















Examiner: Sebastian Stich  
Optimization for Machine Learning  
04.08.2022 from 09h30 to 12h00  
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, 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 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 <b>PAS</b> faire   what should <b>NOT</b> be done   was man <b>NICHT</b> tun sollte		
     		



## First part, multiple choice

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

### Convexity

**Question 1** Let  $f: \mathbb{R} \rightarrow \mathbb{R}$  and  $g: \mathbb{R} \rightarrow \mathbb{R}$  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**?

- ☒ C and F are convex.
- ☐ A, B, C are convex.
- ☐ D and E are non-convex.
- ☐ A, and D are convex.
- ☐ None of the other four choices.

**Question 2** Let  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  be a function that can be written as  $f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$ , where  $n \geq 1$  is an integer and each  $f_i: \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $i = 1, \dots, n$ , is a convex function. Which statement is **true**?

- ☐ The function  $f(\mathbf{x})$  is strongly convex.
- ☐ Let  $\mathbf{x}^* \in \operatorname{argmin} f(\mathbf{x})$  be a optimal solution. Then it holds  $\|\nabla f_i(\mathbf{x}^*)\| = 0$  for at least one  $i \in [n]$ .
- ☒ None of the other four choices.
- ☐ There exists (at least one) minimizer  $\mathbf{x}^* \in \operatorname{argmin} f(\mathbf{x})$ .
- ☐ Let  $\mathbf{x}^* \in \operatorname{argmin} f(\mathbf{x})$  be a optimal solution. Then it holds  $\|\nabla f_i(\mathbf{x}^*)\| = 0$  for all  $i = 1, \dots, n$ .

**Question 3** Let  $\mathbf{A} \in \mathbb{R}^{n \times d}$  denote a matrix (the *data matrix*) and  $\mathbf{b} \in \mathbb{R}^n$  a vector (the *labels*). The  $\ell_2$ -regularized least squares problem can be written as

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

where  $\lambda \geq 0$  is a (fixed) parameter. Which of the following statements is **false**?

- ☐ None of the other four choices.
- ☐ The objective  $f(\mathbf{x})$  can be written as a finite-sum:  $f(\mathbf{x}) = \sum_{i=1}^n (\mathbf{a}_i^\top \mathbf{x} - \mathbf{b}_i)^2 + \lambda \|\mathbf{x}\|_2^2$  for vectors  $\mathbf{a}_i \in \mathbb{R}^d$ .
- ☐ This objective function  $f(\mathbf{x})$  is coordinate-wise smooth along every coordinate direction.
- ☒ This objective function  $f(\mathbf{x})$  is strongly convex.
- ☐ The computation of the gradient  $\nabla f(\mathbf{x})$  requires  $\mathcal{O}(nd)$  arithmetic operations.



## (Stochastic) Gradient Descent

**Question 4** Your friend Bob implemented 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 convex function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ . In Figure 1 he plotted the evolution of the function value in log-scale,  $\log(f(\mathbf{x}_t))$ , while using three different stepsizes:  $\gamma$ ,  $\frac{\gamma}{2}$  and  $\frac{\gamma}{4}$ . Can you help Bob to label the curves correctly?

Which labels (stepsize  $\rightarrow \{A, B, C\}$ ) are **correct**?

- ☐  $\gamma \rightarrow A, \frac{\gamma}{2} \rightarrow B, \frac{\gamma}{4} \rightarrow C$
- ☐ None of the other four choices.
- ☒  $\gamma \rightarrow A, \frac{\gamma}{2} \rightarrow C, \frac{\gamma}{4} \rightarrow B$
- ☐  $\gamma \rightarrow B, \frac{\gamma}{2} \rightarrow C, \frac{\gamma}{4} \rightarrow A$
- ☐  $\gamma \rightarrow C, \frac{\gamma}{2} \rightarrow B, \frac{\gamma}{4} \rightarrow A$

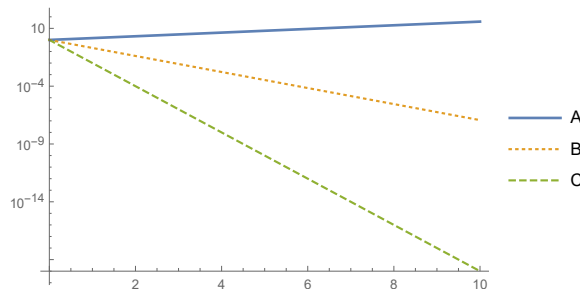


Figure 1: Three runs of gradient descent with different stepsizes,  $\{\gamma, \gamma/2, \gamma/4\}$ , on a convex function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ .  $x$ -axis: # iterations  $t$ ,  $y$ -axis:  $\log(f(\mathbf{x}_t))$ .

**Question 5** Given a function  $f: \mathbb{R} \rightarrow \mathbb{R}$ . We assume that a stochastic oracle is providing us with a stochastic gradient  $\mathbf{g}(\mathbf{x})$  for an input  $\mathbf{x} \in \mathbb{R}$ . We consider the following two stochastic oracles, where  $U \sim [0, 1]$  is a random variable that is sampled uniformly at random from the interval  $[0, 1]$  each time the oracle is called.

$$\mathbf{g}_A(\mathbf{x}) := \begin{cases} 3\nabla f(\mathbf{x}), & \text{w. prob. } \frac{1}{2} \\ -U \cdot \nabla f(\mathbf{x}), & \text{w. prob. } \frac{1}{2} \end{cases} \quad \mathbf{g}_B(\mathbf{x}) := \begin{cases} 2\nabla f(\mathbf{x}) - 1, & \text{w. prob. } \frac{1}{2} \\ \nabla f(\mathbf{x}) + 1, & \text{w. prob. } \frac{1}{2} \end{cases}$$

Which statement is **true**?

- ☐ Oracle A and B are both unbiased.
- ☐ Oracle A is unbiased, oracle B is biased.
- ☒ Oracle A and B are both biased.
- ☐ Oracle A is biased, oracle B is unbiased.

**Question 6** Recall the point estimator  $\hat{\Theta}_\alpha = \alpha(X - Y) + \mathbb{E}[Y]$ , defined for two random variables  $X, Y$  with  $\text{Cov}[X, Y] > 0$  and a parameter  $\alpha \in [0, 1]$ .

Which one of the following statements about  $\hat{\Theta}_\alpha$  is **false**?

- ☐ The variance  $\mathbb{E} \left\| \hat{\Theta}_\alpha - \mathbb{E}[\hat{\Theta}_\alpha] \right\|^2$  increases as  $\alpha$  increases from 0 to 1.
- ☐ If  $\mathbb{E}[Y] = \mathbb{E}[X]$ , the estimator is unbiased for any  $\alpha$ , i.e.  $\mathbb{E}[\hat{\Theta}_\alpha] = \mathbb{E}[X]$ , for all  $\alpha \in [0, 1]$ .
- ☐ If  $\alpha = 1$ , the estimator is unbiased, i.e.  $\mathbb{E}[\hat{\Theta}_1] = \mathbb{E}[X]$ .
- ☒ The bias  $\left\| \mathbb{E}[\hat{\Theta}_\alpha] - \mathbb{E}[X] \right\|$  increases as  $\alpha$  increases from 0 to 1.



## Complexity Estimates

**Question 7** Your friend Alice has developed a new iterative algorithm to minimize the gradient norm of an arbitrary differentiable function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ . She has proven that after performing  $T$  steps of her algorithm the following guarantee holds:

$$\|\nabla f(\mathbf{x}_T)\|^2 \leq \frac{L}{T} + \frac{\sigma^2}{T^2} + \frac{M}{T^{5/2}},$$

where  $L, M, \sigma^2 \geq 0$  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 her to derive the correct complexity estimate, i.e. after which number  $T$  of iterations does it hold  $\|\nabla f(\mathbf{x}_T)\|^2 \leq \varepsilon$ , for any arbitrary  $\varepsilon > 0$ ?

- ☒ For  $T = \mathcal{O}\left(\frac{L}{\varepsilon} + \frac{\sigma}{\sqrt{\varepsilon}} + \frac{M^{2/5}}{\varepsilon^{2/5}}\right)$ .
- ☐ For  $T = \mathcal{O}\left(\frac{L}{\varepsilon} + \frac{\sigma^2}{\varepsilon^2} + \frac{M}{\varepsilon^{5/2}}\right)$ .
- ☐ None of the other four choices.
- ☐ For  $T = \mathcal{O}\left(\frac{L}{\varepsilon} + \frac{\sigma^2}{\sqrt{\varepsilon}} + \frac{M}{\varepsilon^{2/5}}\right)$ .
- ☐ For  $T = \mathcal{O}\left(\frac{L+\sigma+M^{2/5}}{\varepsilon^{2/5}}\right)$

**Question 8** Consider a  $L$ -smooth,  $\mu$ -strongly convex finite-sum optimization problem  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ , with  $f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$  for an integer  $n \geq 1$ . In the lecture we have proven that stochastic gradient descent, defined as the iteration  $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla f_{i_t}(\mathbf{x}_t)$  for a uniformly at random selected index  $i_t \in [n]$ , converges in  $\mathcal{O}\left(\frac{\sigma^2}{\mu\varepsilon} + \kappa \log \frac{1}{\varepsilon}\right)$  iterations to an  $\varepsilon$ -accurate solution, where  $\kappa = \frac{L}{\mu}$  denotes the condition number and  $\sigma^2$  is an upper bound on the stochastic variance  $\mathbb{E}_i \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \sigma^2, \forall \mathbf{x} \in \mathbb{R}^d$ . Which of the following statements is **true**?

- ☒ (Full batch) Gradient descent on  $f$  converges in at most  $\mathcal{O}\left(n\kappa \log \frac{1}{\varepsilon}\right)$  gradient computations.
- ☐ None of the other four choices.
- ☐ No algorithm can converge in fewer iterations than  $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$  on this problem (lower bound of Nemirovski and Yudin).
- ☐ By using a batch size  $b > 1$  in SGD, we obtain a linear speedup in the convergence:  $\mathcal{O}\left(\frac{\sigma^2}{b\mu\varepsilon} + \frac{\kappa}{b} \log \frac{1}{\varepsilon}\right)$ .
- ☐ By using (Nesterov-type) acceleration in SGD, we can improve the convergence rate to  $\mathcal{O}\left(\frac{\sigma^2}{\mu\sqrt{\varepsilon}} + \sqrt{\kappa} \log \frac{1}{\varepsilon}\right)$ .



**Question 9** You have found an old optimization book that defines the class of  $(A, B, C)$ -convex functions  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ , for parameters  $A, B, C \geq 0$ . You tell your roommates, Fred and Lea, about it. After some thoughts, both come up with an algorithm to minimize such functions. For a starting point  $\mathbf{x}_0 \in \mathbb{R}^d$ , the output  $\mathbf{x}_T^{\text{Fred}}$ , obtained after  $T$  steps of Fred's algorithm, has the following property:

$$f(\mathbf{x}_T^{\text{Fred}}) - f(\mathbf{x}^*) + A \|\mathbf{x}_T^{\text{Fred}} - \mathbf{x}^*\|^2 \leq \frac{A}{2T} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{B}{\sqrt{T}} + \frac{C}{T^2},$$

where  $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$  denotes a minimizer of  $f$ . Conversely, Lea's algorithm has the property:

$$f(\mathbf{x}_T^{\text{Lea}}) - f(\mathbf{x}^*) \leq \frac{A \|\mathbf{x}_0 - \mathbf{x}^*\|^2}{T} + \frac{B}{T} + \frac{C}{T^2},$$

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

- ☐ Suppose we would first use Fred's algorithm for  $T/2$  iterations to reduce  $\|\mathbf{x}_{T/2}^{\text{Fred}} - \mathbf{x}^*\|^2$ , and then switch to Lea's algorithm with  $\mathbf{x}_0 = \mathbf{x}_{T/2}^{\text{Fred}}$  as a starting point and run it for another  $T/2$  iterations. This new algorithm is faster than both, Fred's or Lea's original proposal (for convergence in function suboptimality,  $f(\mathbf{x}_T) - f(\mathbf{x}^*)$ ).
- ☐ Suppose we are only interested in a small function value gap,  $f(\mathbf{x}_T) - f(\mathbf{x}^*)$ . If  $C > 0$ , then asymptotically for large  $T$  both algorithms are equally good (up to constant factors).
- ☐ Suppose  $B = C = 0$  and  $A \geq 0$ . Then Fred's algorithm is the preferred choice if we aim to find a solution with high accuracy (say,  $\varepsilon = 10^{-9}$ ), while Lea's algorithm performs better for low accuracy (say,  $\varepsilon = 10^{-1}$ ).
- ☐ None of the other four choices.
- ☒ The function  $|x|^3$  is not  $(A, B, C)$ -convex for any choice of  $A, B, C$ .

**Question 10** Consider a federated 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  $f_i: \mathbb{R}^d \rightarrow \mathbb{R}$  is given in stochastic form, that is, there exists gradient oracles  $\mathbf{g}^{(i)}: \mathbb{R}^d \rightarrow \mathbb{R}^d$ , with the properties  $\mathbb{E} \mathbf{g}^{(i)} = \nabla f_i(\mathbf{x})$ ,  $\forall \mathbf{x} \in \mathbb{R}^d, i \in [n]$  and  $\mathbb{E} \|\mathbf{g}^{(i)}(\mathbf{x}) - \nabla f_i(\mathbf{x})\|^2 \leq \sigma^2$ ,  $\forall \mathbf{x} \in \mathbb{R}^d, i \in [n]$ . Which of the following statements is **true**?

- ☐ Consider mini-batch SGD with client sampling that computes a mini-batch gradient of the form  $\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathbf{g}^{(i)}(\mathbf{x})$  for a randomly chosen subset  $\mathcal{S} \subsetneq [n]$ . This algorithm does not suffer from client drift and converges linearly.
- ☐ When each  $f_i, i \in [n]$  is a quadratic function then there is no client drift (i.e. every minimizer of an  $f_i$  component is also a global minimizer).
- ☐ Local SGD converges linearly when the number of local steps  $\tau \leq n$ .
- ☒ None of the other four choices.
- ☐ SCAFFOLD can only be applied to problems with this structure when  $\sigma^2 = 0$ .



## Second part, true/false questions

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

**Question 11** (Coordinate-wise Lipschitz) Let  $L_i$ ,  $i = 1, \dots, d$ , denote the coordinate-wise Lipschitz constants of a differentiable  $L$ -smooth function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ . It holds  $\sum_{i=1}^d L_i \leq L$ .

☐ TRUE ☒ FALSE

**Question 12** (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  $L = \|\mathbf{Q}\|_F$ .

☒ TRUE ☐ FALSE

**Question 13** (Compressor) Consider the operator  $\mathcal{C}: \mathbb{R} \rightarrow \mathbb{R}$  defined as

$$\mathcal{C}(\mathbf{x}) = \begin{cases} -\mathbf{x} & \text{w. prob. } \frac{1}{3} \\ \mathbf{0} & \text{w. prob. } \frac{1}{3} \\ \mathbf{x} & \text{w. prob. } \frac{1}{3} \end{cases}$$

$\mathcal{C}$  is a  $\delta$ -compressor for  $\delta = \frac{5}{6}$ .

☐ TRUE ☒ FALSE

**Question 14** (Projection) Let  $C$  be the unit-level set of a star-convex function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$ , that is,  $C = \{\mathbf{x} \in \mathbb{R}^d \mid f(\mathbf{x}) \leq 1\}$ . The projection onto  $C$ ,  $\text{proj}_C: \mathbb{R}^d \rightarrow C \subset \mathbb{R}^d$ , can be written as a convex optimization problem:

$$\text{proj}_C(\mathbf{x}) = \underset{\mathbf{y} \in C}{\text{argmin}} \|\mathbf{x} - \mathbf{y}\|.$$

☐ TRUE ☒ FALSE

**Question 15** (Strong Convexity) Let  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  be a  $\mu$ -strongly convex function, for a parameter  $\mu > 0$ . Then

$$f(\mathbf{x}) \geq \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 + f(\mathbf{x}^*), \quad \forall \mathbf{x} \in \mathbb{R}^d,$$

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

☒ TRUE ☐ FALSE

**Question 16** (Convexity) A function  $f(x)$  is *convex* if and only if  $g(x) = -f(x)$  is *non-convex*.

☐ TRUE ☒ FALSE

**Question 17** (Convexity) Any critical point of a convex differentiable function on an open domain is a global minimizer of the function.

☒ 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.

#### Descent and Convergence

**Question 18:** 3 points. Let  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  be a  $L$ -smooth function. Consider one step of gradient descent with an exact line search, that is, given  $\mathbf{x}_t \in \mathbb{R}^d$ , define  $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t^* \nabla f(\mathbf{x}_t)$ , where  $\gamma_t^*$  is chosen in the optimal way:  $\gamma_t^* \in \operatorname{argmin}_{\gamma \geq 0} f(\mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t))$ , i.e. to minimize the function value along the negative gradient direction. Can you derive a sufficient decrease lemma for gradient descent with exact line search?

Derive the the best (i.e. smallest) possible upper bound on  $f(\mathbf{x}_{t+1})$  and argue why your bound is tight.

☐ 0 ☐ 1 ☐ 2 ☒ 3

**Solution:** Observe that the stepsize  $\gamma = \frac{1}{L}$  used in the standard decrease lemma is best possible in general ( $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{x}$ ) and so the decrease lemma  $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2$  cannot be improved in general. (1 point for applying smoothness quadratic upper bound, 1 point for correct sufficient decrease bound with the right constant, 1 point for arguing tightness).

**Question 19:** 2 points. Let  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  be of the form  $f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$  for an integer  $n \geq 1$  and  $f_i: \mathbb{R}^d \rightarrow \mathbb{R}$  convex and  $L_i$ -smooth for  $L_i \geq 0$ ,  $i = 1, \dots, n$ . Let  $\mathbf{x}^* \in \mathbb{R}^d$  denote a global minimum of the function  $f$ . Prove that

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

for  $L = \max\{L_1, \dots, L_n\}$ .

☐ 0 ☐ 1 ☒ 2

**Solution:** This is Exercise 10.1. 1 point for (deriving/citing/recalling/...) the inequality  $f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \nabla f_i(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq \frac{1}{2L_i} \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|^2$  that holds for convex smooth functions. 1 point for deriving the conclusion with  $L_i \leq L_{\max}$  and  $\nabla f(\mathbf{x}^*) = 0$ .

**Question 20:** 2 points. Let  $\mathbf{x}_t \in \mathbb{R}^d$ ,  $t = 0, \dots, T$  denote the iterates generated by an iterative algorithm. The iterates satisfy the following relation:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \leq (1 - \gamma) \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \gamma \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle + \gamma^2 A ,$$

where  $A \geq 0, \gamma \geq 0$  are parameters and  $\mathbf{x}^* \in \mathbb{R}^d$ . Prove that it holds

$$\frac{1}{T} \sum_{t=0}^{T-1} \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{\gamma T} + \gamma A .$$

Which parameter  $\gamma \geq 0$  minimizes this upper bound?

☐ 0 ☐ 1 ☒ 2

**Solution:** Note that  $(1 - \gamma) \|\mathbf{x}_t - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^*\|^2$  and use the standard proof technique from the lecture (1 point). We choose  $\gamma = \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|}{\sqrt{AT}}$  (1 point).



## Sigma-star Assumption

For the next three questions, let  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  be a  $\mu$ -strongly convex function of the form  $f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$ ,  $n \geq 1$ , where each  $f_i: \mathbb{R}^d \rightarrow \mathbb{R}$  is a  $L$ -smooth function. Let  $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$ .

**Question 21:** 2 points. Consider the stochastic gradient oracle  $\nabla f_i(\mathbf{x})$ , where the index  $i$  is sampled uniformly at random with probability  $\frac{1}{n}$ . Prove that its second moment can be bounded as follows:

$$\mathbb{E}_i \|\nabla f_i(\mathbf{x})\|^2 \leq 4L(f(\mathbf{x}) - f(\mathbf{x}^*)) + 2\sigma_*^2,$$

where  $\sigma_*^2 := \mathbb{E} \|\nabla f_i(\mathbf{x}_*)\|^2$ .

☐ 0 ☐ 1 ☒ 2

**Solution:** Add and subtract  $\nabla f_i(\mathbf{x}^*)$ ,  $\|\nabla f_i(\mathbf{x})\|^2 = \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*) + \nabla f_i(\mathbf{x}^*)\|^2$ , with the inequality  $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$  (1 point) and use smoothness (inequality stated in Question 19, 1 point).

**Question 22:** 2 points. Consider the stochastic gradient algorithm  $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla f_{i_t}(\mathbf{x}_t)$ , where the index  $i_t$  is sampled uniformly at random. By using the result stated in the previous question, prove that (under appropriate conditions on  $\gamma$ ) the one step progress of SGD can be upper bounded as follows (recall that  $f$  is assumed to be  $\mu$ -strongly convex):

$$\mathbb{E}_{i_t} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \leq (1 - \mu\gamma) \|\mathbf{x}_t - \mathbf{x}^*\|^2 - \gamma(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + 2\gamma^2\sigma_*^2.$$

and state the necessary condition on the stepsize  $\gamma \geq 0$ .

☐ 0 ☐ 1 ☒ 2

**Solution:** Use strong convexity (+1), and the inequality stated in the previous question with  $\gamma \leq \frac{1}{4L}$ . (+1).

**Question 23:** 1 point. Compare this result with what you have seen in the course.

☐ 0 ☒ 1

**Solution:** No bounded gradient or bounded variance assumption needed!

## Consensus Optimization with Gossip

We are given  $n$  nodes that each hold a (private) value  $x_i \in \mathbb{R}$ , for  $i = 1, \dots, n$ .

The nodes aim to compute the average  $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$  of their values, while avoiding central communication. For this, they invent the following iterative protocol: at each iteration (or time step)  $t$ , a coordinator selects two different nodes  $a \in [n]$ ,  $b \in [n]$ ,  $a \neq b$  uniformly at random. The selected nodes then compute the average of their parameters and replace their parameters with the computed average. We call this process *randomized pairwise gossip averaging*.

**Question 24:** 1 point. In the lecture we had seen that this process can be written in matrix notation, for instance as  $\mathbf{x}_{t+1} = \mathbf{W}\mathbf{x}_t$  for  $\mathbf{x}_t, \mathbf{x}_{t+1} \in \mathbb{R}^n$  and a mixing matrix  $\mathbf{W} \in \mathbb{R}^{n \times n}$ . In this notation, the  $i$ -th coordinate of  $\mathbf{x}_t$  corresponds to the state on node  $i$  at time step  $t$ .

Suppose that in iteration  $t$  the nodes  $(a, b)$ ,  $a < b$ , have been picked. Derive the mixing matrix  $\mathbf{W}_{(a,b)}$  that corresponds to averaging between nodes  $a$  and  $b$  (as described in the process above).

☐ 0 ☒ 1

**Solution:** We have  $(\mathbf{W}_{(a,b)})_{i,j} = \frac{1}{2}$  for  $i \in \{a, b\}$ ,  $j \in \{a, b\}$  (all 4 combinations),  $(\mathbf{W}_{(a,b)})_{i,i} = 1$  for  $i \notin \{a, b\}$ , and zero otherwise.





We now derive an alternative formulation of this process. Let  $\mathcal{E} := \{(a, b) \mid a \in [n], b \in [n], a < b\}$  denote all (ordered) pairs of nodes. Then we can define  $f: \mathbb{R}^n \rightarrow \mathbb{R}$ :

$$f(\mathbf{x}) := \frac{2}{n(n-1)} \sum_{(a,b) \in \mathcal{E}} \left[ f_{(a,b)}(\mathbf{x}) := \frac{1}{2}(\mathbf{x}_a - \mathbf{x}_b)^2 \right].$$

Note that  $\frac{n(n-1)}{2} = |\mathcal{E}|$  and is just the normalization by the number of terms in the sum.

**Question 25:** 3 points. Show that each  $f_{(a,b)}(\mathbf{x}) = \frac{1}{2}(\mathbf{x}_a - \mathbf{x}_b)^2$  is smooth and strongly convex and find the best possible parameters smoothness and strong-convexity parameters. Here  $\mathbf{x}_a$  denotes the  $a$ -th coordinate of  $\mathbf{x}$ , and  $\mathbf{x}_b$  the  $b$ -th coordinate of  $\mathbf{x}$ .

What does follow for  $f$ , in terms of smoothness and convexity? Can you estimate its parameters?

0  1  2  3

**Solution:** Note that it must hold  $\mu_{(a,b)} = L_{(a,b)}$  (each  $f_{(a,b)}$  is a quadratic function, 1 point). We can compute  $L_{(a,b)} = 2$  (1 point). From the sum structure we can conclude  $\mu \geq \frac{2}{n(n-1)}\mu_{(a,b)}$ ,  $L \leq L_{(a,b)} \leq 2$  (alternatively, we could also compute the Hessian of  $f$ ). (1 point for the correct order of magnitude of  $L$  and  $\mu$ .)

**Question 26:** 1 point. Suppose we pick one pair  $(a, b) \in \mathcal{E}$  uniformly at random and perform a (stochastic) gradient step with stepsize  $\gamma = \frac{1}{2}$ :

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{2} \nabla f_{(a,b)}(\mathbf{x}_t)$$

Show that this update is equivalent to the gossip averaging step  $\mathbf{x}_{t+1} = \mathbf{W}_{(a,b)} \mathbf{x}_t$ .

0  1

**Solution:** Note that the gradient update gives  $(\mathbf{x}_{t+1})_a = \frac{1}{2}(\mathbf{x}_t)_a + \frac{1}{2}(\mathbf{x}_t)_b = (\mathbf{x}_{t+1})_b$ .

**Question 27:** 6 points. Derive the convergence rate of randomized pairwise gossip averaging. How many iterations  $T$  (in big- $\mathcal{O}$  notation) are needed to reduce the optimization error ( $\|\mathbf{x}_t - \mathbf{x}^*\|^2 \leq \varepsilon$ , where  $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$ ) to less than  $\varepsilon \geq 0$ ? (Justification needed to earn points).

0  1  2  3  4  5  6

**Solution:** We have proven so far that gossip averaging is equivalent to picking one  $f_{(a,b)}$  uniformly at random and performing a (stochastic) gradient update:  $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{2} \nabla f_{(a,b)}$ . We therefore need to derive a convergence rate for  $\|\mathbf{x}_t - \mathbf{x}^*\|^2$ , or  $f(\mathbf{x}_t) - f(\mathbf{x}^*)$ .

A possible proof:

- 1 point for observing what needs to be proven (or noting  $f(\mathbf{x}^*) = 0$ , what is  $\mathbf{x}^*$ , etc., or similar structural insights)
- $\mathbb{E} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \leq (1 - \mu\gamma) \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\gamma(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \gamma^2 \mathbb{E} \|\nabla f_{(a,b)}(\mathbf{x}_t)\|^2$  (usual expansion and strong convexity applied, 1 point)
- Observe  $\mathbb{E} \|\nabla f_{(a,b)}(\mathbf{x}_t)\|^2 = 4f(\mathbf{x}_t)$  (+ 2 points, full points for any other (reasonable) bound  $\mathbb{E} \|\nabla f_{(a,b)}(\mathbf{x}_t)\|^2 \leq \mathcal{O}(1) \cdot f(\mathbf{x}_t)$ , 1 point for a valid bound of the form  $\mathbb{E} \|\nabla f_{(a,b)}(\mathbf{x}_t)\|^2 \leq A \|\nabla f(\mathbf{x}_t)\|^2 + B$ ).
- 1 point for picking a stepsize small enough (e.g.  $\gamma \leq \frac{1}{2}$ ), to obtain

$$\mathbb{E} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \leq (1 - \mu\gamma) \|\mathbf{x}_t - \mathbf{x}^*\|^2$$

(or similar).

- 1 point for unrolling and complexity  $T = \mathcal{O}(\frac{1}{\mu} \log \frac{1}{\varepsilon}) = \mathcal{O}(n^2 \log \frac{1}{\varepsilon})$ . Full points if rate is stated in terms of  $\mu, L$  without plugging in their values (see Exercise 25).