Subject Part II Optimization & Linear Algebra

Attendants 2/A_Q2

Date /

$$\theta^* \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} l(f_\theta(x_i), y_i)$$

where $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $f_i(\theta) := l(f_\theta(x_i), y_i)$

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$$

We consider the sequence (θ_t) defined with the stochastic update rule.

Q1. In general, is θ^* unique?

e.g. $f_\theta(x_i) = w^T x_i + b \qquad \theta = \{w, b\}$

let loss function

$$f_i(\theta) = l(f_\theta(x_i), y_i) = (y_i - f_\theta(x_i))^2$$

$$= (y_i - b)^2 - 2w^T x_i (y_i - b) + (w^T x_i)^2$$

$$\nabla_w f_i(\theta) = 0 - 2x_i(y_i - b) + 2(w^T x_i) x_i$$

$$\nabla_w^2 f_i(\theta) = 2 x_i x_i$$

$$f_i(\theta) = (y_i - w^T x_i - b)^2$$

Assume θ^* is unique minimum,

Take $u \in \mathbb{R}^p$ such that

$$u^T x_i = 0$$

orthogonal vector

(as long as $p > 1$ this is always possible)

if $\theta^* = \{w^*, b^*\}$ then

$$\theta' = \{w^* + u, b^*\} \quad \theta' \neq \theta^*$$

$$f_i(\theta^*) = (y_i - (w^*)^T x_i - b^*)^2$$

$$f_i(\theta') = (y_i - (w^* + u)^T x_i - b^*)^2$$

$$= (y_i - (w^*)^T x_i - b^*)^2 \text{ since}$$

$$u x_i = 0$$

$$f_i(\theta^*) = f_i(\theta')$$

Our assumption cannot hold we have more than 1

Subject

Date 3-A

Attendants

Q2. Show the necessary condition for θ^* to be minimizer of (1) is $\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\theta^*) = 0$

Is this condition sufficient?

$$\theta^* \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i), y_i) \quad \cdots (1)$$

DL Book 線形回帰 (linear Regression)

e.g. $\hat{y} = w^T x$, $w \in \mathbb{R}^n$ param

linear Regression の場合

$$MSE_{train} = \frac{1}{m} |\hat{y}_{train} - y_{train}|_2^2$$

最適性条件 x^*が unconstrained minimization problem の local minima で continuous differentiable

一次の必要条件 $\nabla f(x^*) = 0$

二次の必要条件 $\nabla^2 f(x^*)$ is positive semi-definite.

Subject　　　　　　　　　　　　Date 3-B

Attendants

$f:\mathbb{R}^n \to \mathbb{R}$ が x^* の近傍で
continuously differentiable
for 2-times

2次の十分条件

x^* が $\nabla f(x^*) = 0$ かつ
$\nabla^2 f(x^*)$ が positive definite

Q2, cont...

if l is convex, (as additional assumption)
local optimal = global optimal

$$\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\theta^*) = 0 \iff \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta}\left(l(f_\theta(x_i), y_i)\right) = 0$$

is necessity condition of

$$\theta^* \in \underset{\theta\in\mathbb{R}^d}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i), y_i)$$

(maybe also sufficient condition)

Subject

Date 3 - C

Attendants

Q2 cont ...

otherwise there are no additional constrains (assumption)

we assume loss as MSE.

if $\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\theta^*) \neq 0$

we use iterative method of stochastic gradient

$$\theta' = \theta^* - \eta \nabla f_i(\theta^*)$$

where η is stepsize and enough small.

$$f_i(\theta^*) > f_i(\theta')$$

this cannot hold assumption that $\theta^* \in \underset{\theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$

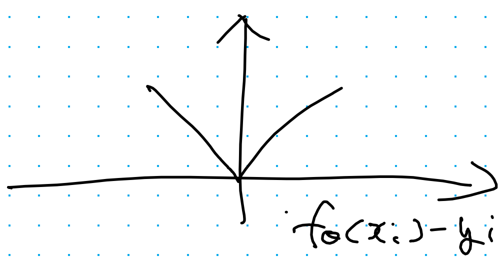
$$z = f_{\theta^*}(x_i) = y_i$$

then

$$\frac{d}{dz} l(y_i, z)\Big|_{z=y_i} = 0 \qquad = \nabla_\theta f_i(\theta^*)$$

This condition is sufficient.

Q3_ Provide an example of functions f_i where $\nabla f_i(\theta^*) = 0$, $\forall i \in \{1, \cdots n\}$ is not true.



$$f_i(\theta) = |f_\theta(x_i) - y_i|$$

which holds assumption

$$\left(\begin{array}{l} f_i(\theta) = l(f_\theta(x_i), y_i) \geq 0 \\ f_i(\theta^*) = 0 \end{array} \right)$$

but cannot hold

$$\nabla f_i(\theta^*) \neq 0$$

Subject Date 5

Attendants

Q4, In the context of classification for deep learning with $\ell(\hat{y}, y) = (\hat{y} - y)^2$ why is this assumption reasonable?

Assumption $i \in \{1, \dots, n\}$

$f_i(\theta) \geq 0, \ \forall \theta \in \mathbb{R}, \ f_i(\theta^*) = 0,$

$\nabla_\theta f_i(\theta^*) = 0$

wide enough

$\Downarrow$

we can interporate dataset

$\Updownarrow$

Universal Approximation Theorem

(普遍性定理)

<from Youtube Private>

square loss is valid

cuz $\forall x \quad x^2 \geq 0$

Subject Q4 Discussion

Date 6

Attendants

$\exists \theta^*$ s.t $f_i(\theta^*) = 0 \quad \forall i = 1, \dots, n$

$\Rightarrow$ interporation (zero train error)

$\downarrow$

possible with NN (at least 2 layers)
if we make them wide enough

$$\nabla f_i(\theta) = \nabla_\theta l(y_i, f_\theta(x_i))$$

$$= l'(y_i, f_\theta(x_i)) \cdot \nabla_\theta f_\theta(x_i)$$

then $\theta = \theta^*$ $\qquad$ 0 $\qquad$ vector (we don't know)

$$\nabla f_i(\theta^*) = \underbrace{l'(y_i, f_{\theta^*}(x_i))}_{=0} \underbrace{\nabla_\theta f_\theta(x_i)}_{\text{vector}}$$

$$\boxed{\frac{\partial}{\partial z} l(y_i, z)}\Big|_{z = f_{\theta^*}(x_i)} = \frac{\partial}{\partial z}(y_i - z)^2 \Big|_{z = f_{\theta^*}(x_i)}$$

when we use

$$= 2(z - y_i)\Big|_{z = f_{\theta^*}(x_i) - y_i} = 0$$

square loss, and assume interporation

which hold Assumption 1.

based on
wide-enough NN
$f_i(\theta^*) = 0$

Subject

Date 7

Attendants

Simple One

Gradient exists everywhere,

if grad ≠ 0, move θ which minimize L

then grad should have to zero.

Q5

Assumption 2 In addition to Assumption 1,

$$f(\theta) \geq f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + \frac{\mu}{2} \|\theta - \theta'\|^2$$

(strong convex)

$$\| f_i(\theta') - f_i(\theta) \leq L \|\theta' - \theta| \quad \forall_i \in \{1, \cdots, n\}$$

(Lipschitz continuity)

A iff (if and only if) B

BのときAが成立（Bのときしか成立しない

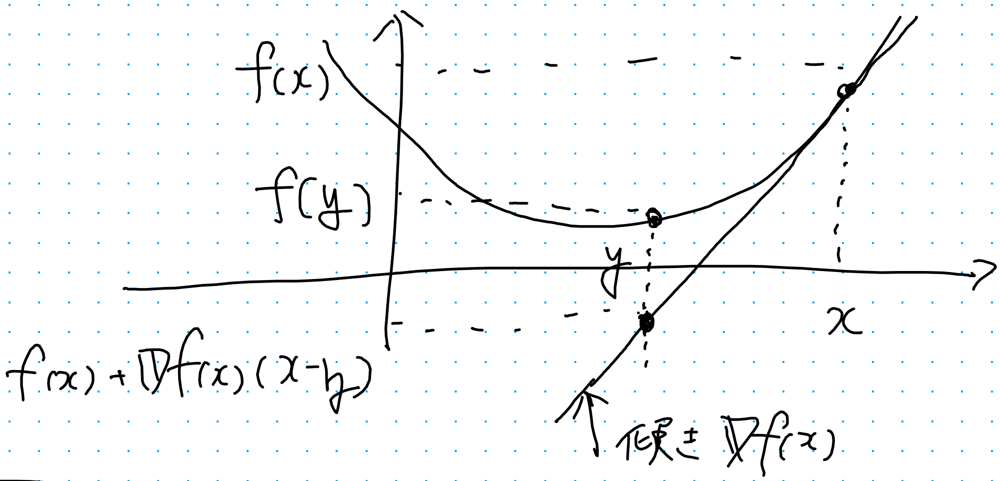
（必要十分条件）

Subject IFT 6085 L-2 Date 8

Attendants

Lemma 2 First Order condition for convexity

$$f(y) \geq f(x) + \nabla f(x)^T (x-y)$$



Lemma 4 Second order condition for conectivity

A twice differentiable function f is convex iff dom(f) is a convex set and $\forall x \in$ dom(f)

$$\nabla^2 f(x) \succeq 0$$

$\Rightarrow$ If f is convex, then by Lemma 2

$\forall x \in$ dom(f), $h > 0$

$$f(x+h) - f(x) \geq \nabla f(x)(x+h-x)$$

$$f(x) - f(x+h) \geq \nabla f(x+h)(x-x-h)$$

Subject Date 9

Attendants

$$f(x+h) - f(x) \geq \nabla f(x)(x+h-x)$$
$$f(x) - f(x+h) \geq \nabla f(x+h)(x-x-h)$$

Hence

$$\nabla f(x+h) \geq \frac{f(x+h) - f(x)}{h} \geq \nabla f(x)$$

which means, ∇f is non decreasing

hence $\nabla^2 f$ is non negative

$\Leftarrow$ Suppose f'' is nonnegative — $\boxed{\text{Hessian } \nabla^2 f \succeq 0}$

By Taylor Theorem of Order 2

we have $\forall x, y \in \mathrm{dom}(f)$, $\exists z \in [x, y]$

such ~~that~~

$$f(y) = f(x) + f'(x)(y-x) + \underbrace{\frac{1}{2} f''(z)(y-x)^2}_{\text{remove } \geq 0}$$

Therefore $f(y) \geq f(x) + f'(x)(y-x)$

which has the first order condition of convexity

Subject

Date 10

Attendants

Applying $\underbrace{\text{Lemma 3}}$,

$$\boxed{\begin{array}{l}\text{A function } f \text{ is convex iff } dom(f) \text{ is a} \\ \text{convex set and } \forall x \in dom(f),\ u \in \mathbb{R} \\ \text{the real function } g_{x,u} \text{ is convex, where} \\ \qquad g_{x,u} := f(x+tu) \\ \text{for all } t \in \mathbb{R}, \text{ such that } x+tu \in dom(f)\end{array}}$$

f is convex iff $\forall x \in dom(f)$, $u \in \mathbb{R}^d$,
the real function $g_{x,u}$, defined by

$$g_{x,u}(t) = f(x+tu) \text{ for } t$$

such that $x+tu \in dom(f)$ is convex

$$g''_{x,u}(t) = u^T \underbrace{\nabla^2 f(x+tu)}\, u \geq 0$$

Since, this is for all $u \in \mathbb{R}$

$x+tu \in dom(f)$

by taking $t=0$

$$g''_{x,u}(0) = u^T \nabla^2 f(x)\, u \geq 0$$

we can see that it is equivalent $\nabla^2 f(x) \succeq 0$

Subject

Date / /

Attendants

The lemma states that for a twice differentiable function to be convex, its Hessian must be positive semi-definite

Go Back to QT Autumn 2021 Part 2

From Assumption 2,

$$f(\theta) \geq f(\theta') + \langle \nabla f(\theta'), \theta - \theta' \rangle + \frac{\mu}{2} \|\theta - \theta'\|^2$$

as μ-strong convexity

$\Rightarrow f(\theta) = g(\theta) + \frac{\mu}{2}\|\theta\|^2$ for some convex g

$$\Rightarrow \nabla^2 f(\theta) = \underbrace{\nabla^2 g(\theta)}_{} + \mu I$$

convex functions have all ≥ 0 eigen values

$$\Rightarrow \nabla^2 f(\theta) \succeq 0 + \mu I$$

Subject Date 12

Attendants

$\boxed{Q6}$

$$\| \underbrace{\theta_{t+1}} - \theta^* \|^2 = \| \underbrace{\theta_t - \eta \nabla f_{i_t}(\theta_t)} - \theta^* \|^2$$

$$= \|\theta_t - \theta^*\|^2 - 2\langle \theta_t - \theta^*, \eta \nabla f_{i_t}(\theta_t) \rangle$$

$$+ \eta^2 \|\nabla f_{i_t}(\theta_t)\|^2$$

$\square$

$\boxed{Q7}$ show

$$\mathbb{E}_{i_t}\left[\|\theta_{t+1} - \theta^*\|^2 \,\middle|\, \theta_t\right] \le (1 - \eta\mu)\|\theta_t - \theta^*\|^2$$

$$+ \eta\left(\eta \mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(\theta_t)\|^2 \,\middle|\, \theta_t\right] - 2\mathbb{E}_{i_t}\left[f_{i_t}(\theta_t) \,\middle|\, \theta_t\right]\right)$$

using Q6

$$\mathbb{E}_{i_t}\left[\|\theta_{t+1} - \theta^*\|^2 \,\middle|\, \theta_t\right]$$

$$= \mathbb{E}_{i_t}\left[\|\theta_t - \theta^*\|^2 \,\middle|\, \theta_t\right] + \eta\Big(\mathbb{E}_{i_t}\left[\eta \|\nabla f_{i_t}(\theta_t)\|^2 \,\middle|\, \theta_t\right]$$

$$- 2\mathbb{E}_{i_t}\left[\nabla f_{i_t}(\theta_t)(\theta_t - \theta^*) \,\middle|\, \theta_t\right]\Big)$$

Subject Date 13

Attendants

$$\underbrace{\mathbb{E}_{it}\left[\|\theta_t-\theta^*\|^2 \mid \theta_t\right]}_{\| \atop \|\theta_t-\theta^*\|^2} + \eta\left(\mathbb{E}_{it}\left[\eta\|\nabla f_{it}(\theta_t)\|^2 \mid \theta_t\right] - 2\,\mathbb{E}_{it}\left[\nabla f_{it}(\theta_t)(\theta_t-\theta^*) \mid \theta_t\right]\right)$$

$$\wedge|$$

$$\boxed{\mathbb{E}_{it}\left[\nabla f_{it}(\theta_t) \mid \theta_t\right] \quad + \mathbb{E}_{it}\left[\nabla f_{it}(\theta_t)(\theta_t-\theta^*) \mid \theta_t\right]}$$

because, let assume $\nabla f_{it}(\theta_t)$ is positive

$X = \nabla f_{it}(\theta_t)$

$$X(\theta_t-\theta^*) = X\theta_t - X\theta^*$$

$$= X + X\theta_t - X\theta^* - X$$

$$= X + X(\theta_t - \theta^* - 1)$$

$$\le X + X(\theta_t-\theta^*)$$

drop -1

Subject

Date 14

Attendants

$$\mathbb{E}_{i_t}[\|\theta_{t+1}-\theta^*\|^2|\theta_t]$$

$$\leq \mathbb{E}_{i_t}[\|\theta_t-\theta^*\|^2|\theta_t] + \eta\Big(\eta\,\mathbb{E}_{i_t}[\|\nabla f_{i_t}(\theta_t)\|^2|\theta_t]$$

$$-2\mathbb{E}_{i_t}[f_{i_t}(\theta_t)|\theta_t]\Big)$$

$$-2\eta\,\mathbb{E}_{i_t}[\underbrace{\nabla f_{i_t}(\theta_t)(\theta_t-\theta^*)}|\theta_t]$$

from strong convex

$$f(\theta^*) \geq f(\theta_t) + \underbrace{\nabla f_{i_t}(\theta_t)(\theta^*-\theta_t)} + \frac{\mu}{2}\|\theta^*-\theta_t\|^2$$

$$-\frac{\mu}{2}\|\theta_t-\theta^*\|^2 \leq \underbrace{f(\theta_t)}-\underbrace{f(\theta^*)}_{\substack{\shortparallel \\ 0}} + \nabla f_{i_t}(\theta_t)(\theta^*-\theta_t)$$

from Assumption 1

$f_i(\theta) \geq 0$

$$\leq \nabla f_{i_t}(\theta_t)(\theta^*-\theta_t)$$

then $\nabla f_{i_t}(\theta_t)(\theta_t-\theta^*) \leq \frac{\mu}{2}\|\theta_t-\theta^*\|^2$

Subject

Date 15

Attendants

$$\mathbb{E}_{i_t}[\|\theta_{t+1}-\theta^*\|^2|\theta_t]$$

$$\leq \mathbb{E}_{i_t}[\|\theta_t-\theta^*\|^2|\theta_t] + \eta\Big(\eta\,\mathbb{E}_{i_t}[\|\nabla f_{i_t}(\theta_t)\|^2|\theta_t]$$

$$-2\,\mathbb{E}_{i_t}[f_{i_t}(\theta_t)|\theta_t]\Big)$$

$$-2\eta\,\mathbb{E}_{i_t}[\nabla f_{i_t}(\theta_t)(\theta_t-\theta^*)|\theta_t]$$

$$\leq \underbrace{\mathbb{E}_{i_t}[\|\theta_t-\theta^*\|^2|\theta_t]}_{\|\theta_t-\theta^*\|^2} + \eta\Big(\eta\,\mathbb{E}_{i_t}[\|\nabla f_{i_t}(\theta_t)\|^2|\theta_t]$$

$$-2\,\mathbb{E}_{i_t}[f_{i_t}(\theta_t)|\theta_t]\Big)$$

$$-2\eta\times\frac{\mu}{2}\|\theta_t-\theta^*\|^2$$

$$=(1-\mu\eta)\|\theta_t-\theta^*\|^2+\eta\Big(\eta\,\mathbb{E}_{i_t}[\|\nabla f_{i_t}(\theta_t)\|^2|\theta_t]$$

$$-2\,\mathbb{E}_{i_t}[f_{i_t}(\theta_t)|\theta_t]\Big) \quad \square$$

Subject | Date | 16

Attendants $\boxed{Q_6 f}$ show following formula

$$\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right] \leq (1-\eta\mu+L\eta^2)^t \|\theta_0-\theta^*\|^2$$

$$\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right] = \mathbb{E}\left[\mathbb{E}_{i_t}\left[\|\theta_t-\theta^*\|^2 \mid \theta_t\right]\right]$$

use Q7

$$\leq \mathbb{E}\left[(1-\eta\mu)\|\theta_t-\theta^*\|^2+\eta\left(\eta\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(\theta_t)\|^2 \mid \theta_t\right] - 2\mathbb{E}_{i_t}\left[f_{i_t}(\theta_t) \mid \theta_t\right]\right)\right]$$

$$= (1-\eta\mu)\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right] + \underbrace{\eta^2\mathbb{E}\left[\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(\theta_t)\|^2 \mid \theta_t\right]\right]}_{②}$$

$$\underbrace{-2\mathbb{E}\left[\mathbb{E}_{i_t}\left[f_{i_t}(\theta_t) \mid \theta_t\right]\right]}_{③}$$

$$② \; \mathbb{E}\left[\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(\theta_t) - \underbrace{\nabla f_{i_t}(\theta^*)}_{=0}\|^2 \mid \theta_t\right]\right] \leq \mathbb{E}\left[\mathbb{E}_{i_t}\left[L\|\theta_t-\theta^*\|^2 \mid \theta_t\right]\right]$$

(from Assumption 7) ← from Lipschitz constant

Subject

Date 17

Attendants

③ $\mathbb{E}\left[\mathbb{E}_{i_t}\left[f_{i_t}(\theta_t) \mid \theta_t\right]\right] \geq 0$ from Assumption 1

$$(1-\eta\mu)\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right] + \eta^2 \underbrace{\mathbb{E}\left[\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(\theta_t)\|^2 \mid \theta_t\right]\right]}_{\leftarrow\ \mathbb{E}\left[2L\|\theta_t-\theta^*\|^2\right]\times\eta^2}$$

$$-2\underbrace{\mathbb{E}\left[\mathbb{E}_{i_t}\left[f_{i_t}(\theta_t) \mid \theta_t\right]\right]}_{\substack{\geq \\ 0}}$$

$\Leftrightarrow -2\cdots < 0$

So, $\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right] \leq 1-\eta\mu\,\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right] + 2L\eta^2\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right] + 0$

$$= \left((1-\eta\mu+2L\eta^2\right)\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right]$$

repeat for t steps (unroll recursion)

$$\mathbb{E}\left[\|\theta_t-\theta^*\|^2\right] \leq (1-\mu\eta+2L\eta^2)^t\,\mathbb{E}\left[\|\theta_0-\theta^*\|^2\right]$$