$e^{\langle w,x\rangle}$, $\langle w,x\rangle^2$, $(\langle w,x\rangle - a)^2$, $|\langle w,x\rangle|$

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

ℓ: $a > 0$; Sum of convex

So only prove $\ell\left(h_\theta(x_i), y_i\right)$ is convex.

* Least Squares Regression:

Loss: $(\langle \theta, x_i\rangle - y_i)^2$

$\downarrow$

$(x-a)^2$

$\downarrow$

Convex.

PROVE CONVEX:

* Remove positive values multiplied.
* If sum, prove individually convex.
* If $\langle \theta, x_i\rangle$, prove for just x_i.
* Know e^x, x^2, $(x-a)^2$, $|x|$ are convex.

Convex - GD Convergence:

If f is β-smooth and convex, then

if $\eta \leq 1/\beta$

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

Note: If β is low, good function,
So converge faster

If starting point close to optimum,
converge faster.

x_* → Global Optimum

x_0 → Starting point

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

So Distance to optimum $\propto 1/k$

To get with ε of of $x_k \rightarrow 1/\varepsilon$ steps.

- If f was L-Lipschitz

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

Distance to optimum $\propto 1/\sqrt{k}$ $\quad 1/\varepsilon^2$ steps

PROOF :

To Prove:

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

Given:

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

* $f(u) + \langle \nabla f(u), v - u \rangle \leq f(v)$ [CONVEX]

* **Properties of vectors:**

  → $\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2\langle u, v \rangle$

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

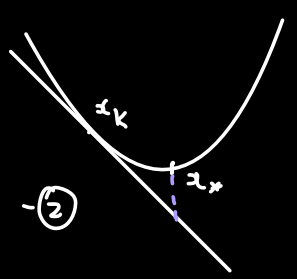
  → $|\langle u, v \rangle| \leq \|u\| \cdot \|v\|$ [Cauchy - Schwarz]

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

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

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

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



Combine ①, ②

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

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

$$f(x_{k+1}) \le f(x_*) + \beta \langle x_k - x_{k+1}, x_k - x_* \rangle - \frac{\beta}{2} \|(x_k - x_{k+1})\|_2^2$$

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

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

$$\Rightarrow \quad f(x_1) - f(x_*) \le \frac{\beta}{2} \|x_0 - x_*\|^2 - \frac{\beta}{2} \|x_* - x_1\|^2$$

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

$$\vdots$$

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

Add them

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

As GD is monotonic

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

$$f(x_{k+1}) - f(x_*) \leq f(x_i) - f(x_*)$$

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

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

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

NESTEROV'S ACCELERATED GRADIENT DESCENT (NAGD):

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

$\eta \leq 1/\beta$

$\eta_i = \frac{(i+1)\eta}{2}$

$\alpha_i = 2/(i+3)$

To be within $\varepsilon = 0.01$ of optimum:

Convex:

* Lipschitz + GD $\Rightarrow$ $1/\varepsilon^2 = 1/10^{-4} = 10{,}000$ iterations
* Smooth + GD $\Rightarrow$ $1/\varepsilon = 100$ iterations
* Smooth + NAGD $\Rightarrow$ $1/\sqrt{\varepsilon} = 10$ iterations.

What about time taken for each iteration!

For LSR:

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

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

(iterate over d dimensions; iterate over n examples.)

So each iteration takes $O(nd)$ time.

So to be within ε of answer

$1/\varepsilon$ iterations $\times$ $O(nd)$ per iteration $\Rightarrow O(nd/\varepsilon)$ time.

STOCHASTIC GRADIENT DESCENT:

k elements → Gradient

So $x_{t+1} = x_t - \eta\, G(x_t)$

↳ Gradient estimator

* G should have low variance

$$\text{var}(G) = \sum_{i=1}^{d} \text{var}(G_i) \quad \text{and} \quad \mathbb{E}[G(x)] = \nabla f(x)$$

SGD CONVERGENCE:

f : convex and β-smooth

$\eta \leq 1/\beta$, $\text{var}(G(x)) \leq \sigma^2$

$$\mathbb{E}[f(\tilde{x}_k)] \leq f(x_*) + \frac{\|x_0 - x_*\|_2^2}{2\eta k} + \eta\sigma^2$$

only difference.

$\frac{1}{k}(x_1 + \dots + x_k)$

average of all the iteration values.

equate

$$\eta = \frac{\|x_0 - x_*\|_2^2}{\sigma\sqrt{k}}$$

=> So take $\eta_k = 1/\sqrt{k}$, not $1/\beta$.

Monotonicity of SGD:

Given Smoothness

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

Proof:

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

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

[$f(b) \leq f(a) + \langle \nabla f(a), b-a \rangle + \frac{\beta}{2}\|b-a\|_2^2$ - Smoothness Upperbound]

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

Apply Expectation

$$E[f(x_{i+1})] \leq E[f(x_i)] - \eta \langle \underbrace{E[\nabla f(x_i)]}_{\nabla f(x_i)}, \underbrace{E[G(x_i)]}_{\nabla f(x_i)} \rangle + \frac{\eta^2\beta}{2} E\|[G(x_i)]\|_2^2$$

$$\leq E[f(x_i)] - \eta E[\|\nabla f(x_i)\|_2^2] + \frac{\eta^2\beta}{2} E\|\nabla f(x_i)\|_2^2 + \frac{\eta\sigma^2}{2}$$

$$\mathrm{var}(G(x,\cdot)) = \mathbb{E}\left[\|G(x) - \mathbb{E}[G(x)]\|_2^2\right]$$

$$= \mathbb{E}\left[\|G(x)\|_2^2\right] - \|\mathbb{E}[G(x)]\|_2^2$$

$$= \mathbb{E}\left[\|\nabla f(x)\|_2^2\right] - \|\nabla f(x)\|_2^2$$

Summary:

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

2. Lipschitzness: $|f(x) - f(y)| \leq L \|x-y\|_2$

3. Smoothness: $\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x-y\|_2$

Upper-Bound:

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

4. Monotonicity of GD given Smooth:

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

Proof: GD + S. Upper Bound

$f(x_k) \leq f(x_k) \leq f(x_i) \quad \forall \; i \leq k$

5. Convex Functions:

$$f\left(\frac{u+v}{2}\right) \leq \frac{f(u) + f(v)}{2}$$

$$f[\lambda u + (1-\lambda) v] \leq \lambda f(u) + (1-\lambda) f(v)$$

* $f(u) + \langle \nabla f(u), v-u \rangle \leq f(v)$ •

6. PROVE CONVEX:

* Remove positive values multiplied.
* If sum, prove individually convex.
* If $\langle \theta, x_i \rangle$, prove for just x_i.
* Know e^x, x^2, $(x-a)^2$, $|x|$ are convex.

7. GD CONVERGENCE: Convex + Smooth

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

Proof: Monotonicity - ①

Convex Definition (x_k, x_*) - ②

① + ② + GD Replace $\nabla f(x)$ + V2 + Sum trick

$f(x_k) \leq f(x_i) \quad \forall \; i \leq k$

8. NAGD CONVERGENCE: Convex + Smooth

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

9. SGD: $x_{t+1} = x_t - \eta G(x_t)$

$E[G(x)] = \nabla f(x)$, $\text{var}(G(x)) \leq \sigma^2$

10. MONOTONICITY (Smooth):

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

11. CONVERHENCE:

$$E[f(\tilde{x}_k)] \leq f(x_*) + \frac{\|x_0 - x_*\|_2^2}{2\eta k} + \eta\sigma^2$$

12. SGD : ↓ Iteration time

PGD: Constrained Optimization, new point outside convex set.

Project it.

$$x_{t+1} = \text{Proj}_C(x_t - \eta \nabla f(x_t))$$

=> VECTOR PROPERTIES:

V1: $\langle x, x \rangle = \|x\|_2^2$

V2: $2\langle u, v \rangle = \|u\|^2 + \|v\|^2 - \|u - v\|^2$

V3: $|\langle u, v \rangle| \leq \|u\| \cdot \|v\|$ [Cauchy - Schwarz)

HOMEWORK

$\exists \geq \|\nabla f(w)\|$

- $f(w_{i+1}) \leq f(w_i) - \frac{\eta}{2} \|\nabla f(w_i)\|^2$
- $f(b) \leq f(a) + \langle \nabla f(a), b-a \rangle + \frac{\beta}{2} \|b-a\|_2^2$
- $w_{i+1} = w_i - \eta \nabla f(w_i)$

$$f(w_{k+1}) \leq f(w_0) - \frac{1}{2\beta} \left(\sum_{i=0}^{k} \|\nabla f(w_i)\|_2^2 \right.$$

$$\sum_{i=0}^{k-1} \|\nabla f(w_i)\|_2^2 \leq 2\beta \left(f(w_0) - f(w_k) \right)$$

$$\leq 2\beta \left(f(w_0) - f(w_*) \right)$$

We have:

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

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

$$\rightarrow f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|_2^2$$

Monotonicity - ①

Convex Definition (x_k, x_*) - ②

① + ② + GD Replace $\nabla f(x)$ + V2 + Sum trick

$f(x_k) \leq f(x_i) \quad \forall \; i \leq k$

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

$$f(x^*) \geq f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle + \frac{\alpha}{2} \|x^* - x_k\|_2^2$$

$$f(x_{k+1}) \leq f(x^*) + \beta \langle x_k - x_{k+1}, x_k - x^* \rangle - \frac{\alpha}{2} \|x^* - x_k\|_2^2 - \frac{\beta}{2} \|x_k - x_{k+1}\|_2^2$$

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

$$+ \frac{\beta}{2}\|x_k - x^*\|_2^2 - \frac{\beta}{2}\|x^* - x_{k+1}\|^2$$

$$+ \frac{\alpha}{2}\|x^* - x_k\|_2^2 - \frac{\beta}{2}\|x_k - x_{k+1}\|_2^2$$

$$f(x_{k+1}) \leq f(x^*) + \frac{(\beta-\alpha)}{2}\|x_k - x^*\|_2^2$$

$$- \frac{\beta}{2}\|x^* - x_{k+1}\|^2$$

$$f(x_{k+1}) \geq f(x^*) + \underbrace{\langle \nabla f(x^*), x_{k+1} - x^* \rangle}_{0} + \frac{\alpha}{2}\|x_{k+1} - x^*\|_2^2$$

$$\geq f(x^*) + \frac{\alpha}{2}\|x_{k+1} - x^*\|_2^2$$

$$f(x^*) + \frac{\alpha}{2}\|x_{k+1} - x^*\|_2^2 \leq f(x^*) + \frac{(\beta-\alpha)}{2}\|x_k - x^*\|$$

$$- \frac{\beta}{2}\|x_{k+1} - x^*\|_2^2$$

$$\frac{(\alpha+\beta)}{2} \|x_{k+1} - x^*\|_2^2 \leq \frac{(\beta-\alpha)}{2} \|x_k - x^*\|_2^2$$

$$\frac{\beta - \alpha}{\beta}$$