EPFL

Nicolas Flammarion
Optimization for Machine Learning — CS-439 - MA
20.06.2025 from 15h15 to 18h15
Duration : 180 minutes

1

Student 1

SCIPER : **999000**

Wait for the start of the exam before turning to the next page. This document is printed double sided, 20 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.
- You each have a different exam.
- This exam has many questions. We do *not* expect you to solve all of them even for the best grade.
- Only answers in this booklet count. No extra loose answer sheets. You can use the last two pages as scrap paper.
- For the **multiple choice** questions, we give +2 points if your answer is correct, and 0 points for incorrect or no answer.
- For the **true/false** questions, we give +1.5 points if your answer is correct, and 0 points for incorrect or no answer.
- Use a **black or dark blue ballpen** and clearly erase with **correction fluid** if necessary.
- If a question turns out to be wrong or ambiguous, we may decide to nullify it.

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 questions

For each question, mark the box corresponding to the correct answer. Each question has **exactly one correct answer**.

Question 1 (Gradient Descent) Let $f(x) = \frac{1}{2}\mathbf{x}^\top Q\mathbf{x}$, where $Q \in \mathbb{R}^{d\times d}$ is a positive-definite matrix. Let $\lambda_1 > \lambda_2 > \cdots > \lambda_d > 0$ denote the eigenvalues of Q. Consider applying gradient descent with a fixed step size γ. There exists a threshold $\tilde{\gamma} > 0$ such that for any $\gamma > \tilde{\gamma}$, the algorithm diverges for certain initialization points. Determine the smallest such $\tilde{\gamma}$.

- ☐ $\frac{1}{\sum_{i=1}^d \lambda_i}$
- ☐ $\frac{1}{\lambda_1}$
- ☐ $\frac{2d}{\sum_{i=1}^d \lambda_i}$
- ☐ $\frac{2}{\sum_{i=1}^d \lambda_i}$
- ☐ $\frac{2}{\lambda_d}$
- ☒ $\frac{2}{\lambda_1}$
- ☐ $\frac{1}{\lambda_d}$
- ☐ $\frac{d}{\sum_{i=1}^d \lambda_i}$

Solution: The update rule is $\mathbf{x}_{k+1} = (I - \gamma Q)\mathbf{x}_k$. Convergence requires $|1 - \gamma\lambda_i| < 1$ for all i, which implies:

$$0 < \gamma < \frac{2}{\lambda_1}.$$

Thus, the smallest $\tilde{\gamma}$ such that divergence occurs for $\gamma > \tilde{\gamma}$ is $\frac{2}{\lambda_1}$.

Question 2 (Gradient Descent) Consider the problem of estimating a fixed but unknown vector $\mathbf{x} \in \mathbb{R}^d$. We are given a dataset of T observations, $\mathcal{D} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T\}$. Each observation $\mathbf{y}_t$ is generated independently by the following process:

$$\mathbf{y}_t = \alpha_t \mathbf{x} + \beta_t \boldsymbol{\varepsilon}_t$$

where α_t, β_t are known non-zero scalar coefficients and the noise is drawn independently from each other and $\mathbf{x}$ as $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, I_d)$ where I_d is the d by d identity matrix.
For a dataset $\mathcal{D}$ and model parameters $\boldsymbol{\theta}$, the likelihood function $q(\mathcal{D}|\boldsymbol{\theta})$ represents the probability of observing the data given the parameters. The maximum likelihood estimator (MLE) is the value of $\boldsymbol{\theta}$ that maximizes this function:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} q(\mathcal{D}|\boldsymbol{\theta})$$

Recall that for $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_d)$, the probability density function is given by $f(\mathbf{z}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\mathbf{z}-\boldsymbol{\mu}\|_2^2}{2\sigma^2}\right)$. Which of the following optimization problems correctly formulates the MLE for the unknown parameter $\boldsymbol{\theta} = \mathbf{x}$ given the entire dataset $\mathcal{D}$ and parameters $\{(\alpha_t, \beta_t) | t \in \{1, \dots, T\}\}$?

- [ ] $\max_{\mathbf{x}\in\mathbb{R}^d} \sum_{t=1}^T \frac{1}{\beta_t^2} \|\mathbf{y}_t - \alpha_t \mathbf{x}\|_2^2$.
- [ ] $\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{t=1}^T \|\mathbf{y}_t - \alpha_t \mathbf{x}\|_2^2$.
- [ ] $\max_{\mathbf{x}\in\mathbb{R}^d} \sum_{t=1}^T \exp\left(-\frac{\|\mathbf{y}_t - \alpha_t \mathbf{x}\|_2^2}{2\beta_t^2}\right)$.
- [x] $\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{t=1}^T \frac{1}{\beta_t^2} \|\mathbf{y}_t - \alpha_t \mathbf{x}\|_2^2$.
- [ ] $\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{t=1}^T \beta_t^2 \|\mathbf{y}_t - \alpha_t \mathbf{x}\|_2^2$.
- [ ] $\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{t=1}^T \frac{1}{\beta_t} \|\mathbf{y}_t - \alpha_t \mathbf{x}\|_1$.
- [ ] $\max_{\mathbf{x}\in\mathbb{R}^d} \log\left(\sum_{t=1}^T \frac{1}{\beta_t} \exp(-\|\mathbf{y}_t - \alpha_t \mathbf{x}\|_2^2)\right)$.

Solution: The correct answer is $\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{t=1}^T \frac{1}{\beta_t^2} \|\mathbf{y}_t - \alpha_t \mathbf{x}\|_2^2$.
Since the observations are independent, the joint likelihood is the product of the individual likelihoods

$$q(\mathcal{D}|\mathbf{x}) = \prod_{t=1}^T q(y_t|\mathbf{x})$$

hence

$$\hat{\mathbf{x}}_{\text{MLE}} = \arg\max_{\mathbf{x}} \prod_{t=1}^T q(y_t|\mathbf{x}) = \arg\min_{\mathbf{x}} \left(-\sum_{t=1}^T \log q(y_t|\mathbf{x})\right)$$

For our Gaussian model we have

$$\log q(y_t|\mathbf{x}) = -\frac{\|y_t - \alpha_t \mathbf{x}\|_2^2}{2\beta_t^2} - \underbrace{\frac{d}{2}\log(2\pi\beta_t^2)}_{\text{Constant w.r.t. } \mathbf{x}}$$

Therefore

$$\hat{\mathbf{x}}_{\text{MLE}} = \arg\min_{\mathbf{x}} \sum_{t=1}^T \frac{\|y_t - \alpha_t \mathbf{x}\|_2^2}{2\beta_t^2}.$$

Question 3 (Convexity) Let $A \in \mathbb{R}^{m\times d}$ and $\mathbf{w} \in \mathbb{R}^m$. Define the mapping

$$f : \mathbb{R}^d \longrightarrow \mathbb{R}, \qquad f(\mathbf{x}) \;=\; \mathbf{w}^\top \sigma(A\mathbf{x}),$$

where the ReLU activation $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ is applied coordinate-wise; that is, for $\mathbf{z} \in \mathbb{R}^m$,

$$\big[\sigma(\mathbf{z})\big]_i \;=\; \max\{0,\, z_i\}, \qquad i = 1, \ldots, m.$$

Thus the network consists of an input linear layer $\mathbf{x} \mapsto A\mathbf{x}$, the ReLU activation, and an output linear layer $\sigma \mapsto \mathbf{w}^\top\sigma$. Under which of the following conditions is the mapping $\mathbf{x} \mapsto f(\mathbf{x})$ **convex**?

- [ ] All entries of the output weights must be positive.
- [ ] The output weight matrix must be positive semi-definite.
- [ ] The input weight matrix must be positive definite.
- [x] All entries of the output weights must be non-negative.
- [ ] All entries of the input weights must be positive.
- [ ] The input weight matrix must be positive semi-definite.
- [ ] The output weight matrix must be positive definite.
- [ ] All entries of the input weights must be non-negative.

Solution: All entries of the output weights must be non-negative.

Question 4 (Convexity) Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Consider the hypercube C in $\mathrm{dom}(f)$:

$$C \;=\; [l_1, u_1] \times [l_2, u_2] \times \ldots \times [l_d, u_d], \quad \text{with } l_i < u_i, \forall i$$

Define the set of the 2^d vertices of C by

$$V \;=\; \big\{\mathbf{v} \in \mathbb{R}^d \;:\; v_i \in \{l_i, u_i\} \text{ for every } i\big\}.$$

Pick an arbitrary point $\mathbf{x} \in C$. Which of the following statements is **necessarily true**?

- [ ] The minimum of f over C is attained only at a single point in V.
- [x] The maximum of f over C can be attained at one or more points in V.
- [ ] The maximum of f over C is attained only at a single point in V.
- [ ] The minimum of f over C can occur on the boundary of C without ever occurring at a point in V.
- [ ] The minimum of f over C can be attained at one or more points in V.
- [ ] The maximum of f over C can occur on the boundary of C without ever occurring at a point in V.

Solution: The maximum of f over C is attained at one or more vertices of C. The boundary can have equal values as $\mathbf{x}$.

Question 5 (Projected Gradient Descent) Consider the following function $f : [1, 3] \to \mathbb{R}$ defined as $f(x) = x^2 - 4x + 3$. Note that the function is defined on the closed interval $\mathcal{X} = [1, 3]$. We minimize f with the projected gradient descent algorithm, which is defined as follows:

$$x_{k+1} = P_{\mathcal{X}}(x_k - \gamma \nabla f(x_k)),$$

where $P_{\mathcal{X}}$ is the projection operator onto the set $\mathcal{X}$, and $\gamma > 0$ is a fixed step size. Over the choice of initialization x_0 and step size γ, which of the following scenarios are **not possible**?

- ☐ $x_7 = 3$ and $x_{14} = 1$
- ☐ $x_7 = 1.2$ and $x_{14} = 3$ and $x_{21} = 1$
- ☐ $x_7 = 1.2$ and $x_{14} = 1.6$ and $x_{21} = 1.8$
- ☐ $x_7 = 1.5$ and $x_{14} = 2.5$
- ■ $x_7 = 1.2$ and $x_{14} = 1.8$ and $x_{21} = 2.4$

Solution: When all the iterates are in the interval $[1, 3]$, the projected gradient descent algorithm can be written as follows:

$$\mathbf{x}_{2t} - 2 = (1 - \gamma)^t (\mathbf{x}_t - 2)$$

Using the above relation, it can be seen that if $(1 - \gamma)$ is positive then $x_{t'} - 2$ cannot change sign, hence, the sequence $\mathbf{x}_7 - 2 = -.8$ and $\mathbf{x}_{14} - 2 = -.2$ and $\mathbf{x}_{21} - 2 = 0.4$ cannot change sign. Even in the case when the iterates does not stay in the interval $[1, 3]$, it implies that $|(1 - \gamma)| > 1$ and the iterates diverge away from 2 and $x_{14} = 1.8$ is not possible.

Question 6 (Proximal Gradient Descent) Consider the functions $g : \mathbb{R}^d \to \mathbb{R}$ and $h : \mathbb{R}^d \to \mathbb{R}$ defined as follows:

$$g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x}, \quad h(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top B\mathbf{x}$$

where A, B are positive-definite diagonal matrices. Let $f = g + h$. We are interested in minimizing f using the following algorithms:

PGD-1 : With a fixed step size $\gamma > 0$, the proximal gradient descent algorithm is defined as

$$\mathbf{x}_{t+1}^{\text{PGD-1}} = \text{Prox}_{h,\gamma}\left(\mathbf{x}_t^{\text{PGD-1}} - \gamma\nabla g(\mathbf{x}_t^{\text{PGD-1}})\right)$$

PGD-2 : With a fixed step size $\gamma > 0$, the second proximal gradient descent algorithm with the proximal orcale defined with g is defined as

$$\mathbf{x}_{t+1}^{\text{PGD-2}} = \text{Prox}_{g,\gamma}\left(\mathbf{x}_t^{\text{PGD-2}} - \gamma\nabla h(\mathbf{x}_t^{\text{PGD-2}})\right)$$

GD : With a fixed step size $\gamma > 0$, the gradient descent algorithm is defined as

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

Recall the proximal operator $\text{Prox}_{g,\gamma}(\mathbf{z})$ is defined as

$$\text{Prox}_{g,\gamma}(\mathbf{z}) = \arg\min_{\mathbf{y}}\left\{\frac{1}{2\gamma}||\mathbf{y} - \mathbf{z}||^2 + g(\mathbf{y})\right\}.$$

where $||\cdot||$ is the Euclidean norm. When the three above algorithms are initialized at $\mathbf{x}_0$ and are run with a same step size $0 < \gamma < \frac{1}{L}$, where L is the smoothness constant of f. For $t > 0$, which of the following statements is **always true** regarding the rate of convergence of function f?

- [ ] $f(\mathbf{x}_t^{\text{PGD-1}}) = f(\mathbf{x}_t^{\text{PGD-2}}) \leqslant f(\mathbf{x}_t^{\text{GD}})$
- [x] $f(\mathbf{x}_t^{\text{GD}}) \leqslant f(\mathbf{x}_t^{\text{PGD-1}})$
- [ ] $f(\mathbf{x}_t^{\text{PGD-1}}) \leqslant f(\mathbf{x}_t^{\text{GD}})$ and $f(\mathbf{x}_t^{\text{PGD-2}}) \leqslant f(\mathbf{x}_t^{\text{GD}})$
- [ ] None of the remaining statements are always true.
- [ ] $f(\mathbf{x}_t^{\text{PGD-1}}) = f(\mathbf{x}_t^{\text{PGD-2}}) \geqslant f(\mathbf{x}_t^{\text{GD}})$

Solution: The proximal operator for g and h are given by $\text{Prox}_{g,\gamma}(\mathbf{z}) = (1 + \gamma A)^{-1}\mathbf{z}$ and $\text{Prox}_{h,\gamma}(\mathbf{x}) = (1 + \gamma C)^{-1}\mathbf{z}$. Using this the update rule for PGD-1, PGD-2 and GD can be written as

$$\begin{aligned}
\mathbf{x}_{t+1}^{\text{PGD-1}} &= (1 + \gamma A)^{-1}(1 - \gamma B)\mathbf{x}_t^{\text{PGD-1}} \\
\mathbf{x}_{t+1}^{\text{PGD-2}} &= (1 + \gamma B)^{-1}(1 - \gamma A)\mathbf{x}_t^{\text{PGD-2}} \\
\mathbf{x}_{t+1}^{\text{GD}} &= (1 - \gamma(A + B))\mathbf{x}_t^{\text{GD}}.
\end{aligned}$$

It can be seen that $(1 - \gamma(A + B)) \preccurlyeq (1 + \gamma A)^{-1}(1 - \gamma B)$ and $(1 - \gamma(A + B)) \preccurlyeq (1 + \gamma B)^{-1}(1 - \gamma A)$ for $\gamma \leq \|A + B\|$.

Question 7 (Stochastic Gradient Descent) Let $f : \mathbb{R} \to \mathbb{R}$ be μ-*strongly convex* and L-*smooth*, with $\mu > 0$ and $L > 0$. Let $\mathbf{x}^\star$ denote the unique minimizer of f, and define the minimum value as $f^\star = f(\mathbf{x}^\star)$. At each iteration $t = 0, 1, 2, \ldots$, we observe a stochastic gradient $\mathbf{g}_t$ satisfying:

$$\mathbb{E}[\,\mathbf{g}_t \mid \mathbf{x}_t] = f'(\mathbf{x}_t), \qquad \mathbb{E}\big[(\mathbf{g}_t - f'(\mathbf{x}_t))^2 \mid \mathbf{x}_t\big] = \sigma^2, \quad \text{with } \sigma^2 > 0.$$

For a fixed step size $0 < \gamma < \frac{1}{L}$, consider the following updates:

$$\text{GD: } \mathbf{x}_{t+1} = \mathbf{x}_t - \gamma\, f'(\mathbf{x}_t), \qquad \text{SGD: } \mathbf{x}_{t+1} = \mathbf{x}_t - \gamma\, \mathbf{g}_t.$$

We say that a method "oscillates around $f^\star$ within range $\mathcal{O}(\xi)$" if there exists a $T_0 \in \mathbb{N}$ and a constant $C > 0$ that is independent of ξ such that:

$$\mathbb{E}[f(\mathbf{x}_t)] - f^\star \leq C\xi \quad \text{for all} \quad t \geq T_0\,.$$

Choose the statement that is **always true** under the conditions above:

- [ ] GD oscillates around $f^\star$ within $\mathcal{O}(\gamma\sigma^2)$ while SGD converges to $f^\star$.
- [x] SGD oscillates around $f^\star$ within range $\mathcal{O}(\gamma\sigma^2)$, while GD is guaranteed to converge to $f^\star$.
- [ ] Both methods do not converge and oscillate around $f^\star$ within range $\mathcal{O}(\gamma\sigma^2)$.
- [ ] Both methods are guaranteed to converge to $f^\star$.

Solution: SGD oscillates around$f^\star$ within range $\mathcal{O}(\gamma\sigma^2)$, while GD is guaranteed to converge to $f^\star$. Test the concept of SGD to see that it is not a noise-free process. No test on the exact form of the noise. Proof: Let

$$x_{t+1} = x_t - \gamma\, g_t, \qquad \mathbb{E}[g_t \mid x_t] = f'(x_t), \qquad \mathbb{E}\big[(g_t - f'(x_t))^2 \mid x_t\big] = \sigma^2.$$

Define the expected value: $\Delta_t := \mathbb{E}[f(x_t)] - f^\star$. Because f is L-smooth,

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + f'(x_t)(x_{t+1} - x_t) + \frac{L}{2}(x_{t+1} - x_t)^2 \\ &= f(x_t) - \gamma\, f'(x_t) g_t + \frac{L\gamma^2}{2} g_t^2. \end{aligned}$$

Taking conditional expectation given x_t:

$$\mathbb{E}(f(x_{t+1})|x_t) \leq f(x_t) - \gamma(f'(x_t))^2 + \frac{L\gamma^2}{2}(f'(x_t)^2 + \sigma^2)$$

Take the total expectation and subtract the optimal $f^\star$:

$$\Delta_{t+1} \leq \Delta_t - (\gamma - \frac{L\gamma^2}{2})\mathbb{E}(f'(x_t)^2) + \frac{L\gamma^2\sigma^2}{2}$$

Because the function is μ-strongly convex (this question is not about PL so use stronger condition to replace),

$$f'(x_t)^2 \geq 2\mu(f(x_t) - f^\star) = 2\mu\Delta_t$$

Plug in this inequality:

$$\Delta_{t+1} \leq (1 - 2\gamma\mu - \gamma^2 L\mu)\Delta_t + \frac{L\gamma^2\sigma^2}{2} \leq (1 - \gamma\mu)\Delta_t + \frac{L\gamma^2\sigma^2}{2}$$

Apply the recursion:

$$\lim_{t\to\infty} \Delta_t \leq \frac{L\gamma^2\sigma^2}{2\gamma\mu} = \frac{L}{2\mu}\gamma\sigma^2$$

Hence the resulting upper bound.

Question 8 (Non-convex Optimization) Consider the non-convex function $f : \mathbb{R} \to \mathbb{R}$ defined via the mapping $f(x) = -e^{-x^2}$. In order to minimize f, we run gradient descent (GD) on f with step size $0 < \gamma < \infty$ starting from some $x_0 \neq 0$. Which of the following is true?

- ☐ We have no convergence guarantees for GD since f is non-smooth.
- ☐ Even for small step sizes, there exist some points of initialization such that GD diverges to either $-\infty$ or $+\infty$ since $\lim_{x\to\pm\infty} \nabla f(x) = 0$.
- ■ GD converges to 0 for step size $1/2$, but not for step size $\gamma = 2$.
- ☐ Depending on initialization and step-size, GD can converge to either one of the three points of inflection $\{0, \pm\sqrt{1/2}\}$ of f.
- ☐ None of the other options is true.
- ☐ GD converges to 0 for step sizes $\gamma = 1/2$ and $\gamma = 2$.

Solution: The correct answer is: "GD converges to 0 for step size $1/2$."

- It holds that $f'(x) = 2xe^{-x^2}, f''(x) = (2 - 4x^2)e^{-x^2}$, hence the smoothness parameter of f is $L = \max_x |f''(x)| = |f''(0)| = 2$.
- Step-size 2: This step-size is too large and will lead to oscillations around 0.
- Step-size $1/2$: This step-size will lead to convergence of GD to only critical point (and global minimum) 0, see Theorem 6.2.
- At the points $\{\pm\sqrt{1/2}\}$ there exist decreasing directions, hence GD will never converge to them, no matter the step-size.

Question 9 (Polyak-Lojasiewicz Inequality) Assume $f : \mathbb{R}^d \to \mathbb{R}$ fulfils the Polyak-Lojasiewicz (PL) inequality. Which of the following functions are PL (with some arbitrary parameter $0 < \mu < +\infty$)?

- $h_1(\mathbf{x}) := f(A\mathbf{x})$, where A is a square invertible matrix.
- $h_2(\mathbf{x}) := [f(\mathbf{x})]^2$.
- $h_3(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})$ where g also satisfies PL.

- ☐ Only h_2.
- ☐ Only h_1 and h_3.
- ☐ None of the functions has the PL property.
- ■ Only h_1.
- ☐ Only h_1 and h_2.
- ☐ Only h_2 and h_3.
- ☐ Only h_3.

Solution: The correct answer is: "Only h_1."

- h_1: By the chain rule, the gradient is $\nabla h(x) = A^\mathsf{T} \nabla f(Ax)$. Using the fact $\left\|A^\mathsf{T} \nabla f(Ax)\right\|^2 \geq \sigma^2_{\min}(A) \left\|\nabla f(Ax)\right\|^2 \geq 2\sigma^2_{\min}(A) \left(\mu_f(f(Ax) - f^*)\right)$ (the last inequality follows from f being PL), it follows that h_1 is PL.
- h_2: take for example $f(x) = x^2$. Then, f is PL, but it is easy to show that $h_2(x) = x^4$ is not PL.
- h_3: The gradient of the sum is $\nabla h(x) = \nabla f(x) + \nabla g(x)$. The individual PL properties of f and g do not prevent a scenario where at some non-optimal point x_0 (i.e., $h(x_0) > h^*$), their gradients are equal and opposite: $\nabla f(x_0) = -\nabla g(x_0)$ This cancellation results in a stationary point for the sum, $\nabla h(x_0) = \nabla f(x_0) + \nabla g(x_0) = 0$, where the function value is not the global minimum. Hence, $0 = \frac{1}{2} \left\|\nabla h(x_0)\right\|^2 \geq \mu_h(h(x_0) - h^*) > 0$ which is a contradiction.

Question 10 (Linear Minimization Oracles) We consider the computational complexity of linear minimization oracles (LMOs) for the constraint set $\mathcal{X}$ being unit radius $\mathcal{L}_p$ balls in $\mathbb{R}^d$.
Recall that $\mathcal{B}_p = \left\{ \mathbf{x} \in \mathbb{R}^d \,\middle|\, \left(\sum_{i=1}^d |x_i|^p\right)^{1/p} \leq 1 \right\}$, for $p \in \{1, 2\}$ and $\mathcal{B}_\infty = \left\{ \mathbf{x} \in \mathbb{R}^d \mid \max_{1 \leq i \leq d} |x_i| \leq 1 \right\}$.
Which of the following statements are the only true ones?

A) There exist LMOs for $\mathcal{B}_1, \mathcal{B}_2$, and $\mathcal{B}_\infty$ that can all be evaluated in $\mathcal{O}(d)$ time.

B) Any LMO for $\mathcal{B}_\infty$ has time complexity $\Omega(2^d)$ because $\mathcal{B}_\infty$ has 2^d extremal points.

C) There exist an LMO for $\mathcal{B}_2$ that has time complexity $\mathcal{O}(1)$, since it suffices for the LMO to simply rescale the gradient.

D) In terms of per-iteration cost, Frank-Wolfe is preferable over projected gradient descent for $\mathcal{X} = \mathcal{B}_1$.

E) In terms of per-iteration cost, Frank-Wolfe is preferable over projected gradient descent for $\mathcal{X} = \mathcal{B}_2$.

F) In terms of per-iteration cost, Frank-Wolfe is preferable over projected gradient descent for $\mathcal{X} = \mathcal{B}_\infty$.

- ☐ B), C), D) and E)
- ■ A) and D)
- ☐ B), D) and E)
- ☐ C) and E)
- ☐ C), D) and E)
- ☐ A), D) and F)
- ☐ B) and E)
- ☐ A), C), E) and F)

Solution: The correct answer is "A) and D)".

- a) is true. As seen in the lecture, there is an LMO for $\mathcal{B}_1$ with linear time complexity since the cardinality of extremal points is linear in d. For $\mathcal{B}_2$, the LMO can compute the projection $-\nabla f / \|\nabla f\|_2$, which has time complexity d due to the computation of the euclidean norm. For $\mathcal{B}_\infty$, the projection involves also only linearly many operations since it amounts to computing $\text{sign}(x_i) \min(|x_i|, 1)$ for each coordinate.
- As you were told at the beginning of the exam, the option d) is true. As discussed on the forum, while there exist linear time projections onto $\mathcal{B}_1$, the linear time LMO is faster by an appreciable constant factor.
- On the other hand, the projections onto $\mathcal{B}_1$ and $\mathcal{B}_\infty$ are already very cheap (linear in d), and there are no LMOs with strictly sub-linear runtime.

Question 11 (Subgradients) Given a convex function $f : \mathrm{dom}(f) \to \mathbb{R}$, a point $\mathbf{x}_0 \in \mathrm{dom}(f) \subseteq \mathbb{R}^d$ and a subgradient $\mathbf{g} \in \partial f(\mathbf{x}_0)$, which of the following statements are true?

A) There exists a point $\mathbf{x}_1 \neq \mathbf{x}_0$ in the domain of f such that $\mathbf{0} \in \partial f(\mathbf{x}_1)$.

B) There exists a point $\mathbf{x}_2 \neq \mathbf{x}_0$ in the domain of f such that $\mathbf{g} \in \partial f(\mathbf{x}_2)$.

C) Given a point $\mathbf{x}_3 \neq \mathbf{x}_0$ in the domain of f such that $\|\mathbf{x}_3 - \mathbf{x}_0\| \leq R$, then $f(\mathbf{x}_3) - f(\mathbf{x}_0) \leq R\|\mathbf{g}\|$.

D) Given a subgradient $\mathbf{g}' \in \partial f(\mathbf{x}_0)$, then $\lambda\mathbf{g} + (1-\lambda)\mathbf{g}'$ for any $\lambda \in [0,1]$ is also a subgradient of f at $\mathbf{x}_0$.

E) Assuming f is differentiable at $\mathbf{x}_0$, the set $\{\lambda\mathbf{g} + (1-\lambda)\nabla f(\mathbf{x}_0) \mid \lambda \in [0,1]\}$ consists of subgradients and is not a singleton.

- [ ] C), D) and E)
- [ ] B)
- [ ] A), B) and C)
- [ ] B), C) and D)
- [x] D)
- [ ] A), B) and E)
- [ ] C) and E)
- [ ] A) and B)
- [ ] D) and E)

Solution: The correct answer is D).

- A) is false, as for linear functions, there is no point where the subgradient is zero.
- B) is false, for the univariate convex function $f = 1/x$, the subgradient is unique in the domain $\mathbb{R}^+$.
- C) is false, as Cauchy-Schwarz inequality gives $f(x_3) - f(x_0) \geq -R\|g\|$. If g was a subgradient at x_3, then the desired inequality would hold.
- D) is true, as it follows from the definition of subgradients.
- E) is false, as if f is differentiable at x_0, then the set of subgradients at that point is a singleton containing only $\nabla f(x_0)$.

Question 12 (Newton's method) Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined as $f(x) = x^3/3 - x$. We run gradient descent with fixed step size $\gamma > 0$ and Newton's method both from some initial point $x_0 \in \mathbb{R}$. Which of the following statements are true?

A) Assume $x_0 \gg 0$, the gradient descent algorithm converges to a local minimum of f.

B) Regardless of the initial point x_0, the iterates of gradient descent with large enough step size γ diverges.

C) For any initial point $x_0 \neq 0$, Newton's method converges to a critical point of f.

D) For any initial point $x_0 \neq 0$, Newton's method has at most single iterate that is inside the interval $(-1, 1)$.

- [ ] A), B) and D)
- [ ] B) and D)
- [ ] B), C) and D)
- [ ] A), B) and C)
- [ ] A) and B)
- [ ] A) and D)
- [ ] A), C) and D)
- [x] C) and D)
- [ ] A) and C)
- [ ] B) and D)

Solution: The correct answer is C) and D).

- A) is false, as with a large step size, the function can jump to $x_0 \ll 0$ and the gradient descent will not converge to a local minimum.
- B) is false, as the gradient descent initialized at $x_0 = -1$ stays at the local minimum $x = -1$ and does not diverge.
- C) is true, as Newton's method converges to a critical point of f regardless of the initial point $x_0 \neq 0$. This can be seen from Theorem 7.4.
- D) is true, as the update of Newton's method is given by

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} = x_k - \frac{x_k^2 - 1}{2x_k} = \frac{x_k^2 + 1}{2x_k}.$$

  In absolute value, $|\frac{x_k^2+1}{2x_k}| \geq 1$.

Second part: true/false questions

For each question, mark the box (without erasing) TRUE if the statement is **always true** and the box FALSE if it is **not always true** (i.e., it is sometimes false).

Question 13 (Gradient Descent) The direction of the gradient points to the steepest descent direction, which is the direction pointing to a local or global minimum.

☐ TRUE ■ FALSE

Solution: False. A point on an ellipsoid loss landscape optimization.

Question 14 (Subgradients) Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ such that for any $\mathbf{x} \in \mathbb{R}^d$ and for any $\mathbf{g} \in \partial f(\mathbf{x})$, $\|\mathbf{g}\| \leq B$. Then, f is Lipschitz continuous with constant $L = B$.

☐ TRUE ■ FALSE

Solution: False. This is true only if f is convex.

Question 15 (Lower Bound) There always exists a B-Lipschitz convex function $f \colon \mathbb{R}^d \to \mathbb{R}$ with the following property: For any (sub)gradient-based method initialized at $\mathbf{x}_0$ and run for $T \geq 1$ iterations, the objective error satisfies

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \geq c\left(\frac{B}{\sqrt{T}}\right),$$

for some constant $c > 0$, where $\mathbf{x}^\star$ is the global minimum of f.

☐ TRUE ■ FALSE

Solution: False. It is true only for $T < d$.

Question 16 (Convexity) Any univariate differentiable function with a convex gradient is convex.

☐ TRUE ■ FALSE

Solution: False. $f(x) = x^3$

Question 17 (Smooth Functions) Let f be a L-smooth convex function. Consider the gradient descent algorithm with stepsize $\gamma = \frac{1}{L}$ from an initial point $\mathbf{x}_0$ such that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$, where $\mathbf{x}^\star$ is the global minimum of f. Then, after T iterations,

$$f(\frac{1}{T}\sum_{i=1}^{T}\mathbf{x}_i) - f(\mathbf{x}^\star) = \mathcal{O}(1/T).$$

■ TRUE ☐ FALSE

Solution: $f(\frac{1}{T}\sum_{i=1}^{T}\mathbf{x}_i) - f(\mathbf{x}^\star) \leq \frac{1}{T}\left[\sum_{i=1}^{T}(f(\mathbf{x}_i) - f(\mathbf{x}^\star))\right]$

Question 18 (Convexity) A convex function is continuous.

☐ TRUE ■ FALSE

Solution: False. $dom(f)$ needs to be open.

Question 19 (Subgradients) Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(\mathbf{x}) = \max_{1 \le j \le d} x_j + \frac{1}{2}\|\mathbf{x} - \mathbf{a}\|^2,$$

for some $\mathbf{a} \in \mathbb{R}^d$. There exist a unique point $\mathbf{x}^\star$ such that $\mathbf{0} \in \partial f(\mathbf{x}^\star)$.

■ TRUE ☐ FALSE

Solution: The function is strongly convex, hence a unique minimizer.

Question 20 (Quasi-Newton Methods) Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Then, the Quasi-Newton methods differ in their choice of $H_t \in \mathbb{R}^{d\times d}$ that verifies the following d-dimensional secant condition:

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

■ TRUE ☐ FALSE

Solution: H_t approximates $\nabla^2 f(\mathbf{x}_t)$ by solving the given secant condition. Since, there are many choices when $d > 1$, we have different Quasi-Newton approaches.

Question 21 (SGD) Averaging the stochastic gradients over a mini-batch of size m can reduce the variance by a factor of $\frac{1}{m^2}$ compared to the single-sample variance.

☐ TRUE ■ FALSE

Solution: False. $\frac{1}{m}$.

Question 22 (Convexity) Suppose there is a two-layer neural network: $f : \mathbb{R}^d \to \mathbb{R}$, $f(\mathbf{x}) = W_2 W_1 \mathbf{x}$, with two linear mappings W_1, W_2 and no activation function in between. The neural network is trained using a mean-squared error loss $\mathcal{L}$. $\mathcal{L}$ is convex in (W_1, W_2).

☐ TRUE ■ FALSE

Solution: False. Bilinear optimization.

Question 23 (Newton's Method) Newton's method is invariant under any invertible affine transformation.

■ TRUE ☐ FALSE

Solution: See Lemma 7.2. of lecture notes

Question 24 (ClippedSGD) A clipped stochastic gradient is an unbiased estimator of the true gradient.

☐ TRUE ■ FALSE

Solution: False. Clipping introduces a bias in the stochastic gradient.

Third part, open questions

Answer in the empty space below. Your answer should be carefully justified, and all the steps of your argument should be discussed in details. Leave the check-boxes empty, they are used for the grading.

1 Affine Invariance of Frank-Wolfe

Question 25: (*3 points*) Show that the Frank-Wolfe algorithm is *affine invariant*, which formally means the following:

Let $\{\mathbf{x}_k\}$, $\mathbf{x}_k \in \mathbb{R}^d$, be the sequence of iterates generated by applying the Frank-Wolfe algorithm to the problem $\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$ starting from $\mathbf{x}_0$. Let $T(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$ with A invertible. If the algorithm is now applied to the transformed problem $\min_{\mathbf{y}\in T(\mathcal{C})} f(T^{-1}(\mathbf{y}))$ starting from the transformed point $\mathbf{y}_0 = T(\mathbf{x}_0)$, the resulting sequence of iterates $\{\mathbf{y}_k\}$ will satisfy $\mathbf{y}_k = T(\mathbf{x}_k)$ for all $k \geq 0$.

☐0 ☐1 ☐2 ■3

Solution: By the chain rule, we have for $g(\mathbf{y}) = f(T^{-1}(\mathbf{y}))$ that $\nabla g(\mathbf{y}_k) = (A^{-1})^T \nabla f(\mathbf{x}_k)$.
The solution to the LMO in the transformed space, $\mathbf{z}_k$, is precisely the transformed LMO solution from the original space, $\mathbf{s}_k$:

$$\begin{aligned}
\mathbf{z}_k &= \arg\min_{\mathbf{z}\in T(\mathcal{C})} \langle \nabla g(\mathbf{y}_k), \mathbf{z}\rangle \\
&= T\left(\arg\min_{\mathbf{s}\in\mathcal{C}} \langle \nabla g(\mathbf{y}_k), T(\mathbf{s})\rangle\right) \\
&= T\left(\arg\min_{\mathbf{s}\in\mathcal{C}} \langle (A^{-1})^T \nabla f(\mathbf{x}_k), A\mathbf{s} + \mathbf{b}\rangle\right) \\
&= T\left(\arg\min_{\mathbf{s}\in\mathcal{C}} \left(\langle (A^{-1}A)^T \nabla f(\mathbf{x}_k), \mathbf{s}\rangle + \text{const}\right)\right) \\
&= T\left(\arg\min_{\mathbf{s}\in\mathcal{C}} \left(\langle \nabla f(\mathbf{x}_k), \mathbf{s}\rangle + \text{const}\right)\right) \\
&= T(\mathbf{s}_k).
\end{aligned}$$

Therefore,

$$\mathbf{y}_{k+1} = (1-\gamma_k)\mathbf{y}_k + \gamma_k \mathbf{z}_k = (1-\gamma_k)T(\mathbf{x}_k) + \gamma_k T(\mathbf{s}_k) = T((1-\gamma_k)\mathbf{x}_k + \gamma_k \mathbf{s}_k) = T(\mathbf{x}_{k+1}).$$

2 Cocoercivity

The goal of the exercise is to establish the cocoercivity inequality

$$\frac{1}{2L}\|\nabla f(\mathbf{z}) - \nabla f(\mathbf{y})\|^2 \leq f(\mathbf{z}) - f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{y} - \mathbf{z}\rangle \tag{Coco}$$

characterizing smooth convex functions.

Question 26: (*3 points*) Show that f is convex and L-smooth if and only if, $\forall(\mathbf{x}, \mathbf{y}, \mathbf{z})$

$$f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \leq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z}\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{z}\|^2. \tag{1}$$

☐0 ☐1 ☐2 ■3

Solution: If f is said convex and L-smooth, then using the definitions, we get $\forall(\mathbf{x}, \mathbf{y}, \mathbf{z})$:

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

and thus (1). Reciprocally, if $\forall(\mathbf{x},\mathbf{y},\mathbf{z})$

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

then with $\mathbf{x}=\mathbf{y}$, we get, $\forall(\mathbf{x},\mathbf{z})$:

$$f(\mathbf{x})\leq f(\mathbf{z})+\langle\nabla f(\mathbf{z}),\mathbf{x}-\mathbf{z}\rangle+\frac{L}{2}\|\mathbf{x}-\mathbf{z}\|^2.$$

And with $\mathbf{x}=\mathbf{z}$, we get, $\forall(\mathbf{x},\mathbf{y})$:

$$f(\mathbf{y})+\langle\nabla f(\mathbf{y}),\mathbf{x}-\mathbf{y}\rangle\leq f(\mathbf{x}).$$

Question 27: (*3 points*) Show that f is convex and L-smooth if and only if, $\forall(\mathbf{y},\mathbf{z})$

$$0\leq f(\mathbf{z})-f(\mathbf{y})+\langle\nabla f(\mathbf{y}),\mathbf{y}-\mathbf{z}\rangle-\frac{1}{2L}\|\nabla f(\mathbf{z})-\nabla f(\mathbf{y})\|^2, \tag{2}$$

i.e.,

$$\frac{1}{2L}\|\nabla f(\mathbf{z})-\nabla f(\mathbf{y})\|^2\leq f(\mathbf{z})-f(\mathbf{y})+\langle\nabla f(\mathbf{y}),\mathbf{y}-\mathbf{z}\rangle. \tag{3}$$

☐0 ☐1 ☐2 ■3

Solution: By (1), f is convex and L-smooth if and only if, $\forall(\mathbf{x},\mathbf{y},\mathbf{z})$

$$0\leq f(\mathbf{z})-f(\mathbf{y})+\langle\nabla f(\mathbf{y}),\mathbf{y}\rangle-\langle\nabla f(\mathbf{z}),\mathbf{z}\rangle+\langle\nabla f(\mathbf{z})-\nabla f(\mathbf{y}),\mathbf{x}\rangle+\frac{L}{2}\|\mathbf{x}-\mathbf{z}\|^2.$$

thus if and only if, $\forall(\mathbf{y},\mathbf{z})$

$$0\leq f(z)-f(y)+\langle\nabla f(y),y-z\rangle+\underbrace{\min_{x\in\mathbb{R}^d}\left(\langle\nabla f(z)-\nabla f(y),x-z\rangle+\frac{L}{2}\|x-z\|^2\right)}_{=-\frac{1}{2L}\|\nabla f(z)-\nabla f(y)\|^2}.$$

which gives (2).

Question 28: (*3 points*) Show that f is convex and L-smooth if and only if, $\forall(\mathbf{y},\mathbf{z})$

$$\frac{1}{L}\|\nabla f(\mathbf{z})-\nabla f(\mathbf{y})\|^2\leq\langle\nabla f(\mathbf{y})-\nabla f(\mathbf{z}),\mathbf{y}-\mathbf{z}\rangle. \tag{4}$$

Hint: for the reverse direction, show that f satisfies for any $\boldsymbol{\eta},\boldsymbol{\theta}\in\mathbb{R}^d$, $\langle\boldsymbol{\eta}-\boldsymbol{\theta},\nabla f(\boldsymbol{\eta})-\nabla f(\boldsymbol{\theta})\rangle\geq 0$, iif f is convex. Consider $g(t)=f(\boldsymbol{\theta}+t(\boldsymbol{\eta}-\boldsymbol{\theta}))$. Show that $t\geq 0$, $g'(t)\geq g'(0)$ and prove $f(\boldsymbol{\eta})\geq f(\boldsymbol{\theta})+\langle\nabla f(\boldsymbol{\theta}),\boldsymbol{\eta}-\boldsymbol{\theta}\rangle$.

☐0 ☐1 ☐2 ■3

Solution: There was an obvious typo in the hint (previously, it was $\langle\nabla f(\boldsymbol{\eta})-\nabla f(\boldsymbol{\theta}),\boldsymbol{\theta}-\boldsymbol{\eta}\rangle\geq 0$), which is now corrected.

For the first direction, (3) summed with the same inequality with y,z permuted gives (4).
For the other direction:

(a) (4) implies that ∇f is Lipschitz by Cauchy Schwartz.

(b) $g'(t)=\langle\nabla f(\boldsymbol{\theta}+t(\boldsymbol{\eta}-\boldsymbol{\theta})),\boldsymbol{\eta}-\boldsymbol{\theta}\rangle$ and thus for all $t>0$, we have

$$g'(t)-g'(0)=\frac{1}{t}\langle\nabla f(\boldsymbol{\theta}+t(\boldsymbol{\eta}-\boldsymbol{\theta}))-\nabla f(\boldsymbol{\theta}),t(\boldsymbol{\eta}-\boldsymbol{\theta})\rangle\geq 0$$

by (4). Writing $g(1)=g(0)+\int_0^1 g'(t)\mathrm{d}t\geq g(0)+g'(0)$ we get $f(\boldsymbol{\eta})-f(\boldsymbol{\theta})=\int_{t=0}^1 g'(t)\mathrm{d}t\geq\langle\nabla f(\boldsymbol{\theta}),(\boldsymbol{\eta}-\boldsymbol{\theta})\rangle$ which implies convexity of f.

Remark: another solution, is too show that $g(\mathbf{x}) := \frac{L}{2}\|\mathbf{x}\|^2 - f(\mathbf{x})$ is L-smooth and to conclude that f is thus convex.
To do so, we observe that:

$$\begin{aligned}
& \frac{1}{L}\|\nabla g(\mathbf{y}) - \nabla g(\mathbf{z})\|^2 \leq \langle \nabla g(\mathbf{y}) - \nabla g(\mathbf{z}), \mathbf{y} - \mathbf{z}\rangle \\
\Leftrightarrow \quad & \frac{1}{L}\|L(\mathbf{y} - \mathbf{z}) - (\nabla f(\mathbf{y}) - \nabla f(\mathbf{z}))\|^2 \leq \langle L(\mathbf{y} - \mathbf{z}) - (\nabla f(\mathbf{y}) - \nabla f(\mathbf{z})), \mathbf{y} - \mathbf{z}\rangle \,. \\
\Leftrightarrow \quad & \frac{1}{L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{z})\|^2 \leq \langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{z}\rangle \,.
\end{aligned}$$

Thus (4) for f is equivalent to (4) for g! And as remarked above, (4) implies L-smoothness.

Question 29: (*2 points*) Find a inequality similar to (Coco) characterizing smooth strongly convex functions.

☐ 0 ☐ 1 ■ 2

Solution: f is L-smooth and μ-strongly convex if and only if $f - \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2$ is $L - \mu$-smooth and convex. Applying (Coco) to $f - \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2$ therefore answers the question.

- Assume that f is L-smooth and μ-strongly convex. Then, $f - \frac{\mu}{2}\mathbf{x}^\top\mathbf{x}$ is convex by Lemma 2.11. Thus, $f - \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2$ is convex for any choice of $x^\star$ as only a linear term and and a constant term is added. On the other hand, $\frac{L}{2}\mathbf{x}^\top\mathbf{x} - f$ is convex by Lemma 2.3. That is, $\frac{L-\mu}{2}\mathbf{x}^\top\mathbf{x} - (f - \frac{\mu}{2}\mathbf{x}^\top\mathbf{x})$ is convex. Thus, $\frac{L-\mu}{2}\mathbf{x}^\top\mathbf{x} - (f - \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2)$ is convex. Again using Lemma 2.3, we conclude that $f - \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2$ is $L - \mu$-smooth.

- Assume that $f - \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2$ is $L - \mu$-smooth and convex. Similarly, by Lemma 2.11 and Lemma 2.3, f is μ-strongly convex and L-smooth.

Another method is to rewrite the lower bound to $f(\mathbf{x})$ by strong convexity:

$$f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2 \leq f(\mathbf{x})\,,$$

to obtain an analog of (1):

$$f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2 \leq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z}\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{z}\|^2 - \frac{\mu}{2}\|\mathbf{x} - \mathbf{z}\|^2.$$

Then, we need to follow Q27 and minimize the right-hand side with respect to $\mathbf{x}$. This leads to the same conclusion as above.

Simplest proof of Nesterov Accelerated Gradient

Consider the problem of minimizing a L-smooth convex function $f : \mathbb{R}^d \to \mathbb{R}$. For this, we consider the Nesterov accelerated gradient method which is defined by the update rule

$$\begin{cases} \mathbf{y}_t &= \mathbf{x}_t + \beta_t(\mathbf{x}_t - \mathbf{x}_{t-1}) \\ \mathbf{x}_{t+1} &= \mathbf{y}_t - \frac{1}{L}\nabla f(\mathbf{y}_t) \end{cases} \tag{NAG}$$

where $(\beta_t)_{t\in\mathbb{N}}$ is called momentum parameter. For the sake of convenience, let $\mathbf{x}_{-1} = \mathbf{x}_0$ and for $t \geq 0$, define

$$V_t \triangleq \lambda_t^2(f(\mathbf{x}_t) - f^\star) + \frac{L}{2}\|\lambda_t(\mathbf{x}_t - \mathbf{x}^\star) + (1 - \lambda_t)(\mathbf{x}_{t-1} - \mathbf{x}^\star)\|^2. \tag{5}$$

In the following questions, we will establish a rate of convergence for the NAG method using V_t.

Question 30: (*4 points*) Using (Coco), show that we have $V_{t+1} \leq V_t$ for any $t \geq 0$, with $\lambda_{t+1}^2 - \lambda_{t+1} = \lambda_t^2$ ($\lambda_0 = 0$) and $\beta_t = \frac{\lambda_t - 1}{\lambda_{t+1}}$.

□0 □1 □2 □3 ■4

Solution: First we compute

$$\begin{aligned}
V_{t+1} - V_t \le\ & \lambda_{t+1}^2 (f(\mathbf{x}_{t+1}) - f^\star) - \lambda_t^2 (f(\mathbf{x}_t) - f^\star) \\
& + \frac{L}{2}\|\lambda_{t+1}(\mathbf{x}_{t+1} - \mathbf{x}^\star) + (1-\lambda_{t+1})(\mathbf{x}_t - \mathbf{x}^\star)\|^2 - \frac{L}{2}\|\lambda_t(\mathbf{x}_t - \mathbf{x}^\star) + (1-\lambda_t)(\mathbf{x}_{t-1} - \mathbf{x}^\star)\|^2 \\
=\ & \lambda_{t+1}^2 (f(\mathbf{x}_{t+1}) - f(\mathbf{y}_t)) + \lambda_{t+1}(f(\mathbf{y}_t) - f^\star) + \lambda_t^2 (f(\mathbf{y}_t) - f(\mathbf{x}_t)) \\
& + \frac{L}{2}\|\lambda_{t+1}(\mathbf{x}_{t+1} - \mathbf{x}^\star) + (1-\lambda_{t+1})(\mathbf{x}_t - \mathbf{x}^\star)\|^2 - \frac{L}{2}\|\lambda_t(\mathbf{x}_t - \mathbf{x}^\star) + (1-\lambda_t)(\mathbf{x}_{t-1} - \mathbf{x}^\star)\|^2 \\
\le\ & -\frac{\lambda_{t+1}^2}{2L}\|\nabla f(\mathbf{y}_t)\|^2 + \lambda_{t+1}\langle \nabla f(\mathbf{y}_t), \mathbf{y}_t - \mathbf{x}^\star\rangle + \lambda_t^2 \langle \nabla f(\mathbf{y}_t), \mathbf{y}_t - \mathbf{x}_t\rangle \\
& + \frac{L}{2}\|\lambda_{t+1}(\mathbf{x}_{t+1} - \mathbf{x}^\star) + (1-\lambda_{t+1})(\mathbf{x}_t - \mathbf{x}^\star)\|^2 - \frac{L}{2}\|\lambda_t(\mathbf{x}_t - \mathbf{x}^\star) + (1-\lambda_t)(\mathbf{x}_{t-1} - \mathbf{x}^\star)\|^2
\end{aligned}$$

Then, by rearranging terms and using (NAG), we conclude

$$V_{t+1} - V_t \le -\frac{1}{2L}\left[\lambda_{t+1}^2\|\nabla f(\mathbf{x}_{t+1})\|^2 + \lambda_{t+1}\|\nabla f(\mathbf{y}_t)\| + \lambda_t^2\|\nabla f(\mathbf{y}_t) - \nabla f(\mathbf{x}_t)\|^2\right].$$

Which is the wanted inequality $V_{t+1} - V_t \le -\Delta_t$, with

$$\Delta_t = \frac{1}{2L}\left[\lambda_{t+1}^2\|\nabla f(\mathbf{x}_{t+1})\|^2 + \lambda_{t+1}\|\nabla f(\mathbf{y}_t)\| + \lambda_t^2\|\nabla f(\mathbf{y}_t) - \nabla f(\mathbf{x}_t)\|^2\right]. \tag{6}$$

Question 31: (*2 points*) Conclude by providing the convergence rate of the function value.

□0 □1 ■2

Solution: We use that $\lambda_{t+1}^2 - \lambda_{t+1} = \lambda_t^2$ can also be written as $\lambda_{t+1} = \frac{1}{2} + \sqrt{\lambda_t^2 + \frac{1}{4}}$. Then, $\lambda_{t+1} \ge \frac{1}{2} + \lambda_t$, hence $\lambda_t \ge \frac{t}{2}$.

Question 32: (*2 points*) You have obtained from Question 30 that $V_{t+1} - V_t \le -\Delta_t$ for a specific non-negative Δ_t to specify. Prove that the sequence $V_t + \sum_{s=0}^{t-1} \Delta_s$ is non-increasing.

□0 □1 ■2

Solution: We compute the difference

$$\left(V_{t+1} + \sum_{s=0}^{t} \Delta_s\right) - \left(V_t + \sum_{s=0}^{t-1} \Delta_s\right) = V_{t+1} - V_t + \Delta_t \le 0.$$

Question 33: (*3 points*) Conclude on a convergence guarantee of the smallest observed gradient norm.

□0 □1 □2 ■3

Solution: We have that for any $t \ge 0$,

$$\frac{1}{2L}\sum_{s=0}^{t-1} \lambda_{s+1}^2 \|\nabla f(x_{s+1})\|^2 \le V_t + \sum_{s=0}^{t-1}\Delta_s \le V_0 = \frac{L}{2}\|x_0 - \mathbf{x}^\star\|^2. \tag{7}$$

We conclude that

$$\min_{0 \le s \le t-1} \|\nabla f(x_{s+1})\|^2 \le \frac{L^2\|x_0 - \mathbf{x}^\star\|^2}{\sum_{s=0}^{t-1}\lambda_{s+1}^2} = O\left(\frac{1}{t^3}\right), \tag{8}$$

where we used that $\lambda_{t+1}^2 - \lambda_{t+1} = \lambda_t^2$ can also be written as $\lambda_{t+1} = \frac{1}{2} + \sqrt{\lambda_t^2 + \frac{1}{4}}$. Then, $\lambda_{t+1} \ge \frac{1}{2} + \lambda_t$, hence $\lambda_t \ge \frac{t}{2}$.