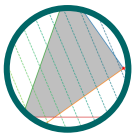
7

Continuous Optimization



Since machine learning algorithms are implemented on a computer, the mathematical formulations are expressed as numerical optimization methods. This chapter describes the basic numerical methods for training machine learning models. Training a machine learning model often boils down to finding a good set of parameters. The notion of "good" is determined by the objective function or the probabilistic model, which we will see examples of in the second part of this book. Given an objective function, finding the best value is done using optimization algorithms.

Since we consider data and models in $\mathbb{R}^D$, the optimization problems we face are *continuous* optimization problems, as opposed to *combinatorial* optimization problems for discrete variables.

This chapter covers two main branches of continuous optimization (Figure 7.1): unconstrained and constrained optimization. We will assume in this chapter that our objective function is differentiable (see Chapter 5), hence we have access to a gradient at each location in the space to help us find the optimum value. By convention, most objective functions in machine learning are intended to be minimized, that is, the best value is the minimum value. Intuitively finding the best value is like finding the valleys of the objective function, and the gradients point us uphill. The idea is to move downhill (opposite to the gradient) and hope to find the deepest point. For unconstrained optimization, this is the only concept we need, but there are several design choices, which we discuss in Section 7.1. For constrained optimization, we need to introduce other concepts to manage the constraints (Section 7.2). We will also introduce a special class of problems (convex optimization problems in Section 7.3) where we can make statements about reaching the global optimum.

global minimum

Consider the function in Figure 7.2. The function has a *global minimum* around $x = -4.5$, with a function value of approximately -47. Since the function is "smooth," the gradients can be used to help find the minimum by indicating whether we should take a step to the right or left. This assumes that we are in the correct bowl, as there exists another *local minimum* around $x = 0.7$. Recall that we can solve for all the stationary points of a function by calculating its derivative and setting it to zero. For

local minimum

Stationary points are the real roots of the derivative, that is, points that have zero gradient.

$$\ell(x) = x^4 + 7x^3 + 5x^2 - 17x + 3\,, \tag{7.1}$$

we obtain the corresponding gradient as

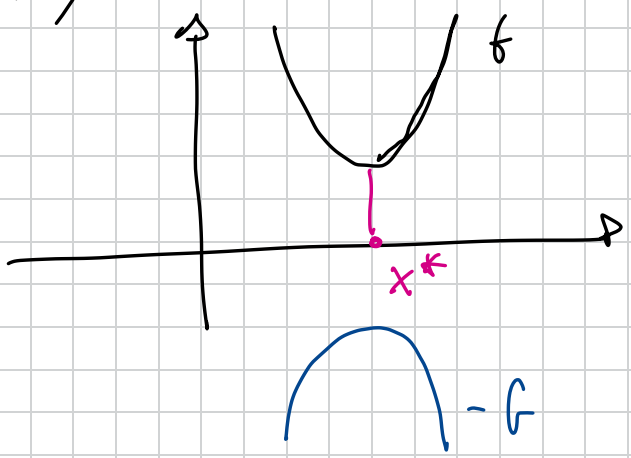
$$\frac{\mathrm{d}\ell(x)}{\mathrm{d}x} = 4x^3 + 21x^2 + 10x - 17\,. \tag{7.2}$$

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

$$f: \mathbb{R}^n \to \mathbb{R} \qquad \underline{x} \in \mathbb{R}^n$$

$\rightarrow$ $\max f(x)$

$$\arg\max_{x \in \mathbb{R}} f(x) =$$



$$= \arg\min_{x \in \mathbb{R}} f(x)$$

Definitions

1) x^* is a point of strict local minimum for f if it exists $\varepsilon > 0$:

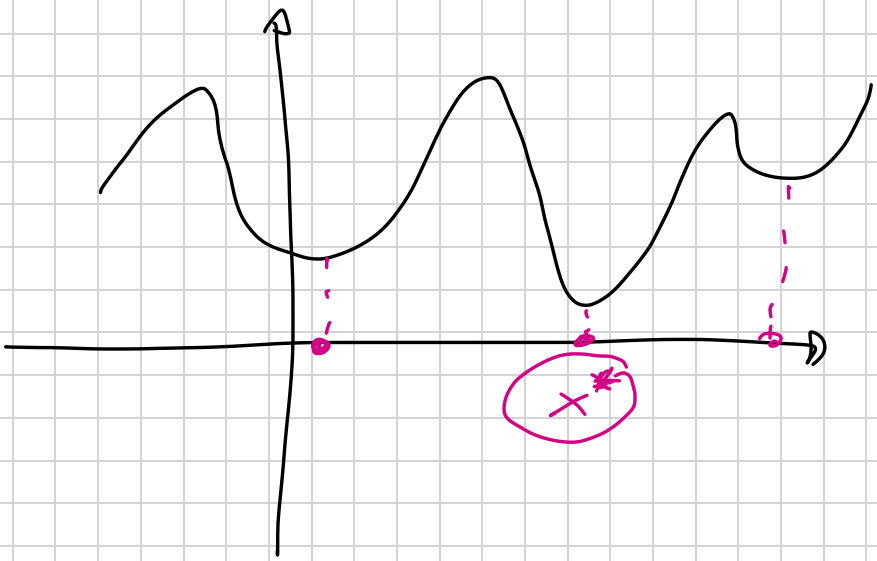
$$f(x^*) \leq f(x) \quad \forall \in \|\underline{x} - x^*\| < \varepsilon$$

$$f(x^*) < f(x)$$

2) x^* is a point of strict global minimum for f if

$$f(x^*) \leq f(x) \quad \forall x \in \mathbb{R}^n$$

$$f(x^*) < f(x) \quad \forall x \in \mathbb{R}^n$$



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

$\min_{a \leq x \leq b} f(x)$ constrained

Optimality conditions

FIRST ORDER CONDITIONS

If x^* is a point of local minimum and f is differentiable with continuity around x^*, then

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

NECESSARY CONDITION

x^* LOCAL MINIMUM $\Rightarrow \nabla f(x^*) = 0$

$\not\Leftarrow$

$x^* : \nabla f(x^*) = 0$ called STATIONARY POINT

SECOND ORDER CONDITIONS

If x^* is a local minimum for f and f is twice differentiable around x^*, then

$$\nabla f(x^*) = 0 \quad \& \quad \nabla^2 f(x^*)$$

is positive semidefinite

$\nabla^2 f(x^*)$ symmetric matrix $n \times n$

$$\nabla^2 f_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} \qquad i, j = 1 \dots n$$

$$\| \qquad \frac{\partial^2 f}{\partial x_j \partial x_i}$$

NECESSARY CONDITION

x^* LOCAL MINIMUM $\Rightarrow \nabla f(x^*) = 0$ & $\nabla^2 f(x^*)$ is pos. semid.

SUFFICIENT CONDITION OF THE SECOND ORDER

If f is twice differentiable with continuity around x^* and $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite then

x^* is a strict local minimum of f.

SUFFICIENT CONDITION

$\nabla f(x^*) = 0$ & $\nabla^2 f(x^*)$ pos def $\Rightarrow$ x^* strict local minimum

REMARK:

If $\nabla f(x^*) = 0$ then x^* can be:

a) a minimum for f

b) a maximum for f

c) a saddle point for f

Example

$$f(x_1, x_2) = (x_1^2 + x_2^2 - 1)^2 + (x_2^2 - 1)^2$$

find the stationary points (classify them)

$\frac{\partial f}{\partial x_1}$

$$\nabla f(x_1, x_2) = \Big((x_1^2 + x_2^2 - 1) \cdot 2x_1 ,$$

$$(x_1^2 + x_2^2 - 1) \cdot 2x_2 + 2(x_2^2 - 1) \cdot 2x_2 \Big)$$

$$\nabla f(x_1, x_2) = (0, 0)$$

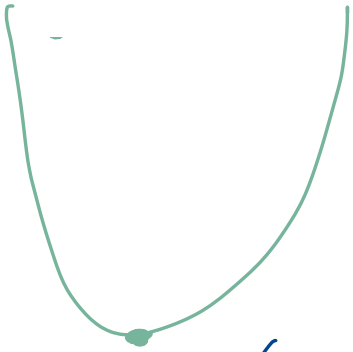
$$\begin{cases} 2x_1(x_1^2 + x_2^2 - 1) = 0 \\ 2x_2(x_1^2 + x_2^2 - 1) + 4x^2(x_2^2 - 1) = 0 \end{cases}$$

$\nabla^2 f(x_1, x_2)$ 2 x 2 matrix

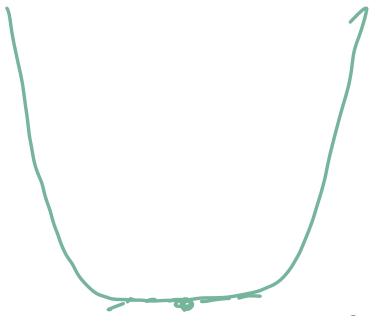
$\nabla^2 f(x_1^*, x_2^*)$ (x_1^*, x_2^*) is a stationary point

is positive definite?

$\Rightarrow (x_1^*, x_2^*)$ strict local minimum

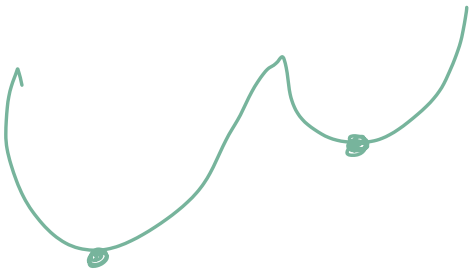


STRICT (LOCAL)



x^* LOCAL

$$f(x^*) \leq f(x) \qquad x \in [x^* - \varepsilon, x^* + \varepsilon]$$



Def.

A function $f: \mathbb{R}^n \to \mathbb{R}$ is said to be **differentiable at a point x_0** if there exists a linear mapping $J: \mathbb{R}^n \to \mathbb{R}^m$

$$\lim_{\underline{h} \to 0} \frac{\| f(\underline{x}_0 + \underline{h}) - f(\underline{x}_0) - J(\underline{h}) \|}{\|\underline{h}\|} = 0$$

this means that f can be approximated by a linear mapping around x_0.

f is differentiable in a domain D if it is differentiable in each point of D.

Proposition

If A is an open set of $\mathbb{R}^n$, $f: A \to \mathbb{R}$, and all the partial derivatives of f in x_0, $\left(\frac{\partial f}{\partial x_i}(\underline{x}_0), i = 1 \dots n \right)$ exist and are continuous, then f is differentiable in x_0.

N.B. NECESSARY CONDITION

$\frac{\partial f}{\partial x_i}(\underline{x}_0)$ exist and is continuous

$\forall i = 1 \dots n$

$\Downarrow \not\Rightarrow$

f is differentiable in x_0

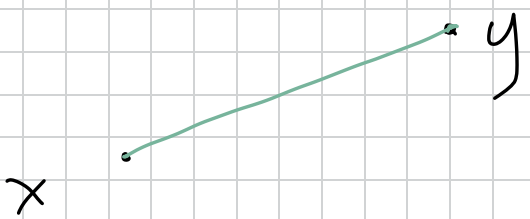
CONVEX FUNCTIONS

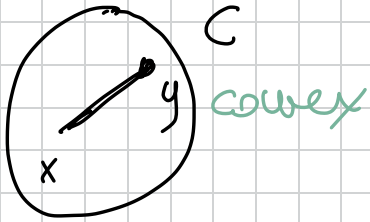
Def. CONVEX SET

A set $C \subset \mathbb{R}^n$ is CONVEX if $\forall x, y \in C$ and $\forall \theta \in [0,1] \subset \mathbb{R}$ we have:

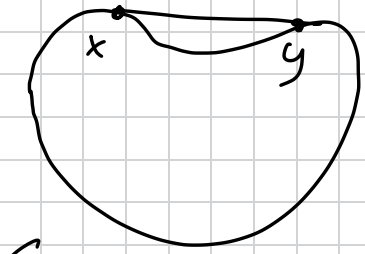
$$\theta x + (1-\theta) y \in C$$

↓ equation of the line connecting x and y





convex



C not convex

Def. CONVEX FUNCTION

$f: C \subset \mathbb{R}^n \to \mathbb{R}$, C convex set

f is a convex function if

$\forall x, y \in C$ and $\theta \in [0,1] \subset \mathbb{R}$

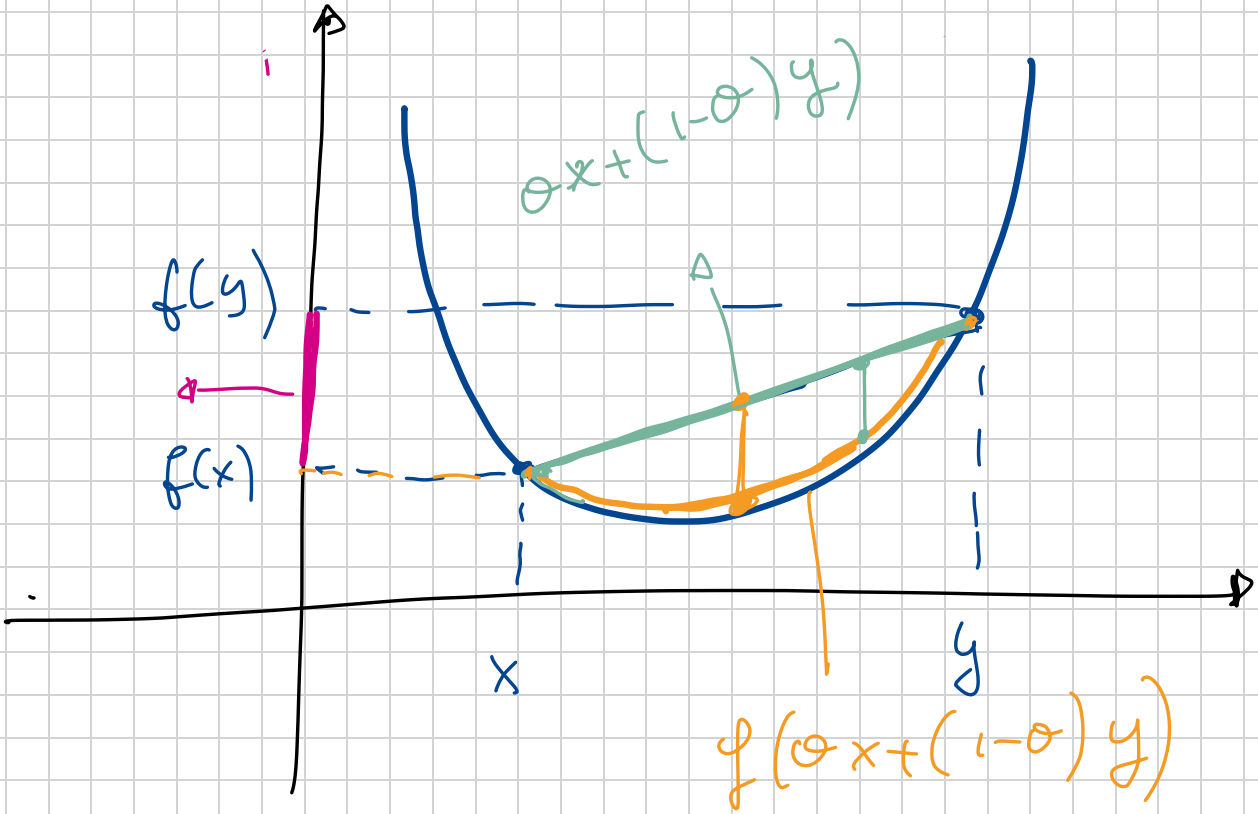
we have:

$$f(\theta x + (1-\theta) y) \leq \theta f(x) + (1-\theta) f(y)$$

$\theta x + (1-\theta) y$ ↓ $f(z) \leq \theta$

value connecting $f(x)$ and $f(y)$

a line connecting x and y lies above the function in x and y



Property

- If f is a convex function, then each local minimum is also a global minimum.

- If f is convex and differentiable then each stationary point is a global minimum of f.

$$\nabla f(x^*) = 0 \Longrightarrow x^* \text{ global minimum}$$

$F: \|Ax - b\|_2^2 \longrightarrow$ convex function

$$\|Ax - b\|_2^2 : \mathbb{R}^n \longrightarrow \mathbb{R}$$

$$\nabla F(x) = 2A^TAx - 2A^Tb$$

$$\nabla F(x) = 0 \Rightarrow 2A^TAx - 2A^Tb = 0$$

$$\Rightarrow A^TAx = A^Tb \quad \text{LINEAR SYSTEM}$$

The solution x's the global minimum of F.

RESUME

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \quad \rightarrow \text{LEAST SQUARES PROBLEM}$$

if A has full rank has a unique solution:

1) by solving the linear system obtained from the first order conditions ($\nabla f(x) = 0$)

$$A^T A x = A^T b \qquad (n \times n)$$

system of normal equations

$A^T A$ is symmetric & positive definite $\Rightarrow$ apply the Cholesky decomposition

NOTE that A can be of size $m \times n$, $m \geq n$

$A^T A$ of size $n \times n$

For a general function $f: \mathbb{R}^n \to \mathbb{R}$, how can we solve

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

→ ITERATIVE ALGORITHMS THAT COMPUTE A LOCAL MINIMUM of f.

create a sequence of "iterative solutions" of the problem

$$\underline{x}_0, \underline{x}_1, \underline{x}_2 \dots \underline{x}_k, \underline{x}_{k+1} \dots$$

with the property:

$$\lim_{k \to \infty} \underline{x}_k = x^*$$

x^* "SOLUTION" of the problem

Starting from a given $\underline{x}_0$

$$x_k = \underline{G}(x_{k-1})$$

$$\left(x_k = G(x_{k-1}, x_{k-2} \dots x_{k-p}) \right)$$

$$\{x_0, x_1 \dots x_{k^*}\} \dots \quad <>$$

TRUNCATION ERROR

- k^* is OBTAINED BY

STOPPING CRITERIA

→ $\|\nabla f(x_k)\| < \tau$ τ = tolerance

A PARTICULAR CLASS OF ITERATIVE ALGORITHMS TO COMPUTE

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

↳ objective function

DESCENT METHODS

$$x_k = G(x_{k-1}) \rightarrow$$

$$\underline{x}_k = \underline{x}_{k-1} + \alpha_{k-1} \underline{p}_{k-1}$$

x_{k-1} previous iterate $\in \mathbb{R}^n$

$\underline{p}_{k-1} \in \mathbb{R}^n$ → search direction

$\alpha_{k-1} \in \mathbb{R}$ → step length

p_{k-1} → also called descent direction