

Week 2: Single Variable Linear Regression

Ekarat Rattagan

July 28, 2025

Notations

Let a dataset $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ be a set of N pairs (x_i, y_i) , where $x_i \in \mathbb{R}^d$ is a feature vector (independent variable) and $y_i \in \mathbb{R}$ (dependent variable).

Function approximation or hypothesis (model):

$$f : X \rightarrow Y, \quad f : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^{n \times 1}$$

Model Representation

- For a single variable linear regression: $d = 1$
- Model: $h_{\theta}(x) = \theta_0 + \theta_1 x$
- θ : parameters, $\theta_j \in \mathbb{R}$, and $|\theta| = d + 1$
- Loss/Cost function (Mean Squared Error)

$$J(\theta_0, \theta_1) = \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i)^2$$

- Objective

$$\hat{\theta}_0, \hat{\theta}_1 = \arg \min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$$

Example

x	y
0	1
2	1
3	4

Two possible hypotheses:

$$h_{\theta}(x) = 3, \text{ and } \text{MSE} = 3$$

$$h_{\theta}(x) = 1 + x, \text{ and } \text{MSE} = 1.33$$

Solution approach

- 1 **Analytical approach:** Normal equation

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

- 2 **Iterative approach:** Gradient Descent

Normal Equation Issues (1/2)

- Problem 1: $(X^T X)$ non-invertible
 - Example: redundant features
- Solution:
 - SVD (Singular Value Decomposition)
 - Moore–Penrose Pseudo-inverse

Normal Equation Issues (2/2)

- Problem 2: Computational complexity $O(n^3)$ for large feature sets (Why?).
- Solution: Use iterative approach (Gradient Descent)

Iterative approach: (Batch) Gradient Descent

Hypothesis: $h_{\theta}(x) = \theta_0 + \theta_1 x$

Objective function (Mean Squared Error):

$$J(\theta_0, \theta_1) = \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i)^2$$

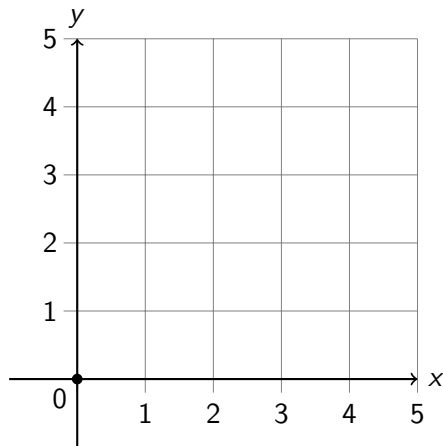
Goal:

$$\arg \min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$$

Batch Gradient Descent (GD) is a way to minimize an objective function $J(\theta)$ by updating θ in the opposite direction of the gradient of J , i.e., $\nabla J(\theta)$.

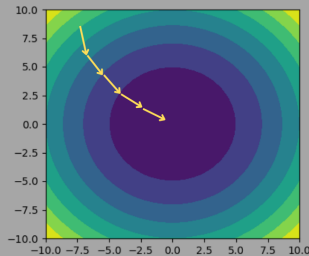
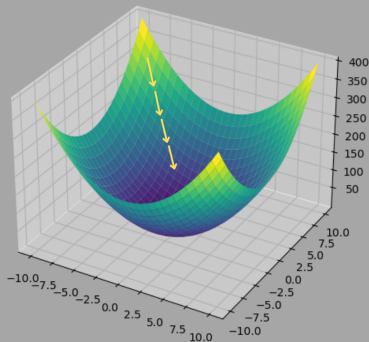
Relation between J and one θ

How J changes with θ :



Relation between J and two θ

Gradient Descent



Datamapu

Figure: Relation between J and θ_0, θ_1

Batch GD equation

Loop until convergence:

$$\theta_j := \theta_j - \eta \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1), \text{ where } j = 0, 1 \text{ and } \eta \text{ is a learning rate}$$

Take a look at $\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$. We get,

$$\theta_j := \theta_j - \eta \frac{\frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i)^2}{\partial \theta_j}, \quad j = 0, 1$$

$$\theta_j := \theta_j - \eta \frac{1}{N} \sum_{i=1}^N \frac{\partial (h_{\theta}(x_i) - y_i)^2}{\partial \theta_j}, \quad j = 0, 1$$

Update equation of Batch GD (I)

Loop until convergence:

$$\theta_j := \theta_j - \eta \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i) x_i, \quad j = 0, 1$$

Stopping criteria:

- Max iterations or
- $|MSE_t - MSE_{t+1}| < 0.001$

Update equation of Batch GD (II)

Loop until convergence:

$$\theta_0 := \theta_0 - \eta \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i)1$$

$$\theta_1 := \theta_1 - \eta \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x_i) - y_i)x_i$$

Limitations:

- Need to define η
- Slow for large datasets

Iterative VS Analytical approach

Iterative	Analytical
Need η Workable for large d Need feature scaling No invertibility issue $O(n^2)$	No need η Not workable for large d No feature scaling Invertibility issue $O(n^3)$

Exercise 1: Solve Both Approaches

x	y
2	12
5	9
1	6

- 1 Analytical: $\theta = (X^T X)^{-1} X^T y$
- 2 Iterative: Batch GD, define $\theta_0, \theta_1 = 0.1, \eta = 0.01, \#iter = 3$.

- Marill, K.A. (2004). Advanced statistics: linear regression, part I: simple linear regression. Academic emergency medicine, 11(1), 87-93.