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 1 — FS23

- The solution is due on **March 28, 2023** by **11:59 pm**. Please submit your solution as a PDF on Moodle. The name of the file should follow the format `GA4-{Legi number}`, e.g., `GA1-19-123-456`. 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.

Separating Points on the Unit Interval (20 points)

Consider a learning problem where the data source $\mathcal{X} = [0, 1]$ is the unit interval and each sample point $X \in \mathcal{X}$ is drawn uniformly from $\mathcal{X}$ and is labeled as zero if $X < p^*$ and labeled as 1 otherwise, where p^* is an unknown parameter. Suppose we want to model finding p^* with 0-1-loss and the class of hypotheses is $\mathcal{H} = [0, 1]$. Provide a function $f(\cdot, \cdot)$ such that for any $0 \leqslant \varepsilon, \delta \leqslant 1$ and given $n \geqslant f(\varepsilon, \delta)$ many samples, any hypothesis $H \in \mathcal{H}$ with zero empirical risk[1] has low expected risk with probability at least $1 - \delta$. That is:

$$\ell(H) \leqslant \varepsilon.$$

In other words, if $n \geqslant f(\varepsilon, \delta)$, then the probability of existence of a hypothesis with zero empirical risk but with expected risk more than ε is at most δ.

Continuous Convex Functions (35 points)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous function. Show that the following are equivalent:

(a) f is a convex function.

(b) For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the following inequality holds:

$$\int_0^1 f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y}))\, d\lambda \leqslant \frac{f(\mathbf{x}) + f(\mathbf{y})}{2}.$$

Gradient Descent with Inexact Gradient Oracle (45 points)

Consider an unconstrained optimization problem $\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$. Assume f is μ-strongly convex and L-Lipschitz smooth. Now we only have access to an inexact gradient $g(\mathbf{x})$ at each point $\mathbf{x}$ such that $\|g(\mathbf{x}) - \nabla f(\mathbf{x})\| \leqslant \delta$ with $\delta > 0$. Consider gradient descent with this inexact gradient:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma g(\mathbf{x}_t),$$

where $\gamma > 0$ is the step-size. Define $\mathbf{x}^* \triangleq \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(x)$ and $f^* = f(\mathbf{x}^*)$.

(a) Show that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$f(\mathbf{x}) \geqslant f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{4}\|\mathbf{x} - \mathbf{y}\|^2 - \frac{\delta^2}{\mu},$$

and moreover,

$$\frac{1}{\mu}\|g(\mathbf{y})\|^2 \geqslant f(\mathbf{y}) - f^* - \frac{\delta^2}{\mu}.$$

[1] Observe that there is always at least one such hypothesis.

(b) Show that for any $x, y \in \mathbb{R}^d$ we have

$$f(x) \leqslant f(y) + \langle g(y), x - y \rangle + L\|x - y\|^2 + \frac{\delta^2}{2L}.$$

(c) Show that by running gradient descent with inexact gradient and setting $\gamma = \frac{1}{2L}$, we have

$$f(x_{t+1}) - f^* \leqslant \left(1 - \frac{\mu}{4L}\right)(f(x_t) - f^*) + \frac{3\delta^2}{4L}.$$

This directly implies

$$f(x_T) - f^* \leqslant \left(1 - \frac{\mu}{4L}\right)^T (f(x_0) - f^*) + \frac{3\delta^2}{\mu}.$$

(d) Find a function that is μ-strongly-convex and show that the algorithm above can not guarantee to find a point x such that $f(x) - f^* < \frac{\delta^2}{2\mu}$.

Optimization for Data Science, FS23 (Bernd Gärtner and Niao He)
Graded Assignment 1

Student Name, Legi Number: Jiaqing Xie, 22-943-914

Exercise 1

Solution for Exercise 1

Define a set of worse hypotheses s.t. the expected risk is more than ϵ: $\mathcal{H}_w = \{h_w \in \mathcal{H} \mid l(h_w) \geq \epsilon\}$. Define a set of points s.t. their empirical loss is zero under the scheme of the worse hypotheses: $X_w = \{X' \in X \mid \exists h_w \in \mathcal{H}_w, l_{X'}(h_w) = 0\}$. There exist a hypothesis $h \in \mathcal{H}$, such that $l(h) = 0$ and $l_{X_w}(h) = 0$ (optimal hypothesis).

Under this circumstance, we find that the probability of a hypothesis with zero empirical risk but with expected risk more than ϵ is equivalent to the statement that: the probability that we have sampled such "worse" points X' from our data source s.t. 1) their empirical risk $l_{X'}(h_w) = 0$, 2) the hypothesis h_w itself has the expected loss more than ϵ.

Therefore, we could define the probability as mentioned above:

$$
\begin{aligned}
\mathcal{P}_{X_w \sim \mathcal{X}}(X_w \in \mathcal{X}) &= \mathcal{P}_{X_w \sim \mathcal{X}}(\sum_{h_w \in \mathcal{H}} l_{X_w}(h_w) = 0) \\
&\leq \sum_{h_w \in \mathcal{H}} \mathcal{P}_{X_w \sim \mathcal{X}}(l_{X_w}(h_w) = 0) \\
&= \sum_{h_w \in \mathcal{H}} \prod_{i=1}^{n} \mathcal{P}_{X_w \sim \mathcal{X}}(\underbrace{h_w(x_i) = h^*(x_i)}_{\text{0-1 loss}}) \\
&= \sum_{h_w \in \mathcal{H}} \prod_{i=1}^{n} (1 - \underbrace{\mathcal{P}_{X_w \sim \mathcal{X}}(h_w(x_i) \neq h^*(x_i))}_{l(h_w) \geq \epsilon}) \\
&\leq |\mathcal{H}|(1-\epsilon)^n \\
&\leq |\mathcal{H}|(1-n\epsilon) \\
&\leq |\mathcal{H}|e^{-n\epsilon}, \quad \text{using the fact that:} e^x \geq 1 + x
\end{aligned}
\tag{1}
$$

We know that this probability is at most δ, so that

$$
\begin{aligned}
&|\mathcal{H}|e^{-n\epsilon} \leq \delta \\
&\Rightarrow n\epsilon \geq -log(\frac{\delta}{|\mathcal{H}|}) \\
&\Rightarrow n \geq \frac{1}{\epsilon} log(\frac{|\mathcal{H}|}{\delta})
\end{aligned}
\tag{2}
$$

Therefore we could define a function $f(\delta, \epsilon) = \frac{1}{\epsilon} log(\frac{|\mathcal{H}|}{\delta})$, such that for each n (number of sampled points) $\geq \frac{1}{\epsilon} log(\frac{|\mathcal{H}|}{\delta})$, the probability of existence of a hypothesis with zero empirical risk but with expected risk more than ϵ is at most δ.

Exercise 2

Solution for Exercise 2

1. $\Rightarrow$ Suppose f: $\mathbb{R}^d \to \mathbb{R}$ is a convex function, then for $\lambda \in [0,1], \mathbf{y} + \lambda(\mathbf{x} - \mathbf{y}) \in \mathbb{R}^d$ because of convex sets. The definition of convex:

$$f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y})) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) \tag{3}$$

Then we do integral over λ in the range [0, 1]:

$$\int_0^1 f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y}))\, d\lambda \leq \int_0^1 \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})\, d\lambda \tag{4}$$

Left integrand is always smaller than the right integrand (4), so integrating it will not change the sign between two integrals. Under this circumstance, we only need to calcualte the integral on the RHS.

$$\begin{aligned} \int_0^1 \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})\, d\lambda &= f(\mathbf{x}) \int_0^1 \lambda\, d\lambda + f(\mathbf{y}) \int_0^1 (1 - \lambda)\, d\lambda \quad \textit{linearity} \\ &= f(\mathbf{x}) \tfrac{1}{2}\lambda^2 \Big|_0^1 + f(\mathbf{y})(\lambda - \tfrac{1}{2}\lambda^2) \Big|_0^1 \\ &= \frac{f(\mathbf{x}) + f(\mathbf{y})}{2} \end{aligned} \tag{5}$$

Therefore, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the inequality holds:

$$\int_0^1 f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y}))\, d\lambda \leq \frac{f(\mathbf{x}) + f(\mathbf{y})}{2} \tag{6}$$

2. $\Leftarrow$ We're given

$$\int_0^1 f(\mathbf{y} + \lambda(\mathbf{x} - \mathbf{y}))\, d\lambda \leq \frac{f(\mathbf{x}) + f(\mathbf{y})}{2} \tag{7}$$

, prove f is convex.

Suppose f is not convex. We could find a triple of ($\mathbf{x_1}$, $\mathbf{y_1}$, $\lambda_\mathbf{1}$) s.t.

$$f(\mathbf{y_1} + \lambda_\mathbf{1}(\mathbf{x_1} - \mathbf{y_1})) > \lambda_\mathbf{1} f(\mathbf{x_1}) + (1 - \lambda_\mathbf{1}) f(\mathbf{y_1}), \quad \mathbf{x_1}, \mathbf{y_1} \in \mathbb{R}^d, \lambda_\mathbf{1} \in [0,1] \tag{8}$$

LHS - RHS:

$$f(\mathbf{y_1} + \lambda_\mathbf{1}(\mathbf{x_1} - \mathbf{y_1})) - \lambda_\mathbf{1} f(\mathbf{x_1}) - (1 - \lambda_\mathbf{1}) f(\mathbf{y_1}) > 0 \tag{9}$$

We could let $F(\lambda) = f(\mathbf{y_1} + \lambda(\mathbf{x_1} - \mathbf{y_1})) - \lambda f(\mathbf{x_1}) - (1 - \lambda) f(\mathbf{y_1})$. We already know that f is continuous, then $F(\lambda)$ is also continuous. So there must exist two different λ: $\lambda_\mathbf{0}$ and $\lambda_\mathbf{2}$ s.t. $F(\lambda_\mathbf{0}) > 0$, $F(\lambda_\mathbf{2}) > 0$, $0 \leq \lambda_\mathbf{0} < \lambda_\mathbf{1} < \lambda_\mathbf{2} \leq 1$.

Therefore consider two boundary points based on λ_0 and λ_2: $\mathbf{z_1} = \mathbf{y_1} + \lambda_0(\mathbf{x_1} - \mathbf{y_1})$ and $\mathbf{z_2} = \mathbf{y_1} + \lambda_2(\mathbf{x_1} - \mathbf{y_1})$. Then,

$$\begin{aligned} \mathbf{z_2} + \lambda(\mathbf{z_1} - \mathbf{z_2}) &= \lambda(\mathbf{y_1} + \lambda_0(\mathbf{x_1} - \mathbf{y_1})) + (1-\lambda)(\mathbf{y_1} + \lambda_2(\mathbf{x_1} - \mathbf{y_1})) \\ &= \mathbf{y_1} + (\lambda_2 + \lambda(\lambda_0 - \lambda_2))(\mathbf{x_1} - \mathbf{y_1}) \end{aligned} \tag{10}$$

Therefore, for a $\lambda \in [0, 1]$, we could find a corresponding $\lambda_1 : \lambda_2 + \lambda(\lambda_0 - \lambda_2) \in [\lambda_0, \lambda_2]$, so that it satisfied the boundary condition that within the range [0, 1]. And this will lead to the statement: exist $\lambda \in [0, 1]$, such that:

$$f(\mathbf{z_2} + \lambda(\mathbf{z_1} - \mathbf{z_2})) \geq \lambda f(\mathbf{z_1}) + (1-\lambda) f(\mathbf{z_2}) \tag{11}$$

And We could express the equation into the origin $(\mathbf{x}, \mathbf{y})$ space in order to check the correctness of (11):

$$\begin{aligned} RHS &= \lambda f(\mathbf{y_1} + \lambda_0(\mathbf{x_1} - \mathbf{y_1})) + (1-\lambda) f(\mathbf{y_1} + \lambda_2(\mathbf{x_1} - \mathbf{y_1})) \\ &> \lambda\lambda_0 f(\mathbf{x_1}) + \lambda(1-\lambda_0) f(\mathbf{y_1}) + (1-\lambda)\lambda_2 f(\mathbf{x_1}) + (1-\lambda)(1-\lambda_2) f(\mathbf{y_1}) \\ &= (\lambda_2 + \lambda(\lambda_0 - \lambda_2)) f(\mathbf{x_1}) + (1 - (\lambda_2 + \lambda(\lambda_0 - \lambda_2))) f(\mathbf{y_1}) \end{aligned} \tag{12}$$

Therefore we have: $f(\mathbf{y_1} + (\lambda_2 + \lambda(\lambda_0 - \lambda_2))(\mathbf{x_1} - \mathbf{y_1})) > (\lambda_2 + \lambda(\lambda_0 - \lambda_2)) f(\mathbf{x_1}) + (1 - (\lambda_2 + \lambda(\lambda_0 - \lambda_2))) f(\mathbf{y_1})$. This inequality is true according to (9), if we set $\lambda_1 = \lambda_2 + \lambda(\lambda_0 - \lambda_2)$ which proves the statement (11) above.

Integrating on both sides of (11) over λ from 0 to 1, we will have:

$$\begin{aligned} \int_0^1 f(\mathbf{z_2} + \lambda(\mathbf{z_1} - \mathbf{z_2}))\, d\lambda &> \int_0^1 \lambda f(\mathbf{z_1}) + (1-\lambda) f(\mathbf{z_2})\, d\lambda \\ &= f(\mathbf{z_1}) \tfrac{1}{2}\lambda^2 \Big|_0^1 + f(\mathbf{z_2})(\lambda - \tfrac{1}{2}\lambda^2) \Big|_0^1 \\ &= \frac{f(\mathbf{z_1}) + f(\mathbf{z_2})}{2}. \end{aligned} \tag{13}$$

It contradicts with the statement that for **all** $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the following inequality holds:

$$\int_0^1 f(\mathbf{x} + \lambda(\mathbf{x} - \mathbf{y}))\, d\lambda < \frac{f(\mathbf{x}) + f(\mathbf{y})}{2} \tag{14}$$

Therefore f must be a **convex** function.

If and only if conditions are satisfied so we've proved the equivalence of two statements.

Exercise 3

Solution for Exercise 3

(a) f is μ-strictly convex, where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d = \text{dom(f)}$, which means that

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) + \frac{\mu}{2}||\mathbf{x} - \mathbf{y}||^2, \quad \mu \in \mathbb{R}_+ \tag{15}$$

We're given that $\|g(\mathbf{y}) - \nabla f(\mathbf{y})\| \leq \delta$, $\delta > 0$. $\nabla f(\mathbf{y})$ is bounded from below by $g(\mathbf{y}) - \delta \iff \nabla f(\mathbf{y}) \geq g(\mathbf{y}) - \delta$. We could treat the second order term $\frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2$ as $\frac{\mu}{4}\|\mathbf{x} - \mathbf{y}\|^2 + \frac{\mu}{4}\|\|\mathbf{x} - \mathbf{y}\|^2$. Therefore, the inequality could be reconstructed as:

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{y}) + (g(\mathbf{y}) - \delta)^T(\mathbf{x} - \mathbf{y}) + \frac{\mu}{4}\|\mathbf{x} - \mathbf{y}\|^2 + \frac{\mu}{4}\|\mathbf{x} - \mathbf{y}\|^2 \\ &= f(\mathbf{y}) + g(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) + \frac{\mu}{4}\|\mathbf{x} - \mathbf{y}\|^2 + (\frac{\mu}{4}\|\mathbf{x} - \mathbf{y}\|^2 - \delta(\mathbf{x} - \mathbf{y})) \\ &= f(\mathbf{y}) + g(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) + \frac{\mu}{4}\|\mathbf{x} - \mathbf{y}\|^2 + \frac{\mu}{4}(\|\mathbf{x} - \mathbf{y}\| - \frac{2\delta}{\mu})^2 - \frac{\delta^2}{\mu} \end{aligned} \tag{16}$$

We have inner product $\langle g(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle = g(\mathbf{y})^T(\mathbf{x} - \mathbf{y})$, and $\frac{\mu}{4}(\|\mathbf{x} - \mathbf{y}\| - \frac{2\delta}{\mu})^2 \geq 0$. Then we could achieve the desired inequality:

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle g(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle + \frac{\mu}{4}\|\mathbf{x} - \mathbf{y}\|^2 - \frac{\delta^2}{\mu} \tag{17}$$

Set $\mathbf{x} = x^*$ and replace the origin $\mathbf{x}$ in the inequality (17) we get:

$$f(\mathbf{x}^*) \geq f(\mathbf{y}) + g(\mathbf{y})^T(\mathbf{x}^* - \mathbf{y}) + \frac{\mu}{4}\|\mathbf{x}^* - \mathbf{y}\|^2 - \frac{\delta^2}{\mu} \tag{18}$$

Rearrange the inequality on both sides we get:

$$-g(\mathbf{y})^T(\mathbf{x}^* - \mathbf{y}) - \frac{\mu}{4}\|\mathbf{x}^* - \mathbf{y}\|^2 \geq f(\mathbf{y}) - f(\mathbf{x}^*) - \frac{\delta^2}{\mu} \tag{19}$$

$\frac{1}{\mu}(g(\mathbf{y}) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{y}\|)^2 \geq 0 \iff \frac{1}{\mu}\|g(\mathbf{y})\|^2 \geq -g(\mathbf{y})^T(\mathbf{x}^* - \mathbf{y}) - \frac{\mu}{4}\|\mathbf{x}^* - \mathbf{y}\|^2$

So that $\frac{1}{\mu}\|g(\mathbf{y})\|^2 \geq -g(\mathbf{y})^T(\mathbf{x}^* - \mathbf{y}) - \frac{\mu}{4}\|\mathbf{x}^* - \mathbf{y}\|^2 \geq f(\mathbf{y}) - f(\mathbf{x}^*) - \frac{\delta^2}{\mu} \Rightarrow \frac{1}{\mu}\|g(\mathbf{y})\|^2 \geq f(\mathbf{y}) - f(\mathbf{x}^*) - \frac{\delta^2}{\mu}$

(b) for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, since function f is L-Lipschitz smooth, we have

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \tag{20}$$

Likewise, $\nabla f(\mathbf{y})$ is bounded from above by $g(\mathbf{y}) + \delta \iff \nabla f(\mathbf{y}) \leq g(\mathbf{y}) + \delta$. We could also rewrite the term $\frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 = L\|\mathbf{x} - \mathbf{y}\|^2 - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$. Using these tricks, the inequality then becomes:

$$\begin{aligned} f(\mathbf{x}) &\leq f(\mathbf{y}) + (g(\mathbf{y}) + \delta)^T(\mathbf{x} - \mathbf{y}) + L\|\mathbf{x} - \mathbf{y}\|^2 - \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \\ &= f(\mathbf{y}) + g(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) + L\|\mathbf{x} - \mathbf{y}\|^2 - (\frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 - \delta(\mathbf{x} - \mathbf{y})) \\ &= f(\mathbf{y}) + g(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) + L\|\mathbf{x} - \mathbf{y}\|^2 - \frac{L}{2}(\|\mathbf{x} - \mathbf{y}\|^2 - \frac{1}{L}\delta)^2 + \frac{\delta^2}{2L} \end{aligned} \tag{21}$$

We have inner product $\langle g(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle = g(\mathbf{y})^T(\mathbf{x}-\mathbf{y})$, and $-\frac{L}{2}(\|\mathbf{x}-\mathbf{y}\|^2 - \frac{1}{L}\delta)^2 < 0$. Then we could achieve the desired inequality:

$$f(\mathbf{x}) \leq f(\mathbf{y}) + g(\mathbf{y})^T(\mathbf{x}-\mathbf{y}) + L\|\mathbf{x}-\mathbf{y}\|^2 + \frac{\delta^2}{2L} \tag{22}$$

(c) we use the conclusion from (b), set $\mathbf{x} = \mathbf{x}_{t+1}$ and $\mathbf{y} = \mathbf{x}_t$:

$$f(\mathbf{x_{t+1}}) \leq f(\mathbf{x_t}) + g(\mathbf{x}_t)^T(\mathbf{x}_{t+1}-\mathbf{x}_t) + L\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \frac{\delta^2}{2L} \tag{23}$$

We know that $\mathbf{x}_t - \mathbf{x}_{t+1} = \gamma g(\mathbf{x}_t)$ and we susbtitute it with $\gamma = \frac{1}{2L}$ into the function:

$$\begin{aligned} f(\mathbf{x_{t+1}}) &\leq f(\mathbf{x_t}) + g(\mathbf{x}_t)^T(\mathbf{x}_{t+1}-\mathbf{x}_t) + L\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \frac{\delta^2}{2L} \\ &= f(\mathbf{x_t}) - \gamma\|g(\mathbf{x}_t)\|^2 + L\|\gamma g(\mathbf{x}_t)\|^2 + \frac{\delta^2}{2L} \\ &= f(\mathbf{x_t}) - \frac{1}{4L}\|g(\mathbf{x}_t)\|^2 + \frac{\delta^2}{2L} \end{aligned} \tag{24}$$

We have $g(\mathbf{x}_t)$ bounded: $g(\mathbf{x}_t) \geq \nabla f(\mathbf{x}_t) - \delta$, therefore:

$$\begin{aligned} f(\mathbf{x_{t+1}}) &\leq f(\mathbf{x_t}) - \frac{1}{4L}\|g(\mathbf{x}_t)\|^2 + \frac{\delta^2}{2L} \\ &\leq f(\mathbf{x_t}) - \frac{1}{4L}\|\nabla f(\mathbf{x}_t) - \delta\|^2 + \frac{\delta^2}{2L} \\ &= f(\mathbf{x_t}) - \frac{1}{4L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\delta\nabla f(\mathbf{x}_t)}{2L} + \frac{\delta^2}{4L} \end{aligned} \tag{25}$$

Imporantly we know that: if f is strongly convex, it must fulfill Polyak-Łojasiewicz Inequality: $\frac{1}{2}\|\nabla f(\mathbf{x}_t)\|^2 \geq \mu(f(\mathbf{x}_t) - f(\mathbf{x}^*))$ So we will have $\frac{1}{8L}\|\nabla f(\mathbf{x}_t)\|^2 \geq \frac{\mu}{4L}(f(\mathbf{x}_t) - f(\mathbf{x}^*))$.

We also use two tricks 1) $-\frac{1}{4L}\|\nabla f(\mathbf{x}_t)\|^2 = -\frac{1}{8L}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{1}{8L}\|\nabla f(\mathbf{x}_t)\|^2$ 2) $\frac{\delta^2}{4L} = \frac{3\delta^2}{4L} - \frac{\delta^2}{2L}$. Substituting all these into the function, we get:

$$\begin{aligned} f(\mathbf{x_{t+1}}) &\leq f(\mathbf{x_t}) - \frac{1}{4L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\delta\nabla f(\mathbf{x}_t)}{2L} + \frac{\delta^2}{4L} \\ &= f(\mathbf{x_t}) - \frac{1}{8L}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{1}{8L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\delta\nabla f(\mathbf{x}_t)}{2L} + \frac{3\delta^2}{4L} - \frac{\delta^2}{2L} \\ &\leq f(\mathbf{x_t}) - \frac{\mu}{4L}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) - \frac{1}{8L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\delta\nabla f(\mathbf{x}_t)}{2L} - \frac{\delta^2}{2L} + \frac{3\delta^2}{4L} \\ &= f(\mathbf{x_t}) - \frac{\mu}{4L}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) - \frac{1}{8L}\|\nabla f(\mathbf{x}_t) - 2\delta\|^2 + \frac{3\delta^2}{4L} \end{aligned} \tag{26}$$

Since $-\frac{1}{8L}\|\nabla f(\mathbf{x}_t) - 2\delta\|^2 \leq 0$, we must have:

$$\begin{aligned} f(\mathbf{x_{t+1}}) &\leq f(\mathbf{x_t}) - \frac{\mu}{4L}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) - \frac{1}{8L}\|\nabla f(\mathbf{x}_t) - 2\delta\|^2 + \frac{3\delta^2}{4L} \\ &\leq f(\mathbf{x_t}) - \frac{\mu}{4L}(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \frac{3\delta^2}{4L} \end{aligned} \tag{27}$$

Substract $f(\mathbf{x}^*)$ on both sides, we'll have the desired inequality:

$$f(\mathbf{x_{t+1}}) - f(\mathbf{x}^*) \leq (1 - \frac{\mu}{4L})(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \frac{3\delta^2}{4L} \tag{28}$$

We could easily observe from this inequality (by substition for T times) that:

$$\begin{aligned} f(\mathbf{x_T}) - f(\mathbf{x}^*) &\leq \underbrace{((1 - \frac{\mu}{4L})((1 - \frac{\mu}{4L})...((1 - \frac{\mu}{4L})}_{\text{T times}}(f(\mathbf{x_0}) - f(\mathbf{x}^*)) + \underbrace{\frac{3\delta^2}{4L}) + ...) + \frac{3\delta^2}{4L})}_{\text{T times}} \\ &= (1 - \frac{\mu}{4L})^T(f(\mathbf{x_0}) - f(\mathbf{x}^*)) + \frac{3\delta^2}{4L} + \frac{3\delta^2}{4L} * (1 - \frac{\mu}{4L}) + ... + \frac{3\delta^2}{4L} * (1 - \frac{\mu}{4L})^{T-1} \\ &= (1 - \frac{\mu}{4L})^T(f(\mathbf{x_0}) - f(\mathbf{x}^*)) + \frac{3\delta^2}{4L}(1 + (1 - \frac{\mu}{4L}) + ... + (1 - \frac{\mu}{4L})^{T-1}) \end{aligned} \tag{29}$$

We know that actually

$$\begin{aligned} \frac{3\delta^2}{4L}(1 + (1 - \frac{\mu}{4L}) + ... + (1 - \frac{\mu}{4L})^{T-1}) &\leq \lim_{T\to\infty} \frac{3\delta^2}{4L}(1 + (1 - \frac{\mu}{4L}) + ... + (1 - \frac{\mu}{4L})^{T-1}) \\ &= \frac{3\delta^2}{4L} * \frac{1}{1 - (1 - \frac{\mu}{4L})} \\ &= \frac{3\delta^2}{\mu} \end{aligned} \tag{30}$$

Overall, this will directly implies

$$f(\mathbf{x_T}) - f(\mathbf{x}^*) \leq (1 - \frac{\mu}{4L})^T(f(\mathbf{x_0}) - f(\mathbf{x}^*)) + \frac{3\delta^2}{\mu} \tag{31}$$

(d) For example $f(\mathbf{x}) = \frac{\mu}{2}(\mathbf{x} - \mathbf{x}_*)^2$. Observe that $f(\mathbf{x}_*) = 0$. Therefore $f(\mathbf{x}) - f(\mathbf{x}_*) = \frac{\mu}{2}(\mathbf{x} - \mathbf{x}_*)^2$. So it satisfied with the condition that the function f is μ-strongly-convex. We already have the inexact gradient iteration: $\mathbf{x_{t+1}} = \mathbf{x_t} - \gamma g(\mathbf{x_t})$. If we let $\mathbf{x_t} = \mathbf{x}_*$, then $\mathbf{x_{t+1}} = \mathbf{x}_* - \gamma g(\mathbf{x}_*)$.

Note that the gradient $\nabla f(\mathbf{x}_*)$ is equal to zero. And according to the algorithm, $\|g(\mathbf{x}_*) - \nabla f(\mathbf{x}_*)\| \leq \delta$. Therefore we have $\|g(\mathbf{x}_*)\| \leq \delta$. So we will have:

$$\begin{aligned} \mathbf{x_{t+1}} - \mathbf{x}^* &= -\gamma g(\mathbf{x}_*) \\ &\leq \gamma\delta \end{aligned} \tag{32}$$

So,

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{x}_*) &= \frac{\mu}{2}(\mathbf{x} - \mathbf{x}_*)^2 \\ &\leq \frac{\mu}{2}(\gamma\delta)^2 \end{aligned} \tag{33}$$

If we want $f(\mathbf{x}) - f(\mathbf{x}_*) < \frac{\delta^2}{2\mu}$ always hold, we must have $\frac{\mu}{2}(\gamma\delta)^2 < \frac{\delta^2}{2\mu}$, which means that $\mu^2\gamma^2 < 1$. The question itself does not declare clearly whether f should a L-smooth

function or not, since it just mentioned "the function is μ-strongly convex". We then discuss either two conditions for this inequality to be hold.

1. If function f is not L-smooth. It is obvious that we for any positive step size $\gamma > 0$, we could always find a $\mu \geq \frac{1}{\gamma}$, s.t. $\mu\gamma \geq 1$.

2. If function f is L-smooth. If we choose $\gamma = \frac{2}{L}$ for example, then μ must be less than $\frac{L}{2}$. However, we know that $\mu \leq L$. If we choose $\frac{L}{2} \leq \mu \leq L$, then it does not follow the constraint. Therefore it's not guaranteed to find a point x such that $f(\mathbf{x}) - f(\mathbf{x}_*) < \frac{\delta^2}{2\mu}$.

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 2 — FS23

- The solution is due on **May 1, 2023** by **11:59 pm**. Please submit your solution as a PDF on Moodle. The name of the file should follow the format `GA4-{Legi number}`, e.g., `GA2-19-123-456`. 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.

CGD with Unknown Smooth Parameter (25 points)

Suppose a function $f : \mathbb{R}^d \to \mathbb{R}$ (for $d \geqslant 2$) that is differentiable and coordinate-wise smooth with parameter $\mathcal{L} = (L_1, L_2, \ldots, L_d)$ where exactly one of L_j-s is equal to β ($\beta > 1$) and the others are all equal to 1. Moreover, suppose that f is μ-strongly convex ($\mu \leqslant 1$). Now we know β and μ, but we do not know which coordinate is β-smooth. We consider Algorithm 1: starting with a guess $\tilde{\mathcal{L}}^{(0)} = (\tilde{L}_1 = 1, \tilde{L}_2 = 1, \ldots, \tilde{L}_d = 1)$, when a coordinate-wise sufficient decrease criterion

$$f\left(x_t - \frac{1}{\tilde{L}_i^{(t)}} \nabla_i f(x_t) e_i\right) \leqslant f(x_t) - \frac{1}{2\tilde{L}_i^{(t)}} \|\nabla_i f(x_t)\|^2 \tag{1}$$

is not satisfied, we update our guess of the smoothness parameter of this coordinate to be β. In Algorithm 1, $\mathcal{D}_{IS}(L_1, L_2, \ldots, L_d)$ represents the probability distribution with the following mass function:

$$\mathbb{P}[i = k] = \frac{L_k}{\sum_{j=1}^{d} L_j}, \quad \forall k \in [d].$$

Algorithm 1 FUNNY-CGD $(x_0, \varepsilon, \beta, \mu)$

$\tilde{\mathcal{L}}^{(0)} := \left(\tilde{L}_1^{(0)}, \tilde{L}_2^{(0)}, \ldots, \tilde{L}_d^{(0)}\right) = (1, 1, \ldots, 1)$
$T = \left\lceil \frac{2(\beta+(d-1))}{\mu} \ln \frac{1}{\varepsilon} \right\rceil$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  Sample i from $\mathcal{D}_{IS}\left(\tilde{L}_1^{(t)}, \tilde{L}_2^{(t)}, \ldots, \tilde{L}_d^{(t)}\right)$
  if $f\left(x_t - \frac{1}{\tilde{L}_i^{(t)}} \nabla_i f(x_t) e_i\right) \leqslant f(x_t) - \frac{1}{2\tilde{L}_i^{(t)}} \|\nabla_i f(x_t)\|^2$ **then**
    $\tilde{\mathcal{L}}^{(t+1)} = \tilde{\mathcal{L}}^{(t)}$
    $x_{t+1} = x_t - \frac{1}{\tilde{L}_i^{(t)}} \nabla_i f(x_t)\, e_i$
  else
    $\tilde{\mathcal{L}}^{(t+1)} = \left(\tilde{L}_1^{(t)}, \ldots \tilde{L}_{i-1}^{(t)}, \beta, \tilde{L}_{i+1}^{(t)}, \ldots, \tilde{L}_d^{(t)}\right)$
    $x_{t+1} = \text{CGD-RAND}\left(x_t, \left\lceil \frac{\beta+(d-1)}{\mu} \ln 2 \right\rceil, \mathcal{D}_{IS}\left(\tilde{\mathcal{L}}^{(t+1)}\right), \gamma\right)$ with
γ equals to $\left(\frac{1}{\tilde{L}_1^{(t+1)}}, \ldots, \frac{1}{\tilde{L}_d^{(t+1)}}\right)$
  end if
end for
return x_T

Algorithm 2 CGD-RAND $(x_0, T, \mathcal{D}, \gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_d\})$

for $t = 0, 1, 2, \ldots, T-1$ **do**
  Sample i from Distribution $\mathcal{D}$
  $x_{t+1} = x_t - \gamma_i \nabla_i f(x_t)\, e_i$
end for
return x_T

Prove the followings for Algorithm 1.

(a) Show that when Algorithm 1 stops, it queries at most $O\left(\frac{d\bar{L}}{\mu}\ln\frac{1}{\varepsilon}\right)$ numbers of partial derivatives of f with $\bar{L}=\frac{1}{d}\sum_{j=1}^{d}L_j$.

(b) Show that the output from Algorithm 1 satisfies

$$\mathbb{E}\left[f\left(x_T\right)-f\left(x^*\right)\right]\leqslant\varepsilon\left(f\left(x_0\right)-f\left(x^*\right)\right).$$

Normalized GD for Nonconvex Optimization (25 points)

Consider a L-smooth function $f:\mathbb{R}^d\to\mathbb{R}$, which could be nonconvex. In addition, we assume the function is differentiable and has a global minimum x^*. Our goal is to find a stationary point x such that $\|\nabla f(x)\|$ is small. Instead of using conventional gradient descent, we consider a normalized version as follows:

$$x_{t+1}=x_t-\eta_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|+\beta_t},$$

where $\eta_t,\beta_t>0$.

(a) In this part, we consider fixed η_t and β_t, i.e., $\eta_t=\eta$ and $\beta_t=\beta$ for $t\geqslant 0$. Find a stepsize η such that $f(x_{t+1})\leqslant f(x_t)$ for all iterations.

(b) Under the same setting as part (a), show that the algorithm converges with rate $\mathcal{O}\left(T^{-1}\right)$, i.e.,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2=\mathcal{O}\left(T^{-1}\right).$$

(c) Now we consider the more general case where η_t and β_t are allowed to change over time. Design η_t and β_t such that without knowing the smoothness parameter, i.e., η_t and β_t should not depend on L, the algorithm provides the following guarantee:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|=\tilde{\mathcal{O}}\left(T^{-1/2}\right).$$

where $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic terms of T. Note that now we bound the average of gradient norms (instead of squared norms), and that is why the right-hand side depends on $\sqrt{T}$ rather than T.

Frank-Wolfe with an Approximation Oracle (20 points)

Recall that in the Frank-Wolfe algorithm, we assume there is a linear minimization oracle (LMO) that can return the exact minimizer. However, this minimization problem can itself be challenging to solve. In this exercise, we analyze the convergence of a variant of the Frank-Wolfe algorithm with an approximation LMO.

We consider the optimization problem $\min_{x \in X} f(x)$ for a convex function $f : \mathbb{R}^d \to \mathbb{R}$ with smooth parameter 1, and for X equals to $[-1/2, 1/2]^d$ (i.e., X is a unit d-dimensional cube). We assume a minimizer $x^* \in X$ exists. For any accuracy $\alpha \geqslant 0$, let $\text{APPROX-LMO}_X^\alpha(g)$ be an oracle that computes a vector in X such that

$$\nabla f(x_t)^\top \text{APPROX-LMO}_X^\alpha(g) \leqslant \min_{z \in X} \nabla f(x_t)^\top z + \alpha C_{(f,X)}$$

where $C_{(f,X)}$ is the curvature constant. So for α being zero, APPROX-LMO_X^0 is an exact oracle.

Algorithm 3 shows the modified Frank-Wolfe algorithm,

Algorithm 3 APPROX-FW (x_0, T)

for $t = 0, 1, 2, \ldots, T-1$ **do**
  $\gamma_t = \frac{2}{t+2}$
  $s_t = \text{APPROX-LMO}_X^{\gamma_t}(\nabla f(x_t))$
  $x_{t+1} = (1-\gamma_t)x_t + \gamma_t s_t$
end for
return x_T

(a) Show that

$$f(x_{t+1}) \leqslant f(x_t) - \gamma_t g(x_t) + \gamma_t^2 d \tag{2}$$

where $g(x_t)$ is the duality gap that is defined in the lectures.

(b) Show that for any $\varepsilon > 0$, and for any $T \geqslant 4d/\varepsilon$, we have:

$$f(x_T) - f(x^*) \leqslant \varepsilon.$$

Modified Newton's Method (30 points)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex twice-differentiable function with Lipschitz Hessian. In order to minimize the given function we consider a modified version of Newton's Method.

Algorithm 4 MODIFIED-NEWTON (x_0, H, T)

for $t = 0, 1, 2, \ldots, T-1$ **do**
  $\lambda_t = \sqrt{H\|\nabla f(x_t)\|}$
  $x_{t+1} = x_t - (\nabla^2 f(x_t) + \lambda_t I)^{-1} \nabla f(x_t)$
end for
return x_T

Assume the following for this exercise:

1. There is a $x^* \in \mathbb{R}^d$ such that $f(x^*) = \min_{x \in \mathbb{R}^d} f(x)$.
2. There is a constant $B \in \mathbb{R}$ such that for all $x \in \mathbb{R}^d$, if $f(x) \leqslant f(x_0)$, then $\|x - x^*\| \leqslant B$.
3. There is a constant $H > 0$ such that for all $x, y \in \mathbb{R}^d$, we have:

$$f(y) \leqslant f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(x)(y - x) + \frac{H}{3}\|y - x\|^3$$

and

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\| \leqslant H\|y - x\|^2.$$

In the following steps, we derive a convergence result for Algorithm 4 given the three assumptions above on f.

(a) Show that the following relations holds for all λ_t in Algorithm 4:

$$\lambda_t(x_{t+1} - x_t) = -\nabla f(x_t) - \nabla^2 f(x_t)(x_{t+1} - x_t). \tag{3}$$

(b) Show that for all iterations, the followings hold:

$$H\|x_{t+1} - x_t\| \leqslant \lambda_t,$$
$$\|\nabla f(x_{t+1})\| \leqslant 2\lambda_t\|x_{t+1} - x_t\| \leqslant 2\|\nabla f(x_t)\|.$$

(c) Prove the following descent lemma for Algorithm 4:

$$f(x_{t+1}) \leqslant f(x_t) - \frac{2}{3}\lambda_t\|x_{t+1} - x_t\|^2.$$

(d) Let $\mathcal{I}_\infty = \{i \in \mathbb{N} : \|\nabla f(x_{i+1})\| \geqslant \frac{1}{4}\|\nabla f(x_i)\|\}$ be the set of iterations at which the norm of gradient shrinks by at least a factor four.

Show that for all $t \in \mathcal{I}_\infty$, we have:

$$f(x_{t+1}) - f(x_t) \leqslant -\frac{1}{96B^{3/2}\sqrt{H}}(f(x_t) - f(x^*))^{3/2}.$$

Hint: First show that for any iteration, we have $f(x_t) - f(x^*) \leqslant B\|\nabla f(x_t)\|$.

Remark: Using the properties above, we can show that for all iterations (note that this holds for all t and not just members of $\mathcal{I}_\infty$):

$$f(x_t) - f(x^*) = \mathcal{O}(1/t^2).$$

Let us emphasize that proving this bound is not part of the exercise.

Optimization for Data Science, FS23 (Bernd Gärtner and Niao He)
Graded Assignment 2

Student Name, Legi Number: Jiaqing Xie, 22-943-914

CGD with Unknown Smooth Parameter

Solution for Exercise 1

(a) The algorithm `FUNNY-CGD` queries the partial derivatives of f when it executes the condition (either inside `if` or `else`). It executes the `else` condition only once because there's exactly only one of L_j equal to β, where $1 \leq j \leq d$. For those $i \neq j$ and $1 \leq i \leq d$, it should satisfy the following condition as specified in the algorithm:

$$f\left(x_t - \frac{1}{\widetilde{L}_i^{(t)}} \nabla_i f(x_t) e_i\right) \leq f(x_t) - \frac{1}{2\widetilde{L}_i^{(t)}} \left\|\nabla_i f(x_t)\right\|^2 \tag{1}$$

We could observe that before the k-th iteration which would execute `else` condition, it only executes `if` condition (for each 0, 1, ... k-1 iteration), where $\widetilde{L}^{(0)}$, $\widetilde{L}^{(1)}$, ..., $\widetilde{L}^{(k-1)}$ are equal to (1, 1, ..., 1) $\in \mathbb{R}^d$. Therefore $\bar{L} = \frac{1}{d}\sum_{j=1}^{d} L_j = 1$. At the k-th iteration, it will execute `else` condition so that one of the dimension of $\widetilde{L}^{(k)}$ will become β, which leads a new $\bar{L}' = \frac{1}{d}\sum_{j=1}^{d} L_j = \frac{1}{d}\left(\beta + (d-1)\right) > 1 = \bar{L}$. We could further state that $\bar{L}$ is upper bounded by $\frac{1}{d}(\beta + (d-1))$. After that $\bar{L}$ will keep unchanged until it reaches $\mathbf{T} = \left\lceil \frac{2(\beta+(d-1))}{\mu} \ln(\frac{1}{\epsilon}) \right\rceil$. So if execute the `if` condition, it will queries $\left\lceil \frac{2(\beta+(d-1))}{\mu} \ln(\frac{1}{\epsilon}) \right\rceil - 1$ times of the partial derivatives of f.

If it executes the `else` condition at the k-th iteration, it will execute `CGD-RAND` algorithm. There's also a loop in this algorithm which queries **T'** $= \left\lceil \frac{\beta+(d-1)}{\mu} \ln(2) \right\rceil$ times of the partial derivatives of f. Since we execute the `else` condition only once as we mention that only one of the dimension of L_j could be changed to β. So the total queries are 1 * $\left\lceil \frac{\beta+(d-1)}{\mu} \ln(2) \right\rceil = \left\lceil \frac{\beta+(d-1)}{\mu} \ln(2) \right\rceil$.

We know that if $\bar{L} = 1$, then $\beta + (d-1) = (\beta + (d-1))\bar{L} = d\bar{L}'$ where $\bar{L}' = \frac{1}{d}(\beta + d - 1)$, which is corresponding to the new $\bar{L}$ where the algorithm stopped. We can conclude that $d\bar{L} = \beta + d - 1$ when the algorithm stopped. We put all of the things mentioned together, the total number of iterations to query partial derivatives of f `t'` is equal to:

$$\begin{aligned} \mathtt{t}' &= \left\lceil \frac{2(\beta + (d-1))}{\mu} \ln(\frac{1}{\epsilon}) \right\rceil - 1 + \left\lceil \frac{\beta + (d-1)}{\mu} \ln(2) \right\rceil \\ &= \left\lceil \frac{2d\bar{L}}{\mu} \ln(\frac{1}{\epsilon}) \right\rceil - 1 + \left\lceil \frac{d\bar{L}}{\mu} \ln(2) \right\rceil \\ &\approx \frac{d\bar{L}}{\mu}\left(2\ln\left(\frac{1}{\epsilon}\right) + \ln(2)\right) \end{aligned} \tag{2}$$

According to big-O notation, we know that $2\ln\left(\frac{1}{\epsilon}\right) + \ln(2) = O\left(\ln\left(\frac{1}{\epsilon}\right)\right)$ if ϵ is small enough (close to zero). Even if ϵ is not too big which means that $\ln(2)$ could lead, it asks to find the number of partial derivatives **at most**, therefore, if ϵ is small enough, $\max(\ln(\frac{1}{\epsilon}), \ln(2)) = \ln(\frac{1}{\epsilon})$. We then conclude that $2\ln\left(\frac{1}{\epsilon}\right) + \ln(2) = O\left(\ln\left(\frac{1}{\epsilon}\right)\right)$.

And according to the another property of big-O notation: $f * (O(g)) = O(f * g)$, we get:

$$\begin{aligned} \mathtt{t}' &= \frac{d\bar{L}}{\mu} O\left(\ln\left(\frac{1}{\epsilon}\right)\right) \\ &= O\left(\frac{d\bar{L}}{\mu}\ln\left(\frac{1}{\epsilon}\right)\right) \end{aligned} \tag{3}$$

Therefore it proves the statement that when algorithm 1 stops, it queries at most $O\left(\frac{d\bar{L}}{\mu}\ln\left(\frac{1}{\epsilon}\right)\right)$ numbers of partial derivatives of f with $\bar{L} = \frac{1}{d}\sum_{j=1}^{d} L_j = \frac{1}{d}(\beta + d - 1)$.

(b) Suppose that before t-th iteration, $\bar{L}$ didn't change and always equal to 1 according to algorithm 1. We must have:

$$\mathbb{E}[f(x_t) - f(x^*)] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^t [f(x_0) - f(x^*)] \tag{4}$$

. This is theorem 5.7 with importance sampling and it has been proved in the exercise. After t-th iteration, $\bar{L}$ changed to $\bar{L}'$ which is equal to $\frac{1}{d}(\beta + d - 1)$, and we have:

$$\mathbb{E}[f(x_T) - f(x^*)] \leq \left(1 - \frac{\mu}{d\bar{L}'}\right)^{T-t} [f(x_t) - f(x^*)] \tag{5}$$

. We can regard x_t as the initial point after it executes the else branch (since after that $\bar{L}'$ didn't change and perform as before). And we combine equation 4 and equation 5:

$$\begin{aligned} \mathbb{E}[f(x_T) - f(x^*)] &\leq \left(1 - \frac{\mu}{d\bar{L}'}\right)^{T-t} [f(x_t) - f(x^*)] \\ &\leq \left(1 - \frac{\mu}{d\bar{L}'}\right)^{T-t} \left(1 - \frac{\mu}{d\bar{L}}\right)^t [f(x_0) - f(x^*)] \\ &= \left(\frac{1 - \frac{\mu}{d}}{1 - \frac{\mu}{\beta + d - 1}}\right)^t * \left(1 - \frac{\mu}{\beta + d - 1}\right)^T [f(x_0) - f(x^*)] \end{aligned} \tag{6}$$

We know that $\beta > 1$, so $1 - \frac{\mu}{d} \leq 1 - \frac{\mu}{\beta + d - 1}$. So the first term in the inequality 6 is less than 1. We have $\mathbf{T} = \left\lceil \frac{2(\beta + (d-1))}{\mu} \ln(\frac{1}{\epsilon}) \right\rceil$ Therefore we have:

$$
\begin{aligned}
\mathbb{E}[f(x_T) - f(x^*)] &\leq \left(1 - \frac{\mu}{\beta + d - 1}\right)^T [f(x_0) - f(x^*)] \\
&\leq \left(1 - \frac{\mu}{\beta + d - 1} * T\right) [f(x_0) - f(x^*)] \\
&\leq \left(e^{-\frac{\mu}{\beta+d-1} * T}\right) [f(x_0) - f(x^*)] \\
&\leq \left(e^{-\frac{\mu}{\beta+d-1} * \left\lceil \frac{2(\beta+(d-1))}{\mu} \ln(\frac{1}{\epsilon}) \right\rceil}\right) [f(x_0) - f(x^*)] \\
&= \left(e^{-2\ln\frac{1}{\epsilon}}\right) [f(x_0) - f(x^*)] \\
&= \epsilon^2 [f(x_0) - f(x^*)]
\end{aligned} \tag{7}
$$

$\ln \frac{1}{\epsilon}$ is valid, so that $0 < \epsilon \leq 1$, so that $\epsilon^2 \leq \epsilon$.
Therefore,

$$
\begin{aligned}
\mathbb{E}[f(x_T) - f(x^*)] &\leq \epsilon^2 [f(x_0) - f(x^*)] \\
&\leq \epsilon [f(x_0) - f(x^*)]
\end{aligned} \tag{8}
$$

Normalized GD for nonconvex Optimization

Solution for Exercise 2

(a) We let $\eta_t' = \frac{\eta_t}{\|\nabla f(x_t)\| + \beta_t}$, then $x_{t+1} = x_t - \eta_t' \nabla f(x_t)$. And we let $\Delta_t = f(x_t) - f(x^*)$.

Since f is L-smooth, it satisfied the inequality:

$$
f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \tag{9}
$$

We take $x_{t+1} - x_t$ w.r.t η_t' into the inequality and 'll have:

$$
\begin{aligned}
& f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
\Rightarrow & f(x_{t+1}) - f(x^*) \leq f(x_t) - f(x^*) + \langle \nabla f(x_t), -\eta_t' \nabla f(x_t) \rangle + \frac{L}{2} \| - \eta_t' \nabla f(x_t)\|^2 \\
\Rightarrow & \Delta_{t+1} \leq \Delta_t + \langle \nabla f(x_t), -\eta_t' \nabla f(x_t) \rangle + \frac{L}{2} \| - \eta_t' \nabla f(x_t)\|^2 \\
& \leq \Delta_t + (-\eta_t' + \frac{L}{2} \eta_t'^2) \|\nabla f(x_t)\|^2
\end{aligned} \tag{10}
$$

Since for all $t \geq 0$, $f(x_{t+1}) \leq f(x_t) \Rightarrow \Delta_{t+1} \leq \Delta_t \Rightarrow (-\eta_t' + \frac{L}{2}\eta_t'^2) \leq 0$. So that η_t' must less than $\frac{2}{L}$. This means that $\frac{\eta_t}{\|\nabla f(x_t)\| + \beta_t} \leq \frac{2}{L}$. We could choose $\eta_t = 1$ and $\beta_t = L$ s.t. $\frac{1}{\|\nabla f(x_t)\| + L} \leq \frac{2}{L}$. Since the norm term $\| \cdot \| \geq 0$, the inequality alway holds, where we found fixed $\eta_t = \eta = 1$, and fixed $\beta_t = \beta = L$. (Important for b)

(b) We use the same settings as(a), which means that $\eta_t = \eta = 1$, and fixed $\beta_t = \beta = L$. Therefore, $\eta'_t = \frac{\eta_t}{\|\nabla f(x_t)\|+\beta_t} = \frac{1}{\|\nabla f(x_t)\|+L}$. It is important to find that $-\eta'_t + \frac{L}{2}\eta'^2_t \leq -\frac{\eta'_t}{2}$ (Since $\eta'_t \leq \frac{1}{L}$). So the last inequality in (10) could be rewritten:

$$\begin{aligned} \Delta_{t+1} &\leq \Delta_t + (-\eta'_t + \frac{L}{2}\eta'^2_t)\|\nabla f(x_t)\|^2 \\ &\leq \Delta_t - \frac{\eta'_t}{2}\|\nabla f(x_t)\|^2 \end{aligned} \tag{11}$$

And it is equivalent to say that:

$$\begin{aligned} \|\nabla f(x_t)\|^2 &\leq \frac{2}{\eta'_t}(\Delta_t - \Delta_{t+1}) \\ &= 2(\|\nabla f(x_t)\| + L)(\Delta_t - \Delta_{t+1}) \end{aligned} \tag{12}$$

We use two properties in the following proof: 1) $\|\nabla f(x_t)\|^2 \leq \sum_{t=0}^{t} \|\nabla f(x_t)\|^2$, 2) $f(x^*) \leq f(x_t)$ for all $t \geq 0$. The aim of the property 1 is to create a non-decreasing sequence. If we take sum over $\|\nabla f(x_t)\|^2$, we'll have:

$$\begin{aligned} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 &\leq \sum_{t=0}^{T-1} 2(\|\nabla f(x_t)\| + L)(\Delta_t - \Delta_{t+1}) \\ &= \sum_{t=0}^{T-1} 2\|\nabla f(x_t)\|(\Delta_t - \Delta_{t+1}) + \sum_{t=0}^{T-1} L(\Delta_t - \Delta_{t+1}) \\ &\leq \left(2\sum_{t=0}^{T-1}\sqrt{\sum_{t=0}^{t}\|\nabla f(x_t)\|^2}(\Delta_t - \Delta_{t+1})\right) + L(\Delta_0 - \Delta_T) \\ &= \left(2\sum_{t=0}^{T-1}\sqrt{\sum_{t=0}^{t}\|\nabla f(x_t)\|^2}(\Delta_t - \Delta_{t+1})\right) + L(f(x_0) - f(x_T)) \\ &\leq \left(2\sum_{t=0}^{T-1}\sqrt{\sum_{t=0}^{t}\|\nabla f(x_t)\|^2}(\Delta_t - \Delta_{t+1})\right) + L\left(f(x_0) - f(x^*)\right) \\ &= \left(2\sum_{t=0}^{T-1}\sqrt{\sum_{t=0}^{t}\|\nabla f(x_t)\|^2}(\Delta_t - \Delta_{t+1})\right) + L\Delta_0 \end{aligned} \tag{13}$$

We let $b_t = \frac{1}{\sqrt{\sum_{t=0}^{t}\|\nabla f(x_t)\|^2}}$. We can use the property that 1) $\Delta_0 \geq \Delta_t$ since $f(x_0) \geq$

$f(x_t)$ and 2) $\Delta_t \geq 0$ since $f(x_t) \geq f(x^*)$. The last inequality in (13) will become:

$$\begin{aligned}
\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 &\leq \left(2\sum_{t=0}^{T-1} \frac{1}{b_t}(\Delta_t - \Delta_{t+1})\right) + L\Delta_0 \\
&\leq 2\left(\frac{\Delta_0}{b_0} + \sum_{t=1}^{T-1} \Delta_t(\frac{1}{b_t} - \frac{1}{b_{t-1}}) - \frac{\Delta_T}{b_{T-1}}\right) + L\Delta_0 \\
&\leq 2\left(\frac{\Delta_0}{b_0} + \sum_{t=1}^{T-1} \Delta_0(\frac{1}{b_t} - \frac{1}{b_{t-1}}) - \frac{\Delta_T}{b_{T-1}}\right) + L\Delta_0 \\
&= 2\left(\frac{\Delta_0}{b_0} + \Delta_0(\frac{1}{b_{T-1}} - \frac{1}{b_0}) - \frac{\Delta_T}{b_{T-1}}\right) + L\Delta_0 \\
&= 2\left(\frac{\Delta_0}{b_{T-1}} - \frac{\Delta_T}{b_{T-1}}\right) + L\Delta_0 \\
&\leq 2\frac{\Delta_0}{b_T} + L\Delta_0 \\
&\leq 2\Delta_0\sqrt{\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2} + L\Delta_0
\end{aligned} \tag{14}$$

It could be further found that only the term $\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2$ is on both sides of the inequality (14). And all the other terms could be regarded as constants. It's possible to solve the inequality:

$$\begin{aligned}
&\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \leq 2\Delta_0\sqrt{\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2} + L\Delta_0 \\
\Rightarrow &\left(\sqrt{\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2} - \Delta_0\right)^2 \leq \Delta_0^2 + L\Delta_0 \\
&\Rightarrow \sqrt{\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2} \leq \Delta_0 + \sqrt{\Delta_0^2 + L\Delta_0} \\
&\Rightarrow \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \leq \left(\Delta_0 + \sqrt{\Delta_0^2 + L\Delta_0}\right)^2 \\
&\Rightarrow \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \leq \frac{1}{T}\left(\Delta_0 + \sqrt{\Delta_0^2 + L\Delta_0}\right)^2
\end{aligned} \tag{15}$$

Since $\Delta_0 = f(x_0) - f(x^*)$ is a constant and L is also a constant, $\frac{1}{T}\left(\Delta_0 + \sqrt{\Delta_0^2 + L\Delta_0}\right)^2 =$

$\frac{C}{T}$, where C is a constant. Therefore,

$$\begin{aligned}\frac{1}{T-1}\sum_{t=0}^{T}\|\nabla f(x_t)\|^2 &\leq \frac{C}{T} \\ &= O(\frac{1}{T})\end{aligned} \tag{16}$$

(c) We could design

$$\eta_t = \frac{1}{\sum_{n=0}^{t}\frac{1}{n+1}}\|\nabla f(x_t)\| \tag{17}$$

$$\beta_t = \|\nabla f(x_t)\| \tag{18}$$

such that

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(x_t)\| = \widetilde{O}(\frac{1}{\sqrt{T}}) \tag{19}$$

, where $\widetilde{O}(\cdot)$ hides logarithmic terms of T. We will verify this below.

According to the update rule of x_t:

$$\begin{aligned}x_{t+1} &= x_t - \eta_t\frac{\nabla f(x_t)}{\|\nabla f(x_t)\| + \beta_t} \\ &= x_t - \frac{1}{\sum_{n=0}^{t}\frac{1}{n+1}}\|\nabla f(x_t)\|\frac{\nabla f(x_t)}{\|\nabla f(x_t)\| + \|\nabla f(x_t)\|} \\ &= x_t - \frac{1}{\sum_{n=0}^{t}\frac{2}{n+1}}\nabla f(x_t)\end{aligned} \tag{20}$$

Recall the L-smooth property:

$$\begin{aligned}f(x_{t+1}) &\leq f(x_t) + \langle\nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &\leq f(x_t) + \langle\nabla f(x_t), -\frac{1}{\sum_{n=0}^{t}\frac{2}{n+1}}\nabla f(x_t)\rangle + \frac{L}{2}\|-\frac{1}{\sum_{n=0}^{t}\frac{2}{n+1}}\nabla f(x_t)\|^2 \\ &\leq f(x_t) + \frac{\|\nabla f(x_t)\|^2}{(\sum_{n=0}^{t}\frac{2}{n+1})^2}\left(\frac{L}{2} - \sum_{n=0}^{t}\frac{2}{n+1}\right)\end{aligned} \tag{21}$$

We know that $\sum_{n=0}^{t}\frac{2}{n+1}$ is the divergent series which is equal to $\Theta(\log t)$. There must $\exists k \geq 0$, so that $\forall t \geq k, \left(\frac{L}{2} - \sum_{n=0}^{t}\frac{2}{n+1}\right) \leq 0$. so $\forall t \geq k$, we have:

$$\begin{aligned}&f(x_{t+1}) \leq f(x_t) + \frac{\|\nabla f(x_t)\|^2}{(\sum_{n=0}^{t}\frac{2}{n+1})^2}\left(\frac{L}{2} - \sum_{n=0}^{t}\frac{2}{n+1}\right) \\ \Rightarrow &\left(\frac{1}{\sum_{n=0}^{t}\frac{2}{n+1}} - \frac{L}{2}\frac{1}{(\sum_{n=0}^{t}\frac{2}{n+1})^2}\right)\|\nabla f(x_t)\|^2 \leq f(x_t) - f(x_{t+1}) \\ \Rightarrow &\|\nabla f(x_t)\|^2 \leq \left(\frac{1}{\sum_{n=0}^{t}\frac{2}{n+1}} - \frac{L}{2}\frac{1}{(\sum_{n=0}^{t}\frac{2}{n+1})^2}\right)^{-1}(f(x_t) - f(x_{t+1}))\end{aligned} \tag{22}$$

We could let

$$a_t = \left(\frac{1}{\sum_{n=0}^{t} \frac{2}{n+1}} - \frac{L}{2} \frac{1}{(\sum_{n=0}^{t} \frac{2}{n+1})^2} \right)^{-1} \tag{23}$$

. It has the property that $a_t \geq a_{t-1}$. Because we could regard the function inside the outer brackets as a second order concave function with initial point on the right side of the y-axis, with a decreasing point to zero. Therefore its reciprocal must be increasing. We could further observe that:

$$\begin{aligned} a_t &= \left(\frac{1}{\Theta(\log(t))} - \frac{L}{2} \frac{1}{\Theta(\log(t))^2} \right)^{-1} \\ &= \frac{\Theta(\log(t))^2}{\Theta(\log(t)) - \frac{L}{2}} \\ &= \Theta(\log(t)) \\ &= O(\log(t)) \end{aligned} \tag{24}$$

Now we sum over $\|\nabla f(x_t)\|^2$ to bound the 2-norm in order to bound the 1-norm, where we take the sum of first k 2-norm as a constant C_1:

$$\begin{aligned} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 &= \sum_{t=0}^{k-1} \|\nabla f(x_t)\|^2 + \sum_{t=k}^{T-1} \|\nabla f(x_t)\|^2 \\ &\leq C_1 + \sum_{t=k}^{T-1} a_t \left(f(x_t) - f(x_{t+1}) \right) \\ &= C_1 + \left(\sum_{t=k+1}^{T-1} (a_t - a_{t-1}) f(x_t) \right) + a_k f(x_k) - a_{T-1} f(x_T) \\ &\leq C_1 + \left(\sum_{t=k+1}^{T-1} (a_t - a_{t-1}) f(x_k) \right) + a_k f(x_k) - a_{T-1} f(x_T) \\ &= C_1 + a_{T-1} f(x_k) - a_k f(x_k) + a_k f(x_k) - a_{T-1} f(x_T) \\ &= C_1 + a_{T-1} [f(x_k) - f(x_T)] \\ &\leq C_1 + a_{T-1} [f(x_k) - f(x^*)] \\ &= C_1 + O(\log(T)) [f(x_k) - f(x^*)] \\ &= O(\log(T)) \end{aligned} \tag{25}$$

We could then bound the average of 1-norm:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(x_t)\| &= \sum_{t=0}^{T-1}\frac{1}{T}\|\nabla f(x_t)\| \\
&\leq \sqrt{\left(\sum_{t=0}^{T-1}\frac{1}{T^2}\right)\left(\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2\right)} \\
&\leq \sqrt{\left(T\cdot\frac{1}{T^2}\right)}\sqrt{C_1 + O(\log(T))[f(x_k) - f(x^*)]} \\
&\leq \sqrt{\frac{1}{T}}\sqrt{O(\log T)} \\
&= O(\frac{\sqrt{\log T}}{\sqrt{T}}) \\
&= \widetilde{O}(\frac{1}{\sqrt{T}})
\end{aligned} \tag{26}$$

, where $\widetilde{O}(\cdot)$ hides the logarithmic terms of T.

Frank-Wolfe with an Approximation Oracle

Solution for Exercise 3

(a) We have that $x_{t+1} = (1-\gamma_t)x_t + \gamma_t s_t$, therefore $x_{t+1} - x_t = \gamma_t(s_t - x_t)$. Since f is convex, we have that

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \langle x_{t+1} - x_t, \nabla f(x_t)\rangle + \frac{\gamma_t^2}{2}\|s_t - x_t\|^2 \\
&\leq f(x_t) + \gamma_t\langle(s_t - x_t), \nabla f(x_t)\rangle + \gamma_t^2 C_{f,X}
\end{aligned} \tag{27}$$

This is already proven in the lecture notes that $C_{f,X}$ could be replaced by its definition which is related with the norm term.

We are given that $\langle\nabla f(x_t), s_t\rangle \leq \min_{z\in X}\langle\nabla f(x_t), z\rangle + \alpha C_{f,X}$, which in other words,

$$\begin{aligned}
\langle\nabla f(x_t), s_t - x_t\rangle &\leq \min_{z\in X}\langle\nabla f(x_t), z - x_t\rangle + \alpha C_{f,X} \\
&= -g(x_t) + \alpha C_{f,X}
\end{aligned} \tag{28}$$

We substitute it to the above inequality, we'll have:

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \gamma_t\langle(s_t - x_t), \nabla f(x_t)\rangle + \gamma_t^2 C_{f,X} \\
&\leq f(x_t) + \gamma_t(-g(x_t) + \alpha C_f, X) + \gamma_t^2 C_{f,X} \\
&= f(x_t) - \gamma_t g(x_t) + (\gamma_t\alpha + \gamma_t^2)C_{f,X}
\end{aligned} \tag{29}$$

where it holds for all $\alpha \geq 0$. The diameter of a n-dimensional cube $[-1/2, 1/2]^d$ is equal to $\sqrt{d}$. According the Lemma 7.6 which is proved in the exercise, $C_{f,X} \leq \frac{1}{2} diam(X)^2 \leq \frac{1}{2} d$. Therefore, if we let $\alpha = \gamma_t$ (algorithm 3 line 3, where $\alpha = \gamma_t$),

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \gamma_t g(x_t) + (\gamma_t \alpha + \gamma_t^2) C_{f,X} \\ &\leq f(x_t) - \gamma_t g(x_t) + (\gamma_t \alpha + \gamma_t^2) \frac{d}{2} \\ &\leq f(x_t) - \gamma_t g(x_t) + (\gamma_t * \gamma_t + \gamma_t^2) \frac{d}{2} \\ &\leq f(x_t) - \gamma_t g(x_t) + \gamma_t^2 d \end{aligned} \tag{30}$$

(b) We first prove the inequality: $h(x_{t+1}) \leq (1 - \gamma_t) h(x_t) + \gamma_t^2 d$
We let $h(x_t) = f(x_t) - f(x^*)$ and we substitute it into the conclusion of (a):

$$\begin{aligned} f(x_{t+1}) - f(x^*) &\leq f(x_t) - f(x^*) - \gamma_t g(x_t) + \gamma_t^2 d \\ h(x_{t+1}) &\leq h(x_t) - \gamma_t g(x_t) + \gamma_t^2 d \\ &\leq h(x_t) - \gamma_t h(x_t) + \gamma_t^2 d \quad (\text{since } g(x_t) \geq f(x_t) - f(x^*)) \\ &= (1 - \gamma_t) h(x_t) + \gamma_t^2 d \end{aligned} \tag{31}$$

To show the result, we need to prove that $h(x_t) \leq \frac{4d}{t+2}$. We then use the induction method. when t is equal to 0, $\gamma_t = 1$ it follows by the above inequality that $h(x_1) \leq 0 + d \leq 2d$, which is correct. Suppose at $t = k - 1$ we have $h(x_k) \leq \frac{4d}{k+2}$. Then when $t = k$, we have

$$\begin{aligned} h(x_{k+1}) &\leq (1 - \gamma_k) h(x_k) + \gamma_k^2 d \\ &\leq \left(1 - \frac{2}{k+2}\right) \frac{4d}{k+2} + \left(\frac{2}{k+2}\right)^2 d \\ &\leq \frac{4d}{k+2} \left(1 - \frac{1}{k+2}\right) \\ &= \frac{4d}{k+2} \left(\frac{k+1}{k+2}\right) \\ &\leq \frac{4d}{k+2} \left(\frac{k+2}{k+3}\right) \\ &= \frac{4d}{k+3} \end{aligned} \tag{32}$$

, which is the claimed bound.

In order to achieve $f(x_T) - f(x^*) \leq \epsilon \quad \forall \epsilon > 0$, we must have $f(x_T) - f(x^*) \leq \frac{4d}{T+2} \leq \epsilon$. Therefore, we must have $T \geq \lceil \frac{4d}{\epsilon} \rceil - 2$. So for $\forall \epsilon > 0$ and $\forall T \geq 4d/\epsilon$, we have $h(x_T) = f(x_T) - f(x^*) \leq \epsilon$

Modified Newton's Method

Solution for Exercise 4
(a) We first look from the left hand side (LHS) and right hand side (RHS) of the equation that we want to prove. We are given that: $x_{t+1} = x_t - \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \nabla f(x_t)$ and the property of identity matrix that: $I = \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)\left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1}$

$$
\begin{aligned}
LHS &= \lambda_t \left(x_{t+1} - x_t\right) = -\lambda_t \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \nabla f(x_t) \\
RHS &= -\nabla f(x_t) - \nabla^2 f(x_t)\left(x_{t+1} - x_t\right) \\
&= -\nabla f(x_t) + \nabla^2 f(x_t)\left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \nabla f(x_t) \\
&= \left(\nabla^2 f(x_t)\left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} - \mathbb{I}\right) \nabla f(x_t) \\
&= \left(\nabla^2 f(x_t)\left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} - \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)\left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1}\right) \nabla f(x_t) \\
&= \left(\nabla^2 f(x_t) - \nabla^2 f(x_t) - \lambda_t \mathbb{I}\right)\left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \nabla f(x_t) \\
&= -\lambda_t \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \nabla f(x_t)
\end{aligned}
\tag{33}
$$

And we observe that LHS = RHS, so that the following equation stands:

$$
\lambda_t \left(x_{t+1} - x_t\right) = -\nabla f(x_t) - \nabla^2 f(x_t)\left(x_{t+1} - x_t\right) \tag{34}
$$

(b) We are given that $\lambda_t = \sqrt{\mathcal{H}\|\nabla f(x_t)\|}$ and $x_{t+1} = x_t - \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \nabla f(x_t)$. For a matrix K and a matrix A, the sub-multiplicative property of matrix multiplication is given by $\|AK\| \leq \|A\|\|K\|$.

Therefore, we have:

$$
\begin{aligned}
\mathcal{H}\|x_{t+1} - x_t\| &= \mathcal{H}\| - \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \nabla f(x_t)\| \\
&\leq \mathcal{H}\| \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \|\|\nabla f(x_t)\| \\
&= \lambda_t^2 \| \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \|
\end{aligned}
\tag{35}
$$

We know that the norm of a matrix could be related to its eigenvalues. For a matrix A, $\|A\| = max\|\lambda\| = \lambda_{max}$. Similarly, the norm of its inverse $\|A^{-1}\| = max\|1/\lambda\| = 1/\lambda_{min}$. Since we know f is a convex twice differentiable function, so that its Hessian constructed by $\nabla^2 f(x_t)$ has all eigenvalues larger or equal to zero: $\lambda' \geq 0$. A non-negative λ_t is added to its diagonal, which increases the eigenvalues of the matrix by λ_t. Therefore we have: $\| \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \| = max\|1/(\lambda' + \lambda_t)\| = 1/(\lambda'_{min} + \lambda_t) \leq 1/\lambda_t$. Since all $\lambda' \geq 0$, $\lambda'_{min} \geq 0$.

$$
\begin{aligned}
\mathcal{H}\|x_{t+1} - x_t\| &\leq \lambda_t^2 \| \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \| \\
&\leq \lambda_t^2 \cdot 1/\lambda_t \\
&= \lambda_t
\end{aligned}
\tag{36}
$$

We proved the inequality: $\mathcal{H}\|x_{t+1} - x_t\| \leq \lambda_t$.

For two vectors a and b, we have triangle inequality that $\|a + b\| \leq \|a\| + \|b\|$. We are given that $\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\| \leq \mathcal{H}\|y - x\|^2$. We use the conclusion that we proved in (a), we can get

$$\begin{aligned}\|\nabla f(x_{t+1})\| &\leq \|\nabla f(x_{t+1}) - \nabla f(x_t) - \nabla^2 f(x_t)\,(x_{t+1} - x_t)\,\| + \|\nabla f(x_t) + \nabla^2 f(x_t)\,(x_{t+1} - x_t)\,\| \\ &\leq \mathcal{H}\|x_{t+1} - x_t\|^2 + \|\lambda_t(x_{t+1} - x_t)\| \\ &\leq \lambda_t\|x_{t+1} - x_t\| + \|\lambda_t(x_{t+1} - x_t)\| \\ &= 2\lambda_t\|x_{t+1} - x_t\|\end{aligned} \tag{37}$$

The left inequality is proved. It's easier to prove the right inequality, since

$$\begin{aligned}2\lambda_t\|x_{t+1} - x_t\| &\leq 2\lambda_t \cdot \lambda_t/\mathcal{H} \\ &= 2\mathcal{H}\|\nabla f(x_t)\|/\mathcal{H} \quad \text{use the definition of } \lambda_t \\ &= 2\|\nabla f(x_t)\|\end{aligned} \tag{38}$$

So the right inequality holds. Together we get the double inequalities: $\|\nabla f(x_{t+1})\| \leq 2\lambda_t\|x_{t+1} - x_t\| \leq 2\|\nabla f(x_t)\|$.

(c) Since we proved in the (a) that $\lambda_t\,(x_{t+1} - x_t) = -\nabla f(x_t) - \nabla^2 f(x_t)\,(x_{t+1} - x_t)$.

$$\begin{aligned}\lambda_t\|x_{t+1} - x_t\|^2 &= (x_{t+1} - x_t)^T\,\lambda_t\,(x_{t+1} - x_t) \\ &= (x_{t+1} - x_t)^T\left(-\nabla f(x_t) - \nabla^2 f(x_t)\,(x_{t+1} - x_t)\right) \\ &= -\nabla^2 f(x_t)^T\,(x_{t+1} - x_t) - (x_{t+1} - x_t)^T\,\nabla^2 f(x_t)\,(x_{t+1} - x_t)\end{aligned} \tag{39}$$

To prove the lemma, we use the given inequality given in 3, which stated that:

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}\,(y - x)\,\nabla^2 f(x)\,(y - x) + \frac{\mathcal{H}}{3}\|y - x\|^3. \tag{40}$$

Let y $= x_{t+1}$ and x $= x_t$, we get:

$$\begin{aligned}f(x_{t+1}) &\leq f(x_t) + \nabla f(x_t)^T(x_{t+1} - x_t) + \frac{1}{2}\,(x_{t+1} - x_t)\,\nabla^2 f(x_t)\,(x_{t+1} - x_t) + \frac{\mathcal{H}}{3}\|x_{t+1} - x_t\|^3 \\ &= f(x_t) + \frac{1}{2}\nabla f(x_t)^T(x_{t+1} - x_t) + \frac{1}{2}\left(\nabla f(x_t)^T(x_{t+1} - x_t) + (x_{t+1} - x_t)\,\nabla^2 f(x_t)\,(x_{t+1} - x_t)\right) \\ &+ \frac{\mathcal{H}}{3}\|x_{t+1} - x_t\|^3 \\ &= f(x_t) + \frac{1}{2}\nabla f(x_t)^T(x_{t+1} - x_t) - \frac{1}{2}\lambda_t\|x_{t+1} - x_t\|^2 + \frac{\mathcal{H}}{3}\|x_{t+1} - x_t\|^3 \\ &\leq f(x_t) + \frac{1}{2}\nabla f(x_t)^T(x_{t+1} - x_t) - \frac{1}{2}\lambda_t\|x_{t+1} - x_t\|^2 + \frac{\lambda_t}{3}\|x_{t+1} - x_t\|^2 \\ &\quad \text{since}(\mathcal{H}\|x_{t+1} - x_t\| \leq \lambda_t) \\ &\leq f(x_t) + \frac{1}{2}\nabla f(x_t)^T(x_{t+1} - x_t) - \frac{\lambda_t}{6}\|x_{t+1} - x_t\|^2\end{aligned} \tag{41}$$

We are given that $x_{t+1} = x_t - \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)^{-1} \nabla f(x_t)$, so that $\nabla f(x_t) = -\left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)(x_{t+1} - x_t)$. For a semi-positive definite matrix A, for any given test vector x, $x^T A x \geq \|A\|\|x\|^2$. $\|\nabla^2 f(x_t)\| = \lambda'_{max} \geq \lambda'_{min} \geq 0$. So

$$\begin{aligned}
\frac{1}{2}\nabla f(x_t)^T (x_{t+1} - x_t) &= -\frac{1}{2}(x_{t+1} - x_t)^T \left(\nabla^2 f(x_t) + \lambda_t \mathbb{I}\right)(x_{t+1} - x_t) \\
&\leq -\frac{1}{2}\|\nabla^2 f(x_t) + \lambda_t \mathbb{I}\|\|x_{t+1} - x_t\|^2 \\
&\leq -\frac{1}{2}(\lambda'_{min} + \lambda_t)\|x_{t+1} - x_t\|^2 \\
&\leq -\frac{1}{2}\lambda_t \|x_{t+1} - x_t\|^2
\end{aligned} \tag{42}$$

We substitute this into the inequality above we get:

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \frac{1}{2}\nabla f(x_t)^T (x_{t+1} - x_t) - \frac{\lambda_t}{6}\|x_{t+1} - x_t\|^2 \\
&\leq f(x_t) - \frac{1}{2}\lambda_t \|x_{t+1} - x_t\|^2 - \frac{\lambda_t}{6}\|x_{t+1} - x_t\|^2 \\
&= f(x_t) - \frac{2\lambda_t}{3}\|x_{t+1} - x_t\|^2
\end{aligned} \tag{43}$$

So we've proved the inequality: $f(x_{t+1}) \leq f(x_t) - \frac{2}{3}\lambda_t \|x_{t+1} - x_t\|^2$.

(d) We need to first prove that $f(x_t) - f(x^*) \leq B\|\nabla f(x_t)\|$. Since f is convex is twice-differentiable, it satisfies the first order characteristics: $f(x^*) \geq f(x_t) + \nabla f(x_t)(x^* - x_t)$. Therefore $f(x_t) - f(x^*) \leq \nabla f(x_t)(x_t - x^*) \leq \|\nabla f(x_t)\| \cdot \|(x_t - x^*)\|$ (Cauchy-Schwarz inequality). From (c) we know that $f(x_t)$ is a non-increasing function so that we always have $f(x_t) \leq f(x_0)$. From the given condition in 2 that there exists B such that for all x_t, if $f(x_t) \leq f(x_0)$, then $\|x_t - x^*\| \leq B$. So that we proved $f(x_t) - f(x^*) \leq B\|\nabla f(x_t)\|$.

In order to prove

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{96B^{3/2}\sqrt{\mathcal{H}}}(f(x_t) - f(x^*))^{3/2} \tag{44}$$

, we first prove:

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{96\sqrt{\mathcal{H}}}(\|\nabla f(x_t)\|)^{3/2} \tag{45}$$

. If we prove that the stuff on the left hand side is less or equal than a smaller value, then it should be always less than a larger value, since:

$$-\frac{1}{96\sqrt{\mathcal{H}}}(\|\nabla f(x_t)\|)^{3/2} \leq -\frac{1}{96B^{3/2}\sqrt{\mathcal{H}}}(B\|\nabla f(x_t)\|)^{3/2} \tag{46}$$

To prove this, we start with the conclusion in (c) and (b), which are 1) $f(x_{t+1}) - f(x_t) \leq -\frac{2}{3}\lambda_t \|x_{t+1} - x_t\|^2$, and 2) $\|\nabla f(x_{t+1})\| \leq 2\lambda_t \|x_{t+1} - x_t\|$, and 3) $\|\nabla f(x_{t+1})\| \geq \frac{1}{4}\|\nabla f(x_t)\|$

we will have:

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq -\frac{2}{3}\lambda_t \|x_{t+1} - x_t\|^2 \\ &\leq -\frac{2}{3}\lambda_t \left(\frac{\|\nabla f(x_{t+1})\|}{2\lambda_t}\right)^2 \\ &\leq -\frac{2}{3}\lambda_t \left(\frac{\|\nabla f(x_t)\|}{4 * 2\lambda_t}\right)^2 \\ &\leq -\frac{1}{96}\frac{\|\nabla f(x_t)\|^2}{\lambda_t} \end{aligned} \tag{47}$$

We have $\lambda_t = \sqrt{\mathcal{H}\|\nabla f(x_t)\|}$, and we substitute into the inequality we'll have:

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq -\frac{1}{96}\frac{\|\nabla f(x_t)\|^2}{\lambda_t} \\ &= -\frac{1}{96}\frac{\|\nabla f(x_t)\|^2}{\sqrt{\mathcal{H}\|\nabla f(x_t)\|}} \\ &= -\frac{1}{96}\frac{\|\nabla f(x_t)\|^{3/2}}{\sqrt{\mathcal{H}}} \\ &\leq -\frac{1}{96}\frac{1}{\sqrt{\mathcal{H}}}\|\frac{f(x_t) - f(x^*)}{B}\|^{3/2} \quad \text{use } f(x_t) - f(x^*) \leq B\|\nabla f(x_t)\| \\ &= -\frac{1}{96B^{3/2}\sqrt{H}}\left(f(x_t) - f(x^*)\right)^{3/2} \end{aligned} \tag{48}$$

We could remove norm since $f(x_t) \geq f(x^*)$ for any x_t. Therefore we conclude that $f(x_{t+1}) - f(x_t) \leq -\frac{1}{96B^{3/2}\sqrt{H}}\left(f(x_t) - f(x^*)\right)^{3/2}$

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 3 — FS23

- The solution is due on **May 29, 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.

Subgradients on Convex Sets (15 points)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a non-empty convex set. Prove $y^* \in \mathcal{K}$ is a minimizer of f over $\mathcal{K}$ if and only if there exists a subgradient $g \in \partial f(y^*)$ such that

$$\langle y - y^*, g \rangle \geqslant 0 \quad \forall y \in \mathcal{K}.$$

Smoothed Function (25 points)

Consider the following composite optimization problem:

$$\min_{x \in \mathcal{X}} \left[\Phi(x) := f(x) + g(x) \right],$$

where $f : \mathcal{X} \to \mathbb{R}$ is L-smooth and convex function, and $g : \mathcal{X} \to \mathbb{R}$ is a convex and possible non-smooth function. Assume g can be smoothed with constant α and β. This means that for any $\mu > 0$, there exists a continuously differentiable convex function $g_\mu : \mathcal{X} \to \mathbb{R}$, that satisfies:

- $g(x) - \beta\mu \leqslant g_\mu(x) \leqslant g(x) + \beta\mu$, $\forall x \in \mathcal{X}$;
- g_μ is $\frac{\alpha}{\mu}$-smooth, i.e., $\|\nabla g_\mu(x) - \nabla g_\mu(y)\| \leqslant \frac{\alpha}{\mu}\|x - y\|$, $\forall x, y \in \mathcal{X}$.

We further assume that we have an algorithm $\mathcal{A}$ that can minimize any $\hat{L}$-smooth and convex function h over domain $\mathcal{X}$ with the guarantee: after t iterations, $h(x_t) - \min_{x \in \mathcal{X}} h(x) \leqslant \frac{c\hat{L}}{t^2}$ for some constant $c > 0$.

Now we apply $\mathcal{A}$ to minimize the smoothed composite function:

$$\min_{x \in \mathcal{X}} \left[\Phi_\mu(x) := f(x) + g_\mu(x) \right].$$

Show that with some choice of $\mu > 0$ (which can depend on the total number of iterations t), after t iterations, we have

$$\Phi(x_t) - \min_{x \in \mathcal{X}} \Phi(x) \leqslant \frac{Lc}{t^2} + \frac{2\sqrt{2\alpha\beta c}}{t}.$$

Then continue to show that for $\varepsilon > 0$, with

$$\mu = \sqrt{\frac{\alpha}{2\beta}} \frac{\varepsilon}{\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\varepsilon}}$$

and

$$t \geqslant \frac{2\sqrt{2\alpha\beta c}}{\varepsilon} + \frac{\sqrt{Lc}}{\sqrt{\varepsilon}},$$

it holds that $\Phi(x_t) - \min_{x \in \mathcal{X}} \Phi(x) \leqslant \varepsilon$.

Remark Compare this result with subgradient descent. You may use accelerated gradient methods as $\mathcal{A}$ and think about how smoothing techniques introduced in class are related to the conditions for the smoothed function here.

Proximal Non-Convex SGD (30 points)

Consider the following composite stochastic optimization problem:

$$\min_{x\in\mathbb{R}^d} \left[\Phi(x) := f(x) + r(x)\right], \qquad f(x) := \mathbb{E}\left[f(x,\xi)\right],$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a differentiable, L-smooth and has L-Lipschitz continuous gradient, and (possibly) non-convex function; $r : \mathbb{R}^d \to \mathbb{R}$ is a convex proximal-friendly function; the function $\Phi(x) := f(x) + r(x)$ is lower bounded by Φ^* for all $x \in \mathbb{R}^d$; the random variable ξ is distributed according some distribution $\mathcal{D}$. We are given an unbiased stochastic gradient oracle with bounded variance, i.e., at any point $x \in \mathbb{R}^d$, we can query $\nabla f(x,\xi) \in \mathbb{R}^d$ such that

$$\mathbb{E}\left[\nabla f(x,\xi)\right] = \nabla f(x), \qquad \mathbb{E}\left[\|\nabla f(x,\xi) - \nabla f(x)\|^2\right] \leqslant \sigma^2. \tag{1}$$

Consider the following method (Proximal Stochastic Gradient Descent)

$$x_{t+1} = \mathrm{prox}_{\eta r}(x_t - \eta \nabla f(x_t, \xi_{t+1})), \tag{2}$$

where ξ_{t+1} are independent for all $t \geqslant 0$, $\eta > 0$ is the step-size.

Recall that for any $\rho > 0$, the Moreau envelope of a function $\Phi(x)$ is given by

$$\Phi_\rho(x) := \min_{y\in\mathbb{R}^d} \left\{\Phi(y) + \frac{\rho}{2}\|y - x\|^2\right\}. \tag{3}$$

For any $\rho > 0$ and $x \in \mathbb{R}^d$, the proximal operator is defined as

$$\hat{x} := \mathrm{prox}_{\Phi/\rho}(x) := \operatorname*{argmin}_{y\in\mathbb{R}^d} \left\{\Phi(y) + \frac{\rho}{2}\|y - x\|^2\right\}. \tag{4}$$

Remark 1 Assume everywhere that $\rho > 0$ is large enough, so that the value of the proximal operator is unique. In fact, this will be satisfied if we take $\rho > L$.

(a) Let for any $x_t \in \mathbb{R}^d$, we have $\hat{x}_t := \mathrm{prox}_{\Phi/\rho}(x_t)$. Prove that

$$\hat{x}_t = \mathrm{prox}_{\eta r}\left(\eta\rho x_t - \eta\nabla f\left(\hat{x}_t\right) + (1 - \eta\rho)\hat{x}_t\right).$$

(b) Let $\rho = 4L$, $\eta \leqslant \frac{2}{9L}$, and x^{t+1} is given by (2). Prove that for all $t \geqslant 0$

$$\mathbb{E}\left[\|x_{t+1} - \hat{x}_t\|^2 \mid x_t\right] \leqslant (1 - \eta\rho)\|x_t - \hat{x}_t\|^2 + \sigma^2\eta^2.$$

(c) Let $\rho = 4L$, $\eta \leqslant \frac{2}{9L}$, and x^{t+1} is given by (2). Prove that for all $t \geqslant 0$

$$\mathbb{E}\left[\Phi_\rho\left(x_{t+1}\right)\right] \leqslant \mathbb{E}\left[\Phi_\rho\left(x_t\right)\right] - \frac{\eta}{2}\mathbb{E}\left[\left\|\rho(x_t - \hat{x}_t)\right\|^2\right] + \frac{\rho\eta^2\sigma^2}{2}.$$

Let index τ be chosen uniformly at random from the set $\{0, 1, \ldots, T-1\}$, prove that

$$\mathbb{E}\left[\left\|\rho(x_\tau - \hat{x}_\tau)\right\|^2\right] \leqslant \frac{2(\Phi_{4L}\left(x_0\right) - \inf_x \Phi_{4L}(x))}{\eta T} + 4L\eta\sigma^2.$$

Remark 2 Recall from the lecture (Handout 10, slide 54), the notion of generalized gradient. For any $\eta > 0$, $x \in \mathbb{R}^d$, generalized gradient is given by

$$G_\eta(x) := \frac{1}{\eta}(x - \operatorname{prox}_{\eta r}(x - \eta\nabla f(x))).$$

The following chain of inequalities relates the norm of generalized gradient with the quantity $\|x - \hat{x}\|$, for which we need to derive convergence in part (c). Let $\hat{x} := \operatorname{prox}_{\eta\Phi}(x)$. Then for any $\eta < 1/L$, and all $x \in \mathbb{R}^d$ it holds

$$\eta(1 - \eta L)\|G_\eta(x)\| \leqslant \|x - \hat{x}\| \leqslant \eta(1 + \eta L)\|G_\eta(x)\|.$$

We note that you do not need to prove any part of this remark.

Mirror Descent (30 points)

Let $f : \Omega \to \mathbb{R}$ be a convex and differentiable function. Assume that f is L-smooth ($L > 0$) with respect to some norm $\|\cdot\|$ (note that this does not need to be ℓ_2-norm). For any two $x, y \in \Omega$, We restate the Bregman divergence as seen in the lecture below:

$$V_\omega(x, y) := \omega(x) - \omega(y) - \nabla\omega(y)^\top(x - y)$$

Prove the followings for Algorithm 1:

Algorithm 1 GD-MD (Ω, $x_0 \in \Omega$, $y_0 \in \Omega$, $z_0 \in \Omega$, $L > 0$, $T \in \mathbb{N}$)

for $t = 0, 1, 2, \ldots, T-1$ **do**
Define $\gamma_{t+1} = \frac{t+2}{2L}$
Define $\eta_t = \frac{1}{\gamma_{t+1}L}$
Update

$$x_{t+1} = \eta_t z_t + (1 - \eta_t)y_t$$
$$y_{t+1} = \operatorname*{argmin}_{y\in\Omega}\left\{\frac{L}{2}\|y - x_{t+1}\|^2 + \langle\nabla f(x_{t+1}), y - x_{t+1}\rangle\right\}$$
$$z_{t+1} = \operatorname*{argmin}_{z\in\Omega}\{V_\omega(z, z_t) + \gamma_{t+1}\langle\nabla f(x_{t+1}), z - z_t\rangle\}$$

end for

(a) For any $\mathbf{u} \in \Omega$, show that

$$\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{u} \rangle \leqslant (\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle) + V_\omega(\mathbf{u}, \mathbf{z}_t) - V_\omega(\mathbf{u}, \mathbf{z}_{t+1}).$$

(b) Prove that

$$\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle - \frac{1}{2}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 \leqslant \gamma_{t+1}^2 L \left(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1})\right).$$

Next, we can observe that by combining a bit stronger inequality than part (a) (you do not need to prove it)

$$\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{u} \rangle \leqslant (\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle) - \frac{1}{2}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 + V_\omega(\mathbf{u}, \mathbf{z}_t) - V_\omega(\mathbf{u}, \mathbf{z}_{t+1})$$

The following equation holds (you do not need to prove it):

$$\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{u} \rangle \leqslant \gamma_{t+1}^2 L \left(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1})\right) + V_\omega(\mathbf{u}, \mathbf{z}_t) - V_\omega(\mathbf{u}, \mathbf{z}_{t+1}).$$

Hint You might find the relation $z_t - z_{t+1} = \eta_t^{-1}(\mathbf{x}_{t+1} - \mathbf{v}_t)$ useful in which $\mathbf{v}_t = \eta_t \mathbf{z}_{t+1} + (1 - \eta_t)\mathbf{y}_t$. You do not need to prove this relation.

(c) Next, show that for any $\mathbf{u} \in \Omega$, we have:

$$\gamma_{t+1}^2 L f(\mathbf{y}_{t+1}) - (\gamma_{t+1}^2 L - \gamma_{t+1}) f(\mathbf{y}_t) + V_\omega(\mathbf{u}, \mathbf{z}_{t+1}) - V_\omega(\mathbf{u}, \mathbf{z}_t) \leqslant \gamma_{t+1} f(\mathbf{u}).$$

Hint You might find the relation

$$\eta_t(\mathbf{x}_{t+1} - \mathbf{z}_t) = (1 - \eta_t)(\mathbf{y}_t - \mathbf{x}_{t+1}).$$

useful. You do not need to prove it, but it can be simply derived from the definition of $\mathbf{x}_{t+1}$.

(d) Assume there exists a minimizer $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x} \in \Omega} f(\mathbf{x})$ and for any choice of starting point $\mathbf{x}_0 \in \Omega$, we have $V_\omega(\mathbf{x}^*, \mathbf{x}_0) \leqslant R$, with $R \geqslant 0$. Prove that

$$f(\mathbf{y}_T) - f(\mathbf{x}^*) \leqslant \frac{4RL}{(T+1)^2}$$

Optimization for Data Science, FS23 (Bernd Gärtner and Niao He)
Graded Assignment 3

Student Name, Legi Number: Jiaqing Xie, 22-943-914

Subgradients on Convex Sets

Solution for Exercise 1
$\Leftarrow$: Suppose that there exist a subgradient $\mathbf{g} \in \partial f(\mathbf{y}^*)$ such that:

$$\langle \mathbf{y} - \mathbf{y}^*, \mathbf{g} \rangle \geq 0 \Leftrightarrow \langle \mathbf{g}, \mathbf{y} - \mathbf{y}^* \rangle \geq 0 \text{ for any } \mathbf{y} \in \mathcal{K} \tag{1}$$

. Using the definition of the subgradient of a convex function f, we have:

$$\begin{aligned} f(\mathbf{y}) - f(\mathbf{y}^*) &\geq \mathbf{g}^T(\mathbf{y} - \mathbf{y}^*) \\ &= \langle \mathbf{g}, \mathbf{y} - \mathbf{y}^* \rangle \geq 0 \end{aligned} \tag{2}$$

for any $\mathbf{y} \in \mathcal{K}$. So that $\mathbf{y}^*$ is the minimizer of f over $\mathcal{K}$.
$\Rightarrow$:

We know that $\mathbf{y}^*$ is a minimizer of g over $\mathcal{K}$, therefore:

$$\lim_{t \to 0} \frac{f(\mathbf{y}^* + t(\mathbf{y} - \mathbf{y}^*)) - f(\mathbf{y}^*)}{t} = \sup_{\mathbf{g} \in \partial f(\mathbf{y}^*)} \mathbf{g}^T(\mathbf{y} - \mathbf{y}^*) \geq 0 \tag{3}$$

for any $\mathbf{y} \in \mathcal{K}$. The last equation holds since $f(\mathbf{y}^* + t(\mathbf{y} - \mathbf{y}^*)) - f(\mathbf{y}^*) \geq t\mathbf{g}^T(\mathbf{y} - \mathbf{y}^*)$ for some $t \in \mathbb{R}^+$ with the first order characterization of convex function, together with the definition of minimizer $\mathbf{y}^*$. We could define a ball $B = \{\mathbf{y} + \mathbf{y}^* \mid \|\mathbf{y}\| \leq \epsilon\}$, where $\epsilon > 0$. So we must have:

$$\inf_{\mathbf{y} \in \mathcal{K} \bigcap B} \sup_{\mathbf{g} \in \partial f(\mathbf{y}^*)} \mathbf{g}^T(\mathbf{y} - \mathbf{y}^*) \geq 0 \tag{4}$$

, according to correctness of inequality (3). Since $\partial f(\mathbf{y}^*)$ is bounded, we could change the position of sup and inf. Therefore there exists at least a $\mathbf{g} \in \partial f(\mathbf{y}^*)$ such that:

$$\inf_{\mathbf{y} \in \mathcal{K} \bigcap B} \mathbf{g}^T(\mathbf{y} - \mathbf{y}^*) \geq 0 \tag{5}$$

. Otherwise the supremum should be less than zero, which leads to a contradiction.

Therefore, for any $\mathbf{y} \in \mathcal{K}$, it could be rewritten as $\mathbf{y}^* + t(\mathbf{x} - \mathbf{y}^*) \in \mathcal{K}$ (since set $\mathcal{K}$ convex), where t ≥ 0 and x $\in \mathcal{K} \bigcap B$, such that $\langle \mathbf{y} - \mathbf{y}^*, \mathbf{g} \rangle = t\langle \mathbf{g}, \mathbf{x} - \mathbf{y}^* \rangle \geq 0$. This inequality holds when looking at inequality (5).

Alternative: Instead of defining the ball to deal with sup and inf, we could also prove the inverse direction by contradiction and is much easier to derive that. Suppose there doesn't exist any $\mathbf{g} \in \partial f(\mathbf{y}^*)$ such that $\langle \mathbf{y} - \mathbf{y}^*, \mathbf{g} \rangle \geq 0$, which is equivalent to the statement $\forall \mathbf{g}, \langle \mathbf{y} - \mathbf{y}^*, \mathbf{g} \rangle < 0$, therefore $\sup_{\mathbf{g} \in \partial f(\mathbf{y}^*)} \mathbf{g}^T(\mathbf{y} - \mathbf{y}^*)$ in the inequality (3) is always less than 0. Then we observe the left hand side: $\lim_{t \to 0} f(\mathbf{y}^* + t(\mathbf{y} - \mathbf{y}^*)) - f(\mathbf{y}^*) < 0$. But it contradicts with the statement that $\mathbf{y}^*$ is the minimizer of f over $\mathcal{K}$. Therefore, there must **exist a** subgradient $\mathbf{g} \in \partial f(\mathbf{y}^*)$ such that $\langle \mathbf{y} - \mathbf{y}^*, \mathbf{g} \rangle \geq 0$.

Smoothed Function

Solution for Exercise 2
The definition of $\Phi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})$. The definition of $\Phi_\mu(\mathbf{x}) := f(\mathbf{x}) + g_\mu(\mathbf{x})$.
Firstly, we show that $\Phi_\mu(\mathbf{x})$ is a smooth function. We know that $f(\mathbf{x})$ is a L-smooth function where $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$. $g_\mu(\mathbf{x})$ is a $\frac{\alpha}{\mu}$-smooth function where $\|\nabla g_\mu(\mathbf{x}) - \nabla g_\mu(\mathbf{y})\| \leq \frac{\alpha}{\mu}\|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$. Therefore:

$$\begin{aligned}
\|\nabla\Phi_\mu(\mathbf{x}) - \nabla\Phi_\mu(\mathbf{y})\| &= \|\nabla f(\mathbf{x}) + \nabla g_\mu(\mathbf{x}) - \nabla f(\mathbf{y}) - \nabla g_\mu(\mathbf{y})\| \\
&= \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) + \nabla g_\mu(\mathbf{x}) - \nabla g_\mu(\mathbf{y})\| \\
&\leq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| + \|g_\mu(\mathbf{x}) - \nabla g_\mu(\mathbf{y})\| \text{ tri. ineq.} \\
&\leq L\|\mathbf{x} - \mathbf{y}\| + \frac{\alpha}{\mu}\|\mathbf{x} - \mathbf{y}\| \\
&= \left(L + \frac{\alpha}{\mu}\right)\|\mathbf{x} - \mathbf{y}\|
\end{aligned}$$

Therefore we could conclude that $\Phi_\mu(\mathbf{x})$ is $\left(L + \frac{\alpha}{\mu}\right)$-smooth. The addition of two convex functions are also convex, which means that $\Phi_\mu(\mathbf{x})$ is convex. A simple proof here: suppose f and g are two convex functions:

$$\begin{aligned}
(f+g)(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) &= f(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) + g(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \\
&\leq t(f(\mathbf{x}_1) + g(\mathbf{x}_1)) + (1-t)(f(\mathbf{x}_2) + g(\mathbf{x}_2)) \\
&= t(f+g)(x1) + (1-t)(f+g)(x2)
\end{aligned} \tag{6}$$

therefore f+g is a convex function.

We expand the expression of $\Phi(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x})$:

$$\begin{aligned}
&\Phi(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}) = f(\mathbf{x}_t) + g(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} (f(\mathbf{x}) + g(\mathbf{x})) \\
&= f(\mathbf{x}_t) + g_\mu(\mathbf{x}_t) + g(\mathbf{x}_t) - g_\mu(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}}(f(\mathbf{x}) + g_\mu(\mathbf{x}) + g(\mathbf{x}) - g_\mu(\mathbf{x})) \\
&= \Phi_\mu(\mathbf{x}_t) + g(\mathbf{x}_t) - g_\mu(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}}(\Phi_\mu(\mathbf{x}) + g(\mathbf{x}) - g_\mu(\mathbf{x}))
\end{aligned} \tag{7}$$

We are given that $\min_{\mathbf{x}\in\mathcal{X}}(g(\mathbf{x}) - g_\mu(\mathbf{x})) = -\beta\mu$ and $\max_{\mathbf{x}\in\mathcal{X}}(g(\mathbf{x}) - g_\mu(\mathbf{x})) = \beta\mu$. We have the property of min : $\min(f(\mathbf{x}) + g(\mathbf{x})) \geq \min(f(\mathbf{x})) + \min(g(\mathbf{x}))$. Therefore we could furthermore simplify the equation in (6):

$$\begin{aligned}
\Phi(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}) &= \Phi_\mu(\mathbf{x}_t) + g(\mathbf{x}_t) - g_\mu(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}}(\Phi_\mu(\mathbf{x}) + g(\mathbf{x}) - g_\mu(\mathbf{x})) \\
&\leq \Phi_\mu(\mathbf{x}_t) + \beta\mu - \min_{\mathbf{x}\in\mathcal{X}} \Phi_\mu(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}}(g(\mathbf{x}) - g_\mu(\mathbf{x}))) \\
&\leq \Phi_\mu(\mathbf{x}_t) + \beta\mu - \min_{\mathbf{x}\in\mathcal{X}} \Phi_\mu(\mathbf{x}) + \beta\mu \\
&= \Phi_\mu(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \Phi_\mu(\mathbf{x}) + 2\beta\mu
\end{aligned} \tag{8}$$

We are given an algorithm A, which can minimize a $\hat{L}$-smooth and convex function h: after t iterations, $h(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} h(\mathbf{x}) \leq \frac{c\hat{L}}{t^2}$. In this exercise, $\hat{L} = L + \frac{\alpha}{\mu}$, and $h(\mathbf{x}_t) = \Phi_\mu(\mathbf{x}_t)$. Therefore, we have:

$$\begin{aligned}\Phi(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}) &\leq \Phi_\mu(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \Phi_\mu(\mathbf{x}) + 2\beta\mu \\ &\leq \frac{c}{t^2}(L + \frac{\alpha}{\mu}) + 2\beta\mu \\ &= \frac{Lc}{t^2} + \frac{\alpha c}{\mu t^2} + 2\beta\mu\end{aligned} \tag{9}$$

We minimize $\frac{Lc}{t^2} + \frac{\alpha c}{\mu t^2} + 2\beta\mu$ w.r.t μ. The smallest value is achieved when $\frac{\alpha c}{\mu t^2} = 2\beta\mu$, which means that $\mu = \frac{1}{t}\sqrt{\frac{\alpha c}{2\beta}}$, where μ must be a function of t. We replace the μ in the inequality with this specific μ:

$$\begin{aligned}\Phi(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}) &\leq \frac{Lc}{t^2} + \frac{\alpha c}{\mu t^2} + 2\beta\mu \\ &= \frac{Lc}{t^2} + \frac{\alpha c}{t^2} t\sqrt{\frac{2\beta}{\alpha c}} + 2\beta\frac{1}{t}\sqrt{\frac{\alpha c}{2\beta}} \\ &= \frac{Lc}{t^2} + \frac{2\sqrt{2\alpha\beta c}}{t}\end{aligned} \tag{10}$$

To obtain the inequality that $\Phi(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}) \leq \epsilon$, the following inequality should hold:

$$\frac{Lc}{t^2} + \frac{2\sqrt{2\alpha\beta c}}{t} \leq \epsilon \tag{11}$$

It is equivalent to solve the following inequality:

$$\begin{aligned}&\epsilon t^2 - 2\sqrt{2\alpha\beta c}t - Lc \geq 0 \\ &\Rightarrow t \geq \frac{2\sqrt{2\alpha\beta c} + \sqrt{8\alpha\beta c + 4\epsilon Lc}}{2\epsilon} \\ &\Rightarrow t \geq \frac{\sqrt{2\alpha\beta c} + \sqrt{2\alpha\beta c + \epsilon Lc}}{\epsilon} \\ &\Rightarrow \frac{1}{t} \leq \frac{\epsilon}{\sqrt{2\alpha\beta c} + \sqrt{2\alpha\beta c + \epsilon Lc}}\end{aligned} \tag{12}$$

Since $\mu = \frac{1}{t}\sqrt{\frac{\alpha c}{2\beta}}$, $\mu \leq \sqrt{\frac{\alpha c}{2\beta}}\frac{\epsilon}{\sqrt{2\alpha\beta c}+\sqrt{2\alpha\beta c+\epsilon Lc}} \Rightarrow \mu \leq \sqrt{\frac{\alpha}{2\beta}}\frac{\epsilon}{\sqrt{2\alpha\beta}+\sqrt{2\alpha\beta+\epsilon L}}$. Therefore we could take $\mu = \sqrt{\frac{\alpha}{2\beta}}\frac{\epsilon}{\sqrt{2\alpha\beta}+\sqrt{2\alpha\beta+\epsilon L}}$.

We have the inequality of $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, $\forall a, b \geq 0$. Therefore:

$$\sqrt{2\alpha\beta c + \epsilon Lc} \leq \sqrt{2\alpha\beta c} + \sqrt{\epsilon Lc}$$

Therefore

$$\frac{\sqrt{2\alpha\beta c} + \sqrt{2\alpha\beta c + \epsilon Lc}}{\epsilon} \leq \frac{2\sqrt{2\alpha\beta c}}{\epsilon} + \frac{\sqrt{Lc}}{\sqrt{\epsilon}}$$

. So if $t \geq \frac{2\sqrt{2\alpha\beta c}}{\epsilon} + \frac{\sqrt{Lc}}{\sqrt{\epsilon}}$, it will definitely satisfy the inequality in inequality (12) since the value in (12) is much smaller. Therefore, if $t \geq \frac{2\sqrt{2\alpha\beta c}}{\epsilon} + \frac{\sqrt{Lc}}{\sqrt{\epsilon}}$, it holds that $\Phi(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}) \leq \epsilon$.

Proximal Non-Convex SGD

Solution for Exercise 3
The definition of proximal method is given by:

$$\mathbf{prox}_{\eta r}(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathbb{R}^d} \eta r(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2 \tag{13}$$

It is also equivalent to let $\nabla(\eta r(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2) = 0$, which means that: $\eta\nabla r(\mathbf{y}) + (\mathbf{y} - \mathbf{x}) = 0$. This is the property that we'll use in the proof.

(a) We let $\mathbf{y} = \mathbf{prox}_{\eta r}(\eta\rho\mathbf{x}_t - \eta\nabla f(\hat{\mathbf{x}}_t) + (1 - \eta\rho)\hat{\mathbf{x}}_t)$. Therefore, we take this into the above property and we'll have:

$$\begin{aligned} &\eta\nabla r(\mathbf{y}) + \mathbf{y} - \eta\rho\mathbf{x}_t + \eta\nabla f(\hat{\mathbf{x}}_t) - (1 - \eta\rho)\hat{\mathbf{x}}_t = 0 \\ \Leftrightarrow\ &\eta\nabla r(\mathbf{y}) + \mathbf{y} + \eta\rho(\hat{\mathbf{x}}_t - \mathbf{x}_t) + \eta\nabla f(\hat{\mathbf{x}}_t) - \hat{\mathbf{x}}_t = 0 \end{aligned} \tag{14}$$

Given that $\hat{\mathbf{x}}_t := \mathbf{prox}_{\Phi/\rho}(\mathbf{x}_t)$, according to the above property, we'll have:

$$\begin{aligned} &\nabla\Phi(\hat{\mathbf{x}}_t) + \rho(\hat{\mathbf{x}}_t - \mathbf{x}_t) = 0 \\ \Leftrightarrow\ &\eta\nabla\Phi(\hat{\mathbf{x}}_t) + \eta\rho(\hat{\mathbf{x}}_t - \mathbf{x}_t) = 0 \end{aligned} \tag{15}$$

(14) substracts (15), and using the fact that $\nabla\Phi(\mathbf{x}) = \nabla f(\mathbf{x}) + \nabla r(\mathbf{x})$:

$$\begin{aligned} &\eta\nabla r(\mathbf{y}) + \mathbf{y} + \eta\nabla f(\hat{\mathbf{x}}_t) - \hat{\mathbf{x}}_t - \eta\nabla\Phi(\hat{\mathbf{x}}_t) = 0 \\ \Rightarrow\ &\eta(\nabla r(\mathbf{y}) - \nabla r(\hat{\mathbf{x}}_t)) + \mathbf{y} - \hat{\mathbf{x}}_t = 0 \end{aligned} \tag{16}$$

Therefore $\mathbf{y} = \hat{\mathbf{x}}_t$ is a solution of the equation in (16). We have defined $\mathbf{y} = \mathbf{prox}_{\eta r}(\eta\rho\mathbf{x}_t - \eta\nabla f(\hat{\mathbf{x}}_t) + (1 - \eta\rho)\hat{\mathbf{x}}_t)$ in the beginning. So:

$$\hat{\mathbf{x}}_t = \mathbf{prox}_{\eta r}(\eta\rho\mathbf{x}_t - \eta\nabla f(\hat{\mathbf{x}}_t) + (1 - \eta\rho)\hat{\mathbf{x}}_t) \tag{17}$$

(b)We have:

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{prox}_{\eta r}(\mathbf{x}_t - \eta\nabla f(\mathbf{x}_t, \xi_{t+1})) \\ \hat{\mathbf{x}}_t &= \mathbf{prox}_{\eta r}(\eta\rho\mathbf{x}_t - \eta\nabla f(\hat{\mathbf{x}}_t) + (1 - \eta\rho)\hat{\mathbf{x}}_t) \end{aligned} \tag{18}$$

One property of proximal operators is called non-expansiveness, which is given by:

$$\|\mathbf{prox}_g(\mathbf{x}) - \mathbf{prox}_g(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$$

, which is given in the lecture notes. Therefore, we can further bound $\|\mathbf{x}_{t+1} - \hat{\mathbf{x}}_t\|^2$:

$$\begin{aligned}\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}_t\|^2 &= \|\mathbf{prox}_{\eta r}(\mathbf{x}_t-\eta\nabla f(\mathbf{x}_t,\xi_{t+1}))-\mathbf{prox}_{\eta r}(\eta\rho\mathbf{x}_t-\eta\nabla f(\hat{\mathbf{x}}_t)+(1-\eta\rho)\hat{\mathbf{x}}_t)\|^2\\ &\le \|\mathbf{x}_t-\eta\nabla f(\mathbf{x}_t,\xi_{t+1})-(\eta\rho\mathbf{x}_t-\eta\nabla f(\hat{\mathbf{x}}_t)+(1-\eta\rho)\hat{\mathbf{x}}_t)\|^2\\ &= \|(1-\eta\rho)(\mathbf{x}_t-\hat{\mathbf{x}}_t)+\eta\nabla f(\hat{\mathbf{x}}_t)-\eta\nabla f(\mathbf{x}_t,\xi_{t+1})\|^2\\ &= \|(1-\eta\rho)(\mathbf{x}_t-\hat{\mathbf{x}}_t)+\eta(\nabla f(\hat{\mathbf{x}}_t)-\nabla f(\mathbf{x}_t))+\eta(\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_t,\xi_{t+1}))\|^2\end{aligned} \tag{19}$$

We take expectation on both sides. Since we have $\mathbb{E}(\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_t,\xi_{t+1}))=0$, and $\mathbb{E}(\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_t,\xi_{t+1}\|^2)\le\sigma^2$. We know that f is L-smooth, therefore $\|\nabla f(\mathbf{x}_t)-\nabla f(\hat{\mathbf{x}}_t)\|\le L\|\mathbf{x}_t-\hat{\mathbf{x}}_t\|$

$$\begin{aligned}\mathbb{E}(\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}_t\|^2) &\le (1-\eta\rho)^2\|\mathbf{x}_t-\hat{\mathbf{x}}_t\|^2+\eta^2\|\nabla f(\hat{\mathbf{x}}_t)-\nabla f(\mathbf{x}_t)\|^2+\eta^2\sigma^2+\\ &\quad 2\eta(1-\eta\mu)(\mathbf{x}_t-\hat{\mathbf{x}}_t)(\nabla f(\hat{\mathbf{x}}_t)-\nabla f(\mathbf{x}_t))\\ &\le ((1-\eta\rho)^2+\eta^2L^2+2\eta(1-\eta\mu)L)\|\mathbf{x}_t-\hat{\mathbf{x}}_t\|^2+\eta^2\sigma^2\end{aligned} \tag{20}$$

Since $\rho=4L$, $L=\frac{1}{4}\rho$. Therefore:

$$\begin{aligned}\mathbb{E}(\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}_t\|^2) &\le ((1-\eta\rho)^2+\eta^2\left(\frac{\rho}{4}\right)^2+2\eta(1-\eta\mu)\frac{\rho}{4})\|\mathbf{x}_t-\hat{\mathbf{x}}_t\|^2+\eta^2\sigma^2\\ &= (1-\frac{3}{2}\eta\rho+\frac{9}{16}\eta^2\rho^2)\|\mathbf{x}_t-\hat{\mathbf{x}}_t\|^2+\eta^2\sigma^2\end{aligned} \tag{21}$$

We are also given that $\eta\le\frac{2}{9L}$, therefore $0<\eta\rho\le\frac{8}{9}\Rightarrow\frac{9}{16}\eta^2\rho^2\le\frac{1}{2}\eta\rho$. So $1-\frac{3}{2}\eta\rho+\frac{9}{16}\eta^2\rho^2\le 1-\frac{3}{2}\eta\rho+\frac{1}{2}\eta\rho=1-\eta\rho$. So:

$$\begin{aligned}\mathbb{E}(\|\mathbf{x}_{t+1}-\hat{\mathbf{x}}_t\|^2) &\le (1-\frac{3}{2}\eta\rho+\frac{9}{16}\eta^2\rho^2)\|\mathbf{x}_t-\hat{\mathbf{x}}_t\|^2+\eta^2\sigma^2\\ &\le (1-\eta\rho)\|\mathbf{x}_t-\hat{\mathbf{x}}_t\|^2+\eta^2\sigma^2\end{aligned} \tag{22}$$

(c) We know that $\Phi_\rho(\mathbf{x}_{t+1})=\min_{\mathbf{y}\in\mathbb{R}^d}\Phi(\mathbf{y})+\frac{\rho}{2}\|\mathbf{y}-\mathbf{x}_{t+1}\|^2$. Therefore, $\Phi_\rho(\mathbf{x}_{t+1})\le\Phi(\hat{\mathbf{x}}_t)+\frac{\rho}{2}\|\hat{\mathbf{x}}_t-\mathbf{x}_{t+1}\|^2$ (let $\mathbf{y}=\hat{\mathbf{x}}_t$ in the function which will be always larger than min). We use the conclusion in (b) and take the expectation of $\Phi_\rho(\mathbf{x}_{t+1})$ (since the settings of the parameters are the same):

$$\begin{aligned}\mathbb{E}[\Phi_\rho(\mathbf{x}_{t+1})] &\le \mathbb{E}[\Phi(\hat{\mathbf{x}}_t)+\frac{\rho}{2}\|\hat{\mathbf{x}}_t-\mathbf{x}_{t+1}\|^2]\\ &= \mathbb{E}[\Phi(\hat{\mathbf{x}}_t)]+\frac{\rho}{2}\mathbb{E}[\|\hat{\mathbf{x}}_t-\mathbf{x}_{t+1}\|^2]\\ &\le \mathbb{E}[\Phi(\hat{\mathbf{x}}_t)]+\frac{\rho}{2}\left((1-\eta\rho)\mathbb{E}[\|\mathbf{x}_t-\hat{\mathbf{x}}_t\|^2]+\eta^2\sigma^2\right)\\ &= \mathbb{E}[\Phi(\hat{\mathbf{x}}_t)+\frac{\rho}{2}\|\mathbf{x}_t-\hat{\mathbf{x}}_t\|^2]-\frac{\eta}{2}\mathbb{E}[\|\rho(\mathbf{x}_t-\hat{\mathbf{x}}_t)\|^2]+\frac{\rho\eta^2\sigma^2}{2}\end{aligned} \tag{23}$$

We know that $\hat{\mathbf{x}}_t=\mathbf{prox}_{\Phi/\rho}(\mathbf{x}_t)$, and combined with the definition of $\Phi_\rho(\mathbf{x})$, we have:

$$\Phi_\rho(\mathbf{x}_t)=\Phi(\hat{\mathbf{x}}_t)+\frac{\rho}{2}\|\mathbf{x}_t-\hat{\mathbf{x}}_t\|^2 \tag{24}$$

therefore,

$$\mathbb{E}[\Phi_\rho(\mathbf{x}_{t+1})] \leq \mathbb{E}[\Phi_\rho(\mathbf{x}_t)] - \frac{\eta}{2}\mathbb{E}[\|\rho(\mathbf{x}_t - \hat{\mathbf{x}}_t)\|^2] + \frac{\rho\eta^2\sigma^2}{2} \tag{25}$$

We uniformly sample index τ from the set {0, 1, 2, ..., T - 1}. We want to compute $\mathbb{E}[\|\rho(\mathbf{x}_\tau - \hat{\mathbf{x}_\tau})\|^2]$ w.r.t. τ. Uniform sampling means that calculating this expectation is equivalent to calculating the average of all possible values it could take, mathematically speaking:

$$\begin{aligned}\mathbb{E}[\|\rho(\mathbf{x}_\tau - \hat{\mathbf{x}_\tau})\|^2] &= \frac{1}{T}\sum_{t=0}^{T-1}\|\rho(\mathbf{x}_t - \hat{\mathbf{x}}_t)\|^2 \\ &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\rho(\mathbf{x}_t - \hat{\mathbf{x}}_t)\|^2] \quad (\mathbb{E}[C] = C)\end{aligned} \tag{26}$$

Rewrite the inequality (25):

$$\begin{aligned}&\frac{\eta}{2}\mathbb{E}[\|\rho(\mathbf{x}_t - \hat{\mathbf{x}}_t)\|^2] \leq \mathbb{E}[\Phi_\rho(\mathbf{x}_t)] - \mathbb{E}[\Phi_\rho(\mathbf{x}_{t+1})] + \frac{\rho\eta^2\sigma^2}{2} \\ \Leftrightarrow\ &\mathbb{E}[\|\rho(\mathbf{x}_t - \hat{\mathbf{x}}_t)\|^2] \leq \frac{2}{\eta}(\mathbb{E}[\Phi_\rho(\mathbf{x}_t)] - \mathbb{E}[\Phi_\rho(\mathbf{x}_{t+1})]) + \rho\eta\sigma^2\end{aligned} \tag{27}$$

We take this into inequality (26):

$$\begin{aligned}\mathbb{E}[\|\rho(\mathbf{x}_\tau - \hat{\mathbf{x}_\tau})\|^2] &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\rho(\mathbf{x}_t - \hat{\mathbf{x}}_t)\|^2] \\ &\leq \frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{2}{\eta}(\mathbb{E}[\Phi_\rho(\mathbf{x}_t)] - \mathbb{E}[\Phi_\rho(\mathbf{x}_{t+1})]) + \rho\eta\sigma^2\right) \\ &= \frac{2}{\eta T}\left(\mathbb{E}[\Phi_\rho(\mathbf{x}_0)] - \mathbb{E}[\Phi_\rho(\mathbf{x}_T)]\right) + \rho\eta\sigma^2 \\ &= \frac{2}{\eta T}\left(\Phi_\rho(\mathbf{x}_0) - \Phi_\rho(\mathbf{x}_T)\right) + \rho\eta\sigma^2\end{aligned} \tag{28}$$

We have $\rho = 4L$ and the fact that the infimum of a function f: $\inf_{x\in\mathbb{R}^d} f(\mathbf{x}) \leq f(\mathbf{x}_\tau)$, we'll have:

$$\begin{aligned}\mathbb{E}[\|\rho(\mathbf{x}_\tau - \hat{\mathbf{x}_\tau})\|^2] &\leq \frac{2}{\eta T}\left(\Phi_\rho(\mathbf{x}_0) - \Phi_\rho(\mathbf{x}_T)\right) + \rho\eta\sigma^2 \\ &\leq \frac{2}{\eta T}\left(\Phi_{4L}(\mathbf{x}_0) - \inf_{\mathbf{x}}\Phi_{4L}(\mathbf{x})\right) + 4L\eta\sigma^2\end{aligned} \tag{29}$$

Mirror Descent

Solution for Exercise 4

(a) The Bregman divergence is defined as follows:

$$\mathcal{V}_w(\mathbf{x}, \mathbf{y}) := w(\mathbf{x}) - w(\mathbf{y}) - \nabla w(\mathbf{y})^T(\mathbf{x} - \mathbf{y}) \tag{30}$$

. And its partial derivative of $\mathbf{x}$ is given by:

$$\frac{\partial}{\partial \mathbf{x}} \mathcal{V}_w(\mathbf{x}, \mathbf{y}) := \nabla w(\mathbf{x}) - \nabla w(\mathbf{y}) \tag{31}$$

, which is used later. The update rule for z_{t+1} in the algorithm is given by:

$$\mathbf{z}_{t+1} = \arg\min_{\mathbf{z} \in \Omega} \{\mathcal{V}_w(\mathbf{z}, \mathbf{z}_t) + \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z} - \mathbf{z}_t \rangle\} \tag{32}$$

Therefore, according to the definition of argmin:

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{z}} \left(\mathcal{V}_w(\mathbf{z}, \mathbf{z}_t) + \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z} - \mathbf{z}_t \rangle \right) = 0 \\ &\quad \Rightarrow \nabla w(\mathbf{z}) - \nabla w(\mathbf{z}_t) + \gamma_{t+1} \nabla f(\mathbf{x}_{t+1}) = 0 \\ &\quad \Leftrightarrow \gamma_{t+1} \nabla f(\mathbf{x}_{t+1}) = \nabla w(\mathbf{z}_t) - \nabla w(\mathbf{z}) \end{aligned} \tag{33}$$

Here, $\mathbf{z} = \mathbf{z}_{t+1}$.

Alternatively, the LHS of the inequality that requires to be proved could be rewritten as follows combined with three point identity (key property of Bregman divergence), which is proved in the lecture slides:

$$\begin{aligned} \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{u} \rangle &= \langle \nabla w(\mathbf{z}_t) - \nabla w(\mathbf{z}_{t+1}), \mathbf{z}_t - \mathbf{u} \rangle \\ &= -\langle \nabla w(\mathbf{z}_t) - \nabla w(\mathbf{z}_{t+1}), \mathbf{u} - \mathbf{z}_t \rangle \\ &= -\left(\mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1}) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \mathcal{V}_w(\mathbf{z}_t, \mathbf{z}_{t+1}) \right) \\ &= \mathcal{V}_w(\mathbf{z}_t, \mathbf{z}_{t+1}) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1}) \end{aligned} \tag{34}$$

The RHS of the inequality is given by:

$$(\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1}) \tag{35}$$

Therefore it is equivalent to prove that:

$$\mathcal{V}_w(\mathbf{z}_t, \mathbf{z}_{t+1}) \leq (\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle) \tag{36}$$

. We should prove this. Recall that $\gamma_{t+1} \nabla f(\mathbf{x}_{t+1}) = \nabla w(\mathbf{z}_t) - \nabla w(\mathbf{z}_{t+1})$ and with the definition of Bregman divergence we will have:

$$\begin{aligned} \mathcal{V}_w(\mathbf{z}_t, \mathbf{z}_{t+1}) - \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle &= \mathcal{V}_w(\mathbf{z}_t, \mathbf{z}_{t+1}) - \langle \nabla w(\mathbf{z}_t) - \nabla w(\mathbf{z}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle \\ &= w(\mathbf{z}_t) - w(\mathbf{z}_{t+1}) - \nabla w(\mathbf{z}_{t+1})^T (\mathbf{z}_t - \mathbf{z}_{t+1}) - \langle \nabla w(\mathbf{z}_t) - \nabla w(\mathbf{z}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle \\ &= w(\mathbf{z}_t) - w(\mathbf{z}_{t+1}) - \langle \nabla w(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle \\ &= -(w(\mathbf{z}_{t+1}) - w(\mathbf{z}_t) - \langle \nabla w(\mathbf{z}_t), \mathbf{z}_{t+1} - \mathbf{z}_t \rangle) \\ &= -\mathcal{V}_w(\mathbf{z}_{t+1}, \mathbf{z}_t) \end{aligned} \tag{37}$$

Since the property of non-negativity of Bregman divergence, which means $\mathcal{V}_w(\mathbf{z}_{t+1}, \mathbf{z}_t) \geq 0$, $-\mathcal{V}_w(\mathbf{z}_{t+1}, \mathbf{z}_t) \leq 0$, which means that $\mathcal{V}_w(\mathbf{z}_t, \mathbf{z}_{t+1}) - \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle \leq 0$. Therefore $\mathcal{V}_w(\mathbf{z}_t, \mathbf{z}_{t+1}) \leq (\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle)$. Therefore:

$$\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{u} \rangle \leq (\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1}) \tag{38}$$

(b) In (a), we have two equations proven:

$$\gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{u}\rangle = \mathcal{V}_w(\mathbf{z}_t, \mathbf{z}_{t+1}) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1}) \tag{39}$$

and

$$\mathcal{V}_w(\mathbf{z}_t, \mathbf{z}_{t+1}) - \gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1}\rangle = -\mathcal{V}_w(\mathbf{z}_{t+1}, \mathbf{z}_t) \tag{40}$$

, then:

$$\gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{u}\rangle = \gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1}\rangle - \mathcal{V}_w(\mathbf{z}_{t+1}, \mathbf{z}_t) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1}) \tag{41}$$

. We then extend $\gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1}\rangle$ and use the relation that $\mathbf{z}_t - \mathbf{z}_{t+1} = \frac{1}{\eta_t}(\mathbf{x}_{t+1} - \mathbf{v}_t)$ with the definition of $\mathbf{y}_{t+1}$:

$$\begin{aligned}
&\gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1}\rangle = \frac{\gamma_{t+1}}{\eta_t}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{v}_t\rangle \\
&= \gamma_{t+1}^2 L\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{v}_t\rangle \\
&= \gamma_{t+1}^2 L\langle -\nabla f(\mathbf{x}_{t+1}), \mathbf{v}_t - \mathbf{x}_{t+1}\rangle \\
&= \gamma_{t+1}^2 L\left(\langle -\nabla f(\mathbf{x}_{t+1}), \mathbf{v}_t - \mathbf{x}_{t+1}\rangle - \frac{L}{2}\|\mathbf{v}_t - \mathbf{x}_{t+1}\|^2\right) + \frac{\gamma_{t+1}^2 L^2}{2}\|\mathbf{v}_t - \mathbf{x}_{t+1}\|^2 \\
&= \gamma_{t+1}^2 L\left(\langle -\nabla f(\mathbf{x}_{t+1}), \mathbf{y}_{t+1} - \mathbf{x}_{t+1}\rangle - \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2\right) + \frac{1}{2}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2
\end{aligned} \tag{42}$$

We use the following property: f is L-smooth which means that $f(\mathbf{y}_{t+1}) \leq f(\mathbf{x}_{t+1}) + \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{y}_{t+1} - \mathbf{x}_{t+1}\rangle + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$ Therefore:

$$\begin{aligned}
&\gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1}\rangle \leq \gamma_{t+1}^2 L\left(f(\mathbf{x}_{t+1}) - f(\mathbf{x}_{t+1}) + \langle -\nabla f(\mathbf{x}_{t+1}), \mathbf{y}_{t+1} - \mathbf{x}_{t+1}\rangle - \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2\right) \\
&+ \frac{1}{2}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 \\
&\leq \gamma_{t+1}^2 L\left(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1})\right) + \frac{1}{2}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2
\end{aligned} \tag{43}$$

The conclusion of (b) should be:

$$\gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1}\rangle \leq \gamma_{t+1}^2 L\left(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1})\right) + \frac{1}{2}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 \tag{44}$$

. Combine with equation (41) and the fact of 1-strongly convex of w (from Bregman divergence) which means that $\frac{1}{2}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 \leq \mathcal{V}_w(\mathbf{z}_{t+1}, \mathbf{z}_t)$:

$$\begin{aligned}
\gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{u}\rangle &= \gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1}\rangle - \mathcal{V}_w(\mathbf{z}_{t+1}, \mathbf{z}_t) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1}) \\
&\leq \gamma_{t+1}^2 L\left(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1})\right) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1})
\end{aligned} \tag{45}$$

(c) Observe the relation and the conclusion in (b), we could expand $\eta_{t+1}(f(\mathbf{x}_{t+1}) - f(\mathbf{u}_t))$ with the property of convex function f:

$$\begin{aligned}\gamma_{t+1}(f(\mathbf{x}_{t+1}) - f(\mathbf{u})) &\leq \gamma_{t+1}\nabla f(\mathbf{x}_{t+1})(\mathbf{x}_{t+1} - \mathbf{u}) \\ &= \gamma_{t+1}\nabla f(\mathbf{x}_{t+1})(\mathbf{x}_{t+1} - \mathbf{z}_t + \mathbf{z}_t - \mathbf{u}) \\ &= \frac{\gamma_{t+1}(1-\eta_t)}{\eta_t}\nabla f(\mathbf{x}_{t+1})(\mathbf{y}_t - \mathbf{x}_{t+1}) + \gamma_{t+1}^2 L(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1}))+ \\ &\mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1}) \\ &\leq \frac{\gamma_{t+1}(1-\eta_t)}{\eta_t}(f(\mathbf{y}_t) - f(\mathbf{x}_{t+1})) + \gamma_{t+1}^2 L(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1}))+ \\ &\mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1})\end{aligned} \tag{46}$$

Since $\eta_t = \frac{1}{\gamma_{t+1}L}$,

$$\begin{aligned}\gamma_{t+1}(f(\mathbf{x}_{t+1}) - f(\mathbf{u})) &\leq (\gamma_{t+1}^2 L - \gamma_{t+1})(f(\mathbf{y}_t) - f(\mathbf{x}_{t+1})) + \gamma_{t+1}^2 L(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1}))+ \\ &\mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1}) \\ &\leq (\gamma_{t+1}^2 L - \gamma_{t+1})f(\mathbf{y}_t) + \gamma_{t+1}f(\mathbf{x}_{t+1}) - \gamma_{t+1}^2 L f(\mathbf{y}_{t+1}) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) - \\ &\mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1})\end{aligned} \tag{47}$$

We could eliminate $\gamma_{t+1}f(\mathbf{x}_{t+1})$ on both sides, and rewrite this inequality into:

$$\gamma_{t+1}^2 L f(\mathbf{y}_{t+1}) - (\gamma_{t+1}^2 L - \gamma_{t+1})f(\mathbf{y}_t) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1}) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) \leq \gamma_{t+1}f(\mathbf{u}) \tag{48}$$

(d) We observe the difference between $\gamma_{t+1}^2 L - \gamma_{t+1}$ and $\gamma_t^2 L$. $\gamma_{t+1}^2 L = \frac{(t+2)^2}{4L} \Rightarrow \gamma_{t+1}^2 L - \gamma_{t+1} = \frac{t^2+2t}{4L}$. $\gamma_t^2 L = \frac{t^2+2t+1}{4L}$. Therefore,

$$\gamma_t^2 L - \left(\gamma_{t+1}^2 L - \gamma_{t+1}\right) = \frac{1}{4L} \tag{49}$$

We sum both sides w.r.t. t from 0 to T-1:

$$\begin{aligned}&\sum_{t=0}^{T-1} \gamma_{t+1}^2 L f(\mathbf{y}_{t+1}) - (\gamma_{t+1}^2 L - \gamma_{t+1})f(\mathbf{y}_t) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_{t+1}) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_t) \leq \sum_{t=0}^{T-1} \gamma_{t+1}f(\mathbf{u}) \\ &\Rightarrow \gamma_T^2 L f(\mathbf{y}_T) + \sum_{t=1}^{T-1} \frac{1}{4L} f(\mathbf{y}_t) - (\gamma_1^2 L - \gamma_1)f(\mathbf{y}_0) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_T) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_0) \leq \sum_{t=0}^{T-1} \gamma_{t+1}f(\mathbf{u}) \\ &\Rightarrow \gamma_T^2 L f(\mathbf{y}_T) + \sum_{t=1}^{T-1} \frac{1}{4L} f(\mathbf{y}_t) + \mathcal{V}_w(\mathbf{u}, \mathbf{z}_T) - \mathcal{V}_w(\mathbf{u}, \mathbf{z}_0) \leq \sum_{t=0}^{T-1} \gamma_{t+1}f(\mathbf{u})\end{aligned} \tag{50}$$

We could use/reuse the following assumptions and properties 1) the Bregman divergence is non-negative: $\mathcal{V}_w(\mathbf{u}, \mathbf{z}_T) \geq 0$ 2) $\mathbf{u} = \mathbf{x}^*$, 3) $\mathbf{x}^*$ is the minimizer of f, which means

that $\forall \mathbf{x}, f(\mathbf{x}) \geq f(\mathbf{x}^*)$ 4) $\mathcal{V}_w(\mathbf{x}^*, \mathbf{x}_0) \leq \mathcal{R}$. Then:

$$\begin{aligned}
&\gamma_T^2 L f(\mathbf{y}_T) + \sum_{t=1}^{T-1} \frac{1}{4L} f(\mathbf{y}_t) + \mathcal{V}_w(\mathbf{x}^*, \mathbf{z}_T) - \mathcal{V}_w(\mathbf{x}^*, \mathbf{z}_0) \leq \sum_{t=0}^{T-1} \gamma_{t+1} f(\mathbf{x}^*) \\
&\Rightarrow \gamma_T^2 L f(\mathbf{y}_T) \leq \sum_{t=0}^{T-1} \gamma_{t+1} f(\mathbf{x}^*) - \sum_{t=1}^{T-1} \frac{1}{4L} f(\mathbf{y}_t) - \mathcal{V}_w(\mathbf{x}^*, \mathbf{z}_T) + \mathcal{V}_w(\mathbf{x}^*, \mathbf{z}_0) \\
&\Rightarrow \gamma_T^2 L f(\mathbf{y}_T) \leq f(\mathbf{x}^*) \sum_{t=0}^{T-1} \gamma_{t+1} - f(\mathbf{x}^*) \sum_{t=1}^{T-1} \frac{1}{4L} - 0 + \mathcal{R}
\end{aligned} \tag{51}$$

Since $\sum_{t=0}^{T-1} \gamma_{t+1} = \sum_{t=0}^{T-1} \frac{t+2}{2L} = \frac{T(T+3)}{4L}$,

$$\begin{aligned}
&\gamma_T^2 L f(\mathbf{y}_T) \leq f(\mathbf{x}^*) \sum_{t=0}^{T-1} \gamma_{t+1} - f(\mathbf{x}^*) \sum_{t=1}^{T-1} \frac{1}{4L} - 0 + \mathcal{R} \\
&\Rightarrow \gamma_T^2 L f(\mathbf{y}_T) \leq \left(\frac{T(T+3)}{4L} - \frac{T-1}{4L} \right) f(\mathbf{x}^*) + \mathcal{R} \\
&\Rightarrow \left(\frac{T+1}{2L} \right)^2 L f(\mathbf{y}_T) \leq \left(\frac{T^2+2T+1}{4L} \right) f(\mathbf{x}^*) + \mathcal{R} \\
&\Rightarrow \frac{(T+1)^2}{4L} f(\mathbf{y}_T) \leq \frac{(T+1)^2}{4L} f(\mathbf{x}^*) + \mathcal{R} \\
&\Rightarrow f(\mathbf{y}_T) \leq f(\mathbf{x}^*) + \frac{4L}{(T+1)^2} \mathcal{R}
\end{aligned} \tag{52}$$

The proof ends.

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.

Stochastic Gradient Descent (40 points)

Algorithm 1 SGD

Input: $x_0 \in \mathbb{R}^d$
for $t = 0, 1, 2, \ldots T-1$ **do**
  sample new ξ_t from a distribution $P(\xi)$
  $x_{t+1} \leftarrow x_t - \gamma_t \nabla f(x_t, \xi_t)$
end for

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.

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:

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

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 $t \geqslant 2$

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

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.

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:

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

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:

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

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

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

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

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

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

Student Name, Legi Number: Jiaqing Xie, 22-943-914

Min-Max for Smooth Functions

Consider the optimization problem:

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

, where f_i is 1-smooth lipschitz convex. Also, $\|\nabla f_i(x)\| \leq 1, \forall 1 \leq i \leq 10$. Give an algorithm that computes a value $\hat{f}$ s.t. $\hat{f} - f^* < \epsilon$. The algorithm evalutes the gradient of each of f_i for at most $O(1/\epsilon)$ many times.

We define a new function G(x):

$$G(x) = (f_1(x), f_1(x), f_2(x), ..., f_{10}(x))^T \in \mathbb{R}^{10\times 1} \tag{2}$$

Then maximization problem $\max_{1\leq i\leq 10} f_i(x)$ is then equivalent to:

$$f(x) = \max_{\text{y is a simplex}} y^T G(x), y \in \mathbb{R}^{10\times 1} \tag{3}$$

. According to the definition of simplex, the sum of all elements in y is equal to 1 with all element non-negative. The solution to the maximum problem is given by a sparse y $= [-------, 1, -------]^T$ where each element is equal to zero except the position where the function had the greatest value at x. So the functionality is the same as selecting the maximum function. Therefore, the original question is equivalent to the following question:

$$f^* = \min_{x\in\mathbb{R}^{20},\|x\|_2\leq 1} \max_{y\in\mathbb{R}^{10},\text{y is a simplex}} y^T G(x) \tag{4}$$

. If we let a new function $\Phi(x, y) = y^T G(x)$, then it becomes the standard min max optimization problem as mentioned in the lecture:

$$f^* = \min_{x\in\mathbb{R}^{20},\|x\|_2\leq 1} \max_{y\in\mathbb{R}^{10},\text{y is a simplex}} \Phi(x, y) \tag{5}$$

We know from the theorem 13.6 in the lecture, if $\Phi(x, y)$ is convex-concave, L-Lipschitz smooth, $\mathcal{X}$ has diameter $D_{\mathcal{X}}$, $\mathcal{Y}$ has diameter $D_{\mathcal{Y}}$, then applying Extra Gradient (EG) methods with stepsize $\eta \leq \frac{1}{2L}$ satisfies:

$$\max_{y\in\mathcal{Y}} \Phi(\frac{1}{T}\sum_{t=1}^{T} x_{t+\frac{1}{2}}, y) - \min_{x\in\mathcal{X}} \Phi(x, \frac{1}{T}\sum_{t=1}^{T} y_{t+\frac{1}{2}}) \leq \frac{D_{\mathcal{X}}^2 + D_{\mathcal{Y}}^2}{2\eta T} \tag{6}$$

We know that x, y are bounded by a constant, therefore, $D_{\mathcal{X}}^2 + D_{\mathcal{Y}}^2$ is bounded by a constant. The left handside is the duality gap. In this question, we know that for each computed $\hat{f}, \hat{f} - f^* < \epsilon$. So indeed we must have:

$$\max_{y \in \mathcal{Y}} \Phi(\frac{1}{T}\sum_{t=1}^{T} x_{t+\frac{1}{2}}, y) - \min_{x \in \mathcal{X}} \Phi(x, \frac{1}{T}\sum_{t=1}^{T} y_{t+\frac{1}{2}}) \leq \epsilon \leq \frac{D_{\mathcal{X}}^2 + D_{\mathcal{Y}}^2}{2\eta T} \tag{7}$$

so that the queries for each gradient is $O(\frac{1}{\epsilon})$. In order to achieve this, we need to prove two things 1) convex-concave of Φ 2) L-Lipschitz smooth and give an exact stepsize w.r.t. L in this question.

EG setting (the same as the setting in the lecture 13 page 53):

$$\hat{x}_{t+\frac{1}{2}} = x_t - \eta \nabla_x \Phi(x_t, y_t), \quad \hat{y}_{t+\frac{1}{2}} = y_t + \eta \nabla_y \Phi(x_t, y_t) \tag{8}$$

$$x_{t+\frac{1}{2}} = \Pi_{\mathcal{X}}(\hat{x}_{t+\frac{1}{2}}), \quad y_{t+\frac{1}{2}} = \Pi_{\mathcal{Y}}(\hat{y}_{t+\frac{1}{2}}) \tag{9}$$

$$\hat{x}_{t+1} = x_t - \eta \nabla_x \Phi(x_{t+\frac{1}{2}}, y_{t+\frac{1}{2}}), \quad \hat{y}_{t+1} = y_t + \eta \nabla_y \Phi(x_{t+\frac{1}{2}}, y_{t+\frac{1}{2}}) \tag{10}$$

$$x_{t+1} = \Pi_{\mathcal{X}}(\hat{x}_{t+1}), \quad y_{t+1} = \Pi_{\mathcal{Y}}(\hat{y}_{t+1}) \tag{11}$$

1) For every fixed y, since $f_i(x)$ is convex, and all elements of y is non-negative, the linear combination of convex function $y^T G(x)$ would result in a convex function. For every fixed x, it would result in a linear function which could be regarded as a concave (not strongly) function. Therefore Φ is convex-concave.

2) We first explore how $\nabla_x G(x)$ look like: $\begin{pmatrix} \nabla_{x_1} f_1(x) & \nabla_{x_2} f_1(x) & ... & \nabla_{x_{20}} f_1(x) \\ \nabla_{x_1} f_2(x) & \nabla_{x_2} f_2(x) & ... & \nabla_{x_{20}} f_2(x) \\ ... & ... & ... & ... \\ \nabla_{x_1} f_{10}(x) & \nabla_{x_2} f_{10}(x) & ... & \nabla_{x_{20}} f_{10}(x) \end{pmatrix} \in \mathbb{R}^{10\times 20}$. $\nabla_x G(x^*)$ look like: $\begin{pmatrix} \nabla_{x_1} f_1(x^*) & \nabla_{x_2} f_1(x^*) & ... & \nabla_{x_{20}} f_1(x^*) \\ \nabla_{x_1} f_2(x^*) & \nabla_{x_2} f_2(x^*) & ... & \nabla_{x_{20}} f_2(x^*) \\ ... & ... & ... & ... \\ \nabla_{x_1} f_{10}(x^*) & \nabla_{x_2} f_{10}(x^*) & ... & \nabla_{x_{20}} f_{10}(x^*) \end{pmatrix} \in \mathbb{R}^{10\times 20}$.

Then $y^T \nabla_x G(x) = \begin{pmatrix} y_1 * \nabla_{x_1} f_1(x) + ... + y_{10} * \nabla_{x_1} f_{10}(x) \\ y_1 * \nabla_{x_2} f_1(x) + ... + y_{10} * \nabla_{x_2} f_{10}(x) \\ ... \\ y_1 * \nabla_{x_1} f_{20}(x) + ... + y_{10} * \nabla_{x_{20}} f_{10}(x) \end{pmatrix}^T$.

$$\nabla_x y^T G(x) - \nabla_x y^{*T} G(x) = \begin{pmatrix} (y_1 - y_1^*) * \nabla_{x_1} f_1(x) + ... + (y_{10} - y_{10}^*) * \nabla_{x_1} f_{10}(x) \\ (y_1 - y_1^*) * \nabla_{x_2} f_1(x) + ... + (y_{10} - y_{10}^*) * \nabla_{x_2} f_{10}(x) \\ ... \\ (y_1 - y_1^*) * \nabla_{x_1} f_{20}(x) + ... + (y_{10} - y_{10}^*) * \nabla_{x_{20}} f_{10}(x) \end{pmatrix}^T.$$

Since $\|\nabla_x f_i(x)\| \leq 1$, then $\|\nabla_x y^T G(x) - \nabla_x y^{*T} G(x)\| \leq \left\| \begin{pmatrix} \|y_1 - y_1^*\| + ... + \|y_{10} - y_{10}^*\| \\ \|y_1 - y_1^*\| + ... + \|y_{10} - y_{10}^*\| \\ ... \\ \|y_1 - y_1^*\| + ... + \|y_{10} - y_{10}^*\| \end{pmatrix}^T \right\| =$

$\sqrt{20}\|y - y^*\|$

$$y^{*T}\nabla_x G(x) - y^{*T}\nabla_x G(x^*) = \begin{pmatrix} y_1^* * (\nabla_{x_1} f_1(x) - \nabla_{x_1} f_1(x^*)) + ... + y_{10}^* * (\nabla_{x_1} f_{10}(x) - \nabla_{x_1} f_{10}(x^*)) \\ y_1^* * (\nabla_{x_2} f_1(x) - \nabla_{x_2} f_1(x^*)) + ... + y_{10}^* * (\nabla_{x_2} f_{10}(x) - \nabla_{x_2} f_{10}(x^*)) \\ ... \\ y_1^* * (\nabla_{x_{20}} f_1(x) - \nabla_{x_{20}} f_1(x^*)) + ... + y_{10}^* * (\nabla_{x_{20}} f_{10}(x) - \nabla_{x_{20}} f_{10}(x^*)) \end{pmatrix}^T$$

We know that $f_i(x)$ is 1 Lipschitz smooth, then according to the definition we have: $\|\nabla_x f_i(x) - \nabla_x f_i(x^*)\| \leq \|x - x^*\|$.

$$\|y^{*T}\nabla_x G(x) - y^{*T}\nabla_x G(x^*)\| \leq \left\| \begin{pmatrix} \|y_1^*\| * \|x - x^*\| + ... + \|y_{10}^*\| * \|x - x^*\| \\ \|y_1^*\| * \|x - x^*\| + ... + \|y_{10}^*\| * \|x - x^*\| \\ ... \\ \|y_1^*\| * \|x - x^*\| + ... + \|y_{10}^*\| * \|x - x^*\| \end{pmatrix}^T \right\| = \sqrt{20}\|y^*\|\|x - x^*\| \leq \sqrt{20}\|x - x^*\|$$

Then we explore the Lipschitz property of $\Phi(x, y) = y^T G(x)$:

$$\begin{aligned} \|\nabla_x y^T G(x) - \nabla_x y^{*T} G(x^*)\| &= \|y^T \nabla_x G(x) - y^{*T}\nabla_x G(x^*)\| \\ &= \|y^T \nabla_x G(x) - y^{*T}\nabla_x G(x) + y^{*T}\nabla_x G(x) - y^{*T}\nabla_x G(x^*)\| \\ &\leq \|y^T \nabla_x G(x) - y^{*T}\nabla_x G(x)\| + \|y^{*T}\nabla_x G(x) - y^{*T}\nabla_x G(x^*)\| \\ &\leq \sqrt{20}\|y - y^*\| + \sqrt{20}\|x - x^*\| \end{aligned} \tag{12}$$

$$\nabla_y y^T G(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ ... \\ f_{10}(x) \end{pmatrix}, \nabla_{y^*} y^T G(x^*) = \begin{pmatrix} f_1(x^*) \\ f_2(x^*) \\ ... \\ f_{10}(x^*) \end{pmatrix}$$

According to 1-Lipschitz smooth condition, we also have the equivalent expression: $f_i(x) - f_i(x^*) \leq \nabla_x f_i(x^*)(x - x^*) + \frac{1}{2}\|x - x^*\|^2 \leq \|x - x^*\| + \frac{1}{2}\|x - x^*\|^2 \leq 2\|x - x^*\|$, therefore

$$\begin{aligned} \|\nabla_y y^T G(x) - \nabla_y y^{*T} G(x^*)\| &= \|G(x) - G(x^*)\| \\ &= \left\| \begin{pmatrix} f_1(x) - f_1(x^*) \\ f_2(x) - f_2(x^*) \\ ... \\ f_{10}(x) - f_{10}(x^*) \end{pmatrix} \right\| \\ &\leq \left\| \begin{pmatrix} 2\|x - x^*\| \\ 2\|x - x^*\| \\ ... \\ 2\|x - x^*\| \end{pmatrix} \right\| \\ &\leq \sqrt{40}\|x - x^*\| \end{aligned} \tag{13}$$

. Combining two results, Φ is $2\sqrt{10}$-Lipschitz smooth function. (A larger L also applies) We know that $\eta \leq \frac{1}{2L}$, therefore we could let $\eta = \frac{1}{4\sqrt{10}} \approx 0.079$, to make the inequality

(7) stands. Then

$$\begin{aligned} T &\leq \frac{D_{\mathcal{X}}^2 + D_{\mathcal{Y}}^2}{2\epsilon\eta} \\ &= \frac{2^2 + 2^2}{2 * \epsilon * \frac{1}{4\sqrt{10}}} \\ &= \frac{16\sqrt{10}}{\epsilon} = O(\frac{1}{\epsilon}) \end{aligned} \tag{14}$$

Proof ends for question 1.

Stochastic Gradient Descent

Question 1:

$$F(x) = \frac{x^2}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \tag{15}$$

with 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** $F(x_0) - \min_x F(x) \leq \Delta$

We could find the first derivative of F(x):

$$\nabla_x F(x) = \frac{x}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \tag{16}$$

If we have x_1 and x_2, then:

$$\frac{\|\nabla_x F(x_1) - \nabla_x F(x_2)\|}{\|x_1 - x_2\|} = \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \tag{17}$$

1) $\frac{1}{L} \geq 2\sum_{t=0}^{T-1}\gamma_t$, then $\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\} = \frac{1}{L}$. Therefore,

$$\frac{\|\nabla_x F(x_1) - \nabla_x F(x_2)\|}{\|x_1 - x_2\|} = \frac{1}{1/L} = L \tag{18}$$

which indicates that

$$\|\nabla_x F(x_1) - \nabla_x F(x_2)\| \leq L\|x_1 - x_2\| \tag{19}$$

2) $2\sum_{t=0}^{T-1}\gamma_t \geq \frac{1}{L}$, therefore $\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\} = 2\sum_{t=0}^{T-1}\gamma_t \geq \frac{1}{L}$. Equivalently,

$$\frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \leq L \tag{20}$$

So

$$\frac{\|\nabla_x F(x_1) - \nabla_x F(x_2)\|}{\|x_1 - x_2\|} = \frac{1}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \leq L \tag{21}$$

Therefore F is L-smooth.

On the other hand,

$$
\begin{aligned}
F(x_0) &= \frac{\sqrt{2\Delta \max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}^2}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \\
&= \Delta
\end{aligned} \tag{22}
$$

We know that L $>$ 0 so that F(x) is always non-negative. $\min_x F(x) = 0$ when x $= 0$. Otherwise $\min_x F(x) \geq 0$, therefore, $F(x_0) - \min_x F(x) \leq \Delta$

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

$$
\|\nabla F(x_t)\| \geq \sqrt{\frac{\Delta}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}} \tag{23}
$$

We know the update function from the algorithm 1 SGD that: $x_{t+1} = x_t - \gamma_t \nabla f(x_t, \xi_t)$, then:

$$
\begin{aligned}
x_{t+1} &= x_t - \gamma_t \nabla f(x_t, \xi_t) \\
&= x_t - \gamma_t \nabla F(x_t) \\
&= x_t - \gamma_t \frac{2x_t}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \\
&= \left(1 - \frac{\gamma_t}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\right) x_t \\
&= \prod_{k=0}^{t} \left(1 - \frac{\gamma_k}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}\right) x_0 \\
&= \prod_{k=0}^{t} \left(1 - \frac{\gamma_k}{2\max\{\frac{1}{2L}, \sum_{t=0}^{T-1}\gamma_t\}}\right) x_0
\end{aligned} \tag{24}
$$

We know that if $a \leq 1$, then $\frac{1}{2}a \leq 1 - \frac{1}{2}a \Rightarrow \exp\ln\frac{1}{2}a \leq 1 - \frac{1}{2}a$. If $\sum_{t=0}^{T-1}\gamma_t \geq \frac{1}{2L}$, then $\frac{\gamma_k}{\max\{\frac{1}{2L}, \sum_{t=0}^{T-1}\gamma_t}\} \leq 1$. This inequality also stands when $\sum_{t=0}^{T-1}\gamma_t \leq \frac{1}{2L}$. Then we rewrite equation (24):

$$
\begin{aligned}
x_{t+1} &= \prod_{k=0}^{t} \left(1 - \frac{\gamma_k}{2\max\{\frac{1}{2L}, \sum_{t=0}^{T-1}\gamma_t\}}\right) x_0 \\
&\geq \prod_{k=0}^{t} \exp\ln \frac{\gamma_k}{2\max\{\frac{1}{2L}, \sum_{t=0}^{T-1}\gamma_t\}} x_0 \\
&= \exp\left(\ln\frac{1}{2} \cdot \sum_{k=0}^{t} \frac{\gamma_k}{\max\{\frac{1}{2L}, \sum_{t=0}^{T-1}\gamma_t\}}\right) x_0 \geq \frac{1}{2}x_0
\end{aligned} \tag{25}
$$

Therefore,

$$\begin{aligned}\|\nabla F(x_t)\| &= \frac{x_t}{\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \\ &\geq \frac{x_0}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \\ &= \frac{\sqrt{2\Delta\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}} \\ &= \sqrt{\frac{\Delta}{2\max\{1/L, 2\sum_{t=0}^{T-1}\gamma_t\}}}\end{aligned} \tag{26}$$

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

$$F(x) = \frac{L}{2}\|x\|^2 \tag{27}$$

. We pick an inital point x_0 s.t. $\|x_0\| = \sqrt{\Delta/L}$. Consider another algorithm but with $\nabla f(x,\xi) = \nabla F(x) + \xi$. First show that for all t ≥ 2, we have:

$$x_t = \prod_{j=0}^{t-1}(1 - L\gamma_j)x_0 - \sum_{j=0}^{t-2}\gamma_j \prod_{i=j+1}^{t-1}(1 - L\gamma_j)\xi_j - \gamma_{t-1}\xi_{t-1} \tag{28}$$

From the definition, we have:

$$\begin{aligned}x_t &= x_{t-1} - \gamma_{t-1}(\nabla F(x_{t-1}) + \xi_{t-1}) \\ &= x_{t-1} - \gamma_{t-1}(Lx_{t-1} + \xi_{t-1}) \\ &= (1 - L\gamma_{t-1})x_{t-1} - \gamma_{t-1}\xi_{t-1}\end{aligned} \tag{29}$$

$$x_{t-1} = (1 - L\gamma_{t-2})x_{t-2} - \gamma_{t-2}\xi_{t-2} \tag{30}$$

Therefore we have

$$\begin{aligned}x_t &= (1 - L\gamma_{t-1})((1 - L\gamma_{t-2})x_{t-2} - \gamma_{t-2}\xi_{t-2}) - \gamma_{t-1}\xi_{t-1} \\ &= (1 - L\gamma_{t-1})(1 - L\gamma_{t-2})x_{t-2} - (1 - L\gamma_{t-1})\gamma_{t-2}\xi_{t-2} - \gamma_{t-1}\xi_{t-1} \\ &= (1 - L\gamma_{t-1})(1 - L\gamma_{t-2})(1 - L\gamma_{t-3})x_{t-3} - (1 - L\gamma_{t-1})(1 - L\gamma_{t-2})\gamma_{t-3}\xi_{t-3} - \\ &(1 - L\gamma_{t-1})\gamma_{t-2}\xi_{t-2} - \gamma_{t-1}\xi_{t-1} \\ &= (1 - L\gamma_{t-1})(1 - L\gamma_{t-2})(1 - L\gamma_{t-3})(1 - L\gamma_{t-4})x_{t-4} \\ &- (1 - L\gamma_{t-1})(1 - L\gamma_{t-2})(1 - L\gamma_{t-3})\gamma_{t-4}\xi_{t-4} \\ &- (1 - L\gamma_{t-1})(1 - L\gamma_{t-2})\gamma_{t-3}\xi_{t-3} - (1 - L\gamma_{t-1})\gamma_{t-2}\xi_{t-2} - \gamma_{t-1}\xi_{t-1} \\ &= ...\end{aligned} \tag{31}$$

From the observation we find the rule for the exact expansion: 1) for the first term in equation (31), it's just the product from $1-L\gamma_{t-1}$ to $1-L\gamma_0$ and then multiplied by x_0. 2) for the rest sum of the product terms, the term $\gamma_j\xi_j$ has coefficient producted from $1-L\gamma_{j+1}$ to $1-L\gamma_{t-1}$, where j is not equal to t-1. 3) $\gamma_{t-1}\xi_{t-1}$ has coefficient 1, therefore it left itself in the end. So,

$$\begin{aligned} x_t &= \prod_{j=0}^{t-1}(1-L\gamma_j)x_0 - \sum_{j=0}^{t-2}\gamma_j\xi_j \prod_{i=j+1}^{t-1}(1-L\gamma_i) - \gamma_{t-1}\xi_{t-1} \\ &= \prod_{j=0}^{t-1}(1-L\gamma_j)x_0 - \sum_{j=0}^{t-2}\gamma_j \prod_{i=j+1}^{t-1}(1-L\gamma_i)\xi_j - \gamma_{t-1}\xi_{t-1} \end{aligned} \tag{32}$$

Question 4: $\xi \sim \mathcal{N}(0, \frac{\sigma^2}{d}\mathbb{I}_d)$. We have fixed $\delta \in (0,1)$. Show that with dimension d $\geq d_0 = O(\log T/\delta)$, for any $2 \leq t \leq T$, we have:

$$\|\nabla F(x_t)\|^2 \geq \frac{L}{2}\left(\Delta\prod_{j=0}^{t-1}(1-L\gamma_j)^2 + L\sigma^2\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1-L\gamma_i)^2 + L\sigma^2\gamma_{t-1}^2\right) \tag{33}$$

From the last question, we have x_t also serves to a Gaussian distribution, where its mean is equal to $\prod_{j=0}^{t-1}(1-L\gamma_j)x_0$, and its variance is equal to $\frac{\eta_t}{d}\mathbb{I}_d$, where η_t is equal to $\sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1-L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2$. From the hint we know that if x $\sim \mathcal{N}(y, \frac{\eta}{d}\mathbb{I}_d)$:

$$Pr\left(\left|\frac{\|x\|^2}{\|y\|^2+\eta} - 1\right| \leq \hat{\delta}\right) \geq 1 - 4\exp\left(-\frac{d\hat{\delta}^2}{24}\right), \quad \hat{\delta} \in (0,1) \tag{34}$$

The inequality inside Pr() could be further rewritten as: $\|x\|^2 - (\|y\|^2+\eta) \geq -\hat{\delta}(\|y\|^2+\eta) \Rightarrow \|x\|^2 \geq (1-\hat{\delta})(\|y\|^2+\eta)$. We insert y (mean) and η (η_t), which are achieved above, into this inequality, we'll have:

$$\|x_t\|^2 \geq (1-\hat{\delta})\left(\prod_{j=0}^{t-1}(1-L\gamma_j)^2x_0^2 + \sum_{j=0}^{t-2}\gamma_j^2\prod_{i=j+1}^{t-1}(1-L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2\right) \tag{35}$$

Since 1) $x_0 = \sqrt{\Delta/L}$ and 2) $\nabla F(x_t) = Lx_t$, then:

$$
\begin{aligned}
&\|x_t\|^2 \geq (1-\hat{\delta})\left(\prod_{j=0}^{t-1}(1-L\gamma_j)^2 x_0^2 + \sum_{j=0}^{t-2}\gamma_j^2 \prod_{i=j+1}^{t-1}(1-L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2\right) \\
&\Leftrightarrow \|x_t\|^2 \geq (1-\hat{\delta})\left(\prod_{j=0}^{t-1}(1-L\gamma_j)^2 \frac{\Delta}{L} + \sum_{j=0}^{t-2}\gamma_j^2 \prod_{i=j+1}^{t-1}(1-L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2\right) \\
&\Leftrightarrow L^2\|x_t\|^2 \geq L^2(1-\hat{\delta})\left(\prod_{j=0}^{t-1}(1-L\gamma_j)^2 \frac{\Delta}{L} + \sum_{j=0}^{t-2}\gamma_j^2 \prod_{i=j+1}^{t-1}(1-L\gamma_i)^2\sigma^2 + \gamma_{t-1}^2\sigma^2\right) \\
&\Leftrightarrow \|\nabla F(x_t)\|^2 \geq L(1-\hat{\delta})\left(\prod_{j=0}^{t-1}(1-L\gamma_j)^2 \Delta + L\sum_{j=0}^{t-2}\gamma_j^2 \prod_{i=j+1}^{t-1}(1-L\gamma_i)^2\sigma^2 + L\gamma_{t-1}^2\sigma^2\right)
\end{aligned} \tag{36}
$$

We could let $\hat{\delta} = \frac{1}{2}$, then the probability of

$$
\|\nabla F(x_t)\|^2 \geq \frac{L}{2}\left(\prod_{j=0}^{t-1}(1-L\gamma_j)^2 \Delta + L\sum_{j=0}^{t-2}\gamma_j^2 \prod_{i=j+1}^{t-1}(1-L\gamma_i)^2\sigma^2 + L\gamma_{t-1}^2\sigma^2\right) \tag{37}
$$

is at least $1 - 4\exp\left(-\frac{d*(1/2)^2}{24}\right) = 1 - 4\exp\left(-\frac{d}{96}\right)$. We should show that the following inequality holds with probability at least $1 - \frac{\delta}{T}$. It is equivalent to say that: $1 - 4\exp\left(-\frac{d}{96}\right) \geq 1 - \frac{\delta}{T}$ to achieve a looser lower bound. Then,

$$
\begin{aligned}
&1 - 4\exp\left(-\frac{d}{96}\right) \geq 1 - \frac{\delta}{T} \\
&\Leftrightarrow 4\exp\left(-\frac{d}{96}\right) \leq \frac{\delta}{T} \\
&\Leftrightarrow \exp\left(\frac{d}{96}\right) \geq \frac{4T}{\delta} \\
&\Leftrightarrow d \geq 96\log\left(\frac{4T}{\delta}\right)
\end{aligned} \tag{38}
$$

Therefore we could state that when $\hat{\delta} = \frac{1}{2}$, we let dimension $d \geq d_0 = 96\log\left(\frac{4T}{\delta}\right) = O(log(T/\delta))$, for any $2 \leq t \leq T$, we have:

$$
\|\nabla F(x_t)\|^2 \geq \frac{L}{2}\left(\Delta\prod_{j=0}^{t-1}(1-L\gamma_j)^2 + L\sigma^2\sum_{j=0}^{t-2}\gamma_j^2 \prod_{i=j+1}^{t-1}(1-L\gamma_i)^2 + L\sigma^2\gamma_{t-1}^2\right) \tag{39}
$$

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

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

$$\|\nabla F(x_t)\|^2 \geq \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2 - L\gamma)}\} \tag{40}$$

We first fix $\hat{\delta}$ (the same as the setting in the last sub question), then substitute a fixed $\gamma \in (0, 1/L)$ in this sub-question. Then inequality (39) could be rewritten as:

$$\begin{aligned}
\|\nabla F(x_t)\|^2 &\geq \frac{L}{2}\left(\Delta\prod_{j=0}^{t-1}(1 - L\gamma)^2 + L\sigma^2\sum_{j=0}^{t-2}\gamma^2\prod_{i=j+1}^{t-1}(1 - L\gamma)^2 + L\sigma^2\gamma^2\right) \\
&= \frac{L}{2}\left(\Delta(1 - L\gamma)^{2t} + L\sigma^2\sum_{j=0}^{t-2}\gamma^2(1 - L\gamma)^{2(t-j-1)} + L\sigma^2\gamma^2\right) \\
&= \frac{L}{2}\left(\Delta(1 - L\gamma)^{2t} + L\sigma^2\sum_{j=0}^{t-1}\gamma^2(1 - L\gamma)^{2(t-j-1)}\right) \\
&= \frac{L}{2}\left(\Delta(1 - L\gamma)^{2t} + L\sigma^2\gamma^2\frac{1 - (1 - L\gamma)^{2t}}{1 - (1 - L\gamma)^2}\right) \\
&= \frac{L}{2}\left(\Delta(1 - L\gamma)^{2t} + L\sigma^2\gamma^2\frac{1 - (1 - L\gamma)^{2t}}{2L\gamma - L^2\gamma^2}\right) \\
&= \frac{L}{2}\left(\Delta(1 - L\gamma)^{2t} + \frac{\gamma\sigma^2}{2 - L\gamma}\left(1 - (1 - L\gamma)^{2t}\right)\right) \\
&= \frac{L\Delta}{2}(1 - L\gamma)^{2t} + \frac{L\gamma\sigma^2}{2(2 - L\gamma)}\left(1 - (1 - L\gamma)^{2t}\right)
\end{aligned} \tag{41}$$

1) if $\frac{L\Delta}{2} \geq \frac{L\gamma\sigma^2}{2(2-L\gamma)}$, then

$$\begin{aligned}
\|\nabla F(x_t)\|^2 &\geq \frac{L\gamma\sigma^2}{2(2 - L\gamma)}(1 - L\gamma)^{2t} + \frac{L\gamma\sigma^2}{2(2 - L\gamma)}\left(1 - (1 - L\gamma)^{2t}\right) \\
&= \frac{L\gamma\sigma^2}{2(2 - L\gamma)} \\
&= \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2 - L\gamma)}\}
\end{aligned} \tag{42}$$

2) if $\frac{L\Delta}{2} \leq \frac{L\gamma\sigma^2}{2(2-L\gamma)}$, then

$$\begin{aligned}
\|\nabla F(x_t)\|^2 &\geq \frac{L\Delta}{2}(1 - L\gamma)^{2t} + \frac{L\Delta}{2}\left(1 - (1 - L\gamma)^{2t}\right) \\
&= \frac{L\Delta}{2} \\
&= \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2 - L\gamma)}\}
\end{aligned} \tag{43}$$

Then we have:

$$\|\nabla F(x_t)\|^2 \geq \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2-L\gamma)}\} \tag{44}$$

At the end, $1 - 4\exp\left(-\frac{d}{96}\right) \geq 1 - \delta$ to achieve a looser lower bound. Then,

$$\begin{aligned} & 1 - 4\exp\left(-\frac{d}{96}\right) \geq 1 - \delta \\ \Leftrightarrow\ & 4\exp\left(-\frac{d}{96}\right) \leq \delta \\ \Leftrightarrow\ & \exp\left(\frac{d}{96}\right) \geq \frac{4}{\delta} \\ \Leftrightarrow\ & d \geq 96\log\left(\frac{4}{\delta}\right) \end{aligned} \tag{45}$$

Since in the last question, we get the setting $d \geq 96\log\left(\frac{4T}{\delta}\right) \geq 96\log\left(\frac{4}{\delta}\right)$ since $T \geq 2$, it definitely satisfied the setting in Question 4, therefore we keep the same d as last question, with probability at least $1 - \delta$, we have for all $2 \leq t \leq T$

$$\|\nabla F(x_t)\|^2 \geq \min\{\frac{L\Delta}{2}, \frac{L\gamma\sigma^2}{2(2-L\gamma)}\} \tag{46}$$

Modified Extragradient

Question 1: Show that for any t $\geq$ 0, A_t and B_t are non-negative

We are given that:

$$A_t = \langle F(z_{t+1}) - F(z_t), \frac{1}{t+1}(z_0 - z_t) - \eta F(z_{t+\frac{1}{2}})\rangle. \tag{47}$$

and

$$z_{t+1} = z_t - \eta F(z_{t+\frac{1}{2}}) + \frac{1}{t+1}(z_0 - z_t) \tag{48}$$

Therefore, we could rewrite A_t:

$$A_t = \langle F(z_{t+1}) - F(z_t), z_{t+1} - z_t\rangle \tag{49}$$

Since f is convex-concave and smooth, which means that for any $z, \hat{z} \in \mathbb{R}^{d_1+d_2}$, $\langle F(z) - F(\hat{z}), z - \hat{z}\rangle \geq 0$. If we choose $z = z_{t+1}$ and $\hat{z} = z_t$, $A_t \geq 0$.

We are given that:

$$B_t = \|\eta F(z_t) - \eta F(z_{t+\frac{1}{2}})\|^2 - \frac{1}{L^2}\|F(z_{t+\frac{1}{2}}) - F(z_{t+1})\|^2 \tag{50}$$

and

$$z_{t+\frac{1}{2}} = z_t - \eta F(z_t) + \frac{1}{t+1}(z_0 - z_t) \tag{51}$$

We could then build the following relationship by the elimination of the term $\frac{1}{t+1}(z_0 - z_t)$:

$$z_{t+\frac{1}{2}} - z_{t+1} = \eta(F(z_{t+\frac{1}{2}}) - F(z_t)) \tag{52}$$

. Therefore, B_t can be rewritten as:

$$B_t = \|z_{t+\frac{1}{2}} - z_{t+1}\|^2 - \frac{1}{L^2}\|F(z_{t+\frac{1}{2}}) - F(z_{t+1})\|^2 \tag{53}$$

Since function f is L-smooth, we have $\|F(z_{t+\frac{1}{2}}) - F(z_{t+1})\| \leq L\|z_{t+\frac{1}{2}} - z_{t+1}\|$. So $B_t \geq \|z_{t+\frac{1}{2}} - z_{t+1}\|^2 - \frac{1}{L^2}L^2\|z_{t+\frac{1}{2}} - z_{t+1}\|^2 = 0$. Finish the proof for question 1.

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

$$V_{t+1} - V_t \leq \frac{(t+1)\eta^2 L^2}{2t(1-\eta^2 L^2)}\|\eta F(z_{t+1})\|^2 \tag{54}$$

We could use the inequality:

$$V_{t+1} - V_t \leq V_{t+1} - V_t + \eta t(t+1)A_t + \frac{t(t+1)}{2}B_t \tag{55}$$

since we know that $A_t, B_t \geq 0$, where the potential function V_t is defined by:

$$V_t = \frac{t(t+1)}{2}\|\eta F(z_t)\|^2 + t\langle \eta F(z_t), z_t - z_0\rangle \tag{56}$$

.

We need to upper bound the RHS of this inequality. We split the calculation of the RHS of this inequality into three terms: 1) $V_{t+1} - V_t$ 2) $\eta t(t+1)A_t$ 3) $\frac{t(t+1)}{2}B_t$, then sum them up.

1) $V_{t+1} - V_t$.

$$
\begin{aligned}
V_{t+1} - V_t = & \frac{(t+1)(t+2)}{2}\|\eta F(z_{t+1})\|^2 + (t+1)\langle \eta F(z_{t+1}), z_{t+1} - z_0\rangle - \frac{t(t+1)}{2}\|\eta F(z_t)\|^2 - \\
& t\langle \eta F(z_t), z_t - z_0\rangle \\
= & \frac{(t+1)(t+2)}{2}\|\eta F(z_{t+1})\|^2 + (t+1)\langle \eta F(z_{t+1}), z_{t+1} - z_t + z_t - z_0\rangle \\
& - \frac{t(t+1)}{2}\|\eta F(z_t)\|^2 - t\langle \eta F(z_t), z_t - z_0\rangle \\
= & \frac{(t+1)(t+2)}{2}\|\eta F(z_{t+1})\|^2 + (t+1)\langle \eta F(z_{t+1}), -\eta F(z_{t+\frac{1}{2}}) + \frac{1}{t+1}(z_0 - z_t) + z_t - z_0\rangle \\
& - \frac{t(t+1)}{2}\|\eta F(z_t)\|^2 - t\langle \eta F(z_t), z_t - z_0\rangle \\
= & \frac{(t+1)(t+2)}{2}\|\eta F(z_{t+1})\|^2 - \frac{t(t+1)}{2}\|\eta F(z_t)\|^2 - \eta^2(t+1)\langle F(z_{t+1}), F(z_{t+\frac{1}{2}})\rangle \\
& + \eta t\langle F(z_{t+1}) - F(z_t), z_t - z_0\rangle
\end{aligned} \tag{57}
$$

2) $\eta t(t+1)A_t$

$$
\begin{aligned}
\eta t(t+1)A_t &= \eta t(t+1)\langle F(z_{t+1}) - F(z_t), \frac{1}{t+1}(z_0 - z_t) - \eta F(z_{t+\frac{1}{2}})\rangle \\
&= \eta t\langle F(z_{t+1}) - F(z_t), z_0 - z_t\rangle - \eta^2 t(t+1)\langle F(z_{t+1}), F(z_{t+\frac{1}{2}})\rangle \\
&\quad + \eta^2 t(t+1)\langle F(z_t), F(z_{t+\frac{1}{2}})\rangle
\end{aligned} \tag{58}
$$

3) $\frac{t(t+1)}{2}B_t$

$$
\begin{aligned}
\eta\frac{t(t+1)}{2}B_t &= \frac{t(t+1)}{2}\left(\|\eta F(z_t) - \eta F(z_{t+\frac{1}{2}})\|^2 - \frac{1}{L^2}\|F(z_{t+\frac{1}{2}}) - F(z_{t+1})\|^2\right) \\
&= \eta^2\frac{t(t+1)}{2}\|F(z_t)\|^2 - \eta^2 t(t+1)\langle F(z_t), F(z_{t+\frac{1}{2}})\rangle + \eta^2\frac{t(t+1)}{2}\|F(z_{t+\frac{1}{2}})\|^2 \\
&\quad - \frac{t(t+1)}{2L^2}\|F(z_{t+\frac{1}{2}})\|^2 + \frac{t(t+1)}{L^2}\langle F(z_{t+\frac{1}{2}}), F(z_{t+1})\rangle - \frac{t(t+1)}{2L^2}\|F(z_{t+1})\|^2
\end{aligned} \tag{59}
$$

We then sum up the three items and do the elimination:

$$
\begin{aligned}
&V_{t+1} - V_t + \eta t(t+1)A_t + \frac{t(t+1)}{2}B_t = \frac{(t+1)(t+2)}{2}\|\eta F(z_{t+1})\|^2 - \frac{t(t+1)}{2}\|\eta F(z_t)\|^2 \\
&- \eta^2(t+1)\langle F(z_{t+1}), F(z_{t+\frac{1}{2}})\rangle + \eta t\langle F(z_{t+1}) - F(z_t), z_t - z_0\rangle + \eta t\langle F(z_{t+1}) - F(z_t), z_0 - z_t\rangle \\
&- \eta^2 t(t+1)\langle F(z_{t+1}), F(z_{t+\frac{1}{2}})\rangle + \eta^2 t(t+1)\langle F(z_t), F(z_{t+\frac{1}{2}})\rangle + \eta^2\frac{t(t+1)}{2}\|F(z_t)\|^2 \\
&-\eta^2 t(t+1)\langle F(z_t), F(z_{t+\frac{1}{2}})\rangle + \eta^2\frac{t(t+1)}{2}\|F(z_{t+\frac{1}{2}})\|^2 - \frac{t(t+1)}{2L^2}\|F(z_{t+\frac{1}{2}})\|^2 \\
&+ \frac{t(t+1)}{L^2}\langle F(z_{t+\frac{1}{2}}), F(z_{t+1})\rangle - \frac{t(t+1)}{2L^2}\|F(z_{t+1})\|^2 \\
&= \left(\eta^2\frac{(t+1)(t+2)}{2} - \frac{t(t+1)}{2L^2}\right)\|F(z_{t+1})\|^2 + \left(\frac{t(t+1)}{L^2} - \eta^2(t+1)^2\right)\langle F(z_{t+1}), F(z_{t+\frac{1}{2}})\rangle \\
&+ \left(\eta^2\frac{t(t+1)}{2} - \frac{t(t+1)}{2L^2}\right)\|F(z_{t+\frac{1}{2}})\|^2
\end{aligned}
\tag{60}
$$

We then rewrite those coefficient terms in order to generate a squared term plus the result we want.

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

$$
\begin{aligned}
&\eta^2\frac{(t+1)(t+2)}{2} - \frac{t(t+1)}{2L^2} = (\eta^2L^2 - 1)\frac{t(t+1)}{2L^2} + \eta^2(t+1) \\
&= (\eta^2L^2 - 1)\frac{t(t+1)}{2L^2} + \eta^2(t+1) + \frac{\eta^4L^2(t+1)}{2(\eta^2L^2-1)t} - \frac{\eta^4L^2(t+1)}{2(\eta^2L^2-1)t} \\
&= \frac{t+1}{2L^2}\frac{(\eta^2L^2-1)^2t^2 + 2\eta^2L^2(\eta^2L^2-1)t + \eta^4L^4}{(\eta^2L^2-1)t} - \frac{\eta^4L^4(t+1)}{2L^2(\eta^2L^2-1)t} \\
&= \frac{t+1}{2L^2}\frac{(\eta^2L^2 + (\eta^2L^2-1)t)^2}{(\eta^2L^2-1)t} - \frac{\eta^4L^4(t+1)}{2L^2(\eta^2L^2-1)t}
\end{aligned}
\tag{61}
$$

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

$$
\begin{aligned}
&\frac{t(t+1)}{L^2} - \eta^2(t+1)^2 = \frac{t(t+1)}{L^2} - \eta^2\frac{t+1}{L^2}(tL^2 + L^2) \\
&= \frac{t+1}{L^2}\left(t - t\eta^2L^2 - \eta^2L^2\right) \\
&= -\frac{t+1}{L^2}\left((\eta^2L^2-1)t + \eta^2L^2\right)
\end{aligned}
\tag{62}
$$

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

$$
\begin{aligned}
\eta^2\frac{t(t+1)}{2} - \frac{t(t+1)}{2L^2} &= \frac{t(t+1)}{2}\left(\eta^2 - \frac{1}{L^2}\right) \\
&= \frac{t+1}{2L^2}\left(\eta^2L^2 - 1\right)t
\end{aligned}
\tag{63}
$$

Therefore,

$$
\begin{aligned}
&V_{t+1} - V_t + \eta t(t+1)A_t + \frac{t(t+1)}{2}B_t \\
&= \left(\frac{t+1}{2L^2}\frac{(\eta^2L^2 + (\eta^2L^2-1)t)^2}{(\eta^2L^2-1)t} - \frac{\eta^4L^4(t+1)}{2L^2(\eta^2L^2-1)t}\right)\|F(z_{t+1})\|^2 \\
&- \frac{t+1}{L^2}\left((\eta^2L^2-1)t + \eta^2L^2\right)\langle F(z_{t+1}), F(z_{t+\frac{1}{2}})\rangle + \frac{t+1}{2L^2}\left(\eta^2L^2-1\right)t\|F(z_{t+\frac{1}{2}})\|^2 \\
&= -\frac{t+1}{2L^2}\|\frac{\eta^2L^2 + (\eta^2L^2-1)t}{\sqrt{(1-\eta^2L^2)t}}F(z_{t+1}) - \sqrt{(1-\eta^2L^2)t}F(z_{t+\frac{1}{2}})\|^2 \\
&+ \frac{\eta^4L^4(t+1)}{2L^2(1-\eta^2L^2)t}\|F(z_{t+1})\|^2 \quad \text{first norm term is achieved by perfect square trinomial} \\
&\leq \frac{\eta^4L^4(t+1)}{2L^2(1-\eta^2L^2)t}\|F(z_{t+1})\|^2 \quad \text{non positiveness of norm} \\
&= \frac{(t+1)\eta^2L^2}{2t(1-\eta^2L^2)}\|\eta F(z_{t+1})\|^2
\end{aligned}
\tag{64}
$$

Notice that since $\eta < \frac{1}{\sqrt{3}L}$, so $\eta^2L^2 - 1 < 0$, we need to rewrite $\eta^2L^2 - 1$ as $-(1-\eta^2L^2)$ in the equation to get the squared root $\sqrt{1-\eta^2L^2}$. That's why there's a negative symbol before the squared norm term. Finish the proof for question 2.

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

$$
\frac{T^2}{4}\|\eta F(z_T)\|^2 \leq (1+\eta L)^2\|z^* - z_0\|^2 + \sum_{t=2}^{T-1}\frac{\eta^2L^2}{1-\eta^2L^2}\|\eta F(z_t)\|^2 \tag{65}
$$

We have assumed the existence of $z^* = (x^*, y^*)$ such that $F(z^*) = 0$. Given z_t and z^*, we have: $\langle F(z_t) - F(z^*), z_t - z^*\rangle \geq 0 \Leftrightarrow \langle F(z_t), z_t - z^*\rangle \geq 0$, which also means that $\langle F(z_t), z^* - z_0\rangle \leq \langle F(z_t), z_t - z_0\rangle$. According to the definition of V_t and the inequality $\langle a, b\rangle \geq -\frac{\lambda}{4}\|a\|^2 - \frac{1}{\lambda}\|b\|^2$, we have:

$$
\begin{aligned}
V_t &= \frac{t(t+1)}{2}\|\eta F(z_t)\|^2 + t\langle \eta F(z_t), z_t - z_0\rangle \\
&\geq \frac{t(t+1)}{2}\|\eta F(z_t)\|^2 + t\langle \eta F(z_t), z^* - z_0\rangle \\
&\geq \frac{t(t+1)}{2}\|\eta F(z_t)\|^2 - \frac{t(t+1)}{4}\|\eta F(z_t)\|^2 - \frac{t}{t+1}\|z^* - z_0\|^2 \\
&= \frac{t(t+1)}{4}\|\eta F(z_t)\|^2 - \frac{t}{t+1}\|z^* - z_0\|^2 \\
&\geq \frac{t(t+1)}{4}\|\eta F(z_t)\|^2 - \|z^* - z_0\|^2 \\
&\Rightarrow \frac{T(T+1)}{4}\|\eta F(z_T)\|^2 \leq \|z^* - z_0\|^2 + V_T
\end{aligned}
\tag{66}
$$

In the last question, we had proved that:

$$V_{t+1} - V_t \leq \frac{(t+1)\eta^2 L^2}{2t(1-\eta^2 L^2)} \|\eta F(z_{t+1})\|^2 \tag{67}$$

We know that for all $t \geq 1$, $\frac{t+1}{2t} \leq 1$, therefore if we perform telescoping

$$\begin{aligned} V_T - V_1 &\leq \sum_{t=1}^{T-1} \frac{(t+1)\eta^2 L^2}{2t(1-\eta^2 L^2)} \|\eta F(z_{t+1})\|^2 \\ &\leq \sum_{t=1}^{T-1} \frac{\eta^2 L^2}{(1-\eta^2 L^2)} \|\eta F(z_{t+1})\|^2 \end{aligned} \tag{68}$$

Combined with above inequality related with T and the inequality with regard to V_1: $V_1 \leq (2\eta L + \eta^2 L^2)\|z_0 - z^*\|^2$, we'll have:

$$\begin{aligned} \frac{T(T+1)}{4} \|\eta F(z_T)\|^2 &\leq \|z^* - z_0\|^2 + V_T \\ &\leq \|z^* - z_0\|^2 + \sum_{t=1}^{T-1} \frac{\eta^2 L^2}{(1-\eta^2 L^2)} \|\eta F(z_{t+1})\|^2 + V_1 \\ \Rightarrow \left(\frac{T(T+1)}{4} - \frac{\eta^2 L^2}{1-\eta^2 L^2} \right) \|\eta F(z_T)\|^2 &\leq (1 + 2\eta L + \eta^2 L^2)\|z_0 - z^*\|^2 + \sum_{t=2}^{T-1} \frac{\eta^2 L^2}{(1-\eta^2 L^2)} \|\eta F(z_t)\|^2 \end{aligned} \tag{69}$$

Since $\eta < \frac{1}{\sqrt{3}L}$, so $\frac{\eta^2 L^2}{1-\eta^2 L^2} \leq \frac{1}{2} \leq \frac{t}{4}, \quad \forall t \geq 2,$

$$\begin{aligned} \left(\frac{T(T+1)}{4} - \frac{T}{4} \right) \|\eta F(z_T)\|^2 &\leq \left(\frac{T(T+1)}{4} - \frac{\eta^2 L^2}{1-\eta^2 L^2} \right) \|\eta F(z_T)\|^2 \\ &\leq (1 + 2\eta L + \eta^2 L^2)\|z_0 - z^*\|^2 + \sum_{t=2}^{T-1} \frac{\eta^2 L^2}{(1-\eta^2 L^2)} \|\eta F(z_t)\|^2 \\ \Rightarrow \frac{T^2}{4} \|\eta F(z_T)\|^2 &\leq (1 + \eta L)^2 \|z_0 - z^*\|^2 + \sum_{t=2}^{T-1} \frac{\eta^2 L^2}{(1-\eta^2 L^2)} \|\eta F(z_t)\|^2 \end{aligned} \tag{70}$$

Finish the proof for question 3.

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

$$\|F(z_T)\|^2 \leq \frac{4(1+\eta L)^2}{\eta^2(1-3\eta^2 L^2)T^2} \|z^* - z_0\|^2 \tag{71}$$

Let $a_T = \frac{\eta^2 \|F(z_T)\|^2}{\|z^* - z_0\|^2}$, then it is equivalent to show that:

$$a_T \leq \frac{4(1+\eta L)^2}{(1-3\eta^2 L^2)T^2} \tag{72}$$

By dividing $\|z^* - z_0\|^2$ on both sides of the inequality in question 3, and using the definition of a_T, we have:

$$\frac{T^2}{4} \cdot a_T \leq (1+\eta L)^2 + \frac{\eta^2 L^2}{1-\eta^2 L^2} \sum_{t=2}^{T-1} a_t \tag{73}$$

We could prove the second inequality (72) by induction: 1) when T = 2, we have:

$$\frac{2^2}{4} a_2 \leq (1+\eta L)^2 \leq \frac{(1+\eta L)^2}{(1-3\eta^2 L^2)} \tag{74}$$

2) Suppose T = k - 1,

$$a_{k-1} \leq \frac{4(1+\eta L)^2}{(1-3\eta^2 L^2)(k-1)^2} \tag{75}$$

3) When T = k,

$$\begin{aligned}
\frac{k^2}{4} \cdot a_k &\leq (1+\eta L)^2 + \frac{\eta^2 L^2}{1-\eta^2 L^2} \sum_{t=2}^{k-1} a_t \\
&\leq (1+\eta L)^2 + \frac{\eta^2 L^2}{1-\eta^2 L^2} a_2 + \frac{\eta^2 L^2}{1-\eta^2 L^2} \sum_{t=3}^{k-1} a_t \\
&\leq (1+\eta L)^2 + \frac{\eta^2 L^2}{1-\eta^2 L^2} (1+\eta L)^2 + \frac{\eta^2 L^2}{1-\eta^2 L^2} \sum_{t=3}^{k-1} a_t \\
&\leq \frac{(1+\eta L)^2}{1-\eta^2 L^2} + \frac{\eta^2 L^2}{1-\eta^2 L^2} \sum_{t=3}^{k-1} \frac{4(1+\eta L)^2}{(1-3\eta^2 L^2) t^2} \\
&\leq \frac{(1+\eta L)^2}{1-\eta^2 L^2} + \frac{\eta^2 L^2}{1-\eta^2 L^2} \frac{4(1+\eta L)^2}{(1-3\eta^2 L^2)} \sum_{t=3}^{k-1} \frac{1}{t^2}
\end{aligned} \tag{76}$$

We know that $\sum_{t=3}^{k-1} \frac{1}{t^2} \leq \sum_{t=3}^{\infty} \frac{1}{t^2} \leq \frac{\pi^2}{6} - \frac{1}{1} - \frac{1}{4} \leq \frac{1}{2}$. Therefore,

$$\begin{aligned}
\frac{k^2}{4} \cdot a_k &\leq \frac{(1+\eta L)^2}{1-\eta^2 L^2} + \frac{\eta^2 L^2}{1-\eta^2 L^2} \frac{4(1+\eta L)^2}{(1-3\eta^2 L^2)} \sum_{t=3}^{k-1} \frac{1}{t^2} \\
&\leq \frac{(1+\eta L)^2}{1-\eta^2 L^2} + \frac{\eta^2 L^2}{1-\eta^2 L^2} \frac{2(1+\eta L)^2}{(1-3\eta^2 L^2)} \\
&= (1+\eta L)^2 \left(\frac{1-3\eta^2 L^2 + 2\eta^2 L^2}{(1-3\eta^2 L^2)(1-\eta^2 L^2)} \right) \\
&= \frac{(1+\eta L)^2}{1-3\eta^2 L^2} \\
&\Rightarrow a_k \leq \frac{4(1+\eta L)^2}{(1-3\eta^2 L^2) k^2}
\end{aligned} \tag{77}$$

The induction is finished.
We applied back $a_k = \frac{\eta^2\|F(z_k)\|^2}{\|z^*-z_0\|^2} \Rightarrow a_T = \frac{\eta^2\|F(z_T)\|^2}{\|z^*-z_0\|^2} \leq \frac{4(1+\eta L)^2}{(1-3\eta^2L^2)T^2}$, therefore,

$$\|F(z_T)\|^2 \leq \frac{4(1+\eta L)^2}{\eta^2(1-3\eta^2L^2)T^2}\|z^* - z_0\|^2 \tag{78}$$

Question 5: Let $\mathcal{X} = \mathcal{B}^{d1}(x^*, \|z_0 - z^*\|)$ and $\mathcal{Y} = \mathcal{B}^{d2}(y^*, \|z_0 - z^*\|)$, 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:

$$\max_{y\in\mathcal{Y}} f(x_T, y) - \min_{x\in\mathcal{X}} f(x, y_T) \leq \frac{2\sqrt{2}(1+\eta L)}{\eta\sqrt{1-3\eta^2L^2}T}\|z^* - z_0\|^2 \tag{79}$$

We know that f is concave on y, we have:

$$f(x_T, y) - f(x_T, y_T) \leq \langle \nabla_y f(x_T, y_T), y - y_T\rangle \tag{80}$$

And we know that f is convex on x, we have:

$$f(x, y_T) - f(x_T, y_T) \geq \langle \nabla_x f(x_T, y_T), x - x_T\rangle \tag{81}$$

Therefore,

$$\begin{aligned}
\max_{y\in\mathcal{Y}} f(x_T, y) - \min_{x\in\mathcal{X}} f(x, y_T) &= \max_{y\in\mathcal{Y}}(f(x_T, y) - f(x_T, y_T)) - \min_{x\in\mathcal{X}}(f(x, y_T) - f(x_T, y_T)) \\
&\leq \max_{y\in\mathcal{Y}}\langle \nabla_y f(x_T, y_T), y - y_T\rangle - \min_{x\in\mathcal{X}}\langle \nabla_x f(x_T, y_T), x - x_T\rangle \\
&\leq \max_{y\in\mathcal{Y}}\langle \nabla_y f(x_T, y_T), y - y_T\rangle + \max_{x\in\mathcal{X}}\langle \nabla_x f(x_T, y_T), x_T - x\rangle
\end{aligned} \tag{82}$$

We know that $z_T = (x_T, y_T)$, and $F(z_T) = (\nabla_x f(x_T, y_T), -\nabla_y f(x_T, y_T))$. We know that $z = (x, y)$, then: $z - z_T = (x - x_T, y - y_T)$, so

$$\max_{y\in\mathcal{Y}} f(x_T, y) - \min_{x\in\mathcal{X}} f(x, y_T) \leq \max_{z\in\mathcal{Z}}\langle F(z_T), z_T - z\rangle \tag{83}$$

$\|z_T - z\| = \sqrt{\|x_T - x\|^2 + \|y_T - y\|^2}$. $\max_{z\in\mathcal{Z}}\|z_T - z\| = \sqrt{2R^2} = \sqrt{2\|z_0 - z^*\|^2}$
Therefore, using the result from question 4 with Cauchy-Schwarz inequality:

$$\begin{aligned}
\max_{y\in\mathcal{Y}} f(x_T, y) - \min_{x\in\mathcal{X}} f(x, y_T) &\leq \max_{z\in\mathcal{Z}}\langle F(z_T), z_T - z\rangle \\
&\leq \max_{z\in\mathcal{Z}}\|F(z_T)\|\|z_T - z\| \\
&\leq \sqrt{\frac{4(1+\eta L)^2}{\eta^2(1-3\eta^2L^2)T^2}\|z^* - z_0\|^2} * \sqrt{2\|z_0 - z^*\|^2} \\
&= \frac{2(1+\eta L)}{\eta\sqrt{1-3\eta^2L^2}T}\|z_0 - z^*\| * \sqrt{2}\|z_0 - z^*\| \\
&= \frac{2\sqrt{2}(1+\eta L)}{\eta\sqrt{1-3\eta^2L^2}T}\|z_0 - z^*\|^2
\end{aligned} \tag{84}$$

Proof for question 5 ends.