Profs. Martin Jaggi and Nicolas Flammarion
Optimization for Machine Learning – CS-439 - IC
11.08.2020 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 (handwritten or 11pt min font size) 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

▨ ☒ ○ ⊡ ▨ □

Pour votre examen, imprimez de préférence les documents compilés à l'aide de auto-multiple-choice.

First part, multiple choice

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

Convexity and Smoothness

For each of the functions below, verify whether they are (1) convex, (2) strictly convex, (3) strongly convex, and (4) smooth, in the sense of the definitions used in the course:

A. $f(x) = -2x, \quad x \in \mathbb{R}$

B. $f(x) = \sin(x), \quad x \in (\pi, 2\pi)$

C. $f(x) = \tanh(ax + b), \quad x \in \mathbb{R}$

D. $f(x) = x^4, \quad x \in \mathbb{R}$

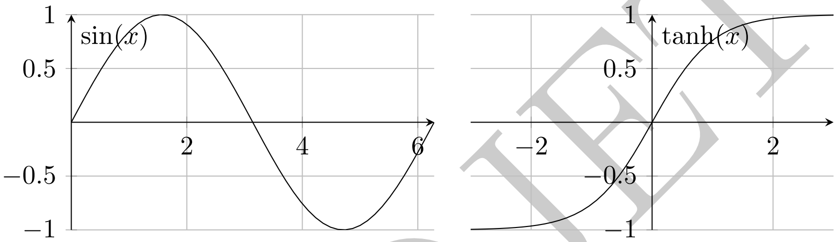
E. $f(x) = -\log(x), \quad x \in \mathbb{R}_{>0}$

F. $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2, \quad \mathbf{x} \in \mathbb{R}^2$

G. $f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^2,$

where

$$A := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \tanh(x) := \frac{e^{2x} - 1}{e^{2x} + 1}, \qquad a, b, a_i, b_i \in \mathbb{R}, \qquad \mathbf{a}, \mathbf{b} \in \mathbb{R}^2 .$$



Question 1 Given the function **A.** above, which are all of its properties?

□ convex + strictly convex

□ convex + strictly convex + strongly convex

□ convex + strictly convex + strongly convex + smooth

□ convex

□ convex + strictly convex + smooth

□ smooth

□ convex + smooth

□ none of these properties

Question 2 Given the function **B.** above, which are all of its properties?

□ convex

□ convex + strictly convex + strongly convex

□ smooth

□ convex + smooth

□ convex + strictly convex

□ convex + strictly convex + strongly convex + smooth

□ convex + strictly convex + smooth

□ none of these properties

Question 3 Given the function **C.** above, which are all of its properties?

- □ convex + strictly convex
- □ convex + strictly convex + strongly convex + smooth
- □ convex
- □ convex + smooth
- □ convex + strictly convex + smooth
- □ smooth
- □ convex + strictly convex + strongly convex
- □ none of these properties

Question 4 Given the function **D.** above, which are all of its properties?

- □ convex + strictly convex
- □ smooth
- □ convex + strictly convex + strongly convex
- □ convex + strictly convex + smooth
- □ convex
- □ convex + strictly convex + strongly convex + smooth
- □ convex + smooth
- □ none of these properties

Question 5 Given the function **E.** above, which are all of its properties?

- □ convex + smooth
- □ convex + strictly convex + strongly convex + smooth
- □ convex + strictly convex
- □ smooth
- □ convex + strictly convex + smooth
- □ convex + strictly convex + strongly convex
- □ convex
- □ none of these properties

Question 6 Given the function **F.** above, which are all of its properties?

- □ convex + strictly convex + smooth
- □ convex + strictly convex
- □ smooth
- □ convex + strictly convex + strongly convex + smooth
- □ convex + strictly convex + strongly convex
- □ convex
- □ convex + smooth
- □ none of these properties

Question 7 Given the function **G.** above, which are all of its properties?

□ convex + strictly convex + strongly convex + smooth
□ convex
□ convex + strictly convex
□ smooth
□ convex + strictly convex + smooth
□ convex + strictly convex + strongly convex
□ convex + smooth
□ none of these properties

Deep linear neural networks

Question 8 The output of a *linear network with more than one layer*, for a given input, as a function of the weight matrices,

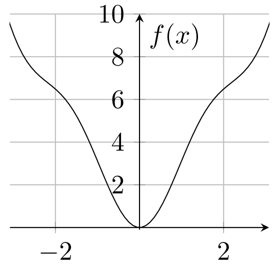
□ is a non-convex function, and equally or less expressive than a one-layer linear network.
□ is a non-convex function, and more expressive than a one-layer linear network.
□ is a convex function, and equally or less expressive than a one-layer linear network.
□ is a convex function, and more expressive than a one-layer linear network.

Question 9 Consider a deep linear neural network with 1-dim weights, input & output, with squared loss. Let $c \geq 1$ and $\delta > 0$ such that the initial point $\mathbf{x}_0 > \mathbf{0}$ is c-balanced with $\delta \leq \prod_k (\mathbf{x}_0)_k < 1$. Then the error $f(\mathbf{x}_t) - f^\star$ of gradient descent

□ converges to 0 as $\Theta(1/t)$, for an appropriate choice of step-size.
□ converges to 0 as $\Theta(1/t)$, for a constant step-size.
□ converges to 0 as $\Theta(1/\sqrt{t})$, for an appropriate choice of step-size, due to non-convexity.
□ converges to 0 exponentially fast, for a constant step-size.

Smoothness and gradient descent

Consider the function $f(x) = x^2 + 3\sin^2(x)$ for the next two questions plotted below.



Question 10 Which of the following properties does $f(x)$ satisfy?

□ Strongly convex and smooth with $L = 7$
□ Convex and smooth with $L = 5$
□ Convex and smooth with $L = 8$
□ Non-convex and smooth with $L = 8$
□ Non-convex and smooth with $L = 5$

Question 11 Suppose we run 1000 steps of gradient descent on $f(x)$ as above, with correct stepsizes. Which of the following is true about the error $f(\mathbf{x}_t) - f^\star$, relative to $f(\mathbf{x}_0) - f^\star$? Assume that the following inequality holds: $x^2 + 3\sin^2(x) \leq 16(2x + 3\sin(2x))^2$.

□ The error becomes $\frac{128}{1000}$ since f is non-convex

□ The error is $\frac{128}{1000}$ since f satisfies the Polyak-Lojasiewicz Inequality

□ The error is $(1 - \frac{1}{256})^{1000}$ since f satisfies the Polyak-Lojasiewicz Inequality

□ The error is $(1 - \frac{1}{128})^{1000}$ since f is strongly convex

□ The error is $\frac{16}{1000}$ since f is non-convex

Adaptive methods

Question 12 Consider the practical implementation of the three algorithms *Adagrad, Adam and SignSGD*. After computing a fresh stochastic gradient in every iteration, the practical *memory requirement* for the three variants is, for reasonably large machine learning models,

□ SignSGD $\ll$ Adam $\ll$ Adagrad

□ SignSGD $\ll$ Adagrad $\approx$ Adam

□ SignSGD $\ll$ Adagrad $\ll$ Adam

□ similar for all three variants

Non-smooth optimization

Question 13 For the composite objective function $f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x})$, where $g(\mathbf{x})$ is convex and L-smooth, $h(\mathbf{x})$ is convex, define $x^\star$ as a global minimum of $f(\mathbf{x})$, which of the following statements is true in general?

□ $\mathbf{x}^\star = \mathbf{x}^\star - \frac{1}{L}\nabla g(\mathbf{x}^\star)$

□ $\mathbf{x}^\star = \text{prox}_{h,1}(\mathbf{x}^\star - \frac{1}{L}\nabla g(\mathbf{x}^\star))$

□ $\mathbf{x}^\star = \text{prox}_{h,\frac{1}{L}}(\mathbf{x}^\star - \frac{1}{L}\nabla g(\mathbf{x}^\star))$

□ $\mathbf{x}^\star = \text{prox}_{h,\frac{1}{L}}(\mathbf{x}^\star + \frac{1}{L}\nabla g(\mathbf{x}^\star))$

□ $\mathbf{x}^\star = \mathbf{x}^\star - \frac{1}{L}\nabla f(\mathbf{x}^\star)$

Empirical comparison of different methods

Suppose that your roommate wanted to minimize a **linear regression problem with ℓ_2 regularization**. Last night, she overheard you mumbling in your sleep something about "SGD", "gradient descent" and "stepsizes", and was curious to try it out. Can you identify the algorithms she ran by looking at their performance? Note that the scale on the y-axis is **logarithmic**.

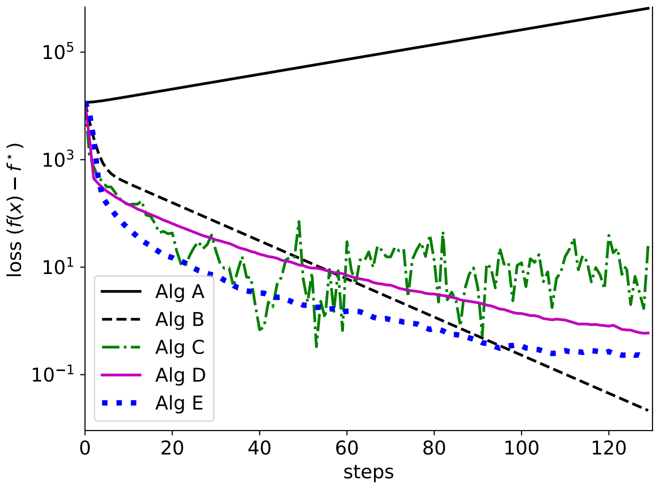


Figure 1: Performance of different optimization algorithms.

Question 14 Which of these algorithms were **gradient descent** (not SGD)?

□ None of them

□ Algorithm A, Algorithm B, and Algorithm E

□ Algorithm A and Algorithm B

□ Only Algorithm B

□ All of them

Question 15 Which optimization method corresponds to the error-curve for **Algorithm C**?

□ SGD with constant stepsize

□ Gradient descent with stepsize $1/L$

□ Gradient descent with incorrect stepsize

□ SGD with stepsize decreasing as $\mathcal{O}(1/\sqrt{t})$

□ SGD with stepsize decreasing as $\mathcal{O}(1/t)$

Question 16 Which optimization method corresponds to the error-curve for **Algorithm D**?

□ SGD with stepsize decreasing as $\mathcal{O}(1/t)$

□ Gradient descent with stepsize $1/L$

□ SGD with stepsize decreasing as $\mathcal{O}(1/\sqrt{t})$

□ Gradient descent with incorrect stepsize

□ SGD with constant stepsize

Question 17 Which optimization method corresponds to the error-curve for **Algorithm E**?

□ Gradient descent with stepsize $1/L$

□ SGD with stepsize decreasing as $\mathcal{O}(1/t)$

□ Gradient descent with incorrect stepsize

□ SGD with stepsize decreasing as $\mathcal{O}(1/\sqrt{t})$

□ SGD with constant stepsize

Second part, true/false questions

Question 18 (Frank-Wolfe convergence in duality gap) On a convex and smooth function, and a bounded and convex constraint set, let $\mathbf{x}_0, \mathbf{x}_1, \ldots$ be the iterates of the Frank-Wolfe algorithm.
The duality gap (or Hearn gap) $g(\mathbf{x}_t) := \langle \mathbf{x}_t - \mathbf{s}, \nabla f(\mathbf{x}_t) \rangle$ of the iterates satisfies $g(\mathbf{x}_t) \leq \mathcal{O}(1/t)$.

□ TRUE □ FALSE

Question 19 (Lower Bounds for Iteration Complexity) Every first-order optimization method needs in the worst case $\Omega(1/\sqrt{\varepsilon})$ steps (gradient evaluations) in order to achieve an additive error of ε on smooth functions.

□ TRUE □ FALSE

Question 20 (GD non-convex) Gradient descent with stepsize $1/L$ converges to an optimum function value on any smooth possibly non-convex function.

□ TRUE □ FALSE

Question 21 (Random search) Consider derivative-free random search as discussed in the lecture. For L-smooth convex functions, using random directions with line-search, converges as $\mathcal{O}(dL/\varepsilon)$.

□ TRUE □ FALSE

Question 22 (Adaptive methods) The three algorithm variants *Adagrad, Adam and SignSGD* have a comparable *computational complexity* per iteration, for deep learning applications

□ 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.

Variance reduction for sum of smooth and strongly convex functions

We are here interested in the unconstrained minimization of the function:

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}),$$

where $f_1, \cdots, f_n$ are L-smooth and convex functions. In addition we assume that the function f is μ-strongly convex. We denote by $\mathbf{x}^\star$ the global minimum of f.

Question 23: *3 points.* Let i be a random variable uniformly distributed in $\{1, \cdots, n\}$. Then prove that

$$\mathbb{E}[\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^\star)\|_2^2] \leq 2L(f(\mathbf{x}) - f(\mathbf{x}^\star)), \tag{S}$$

where the expectation is taken with respect to the randomness of i.
Hint: You can assume that for any L-smooth convex function f the following holds

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \leq 2L\left(f(\mathbf{x}) - f(\mathbf{y}) - \nabla f(\mathbf{y})^\top(\mathbf{x} - \mathbf{y})\right) \quad \text{for all vectors } \mathbf{x}, \mathbf{y}$$

□0 □1 □2 □3

Let $\mathbf{x}_1 \in \mathbb{R}^d$ be an arbitrary initial point and consider the following iterates defined for $t \geq 1$ as:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma(\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)),$$

where i_t is drawn uniformly at random and independently in $\{1, \cdots, n\}$.

Question 24: *1 point.* Let us denote by $\mathbf{v}_t := \nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\mathbf{x}_1) + \nabla f(\mathbf{x}_1)$. Give a closed-form expression for $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|_2^2$ as a function of $\|\mathbf{x}_t - \mathbf{x}^\star\|$, $\|\mathbf{v}_t\|_2^2$, $\mathbf{v}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$ and γ.

□0 □1

Question 25: *3 points.*

Give a lower bound on $\mathbb{E}_{i_t}[\mathbf{v}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)]$ depending on the function values $f(\mathbf{x}_t)$ and $f(\mathbf{x}^\star)$.

□0 □1 □2 □3

Question 26: *4 points.* Prove an upper bound of the form

$$\mathbb{E}_{i_t}[\|\mathbf{v}_t\|_2^2] \leq C_1 L\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) + C_2 L\left(f(\mathbf{x}_1) - f(\mathbf{x}^\star)\right)$$

where C_1 and C_2 are constants.

Hint: You may want to use inequality (S) from Question 23, and that

$$\|\mathbf{a}+\mathbf{b}+\mathbf{c}\|_2^2 \leq 3\left\|\mathbf{a}\right\|^2 + 3\left\|\mathbf{b}\right\|^2 + 3\left\|\mathbf{c}\right\|^2 \quad \text{for all vectors } \mathbf{a}, \mathbf{b}, \mathbf{c}.$$

□0 □1 □2 □3 □4

Question 27: *3 points.* Combine the answers to the previous questions to obtain an upper bound on $E_{i_t}[\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|_2^2$ depending on $\|\mathbf{x}_t - \mathbf{x}^\star\|_2^2$, γ, L and the function values $f(\mathbf{x}_t)$, $f(\mathbf{x}^\star)$ and $f(\mathbf{x}_1)$.

Comment: if you did not solve Question 26, you can instead use the general expression from there.

□0 □1 □2 □3

Question 28: *4 points.* Unroll the recursion proven in previous question for $t = 1, \cdots, T$ to get a upper bound on $E[\|\mathbf{x}_{T+1} - \mathbf{x}^\star\|_2^2$ depending on $\|\mathbf{x}_1 - \mathbf{x}^\star\|_2^2$, γ, L and the function values $(f(\mathbf{x}_t))_{t=1}^T$, $f(\mathbf{x}^\star)$ and $f(\mathbf{x}_1)$.

□0 □1 □2 □3 □4

Question 29: *4 points.* Using the properties of the function f and the previous inequality show that for a certain $c \geq 0$, for which you will give the precise expression, we have

$$\mathbb{E}\Big[f\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\Big)\Big] - f(x^\star) \;\leq\; c \cdot \big(f(x_1) - f(x^\star)\big).$$

In addition show that $c \leq 0.9$ when used with $\gamma = \frac{1}{10L}$ and $T = \frac{20L}{\mu}$.

□0 □1 □2 □3 □4

Averaged SGD for Quadratic Functions

Throughout this exercise we consider minimizing a convex quadratic function:

$$f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top H\mathbf{x} - \mathbf{q}^\top \mathbf{x},$$

where $H \in \mathbf{R}^{d\times d}$ is an invertible, symmetric positive semi-definite matrix and $\mathbf{q} \in \mathbf{R}^d$.

Question 30: *3 points.* Show that the function f is convex and that it admits a global minimum $\mathbf{x}^\star$ on $\mathbf{R}^d$. Give a closed-form expression for this minimum $\mathbf{x}^\star$. Give also a closed-form expression for the excess cost function $f(\mathbf{x}) - f(\mathbf{x}^\star)$ depending only on H, $\mathbf{x}$ and $\mathbf{x}^\star$. Then give the expression of the gradient of f, first in function of H, $\mathbf{x}$ and $\mathbf{q}$ and then in function of H, $\mathbf{x}$ and $\mathbf{x}^\star$.

□0 □1 □2 □3

Now we assume that the true gradient of f is not available and rather that we have access to a noisy oracle for the gradient $\mathbf{g}_t = \nabla f(\mathbf{x}_t) + \boldsymbol{\varepsilon}_{t+1}$. The noise $(\boldsymbol{\varepsilon}_t)$ is assumed to be uncorrelated zero-mean with bounded covariance: $\mathbb{E}[\boldsymbol{\varepsilon}_t] = \mathbf{0}$, $\mathbb{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_{t'}^\top] = \mathbf{0} \in \mathbb{R}^{d\times d}$ for all $t \neq t'$ and $\mathbb{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\top] \preccurlyeq \sigma^2 H$, where $\sigma \geq 0$.

Question 31: *2 points.* Write the stochastic gradient descent iteration with the stochastic gradient $\mathbf{g}_t$ with step-size γ, where you will denote the iterate by $\mathbf{x}_t$. Then writing $\boldsymbol{\alpha}_t := \mathbf{x}_t - \mathbf{x}^\star$, you are asked to state the recursion satisfied by $\boldsymbol{\alpha}_t$. It should only depend on $\boldsymbol{\alpha}_{t-1}$, H, $\boldsymbol{\varepsilon}_t$ and the step-size γ.

□0 □1 □2

Question 32: *4 points.* Compute a closed-form expression for $\boldsymbol{\alpha}_t$ in function of t, γ, H, the initial iterate $\boldsymbol{\alpha}_0$ and the noise vectors $(\boldsymbol{\varepsilon}_k)_{k=1}^t$.

□0 □1 □2 □3 □4

Question 33: *4 points.* Prove that $\bar{\boldsymbol{\alpha}}_t := \frac{1}{t}\sum_{i=0}^{t-1}\boldsymbol{\alpha}_i$ satisfies:

$$\bar{\boldsymbol{\alpha}}_t = \frac{1}{t}(I-(I-\gamma H)^t)(\gamma H)^{-1}\boldsymbol{\alpha}_0 + \frac{\gamma}{t}\sum_{j=1}^{t-1}(I-(I-\gamma H)^{t-j})(\gamma H)^{-1}\boldsymbol{\varepsilon}_j.$$

You will need the identity $\sum_{k=0}^{t-1}(I-\gamma H)^k = (I-(I-\gamma H)^t)(\gamma H)^{-1}$.

□0 □1 □2 □3 □4

Question 34: *4 points.*

Using the properties given on the noise $\boldsymbol{\varepsilon}_t$ and the expressions obtained above compute the value of

$$\mathbb{E}[(\bar{\boldsymbol{\alpha}}_t)^\top H \bar{\boldsymbol{\alpha}}_t].$$

□0 □1 □2 □3 □4

Question 35: *4 points.* Using that $\frac{(1-(1-u)^t)^2}{tu} \leq 1$ for all $u \in [0,1]$, give an upper bound on $\mathbb{E}[(\bar{\boldsymbol{\alpha}}_t)^\top H \bar{\boldsymbol{\alpha}}_t]$, which only depends on γ, $\boldsymbol{\alpha}_0$, σ^2, t and the dimension d.

□0 □1 □2 □3 □4

Question 36: *2 points.* Give two differences between the convergence result you just proved and the classical result known for SGD on strongly convex functions.

□0 □1 □2

Question 37: *4 points (BONUS, optional question).* Prove the inequality $\frac{(1-(1-u)^t)^2}{tu} \leq 1$, for all $u \in [0,1]$. (We have used this in Question 35. Previous questions are not necessary to prove the inequality.)

□0 □1 □2 □3 □4