Optimization for Data Science, FS23 (Bernd Gärtner and Niao He)
Graded Assignment 4! Last one!

Student Name: Dania Sana

Exercise 1: Min-Max for Smooth Functions

Consider the optimization problem:

$$f^* = \min_{x \in \mathbb{R}^{20},\ ||x||_2 \leq 1} f(x)$$

$$f(x) = \max_{1 \leq i \leq 10} f_i(x)$$

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

Solution 1): Consider the following function:

$$h(x) = \frac{ln(e^{f_1(x)\epsilon'} + e^{f_2(x)\epsilon'} + ... + e^{f_{10}(x)\epsilon'})}{\epsilon'}$$

with $\epsilon' = T$, where T is the total number of iterations. It then holds:

$$h(x) = \frac{ln(e^{f_1(x)\epsilon'} + e^{f_2(x)\epsilon'} + ... + e^{f_{10}(x)\epsilon'})}{\epsilon'} \geq \frac{ln(e^{\max_{i \in \{1,2,...,10\}} f_i(x)\epsilon'})}{\epsilon'} = \max_{i \in \{1,2,...,10\}} f_i(x)$$

Moreover, it holds that:

$$h(x) = \frac{ln(e^{f_1(x)\epsilon'} + e^{f_2(x)\epsilon'} + ... + e^{f_{10}(x)\epsilon'})}{\epsilon'} \leq \frac{ln(e^{\max_{i \in \{1,2,...,10\}} f_i(x)\epsilon'} + ... + e^{\max_{i \in \{1,2,...,10\}} f_i(x)\epsilon'})}{\epsilon'}$$

$$\leq \frac{ln(10 \cdot e^{\max_{i \in \{1,2,...,10\}} f_i(x)\epsilon'})}{\epsilon'} = \frac{ln(10)}{\epsilon'} + \frac{\max_{i \in \{1,2,...,10\}} f_i(x)\epsilon'}{\epsilon'} = \frac{ln(10)}{\epsilon'} + \max_{i \in \{1,2,...,10\}} f_i(x)$$

Thus, letting $\epsilon' = T$, we get the following very important result:

$$\max_{i \in \{1,2,...,10\}} f_i(x) \leq h(x) \leq \max_{i \in \{1,2,...,10\}} f_i(x) + \frac{ln(10)}{T} \tag{1}$$

Now, we show that $h(x)$ is convex and L-smooth, given that each f_i is convex and 1-smooth. To conclude that it is convex, it is enough to show that for $\theta \in [0, 1]$, we have:

$$h(\theta x + (1 - \theta)y) \leq \theta h(x) + (1 - \theta)h(y) \tag{2}$$

We get the following:

$$h(\theta x+(1-\theta)y) = \frac{1}{\epsilon'}[ln(e^{f_1(\theta x+(1-\theta)y)\epsilon'}+...+e^{f_{10}(\theta x+(1-\theta)y)\epsilon'})] = \frac{1}{\epsilon'}\left[ln(\sum_{i=1}^{10} e^{f_i(\theta x+(1-\theta)y)\epsilon'})\right] \leq$$

$$\leq \frac{1}{\epsilon'}\left[ln(\sum_{i=1}^{10} e^{\epsilon'[\theta f_i(x)+(1-\theta)f_i(y)]})\right] \quad f_i \text{ is convex, thus } f_i(\theta x+(1-\theta)y) \leq \theta f_i(x)+(1-\theta)f_i(y)$$

$$= \frac{1}{\epsilon'}\left[ln(\sum_{i=1}^{10} e^{\theta\epsilon' f_i(x)} \cdot e^{(1-\theta)\epsilon' f_i(y)})\right] \quad (*)$$

Apply the Hölder's inequality $\sum_{i=1}^{n} x_i y_i \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \cdot \left(\sum_{i=1}^{n} |y_i|^q\right)^{\frac{1}{q}}$ with $\frac{1}{p}+\frac{1}{q}=1$ to the above inequality by letting, $p=\frac{1}{\theta}$ and $q=\frac{1}{1-\theta}$. Therefore, it holds $\sum_{i=1}^{10} e^{\theta\epsilon' f_i(x)} \cdot e^{(1-\theta)\epsilon' f_i(y)} \leq \left(\sum_{1=1}^{10} e^{\epsilon' f_i(x)}\right)^{\theta} \cdot \left(\sum_{1=1}^{10} e^{\epsilon' f_i(y)}\right)^{1-\theta}$. Therefore the above expression $(*)$ is bounded by:

$$\leq \frac{1}{\epsilon'}\left[ln([\sum_{i=1}^{10} e^{\epsilon' f_i(x)}]^{\theta})+ln([\sum_{i=1}^{10} e^{\epsilon' f_i(y)}]^{1-\theta})\right] = \frac{1}{\epsilon'}\theta \cdot ln[\sum_{i=1}^{10} e^{\epsilon' f_i(x)}]+\frac{1}{\epsilon'}(1-\theta)\cdot ln[\sum_{i=1}^{10} e^{\epsilon' f_i(y)}]$$

$$= \theta h(x)+(1-\theta)h(y)$$

Thus we just showed that $h(\theta x+(1-\theta y)) \leq \theta h(x)+(1-\theta)h(y)$, concluding that h is convex. In particular, h is convex over the 20-dimensional unit ball set.
Now we need to show that h is smooth over the 20-dimensional unit ball.

From Lemma 3.5, since each f_i is 1-smooth and convex, the gradient of each f_i is 1-lipschitz continuous, meaning that each of their hessians is bounded by the identity matrix I (Theorem 2.10). It would be enough to show that $||\nabla^2 h(x)|| \leq L \ \forall x \in \mathcal{X}$ ($\mathcal{X}$ is the 20-dimensional unit ball), which would imply that h is smooth with parameter L ([1], Definition 5.1 and Theorem 2.10 of the lecture notes). Let us find the partial derivatives of h with respect to x_i ($i \in \{1,2,...,20\}$):

$$\frac{\partial h}{\partial x_i}(x) = \frac{1}{\epsilon'}\cdot\epsilon'\left[\frac{e^{\epsilon' f_1(x)}\frac{\partial f_1}{\partial x_i}(x)+...+e^{\epsilon' f_{10}(x)}\frac{\partial f_{10}}{\partial x_i}(x)}{e^{\epsilon' f_1(x)}+...+e^{\epsilon' f_{10}(x)}}\right] = \frac{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_i}(x)}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}}$$

Now we find the elements of the Hessian matrix of this function, in the $j-th$ column, $i-th$ row (or vice versa):

$$\frac{\partial^2 h}{\partial x_j \partial x_i}(x) = \frac{\partial}{\partial x_j}\left[\frac{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_i}(x)}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}}\right] =$$

$$= \frac{\frac{\partial}{\partial x_j}(\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_i}(x)) \cdot \sum_{k=1}^{10} e^{\epsilon' f_k(x)} - (\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_i}(x)) \cdot (\frac{\partial}{\partial x_j}\sum_{k=1}^{10} e^{\epsilon' f_k(x)})}{(\sum_{k=1}^{10} e^{\epsilon' f_k(x)})^2}$$

We commit the calculations component by component:

$$\frac{\partial}{\partial x_j}\left[\sum_{k=1}^{10} e^{\epsilon' f_k(x)} \cdot \frac{\partial f_k}{\partial x_i}(x)\right] \cdot \sum_{k=1}^{10} e^{\epsilon' f_k(x)} = \sum_{k=1}^{10} \frac{\partial}{\partial x_j}(e^{\epsilon' f_k(x)} \cdot \frac{\partial f_k}{\partial x_i}(x)) \cdot \sum_{k=1}^{10} e^{\epsilon' f_k(x)}$$

$$= \sum_{k=1}^{10}\left(\frac{\partial}{\partial x_j}(e^{\epsilon' f_k(x)}) \cdot \frac{\partial f_k}{\partial x_i}(x) + e^{\epsilon' f_k(x)}\frac{\partial^2 f_k}{\partial x_j \partial x_i}(x)\right) \cdot \sum_{k=1}^{10} e^{\epsilon' f_k(x)}$$

$$= \sum_{k=1}^{10}\left(\epsilon' e^{\epsilon' f_k(x)} \cdot \frac{\partial f_k}{\partial x_j}(x)\frac{\partial f_k}{\partial x_i}(x) + e^{\epsilon' f_k(x)}\frac{\partial^2 f_k}{\partial x_j \partial x_i}(x)\right) \sum_{k=1}^{10} e^{\epsilon' f_k(x)}$$

$$= \sum_{k=1}^{10} \epsilon' e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_j}(x)\frac{\partial f_k}{\partial x_i}(x) \cdot \sum_{k=1}^{10} e^{\epsilon' f_k(x)} + \sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial^2 f_k}{\partial x_j \partial x_i}(x)\sum_{k=1}^{10} e^{\epsilon' f_k(x)}$$

After taking the ratio to $(\sum_{k=1}^{10} e^{\epsilon' f_k(x)})^2$, we get the first component as:

$$\frac{\sum_{k=1}^{10} \epsilon' e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_j}(x)\frac{\partial f_k}{\partial x_i}(x)}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} + \frac{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial^2 f_k}{\partial x_j \partial x_i}(x)}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} \tag{3}$$

Now we calculate the other component as below:

$$\left(\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_i}\right) \cdot \frac{\partial}{\partial x_j}\left(\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\right) = \epsilon\left(\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_i}(x)\right) \cdot \left(\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_j}(x)\right)$$

Finally, the second component is:

$$\frac{-\epsilon\left(\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_i}(x)\right) \cdot \left(\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_j}(x)\right)}{(\sum_{k=1}^{10} e^{\epsilon' f_k(x)})^2} \tag{4}$$

Using (3) and (4), we get:

$$\frac{\partial^2 h}{\partial x_j \partial x_i}(x) =$$

$$= \frac{\sum_{k=1}^{10} \epsilon' e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_j}(x)\frac{\partial f_k}{\partial x_i}(x)}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} + \frac{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial^2 f_k}{\partial x_j \partial x_i}(x)}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} - \frac{\epsilon'\left(\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_i}(x)\right) \cdot \left(\sum_{k=1}^{10} e^{\epsilon' f_k(x)}\frac{\partial f_k}{\partial x_j}(x)\right)}{(\sum_{k=1}^{10} e^{\epsilon' f_k(x)})^2}$$

Carefully observe that the hessian matrix of function h is symmetric (because if we interchange i and j, the above result remains unchanged). Now, we need to figure out how the hessian matrix looks like, based on the i, j-the entry we just calculated above. My claim is that H, the hessian matrix of function h can be expressed as $H = \epsilon' \sum_{k=1}^{10} A_k +$

$\sum_{m=1}^{10} B_m - \epsilon' CC^\top$ where A_k, B_m are matrices of dimension 20×20 and C is a matrix of dimension 20×10. Indeed A_k can be represented as below:

$$\begin{pmatrix} (A_k)_{1,1} & (A_k)_{1,2} & \cdots & (A_k)_{1,20} \\ (A_k)_{2,1} & (A_k)_{2,2} & \cdots & (A_k)_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ (A_k)_{20,1} & (A_k)_{20,2} & \cdots & (A_k)_{20,20} \end{pmatrix}$$

with the ij-th entry as below:

$$(A_k)_{i,j} = \frac{e^{\epsilon' f_k(x)} \frac{\partial f_k}{\partial x_j}(x) \frac{\partial f_k}{\partial x_i}(x)}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}}$$

Similarly, B_m ca represented as below:

$$\begin{pmatrix} (B_m)_{1,1} & (B_m)_{1,2} & \cdots & (B_m)_{1,20} \\ (B_m)_{2,1} & (B_m)_{2,2} & \cdots & (B_m)_{2,20} \\ \vdots & \vdots & \vdots & \vdots \\ (B_m)_{20,1} & (B_m)_{20,2} & \cdots & (B_m)_{20,20} \end{pmatrix}$$

with entries as below:

$$(B_m)_{i,j} = \frac{e^{\epsilon' f_m(x)} \frac{\partial^2 f_m}{\partial x_j \partial x_i}(x)}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}}$$

On the other hand, matrix C can be written as:

$$\begin{pmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,10} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,10} \\ \vdots & \vdots & \vdots & \vdots \\ C_{20,1} & C_{20,2} & \cdots & C_{20,20} \end{pmatrix}$$

with entries:

$$C_{i,j} = \frac{e^{\epsilon' f_j(x)} \frac{\partial f_j}{\partial x_i}(x)}{(\sum_{k=1}^{10} e^{\epsilon' f_k(x)})}$$

By observing carefully, we can conclude that

$$(B_m)_{i,j} = \frac{e^{\epsilon' f_m(x)}}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} \cdot \frac{\partial^2 f_m}{\partial x_j \partial x_i}(x) \text{ meaning that } B_m = \frac{e^{\epsilon' f_m(x)}}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} \cdot \nabla^2 f_m(x)$$

Where $\nabla^2 f_m(x)$ is the hessian matrix of function f_m. Since $= \frac{e^{\epsilon' f_m(x)}}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} > 0$, and f_m is an 1-smooth function (i.e., $\nabla^2 f_m(x) \leq I$) then $B_m \leq \frac{e^{\epsilon' f_m(x)}}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} \cdot I$. This further implies that

$$\sum_{m=1}^{10} B_m \leq \frac{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} \cdot I = I \tag{5}$$

Moreover, observe that for any matrix C, $CC^\top$ is positive semi-definite. In order to show this it is enough to prove that $\langle x, CC^\top x\rangle \geq 0$, $\forall x \in \mathbb{R}^{20}$. This holds indeed since we have $\langle x, CC^\top x\rangle = \langle C^\top x, C^\top x\rangle = ||C^\top x||^2 \geq 0$. This implies that $CC^\top \geq 0$ which immediately infers that:

$$-\epsilon' CC^\top \leq 0 \tag{6}$$

Now we work with matrix A. First note that by the problem's condition, $||\nabla f_i(x)|| \leq 1$, $\forall x \in \mathcal{X}$ (20-dimensional unit ball) $\forall i \in \{1, 2, ..., 10\}$, which is equivalent to $\sum_{j=1}^{20} \left|\frac{\partial f_i}{\partial x_j}(x)\right|^2 \leq 1$ which directly implies that $\forall j \in \{1, 2, ..., 20\}$, $\left|\frac{\partial f_i}{\partial x_j}(x)\right| \leq 1$ (otherwise, the sum would be greater than 1). Moreover $\forall k \in \{1, 2, ..., 10\}$ and $\forall i, j \in \{1, 2, ..., 20\}$, using Cauchy-Schwartz inequality it holds that:

$$\frac{\partial f_k(x)}{\partial x_j} \cdot \frac{\partial f_k(x)}{\partial x_i} \leq \left|\frac{\partial f_k(x)}{\partial x_j}\right| \cdot \left|\frac{\partial f_k(x)}{\partial x_i}\right| \leq 1 \tag{7}$$

I claim that $A_k \leq 50 \cdot I$ where $k \in \{1, 2, ..., 10\}$. This is indeed true because if we find the entries of the matrix $50 \cdot I - A_k$ (this matrix is of dimension 20×20) they are of the form:

$$(50 \cdot I - A_k)_{i,i} = 50 - \frac{e^{\epsilon' f_k(x)} \left(\frac{\partial f_k}{\partial x_i}(x)\right)^2}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}}$$

$$(50 \cdot I - A_k)_{i,j} = -\frac{e^{\epsilon' f_k(x)} \frac{\partial f_k}{\partial x_i} \cdot \frac{\partial f_k}{\partial x_j}}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} \quad \text{for } i \neq j$$

Let us bound each diagonal entry of the matrix $50 \cdot I - A_k$ and use (7) to infer that $(\frac{\partial f_k}{\partial x_i})^2 \leq 1$:

$$(50 \cdot I - A_k)_{i,i} = (50 - \frac{e^{\epsilon' f_k(x)} \left(\frac{\partial f_k}{\partial x_i}\right)^2}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}}) \geq 50 - \frac{e^{\epsilon' f_k(x)}}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} \geq 50 - 1 = 49 \tag{8}$$

In the last inequality, I use the fact that $\frac{e^{\epsilon' f_k(x)}}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} \leq 1$ Now, let us use this together with (7) and add the absolute values of the off diagonal entries in row s

$$R_s = \sum_{j=1, j\neq s}^{20} |(50 \cdot I - A_k)_{s,j}| \leq \sum_{j=1}^{20} \frac{e^{\epsilon' f_k(x)} |\frac{\partial f_k}{\partial x_s}| \cdot |\frac{\partial f_k}{\partial x_j}|}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} \leq \sum_{j=1}^{20} \frac{e^{\epsilon' f_k(x)}}{\sum_{k=1}^{10} e^{\epsilon' f_k(x)}} \leq 20 \tag{9}$$

Moreover, we observe that $(50 \cdot I - A_k)$ is symmetric. Finally, using (8) and (9), $(50 \cdot I - A_k)$ is a symmetric matrix with positive diagonal entries such that each diagonal entry is greater then the sum of the absolute values of the off diagonal entries in its row (since each diagonal entry ≥ 49, whereas the sum of the absolute values of the off-diagonal entries in each row is ≤ 20). We now use the 'Gerschgorin circle theorem' which would infer that every eigenvalues of the matrix $50 \cdot I - A_k$ lies within at least one of the Gershgorin discs $D((50 \cdot I - A_k)_{i,i}, R_i)$. Since all of these discs lie in the positive quadrant of the coordinate axes (since we found that $R_i < (50 \cdot I - A_k)_{i,i}$), this directly implies that all the eigenvalues of this matrix are positive and thus $(50 \cdot I - A_k)$ is positive definite [[4]: Find a proof of this theorem] . This implies that:

$$50 \cdot I - A_k \geq 0 \quad \to \quad A_k \leq 50 \cdot I$$

This further implies that:

$$\epsilon' \sum_{k=1}^{10} A_k \leq \epsilon' 500 \tag{10}$$

Finally using (5), (6) and (10), we get:

$$H = \epsilon' \sum_{k=1}^{10} A_k + \sum_{m=1}^{10} B_m - \epsilon' CC^\top \leq \epsilon' 500 \cdot I + I = (500\epsilon' + 1) \cdot I$$

directly implying that h is $(500\epsilon' + 1)$-smooth.

We aim to design an accelerated gradient descent algorithm which achieves the desired convergence rate. For this we need the following two lemmas:

Lemma 1. *Let $x \in \mathcal{X}$ with $\mathcal{X}$ being a convex and closed set and $y \in \mathbb{R}^n$, then:*

$$(\Pi_{\mathcal{X}}(y) - x)^\top (\Pi_{\mathcal{X}}(y) - y) \leq 0$$

which also implies that $||\Pi_{\mathcal{X}}(y) - x||^2 + ||y - \Pi_{\mathcal{X}}(y)||^2 \leq ||x - y||^2$

Proof. *The proof can be found in the lecture notes page 115, Fact 4.1*

Lemma 2. *Let $x, y \in \mathcal{X} \subseteq \mathbb{R}^n$, with $\mathcal{X}$ being a convex and closed set and let f be a convex and L-smooth function. Define $x^+ = \Pi_{\mathcal{X}}(x - \frac{1}{L}\nabla f(x))$ and $g_{\mathcal{X}}(x) = L(x - x^+)$ Then the following holds true:*

$$f(x^+) - f(y) \leq g_{\mathcal{X}}(x)^\top (x - y) - \frac{1}{2L}||g_{\mathcal{X}}(x)||^2$$

Proof. Observe that

$$\nabla f(x)^\top (x^+ - y) \leq g_{\mathcal{X}}(x)^\top (x^+ - y) \tag{11}$$

Indeed this holds as it is equivalent to $(\nabla f(x) - g_{\mathcal{X}}(x))^\top (x^+ - y) \leq 0$. We substitute $g_{\mathcal{X}}(x) = L(x - x^+)$ and this becomes equivalent to the inequalities $(\nabla f(x) - L(x - x^+))^\top (x^+ - y) \leq 0 \to L(x^+ - (x - \frac{\nabla f(x)}{L}))^\top (x^+ - y) \leq 0 \to$ $\to (x^+ - (x - \frac{\nabla f(x)}{L}))^\top (x^+ - y) \leq 0$. Since $x - \frac{\nabla f(x)}{L} \in \mathbb{R}^n$ and $y \in \mathcal{X}$, this inequality holds because of **Lemma 1**. We use (11) to prove Lemma 2 as below:

$$f(x^+) - f(y) = f(x^+) - f(x) + f(x) - f(y) \leq \tag{12}$$

$$\leq \nabla f(x)^\top (x^+ - x) + \frac{L}{2}||x^+ - x||^2 + \nabla f(x)^\top (x - y) = \tag{13}$$

$$= \nabla f(x)^\top (x^+ - y) + \frac{1}{2L}||g_{\mathcal{X}}(x)||^2 \leq \tag{14}$$

$$\leq g_{\mathcal{X}}(x)^\top (x^+ - y) + \frac{1}{2L}||g_{\mathcal{X}}(x)||^2 = \tag{15}$$

$$= g_{\mathcal{X}}(x)^\top (x - y) - \frac{1}{2L}||g_{\mathcal{X}}(x)||^2 \tag{16}$$

Explanations about the steps (12) – (16):
In (12)-(13) we use the facts that f is L-smooth and convex. This implies that $f(x^+) \leq f(x) + \nabla f(x)^\top (x^+ - x) + \frac{L}{2}||x^+ - x||^2$ (smoothness) and $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$ (convexity). In (14), we just do the respective calculations
$\nabla f(x)^\top (x^+ - x) + \nabla f(x)^\top (x - y) = \nabla f(x)^\top (x^+ - y)$ and $||g_{\mathcal{X}}(x)||^2 = L^2 ||x - x^+||^2$. In (15) we use the result we showed in (11). In (16) we use the following expansion $g_{\mathcal{X}}(x)^\top (x^+ - y) = g_{\mathcal{X}}(x)^\top (x^+ - x + x - y) = g_{\mathcal{X}}(x)^\top (x^+ - x) + g_{\mathcal{X}}(x)^\top (x - y) = -\frac{1}{L}||g_{\mathcal{X}}(x)||^2 + g_{\mathcal{X}}(x)^\top (x - y)$ given that $g_{\mathcal{X}}(x) = L(x - x^+)$.

Note: This proof and most of the proof of the next theorem are taken from one of the books this course is based on, also apparent in the website [2]. (We can find them in this book as Lemma 3.4 and Theorem 3.12. respectively)

Now, we need to prove the following theorem which will help us on constructing a suitable algorithm for our problem:

Theorem 1. Nesterove Accelerated Gradient Descent. Let f be an L-smooth and convex function. Moreover, let $\mathcal{X}$ be the unit ball (we get for free that is a convex and closed set). Define the sequences:

$$\lambda_0 = 0, \quad \lambda_s = \frac{1 + \sqrt{1 + 4\lambda_{s-1}^2}}{2}, \text{ and } \gamma_s = \frac{1 - \lambda_s}{\lambda_{s+1}}$$

Let $y_1 = x_1 \in \mathcal{X}$ and design an algorithm as below:

$$y'_{s+1} = x_s - \frac{1}{L}\nabla f(x_s)$$
$$y_{s+1} = \Pi_{\mathcal{X}}(y'_{s+1})$$
$$x'_{s+1} = (1 - \gamma_s)y_{s+1} + \gamma_s y_s$$
$$x_{s+1} = \Pi_{\mathcal{X}}(x'_{s+1})$$

If $x^* \in \mathcal{X}$ such that $f(x^*) = \min_{x \in \mathcal{X}} f(x)$, we then have:

$$f(y_t) - f(x^*) \leq \frac{2L||x_1 - x^*||^2}{t^2} \tag{17}$$

Proof. $y_{s+1} = \Pi_{\mathcal{X}}(x_s - \frac{1}{L}\nabla f(x_s))$ and $y_s \in \mathcal{X}$. We can therefore apply lemma 2 to upper bound $f(y_{s+1}) - f(y_s)$, as:

$$f(y_{s+1}) - f(y_s) \leq g_{\mathcal{X}}(x_s)^\top (x_s - y_s) - \frac{L}{2}||x_s - y_{s+1}||^2 = L(x_s - y_{s+1})^\top (x_s - y_s) - \frac{L}{2}||x_s - y_{s+1}||^2 \tag{18}$$

Similarly, using **Lemma 2**, and the fact that $y_{s+1} = \Pi_{\mathcal{X}}(x_s - \frac{1}{L}\nabla f(x_s))$ and $x^* \in \mathcal{X}$, let $x^+ = y_{s+1}, y = x^*, x = x_s$ we get the following inequality concerning $f(y_{s+1}) - f(x^*)$:

$$f(y_{s+1}) - f(x^*) \leq g_{\mathcal{X}}(x_s)^\top (x_s - x^*) - \frac{1}{2L}||g_{\mathcal{X}}(x_s)||^2 = L(x_s - y_{s+1})^\top (x_s - x^*) - \frac{L}{2}||x_s - y_{s+1}||^2 \tag{19}$$

Now, observe that $\lambda_s \geq 1$. (Why? - Because $\lambda_{s-1}^2 \geq 0 \rightarrow \sqrt{1+4\lambda_{s-1}^2} \geq 1 \rightarrow \lambda_s = \frac{1+\sqrt{1+4\lambda_{s-1}^2}}{2} \geq \frac{2}{2} = 1$). Therefore, we multiply both sides of inequality (18) by $\lambda_s - 1$ and adding the result to (19), we obtain by letting $\delta_s = f(y_s) - f(x^*)$:

$$\lambda_s \delta_{s+1} - (\lambda_s - 1)\delta_s \leq$$

$$\leq L(x_s - y_{s+1})^\top (\lambda_s x_s - (\lambda_s - 1)y_s - x^*) - \frac{L}{2}\lambda_s ||x_s - y_{s+1}||^2$$

Multiply this inequality by λ_s and using that by definition $\lambda_{s-1}^2 = \lambda_s^2 - \lambda_s$ and the inequality $2a^\top b - ||a||^2 = ||b||^2 - ||b-a||^2$:

$$= \lambda_s^2 \delta_{s+1} - \lambda_{s-1}^2 \delta_s \leq \frac{L}{2}\left(2\lambda_s(x_s - y_{s+1})^\top(\lambda_s x_s - (\lambda_s - 1)y_s - x^*) - ||\lambda_s(y_{s+1} - x_s)||^2\right)$$

$$= \frac{L}{2}\left(||\lambda_s x_s - (\lambda_s - 1)y_s - x^*||^2 - ||\lambda_s y_{s+1} - (\lambda_s - 1)y_s - x^*||^2\right) \tag{20}$$

From Lemma 1, it holds that given $y \in \mathbb{R}^n$ and $x \in \mathcal{X}$, (where $\mathcal{X}$ is a closed and convex set), then $||\Pi_{\mathcal{X}}(y) - x||^2 + ||y - \Pi_{\mathcal{X}}(y)|| \leq ||y - x||^2$. This directly shows that also the inequality $||y - x||^2 \geq ||\Pi_{\mathcal{X}}(y) - x||^2$ holds. Thus, let $\mathcal{Y}$ be a convex closed set such that $x^* \in \mathcal{X}$. And let $y = \lambda_s y_{s+1} - (\lambda_s - 1)y_s)$. Then it holds:

$$||(\lambda_s y_{s+1} - (\lambda_s - 1)y_s) - x^*|| \geq ||\Pi_{\mathcal{Y}}(\lambda_s y_{s+1} - (\lambda_s - 1)y_s) - x^*|| \tag{21}$$

Need to find a convex closed set $\mathcal{Y}$ with $x^* \in \mathcal{Y}$ such that

$$\Pi_{\mathcal{Y}}(\lambda_s y_{s+1} - (\lambda_s - 1)y_s) = \lambda_{s+1} x_{s+1} - (\lambda_{s+1} - 1)y_{s+1} \tag{22}$$

By definition, it holds that $x_{s+1} = \Pi_{\mathcal{X}}((1-\gamma_s)y_{s+1} + \gamma_s y_s)$, implying that $\forall x \in \mathcal{X}$: $||(1-\gamma_s)y_{s+1} + \gamma_s y_s - x_{s+1}||^2 \leq ||(1-\gamma_s)y_{s+1} + \gamma_s y_s - x||^2$. Substituting $\gamma_s = \frac{(1-\lambda_s)}{\lambda_{s+1}}$ and $1 - \gamma_s = \frac{\lambda_{s+1}+\lambda_s - 1}{\lambda_{s+1}}$, the inequality becomes ($\forall x \in \mathcal{X}$):

$$||(\lambda_s y_{s+1} - (\lambda_s - 1)y_s) - (\lambda_{s+1}x_{s+1} - (\lambda_{s+1}-1)y_{s+1})||^2 \leq ||(\lambda_s y_{s+1} - (\lambda_s - 1)y_s) - (\lambda_{s+1}x - (\lambda_{s+1}-1)y_{s+1})||^2 \tag{23}$$

Let $\mathcal{Y} = \lambda_{s+1}\mathcal{X} - (\lambda_{s+1} - 1)y_{s+1}$. Observe that is closed (Why? - Because $\mathcal{X}$ is compact, $\lambda_{s+1}\mathcal{X}$ is compact - since it is a set obtained by applying a continuous map to the set $\mathcal{X}$; thus since $\mathcal{X}$ is compact $\rightarrow \lambda_{s+1}\mathcal{X}$ is compact as well. Moreover, observe that $(1 - \lambda_{s+1})y_{s+1}$ as a single point constitutes a closed set. Finally from basic analysis, the algebraic sum of closed set and compact set is closed, concluding that $\mathcal{Y}$ is closed.) Moreover, $\mathcal{Y}$ is convex, since $\mathcal{X}$ is convex. (Why? - let $y_1, y_2 \in \mathcal{Y}$. Then $\exists x_1, x_2 \in \mathcal{X}$, such that $y_1 = \lambda_{s+1}x_1 - (\lambda_{s+1} - 1)y_{s+1}$, and $y_2 = \lambda_{s+1}x_2 - (\lambda_{s+1} - 1)y_{s+1}$. Need to show that for $\lambda \in [0,1]$, $\lambda y_1 + (1-\lambda)y_2 \in \mathcal{Y}$. Indeed $\lambda y_1 + (1-\lambda)y_2 = \lambda(\lambda_{s+1}x_1 - (\lambda_{s+1} - 1)y_{s+1} + (1-\lambda)(\lambda_{s+1}x_2 - (\lambda_{s+1} - 1)y_{s+1}) = \lambda_{s+1}(\lambda x_1 + (1-\lambda)x_2) - (\lambda_{s+1} - 1)y_{s+1} \in \mathcal{Y}$, since $\lambda x_1 + (1-\lambda)x_2 \in \mathcal{X}$, as $\mathcal{X}$ is convex.) Next we need to show that $x^* \in \mathcal{Y}$.

For this it is enough to prove that $\exists x \in \mathcal{X}$ such that $x^* = \lambda_{s+1}x - (\lambda_{s+1} - 1)y_{s+1}$, Indeed $x = \frac{x^*+(\lambda_{s+1}-1)y_{s+1}}{\lambda_{s+1}}$ satisfies the equality and I claim that this $x \in \mathcal{X}$. Indeed $||x|| = ||\frac{x^*+(\lambda_{s+1}-1)y_{s+1}}{\lambda_{s+1}}|| \leq \frac{||x^*||}{\lambda_{s+1}} + \frac{\lambda_{s+1}-1}{\lambda_{s+1}}||y_{s+1}|| \leq \frac{1}{\lambda_{s+1}} + \frac{\lambda_{s+1}-1}{\lambda_{s+1}} = 1$. This holds because $x^*, y_{s+1} \in \mathcal{X}$, and thus $||x^*||, ||y_{s+1}|| \leq 1$. Thus, we found a set $\mathcal{Y}$ such that (22) holds. We go back to (21), which then becomes (using (22)):

$$||(\lambda_s y_{s+1} - (\lambda_s - 1)y_s) - x^*|| \geq ||\lambda_{s+1}x_{s+1} - (\lambda_{s+1} - 1)y_{s+1} - x^*|| \tag{24}$$

Using (24) in (20), and letting $u_s = \lambda_s x_s - (\lambda_s - 1)y_s - x^*$, we obtain:

$$\lambda_s^2\delta_{s+1} - \lambda_{s-1}^2\delta_s \leq \frac{L}{2}\Big(||u_s||^2 - ||u_{s+1}||^2\Big) \tag{25}$$

Sum these inequalities from $s = 1$ to $s = t - 1$ and get:

$$\lambda_{t-1}^2\delta_t - \lambda_0^2\delta_1 = \sum_{s=1}^{t-1}(\lambda_s^2\delta_{s+1} - \lambda_{s-1}^2\delta_s) \leq \sum_{s=1}^{t-1}\frac{L}{2}\Big(||u_s||^2 - ||u_{s+1}||^2\Big) = \frac{L}{2}(||u_1||^2 - ||u_t||^2) \leq \frac{L}{2}||u_1||^2$$

Finally since $\lambda_0 = 0$, we get:

$$\delta_t \leq \frac{L||u_1||}{2\lambda_{t-1}^2} \tag{26}$$

Last, we are left to show that $\lambda_{t-1} \geq \frac{t}{2}$ for all $t \geq 2$. We prove it by induction. Let us show it for the base case, mainly for $t_0 = 2$. Indeed it is true since:

$$\lambda_1 = 1 \geq \frac{t_0}{2} = 1$$

Now, suppose that the inequality holds for some $t = k$, i.e.,:

$$\lambda_{k-1} \geq \frac{k}{2}$$

This directly implies that $\lambda_{k-1}^2 \geq \frac{k^2}{4}$. Let us now prove that the inequality holds for $t = k + 1$, i.e:

$$\lambda_k \geq \frac{(k+1)}{2} \tag{27}$$

Well, using the inductive hypothesis, we get:

$$\lambda_k = \frac{1 + \sqrt{1 + 4\lambda_{k-1}^2}}{2} \geq \frac{1 + \sqrt{1 + 4\frac{k^2}{4}}}{2} = \frac{1 + \sqrt{1 + k^2}}{2} \geq \frac{1 + k}{2} \quad \text{b.c.s } \sqrt{1 + k^2} \geq \sqrt{k^2} = k$$

We just showed (27), thus we can conclude that $\lambda_{t-1} \geq \frac{t}{2}$. We now go back to inequality (26), and use this fact to finally get:

$$\delta_t \leq \frac{L||u_1||^2}{2\lambda_{t-1}^2} \leq \frac{2L||u_1||^2}{t^2} \tag{28}$$

Substitute $\delta_t = f(y_t) - f(x^*)$, and $u_1 = \lambda_1 x_1 - \lambda_1 y_1 + y_1 - x^* = x_1 - x^*$ (since $x_1 = y_1$) and get the desired result:

$$f(y_t) - f(x^*) \leq \frac{2L||x_1 - x^*||^2}{t^2} \tag{29}$$

The rough idea of our designed algorithm, is to apply the above accelerated projected gradient descent on the function h. In other words we get the following complete algorithm:

Algorithm 1 Accelerated Projected Gradient Descent

```
L→ the smoothness parameter of function $h$
$\mathcal{X}$ is the 20-dimensional unit ball
$\epsilon' = T$
Pick $x_1 = y_1 \in \mathcal{X}$
$\lambda_0 = 0$
$\lambda - list = zeros[T+1]$    ▷ Stores the values of $\lambda_k$: It has $T+1$ elements at the end: $\lambda_0, ..., \lambda_T$; $\lambda_0$ is at position 0
for $t = 1, 2, 3, ..., T$ do
    $\lambda_{t-1} = \lambda - list[t-1]$
    $\lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$
    $\lambda - list[t] \leftarrow \lambda_t$
    $\lambda_{t+1} = \frac{1+\sqrt{1+4\lambda_t^2}}{2}$
    $\gamma_t = \frac{1-\lambda_t}{\lambda_{t+1}}$
    $\nabla f_1(x_t) = \left( \frac{\partial f_1(x_t)}{\partial x_1}, \frac{\partial f_1(x_t)}{\partial x_2}, ..., \frac{\partial f_1(x_t)}{\partial x_{20}} \right)$    ▷ The gradient of each $f_i$ is evaluated exactly once per iteration; at the value $x_t$
    $\nabla f_2(x_t) = \left( \frac{\partial f_2(x_t)}{\partial x_1}, \frac{\partial f_2(x_t)}{\partial x_2}, ..., \frac{\partial f_2(x_t)}{\partial x_{20}} \right)$
    $\nabla f_3(x_t) = \left( \frac{\partial f_3(x_t)}{\partial x_1}, \frac{\partial f_3(x_t)}{\partial x_2}, ..., \frac{\partial f_3(x_t)}{\partial x_{20}} \right)$
    $\vdots$
    $\nabla f_{10}(x_t) = \left( \frac{\partial f_{10}(x_t)}{\partial x_1}, \frac{\partial f_{10}(x_t)}{\partial x_2}, ..., \frac{\partial f_{10}(x_t)}{\partial x_{20}} \right)$
    $all - gradients = cbind(\nabla f_1(x_t), \nabla f_2(x_t), ..., \nabla f_{10}(x_t))$    ▷ This matrix consists of all the gradients as columns
    $vector - exp = c(e^{\epsilon' f_1(x_t)}, e^{\epsilon' f_2(x_t)}, ..., e^{\epsilon' f_{10}(x_t)})$
    $\nabla h(x_t) = \left( \frac{\langle vector{-}exp,\ all{-}gradients[1,] \rangle}{sum(vector{-}exp)}, ..., \frac{\langle vector{-}exp,\ all{-}gradients[20,] \rangle}{sum(vector{-}exp)} \right)$    ▷ X[1,] denotes the first row of matrix X
    $y'_{t+1} = x_t - \frac{1}{L}\nabla h(x_t)$
    $y_{t+1} = \Pi_{\mathcal{X}}(y'_{t+1})$
    $x'_{t+1} = (1-\gamma_t)y_{t+1} + \gamma_t y_t$
    $x_{t+1} = \Pi_{\mathcal{X}}(x'_{t+1})$
end for
return $y_{T+1}, \hat{f} = \max_{i \in \{1,2,...,10\}} f_i(y_{T+1})$
```

Note that $\mathcal{X}$ is a convex set ($\mathcal{X} = \{x \in \mathbb{R}^{20} : ||x||_2 \leq 1\}$. Let $x_1, x_2 \in \mathcal{X} \rightarrow ||x_1||_2 \leq 1$ and $||x_2||_2 \leq 1$. This means that given $\lambda \in [0,1]$, $||\lambda x + (1-\lambda)y||_2 \leq \lambda ||x||_2 + (1-\lambda)||y||_2 \leq \lambda + (1-\lambda) = 1$). Moreover, $x_1 = y_1 \in \mathcal{X}$, and $x^* = argmin_{x \in \mathcal{X}} h(x)$, thus $||x^* - x_1||^2 \leq (||x^*|| + ||x_1||)^2 \leq 4$. We have shown that (2) holds indicating that h is convex. Moreover, in (10), we have shown that h is smooth with smoothness parameter $L = (500\epsilon' + 1) = (500T + 1)$. We use Algorithm 1, Accelerated Projected Gradient Descent and since we the conditions of theorem 1 are satisfied, we get a convergence rate for h as in (17):

$$h(y_{T+1}) - h(x^*) \leq \frac{2L||x_1 - x^*||^2}{(T+1)^2} = \frac{8(500T+1)}{(T+1)^2} \quad (\textbf{Result})$$

It also holds that $500T + 1 \leq 500(T+1)$, thus $\frac{8(500T+1)}{(T+1)^2} \leq \frac{4000}{T+1}$. Moreover it is trivially true that $\frac{4000}{T+1} \leq \frac{4001}{T}$, which finally concludes that:

$$h(y_{T+1}) - h(x^*) \leq \frac{4001}{T} \tag{30}$$

Now, we use the inequality we used right in the beginning of this exercise mainly (1):

$$\max_{i \in \{1,2,\dots,10\}} f_i(x) \leq h(x) \leq \max_{i \in \{1,2,\dots,10\}} f_i(x) + \frac{ln(10)}{T}$$

Since $y_{T+1} \in \mathcal{X}$ (since it is defined as the projection on $\mathcal{X}$), it holds that

$$\max_{i \in \{1,2,\dots,10\}} f_i(y_{T+1}) \leq h(y_{T+1}) \tag{31}$$

Moreover, it also holds:

$$\min_{x \in \mathcal{X}} h(x) \leq \min_{x \in \mathcal{X}} \max_{i \in \{1,2,..,10\}} f_i(x) + \frac{ln(10)}{T} \;\rightarrow\; -h(x^*) = -\min_{x \in \mathcal{X}} h(x) \geq -\min_{x \in \mathcal{X}} \max_{i \in \{1,2,..,10\}} f_i(x) - \frac{ln(10)}{T} \tag{32}$$

Thus, using (30), (31), (32), we get the following:

$$\max_{i \in \{1,2,\dots,10\}} f_i(y_{T+1}) - \min_{x \in \mathcal{X}} \max_{i \in \{1,2,..,10\}} f_i(x) - \frac{ln(10)}{T} \leq h(y_{T+1}) - h(x^*) \leq \frac{4001}{T}$$

Finally:

$$\max_{i \in \{1,2,\dots,10\}} f_i(y_{T+1}) - \min_{x \in \mathcal{X}} \max_{i \in \{1,2,..,10\}} f_i(x) \leq \frac{4001 + ln(10)}{T}$$

Choose, $O(\frac{1}{\epsilon}) = T = \lceil \frac{4001 + ln(10)}{\epsilon} \rceil \leq \frac{2(4001 + ln(10))}{\epsilon}$ (Since $\lceil x \rceil \leq x + 1 \leq 2x$, for $x \geq 1$: we suppose ϵ is suffieciently small, i.e, $\frac{4001 + ln(10)}{\epsilon} \geq 1$), assuring that $\hat{f} - f^* \leq \epsilon$. Thus we need $O(\frac{1}{\epsilon})$ steps to achieve that $\rightarrow$ the gradient of each $f_i(\cdot)$ is evaluated for at most $O(1/\epsilon)$ times (since each such gradient is evaluated once per iteration and we have $O(1/\epsilon)$ iterations).

Note. Here, I considered that f_i are twice differentiable, however, since f_i is $1-smooth$, we interpret $\nabla^2 f \leq I$ as holding almost everywhere. Then, one can use this argument shown in the forum ([3]: Answer of Question 2) to proceed.

Exercise 2: Stochastic Gradient Descent

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

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

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

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

Solution 1):
In order to show that F is an L-smooth function, it is enough to show that (lecture notes, page 99, definition 3.2), $\forall x, y \in \mathbb{R}$:

$$F(x) \leq F(y) + \nabla F(y)^\top(x-y) + \frac{L}{2}||x-y||^2$$

or equivalently, it is enough to show that the following holds:

$$F(x) - F(y) - \nabla F(y)^\top(x-y) \leq \frac{L}{2}||x-y||^2 \tag{33}$$

$\nabla F(y) = \frac{y}{\max\{\frac{1}{L}, 2\sum_{t=0}^{T-1}\gamma_t\}}$, since F is a quadratic function and $\max\{\frac{1}{L}, 2\sum_{t=0}^{T-1}\gamma_t\}$ is just a scalar. We then get the following calculations:

$$F(x) - F(y) - \nabla F(y)^\top(x-y) = \frac{x^2}{2\max\{\frac{1}{L}, 2\sum_{t=0}^{T-1}\gamma_t\}} - \frac{y^2}{2\max\{\frac{1}{L}, 2\sum_{t=0}^{T-1}\gamma_t\}} -$$

$$-\frac{y^\top}{\max\{\frac{1}{L}, 2\sum_{t=0}^{T-1}\gamma_t\}}(x-y) = \frac{x^2 - y^2 - 2y^\top(x-y)}{2\max\{\frac{1}{L}, 2\sum_{t=0}^{T-1}\gamma_t\}} = \frac{(x-y)^2}{2\max\{\frac{1}{L}, 2\sum_{t=0}^{T-1}\gamma_t\}}$$

$$\leq \frac{(x-y)^2}{\frac{2}{L}} = \frac{L}{2}||x-y||^2 \text{ Since } \max\{\frac{1}{L}, 2\sum_{t=0}^{T-1}\gamma_t\} \geq \frac{1}{L} \to \frac{1}{\max\{\frac{1}{L}, 2\sum_{t=0}^{T-1}\gamma_t\}} \leq L$$

We just showed (33) holds meaning that F is L-smooth. Now we show that $F(x_0) - \min_x F(x) \leq \Delta$. Since $x_0 = \sqrt{2\Delta\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}$, $F(x_0)$ is given as below:

$$F(x_0) = \frac{x_0^2}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} = \frac{2\Delta\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} = \Delta$$

whereas, we have the following value for $\min_x F(x)$:

$$\min_x F(x) = \min_x \frac{x^2}{2\max\{\frac{1}{L}, 2\sum_{t=0}^{T-1}\gamma_t\}} \geq 0 \text{ since } x^2 \geq 0,\ \forall x \in \mathbb{R} \text{ and } \max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\} > 0$$

(Indeed if we minimize over $\mathbb{R}$, $\min_x F(x) = 0$) Therefore it holds that $F(x_0) - \min_x F(x) \leq \Delta - 0 = \Delta$

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 $0 \leq t \leq T$, we have $x_t \geq x_0/2$. This implies:

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

Solution 2):
Based on the way $F(x)$ is defined, it holds true that $\nabla F(x) = \frac{x}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}$. We then apply the SGD-algorithm and obtain an iterate as below:

$$x_t = x_{t-1} - \gamma_{t-1} \cdot \frac{x_{t-1}}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} = \left(1 - \frac{\gamma_{t-1}}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\right)x_{t-1} \quad (34)$$

Iterating further on $x_{t-1}, x_{t-2}, ..$ we obtain:

$$x_t = \left(1 - \frac{\gamma_{t-1}}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\right)x_{t-1} =$$

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

$$= \prod_{k=0}^{t-1}\left(1 - \frac{\gamma_k}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\right)x_0$$

Rigorous proof of why the above expansion holds: We show it by induction. Indeed, we need to prove that $\forall t \geq 1$, $x_t = \prod_{k=0}^{t-1}(1 - \frac{\gamma_k}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}})x_0$. Let us prove the base case, i.e., $t = 1$. Using (34), we obtain:

$$x_1 = \left(1 - \frac{\gamma_0}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\right)x_0 = \prod_{k=0}^{t-1=1-1=0}\left(1 - \frac{\gamma_k}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\right)x_0$$

Now, suppose it holds for some $t = s \geq 1$, i.e., $x_s = \prod_{k=0}^{s-1}\left(1 - \frac{\gamma_k}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\right)x_0$. We are now left to proving the equality for $t = s+1$. Using the inductive hypothesis and (34), we get the following expression for x_{s+1}:

$$x_{s+1} = \left(1 - \frac{\gamma_s}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\right)x_s =$$

$$\Big(1-\frac{\gamma_s}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}}\Big)\Big(\prod_{k=0}^{s-1}\Big(1-\frac{\gamma_k}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}}\Big)x_0\Big)=$$

$$\prod_{k=0}^{s}\Big(1-\frac{\gamma_k}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}}\Big)x_0$$

This concludes the proof for the equality:

$$x_t=\prod_{k=0}^{t-1}\Big(1-\frac{\gamma_k}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}}\Big)x_0 \;\; \forall t\geq 1 \tag{35}$$

If we further expand the above last product, we obtain:

$$\prod_{k=0}^{t-1}\Big(1-\frac{\gamma_k}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}}\Big)x_0=$$

$$=\Bigg[1-\frac{1}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}}\sum_{k=0}^{t-1}\gamma_k+\frac{1}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}^2}\sum_{k\neq j\in\{0,1,..,t-1\}}\gamma_k\gamma_j$$

$$-\frac{1}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}^3}\sum_{m\neq k\neq j\neq m\in\{0,1,..,t-1\}}\gamma_k\gamma_j\gamma_m+...(+/-)\frac{1}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}^t}\prod_{i=0}^{t-1}\gamma_i\Bigg]x_0 \tag{36}$$

where the last expression is either negative (it t is odd) or positive (if t is even).
Let us prove this by induction. Base case: $t=1$ it is true that
$\prod_{k=0}^{t-1=0}\big(1-\frac{\gamma_k}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}}\big)x_0=\big(1-\frac{\gamma_0}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}}\big)x_0=\big(1-\frac{\gamma_0}{(\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\})^1}\big)x_0$
Suppose the expansion holds for $t=s$. Thus we have:

$$\prod_{k=0}^{s-1}\Big(1-\frac{\gamma_k}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_k\}}\Big)x_0=$$

$$=\Bigg[1-\frac{1}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}}\sum_{k=0}^{s-1}\gamma_k+\frac{1}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}^2}\sum_{k\neq j\in\{0,1,..,s-1\}}\gamma_k\gamma_j$$

$$-\frac{1}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}^3}\sum_{m\neq k\neq j\neq m\in\{0,1,..,s-1\}}\gamma_k\gamma_j\gamma_m+...(+/-)\frac{1}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_t\}^s}\prod_{i=0}^{s-1}\gamma_i\Bigg]x_0$$

We need to prove it for $t=s+1$:

$$\prod_{k=0}^{s}\Big(1-\frac{\gamma_k}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_k\}}\Big)x_0=\Big(1-\frac{\gamma_s}{\max\{1/L,2\sum_{t=0}^{T-1}\}}\Big)\prod_{k=0}^{s-1}\Big(1-\frac{\gamma_k}{\max\{1/L,2\sum_{t=0}^{T-1}\gamma_k\}}\Big)x_0=$$

$$= \left(1 - \frac{\gamma_s}{\max\{1/L, 2\sum_{t=0}^{T-1}\}}\right)\Bigg[1 - \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\sum_{k=0}^{s-1}\gamma_k +$$

$$+ \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^2}\sum_{k\neq j\in\{0,1,..,s-1\}}\gamma_k\gamma_j - \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^3}\sum_{m\neq k\neq j\neq m\in\{0,1,..,s-1\}}\gamma_k\gamma_j\gamma_m +$$

$$+ ...(+/-)\frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^s}\prod_{i=0}^{s-1}\gamma_i\Bigg]x_0 = \Bigg[1 - \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\sum_{k=0}^{s}\gamma_k +$$

$$+ \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^2}\sum_{k\neq j\in\{0,1,..,s\}}\gamma_k\gamma_j - \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^3}\sum_{m\neq k\neq j\neq m\in\{0,1,..,s\}}\gamma_k\gamma_j\gamma_m +$$

$$+ ...(+/-)\frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^{s+1}}\prod_{i=0}^{s}\gamma_i\Bigg]x_0$$

Thus, (36) holds.

If we manage to show that the expression in the square parantheses of (36) is $\geq \frac{1}{2}$, we are done, because since $x_0 \geq 0$, this would imply that $x_t \geq \frac{1}{2}x_0$. Now, observe that $\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\} \geq 2\sum_{t=0}^{T-1}\gamma_t \rightarrow \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \leq \frac{1}{2}\cdot\frac{1}{\sum_{t=0}^{T-1}\gamma_t} \rightarrow -\frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \geq -\frac{1}{2}\cdot\frac{1}{\sum_{t=0}^{T-1}\gamma_t}$. Moreover, observe that $\sum_{k\neq m\neq l}\gamma_k\gamma_m\gamma_l \leq \left(\sum_{k\neq m}\gamma_k\gamma_m\right)\left(\sum_{t=0}^{T-1}\gamma_t\right)$. Moreover, we group the expanded elements of the product in (36) two by two. Let us now show that the first component of the expansion in (36): $[1 - \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\sum_{k=0}^{t-1}\gamma_k] \geq \frac{1}{2}$, Indeed it holds that:

$$[1 - \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\sum_{k=0}^{t-1}\gamma_k] \geq [1 - \frac{1}{2}\cdot\frac{1}{\sum_{t=0}^{T-1}\gamma_t}\sum_{k=0}^{t-1}\gamma_k] \geq 1 - \frac{1}{2} = \frac{1}{2}$$

where the last inequality holds since $\sum_{t=0}^{T-1}\gamma_t \geq \sum_{k=0}^{t-1}\gamma_k$ since T is the last iterate and $\gamma_i \geq 0\ \forall i$. Now we are left to show that each remaining group of the product value is non-negative. This is indeed true. One group is represented as (k-even, $k > 0$) (Note that each pair of $a_j \neq a_i$, $\forall i \neq j$):

$$\frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^k}\sum_{a_1\neq a_2\neq...\neq a_k}\prod_{i=a_1}^{a_k}\gamma_i - \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^{k+1}}\sum_{a_1\neq a_2\neq...\neq a_k\neq a_{k+1}}\prod_{i=a_1}^{a_{k+1}}\gamma_i \tag{37}$$

The following holds:

$$\frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^{k+1}}\sum_{a_1\neq a_2\neq...\neq a_k\neq a_{k+1}}\prod_{i=a_1}^{a_{k+1}}\gamma_i \leq$$

$$\leq \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^{k+1}}(\sum_{a_1\neq a_2\neq\ldots\neq a_k}\prod_{i=a_1}^{a_k}\gamma_i)\cdot(\sum_{i=0}^{T-1}\gamma_i)$$

$$\leq \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^{k}}(\sum_{a_1\neq a_2\neq\ldots\neq a_k}\prod_{i=a_1}^{a_k}\gamma_i)\frac{1}{2} \quad \text{Since } \max\{1/L, 2\sum_{t=0}^{T-1}\gamma_k\} \geq 2\sum_{t=0}^{T-1}\gamma_k$$

This further indicates that:

$$\frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^{k}}\sum_{a_1\neq a_2\neq\ldots\neq a_k}\prod_{i=a_1}^{a_k}\gamma_i - \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^{k+1}}\sum_{a_1\neq a_2\neq\ldots\neq a_k\neq a_{k+1}}\prod_{i=a_1}^{a_{k+1}}\gamma_i \tag{38}$$

$$\geq \frac{1}{2}\frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}^{k}}(\sum_{a_1\neq a_2\neq\ldots\neq a_k}\prod_{i=a_1}^{a_k}\gamma_i) \geq 0$$

Moreover, observe that the product in (36) has $t+1$ terms inside the square brackets. If t is odd, $t+1$ is even, indicating that that there are a total of $(t+1)/2$ groups and we apply the above argument each group except the first group (where $k=0$). However, if t is even, the last element is remained outside a group, however, it does not ruin our non-negativeness argument since is is accompanied by a $+$ sign, making it non-negative (since t is even). All these arguments conclude that $x_t \geq \frac{1}{2}x_0$.

Since $\nabla F(x) = \frac{x}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}$, and using the facts that $x_t \geq x_0/2$, and $x_0 = \sqrt{2\Delta\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \geq 0$, then:

$$\nabla F(x_t) = \frac{x_t}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \geq \frac{x_0}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} = \frac{\sqrt{2\Delta\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} =$$

$$= \frac{\sqrt{\Delta}}{\sqrt{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}}$$

which is what we wanted to show, respectively that $|\nabla F(x_t)| \geq \sqrt{\frac{\Delta}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}}$

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

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

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 $t \geq 2$:

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

Solution 3):
We show the inequality by induction. It holds that $\nabla f(x,\xi) = Lx + \xi$. It is true that $x_{t+1} = x_t - \gamma_t \nabla f(x_t, \xi_t) = (1-\gamma_t L)x_t - \gamma_t \xi_t$. Therefore, $x_1 = (1-\gamma_0 L)x_0 - \gamma_0 \xi_0$ and $x_2 = (1-\gamma_1 L)x_1 - \gamma_1 \xi_1 = (1-\gamma_1 L)[(1-\gamma_0 L)x_0 - \gamma_0 \xi_0] - \gamma_1 \xi_1 =$
$= (1-\gamma_1 L)(1-\gamma_0 L)x_0 - \gamma_0(1-\gamma_1 L)\xi_0 - \gamma_1 \xi_1$. In order to prove the inequality by induction, we need to show it for the base case $t_0 = 2$ first. Based on these calculations, it holds that:

$$x_2 = x_{t_0} = \prod_{j=0}^{t_0-1=1} (1-\gamma_j L)x_0 - \sum_{j=0}^{t_0-2=0} \gamma_j \prod_{i=j+1=1}^{t_0-1=1} (1-L\gamma_i)\xi_j - \gamma_{t_0-1}\xi_{t_0-1}$$

Thus for $t_0 = 2$, the equality holds. Suppose that it holds for some $t = k$. If we manage to show that the equality holds for $t = k+1$, then we are done. Thus, it is true that:

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

Need to show that:

$$x_{k+1} = \prod_{j=0}^{k+1-1}(1-L\gamma_j)x_0 - \sum_{j=0}^{k+1-2}\gamma_j \prod_{i=j+1}^{k+1-1}(1-L\gamma_i)\xi_j - \gamma_k\xi_k \tag{39}$$

From the algorithm it holds that $x_{k+1} = (1-\gamma_k L)x_k - \gamma_k \xi_k$. Moreover, using the inductive hypothesis for for x_k, we obtain:

$$x_{k+1} = (1-\gamma_k L)x_k - \gamma_k\xi_k =$$

$$= (1-\gamma_k L)\left[\prod_{j=0}^{k-1}(1-L\gamma_j)x_0 - \sum_{j=0}^{k-2}\gamma_j\xi_j \prod_{i=j+1}^{k-1}(1-L\gamma_i) - \gamma_{k-1}\xi_{k-1}\right] - \gamma_k\xi_k =$$

$$= \prod_{j=0}^{k-1}(1-L\gamma_j)x_0(1-\gamma_k L) - \sum_{j=0}^{k-2}\gamma_j\xi_j \prod_{i=j+1}^{k}(1-L\gamma_i) - (1-\gamma_k L)\gamma_{k-1}\xi_{k-1} - \gamma_k\xi_k =$$

$$= \prod_{j=0}^{k}(1-L\gamma_j)x_0 - \sum_{j=0}^{k-1}\gamma_j\xi_j \prod_{i=j+1}^{k}(1-L\gamma_i) - \gamma_k\xi_k$$

We just showed what we wanted to prove, mainly (39).

4) Fix $\delta \in (0,1)$. Show that with dimension $d \geq d_0 = O(log(T/\delta)$, for any $2 \leq t \leq T$, we have:

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

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

Solution 4):
Let $t \geq 2$. From Solution 3, it holds that $x_t = \prod_{j=0}^{t-1}(1-L\gamma_j)x_0-\sum_{j=0}^{t-2}\gamma_j\prod_{i=j+1}^{t-1}(1-L\gamma_i)\xi_j-\gamma_{t-1}\xi_{t-1}$. Since $\xi_j \sim N(0, \frac{\sigma^2}{d}I_d)$, then also x_t has a normal distribution since it is represented as the linear combination of random variables with normal distribution. Moreover, since $\mathbb{E}[\xi_j] = 0$, $\forall j$, then $\mathbb{E}[x_t] = \prod_{j=0}^{t-1}(1 - L\gamma_j)x_0$. Moreover, since ξ_i is independently sampled from ξ_j ($\forall j \neq i$) (Handout 11, page 12), then $Var(\sum_{i=0}^{t-2}\xi_i) = \sum_{i=0}^{t-2}Var(\xi_i)$, and since $Var(\alpha X) = \alpha^2 Var(X)$, where α is a constant, (and also because $Cov(\xi_{t,j}, \xi_{t,i}) = 0$, $\forall j \neq i, \forall t \geq 0$, where $\xi_{t,j}$ is the j-th component of ξ_t) then the covariance matrix of x_t is given by $\big(\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1 - L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2\big)\frac{I_d}{d}$. Thus $x_t \sim N(\prod_{j=0}^{t-1}(1 - L\gamma_j)x_0, \big(\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1 - L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2\big)\frac{I_d}{d})$. Now we use the hint on this normal distributed random variable by setting $\bar{\delta} = \frac{1}{2}$ (also $y = \prod_{j=0}^{t-1}(1 - L\gamma_j)x_0, x = x_t, \eta = \big(\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1 - L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2\big)$ like the notations in the hint) and finally obtain:

$$Pr\bigg(\bigg|\frac{||x_t||^2}{||\prod_{j=0}^{t-1}(1 - L\gamma_j)x_0||^2 + (\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1 - L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2)} - 1\bigg| \leq \frac{1}{2}\bigg)$$

$$\geq 1 - 4exp\bigg(-\frac{d}{96}\bigg) \tag{40}$$

Observe that the following holds:

$$\{\bigg|\frac{||x_t||^2}{||\prod_{j=0}^{t-1}(1 - L\gamma_j)x_0||^2 + (\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1 - L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2)} - 1\bigg| \leq \frac{1}{2}\} =$$

$$= \{\frac{1}{2} \leq \frac{||x_t||^2}{||\prod_{j=0}^{t-1}(1 - L\gamma_j)x_0||^2 + (\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1 - L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2)} \leq \frac{3}{2}\} \subseteq$$

$$\subseteq \{\frac{1}{2} \leq \frac{||x_t||^2}{||\prod_{j=0}^{t-1}(1 - L\gamma_j)x_0||^2 + (\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1 - L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2)}\} \tag{41}$$

This property (41) and the above inequality (40) imply the following:

$$Pr\bigg(\frac{1}{2} \leq \frac{||x_t||^2}{||\prod_{j=0}^{t-1}(1 - L\gamma_j)x_0||^2 + (\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1 - L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2)}\bigg) \geq$$

$$\geq Pr\bigg(\bigg|\frac{||x_t||^2}{||\prod_{j=0}^{t-1}(1 - L\gamma_j)x_0||^2 + (\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1 - L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2)} - 1\bigg| \leq \frac{1}{2}\bigg) \geq$$

$$\geq 1 - 4exp\bigg(-\frac{d}{96}\bigg)$$

In other words, the following inequality holds:

$$Pr\left(\frac{1}{2} \leq \frac{||x_t||^2}{||\prod_{j=0}^{t-1}(1-L\gamma_j)x_0||^2 + (\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1-L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2)}\right) \geq$$

$$\geq 1 - 4exp\left(-\frac{d}{96}\right) \tag{42}$$

Using the fact that $||x_0||^2 = \Delta/L$ The following events are equivalent:

$$\{\frac{1}{2} \leq \frac{||x_t||^2}{||\prod_{j=0}^{t-1}(1-L\gamma_j)x_0||^2 + (\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1-L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2)}\} =$$

$$= \{L^2||x_t||^2 \geq \frac{L}{2}(\prod_{j=0}^{t-1}(1-L\gamma_j)^2\Delta + L\sigma^2\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1-L\gamma_i)^2 + L\sigma^2\gamma_{t-1}^2)\} \tag{43}$$

Moreover, using the fact that $\nabla F(x_t) = Lx_t \rightarrow ||\nabla F(x_t)||^2 = L^2||x_t||^2$ and the above results (42) and (43) we get:

$$Pr[||\nabla F(x_t)||^2 \geq \frac{L}{2}(\prod_{j=0}^{t-1}(1-L\gamma_j)^2\Delta+L\sigma^2\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1-L\gamma_i)^2+L\sigma^2\gamma_{t-1}^2] \geq 1-4exp\left(-\frac{d}{96}\right) \tag{44}$$

Choose $d_0 = 96 \cdot log\big(\frac{4T}{\delta}\big) = O(log(T/\delta))$ (Why? - Because for $\frac{T}{\delta} \geq 4$, $log(\frac{4T}{\delta}) = log(4) + log(\frac{T}{\delta}) \leq log(\frac{T}{\delta}) + log(\frac{T}{\delta}) = 2log(\frac{T}{\delta}) \rightarrow d_0 = 96 \cdot log\big(\frac{4T}{\delta}\big) \leq 96 \cdot 2log(\frac{T}{\delta})$). Since $d \geq d_0 = 96 \cdot log\big(\frac{4T}{\delta}\big)$, the following holds:

$$-\frac{d}{96} \leq -\frac{d_0}{96} \quad \rightarrow \quad 1-4exp\left(-\frac{d}{96}\right) \geq 1-4exp\left(-\frac{d_0}{96}\right) = 1-4\cdot exp\left(log\left(\frac{\delta}{4T}\right)\right) = 1-\frac{\delta}{T}$$

We can use this property, i.e., that $1 - 4exp\big(-\frac{d}{96}\big) \geq 1 - \frac{\delta}{T}$, which holds $\forall d \geq d_0$ in (44) to conclude that by choosing $d_0 = 96 \cdot log\big(\frac{4T}{\delta}\big)$, we get:

$$Pr[||\nabla F(x_t)||^2 \geq \frac{L}{2}(\prod_{j=0}^{t-1}(1-L\gamma_j)^2\Delta+L\sigma^2\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1-L\gamma_i)^2+L\sigma^2\gamma_{t-1}^2] \geq 1-\frac{\delta}{T} \tag{45}$$

which is what we wanted to show.

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 that for all $2 \leq t \leq T$:

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

Solution 5):
From Question 4,(45) holds. We let $\gamma_t = \gamma \in (0, 1/L)$, and choose the same d as in 4 (i.e., $d \geq d_0$ with d_0 defined in Question 4). Thus using (45):

$$||\nabla F(x_t)||^2 \geq \frac{L}{2}\bigg(\Delta(1-L\gamma)^{2t} + L\sigma^2 \sum_{j=0}^{t-2} \gamma^2 \prod_{i=j+1}^{t-1} (1-L\gamma)^2 + L\sigma^2\gamma^2\bigg) \quad (\alpha)$$

with probability at least $1 - \frac{\delta}{T}$. Moreover, it holds that the following inequalities are equivalent:

$$||\nabla F(x_t)||^2 \geq \frac{L}{2}\bigg(\Delta(1-L\gamma)^{2t} + L\sigma^2 \sum_{j=0}^{t-2} \gamma^2 \prod_{i=j+1}^{t-1} (1-L\gamma)^2 + L\sigma^2\gamma^2\bigg) \leftrightarrow$$

$$\leftrightarrow ||\nabla F(x_t)||^2 \geq \frac{L}{2}\bigg(\Delta(1-L\gamma)^{2t} + L\sigma^2\gamma^2 \sum_{j=0}^{t-2} (1-L\gamma)^{2(t-j-1)} + L\sigma^2\gamma^2\bigg)$$

Let us now calculate the expression: $\sum_{j=0}^{t-2}(1-L\gamma)^{2(t-j-1)}$ as below (Note, I calculate $\sum_{t=0}^{t-2}\frac{1}{(1-L\gamma)^{2j}}$, easily by letting $S = \sum_{t=0}^{t-2}\frac{1}{(1-L\gamma)^{2j}}$ and $\frac{1}{(1-L\gamma)^2}S = \sum_{t=0}^{t-2}\frac{1}{(1-L\gamma)^{2(j+1)}}$. Substracting the two sides of the second equality from the first one, we have $(1 - \frac{1}{(1-L\gamma)^{2(t-1)}}) = (1 - \frac{1}{(1-L\gamma)^2})S$. Finally, doing the calculations we end up with $S = \frac{(1-L\gamma)^{2t}-(1-L\gamma)^2}{L\gamma(L\gamma-2)(1-L\gamma)^{2(t-1)}}$):

$$\sum_{j=0}^{t-2}(1-L\gamma)^{2(t-j-1)} = (1-L\gamma)^{2(t-1)} \cdot \sum_{j=0}^{t-2}(1-L\gamma)^{-2j} = (1-L\gamma)^{2(t-1)} \frac{((1-L\gamma)^{2(t-1)}-1)\cdot(1-L\gamma)^2}{(1-L\gamma)^{2(t-1)}\cdot L\gamma(L\gamma-2)}$$

$$= \frac{(1-L\gamma)^{2t}-(1-L\gamma)^2}{L\gamma(L\gamma-2)}$$

We obtain therefrom the equalities below:

$$\frac{L}{2}\bigg(\Delta(1-L\gamma)^{2t} + L\sigma^2\gamma^2 \sum_{j=0}^{t-2}(1-L\gamma)^{2(t-j-1)} + L\sigma^2\gamma^2\bigg) = \tag{46}$$

$$= \frac{L}{2}\bigg(\Delta(1-L\gamma)^{2t} + L\sigma^2\gamma^2(1-L\gamma)^{2(t-1)}\frac{[(1-L\gamma)^{2t}-(1-L\gamma)^2]}{L\gamma(L\gamma-2)(1-L\gamma)^{2(t-1)}} + L\sigma^2\gamma^2\bigg) = \tag{47}$$

$$= \frac{L\Delta}{2}(1-L\gamma)^{2t} + \frac{L\gamma\sigma^2}{2}\bigg[\frac{(1-L\gamma)^{2t}-(1-2L\gamma+L^2\gamma^2)}{(L\gamma-2)} + \frac{L^2\gamma^2-2L\gamma}{(L\gamma-2)}\bigg] = \tag{48}$$

$$= \frac{L}{2}\Delta(1-L\gamma)^{2t} + \frac{L\gamma\sigma^2}{2(2-L\gamma)}[1-(1-L\gamma)^{2t}] \tag{49}$$

Next, observe that the event $\{||\nabla F(x_t)||^2 \leq \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2-L\gamma)}\}\}$ is the same as the event $\{||\nabla F(x_t)||^2 \leq \frac{L\Delta}{2}$ and $||\nabla F(x_t)||^2 \leq \frac{L\gamma\sigma^2}{2(2-L\gamma)}\}$. Moreover. since $\gamma < \frac{1}{L}$ ($\gamma L < 1 \rightarrow (1-L\gamma) < 1$. Also, $(1-L\lambda)^{2t} > 0$), it holds that $[1-(1-L\gamma)^{2t}] > 0$. Consequently we get the following relations among events:

$$\{||\nabla F(x_t)|| \leq \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2-L\gamma)}\}\} = \{||\nabla F(x_t)||^2 \leq \frac{L\Delta}{2} \text{ and } ||\nabla F(x_t)||^2 \leq \frac{L\gamma\sigma^2}{2(2-L\gamma)}\}$$

$$= \{(1-L\gamma)^{2t}||\nabla F(x_t)||^2 \leq (1-L\gamma)^{2t}\frac{L\Delta}{2} \text{ and } [1-(1-L\gamma)^{2t}]\cdot||F(x_t)||^2 \leq [1-(1-L\gamma)^{2t}]\frac{L\gamma\sigma^2}{2(2-L\gamma)}\}$$

Add both sides of the two inequalities → direct result of this event, thus this event is a subset of:

$$\subseteq \{||\nabla F(x_t)||^2 = ([1-L\gamma]^{2t}+[1-(1-L\gamma)^{2t}])\cdot||\nabla F(x_t)||^2 \leq (1-L\gamma)^{2t}\frac{L\Delta}{2}+[1-(1-L\gamma)^{2t}]\frac{L\gamma\sigma^2}{2(2-L\gamma)}\} \tag{50}$$

We know from (46-49) and (α) that:

$$Pr[||\nabla F(x_t)||^2 \geq (1-L\gamma)^{2t}\frac{L\Delta}{2} + [1-(1-L\gamma)^{2t}]\frac{L\gamma\sigma^2}{2(2-L\gamma)}] \geq 1 - \frac{\delta}{T}$$

This suggest that the following is true, since $\{||\nabla F(x_t)|| \geq (1-L\gamma)^{2t}\frac{L\Delta}{2} + [1-(1-L\gamma)^{2t}]\frac{L\gamma\sigma^2}{2(2-L\gamma)}\}$ is the complements event of $\{||\nabla F(x_t)|| < (1-L\gamma)^{2t}\frac{L\Delta}{2} + [1-(1-L\gamma)^{2t}]\frac{L\gamma\sigma^2}{2(2-L\gamma)}\}$ (Moreover, since $||\nabla F(x_t)|| = Lx_t$ is a continuous random variable (it is normal) it holds true that
$Pr[\{||\nabla F(x_t)|| = (1-L\gamma)^{2t}\frac{L\Delta}{2} + [1-(1-L\gamma)^{2t}]\frac{L\gamma\sigma^2}{2(2-L\gamma)}\}] = 0$, Finally using the property that if A is an event with complement A^c and $Pr(A) \geq a$, then $Pr(A^c) = 1 - Pr(A) \leq 1 - a$, we get):

$$Pr[||\nabla F(x_t)|| \leq (1-L\gamma)^{2t}\frac{L\Delta}{2} + [1-(1-L\gamma)^{2t}]\frac{L\gamma\sigma^2}{2(2-L\gamma)}] \leq \frac{\delta}{T}$$

Therefore, using this and (50), it holds:

$$Pr[\{||\nabla F(x_t)|| \leq \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2-L\gamma)}\}] \leq$$

$$\leq Pr[||\nabla F(x_t)|| \leq (1-L\gamma)^{2t}\frac{L\Delta}{2} + [1-(1-L\gamma)^{2t}]\frac{L\gamma\sigma^2}{2(2-L\gamma)}] \leq \frac{\delta}{T} \quad (\beta)$$

Moreover, using this and the De-Morgan law which states that if $A_1, A_2, ..., A_k$ are events and $A_1^c, A_2^c, ..., A_k^c$ are the respective complement sets of $A_1, A_2, ..., A_k$, it is true that $\cap_{i=1}^{k} A_i^c = (\cup_{i=1}^{k} A_i)^c$. Finally we have:

$$Pr[\{\forall t : 2 \leq t \leq T;\ \ ||\nabla F(x_t)||^2 \geq \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2-L\gamma)}\}] = \tag{51}$$

$$= Pr[\cap_{t=2}^{T}\{||\nabla F(x_t)||^2 \geq \min \frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2-L\gamma)}\}] = \tag{52}$$

$$= Pr[(\cup_{t=2}^{T}\{||\nabla F(x_t)||^2 \leq \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2-L\gamma)}\}\})^c] = \tag{53}$$

$$= 1 - Pr[(\cup_{t=2}^{T}\{||\nabla F(x_t)||^2 \leq \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2-L\gamma)}\}\})] \geq \tag{54}$$

$$\geq 1 - \sum_{t=2}^{T} Pr[||\nabla F(x_t)||^2 \leq \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2-L\gamma)}] \geq 1 - \frac{(T-1)}{T}\delta \geq 1 - \delta \tag{55}$$

In (52)-(53) we apply the De-Morgan law. In (54) I use the property that given A an event and A^c its complement, $Pr[A] = 1 - Pr[A^c]$. In the transition of (54) to (55), I use the fact that $Pr[\cup_{i=1}^{k} A_i] \leq \sum_{i=1}^{k} Pr[A_i]$. Moreover in the second and third inequality of (55) we use the result in (β).
We just proved the desired result, mainly that $2 \leq \forall t \leq T$:

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

with probability at least $1 - \delta$.

Exercise 3: Modified Extragradient

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 the 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 $O(\frac{1}{T})$ for averaged iterates, but only $O(\frac{1}{\sqrt{T}})$ 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 $t \geq 0$ are given by:

$$z_{t+\frac{1}{2}} = z_t - \eta F(z_t) + \frac{1}{t+1}(z_0 - z_t)$$

$$z_{t+1} = z_t - \eta F(z_{t+\frac{1}{2}}) + \frac{1}{t+1}(z_0 - z_t)$$

where $\eta > 0$ is the stepsize.
Since f is convex-concave and smooth, we have the following properties for any $z, \hat{z} \in \mathbb{R}^{d_1+d_2}$:

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

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

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

1) Define:

$$A_t = \langle F(z_{t+1}) - F(z_t), \frac{1}{t+1}(z_0 - z_t) - \eta F(z_{t+\frac{1}{2}}) \rangle;$$

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

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

Solution 1):

- Based on the conditions of the problem, the following holds:

$$z_{t+1} = z_t - \eta F(z_{t+\frac{1}{2}}) + \frac{1}{t+1}(z_0 - z_t) \quad \leftrightarrow \quad \frac{1}{t+1}(z_0 - z_t) - \eta F(z_{t+\frac{1}{2}}) = z_{t+1} - z_t$$

We use the above result on the right and get:

$$A_t = \langle F(z_{t+1}) - F(z_t), \frac{1}{t+1}(z_0 - z_t) - \eta F(z_{t+\frac{1}{2}}) \rangle = \langle F(z_{t+1}) - F(z_t), z_{t+1} - z_t \rangle$$

Since for any $z, \hat{z} \in \mathbb{R}^{d_1+d_2}$, it holds that $\langle F(z) - F(\hat{z}), z - \hat{z} \rangle \geq 0$, then given $z_{t+1}, z_t \in \mathbb{R}^{d_1+d_1}$, it holds that $A_t = \langle F(z_{t+1}) - F(z_t), z_{t+1} - z_t \rangle \geq 0$. We just showed that A_t is non-negative.

- Now, we show that B_t is non-negative. It is given that the following hold:

$$z_{t+\frac{1}{2}} = z_t - \eta F(z_t) + \frac{1}{t+1}(z_0 - z_t) \tag{56}$$

$$z_{t+1} = z_t - \eta F(z_{t+\frac{1}{2}}) + \frac{1}{t+1}(z_0 - z_t) \tag{57}$$

By substracting from the both sides of (57) the sides of (56), we get:

$$\eta F(z_t) - \eta F(z_{t+\frac{1}{2}}) = z_{t+1} - z_{t+\frac{1}{2}} \tag{58}$$

Next, since it is true that $||F(z_{t+\frac{1}{2}}) - F(z_{t+1})||^2 \leq L^2||z_{t+\frac{1}{2}} - z_{t+1}||^2$, the following is valid:

$$-\frac{1}{L^2}||F(z_{t+\frac{1}{2}}) - F(z_{t+1})||^2 \geq -||z_{t+\frac{1}{2}} - z_{t+1}||^2$$

Using this result and (58) we obtain:

$$B_t = ||\eta F(z_t) - \eta F(z_{t+\frac{1}{2}})||^2 - \frac{1}{L^2}||F(z_{t+\frac{1}{2}}) - F(z_{t+1})||^2 \geq ||z_{t+1} - z_{t+\frac{1}{2}}||^2 - ||z_{t+\frac{1}{2}} - z_{t+1}||^2 = 0$$

Thus we just showed that $B_t \geq 0$, i.e., B_t is non-negative.

2) Show that for any $t \geq 1$, it holds that:

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

Solution 2):

Since, $A_t, B_t \geq 0$, we use the hint:

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

We expand the right hand side of this inequality as below (detailed equalities):

$$V_{t+1} - V_t + \eta t(t+1)A_t + \frac{t(t+1)}{2}B_t = \frac{(t+1)(t+2)}{2}\eta^2||F(z_{t+1})||^2 + (t+1)\langle \eta F(z_{t+1}), z_{t+1} - z_0\rangle -$$

$$-\frac{t(t+1)}{2}\eta^2||F(z_t)||^2 - t\eta\langle F(z_t), z_t - z_0\rangle + \eta t(t+1)\langle F(z_{t+1}) - F(z_t), \frac{1}{t+1}(z_0 - z_t) - \eta F(z_{t+\frac{1}{2}})\rangle +$$

$$+\frac{t(t+1)}{2}[\eta^2||F(z_t) - F(z_{t+\frac{1}{2}})||^2 - \frac{1}{L^2}||F(z_{t+\frac{1}{2}}) - F(z_{t+1})||^2] =$$

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

$$-t\eta\langle F(z_t), z_t - z_0\rangle + \eta t\langle F(z_{t+1}), z_0 - z_t\rangle - \eta t\langle F(z_t), z_0 - z_t\rangle - \eta^2 t(t+1)\langle F(z_{t+1}), F(z_{t+\frac{1}{2}})\rangle +$$

$$+\eta^2 t(t+1)\langle F(z_t), F(z_{t+\frac{1}{2}})\rangle + \frac{t(t+1)}{2}\eta^2||F(z_t)||^2 - \eta^2 t(t+1)\langle F(z_t), F(z_{t+\frac{1}{2}})\rangle + \frac{\eta^2 t(t+1)}{2}||F(z_{t+\frac{1}{2}})||^2 -$$

$$-\frac{t(t+1)}{2L^2}||F(z_{t+\frac{1}{2}})||^2 + \frac{t(t+1)}{L^2}\langle F(z_{t+\frac{1}{2}}), F(z_{t+1})\rangle - \frac{t(t+1)}{2L^2}||F(z_{t+1})||^2$$

$$= \left(\frac{(t+1)(t+2)}{2}\eta^2 - \frac{t(t+1)}{2L^2}\right)||F(z_{t+1})||^2 + \langle F(z_{t+1}), F(z_{t+\frac{1}{2}})\rangle[-\eta^2 t(t+1) + \frac{t(t+1)}{L^2}] +$$

$$+||F(z_{t+\frac{1}{2}})||^2\left[\frac{\eta^2 t(t+1)}{2} - \frac{t(t+1)}{2L^2}\right] + \langle \eta F(z_{t+1}), (t+1)z_{t+1} - tz_t - z_0\rangle$$

- Next use $z_{t+1} = z_t - \eta F(z_{t+\frac{1}{2}}) + \frac{1}{t+1}(z_0 - z_t) \rightarrow (t+1)z_{t+1} - tz_t - z_0 = -\eta t F(z_{t+\frac{1}{2}})$

$$= \left(\frac{(t+1)(t+2)}{2}\eta^2 - \frac{t(t+1)}{2L^2}\right)||F(z_{t+1})||^2 + \langle F(z_{t+1}), F(z_{t+\frac{1}{2}})\rangle[-\eta^2 t(t+1) + \frac{t(t+1)}{L^2} - \eta^2(t+1)] +$$

$$+||F(z_{t+\frac{1}{2}})||^2\left[\frac{\eta^2 t(t+1)}{2}-\frac{t(t+1)}{2L^2}\right]$$

Now, we use the equality: $\frac{(t+1)(t+2)}{2}\eta^2-\frac{t(t+1)}{2L^2}=\frac{t(t+1)}{2}\eta^2+\frac{\eta^2(t+1)}{2}-\frac{t(t+1)}{2L^2}+\frac{(t+1)\eta^2}{2}$ and the above last expression becomes:

$$\left[\frac{t(t+1)}{2}\eta^2+\frac{(t+1)}{2}\eta^2-\frac{t(t+1)}{2L^2}\right]||F(z_{t+1})||^2+\left[\frac{t(t+1)}{L^2}-\eta^2 t(t+1)-\eta^2(t+1)\right]\langle F(z_{t+1}),F(z_{t+\frac{1}{2}})\rangle+$$

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

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

$$+\frac{\eta^2(t+1)}{2}||F(z_{t+1})||^2 \tag{59}$$

In the last expression of the above equalities, we use the fact that $||F(z_{t+1})-F(z_{t+\frac{1}{2}})||^2=||F(z_{t+1})||^2-2\langle F(z_{t+\frac{1}{2}}),F(z_{t+1})\rangle+||F(z_{t+\frac{1}{2}})||^2$. Using all the above calculations we get an upper bound for $V_{t+1}-V_t$ as below:

$$V_{t+1}-V_t\le V_{t+1}-V_t+\eta t(t+1)A_t+\frac{t(t+1)}{2}B_t= \tag{60}$$

$$=\left(\frac{\eta^2 t(t+1)}{2}-\frac{t(t+1)}{2L^2}+\frac{\eta^2(t+1)}{2}\right)||F(z_{t+\frac{1}{2}})-F(z_{t+1})||^2- \tag{61}$$

$$-\frac{\eta^2(t+1)}{2}||F(z_{t+\frac{1}{2}})||^2+\frac{\eta^2(t+1)}{2}||F(z_{t+1})||^2 \tag{62}$$

It holds that

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

$$\le\max_{x\in\mathbb{R}^{d_1+d_2}}\left[\left(\frac{\eta^2 t(t+1)}{2}-\frac{t(t+1)}{2L^2}+\frac{\eta^2(t+1)}{2}\right)||x-F(z_{t+1})||^2-\frac{\eta^2(t+1)}{2}||x||^2+\frac{\eta^2(t+1)}{2}||F(z_{t+1})||^2\right]$$

The function $g(x)=\left(\frac{\eta^2 t(t+1)}{2}-\frac{t(t+1)}{2L^2}+\frac{\eta^2(t+1)}{2}\right)||x-F(z_{t+1})||^2-\frac{\eta^2(t+1)}{2}||x||^2+\frac{\eta^2(t+1)}{2}||F(z_{t+1})||^2$ is concave (w.r.t x) because $\nabla_x g(x)=(\eta^2 t(t+1)-\frac{t(t+1)}{L^2}+\eta^2(t+1))(x-F(z_{t+1}))-\eta^2(t+1)x\rightarrow\rightarrow\nabla_x^2 g(x)=\eta^2 t(t+1)-\frac{t(t+1)}{L^2}\le\frac{1}{3L^2}t(t+1)-\frac{t(t+1)}{L^2}=-\frac{2t(t+1)}{3L^2}<0$ (we used the fact that $\eta<\frac{1}{\sqrt{3}L}$), thus it reaches its maximum (w.r.t the value of x) at x_0, when $\nabla g(x_0)=0$. In other words, x_0 satisfies:

$$(\eta^2 t(t+1)-\frac{t(t+1)}{L^2}+\eta^2(t+1))(x_0-F(z_{t+1}))-\eta^2(t+1)x_0=0$$

By doing the calculations we get that the value of x_0, should be:

$$x_0 = \frac{\left(-\frac{t(t+1)}{L^2} + \eta^2 t(t+1) + \eta^2(t+1)\right)}{\left(\eta^2 t(t+1) - \frac{t(t+1)}{L^2}\right)} F(z_{t+1}) = \left(1 - \frac{\eta^2(t+1)}{\frac{t(t+1)}{L^2} - \eta^2 t(t+1)}\right) F(z_{t+1})$$

Denote $y = \frac{\eta^2(t+1)}{\frac{t(t+1)}{L^2} - \eta^2 t(t+1)}$. Since $\eta < \frac{1}{\sqrt{3}L}$, it is true that $\frac{1}{L^2} > 3\eta^2$. Therefore it holds that $\frac{t(t+1)}{L^2} - \eta^2 t(t+1) > 3\eta^2 t(t+1) - \eta^2 t(t+1) = 2\eta^2 t(t+1) \geq 0$. Moreover, since $\eta^2(t+1) \geq 0$, it holds that $y \geq 0$. Furthermore, it holds that $y < 1$ since it is equivalent to showing that $\frac{t(t+1)}{L^2} - \eta^2 t(t+1) - \eta^2(t+1) > 0$. This holds because $\frac{t(t+1)}{L^2} - \eta^2 t(t+1) - \eta^2(t+1) > 3\eta^2 t(t+1) - \eta^2 t(t+1) - \eta^2(t+1) = 2\eta^2 t(t+1) - \eta^2(t+1) \geq 2\eta^2(t+1) - \eta^2(t+1) = \eta^2(t+1) \geq 0$, given that $t \geq 1$ and $\frac{1}{L^2} > 3\eta^2$. We therefore conclude that

$$\left(\frac{\eta^2 t(t+1)}{2} - \frac{t(t+1)}{2L^2} + \frac{\eta^2(t+1)}{2}\right)||x - F(z_{t+1})||^2 - \frac{\eta^2(t+1)}{2}||x||^2 + \frac{\eta^2(t+1)}{2}||F(z_{t+1})||^2 \tag{63}$$

achieves its maximum value for $x = (1-y)F(z_{t+1})$ with y defined as above ($0 \leq y < 1$). We substitute $x = (1-y)F(z_{t+1})$ in (63) and get its maximum value *w.r.t* x to be:

$$\begin{aligned}
&\left(\frac{\eta^2 t(t+1)}{2} - \frac{t(t+1)}{2L^2} + \frac{\eta^2(t+1)}{2}\right)||yF(z_{t+1})||^2 - \frac{\eta^2(t+1)}{2}(1-y)^2||F(z_{t+1})||^2 + \\
&+ \frac{\eta^2(t+1)}{2}||F(z_{t+1})||^2 = ||F(z_{t+1})||^2\left[\frac{y^2\eta^2 t(t+1)}{2} - \frac{y^2 t(t+1)}{2L^2} + \eta^2(t+1)y\right] \\
&= \frac{(t+1)}{2}||\eta F(z_{t+1})||^2\left[y^2 t - \frac{y^2 t}{L^2\eta^2} + 2y\right] \qquad (64)
\end{aligned}$$

It holds that $y^2 t - \frac{y^2 t}{L^2\eta^2} + 2y = 2y - \frac{(1-\eta^2L^2)y^2 t}{L^2\eta^2}$. Moreover, $Arithmetic - Mean - Geometric - Mean$ inequality states that if $a \geq 0$, $b \geq 0$, then, $a + b \geq 2\sqrt{ab}$. Thus since $\frac{(1-\eta^2L^2)y^2 t}{L^2\eta^2} > 0$, and $\frac{L^2\eta^2}{(1-\eta^2L^2)t} > 0$ (since $\eta L < \frac{1}{\sqrt{3}}$, and $y \geq 0$, and $t, L > 0$) then $\frac{(1-\eta^2L^2)y^2 t}{L^2\eta^2} + \frac{L^2\eta^2}{(1-\eta^2L^2)t} \geq 2y \rightarrow 2y - \frac{(1-\eta^2L^2)y^2 t}{L^2\eta^2} \leq \frac{\eta^2L^2}{t(1-\eta^2L^2)}$. Therefore, we could further upper bound (64) by:

$$\frac{t+1}{2}||\eta F(z_{t+1})||^2 \frac{L^2\eta^2}{(1-\eta^2L^2)t} = \frac{(t+1)\eta^2L^2}{2t(1-\eta^2L^2)}||\eta F(z_{t+1})||^2$$

Combining all the previous arguments we conclude that: $V_{t+1} - V_t \leq \frac{(t+1)\eta^2L^2}{2t(1-\eta^2L^2)}||\eta F(z_{t+1})||^2$, which is what we wanted to show.

3) By the recursion of V_t, show that for $T \geq 2$,

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

You can use the inequality $V_1 \leq (2\eta L + \eta^2 L^2)||z_0 - z^*||^2$ without proof.
Solution 3):
From Question 2, it is true that for $t \geq 1$:

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

We sum both sides of the above inequality over $t = 1, ..., T-1$ and get:

$$\sum_{t=1}^{T-1}(V_{t+1} - V_t) \leq \sum_{t=1}^{T-1} \frac{(t+1)\eta^2 L^2}{2t(1-\eta^2 L^2)}||\eta F(z_{t+1})||^2$$

$$\sum_{t=2}^{T} V_t - \sum_{t=1}^{T-1} V_t \leq \sum_{t=2}^{T} \frac{t\eta^2 L^2}{2(t-1)(1-\eta^2 L^2)}||\eta F(z_t)||^2$$

$$V_T - V_1 \leq \sum_{t=2}^{T-1} \frac{t\eta^2 L^2}{2(t-1)(1-\eta^2 L^2)}||\eta F(z_t)||^2 + \frac{T(\eta^2 L^2)}{2(T-1)(1-\eta^2 L^2)}||\eta F(z_T)||^2 \quad (65)$$

It holds that $\eta < \frac{1}{\sqrt{3}}$, which implies that $\eta^2 L^2 < \frac{1}{3} \rightarrow -\eta^2 L^2 > -\frac{1}{3} \rightarrow$
$\rightarrow 1 - \eta^2 L^2 > \frac{2}{3} \rightarrow \frac{1}{(1-\eta^2 L^2)} < \frac{3}{2} \rightarrow \frac{\eta^2 L^2}{1-\eta^2 L^2} < \frac{1}{3} \cdot \frac{3}{2} = \frac{1}{2}$. Thus, finally it holds that $\frac{\eta^2 L^2}{1-\eta^2 L^2} < \frac{1}{2}$. Since $T \geq 2$, it holds that $\frac{T}{2(T-1)} \leq 1$. Combining these two results, it holds that $\frac{T(\eta^2 L^2)}{2(T-1)(1-\eta^2 L^2)}||\eta F(z_T)||^2 \leq \frac{1}{2}||\eta F(z_T)||^2$. Finally using this inequality and (65), we obtain:

$$V_T - V_1 \leq \sum_{t=2}^{T-1} \frac{t\eta^2 L^2}{2(t-1)(1-\eta^2 L^2)}||\eta F(z_t)||^2 + \frac{1}{2}||\eta F(z_T)||^2 \quad (66)$$

Given that $V_T = \frac{T(T+1)}{2}||\eta F(z_T)||^2 + T\langle \eta F(z_T), z_T - z_0\rangle$ and rearranging terms of inequality (66), we get:

$$\frac{T(T+1)}{2}||\eta F(z_T)||^2 + T\langle \eta F(z_T), z_T - z_0\rangle - \frac{1}{2}||\eta F(z_T)||^2 \leq \sum_{t=2}^{T-1} \frac{t\eta^2 L^2}{2(t-1)(1-\eta^2 L^2)}||\eta F(z_t)||^2 + V_1$$

$$\leq \sum_{t=2}^{T-1} \frac{t\eta^2 L^2}{2(t-1)(1-\eta^2 L^2)}||\eta F(z_t)||^2 + (2\eta L + \eta^2 L^2)||z_0 - z^*||^2 \quad (67)$$

In (67), we used the hint $V_1 \leq (2\eta L + \eta^2 L^2)||z_0 - z^*||^2$. Using (67), we get:

$$\left(\frac{T(T+1)}{2} - \frac{1}{2}\right)||\eta F(z_T)||^2 \leq \tag{68}$$

$$\leq \sum_{t=2}^{T-1} \frac{t\eta^2 L^2}{2(t-1)(1-\eta^2 L^2)}||\eta F(z_t)||^2 + (2\eta L + \eta^2 L^2)||z_0 - z^*||^2 - \langle \eta T F(z_T), z_T - z_0\rangle \tag{69}$$

Now, we expand further on the term $-\langle \eta T F(z_T), z_T - z_0\rangle$. Using the fact that $F(z^*) = 0$, we obtain:

$$-\langle \eta T F(z_T), z_T - z_0\rangle = -\eta T\langle F(z_T), z_T - z_0\rangle =$$

$$= -\eta T\langle F(z_T) - F(z^*), z_T - z^*\rangle - \langle \eta T F(z_T), z^* - z_0\rangle \quad (\S)$$

It holds from the condition of the problem that $\langle F(z_T) - F(z^*), z_T - z^*\rangle \geq 0$, which implies that $-\eta T\langle F(z_T) - F(z^*), z_T - z^*\rangle \leq 0$. Moreover, using the hint of this subquestion, and letting $\lambda = 1$, we obtain the following valid inequalities:
$\langle \eta T F(z_T), z^* - z_0\rangle \geq -\frac{1}{4}||\eta T F(z_T)||^2 - ||z^* - z_0||^2 \leftrightarrow$
$\leftrightarrow -\langle \eta T F(z_T), z^* - z_0\rangle \leq \frac{T^2}{4}||\eta F(z_T)||^2 + ||z^* - z_0||^2$. Finally, these arguments show that using (§), we get the inequality result:

$$-\langle \eta T F(z_T), z_T - z_0\rangle \leq \frac{T^2}{4}||\eta F(z_T)||^2 + ||z^* - z_0||^2 \tag{70}$$

Using this inequality in (69), and the property that $\frac{t}{2(t-1)} \leq 1$ inside the sum (since inside the sum, we consider $t \geq 2$) we finally get:

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

which is equivalent to:

$$\left[\frac{2T(T+1) - 2 - T^2}{4}\right]||\eta F(z_T)||^2 \leq (1+\eta L)^2||z_0 - z^*||^2 + \sum_{t=2}^{T-1} \frac{\eta^2 L^2}{(1-\eta^2 L^2)}||\eta F(z_t)||^2$$

Observe that $2T(T+1) - 2 - T^2 = T^2 + 2T - 2 \geq T^2$ Since $2T - 2 \geq 0$ (since $T \geq 2$). Thus, finally we get that $\frac{T^2}{4} \leq \frac{[2T(T+1)-2-T^2]}{4}$. We substitute this result in the above inequality and get:

$$\frac{T^2}{4}||\eta F(z_T)||^2 \leq \left[\frac{2T(T+1) - 2 - T^2}{4}\right]||\eta F(z_T)||^2 \leq$$

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

From (§§), we just obtained what we wanted to show, mainly that $\frac{T^2}{4}||\eta F(z_T)||^2 \leq (1+\eta L)^2||z_0 - z^*||^2 + \sum_{t=2}^{T-1} \frac{\eta^2 L^2}{(1-\eta^2 L^2)}||\eta F(z_t)||^2$

4) Show by induction that for $T \geq 2$, we have:

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

Solution 4):
We show it by induction. Start with $T = 2$, we need to show that:

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

From Question 2, it holds that:

$$V_2 - V_1 \leq \frac{\eta^2 L^2}{(1-\eta^2 L^2)}||\eta F(z_2)||^2 \tag{71}$$

Since $V_2 = 3||\eta F(z_2)||^2 + 2\langle \eta F(z_2), z_2 - z_0\rangle$, using (71), we obtain:

$$3||\eta F(z_2)||^2 - \frac{\eta^2 L^2}{(1-\eta^2 L^2)}||\eta F(z_2)||^2 \leq V_1 - 2\langle \eta F(z_2), z_2 - z_0\rangle \tag{72}$$

From Question 3, the hint suggests that $V_1 \leq (2\eta L + \eta^2 L^2)||z_0 - z^*||^2$. Moreover, from the previous subexercise, in the inequality (70), we substitute $T = 2$ and it holds that $-2\langle \eta F(z_2), z_2 - z_0\rangle \leq ||\eta F(z_2)||^2 + ||z^* - z_0||^2$. We use these results in (72) to get:

$$2||\eta F(z_2)||^2 - \frac{\eta^2 L^2}{(1-\eta^2 L^2)}||\eta F(z_2)||^2 \leq (1+\eta L)^2||z^* - z_0||^2$$

equivalent to the inequality:

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

In order to prove our aim, it is enough to show that:

$$\frac{(1+\eta L)^2(1-\eta^2 L^2)}{\eta^2(2-3\eta^2 L^2)} \leq \frac{(1+\eta L)^2}{\eta^2(1-3\eta^2 L^2)}$$

equivalent to showing the following:

$$1 - 4\eta^2 L^2 + 3\eta^4 L^4 \leq 2 - 3\eta^2 L^2 \quad \leftrightarrow \quad 3\eta^4 L^4 \leq \eta^2 L^2 + 1$$

The inequality on the right surely holds since $\eta < \frac{1}{\sqrt{3}L}$ implying that $3\eta^2 L^2 \leq 1$ which further implies that $3\eta^4 L^4 \leq \eta^2 L^2 \leq \eta^2 L^2 + 1$. Therefore, this completes

the proof that $||F(z_2)||^2 \leq \frac{(1+\eta L)^2}{\eta^2(1-3\eta^2L^2)}||z^* - z_0||^2$.
Now suppose that the inequality holds $\forall t = k$, for $k \geq 2$ and we must then show it for $T = k+1$. From question 3, it holds that:

$$\frac{(k+1)^2}{4}||\eta F(z_{k+1})||^2 \leq (1+\eta L)^2||z^* - z_0||^2 + \sum_{t=2}^{k} \frac{\eta^2 L^2}{(1-\eta^2L^2)}||\eta F(z_t)||^2 \quad (73)$$

We now use the inductive hypothesis for all $t = 2, ..., k$, mainly that

$$||F(z_t)||^2 \leq \frac{4(1+\eta L)^2}{\eta^2(1-3\eta^2L^2)t^2}||z^* - z_0||^2 \ \leftrightarrow \ ||\eta F(z_t)||^2 \leq \frac{4(1+\eta L)^2}{(1-3\eta^2L^2)t^2}||z^* - z_0||^2$$

We use the above result on the right in (73) and finally get:

$$\frac{(k+1)^2}{4}||\eta F(z_{k+1})||^2 \leq (1+\eta L)^2||z^* - z_0||^2 + \sum_{t=2}^{k} \frac{\eta^2 L^2}{(1-\eta^2L^2)} \frac{4(1+\eta L)^2}{(1-3\eta^2L^2)t^2}||z^* - z_0||^2 = \quad (74)$$

$$= ||z^* - z_0||^2 \left[(1+\eta L)^2 + \frac{\eta^2 L^2 4(1+\eta L)^2}{(1-\eta^2L^2)(1-3\eta^2L^2)} \sum_{t=2}^{k} \frac{1}{t^2}\right] = \quad (75)$$

$$= ||z^* - z_0||^2 \frac{(1+\eta L)^2}{(1-3\eta^2L^2)} \left[(1-3\eta^2L^2) + \frac{4\eta^2L^2}{(1-\eta^2L^2)} \sum_{t=2}^{k} \frac{1}{t^2}\right] \quad (76)$$

We aim now to bound $\sum_{t=2}^{k} \frac{1}{t^2}$. Well, observe that $t^2 \geq t(t-1) \rightarrow \frac{1}{t^2} \leq \frac{1}{t(t-1)} \rightarrow \sum_{t=2}^{k} \frac{1}{t^2} \leq \sum_{t=2}^{k} \frac{1}{t(t-1)} = \sum_{t=2}^{k} (\frac{1}{t-1} - \frac{1}{t}) = 1 - \frac{1}{k} \leq 1$. In this way, we obtain an upper bound for the expression inside the square brackets in (76) as below:

$$\left[(1-3\eta^2L^2) + \frac{4\eta^2L^2}{(1-\eta^2L^2)} \sum_{t=2}^{k} \frac{1}{t^2}\right] \leq [1-3\eta^2L^2 + \frac{4\eta^2L^2}{(1-\eta^2L^2)}] = \frac{(1-3\eta^2L^2)(1-\eta^2L^2) + 4\eta^2L^2}{(1-\eta^2L^2)} =$$

$$= \frac{1+3\eta^4L^4}{1-\eta^2L^2}$$

We can further bound $\frac{1+3\eta^4L^4}{1-\eta^2L^2}$. It holds that $\eta^2L^2 < \frac{1}{3} \rightarrow \eta^4L^4 < \frac{1}{9} \rightarrow 3\eta^4L^4 < \frac{1}{3}$. It also hold that $-\eta^2L^2 > -\frac{1}{3} \rightarrow 1 - \eta^2L^2 > \frac{2}{3} \rightarrow \frac{1}{1-\eta^2L^2} < \frac{3}{2}$. Combining all of these inequalities, we get $\frac{1+3\eta^4L^4}{1-\eta^2L^2} < \frac{(1+\frac{1}{3})}{\frac{3}{2}} = \frac{8}{9} < 1$ Thus, we finally get:

$$\left[(1-3\eta^2L^2) + \frac{4\eta^2L^2}{(1-\eta^2L^2)} \sum_{t=2}^{k} \frac{1}{t^2}\right] \leq \frac{1+3\eta^4L^4}{1-\eta^2L^2} < 1$$

We use this inequality in (76) to get:

$$\frac{(k+1)^2}{4}||\eta F(z_{k+1})||^2 \leq ||z^* - z_0||^2 \frac{(1+\eta L)^2}{(1-3\eta^2 L^2)}$$

which in turn directly shows the desired claim:

$$||F(z_{k+1})||^2 \leq \frac{4(1+\eta L)^2}{\eta^2 (k+1)^2 (1-3\eta^2 L^2)}||z^* - z_0||^2$$

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 $T \geq 2$, we have:

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

Solution 5):

Since $f(\cdot, y)$ is convex and $f(x, \cdot)$ is concave, we have:

$$f(x_1, y) \geq f(x_2, y) + \nabla_x f(x_2, y)^\top (x_1 - x_2) \tag{77}$$

$$-f(x, y_1) \geq -f(x, y_2) - \nabla_y f(x, y_2)^\top (y_1 - y_2) \tag{78}$$

We let $x_1 = x$, $x_2 = x_T$ and $y = y_T$ in (77), and multiply both sides by -1, to get:

$$-f(x, y_T) \leq -f(x_T, y_T) - \nabla_x f(x_T, y_T)^\top (x - x_T) \tag{79}$$

Let $x = x_T$, $y_1 = y$ and $y_2 = y_T$ in (78) and multiply both sides by -1, to finally get:

$$f(x_T, y) \leq f(x_T, y_T) + \nabla_y f(x_T, y_T)^\top (y - y_T) \tag{80}$$

Let $y \in \mathcal{Y}$ and $x \in \mathcal{X}$ and we then have, using (79) and (80):

$$f(x_T, y) - f(x, y_T) \leq \tag{81}$$

$$\leq \nabla_y f(x_T, y_T)^\top (y - y_T) - \nabla_x f(x_T, y_T)^\top (x - x_T) \quad \text{Sum up (79) and (80)} \tag{82}$$

$$\leq ||\nabla_y f(x_T, y_T)|| \cdot ||y - y_T|| + ||\nabla_x f(x_T, y_T)|| \cdot ||x - x_T|| \quad \text{Use Cauchy-Schwarz inequality} \tag{83}$$

$$\leq ||z_0 - z^*|| \cdot (||\nabla_y f(x_T, y_T)|| + ||\nabla_x f(x_T, y_T)||) \tag{84}$$

In (84), we use the fact that since $y \in \mathcal{Y}$, it holds that $||y - y_T|| \leq ||z_0 - z^*||$ because the distance among the point in a ball and its center can't be larger than the radius which is $||z_0 - z^*||$. The same argument holds for $||x - x_T|| \leq ||z_0 - z^*||$. It holds that $||\nabla_x f(x_T, y_T)|| = \sqrt{x_1^2 + x_2^2 + ... + x_{d_1}^2}$ where $x_i = \frac{\partial f}{\partial x_i}(x_T, y_T)$ and

$||\nabla_y f(x_T, y_T)|| = \sqrt{y_1^2 + y_2^2 + ... + y_{d_2}^2}$ where $y_i = \frac{\partial f}{\partial y_i}(x_T, y_T)$. Moreover, it holds that $\sqrt{x_1^2 + ... + x_{d_1}^2} + \sqrt{y_1^2 + ... + y_{d_2}^2} \leq \sqrt{2}\sqrt{x_1^2 + ... + x_{d_1}^2 + y_1^2 + ... + y_{d_2}^2}$ (Why?
- Because their equivalent inequalities hold
$(\sqrt{x_1^2 + ... + x_{d_1}^2} + \sqrt{y_1^2 + ... + y_{d_2}^2})^2 \leq (\sqrt{2}\sqrt{x_1^2 + ... + x_{d_1}^2 + y_1^2 + ... + y_{d_2}^2})^2 \rightarrow$
$(x_1^2 + ... + x_{d_1})^2 + (y_1^2 + ... + y_{d_2}^2) + 2\sqrt{(x_1^2 + ... + x_{d_1}^2)(y_1^2 + ... + y_{d_2}^2)} \leq 2(x_1^2 + .. + x_{d_1}^2 + y_1^2 + ... + y_{d_2}^2) \rightarrow 2\sqrt{(x_1^2 + ... + x_{d_1}^2)(y_1^2 + ... + y_{d_2}^2)} \leq (x_1^2 + .. + x_{d_1}^2) + (y_1^2 + ... + y_{d_2}^2)$,
which holds because of the $Gm - AM$ inequality (Geometric Mean- Arithmetic Mean inequality.))
Moreover, since $F(z_T) = (x_1, ..., x_{d_1}, y_1, ..., y_{d_2})$, then $||F(z_T)|| = \sqrt{x_1^2 + .. + x_{d_1}^2 + y_1^2 + ... + y_{d_2}^2}$
Therefore, $||\nabla_y f(x_T, y_T)|| + ||\nabla_x f(x_T, y_T)||) \leq \sqrt{2}||F(z_T)||$. Using this property in (84), we get:

$$f(x_T, y) - f(x, y_T) \leq ||z_0 - z^*||\sqrt{2}||F(z_T)|| \leq \frac{2\sqrt{2}(1 + \eta L)}{\eta T\sqrt{1 - 3\eta^2 L^2}}||z^* - z_0||^2 \qquad (85)$$

where in the last inequality we use the result from Question 4. Since (85) holds $\forall x \in \mathcal{X}$ it will also hold if we take the maximum of both sides of inequality (85) w.r.t x:

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

Since the above inequality holds $\forall y \in \mathcal{Y}$, it will also hold if we take the maximum of both sides over y:

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

which is exactly what we wanted to show!

References

[1] Amir Beck. *First-order methods in optimization.* SIAM, 2017.

[2] Sébastien Bubeck et al. "Convex optimization: Algorithms and complexity." In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[3] StackExchange. *Lipschitz smoothness, strong convexity and the Hessian.* [Online; accessed 30-June-2023]. URL: `%5Curl%7Bhttps://math.stackexchange.com/questions/673898/lipschitz-smoothness-strong-convexity-and-the-hessian%7D`.

[4] Wikipedia. *Gershgorin circle theorem.* [Online; accessed 30-June-2023]. URL: `%5Curl%7Bhttps://en.wikipedia.org/wiki/Gershgorin_circle_theorem)%7D`.