Basics

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

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

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

Gradient

Def. **Directional Derivative**

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

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

- 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.⇒ $\begin{array}{l} \text{let } g(h) = f(x + hu) \\ \nabla_u f(x) = g'(0) = \lim\limits_{h \to 0} \frac{f(x+hu) - g(0)}{h} \\ \because g'(h) = \sum_{i=1}^{n} f_i'(x) \frac{d}{dh}(x_i + hu_i) = \sum_{i=1}^{n} f_i'(x) u_i \\ \text{let } h = 0 \therefore \nabla_u f(x) = \sum_{i=1}^{n} f_i'(x) u_i \end{array}$

Def. **Gradient**

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

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

- $\nabla_u f(x) = \nabla f(x) \cdot u = ||\nabla f(x)|| \cos a$ Where u is a unit vector.
- When $u = \nabla f(x)$ 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 converence criterion is satisfied
  Calculate gradient \nabla J(\theta)
  Update \theta \leftarrow \theta - \alpha\nabla J(\theta)
```

θ Is usually initialized randomly, and α is so-called learning rate.

- For linear regression,

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

- 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:

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

trace

Def. $trA = \sum_{i=1}^{n} A_i i$

- $trABCD = trDABC = trCDAB = trBCDA$
- $trA = trA^T, tr(A+B) = trA + trB, tr(aA) = a \cdot trA$
- $\nabla_A trAB = B^T, \nabla_{A^T} f(A) = (\nabla_A f(A))^T$
- $\nabla_A trABA^TC = CAB + C^TAB^T, \nabla_A |A| = |A|(A^{-1})^T$
- Funky trace derivative $\nabla_{A^T} trABA^TC = B^TA^TC^T + BA^TC$

Jacobian Matrix

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

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

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

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

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

- 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.

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