

# Convex Optimization Exam

14/06/2024

3 hours – documents allowed  
Answers in English or French

The exam is made of 2 independent exercises, and a problem. If necessary, you can admit the results of previous questions. When using the recalls, cite them. “Classifying” an optimization problem consists in precisizing in which of the category presented in chapter 5 it falls (LP, QP, QCQP, SOCP, SDP, unconstrained or not, differentiable or not, continuous or not, convex or not).

Total number of points 24

## Some useful recalls

- i) The Fenchel transform of a function  $f$  is defined as  $f^* : x^* \mapsto \sup_{x \in \mathbb{R}^n} \langle x^*, x \rangle - f(x)$ .
- ii) If  $f$  is a proper, convex, lsc function of  $\mathbb{R}^n$  to  $\mathbb{R} \cup \{+\infty\}$ , then for all  $x, x^* \in \mathbb{R}^n$ , the following assertion are equivalent:
  - a)  $x^* \in \partial f(x)$
  - b)  $x \in \partial f^*(x^*)$
  - c)  $f(x) + f^*(x^*) = \langle x^*, x \rangle$
  - d)  $x \in \arg \max_{y \in \mathbb{R}^n} \langle x^*, y \rangle - f(y)$
  - e)  $x^* \in \arg \max_{y^* \in \mathbb{R}^n} \langle y^*, x \rangle - f^*(y^*)$
- iii) A function  $\Psi$  is  $\rho$ -convex if, for all  $x, y$ , we have  $\Psi(y) \geq \Psi(x) + \langle \nabla \Psi(x), y - x \rangle + \frac{\rho}{2} \|y - x\|^2$ .

## Exercise 1: Breakfast

2 points

- (a) ( $1/2$  point) Give 2 general ideas that speed-up the convergence of (variants of) the stochastic gradient algorithm.
- (b) ( $1 1/2$  points) Consider the following problem. Is it qualified? Give the KKT conditions. Are they enough to solve the problem? Solve the problem without solving the KKT conditions.

$$\begin{aligned} \min_{(x,y) \in \mathbb{R}^2} \quad & xy \\ \text{s.t.} \quad & y \geq x^2 \\ & y \leq x^3 \end{aligned}$$

**Exercise 2: How to save cloud costs?**

5 points

You are working for a company that conceive car design. You have been recruited, as an optimization expert, in the team that optimize a given design. The design is described by a vector  $x \in \mathbb{R}^{250}$  of parameters. We are looking for parameters  $x$  that minimize the drag coefficient  $D(x)$  of the car, under some mass constraints  $M(x) \leq \bar{M}$ , and budget constraint  $B(x) \leq \bar{B}$ . The function  $D$ ,  $M$  and  $B$  are "smooth" with respect to the parameter  $x$ , but their evaluation is numerically expensive (about 1 minute per evaluation per function). There is also a method to evaluate the sensitivity of each function with respect to the parameters which require about (10 minutes per evaluation per function).

All these evaluations are running in the cloud, which is costly. Your boss want to reduce this cost, and count on your knowledge to do so.

Currently, the optimization is done as follows: take the initial design  $x^0$ , choose parameter  $\alpha$  and  $\beta$ , compute  $D, M$  and  $B$  at  $x^0$ , and the sensitivities  $s_D, s_M$ , and  $s_B$ . Then update the design as  $x^{k+1} = x^k - \frac{1}{\tau}(s_D + \alpha s_M + \beta s_B)$ , and repeat the process until the sensitivities are small. If the final design is not satisfying the constraints, the process is restarted by adjusting  $\alpha$  and  $\beta$  by hand.

- (1/2 point) Classify the type of problem that is solved by the optimization process. Describe the current methodological approach with the language of optimization.
- (1 point) Currently, the parameter  $\tau$  is hard-coded based on some physical constant. Your boss want to increase it aggressively to reduce the number of iterations. Is it the right approach? What is the risk of doing so? And if you decrease it?
- (1 point) He has heard that there might be a way to automatically adjust  $\tau$  during the optimization process. Give the pseudo-code of such a method you would like to try, and an estimation of the computing time if you try 5  $\tau$  before accepting one.
- (1 point) Give two other ideas you could try to reduce the cost of the optimization process (for given  $\alpha$  and  $\beta$ ).
- (1 point) Another pain-point is adjusting the parameters  $\alpha$  and  $\beta$  by hand. Identify a method to adjust them automatically, and give its pseudo-code (assuming that the method solving the problem for given  $\alpha$  and  $\beta$  is given).
- (1/2 point) With the new ecological malus, the constraint on the mass is now a hard constraint (satisfied by the initial design). Suggest a new approach which ensure that the mass constraint is satisfied at each iteration.

**Exercise 3: Mirror descent and Bregman divergence**

17 points

In this exercise we are going to study a new class of optimization algorithm, called mirror descent, which leverage the Bregman divergence.

The exercise is divided in several parts. Each part of the exercise can be tackled independently of the others, but the results of the previous questions can be used if clearly cited. The part are roughly in increasing order of difficulty, but the first questions of each should be easy to answer.

We say that a function  $\Psi : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$  is a *mirror map* if:

- $\Psi$  is proper, strictly convex, lower semicontinuous, and differentiable on the interior of its domain;
- the gradient of  $\Psi$  diverges on the boundary of its domain (i.e.  $\lim_{x \rightarrow bd(\text{dom } \Psi)} \|\nabla \Psi(x)\| = +\infty$ ).

Throughout the exercise we consider that  $\Psi$  is a mirror map.

We also consider a proper, convex, lower semicontinuous objective function  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ , differentiable on its domain, and  $x^{(0)} \in \text{dom}(f)$ .

- We define the associated *Bregman Divergence* associated to  $\Psi$  as

$$D_{\Psi}(x, x_0) = \Psi(x) - \left( \Psi(x_0) + \langle \nabla \Psi(x_0), x - x_0 \rangle \right). \quad (1)$$

- (1/2 point) Give a geometric interpretation of the Bregman divergence (a graph can be useful).

- ii. (1 point) Show that, for all  $x, x_0 \in \text{dom}(\Psi)$ ,  $D_\Psi(x, x_0) \geq 0$ , and that  $D_\Psi(x, x_0) = 0$  if and only if  $x = x_0$ .
- iii. ( $1/2$  point) We admit that  $\Psi^*$  is differentiable on the interior of its domain. Show that, for  $x \in \text{dom}(\Psi)$ , and  $x^* \in \text{dom}(\Psi^*)$ , the following relations hold:

$$\nabla \Psi^*(\nabla \Psi(x)) = x \quad \text{and} \quad \nabla \Psi(\nabla \Psi^*(x^*)) = x^*$$

(b) We now discuss the mirror descent algorithm, in the unconstrained case.

- i. ( $1/2$  point) We consider the following algorithm, for  $\gamma > 0$ ,

$$x^{(k+1)} = \arg \min_{x \in \mathbb{R}^n} f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{2\gamma} \|x - x^{(k)}\|_2^2.$$

Is the algorithm well-defined? Recognize the algorithm.

- ii. (1 point) In its simplest form, the mirror descent algorithm, for unconstrained minimization, is defined as

$$x^{(k+1)} = \arg \min_{x \in \mathbb{R}^n} f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{\gamma} D_\Psi(x, x^{(k)}).$$

Justify that an iteration of the mirror descent algorithm can be written as

$$x^{(k+1)} = \nabla \Psi^* \left( \nabla \Psi(x^{(k)}) - \gamma \nabla f(x^{(k)}) \right).$$

- iii. ( $1/2$  point) Show that  $\Psi = \frac{1}{2} \|\cdot\|_2^2$  is a mirror map. What is the mirror descent algorithm associated with this choice of  $\Psi$ ?
  - iv. (2 points) We now consider  $\Psi = \sum x_i \ln x_i$ , for  $x \in \mathbb{R}_+^n$ , with the convention  $0 \ln(0) = 0$ , and  $+\infty$  otherwise. Show that  $\Psi$  is a mirror map. Give an explicit formula for the mirror descent algorithm associated with this choice of  $\Psi$ .
- (c) To address the constrained case, we introduce the notion of Bregman projection. Let  $C \subset \text{dom}(\Psi)$  be a closed convex compact set, and  $x_0 \in \text{dom}(\Psi)$ . The Bregman projection of  $x_0$  on  $C$  is defined as

$$\Pi_C(x_0) = \arg \min_{x \in C} D_\Psi(x, x_0). \tag{2}$$

- i. ( $1/2$  point) Show that the Bregman projection is well-defined.
- ii. ( $1/2$  point) Show that  $\Pi_C(x) = x$  if and only if  $x \in C$ . Further, show that  $\Pi_C \circ \Pi_C = \Pi_C$ .
- iii. (1 point) Show that, for all  $x \in C, y \in \text{dom}(\Psi)$ , the following inequality holds:

$$\left\langle \nabla \Psi(y) - \nabla \Psi(\Pi_C(y)), x - \Pi_C(y) \right\rangle \leq 0. \tag{3}$$

- iv. ( $1/2$  point) Deduce that, for all  $x \in C, y \in \text{dom}(\Psi)$ , the following inequality holds:

$$D_\Psi(x, \Pi_C(y)) + D_\Psi(\Pi_C(y), y) \leq D_\Psi(x, y). \tag{4}$$

- v. ( $1/2$  point) Show that, for  $x, y, z \in \text{dom}(\Psi)$ , the following equality holds:

$$\langle \nabla \Psi(x) - \nabla \Psi(y), x - z \rangle = D_\Psi(x, y) + D_\Psi(z, x) - D_\Psi(z, y). \tag{5}$$

(d) We now consider the mirror descent algorithm in the constrained case. Question marked with ( $\spadesuit$ ) require more computation.

Assume that  $\Psi$  is  $\rho$ -convex, that  $C$  is a closed convex compact set, and  $f$  is  $L$ -Lipschitz with respect to  $\|\cdot\|$ . The mirror descent algorithm in the constrained case is defined as

$$y^{(k+1)} = \nabla \Psi^* \left( \nabla \Psi(x^{(k)}) - \gamma \nabla f(x^{(k)}) \right) \tag{6a}$$

$$x^{(k+1)} = \Pi_C(y^{(k+1)}). \tag{6b}$$

Let  $x^\sharp \in \arg \min_{x \in C} f$ , and  $k \in \mathbb{N}$ .

- i. (1/2 point) Justify that  $\nabla\Psi(y^{(k+1)}) = \nabla\Psi(x^{(k)}) - \gamma\nabla f(x^{(k)})$ .  
ii. (1 point) Show that the following inequality holds:

$$f(x^{(k)}) - f(x^\#) \leq \frac{1}{\gamma} \left\langle \nabla\Psi(x^{(k)}) - \nabla\Psi(y^{(k+1)}), x^{(k)} - x^\# \right\rangle. \quad (7)$$

- iii. (1 point) Using the results of c), deduce that the following inequality holds ( $\spadesuit$ ):

$$f(x^{(k)}) - f(x^\#) \leq \frac{1}{\gamma} \left( D_\psi(x^\#, x^{(k)}) + D_\psi(x^{(k)}, y^{(k+1)}) - D_\psi(x^\#, x^{(k+1)}) - D_\psi(x^{(k+1)}, y^{(k+1)}) \right). \quad (8)$$

- iv. (2 points) Assume that  $\Psi$  is  $\rho$ -convex, and show that the following inequality holds ( $\spadesuit$ ):

$$D_\Psi(x^{(k)}, y^{(k+1)}) - D_\Psi(x^{(k+1)}, y^{(k+1)}) \leq \left\langle \nabla\Psi(y^{(k+1)}) - \nabla\Psi(x^{(k)}), x^{(k+1)} - x^{(k)} \right\rangle - \frac{\rho}{2} \|x^{(k+1)} - x^{(k)}\|^2. \quad (9)$$

- v. (1 point) Deduce that the following inequality holds:

$$D_\Psi(x^{(k)}, y^{(k+1)}) - D_\Psi(x^{(k+1)}, y^{(k+1)}) \leq \gamma L \|x^{(k+1)} - x^{(k)}\| - \frac{\rho}{2} \|x^{(k+1)} - x^{(k)}\|^2. \quad (10)$$

- vi. (1/2 point) Deduce that the following inequality holds:

$$D_\Psi(x^{(k)}, y^{(k+1)}) - D_\Psi(x^{(k+1)}, y^{(k+1)}) \leq \frac{(\gamma L)^2}{2\rho}. \quad (11)$$

- vii. (1 point) Deduce that the following inequality holds:

$$\sum_{k=1}^K (f(x^{(k)}) - f(x^\#)) \leq \frac{D_\Psi(x^\#, x^{(1)})}{\gamma} + \gamma \frac{L^2 K}{2\rho}. \quad (12)$$

- viii. (1 point) We denote  $\bar{x}_K = \frac{1}{K} \sum_{k=1}^K x^{(k)}$ . Show that

$$f(\bar{x}_K) - f(x^\#) \leq \frac{D_\Psi(x^\#, x^{(1)})}{\gamma K} + \gamma \frac{L^2}{2\rho}. \quad (13)$$