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

Student Name: Dania Sana

Exercise 1: Subgradients on Convex Sets

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 that $y^* \in \mathcal{K}$ is a minimizer of f over $\mathcal{K}$ is and only if there exists a subgradient $g \in \partial f(y^*)$ such that:

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

Solution 1):

- Let us first prove the easiest direction of this statement, mainly, that if there exists a subgradient $g \in \partial f(y^*)$ such that $\langle y - y^*, g \rangle \geq 0$ for every $y \in \mathcal{K}$, then y^* is a minimizer of f over $\mathcal{K}$. Since $g \in \partial f(y^*)$, from the definition of the subgradient at the point y^* (handout 9, page 8), the following inequality holds, for every $y \in \mathbb{R}^d$:

  $$f(y) \geq f(y^*) + \langle g, y - y^* \rangle$$

  Since the above inequality holds for every $y \in \mathbb{R}^d$, it holds in particular for every $y \in \mathcal{K} \subseteq \mathbb{R}^d$. Moreover, for $y \in \mathcal{K}$, it is true that $\langle g, y - y^* \rangle = \langle y - y^*, g \rangle \geq 0$. Thus, using these two inequalities we obtain the following valid result for every $y \in \mathcal{K}$:

  $$f(y) \geq f(y^*) + \langle g, y - y^* \rangle \geq f(y^*) + 0 = f(y^*)$$

  We just showed that $\forall y \in \mathcal{K}$, $f(y) \geq f(y^*)$ implying that y^* is a minimizer of f over $\mathcal{K}$.

- Let us now prove the other direction of this statement, mainly, that if $y^* \in \mathcal{K}$ (in its interior; if $\mathcal{K}$ is open, this is always the case) is a minimizer of f over $\mathcal{K}$, then there exists a subgradient $g \in \partial f(y^*)$ such that:

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

  Since $y^* \in \mathcal{K}$ is a minimizer of f over $\mathcal{K}$, then $f(y) \geq f(y^*)$ for all $y \in \mathcal{K}$. In order to prove this result, we make use of the property of the directional derivative at the point y^* in the direction $y - y^*$, where $y \in \mathcal{K}$, stated in handout 9, page 43 as below (since f is a convex function, and $y^* \in \mathcal{K} \subseteq \mathbb{R}^d$ i.e., $y^* \in int(dom(f)) = \mathbb{R}^d$ as $\mathbb{R}^d$ is an open set; theorem proved in the book "Convex Analysis", by R.Tyrrell Rockafellar, Theorem 23.4, as indicated by handout 9):

  $$f'(y^*, y - y^*) = \lim_{t \to 0^+} \frac{f(y^* + t(y - y^*)) - f(y^*)}{t} = \max_{g \in \partial f(y^*)} \langle y - y^*, g \rangle = \max_{g \in \partial f(y^*)} g^\top (y - y^*)$$

  Since $\mathcal{K}$ is a convex set, and $y, y^* \in \mathcal{K}$, then $y^* + t(y - y^*) = t \cdot y + (1 - t)y^* \in \mathcal{K}$ for sufficiently small, positive t (indeed if $t \in [0, 1]$). Thus for $t \to 0^+$, $y^* + t(y - y^*) \in \mathcal{K}$

and since y^* is a minimizer of f over $\mathcal{K}$, then as $t \to 0^+$ $f(y^* + t(y - y^*)) \geq f(y^*)$. This leads to the following inequality

$$\max_{g \in \partial f(y^*)} \langle y - y^*, g \rangle = \lim_{t \to 0^+} \frac{f(y^* + t(y - y^*)) - f(y^*)}{t} \geq 0$$

We just showed a very important result, mainly that $\max_{g \in \partial f(y^*)} \langle y - y^*, g \rangle \geq 0$, $\forall y \in \mathcal{K}$. Let us now define the closed set:

$$\mathcal{B}_\epsilon = \{x + y^* \in \mathbb{R}^d \mid ||x||_2 \leq \epsilon\}$$

for some $\epsilon \geq 0$. First, we need to show that $\mathcal{B}_\epsilon$ is a convex set. For this, it is necessary and sufficient to show that if $y_1, y_2 \in \mathcal{B}_\epsilon$, and $\lambda \in [0, 1]$, then $\lambda y_1 + (1 - \lambda) y_2 \in \mathcal{B}_\epsilon$. Since $y_1, y_2 \in \mathcal{B}_\epsilon$, then $y_1 = x_1 + y^*$ and $y_2 = x_2 + y^*$ with $||x_1||_2 \leq \epsilon$ and $||x_2||_2 \leq \epsilon$. In this case $\lambda y_1 + (1 - \lambda) y_2 = \lambda(x_1 + y^*) + (1 - \lambda)(x_2 + y^*) = [\lambda x_1 + (1 - \lambda) x_2] + y^*$. Since $\lambda x_1 + (1 - \lambda) x_2 \in \mathbb{R}^d$ and $||\lambda x_1 + (1 - \lambda) x_2||_2 \leq \lambda ||x_1||_2 + (1 - \lambda)||x_2||_2 \leq \lambda \cdot \epsilon + (1 - \lambda)\epsilon = \epsilon$ (given that $||x_1||_2 \leq \epsilon$ and $||x_2||_2 \leq \epsilon$, since $y_1, y_2 \in \mathcal{B}_\epsilon$). Thus $\lambda y_1 + (1 - \lambda) y_2 \in \mathcal{B}_\epsilon$ and this concludes that $\mathcal{B}_\epsilon$ is a convex set. Finally, $\mathcal{B}_\epsilon$ is a closed and convex set. Since $\mathcal{K}$ is a convex set and $\mathcal{B}_\epsilon$ are convex, then $\mathcal{B}_\epsilon \cap \mathcal{K}$ is convex as well (since if $p_1, p_2 \in \mathcal{B}_\epsilon \cap \mathcal{K}$, it implies that $p_1, p_2 \in \mathcal{B}_\epsilon$ and $p_1, p_2 \in \mathcal{K}$. Since $\mathcal{K}$ and $\mathcal{B}_\epsilon$ are convex, this implies that $\lambda p_1 + (1 - \lambda) p_2 \in \mathcal{B}_\epsilon$ and $\lambda p_1 + (1 - \lambda) p_2 \in \mathcal{K}$ meaning that $\lambda p_1 + (1 - \lambda) p_2 \in \mathcal{B}_\epsilon \cap \mathcal{K}$, thus $\mathcal{B}_\epsilon \cap \mathcal{K}$ is a convex set). Moreover, if there are at least three points in $\mathcal{K}$ that do not lie in a line, then we choose ϵ to be the largest possible value such as $\mathcal{B}_\epsilon \subseteq \mathcal{K}$. In this case $\mathcal{K} \cap \mathcal{B}_\epsilon = \mathcal{B}_\epsilon$ implying that $\mathcal{K} \cap \mathcal{B}_\epsilon$ is closed since $\mathcal{B}_\epsilon$ is a closed set. On the other hand, if all the points in $\mathcal{K}$ lie on a line (here included also the case when $\mathcal{K}$ contains a single point), then choose ϵ small enough such that $\exists m_1, m_2 \in \mathcal{K}$ which satisfy $||m_1 - y^*||_2 \leq \epsilon$ and $||m_2 - y^*||_2 \leq \epsilon$ in a way that "the closed interval (segment) centered at y^* and boundary points m_1 and m_2 $\subseteq \mathcal{K}$, then $\mathcal{K} \cap \mathcal{B}_\epsilon$ is exactly this closed interval centered at y^* which is a closed set as well. Thus, we choose ϵ such that $\mathcal{K} \cap \mathcal{B}_\epsilon$ is a closed set.

Moreover, $\partial f(y^*)$ is also a closed convex set. It is closed, since it is the intersection of an infinite set of halfspaces as demonstrated below:

$$\partial f(y^*) = \bigcap_{z \in dom(f)} \{g \mid f(z) \geq f(y^*) + g^\top (z - y^*)\}$$

To show that $\partial f(y^*)$ is convex, let $g_1, g_2 \in \partial f(y^*)$ and $\lambda \in [0, 1]$. Then, the inequalities hold $\forall z \in dom(f)$:

$$f(z) \geq f(y^*) + g_1^\top (z - y^*) \quad \leftrightarrow \quad \lambda f(z) \geq \lambda f(y^*) + \lambda g_1^\top (z - y^*)$$

$$f(z) \geq f(y^*) + g_2^\top (z - y^*) \quad \leftrightarrow \quad (1 - \lambda) f(z) \geq (1 - \lambda) f(y^*) + (1 - \lambda) g_2^\top (z - y^*)$$

Therefore we add both sides of the inequalities on the right to finally get: $f(z) = \lambda f(z) + (1 - \lambda) f(z) \geq \lambda f(y^*) + (1 - \lambda) f(y^*) + (\lambda g_1 + (1 - \lambda) g_2)^\top (z - y^*) = f(y^*) +$

$(\lambda g_1 + (1-\lambda)g_2)^\top (z - y^*)$. Therefore, we just proved that $\lambda g_1 + (1-\lambda)g_2 \in \partial f(y^*)$ implying that $\partial f(y^*)$ is convex. Finally, we just finished showing that $\partial f(y^*)$ is closed and convex.
Next, we show that $\partial f(y^*)$ is bounded. (Let us note that since f is convex, with $dom(f) = \mathbb{R}^d$ open, then it is also continuous: proven in exercise 1, homework 2.) Choose some $\epsilon' > 0$, finite and define the set $A = \{z \in dom(f) : ||z - y^*|| \leq \epsilon'\}$ (with y^* being the minimizer of f over $\mathcal{K}$) which is bounded and because f is continuous, then the values of f in this set A will be bounded. To show that $\partial f(y^*)$is bounded let us suppose the contrary, mainly that $\partial f(y^*)$ is unbounded. This implies, that there is a sequence $k_n \in \partial f(y^*)$ such that $||k_n||_2 \to \infty^+$. Take the sequence $y_n = y^* + \epsilon' k_n / ||k_n||_2$. It is clear that $y_n \in A$ (because $||y_n - y^*||_2 = \epsilon'$). Since $k_n \in \partial f(y^*)$, then $f(y_n) \geq f(y^*) + k_n^\top (y_n - y^*) = f(y^*) + \epsilon' ||k_n||_2 \to \infty$. This shows that $f(y_n)$ is unbounded, which contradicts the fact that f is bounded on a bounded domain, in this case in the domain set A. Therefore, this contradicts the claim that $\partial f(y^*)$ is unbounded, which means that we just showed that $\partial f(y^*)$ is bounded.
Previously, I showed the important result that $\forall y \in \mathcal{K}$ and y^* being the minimizer of f over $\mathcal{K}$, we have:

$$\max_{g \in \partial f(y^*)} \langle y - y^*, g \rangle \geq 0.$$

This implies that

$$\min_{p \in \mathcal{B}_\epsilon \cap \mathcal{K},\, g \in \partial f(y^*)} \langle p - y^*, g \rangle \geq 0$$

We proved that $\partial f(y^*)$ is closed, convex and bounded and we reasoned that by choosing ϵ properly, $\mathcal{B}_\epsilon \cap \mathcal{K}$ is closed and convex. Moreover $\phi(p, g) = \langle p - y^*, g \rangle$ is a (bi)linear function in both arguments, thus it is a convex-concave continuous function (convex in p and concave in g). Using all these facts, one can apply theorem 13.4, in handout 13, page 21 (whose proof can be found in the book *"Zur Theorie der Gesellschaftsspiele"* by John von Neumann, published in 1928.) which implies that:

$$\min_{p \in \mathcal{B}_\epsilon \cap \mathcal{K}} \max_{g \in \partial f(y^*)} \langle p - y^*, g \rangle = \max_{g \in \partial f(y^*)} \min_{p \in \mathcal{B}_\epsilon \cap \mathcal{K}} \langle p - y^*, g \rangle \geq 0$$

This result indicates that $\exists g \in \partial f(y^*)$ such that:

$$\min_{p \in \mathcal{B}_\epsilon \cap \mathcal{K}} \langle p - y^*, g \rangle \geq 0$$

(This holds because if for every $g \in \partial f(y^*)$, $\min_{p \in \mathcal{B}_\epsilon \cap \mathcal{K}} \langle p - y^*, g \rangle \leq 0$, then also $\max_{g \in \partial f(y^*)} \min_{p \in \mathcal{B}_\epsilon \cap \mathcal{K}} \langle p - y^*, g \rangle \leq 0$ which is not true as shown before.) In other words, for every $p \in \mathcal{B}_\epsilon \cap \mathcal{K}$ it is true that $\langle p - y^*, g \rangle \geq 0$. Every $y \in \mathcal{K}$ can be written as $tp + (1-t)y^* = t(p - y^*) + y^*$ for some $t \geq 0$ and $y \in \mathcal{B}_\epsilon \cap \mathcal{K} \subseteq \mathcal{K}$ (why? - this mainly suggests that given $y \in \mathcal{K}$, $\exists p \in \mathcal{B}_\epsilon \cap \mathcal{K}$ such that y is on the line connecting y^* and p. Well this is true, because y^* is in the interior of $\mathcal{K} \cap \mathcal{B}_\epsilon$ as proposed before, because of the way we chose ϵ. Thus the line connecting y with y^*

will go through the set $\mathcal{K} \cap \mathcal{B}_\epsilon$ and touches it at some point p). Thus given $y \in \mathcal{K}$, and $t \geq 0, p \in \mathcal{B}_\epsilon \cap \mathcal{K}$ such that $y = t(p - y^*) + y^*$ the following holds:

$$\langle y - y^*, g \rangle = \langle t(p - y^*) + y^* - y^*, g \rangle = t\langle p - y^*, g \rangle \geq 0$$

the inequality holds because as shown previously $\langle p - y^*, g \rangle \geq 0$ for every $p \in \mathcal{B}_\epsilon \cap \mathcal{K}$ and $t \geq 0$. We just proved the statement of this direction, i.e., that if y^* is a minimizer of f over $\mathcal{K}$, then there exists a subgradient $g \in \partial f(y^*)$ such that $\langle y - y^*, g \rangle \geq 0$ for every $y \in \mathcal{K}$.

Exercise 2: Smoothed Function

Consider the following composite optimization problem:

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

where $f : \mathcal{X} \to \mathbb{R}$ is L-smooth and convex function, and $g : \mathcal{X} \to \mathbb{R}$ is convex and possible non-smooth function. Assume g can be smoothed with constants 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:

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

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, $h(x_t) - \min_{x \in \mathcal{X}} h(x) \leq \frac{c\hat{L}}{t^2}$ for some constant $c > 0$. Now we apply $\mathcal{A}$ to minimize the smoothed composite function:

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

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

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

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

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

and

$$t \geq \frac{2\sqrt{2\alpha\beta c}}{\epsilon} + \frac{\sqrt{Lc}}{\sqrt{\epsilon}},$$

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

Solution 2):

$\Phi_\mu(x) = f(x) + g_\mu(x)$ is a smooth and convex function since the sum of two convex functions is convex (lecture notes lemma 2.19) and the sum of two smooth functions is smooth (lecture notes lemma 3.6) (g_μ and f are both convex and smooth according to the conditions of the problem). Moreover, Φ_μ is smooth with parameter $\hat{L} = L + \frac{\alpha}{\mu}$ since f is L-smooth and g_μ is $\frac{\alpha}{\mu}$-smooth (lemma 3.6 i) (proved in one of the exercise sessions) of the lecture notes states that the smoothness parameter of the sum of two smooth functions can be expressed as the sum of the corresponding smoothness parameters of each function). Now, we apply algorithm $\mathcal{A}$ to minimize $\Phi_\mu(x)$ and for some c, the following holds:

$$\Phi_\mu(x_t) - \min_{x\in\mathcal{X}} \Phi_\mu(x) \leq \frac{c\hat{L}}{t^2} = \frac{c}{t^2}[L + \frac{\alpha}{\mu}] = \frac{cL}{t^2} + \frac{c\alpha}{t^2\mu} \tag{1}$$

Since $g(x) - \beta\mu \leq g_\mu(x)$, $\forall x \in \mathcal{X}$, then in particular $g(x_t) - \beta\mu \leq g_\mu(x_t)$, and since $\Phi(x_t) = f(x_t) + g(x_t)$, the following holds:

$$\Phi(x_t) - \beta\mu = f(x_t) + g(x_t) - \beta\mu \leq f(x_t) + g_\mu(x_t) = \Phi_\mu(x_t) \tag{2}$$

Moreover, since $g_\mu(x) \leq g(x) + \beta\mu$, $\forall x \in \mathcal{X}$, we have the following:

$$\Phi_\mu(x) = g_\mu(x) + f(x) \leq g(x) + \beta\mu + f(x) = \Phi(x) + \beta\mu \tag{3}$$

This implies that $\min_{x\in\mathcal{X}} \Phi_\mu(x) \leq \min_{x\in\mathcal{X}} \Phi(x) + \beta\mu$ (why? - Since it holds that $\min_{x\in\mathcal{X}} \Phi_\mu(x) \leq \Phi_\mu(x) \leftrightarrow -\min_{x\in\mathcal{X}} \Phi_\mu(x) \geq -\Phi_\mu(x) \rightarrow \Phi(x) - \min_{x\in\mathcal{X}} \Phi_\mu(x) + \beta\mu \geq \Phi(x) - \Phi_\mu(x) + \beta\mu$. Since it is given that $\forall x \in \mathcal{X}, \Phi(x) - \Phi_\mu(x) + \beta\mu \geq 0$ from (3) this implies that $\Phi(x) - \min_{x\in\mathcal{X}} \Phi_\mu(x) + \beta\mu \geq \Phi(x) - \Phi_\mu(x) + \beta\mu \geq 0$, in particular $\Phi(x) - \min_{x\in\mathcal{X}} \Phi_\mu(x) + \beta\mu \geq 0$. Since this inequality holds $\forall x \in \mathcal{X}$, it holds in particular for that x for which $\Phi(x)$ reaches its minimum. Therefore, $\min_{x\in\mathcal{X}} \Phi(x) + \beta\mu \geq \min_{x\in\mathcal{X}} \Phi_\mu(x)$), which on the other hand, is equivalent to the following important result:

$$-\min_{x\in\mathcal{X}} \Phi(x) - \beta\mu \leq -\min_{x\in\mathcal{X}} \Phi_\mu(x) \tag{4}$$

We combine the inqualities (2), and (4) to get:

$$\Phi(x_t) - \min_{x\in\mathcal{X}} \Phi(x) - 2\beta\mu \leq \Phi_\mu(x_t) - \min_{x\in\mathcal{X}} \Phi_\mu(x) \quad \leftrightarrow \quad \Phi(x_t) - \min_{x\in\mathcal{X}} \Phi(x) \leq \Phi_\mu(x_t) - \min_{x\in\mathcal{X}} \Phi_\mu(x) + 2\beta\mu$$

Now I use the above inequality on the right and together with (1), I obtain the following result:

$$\Phi(x_t) - \min_{x\in\mathcal{X}} \Phi(x) \leq \Phi_\mu(x_t) - \min_{x\in\mathcal{X}} \Phi_\mu(x) + 2\beta\mu \leq \frac{cL}{t^2} + \frac{c\alpha}{t^2\mu} + 2\beta\mu \tag{5}$$

Finally, we showed that: $\Phi(x_t) - \min_{x\in\mathcal{X}} \Phi(x) \leq \frac{cL}{t^2} + \frac{c\alpha}{t^2\mu} + 2\beta\mu$. We need to find a $\mu > 0$ such that

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

Since 5 holds, we could just let (and solve with respect to μ):

$$\frac{cL}{t^2} + \frac{c\alpha}{t^2\mu} + 2\beta\mu = \frac{Lc}{t^2} + \frac{2\sqrt{2\alpha\beta c}}{t}$$

equivalent to the equalities:

$$\frac{\alpha c}{t^2\mu} + 2\beta\mu = \frac{2\sqrt{2\alpha\beta c}}{t} \quad \leftrightarrow \quad \alpha c + 2\beta t^2\mu^2 = 2\mu t\sqrt{2\alpha\beta c}$$

In other words we need to solve the equation $\alpha c + 2\beta t^2\mu^2 - 2\mu t\sqrt{2\alpha\beta c} = 0$ with respect to μ (it is a second degree equation which could be solved using the discriminant $D = (-2t\sqrt{2\alpha\beta c})^2 - 4\cdot 2\beta t^2\alpha c = 0$. In this case the solution for μ is $\frac{2t\sqrt{2\alpha\beta c}}{2\cdot 2\beta t^2} = \frac{\sqrt{2\alpha\beta c}}{2\beta t}$). Thus, if one chooses $\mu = \frac{\sqrt{2\alpha\beta c}}{2\beta t}$, it holds that:

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

Next, we need to prove that for

$$\mu = \sqrt{\frac{\alpha}{2\beta}} \cdot \frac{\epsilon}{\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\epsilon}} \quad \text{and} \quad t \geq \frac{2\sqrt{2\alpha\beta c}}{\epsilon} + \frac{\sqrt{Lc}}{\sqrt{\epsilon}} \tag{6}$$

it holds that:

$$\Phi(x_t) - \min_{x\in\mathcal{X}} \Phi(x) \leq \epsilon$$

We already know from (5) that:

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

Thus, it would be sufficient to show that for the values of t and μ as in (6), the below equivalent inequalities hold:

$$\frac{Lc}{t^2} + \frac{\alpha c}{t^2\mu} + 2\beta\mu \leq \epsilon \quad \leftrightarrow \quad 0 \leq t^2(\epsilon\mu - 2\beta\mu^2) - (Lc\mu + \alpha c)$$

Observe the function: $k(t) = t^2(\epsilon\mu - 2\beta\mu^2) - (Lc\mu + \alpha c)$. Let us calculate $\epsilon\mu - 2\beta\mu^2$ with μ as in (6):

$$\epsilon\mu - 2\beta\mu^2 = \frac{\epsilon^2\sqrt{\alpha}}{\sqrt{2\beta}(\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\epsilon})} - \frac{2\beta\epsilon^2\alpha}{(\sqrt{2\beta}(\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\epsilon}))^2} = \tag{7}$$

$$= \frac{\epsilon^2\sqrt{2\alpha\beta}\sqrt{2\alpha\beta + L\epsilon}}{(\sqrt{2\beta}(\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\epsilon}))^2} > 0 \tag{8}$$

Observe that the above expression is strictly positive because $\epsilon > 0$, $\beta > 0$ (otherwise, μ could not be defined as in (6)) and $\alpha > 0$ (otherwise μ could not be defined as in (6) for $\alpha < 0$ and if $\alpha = 0$, then $\mu = 0$, however, $\mu > 0$). Moreover, since $\sqrt{Lc}$ is valid, as we see in (6) it means that $L \geq 0$ since $c > 0$. These arguments, make clear that $\epsilon\mu - 2\beta\mu^2 > 0$.

Let $t_0 \geq 0$ be the value for t such that $k(t_0) = t_0^2(\epsilon\mu - 2\beta\mu^2) - (Lc\mu + \alpha c) = 0$. Since $\epsilon\mu - 2\beta\mu^2 > 0$, $k(t)$ is monotonically increasing for $t \in [t_0, \infty^+)$ (because $k'(t) = 2t(\epsilon\mu - 2\beta\mu^2)$ and since $\epsilon\mu - 2\beta\mu^2 > 0$, and $t_0 \geq 0$, it is true that $\forall t \in (t_0, \infty^+)$, $k'(t) > 0$ and thus from basic calculus this means that $k(t)$ is increasing in $[t_0, \infty^+)$). Thus, we prove the desired statement if we manage to show that:

$$\frac{2\sqrt{2\alpha\beta c}}{\epsilon} + \frac{\sqrt{Lc}}{\sqrt{\epsilon}} \geq t_0$$

because in that case, $\forall t \geq \frac{2\sqrt{2\alpha\beta c}}{\epsilon} + \frac{\sqrt{Lc}}{\sqrt{\epsilon}}$, it will hold that $k(t) \geq k(t_0) = 0$ and thus $t^2(\epsilon\mu - 2\beta\mu^2) - (Lc\mu + \alpha c) \geq 0 \leftrightarrow \frac{Lc}{t^2} + \frac{\alpha c}{t^2\mu} + 2\beta\mu \leq \epsilon$ which is what we want to show. The inequality above is equivalent to the following inequality (since both sides of it are positive, squaring is elligible):

$$\frac{8\alpha\beta c}{\epsilon^2} + \frac{Lc}{\epsilon} + \frac{4c\sqrt{2\alpha\beta L}}{\epsilon\sqrt{\epsilon}} = (\frac{2\sqrt{2\alpha\beta c}}{\epsilon} + \frac{\sqrt{Lc}}{\sqrt{\epsilon}})^2 \geq t_0^2 \qquad (*)$$

Let us estimate $t_0^2 = \frac{Lc\mu + \alpha c}{\epsilon\mu - 2\beta\mu^2}$, with μ as in (6). We start by calculating $Lc\mu + \alpha c$ as below:

$$Lc\mu + \alpha c = \frac{Lc\epsilon\sqrt{\alpha}}{\sqrt{2\beta}(\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\epsilon})} + \alpha c = \frac{Lc\epsilon\sqrt{\alpha} + \alpha c\sqrt{2\beta}(\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\epsilon})}{\sqrt{2\beta}(\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\epsilon})}$$

Finally:

$$Lc\mu + \alpha c = \frac{Lc\epsilon\sqrt{\alpha} + 2\alpha c\beta\sqrt{\alpha} + \alpha c\sqrt{2\beta}\sqrt{2\alpha\beta + L\epsilon}}{\sqrt{2\beta}(\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\epsilon})} \qquad (9)$$

In (8), we calculated $\epsilon\mu - 2\beta\mu^2$. Thus using (8) and (9) we get the following solution for t_0^2:

$$t_0^2 = \frac{Lc\mu + \alpha c}{\epsilon\mu - 2\beta\mu^2} = \frac{(Lc\epsilon\sqrt{\alpha} + 2\alpha c\beta\sqrt{\alpha} + \alpha c\sqrt{2\beta}\sqrt{2\alpha\beta + L\epsilon}) \cdot (\sqrt{2\beta}(\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\epsilon}))^2}{\sqrt{2\beta}(\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\epsilon}) \cdot \epsilon^2\sqrt{2\alpha\beta}\sqrt{2\alpha\beta + L\epsilon}} =$$

$$= \frac{(Lc\epsilon\sqrt{\alpha} + 2\alpha c\beta\sqrt{\alpha} + \alpha c\sqrt{2\beta}\sqrt{2\alpha\beta + L\epsilon}) \cdot (\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\epsilon})}{\epsilon^2\sqrt{\alpha}\sqrt{2\alpha\beta + L\epsilon}} =$$

$$= \frac{Lc\epsilon\sqrt{2\alpha\beta}}{\epsilon^2\sqrt{2\alpha\beta + L\epsilon}} + \frac{2\alpha\beta c\sqrt{2\alpha\beta}}{\epsilon^2\sqrt{2\alpha\beta + L\epsilon}} + \frac{Lc}{\epsilon} + \frac{2\alpha\beta c}{\epsilon^2} + \frac{2\alpha\beta c}{\epsilon^2} + \frac{c\sqrt{\alpha}\sqrt{2\beta}\sqrt{2\alpha\beta + L\epsilon}}{\epsilon^2}$$

Since $\sqrt{2\alpha\beta + L\epsilon} \geq \sqrt{L\epsilon} \rightarrow \frac{Lc\epsilon\sqrt{2\alpha\beta}}{\epsilon^2\sqrt{2\alpha\beta + L\epsilon}} \leq \frac{Lc\sqrt{2\alpha\beta}}{\epsilon\sqrt{L\epsilon}} = \frac{c\sqrt{2L\alpha\beta}}{\epsilon\sqrt{\epsilon}}$; since $\frac{\sqrt{2\alpha\beta}}{\sqrt{2\alpha\beta + L\epsilon}} \leq 1$, $\rightarrow \frac{2\alpha\beta c\sqrt{2\alpha\beta}}{\epsilon^2\sqrt{2\alpha\beta + L\epsilon}} \leq \frac{2\alpha\beta c}{\epsilon^2}$ and last, since $\sqrt{2\alpha\beta + L\epsilon} \leq \sqrt{2\alpha\beta} + \sqrt{L\epsilon} \rightarrow \frac{c\sqrt{\alpha}\sqrt{2\beta}\sqrt{2\alpha\beta + L\epsilon}}{\epsilon^2} \leq$

$\frac{c\sqrt{2\alpha\beta}(\sqrt{2\alpha\beta}+\sqrt{L\epsilon})}{\epsilon^2} = \frac{2\alpha\beta c}{\epsilon^2} + \frac{c\sqrt{2\alpha\beta L}}{\epsilon\sqrt{\epsilon}}$ We apply all these inequalities to provide an upper bound for t_0^2 as below:

$$t_0^2 = \frac{Lc\epsilon\sqrt{2\alpha\beta}}{\epsilon^2\sqrt{2\alpha\beta+L\epsilon}} + \frac{2\alpha\beta c\sqrt{2\alpha\beta}}{\epsilon^2\sqrt{2\alpha\beta+L\epsilon}} + \frac{Lc}{\epsilon} + \frac{2\alpha\beta c}{\epsilon^2} + \frac{2\alpha\beta c}{\epsilon^2} + \frac{c\sqrt{\alpha}\sqrt{2\beta}\sqrt{2\alpha\beta+L\epsilon}}{\epsilon^2}$$

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

$$= \frac{Lc}{\epsilon} + 4\frac{2\alpha\beta c}{\epsilon^2} + 2\frac{c\sqrt{2L\alpha\beta}}{\epsilon\sqrt{\epsilon}} \leq \frac{Lc}{\epsilon} + 4\frac{2\alpha\beta c}{\epsilon^2} + 4\frac{c\sqrt{2L\alpha\beta}}{\epsilon\sqrt{\epsilon}} = \left(\frac{\sqrt{Lc}}{\sqrt{\epsilon}} + \frac{2\sqrt{2\alpha\beta c}}{\epsilon}\right)^2$$

We just proved our goal inequality (*), that $t_0^2 \leq \left(\frac{\sqrt{Lc}}{\sqrt{\epsilon}} + \frac{2\sqrt{2\alpha\beta c}}{\epsilon}\right)^2$. Therefore, we can conclude that for t and μ as in (6), $\Phi(x_t) - \min_{x\in\mathcal{X}} \Phi(x) \leq \epsilon$.

Exercise 3: Proximal Non-Convex SGD

Consider the following stochastic optimization problem:

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

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 to 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:

$$\mathbb{E}[\nabla f(x,\xi)] = \nabla f(x), \qquad \mathbb{E}[||\nabla f(x,\xi) - \nabla f(x)||^2] \leq \sigma^2 \tag{10}$$

Consider the following method (Proximal Stochastic Gradient Descent)

$$x_{t+1} = prox_{\eta r}(x_t - \eta\nabla f(x_t, \xi_{t+1})), \tag{11}$$

where ξ_{t+1} are independent for all $t \geq 0$, $\eta > 0$ is the step-size. Recall that for any $\rho > 0$, the Moreau envelope of a function $\Phi(x)$ is given by

$$\Phi_\rho(x) := min_{y\in\mathbb{R}^d}\{\Phi(y) + \frac{\rho}{2}||y-x||^2\}. \tag{12}$$

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

$$\hat{x} := prox_{\Phi/\rho}(x) := argmin_{y\in\mathbb{R}^d}\{\Phi(y) + \frac{\rho}{2}||y-x||^2\} \tag{13}$$

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 $\hat{x}_t := prox_{\Phi/\rho}(x_t)$. Prove that:

$$\hat{x}_t = prox_{\eta r}(\eta\rho x_t - \eta\nabla f(\hat{x}_t) + (1-\eta\rho)\hat{x}_t).$$

Solution 3a)
It is given that $\hat{x}_t = prox_{\Phi/\rho}(x_t) = argmin_{y\in\mathbb{R}^d}\{\Phi(y) + \frac{\rho}{2}||y - x_t||^2\}$ which implies that $\forall z \in \mathbb{R}^d$ it holds:

$$\Phi(z) + \frac{\rho}{2}||z - x_t||^2 \geq \Phi(\hat{x}_t) + \frac{\rho}{2}||\hat{x}_t - x_t||^2 = \Phi(\hat{x}_t) + \frac{\rho}{2}||\hat{x}_t - x_t||^2 + 0^\top(z - \hat{x}_t)$$

which implies that $0 \in \partial k(\hat{x}_t)$ where k is the function given by: $k(y) = \Phi(y) + \frac{\rho}{2}||y-x_t||^2 = f(y)+r(y)+\frac{\rho}{2}||y-x_t||^2$. (Handout 9, page 16, suggests that if $h(x) = \beta_1 f_1(x)+\beta_2 f_2(x)$, then $\partial h(x) = \beta_1\partial f_1(x)+\beta_2\partial f_2(x)$. This moreover implies that if $h(x) = \beta_1 f_1(x)+\beta_2 f_2(x)+\beta_3 f_3(x)$, then $\partial h(x) = \beta_1\partial f_1(x)+\partial(\beta_2 f_2(x)+\beta_3 f_3(x)) = \beta_1\partial f_1(x) + \beta_2\partial f_2(x) + \beta_3\partial f_3(x)$. Thus given that $k(y) = f(y) + r(y) + \frac{\rho}{2}||y - x_t||^2$, it holds that $\partial k(y) = \partial f(y) + \partial r(y) + \partial\frac{\rho}{2}||y - x_t||^2)$. This suggests that

$$0 \in \partial k(\hat{x}_t) = \partial f(\hat{x}_t) + \partial r(\hat{x}_t) + \partial(\frac{\rho}{2}||y - x_t||^2 \mid_{y=\hat{x}_t}). \tag{14}$$

Since $\frac{\rho}{2}||y - x_t||^2$ is convex, it has subgradients for every y (Lemma 9.3, handout 9). Moreover, since it is differentiable, $\partial\frac{\rho}{2}||y - x_t||^2 \subseteq \nabla_y(\frac{\rho}{2}||y - x_t||^2) = \rho(y - x_t)$ (proved in exercise 1, exercise set 8), thus $\partial(\frac{\rho}{2}||y - x_t||^2 \mid_{y=\hat{x}_t}) = \rho(\hat{x}_t - x_t)$. Therefore, using (14), and this result, it holds that $\exists p_1 \in \partial f(\hat{x}_t)$ and $\exists p_2 \in \partial r(\hat{x}_t)$, such that:

$$0 = p_1 + p_2 + \rho(\hat{x}_t - x_t) \qquad (**)$$

However, since f is differentiable, it holds that $\partial f(\hat{x}_t) \subseteq \nabla f(\hat{x}_t)$ and since there is such $p_1 \in \partial f(\hat{x}_t)$ (given that (**) is satisfied), this implies that $p_1 = \nabla f(\hat{x}_t)$. Finally using (**), this means that $\exists p_2 \in \partial r(\hat{x}_t)$ such that:

$$0 = \nabla f(\hat{x}_t) + p_2 + \rho(\hat{x}_t - x_t) \quad \leftrightarrow \quad -\nabla f(\hat{x}_t) + \rho(x_t - \hat{x}_t) = p_2 \in \partial r(\hat{x}_t) \quad (***)$$

In order to show that $\hat{x}_t = prox_{\eta r}(\eta\rho x_t - \eta\nabla f(\hat{x}_t) + (1-\eta\rho)\hat{x}_t)$, it is sufficient to show that $\eta\rho x_t - \eta\nabla f(\hat{x}_t) + \hat{x}_t - \eta\rho\hat{x}_t - \hat{x}_t \in \partial\eta r(\hat{x}_t)$. (Why? - Exercise 3a from exercise set 9 states that if g is a convex function with $dom(g) = \mathbb{R}^d$, $y = prox_g(x) \leftrightarrow x - y \in \partial g(y)$. Given that r is convex implies that ηr is convex with $dom(\eta r) = \mathbb{R}^d$. Therefore, using the result from exercise 1, set 9, $\hat{x}_t = prox_{\eta r}(\eta\rho x_t - \eta\nabla f(\hat{x}_t) + (1-\eta\rho)\hat{x}_t) \leftrightarrow \eta\rho x_t - \eta\nabla f(\hat{x}_t) + \hat{x}_t - \eta\rho\hat{x}_t - \hat{x}_t \in \partial\eta r(\hat{x}_t)$. Moreover, $m \in \partial\eta r(\hat{x}_t)$ iff $\forall z \in \mathbb{R}^d$, $\eta r(z) \geq \eta r(\hat{x}_t) + m^\top(z - \hat{x}_t)$. This is equivalent to $r(z) \geq r(\hat{x}_t) + (\frac{m}{\eta})^\top(z - \hat{x}_t) \leftrightarrow \frac{m}{\eta} \in \partial r(\hat{x}_t))$. Thus this is equivalent to prove that:

$$\rho(x_t - \hat{x}_t) - \nabla f(\hat{x}_t) \in \partial r(\hat{x}_t)$$

which is already shown before in $(***)$.

(b) Let $\rho = 4L$, $\eta \leq \frac{2}{9L}$, and x^{t+1} is given by (11). Prove that for all $t \geq 0$:

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

Solution 3b):

x_{t+1} and $\hat{x}_t$ can be expressed as below:

$$x_{t+1} = prox_{\eta r}(x_t - \eta \nabla f(x_t, \xi_{t+1})) \tag{15}$$

$$\hat{x}_t = prox_{\eta r}(\eta\rho x_t - \eta \nabla f(\hat{x}_t) + \hat{x}_t - \rho\hat{x}_t) \quad \text{from 3a)} \tag{16}$$

In order to provide an upper bound for $||x_{t+1} - \hat{x}_t||^2$, we use the property in Lemma 10.9, handout 10 which was proven in the solution of homework 9, exercise 3c, which states that if g is a convex function with $dom(g) = \mathbb{R}^d$, it holds that:

$$||prox_g(x) - prox_g(y)||_2 \leq ||x - y||_2$$

Now on the norm $||\cdot||$, indicates the ℓ_2-norm $||\cdot||_2$. Thus using (15) and (16) and the fact that ηr is convex with $dom(\eta r) = \mathbb{R}^d$, we get the following series of inequalities/equalities:

$$||x_{t+1} - \hat{x}_t||^2 = ||prox_{\eta r}(x_t - \eta \nabla f(x_t, \xi_{t+1})) - prox_{\eta r}(\eta\rho x_t - \eta \nabla f(\hat{x}_t) + \hat{x}_t - \eta\rho\hat{x}_t)||^2 \tag{17}$$

$$\leq ||x_t - \eta \nabla f(x_t, \xi_{t+1}) - \eta\rho x_t + \eta \nabla f(\hat{x}_t) - \hat{x}_t + \eta\rho\hat{x}_t||^2 \tag{18}$$

$$= ||(1 - \eta\rho)(x_t - \hat{x}_t) + \eta(\nabla f(\hat{x}_t) - \nabla f(x_t, \xi_{t+1}))||^2 \tag{19}$$

$$= (1 - \eta\rho)^2||x_t - \hat{x}_t||^2 + \eta^2||\nabla f(\hat{x}_t) - \nabla f(x_t, \xi_{t+1})||^2 + \tag{20}$$

$$+2\eta(1 - \eta\rho)\langle x_t - \hat{x}_t, \nabla f(\hat{x}_t) - \nabla f(x_t, \xi_{t+1})\rangle \tag{21}$$

Consider the expression $||\nabla f(\hat{x}_t) - \nabla f(x_t, \xi_{t+1})||^2$ and take expectation with respect to ξ_{t+1} to get:

$$\mathbb{E}[||\nabla f(\hat{x}_t) - \nabla f(x_t, \xi_{t+1})||^2 | x_t] = \mathbb{E}[||\nabla f(\hat{x}_t)||^2 | x_t] + \mathbb{E}[||\nabla f(x_t, \xi_{t+1})||^2 | x_t] - \tag{22}$$

$$-2\mathbb{E}[\nabla f(\hat{x}_t)^\top \nabla f(x_t, \xi_{t+1}) | x_t] \tag{23}$$

$$= ||\nabla f(\hat{x}_t)||^2 + \mathbb{E}[||\nabla f(x_t, \xi_{t+1})||^2 | x_t] - 2\nabla f(\hat{x}_t)^\top \mathbb{E}[\nabla f(x_t, \xi_{t+1}) | x_t] \tag{24}$$

$$= ||\nabla f(\hat{x}_t)||^2 + \mathbb{E}[||\nabla f(x_t, \xi_{t+1})||^2 | x_t] - 2\nabla f(\hat{x}_t)^\top \nabla f(x_t) \tag{25}$$

In (23)-(24) I use the property (10), that $\mathbb{E}[\nabla f(x_t, \xi_{t+1})|x_t] = \nabla f(x_t)$.

From (10), it holds that $\mathbb{E}[\nabla f(x_t, \xi_{t+1})|x_t] = \nabla f(x_t)$ and $\mathbb{E}[||\nabla f(x_t, \xi_{t+1}) - \nabla f(x_t)||^2|x_t] \leq \sigma^2$. First we expand $\mathbb{E}[||\nabla f(x_t, \xi_{t+1}) - \nabla f(x_t)||^2|x_t]$ and by using these properties we obtain (we are taking expectation with respect to ξ_{t+1}):

$$\mathbb{E}[||\nabla f(x_t, \xi_{t+1}) - \nabla f(x_t)||^2|x_t] = \mathbb{E}[||\nabla f(x_t, \xi_{t+1})||^2|x_t] + ||\nabla f(x_t)||^2 - 2\nabla f(x_t)^\top \mathbb{E}[\nabla f(x_t, \xi_{t+1})|x_t]$$
$$= \mathbb{E}[||\nabla f(x_t, \xi_{t+1})||^2|x_t] + ||\nabla f(x_t)||^2 - 2\nabla f(x_t)^\top \nabla f(x_t) = \mathbb{E}[||\nabla f(x_t, \xi_{t+1})||^2|x_t] - ||\nabla f(x_t)||^2$$
$$\leq \sigma^2$$

Therefore, from the last above inequality, it holds that

$$\mathbb{E}[||\nabla f(x_t, \xi_{t+1})||^2|x_t] \leq \sigma^2 + ||\nabla f(x_t)||^2$$

We use this result in (25), to get:

$$\mathbb{E}[||\nabla f(x_t, \xi_{t+1}) - \nabla f(\hat{x}_t)||^2|x_t] \leq ||\nabla f(\hat{x}_t)||^2 + \sigma^2 + ||\nabla f(x_t)||^2 - 2\nabla f(\hat{x}_t)^\top \nabla f(x_t)$$
$$= ||\nabla f(\hat{x}_t) - \nabla f(x_t)||^2 + \sigma^2$$

Finally, we get:

$$\mathbb{E}[||\nabla f(x_t, \xi_{t+1}) - \nabla f(\hat{x}_t)||^2|x_t] \leq ||\nabla f(\hat{x}_t) - \nabla f(x_t)||^2 + \sigma^2$$

By using this result, and the fact that $\mathbb{E}[\nabla f(x_t, \xi_{t+1})|x_t] = \nabla f(x_t)$ we take the expectation with respect to ξ_{t+1} of the both sides of the inequality in (21):

$$\mathbb{E}[||x_{t+1} - \hat{x}_t||^2|x_t] \leq (1 - \eta\rho)^2 \mathbb{E}[||x_t - \hat{x}_t||^2|x_t] + \eta^2 \mathbb{E}[||\nabla f(\hat{x}_t) - \nabla f(x_t, \xi_{t+1})||^2|x_t] +$$
$$+ 2\eta(1 - \eta\rho)\mathbb{E}[\langle x_t - \hat{x}_t, \nabla f(\hat{x}_t) - \nabla f(x_t, \xi_{t+1})\rangle|x_t]$$
$$\leq (1 - \eta\rho)^2 ||x_t - \hat{x}_t||^2 + \eta^2(||\nabla f(\hat{x}_t) - \nabla f(x_t)||^2 + \sigma^2) +$$
$$+ 2\eta(1 - \eta\rho)\langle x_t - \hat{x}_t, \nabla f(\hat{x}_t) - \nabla f(x_t)\rangle$$

Since f has L-Lipshitz continuous gradient, it holds that $||\nabla f(\hat{x}_t) - \nabla f(x_t)||^2 \leq L^2||\hat{x}_t - x_t||^2$. Moreover, from Cauchy-Schwarz inequality, it holds that $\langle x_t - \hat{x}_t, \nabla f(\hat{x}_t) - \nabla f(x_t)\rangle \leq ||x_t - \hat{x}_t|| \cdot ||\nabla f(\hat{x}_t) - \nabla f(x_t)||$. Using the fact that f has L-Lipschitz continuous gradient, it holds that $\langle x_t - \hat{x}_t, \nabla f(\hat{x}_t) - \nabla f(x_t)\rangle \leq ||x_t - \hat{x}_t|| \cdot ||\nabla f(\hat{x}_t) - \nabla f(x_t)|| \leq L||x_t - \hat{x}_t||^2$. We use these results in the above inequality and obtain:

$$\mathbb{E}[||x_{t+1} - \hat{x}_t||^2|x_t] \leq [(1 - \eta\rho)^2 + \eta^2 L^2 + 2\eta(1 - \eta\rho)L] \cdot ||x_t - \hat{x}_t||^2 + \sigma^2\eta^2 \quad (26)$$

If we manage to show that $[(1 - \eta\rho)^2 + \eta^2 L^2 + 2\eta(1 - \eta\rho)L] \leq (1 - \eta\rho)$, we would be done, because in that case from (26) it is true that:

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

Well, indeed $[(1-\eta\rho)^2+\eta^2L^2+2\eta(1-\eta\rho)L] \leq (1-\eta\rho)$ holds. We plug in $\rho = 4L$ and get $[(1-\eta\rho)^2+\eta^2L^2+2\eta(1-\eta\rho)L] = 1-6\eta L+9L^2\eta^2$. Now, it is left to show that $1-6\eta L+9L^2\eta^2 \leq (1-\eta\rho) = (1-4L\eta)$. This holds iff $9L^2\eta^2 \leq 2\eta L$ which is naturally true since $\eta \leq \frac{2}{9L}$, given in the problem statement. Thus we proved the desired property:

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

(c) Let $\rho = 4L$, $\eta \leq \frac{2}{9L}$, and x^{t+1} is given by (11). Prove that for all $t \geq 0$:

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

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

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

Solution 3c):

Using (12), $\Phi_\rho(x_{t+1})$ and $\Phi_\rho(x_t)$ are defined as below:

i) $\Phi_\rho(x_{t+1}) = \min_y\{\Phi(y)+\frac{\rho}{2}||y-x_{t+1}||^2\}$

ii) $\Phi_\rho(x_t) = \min_y\{\Phi(y)+\frac{\rho}{2}||y-x_t||^2\}$

Moreover since $\hat{x}_t = argmin_y\{\Phi(y)+\frac{\rho}{2}||y-x_t||^2\}$, the following holds:

$$\Phi_\rho(x_t) = \Phi(\hat{x}_t)+\frac{\rho}{2}||\hat{x}_t-x_t||^2 \tag{27}$$

Since $\Phi_\rho(x_{t+1})$ is the minimal value of $\Phi(y)+\frac{\rho}{2}||y-x_{t+1}||^2$ over y, then it holds that:

$$\Phi_\rho(x_{t+1}) \leq \Phi(\hat{x}_t)+\frac{\rho}{2}||\hat{x}_t-x_{t+1}||^2 \tag{28}$$

In order to express the right hand side of inequality in (28) in terms of $\Phi_\rho(x_t)$, we need to elaborate more the term $||\hat{x}_t-x_{t+1}||^2$ as below:

$$||\hat{x}_t-x_{t+1}||^2 = ||(\hat{x}_t-x_t)+(x_t-x_{t+1})||^2 = ||\hat{x}_t-x_t||^2+||x_t-x_{t+1}||^2+2\langle\hat{x}_t-x_t, x_t-x_{t+1}\rangle$$

We plug this result in (28) and get the following:

$$\Phi_\rho(x_{t+1}) \leq (\Phi(\hat{x}_t)+\frac{\rho}{2}||\hat{x}_t-x_t||^2)+\frac{\rho}{2}||x_t-x_{t+1}||^2+\rho\langle\hat{x}_t-x_t, x_t-x_{t+1}\rangle \tag{29}$$

Using (27) in (29), one can conclude that:

$$\Phi_\rho(x_{t+1}) \leq \Phi_\rho(x_t)+\frac{\rho}{2}||x_t-x_{t+1}||^2+\rho\langle\hat{x}_t-x_t, x_t-x_{t+1}\rangle \tag{30}$$

Moreover, we can elaborate more on the terms $||x_t - x_{t+1}||^2$ and $\langle \hat{x}_t - x_t, x_t - x_{t+1} \rangle$ as below:

$$||x_t - x_{t+1}||^2 = \langle x_t - x_{t+1}, x_t - x_{t+1} \rangle = \langle (x_t - \hat{x}_t) + (\hat{x}_t - x_{t+1}), (x_t - \hat{x}_t) + (\hat{x}_t - x_{t+1}) \rangle$$
$$= ||x_t - \hat{x}_t||^2 + ||\hat{x}_t - x_{t+1}||^2 + 2\langle x_t - \hat{x}_t, \hat{x}_t - x_{t+1} \rangle$$

And:

$$\langle \hat{x}_t - x_t, x_t - x_{t+1} \rangle = \langle \hat{x}_t - x_t, (x_t - \hat{x}_t) + (\hat{x}_t - x_{t+1}) \rangle = -||\hat{x}_t - x_t||^2 + \langle \hat{x}_t - x_t, \hat{x}_t - x_{t+1} \rangle$$

We now combine these two results and use them in (30) and we get the following:

$$\Phi_\rho(x_{t+1}) \leq \Phi_\rho(x_t) + \frac{\rho}{2}||x_t - x_{t+1}||^2 + \rho\langle \hat{x}_t - x_t, x_t - x_{t+1} \rangle \tag{31}$$
$$= \Phi_\rho(x_t) + \frac{\rho}{2}\big(||x_t - \hat{x}_t||^2 + ||\hat{x}_t - x_{t+1}||^2 + 2\langle x_t - \hat{x}_t, \hat{x}_t - x_{t+1} \rangle\big) + \tag{32}$$
$$+ \rho(-||\hat{x}_t - x_t||^2 + \langle \hat{x}_t - x_t, \hat{x}_t - x_{t+1} \rangle) \tag{33}$$
$$= \Phi_\rho(x_t) - \frac{\rho}{2}||x_t - \hat{x}_t||^2 + \frac{\rho}{2}||\hat{x}_t - x_{t+1}||^2 \tag{34}$$

Finally, we proved that:

$$\Phi_\rho(x_{t+1}) \leq \Phi_\rho(x_t) - \frac{\rho}{2}||x_t - \hat{x}_t||^2 + \frac{\rho}{2}||\hat{x}_t - x_{t+1}||^2$$

Now, we take expectation w.r.t to the whole randomness involved in our setting $(\xi_1, ..., \xi_{t+1})$ on both sides of the above inequality to get the following important result:

$$\mathbb{E}[\Phi_\rho(x_{t+1})] \leq \mathbb{E}[\Phi_\rho(x_t)] - \mathbb{E}[\frac{\rho}{2}||x_t - \hat{x}_t||^2] + \mathbb{E}[\frac{\rho}{2}||\hat{x}_t - x_{t+1}||^2] \tag{35}$$

Since we are under the same setting as 3b), we can use the inequality proven in 3b) $\mathbb{E}[||x_{t+1} - \hat{x}_t||^2 | x_t] \leq (1 - \eta\rho)||x_t - \hat{x}_t|| + \sigma^2\eta^2$. Moreover, we take the expectations with respect to the whole randomness $(\xi_1, .., \xi_t)$ of both sides of this proven inequality and get $\mathbb{E}[||x_{t+1} - \hat{x}_t||^2] \leq (1 - \eta\rho)\mathbb{E}[||x_t - \hat{x}_t||^2] + \sigma^2\eta^2$. We use this result to obtain the following inequality:

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

We use this inequality in (35), to finally get:

$$\mathbb{E}[\Phi_\rho(x_{t+1})] \leq \mathbb{E}[\Phi_\rho(x_t)] - \mathbb{E}[\frac{\rho}{2}||x_t - \hat{x}_t||^2] + \mathbb{E}[\frac{\rho}{2}||\hat{x}_t - x_{t+1}||^2] \tag{36}$$
$$\leq \mathbb{E}[\Phi_\rho(x_t)] - \frac{\rho}{2}\mathbb{E}[||x_t - \hat{x}_t||^2] - \frac{\eta\rho^2}{2}\mathbb{E}[||x_t - \hat{x}_t||^2] + \frac{\rho}{2}\mathbb{E}[||\hat{x}_t - x_t||^2] + \frac{\rho}{2}\sigma^2\eta^2 \tag{37}$$
$$= \mathbb{E}[\Phi_\rho(x_t)] - \frac{\eta}{2}\mathbb{E}[||\rho(x_t - \hat{x}_t)||^2] + \frac{\rho\sigma^2\eta^2}{2} \tag{38}$$

Thus, we just showed the desired property; mainly that the following holds:

$$\mathbb{E}[\Phi_\rho(x_{t+1})] \leq \mathbb{E}[\Phi_\rho(x_t)] - \frac{\eta}{2}\mathbb{E}[||\rho(x_t - \hat{x}_t)||^2] + \frac{\rho\sigma^2\eta^2}{2} \tag{39}$$

Now we need to prove the second part of this problem, i.e., that for τ, chosen uniformly at random from the set $\{0, 1, ..., T-1\}$, it holds that:

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

Using (39), the following inequality holds:

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

Now, given that τ is chosen uniformly at random from the set $\{0, 1, 2, ..., T-1\}$, using the previous result and substituting $\rho = 4L$ it then holds:

$$\mathbb{E}[||\rho(x_\tau - \hat{x}_\tau)||^2|\tau = t] \leq \frac{2}{\eta}[\mathbb{E}[\Phi_\rho(x_\tau)|\tau = t] - \mathbb{E}[\Phi_\rho(x_{\tau+1})|\tau = t]] + 4L\eta\sigma^2 \ (***)$$

From any probability course, we know that $\mathbb{E}[X] = \sum_i \mathbb{E}[X|A_i]P(A_i)$. Therefore, in our setting $\mathbb{E}[||\rho(x_\tau - \hat{x}_\tau)||^2] = \sum_{t=0}^{T-1}\mathbb{E}[||\rho(x_t - \hat{x}_t)||^2|\tau = t] \cdot P(\tau = t)$. Since τ is a uniform random variable from the set $\{0, 1, ..., T-1\}$, $P(\tau = t) = \frac{1}{T}$
$\forall t \in \{0, 1, ..., T-1\}$. Thus, the equality holds:
$\mathbb{E}[||\rho(x_\tau - \hat{x}_\tau)||^2] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[||\rho(x_\tau - \hat{x}_\tau)||^2|\tau = t]$. Using this result and the inequality $(***)$ we have:

$$\mathbb{E}[||\rho(x_\tau - \hat{x}_\tau)||^2] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[||\rho(x_\tau - \hat{x}_\tau)||^2|\tau = t] \leq$$

$$\leq \frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{2}{\eta}[\mathbb{E}[\Phi_\rho(x_\tau)|\tau = t] - \mathbb{E}[\Phi_\rho(x_{\tau+1})|\tau = t]] + 4L\eta\sigma^2\right)$$

$$= \frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{\eta}[\mathbb{E}[\Phi_\rho(x_t)] - \mathbb{E}[\Phi_\rho(x_{t+1})]] + 4L\eta\sigma^2 = \frac{2}{\eta T}\sum_{\tau=0}^{T-1}(\mathbb{E}[\Phi_\rho(x_t)] - \mathbb{E}[\Phi_\rho(x_{t+1})]) + 4L\eta\sigma^2 =$$

$$= \frac{2}{\eta T}(\Phi_{4L}(x_0) - \mathbb{E}[\Phi_{4L}(x_T)]) + 4L\eta\sigma^2 \leq \frac{2(\Phi_{4L}(x_0) - inf_x\Phi_{4L}(x))}{\eta T} + 4L\eta\sigma^2$$

where in the last inequality, I use the fact that
$\Phi_{4L}(x_T) \geq inf_x\Phi_{4L}(x) \leftrightarrow -\Phi_{4L}(x_T) \leq -inf_x\Phi_{4L}(x) \rightarrow -\mathbb{E}[\Phi_{4L}(x_T)] \leq -inf_x\Phi_{4L}(x)$.
Thus we just showed what we wanted to prove.

Exercise 4: Mirror Descent

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 the ℓ_2-norm). For any two $\mathbf{x}, \mathbf{y} \in \Omega$. We restate the Bregman divergence as seen in the lecture below:

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

Prove the following for Algorithm 1:

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

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

Solution 4a):

From algorithm 1, $\mathbf{z}_{t+1} = argmin_{\mathbf{z}\in\Omega}\{V_\omega(\mathbf{z}, \mathbf{z}_t) + \gamma_{t+1}\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{z} - \mathbf{z}_t\rangle\}$. Let us define the function $g(\mathbf{z}) = V_\omega(\mathbf{z}, \mathbf{z}_t) + \gamma_{t+1}\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{z} - \mathbf{z}_t\rangle$ and by substituting $V_\omega(\mathbf{z}, \mathbf{z}_t) = \omega(\mathbf{z}) - \omega(\mathbf{z}_t) - \nabla\omega(\mathbf{z}_t)^\top(\mathbf{z} - \mathbf{z}_t)$ we get $g(\mathbf{z}) = \omega(\mathbf{z}) - \omega(\mathbf{z}_t) - \nabla\omega(\mathbf{z}_t)^\top(\mathbf{z} - \mathbf{z}_t) + \gamma_{t+1}\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{z} - \mathbf{z}_t\rangle$. Moreover, $\nabla g(\mathbf{z}_{t+1}) = \nabla\omega(\mathbf{z}_{t+1}) - \nabla\omega(\mathbf{z}_t) + \gamma_{t+1}\nabla f(\mathbf{x}_{t+1})$. Since $\mathbf{z}_{t+1}$ is the minimizer of $g(\mathbf{z})$, by the optimality condition, it holds that for any $\mathbf{u} \in \Omega$:

$$\langle\nabla\omega(\mathbf{z}_{t+1}) - \nabla\omega(\mathbf{z}_t) + \gamma_{t+1}\nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_{t+1}\rangle = \quad (40)$$

$$= \langle\nabla\omega(\mathbf{z}_{t+1}) - \nabla\omega(\mathbf{z}_t), \mathbf{u} - \mathbf{z}_{t+1}\rangle + \gamma_{t+1}\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_{t+1}\rangle \geq 0 \quad (41)$$

We now use the three-point identity, lemma 10.1, handout 10, page 11, which states that:

$$V_\omega(\mathbf{x}, \mathbf{z}) - V_\omega(\mathbf{x}, \mathbf{y}) - V_\omega(\mathbf{y}, \mathbf{z}) = \langle\nabla\omega(\mathbf{y}) - \nabla\omega(\mathbf{z}), \mathbf{x} - \mathbf{y}\rangle$$

This implies that $\langle\nabla\omega(\mathbf{z}_{t+1}) - \nabla\omega(\mathbf{z}_t), \mathbf{u} - \mathbf{z}_{t+1}\rangle = V_\omega(\mathbf{u}, \mathbf{z}_t) - V_\omega(\mathbf{u}, \mathbf{z}_{t+1}) - V_\omega(\mathbf{z}_{t+1}, \mathbf{z}_t)$
Thus using this property and inequality (41), it holds that:

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

$$= V_\omega(\mathbf{u}, \mathbf{z}_t) - V_\omega(\mathbf{u}, \mathbf{z}_{t+1}) - V_\omega(\mathbf{z}_{t+1}, \mathbf{z}_t) + \gamma_{t+1}\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_{t+1}\rangle \geq 0$$

Since $\mathbf{u} - \mathbf{z}_{t+1} = (\mathbf{u} - \mathbf{z}_t) + (\mathbf{z}_t - \mathbf{z}_{t+1})$ the above inequality becomes:

$$V_\omega(\mathbf{u}, \mathbf{z}_t) - V_\omega(\mathbf{u}, \mathbf{z}_{t+1}) - V_\omega(\mathbf{z}_{t+1}, \mathbf{z}_t) + \gamma_{t+1}\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_t\rangle + \gamma_{t+1}\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1}\rangle \geq 0$$

Since $V_\omega(m, n) \geq 0, \ \forall m, n$ (Handout 10, page 11), in particular $V_\omega(\mathbf{z}_{t+1}, \mathbf{z}_t) \geq 0$ implying that $-V_\omega(\mathbf{z}_{t+1}, \mathbf{z}_t) \leq 0$. Applying this result in the above inequality we get:

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

$$V_\omega(\mathbf{u}, \mathbf{z}_t) - V_\omega(\mathbf{u}, \mathbf{z}_{t+1}) - V_\omega(\mathbf{z}_{t+1}, \mathbf{z}_t) + \gamma_{t+1}\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_t\rangle + \gamma_{t+1}\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1}\rangle \geq 0$$

which is equivalent to what we wanted to prove, mainly that:

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

(b) Prove that:

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

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

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

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

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

Hint. You might find the relation $\mathbf{z}_t - \mathbf{z}_{t+1} = \eta_t^{-1}(\mathbf{x}_{t+1} - \mathbf{v}_t)$ 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.

Solution 4b):
From algorithm 1, $\eta_t = \frac{1}{\gamma_{t+1}L}$ which implies that $\eta_t^{-1} = \gamma_{t+1}L$. Using the hint,

$$\mathbf{z}_t - \mathbf{z}_{t+1} = \eta_t^{-1}(\mathbf{x}_{t+1} - \mathbf{v}_t) = \gamma_t L(\mathbf{x}_{t+1} - \mathbf{v}_t)$$

We use this property in the following:

$$\gamma_{t+1}\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1}\rangle = \gamma_t^2 L\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{v}_t\rangle$$

Using this representation and properties of function f, we can obtain the following set of equalities/inequalities:

$$\begin{aligned}
\gamma_{t+1}\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{z}_{t+1}\rangle &= \gamma_{t+1}^2 L\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{v}_t\rangle && (42)\\
&= \gamma_{t+1}^2 L[\langle\nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{y}_{t+1}\rangle + \langle\nabla f(\mathbf{x}_{t+1}), \mathbf{y}_{t+1} - \mathbf{v}_t\rangle] && (43)\\
&\leq \gamma_{t+1}^2 L[f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1}) + \frac{L}{2}||\mathbf{x}_{t+1} - \mathbf{y}_{t+1}||^2] + && (44)\\
&+ \gamma_{t+1}^2 L[\nabla f(\mathbf{x}_{t+1})^\top(\mathbf{y}_{t+1} - \mathbf{v}_t)] && (45)\\
&\leq \gamma_{t+1}^2 L[f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1}) + \frac{L}{2}||\mathbf{x}_{t+1} - \mathbf{y}_{t+1}||^2] + && (46)\\
&+ \gamma_{t+1}^2 L[\frac{L}{2}||\mathbf{v}_t - \mathbf{x}_{t+1}||^2 - \frac{L}{2}||\mathbf{y}_{t+1} - \mathbf{x}_{t+1}||^2] && (47)\\
&= \gamma_{t+1}^2 L[f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1})] + \frac{L}{2}\gamma_{t+1}^2 L||\mathbf{v}_t - \mathbf{x}_{t+1}||^2 && (48)\\
&= \gamma_{t+1}^2 L[f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1})] + \frac{||\mathbf{z}_t - \mathbf{z}_{t+1}||^2}{2} && (49)
\end{aligned}$$

Explanations about the steps (42)-(49):
In the steps (42)-(43) I use the trivial equality $\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{v}_t \rangle = \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{y}_{t+1} \rangle + \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{y}_{t+1} - \mathbf{v}_t \rangle$.
In (44)-(45), I use the fact that the function f is L-smooth, and thus the following holds $f(\mathbf{y}_{t+1}) \leq f(\mathbf{x}_{t+1}) + \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{y}_{t+1} - \mathbf{x}_{t+1} \rangle + \frac{L}{2}||\mathbf{y}_{t+1} - \mathbf{x}_{t+1}||^2$ which implies that the inequality we used is true, mainly that: $\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{y}_{t+1} \rangle \leq f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1}) + \frac{L}{2}||\mathbf{y}_{t+1} - \mathbf{x}_{t+1}||^2$.
In (46)-(47) I use the fact that since $\mathbf{v}_t = \eta_t \mathbf{z}_{t+1} + (1 - \eta_t)\mathbf{y}_t$ and Ω is a convex set, where $\mathbf{z}_{t+1}, \mathbf{y}_t \in \Omega$ then $\mathbf{v}_t \in \Omega$. Therefore since:

$$\mathbf{y}_{t+1} = argmin_{\mathbf{y} \in \Omega} \{ \frac{L}{2}||\mathbf{y} - \mathbf{x}_{t+1}||^2 \} + \langle \nabla f(\mathbf{x}_{t+1}, \mathbf{y} - \mathbf{x}_{t+1}) \rangle$$

in particular the below equivalent inequalities hold:

$$\frac{L}{2}||\mathbf{y}_{t+1} - \mathbf{x}_{t+1}||^2 + \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{y}_{t+1} - \mathbf{x}_{t+1} \rangle \leq \frac{L}{2}||\mathbf{v}_t - \mathbf{x}_{t+1}||^2 + \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{v}_t - \mathbf{x}_{t+1} \rangle$$

$$\leftrightarrow \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{y}_{t+1} - \mathbf{v}_t \rangle \leq \frac{L}{2}||\mathbf{v}_t - \mathbf{x}_{t+1}||^2 - \frac{L}{2}||\mathbf{y}_{t+1} - \mathbf{x}_{t+1}||^2$$

Thus we use this last equation in (46)-(47) to get the expression in (48). Finally in (49), I use the property that $\eta_t(\mathbf{z}_t - \mathbf{z}_{t+1}) = (\mathbf{x}_{t+1} - \mathbf{v}_t) \rightarrow \frac{1}{\gamma_{t+1}^2 L^2}||\mathbf{z}_t - \mathbf{z}_{t+1}||^2 = ||\mathbf{x}_{t+1} - \mathbf{v}_t||^2 \leftrightarrow ||\mathbf{z}_t - \mathbf{z}_{t+1}||^2 = \gamma_{t+1}^2 L^2 ||\mathbf{x}_{t+1} - \mathbf{v}_t||^2$ given that $\eta_t = \frac{1}{\gamma_{t+1} L}$. Finally using the result in (49), we obtain the following desired inequality:

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

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

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

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 $\mathbf{x}_{t+1}$

Solution 4c):

From 4b), it is given that the following holds $\forall \mathbf{u} \in \Omega$:

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

In other words, it holds that:

$$\gamma_{t+1}^2 L f(\mathbf{y}_{t+1}) + V_\omega(\mathbf{u}.\mathbf{z}_{t+1}) - V_\omega(\mathbf{u}, \mathbf{z}_t) \leq \gamma_{t+1}^2 L f(\mathbf{x}_{t+1}) + \gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_t \rangle \tag{50}$$

Now, we need to show the following:

$$\gamma_{t+1}^2 Lf(\mathbf{y}_{t+1}) - \gamma_{t+1}^2 Lf(\mathbf{y}_t) + \gamma_{t+1} f(\mathbf{y}_t) + V_\omega(\mathbf{u}, \mathbf{z}_{t+1}) - V_\omega(\mathbf{u}, \mathbf{z}_t) \leq \gamma_{t+1} f(\mathbf{u}) \quad (51)$$

In order to prove (51), it is sufficient to prove that the following holds, mainly that:

$$-\gamma_{t+1}^2 Lf(\mathbf{y}_t) + \gamma_{t+1} f(\mathbf{y}_t) \leq \gamma_{t+1} f(\mathbf{u}) - \gamma_{t+1}^2 Lf(\mathbf{x}_{t+1}) - \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_t \rangle \quad (52)$$

Proving (52), immediately proves (51), since we could then just sum up each of the sides of inequality (50) and (52) and then the resulting holding inequality will be (51), what we want to show. On the other hand, proving (52) is equivalent to proving the following:

$$-\gamma_{t+1}^2 Lf(\mathbf{y}_t) + \gamma_{t+1}^2 Lf(\mathbf{x}_{t+1}) + \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_t \rangle \leq \gamma_{t+1} [f(\mathbf{u}) - f(\mathbf{y}_t)] \quad (53)$$

We show this by the following series of inequalities/equalities:

$$-\gamma_{t+1}^2 Lf(\mathbf{y}_t) + \gamma_{t+1}^2 Lf(\mathbf{x}_{t+1}) + \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_t \rangle = \quad (54)$$

$$= \gamma_{t+1}^2 L(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_t)) + \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{x}_{t+1} \rangle + \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{z}_t \rangle \quad (55)$$

$$\leq \gamma_{t+1}^2 L(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_t)) + \gamma_{t+1} (f(\mathbf{u}) - f(\mathbf{x}_{t+1})) + \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{z}_t \rangle \quad (56)$$

Explanations about the steps (54)-(56):
In (54)-(55), I use the trivial equality $\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_t \rangle = \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{x}_{t+1} \rangle + \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{z}_t \rangle$.
In (56), I use the property that f is convex, and therefore it holds that $\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{x}_{t+1} \rangle \leq f(\mathbf{u}) - f(\mathbf{x}_{t+1})$
Guided by the hint, since $\eta_t(\mathbf{x}_{t+1} - \mathbf{z}_t) = (1 - \eta_t)(\mathbf{y}_t - \mathbf{x}_{t+1})$, it holds that $\mathbf{x}_{t+1} - \mathbf{z}_t = \frac{1-\eta_t}{\eta_t}(\mathbf{y}_t - \mathbf{x}_{t+1})$. The last expression in (56) becomes:

$$\gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{z}_t \rangle = \gamma_{t+1} \langle \nabla f(\mathbf{x}_{t+1}), \frac{1 - \eta_t}{\eta_t}(\mathbf{y}_t - \mathbf{x}_{t+1}) \rangle = \gamma_{t+1} \frac{1 - \eta_t}{\eta_t} \langle \nabla f(\mathbf{x}_{t+1}), \mathbf{y}_t - \mathbf{x}_{t+1} \rangle$$

We substitute this result in (56), and get the following series of inequalities/equalities:

$$-\gamma_{t+1}^2 L f(\mathbf{y}_t) + \gamma_{t+1}^2 L f(\mathbf{x}_{t+1}) + \gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_t\rangle \tag{57}$$

$$\leq \gamma_{t+1}^2 L(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_t)) + \gamma_{t+1}(f(\mathbf{u}) - f(\mathbf{x}_{t+1})) + \gamma_{t+1}\frac{1-\eta_t}{\eta_t}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{y}_t - \mathbf{x}_{t+1}\rangle \tag{58}$$

$$\leq \gamma_{t+1}^2 L(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_t)) + \gamma_{t+1}(f(\mathbf{u}) - f(\mathbf{x}_{t+1})) + \gamma_{t+1}\frac{1-\eta_t}{\eta_t}[f(\mathbf{y}_t) - f(\mathbf{x}_{t+1})] \tag{59}$$

$$= \frac{\gamma_{t+1}}{\eta_t}(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_t)) + \gamma_{t+1}[f(\mathbf{u}) - f(\mathbf{x}_{t+1})] + \frac{\gamma_{t+1}}{\eta_t} f(\mathbf{y}_t) - \gamma_{t+1} f(\mathbf{y}_t) - \tag{60}$$

$$-\frac{\gamma_{t+1}}{\eta_t} f(\mathbf{x}_{t+1}) + \gamma_{t+1} f(\mathbf{x}_{t+1}) \tag{61}$$

$$= \gamma_{t+1}(f(\mathbf{u}) - f(\mathbf{y}_t)) \tag{62}$$

Explanations about the steps (57)-(62):
In (57)-(58)-(59), I make use of the fact that f is convex, and therefore the inequality $\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{y}_t - \mathbf{x}_{t+1}\rangle \leq f(\mathbf{y}_t) - f(\mathbf{x}_{t+1})$ holds. In (60), I use the fact that $\gamma_{t+1}L = \frac{1}{\eta_t}$. In steps (61)-(62), I only do calculations. Thus, this concludes the proof of the desired inequality:

$$-\gamma_{t+1}^2 L f(\mathbf{y}_t) + \gamma_{t+1}^2 L f(\mathbf{x}_{t+1}) + \gamma_{t+1}\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{u} - \mathbf{z}_t\rangle \leq \gamma_{t+1}(f(\mathbf{u}) - f(\mathbf{y}_t))$$

which is what we wanted to show (53).

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

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

Solution 4d):

From 4c) the following inequality holds:

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

The above inequality is equivalent to the following:

$$\gamma_{t+1}^2 L(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)) + \gamma_{t+1}(f(\mathbf{y}_t) - f(\mathbf{u})) \leq V_\omega(\mathbf{u}, \mathbf{z}_t) - V_\omega(\mathbf{u}, \mathbf{z}_{t+1}) \tag{63}$$

Since (63) holds for every $\mathbf{u} \in \Omega$, take $\mathbf{u} = \mathbf{x}^*$ and get the following inequality:

$$\gamma_{t+1}^2 L(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)) + \gamma_{t+1}(f(\mathbf{y}_t) - f(\mathbf{x}^*)) \leq V_\omega(\mathbf{x}^*, \mathbf{z}_t) - V_\omega(\mathbf{x}^*, \mathbf{z}_{t+1})$$

Let $t = 0, 1, ..., T-1$ and sum over both sides of the above inequality to get the following:

$$\sum_{t=0}^{T-1} \gamma_{t+1}^2 L(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)) + \sum_{t=0}^{T-1} \gamma_{t+1}(f(\mathbf{y}_t) - f(\mathbf{x}^*)) \leq \sum_{t=0}^{T-1} (V_\omega(\mathbf{x}^*, \mathbf{z}_t) - V_\omega(\mathbf{x}^*, \mathbf{z}_{t+1})) \tag{64}$$

$$= V_\omega(\mathbf{x}^*, \mathbf{z}_0) - V_\omega(\mathbf{x}^*, \mathbf{z}_T) \leq V_\omega(\mathbf{x}^*, \mathbf{z}_0) \leq R \tag{65}$$

The inequality $V_\omega(\mathbf{x}^*, \mathbf{z}_0) - V_\omega(\mathbf{x}^*, \mathbf{z}_T) \leq V_\omega(\mathbf{x}^*, \mathbf{z}_0)$ holds since $V_\omega(\mathbf{x}^*, \mathbf{z}_T) \geq 0 \leftrightarrow -V_\omega(\mathbf{x}^*, \mathbf{z}_T) \leq 0$ holds. Moreover, $V_\omega(\mathbf{x}^*, \mathbf{z}_0) \leq R$ since it is given that for any choice of starting point $\mathbf{x}_0 \in \Omega$, we have $V_\omega(\mathbf{x}^*, \mathbf{x}_0) \leq R$, thus it holds also for $\mathbf{x}_0 = \mathbf{z}_0$. Now on, we work with the left-hand side of the above inequality so that we could obtain our desired inequality. Let $\gamma_{t+1} = \frac{t+2}{2L}$ as defined by the algorithm

1 and get:

$$\sum_{t=0}^{T-1} \gamma_{t+1}^2 L(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)) + \sum_{t=0}^{T-1} \gamma_{t+1}(f(\mathbf{y}_t) - f(\mathbf{x}^*)) \tag{66}$$

$$= \sum_{t=0}^{T-1} \frac{(t+2)^2}{4L}(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)) + \sum_{t=0}^{T-1} (\frac{t+2}{2L})(f(\mathbf{y}_t) - f(\mathbf{x}^*)) \tag{67}$$

$$= \sum_{t=0}^{T-1} \frac{(t+2)^2}{4L} f(\mathbf{y}_{t+1}) - \sum_{t=0}^{T-1} \frac{(t+2)^2}{4L} f(\mathbf{y}_t) + \sum_{t=0}^{T-1} \frac{(t+2)}{2L} f(\mathbf{y}_t) - \sum_{t=0}^{T-1} \frac{(t+2)}{2L} f(\mathbf{x}^*) \tag{68}$$

$$= \sum_{t=1}^{T} \frac{(t+1)^2}{4L} f(\mathbf{y}_t) - \sum_{t=0}^{T-1} \frac{(t+2)^2}{4L} f(\mathbf{y}_t) + \sum_{t=0}^{T-1} \frac{(t+2)}{2L} f(\mathbf{y}_t) - \sum_{t=0}^{T-1} \frac{(t+2)}{2L} f(\mathbf{x}^*) \tag{69}$$

$$= \sum_{t=1}^{T-1} \frac{(t+1)^2}{4L} f(\mathbf{y}_t) + \frac{(T+1)^2}{4L} f(\mathbf{y}_T) - \sum_{t=1}^{T-1} \frac{(t+2)^2}{4L} f(\mathbf{y}_t) - \frac{f(\mathbf{y}_0)}{L} + \tag{70}$$

$$+ \sum_{t=1}^{T-1} \frac{(t+2)}{2L} f(\mathbf{y}_t) + \frac{f(\mathbf{y}_0)}{L} - \frac{f(\mathbf{x}^*)}{2L} \sum_{t=0}^{T-1} (t+2) \tag{71}$$

$$= \sum_{t=1}^{T-1} \left(\frac{(t+1)^2}{4L} - \frac{(t+2)^2}{4L} + \frac{t+2}{2L} \right) f(\mathbf{y}_t) + \frac{(T+1)^2}{4L} f(\mathbf{y}_T) - \frac{f(\mathbf{x}^*)}{2L} \sum_{t=0}^{T-1} (t+2) \tag{72}$$

$$= \sum_{t=1}^{T-1} \frac{1}{4L} f(\mathbf{y}_t) + \frac{(T+1)^2}{4L} f(\mathbf{y}_T) - \frac{f(\mathbf{x}^*)}{2L} \sum_{t=0}^{T-1} (t+2) \tag{73}$$

$$= \sum_{t=1}^{T-1} \frac{1}{4L} (f(\mathbf{y}_t) - f(\mathbf{x}^*)) + \sum_{t=1}^{T-1} \frac{1}{4L} f(\mathbf{x}^*) + \frac{(T+1)^2}{4L} f(\mathbf{y}_T) - \frac{f(\mathbf{x}^*)}{2L} \sum_{t=0}^{T-1} (t+2) \tag{74}$$

Let us now estimate

$$\sum_{t=1}^{T-1} \frac{1}{4L} f(\mathbf{x}^*) - \frac{f(\mathbf{x}^*)}{2L} \sum_{t=0}^{T-1} (t+2)$$

to plug the resulting expression in (74).

$$\sum_{t=1}^{T-1} \frac{1}{4L} f(\mathbf{x}^*) - \frac{f(\mathbf{x}^*)}{2L} \sum_{t=0}^{T-1} (t+2) = \frac{(T-1) f(\mathbf{x}^*)}{4L} - \left[\frac{f(\mathbf{x}^*)}{2L} \left(\sum_{t=0}^{T-1} t + \sum_{t=0}^{T-1} 2 \right) \right]$$

$$= \frac{(T-1)f(\mathbf{x}^*)}{4L} - \Big[\frac{f(\mathbf{x}^*)}{2L}\big(\sum_{t=1}^{T} t - \sum_{t=1}^{T} 1 + \sum_{t=0}^{T-1} 2\big)\Big] = \frac{(T-1)f(\mathbf{x}^*)}{4L} - \Big[\frac{f(\mathbf{x}^*)}{2L}\big(\frac{T(T+1)}{2} + T\big)\Big]$$

$$= \frac{(T-1)f(\mathbf{x}^*)}{4L} - \frac{f(\mathbf{x}^*)(T^2+3T)}{4L} = \frac{f(\mathbf{x}^*)}{4L}[(T-1) - (T^2+3T)]$$

$$= -\frac{f(\mathbf{x}^*)}{4L}(T+1)^2$$

Therefore, using this result in the equality (74), we get:

$$\sum_{t=0}^{T-1} \gamma_{t+1}^2 L(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)) + \sum_{t=0}^{T-1} \gamma_{t+1}(f(\mathbf{y}_t) - f(\mathbf{x}^*)) \tag{75}$$

$$= \sum_{t=1}^{T-1} \frac{1}{4L}(f(\mathbf{y}_t) - f(\mathbf{x}^*)) + \sum_{t=1}^{T-1} \frac{1}{4L} f(\mathbf{x}^*) + \frac{(T+1)^2}{4L} f(\mathbf{y}_T) - \frac{f(\mathbf{x}^*)}{2L} \sum_{t=0}^{T-1} (t+2) \tag{76}$$

$$= \sum_{t=1}^{T-1} \frac{1}{4L}(f(\mathbf{y}_t) - f(\mathbf{x}^*)) + \frac{(T+1)^2}{4L} f(\mathbf{y}_T) - \frac{(T+1)^2}{4L} f(\mathbf{x}^*) \tag{77}$$

$$= \sum_{t=1}^{T-1} \frac{1}{4L}(f(\mathbf{y}_t) - f(\mathbf{x}^*)) + \frac{(T+1)^2}{4L}\big[f(\mathbf{y}_T) - f(\mathbf{x}^*)\big] \tag{78}$$

Moreover using this result as the left hand side of inequality (65), we get:

$$\sum_{t=1}^{T-1} \frac{1}{4L}(f(\mathbf{y}_t) - f(\mathbf{x}^*)) + \frac{(T+1)^2}{4L}\big[f(\mathbf{y}_T) - f(\mathbf{x}^*)\big] \leq R \tag{79}$$

Since $\mathbf{x}^*$ is a minimizer of f, then $f(\mathbf{y}_t) - f(\mathbf{x}^*) \geq 0 \ \forall t \in \{1, .., T-1\}$, therefore $\sum_{t=1}^{T-1} \frac{1}{4L}(f(\mathbf{y}_t) - f(\mathbf{x}^*)) \geq 0$. We use this inequality in (79) to finally get:

$$\frac{(T+1)^2}{4L}\big[f(\mathbf{y}_T) - f(\mathbf{x}^*)\big] \leq \sum_{t=1}^{T-1} \frac{1}{4L}(f(\mathbf{y}_t) - f(\mathbf{x}^*)) + \frac{(T+1)^2}{4L}\big[f(\mathbf{y}_T) - f(\mathbf{x}^*)\big] \leq R$$

This implies:

$$\frac{(T+1)^2}{4L}\big[f(\mathbf{y}_T) - f(\mathbf{x}^*)\big] \leq R \quad \leftrightarrow \quad f(\mathbf{y}_T) - f(\mathbf{x}^*) \leq \frac{4RL}{(T+1)^2}$$

which is what we wanted to show!