

## Basics

- Linear hypothesis:  $h(x) = \theta_1 x + \theta_0$ ,  $\theta_i (i = 1, 2 \text{ for 2D cases})$ .
- cost function:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2, \quad h_{\theta}(x) = \sum_{i=0}^n \theta_i x_i = \theta^T x$$

- best choice for  $\theta = \underset{\theta}{\operatorname{argmin}} J(\theta)$

## Gradient

### Def. Directional Derivative

The directional derivative of function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  in the direction  $u$  is

$$\nabla_u f(x) = \lim_{h \rightarrow 0} \frac{f(x + hu) - f(x)}{h}$$

- When  $u$  is the  $i$ -th standard unit vector  $e_i$ , then  $\nabla_u f(x) = f'_i(x) = \frac{\partial f(x)}{\partial x_i}$ .
- For any  $n$ -dimensional vector  $u$ , the directional derivative of  $f$  in the direction of  $u$  can be represented as  $\nabla_u f(x) = \sum_{i=1}^n f'_i(x) \cdot u_i$ .

– Proof.  $\Rightarrow$

$$\begin{aligned} \text{let } g(h) &= f(x + hu) \\ \nabla_u f(x) &= g'(0) = \lim_{h \rightarrow 0} \frac{f(x + hu) - g(0)}{h} \\ \therefore g'(h) &= \sum_{i=1}^n f'_i(x) \frac{d}{dh}(x_i + hu_i) = \sum_{i=1}^n f'_i(x) u_i \\ \text{let } h &= 0 \therefore \nabla_u f(x) = \sum_{i=1}^n f'_i(x) u_i \end{aligned}$$

### Def. Gradient

The gradient of  $f$  is a vector function  $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  defined by

$$\begin{aligned} \nabla f(x) &= \sum_{i=1}^n \frac{\partial f}{\partial x_i} e_i \\ \Rightarrow \nabla f(x) &= \left[ \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right] \end{aligned}$$

- $\nabla_u f(x) = \nabla f(x) \cdot u = \|\nabla f(x)\| \cos a$  Where  $u$  is a unit vector.
- When  $u = \frac{\nabla f(x)}{\|\nabla f(x)\|}$  such that  $a = 0$ , we have the maximum directional derivative of  $f$ .

## Gradient Descent (GD) Algorithm

Algorithm.

```
Given a starting point \theta in dom J
while convergence criterion is satisfied
    Calculate gradient \nabla J(\theta)
    Update \theta \leftarrow \theta - \alpha \nabla J(\theta)
```

$\theta$  Is usually initialized randomly, and  $\alpha$  is so-called learning rate.

- For linear regression,

$$\begin{aligned}\theta_j &\leftarrow \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}, \quad \forall j = 0, 1, \dots, n, \quad x_0^{(i)} = 1 \\ \frac{\partial J(\theta)}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)})^2 \\ &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \sum_{i=1}^m \left( \sum_{j=0}^n \theta_j x_j^{(i)} - y^{(i)} \right)^2 \\ &= \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)}) x_j^{(i)}\end{aligned}$$

- Another commonly used form  $J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$ .
- $m$  is introduced to scale the objective function to deal with differently sized training set.

## Matrix Derivatives

- The derivative of  $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$  with respect to  $A$  is defined as:

$$\nabla f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \dots & \frac{\partial f}{\partial A_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{m1}} & \dots & \frac{\partial f}{\partial A_{mn}} \end{bmatrix}$$

### trace

$$\text{Def. } \text{tr} A = \sum_{i=1}^n A_{ii}$$

- $\text{tr} ABCD = \text{tr} DABC = \text{tr} CDAB = \text{tr} BCDA$
- $\text{tr} A = \text{tr} A^T, \text{tr}(A + B) = \text{tr} A + \text{tr} B, \text{tr}(aA) = a \cdot \text{tr} A$
- $\nabla_A \text{tr} AB = B^T, \nabla_{A^T} f(A) = (\nabla_A f(A))^T$
- $\nabla_A \text{tr} ABA^T C = CAB + C^T AB^T, \nabla_A |A| = |A| (A^{-1})^T$
- Funky trace derivative  $\nabla_{A^T} \text{tr} ABA^T C = B^T A^T C^T + BA^T C$

### Jacobian Matrix

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \dots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

### Hesse Matrix

$$G(x_0) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}_{x_0}$$

- $H(f) = J(\nabla f)$

### Revisiting Least Square with Matrix Form

$$X = \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix}, Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} \quad J(\theta) = \frac{1}{2} \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)})^2 = \frac{1}{2} (X\theta - Y)^T (X\theta - Y)$$

- Minimize  $J(\theta) = \frac{1}{2} (Y - X\theta)^T (Y - X\theta)$

$$\begin{aligned} \nabla_{\theta} J(\theta) &= \nabla_{\theta} \frac{1}{2} (Y - X\theta)^T (Y - X\theta) \\ &= \frac{1}{2} \nabla_{\theta} \text{tr}(Y^T Y - Y^T X\theta - \theta^T X^T Y + \theta^T X^T X\theta) \\ &= \frac{1}{2} \nabla_{\theta} \text{tr}(\theta^T X^T X\theta) - X^T Y \\ &= \frac{1}{2} (X^T X\theta + X^T X\theta) - X^T Y \\ &= X^T X\theta - X^T Y \end{aligned}$$

- Theorem. **Normal Equation**

The matrix  $A^T A$  is invertible if and only if the columns of  $A$  are linearly independent. In this case, there exists only one least-squares solution.

$$\theta = (X^T X)^{-1} X^T Y$$