Basics

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

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

- best choice for $\theta = \arg\min_{\theta}\ J(\theta)$

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,

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

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

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

- Theorem. **Normal Equation**

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

  $$\theta = (X^TX)^{-1}X^TY$$

test by rookie cui