# Optimization Methods for Neural Networks

Danilo Comminiello

Neural Networks 2023/2024

October 4, 2023

Sapienza
Università di Roma

# Table of contents

## 1 Elements of Unconstrained Convex Optimization

Preliminary definitions

Existence of Minima

# Optimization problem

An Optimization problem is defined as the minimization or maximization of a real-valued function called the cost function (CF).

The CF may be subject to equality and inequality constraints which delimit the space of the feasible region of solutions.

The combination of CF and constraints determines a system of equations and inequalities that describe the OP.

Based on characteristics and properties of the CF and its constraints, an OP can be solved by using several optimization methods, which can be: *linear* or *nonlinear*, *convex* or *non-convex*, *continuous* or *discrete*, *integer* or *non-integer*, *derivative* or *derivative-free*, *constrained* or *unconstrained*, *single-* or *multi-objective*.

# Loss and cost functions

The loss function quantifies the deviation/error between the measured value of $y$ and that which is predicted, using the corresponding measurement $\mathbf{x}$, i.e., $f_{\boldsymbol{\theta}}(\mathbf{x})$.

In a more formal way, we first adopt a nonnegative (loss) function: $\mathcal{L}(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \longmapsto [0, \infty)$.

Then, $\boldsymbol{\theta}_*$ is computed in order to minimize the total loss, or also called the **cost function**, over all the data points, i.e.:

$$f(\cdot) = f_{\boldsymbol{\theta}_*}(\cdot) \Rightarrow \boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta} \in \mathcal{A}} J(\boldsymbol{\theta}),$$

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} \mathcal{L}\left(y_n, f_{\boldsymbol{\theta}}(\mathbf{x}_n)\right),$$

assuming that a minimum exists.

In general, there may be more than one optimal values $\boldsymbol{\theta}_*$, depending on the shape of $J(\boldsymbol{\theta})$.

# Nonlinear programming

Among the most popular procedures for the determining the solution of an OP we find:

- linear programming, whose CF is linear and its constraints define a polytope;
- convex programming, whose CF and constraints are both convex.

We focus on a more general case of CFs not necessarily convex.

In particular, we refer to a class of optimization methods denoted as **nonlinear programming** (NLP), which is characterized by the following properties:

- the minimization (or maximization) of a CF is defined over real variables;
- variables are subject to a set of equalities and inequalities;
- the CF or some of the constraints are nonlinear.

# Definition of an optimization problem

Let us consider a cost function (CF) (or *objective function*) $J(\cdot) : \Omega \subseteq \mathbb{R}^M \rightarrow \mathbb{R}$, and an $M$-dimensional vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^M$.

An (unconstrained) optimization problem is defined as:

$$\boldsymbol{\theta}_* = \min_{\boldsymbol{\theta} \in \Omega} J(\boldsymbol{\theta}) \tag{1}$$

where $\Omega$ represents the *feasible region* or *feasible set* containing all the possible OP solutions and delimited by the set of OP constraints.

Minimizing a function is equivalent to maximizing it:

$$\max_{\boldsymbol{\theta} \in \mathbb{R}^M} J(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta} \in \mathbb{R}^M} -J(\boldsymbol{\theta})$$

# Unconstrained and constrained optimization

If the set $\Omega$ coincides with the entire space $\mathbb{R}^M$, i.e., the feasible region is an open set, the OP is said to be **unconstrained** and defined as:

$$\boldsymbol{\theta}_* = \min_{\boldsymbol{\theta} \in \mathbb{R}^M} J(\boldsymbol{\theta}) \tag{2}$$

If $\Omega \subset \mathbb{R}^M$, the region of feasible solutions is delimited by a set of *equality* and/or *inequality* constraints on the decision variables.
In this case, the OP is said to be **constrained** and defined as:

$$\begin{aligned} \boldsymbol{\theta}_* = &\min_{\boldsymbol{\theta} \in \Omega} \quad J(\boldsymbol{\theta}) \\ \text{s.t.} \quad &\mathbf{g}(\boldsymbol{\theta}) \leq 0 \\ &\mathbf{h}(\boldsymbol{\theta}) = 0 \end{aligned} \tag{3}$$

where $\mathbf{g}(\cdot) : \mathbb{R}^M \to \mathbb{R}^P$ refers to the set of *inequality constraints* and $\mathbf{h}(\cdot) : \mathbb{R}^M \to \mathbb{R}^Q$ denotes the set of *equality constraints*.

# Convex functions

A fundamental concept for the solution of an OP is the study of its convexity.

A function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ is **convex** if, chosen any two points of the function, $x_1, x_2 \in \mathbb{R}^n$, any point of that function between the two chosen extremes always lies *below* the segment line connecting them:

$$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2), \qquad \forall \lambda \in [0,1] \tag{4}$$

A function is *strictly convex* if $\forall x_1 \neq x_2 \in \mathbb{R}^n$, $\forall \lambda \in [0,1]$:

$$f(\lambda x_1 + (1-\lambda) x_2) < \lambda f(x_1) + (1-\lambda) f(x_2)$$

A function $f$ is *strongly convex* if $\forall x_1, x_2 \in \mathbb{R}^n$, $\forall m > 0, \lambda \in [0,1]$:

$$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) - \frac{1}{2} m \lambda (1-\lambda) \|x_1 - x_2\|_2^2 .$$

# Concave functions

Concave functions are simply the negative of convex functions, i.e., their definition comes out simply by reversing the direction of the inequality. Strict concavity is defined analogously.

A function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ is **concave** if, chosen any two points of the function, $x_1, x_2 \in \mathbb{R}^n$, any point of that function between the two chosen extremes always lies *above* the segment line connecting them:

$$f(\lambda x_1 + (1-\lambda) x_2) \geq \lambda f(x_1) + (1-\lambda) f(x_2), \qquad \forall \lambda \in [0,1] \tag{5}$$

A function is *strictly concave* if $\forall x_1 \neq x_2 \in \mathbb{R}^n$, $\forall \lambda \in [0,1]$:

$$f(\lambda x_1 + (1-\lambda) x_2) > \lambda f(x_1) + (1-\lambda) f(x_2)$$

# Examples of convex and concave functions

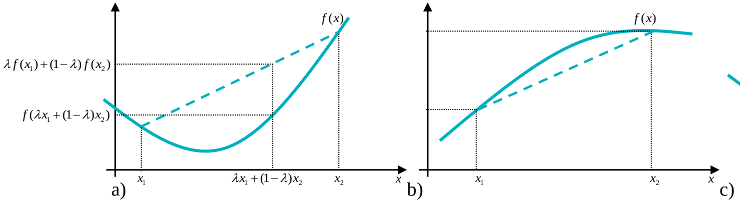


Figure 1: Examples of function: a) convex, b) concave, c) nonconvex (and nonconcave) [1].

Simple examples of convex functions are $x^{2p}$, $p = 1, 2, ...$; $e^x$, $e^{-x}$ or $-\lg x$.

Moreover, multiplying each example by $-1$ one gives a concave function.

The definition of convexity implies that the sum of convex functions is convex and that any nonnegative multiple of a convex function also is convex.

# Global and local minima

## Definition 1 (Global minimum)

A point $\boldsymbol{\theta}_*$ is a global minimum for function $J(\boldsymbol{\theta})$ if:

$$J(\boldsymbol{\theta}_*) \leq J(\boldsymbol{\theta}), \ \forall \boldsymbol{\theta} \in \mathbb{R}^M \tag{6}$$

It is a local minimum if (6) holds only for an $\varepsilon$-radious ball centered in $\boldsymbol{\theta}_*$. It is a strict minimizer if (6) holds without equality.

Without assuming a specific structure of $J(\cdot)$, an algorithm can only search for local minima. In our case, we only assume that the CF is twice differentiable and bounded.

## 2 Gradient and Hessian

# Gradient and Hessian

## Definition 2 (Gradient)

The gradient $\nabla J(\boldsymbol{\theta}) \in \mathbb{R}^M$ of $J(\boldsymbol{\theta})$ is defined as:

$$\nabla J(\boldsymbol{\theta}) = \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial J(\boldsymbol{\theta})}{\partial \theta_M}\right]^{\mathsf{T}} \tag{7}$$

## Definition 3 (Hessian)

The Hessian matrix $\nabla^2 J(\boldsymbol{\theta}) \in \mathbb{R}^{M \times M}$ of $J(\boldsymbol{\theta})$ is defined as:

$$\nabla^2 J(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \left[\nabla J(\boldsymbol{\theta})\right]^{\mathsf{T}} = \begin{bmatrix} \frac{\partial^2 J(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 J(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_M} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 J(\boldsymbol{\theta})}{\partial \theta_M \partial \theta_1} & \cdots & \frac{\partial^2 J(\boldsymbol{\theta})}{\partial \theta_M \partial \theta_M} \end{bmatrix} \tag{8}$$

# Example of gradient and Hessian computation

## Example 1

**Problem:** Given $f(x_1, x_2) = x_1^2 + 3x_1x_2$, find $\nabla f$ and $\nabla^2 f$.

**Solution:**

$$\nabla f(x_1, x_2) = \begin{bmatrix} 2x_1 + 3x_2 & 3x_1 \end{bmatrix}^\mathsf{T}$$

$$\nabla^2 f(x_1, x_2) = \begin{bmatrix} 2 & 3 \\ 3 & 0 \end{bmatrix}$$

# Stationarity and nonnegativity

## Definition 4 (Stationary point)

A point $\boldsymbol{\theta}_*$ is a stationary point of $J(\boldsymbol{\theta})$ if:

$$\nabla J(\boldsymbol{\theta}_*) = 0 \tag{9}$$

## Definition 5 (Positive definiteness)

A matrix $\mathbf{S}$ is positive semidefinite if:

$$\mathbf{a}^{\mathsf{T}}\mathbf{S}\mathbf{a} \geq 0,\ \forall \mathbf{a} \in \mathbb{R}^{M} \tag{10}$$

If (10) holds without equality, $\mathbf{S}$ is positive definite.

# Necessary and sufficient conditions for optimality

**Theorem 1 (Necessary optimality conditions)**

*If a point* $\boldsymbol{\theta}_*$ *is a local minimum then it is a stationary point and the Hessian matrix evaluated at* $\boldsymbol{\theta}_*$ *is positive semidefinite.*

**Theorem 2 (Sufficient optimality conditions)**

*If a point* $\boldsymbol{\theta}_* \in \mathbb{R}^M$ *is a stationary point and the Hessian matrix evaluated at* $\boldsymbol{\theta}_*$ *is positive definite, then* $\boldsymbol{\theta}_*$ *is a strict local minimum.*

**Proof.**

For both the above conditions, see proofs in [2, Section 2.1] □

$$f(\mathbf{w}) = \frac{w_1^2 + w_2^2}{20}$$

$$\frac{\partial f(\mathbf{w})}{\partial w_1} = \frac{w_1}{10} \qquad \frac{\partial f(\mathbf{w})}{\partial w_2} = \frac{w_2}{10}$$

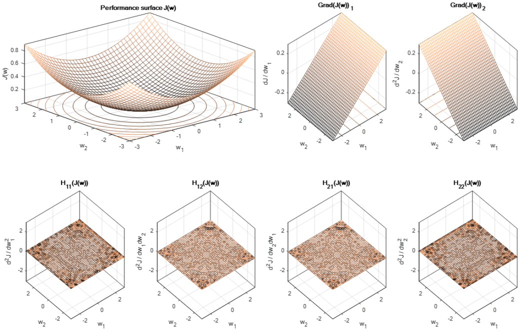


Figure 2: Gradient and Hessian of a simple quadratic function.

# Examples of gradient and Hessian II

$$J(\mathbf{w}) = \frac{w_1^2 + w_2^2}{20} + \sin(w_1)\cos(w_2)$$

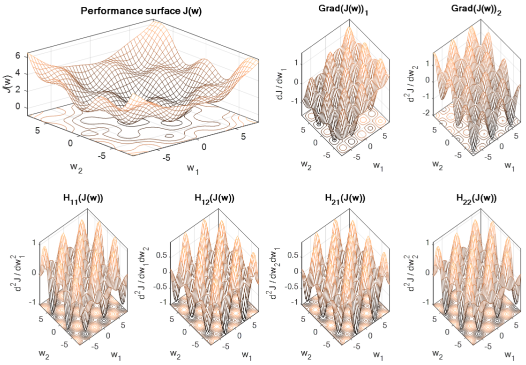


Figure 3: Gradient and Hessian of a complex quadratic function.

$$J(\mathbf{w}) = \frac{w_1^2 + w_2^2}{20} + w_1 e^{-(w_1^2 + w_2^2)}$$

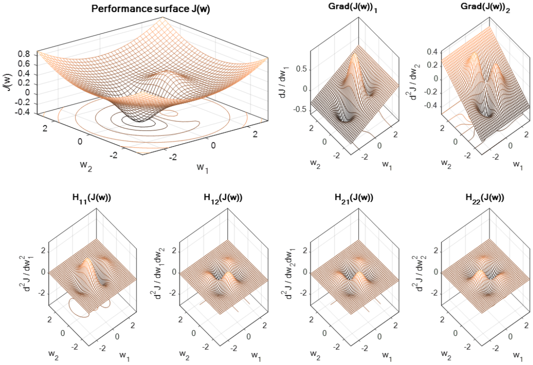


Figure 4: Gradient and Hessian of a complex quadratic function.

# Examples of gradient and Hessian IV

$$J(\mathbf{w}) = 3w_1^2 + 2w_1 w_2 + w_2^2 - 4w_1 + 5w_2 - w_1^2 w_2 + 0.3 w_1 w_2^2$$

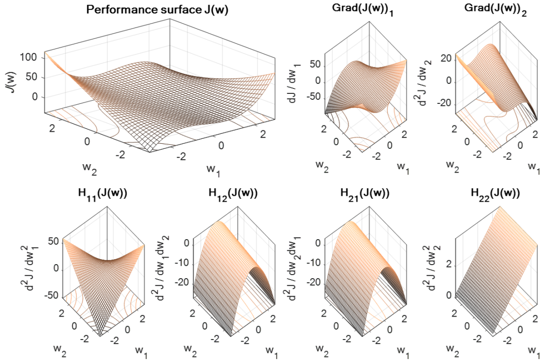


Figure 5: Gradient and Hessian of a complex quadratic function.

## 3 Least-Square Cost Function

Definition of the Least-Square Cost Function

Advantages of the Least-Square Method

Optimization Algorithms

# The squared error loss function

The squared error loss function is defined as:

$$\mathcal{L}\left(y, f_{\boldsymbol{\theta}}\left(\mathbf{x}\right)\right) = \left(y - f_{\boldsymbol{\theta}}\left(\mathbf{x}\right)\right)^2 \tag{11}$$

and it gives rise to the cost function corresponding to the total (over all data points) squared-error cost function:

$$J\left(\boldsymbol{\theta}\right) = \sum_{k=1}^{K} \left(y_k - f_{\boldsymbol{\theta}}\left(\mathbf{x}_k\right)\right)^2$$

The minimization approach based on the previous cost function is known as the **Least-Square** (LS) **method**, which was first introduced and used by Gauss.

# Uniqueness of the solution

The most important characteristic of the LS loss is the uniqueness of the minimization solution, which is due to the strict convexity of its parabolic shape.

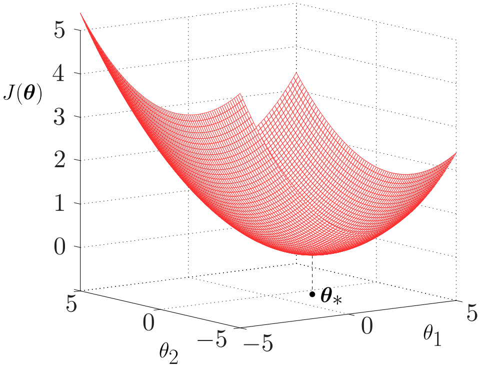


Figure 6: The least-square loss function has a unique minimum at the point $\boldsymbol{\theta}_*$ [3]. It is readily observed that the graph has a unique minimum.

# Least-square method and linear models

The use of the LS method together with linear models has a number of *computational advantages* that makes it one among the most popular techniques in machine learning.

More specifically:

- The minimization leads to a unique solution in the parameters' space.
- The optimal set of the unknown parameters is given by the solution of a linear system of equations.

Moreover, understanding linearity is very important.
Treating nonlinear tasks, most often, can be turned out to finally resort to a linear problem.

# Optimization algorithms

- Gradient Descent
- Stochastic Gradient Descent
- Mini-Batch Gradient Descent
- Momentum
- Nesterov Accelerated Gradient
- AdaDelta
- Adam (Adaptive Momentum Estimation)

# Next lecture

- We focus on supervised learning
- We will introduce the linear regression and linear classification.
- Fully-connected networks will be also discussed.
- We will see how we optimize the training of a network with automatic differentiation.

# References


[1] A. Uncini, "Introduction to adaptive algorithms and machine learning." 2018.

[2] J. Nocedal and S. Wright, *Numerical optimization*.
2nd ed., 2006.

[3] S. Theodoridis, *Machine Learning. A Bayesian and Optimization Perspective*.
Elsevier, 2nd ed., 2020.

[4] D. P. Bertsekas, *Nonlinear Programming*.
Athena Scientific, 2nd ed., 1999.

# Optimization Methods for Neural Networks

Neural Networks 2023/2024

**Danilo Comminiello**

https://sites.google.com/uniroma1.it/neuralnetworks2023

{danilo.comminiello, simone.scardapane}@uniroma1.it