Optimisation for Machine Learning

Ludovic De Matteis
LAAS-CNRS
ldematteis@laas.fr

Table of contents

1 Motivations in Machine Learning

Optimization describe a mathematical theory focussing on the problem of finding the minimum (or equivalently the maximum) of a function. This problems occurs in almost every domain, from investment portfolios built to maximize the return to minimizing the error in a weather forecast model. This introductory course aims at teaching the basics of optimization and how it is applied to solve machine learning problem.

1.a Unconstraint Optimization

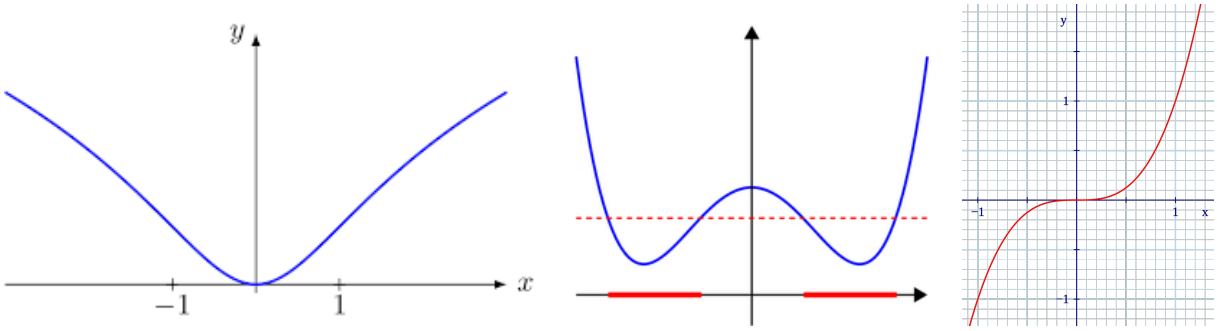
In unconstraint optimization, we consider problems of the form

$$\inf_{x\in\mathbb{R}^n} f(x) \tag{1}$$

we define the set of **global minimizers** of the function f as

$$\arg\min_{x\in\mathbb{R}^n} f(x) \stackrel{\text{def}}{=} \{x_0 \in \mathbb{R}^n \mid \forall x \in \mathbb{R}^n, f(x_0) \leq f(x)\} \tag{2}$$

The minimizer of a function does not necessarily exists and if it does, it can be non unique.



(a) Unique Minimizer (b) Multiple Minimizer (c) Zero Minimizer

Figure 1: Different functions shapes and corresponding number of minimizers

Figure 1 shows different function shapes corresponding to different number of minimizers. In general, when a minimizer exists - and this will be our focus for the rest of the course - we denote the problem Equation 1 as

$$\min_{x\in\mathbb{R}^n} f(x) \tag{3}$$

1.b Regression

We will start by considering a linear regression problem, in which

$$f(x) = \frac{1}{2}\sum_{i-1}^{N}(y_i - < x, a_i >)^2 = \frac{1}{2} \parallel Ax - y \parallel^2 \tag{4}$$

is the least square quadratic risk function.

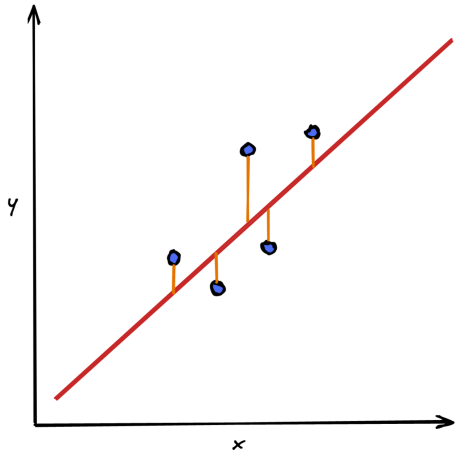


Figure B: Linear regression problem.
The linear function $l(x) = Ax$ in red aims to fit as well as possible the given data points

An illustration of the linear regression problem is shown in Figure B.

The regression problem can be extended to different function, by considering for instance a quadratic function

$$f(x) = \frac{1}{2} \|(x^T Ax + Bx) - y\|^2 \tag{5}$$

or an exponential function

$$f(x) = \frac{1}{2} \parallel e^{Ax} - y \parallel^2 \tag{6}$$

1.c Classification

For binary classification, the data points y_i are **labelled** with a value 1 or -1, defining a class. In this problem, we aim the minimize the function

$$f(x) = \sum_{i-1}^{n} l(-y_i < x, a_i >) = L(-\text{diag}(y)Ax) \tag{7}$$

where l is the 0-1 loss function $1_{\mathbb{R}^+}$ or a smooth approximation of it, giving a value of 1 if the signs of y_i and $< a_i, x >$ are opposed and zero otherwise.

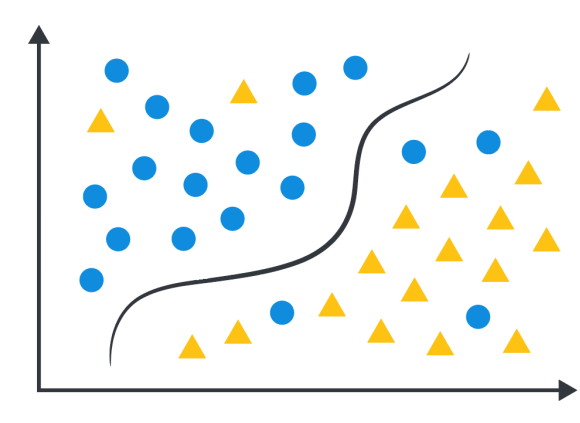


Figure C: Binary Classification problem.
The function in black aims to separate as well as possible the different classes

As for regression, the classification problem can use a different separating function. Figure C presents a binary classification problem with an arbitrary separating function.

2 Derivatives and Gradients

2.a Reminder on derivatives

A function is said to be differentiable at a point x_0 if the following limit exists:

$$\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{8}$$

The value of this limit when it exists is called the derivative of f at x_0 and is denoted $f'(x_0)$ or $\frac{df}{dx}(x_0)$.

Examples - Let's illustrate this idea on several classical functions

- $f(x) = 3x^2 - x$

$$\begin{aligned} f(x+h) - f(x) &= 3(x+h)^2 - x - h - 3x^2 + x \\ &= 3(h^2 + 2xh) - h \end{aligned} \tag{9}$$

which gives

$$\frac{f(x+h) - f(x)}{h} = 3h + 6x - 1 \tag{10}$$

the limit when $h \to 0$ as this expression is defined for all points $x \in \mathbb{R}$ and gives the derivative

$$f'(x) = 6x - 1 \tag{11}$$

- $f(x) = e^{-2x}$

$$\begin{aligned}\frac{f(x+h)-f(x)}{h} &= \frac{e^{-2x-2h}-e^{-2x}}{h} \\ &= \frac{e^{-2x}\left(e^{-2h}-1\right)}{h} \\ &= -2e^{-2x}\frac{\left(e^{-2h}-1\right)}{-2h}\end{aligned} \tag{12}$$

Moreover

2.b First order optimality conditions

2.c Derivative of classification cost

2.d The chain rule

3 Gradient Descent

3.a The algorithm

3.b Convergence Analysis

3.c Regularization

4 Newton method

4.a Comparison the Gradient Descent

4.b Advantages

4.c Limits

5 Stochastic Optimization

5.a Problem formulation

5.b Batch gradient descent

5.c Stochastic gradient descent

5.d Backpropagation

6 Neural Networks

6.a Perceptron

6.b Multi-Layer Perceptron (MLP)

6.c Additional structures

7 Opening

7.a Introduction to Large Language Models