Week 2: Optimization Problems in Machine Learning

Gregory Glickert

January 27, 2026

Introduction

This document explores two 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.

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 θ:

$$\frac{dJ}{d\theta} = 8\theta - 40.$$

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:

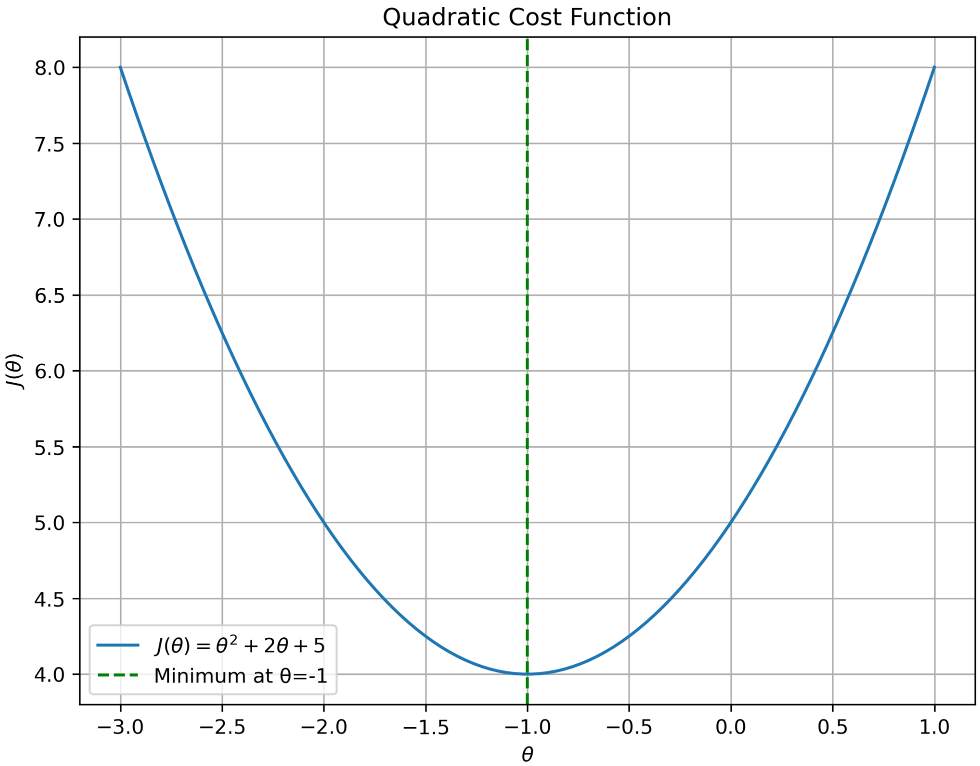
$$J(\theta) = \theta^2 + 2\theta + 5.$$

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

2. Plotting $J(\theta)$ vs. θ

The function $J(\theta)$ is a quadratic polynomial with a positive coefficient on θ^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 θ:

$$\frac{dJ}{d\theta} = 2\theta + 2.$$

Gradient Descent Update Equation

The gradient descent update rule is given by:

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

Here, α 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:

$$\boxed{\theta^* = -1}$$

The minimum value of the function is:

$$\boxed{J(\theta^*) = 4}$$

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: