















Examiner: Sebastian Stich
Optimization for Machine Learning
04.06.2024 from 16h15 to 17h15
Duration : 60 minutes

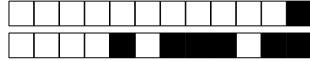
Name : _____

Student ID : _____

Wait for the start of the exam before turning to the next page. This document is printed double sided, 10 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.
- Place out of reach: Please put your **mobile phone in flight mode** (or silent—no vibration) and put it on the desk (but out of reach—e.g. two seats to your left).
- For technical reasons, **do use black or blue pens for the MCQ part, no pencils!** Use white corrector if necessary.
- You find two scratch papers for notes on your desk (you can ask for more). Do not hand in scratch papers, only the answers on the exam sheets count.

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		
     		



Solution:

First part, short questions

Answer in the space provided! Do not cross any checkboxes, they are reserved for correction.

Give a clear and short answer to the question (a “yes”/“no” suffices, 1 point), and a short justification why you selected this answer (1 point).

Recall the definition of a smooth function:

Definition A A differentiable function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Question 1: 2 points. Is every convex function smooth (for a suitably chosen parameter L)?

☐ 0 ☐ 1 ☒ 2

Solution: No. Consider the function $|x|$.

Question 2: 2 points. Suppose f_1, f_2, \dots, f_n are n convex functions on \mathbb{R}^d where $n \geq 2$. Is $f := \min(f_1, \dots, f_n)$ a convex function?

If yes, please provide a short justification. If not, please provide an example.

☐ 0 ☐ 1 ☒ 2

Solution: f may not be convex. Consider $\min\{x, -x\}$.

Question 3: 2 points. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be defined as $f(\mathbf{x}) := ax^2 + bx + c$ where $a > 0$ and $b, c \in \mathbb{R}$. Let us use Newton's method to minimize f :

$$\mathbf{x}_{t+1} = \mathbf{x}_t - (\nabla^2 f(\mathbf{x}_t))^{-1} \nabla f(\mathbf{x}_t).$$

Is it true that we can always find the minimizer of f after one iteration?

☐ 0 ☐ 1 ☒ 2

Solution: Yes. One step of NM is $x_+ = x - (2a)^{-1}(2ax + b) = -b/(2a)$, which is the optimal solution.

Question 4: 2 points. Consider $f(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$ where each f_i is L_i -smooth. Is f a L -smooth function with $L \geq 0$? What can you say about the relation between L and L_{\min} where $L_{\min} := \min_i \{L_i\}$?

☐ 0 ☐ 1 ☒ 2

Solution: Yes. There is no relation between them.



Question 5: 2 points. For an optimization algorithm it has been proven that the output \mathbf{x}_T after T steps of the method satisfies

$$\|\nabla f(\mathbf{x}_T)\| \leq \frac{A}{\sqrt{T}} + \frac{B}{T^2}.$$

for absolute constants $A, B \geq 0$. Can we say that the complexity, i.e. the number of steps to find an output \mathbf{x}_{out} with $\|\nabla f(\mathbf{x}_{\text{out}})\| \leq \varepsilon$, is $\mathcal{O}\left(\frac{A}{\varepsilon^2} + \frac{B}{\sqrt{\varepsilon}}\right)$?

☐ ₀ ☐ ₁ ☒ ₂

Solution: No. It should be $\mathcal{O}\left(\frac{A^2}{\varepsilon^2} + \frac{\sqrt{B}}{\sqrt{\varepsilon}}\right)$.



Second part, open questions

Answer in the space provided! Do not cross any checkboxes, they are reserved for correction.

Your answer must be justified with all steps.

Useful inequalities

– Recall the quadratic upper bound, which holds for smooth functions:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2 \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

– For two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, it holds

$$\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\| \quad \text{and} \quad \|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2.$$

Questions

Question 6: 4 points. Your friend Alice has implemented gradient descent, i.e.

the iteration $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$. She wants to use this method to minimize a function $f: \mathbb{R}^d \rightarrow \mathbb{R}$.

She knows that the function is L -smooth.

Which of the following two stepsize choices would you recommend her to use: either the stepsize $\gamma = \frac{3}{4L}$ or the stepsize $\gamma = \frac{5}{4L}$? . And why?

Hint: Estimate the function value decrease for both stepsize choices!

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☒ 4

Solution: Use the quadratic upper bound to calculate:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \left(\gamma - \frac{L}{2}\gamma^2\right) \|\nabla f(\mathbf{x}_t)\|^2 = f(\mathbf{x}_t) - \frac{15}{32L} \|\nabla f(\mathbf{x}_t)\|^2.$$

1 point for the correct $-\frac{15}{32L} \|\nabla f(\mathbf{x}_t)\|^2$ decrease for stepsize $\frac{3}{4L}$, 1 more point for finding that the for the stepsize $\frac{5}{4L}$ the same decrease occurs (even if the calculation of the exact value of the decrease contains a mistake), 1 point for observing that the progress with $\frac{5}{4L}$ is always at least as good as for $\frac{3}{4L}$ (but not the other way around). 1 point for arguing why this choice is therefore better (for example with an example). Example: consider a function with smoothness $L/2$ (that is also an L -smooth function).

If there were mistakes in the derivations (or no derivations), then correct follow up conclusions (including arguments that mentioned large stepsizes might be problematic/could overshoot) are also given 1 point.

Answers that derive the correct decrease, but conclude ‘selecting either of the stepsizes is fine’ are given 2 points (in total).

Question 7: 2 points. Given a function f that is L -smooth with a minima \mathbf{x}^* , prove the following:

$$\|\nabla f(\mathbf{x})\|^2 \leq 2L[f(\mathbf{x}) - f(\mathbf{x}^*)]$$

Hint: Use the sufficient decrease lemma.

☐ 0 ☐ 1 ☒ 2

Solution: The lemma says: for $\mathbf{x}_+ = \mathbf{x} - \frac{1}{2L} \nabla f(\mathbf{x})$, it holds $f(\mathbf{x}_+) \leq f(\mathbf{x}) - \frac{1}{L} \|\nabla f(\mathbf{x})\|^2$. Therefore $\|\nabla f(\mathbf{x})\|^2 \leq 2L(f(\mathbf{x}) - f(\mathbf{x}_+)) \leq 2L(f(\mathbf{x}) - f(\mathbf{x}^*))$.



Question 8: 5 points. Consider the following implementation of Gradient Descent with momentum.

Algorithm 1 Gradient descent with momentum

Input: $\mathbf{x}_0 \in \mathbb{R}^d$, $T > 0$, $f: \mathbb{R}^d \rightarrow \mathbb{R}$, $\gamma \geq 0$, $\alpha \geq 0$.

Initialization: $\mathbf{m}_0 = \nabla f(\mathbf{x}_0)$

for $t = \{0, \dots, T-1\}$ **do**

$\mathbf{m}_{t+1} = (1 - \alpha)\mathbf{m}_t + \alpha \nabla f(\mathbf{x}_t)$

$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \mathbf{m}_t$

end for

Output: \mathbf{x}_T

In the lecture notes you found the following convergence guarantee for this algorithm when used to minimize a convex function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ (if run with a good choice of parameters γ, α):

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

Here $f^* = f(\mathbf{x}^*)$, for \mathbf{x}^* a minimizer of f .

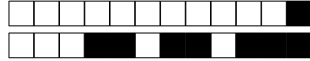
- 1) How can you fix the algorithm so that it outputs a point for which the convergence guarantee applies?
- 2) Analyze the time- and memory complexity of your proposed solution. (Big- \mathcal{O} notation suffices.)
- 3) Is your proposed solution optimal (in order, i.e. ignoring constants)?
If yes, please argue why. If no, can you propose a more efficient solution?

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☒ 5

Solution: 1) By convexity: note that $\bar{\mathbf{x}} := \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{x}_t$ is a valid output for which $f(\bar{\mathbf{x}}) \leq \frac{1}{T} \sum_{t=0}^{T-1} f(\mathbf{x}_t)$. Alternative solutions: keep track of a variable \mathbf{x}_{\min} with the property $f(\mathbf{x}_{\min}) \leq f(\mathbf{x}_t)$, $\forall t$, or output a randomly sampled \mathbf{x}_i (this can be implemented efficiently).

2) One additional variable needs to be stored: $\mathcal{O}(d)$ memory, which can be updated in $\mathcal{O}(d)$ time (plus eventually a function evaluation for one of the possible solutions).

3) Note that to output the average $\bar{\mathbf{x}}$, $\mathcal{O}(nd)$ memory is not required. The average can be tracked by defining $\bar{\mathbf{x}}_t = (1 - \frac{1}{t})\bar{\mathbf{x}}_{t-1} + \frac{1}{t}\mathbf{x}_t$. This is optimal, as the algorithm also needs $\mathcal{O}(d)$ memory and $\mathcal{O}(d)$ time per iteration.



Recall the definition of *the weak growth condition*:

Definition B A function f is L -smooth and has a minima at \mathbf{x}^* . We say the stochastic gradient $\nabla f(\mathbf{x}, \xi)$ satisfies the weak growth condition with constant c if

$$\mathbb{E} \left[\|\nabla f(\mathbf{x}, \xi)\|^2 \right] \leq 2cL[f(\mathbf{x}) - f(\mathbf{x}^*)].$$

We say a function satisfies the *interpolation* condition if the following holds:

Definition C For $f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$, we say that f satisfies the interpolation condition if there exists a \mathbf{x}^* that minimizes all f_i (i.e. $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}} f_i(\mathbf{x})$ for $i = \{1, \dots, n\}$).

Question 9: 2 + 2 points. Consider $f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$ where each f_i is L -smooth. A stochastic gradient is given by $\nabla f(\mathbf{x}, \xi) = \nabla f_\xi(\mathbf{x})$ where $\xi \in [n]$ denotes an index sampled uniformly at random.

1) (2 points.) Show that the interpolation condition implies the weak growth condition.

2) (Bonus, 2 points.) Suppose the interpolation condition does not hold. Does weak growth still hold? If yes, prove it. If not, please propose (and prove) an alternative growth condition that holds (without introducing new assumptions).



Solution: As every f_ξ is smooth, we have by the sufficient decrease lemma (or Question 7): $\|\nabla f_i(\mathbf{x})\|^2 \leq 2L(f_i(\mathbf{x}) - f_i(\mathbf{x}^*))$. (Here we use the property that \mathbf{x}^* is a minimizer of f_i .) Therefore,

$$\mathbb{E} \left[\|\nabla f_i(\mathbf{x})\|^2 \right] \leq \frac{1}{n} \sum_{i=1}^n 2L[f_i(\mathbf{x}) - f_i(\mathbf{x}^*)] = 2L[f(\mathbf{x}) - f(\mathbf{x}^*)].$$