Linear Regression — Normal Equation Method

1. Hypothesis

In linear regression, the prediction function is:

$$\hat{y} = \boldsymbol{w}^T \boldsymbol{x} + b$$

where:

- $\boldsymbol{x} \in \mathbb{R}^n$: feature vector
- $\boldsymbol{w} \in \mathbb{R}^n$: weight vector
- $b \in \mathbb{R}$: bias (intercept)

2. Vectorized Form

To include the bias term neatly, we augment the feature matrix with a column of ones:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \quad \theta = \begin{bmatrix} b \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

- $X \in \mathbb{R}^{m \times (n+1)}$: design matrix, with m training examples
- $\theta \in \mathbb{R}^{n+1}$: parameter vector (bias + weights)

Thus, the prediction for all training examples is:

$$\hat{y} = X\theta$$

3. Cost Function

We minimize the Mean Squared Error (MSE):

$$J(\theta) = \frac{1}{2m} \|X\theta - y\|^2$$

4. Derivation of the Normal Equation

To minimize $J(\theta)$, set its gradient to zero:

$$\nabla_\theta J(\theta) = \frac{1}{m} X^T (X\theta - y) = 0$$

$$\implies X^T X\theta = X^T y$$

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

5. Extracting Parameters

Once θ is computed:

- Bias: $b = \theta_0$ (first element)
- Weights: $\boldsymbol{w} = [\theta_1, \theta_2, \ldots, \theta_n]^T$

Summary

The Normal Equation gives a closed-form solution for linear regression parameters:

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

- Avoids iterative methods like gradient descent.
- Computationally expensive when n (number of features) is very large, due to matrix inversion cost $\mathcal{O}(n^3)$.