Optimisation for machine learning

Ludovic De Matteis
ldematteis@laas.fr

Table of contents

I. 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.

I.A. Unconstraint Optimization

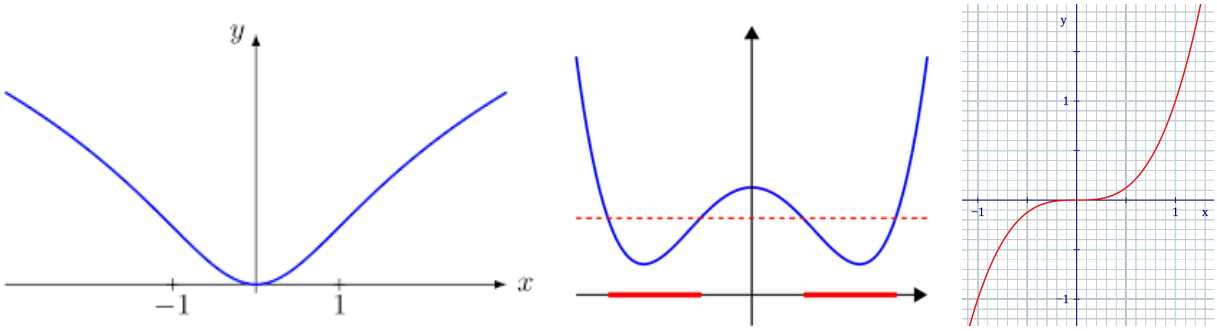
In unconstraint optimization, we consider problems of the form

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

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 global 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 global minimizer exists - and this will be our focus for the rest of the course - we denote the problem Equation 1 as

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

We will also define the notion of **local minimizer** as follows. The point x^* is a local minimizer of the function f if there exists a radius $r > 0$ such that for all $x \in \mathbb{R}^n$ such that $\|x - x^*\| \leq r$, we have $f(x^*) \leq f(x)$. Note that a local minimizer is not necessarily a global minimizer (but a global minimizer is necessarily a local minimizer).

I.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.

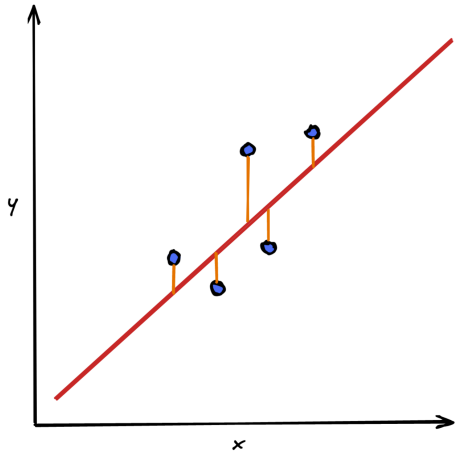


Figure B: Linear regression problem.
The linear function $l(x) = Ax$ in red aims to fit as well as possible the given data points

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

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

I.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.

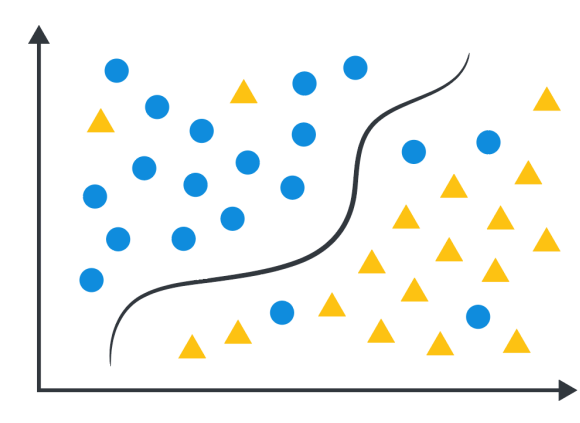


Figure C: Binary Classification problem.
The function in black aims to separate as well as possible the different classes

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

II. Derivatives and Gradients

II.A. Reminder on derivatives

A function $f : \mathbb{R} \to \mathbb{R}$ 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

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

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

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

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

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

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

Moreover, the limit when $h \to 0$ of $\frac{(e^{-2h}-1)}{-2h}$ is 1 (it can be shown using the Taylor expansion of the exponential function). This gives

$$f'(x) = -2e^{-2x} \tag{13}$$

II.B. On Gradient and Jacobian

The definition of a derivative can be extended to functions with multiple variables. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function of d variables. The function f is said to be differentiable at a point $x_0 \in \mathbb{R}^d$ if there exists a vector $g \in \mathbb{R}^d$ such that

$$f(x_0 + h) = f(x_0) + g^T h + o(\|h\|) \tag{14}$$

where $o(\|h\|)$ is a function such that $\frac{o(\|h\|)}{\|h\|} \to 0$ when $\|h\| \to 0$. The vector g is called the gradient of f at point x_0 and is denoted $\nabla f(x_0)$ or $\frac{df}{dx}(x_0)$. Note that the vector g is unique (if it exists).

The gradient vector is related to the derivatives of the function by

$$\nabla f(x) = \left(\frac{df}{dx_1}(x), \frac{df}{dx_2}(x), ..., \frac{df}{dx_d}(x)\right)^T \tag{15}$$

where x_i is the i-th component of vector x.

We also define the **Jacobian** of a function $f : \mathbb{R}^d \to \mathbb{R}^p$ as the matrix $J \in \mathbb{R}^{p \times d}$ writen as

$$\begin{aligned} J &= \begin{pmatrix} \frac{df_1}{dx_1} & \frac{df_1}{dx_2} & \cdots & \frac{df_1}{dx_d} \\ \frac{df_2}{dx_1} & \frac{df_2}{dx_2} & \cdots & \frac{df_2}{dx_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{df_p}{dx_1} & \frac{df_p}{dx_2} & \cdots & \frac{df_p}{dx_d} \end{pmatrix} \\ &= \begin{pmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \vdots \\ \nabla f_p(x)^T \end{pmatrix} \end{aligned} \tag{16}$$

where f_i is the i-th component of the vector-valued function f.

Note: We observe that for a scalar-valued function, the Jacobian equal the transpose of the gradient.

II.C. Derivative of classification cost

Let's consider the classification cost function defined as

$$f(x) = \sum_{i-1}^{n} l(-y_i < x, a_i >) = L(-\text{diag}(y)Ax) \tag{17}$$

We can compute the gradient of this function as follows

$$\begin{aligned} f'(x) &= \sum_{i-1}^{n} l'(-y_i < x, a_i >)(-y_i a_i) \\ &= -A^T \ \text{diag}(y) L'(-\text{diag}(y)Ax) \end{aligned} \tag{18}$$

where L' is the vector of derivatives of the function l applied component-wise.

II.D. The chain rule

The chain rule allows to compute the derivative of a composition of functions. Let $f : \mathbb{R}^d \to \mathbb{R}^p$ and $g : \mathbb{R}^p \to \mathbb{R}$ be two differentiable functions. The function $h : \mathbb{R}^d \to \mathbb{R}$ defined as

$$h(x) = g(f(x)) \tag{19}$$

can be shown to be differentiable at point $x_0 \in \mathbb{R}^d$ if f is differentiable at x_0 and g is differentiable at $f(x_0)$ and its gradient is given by

$$\nabla h(x_0) = J_{f(x_0)}^T \nabla g(f(x_0)) \tag{20}$$

where $J_{f(x_0)}$ is the Jacobian of function f at point x_0 and $\nabla g(f(x_0))$ is the gradient of function g at point $f(x_0)$.

II.E. First order optimality conditions

Let's consider again an unconstrained minimization problem of the form of Equation 1. We can show that if f is differentiable and x^* is a local minimizer of f, then the following condition holds

$$\nabla f(x^*) = 0 \tag{21}$$

Proof -
Let's consider a point x^* which is a local minimizer of f.
By definition of a local minimizer, there exists a radius $r > 0$ such that for all $x \in \mathbb{R}^d$ such that $\|x - x^*\| \leq r$, we have $f(x^*) \leq f(x)$.
This means that for $h \in \mathbb{R}^d$ such that $\|h\| \leq r$, we have

$$f(x^*) \leq f(x^* + h) = f(x^*) + \nabla f(x^*)^T h + o(\|h\|)$$

Simplifying by $f(x^*)$ and dividing by $\|h\|$ (which is non zero as h is non zero), we obtain,

$$\nabla f(x^*)^T \overline{h} \geq 0$$

with $\overline{h} = \frac{h}{\|h\|}$.
We can apply the same reasoning with $-h$ instead of h and obtain $\nabla f(x^*)^T (-|(h)) \geq 0$, which gives

$$\nabla f(x^*)^T \ |(h) \leq 0$$

Eventually, we have $\nabla f(x^*)^T \ |(h) = 0$ for all $|(h)$ such that $\||(h)\| = 1$, and thus

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

■

Note that this is a necessary condition for optimality but is not sufficient (e.g. saddle points).

For constrainted problems, the first order optimality conditions are more complex and will no be considered in this course.

III. Gradient Descent

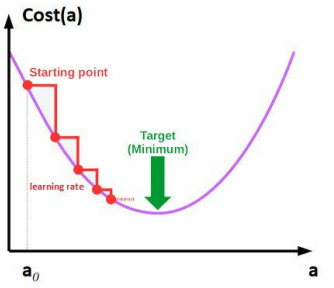
III.A. Algorithm

The gradient descent algorithm is an iterative method to solve the minimization problem Equation 3 when the function f is differentiable. The algorithm starts from an initial point $x_0 \in \mathbb{R}^n$ and iteratively updates the current point x_k using the formula

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \tag{22}$$

where $\alpha_k > 0$ is the step size at iteration k. The algorithm stops when a stopping criterion is met, for instance when the maximum number of iterations is reached or when the norm of the gradient is below a certain threshold. The complete algorithm is summarized below.

procedure Gradient Descent(x_0, f)
    ▷ Initialize the solution
    $x \leftarrow x_0$
    while x_0 is not optimal **do**
        ▷ Compute the gradient at the current solution
        $g \leftarrow \nabla f(x)$
        ▷ Choose a step size
        $\alpha \leftarrow$ **some method**
        ▷ Update the solution
        $x \leftarrow x - \alpha g$
    end
end



,

Algorithm A: Gradient Descent