EPFL

1

Profs. Martin Jaggi and Nicolas Flammarion
Optimization for Machine Learning – CS-439 - IC
01.07.2024 from 15h15 to 18h15
Duration : 180 minutes

Student One

SCIPER : **111111**

Wait for the start of the exam before turning to the next page. This document is printed double sided, 20 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 Antwort auswählen	ne PAS choisir une réponse | NOT select an answer NICHT Antwort auswählen	Corriger une réponse | Correct an answer Antwort korrigieren
ce qu'il ne faut **PAS** faire | what should **NOT** be done | was man **NICHT** tun sollte		

First part, multiple choice

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

Question 1 (Convexity & Smoothness) Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function such that

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \beta\langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{1}{2}\gamma\|\mathbf{x}-\mathbf{y}\|^2 \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d\,,$$

where $\beta > 0, \gamma \geq 0$ are constants. What are all the true statements about f?

A) f is convex.

B) f is Lipschitz.

C) If $\gamma > 0$, f is strongly convex with parameter γ.

D) None of these options.

- ☐ D
- ☐ A
- ☐ C
- ☐ B
- ☐ A and B
- ☐ A, B, and C
- ☐ B and C
- ☐ A and C

Question 2 (SGD) Consider the following optimization objective with $x \in \mathbb{R}$:

$$\min_x F(x) := \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[(x-\xi)^2]\,.$$

A stochastic gradient descent update step is defined as

$$x_{t+1} = x_t - 2\gamma_{t+1}(x_t - \xi_t), \qquad \text{where } \xi_t \sim \mathcal{N}(0,1)\,,$$

for all $t \geq 0$. If we run stochastic gradient descent with $x_0 = 5$ and decreasing learning rate $\gamma_t = \frac{1}{t+1}$, which of the following are true?

A) $\lim_{t\to\infty} \mathbb{E}\, F(x_t) = 1$

B) $\lim_{t\to\infty} \mathbb{E}\, F(x_t) = 1 + c$, where $c > 0$ is some constant dependent on x_0.

C) $\lim_{t\to\infty} \mathbb{E}\,|x_t|^2 = 0$

D) We do not have convergence guarantee on $\mathbb{E}\,|x_t|^2$

What are all the true statements?

- ☐ B and D
- ☐ B and C
- ☐ A and D
- ☐ A and C

Question 3 (Convergence Rates) Consider the following function:

$$f(x) = \begin{cases} |x|^3 & \text{for } |x| \leq 1, \\ 3|x| - 2 & \text{otherwise.} \end{cases}$$

with domain $D_f = \mathbb{R}$. Assume we want to find a point x_T with $f(x_T) \leq \varepsilon$ starting from an unknown $x_0 \in \mathbb{R}$. Which algorithm provides the tightest applicable bound (assuming appropriate hyperparameters)?

- ☐ Gradient Descent, $T \in \mathcal{O}(\log(\frac{1}{\varepsilon}))$
- ☐ Nesterov acceleration, $T \in \mathcal{O}(\frac{1}{\sqrt{\varepsilon}})$
- ☐ Gradient Descent, $T \in \mathcal{O}(\frac{1}{\varepsilon^2})$
- ☐ Newton's method, $T \in \mathcal{O}(\log \log \frac{1}{\varepsilon})$
- ☐ Gradient Descent, $T \in \mathcal{O}(\frac{1}{\varepsilon})$

Question 4 (PowerSGD) PowerSGD approximates a matrix $M \in \mathbb{R}^{m \times n}$ with $PQ^\top$ where $P \in \mathbb{R}^{n \times r}$ and $Q \in \mathbb{R}^{m \times r}$. What is the amount of data transmitted in PowerSGD relative to (divided by) that of standard distributed SGD?

- ☐ 1
- ☐ $\frac{mr+nr}{mn}$
- ☐ $\frac{m+n+r^2}{mn}$
- ☐ $\frac{r}{m+n}$
- ☐ $\frac{r}{mn}$
- ☐ $\frac{m+n-r}{m+n}$
- ☐ $\frac{1}{r}$

Question 5 (Projections) Let $X \subseteq \mathbb{R}^d$ be closed and convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d, \mathbf{z} \in \mathbb{R}^d$ and Π_X be the projection operator onto X. Which of the following statements are true?

A) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{x} - \Pi_X(\mathbf{z})\|^2 \leq \|2\mathbf{x} - (\mathbf{y} + \mathbf{z})\|^2$

B) $\|\mathbf{y} - \Pi_X(\mathbf{z})\| + \|\mathbf{x} - \Pi_X(\mathbf{z})\| \geq \|\mathbf{x} - \mathbf{y}\|$

C) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top(\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$

D) $(\mathbf{x} - \Pi_X(\mathbf{x}))^\top(\mathbf{y} - \Pi_X(\mathbf{z})) \leq 0$

E) $\|\mathbf{x} - \Pi_X(\mathbf{y})\| + \|\mathbf{x} - \Pi_X(\mathbf{z})\| \leq \|2\mathbf{x} - (\mathbf{y} + \mathbf{z})\|$

- ☐ A, D, and E
- ☐ B, C, and D
- ☐ B, C, and E
- ☐ A, B, and D
- ☐ A, C, and E

Question 6 (Non-convex optimization) Let f be a non-convex function on $\mathbb{R}^d$ that is lower bounded by some constant $B \in \mathbb{R}$, i.e., $f(\mathbf{x}) \geq B$ for all $\mathbf{x} \in \mathbb{R}^d$. Further, assume that f is smooth on a set $X \subset \mathbb{R}^d$ with parameter L. Let $x_1, \ldots, x_T$ be the trajectory obtained by running gradient descent with step size $\gamma = \frac{1}{L}$ from a $x_0 \in X$. Assume that for all x_0 and T, the trajectory stays within X. Which of the following statements are necessarily true?

A) $\lim_{T\to\infty} f(\mathbf{x}_T) = f(\mathbf{x}^\star)$ for some local minimum $x^\star$.

B) $\lim_{T\to\infty} \|\nabla f(\mathbf{x}_T)\| = 0$.

C) f has the sufficient decrease property in X.

D) The trajectory $x_0, x_1, \ldots, x_T$ convergences to a critical point.

- ☐ A and C
- ☐ A and B
- ☐ B and D
- ☐ A and D
- ☐ C and D
- ☐ B and C

Question 7 Let f be a L-smooth convex function and recall the Nesterov acceleration algorithm given by

$$\begin{aligned}
\mathbf{y}_{t+1} &:= \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t),\\
\mathbf{z}_{t+1} &:= \mathbf{z}_t - \frac{t+1}{2L}\nabla f(\mathbf{x}_t),\\
\mathbf{x}_{t+1} &:= \frac{t+1}{t+3}\mathbf{y}_{t+1} + \frac{2}{t+3}\mathbf{z}_{t+1}.
\end{aligned}$$

Which of the following statements are true ?

A) $f(\mathbf{y}_{t+1}) \leq f(\mathbf{x}_t)$.

B) $\|\mathbf{y}_{t+1} - \mathbf{x}_*\|^2 \leq \|\mathbf{x}_t - \mathbf{x}_*\|^2$, where $\mathbf{x}_*$ is a global minimum.

C) For any $t \geq 0$, we have $\|\mathbf{z}_t - \mathbf{x}_*\| \leq \|\mathbf{z}_0 - \mathbf{x}_*\|$, where $\mathbf{z}_0$ is the initialization.

- ☐ All of the above
- ☐ B and C
- ☐ A
- ☐ A and B
- ☐ B
- ☐ C
- ☐ A and C

Question 8 (Subgradients) Let $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^d$ be vectors such that none of them can be written as a linear combination of the others. Let X be the convex hull of this point set, i.e.:

$$X = \left\{ \mathbf{x} \in \mathbb{R}^d \mid \mathbf{x} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i, \sum_{i=1}^{N} \alpha_i \leq 1, \alpha_i \geq 0 \right\} .$$

Let $f : X \to \mathbb{R}$ be a function such that

$$f(\mathbf{x}) = \sum_{i=1}^{N} \beta_i b_i \quad \text{for } \mathbf{x} = \sum_{i=1}^{N} \beta_i \mathbf{x}_i \in X ,$$

where $b_1, \ldots, b_N \in \mathbb{R}$ are fixed scalars. Which of the following statements are necessarily true?

A) f is strongly convex.

B) $(b_1, \ldots, b_N) \in \partial f(x)$ for all x.

C) f has exactly one subgradient at all points.

D) None of these options.

- ☐ D
- ☐ A, B and C
- ☐ C
- ☐ B
- ☐ A and B
- ☐ A and C
- ☐ B and C
- ☐ A

Second part, true/false questions

Question 9 (Smoothness) Any twice differentiable function with bounded Hessians over some convex set X is smooth over X.

☐ TRUE ☐ FALSE

Question 10 (Convexity) Let f be a function on a convex set $X = \cup_{i=1}^{n} X_i$ for some fixed $n > 0$ and X_i are also convex. If f is convex on each X_i, then f is convex on X.

☐ TRUE ☐ FALSE

Question 11 (Convexity) Let f be a function on a convex set $X = \cup_{i=1}^{n} X_i$ for some fixed $n > 0$ and X_i are also convex. If f is strictly convex on each X_i, then there are at most n local minimums of f on X.

☐ TRUE ☐ FALSE

Question 12 (Stochastic Gradient Descent) If we perform stochastic gradient descent on a L-smooth function using step size $1/L$, assuming unbiased stochastic gradients, each update step guarantees a non-negative decrease in the training loss in expectation.

☐ TRUE ☐ FALSE

Question 13 (Non-Convex Optimization) Running stochastic gradient descent to minimize non-convex functions that are bounded from below guarantees convergence to a local minimum.

☐ TRUE ☐ FALSE

Question 14 (Newton) Running Newton's method on $\cos(x)$ for $x \in \mathbb{R}$ will result in convergence to a local minimum $(2k+1)\pi$ for some integer k.

☐ TRUE ☐ FALSE

Question 15 (SignSGD) Let $f(\mathbf{x}) := \mathbf{x}^\top \mathbf{u}$ with $\mathbf{x}, \mathbf{u} \in \mathbb{R}^d$. A SignSGD step is defined by $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \cdot \text{sign}(\mathbf{g}_t)$ where $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$ and the operations are performed element-wise.
Claim: There exists a $\mathbf{u} \in \mathbb{R}^d$ and $\gamma \geq 0$ such that a SignSGD update results in $f(\mathbf{x}_{t+1}) > f(\mathbf{x}_t)$.

☐ TRUE ☐ FALSE

Question 16 (Adam) An Adam update step can be formulated as:

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t$$
$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t^2$$
$$\mathbf{x}_t = \mathbf{x}_{t-1} - \gamma \cdot \frac{\mathbf{m}_t/(1-\beta_1^t)}{\sqrt{\mathbf{v}_t/(1-\beta_2^t)} + \varepsilon}$$

where all operations are performed element-wise, $\mathbf{x}_t \in \mathbb{R}^d$ is the parameter vector, $\mathbf{m}_t \in \mathbb{R}^d$ a momentum vector, $\mathbf{v}_t \in \mathbb{R}^d$ a vector of second moment estimates, and $\mathbf{g}_t \in \mathbb{R}^d$ the gradients w.r.t. $\mathbf{x}_t$. The momentum and second moment vectors are initialized to $\mathbf{0}$ at $t = 0$ and we only consider updates for $t \geq 1$.
Claim: In the limit $\varepsilon \to 0$, we can set the values of β_1, β_2 in a way that recovers the SignSGD update.

☐ TRUE ☐ FALSE

Question 17 (Federated Learning) In Federated Learning each worker shares their data with the coordinating server, but not other workers.

☐ TRUE ☐ FALSE

Question 18 (Adversarial Examples) The creation of adversarial examples for neural networks can be formulated as a constrained optimization problem.

☐ TRUE ☐ FALSE

Question 19 (Projected Gradient Descent) Projected gradient descent on the set,

$$B_1(R) := \left\{ \mathbf{x} \in \mathbb{R}^d \;\middle|\; \|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i| \leq R \right\},$$

can be implemented at the same time complexity (in $\mathcal{O}$) up to log factors as unconstrained gradient descent.
Hint: i.e., they have the same dependency on dimension d up to constants and log *terms.*

☐ TRUE ☐ FALSE

Question 20 (Proximal Gradient Descent) Let $g, h : \mathbb{R}^d \to \mathbb{R}$ be convex and smooth with parameter L and let $f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$. Considering minimizing over f with proximal gradient descent given proximal mapping for h and classical gradient descent for appropriately chosen stepsizes. The error as a function of the number of steps T scales the same for classical gradient descent and proximal gradient descent up to constants.

☐ TRUE ☐ FALSE

Question 21 (Proximal Gradient Descent) A step of proximal gradient descent on the function $f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}\|_2^2$ is equivalent to a step of gradient descent on the function $f(\mathbf{x})$ for properly chosen stepsizes.

☐ 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.

For the following, assume f is a function with a L-Lipschitz continuous gradient. Let μ be some positive constant (*not necessarily the same for every question*). We focus on a basic unconstrained optimization problem

$$\operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) \ .$$

Let $\mathbf{x}_p$ be the projection of $\mathbf{x}$ onto the solution set $\mathcal{X}^\star$, here we assume that the projection is well-defined. Let $f^\star = f(\mathbf{x})$, where $\mathbf{x} \in \mathcal{X}^\star$.
In the following few questions, we will study the convergence of standard iterative methods such as gradient decent and coordinate descent. We will slightly change the viewpoint and consider assumptions on the objective functions slightly different from the ones seen in class. The first few exercises in this part will be to better understand these assumptions, before then moving to the algorithm convergence analysis.

Relationship between conditions

We start by recalling the definition of strong convexity.

Strong Convexity. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 \ . \tag{SC}$$

Question 22: *1 point.* Show that strong convexity (SC) implies weak strong convexity (WSC).

Weak Strong Convexity. For any $\mathbf{x} \in \mathbb{R}^d$, we have

$$f^\star \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}_p - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{x}_p - \mathbf{x}\|^2 \ . \tag{WSC}$$

☐0 ☐1

Question 23: *1 point.*

Show that weak strong convexity (WSC) implies the restricted secant inequality (RSI).

Restricted Secant Inequality. For any $\mathbf{x} \in \mathbb{R}^d$, we have

$$\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}_p \rangle \geq \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}_p\|^2 \ . \tag{RSI}$$

☐0 ☐1

Question 24: *1 point.*

Show that the restricted secant inequality (RSI) implies the following error bound (EB).

Error Bound. For any $\mathbf{x} \in \mathbb{R}^d$, we have,

$$\|\nabla f(\mathbf{x})\| \geq \mu \|\mathbf{x} - \mathbf{x}_p\| \ . \tag{EB}$$

☐0 ☐1

Question 25: *2 points.* Assume that PL inequality (PL) implies quadratic growth (QG). Show that the above error bound (EB) is *equivalent* to having the PL inequality.

Polyak-Lojasiewicz Inequality. For any $\mathbf{x} \in \mathbb{R}^d$, we have,

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f^\star) . \tag{PL}$$

Quadratic Growth. For any $\mathbf{x} \in \mathbb{R}^d$, we have,

$$f(\mathbf{x}) - f^\star \geq \frac{\mu}{2}\|\mathbf{x}_p - \mathbf{x}\|^2 . \tag{QG}$$

☐0 ☐1 ☐2

In the following, we are going to show that the PL inequality (PL) implies the quadratic growth (QG) condition. To be able to prove this statement, we break the proof into substeps.

Question 26: *1 point.* Construct $g(\mathbf{x}) := \sqrt{f(\mathbf{x}) - f^\star}$, show that if PL holds for f, $\|\nabla g(\mathbf{x})\|$ is lower bounded for $\mathbf{x} \notin \mathcal{X}^\star$ as follows

$$\|\nabla g(\mathbf{x})\| \geq \sqrt{\frac{\mu}{2}}$$

☐ 0 ☐ 1

Question 27: *3 points.* Note that by construction, we have $g(\mathbf{x}) := \sqrt{f(\mathbf{x}) - f^\star} \geq 0$. Recall the optimal solution set of f, $\mathcal{X}^\star$ such that for all $\mathbf{y} \in \mathcal{X}^\star$, $g(\mathbf{y}) = 0$. Define $\mathbf{x}(t)$, for $t \geq 0$, as the solution of the differential equation, for $\mathbf{x}(t) \notin \mathcal{X}^\star$,

$$\frac{d\mathbf{x}(t)}{dt} = -\nabla g(\mathbf{x}(t)),$$
$$\mathbf{x}(0) = \mathbf{x}_0.$$

Show that $\mathbf{x}(t)$ satisfies the following property,

$$g(\mathbf{x}_0) - g(\mathbf{x}(t)) \geq \frac{\mu}{2} t.$$

Furthermore, let T be the time the curve $\mathbf{x}(t)$ hits $\mathcal{X}^\star$, i.e., $\lim_{t \to T} \mathbf{x}(t) \in \mathcal{X}^\star$. What can be said about T, is it bounded ?

☐0 ☐1 ☐2 ☐3

Question 28: *2 points.* The length of the any curve $(\mathbf{x}(t))_{t\geq 0}$ from time t_1 to t_2 is defined as,

$$\mathcal{L}(\mathbf{x}(t), t_1, t_2) = \int_{t_1}^{t_2} \left\| \frac{d\mathbf{x}(t)}{dt} \right\| dt.$$

Let T be the time the curve hits $\mathcal{X}^\star$. Using the definition of the length of the curve, show that

$$\int_0^T \|\nabla g(\mathbf{x}(t))\| dt \geq \|\mathbf{x}_0 - \mathbf{x}_p\|$$

☐0 ☐1 ☐2

Question 29: *2 points.* Combine the results from the previous questions (no matter if you have solved them or not), to finally prove that

$$g(\mathbf{x}_0) \geq \sqrt{\frac{\mu}{2}} \|\mathbf{x}_0 - \mathbf{x}_p\| \ .$$

As a consequence of this, we have that the quadratic growth condition (QG) holds directly.

☐0 ☐1 ☐2

Question 30: *1 point.* Until now, we can summarize that we have shown that

$$SC \Rightarrow WSC \Rightarrow RSI \Rightarrow EB \equiv PL \Rightarrow QG\ .$$

If we further assume f is convex, we can even claim that $RSI \equiv EB \equiv PL \equiv QG$. Can you prove this statement by showing QG implies RSI?

☐0 ☐1

Convergence Proof for Gradient and Coordinate Descent via the PL Inequality

Gradient Descent

Question 31: *2 points.* Consider the function f with $L-$Lipschitz continuous gradient, a non-empty solution set $\mathcal{X}^\star$, and satisfies the PL inequality with constant μ. Show that gradient method with a stepsize of $1/L$

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$$

has a global linear convergence rate, i.e.

$$f(\mathbf{x}_t) - f^\star \leq \left(1 - \frac{\mu}{L}\right)^t (f(\mathbf{x}_0) - f^\star)\ .$$

☐0 ☐1 ☐2

Randomized Coordinate Descent

Question 32: *3 points.*

Now, consider f with *coordinate-wise* $L-$Lipschitz-continuous gradient, i.e., for every coordinate $i = 1, \ldots, d$, we have,

$$f(\mathbf{x} + \alpha \mathbf{e}_i) \leq f(\mathbf{x}) + \alpha \nabla_i f(\mathbf{x}) + \frac{L}{2}\alpha^2.$$

As before, we continue to assume f has a non-empty solution set $\mathcal{X}^\star$, and satisfies the PL inequality with constant μ. Show that the coordinate descent method with a stepsize of $1/L$

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \frac{1}{L}\nabla_{i_t} f(\mathbf{x}_t)\mathbf{e}_t,$$

has an expected linear convergence rate if we choose the variable to update i_t uniformly at random, i.e.

$$\mathbb{E}[f(\mathbf{x}_t) - f^\star] \leq \left(1 - \frac{\mu}{dL}\right)^t [f(\mathbf{x}_0) - f^\star].$$

☐ 0 ☐ 1 ☐ 2 ☐ 3

Greedy Coordinate Descent

Question 33: *3 points.*

Let f satisfy the same conditions as in the last question, but now we sample i_t according to the rule $i_t = \text{argmax}_j[\nabla_j f(\mathbf{x}_t)]$. Further, assume f satisfies

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu_1}{2} \|\mathbf{y} - \mathbf{x}\|_1^2 \,,$$

which leads to the PL inequality in the ∞-norm:

$$\frac{1}{2} \|\nabla f(\mathbf{x})\|_\infty^2 \geq \mu_1 (f(\mathbf{x}) - f^\star)$$

Show that this new method achieves a linear convergence rate, i.e.

$$f(\mathbf{x}_t) - f^\star \leq \left(1 - \frac{\mu_1}{L}\right)^t [f(\mathbf{x}_0) - f^\star]$$

☐0 ☐1 ☐2 ☐3

General Understanding of the PL Inequality

Question 34: *3 points.* Let A be some non-zero matrix. Define $f(\mathbf{x}) := g(A\mathbf{x})$ for a σ-strongly convex function g. Derive the constant for the PL inequality for f.

☐ 0 ☐ 1 ☐ 2 ☐ 3

Question 35: *2 points.* If f satisfies the PL condition, does this imply that f is strongly convex? Justify your answer.

☐ 0 ☐ 1 ☐ 2