Modern Optimization

Jean-Luc Bouchot

School of Mathematics and Statistics
Beijing Institute of Technology
jlbouchot@bit.edu.cn

Spring 2021

Outline

1. Introduction

2. Calculus-free optimization

Remark 2.1

Optimization is the science of finding a parametrization or arrangement which fits a certain criteria at least as good as any other criteria. This relies on the definition of a *measure of fitness* or *mismatch*. We will denote this measure as a function f_0. An optimization problem is then written

$$\min f_0(x)$$

Here we are dealing with a multi-dimensional variable $x \in \mathbb{R}^d$ and a real-valued function $f_0 : \mathbb{R}^d \to \mathbb{R}$.

Definition 2.1

The function we are trying to optimize is called the objective function. We may refer to it as loss function or cost (both are used in minimization problems). It may also be referred to as fitness function, revenue, utility function in maximization problems. Either way, this refers to our function f_0 *from the previous section.*

Definition 2.2

A problem may be constrained or unconstrained. An unconstrained problem lets the variable x move freely in its space $\mathbb{R}^n$ of definition:

$$\text{minimize}\, f_0(x),$$

where $x \in \mathbb{R}^n$. These problems are, from my experience, rather rare and not the most interesting to us. A constrained problem restricts the movement of the variable x. It may do so mainly via two (equivalent) means:

1. *Restricting its domain to a subset: physical constraints tell us that some variables cannot be larger than something or smaller than something else.*
2. *Restricting the variables due to some outside constraints. For instance, we might have limited financial resources and the costs associated to some materials is given by a function $f_1(x)$. In this case, we will need to enforce $f_1(x) \leq B$ to respect the budget, while minimizing our criteria f_0.*

Of course, looking at the first or the second issue is just a matter of view point and both of them can be interchanged easily. Overall, a constrained optimization problem may be written as

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\, f_0(x), \quad \textit{subject to } f_i(x) \leq 0.$$

The f_i's are called the constraint functions or, simply, the constraints.

Definition 2.3

- *x is generally called the free variable.*
- *An optimization problem can be convex or non convex*
- *It may be linear or non linear (these two terms will be explained soon)*

Example 2.1

Change the (seemingly) unconstrained optimization

$$\underset{x \in B_1^{(2)}(x_0)}{\text{minimize}} \|Ax - b\|_2$$

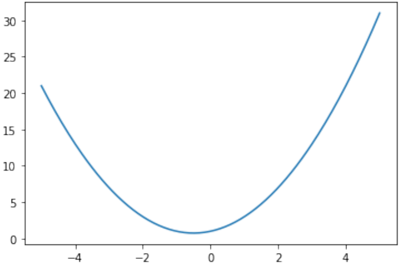
where $B_1^{(2)}(x_0)$ defines the unit ball centered at x_0 and measured in the ℓ_2 norm, into a clearly constrained problem.

Example 2.2

Let us look at the simple case of minimizing a quadratic function:

$$\underset{x \in \mathbb{R}}{\text{minimize}}\, f_0(x)$$

where $f_0(x) = x^2 + x + 1$.



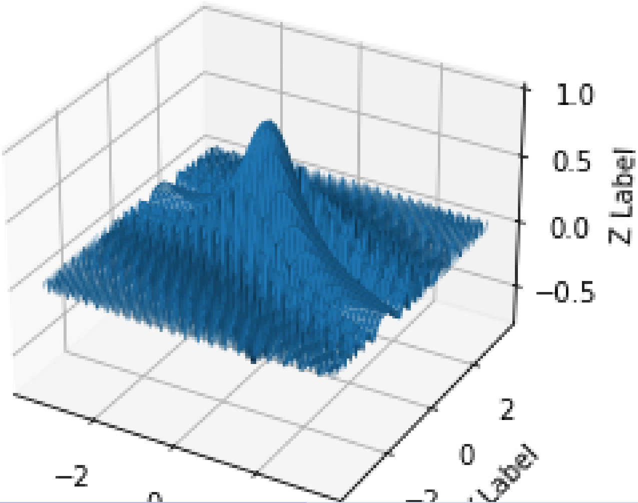
Example 2.3 (In class)

Solve the following problems:

1. Find the minimum of $f(x,y) = \frac{12}{x} + \frac{18}{x} + xy$.
2. Find the maximum of $f(x,y) = xy(72 - 3x + 4y)$.
3. Find the minimum of $f(x,y) = 4x + \frac{x}{y^2} + \frac{4y}{x}$.

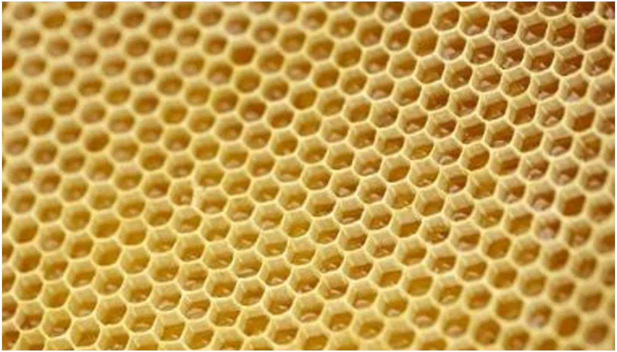
Example 2.4

If we were to look at this



Example 2.5

Honeycombs structure



Example 2.6

(other) Optimal shapes appear in other areas: Soap bubbles minimizing the potential energy !

Example 2.7

And now both of them!

Example 2.8

Optimization is used pretty much every where in our lives, knowingly or not:

- Logistics: a business has warehouses, each having stocks of goods and linked to one another with roads having a certain cost. These goods need to be supplied to the factory sites, and then further to the customers. Which warehouse stores which material? Which factory produce which goods? How many of them? Where should these factories be so that the customers have a faster and / or cheaper access to the goods? (An area of *Operation research* in which some algorithms such as the Hungarian algorithm or the Hopcroft-Karp approach can be used)
- A driver needs to go to a certain location. The GPS computes the fastest route ... or the cheapest ... or the most beautiful. This is usually computed by means of Traveling Salesman Problems for which Dijkstra's algorithm or the A-star algorithm can be used.
- Recommender systems: based on your previous historical data, can we predict your next purchase / movie to watch / webpage to visit? Some approaches related to non-negative matrix factorization can be used, or sparse and low-rank optimization.
- Machine learning: find the optimal parameters of a certain system. This is usually a regression problem (responses to a continuous variable) for which the *classical* methods can be used such as Gradient descent and other

Remark 2.2

This course will be organized along the following topics: (more or less in order)

1. Calculus-free optimization examples.
2. Geometric programming.
3. Local optimization: These methods can be applied to any unconstrained problems. But we'll see their speed of convergence and their limitations.
4. Convex optimization This will be the last *pure maths* topic. We will also review the convergence of the various algorithms introduced above to the convex problems.
5. Constrained optimization (convex and non convex)
6. Proximal methods and projected gradients
7. Coordinate descent
8. Large scale optimization strategies

Outline

Remark 3.1

This could also have been called *high school optimization* as most of the tools used here are pretty much from a mathematical toolbox available in high school.
This goal of this section really is to give a taste for optimization problems and to remind ourselves that hard and unstable algorithms are not all there is.
Some problems require only careful thinking and can be dealt with easily. As a consequence we aim in this chapter at keeping the mathematics to the lowest level possible.

Proposition 3.1 (AGM inequality)

Let $x_1, \cdots, x_n$ be positive numbers. It holds

$$GM = (\prod_{i=1}^{n} x_i)^{1/n} \leq AM = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Moreover $AM = GM$ if and only if $x_1 = x_2 = \cdots = x_n$.

Example 3.1

An application of the AGM inequality: definition of e. Show that

$$\lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = e$$

Definition 3.1

A function $f : \mathbb{R} \to \mathbb{R}$ *is said to be convex if for any* $\alpha \in [0, 1]$ *and any two points in the domain* x_1, x_2*, we have*

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2).$$

Exercise 3.1

Prove the following proposition:

Proposition 3.2 (Homework)

The convexity of a function can equivalently be written as: If f is convex, then for any $\alpha_1, \cdots, \alpha_n$ positive numbers such that $\alpha_1 + \cdots + \alpha_n = 1$, if holds

$$f\left(\sum_i \alpha_i x_i\right) \leq \sum_i \alpha_i f(x_i).$$

Exercise 3.2

Prove the following result:

Proposition 3.3 (Generalized AGM Inequality – Homework)

Let $x_1, \cdots, x_n$ *be positive numbers and let* $\alpha_1, \cdots, \alpha_n$ *be positive numbers such that* $\alpha_1 + \cdots + \alpha_n = 1$*. Then the GAGM inequality reads*

$$\prod_{i=1}^n x_i^{\alpha_i} \leq \sum_{i=1}^n \alpha_i x_i.$$

Example 3.2 (In class)

Solve the following problems:

1. Find the minimum of $f(x,y) = \frac{12}{x} + \frac{18}{x} + xy$.
2. Find the maximum of $f(x,y) = xy(72 - 3x + 4y)$.
3. Find the minimum of $f(x,y) = 4x + \frac{x}{y^2} + \frac{4y}{x}$.

Exercise 3.3

Use the generalized AGM to prove Hölder's inequality:

Proposition 3.4 (Höder's inequality – Homework)

For two sequence of numbers $\{a_k, 1 \leq k \leq n\}$ *and* $\{b_k, 1 \leq k \leq n\}$, *Hölder's inequality reads, for* p, q *such that* $\frac{1}{p} + \frac{1}{q} = 1$ *(we say that* p *and* q *are Hölder conjugates)*

$$\sum_{k=1}^{n} |c_k d_k| \leq \|c\|_p \|d\|_q$$

with equality if and only if

$$\left(\frac{|c_k|}{\|c\|_p}\right)^p = \left(\frac{|d_k|}{\|d\|_q}\right)^q, \qquad \textit{for all } k.$$

And use this result to prove the triangle inequality for general p norms (so called Minkowksi's inequality).

Proposition 3.5 (Cauchy-Schwarz – Admitted)

The case $p = q = 2$ of Hölder's inequality gives rise to Cauchy-Schwarz inequality which reads

$$|\langle u, v, |\rangle \leq \|u\| \|v\|$$

with equality if and only if u and v are colinear.

Example 3.3 (in class)

Find the min and max of $f(x,y,z) = 2x + 3y + 6z$ on the unit sphere.

Example 3.4 (The BLUE – in class)

This example studies the *Best Linear Unbiased Estimator*. Assume we are trying to compute a robust estimate of an unknown value x. Assume we are given n measurements of that value $y_i = x + \varepsilon_i$ where the ε_i are random variables with $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma_i^2 > 0$. Assume that the noise elements are uncorrelated: $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ for $i \neq j$.
A **linear estimator** $\widehat{x}$ of x is one such that there exists coefficients λ_i with $\widehat{x} = \sum \lambda_i y_i$.

1. Show that, for an unbiased estimator ($\mathbb{E}[\widehat{x}] = x$), $\sum \lambda_i = 1$.
2. Find the BLUE estimator which minimizes the variance $\mathbb{E}[(\widehat{x} - x)^2]$.
3. What is the blue estimator in case of i.i.d. noise?