+1/1/60+

EPFL

1

Profs. Martin Jaggi and Nicolas Flammarion
Optimization for Machine Learning – CS-439 - IC
03.07.2023 from 15h15 to 18h15
Duration : 180 minutes

Student One

SCIPER : **111111**

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.
- You each have a different exam.
- For technical reasons, **do use black or blue pens for the MCQ part, no pencils!** Use white corrector if necessary.

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		

For your examination, preferably print documents compiled from auto-multiple-choice.

First part, multiple choice

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

Convexity

Question 1 Let $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}$ be three functions such that $f = h \circ g$, i.e., $f(\mathbf{x}) = h(g(\mathbf{x}))$ for all $\mathbf{x} \in \mathbb{R}^n$.
What are all the true statements?

A) f is concave provided that h is concave and non-decreasing, and g is concave.

B) f is concave provided that h is concave and non-decreasing, and g is convex.

C) f is concave provided that h is concave and non-increasing, and g is concave.

D) f is concave provided that h is concave and non-increasing, and g is convex.

F) f is concave provided that h is convex and non-decreasing, and g is concave.

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

Smoothness and gradient descent

Question 2 Let $f : \mathbb{R} \to \mathbb{R}$ be a convex, differentiable, and L-smooth function. Let G_f be the update function when running gradient descent with learning rate γ so that $x_{t+1} = G_f(x_t)$. Which statement is true for any such function f?

- ☐ G_f is injective if $\frac{1}{2L} < \gamma \leq \frac{1}{L}$.
- ☐ Regardless of how γ is chosen, G_f is always injective.
- ■ G_f is injective if $\gamma \leq \frac{1}{2L}$.
- ☐ Regardless of how γ is chosen, there always exists a function f for which G_f is not injective.

Solution: If $\gamma = \frac{1}{L}$ for $f = \frac{L}{2}x^2$, G_f is not injective since it is equal to zero for all $x \in \mathbb{R}$. If $\gamma \leq \frac{1}{2L}$, $G_f(x) = G_f(y)$ means $x - y = \gamma(\nabla f(x) - \nabla f(y))$. Given the smoothness and convexity we have $\|x - y\| \leq \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\| \leq \frac{1}{2} \|x - y\|$ which can only hold if $\|x - y\| = 0$.

Question 3 Let $f : \mathbb{R} \to \mathbb{R}$ be a function. Assume the sequence $x_0, x_1, \ldots$ exists such that $x_{t-1} := x_t - \gamma \nabla f(x_t)$, i.e. it is the **reverse** of a sequence generated by running gradient descent over f from some starting point with learning rate γ. Assuming $\lim_{t\to\infty} x_t \to \infty$, and that x_0 is a global minimum of f (so the gradient descent converges), which statement is true for any such function f?

- [x] For each of the other statements there is at least one function f for which the statement does not hold.
- [ ] f is either globally smooth for some value L or satisfies Polyak-Lojasiewicz inequality.
- [ ] f satisfies the Polyak-Lojasiewicz inequality.
- [ ] f is $\frac{1}{\gamma}$-smooth.

Solution: Consider $f(x) := \int g(x)dx$ where $g(x) := \begin{cases} x^3 & |x| < 2 \\ 2x & 2k \le |x| < 2k+1 \\ x & 2k+1 \le |x| < 2k+2 \end{cases}$ for all integers $k > 1$. This function is not smooth and does not satisfy PL inequality but gradient descent with a learning rate smaller than $\frac{1}{8}$ converges on it to 0.

Question 4 Which of the following statements is true?

- [ ] $f(x) := x^4$ is 2-smooth.
- [ ] $f(x) := x^4$ is 4-smooth.
- [x] $f(x) := x^2$ is 10-smooth.
- [ ] $f(x) := x^2$ is 1-smooth.

Question 5 Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be L-smooth functions. Which of the following is true about $h(x) := f(g(x))$?

- [x] h may not be smooth.
- [ ] h is never L'-smooth for any $L' < L$.
- [ ] h is always L^2 smooth.
- [ ] h is always L smooth.

Solution: Let $f(x)$ and $g(x)$ both be x^2 which is 2-smooth. The composition is the function x^4 which is not smooth. Also note that $f(x) = x$ and $g(x) = x$ are both 2-smooth (they are also 1 smooth but the assumptions of the problem hold also for $L = 2$). In this case $h(x) = x$ is 1 smooth so h can be smooth for smoothness parameter less than L.

Question 6 Let $f_1, \ldots, f_n$ be functions that are smooth with $L_1, \ldots, L_n$, respectively. Which of the following statements on functions $g = \sum_{i=1}^n \lambda_i f_i$ and $h = \max_i \lambda_i f_i$ for $\lambda_1, \ldots, \lambda_n \in \mathbb{R}^+$ hold generally?

A) g is $\sum_{i=1}^n \lambda_i L_i$ smooth.

B) h is $\sum_{i=1}^n \lambda_i L_i$ smooth.

C) g is $\max_i \lambda_i L_i$ smooth.

D) h is $\max_i \lambda_i L_i$ smooth.

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

Solution: h is not necessarily differentiable and not smooth. (A) is one of the basic lemmas in the course.

Proximal Gradient Descent

Question 7 Assume that the proximal operator is readily available for a convex function $f : \mathbb{R}^d \to \mathbb{R}$ and *any* $\gamma > 0$. Recall that the proximal operator is defined by

$$\text{prox}_{f,\gamma}(\mathbf{v}) = \underset{\mathbf{x}\in\mathbb{R}^d}{\text{argmin}}\, f(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{v}\|^2.$$

For which of the following transformations of f, do we have the proximal operator readily available (with a single call to the proximal oracle of f and any γ)?

A) $g(\mathbf{x}) := af(\mathbf{x}) + b$, for $a, b \in \mathbb{R}$.

B) $g(\mathbf{x}) := f(\mathbf{x}) + \lambda\|\mathbf{x}\|^2$, for $\lambda > 0$.

C) $g(\mathbf{x}) := f(\mathbf{x}) + f(2\mathbf{x})$.

D) $g(\mathbf{x}) := f(\mathbf{x} \odot \mathbf{x})$, where $\mathbf{x} \odot \mathbf{x}$ denotes the coordinate wise multiplication.

- ☐ B, C, D
- ☐ B, C
- ☐ C, D
- ■ A, B
- ☐ A, B, D

Subgradient Descent

Question 8 Which class of functions from $\mathbb{R}$ to $\mathbb{R}$ always have at least one subgradient (i.e., there exist a point x, $\partial f(x) \neq \emptyset$)?

- ☐ Lipschitz continuous
- ■ Convex
- ☐ Bounded
- ☐ Bounded and Lipschitz continuous

Solution: Convex functions have subgradients everywhere. For bounded and Lipschitz continuous, here is a simple counterexample $f(x) = \min\{1, e^{-x}\}$.

Frank-Wolfe

Question 9 For any convex and bounded region $\mathcal{X} \subset \mathbb{R}^d$ and any vector $\mathbf{g} \in \mathbb{R}^d$, define the LMO oracle,

$$\mathrm{LMO}_{\mathcal{X}}(\mathbf{g}) = \operatorname*{argmin}_{\mathbf{z} \in \mathcal{X}} \mathbf{g}^\top \mathbf{z}.$$

Which of the following statements are true ?

A) The $LMO_{\mathcal{X}}(\mathbf{g})$ is unique for any vector $\mathbf{g}$.

B) When $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_1 \leq 1\}$, the computational complexity of evaluating the LMO is less than the projection on the set $\mathcal{X}$.

C) For any $\mathbf{y} \in \mathcal{X}$ and $\lambda > 0$, the linear combination $(1-\lambda)\mathbf{y} + \lambda\,\mathrm{LMO}_{\mathcal{X}}(\mathbf{g})$ stays in the feasible region $\mathcal{X}$.

- [ ] A and C
- [x] B
- [ ] C
- [ ] A and B
- [ ] B and C
- [ ] A

Newton's Method

Question 10 Consider the following strictly convex function

$$f(x) = \sqrt{1 + x^2}$$

Definition: we say that a sequence $(\mathbf{x}_n)$ converges to its limit l at γ-superlinear rate (for a $\gamma > 1$) if γ is the biggest constant such that there exists a constant C such that for all n we have $\|\mathbf{x}_{n+1} - l\| \leq C|\mathbf{x}_n - l\|^\gamma$.

Which of the following statements is **not** true?

A) If $|x_0| < 1$, Newton's method converges with a quadratic superlinear rate ($\gamma = 2$).

B) If $|x_0| < 1$, Newton's method converges with a cubic superlinear rate ($\gamma = 3$).

C) If $|x_0| = 1$, Newton's method oscillates between -1 and 1.

D) If $|x_0| > 1$, Newton's method diverges.

- [x] A
- [ ] B
- [ ] C
- [ ] D

Second part, true/false questions

Question 11 (Subgradients) The number of subgradients $|\partial f(x)|$ for any point $x \in \mathrm{dom}(f)$ and function f is either $0, 1$ or ∞.

■ TRUE ☐ FALSE

Solution: Observe that if there are two distinct subgradients, there are an infinite number of them by linear interpolation.

Question 12 (Subgradients) Consider the function $f_\alpha : \mathbb{R} \to \mathbb{R}$:

$$f_\alpha(x) = \begin{cases} x^\alpha & \text{if } x > 0 \\ 0 & \text{if } x \leq 0. \end{cases}$$

Then for any $\alpha > 1$, f_α is differentiable everywhere.

■ TRUE ☐ FALSE

Question 13 (Subgradients) Again consider the same function $f_\alpha : \mathbb{R} \to \mathbb{R}$:

$$f_\alpha(x) = \begin{cases} x^\alpha & \text{if } x > 0 \\ 0 & \text{if } x \leq 0. \end{cases}$$

Then for any real number $\alpha > 0$, f_α has subgradients everywhere.

☐ TRUE ■ FALSE

Question 14 (Smoothness) Gradient descent with stepsize $\gamma = \frac{1}{L}$ for a smooth function (L being the smoothness constant) converges to a critical point of the function.

☐ TRUE ■ FALSE

Solution: Consider the counterexample e^{-x} which is given in lecture notes.

Question 15 (Convex functions) Consider a function f with a domain that is not necessarily convex. If f has subgradients everywhere then it is necessarily convex.

☐ TRUE ■ FALSE

Question 16 (Semi-norms) A semi-norm is a non-negative function that satisfies i) for any $\alpha \in \mathbb{R}$ $f(\alpha \mathbf{x}) = |\alpha| f(\mathbf{x})$, ii) the triangular inequality $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$.
All semi-norms are convex.

■ TRUE ☐ FALSE

Question 17 (Convex functions) If $f_1, \ldots, f_m$ are all convex functions $\mathbb{R}^d \to \mathbb{R}$, then f defined as $f(\mathbf{x}) := \min(f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ is also convex.

☐ TRUE ■ FALSE

Question 18 (Convex minimum) A convex function has always a minimum.

☐ TRUE ■ FALSE

Question 19 (Strong convexity) A strongly convex function has always exactly one minimizer.

■ TRUE ☐ FALSE

Question 20 (Convex sets) The empty set $\emptyset$ is convex.

■ TRUE ☐ FALSE

Question 21 (Newton) Newton's method always converges faster than gradient descent.

☐ TRUE ■ FALSE

Question 22 (Stochastic Gradient Descent) If the starting point $\mathbf{x}_0$ is chosen too far from the convergence point $\mathbf{x}^\star$, i.e. $\|\mathbf{x}_0 - \mathbf{x}^\star\|$ is larger than a certain threshold R_0, it is not possible to choose a learning rate such that SGD converges in expectation even if the function is convex and differentiable, and the expected norm of the stochastic gradient is bounded, i.e. $\mathbb{E}[\|\mathbf{g}_t\|^2] \leq B^2$.

☐ TRUE ■ FALSE

Question 23 (Projected Gradient Descent) Computing the projection of any vector $\mathbf{z} \in \mathbb{R}^d$ on an Euclidean ball $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq 1\}$ is a $\mathcal{O}(1)$ operation.

☐ TRUE ■ FALSE

Question 24 (Proximal Gradient Descent) For any two convex functions f and g, the proximal operator is additive, i.e., $\text{prox}_{f+g,\gamma} = \text{prox}_{f,\gamma} + \text{prox}_{g,\gamma}$.

☐ 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.

Convex smooth functions

Until the end of this section, we assume that the function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth. We let $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.

Question 25: *2 points.* For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, define the function $g_{\mathbf{x}}(\mathbf{y}) := f(\mathbf{y}) - \nabla f(\mathbf{x})^\top \mathbf{y}$. Show that $g_{\mathbf{x}}$ is convex and has a global minimum at $\mathbf{x}$.

☐$_0$ ☐$_1$ ■$_2$

Solution: We want to show that g lies above its linearization. We use that $\nabla g_{\mathbf{x}}(\mathbf{y}) = \nabla f(\mathbf{y}) - \nabla f(\mathbf{x})$.

$$\begin{aligned} g_{\mathbf{x}}(\mathbf{y}) + \nabla g_{\mathbf{x}}(\mathbf{y})^\top(\mathbf{z}-\mathbf{y}) &= f(\mathbf{y}) - \nabla f(\mathbf{x})^\top \mathbf{y} + \left[\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\right]^\top (\mathbf{z}-\mathbf{y}) \\ &= f(\mathbf{y}) + \nabla f(\mathbf{y})^\top(\mathbf{z}-\mathbf{y}) - \nabla f(\mathbf{x})^\top \mathbf{z} \\ &\leq f(\mathbf{z}) - \nabla f(\mathbf{x})^\top \mathbf{z} \\ &= g_{\mathbf{x}}(\mathbf{z}) \end{aligned}$$

where the last inequality is by convexity of f. For optimality, we see that $\mathbf{y} := \mathbf{x}$ is a point of zero gradient for $g_{\mathbf{x}}$, which implies optimality.

Question 26: *2 points.* Prove that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$$

☐$_0$ ☐$_1$ ■$_2$

Solution: Use the minimum established in the previous result, i.e., $g_{\mathbf{x}}(\mathbf{y} - \frac{1}{L}\nabla g_{\mathbf{x}}(\mathbf{y})) \geq g_{\mathbf{x}}(\mathbf{x})$.

Question 27: *2 points.* Show that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have,

$$\frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq \big[\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\big]^\top (\mathbf{x} - \mathbf{y}).$$

This property is usually referred to as *co-coercivity*.

☐0 ☐1 ■2

Solution: Use the fact that the previous inequality is valid if we exchange $\mathbf{x}$ and $\mathbf{y}$, and combine both.

Gradient descent

Now consider the gradient descent algorithm on the function f :

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t) \qquad \forall\, t \geq 0\,.$$

Question 28: *2 points.* Show that for $\gamma \leq \frac{2}{L}$, the sequence $(\|\mathbf{x}_t - \mathbf{x}^\star\|^2)_{t\geq 0}$ is non-increasing.

☐0 ☐1 ■2

Solution: Use co-coercivity.
Define $\Delta_t := f(\mathbf{x}_t) - f(\mathbf{x}^\star)$.

Question 29: *3 points.* Show that there exists a constant $\alpha > 0$ such that

$$\Delta_t \leq \Delta_{t-1} - \alpha \Delta_{t-1}^2\,.$$

Further show that the above inequality implies

$$\frac{1}{\Delta_{t-1}} \leq \frac{1}{\Delta_t} - \alpha\,.$$

Hint: Use convexity and Cauchy-Schwarz.

☐0 ☐1 ☐2 ■3

Solution: Add the inequality from the previous question twice, once the original and once with $\mathbf{x}$ and $\mathbf{y}$ swapped.

Question 30: *1 point.* Deduce that

$$\Delta_t \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{t+4}\,.$$

☐0 ■1

Solution:

Non-convex optimization

A differentiable function f with a convex $\mathrm{dom}(f)$, is called *invex* with respect to η that is defined over $\mathrm{dom}(\eta) = \mathrm{dom}(f) \times \mathrm{dom}(f)$, if for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$,

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \eta(\mathbf{x}, \mathbf{y})^\top \nabla f(\mathbf{y}). \tag{IVX}$$

If there exist a η such that f is invex with respect to η, f is called *invex*. In the following until the end of this section, assume that the function f is differentiable and $\mathrm{dom}(f)$ is open and convex.

Basics of Invexity

Question 31: *1 point.* Prove that f is invex if it does not have any critical points. *Hint: construct a* η.

☐$_0$ ■$_1$

Solution: Set $\eta(\mathbf{x}, \mathbf{y}) = (f(\mathbf{x}) - f(\mathbf{y})) \frac{\nabla f(\mathbf{y})}{\|\nabla f(\mathbf{y})\|^2}$

Question 32: *2 points.* Prove that f is invex if and only if all of its critical points are global minimum. *Hint: use the previous question for non-critical points.*

☐$_0$ ☐$_1$ ■$_2$

Solution: Set $\eta(\mathbf{x}, \mathbf{y}) = 0$ for critical points.

Rate of convergence

In this subsection, assume that f is invex with respect to function η,

$$\eta(\mathbf{x}, \mathbf{y}) = c_0(\mathbf{x} - \mathbf{y}) + \eta_0(\mathbf{x}, \mathbf{y}),$$

defined through constant $c_0 \geq 0 \in \mathbb{R}$ and function η_0 that is defined over $\text{dom}(f) \times \text{dom}(f)$. In addition, assume that η_0 satisfy the following condition for a fixed constant $N \geq 0$,

$$\|\eta_0(\mathbf{x}, \mathbf{y})\| \leq N \left(\|\nabla f(\mathbf{x})\| + \|\nabla f(\mathbf{y})\|\right), \quad \text{for all} \quad \mathbf{x}, \mathbf{y} \in \text{dom}(f).$$

Recall that when $c_0 = 1$ and $\eta_0 = 0$, the invexity condition given in (IVX) is equivalent to convexity. And, in the convex case, vanilla analysis of gradient descent with constant step size γ and timestep T yields the following bound

$$\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \tag{VB}$$

where x_0 is the initial point and $x_1, \ldots, x_{T-1}$ are iterates obtained by running gradient descent. Our aim in this subsection is to extend this analysis to a more general class of invex functions.

Question 33: *2 points.* Derive the following bound that is reminiscent of (VB) for f,

$$\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{c_0\gamma}{2} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{c_0}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \sum_{t=0}^{T-1} \eta_0(\mathbf{x}^\star, \mathbf{x}_t)^\top \nabla f(\mathbf{x}_t).$$

Hint: treat two terms in η separately.

☐$_0$ ☐$_1$ ■$_2$

Solution: Scale the vanilla analysis by c_0 and sum error terms by η_0:

$$\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}_\star)) \leq \frac{c_0\gamma}{2} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{c_0}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \sum_{t=0}^{T-1} \eta_0(\mathbf{x}^\star, \mathbf{x}_t)^\top \nabla f(\mathbf{x}_t).$$

Question 34: *1 point.* Assume that f is B-Lipschitz and $R = \|x_0 - x^\star\|$ is finite. Give a bound on

$$E_T = \sum_{i=0}^{T-1} f(\mathbf{x}_i) - f(\mathbf{x}^\star)$$

that depends on B, R, N, T, c_0 and γ.

☐$_0$ ■$_1$

Solution: Under the assumption, we have the following bound:

$$\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}_\star)) \leq \frac{c_0\gamma}{2} B^2 T + \frac{c_0}{2\gamma} R^2 + 2TNB^2.$$

Question 35: *1 point.* Does E_T/T convergence to zero as $T \to \infty$ with a proper choice of γ?

☐$_0$ ■$_1$

Solution: The last term linearly scales and NO.

Question 36: *3 points.* Assume now that f is L-smooth, $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R_1$ and $f(\mathbf{x}_0) - f(\mathbf{x}^\star) \leq R_2$ for positive constants L, R_1 and R_2. Also, let $\gamma = \frac{1}{L}$ be the fixed stepsize. Give a bound on

$$E_T = f(\mathbf{x}_T) - f(\mathbf{x}^\star)$$

that depends on L, R, N, T, c_0 and γ.

☐$_0$ ☐$_1$ ☐$_2$ ■$_3$

Solution: From smoothness, we have

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

Therefore, vanilla bound reads

$$\sum_{t=0}^{T-1}\left(f(\mathbf{x}_t) - f(\mathbf{x}_\star)\right) \leq c_0\left(f(\mathbf{x}_0) - f(\mathbf{x}_T)\right) + \frac{Lc_0}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + 2LN\left(f(\mathbf{x}_0) - f(\mathbf{x}_T)\right),$$

where we have applied Cauchy-Schwarz,

$$\eta_0(\mathbf{x}^\star, \mathbf{x}_t)^\top \nabla f(\mathbf{x}_t) \leq \|\eta_0(\mathbf{x}^\star, \mathbf{x}_t)\|\|\nabla f(\mathbf{x}_t)\| \leq N\|\nabla f(\mathbf{x}_t)\|^2.$$

By using that function value is always decreasing,

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{c_0}{T}R_2 + \frac{Lc_0}{2T}R_1^2 + \frac{2LN}{T}R_2.$$

Question 37: *3 points.* Repeat the previous question for η_0 defined over $\mathrm{dom}(f) \times \mathrm{dom}(f)$ that satisfy the following uniform bound,

$$\|\eta_0(\mathbf{x}, \mathbf{y})\| \leq N, \quad \text{for all} \quad \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f).$$

Hint: Cauchy-Schwarz is your friend.

☐0 ☐1 ☐2 ■3

Solution: From smoothness, vanilla bound reads

$$\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}_\star)) \leq c_0 (f(\mathbf{x}_0) - f(\mathbf{x}_T)) + \frac{Lc_0}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \sum_{t=0}^{T-1} \|\eta_0(\mathbf{x}^\star, \mathbf{x}_t)\|\|\nabla f(\mathbf{x}_t)\|.$$

By applying Cauchy-Schwarz,

$$\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\| \leq \sqrt{T}\sqrt{\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2},$$

which leads to

$$\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}_\star)) \leq c_0 (f(\mathbf{x}_0) - f(\mathbf{x}_T)) + \frac{Lc_0}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \sqrt{2TL}\sqrt{f(\mathbf{x}_0) - f(\mathbf{x}_T)}N.$$

By using that function value is always decreasing,

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{c_0}{T}R_2 + \frac{Lc_0}{2T}R_1^2 + \sqrt{\frac{2L}{T}}\sqrt{R_2}N.$$

Question 38: *1 point.* Derive the order of steps required to achieve a small error ε, i.e., $E_T \leq \varepsilon$, for the two previous questions.
Hint: Focus only on the highest order term for the latter.

☐0 ■1

Solution: $\mathcal{O}(1/\varepsilon)$ and $\mathcal{O}(1/\varepsilon^2)$ steps.