Introduction to optimization

Georgios Ropokis (georgios.ropokis@centralesupelec.fr)

CentaraleSupélec, Campus Rennes

Table of contents

What does optimization refer to?

Defining an optimization problem

A first definition of an optimization problem

We define an optimization problem as the mathematical formulation of the problem of selecting the value of some system parameters/decision variables such as to maximize some reward function or minimize some cost function.

What characterizes an optimization problem?

- The objective: Do we wish to maximize or minimize some target function?
- The set or space over which our system parameters/decision variables take values (e.g., binary/integer/real).
- The presence/absence of additional constraints on the values of these system parameters/decision variables.
- The type of functions that formulate the reward function or cost function (e.g., linear functions, quadratic functions, convex functions, concave functions etc) as well as the constraints.

Minimization problems

Definition of a minimization problem

We refer to a problem as a minimization problem when it is of the form:

$$\text{Find } \hat{x} \in D \text{ such that } (\forall x \in D)\, f(\hat{x}) \leq f(x) \tag{1}$$

where $D \subseteq \mathbb{R}^N$.

The cost function

We normally refer to function $f(x)$ as the cost function of the problem or the objective function of the problem.

The decision variables

We refer to the vector x as the vector of decision variables.

The feasibility set and the problem constraints

We refer to the set D as the constraint set or the feasible set. In the majority of optimization problems, the feasible set is determined by a system comprising of equalities and/or inequalities involving the decision variables. We refer to them as the equality/inequality constraints for the problem.

Maximization problems

Definition of a maximization problem

We refer to a problem as a maximization problem when it is of the form:

$$\text{Find } \hat{x} \in D \text{ such that } (\forall x \in D) f(\hat{x}) \geq f(x) \tag{2}$$

where $D \subseteq \mathbb{R}^N$.

The reward or utility function

We often refer to function $f(x)$ as the reward or utility function of the problem.

Equivalent formulation as a minimization problem

$$\text{Find } \hat{x} \in D \text{ such that } (\forall x \in D) - f(\hat{x}) \leq -f(x) \tag{3}$$

We therefore choose to formally treat only minimization problems!

Minimization problems and minimizers

Global minimizers

A point $\hat{x} \in D$ is a global minimizer of f over D if $f(x) \geq f(\hat{x})$ for all $x \in D \backslash \{\hat{x}\}$

We denote all such global minimizers as:

$$Argmin_{x \in D} f(x) \tag{4}$$

Local minimizers

A point $\tilde{x}$ is a local minimizer of f over D if there exists a neighborhood around $\tilde{x}$ such that $f(x) \geq f(\tilde{x})$ for all x that belong to D as well as the specific neighborhood around $\tilde{x}$.

Strict minimizers

If we replace $\geq$ with $>$ in the above definitions we then discuss about strict global and strict local minimizers.

Categories of optimization problems

Unconstrained optimization

The N-dimensional decision vector can take any value in $\mathbb{R}^N$, i.e., $D = \mathbb{R}^N$.

Constrained optimization

- Constrained continuous optimization problems: $D \subset \mathbb{R}^N$. D is normally described by means of a system of equality and inequality constraints.
- Discrete optimization: D is a countable set. The presence of equality and inequality constraints in determining D can also be encountered.
  - If D is finite we refer to the optimization problem as a combinatorial optimization problem.
  - If $D \subset \mathbb{Z}^N$ the problem is characterized as an integer optimization problem.

What makes optimization so interesting?

Applications of optimization

Data processing and machine learning

- Data fitting
- Neural network training
- Reinforcement learning

Signal processing and Communications

- Detection and parameter estimation
- Prediction
- Compression
- Optimization of communications systems

Resource management

- Process planning and decision making

Some examples of optimization problems

Optimal resource management using linear programming [1]

A firm produces n different goods using m different raw materials. Let $b_i, i = 1, ..., m$, be the available amount of the i-th raw material. The j-th good, $j = 1, ..., n$, requires $a_{i,j}$ units of the i-th material and results in a revenue of c_j per unit produced. The firm faces the problem of deciding how much of each good to produce in order to maximize its total revenue.

Let $x_j, j = 1, .., n$, be the amount of the j-th good. The problem can then be formulated as follows:

$$\begin{aligned} &\text{maximize: } c_1 x_1 + \ldots + c_n x_n \\ &\text{subject to: } a_{i,1} x_1 + \ldots a_{i,n} x_n \leq b_i, \quad i = 1, \ldots, n, \\ &x_j \geq 0, \quad j = 1, \ldots, n \end{aligned} \tag{5}$$

[1] Example taken from "Introduction to Linear Optimization" by Dimitris Bertsimas and John N. Tsitsiklis, Athena Scientific Publishing

Non linear constrained programming problems

Non linear constrained programming

In case of a non linear objective function, and/or a constraint set that is specified with the aid of non linear equalities and inequalities, we refer to the resulting optimization problem as a non linear programming problem.

Understanding the limits of a communications systems

From an information theoretic point of view, one of the most important metrics for the performance of a communications systems is the capacity of the system, defined as the maximum transmission rate that the system can support.

From a theoretical perspective, the capacity is found to be the maximum of the mutual information of the communications system's input and output. Calculating the capacity involves solving a non linear programming problem!

Optimal communication over parallel frequency bands

One of the classical communications system model is the model of a frequency flat fading channel. By adopting such a channel model, when a frequency band of bandwidth B is given, the maximum amount of information that we can transmit is equal to:

$$R = B \log_2 \left(1 + \frac{gP}{\sigma^2}\right) \tag{6}$$

where:

- g: The power attenuation introduced by the channel
- P: The transmit power
- σ^2: The noise variance at the receiver.

Question: Given N parallel frequency channels of bandwidth B, where each one is characterized by a power attenuation $g_n, n = 1, \ldots, N$, and a total power budget P_{tot}, how much power should we use on each one of the frequency bands, such as to maximize the transmit rate? (Assume that the all frequency bands suffer from the presence of noise of the same variance.)

Optimal communication over parallel frequency bands

By transmitting different data streams from each frequency band, and using a different power level on each frequency band, we can achieve a total transmission rate of:

$$R = \sum_{n=1}^{N} B \log_2 \left(1 + \frac{g_n P_n}{\sigma^2}\right) \tag{7}$$

where each one of the terms corresponds to the data rate (in bits/sec) for the corresponding frequency band.

The optimal power allocation problem

$$\begin{aligned} \text{maximize:}_{P_1,\ldots,P_n} \quad & B \sum_{n=1}^{N} \log_2 \left(1 + \frac{g_n P_n}{\sigma^2}\right) \\ \text{subject to:} \quad & -P_n \leq 0, \quad n = 1, \ldots, N \\ & \sum_{n=1}^{N} P_n - P_{tot} \leq 0 \end{aligned} \tag{8}$$

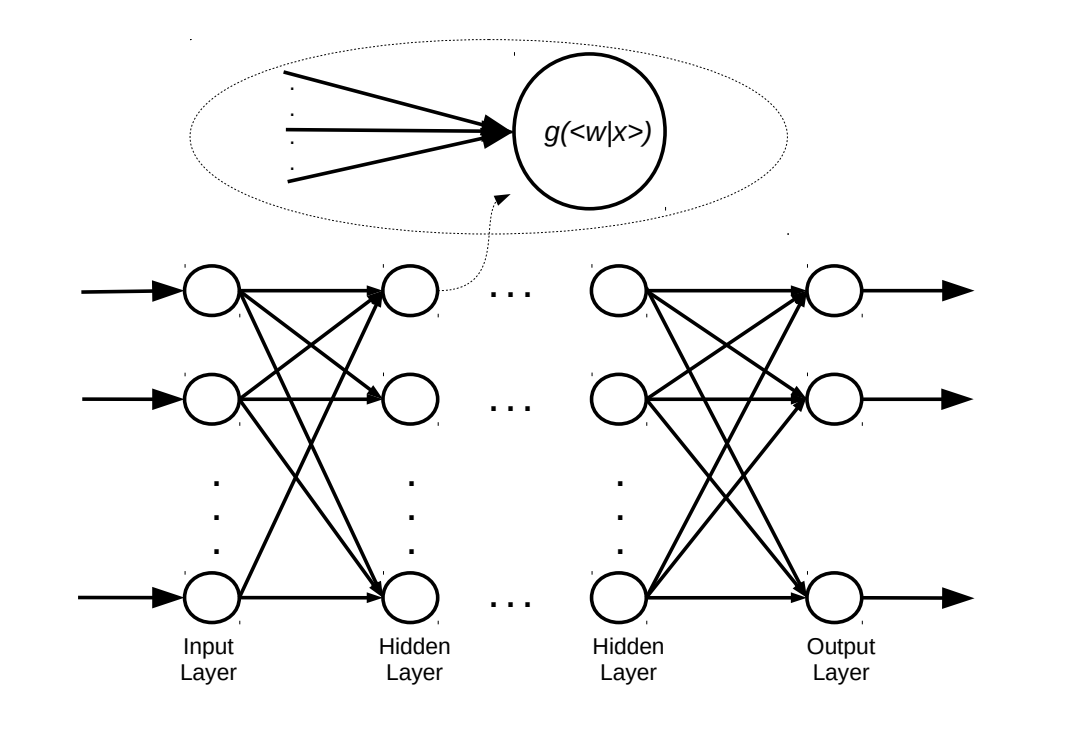
Binary optimization problems in Communications

Optimal caching in wireless/mobile communications systems

Let us consider a library of L files (e.g., the library of Netflix content) as well as the so-called popularity distribution for this library (i.e., the probability that a particular file of this library is requested). Let $\mathcal{S}_{BS}$ be a set of M base stations covering a specific area with K users. Let $d_{k,m,l}$ denote the average download time that user k would experience when downloading file l from base station m. Assuming that each base station can cache $L_c < L$ files (caching takes place in order to avoid unnecessary access to the core network), which data files should be stored at each base station?

Formulating the problem as a binary optimization problem

Let us introduce a matrix of binary decision variables $X \in \{0,1\}^{L \times M}$ where the element of the l-th raw and m-th column takes a value of zero if the l-th file is not cached at the m-th base station and a value of one otherwise. Let $D(X)$ be the average download time corresponding to the caching configuration indicated by matrix X. The optimal caching problem is then formulated as a problem of finding the optimal binary matrix X. The constraint that at most L_c files can be stored at each base station can then be mathematically expressed as the constraint that the sum of all elements of each column is at most equal to L_c.



How do we optimally determine the weight vector w for each one of the neurons? (The backpropagation algorithm)

Linear regression

Given a dataset of the form $(x_i, y_i), i = 1, \ldots, I$, with $x_i \in \mathbb{R}^K, K \geq 1$ and $y_i \in \mathbb{R}$ how can we learn how to predict the value of the parameter y when we simply observe x, using a linear predictor?

Least squares approximation

Design a linear predictor a by minimizing the Mean Square Error, i.e., by solving the following optimization problem:

$$\text{minimize:}_{a \in \mathbb{R}^K} \quad \frac{1}{2I} \sum_{i=1}^{I} \left(y_i - \langle a | x_i \rangle\right)^2 \tag{9}$$

稀疏性.

Sparse linear regression

In several cases we wish to create a sparse linear regression model (for example in order to do parameter/feature selection). A formulation of the regression problem that allows us to learn such sparse models is the following:

$$\text{minimize:}_{a\in\mathbb{R}^K} \quad \sum_{i=1}^{I} \left(y_i - \langle a|x_i\rangle\right)^2, \quad \text{subject to:} \quad \|a\|_1 \leq t \tag{10}$$

where t a predetermined parameter that allows to control sparsity.

Robust linear regression

- In case that a specific data point x_i, y_i deviates a lot from our assumption of a linear relation connecting x_i and y_i, a high absolute error $|y_i - \langle a|x_i\rangle|$ is expected.
- The squaring operation involved in calculating the Mean Square Error results in a strong contribution of this specific term in the total error
- Solution: Use the Mean Absolute Error as a criterion for determining the optimal predictor a, by solving the following problem:

$$\text{minimize:}_{a\in\mathbb{R}^K} \quad \frac{1}{I}\sum_{i=1}^{I} |y_i - \langle a|x_i\rangle| \tag{11}$$

- The resulting problem can be expressed in a linear programming form:

$$\begin{aligned} &\text{minimize:}_{a\in\mathbb{R}^K, t_i\in\mathbb{R}, i=1,\dots I} \quad \frac{1}{I}\sum_{i=1}^{I} t_i \\ &\text{sublect to} \quad t_i \geq 0, \quad -t_i \leq y_i - \langle a|x_i\rangle \leq t_i \end{aligned} \tag{12}$$

Scope of the course

- Can we establish criteria for determining whether a solution exists and if it is unique?
- Can we derive analytical methods for finding the solution to optimization problems?
- If not, can we derive *iterative* algorithms converging to the solution of an optimization problem? 迭代.
- How can we assess an optimization algorithm:
  - Speed of convergence 收敛速率.
  - Robustness to numerical errors
  - Possibility for parallel/distributed implementation

并行/分布.

Course outline

- Introduction
- Existence of minimizers
- Convexity
- Duality
- Linear programming
- Integer linear programming
- Lagrange multipliers method
- Some iterative algorithms
- Stochastic Optimization

Recommended references

1. D. Bertsekas, Nonlinear programming, Athena Scientic, Belmont, Massachussets, 1996.
2. Y. Nesterov, Introductory Lectures on Convex Optimization: A Basic Course, Springer, 2004.
3. S. Boyd and L. Vandenberghe, Convex optimization, Cambridge University Press, 2004.
4. H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer, New York, 2017.
5. H. P. Williams, Model Building in Mathematical Programming, Wiley, 2013.