

# Weeks 2 and 3: Optimization Problems in Machine Learning and Linear Regression via Gradient Descent

Gregory Glickert

February 6, 2026

## Introduction

This document explores fundamental optimization problems in machine learning:

- **Formulating the Mean as an Optimization Problem:** We derive the arithmetic mean by minimizing the sum of squared errors for a set of data points.
- **Optimization of a Quadratic Function Using Gradient Descent:** We minimize a quadratic cost function using the gradient descent algorithm, demonstrating its convergence properties.
- **Linear Regression via Gradient Descent:** We optimize a linear regression model using gradient descent for a simple dataset.

## Formulating the Mean as an Optimization Problem

We are given four data points:

$$x_1 = 2, \quad x_2 = 4, \quad x_3 = 6, \quad x_4 = 8.$$

### 1. Define the Parameter

Let

$$\theta \in \mathbb{R}$$

be a parameter representing a candidate estimate of the central value (mean) of the data.

### 2. Define the Cost Function

We define the objective (cost) function as the sum of squared errors:

$$J(\theta) = \sum_{i=1}^4 (x_i - \theta)^2.$$

Substituting the data values explicitly,

$$J(\theta) = (2 - \theta)^2 + (4 - \theta)^2 + (6 - \theta)^2 + (8 - \theta)^2.$$

### 3. Expand the Cost Function

$$\begin{aligned} J(\theta) &= (\theta^2 - 4\theta + 4) + (\theta^2 - 8\theta + 16) + (\theta^2 - 12\theta + 36) + (\theta^2 - 16\theta + 64) \\ &= 4\theta^2 - 40\theta + 120. \end{aligned}$$

### 4. Compute the First Derivative

We differentiate the cost function with respect to  $\theta$ :

Since  $J(\theta) = 4\theta^2 - 40\theta + 120$ ,

$$\frac{dJ}{d\theta} = \frac{d}{d\theta}(4\theta^2 - 40\theta + 120) = 8\theta - 40.$$

### 5. Find the Critical Point

Setting the derivative equal to zero,

$$8\theta - 40 = 0,$$

which yields

$$\theta^* = 5.$$

### 6. Final Result

$$\boxed{\theta^* = 5}$$

This value coincides with the arithmetic mean of the data:

$$\frac{2 + 4 + 6 + 8}{4} = 5.$$

## Optimization of a Quadratic Function Using Gradient Descent

We consider the problem of finding the minimum of the function

$$y = x^2 + 2x + 5.$$

This problem is solved by formulating it as an optimization problem and applying the gradient descent method.

## 1. Define the Optimization Problem

Let

$$\theta \in \mathbb{R}$$

be the optimization parameter.

Define the cost function:

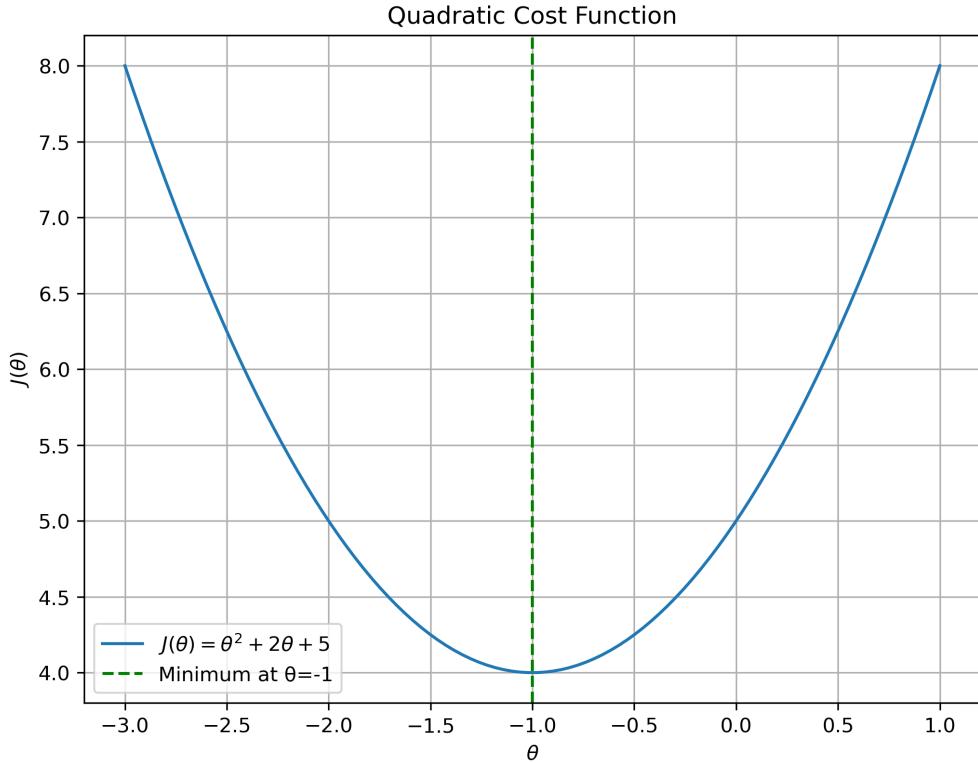
$$J(\theta) = \theta^2 + 2\theta + 5.$$

Our goal is to find the value of  $\theta$  that minimizes  $J(\theta)$ .

## 2. Plotting $J(\theta)$ vs. $\theta$

The function  $J(\theta)$  is a quadratic polynomial with a positive coefficient on  $\theta^2$ , so its graph is a parabola that opens upward and has a single global minimum.

From the graph, we can see that the minimum occurs at  $\theta = -1$  with  $J(-1) = 4$ . We are now going to setup and solve for these using gradient descent.



## 3. Gradient Descent Solution

### Compute the Gradient

Differentiate the cost function with respect to  $\theta$ :

Since  $J(\theta) = \theta^2 + 2\theta + 5$ ,

$$\frac{dJ}{d\theta} = \frac{d}{d\theta}(\theta^2 + 2\theta + 5) = 2\theta + 2.$$

## Gradient Descent Update Equation

The gradient descent update rule is given by:

$$\theta_{k+1} = \theta_k - \alpha \frac{dJ}{d\theta} \Big|_{\theta=\theta_k}.$$

Here,  $\alpha$  is the learning rate that determines the step size in each iteration, and  $\frac{dJ}{d\theta}$  is the gradient of the cost function, which points in the direction of the steepest ascent. Subtracting it moves us towards the minimum.

Substituting the gradient expression, we obtain:

$$\theta_{k+1} = \theta_k - \alpha(2\theta_k + 2).$$

## Choose Parameters

The gradient descent parameters are:

- Initial value:  $\theta_0 = 0$
- Learning rate:  $\alpha = 0.4$

## Iteration 1

Using the update equation,

$$\theta_1 = \theta_0 - 0.4(2\theta_0 + 2).$$

Substituting  $\theta_0 = 0$ ,

$$\theta_1 = 0 - 0.4(2) = -0.8.$$

## Iteration 2

$$\theta_2 = \theta_1 - 0.4(2\theta_1 + 2).$$

Substituting  $\theta_1 = -0.8$ ,

$$\theta_2 = -0.8 - 0.4(0.4) = -0.96.$$

## Iteration 3

$$\theta_3 = \theta_2 - 0.4(2\theta_2 + 2).$$

Substituting  $\theta_2 = -0.96$ ,

$$\theta_3 = -0.96 - 0.4(0.08) = -0.992.$$

## 4. Conclusion

After three iterations of gradient descent, the parameter value is:

$$\theta \approx -0.992,$$

which is very close to the exact analytical minimum:

$$\theta^* = -1$$

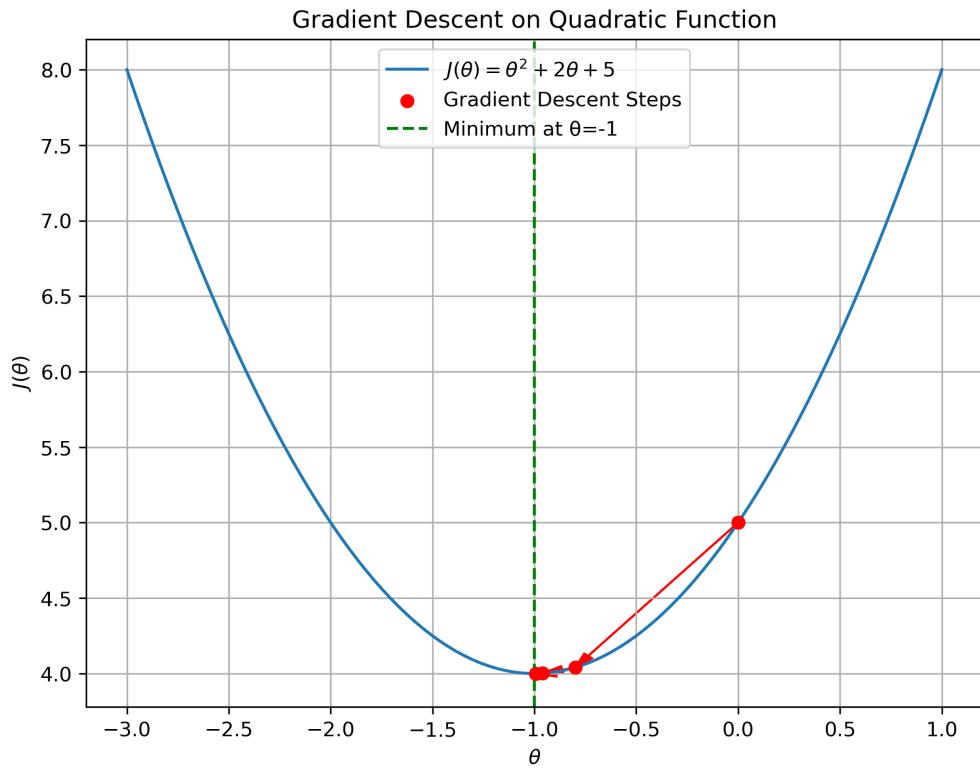
The minimum value of the function is:

$$J(\theta^*) = 4$$

This demonstrates that gradient descent with a sufficiently large learning rate can converge rapidly for a convex quadratic function.

## 5. Visualization of Gradient Descent

The following plot illustrates the gradient descent iterations on the cost function:



## Linear Regression via Gradient Descent

### Introduction

This section explores linear regression optimization using gradient descent for a simple dataset of three data points.

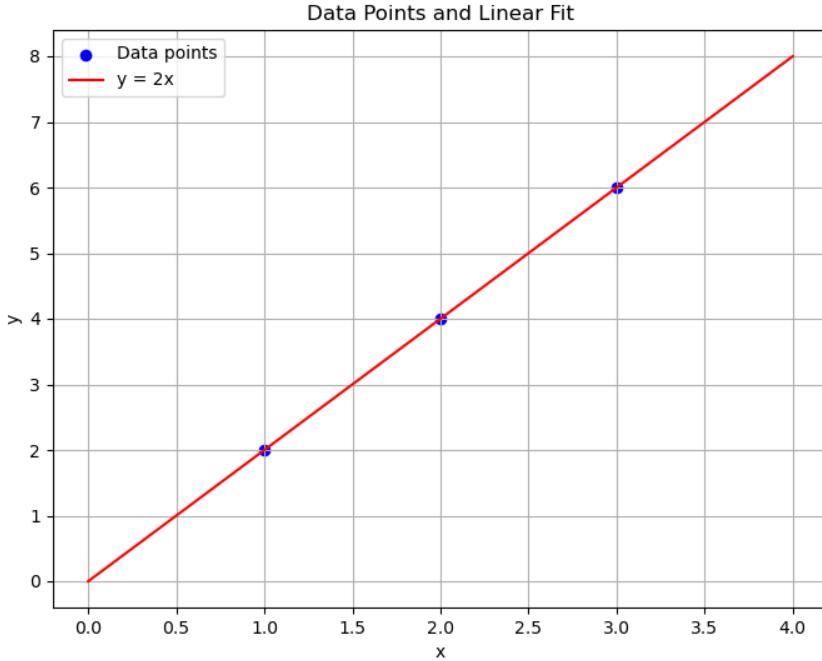
We are given the data points

$$(x, y) = (1, 2), (2, 4), (3, 6).$$

From inspection, the output variable  $y$  is exactly twice the input variable  $x$  for all data points. Therefore, we expect the optimal linear relationship to be

$$y = 2x,$$

which suggests that the optimal parameter value should be  $\theta = 2$ . Next we will verify this result by explicitly formulating and solving a linear regression optimization problem using gradient descent.



## 1. Model Assumption

We use a linear model *without an intercept*:

$$h_{\theta}(x) = \theta x.$$

No intercept term is included because the data lie exactly on a line passing through the origin (and it makes the math a bit easier for this example). For a more complex or completed problem you would be expected to include the intercept

## 2. Define the Cost Function

We use the mean squared error cost function:

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

where  $m = 3$  is the number of data points.

Substituting the data,

$$J(\theta) = \frac{1}{6} [(\theta - 2)^2 + (2\theta - 4)^2 + (3\theta - 6)^2].$$

## 3. Gradient Descent Solution

### Compute the Gradient

To compute the gradient, we need to find the derivative of the cost function  $J(\theta)$  with respect to  $\theta$ .

The cost function is:

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

where  $h_\theta(x_i) = \theta x_i$  is our linear model.

To find  $\frac{dJ}{d\theta}$ , we differentiate each term in the sum using the chain rule:

$$\frac{d}{d\theta} (h_\theta(x_i) - y_i)^2 = 2(h_\theta(x_i) - y_i) \cdot \frac{d}{d\theta} (\theta x_i) = 2(h_\theta(x_i) - y_i)x_i.$$

Differentiating the entire cost function:

$$\frac{dJ}{d\theta} = \frac{1}{2m} \cdot 2 \sum_{i=1}^m (h_\theta(x_i) - y_i)x_i = \frac{1}{m} \sum_{i=1}^m (h_\theta(x_i) - y_i)x_i.$$

For our data with  $m = 3$ , this becomes

$$\frac{dJ}{d\theta} = \frac{1}{3} [(\theta x_1 - y_1)x_1 + (\theta x_2 - y_2)x_2 + (\theta x_3 - y_3)x_3].$$

Substituting the data,

$$\frac{dJ}{d\theta} = \frac{1}{3} [(\theta \cdot 1 - 2) \cdot 1 + (\theta \cdot 2 - 4) \cdot 2 + (\theta \cdot 3 - 6) \cdot 3].$$

Simplifying each term:

$$\begin{aligned} (\theta - 2) \cdot 1 &= \theta - 2, \\ (2\theta - 4) \cdot 2 &= 4\theta - 8, \\ (3\theta - 6) \cdot 3 &= 9\theta - 18. \end{aligned}$$

Summing these terms:

$$(\theta - 2) + (4\theta - 8) + (9\theta - 18) = 14\theta - 28.$$

Dividing by  $m = 3$ :

$$\frac{dJ}{d\theta} = \frac{14\theta - 28}{3} = \frac{14}{3}\theta - \frac{28}{3} = \frac{14}{3}(\theta - 2).$$

### Gradient Descent Update Rule

The gradient descent update equation is

$$\theta_{k+1} = \theta_k - \alpha \frac{dJ}{d\theta} \Big|_{\theta=\theta_k}.$$

Substituting the gradient expression,

$$\theta_{k+1} = \theta_k - \alpha \frac{14}{3}(\theta_k - 2).$$

### Choose Parameters

We choose:

- Initial value:  $\theta_0 = 0$
- Learning rate:  $\alpha = 0.3$

### Iteration 1

$$\theta_1 = 0 - 0.3 \left( \frac{14}{3}(0 - 2) \right) = 2.8.$$

### Iteration 2

$$\theta_2 = 2.8 - 0.3 \left( \frac{14}{3}(2.8 - 2) \right) = 1.68.$$

### Iteration 3

$$\theta_3 = 1.68 - 0.3 \left( \frac{14}{3}(1.68 - 2) \right) \approx 2.13.$$

## 4. Conclusion

After three iterations of gradient descent, the parameter value is

$$\theta \approx 2.13,$$

which is clearly converging toward the expected optimal value

$$\theta^* = 2.$$

The resulting model,

$$h(x) = 2x,$$

exactly captures the linear relationship in the data, yielding zero prediction error.

## 5. Visualization of Gradient Descent

The following plot illustrates the gradient descent iterations on the cost function:

