Exam: Optimization for Machine Learning

M2 IASD/MASH

Monday, January 4, 2021

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.

Part 1 (MASH+IASD)

The following exercises deal with the first part of the course (basics of optimization) and should be addressed by both the IASD and MASH students.

Ex. 1 — 1. If $f : \mathbb{R}^d \to \mathbb{R}$ is convex, show that the set of minimizers of f is a convex set.

2. We consider for $(x, y, a, b) \in \mathbb{R}^4$,

$$f(x) \overset{\text{def}}{=} \frac{1}{2}(x-y)^2 \quad \text{and} \quad g(a,b) \overset{\text{def}}{=} \frac{1}{2}(ab-y)^2$$

   a) Give the set of minimizers of f and g.
   b) Is f convex, coercive ? Same question for g.

3. a) Compute the gradients and Hessians of f and g.
   b) Show that a symmetric 2×2 matrix is positive definite if and only if its trace and its determinant are positive (i.e. > 0).
   c) Compute and simplify the determinant of $\partial^2 g(a,b)$. Assuming $y > 0$, display on the plane $(a,b) \in \mathbb{R}^2$ the set of minimizers of g, and the region where g is locally convex (i.e. where $\partial^2 g(a,b)$ is positive).

4. For $X, Y, A, B \in \mathbb{R}^{n \times n}$ matrices, we consider

$$f(X) \overset{\text{def}}{=} \frac{1}{2}\|X-Y\|^2 \quad \text{and} \quad g(A,B) \overset{\text{def}}{=} \frac{1}{2}\|AB-Y\|^2$$

   where $\|\cdot\|$ is the Euclidean (Frobenius) norm.
   a) Give the set of minimizers of f and g. Is f convex, coercive ? Same question for g. *Hint:* you can show that $\{(A,B) : AB = Y\}$ is non-convex by considering the special case $B = \lambda\text{Id}$ for two different values of λ.
   b) Compute the gradient of f and g (which are represented as matrices).

Ex. 2 — Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function. We assume that

(1) f is convex;

(2) f has a global minimizer x_*;

(3) f is differentiable and, for any $x \in \mathbb{R}^d$,

$$||\nabla f(x)||_2 \leq 1.$$

Starting at some point x_0, we run gradient descent with a sequence of positive stepsizes $(h_k)_{k\in\mathbb{N}}$. This yields a sequence of iterates defined by:

$$x_{k+1} = x_k - h_k \nabla f(x_k), \quad \forall k \in \mathbb{N}.$$

1. a) Show that, for any $k \in \mathbb{N}$,

$$f(x_k) - f(x_*) \leq \langle \nabla f(x_k), x_k - x_* \rangle .$$

b) Show that, for any $k \in \mathbb{N}$,

$$||x_{k+1} - x_*||_2^2 \leq ||x_k - x_*||_2^2 - 2h_k(f(x_k) - f(x_*)) + h_k^2 ||\nabla f(x_k)||_2^2.$$

c) Show that, for any $n \in \mathbb{N}$,

$$2\sum_{k=0}^{n} h_k(f(x_k) - f(x_*)) \leq ||x_0 - x_*||_2^2 - ||x_{n+1} - x_*||_2^2 + \sum_{k=0}^{n} h_k^2 ||\nabla f(x_k)||_2^2.$$

d) For any n, let $k_n \in \{0, \dots, n\}$ be such that

$$f(x_{k_n}) = \min_{k=0,\dots,n} f(x_k).$$

Show that, for any n,

$$2(f(x_{k_n}) - f(x_*)) \left(\sum_{k=0}^{n} h_k \right) \leq ||x_0 - x_*||_2^2 - ||x_{n+1} - x_*||_2^2 + \sum_{k=0}^{n} h_k^2 ||\nabla f(x_k)||_2^2.$$

e) Show that, for any n,

$$2(f(x_{k_n}) - f(x_*)) \left(\sum_{k=0}^{n} h_k \right) \leq ||x_0 - x_*||_2^2 + \sum_{k=0}^{n} h_k^2.$$

2. In this question, we assume that, for any k, $h_k = \frac{1}{\sqrt{k+1}}$. Show that, for any n,

$$f(x_{k_n}) - f(x_*) \leq \frac{||x_0 - x_*||_2^2 + 2 + \log(n)}{\sqrt{n+2}}.$$

Hint: *You can use the fact that, for any n,*

$$\sum_{k=1}^{n+1} \frac{1}{k} \leq 2 + \log(n) \quad \textit{and} \quad \sum_{k=1}^{n+1} \frac{1}{\sqrt{k}} \geq \frac{\sqrt{n+2}}{2}.$$

3. In this question, we assume that the sequence of stepsizes is constant: there exists $\eta > 0$ such that, for any $k \in \mathbb{N}$, $h_k = \eta$.
   Give an example of a function f satisfying properties (1), (2), (3), and a starting point x_0 such that
$$f(x_{k_n}) - f(x_*) \overset{n \to +\infty}{\not\to} 0.$$
   Hint: *Choose $\epsilon > 0$ small enough and consider the following function:*
$$f : x \in \mathbb{R} \mapsto \begin{cases} |x| - \frac{\epsilon}{2} & \textit{if } |x| \geq \epsilon; \\ \frac{x^2}{2\epsilon} & \textit{if } |x| \leq \epsilon. \end{cases}$$

Ex. 3 — We consider a finite-sum optimization problem:

$$\min_{w \in \mathbb{R}^d} f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w), \tag{1}$$

where each function f_i depends on a single item of a dataset consisting in n elements. We suppose that every function f_i is continuously differentiable, and we define a family of algorithms methods based on a starting point $w_0 \in \mathbb{R}^d$ as well as the recursion

$$\forall k \geq 0, \quad w_{k+1} = w_k - \alpha g_k, \tag{2}$$

where $\alpha > 0$ is a given stepsize and g_k is an estimate of the gradient $\nabla f(w_k)$.

1. How should g_k be chosen in order for the recursion (2) to correspond to an instance of
   a) gradient descent?
   b) stochastic gradient?
2. Recall the definition of an epoch: what is the equivalent of this unit in terms of
   a) iterations of gradient descent?
   b) iterations of stochastic gradient?

We now focus on using of batch variants of stochastic gradient. Given a batch size $n_b \in \{1, \dots, n\}$, we draw a batch index set $\mathcal{S}_k \subseteq \{1, \dots, n\}$ based on the following distribution :

$$\forall \mathcal{S} \subseteq \{1, \dots, n\}, \quad \mathbb{P}(\mathcal{S}_k = \mathcal{S}) := \begin{cases} \frac{1}{\binom{n}{n_b}} = \frac{n_b!(n-n_b)!}{n!} & \text{if } |\mathcal{S}| = n_b \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We then set $g_k = \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \nabla f_i(w_k)$ in the recursion (2).

3. Show that $\mathbb{E}_{\mathcal{S}_k}[g_k] = \nabla f(w_k)$.
4. Suppose that the function f is strongly convex; in that case, and under the appropriate assumptions on the problem, one can show that
$$\lim_{k \to \infty} \mathbb{E}[f(w_k) - f^*] \in \left[0, c\alpha \frac{M^2}{n_b}\right], \tag{4}$$
   where $c > 0$ is a problem-dependent constant and $M^2 > 0$ is a bound on $\|\nabla f_i(w)\|$ for any $i \in \{1, \dots, n\}$ and $w \in \mathbb{R}^d$.

a) What property of (batch) stochastic gradient methods does this result illustrate?

b) Describe two modifications of the algorithm (among those covered in the lectures) that can lead to a guarantee of the form $\lim_{k\to\infty} \mathbb{E}\left[f(w_k) - f^*\right] = 0$.

5. *Practical situation :* As in the lab session, suppose that we want to compare several batch sizes. While running on a given problem, we observe that $n_b = 1$ gives better convergence than $n_b = n$, while increasing the batch size from $n_b = 1$ to $n_b = n/10$ consistently improves the results, in that the method converges faster and to a smaller value of f. We then observe that the convergence slows down as we increase the batch size from $n/10$ to n. How can you explain these observations?

Ex. 4 — We consider the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) \overset{\text{def}}{=} \frac{1}{2}n^2 + a_n(x-n), \quad \text{where } a_n \overset{\text{def}}{=} n + \frac{1}{2},$$

for all $x \in \mathbb{R}$, $n \in \mathbb{Z}$ such that $x \in [n, n+1[$.

1. Let $\{\ell_n\}_{n\in\mathbb{Z}}$ be the family of functions defined by $\ell_n\colon x \mapsto \left(\frac{1}{2}n^2 + a_n(x-n)\right)$.

   a) Let $n, p \in \mathbb{Z}$. Prove that $\ell_n \geq \ell_p$ on the interval $[n, n+1]$.
   Hint: *You may draw a figure and invoke a geometric argument. The piecewise affine function f and the functions ℓ_n are built from a well-known smooth convex function. How?*

   b) Deduce that f is convex.

2. Compute $\partial f(x)$ for all $x \in \mathbb{R}$.
   Hint: *You may distinguish the cases $x = n$ and $x \in]n, n+1[$.*

3. a) What is the geometric interpretation of the Legendre-Fenchel conjugate $f^*(s)$?

   b) Compute $f^*(a_n)$.

   c) More generally, compute $f^*(s)$ for $s \in]a_{n-1}, a_n[$.

4. Let $\lambda > 0$, $y \in \mathbb{R}$, and consider the following variational problem

$$\min_{x\in\mathbb{R}} f(x) + \frac{1}{2\lambda}\left|x - y\right|^2 \tag{5}$$

   a) Prove that there exists a unique solution to (5).

   b) Write the optimality conditions and prove that the solution is given by

$$x = \begin{cases} n & \text{for } n(1+\lambda) - \frac{\lambda}{2} \leq y \leq n(1+\lambda) + \frac{\lambda}{2}, \\ y - \lambda\left(n + \frac{1}{2}\right) & \text{for } n(1+\lambda) + \frac{\lambda}{2} < y < (n+1)(1+\lambda) - \frac{\lambda}{2}, \end{cases}$$

   where $n \in \mathbb{Z}$ is uniquely determined.

Part 2 (IASD only)

The following exercises deal with the second part of the course and should only be addressed by IASD students.

Ex. 5 — Consider the functions $F(w) = y \log(\text{sigmoid}(w^\top x))$ and $G(w) = (1-y)\log(1-\text{sigmoid}(w^\top x))$, where $w \in \mathbb{R}^d$, $x \in \mathbb{R}^d$, $y \in [0,1]$ and $\text{sigmoid}(a) = \frac{1}{1+e^{-a}}$.

1. Assume the following primitives: dot product, sigmoid, log and multiplication. Using intermediate variables a, b, c and d, write $d = F(w)$ as a chain of operations. Write the sequence of operations in forward order.
2. Using backpropagation, compute $\frac{\partial d}{\partial w}$. Write the sequence of operations in backward order from $\frac{\partial d}{\partial d} = 1$ to $\frac{\partial d}{\partial w}$. Reminder: sigmoid'$(a) = \text{sigmoid}(a)(1 - \text{sigmoid}(a))$.
3. Assume the following additional primitives: addition and subtraction. Using intermediate variables e, f, g, h, write the operations needed to compute $h = F(w) + G(w)$ in forward order. You should reuse a, b, c and d.
4. Using backpropagation, compute $\frac{\partial h}{\partial w}$. Write the operations in backward order from $\frac{\partial h}{\partial h} = 1$ to $\frac{\partial h}{\partial w}$.

Ex. 6 — We consider a function $L : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$.

1. Prove that the existence of a saddle point (x^*, y^*), i.e. a point in $\mathbb{R}^d \times \mathbb{R}^m$ that satisfies

$$L(x^*, y) \leq L(x, y^*), \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^m,$$

implies strong duality.

2. We consider the special case $L(x, y) = f(x) + y^\top h(x)$, with $f : \mathbb{R}^d \to \mathbb{R}$ convex differentiable and $h : \mathbb{R}^d \to \mathbb{R}^m$ affine. Prove that a point $(x^*, y^*) \in \mathbb{R}^d \times \mathbb{R}^m$ satisfying

$$\nabla_x L(x^*, y^*) = 0, \quad \nabla_y L(x^*, y^*) = 0$$

is a saddle point.

Ex. 7 — Let $f \in C^1(\mathbb{R}^2)$ be an objective function. We want to study the problem

$$\min_{x \in \mathbb{R}^2} f(x) \quad \text{s.t.} \quad \begin{cases} h_1(x) \leq 0, \\ h_2(x) \leq 0. \end{cases} \tag{6}$$

where $h_1(x) \overset{\text{def}}{=} x_2 + x_1^3$ and $h_2(x) \overset{\text{def}}{=} -x_2$.

1. Let A be the feasible set,

$$A \overset{\text{def}}{=} \{x \in \mathbb{R}^2; h_1(x) \leq 0 \text{ and } h_2(x) \leq 0\}.$$

Draw the set A.

2. Let $x^* \in A$. Assume that the Karush-Kuhn-Tucker (KKT) conditions hold at x^*.
   a) Write the KKT conditions at x^* for Problem (6).

b) For x^* in each of the following subsets of A,

$$A_1 \overset{\text{def}}{=} \{x \in \mathbb{R}^2; x_2 = 0, x_1 < 0\}, \qquad A_2 \overset{\text{def}}{=} \{x \in \mathbb{R}^2; x_2 = -x_1^3, x_1 < 0\},$$

$$A_3 \overset{\text{def}}{=} \{x \in \mathbb{R}^2; x_2 > 0, x_2 + x_1^3 < 0\}, \qquad A_4 \overset{\text{def}}{=} \{(0,0)\},$$

what do the KKT conditions tell us about $\nabla f(x^*)$?

3. Let x^{**} be a solution to Problem (6). Do the KKT conditions necessarily hold at x^{**}? *Discuss the different cases, $x^{**} \in A_i$, $i \in \{1,2,3,4\}$.*
Please justify your answer (in the case(s) where the answer is no, provide a counterexample, and propose a "good" necessary condition).

Answer (Ex. 1) — 1. If (x, y) are minimizers and $t + s = 1$, $t, s \geq 0$, then for any z, $f(x), f(y) \leq f(z)$, thus by convexity $f(tx + sy) \leq tf(x) + sf(y) \leq (s + t)f(z) = f(z)$. Thus $tx + sy$ is also a minimizer.

2. a) f and g are positive, and $f(x) = 0$ is equivalent to $x = y$, while $g(a, b) = 0$ is equivalent to $ab = y$.

   b) f is convex and coercive as a square function. The set of minimizer of g is not convex and is not bounded, so that g is neither convex nor coercive.

3. a) $f'(x) = x - y$ and $f''(x) = 1$,

$$\nabla g(a, b) = \begin{pmatrix} b(ab - y) \\ a(ab - y) \end{pmatrix}, \quad \partial^2 g(a, b) = \begin{pmatrix} b^2 & 2ab - y \\ 2ab - y & a^2 \end{pmatrix}$$

   b) We denote as (λ, μ) the (real) eigenvalues of the matrix. Positivity is equivalent to $\lambda, \mu > 0$, which is itself equivalent to $\lambda + \mu > 0$ and $\lambda\mu > 0$.

   c) One has

$$\det(\partial^2 g(a, b)) = b^2 a^2 - (2ab - y)^2 = (y - ab)(3ab - y).$$

   The matrix is positive if either (i) $y - ab > 0$ and $3ab - y > 0$ or if (ii) $y - ab < 0$ and $3ab - y < 0$. (i) defines a region enclosed by two hyperbolas, while (ii) is empty.

4. a) $\text{Argmin}(f) = \{X\}$ and $\text{Argmin}(g) = \{(A, B) : AB = Y\}$. f is convex and coercive because it is a squared Euclidean norm function. g is not coercive because it set of minimizers is not bounded (take $(\lambda A, B/\lambda)$ for $AB = Y$ and $\lambda \to +\infty$. We show that $\text{Argmin}(g)$ is not a convex set by considering (Y, Id) and $(2Y, \text{Id}/2)$ which are minimizers while their mean $(3Y/2, 3\text{Id}/4)$ is not.

   b) One has $\nabla f(X) = X - Y$ and $\partial^2 f(X)$ is the identity mapping. One has

$$\nabla g(a, b) = \begin{pmatrix} (AB - Y)B^\top \\ A^\top (AB - Y) \end{pmatrix},$$

Answer (Ex. 2) — 1. a) Let k be fixed. We apply the characterization of convexity for differentiable functions: at x_*, f is above its tangent at x_k, that is

$$f(x_*) \geq f(x_k) + \langle \nabla f(x_k), x_* - x_k \rangle,$$

which is equivalent to the desired inequality.

b) For any k,

$$\begin{aligned} ||x_{k+1} - x_*||_2^2 &= ||x_k - x_* - h_k \nabla f(x_k)||_2^2 \\ &= ||x_k - x_*||_2^2 - 2h_k \langle \nabla f(x_k), x_k - x_* \rangle + h_k^2 ||\nabla f(x_k)||_2^2 \\ &\overset{1.a)}{\leq} ||x_k - x_*||_2^2 - 2h_k(f(x_k) - f(x_*)) + h_k^2 ||\nabla f(x_k)||_2^2. \end{aligned}$$

c) We deduce from the previous question that, for any $k \in \mathbb{N}$,

$$2h_k(f(x_k) - f(x_*)) \leq ||x_k - x_*||_2^2 - ||x_{k+1} - x_*||_2^2 + h_k^2 ||\nabla f(x_k)||_2^2.$$

Therefore, for any $n \in \mathbb{N}$,

$$\begin{aligned}2\sum_{k=0}^{n} h_k(f(x_k) - f(x_*)) &\leq \sum_{k=0}^{n}(||x_k - x_*||_2^2 - ||x_{k+1} - x_*||_2^2) + \sum_{k=0}^{n} h_k^2||\nabla f(x_k)||_2^2 \\ &= ||x_0 - x_*||_2^2 - ||x_{n+1} - x_*||^2 + \sum_{k=0}^{n} h_k^2||\nabla f(x_k)||_2^2.\end{aligned}$$

d) Let n be fixed. For any $k \leq n$, we have, from the definition of k_n, $f(x_{k_n}) \leq f(x_k)$. As a consequence, for any $k \leq n$,

$$2h_k(f(x_{k_n}) - f(x_*)) \leq 2h_k(f(x_k) - f(x_k)).$$

and

$$\begin{aligned}2(f(x_{k_n}) - f(x_*))\left(\sum_{k=0}^{n} h_k\right) &\leq 2\sum_{k=0}^{n} h_k(f(x_k) - f(x_k)) \\ &\overset{1.c)}{\leq} ||x_0 - x_*||_2^2 - ||x_{n+1} - x_*||^2 + \sum_{k=0}^{n} h_k^2||\nabla f(x_k)||_2^2.\end{aligned}$$

e) From our third assumption on f, $||\nabla f(x_k)||_2 \leq 1$ for any $k \in \mathbb{N}$. Therefore, for any $n \in \mathbb{N}$,

$$\sum_{k=0}^{n} h_k^2||\nabla f(x_k)||_2^2 \leq \sum_{k=0}^{n} h_k^2.$$

Since, in addition, $-||x_{n+1} - x_*||_2^2 \leq 0$, we deduce from question 1.d) that

$$2(f(x_{k_n}) - f(x_*))\left(\sum_{k=0}^{n} h_k\right) \leq ||x_0 - x_*||_2^2 + \sum_{k=0}^{n} h_k^2.$$

2. For any $n \in \mathbb{N}$,

$$\begin{aligned}\sum_{k=0}^{n} h_k^2 &= \sum_{k=1}^{n+1} \frac{1}{k} \leq 2 + \log(n); \\ \sum_{k=0}^{n} h_k &= \sum_{k=1}^{n+1} \frac{1}{\sqrt{k}} \geq \frac{\sqrt{n+2}}{2}.\end{aligned}$$

Plugging these inequalities into the one established at question 1.e) yields

$$\begin{aligned}(f(x_{k_n}) - f(x_*))\sqrt{n+2} &\leq 2(f(x_{k_n}) - f(x_*))\left(\sum_{k=0}^{n} h_k\right) \\ &\leq ||x_0 - x_*||_2^2 + \sum_{k=0}^{n} h_k^2 \\ &\leq ||x_0 - x_*||_2^2 + 2 + \log(n).\end{aligned}$$

Therefore,

$$f(x_{k_n}) - f(x_*) \le \frac{||x_0 - x_*||_2^2 + 2 + \log(n)}{\sqrt{n+2}}.$$

3. We set $\epsilon = \frac{\eta}{2}$ and define f as suggested:

$$\begin{array}{rcll} f : x \in \mathbb{R} & \to & |x| - \frac{\epsilon}{2} & \text{if } |x| \ge \epsilon; \\ & & \frac{x^2}{2\epsilon} & \text{if } |x| \le \epsilon. \end{array}$$

Let us show that f satisfies properties (1), (2), (3).
We start with property (2). For any $x \in \mathbb{R}$ such that $|x| \ge \epsilon$,

$$f(x) \ge \epsilon - \frac{\epsilon}{2} > 0.$$

For any $x \in \mathbb{R}$ such that $|x| < \epsilon$,

$$f(x) = \frac{x^2}{2\epsilon} \ge 0.$$

Therefore, f is nonnegative over $\mathbb{R}$. Since $f(0) = 0$, it implies that $x_* = 0$ is a global minimizer of f.
Let us now show that f is differentiable and compute its derivative. The function $|.|$ is differentiable over $\mathbb{R} - \{0\}$ so f is differentiable over $]-\infty; -\epsilon] \cup [\epsilon; +\infty[$, with derivative

$$\begin{aligned} f'(x) &= -1 \quad \forall x \in]-\infty; -\epsilon]; \\ f'(x) &= 1 \quad \forall x \in [\epsilon; +\infty[. \end{aligned}$$

(The derivative is only a left derivative when $x = -\epsilon$ and a right derivative when $x = \epsilon$.)
The square function is differentiable over $\mathbb{R}$ so f is differentiable over $[-\epsilon; \epsilon]$, with derivative

$$f'(x) = \frac{x}{\epsilon} \quad \forall x \in [-\epsilon; \epsilon].$$

(The derivative is only a right derivative when $x = -\epsilon$ and a left derivative when $x = \epsilon$.)
Since the left and right derivatives coincide in $x = -\epsilon$ and $x = \epsilon$, the function f is differentiable at $-\epsilon$ and ϵ and therefore differentiable over $\mathbb{R}$.
For any x such that $|x| \ge \epsilon$, we have $|f'(x)| = 1$ and, for any x such that $|x| \le \epsilon$, we have $|f'(x)| = \frac{|x|}{\epsilon} \le 1$. As a consequence, the norm of the gradient (that is, in this case, the norm of the derivative), is always at most 1 and Property (3) holds.
Now that we have computed the derivative, we can easily show that f is convex: its derivative is continuous, nondecreasing (actually constant) over $]-\infty; -\epsilon]$, increasing over $[-\epsilon; \epsilon]$, nondecreasing again over $[\epsilon; +\infty[$. Therefore, the derivative is nondecreasing over $\mathbb{R}$ and f is convex.
We consider the starting point $x_0 = \frac{\eta}{2} = \epsilon$. With this definition,

$$\begin{aligned} x_1 &= x_0 - h_0 f'(x_0) \\ &= \epsilon - \eta \times 1 \\ &= -\epsilon \end{aligned}$$

and

$$\begin{aligned} x_2 &= x_1 - h_0 f'(x_1) \\ &= -\epsilon - \eta \times (-1) \\ &= \epsilon. \end{aligned}$$

We can iteratively reapply this reasoning and we obtain that $x_k = -\epsilon$ for all odd k and $x_k = \epsilon$ for all even k. In particular, $x_k \not\to x_* = 0$ when $k \to +\infty$.

Answer (Ex. 3) — 1. a) To obtain an instance of gradient descent, we set $g_k = \nabla f(w_k)$.
   b) To obtain an instance of stochastic gradient, we select an index i_k at random and set $g_k = \nabla f_i(w_k)$.

2. An epoch consists in n accesses to a data point. As a result,
   a) Every iteration of gradient descent corresponds to an epoch;
   b) Every iteration of stochastic gradient corresponds to $\frac{1}{n}$th of an epoch.
3. By definition of the expected value, one has

$$\begin{aligned} \mathbb{E}_{\mathcal{S}_k}[g_k] &= \sum_{\mathcal{S} \subseteq \{1,\dots,n\}} \mathbb{P}(\mathcal{S}_k = \mathcal{S}) \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \nabla_i f(w_k) \\ &= \sum_{\substack{\mathcal{S} \subseteq \{1,\dots,n\} \\ |\mathcal{S}| = n_b}} \frac{1}{\binom{n}{n_b}} \frac{1}{n_b} \sum_{i \in \mathcal{S}} \nabla_i f(w_k) \\ &= \frac{1}{\binom{n}{n_b}} \frac{1}{n_b} \sum_{\substack{\mathcal{S} \subseteq \{1,\dots,n\} \\ |\mathcal{S}| = n_b}} \sum_{i \in \mathcal{S}} \nabla_i f(w_k) \end{aligned}$$

The random set $\mathcal{S}_k$ takes $\binom{n}{n_b}$ possible values. If we consider any index $i \in \{1, \dots, n\}$, this index appears in exactly $\binom{n-1}{n_b-1}$ index sets of cardinality n_b out of the possible $\binom{n}{n_b}$. Therefore,

$$\sum_{\substack{\mathcal{S} \subseteq \{1,\dots,n\} \\ |\mathcal{S}| = n_b}} \sum_{i \in \mathcal{S}} \nabla_i f(w_k) = \binom{n-1}{n_b-1} \sum_{i=1}^{n} \nabla f_i(w_k).$$

Using $\binom{n}{n_b} = \frac{n}{n_b}\binom{n-1}{n_b-1}$, we thus obtain

$$\begin{aligned} \mathbb{E}_{\mathcal{S}_k}[g_k] &= \frac{1}{\binom{n}{n_b}} \frac{1}{n_b} \binom{n-1}{n_b-1} \sum_{i=1}^{n} \nabla f_i(w_k) \\ &= \frac{1}{\binom{n}{n_b}} \frac{1}{n_b} \frac{n_b}{n} \binom{n}{n_b} \sum_{i=1}^{n} \nabla f_i(w_k) \\ &= \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w) = f(w), \end{aligned}$$

which is the desired result.

4. a) This result shows that stochastic gradient methods with a constant step size can only be guaranteed to converge to a neighborhood of the optimal value. It also shows that this neighborhood becomes tighter as n_b grows (and that the bound is overly pessimistic when $n_b = n$). *Note: Bonus point for the students who point out the case $n_b = n$.*

   b) Below is a list of valid answers, that we covered in the lectures:

   - Use a dynamic (growing batch size) instead of a constant one;
   - Use a decreasing step size sequence instead of a constant one;
   - Use a gradient aggregation technique (SAGA, SAG, SVRG) instead of a basic stochastic gradient update;
   - Return the average of the iterates, that possesses better convergence guarantees.

5. If stochastic gradient ($n_b = 1$) improves over gradient descent ($n_b = n$), this indicates there is enough correlation in the data to converge using random subsets of it at every iteration. Using more than one data point yet significantly less than n (mini-batching) can reduce the variance of the gradient estimates while remaining significantly cheaper than a full gradient estimation: this can explain why $n_b = n/10$ yields better performance than $n_b = 1$. When the batch size gets closer to n, its cost also gets closer to that of a full gradient iteration, and the method becomes at risk of suffering from redundancies in the data. This can explain why the behavior of the method worsens when $n_b > n/10$. *Note: This is an open question, I expect the students to be able to a) provide intuition as to why stochastic gradient works better than gradient descent and b) distinguish the mini-batch regime (n_b relatively small) from the regime $n_b \approx n$, where the method tends to behave like gradient descent. We discussed these aspects during our lab session.*

Answer (Ex. 4) — 1. a) Les réponses avec calculs sont bien sûr acceptées, mais voici un argument géométrique.

Introducing the convex function $g\colon x \mapsto \frac{1}{2}x^2$, we note that the graph of ℓ_n is the chord from $(n, g(n))$ to $(n+1, g(n+1))$.

As a consequence the graph of g is above the one of ℓ_p in $\mathbb{R}\backslash]p, p+1[$. In particular, for all $n \in \mathbb{Z}$,

$$\ell_n(n) = g(n) \geq \ell_p(n) \quad \text{and} \quad \ell_n(n+1) = g(n+1) \geq \ell_p(n+1).$$

Since $\ell_p - \ell_n$ is affine (hence convex), the inequality $\ell_p - \ell_n \leq 0$ also holds in $[n, n+1]$.

b) Let $n \in \mathbb{Z}$. We note that $f(x) = \ell_n(x)$ for all $x \in [n, n+1]$. As a, result, $f(x) \leq \sup_{n\in\mathbb{Z}} \ell_n(x)$.

On the other hand, we know that for all $p \in \mathbb{Z}$, $\ell_p(x) \leq \ell_n(x) = f(x)$. As a consequence, for all $x \in \mathbb{R}$,

$$f(x) = \sup_{n\in\mathbb{Z}} \ell_n(x). \tag{7}$$

The function f is convex as the supremum of affine (hence convex) functions.

2. For $x \in]n, n+1[$, f is convex, differentiable at x, hence $\partial f(x) = \{f'(x)\} = \{a_n\}$. For $x = n$, we note that both ℓ_n and ℓ_{n+1} are affine minorants of f which are exact at x. As a result, $\{a_n, a_{n+1}\} \subseteq \partial f(x)$, and since $\partial f(x)$ is convex, $[a_n, a_{n+1}] \subseteq \partial f(x)$. On the other hand, if $s > a_{n+1}$ we see that $f(x+\varepsilon) < f(x) + s(x+\varepsilon-x)$, which contradicts $s \in \partial f(x)$, and similarly if $s < a_n$, $f(x-\varepsilon) < f(x) + s(x-\varepsilon-x)$ contradicts $s \in \partial f(x)$. As a conclusion, $\partial f(x) = [a_n, a_{n+1}]$.

On peut aussi accepter les réponses mentionnant la dérivée à gauche et la dérivée à droite, $\partial f(x) = [f'_g(x), f'_d(x)]$, même si je n'en ai pas explicitement parlé.

3. a) The quantity $-f^*(s)$ is the intercept of the "best" affine minorant of f with slope s.

b) For all $n \in \mathbb{Z}$, we know that ℓ_n is an exact affine minorant of f with slope a_n. As a result, $f^*(a_n) = -\ell_n(0) = \frac{1}{2}n(n+1)$.

c) The function $\ell \colon x \mapsto \frac{1}{2}n^2 + s(x-n)$ for $n \in]a_{n-1}, a_n[$, satisfies $\ell(n) = f(n)$, and it is an affine minorant of f:

$$\forall x < n, \quad \ell(x) \leq \ell_{n-1}(x) \leq f(x), \tag{8}$$

$$\forall x > n, \quad \ell(x) \leq \ell_n(x) \leq f(x). \tag{9}$$

As a result, ℓ is an exact affine minorant of f with slope s, and $f^*(s) = -\ell(0) = -\frac{1}{2}n^2 + sn$.

4. a) We have $h(x) \overset{\text{def}}{=} f(x) + \frac{1}{2}|x-y|^2 \geq \frac{1}{2}|x-y|^2$. Since $x \mapsto \frac{1}{2}|x-y|^2$ is coercive, so is h. Since h is continuous (convex and everywhere defined), we obtain the existence of a minimizer.
That minimizer is unique since h is strictly convex (sum of a convex and a strictly convex function).

b) The function h and $x \mapsto \frac{1}{2}|x-y|^2$ are convex, l.s.c., proper and there is a common point in the relative interior of their respective domains. As a result, for all $x \in \mathbb{R}$,

$$\partial h(x) = \partial f(x) + \partial\left(\frac{1}{2\lambda}|\cdot - y|^2\right)(x) = \partial f(x) + \frac{1}{\lambda}(x-y). \tag{10}$$

A point x is thus solution iff

$$0 \in \partial f(x) + \frac{1}{\lambda}(x-y)$$

$$\text{or equivalently} \quad x = y - \lambda s, \quad s \in \partial f(x). \tag{11}$$

Now, we discuss the form of x. A point of the form $x = n$, $n \in \mathbb{Z}$, satisfies (11) iff

$$n = y - \lambda s, \quad a_{n-1} \leq s \leq a_n$$

$$\iff -\frac{\lambda}{2} \leq y - n(1+\lambda) \leq \frac{\lambda}{2}.$$

On the other hand, a point of the form $x \in]n, n+1[$ is satisfies (11) iff

$$x = y - \lambda a_n, \quad n < x < n+1$$

$$\iff x = y - \lambda\left(n + \frac{1}{2}\right), \quad \frac{\lambda}{2} < y - n(1+\lambda) < \frac{\lambda}{2} + 1.$$

This is a kind of "soft quantization"
The collection of $\{[n(1+\lambda) - \frac{\lambda}{2}, (n+1)(1+\lambda) - \frac{\lambda}{2}[\}_{n \in \mathbb{Z}}$ yields a partition of $\mathbb{R}$. The interval which contains y thus uniquely determines n.

Answer (Ex. 5) — 1. The chain can be decomposed as

- $a = w^\top x$
- $b = \text{sigmoid}(a)$
- $c = \log b$
- $d = y \cdot c$

2. Backpropagation is unrolled as follows
   - $\frac{\partial d}{\partial d} = 1$
   - $\frac{\partial d}{\partial c} = \frac{\partial d}{\partial d}\frac{\partial d}{\partial c} = y$
   - $\frac{\partial d}{\partial b} = \frac{\partial d}{\partial c}\frac{\partial c}{\partial b} = \frac{y}{b}$
   - $\frac{\partial d}{\partial a} = \frac{\partial d}{\partial b}\frac{\partial b}{\partial a} = \frac{y}{b}\text{sigmoid'}(a)$
   - $\frac{\partial d}{\partial w} = \frac{\partial d}{\partial a}\frac{\partial a}{\partial w} = \frac{y}{b}\text{sigmoid'}(a)x$
3. We add the following operations
   - $e = 1 - b$
   - $f = \log e$
   - $g = (1 - y)f$
   - $h = d + g$
4. Backpropagation is now unrolled as follows
   - $\frac{\partial h}{\partial h} = 1$
   - $\frac{\partial h}{\partial g} = \frac{\partial h}{\partial h}\frac{\partial h}{\partial g} = 1$
   - $\frac{\partial h}{\partial f} = \frac{\partial h}{\partial g}\frac{\partial g}{\partial f} = (1 - y)$
   - $\frac{\partial h}{\partial e} = \frac{\partial h}{\partial f}\frac{\partial f}{\partial e} = \frac{1-y}{e}$
   - $\frac{\partial h}{\partial d} = \frac{\partial h}{\partial h}\frac{\partial h}{\partial d} = 1$
   - $\frac{\partial h}{\partial c} = \frac{\partial h}{\partial d}\frac{\partial d}{\partial c} = y$
   - $\frac{\partial h}{\partial b} = \frac{\partial h}{\partial c}\frac{\partial c}{\partial b} + \frac{\partial h}{\partial e}\frac{\partial e}{\partial b} = \frac{y}{b} - \frac{1-y}{e}$
   - $\frac{\partial h}{\partial a} = \frac{\partial h}{\partial b}\frac{\partial b}{\partial a} = [\frac{y}{b} - \frac{1-y}{e}]\text{sigmoid'}(a)$
   - $\frac{\partial h}{\partial w} = \frac{\partial h}{\partial a}\frac{\partial a}{\partial w} = [\frac{y}{b} - \frac{1-y}{e}]\text{sigmoid'}(a)x$

Answer (Ex. 6) — 1. By definition of saddle point we have that

$$\sup_y L(x^*, y) - L(x, y^*) \leq 0, \quad \forall x.$$

Equivalently

$$\sup_y L(x^*, y) \leq L(x, y^*), \quad \forall x.$$

Now, let $g^* = \sup_y L(x^*, y)$, analogously, by the equation above, we have

$$\inf_x L(x, y^*) - g^* \geq 0,$$

then

$$\sup_y L(x^*, y) \leq \inf_x L(x, y^*).$$

The theorem is proven considering that

$$\inf_x \sup_y L(x,y) \leq \sup_y L(x^*,y)$$

and that

$$\sup_y \inf_x L(x,y) \geq \inf_x L(x,y^*).$$

So we obtain

$$\inf_x \sup_y L(x,y) \leq \sup_y \inf_x L(x,y).$$

This, together with weak duality gives strong duality. The students may provide a simpler proof for the case of max and min.

2. Let x^*, y^* satisfying

$$\nabla_x L(x^*,y^*) = 0, \quad , \nabla_y L(x^*,y^*) = 0.$$

Let $g(x) = L(x,y^*)$. g is differentiable and convex since it is the sum of a convex differentiable function and an affine function in x. Then its minimum (if exists) corresponds to the point satisfying $\nabla_x g = 0$, that in our case is exactly x^*. By convexity of g, we have

$$g(x) - g(x^*) \geq (\nabla_x g(x^*))^\top (x - x^*) = 0,$$

Then $L(x,y^*) \geq L(x^*,y^*)$ for all x. Analogously, since $\nabla_y L(x^*,y^*) := h(x^*) = 0$, it means that $L(x^*,y^*) = L(x^*,y)$ for any y, which implies $L(x^*,y^*) \geq L(x^*,y)$, then

$$L(x^*,y) \leq L(x^*,y^*) \leq L(x,y^*).$$

Answer (Ex. 7) — 1.

2. a) The KKT conditions at some point x^* mean that there exists $\mu^* \in \mathbb{R}^2$, such that

$$\begin{cases} \nabla f(x^*) + \mu_1^* \nabla h_1(x_*) + \mu_2^* \nabla h_2(x^*) & = 0 \\ h_1(x^*) \leq 0, \quad h_2(x^*) & \leq 0, \\ \mu_1^* \geq 0, \quad \mu_2^* & \geq 0 \\ \mu_1^* h_1(x^*) = 0, \quad \mu_2^* h_2(x^*) = 0. \end{cases}$$

b) For $x \in A_1$ we have $h_1(x^*) < 0$, hence $\mu_1^* = 0$. It means that $\nabla f(x^*) + \mu_2^* \nabla h_2(x^*) = 0$, that is,

$$\nabla f(x^*) = \begin{pmatrix} 0 \\ \mu_2^* \end{pmatrix}, \quad \mu_2^* \geq 0.$$

For $x^* \in A_2$, we have $h_2(x^*) < 0$ hence $\mu_2^* = 0$. Therefore, $\nabla f(x^*) + \mu_1^* \nabla h_1(x^*) = 0$, that is

$$\nabla f(x^*) = -\mu_1^* \begin{pmatrix} 3x_1^2 \\ 1 \end{pmatrix}, \quad \mu_1^* \geq 0.$$

For $x^* \in A_3$, there is no active constaint: $\mu_1^* = \mu_2^* = 0$. Hence $\nabla f(x^*) = 0$.

For $x^* \in A_4$, we have $\nabla h_1(x^*) = -\nabla h_2(x^*) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, hence

$$\nabla f(x^*) = \nu \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \nu \in \mathbb{R}.$$

3. For $x^{**} \in A_1$, only the constraint 2 is active, and $\nabla h_2(x^*) \neq 0$. Hence the linear independence qualification constraint (LICQ) holds, and the point x^{**} must satisfy the KKT conditions.
   The same argument and conclusion, using the constraint 1, holds for $x^{**} \in A_2$.
   For $x^{**} \in A_3$, there is no active constraint, hence locally the minimization is unconstrained and the well-known condition $\nabla f(x^{**}) = 0$ holds. The KKT conditions therefore hold (with $\mu_1^* = \mu_2^* = 0$). Alternatively, one may invoke the LICQ property with a trivial set of constraints.
   For $x^{**} \in A_4 = \{(0,0)\}$, one may note that the qualification condition LICQ does not hold (nor does the affine constraints condition). As a counterexample, the function $f(x) = -x_1$ reaches its minimum at $x^{**} = (0,0)$, but its gradient is $(-1,0)$, which contradicts the KKT condition. A "clean" necessary condition is that $-\nabla f(x^* *)$ is in the normal cone of A at $(0,0)$. First, we observe that the tangent cone of A at 0 is

$$\mathcal{T}_A(0) = \mathbb{R}_+ \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

   Indeed, if $A \ni x^{(k)} \overset{\text{def}}{=} 0 + t^{(k)} d + o(t^{(k)})$, then $0 \leq t^{(k)} d_2 \leq -(t^{(k)})^3 (d_1)^3 + o(t^{(k)})$ thus $d_2 = 0$. Moreover $d_1 \leq 0$. The converse inclusion follows from taking $x^{(k)} = (-t^{(k)}, 0)$.
   Taking the polar of $\mathcal{T}_A(0)$, we obtain,

$$-\nabla f(x^{**}) \in \mathcal{N}_A(0) = \{x \in \mathbb{R}^2;\ x_1 \geq 0\}.$$