Optimization Techniques for Machine Learning

AMLZC326 · #07 Continuous Optimization I

Anshid Aboobacker

Table of Contents

Why Study Optimization?

Many problems in Machine Learning reduce to the following task:

Find parameters that minimize a loss function

Examples:

- Training a linear regression model
- Training a neural network
- Support Vector Machines
- Logistic regression

Key Idea

Machine learning models are typically obtained by solving an **optimization problem**.

Learning Objectives

By the end of this lecture you should be able to:

- Apply gradient descent and analyse how the learning rate η affects convergence speed and stability
- Formulate constrained optimisation problems with equality and inequality constraints
- Construct the Lagrangian and apply the method of Lagrange multipliers to find constrained optima
- State the necessary conditions for a constrained optimum (stationarity of the Lagrangian)

Learning Objectives

By the end of this lecture you should be able to:

- Apply gradient descent and analyse how the learning rate η affects convergence speed and stability
- Formulate constrained optimisation problems with equality and inequality constraints
- Construct the Lagrangian and apply the method of Lagrange multipliers to find constrained optima
- State the necessary conditions for a constrained optimum (stationarity of the Lagrangian)

From Calculus to Optimization

In calculus we studied: derivatives, gradients, critical points
These tools help us answer the question:

Where does a function attain its minimum or maximum?

Optimization formalizes this idea:

$$\min_{x \in \mathbb{R}^n} f(x)$$

where

- x : parameters or variables
- $f(x)$: objective function (or loss function)

Goal: Find the value of x that minimizes $f(x)$

Unconstrained Vs Constrained Optimization

Unconstrained Optimization

An optimization problem without restrictions on the variables.

General form: $\min_{x \in \mathbb{R}^n} f(x)$

Example: $\min_{x,y} \; x^2 + y^2$

The minimum occurs at $(x, y) = (0, 0)$

Constrained Optimization

Sometimes variables must satisfy constraints.

General form:

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to } g(x) = 0 \text{ or } g(x) \leq 0$$

Example: $\min_{x,y} x^2 + y^2$ subject to $x + y = 1$

Example: Linear Regression

Suppose we have a dataset $\{(x_i, y_i)\}_{i=1}^n$
We model the relationship using a linear function

$$\hat{y}_i = wx_i + b$$

The squared error loss over the dataset is

$$L(w, b) = \sum_{i=1}^{n} (wx_i + b - y_i)^2$$

Training the model corresponds to solving the optimization problem

$$\min_{w,b} \sum_{i=1}^{n} (wx_i + b - y_i)^2$$

Table of Contents

Finding the Minimum of a Function

Consider a function $f(x)$

Goal: Find the value of x that minimizes $f(x)$

In simple cases we can compute this analytically. $f'(x) = 0$
However, in many machine learning problems:

- functions are high dimensional
- derivatives are complicated
- closed-form solutions may not exist

Idea of Iterative Optimization
Instead of solving directly, we improve the solution step by step.
Start with an initial guess: x_0
Then repeatedly update it: $x_0 \to x_1 \to x_2 \to \ldots$
Each step should move us closer to the minimum.

Gradient Descent

Suppose we are currently at point x_k.

The gradient $\nabla f(x_k)$ points in the direction of **steepest increase**.

To minimize the function we move in the opposite direction: $-\nabla f(x_k)$

Update rule: $x_{k+1} = x_k - \eta \nabla f(x_k)$

Algorithm:

1. Start with initial guess x_0
2. Compute gradient $\nabla f(x_k)$
3. Update using gradient descent
4. Repeat until convergence

Learning Rate and Local Minima

The parameter η is called the **learning rate**. It controls how large each step is.

Small learning rate:

- stable updates
- slow convergence

Large learning rate:

- faster updates
- may overshoot the minimum

Many functions have multiple minima:

- local minimum
- global minimum

Gradient descent typically converges to a **local minimum**, and the result may depend on the starting point x_0.

Visual Example

For the function $f(x, y) = x^2 + y^2$

Level curves are circles centered at the origin.

Gradient: $\nabla f(x, y) = \begin{bmatrix} 2x & 2y \end{bmatrix}$

Gradient descent moves toward the origin. The origin is the minimum of the function.

Geometric Interpretation

Gradient descent follows the direction of steepest slope downhill.
Think of standing on a mountain in fog.

- You cannot see the entire landscape.
- You only feel the local slope.
- You step in the direction of the steepest downward slope.

This is exactly what gradient descent does.

Table of Contents

Why Do Constraints Arise?

In many optimization problems, the variables cannot take arbitrary values.
They must satisfy certain restrictions called **constraints.**
Examples:

- Budget constraints in economics
- Physical limitations in engineering
- Probability distributions must sum to 1
- Regularization constraints in machine learning

Therefore we often need to solve optimization problems **with constraints**.

Constrained Optimization Problem

A general constrained optimization problem is

$$\min f(x)$$

subject to

$$g(x) = 0$$

where

- $f(x)$: objective function
- $g(x)$: constraint

The solution must both minimize f and satisfy the constraint.

Example

Consider

$$\min x^2 + y^2$$

subject to

$$x + y = 1$$

The constraint restricts the variables to the line

$$x + y = 1$$

Therefore we minimize the function **along the constraint**.

Key Question

How do we minimize a function when the variables must satisfy a constraint?
In unconstrained optimization, a minimum occurs where

$$\nabla f(x) = 0$$

However, when a constraint such as

$$g(x) = 0$$

is present, we cannot move freely in all directions.
The feasible points lie only on the curve defined by the constraint.

Key Idea: Use the method of **Lagrange multipliers** to convert the constrained problem into an equivalent unconstrained problem.

Table of Contents

Geometric Intuition

The constraint

$$g(x) = 0$$

defines a curve of feasible points.
The objective function has level curves

$$f(x) = c$$

To minimize the function, we look for the **smallest level curve that touches the constraint**.

Key Observation

At the optimal point, the level curve of f and the constraint curve are **tangent**.

Gradient Condition

If two curves are tangent, their normals must be parallel.
The normal to a level curve of f is the gradient

$$\nabla f(x)$$

The normal to the constraint $g(x) = 0$ is

$$\nabla g(x)$$

Lagrange Multiplier Condition

$$\nabla f(x) = \lambda \nabla g(x)$$

Solving the Constrained Problem

To find candidate solutions we solve $\nabla f(x) = \lambda \nabla g(x)$ together with the constraint $g(x) = 0$

Instead of solving $\nabla f(x) = \lambda \nabla g(x)$ directly, we introduce the **Lagrangian**

$$\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$$

This converts the constrained optimization problem into an unconstrained one.
The variable λ is called the **Lagrange multiplier**.

Necessary Conditions

The optimal point satisfies

$$\nabla_x \mathcal{L}(x, \lambda) = 0$$

together with the constraint

$$g(x) = 0$$

Expanding the gradient gives

$$\nabla f(x) + \lambda \nabla g(x) = 0$$

which is equivalent to

$$\nabla f(x) = -\lambda \nabla g(x)$$

Solving Constrained Problems

To solve a constrained optimization problem:

1. Construct the Lagrangian

$$\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$$

2. Compute derivatives

$$\nabla_x \mathcal{L}(x, \lambda)$$

3. Solve the system

$$\nabla_x \mathcal{L}(x, \lambda) = 0$$

$$g(x) = 0$$

Example

Consider the optimization problem

$$\min x^2 + y^2$$

subject to

$$x + y = 1$$

Goal: find the point on the line $x + y = 1$ closest to the origin.

Construct the Lagrangian

$$\mathcal{L}(x, y, \lambda) = x^2 + y^2 + \lambda(x + y - 1)$$

Example

Compute partial derivatives of the Lagrangian.

$$\frac{\partial \mathcal{L}}{\partial x} = 2x + \lambda \qquad \frac{\partial \mathcal{L}}{\partial y} = 2y + \lambda \qquad \frac{\partial \mathcal{L}}{\partial \lambda} = x + y - 1$$

Setting derivatives equal to zero gives

$$2x + \lambda = 0 \qquad 2y + \lambda = 0 \qquad x + y - 1 = 0$$

From the first two equations we obtain

$$2x = 2y \implies x = y$$

Using the constraint

$$x + y = 1 \implies 2x = 1 \implies x = \frac{1}{2}, \quad y = \frac{1}{2}$$

Thus the minimum occurs at $\left(\frac{1}{2}, \frac{1}{2}\right)$

Key Takeaways

- Gradient descent: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \nabla f(\boldsymbol{x}_k)$; η too large $\Rightarrow$ divergence, too small $\Rightarrow$ slow convergence
- Unconstrained optimum: $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ (necessary); $H(\boldsymbol{x}^*) \succ 0$ (sufficient for a minimum)
- Lagrangian: $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_i \lambda_i g_i(\boldsymbol{x}) + \sum_j \mu_j h_j(\boldsymbol{x})$ converts constraints into penalties
- At a constrained optimum: $\nabla_{\boldsymbol{x}} \mathcal{L} = \mathbf{0}$ (stationarity) — leads to KKT conditions in CH09–10

Thank you :)