**Profs. Martin Jaggi and Nicolas Flammarion**
**Optimization for Machine Learning – CS-439 - IC**
**08.0 2022 from 08h15 to 11h15**
**Duration : 180 minutes**

1

# Student One

SCIPER : **111111**

**Wait for the start of the exam before turning to the next page. This document is printed double sided, 16 pages. Do not unstaple.**

- This is a closed book exam. No electronic devices of any kind.
- Place on your desk: your student ID, writing utensils, one double-sided A4 page cheat sheet if you have one; place all other personal items below your desk or on the side.
- You each have a different exam.
- For technical reasons, **do use black or blue pens for the MCQ part, no pencils!** Use white corrector if necessary.

Respectez les consignes suivantes | Observe this guidelines | Beachten Sie bitte die unten stehenden Richtlinien

| choisir une réponse \| select an answer<br>Antwort auswählen | ne PAS choisir une réponse \| NOT select an answer<br>NICHT Antwort auswählen | Corriger une réponse \| Correct an answer<br>Antwort korrigieren |
|---|---|---|
| ☒ ☑ ■ | □ | □ □ |

ce qu'il ne faut **PAS** faire | what should **NOT** be done | was man **NICHT** tun sollte

▨ ☒ ○ ⊡ ▨ □

For your examination, preferably print documents compiled from auto-multiple-choice.

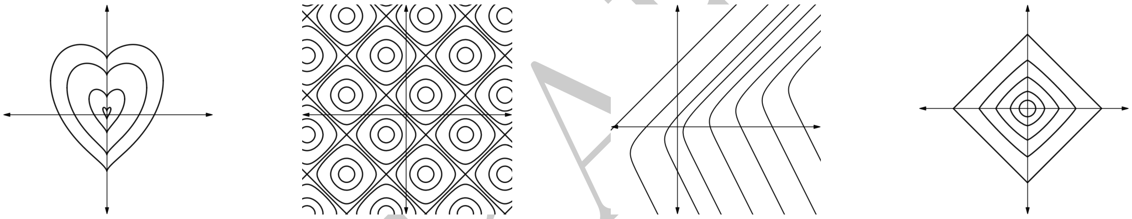
# First part, multiple choice

There is **exactly one** correct answer per question.

## Convexity

**Question 1** ***Definition:*** *For a (not necessarily convex) function $f : \mathbb{R}^2 \to \mathbb{R}$, an $\alpha$-level curve (also known as contour line) corresponds to the set $\{x \in \mathbb{R}^d, f(x) = \alpha\}$ where $\alpha \in \mathbb{R}$. We can represent these curves by drawing them for different values of $\alpha$. In Figure 1 (a) for example, each heart corresponds to an $\alpha$-level line of a certain function $f : \mathbb{R}^2 \to \mathbb{R}$ and for different values of $\alpha \in \mathbb{R}$. The two perpendicular lines with arrows at the end are the axis of the plot **and not level lines**.*
Which of the four plots in Figure 1 **could** correspond to the level curves of a **convex** function $f : \mathbb{R}^2 \to \mathbb{R}$.

- □ B and C.
- □ A and D.
- □ A and C.
- □ A and B.
- □ B and D.
- □ C and D.



(a) Level lines A (b) Level lines B (c) Level lines C (d) Level lines D

Figure 1: Several level lines for four different functions $f : \mathbb{R}^2 \to \mathbb{R}$

**Question 2** Assume we perform constant step-size stochastic gradient descent on $f(x) = \frac{1}{2}(f_1(x)+f_2(x))$, where $f_1(x) = (x-1)^2$ and $f_2(x) = (x+1)^2$ for $x \in \mathbb{R}$, i.e. $x_{t+1} = x_t - \gamma \nabla f_{i_t}(x_t)$ where at each iteration, $i_t$ is chosen uniformly random in $\{1, 2\}$. Which of the following statements is **false**:

- □ For $\gamma = 1$, we cannot guarantee that the iterates stay in a bounded set.
- □ $x = 0$ is the global minimum of $f$.
- □ Whatever the choice of constant step-size $\gamma > 0$, the iterates cannot converge as $t$ goes to infinity.
- □ For $\gamma = 2$, for any starting point $x_0$ and after the first iteration, the iterates will belong to $\{-1, +1\}$.

**Question 3** Considering the same setup as in the question above, but now assuming $f_1(x) = x^2$ and $f_2(x) = e^x$, $x \in \mathbb{R}$. After running $T$ steps of SGD, we find that $\nabla f_{i_T}(x_T) = 0$. Which of the following statements is **true**:

- □ $x_T$ is a local minimum but not a global minimum.
- □ $x_T = 0$.
- □ $x_T$ is a global (and local) minimum.
- □ None of the other choices.

**Question 4** Given a function $f : \mathbb{R}^d \to \mathbb{R}$ we want to minimize. We assume at each iteration $t$ a stochastic oracle is providing us with stochastic gradient $\mathbf{g}(\mathbf{x}_t)$ to run our SGD algorithm: $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \mathbf{g}(\mathbf{x}_t)$. We consider the following two stochastic oracles:

$$\mathbf{g}_{\mathrm{A}}(\mathbf{x}) := \begin{cases} 3\nabla f(\mathbf{x}), & \text{w. prob. } \frac{1}{3} \\ \varepsilon \sim \mathcal{N}(0,1), & \text{w. prob. } \frac{2}{3} \end{cases} \qquad \mathbf{g}_{\mathrm{B}}(\mathbf{x}) := \begin{cases} \nabla f(\mathbf{x}), & \text{w. prob. } \frac{1}{2} \\ -0.5\nabla f(\mathbf{x}), & \text{w. prob. } \frac{1}{2} \end{cases}$$

Which statement is true?

- □ Oracle A and B are both biased.
- □ Oracle A and B are both unbiased.
- □ Oracle A is unbiased, oracle B is biased.
- □ Oracle A is biased, oracle B is unbiased.

**Question 5**
Assume that we want to fit an affine line through a given point $(x_1, y_1) = (1, 1) \in \mathbb{R}^2$ (datapoint illustrated in Figure 2). To do so, we want to minimize the function $f(a, b) = (y_1 - (ax_1 + b))^2$ using gradient descent from a starting point $(a_0, b_0) = (0, 0)$. Using an appropriate and strictly positive step-size, the iterates $(a_t, b_t)_{t \in \mathbb{N}}$, will converge to:

- □ $(a^\star, b^\star) = (0, 1)$.
- □ $(a^\star, b^\star) = (1, 0)$.
- □ $(a^\star, b^\star) = (0.5, 0.5)$.
- □ $(a^\star, b^\star) = (-1, 2)$.

*HINT: do a drawing.*

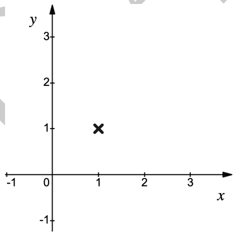


Figure 2: Plot of the datapoint $(x_1, y_1) = (1, 1)$.

## Smoothness and gradient descent

**Question 6** Define $f(x) := ax^2 + b$ for $x \in \mathbb{R}$. Consider running gradient descent with a constant-step size. For which one of the following statements, it is **not** possible to find a combination of starting point, step size and positive real numbers $a$ and $b$ where the statement happens at some step $t$.

- □ $x_{t+1} < 0 < x_{t+2} < x_t$.
- □ $x_{t+1} < x_{t+2} < 0 < x_t$.
- □ $x_{t+1} < 0 < x_t < x_{t+2}$.
- □ $x_{t+1} \neq x_t$ and $x_t = x_{t+2}$.

**Question 7** Define $f(x) := x^4$ with domain $D_f := [-2, 2]$. Assume we want to find a point $x_T$ with $f(x_T) \leq \varepsilon$ starting from $x_0 \in D_f$. Among the following statements, which is true and provides the tightest bound?

□ Using Nesterov acceleration and an appropriate step size, we have $T \in \mathcal{O}(\frac{1}{\varepsilon})$ since $f$ is smooth and convex over $D_f$.

□ Using Nesterov acceleration and an appropriate step size, we have $T \in \mathcal{O}(\frac{1}{\sqrt{\varepsilon}})$ since $f$ is smooth and convex over $D_f$.

□ Using vanilla Gradient Descent and an appropriate step size, we have $T \in \mathcal{O}(\frac{1}{\sqrt{\varepsilon}})$ since $f$ is smooth and convex over $D_f$.

□ Using vanilla Gradient Descent and an appropriate step size, we have $T \in \mathcal{O}(\log(\frac{1}{\varepsilon}))$ since $f$ is smooth and **strongly** convex over $D_f$.

**Question 8** Consider two algorithms $\mathcal{A}_1, \mathcal{A}_2$. For any $L$-smooth, $\mu$-strongly convex function $f$ with global minimum at $\mathbf{x}^\star$, assume the error after $T$ iterations while initialized at $\mathbf{x}_0$ satisfies:

$$Error(\mathcal{A}_1) = \frac{L\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{2T} \qquad Error(\mathcal{A}_2) = \left(1 - \frac{\mu}{L}\right)^T \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{2}.$$

Consider both algorithms for the minimization of a quadratic function $f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^2$ where $\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-3} \end{bmatrix}$ when initialized at $\mathbf{x}_0 = (1, 1)$. Consider the following statements

**A**: To get a target error of $\varepsilon = 10^{-2}$, $\mathcal{A}_2$ takes fewer iterations.

**B**: To get a target error of $\varepsilon = 10^{-7}$, $\mathcal{A}_2$ takes fewer iterations.

□ Only A is true.

□ Both A and B are true.

□ Neither A nor B is true.

□ Only B is true.

## Projected Gradient Descent

**Question 9** Consider the minimization of a $L$-smooth, convex function $f$ over a closed, convex set $\mathcal{X}$ using the projected gradient descent with learning rate $\gamma = \frac{1}{L}$. At iteration $t \geq 0$, we have

$$\mathbf{y}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t), \qquad \mathbf{x}_{t+1} := \Pi_{\mathcal{X}}\left(\mathbf{y}_{t+1}\right).$$

Which of the following properties is **false**?

□ $f(\mathbf{x}_t) \geq f(\mathbf{x}_{t+1})$.

□ $(\mathbf{x}_{t+1} - \mathbf{x}_t)^\top (\mathbf{y}_{t+1} - \mathbf{x}_{t+1}) \geq 0$.

□ $\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \leq \|\mathbf{x}_t - \mathbf{y}_{t+1}\|^2$.

□ None of the other choices.

## Subgradient Descent

**Question 10** Consider the function $f(x) = 4x - x^2$ for $x \in \mathbb{R}$. The gradient and subgradient at $x = 2$ are, respectively

□ 0, $[-1, 1]$.

□ 0, 0.

□ 0, doesn't exist.

□ 2 , 2.

## Frank-Wolfe

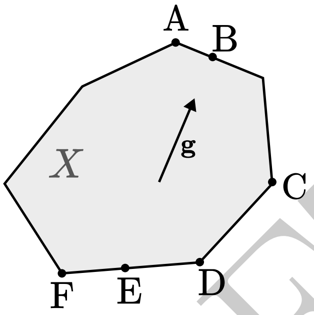


Figure 3: Gradient direction **g** over a convex set X.

**Question 11** Given the gradient direction **g** and the convex set $X$ as depicted in the above figure, what is a solution of the Linear Minimization Oracle $\mathrm{LMO}_X(\mathbf{g})$ ?

□ D

□ A and B

□ F

□ A

□ B

□ C

## Newton's Method and Quasi-Newton

**Question 12** Define $f(x) := x^4$ and $g(x) := x^3$ for $x \in \mathbb{R}$. Consider running Newton's method from an initial point $x_0 \in \mathbb{R}$ on each of these functions. Which one of the following statements is true:

□ Newton's method does not converge to 0 on at least one of the functions.

□ Newton's method converges to 0 on $g(x)$ with the same speed as on $f(x)$.

□ Newton's method converges to 0 on $f(x)$ faster than on $g(x)$.

□ Newton's method converges to 0 on $g(x)$ faster than on $f(x)$.

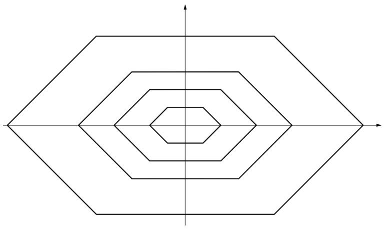
## Second part, true/false questions

**Question 13** (Convexity) If a function is strictly convex, then it is also strongly convex.

□ TRUE □ FALSE

**Question 14** (Strong Convexity) The following level curves (see Question 1 for the definition) **could** correspond to the level lines of a strongly convex function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$.

□ TRUE □ FALSE



**Question 15** (Nonconvex Convergence) For a nonconvex, L-smooth function $f$ with a global minimum, and running gradient descent with stepsize $\gamma := \frac{1}{L}$, we have

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq \mathcal{O}\Big(\frac{1}{T}\Big), \quad T > 0.$$

□ TRUE □ FALSE

**Question 16** (Coordinate-Wise Smoothness) Let $f : \mathbf{x} \in \mathbb{R}^d \rightarrow \mathbb{R}$ be a twice differentiable $L$-smooth function that is also smooth along the $i$-th coordinate with parameter $L_i$, for all $i$. We have $L = \max_{i=1}^d L_i$ if and only if $\nabla^2 f(\mathbf{x})$ is diagonal for all $\mathbf{x} \in \mathbb{R}^d$.

□ TRUE □ FALSE

**Question 17** (Coordinate Descent) Depending on the cost of each iteration, randomized coordinate descent **without** importance sampling can be faster than gradient descent.

□ TRUE □ FALSE

**Question 18** (Proximal-operator) The proximal mapping of any constant function $h$ i.e. $h(\mathbf{x}) = c, \forall \mathbf{x} \in \mathbb{R}^d$ is an identity mapping.

□ TRUE □ FALSE

**Question 19** (Lasso) For any vector $\mathbf{v} \in \mathbb{R}^d$, the projection onto the $\ell_1$-ball $\mathcal{B}$ i.e. $\mathcal{B} = \{\mathbf{x} : \|\mathbf{x}\|_1 \leq 1\}$, always lies on the boundary of the $\ell_1$-ball.

□ TRUE □ FALSE

**Question 20** (Frank-Wolfe) Consider the two constrained optimization problems $\min_{(x_1,x_2)\in[0,1]^2} x_1^2 + x_2^3$ with initial iterate $x_0 = [1,1]^\top$, and $\min_{(x_1,x_2)\in[0,10]\times[0,1]} (x_1/10)^2 + x_2^3$ with initial iterate $x_0 = [10,1]^\top$, after any number of iterations of the Frank-Wolfe algorithm, the optimization error for those two problems will be the same.

□ TRUE □ FALSE

**Question 21** (SGD) For L-smooth and convex functions. If we use an appropriate step-size sequence then Stochastic Gradient Descent (SGD) is guaranteed to strictly reduce the loss at each iteration.

□ TRUE □ FALSE

# Third part, open questions

Answer in the space provided! Your answer must be justified with all steps. Do not cross any checkboxes, they are reserved for correction.

In the whole exercise, we consider a convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ and denote by $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ one of its global minima.

## Convexity Preliminaries

Until the end of this section, we assume that the function $f$ is $L$-smooth and $\mu$-strongly convex.

**Question 22:** *1 point.* Prove the following inequalities for $\mathbf{x} \in \mathbb{R}^d$:

$$\frac{\mu}{2}\|\mathbf{x}-\mathbf{x}^\star\|^2 \le f(\mathbf{x}) - f(\mathbf{x}^\star) \le \frac{L}{2}\|\mathbf{x}-\mathbf{x}^\star\|^2.$$

□0 □1

**Question 23:** *3 point.* Prove the following inequalities for $\mathbf{x} \in \mathbb{R}^d$:

$$\frac{1}{2L}\|\nabla f(\mathbf{x})\|^2 \le f(\mathbf{x}) - f(\mathbf{x}^\star) \le \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2.$$

□0 □1 □2 □3

**Question 24:** *1 point.* Prove finally the following inequalities for $\mathbf{x} \in \mathbb{R}^d$:

$$\frac{1}{L^2}\|\nabla f(\mathbf{x})\|^2 \leq \|\mathbf{x} - \mathbf{x}^\star\|^2 \leq \frac{1}{\mu^2}\|\nabla f(\mathbf{x})\|^2.$$

□0 □1

## The Polyak stepsize rule

In this section, we consider the iterates of the gradient descent algorithm on the function $f$ with stepsize sequence $(\gamma_t)_{t\geq 0}$ defined as:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t) \quad \text{for } t \geq 0, \tag{1}$$

initialized at $\mathbf{x}_0 \in \mathbb{R}^d$. We investigate here a particular stepsize choice which is due to the Russian mathematician Boris Polyak, one of the founding father of modern optimization. Let us start by controlling the decrease of the distance to $\mathbf{x}^\star$, a minimizer of the function $f$.

**Question 25:** *1 point.* Show that the iterates $(\mathbf{x}_t)$ defined in Eq. (1) satisfy for $t \geq 0$:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\gamma_t\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) + \gamma_t^2\|\nabla f(\mathbf{x}_t)\|^2. \tag{2}$$

□0 □1

Boris Polyak argued that the optimal stepsize sequence should be chosen so that it minimizes the previous upper bound on $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2$ defined in Eq. (2). In the next question we derive such a formula.

**Question 26:** *2 points.* Let $t \geq 0$ and define $\gamma_t = \arg\min_{\gamma\geq 0} \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - 2\gamma\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2$. Derive the correct formula for $\gamma_t$.

□0 □1 □2

This stepsize is called the Polyak stepsize. From now on, we consider the iterates of gradient descent defined in Eq. (1) where the sequence $(\gamma_t)$ is defined in the previous question.
In the following sections, we study the rates of convergence of the gradient-descent algorithm with such stepsize. We denote by $\bar{\mathbf{x}}_T$ the iterate which satisfies $f(\bar{\mathbf{x}}_T) = \min_{0 \le t \le T-1} f(\mathbf{x}_t)$.

**Question 27:** *1 point.* Show that the iterates $(\mathbf{x}_t)$ defined with the Polyak stepsize satisfy for $t \geq 0$:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \le \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{(f(\mathbf{x}_t) - f(\mathbf{x}^\star))^2}{\|\nabla f(\mathbf{x}_t)\|^2}. \tag{3}$$

□0 □1

## Analysis under bounded gradients assumption

We assume in this section that the function $f$ has bounded gradients, i.e., there exists $B \geq 0$ such that $\|\nabla f(\mathbf{x})\| \leq B$ for all $\mathbf{x} \in \mathbb{R}^d$.

**Question 28:** *3 points.* Let $T \geq 1$. Show the following inequality:

$$\frac{1}{T}\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{B\|\mathbf{x}_0 - \mathbf{x}^\star\|}{\sqrt{T}}.$$

□0 □1 □2 □3

We assume until the end of the section that the function $f$ is $\mu$-strongly convex.

**Question 29:** *1 point.* Show for $t \geq 0$:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \frac{\mu^2\|\mathbf{x}_t - \mathbf{x}^\star\|^4}{4B^2}.$$

□0 □1

For $t \geq 0$, let us denote by $\beta_t = \frac{\mu^2 \|\mathbf{x}_t - \mathbf{x}^\star\|^2}{4B^2}$. We have then proven that:

$$\beta_{t+1} \leq \beta_t(1 - \beta_t).$$

**Question 30:** *3 points.* Show that for $t \geq 0$:

$$\beta_t \leq \frac{1}{t+1}.$$

□0 □1 □2 □3

Question 30 directly translates into a bound on the function values. But we will get a different one through a different analysis. Let us assume that $T \geq 2$ is even (odd $T$ would lead to similar result).

**Question 31:** *2 points.* Let $T \geq 2$ be an even number. Show that

$$\frac{2}{T} \sum_{t=T/2}^{T-1} \left[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right]^2 \leq \frac{16B^4}{\mu^2 T^2}.$$

□0 □1 □2

**Question 32:** *2 points.* Let $T \geq 2$ be an even number. Compare the bound on $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star)$ implied by Question 30 and 31 and explain which one is tighter.

□0 □1 □2

Although not proven here, a similar bound that the one proved in Question 32 also holds for any odd number $T$.

## Analysis under smoothness assumption

We assume in this section that the function $f$ is $L$-smooth.

**Question 33:** *2 points.* Let $T \geq 1$. Show that

$$\frac{1}{T}\sum_{t=0}^{T-1}\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{T}.$$

□0 □1 □2

We assume until the end of the section that the function $f$ is $\mu$-strongly convex.

**Question 34:** *2 points.* Let $T \geq 1$. Show that

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2(1 - \mu/(4L))^T.$$

□0 □1 □2

## Conclusion

Your results imply that

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^\star) \le \min\left\{ \frac{B\|\mathbf{x}_0 - \mathbf{x}^\star\|}{\sqrt{T}}, \frac{2L\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{T}, \frac{4B^2}{\mu T}, \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 (1 - \mu/(4L))^{T-1} \right\}.$$

**Question 35:** *1 point.* Compare this result with what you have seen in the course.

□0 □1

**Question 36:** *1 point.* Point out a major issue with the applicability of the Polyak stepsize-rule in practice.

□0 □1

**Question 37:** *1 point.* Do you see an application in modern machine learning where this should not be an issue?

□0 □1