Examiner: Sebastian Stich
Optimization for Machine Learning
08.08.2023 from 14h15 to 16h45
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.

First part, multiple choice

There is **exactly one** correct answer per question. *2 points* for each correct answer.

Gradient Descent

For a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$, a starting point $\mathbf{x}_0 \in \mathbb{R}^d$, and a stepsize $\gamma > 0$, the *gradient descent* algorithm generates a sequence $(\mathbf{x}_0, \mathbf{x}_1, \dots)$ of iterates, satisfying:

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

where $\nabla f(\mathbf{x})\colon \mathbb{R}^d \to \mathbb{R}^d$ denotes the *gradient* of the function f.

Question 1 For a vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$, consider the function $f(\mathbf{x}) = x_1^2 + 4x_1x_2 + 4x_2^2$. Which of the following statements is **true**?

- ☐ $\nabla f(\mathbf{x}) = \begin{bmatrix} 2x_1 + 4x_1x_2 \\ 4x_1x_2 + 8x_2 \end{bmatrix}$
- ☐ $\nabla f(\mathbf{x}) = 6x_1 + 12x_2$
- ☐ $\nabla f(\mathbf{x}) = \begin{bmatrix} x_1 + 4x_2 \\ 4x_1 + 4x_2 \end{bmatrix}$
- ☐ None of the other four choices.
- ☐ $\nabla f(\mathbf{x}) = \begin{bmatrix} 2x_1 + 4x_2 \\ 4x_1 + 8x_2 \end{bmatrix}$

Question 2 Consider the function $f\colon \mathbb{R} \to \mathbb{R}$, defined as $f(x) = x^2$. When running gradient descent from $x_0 \in \mathbb{R}$, with a stepsize $\gamma = \frac{1}{8}$, it holds $x_1 = x_0 - \gamma \nabla f(x) = \frac{3}{4}x_0$, and generally:

$$x_t = \left(\frac{3}{4}\right)^t x_0\,.$$

For a parameter $\varepsilon > 0$, we define the *iteration complexity* $\mathcal{T}_\varepsilon$ to be number of iterations it takes to be sure that it holds $|x_t| \leq \varepsilon$, for all $t \geq \mathcal{T}_\varepsilon$.
Which of the following statements is **true**?

- ☐ $\mathcal{T}_\varepsilon = \mathcal{O}\left(\frac{4}{3}\log\left(\frac{\varepsilon}{|x_0|}\right)\right)$
- ☐ $\mathcal{T}_\varepsilon = \frac{\left(\frac{3}{4}\right)^t}{\varepsilon}$
- ☐ None of the other four choices.
- ☐ $\mathcal{T}_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$
- ☐ $\mathcal{T}_\varepsilon = \mathcal{O}\left(4\log\left(\frac{1}{\varepsilon}\right)\right)$
- ☐ $\mathcal{T}_\varepsilon = \mathcal{O}\left(\log\left(\frac{|x_0|}{\varepsilon}\right)\right)$

Question 3 Let $\mathcal{A}$ and $\mathcal{B}$ denote two function classes (i.e. sets of functions). Let $\mathcal{T}_\mathcal{A}$ and $\mathcal{T}_\mathcal{B}$ denote the iteration complexity of gradient descent for finding an ε-accurate solution of problems in class $\mathcal{A}$ and $\mathcal{B}$, respectively. Suppose it holds

$$\mathcal{T}_\mathcal{A} = \mathcal{O}\left(\frac{1}{\varepsilon}\right), \qquad \mathcal{T}_\mathcal{B} = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right),$$

where ε denotes the desired accuracy.
Which of the following statements is **true**?

- ☐ Let $a \in \mathcal{A}$. Then it is not possible that gradient descent finds an ε-accurate solution of the function a in $\frac{42}{\sqrt{\varepsilon}}$ iterations.
- ☐ It cannot hold $\mathcal{T}_\mathcal{B} = \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\right)$.
- ☐ None of the other four choices.
- ☐ Let $a \in \mathcal{A}$ and $b \in \mathcal{B}$. Then gradient descent reaches ε-accuracy strictly faster (in less iterations) on function a than on function b.
- ☐ Let $a \in \mathcal{A}$ and $b \in \mathcal{B}$. Then gradient descent reaches ε-accuracy strictly faster (in less iterations) on function b than on function a.

Question 4 For a function class $\mathcal{F}$ of differentiable functions $f\colon \mathbb{R}^d \to \mathbb{R}$, the following inequality holds after T iterations of gradient descent (with appropriately chosen stepsize):

$$f(\mathbf{x}_T) - f^\star \leq \frac{A}{\sqrt{T}} + \frac{B}{T^3}$$

where $A, B \geq 0$ are parameters (depending on the objective function), $\mathbf{x}_T \in \mathbb{R}^d$ the output of the algorithm after T iterations, and $f^\star$ the optimum value of f.
After which number T of iterations does it hold $f(\mathbf{x}_T) - f^\star \leq \varepsilon$, for any arbitrary $\varepsilon > 0$?

- ☐ For $T = \mathcal{O}\left(\frac{A}{\sqrt{\varepsilon}} + \frac{B}{\varepsilon^3}\right)$.
- ☐ For none of the other four choices.
- ☐ For $T = \mathcal{O}\left(\frac{\sqrt{A}}{\sqrt{\varepsilon}} + \frac{B^3}{\varepsilon^3}\right)$.
- ☐ For $T = \mathcal{O}\left(\frac{A}{\varepsilon^2} + \frac{B}{\varepsilon^{1/3}}\right)$.
- ☐ For $T = \mathcal{O}\left(\frac{A^2}{\varepsilon^2} + \frac{B^{1/3}}{\varepsilon^{1/3}}\right)$.

Convexity

Question 5 Consider the function $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = \sqrt{|x|}$, defined on the interval $I = [-1, 1]$ (see Figure 1 on the next page). Which of the following statements is **true**?

- ☐ The function f is concave in the interval I.
- ☐ The function f is convex in the interval I.
- ☐ The function f is smooth in the interval I.
- ☐ None of the other four choices.
- ☐ The function f is star convex w.r.t $x = 0$ in the interval I.

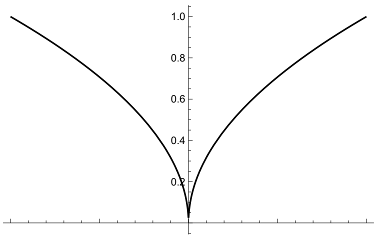

Figure 1: The function value $f(x) = \sqrt{|x|}$ (y-axis) on the interval $x \in [-1, 1]$ (x-axis).

Question 6 Let the differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-smooth. Which of the following statements is **true**?

- ☐ The function $g(\mathbf{x}) = f(\mathbf{x}) + L\,\|\mathbf{x}\|^2$ is also L-smooth.
- ☐ The function $g(\mathbf{x}) = f(\mathbf{x}) - \mu\,\|\mathbf{x}\|^2$ is convex.
- ☐ None of the other four choices.
- ☐ We always have that $L \geq \mu$, and if $L = \mu$, then f must be of the form $f(\mathbf{x}) = \frac{L}{2}\,\|\mathbf{x} - \mathbf{b}\|^2 + c$ for some $\mathbf{b} \in \mathbb{R}^d$ and $c \in \mathbb{R}$.
- ☐ Let $A \in \mathbb{R}^{d\times d}$ be a matrix. If A is negative definite, i.e. $A \prec 0$, then $g(\mathbf{x}) = f(A\mathbf{x})$ is not smooth.

Recall the notation: $A \prec 0$ means that $\mathbf{x}^\top A\mathbf{x} < 0$, for all $\mathbf{x} \in \mathbb{R}^d$.

Nonconvex optimization

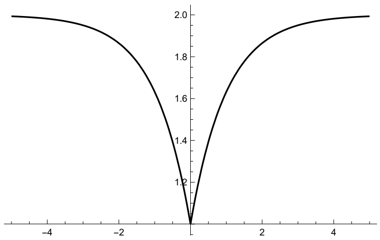

Figure 2: The function value $f(x) = 2 - \exp(-|x|)$ (y-axis) on the interval $x \in [-4, 4]$ (x-axis).

Question 7 Define the univariate function $f(x) = 2 - \exp(-|x|)$ (see Figure 2 above). We consider (any) $x_t \in \mathbb{R}$ and $x_{t+1} = x_t - \nabla f(x_t)$. Which of the following statements is **true**?

- ☐ $f(x_{t+1}) < f(x_t)$
- ☐ $\text{sign}(x_{t+1})$=$\text{sign}(x_t)$
- ☐ $|x_{t+1}| < |x_t|$
- ☐ None of the other four choices.
- ☐ $||\nabla f(x_{t+1})|| \leq ||\nabla f(x_t)||$

Question 8 Consider a differentiable L-smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$.
Which of the following statements is **true**?

- ☐ It holds $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq 2L(f(\mathbf{x}) - f(\mathbf{y}))$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.
- ☐ It holds $\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2 \leq L\|\mathbf{x} - \mathbf{y}\|^2$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.
- ☐ Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, for $d \geq 2$, be two points such that $\|\nabla f(\mathbf{x})\| = 0$, $\|\nabla f(\mathbf{y})\| = 0$. Then it must hold $f(\mathbf{x}) = f(\mathbf{y})$.
- ☐ It holds $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.
- ☐ Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, for $d \geq 2$, be two points such that $\|\nabla f(\mathbf{x})\| = 0$, $\|\nabla f(\mathbf{y})\| = 0$. Then $\|\nabla f(\mathbf{z})\| = 0$, for all $\mathbf{z} = \lambda \mathbf{x} + (1 - \lambda)\mathbf{y}$, $\lambda \in [0, 1]$.

Distributed Optimization

Question 9 Consider a distributed optimization problem of the form $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$ for an integer $n \geq 1$, where each $f_i\colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth. Let $\mathbf{x}^\star \in \mathbb{R}^d$ be such that $\nabla f(\mathbf{x}^\star) = \mathbf{0}$ (i.e. the all-0-vector).
Which of the following statements is **true**?

- ☐ It must hold $\nabla f_i(\mathbf{x}^\star) = \mathbf{0}$ for all $i \in [n]$.
- ☐ For every pair, $i, j \in [n]$, $i \neq j$, it must hold $\nabla f_i(\mathbf{x}^\star) = -\nabla f_j(\mathbf{x}^\star)$.
- ☐ By the optimality condition, the point $\mathbf{x}^\star$ must be a minimizer of the function f, $f(\mathbf{x}^\star) = \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$.
- ☐ It must hold $\|\nabla f_i(\mathbf{x}^\star)\| \leq L$, for all $i \in [n]$.
- ☐ None of the other four choices.

Question 10 Consider a distributed convex optimization problem of the form $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$ for an integer $n \geq 1$, where each $f_i\colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth, f is convex, and there exists stochastic gradient oracles $\mathbf{g}^{(i)}\colon \mathbb{R}^d \to \mathbb{R}^d$ with $\mathbb{E}[\mathbf{g}^{(i)}] = \nabla f_i(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^d$, $i \in [n]$ and $\mathbb{E}[\|\mathbf{g}^{(i)}(\mathbf{x}) - \nabla f_i(\mathbf{x})\|^2] \leq M\|\nabla f_i(\mathbf{x})\|^2 + \sigma^2$, $\forall \mathbf{x} \in \mathbb{R}^d$, $i \in [n]$. Suppose we have n machines and each machine i has access only to $\mathbf{g}^{(i)}$.
Consider the following algorithm: For a stepsize $\gamma > 0$, and $\mathbf{x}_t \in \mathbb{R}^d$,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\gamma}{|S_t|} \sum_{i \in S_t} \left(\frac{1}{B} \sum_{b=1}^{B} \mathbf{g}_b^{(i)} \right)$$

where $\mathbf{g}_b^{(i)}$ for $b = 1, \ldots, B$ denote independent realizations of the random variable $\mathbf{g}^{(i)}(\mathbf{x}_t)$, $B \geq 1$ denotes the local batch size, and $S_t \subseteq [n]$ denotes a set of indices.
Which of the following statements is **true**?

- ☐ To determine a good (maybe optimal) stepsize γ, it suffices to consider problem specific parameters (such as L, M and σ). The best stepsize does not depend on the algorithm's parameters S_t and B.
- ☐ None of the other four choices.
- ☐ When $S_t \neq [n]$, this algorithms suffers from drift which can be addressed by Scaffold or Prox-Skip/Scaffnew.
- ☐ When $\sigma > 0$ and $S_t = [n]$, then the dominant terms (i.e. the terms decreasing slowest in T) in the convergence guarantee after T iterations depend only on the product of T and B, that is, (BT), but not on the individual values of B or T.
- ☐ When $B > 1$, this algorithm is identical to LocalSGD (when $S_t = [n]$), or Federated Averaging (when $S_t \subseteq [n]$.)

Optimization in Machine Learning

Question 11 Consider the logistic regression loss $L\colon \mathbb{R}^d \to \mathbb{R}$ for a binary classification task with data $(\mathbf{a}_i, b_i) \in \mathbb{R}^d \times \{0,1\}$ for $i \in [n]$:

$$L(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(\log\left(1+e^{\mathbf{a}_i^\top \mathbf{x}}\right) - b_i \mathbf{a}_i^\top \mathbf{x}\right).$$

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

- ☐ $\nabla L(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{a}_i \frac{e^{\mathbf{a}_i^\top \mathbf{x}}}{1+e^{\mathbf{a}_i^\top \mathbf{x}}} - b_i \mathbf{a}_i^\top \mathbf{x}\right)$
- ☐ $\nabla L(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{a}_i\left(b_i - \frac{e^{\mathbf{a}_i^\top \mathbf{x}}}{1+e^{\mathbf{a}_i^\top \mathbf{x}}}\right)$
- ☐ $\nabla L(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e^{\mathbf{a}_i^\top \mathbf{x}}}{1+e^{\mathbf{a}_i^\top \mathbf{x}}} - b_i \mathbf{a}_i\right)$
- ☐ $\nabla L(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{a}_i\left(\frac{1}{1+e^{-\mathbf{a}_i^\top \mathbf{x}}} - b_i\right)$
- ☐ None of the other four choices.

Question 12 Consider the least squares objective

$$f(\mathbf{x}) = \frac{1}{2}\left\|A\mathbf{x} - \mathbf{b}\right\|^2$$

for $A \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{x} \in \mathbb{R}^d$, and the following algorithm:
In iteration t, pick an index $i_t \in [d]$, and update:

$$\begin{aligned} s_t &= \mathbf{a}_{i_t}^\top(\mathbf{y}_t - \mathbf{b}) \\ \mathbf{x}_{t+1} &= \mathbf{x}_t - \gamma s_t \mathbf{e}_{i_t} \\ \mathbf{y}_{t+1} &= \mathbf{y}_t - \gamma s_t \mathbf{a}_{i_t} \end{aligned}$$

for variables $\mathbf{x}_t \in \mathbb{R}^d, \mathbf{y}_t \in \mathbb{R}^n$ and a stepsize $\gamma > 0$. Where $\mathbf{a}_i \in \mathbb{R}^n$ denotes the i-th column of A and $\mathbf{e}_i \in \mathbb{R}^d$ denotes the i-th unit vector. One iteration of the algorithm consists of updating all three variables by the equations shown above.
Which of the following statements is **true**?

- ☐ When $\mathbf{y}_0$ is correctly initialized, $\mathbf{y}_0 = \mathbf{0}$, then this algorithm is identical to coordinate descent.
- ☐ Given an index i_t, one iteration of the algorithm can be implemented with $\mathcal{O}(d + \log n)$ arithmetic operations.
- ☐ Given an index i_t, one iteration of the algorithm can be implemented with $\mathcal{O}(n)$ arithmetic operations.
- ☐ None of the other four choices.
- ☐ When $\mathbf{x}_0$ is correctly initialized, $\mathbf{x}_0 = A^\top \mathbf{y}_0$, then this algorithm is identical to coordinate descent.

Second part, true/false questions

There is **exactly one** correct answer per question. *1 point* for each correct answer.

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 (Convexity) Let $g \colon \mathbb{R}^d \to \mathbb{R}$ be a convex and nonnegative function (i.e. $g(\mathbf{x}) \geq 0, \forall \mathbf{x} \in \mathbb{R}^d$). Then $f(\mathbf{x}) = g(\mathbf{x})^2$ is also convex.

☐ TRUE ☐ FALSE

Question 15 (Variance reduction) Consider a convex and smooth finite-sum optimization problem. SGD has worse oracle complexity than SVRG when the target accuracy ε is large.

☐ TRUE ☐ FALSE

Question 16 (Adaptive methods) Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be L-smooth and B-Lipschitz. Let $\mathbf{g}_t$ denote the unbiased stochastic gradient of $\nabla f(\mathbf{x}_t)$ with $||\mathbf{g}_t|| \leq B$. Recall that the stepsize of AdaGrad (scalar version) is defined as $\gamma_t = \frac{c}{\sqrt{\varepsilon^2 + \sum_{i=0}^t ||\mathbf{g}_i||^2}}$. For any fixed c and ε, AdaGrad can always converge.

☐ TRUE ☐ FALSE

Question 17 (Proximal method) Consider the composite objective $f = g + h$ where $g : \mathbb{R}^d \to \mathbb{R}$ and $h : \mathbb{R}^d \to \mathbb{R}$ are convex and g is differentiable. A well known property of gradient descent on f is that for a minimzer $\mathbf{x}^\star$, a gradient step from $\mathbf{x}^\star$ stays at $\mathbf{x}^\star$. This does not hold for a proximal gradient step, i.e. $\mathbf{x}^\star \neq \mathrm{prox}_{h,\gamma}(\mathbf{x}^\star - \gamma \nabla g(\mathbf{x}^\star))$ for some stepsize $\gamma > 0$, where $\mathbf{x}^\star = \mathrm{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$.

☐ TRUE ☐ FALSE

Question 18 (Nonconvex objective) Let $f \colon \mathbb{R}^d \to \mathbb{R}$ be a smooth function. If f is strongly convex, then gradient descent converges to the critical point of f which is also the global minima. However, if f is nonconvex (but still smooth), then gradient descent might not converge to a critical point.

☐ TRUE ☐ FALSE

Question 19 (Compression) Consider $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$ where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is a convex differentiable function. Let $\mathcal{C}_\delta$ be a δ-compressor, i.e. $\mathbb{E}\left\|\mathcal{C}_\delta(\mathbf{x}) - \mathbf{x}\right\|^2 \leq (1-\delta)\left\|\mathbf{x}\right\|^2, \forall \mathbf{x} \in \mathbb{R}^d$. Let $\mathbf{x}^\star$ be a minimizer of f, then $\mathbb{E}\left[\sum_{i=1}^n \mathcal{C}_\delta(\nabla f_i(\mathbf{x}^\star))\right] = \mathbf{0}$.

☐ 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.

Quadratic Upper Bounds

Recall that a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth if f is differentiable and

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

Question 20: *3 points.* Recall the gradient descent algorithm $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$ (as described earlier). In the lecture we have proven that gradient descent converges for the stepsize $\gamma = \frac{1}{L}$. What happens when using other stepsizes?

Concretely, which is the largest value of $\alpha \geq 0$ for which the decrease condition

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \alpha \left\|\nabla f(\mathbf{x}_t)\right\|^2$$

holds, when gradient descent is used with the stepsize $\gamma = \frac{1}{2L}$ instead?

(If no positive α exists, justify your answer by giving an example function.)

☐0 ☐1 ☐2 ☐3

Question 21: *2 points.* Your friend Alice has an L-smooth function f that she wants to minimize. Her implementation of gradient descent does only work support either the stepsize $\gamma = \frac{1}{2L}$ or the stepsize $\gamma = \frac{3}{2L}$. Which of these two stepsizes would you recommend her to use in practice, and why?

☐ 0 ☐ 1 ☐ 2

Extragradient Method and Relative Lipschitzness

In this subsection, we consider a *Variational Inequality.* Suppose we are given an operator $g\colon \mathbb{R}^d \to \mathbb{R}^d$, we say that $\mathbf{x}^\star$ is a solution to the variational inequality if

$$\langle g(\mathbf{x}^\star), \mathbf{x}^\star - \mathbf{x}\rangle \leq 0, \forall \mathbf{x} \in \mathbb{R}^d \ .$$

For some tolerance ε, we want to design an algorithm that outputs some $\hat{\mathbf{x}}$ such that $\langle g(\hat{\mathbf{x}}), \hat{\mathbf{x}} - \mathbf{x}\rangle \leq \varepsilon, \forall \mathbf{x} \in \mathbb{R}^d$. We introduce a class of operators for which this task is easy:

Definition A *We say that an operator $g\colon \mathbb{R}^d \to \mathbb{R}^d$ is L-relatively Lipschitz if for every three $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$, we have:*

$$\langle g(\mathbf{x}) - g(\mathbf{y}), \mathbf{x} - \mathbf{z}\rangle \leq \frac{L}{2}\left(\|\mathbf{x} - \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{z}\|^2\right) \ .$$

Definition B *We say that an operator $g\colon \mathbb{R}^d \to \mathbb{R}^d$ is monotone if*

$$\langle g(\mathbf{x}) - g(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq 0, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \ .$$

Question 22: *3 points.* Prove that if the operator g is L-Lipschitz, then g is also L-relatively Lipschitz.

☐ 0 ☐ 1 ☐ 2 ☐ 3

Question 23: *2 points.*
Prove that if a function $f : \mathbb{R}^d \to \mathbb{R}$ is convex, then $g(\mathbf{x}) := \nabla f(\mathbf{x})$ is monotone.

☐ 0 ☐ 1 ☐ 2

Question 24: *7 points.* Consider Algorithm 1 below. This is called the extragradient method. Assume that g is L-relatively Lipschitz monotone. Show that the iterates $\{\mathbf{y}_t\}$ satisfy for all $\mathbf{u} \in \mathbb{R}^d$:

$$\sum_{t=0}^{T-1} \langle g(\mathbf{y}_t), \mathbf{y}_t - \mathbf{u} \rangle \leq \frac{L}{2} \|\mathbf{x}_0 - \mathbf{u}\|^2 .$$

Hint: Consider $\langle g(\mathbf{x}_t), \mathbf{y}_t - \mathbf{x}_{t+1} \rangle$ and $\langle g(\mathbf{y}_t), \mathbf{x}_{t+1} - \mathbf{u} \rangle$.
Recall the identity $\langle \mathbf{a} - \mathbf{b}, \mathbf{b} - \mathbf{c} \rangle = \frac{1}{2}(\|\mathbf{c} - \mathbf{a}\|^2 - \|\mathbf{c} - \mathbf{b}\|^2 - \|\mathbf{b} - \mathbf{a}\|^2)$.

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

Algorithm 1 Extragradient method

1: **Input:** initial point $\mathbf{x}_0$, L-relatively Lipschitz mootone $g \colon \mathbb{R}^d \to \mathbb{R}^d$
2: **for** $r = 0, 1, 2, \ldots, T$ **do**
3: $\quad \mathbf{y}_t \leftarrow \mathbf{x}_t - \frac{1}{L} g(\mathbf{x}_t)$
4: $\quad \mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \frac{1}{L} g(\mathbf{y}_t)$
5: **end for**

Question 25: *1 point.* Given a convex L-smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$, show that you can minimize f using the extragradient method.

☐ 0 ☐ 1

When Overparameterization Meets Local Stepsizes

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function of the form $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$, $n > 1$, and each $f_i(\mathbf{x}) := \frac{1}{2}(\mathbf{x}-\mathbf{b})^T \mathbf{A}_i(\mathbf{x}-\mathbf{b})$ where $\mathbf{b} \in \mathbb{R}^d$ and $\mathbf{A}_i \in \mathbb{R}^{d\times d}$ for each $i \in [n]$ is a positive definite **diagonal** matrix.

Question 26: *4 points.* Show that each $f_i(\mathbf{x})$ is L_i-smooth and μ_i-strongly convex and therefore $f(\mathbf{x})$ is L-smooth and μ-strongly convex. Give the closed form of the parameters μ_i, μ, L_i, and L.

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4

Question 27: *4 points.* We are given n nodes where each node i has access to $\nabla f_i(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$ (where the f_i are still as defined in the previous question).

Consider full-batch gradient descent with the classical **constant** stepsize $\frac{1}{L}$, i.e.,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{Ln} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_t) \, .$$

Suppose $\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 = 1$. How many iterations does gradient descent need to have $||\mathbf{x}_t - \mathbf{x}^\star|| \leq \varepsilon$? (show the dependence on μ, L and ε.)

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4

Question 28: *3 points.* Assume the same setting as in the previous question: We are given n nodes where each node i has access to $\nabla f_i(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$.

Consider now full-batch gradient descent with the **local** stepsize $\frac{1}{L_i}$, i.e.

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{L_i} \nabla f_i(\mathbf{x}_t) \,.$$

Can this algorithm converge? If yes, please prove and show the convergence rate. If no, please give an example.

☐ 0 ☐ 1 ☐ 2 ☐ 3

Question 29: *2 points.* Are there functions for which local stepsizes can be arbitrarily better than constant stepsize? Can you give an example?

☐ 0 ☐ 1 ☐ 2