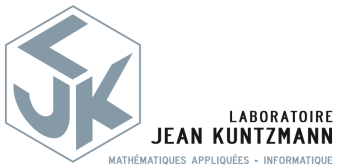

Lecture 1 : Introduction to optimization and refresher course

1 Course logistics

Below are three main points of this course:

1. **Calendar**: 4h30 per week, in which:
   - 1 TP at 9h45 AM in Monday morninng (IM2AG F202).
   - 1 lecture and 1 TD from 9h45 AM to 1h PM on Thursday (IM2AG F321).
2. **Main objectives**: This course aims to cover the most basic elements of numerical (continuous) optimization. Notable components are:
   - Theory of unconstrained and constrained optimization.
   - Convex optimization.
   - Optimization algorithms: (stochastic) gradient descent, acceleration, second-order methods, with theoretical guarantees.
   - Nonsmooth optimization, proximal operators.
3. **Evaluation**: one midterm (1h) and one final exam (2h). Both are written exams. Their contribution are 0.3 and 0.7.

2 Several examples of optimization problems

We consider three examples:

Linear regression We have a statistical serie $\{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} \mid i = 1, \dots, n\}$. We suppose that the relation between x and y is linear, i.e., $y \approx \theta^\top x$, find the best θ to fit the given serie.

$$\underset{\theta \in \mathbb{R}^d}{\text{Minimize}} \quad \mathcal{L}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (x_i^\top \theta - y_i)^2 = \frac{1}{2n} \|\mathbf{X}\theta - y\|_2^2, \tag{1}$$

where

$$\mathbf{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \quad \|v\|_2^2 = \sum_{i=1}^{n} v_i^2, v \in \mathbb{R}^n.$$

Image denoising We have access to a noisy version of an image $\tilde{X} \in [0,1]^{d\times d}$ (assuming that the image is black and white). We would like to denoise $\tilde{X}$. We use a simple observation: in a natural image, close pixels tend to have similar values. Therefore, we can consider the following problems:

$$\underset{X\in[0,1]^{d\times d}}{\text{Minimize}} \quad \mathcal{L}(\mathbf{X}) = \|\mathbf{X} - \tilde{\mathbf{X}}\|_F^2 + \lambda \sum_{i,j} \sum_{(i',j')\in\mathcal{N}(i,j)} |\mathbf{X}_{i,j} - \mathbf{X}_{i',j'}| \tag{2}$$

where:

$$\|\mathbf{X}\|_F^2 = \sum_{i,j} \mathbf{X}_{i,j}^2$$

and $\mathcal{N}(i,j)$ is the set of pixels neighboring the pixel (i,j).

Diet problem - Linear programming We have n types of food, each of which provides m types of nutritions. In detail, each unit of the ith food provides $\mathbf{A}_{i,j}$ gram of the jth nutrition. A normal person needs b_j gram of the jth nutrition to live healthily. The ith food costs c_i euros per unit. Find the cheapest strategy for a man to have enough nutrition.

$$\begin{aligned} \underset{x\in\mathbb{R}^n}{\text{Minimize}} \quad & c^\top x \\ \text{subject to} \quad & \mathbf{A}x \geq b \\ & x \geq 0 \end{aligned}$$

The answer: cabbage, spinach, wheat flour, evaporate milk, beans (Stigler diet).

3 Refresher elements: gradient, hessian matrix, Taylor expansion, properties of optimal solutions

Definition 3.1 (Derivatives of a function)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at a point x if there exists a vector $\nabla f(x)$ such that:

$$\lim_{d\to 0} \frac{\|f(x+d) - f(x) - \langle \nabla f(x), d\rangle\|}{\|d\|} = 0,$$

or equivalently,

$$f(x+d) = f(x) + \langle \nabla f(x), d\rangle + r_f(d) \quad \text{where } \lim_{d\to 0} \frac{\|r_f(d)\|}{\|d\|} = 0.$$

The vector $\nabla f(x)$ is called the gradient of f at $x \in \mathbb{R}$. Moreover, the vector $\nabla f(x)$ is given by:

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_d}(x) \end{pmatrix} \in \mathbb{R}^d,$$

where each coordinate is given by the derivatives of the function $g_i : \mathbb{R} \to \mathbb{R} : x_i \to f(x_1, \ldots, x_i, \ldots, x_n)$, or equivalently:

$$\frac{\partial f}{\partial x_i}(x) = g_i'(x_i) = \lim_{t\to 0} \frac{f(x_1, \ldots, x_i + t, \ldots, x_d) - f(x_1, \ldots, x_i, \ldots, x_n)}{t}.$$

Remark 3.2. If f is differentiable at x, then f has to be continuous at x.

Example 3.3. For example, $f(x) = \|x\|_2^2$ is differentiable at $x = 0$, but the function $f(x) = |x|$ is not.

Definition 3.4 (Differentiable and continuously differentiable (C^1) functions)**.** If a function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at all point $x \in \mathbb{R}^d$, then f is differentiable. If the mapping $x \mapsto \nabla f(x) : \mathbb{R}^d \to \mathbb{R}^d$ is continuous, then we say that f is continuously differentiable, or a C^1 function.

Proposition 3.5 (Properties of derivatives and gradient)**.** *Given two differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$, we have:*

$$\begin{aligned} \nabla(f+g)(x) &= \nabla f(x) + \nabla g(x) \\ \nabla(\alpha f)(x) &= \alpha \nabla f(x), \forall \alpha > 0 \\ \nabla(f \cdot g)(x) &= g(x)\nabla f(x) + f(x)\nabla f(x) \\ \nabla\left(\frac{f}{g}\right) &= \frac{g(x)\nabla f(x) - f(x)\nabla g(x)}{g(x)^2}, \quad \textit{assuming that } g(x) > 0. \end{aligned} \tag{3}$$

Proof. Proof is left as exercise. □

Definition 3.6 (Generalization for vector-valued function)**.** A function $f : \mathbb{R}^d \to \mathbb{R}^\ell : x \mapsto (f_1(x), \ldots, f_\ell(x))$ is differentiable at a point x if $f_i(x)$ is differentiable at x for all $i = 1, \ldots, \ell$. In that case, the derivatives of f at point x given by a matrix (known as Jacobian matrix) whose formulation is:

$$J_f(x) = \begin{pmatrix} \nabla f_1(x)^\top \\ \vdots \\ \nabla f_\ell(x)^\top \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_d}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_\ell}{\partial x_1}(x) & \cdots & \frac{\partial f_\ell}{\partial x_d}(x) \end{pmatrix}.$$

We also have:

$$\lim_{d \to 0} \frac{\|f(x+d) - f(x) - J_f(x)d\|}{\|d\|} = 0.$$

Example 3.7. Consider a vector-valued differentiable function such as $f : \mathbb{R}^2 \to \mathbb{R}^2 : (x, y) \to (x^2 y + x^2, y^3)$.

Proposition 3.8 (Chain rule)**.** *Given two differentiable functions $f : \mathbb{R}^k \to \mathbb{R}^\ell$ and $g : \mathbb{R}^d \to \mathbb{R}^k$, then the composition $f \circ g : \mathbb{R}^d \to \mathbb{R}^\ell$ is also differentiable and its Jacobian matrix is given by:*

$$J_{f \circ g}(x) = J_f(g(x)) J_g(x).$$

Proof. Proof is left as exercise. □

Definition 3.9 (Twice differentiable functions and Hessian matrix)**.** A function $f : \mathbb{R}^d \to \mathbb{R}^d$ is called twice differentiable at a point x if the function $x \to \nabla f(x) : \mathbb{R}^d \to \mathbb{R}^d$ is differentiable (in the sense of a vector-valued function). The hessian matrix is the Jacobian matrix of $x \to \nabla f(x)$, i.e.,

$$H_f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(x) & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_d}(x) \end{pmatrix}$$

where coefficients are given by:

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)(x).$$

f is twice differentiable if f is twice differentiable at all $x \in \mathbb{R}^d$.

f is twice continuously differentiable (or a C^2 function) if the mapping $x \to H_f(x)$ is continuous. In that case, $H_f(x)$ is symmetric for all $x \in \mathbb{R}^d$.

Example 3.10. Consider:

1. $f(x) = \frac{1}{2}\|x\|_2^2$, we have $\nabla f(x) = x$ and $H_f(x) = \mathbf{I}$.
2. $f(x) = \sum_{i=1}^{d} x_i$, we have $\nabla f(x) = x$ and $H_f(x) = 0$.
3. $f(x) = \frac{1}{2}(\sum_{i=1}^{d} x_i)^2$, we have $\nabla f(x) = (\sum_{i=1}^{k} x_i)\mathbf{1}_d$ and $H_f(x) = \mathbf{1}_{d\times d}$.

Proposition 3.11 (Two Taylor formulations)**.** *Given a C^1 (resp. C^2) function $f : \mathbb{R}^d \to \mathbb{R}$, we have:*

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \nabla f(x + t(y-x))^\top (y-x)\, dt && ,\forall x, y \in \mathbb{R}^d \\ (resp.) f(y) &= f(x) + (y-x)^\top \nabla f(x) + \frac{1}{2}(y-x)^\top \nabla^2 f(x)(y-x) + R_2(x-y) && ,\forall x, y \in \mathbb{R}^d, \end{aligned} \tag{4}$$

where $R_2(x-y)$ is a reminder satisfying $\lim_{y\to x} \dfrac{R_2(x-y)}{\|y-x\|^2} = 0$.

Proof. Proof is left as an exercise. □

3.1 Optimal solutions of an optimization problem and their properties

In this section, we consider the following optimization problem:

$$\underset{x\in\mathcal{F}}{\text{Minimize}} \quad f(x), \quad \mathcal{F} \subseteq \mathbb{R}^d. \tag{OP}$$

where $\mathcal{F}$ is called the feasible set and its elements are called admissible points. In the following, we remind the definition of an optimal solution of (OP).

Definition 3.12 (Global solutions)**.** A point $x^\star \in \mathcal{F}$ is called an global optimal solution of (OP) if $f(x^\star)$ is the smallest value of f in $\mathcal{F}$, i.e.:

$$f(x^\star) \leq f(x), \qquad \forall x \in \mathcal{F}.$$

In general, optimal solutions of an optimization problem are difficult to find (unless under certain hypothesis such as convexity). Therefore, we will introduce another notion of optimal solutions, which is more accessible in practice:

Definition 3.13 (Local solutions)**.** A point $x^\star \in \mathcal{F}$ is called an local optimal solution of (OP) if $f(x^\star)$ is the smallest value of f in a admissible neighborhood of $x^\star$, i.e., there exists $\delta > 0$ such that:

$$f(x^\star) \leq f(x), \qquad \forall x \in B_\delta(x^\star) \cap \mathcal{F}.$$

where $B_\delta(x^\star)$ is the open ball centered at $x^\star$ and of radius δ.

It is clear that a global solution is also local. However, a local solution is not necessarily global. In the case where $\mathcal{F} = \mathbb{R}^d$ (unconstrained optimization), we introduce a necessary condition so that a point $x^\star$ is an optimal solution of (OP).

Theorem 3.14 (Necessary conditions)**.** *Consider* (OP) *where $\mathcal{F} = \mathbb{R}^d$. If a point $x^\star$ is a local solution of* (OP)*, then:*

1. *If f is C^1, $\nabla f(x^\star) = 0$.*
2. *If f is C^2, we have in addition that $\nabla^2 f(x^\star) \succeq 0$, i.e., $v^\top \nabla^2 f(x^\star) v \geq 0, \forall v \in \mathbb{R}^d$.*

Proof. Both items are proved using (4):

1. Proof of $\nabla f(x^\star) = 0$: By contradiction, we assume that $d := \nabla f(x^\star) \neq 0$. We consider $y = x^\star - ad$ where a is a constant to be determined. By Taylor formula, we have:

$$f(y) = f(x^\star) - \underbrace{a \int_0^1 d^\top \nabla f(x^\star - tad) dt}_{C}.$$

   Since f is C^1 and $d^\top \nabla f(x^\star) = \|d\|_2 > 2$, for all sufficiently small $a > 0$, we have: $d^\top \nabla f(x^\star - tad) \geq c > 0$. By consequent,

$$f(y) \leq f(x^\star) - ac < f(x^\star)$$

   for all a sufficiently small. This is a contradiction to the assumption that $x^\star$ is a local solution.

2. Proof of $\nabla^2 f(x^\star) \succeq 0$: Assume that $\nabla^2 f(x^\star)$ is not semi-definite positive. In that case, there exists a vector $d \in \mathbb{R}^d$ such that $c := d^\top \nabla^2 f d < 0$ and $\|d\| = 1$. By Taylor formulation:

$$\begin{aligned} f(x^\star - td) &= f(x^\star) + t\nabla f(x^\star)^\top d + \frac{1}{2} t^2 d^\top \nabla^2 f(x^\star) d + R_2(td) \\ &= f(x^\star) + \frac{t^2}{2} c + R_2(td). \end{aligned}$$

   Since $\lim_{t \to 0} \dfrac{R_2(td)}{t^2} = 0$, for all t sufficiently small, we have $f(x^\star - td) < f(x^\star)$. This is also a contraction to $x^\star$ being a local solution.

□

Remark 3.15. The conditions Theorem 3.14 are only necessary and not sufficient. Take $f(x) = \frac{1}{6}x^3$ and $x^\star = 0$. Although $\nabla f(x^\star) = \nabla^2 f(x^\star) = 0$, $x^\star$ is neither a global, nor a local solution of f.

Definition 3.16. A point $x^\star \in \mathcal{F}$ is called a critical point of a function C^1 f if $\nabla f(x^\star) = 0$.

We have the following relations:

$$x^\star \text{ is a global solution } \implies x^\star \text{ is a local solution } \implies x^\star \text{ is a critical point.}$$

The converse is not true in general.

Theorem 3.17 (Sufficient condition)**.** *Consider* (OP) *where* $\mathcal{F} = \mathbb{R}^d$ *and* f *is a* C^2 *function. If a point* $x^\star$ *satisfies :*

1. $\nabla f(x^\star) = 0$.
2. $\nabla^2 f(x^\star) \succ 0$ *(i.e.,* $\forall v \in \mathbb{R}^d, \|v\| = 1,\ v^\top \nabla^2 f(x^\star) v > 0$*),*

then $x^\star$ *is a local solution of* (OP)*.*

Proof. Take $c := \min_{\|v\|=1} v^\top \nabla^2 f(x^\star) v > 0$. Therefore, $\forall \delta \in \mathbb{R}$, we have $\delta^\top \nabla^2 f(x^\star) \delta \geq c\|\delta\|^2$. The Taylor formulation gives us :

$$\begin{aligned} f(x^\star + \delta) &= f(x^\star) + \nabla f(x^\star)^\top \delta + \frac{1}{2} \delta^\top \nabla^2 f(x^\star) \delta + R_2(\delta) \\ &= f(x^\star) + \frac{1}{2} \delta^\top \nabla^2 f(x^\star) \delta + R_2(\delta) \\ &\geq f(x^\star) + \frac{c}{2} \|\delta\|^2 + R_2(\delta). \end{aligned}$$

Since $\lim_{\delta \to 0} \dfrac{R_2(\delta)}{\|\delta\|^2} = 0$, there exists a sufficiently small neighborhood of $x^\star$ where $f(x) > f(x^\star)$. Thus, $x^\star$ is a local solution of (OP). □