EXAM for Optimization for Machine Learning, *Master IASD: AI Systems and Data Science, 2019*

Tuesday January 7th

This is an open book exam, meaning you can consult any written or printed material. Electronic devices are prohibited. You must justify all of your answers. If you cannot solve a question, do not hesitate to admit its result in order to solve the next ones.

Ex. 1 — Let $c > 0$, $A = [a_1, \dots, a_n] \in \mathbb{R}^{d \times n}$ and $V = [v_1, \dots, v_n] \in \mathbb{R}^{d \times n}$. Consider the optimization problem

$$\min_x f(x) = \frac{1}{2n} \sum_{i=1}^{n} (a_i^\top x - v_i^\top x + c)^2. \tag{1}$$

1. What assumption can we make on A and V to guarantee that $f(x)$ is strongly convex and has a unique minimizer?
2. Please recommend an efficient method for solving (1) and justify why this method is efficient. NOTE: There is more than one right answer and you only need to give <u>one</u> reasonable suggestion.

Ex. 2 — Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a $\mathcal{C}^\infty$ function with the following properties:

(a) f is 1-smooth;

(b) f has a finite number of first-order critical points;

(c) for all $(x, y) \in \mathbb{R}^2$,

$$f(x, y) \geq x^2 + y^2 - 2;$$

(d) for all $(x, y) \in [-1; 1] \times [-1; 1]$,

$$f(x, y) = \frac{x^3}{6} + \frac{y^2}{2}.$$

We run gradient descent with step $1/2$ on f, starting from an initial point (x_0, y_0). We denote $((x_n, y_n))_{n \in \mathbb{N}}$ the sequence of iterates. The goal of the exercise is to study the behavior of this sequence.

1. Show that f has at least one minimizer. Deduce that f has at least one critical point.
2. Show that, for almost any (x_0, y_0), the sequence $((x_n, y_n))_{n\in\mathbb{N}}$ converges to a second-order critical point of f.
3. Show that, for any $(x, y) \in [-1; 1] \times [-1; 1]$,
$$\nabla f(x, y) = \left(\frac{x^2}{2}, y\right) \quad \text{and} \quad \text{Hess} f(x, y) = \begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix}.$$
4. Show that $(0, 0)$ is a second-order critical point of f, but not a local minimum.
5. We assume that (x_0, y_0) is in $[0; 1] \times [-1; 1]$. Show that (x_n, y_n) is in $[0; 1] \times [-1; 1]$ for all $n \in \mathbb{N}$, then that $((x_n, y_n))_{n\in\mathbb{N}}$ converges to $(0, 0)$.

Ex. 3 — We consider the function $f(x) \stackrel{\text{def}}{=} \sqrt{\sin(x)} + \sin(x)$.

1. Write the smallest directed acyclic graph corresponding to the evaluation of $f(x)$.
2. Write down the sequence of instructions corresponding to the evaluation of $f'(x)$ using the forward mode of automatic differentiation.
3. Same question but using the backward mode of automatic differentiation.

Ex. 4 — In the context of classification via linear support vector machines (SVM), given input $\{(a_j, b_j)\}_{i=1}^m$ with $a_j \in \mathbb{R}^n$ and $b_j \in \{-1, 1\}$, we seek a vector $x \in \mathbb{R}^n$ and a scalar $y \in \mathbb{R}$ such that

$$\begin{aligned} a_j^\top x - y &\geq 1 \qquad \text{if } b_j = +1; \\ a_j^\top x - y &\leq 1 \qquad \text{if } b_j = -1. \end{aligned}$$

The pair (x, y) describes a separating hyperplane in $\mathbb{R}^n$. Among all separating hyperplanes, we are interested in the one that maximizes the margin between the two classes $b_i = +1$ and $b_i = -1$. The maximum-margin hyperplane can be obtained as a solution of

$$\min_{x\in\mathbb{R}^n,\ y\in\mathbb{R}} \frac{1}{m} \sum_{j=1}^{m} \max(1 - b_j(a_j^\top x - y), 0) + \frac{c}{2}\|x\|_2^2, \tag{2}$$

where $c > 0$ is a small positive value.
The problem of minimizing (2) is unconstrained but involves a nonsmooth function. By introducing m additional variables $z = [z_i]_{i=1}^m$, one obtains a smooth, constrained reformulation of the problem:

$$\left\{ \begin{array}{lll} \min_{x\in\mathbb{R}^n, y\in\mathbb{R}, z\in\mathbb{R}^m} & \frac{1}{m}\sum_{j=1}^m z_j + \frac{c}{2}\|x\|_2^2 & \\ \\ \text{subject to} & z_j \geq 1 - b_j(a_j^\top x - y) & j = 1, \ldots, m, \\ & z_j \geq 0 & j = 1, \ldots, m. \end{array} \right. \tag{3}$$

1. Let $\mathcal{F}$ be the feasible set for problem (3). Find a function $h : \mathbb{R}^{n+1+m} \to \mathbb{R}^{2m}$ such that $\mathcal{F} = \{(x, y, z) \mid h(x, y, z) \leq 0\}$.
2. Justify that constraint qualification holds for problem (3) at any point $(x, y, z) \in \mathbb{R}^{n+1+m}$.
3. Let $\lambda \in (\mathbb{R}_+)^{2m}$ be the vector of Lagrange multipliers associated to the constraints. Write the formula for the Lagrangian function for (3).
4. Write the first-order KKT conditions at an optimal point (x^*, y^*, z^*).

Ex. 5 — We consider the non-smooth formulation of the linear support vector machine problem of **Exercise 4**,

$$\min_{x\in\mathbb{R}^n, y\in\mathbb{R}} f(x, y) \qquad \text{where} \qquad f(x, y) = \frac{1}{m}\sum_{j=1}^{m} \max(1 - b_j(a_j^\top x - y), 0) + \frac{c}{2}\|x\|_2^2, \tag{4}$$

and we want to study its optimality conditions.
In the following, we set $z \overset{\text{def}}{=} (x, y) \in \mathbb{R}^{n+1}$, and we consider a function

$$g : z \mapsto \max\Big(\alpha^\top z + \beta, 0\Big),$$

where $\alpha \in \mathbb{R}^{n+1}$, $\beta \in \mathbb{R}$.

1. Prove that g is convex.
2. Let us define $C \overset{\text{def}}{=} [0, \alpha] \overset{\text{def}}{=} \{\theta 0 + (1 - \theta)\alpha \ ; \ 0 \leq \theta \leq 1\}$. We want to prove that the subdifferential of g satisfies

$$\forall z \in \mathbb{R}^{n+1}, \quad \partial g(z) = \begin{cases} \{\alpha\} & \text{if } \alpha^\top z + \beta > 0, \\ C & \text{if } \alpha^\top z + \beta = 0, \\ \{0\} & \text{otherwise.} \end{cases}$$

   a) Draw a figure (in dimension 1, *i.e.* $z \in \mathbb{R}$) to explain why it should hold.
   b) Prove that $g(z) = \sup_{p\in\mathbb{R}^{n+1}} \big(p^\top z - h(p)\big)$, where

$$\begin{cases} h(p) = +\infty & \text{if } p \notin C, \\ h(\theta 0 + (1 - \theta)\alpha) = -(1 - \theta)\beta & \text{for } 0 \leq \theta \leq 1. \end{cases} .$$

   What is g with respect to h?
   c) Prove that for all $p \in \mathbb{R}^{n+1}$, $g^*(p) = h(p)$. What is the geometric interpretation of $g^*(p)$?
   d) Using the equality case in the Fenchel inequality, prove that $p \in \partial g(z)$ if and only if

$$p = \theta 0 + (1 - \theta)\alpha \text{ (with } 0 \leq \theta \leq 1) \quad \text{and} \quad \max(\alpha^\top z + \beta, 0) = (1 - \theta)(\alpha^\top z + \beta).$$

e) Conclude.

Now, we go back to (4). We set $\alpha_j \overset{\text{def}}{=} \begin{pmatrix} -b_j a_j \\ b_j \end{pmatrix}$, $\beta_j \overset{\text{def}}{=} 1$, and $C_j = [0, \alpha_j] \subset \mathbb{R}^n$.

3. Prove that if $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$ are solutions to (4), then $x_1 = x_2$.
4. a) Describe the subdifferential of $z = (x, y) \mapsto \frac{1}{2}\|x\|^2$.
   b) Prove that the optimality conditions of (4) are equivalent to

$$c \begin{pmatrix} x \\ 0 \end{pmatrix} \in -\frac{1}{m} \sum_{j \in I} \alpha_j - \frac{1}{m} \sum_{j \in J} C_j,$$

where $I = \left\{ j \in \{1, \dots, m\} \; ; \; \alpha_j^\top z + \beta_j > 0 \right\}$ and
$J = \left\{ j \in \{1, \dots, m\} \; ; \; \alpha_j^\top z + \beta_j = 0 \right\}$.
Compare with the question 4 of **Exercise 4**.
Hint: *To describe the subdifferential of a sum of* $(m+1)$ *functions, you may iteratively apply the result for* 2 *functions.*

Ex. 6 — Consider the following problem

$$p^* = \min_{x,z} x^2 - 3xz + z^2 + 2x, \quad \text{s.t.} \quad x + z = 1 \tag{5}$$

1. Write the Lagrangian of the problem above. Prove that p^* is always larger than

$$\min_{x,z} x^2 - 3xz + z^2 - 2z + 2$$

   (hint: use weak duality and bound from below the maximum with a specific value for the dual variable)
2. Prove that the problem in (5) admits a saddle point and that strong duality holds.
3. Write the algorithm to solve (5) by dual gradient ascent.

Answer (Ex. 1) — 1. Differentiating once we have that

$$\nabla f(x) = \frac{1}{n}\sum_{i=1}^{n}(a_i - v_i)(a_i^\top x - v_i^\top x + c)$$
$$= \frac{1}{n}\sum_{i=1}^{n}(a_i - v_i)((a_i - v_i)^\top x + c) = \frac{1}{n}(A - V)\left((A - V)^\top x + \mathbf{1}c\right), \quad (6)$$

where $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^d$ is the vector of all ones. Differentiating again gives

$$\nabla^2 f(x) = \frac{1}{n}(a_i - v_i)(a_i - v_i)^\top = \frac{1}{n}(A - V)(A - V)^\top.$$

Consequently $\nabla^2 f(x)$ is symmetric positive semi-definite since

$$u^\top \nabla^2 f(x) u = \frac{1}{n}\|(A - V)u\|^2 \geq 0.$$

Thus $f(x)$ is convex.

If $(A - V)u \neq 0$ for every $u \neq 0$, then $f(x)$ is strongly convex since then

$$\lambda_{\min}(\nabla^2 f(x)) = \frac{1}{n}\lambda_{\min}((A - V)(A - V)^\top) = \min_{u \neq 0} \frac{\|(A - V)u\|}{\|u\|} > 0.$$

Said in another way, $Au = Vu$ if and only if $u = 0$.

2. Since $f(x)$ is convex, the solution x^* to (1) is given by the solution to $\nabla f(x) = 0$ and thus from (6) we have that

$$(A - V)(A - V)^\top x = -(A - V)\mathbf{1}c.$$

This is a symmetric positive definite linear system. I have now ranked the following answer

1st: The fastest methods are ones that fully exploit the structure of this linear system. Which in this case would be coordinate descent (CD) or the conjugate gradient (CG) algorithm which are both method specialized for solving positive definite linear systems. The CG has an accelerated rate of convergence (though it was not covered in this course) and the CD has a cheap iteration with a rate of convergence of $\lambda_{\min}((A - V)(A - V)^\top)/\text{trace}((A - V)(A - V)^\top)$ amongst the stochastic methods. Another good suggestion would any direct method specialized towards positive definite linear systems, such as a Choleseky decomposition.

2nd: Stochastic gradient descent/ randomized Kaczmarz method would exploit the sum of terms structure and would converge at a rate of $\sigma_{\min}(A-V)^4/\|A-V\|_F^4$ which is slower than CD (and slower in practice). Also any variance reduced method, in particular SAGA since it allows for an efficient implementation. Though this is not as efficient as CD in this case. Another reasonable choice would be any direct method, such as Gaussian elimination or LU decomposition.

Answer (Ex. 2) —

1. Since $f \in C^\infty$, we know that it is proper lower semi-continuous. Moreover, from Property (c), we see that f is coercive, *i.e.* $\lim_{\|(x,y)\|\to+\infty} f(x,y) = +\infty$. As a result, f has a minimizer x_*. Since f is smooth, we must have $\nabla f(x_*) = 0$, that is x_* is a critical point of f.
2. From the lecture of November 26 (see Theorem 3.2 of the corresponding lecture notes), gradient descent on function f converges to a second-order critical point for almost any initial value, provided that the following three conditions hold true:
   1. f has only a finite number of first-order critical points;
   2. for any $M \in \mathbb{R}$, $\{(x,y) \in \mathbb{R}^2, f(x,y) \leq M\}$ is bounded;
   3. the stepsize is constant and belongs to $\left]0; \frac{1}{L}\right[$, where $L > 0$ is such that f is L-smooth.

   The first condition is true, from the second hypothesis on f. The third one is also true, since f is 1-smooth and we run gradient descent with stepsize $1/2$.
   Let us check the second condition. For any $M \in \mathbb{R}$, using the third hypothesis on f,

$$\begin{aligned}\{(x,y) \in \mathbb{R}^2, f(x,y) \leq M\} &\subset \{(x,y) \in \mathbb{R}^2, x^2 + y^2 - 2 \leq M\} \\ &= \{(x,y) \in \mathbb{R}^2, x^2 + y^2 \leq M + 2\}\end{aligned}$$

   Therefore, $\{(x,y) \in \mathbb{R}^2, f(x,y) \leq M\}$ is empty if $M < -2$, and included in the ball centered at zero with radius $\sqrt{M+2}$ otherwise. In any case, it is a bounded set.
3. For any $(x,y) \in [-1;1]^2$,

$$\nabla f(x,y) = \left(\frac{\partial f}{\partial x}(x,y), \frac{\partial f}{\partial y}(x,y)\right) = \left(\frac{x^2}{2}, y\right),$$

   and

$$\begin{aligned}\mathrm{Hess} f(x,y) &= \begin{pmatrix} \frac{\partial^2 f}{\partial x^2}(x,y) & \frac{\partial^2 f}{\partial x \partial y}(x,y) \\ \frac{\partial^2 f}{\partial x \partial y}(x,y) & \frac{\partial^2 f}{\partial y^2}(x,y) \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial}{\partial x}\left((x,y) \to \frac{x^2}{2}\right)(x,y) & \frac{\partial}{\partial x}\left((x,y) \to y\right)(x,y) \\ \frac{\partial}{\partial x}\left((x,y) \to y\right)(x,y) & \frac{\partial}{\partial y}\left((x,y) \to y\right)(x,y) \end{pmatrix} \\ &= \begin{pmatrix} x & 0 \\ 0 & 1 \end{pmatrix}.\end{aligned}$$

4. From the previous question, $\nabla f(0,0) = (0,0)$, hence $(0,0)$ is a first-order critical point of f. At $(0,0)$, the Hessian is $\mathrm{Hess} f(0,0) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$. This matrix is diagonal

with nonnegative coefficients, therefore semidefinite positive. Consequently, $(0,0)$ is a second-order critical point.
For any $x \in [-1;0[$, $f(x,0) = \frac{x^3}{6} < 0 = f(0,0)$. Therefore, in any neighborhood of $(0,0)$, there is a point where f has a stricly smaller value than at $(0,0)$. Consequently, $(0,0)$ is not a local minimum.

5. We prove that (x_n, y_n) is in $[0;1]\times[-1;1]$ by iteration over n. It is true by assumption for $n = 0$. Now, assuming it holds for n, we prove it for $n+1$. We have

$$\begin{aligned}(x_{n+1}, y_{n+1}) &= (x_n, y_n) - \frac{1}{2}\nabla f(x_n, y_n)\\ &= \left(x_n - \frac{x_n^2}{4}, y_n - \frac{y_n}{2}\right)\\ &= \left(x_n\left(1 - \frac{x_n}{4}\right), \frac{y_n}{2}\right).\end{aligned}$$

Since $y_n \in [-1;1]$, $y_{n+1} = \frac{y_n}{2}$ is in $[-1/2;1/2] \subset [-1;1]$. And since $x_n \in [0;1]$,

$$\begin{aligned}&\frac{3}{4} \leq 1 - \frac{x_n}{4} \leq 1;\\ \Rightarrow \qquad &0 = 0 \times \frac{3}{4} \leq x_n\left(1 - \frac{x_n}{4}\right) = x_{n+1} \leq 1 \times 1 = 1.\end{aligned}$$

This concludes the proof for $n+1$.
Now we prove that $((x_n, y_n))_{n\in\mathbb{N}}$ goes to $(0,0)$. We have seen that $y_{n+1} = \frac{y_n}{2}$ for any n, therefore

$$\forall n \in \mathbb{N}, \quad y_n = \frac{y_0}{2^n}$$

and $y_n \to 0$ when $n \to +\infty$.
It is left to prove that $(x_n)_{n\in\mathbb{N}}$ also goes to 0. This sequence is lower bounded by 0 and non-increasing (since $x_{n+1} = x_n - \frac{x_n^2}{4} \leq x_n$ for any n) so it converges to some limit ℓ. As $x_{n+1} = x_n - \frac{x_n^2}{4}$ also goes to ℓ when $n \to +\infty$, we must have $\ell - \frac{\ell^2}{4} = \ell$, that is $\ell = 0$. This proves that $x_n \to 0$ when $n \to +\infty$.

Answer (Ex. 3) — 1. $y = \sin(x)$, $z = \sqrt{y}$, $f = y + z$.

2. As done during the course, with an abuse of notations, I am putting between brackets

operators (as opposed to numerical quantities)

$$\begin{aligned}
\frac{\partial x}{\partial x} &= 1 \\
\frac{\partial y}{\partial x} &= \left(\frac{\partial y}{\partial x}\right)\frac{\partial x}{\partial x} = \cos(x)\frac{\partial x}{\partial x} \\
\frac{\partial z}{\partial x} &= \left(\frac{\partial z}{\partial y}\right)\frac{\partial y}{\partial x} = \frac{1}{2\sqrt{y}}\frac{\partial y}{\partial x} \\
\frac{\partial f}{\partial x} &= \left(\frac{\partial f}{\partial y}\right)\frac{\partial y}{\partial x} + \left(\frac{\partial f}{\partial z}\right)\frac{\partial z}{\partial x} = \frac{\partial y}{\partial x} + \frac{\partial z}{\partial x}
\end{aligned}$$

3.

$$\begin{aligned}
\frac{\partial f}{\partial f} &= 1 \\
\frac{\partial f}{\partial z} &= \frac{\partial f}{\partial f}\left(\frac{\partial f}{\partial z}\right) = \frac{\partial f}{\partial f}1 \\
\frac{\partial f}{\partial y} &= \frac{\partial f}{\partial f}\left(\frac{\partial f}{\partial y}\right) + \frac{\partial f}{\partial z}\left(\frac{\partial z}{\partial y}\right) = \frac{\partial f}{\partial f}1 + \frac{\partial f}{\partial z}\frac{1}{2\sqrt{y}} \\
\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial y}\left(\frac{\partial y}{\partial x}\right) = \frac{\partial f}{\partial y}\cos(x)
\end{aligned}$$

Answer (Ex. 4) — All questions are based on the material from the *Basics of constrained optimization* lecture. An additional question also uses material from the advanced lecture on *second-order* and *interior-point* methods.

1. The feasible set is given by:

$$\mathcal{F} = \left\{(x, y, z) \in \mathbb{R}^{n+1+m} \;\middle|\; \forall j = 1 \dots m,\ z_j \geq 1 - y_j(a_j^\top x - y) \text{ and } z_j \geq 0\right\}.$$

   An equivalent description where all constraints are of the form $g(x, y, s) \leq 0$ is given by

$$\mathcal{F} = \left\{(x, y, z) \in \mathbb{R}^{n+1+m} \;\middle|\; \forall j = 1 \dots m,\ -z_j + 1 - b_j(a_j^\top x - y) \leq 0 \text{ and } -z_j \leq 0\right\}.$$

   Letting $h(x, y, z) = [h_j(x, y, z)]$ with $h_j(x, y, z) = -z_j + 1 - b_j(a_j^\top x - y)$ for $j = 1, \dots, m$ and $h_j(x, y, z) = -z_{j-m}$ for $j = 1, \dots, m$, we obtain $\mathcal{F} = \left\{(x, y, z) \in \mathbb{R}^{n+1+m} \;\middle|\; h(x, y, z) \leq 0\right\}$, which is the desired description.

2. All constraints are linear in the coefficients of x, z, as well as in y: this is one of the form of constraint qualification that was defined in the lecture. Because this property holds independently of the considered point, constraint qualification holds at every $(x, y, z) \in \mathbb{R}^{n+1+m}$.

3. Using the description for the feasible set established in Question 1, we define the Lagrangian function for problem (3) as:

$$\begin{aligned}\mathcal{L}(x,y,z,\lambda) &= \frac{1}{m}\sum_{j=1}^{m} z_j + \frac{c}{2}\|x\|_2^2 + \lambda^\top h(x,y,z) \\ &= \frac{1}{m}\sum_{j=1}^{m} z_j + \frac{c}{2}\|x\|_2^2 + \sum_{j=1}^{m}\lambda_j(-z_j + 1 - b_j(a_j^\top x - y)) - \sum_{j=m+1}^{2m}\lambda_j z_{j-m}.\end{aligned}$$

4. If (x^*, y^*, z^*) is a solution of the problem, then we know from question 2 that constraint qualification holds at (x^*, y^*, z^*). Therefore, there exists $\lambda_* \in (\mathbb{R}_+)^{2m}$ such that:

$$\left\{\begin{array}{lcll} \nabla_x\mathcal{L}(x^*,y^*,z^*,\lambda^*) &=& 0 & \\ \nabla_y\mathcal{L}(x^*,y^*,z^*,\lambda^*) &=& 0 & \\ \nabla_z\mathcal{L}(x^*,y^*,z^*,\lambda^*) &=& 0 & \\ h(x^*,y^*,z^*) &\leq& 0 & \\ \lambda^* &\geq& 0 & \\ \lambda_j^*\, h_j(x^*,y^*,z^*) &=& 0 & \forall j = 1\ldots m, \end{array}\right.$$

where the gradients of the Lagrangian function are given by

$$\begin{aligned}\nabla_x\mathcal{L}(x^*,y^*,z^*,\lambda^*) &= cx^* - \sum_{j=1}^{m}\lambda_j^* b_j a_j, \\ \nabla_y\mathcal{L}(x^*,y^*,z^*,\lambda^*) &= \sum_{j=1}^{m}\lambda_j^* b_j \\ \nabla_{z_j}\mathcal{L}(x^*,y^*,z^*,\lambda^*) &= \frac{1}{m} - \lambda_j^* - \lambda_{m+j}^* \quad \forall j = 1\ldots m.\end{aligned}$$

5. *Additional question: We consider solving problem* (3) *by applying the projected gradient descent method or by calling an interior-point solver. Comment on the interest of both approaches.*
   We propose below some justification for the proposed approaches. The answer is not unique, however, and one simple argument per method was all that was required.
   Projected gradient descent: Since the problem constraints are linear, the projection of a given point onto the feasible set is unique (assuming it is non-empty), and thus projected gradient descent will be well-defined. Moreover, the objective function is convex, which can be leveraged in the analysis of projected gradient descent to establish better convergence rates than in the general nonconvex case. The simplicity of the method, that allows for many enhancements (as in the case of gradient descent), can also be a justification for its use.

Interior-point solver: This framework has been studied in class for linear programs (linear objective+linear constraints), but as discussed, they can be extended to quadratic programs (quadratic objective+linear constraints), which is our setup here. Interior-point algorithms apply Newton's method to a modified version of the KKT equations, thus they are second-order methods, with fast local guarantees, especially on convex problems such as (3). One key aspect of interior-point *solvers* is that they are quite mature, and able to exploit structures in the objective and the constraints. Problem (3) is highly structured, therefore it is likely that popular solvers such as CPLEX will be quite efficient in solving it, even in large dimensions.

Answer (Ex. 5) — 1. g is the maximum of two affine (hence convex) functions. Hence g is convex.

2. a) We expect a figure with the supporting hyperplanes, detailing the slopes of f.
   b)

$$\sup_{p\in\mathbb{R}^{n+1}} \left(p^\top z - h(p)\right) = \sup_{0\le\theta\le 1} (1-\theta)\alpha^\top z + (1-\theta)\beta = \sup_{0\le\theta\le 1} (1-\theta)\left(\alpha^\top z + \beta\right) = g(z).$$

   g is the *convex conjugate* of h, *i.e.* $g = h^*$.
   c) h is convex proper lower semi-continuous, hence $h = h^{**} = g^*$. For all p, $-g^*(p)$ is the intercept of the best affine minorant of g with slope p.
   d)

$$\begin{aligned} p \in \partial g(x) &\Leftrightarrow g(x) + g^*(p) = p^\top x \\ &\Leftrightarrow p = \theta 0 + (1-\theta)\alpha \text{ with } 0 \le \theta \le 1 \text{ and } \max(\alpha^\top z + \beta, 0) = (1-\theta)(\alpha^\top z + \beta). \end{aligned}$$

   If $\alpha^\top z + \beta > 0$ then $\theta = 0$. If $\alpha^\top z + \beta < 0$ then $\theta = 1$. If $\alpha^\top z + \beta = 0$, any $\theta \in [0,1]$ is admissible.

3. If z_1, z_2 are two minimizers,

$$\begin{aligned} f\left(\frac{1}{2}(z_1+z_2)\right) &= r\left(\frac{1}{2}(z_1+z_2)\right) + \frac{c}{2}\|\frac{1}{2}(x_1+x_2)\|^2 \\ &\qquad \text{with } r(x,y) = \tfrac{1}{m}\textstyle\sum_{j=1}^m \max(1-b_j(a_j^\top x - y), 0) \\ &\le \frac{1}{2}r(z_1) + \frac{1}{2}r(z_2) + \frac{c}{2}\left(\frac{1}{2}\|x_1\|^2 + \frac{1}{2}\|x_2\|^2\right) \end{aligned}$$

   by convexity. Since the squared Euclidean norm is strictly convex, the last inequality is strict unless $x_1 = x_2$, which would contradict the optimality of z_1 and z_2.

4. a) The function $v : z = (x,y) \mapsto \frac{1}{2}\|x\|^2$ is convex differentiable. Therefore $\partial v(z) = \{\nabla v(z)\} = \{(x,0)\}$.

b) We let $g_j(z) \stackrel{\text{def}}{=} g(\alpha_j^\top z + \beta_j) = \max(1 - b_j(a_j^\top x - y), 0)$. Since $\text{dom}(g_j) = \mathbb{R}^{n+1}$, we have $\text{ri}(\text{dom}(g_1)) \cap \text{ri}(\text{dom}(g_2)) = \mathbb{R}^{n+1} \cap \mathbb{R}^{n+1} \neq \emptyset$. As a result $\partial(g_1 + g_2) = \partial g_1 + \partial g_2$. Arguing iteratively, we obtain

$$\partial \left(\sum_{j=1}^{m} g_j \right) = \partial \left(\sum_{j=1}^{m-1} g_j \right) + \partial g_m = \sum_{j=1}^{m} \partial g_j.$$

Incorporating the last term, and using the same argument, we get

$$\partial \left(\frac{1}{m} \sum_{j=1}^{m} g_j + \frac{c}{2} v \right) = \frac{1}{m} \partial \left(\sum_{j=1}^{m} g_j \right) + \frac{c}{2} \partial v = \frac{1}{m} \sum_{j=1}^{m} \partial g_j + c \begin{pmatrix} \text{Id} \\ 0 \end{pmatrix}.$$

Now, we apply the optimality condition: z is optimal if and only if

$$\begin{aligned} 0 \in \partial f(z) &\iff 0 \in \frac{1}{m} \sum_{j=1}^{m} \partial g_j(z) + c \begin{pmatrix} x \\ 0 \end{pmatrix} \\ &\iff c \begin{pmatrix} x \\ 0 \end{pmatrix} \in -\frac{1}{m} \sum_{j=1}^{m} \partial g_j(z) \\ &\iff c \begin{pmatrix} x \\ 0 \end{pmatrix} \in -\frac{1}{m} \sum_{j \in I} \alpha_j - \frac{1}{m} \sum_{j \in J} C_j. \end{aligned}$$

The first n equations imply that there exist $\theta_j \in [0, 1]$ $(j \in J)$ such that

$$cx = \frac{1}{m} \left(\sum_{j \in I} b_j a_j + \sum_{j \in J} \theta_j b_j a_j \right)$$

We recover the result of **Exercice 4**, $\nabla_x \mathcal{L}(x, y, z) = 0$, *i.e.* $cx = \sum_{j=1}^{m} \lambda_j^* a_j b_j$, in which $\lambda_j^* = 1$ for $j \in I$, $\lambda_j^* = \theta_j \in [0, 1]$ for $j \in J$, and $\lambda_j^* = 0$ for $j \notin (I \cup J)$. The last equation implies $0 = \sum_{j \in I} b_j + \sum_{j \in J} \theta_j b_j$ which is equivalent to the equation $\nabla_y \mathcal{L}(x, y, z) = 0$.

Answer (Ex. 6) — 1.The Lagrangian is

$$L(x, y, z) = x^2 - 3xz + z^2 + 2x + y(x + z - 1).$$

By weak duality we have

$$\max_y \min_{x,z} L(x, y, z) \leq \min_{x,z} \max_y L(x, y, z) =: p^*$$

If we choose $y = -2$ we obtain

$$\min_{x,z} x^2 - 3xz + z^2 - 2z + 2 = \min_{x,z} L(x, -2, z) \leq \max_y \min_{x,z} L(x, y, z).$$

2.Since the constraint is affine and the function to be minimized is convex and differentiable, then we can apply the KKT conditions. We have

$$\begin{aligned}\nabla_x L(x,y,z) &:= 2x - 3z + 2 + y \\ \nabla_y L(x,y,z) &:= x + z - 1 \\ \nabla_z L(x,y,z) &:= -3x + 2z + y.\end{aligned}$$

By solving the linear system

$$\begin{aligned}2x - 3z + 2 + y &= 0 \\ x + z - 1 &= 0 \\ -3x + 2z + y &= 0,\end{aligned}$$

we obtain the saddle point $x^* = 3/10, y^* = -1/2, z^* = 7/10$. The existence of a saddle point guarantees strong duality.

3.
```
x =0, y = 0, z = 0
for t=1 to T
    #in parallel on machine 1, load y and
    x = argmin_x (x^2 -3xz + 2x + yx)
    #on machine 2, load y and
    z = argmin_z (-3xz + z^2 + yz)
    #collect x, z from machine 1 and 2 and
    y = y + gamma(t) * (x+z-1)
```