ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Department of Computer Science
Institute of Theoretical Computer Science
Bernd Gärtner and Niao He

Optimization for Data Science — Graded Assignment 4 — FS23

- The solution is due on **June 30, 2023** by **11:59 pm**. Please submit your solution as a PDF on Moodle. The name of the file should follow the format `GA3-{Legi number}`, e.g., `GA3-19-123-456`. After uploading your solution, please make sure that the status is "Submitted for grading". You should receive an automatic email that confirms your submission. Please notify us if you don't receive this.
- If you want to submit your solution within six hours before the deadline and a technical problem prevents you from submitting it on Moodle, you can send your solution as PDF to `saeed.ilchi@inf.ethz.ch`. The same submission deadline still applies. If you encounter any trouble with the submission process, complain timely.
- Please solve the exercises carefully and typeset your solution using LaTeX or a similar typesetting program. A tutorial can be found at `http://www.cadmo.ethz.ch/education/thesis/latex`. Handwritten solutions will not be graded! The same applies to solutions written with any kind of tablet device and stylus, etc.
- For geometric drawings that can easily be integrated into LaTeX documents, we recommend the drawing editor IPE, which you can find at `http://ipe7.sourceforge.net/`.
- Keep in mind the following premises:
  - When writing in English, write short and simple sentences.
  - When writing a proof, write precise statements.
- This is a theory course, which means: if an exercise does not explicitly say "you do not need to prove your answer" or "justify intuitively", then a formal proof is **always** required. You can of course refer in your solutions to the lecture notes and to the exercises, if a result you need has already been proved there.
- We would like to stress that the ETH Disciplinary Code applies to this Graded Assignment as it constitutes part of your final grade. The only exception we make to the Code is that we encourage you to verbally discuss the tasks with your colleagues. It is strictly prohibited to share any (hand-)written or electronic (partial) solutions with any of your colleagues. We are obliged to inform the Rector of any violations of the Code.
- As with all exercises, the material of the graded assignments is relevant for the exam.

Min-Max for Smooth Functions (20 points)

Consider the optimization problem

$$f^* = \min_{x \in \mathbb{R}^{20}, \|x\|_2 \leqslant 1} f(x)$$
$$f(x) = \max_{1 \leqslant i \leqslant 10} f_i(x)$$

where each $f_i : \mathbb{R}^{20} \to \mathbb{R}$ is a 1-smooth lipschitz convex function. Moreover, we know that $\|\nabla f_i(x)\| \leqslant 1$ for all x in the 20-dimensional unit ball and for all $1 \leqslant i \leqslant 10$. Design an algorithm that computes a value $\hat{f}$ such that $\hat{f} - f^* < \varepsilon$. Your algorithm should evaluate the gradient of each of $f_i(\cdot)$s for at most $O(1/\varepsilon)$ many times.

Solution: Let $\phi(x, y) = \sum_{i \in [n]} y_i \cdot f_i(x)$, and let $\mathcal{D} = \{y \in \mathbb{R}^d \mid \sum_{i \in [n]} y_i = 1, y_i \geqslant 0\}$. So ϕ is a convex-concave function a saddle point for ϕ has value f^*.

So we are in the C-C setting, and if we run EG with stepsize $1/2L$ for T many iterations, from Theorem 13.6, we can get a point with a duality gap

$$\frac{(1+1)L}{4T} = \frac{L}{2T}.$$

We conclude the proof by setting T to $2L/\varepsilon$. In the following we show that $L = O(1)$ which concludes the proof.

We verify $O(1)$-smoothness of the function $\phi(x, y) = \sum_{i \in [n]} y_i \cdot f_i(x)$. First, note that:

$$\begin{aligned}
&\|\nabla_x \phi(x^{(1)}, y^{(1)}) - \nabla_x \phi(x^{(2)}, y^{(2)})\| \\
&= \|\sum_{i=1}^{10} \left(y_i^{(1)} \nabla f_i(x^{(1)}) - y_i^{(2)} \nabla f_i(x^{(2)}) \right)\| \\
&= \|\sum_{i=1}^{10} \left(y_i^{(1)} \nabla f_i(x^{(1)}) - y_i^{(2)} \nabla f_i(x^{(1)}) + y_i^{(2)} \nabla f_i(x^{(1)}) - y_i^{(2)} \nabla f_i(x^{(2)}) \right)\| \\
&\leqslant \sum_{i=1}^{10} \|\left(y_i^{(1)} - y_i^{(2)} \right) \cdot \nabla f_i(x^{(1)})\| + \sum_{i=1}^{10} |y_i^{(2)}| \cdot \|\nabla f_i(x^{(1)}) - \nabla f_i(x^{(2)})\| \\
&\leqslant \|y^{(1)} - y^{(2)}\|_1 + \cdot \|x^{(1)} - x^{(2)}\| \\
&\leqslant \sqrt{10} \|y^{(1)} - y^{(2)}\|_2 + \|x^{(1)} - x^{(2)}\| \\
&= O(1) \cdot (\|y^{(1)} - y^{(2)}\| + \|x^{(1)} - x^{(2)}\|)
\end{aligned}$$

In the second inequality, we use bounded gradient of f_i for the first term and 1-gradient lipschitzness of f_i for the second term. In the third inequality, we use the general fact that for any $z \in \mathbb{R}^d$, we have $\|z\|_1 \leqslant \sqrt{d}\|z\|_2$.

Moreover, we have:

$$\|\nabla_y \phi(x^{(1)}, y^{(1)}) - \nabla_y \phi(x^{(2)}, y^{(2)})\| = \|\left[f_1(x^{(1)}) - f_1(x^{(1)}), \ldots, f_{10}(x^{(1)}) - f_{10}(x^{(2)}) \right]^T\|$$

$$
\begin{aligned}
&= \sqrt{\sum_{i=1}^{10} \left(f_i(\mathbf{x}^{(1)}) - f_i(\mathbf{x}^{(2)})\right)^2} \\
&\leqslant \sqrt{10\|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|^2} \\
&= O(1) \cdot \|\mathbf{x}^{(1)} - \mathbf{x}^{(2)}\|
\end{aligned}
$$

The inequality holds because f_i is 1-Lipschitz. □

Stochastic Gradient Descent (40 points)

Algorithm 1 SGD

Input: $\mathbf{x}_0 \in \mathbb{R}^d$
for $t = 0, 1, 2, \ldots T-1$ **do**
  sample new ξ_t from a distribution $P(\xi)$
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, \xi_t)$
end for

Consider an unconstrained problem $\min_x F(\mathbf{x}) := \mathbb{E}_\xi[f(\mathbf{x}, \xi)]$, where ξ follow a distribution $P(\xi)$. SGD is presented in Algorithm 1.

Question 1: Prefix the number of iteration T, $L > 0$, $\Delta > 0$ and stepsize sequence $\{\gamma_t\}_{t=0}^{T-1}$. Consider a function $F : \mathbb{R} \to \mathbb{R}$ defined as follows:

$$
F(x) = \frac{x^2}{2\max\left\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\right\}}.
$$

We pick the initial point $x_0 = \sqrt{2\Delta \max\left\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\right\}}$. Show that F is L-smooth and $F(x_0) - \min_x F(x) \leqslant \Delta$.

Question 2: Consider the function in Question 1 and Algorithm 1 with noiseless gradients, i.e., $\nabla f(x, \xi) = \nabla F(x)$ for all x and ξ. Show that for all $0 \leqslant t \leqslant T$, we have $x_t \geqslant x_0/2$. This implies

$$
|\nabla F(x_t)| \geqslant \sqrt{\frac{\Delta}{2\max\left\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\right\}}}.
$$

Question 3: Consider another function $F : \mathbb{R}^d \to \mathbb{R}$ defined as follows:

$$
F(\mathbf{x}) = \frac{L}{2}\|\mathbf{x}\|^2.
$$

We pick an initial point x_0 such that $\|x_0\| = \sqrt{\Delta/L}$. Consider Algorithm 1 with $\nabla f(x, \xi) = \nabla F(x) + \xi$, where ξ is sampled from d-dimensional Gaussian distribution $N(0, \frac{\sigma^2}{d} I_d)$ with $\sigma > 0$ and I_d being the identity matrix. Show that for $t \geqslant 2$

$$x_t = \prod_{j=0}^{t-1} (1 - L\gamma_j)\, x_0 - \sum_{j=0}^{t-2} \gamma_j \prod_{i=j+1}^{t-1} (1 - L\gamma_i)\, \xi_j - \gamma_{t-1}\xi_{t-1}.$$

Question 4: Fix $\delta \in (0, 1)$. Show that with dimension $d \geqslant d_0 = \mathcal{O}(\log(T/\delta))$, for any $2 \leqslant t \leqslant T$, we have

$$\|\nabla F(x_t)\|^2 \geqslant \frac{L}{2} \left(\Delta \prod_{j=0}^{t-1} (1 - L\gamma_j)^2 + L\sigma^2 \sum_{j=0}^{t-2} \gamma_j^2 \prod_{i=j+1}^{t-1} (1 - L\gamma_i)^2 + L\sigma^2 \gamma_{t-1}^2 \right).$$

with probability at least $1 - \delta/T$.

Hint. You can use the following lemma directly:
For a d-dimensional normally distributed random vector $x \sim N\left(y, \frac{\eta}{d} I_d\right)$, where $y \in \mathbb{R}^d$, we have, for any $\bar{\delta} \in (0, 1)$,

$$\Pr\left(\left| \frac{\|x\|^2}{\|y\|^2 + \eta} - 1 \right| \leqslant \bar{\delta} \right) \geqslant 1 - 4\exp\left(-\frac{d\bar{\delta}^2}{24} \right).$$

Question 5: Show that if $\gamma_t = \gamma \in (0, 1/L)$ and we choose the same d as last question, with probability at least $1 - \delta$, we have for all $2 \leqslant t \leqslant T$

$$\|\nabla F(x_t)\|^2 \geqslant \min\left\{ \frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2 - L\gamma)} \right\}.$$

Remark You do not need to prove the remark. Recall that in Lecture 11 we pick stepsize $\gamma_t = \Theta(T^{-1/2})$ in SGD to find near-stationary points. This question gives some explanations why we pick stepsize of this order by providing lower bounds.

- From Question 2, if we choose stepsize $\gamma_t = \frac{c}{(t+1)^\theta}$ or $\gamma_t = \frac{c}{T^\theta}$ with $\theta \in (0, 1)$, there exists some function such that $\|\nabla F(x_t)\| = \Omega(T^{(\theta-1)/2})$ for all t.
- From Question 5, if we choose stepsize $\gamma_t = \frac{1}{LT^\theta}$ with $\theta \in (0, 1)$, there exists some function such that $\|\nabla F(x_t)\| = \Omega(T^{-\theta/2})$ for all t with high probability.

Solution:

Question 1: First, we compute 对x进行求导，得到delta F(x)

$$\nabla F(x) = \frac{x}{\max\left\{1/L, 2\sum_{t=0}^{T-1} \gamma_t\right\}}.$$ 这个是有最大值和最小值的

It is easy to verify that F is ℓ-smooth with $\ell = 1/\max\left\{1/L, 2\sum_{t=0}^{T-1} \gamma_t\right\}$. Since $\ell \leqslant L$, it is also L-smooth. It is trivial to verify that $F(x_0) - \min_x F(x) = F(x_0) = \Delta$.

将x0带入

Question 2: We note that all stepizes are smaller than $1/\ell$, i.e., for all $0 \leqslant t \leqslant T-1$,

step 这个是经典的选择步长的方式

$$\gamma_t \leqslant 1/\ell = \max\left\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\right\}.$$

看算法更新的公式

Then by the update rule of gradient descent, it is easy to verify that $x_t \geqslant 0$ for all $0 \leqslant t \leqslant T$ and $x_{t+1} \leqslant x_t$ for all $0 \leqslant t \leqslant T-1$. This further implies that $\nabla F(x_{t+1}) \leqslant \nabla F(x_t)$. By the update rule,

$$x_{t+1} = x_t - \gamma_t \nabla F(x_t) \geqslant x_t - \gamma_t \nabla F(x_0),$$

which implies, for all $0 \leqslant t \leqslant T$,

$$x_t \geqslant x_0 - \sum_{i=0}^{t-1}\gamma_t \nabla F(x_0) = x_0 - \sum_{i=0}^{t-1}\gamma_t \frac{x_0}{\max\left\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\right\}} \geqslant \frac{x_0}{2}.$$

We reach the conclusion by noting $\nabla f(x_t) \geqslant \nabla f(x_0/2)$.

Question 3: By the update rule, applying the updating rules of it

$$x_{t+1} = x_t - \gamma_t(Lx + \xi_t) = (1 - L\gamma_t)x_t - \gamma_t \xi_t.$$

Recursing this, we have, for $t \geqslant 2$,

$$x_t = \prod_{j=0}^{t-1}(1 - L\gamma_j)\, x_0 - \sum_{j=0}^{t-2}\gamma_j \prod_{i=j+1}^{t-1}(1 - L\gamma_i)\, \xi_j - \gamma_{t-1}\xi_{t-1}.$$

Question 4: We note each ξ_i is sampled from Gaussian distribution $N(0, \frac{\sigma^2}{d} I_d)$ and they are independent. From last question, we know x_t is the sum of scaled x_0 and $\{\xi_i\}_i$, so it follows the Gaussian distribution $N(y_t, \frac{\eta_t}{d} I_d)$ with

$$y_t = \prod_{j=0}^{t-1}(1 - L\gamma_j)\, x_0 \text{ and } \eta_t = \sigma^2 \sum_{j=0}^{t-2}\gamma_j^2 \prod_{i=j+1}^{t-1}(1 - L\gamma_i)^2 + \sigma^2\gamma_{t-1}^2.$$

We directly apply the lemma in the hint with $\bar{\delta} = 1/2$:

$$\Pr\left(\|x_t\|^2 \geqslant \frac{\|y_t\|^2 + \eta_t}{2}\right) \geqslant \Pr\left(\left|\frac{\|x_t\|^2}{\|y_t\|^2 + \eta_t} - 1\right| \leqslant \frac{1}{2}\right) \geqslant 1 - 4\exp\left(-\frac{d}{96}\right).$$

By picking $d \geqslant 96 \log\frac{4T}{\delta}$ and noting that $\|\nabla F(x)\| = L^2\|x\|^2$, we have

$$\|\nabla F(x_t)\|^2 \geqslant \frac{L}{2}\left(\Delta \prod_{j=0}^{t-1}(1 - L\gamma_j)^2 + L\sigma^2 \sum_{j=0}^{t-2}\gamma_j^2 \prod_{i=j+1}^{t-1}(1 - L\gamma_i)^2 + L\sigma^2\gamma_{t-1}^2\right).$$

with probability at least $1 - \delta/T$.

Question 5: From last question, with $\gamma_t = \gamma \in (0, 1/L)$,

$$\begin{aligned}
\|\nabla F(x_t)\|^2 &\geqslant \frac{L}{2}\left(\Delta\,(1-L\gamma)^{2t} + L\sigma^2\gamma^2 \sum_{j=0}^{t-2} (1-L\gamma)^{2(t-j-1)} + L\sigma^2\gamma^2\right) \\
&\qquad \text{this step is not very clear.} \\
&= \frac{L}{2}\left(\Delta\,(1-L\gamma)^{2t} + L\sigma^2\gamma^2 \sum_{i=0}^{t-1} (1-L\gamma)^{2i}\right) \\
&= \frac{L}{2}\left(\Delta\,(1-L\gamma)^{2t} + L\sigma^2\gamma^2 \frac{1-(1-L\gamma)^{2t}}{1-(1-L\gamma)^2}\right) \\
&= \frac{L}{2}\left(\Delta\,(1-L\gamma)^{2t} + \frac{\sigma^2\gamma}{2-L\gamma}\left[1-(1-L\gamma)^{2t}\right]\right) \geqslant \frac{L}{2}\left(\min\left\{\Delta, \frac{\sigma^2\gamma}{2-L\gamma}\right\}\right).
\end{aligned}$$

Then we can use union bound for all $2 \leqslant t \leqslant T$, and we have the desired result.

□

Modified Extragradient (40 points)

We consider the following minimax optimization problem:

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} f(x, y),$$

where function f is smooth for both variables, $f(\cdot, y)$ is convex and $f(x, \cdot)$ is concave. During the course, we have learned about extragradient method, which in this setting achieves a primal-dual gap convergence rate of $\mathcal{O}\left(\frac{1}{T}\right)$ for averaged iterates, but only $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ for the last-iterate. In this exercise, we consider a "regularized" extragradient algorithm, which pushes the iterates towards the initial point and has a better last-iterate convergence guarantee. Denote $z = (x, y)$ and $F(z) = (\nabla_x f(x, y), -\nabla_y f(x, y))$. The updates for $t \geqslant 0$ are given by:

$$\begin{aligned}
z_{t+\frac{1}{2}} &= z_t - \eta F(z_t) + \frac{1}{t+1}(z_0 - z_t), \\
z_{t+1} &= z_t - \eta F\left(z_{t+\frac{1}{2}}\right) + \frac{1}{t+1}(z_0 - z_t),
\end{aligned}$$

where $\eta > 0$ is the stepsize.

Since f is convex-concave and smooth, we have the following properties for any $z, \hat{z} \in \mathbb{R}^{d_1+d_2}$:

- $\langle F(z) - F(\hat{z}), z - \hat{z}\rangle \geqslant 0$,
- $\|F(z) - F(\hat{z})\| \leqslant L\|z - \hat{z}\|$ for some $L > 0$.

You can directly use these properties above. Throughout the exercise, we use a constant stepsize $\eta < \frac{1}{\sqrt{3}L}$, and assume the existence of $z^* = (x^*, y^*)$ such that $F(z^*) = \mathbf{0}$. Our analysis will be based on the potential function:

$$V_t = \frac{t(t+1)}{2}\|\eta F(z_t)\|^2 + t\left\langle \eta F(z_t), z_t - z_0 \right\rangle.$$

Question 1: Define

$$A_t := \left\langle F(z_{t+1}) - F(z_t), \frac{1}{t+1}(z_0 - z_t) - \eta F\left(z_{t+\frac{1}{2}}\right) \right\rangle;$$
$$B_t := \left\|\eta F(z_t) - \eta F\left(z_{t+\frac{1}{2}}\right)\right\|^2 - \frac{1}{L^2}\left\|F\left(z_{t+\frac{1}{2}}\right) - F(z_{t+1})\right\|^2.$$

Show that for any $t \geqslant 0$, A_t and B_t are non-negative.

Question 2: Show that for any $t \geqslant 1$, it holds that

$$V_{t+1} - V_t \leqslant \frac{(t+1)\eta^2 L^2}{2t\left(1-\eta^2 L^2\right)}\|\eta F(z_{t+1})\|^2.$$

hint: You can start by the inequality

$$V_{t+1} - V_t \leqslant V_{t+1} - V_t + \eta t(t+1)A_t + \frac{t(t+1)}{2}B_t,$$

where A_t and B_t are the quantities defined in **Question 1** and are non-negative.

Question 3: By the recursion of V_t, show that for $T \geqslant 2$,

$$\frac{T^2}{4}\|\eta F(z_T)\|^2 \leqslant (1+\eta L)^2\|z^* - z_0\|^2 + \sum_{t=2}^{T-1}\frac{\eta^2 L^2}{1-\eta^2 L^2}\|\eta F(z_t)\|^2.$$

You can use the fact that $V_1 \leqslant (2\eta L + \eta^2 L^2)\|z_0 - z^*\|^2$ without proof.

hint: You may need the inequality:

$$\langle a, b\rangle \geqslant -\frac{\lambda}{4}\|a\|^2 - \frac{1}{\lambda}\|b\|^2, \quad \forall \lambda > 0.$$

Question 4: Show by induction that for $T \geqslant 2$, we have

$$\|F(z_T)\|^2 \leqslant \frac{4(1+\eta L)^2}{\eta^2(1-3\eta^2 L^2)T^2}\|z^* - z_0\|^2.$$

Question 5: Let $\mathcal{X} := \mathcal{B}^{d_1}(x_T, \|z_0 - z^*\|)$ and $\mathcal{Y} := \mathcal{B}^{d_2}(y_T, \|z_0 - z^*\|)$, where $\mathcal{B}^d(c, R)$ denotes a ball in $\mathbb{R}^d$ with center c and radius R. Show that for $T \geqslant 2$, we have

$$\max_{y\in\mathcal{Y}} f(x_T, y) - \min_{x\in\mathcal{X}} f(x, y_T) \leqslant \frac{2\sqrt{2}(1+\eta L)}{\eta\sqrt{1-3\eta^2 L^2 T}}\|z^* - z_0\|^2.$$

Solution:

Question 1:

Since $f(x, y)$ is convex in x and concave in y, we have

$$\langle F(z_{t+1}) - F(z_t), z_{t+1} - z_t \rangle \geqslant 0.$$

Together with the update rule, which can be written as $z_{t+1} - z_t = -\eta F\left(z_{t+\frac{1}{2}}\right) + \frac{1}{t+1}(z_0 - z_t)$, we have

$$\left\langle F(z_{t+1}) - F(z_t), \frac{1}{t+1}(z_0 - z_t) - \eta F\left(z_{t+\frac{1}{2}}\right) \right\rangle \geqslant 0.$$

For the second inequality, note that from the update rule, we have $z_{t+1} - z_{t+\frac{1}{2}} = \eta F(z_t) - \eta F\left(z_{t+\frac{1}{2}}\right)$. Then by the Lipschitz property of the function, we get

$$\left\| F\left(z_{t+\frac{1}{2}}\right) - F(z_{t+1}) \right\|^2 \leqslant L^2 \left\| z_{t+\frac{1}{2}} - z_{t+1} \right\|^2 = L^2 \left\| \eta F(z_t) - \eta F\left(z_{t+\frac{1}{2}}\right) \right\|^2.$$

Question 2:

According to the hint, we have

$$\begin{aligned}
&V_{t+1} - V_t \\
\leqslant\ & V_{t+1} - V_t + t(t+1) \left\langle \eta F(z_{t+1}) - \eta F(z_t), \frac{1}{t+1}(z_0 - z_t) - \eta F\left(z_{t+\frac{1}{2}}\right) \right\rangle \\
& \frac{t(t+1)}{2} \left(\left\| \eta F(z_t) - \eta F\left(z_{t+\frac{1}{2}}\right) \right\|^2 - \frac{1}{L^2} \left\| F\left(z_{t+\frac{1}{2}}\right) - F(z_{t+1}) \right\|^2 \right).
\end{aligned}$$

Plugging in the definition of V_t and combining like terms, we get

$$\begin{aligned}
&\text{RHS} \\
\leqslant\ & \left(\frac{(t+2)(t+1)}{2} - \frac{t(t+1)}{2L^2\eta^2} \right) \|\eta F(z_{t+1})\|^2 + \left(\frac{(t)(t+1)}{2} - \frac{t(t+1)}{2L^2\eta^2} \right) \left\| \eta F\left(z_{t+\frac{1}{2}}\right) \right\|^2 \\
& + \left(\frac{t(t+1)}{L^2\eta^2} - t(t+1) \right) \left\langle \eta F(z_{t+1}), \eta F\left(z_{t+\frac{1}{2}}\right) \right\rangle \\
& + \underbrace{(t+1) \langle \eta F(z_{t+1}), z_{t+1} - z_0 \rangle + t \langle \eta F(z_{t+1}), z_0 - z_t \rangle}_{(A)}.
\end{aligned}$$

We proceed to look at (A):

$$\begin{aligned}
(A) &= (t+1) \langle \eta F(z_{t+1}), z_{t+1} - z_t \rangle - \langle \eta F(z_{t+1}), z_0 - z_t \rangle \\
&= (t+1) \left\langle \eta F(z_{t+1}), \frac{1}{t+1}(z_0 - z_t) - \eta F\left(z_{t+\frac{1}{2}}\right) \right\rangle - \langle \eta F(z_{t+1}), z_0 - z_t \rangle \\
&= -(t+1) \left\langle \eta F(z_{t+1}), \eta F\left(z_{t+\frac{1}{2}}\right) \right\rangle,
\end{aligned}$$

where the second equality comes from the update rule of the algorithm. Putting this term back, we then have

$$
\begin{aligned}
&V_{t+1} - V_t \\
&\leqslant \left(\frac{(t+2)(t+1)}{2} - \frac{t(t+1)}{2L^2\eta^2}\right)\|\eta F(z_{t+1})\|^2 + \left(\frac{(t)(t+1)}{2} - \frac{t(t+1)}{2L^2\eta^2}\right)\left\|\eta F\left(z_{t+\frac{1}{2}}\right)\right\|^2 \\
&\quad + \left(\frac{t(t+1)}{L^2\eta^2} - t(t+1) - (t+1)\right)\left\langle \eta F(z_{t+1}), \eta F\left(z_{t+\frac{1}{2}}\right)\right\rangle \\
&= \frac{(t+1)\eta^2L^2}{2t\left(1-\eta^2L^2\right)}\|\eta F(z_{t+1})\|^2 - \frac{t+1}{2\eta^2L^2}\left\|\frac{\left(\eta^2L^2-1\right)t + \eta^2L^2}{\sqrt{\left(1-\eta^2L^2\right)t}}\eta F(z_{t+1}) + \sqrt{\left(1-\eta^2L^2\right)t}\eta F\left(z_{t+\frac{1}{2}}\right)\right\|^2 \\
&\leqslant \frac{(t+1)\eta^2L^2}{2t\left(1-\eta^2L^2\right)}\|\eta F(z_{t+1})\|^2 .
\end{aligned}
$$

Question 3:

By the recursion we get from last question, for $t \geqslant 2$, we have

$$
V_T \leqslant V_1 + \sum_{t=2}^{T-1}\frac{(t+1)\eta^2L^2}{2t\left(1-\eta^2L^2\right)}\|\eta F(z_{t+1})\|^2 \leqslant V_1 + \sum_{t=2}^{T-1}\frac{\eta^2L^2}{1-\eta^2L^2}\|\eta F(z_{t+1})\|^2 .
$$

By the definition of V_T, we get

$$
\begin{aligned}
V_T &= \frac{T(T+1)}{2}\|\eta F(z_T)\|^2 + T\langle \eta F(z_T), z_T - z_0\rangle \\
&= \frac{T(T+1)}{2}\|\eta F(z_T)\|^2 + T\langle \eta F(z_T), z^* - z_0\rangle + T\langle \eta F(z_T), z_T - z^*\rangle ,
\end{aligned}
$$

which by the the monotonicity of F (since $f(x, y)$ is convex-concave), implies

$$
V_T \geqslant \frac{T(T+1)}{2}\|\eta F(z_T)\|^2 + T\langle \eta F(z_T), z^* - z_0\rangle .
$$

By $\langle a, b\rangle \geqslant -\frac{\lambda}{4}\|a\|^2 - \frac{1}{\lambda}\|b\|^2$ with $\lambda = T+1$, we have

$$
\begin{aligned}
V_T &\geqslant \frac{T(T+1)}{2}\|\eta F(z_T)\|^2 - \frac{T(T+1)}{4}\|\eta F(z_T)\|^2 - \frac{T}{T+1}\|z^* - z_0\|^2 \\
&\geqslant \frac{T(T+1)}{4}\|\eta F(z_T)\|^2 - \|z^* - z_0\|^2 .
\end{aligned}
$$

Combine the result with the upper bound for V_1, we get

$$
\frac{T(T+1)}{4}\|\eta F(z_T)\|^2 \leqslant (1+\eta L)^2\|z^* - z_0\|^2 + \sum_{t=2}^{T-1}\frac{\eta^2L^2}{1-\eta^2L^2}\|\eta F(z_{t+1})\|^2 .
$$

Note that on the RHS, we also have term $\mathcal{O}(\|\eta F(z_T)\|^2)$. Using that $\eta < \frac{1}{\sqrt{3}L}$, we have $\frac{\eta^2L^2}{1-\eta^2L^2} \leqslant \frac{1}{2}$, which is smaller to or equal than $\frac{T}{4}$ for $T \geqslant 2$. Then we have the desired result

$$\frac{T^2}{4}\|\eta F(z_T)\|^2 \leqslant (1+\eta L)^2\|z^* - z_0\|^2 + \sum_{t=2}^{T-1} \frac{\eta^2L^2}{1-\eta^2L^2}\|\eta F(z_t)\|^2.$$

Question 4:

We prove this by induction. When $T = 2$, we have

$$\|F(z_2)\|^2 \leqslant \frac{4(1+\eta L)^2}{\eta^2T^2}\|z^* - z_0\|^2 \leqslant \frac{4(1+\eta L)^2}{\eta^2(1-3\eta^2L^2)T^2}\|z^* - z_0\|^2,$$

where the last inequality holds because $1 - 3\eta^2L^2 > 1$ for $\eta < \frac{1}{\sqrt{3}L}$.

For $T \geqslant 3$, we have

$$\begin{aligned}\frac{T^2}{4}\|\eta F(z_T)\|^2 &\leqslant (1+\eta L)^2\|z^* - z_0\|^2 + \sum_{t=2}^{T-1} \frac{\eta^2L^2}{1-\eta^2L^2}\|\eta F(z_t)\|^2 \\ &\leqslant \frac{(1+\eta L)^2}{1-\eta^2L^2}\|z^* - z_0\|^2 + \sum_{t=3}^{T-1} \frac{\eta^2L^2}{1-\eta^2L^2}\|\eta F(z_t)\|^2,\end{aligned}$$

where we used $\|F(z_2)\|^2 \leqslant \frac{(1+\eta L)^2}{\eta^2}\|z^* - z_0\|^2$ for the second inequality. Now by the induction assumption and the fact that $\sum_{t=3}^{+\infty} \frac{1}{t^2} \leqslant 1/2$, we get

$$\begin{aligned}\frac{T^2}{4}\|\eta F(z_T)\|^2 &\leqslant \frac{(1+\eta L)^2}{1-\eta^2L^2}\|z^* - z_0\|^2 + \frac{4\eta^2L^2(1+\eta L)^2\|z^* - z_0\|^2}{(1-\eta^2L^2)(1-3\eta^2L^2)}\sum_{t=3}^{T-1}\frac{1}{t^2} \\ &\leqslant \frac{(1+\eta L)^2}{1-\eta^2L^2}\|z^* - z_0\|^2 + \frac{2\eta^2L^2(1+\eta L)^2\|z^* - z_0\|^2}{(1-\eta^2L^2)(1-3\eta^2L^2)} \\ &\leqslant \frac{(1+\eta L)^2}{1-3\eta^2L^2}\|z^* - z_0\|^2.\end{aligned}$$

Then we have the required result that

$$\|F(z_T)\|^2 \leqslant \frac{4(1+\eta L)^2}{\eta^2(1-3\eta^2L^2)T^2}\|z^* - z_0\|^2.$$

Question 5:

By the convexity of $f(\cdot, y)$ and the concavity of $f(x, \cdot)$, we have

$$\max_{y\in\mathcal{Y}} f(x_T, y) - \min_{x\in\mathcal{X}} f(x, y_T) \leqslant \max_{x\in\mathcal{X}} \langle \nabla_x f(x_T, y_T), x_T - x\rangle + \max_{y\in\mathcal{Y}} \langle -\nabla_y f(x_T, y_T), y_T - y\rangle$$

$$
\begin{aligned}
&= \max_{x \in \mathcal{X}, y \in \mathcal{Y}} \langle F(z_T), z_T - z \rangle \\
&\leqslant \max_{x \in \mathcal{X}, y \in \mathcal{Y}} \|F(z_T)\| \sqrt{\|x_T - x\|^2 + \|y_T - y\|^2} \\
&\leqslant \|F(z_T)\| \sqrt{2} \|z^* - z_0\| \\
&\leqslant \frac{2\sqrt{2}(1 + \eta L)}{\eta \sqrt{1 - 3\eta^2 L^2 T}} \|z^* - z_0\|^2
\end{aligned}
$$

□