LEC2 DSAS103

- Gradient Method
- Linear Regression

$\boxed{\text{Smooth Optimization Case}}$

1. Gradient Method

a) problem setting

$$\min_{x \in \mathbb{R}^n} f(x) \qquad f \text{ is differentiable}$$

Necessary Condition of Optimality

$$\{\text{global minimizer}\} \subseteq \boxed{\{\text{local minimizer}\} \subseteq \{\text{stationary point}\}}$$

$\nabla f(x) = 0$

optimality (necessary)

$$\boxed{x^* \in \text{local optima}} \Rightarrow \begin{cases} ① \ \nabla f(x^*) = \underline{0} \\ ② \ H_f(x^*) \in PSD \end{cases}$$

b) General Iterative Framework

$$x^{k+1} = x^k + \alpha_k d^k$$

$$\begin{cases} \alpha_k \to \text{step length} \\ d_k \to \text{search direction} \end{cases}$$

Rmk: it is flexible to choose search direction

- ① Newton Direction
- ② Gradient Descent Direction

etc.

Defn: d^k is $\boxed{\text{descent direction}}$ at point x^k

$$\Leftrightarrow \quad \nabla f(x^k)^T d^k < 0$$

Pf from Taylor Expansion, $\exists\, \delta$ small efficient such that

$$f(x^k + \delta d^k) < f(x^k)$$

$$f(x^k + \delta d^k) = f(x^k) + \delta \cdot \nabla f(x^k)^T d^k + o(\delta)$$

Corallary: $d^k = -\nabla f(x^k)$ is always a descent direction

for arbitrary function f, point x^k

↓

General Choice of Descent Direction

→ To check the descent direction d^k at point x^k.

just check $\nabla f(x^k)^T d^k$ $\begin{cases} \geq 0 \Rightarrow \text{not descent direction} \\ < 0 \Rightarrow \text{descent direction} \end{cases}$

c) GD Framework (Steepest Descent)

while $\|\nabla f(x^k)\| > \varepsilon$: → STOP CONDITION

① determine search direction $d_k = -\nabla f(x^k)$

② determine step length α_k

method list $\begin{cases} \text{exact line search} \\ \text{backtracking line search} \\ \text{fix some step length} \end{cases}$

Inexact Method

Some conditions like $\begin{cases} \text{Wolfe cond.} \\ \text{Armijo cond.} \end{cases}$

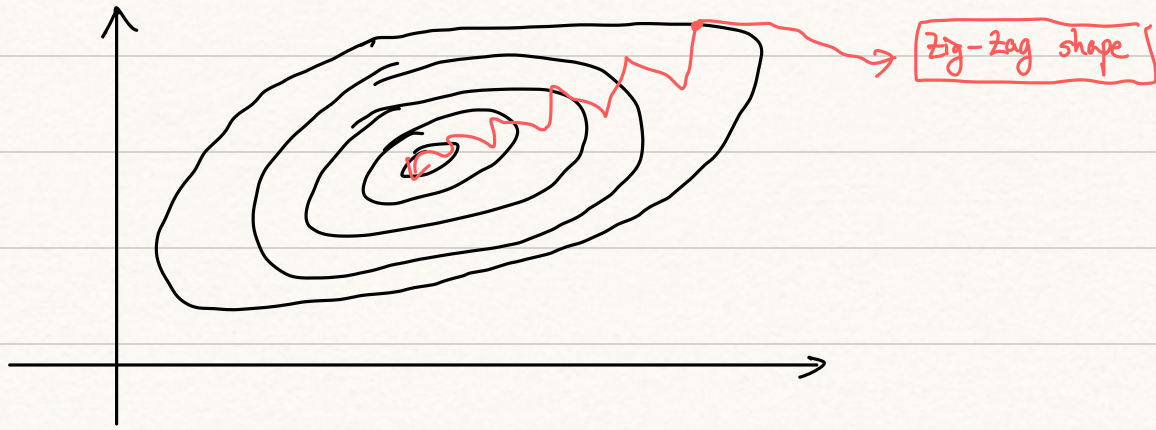
start from big step length & shrink gradually

$x^{k+1} = x^k + \alpha_k d_k$

$k = k+1$

d) $\boxed{\text{Issue}}$ for Gradient-Based Method

$\boxed{\text{Zig-Zag Problem}}$



---

2. Linear Reg (ML Architecture)

$\boxed{\text{Model}}$ $\mathcal{H}_{linear} := \{ f : f(x) = \beta_0 + \beta_1 x \}$

$\Rightarrow$ We want to learn the optimal function $\hat{f} \in \mathcal{H}_{linear}$

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, f(x_i))$$

$$= \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{1}{2} (y_i - f(x_i))^2$$ (if we choose MSE loss)

$$= \underset{\beta_0, \beta_1}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$:= \underset{\beta}{\operatorname{arg\,min}} \ J(\beta)$$

$$J(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$

$$= \frac{1}{2}\| y - X\beta \|_2^2 \qquad X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n\times(d+1)}$$

① Solve for $\nabla J(\hat{\beta}) = X^T X \hat{\beta} - X^T y = 0$

$$\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

② Iterative Method like $\boxed{GD}$

$$\beta^{(k+1)} = \beta^{(k)} - \alpha_k \nabla J(\beta^{(k)})$$

$$\nabla J(\beta) = X^T X\beta - X^T y$$

$$= X^T(X\beta - y)$$

$$= \begin{pmatrix} x_{10} & \cdots & x_{n0} \\ \vdots & & \vdots \\ x_{1d} & \cdots & x_{nd} \end{pmatrix} \begin{pmatrix} x_1^T\beta - y_1 \\ \vdots \\ x_n^T\beta - y_n \end{pmatrix}$$

$$\boxed{X^T = (x_1, \dots, x_n) \in \mathbb{R}^{(d+1)\times n}}$$

$$\Leftrightarrow \beta_j^{(k+1)} = \beta_j^{(k)} + \alpha_k \sum_{i=1}^{n}(y_i - \beta^{(k)T}x_i)\, x_{ij} \qquad j = 0, 1, \dots, d$$

$$\boxed{x_i = \begin{bmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{id} \end{bmatrix} \in \mathbb{R}^{d+1}}$$

3. LR with Statistics Convention

$\boxed{\text{Gauss-Markov Condition}}$

$\boxed{\text{Model}}$

$$y = \beta^T x + \varepsilon \quad \text{where} \quad \varepsilon \sim N(0, \sigma^2)$$

$$\Updownarrow$$

$$\boxed{y \sim \text{Gaussian}(\beta^T x, \sigma^2)}$$

$\boxed{\text{Estimate Parameter}} \leftarrow \boxed{\text{MLE}}$

$\downarrow$

$\boxed{\text{Property of Estimator}} \longrightarrow \begin{cases} \text{var}[\cdot] \ \cdots \\ \mathbb{E}[\cdot] \ \cdots \end{cases}$

$\downarrow$

- Hypothesis Testing
- Goodness of Fit
- Model Diagnosis