Examiner: Sebastian Stich
Optimization for Machine Learning
04.10.2022 from 13h15 to 15h45
Duration : 150 minutes

Name :_________________________________

Student ID : ____

Wait for the start of the exam before turning to the next page. This document is printed double sided, 18 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

▨ ☒ ○ ⊡ ▨ □

First part, multiple choice

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

Convexity

Question 1 Let $f\colon \mathbb{R} \to \mathbb{R}$ and $g\colon \mathbb{R} \to \mathbb{R}$ be two convex functions. Consider the following combinations of f and g:

A	$f(x) + g(x)$	B	$f(x) \cdot g(x)$	C	$\max\{f(x), g(x)\}$
D	$\min\{f(x), g(x)\}$	E	$f(g(x))$	F	$e^{f(x)}$

Which of the following statements is **true**?

□ E is non-convex.
□ A, B and C are convex.
■ A, C and F are convex.
□ None of the other four choices.
□ A, D and F are convex.

Question 2 Consider the function $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^4$, defined on the interval $I = [-1, 1]$. Which of the following statements is **false**?

■ The function f is strongly convex in the interval I.
□ The function f is smooth in the interval I.
□ The function f is convex in the interval I.
□ The function f has a unique minimizer in the interval I.
□ The function f is star convex in the interval I.

Question 3 Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a function that can be written as $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$, where $n \geq 2$ is an integer and each $f_i\colon \mathbb{R}^d \to \mathbb{R}$, $i = 1, \ldots, n$, is a convex and differentiable function. Let $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ be a minimizer of f. Which statement is **true**?

■ Let $\mathbf{x} \in \mathbb{R}^d$ be an arbitrary point. Then $\mathbf{x}^\star \in \{\mathbf{w} \mid \mathbf{w} \in \mathbb{R}^d \text{ and } \nabla f(\mathbf{x})^\top \mathbf{w} \leq \nabla f(\mathbf{x})^\top \mathbf{x}\}$.
□ It holds $\|\nabla f_i(\mathbf{x}^\star)\| = 0$ for all $i = 1, \ldots, n$.
□ It holds $\|\nabla f_i(\mathbf{x}^\star)\| = 0$ for at least one $i \in [n]$.
□ Let $\mathbf{y}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^{n-1} f_i(\mathbf{x})$ be a minimizer of the first $n-1$ components and let $\mathbf{z}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f_n(\mathbf{x})$ be a minimizer of the last component. Then $\mathbf{x}^\star \in \{\lambda\mathbf{y}^\star + (1-\lambda)\mathbf{z}^\star, \lambda \in \mathbb{R}\}$.
□ None of the other four choices.

Question 4 Let $\mathbf{A} \in \mathbb{R}^{d\times n}$ denote a *data matrix* with columns $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$, with $\mathbf{a}_i \in \mathbb{R}^d$. Let $g\colon \mathbb{R}^n \to \mathbb{R}$ denote an arbitrary function and consider the regularized minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left[f(\mathbf{x}) := \left\{ g(\mathbf{A}^\top \mathbf{x}) + \lambda \|\mathbf{x}\|_2^2 \right\} \right] ,$$

where $\lambda \geq 0$ denotes a (fixed) regularization parameter. Which statement is **true**?

□ If $\lambda > 0$, then f is strongly convex.
■ None of the other four choices.
□ If g is strongly convex, then f is strongly convex.
□ If g is convex, then f is strongly convex.
□ If the columns of $\mathbf{A}$ span $\mathbb{R}^d$ ($\operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\} = \mathbb{R}^d$) and g is convex, then f is strongly convex.

Gradient Descent

Question 5 The following Figure 1 depicts the evolution of the function value of iterates generated by gradient descent (i.e. the iteration $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$, for a starting point $\mathbf{x}_0 \in \mathbb{R}^d$ and a fixed stepsize $\gamma \in \mathbb{R}$) on a convex function $f\colon \mathbb{R}^d \to \mathbb{R}$. Which statement is **true**?

□ The function f is μ-strongly convex for $\mu > 0$.

■ None of the other four choices.

□ The convergence of gradient descent on f is linear.

□ The function f is μ-PL for $\mu > 0$.

□ The convergence of gradient descent on f is quadratic.

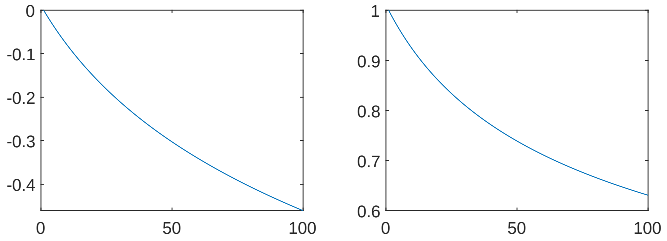

Figure 1: Gradient descent on a convex function f.
Left y-axis: $\log(f(\mathbf{x}_t))$, Right y-axis: $f(\mathbf{x}_t)$. Both x-axis: #iterations t.

Stochastic Gradient Descent

Question 6 Consider $f\colon \mathbb{R} \to \mathbb{R}$, of the form $f(x) = \frac{1}{2}(f_1(x) + f_2(x))$, with $f_1\colon \mathbb{R} \to \mathbb{R}$ defined as $f_1(x) = \frac{1}{2}x^2$ and $f_2\colon \mathbb{R} \to \mathbb{R}$ defined as $f_2(x) = \frac{1}{4}x^4$. We consider stochastic gradient descent, defined as $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla f_{i_t}(\mathbf{x}_t)$ for a starting point $\mathbf{x}_0 \in \mathbb{R}$, a fixed stepsize $\gamma \in \mathbb{R}$ and an index $i_t \in \{1, 2\}$, chosen uniformly at random in each iteration t. Which of the two curves depicted in Figure 2 could correspond to a run of SGD (for an appropriate choice of γ and $\mathbf{x}_0$)?

□ None of the two curves.

■ Curve B but not curve A.

□ Curve A and curve B.

□ Curve A but not curve B.

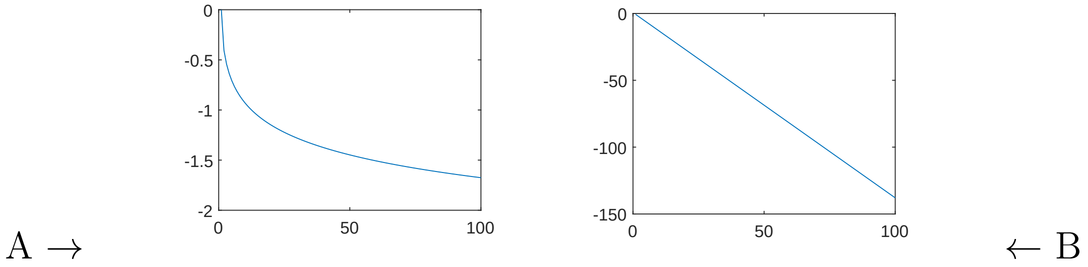

Figure 2: Stochastic gradient descent on the function f specified in Question 6. y-axis: $\log(f(\mathbf{x}_t))$, x-axis: #iterations t.

Question 7 Consider a smooth function $f\colon \mathbb{R} \to \mathbb{R}$. Let $\mathbf{g}_1\colon \mathbb{R} \to \mathbb{R}$ denote an unbiased gradient oracle for f, that is, $\mathbb{E}[\mathbf{g}_1(\mathbf{x})] = \nabla f(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}$. For a parameter $\alpha \in [0,1]$, define $\mathbf{g}_{2,\alpha}\colon (\mathbb{R} \times \mathbb{R}) \to \mathbb{R}$ as $\mathbf{g}_{2,\alpha}(\mathbf{x},\mathbf{y}) = \alpha \mathbf{g}_1(\mathbf{x}) + (1-\alpha)\nabla f(\mathbf{y})$. Which of the following statements is **false**?

□ If $\mathbf{y} = \mathbf{x}$, the estimator $\mathbf{g}_{2,\alpha}$ is unbiased, i.e. $\mathbb{E}[\mathbf{g}_{2,\alpha}(\mathbf{x},\mathbf{y})] = \nabla f(\mathbf{x})$.

□ If $\alpha = 1$, the estimator $\mathbf{g}_{2,\alpha}$ is unbiased, i.e. $\mathbb{E}[\mathbf{g}_{2,\alpha}(\mathbf{x},\mathbf{y})] = \nabla f(\mathbf{x})$.

■ The bias $\|\mathbb{E}[\mathbf{g}_{2,\alpha}(\mathbf{x},\mathbf{y})] - \nabla f(\mathbf{x})\|$ increases as α increases from 0 to 1.

□ The variance $\mathbb{E}\left\|\mathbf{g}_{2,\alpha}(\mathbf{x},\mathbf{y}) - \mathbb{E}[\mathbf{g}_{2,\alpha}(\mathbf{x},\mathbf{y})]\right\|^2$ increases as α increases from 0 to 1.

Complexity Estimates

Question 8 Your friend Bob has developed a new iterative algorithm to minimize the gradient norm of an arbitrary differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$. He has proven that after performing T steps of his algorithm the following guarantee holds:

$$\|\nabla f(\mathbf{x}_T)\|^2 \le \frac{A}{T} + \frac{B}{T^2} + \frac{C}{T^3}\,,$$

where $A, B, C \ge 0$ are parameters (depending on the objective function) and $\mathbf{x}_T \in \mathbb{R}^d$ the output of the algorithm after T iterations. Can you help him to derive the correct complexity estimate, i.e. after which number T of iterations does it hold $\|\nabla f(\mathbf{x}_T)\|^2 \le \varepsilon$, for any arbitrary $\varepsilon > 0$?

□ For $T = \mathcal{O}\left(\frac{A}{\varepsilon} + \frac{B}{\sqrt{\varepsilon}} + \frac{C}{\varepsilon^{1/3}}\right)$.

□ For $T = \mathcal{O}\left(\frac{A}{\varepsilon} + \frac{B}{\varepsilon^2} + \frac{C}{\varepsilon^3}\right)$.

■ For $T = \mathcal{O}\left(\frac{A}{\varepsilon} + \frac{\sqrt{B}}{\sqrt{\varepsilon}} + \frac{C^{1/3}}{\varepsilon^{1/3}}\right)$.

□ For $T = \mathcal{O}\left(\frac{A}{\varepsilon} + \frac{\sqrt{BC}}{\sqrt{\varepsilon}} + \frac{B^{1/3}C^{2/3}}{\varepsilon^{1/3}}\right)$.

□ For none of the other four choices.

Finite-Sum Optimization Problems

For the next two questions, we consider a μ-strongly convex, L-smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$. In the lecture we have proven that stochastic gradient descent, defined as the iteration $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \mathbf{g}(\mathbf{x}_t)$ for a stochastic gradient oracle $\mathbf{g}\colon \mathbb{R}^d \to \mathbb{R}^d$, converges after $\mathcal{O}\left(\frac{\sigma^2}{\mu\varepsilon} + \kappa \log \frac{1}{\varepsilon}\right)$ evaluations of the gradient oracle to an ε-accurate solution, where $\kappa = \frac{L}{\mu}$ denotes the condition number and σ^2 is an upper bound on the stochastic variance $\mathbb{E}\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \le \sigma^2$, $\forall \mathbf{x} \in \mathbb{R}^d$. In the next two questions, we investigate **the oracle complexity** of mini-batch SGD **(note that evaluating a mini-batch stochastic gradient of size $b > 1$ requires b queries to the stochastic gradient oracle and is thus b-times more expensive than a single stochastic gradient)**.

Question 9 Consider mini-batch SGD with batch size $b > 1$ for the above problem, where a mini-batch stochastic gradient is defined as $\mathbf{g}_{MB}(\mathbf{x}) = \frac{1}{b}\sum_{i=1}^{b} \mathbf{g}(\mathbf{x})$, for b independent queries to oracle $\mathbf{g}(\mathbf{x})$. Which of the following statements is **true**?

□ None of the other four choices.

■ Mini-batch SGD with batch size $b > 1$ has always worse oracle complexity than mini-batch SGD with batch size $b = 1$.

□ Mini-batch SGD with batch size $b = 1$ has always worse oracle complexity than mini-batch SGD with batch size $b > 1$.

□ The optimal batch size depends on the target accuracy $\varepsilon > 0$.

□ The optimal batch size is $b = \sigma^2$.

Question 10 Suppose the function $f(\mathbf{x})$ defined above has the following structure: $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$ for $n > 1$ components $f_i \colon \mathbb{R}^d \to \mathbb{R}$. SGD with batch size $b = 1$ can be defined by assuming the the gradient oracle is equal to $\mathbf{g}(\mathbf{x}) = \nabla f_i(\mathbf{x})$ for an index $i \in [n]$ picked uniformly at random. Consider mini-batch SGD with batch size $1 < b \leq n$ for the above problem, where a mini-batch stochastic gradient is defined as $\mathbf{g}_{MB}(\mathbf{x}) = \frac{1}{b}\sum_{i\in\mathcal{S}} \nabla f_i(\mathbf{x})$, where $\mathcal{S}$ denotes a set of b indices sampled uniformly at random without replacement from $[n]$. Which of the following statements is **true**?

□ The optimal batch size is $b = n$.

■ The optimal batch size depends on the target accuracy $\varepsilon > 0$.

□ Mini-batch SGD with batch size $b = 1$ has always worse oracle complexity than mini-batch SGD with batch size $b > 1$.

□ None of the other four choices.

□ Mini-batch SGD with batch size $b > 1$ has always worse oracle complexity than mini-batch SGD with batch size $b = 1$.

Question 11 Researchers at Saarland University developed a new algorithm to optimize neural networks. The algorithm works very well in practice and they also provided a theoretical convergence analysis. For a certain class of functions $f \colon \mathbb{R}^d \to \mathbb{R}$, they prove the following guarantee: For every starting point $\mathbf{x}_0 \in \mathbb{R}^d$, and an arbitrary $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, it holds for the output $\mathbf{x}_T \in \mathbb{R}^d$ generated by their algorithm after any $T \geq 1$ steps:

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) + A\left\|\nabla f(\mathbf{x}_T)\right\|^2 \leq \frac{B}{T^3}\left\|\nabla f(\mathbf{x}_0)\right\|^2 ,$$

where A, B are (problem dependent) constants. Your friend Alice—who did not attend the optimization course—is asking you for which class of functions this theorem might hold. Obviously, the researchers did assume that the function is differentiable (as the gradient appears in their statement), but which additional properties must the function have?
Can you tell her which of the following choices is a **possible option** for which such a statement could hold?

□ For every smooth and convex function.

□ For every smooth (possibly non-convex) function.

■ For every smooth and strongly convex function.

□ None of the other four classes.

□ For every smooth function that satisfies $B\left\|\nabla f(\mathbf{x})\right\|^2 \leq f(\mathbf{x}) - f(\mathbf{x}^\star) + A\left\|\nabla f(\mathbf{x})\right\|^2$, $\forall \mathbf{x} \in \mathbb{R}^d$.

Question 12 Let $\mathcal{Q} \colon \mathbb{R}^d \to \mathbb{R}^d$ denote an unbiased ω-quantizer, that is, a function with the properties $\mathbb{E}_{\mathcal{Q}}[\mathcal{Q}(\mathbf{x})] = \mathbf{x}$, $\forall \mathbf{x} \in \mathbb{R}^d$ and $\mathbb{E}_{\mathcal{Q}}\left\|\mathcal{Q}(\mathbf{x}) - \mathbf{x}\right\|^2 \leq \omega\left\|\mathbf{x}\right\|^2$, $\forall \mathbf{x} \in \mathbb{R}^d$. Which of the following statements is **true**?

□ Every ω-quantizer is also a δ-compressor, i.e. statisfying $\mathbb{E}_{\mathcal{Q}}\left\|\mathcal{Q}(\mathbf{x}) - \mathbf{x}\right\|^2 \leq (1-\delta)\left\|\mathbf{x}\right\|^2$, $\forall \mathbf{x} \in \mathbb{R}^d$, for an appropriate $\delta \in (0,1]$.

□ It must hold $\omega \leq d$.

□ Consider a randomized operator $\mathcal{R} \colon \mathbb{R}^d \to \mathbb{R}^d$, that is defined as $\mathcal{R}(\mathbf{x}) = \mathbf{x}$ with probability $\frac{1}{2}$ and $\mathcal{R}(\mathbf{x}) = 0$ with probability $\frac{1}{2}$. Then $\mathcal{R}$ is a $\omega = \frac{1}{2}$ quantizer.

■ None of the other four choices.

□ Assume that each coordinate entry of $\mathbf{x} \in \mathbb{R}^d$ can be encoded with B bits. Then $\mathcal{Q}(\mathbf{x})$ can be encoded with at most $\lceil \omega B \rceil$ bits.

Second part, true/false questions

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

Question 13 (Lipschitz) Let L_i denote the Lipschitz constant of a function $f_i\colon \mathbb{R}^d \to \mathbb{R}$ for $i \in [n], n \geq 1$. Then the function $f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$ is $\left(\frac{1}{n}\sum_{i=1}^n L_i\right)$-smooth.

■ TRUE □ FALSE

Question 14 (Strong convexity) Let the function $f_i\colon \mathbb{R}^d \to \mathbb{R}$ be μ_i strongly convex, for $i \in [n], n \geq 1$. Then the function $f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$ is $\left(\frac{1}{n}\sum_{i=1}^n \mu_i\right)$-strongly convex.

□ TRUE ■ FALSE

Question 15 (Quadratic functions) Let $\mathbf{Q} \in \mathbb{R}^{d\times d}$ be positive semidefinite, $\mathbf{b} \in \mathbb{R}^d$ and $c \in \mathbb{R}$. The quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$ is L-smooth for parameter $L = \max_{i\in[d]}(\mathbf{Q})_{ii}$.

□ TRUE ■ FALSE

Question 16 (Coordinate Descent) Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be convex and coordinate-wise L_i smooth, for $i \in [d]$. Then coordinate descent with sampling of the coordinates proportional to L_i (and correct step size) converges always faster than coordinate descent with uniform sampling.

■ TRUE □ FALSE

Solution: The special case when both variants are identical caused some ambiguity. Both answers were awarded a point.

Question 17 Consider a (simple) linear neural network with $\ell \geq 1$ layers, that can be written as $f(\mathbf{x}) = \frac{1}{2}\left(\prod_{k=1}^{\ell}(\mathbf{x})_k - 1\right)^2$ for parameter $\mathbf{x} \in \mathbb{R}^\ell$ and $(\mathbf{x})_k$ denoting the k-th coordinate of $\mathbf{x}$. For all critical points ($\nabla f(\mathbf{x}) = \mathbf{0}$) it must hold $\prod_{k=1}^{\ell}(\mathbf{x})_k = 1$.

□ TRUE ■ FALSE

Question 18 (Mixing Matrix) In Federated Learning, a set of n clients average their model parameters $\mathbf{x}^i \in \mathbb{R}^d$, $i \in [n]$, by sending the parameters to a server that computes $\frac{1}{n}\sum_{i=1}^n \mathbf{x}^i$ and sends the result back to all clients. In the matrix notation that we have seen in the lecture, letting $\mathbf{X} \in \mathbb{R}^{d\times n}$ with the columns of $\mathbf{X}$ equal to the parameters $\mathbf{x}^i$, this average computation can be equivalently be written as $\mathbf{X}\mathbf{W}$, for a doubly stochastic mixing matrix $\mathbf{W} \in \mathbb{R}^{n\times n}$ of the following form:

$$\mathbf{W} = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{n-1}{n} & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ \frac{1}{n} & 0 & 0 & \frac{n-1}{n} \end{bmatrix}$$

□ TRUE ■ FALSE

Question 19 (Adaptive Methods) The computational complexity of the adaptive methods ADAM and AdaGrad scales superlinearly in the problem dimension and they are too expensive to run on very large scale deep learning problems (opposed to e.g. stochastic gradient descent that scales linearly in the problem dimension).

□ TRUE ■ FALSE

Solution:

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.

Quadratic Upper Bounds

For the exercises in this subsection, we introduce a new class of functions: $\mathcal{C}_{\mathbf{L}}$ **functions.** A convex function $f\colon \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}_{\mathbf{L}}$ (notation: $f \in \mathcal{C}_{\mathbf{L}}$ or f is $\mathbf{L}$-smooth) if f is differentiable and

$$f(\mathbf{y}) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2} \|\mathbf{y} - \mathbf{x}\|_{\mathbf{L}}^2 \,, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \,,$$

where the norm $\|\mathbf{x}\|_{\mathbf{L}}$ is defined as $\|\mathbf{x}\|_{\mathbf{L}}^2 := \mathbf{x}^\top \mathbf{L} \mathbf{x}$ and $\mathbf{L} \in \mathbb{R}^{d \times d}$ is a positive definite matrix.

Question 20: *4 points.*
Part 1, *2 points*:
Let $g\colon \mathbb{R}^d \to \mathbb{R}$ be a (standard) L-smooth function. Is $g \in \mathcal{C}_{\mathbf{L}}$ for a suitable matrix $\mathbf{L}$?

If yes, give the matrix $\mathbf{L}$ and a proof. If no, prove that no suitable matrix $\mathbf{L}$ exists.

Part 2, *2 points*:
Let $h \in \mathcal{C}_{\mathbf{L}}$ be a $\mathbf{L}$-smooth function. Is h (standard) L-smooth for a suitable parameter L?

If yes, give the parameter L and a proof. If no, prove that no suitable parameter L exists.

□0 □1 □2 □3 ■4

Solution: (1 point for correct answer in each case, 1 point for correct L and $\mathbf{L}$.) Note that $\mathbf{L} = L \cdot \mathbf{I}_d$ and $L = \|\mathbf{L}\|$ are suitable choices.

Question 21: *3 points.* Let $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{x}_+ := \mathbf{x} - \gamma \nabla f(\mathbf{x})$ for a parameter $\gamma > 0$. Let $f \in \mathcal{C}_{\mathbf{L}}$. For which values of γ does it always hold $f(\mathbf{x}_+) \le f(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^d$? Specify the set $\Gamma \subset \{\gamma \mid \gamma \ge 0\}$ of all possible parameters γ with this property.

□0 □1 □2 ■3

Solution: The proof is the same as for the standard sufficient decrease lemma. Plug-in the definition of $\mathbf{x}_+$ (1 point), from the inequality $-\gamma \|\nabla f(\mathbf{x})\|^2 + \frac{\gamma^2}{2} \|\nabla f(\mathbf{x})\|_{\mathbf{L}}^2 \le 0$ conclude that $\gamma \le 2 \|\mathbf{L}\|^{-1}$ is sufficient (and necessary) (2 points).

Question 22: *4 points.* A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is $\mathbf{M}$-strongly convex for a positive definite matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ if

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2} \|\mathbf{y} - \mathbf{x}\|_{\mathbf{M}}^2 \,, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \,.$$

Prove that a $\mathbf{M}$-strongly convex function has an unique minimizer $\mathbf{x}^\star \in \operatorname{argmin} f(\mathbf{x})$ and that it holds

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \le \frac{1}{2} \|\nabla f(\mathbf{x})\|_{\mathbf{M}^{-1}}^2 \,, \qquad \forall \mathbf{x} \in \mathbb{R}^d \,.$$

□0 □1 □2 □3 ■4

Solution: The proof is exactly the same as for μ-strongly convex functions. (1 point for noting that the inequality can be derived by minimizing the given expression in $\mathbf{y}$, 1 point for correct calculations, 2 points for concluding that the inequality (+ convexity) implies that $\mathbf{x}^\star$ must be unique.)

Question 23: *2 points.* Let $\mathbf{x}_t \in \mathbb{R}^d$, $t = 0, \dots, T$ denote the iterates generated by an iterative algorithm on a $\mathbf{M}$-strongly convex function $f\colon \mathbb{R}^d \to \mathbb{R}$. Suppose the iterates satisfy the following relation:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \le f(\mathbf{x}_t) - f(\mathbf{x}^\star) - \gamma \|\nabla f(\mathbf{x}_t)\|^2_{\mathbf{G}} \,,$$

where $\gamma \ge 0$ is a parameter and $\mathbf{G} \in \mathbb{R}^{d\times d}$ a positive definite matrix. By using the properties of $\mathbf{M}$-strongly convex functions stated in Question 22, prove that it holds

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \le (1-\alpha)\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \,,$$

for a parameter α. For which α?

Hint: You can use the fact that $\|\mathbf{x}\|^2_{\mathbf{M}^{-1}} \cdot \frac{1}{\|\mathbf{G}^{-1}\mathbf{M}^{-1}\|} \le \|\mathbf{x}\|^2_{\mathbf{G}} \le \|\mathbf{x}\|^2_{\mathbf{M}^{-1}} \cdot \|\mathbf{G}\mathbf{M}\|$, $\forall \mathbf{x} \in \mathbb{R}^d$.

□0 □1 ■2

Solution: Use the inequality $\|\mathbf{x}\|^2_{\mathbf{G}} \ge \|\mathbf{x}\|^2_{\mathbf{M}^{-1}} \cdot \frac{1}{\|\mathbf{G}^{-1}\mathbf{M}^{-1}\|} \ge \frac{2}{\|\mathbf{G}^{-1}\mathbf{M}^{-1}\|}\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right)$ (1 point), to conclude that $\alpha = \frac{2\gamma}{\|\mathbf{G}^{-1}\mathbf{M}^{-1}\|}$ (1 point).

Question 24: *5 points.* Let $\mathbf{x}_t \in \mathbb{R}^d$, $t = 0, \dots, T$ denote the iterates generated by an iterative algorithm on a function $f \in \mathcal{C}_{\mathbf{L}}$. Suppose that the iterates satisfy the following relation:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star) \le (1-\gamma)\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \,,$$

where $0 \le \gamma \le A$ is a positive parameter that has to be smaller than a (problem dependent constant) $A < 1$.

State the iteration complexity of the algorithm in big-$\mathcal{O}$ notation. Your expression can depend on the problem parameters γ, A, $F_0 := f(\mathbf{x}_0) - f(\mathbf{x}^\star)$, $f(\mathbf{x}_0)$, $f(\mathbf{x}^\star)$, $R_0^2 := \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$, and the target accuracy $\varepsilon \ge 0$. Which value of γ is optimal?

□0 □1 □2 □3 ■4

Solution: The best progress per iteration is obtained for $\gamma = A$ (1 point). Reformulate $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le (1-\gamma)^T F_0$ (1 point), notice that one has to solve $(1-\gamma)^T F_0 \le \varepsilon$ (1 point), and concluding $T \ge \frac{1}{\gamma}\log\frac{F_0}{\varepsilon}$ (1 point + 1 point for correct derivation).

Hessian Similarity

For the next two questions, let $f\colon \mathbb{R}^d \to \mathbb{R}$ be of the form $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$, $n \geq 1$, where each $f_i\colon \mathbb{R}^d \to \mathbb{R}$ is a L-smooth and twice differentiable function.

Question 25: *1 point.* δ-Hessian similarity was defined in the lecture as:

$$\left\|\nabla^2 f_i(\mathbf{x}) - \nabla^2 f(\mathbf{x})\right\| \leq \delta \qquad \forall i \in [n], \forall \mathbf{x} \in \mathbb{R}^d \ .$$

Prove that that the function f specified above satisfies the Hessian similarity property.

□ 0 ■ 1

Solution: See Exercise 11.2

Question 26: *2 points.* Suppose that the function f specified above satiesfies δ-Hessian similarity for a parameter δ. Prove that

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{y}) - \nabla f_i(\mathbf{x}) + \nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq \delta^2 \|\mathbf{y} - \mathbf{x}\|^2 , \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \ .$$

□ 0 □ 1 ■ 2

Solution: See Exercise 11.3

Token (S)GD

In this section we consider (again) a function $f\colon \mathbb{R}^d \to \mathbb{R}$ of the form $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$, $n \geq 1$, where each $f_i\colon \mathbb{R}^d \to \mathbb{R}$ is a L-smooth and twice differentiable function. In addition we assume that each f_i is μ-strongly convex, for a parameter $0 < \mu \leq L$.

Suppose that the components f_i are distributed across n client devices. In the lecture we have discussed gradient descent that computes in each iteration a gradient $\nabla f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x})$ of the function f that requires communication between all clients. We have also discussed the communication efficient local (S)GD algorithm that performs in parallel on all n functions a certain number of gradient descent steps before averaging the parameters. Both algorithms require that all clients are active during the whole optimization process.

We now consider **Token GD**. This algorithm takes two parameters: $\tau \geq 1$ (the number or local steps) and $\gamma \geq 0$ (a stepsize). For a starting point $\mathbf{x}_0 \in \mathbb{R}^d$, the algorithm is defined as follows:

- in each iteration t with ($t \mod \tau \equiv 0$) (that is, in iteration $\{0, \tau, 2\tau, 3\tau, \dots\}$), an *active worker* $i_t \in [n]$ is selected uniformly at random among the n workers.
- for all other iterations, the currently active worker remains active, $i_{t+1} = i_t$.
- at the end of each iteration, a gradient step is performed on the active worker:

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

This means, we can imagine this algorithm as passing a *token* (the current state $\mathbf{x}_t \in \mathbb{R}^d$) between workers, each worker performing τ gradient steps when in possession of the token.

Question 27: *2 points.* Let $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$ denote a minimizer of f and consider token GD with an (integer) $\tau \geq 1$ number of local steps, and stepsize $\gamma \leq \frac{1}{L}$—this stepsize guarantees that GD converges locally on each function f_i.

For which values of $\tau \geq 1$ is $\mathbf{x}^\star$ a fixed point of token GD?

□ 0 □ 1 ■ 2

Solution: None. 2 points for a concrete example, 1 point for vague argumentation but correct answer.

Question 28: *2 points.* Recall that when performing a gradient step with stepsize $\gamma \leq \frac{1}{L}$ on a client i, $\mathbf{x}_+ = \mathbf{x} - \gamma \nabla f_i(\mathbf{x})$, the following decrease condition holds by smoothness:

$$f_i(\mathbf{x}_+) \leq f_i(\mathbf{x}) - \frac{\gamma}{2}\left\|\nabla f_i(\mathbf{x})\right\|^2 .$$

For simplicity, suppose $\tau = 1$ and assume a starting value $\mathbf{x}_0 \in \mathbb{R}^d$ is given.

Derive a recursion for the expected function value $\mathbb{E}\, f(\mathbf{x}_{t+1})$ for the iterate $\mathbf{x}_{t+1}$ of token SGD. The expression should only depend on the previous iterates $\{\mathbf{x}_t, \mathbf{x}_{t-1}, \dots, \mathbf{x}_0\}$.

□ 0 □ 1 ■ 2

Solution: Simply take the average over i.

Question 29: *2 points.* Prove that

$$\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\mathbf{x})\right\|^2 \leq 4L\left(f(\mathbf{x}) - f(\mathbf{x}^\star)\right) + \frac{2}{n}\sum_{i=1}^n \left\|\nabla f_i(\mathbf{x}^\star)\right\|^2 ,$$

where f is as above, and again $\mathbf{x}^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.

□0 □1 ■2

Solution: See Q21 in the exam from August. Add and subtract $\nabla f_i(\mathbf{x}^\star)$, $\|\nabla f_i(\mathbf{x})\|^2 = \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^\star) + \nabla f_i(\mathbf{x}^\star)\|^2$, with the inequality $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$ (1 point) and use smoothness (1 point).

Question 30: *2 points.* Suppose we are in a so-called *overparametrized* setting, where it holds that $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}^\star)\|^2 = 0$ (f and $\mathbf{x}^\star$ as defined above).

Suppose $\tau = 1$. Based on your answers to the previous questions, what can you say about the convergence (or non-convergence) of token GD with stepsize in the overparametrized setting? (In case of convergence you do not need to formally derive the complexity estimate, just point out the reasons for your conclusion. In case of non-convergence no full proof is needed, just give the main argument.)

□0 □1 ■2

Solution: By using Q28 and Q29, it should be obvious that by choosing γ sufficiently small (1 point, $\gamma = \frac{1}{L}$ is not small enough), linear convergence (1 point) can be proven (see Q24 for the details which were not required here).