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

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

Solution: Define indicator function

$$I_{\mathcal{K}}(y) = \begin{cases} 0, & y \in \mathcal{K}, \\ \infty, & y \notin \mathcal{K}. \end{cases}$$

We need two lemmas.

Lemma 1. *For $y \in \mathcal{K}$, one has*

$$\partial I_{\mathcal{K}}(y) = \left\{ g \in \mathbb{R}^d : \langle g, y - y' \rangle \geqslant 0 \ \forall y' \in \mathcal{K} \right\}.$$

Proof. Fix an arbitrary $y \in \mathcal{K}$. Define $\mathcal{N}_{\mathcal{K}}(y) := \left\{ g \in \mathbb{R}^d : \langle g, y - y' \rangle \geqslant 0 \ \forall y' \in \mathcal{K} \right\}$.

"$\subseteq$": Let $g \in \partial I_{\mathcal{K}}(y)$. By definition of subgradients, for any $y' \in \mathcal{K}$, one has

$$I_{\mathcal{K}}(y') \geqslant I_{\mathcal{K}}(y) + \langle g, y' - y \rangle .$$

This implies $\langle g, y - y' \rangle \geqslant 0$ since $I_{\mathcal{K}}(y) = I_{\mathcal{K}}(y') = 0$. Thus $g \in \mathcal{N}_{\mathcal{K}}(y)$.

"$\supseteq$": Assume $g \in \mathcal{N}_{\mathcal{K}}(y)$. For $y' \in \mathcal{K}$, $\langle g, y - y' \rangle \geqslant 0$ by definition of $\mathcal{N}_{\mathcal{K}}(y)$. Then $I_{\mathcal{K}}(y') \geqslant I_{\mathcal{K}}(y) + \langle g, y' - y \rangle$. For $y' \notin \mathcal{K}$, $I_{\mathcal{K}}(y') = \infty \geqslant I_{\mathcal{K}}(y) + \langle g, y' - y \rangle$. Thus $g \in \partial I_{\mathcal{K}}(y)$. □

Lemma 2. *For any function $h : \mathbb{R}^d \to \mathbb{R}$ (convex or not), x is a minimizer of h over $\mathbb{R}^d$ if and only if $0 \in \partial h(x)$.*

Proof. x is a minimizer of h, if and only if $h(y) \geqslant h(x)$ for any y, if and only if $h(y) \geqslant h(x) + \langle 0, y - x \rangle$ for any y, if and only if $0 \in \partial h(x)$. □

Now we apply the above two lemmas. Note y^* is a minimizer of f over $\mathcal{K}$, if and only if y^* is a minimizer of $f + I_{\mathcal{K}}$ over $\mathbb{R}^d$, if and only if $0 \in \partial(f + I_{\mathcal{K}})(y^*) = \partial f(y^*) + \partial I_{\mathcal{K}}(y^*)$, if and only if there exists $g \in \partial f(y^*)$ such that $\langle g, y - y^* \rangle \geqslant 0$ for any $y \in \mathcal{K}$. □

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

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

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 $\Phi\left(x_t\right) - \min_{x\in\mathcal{X}} \Phi(x) \leqslant \varepsilon$.

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.

Solution: It is easy to see that

$$\Phi_\mu\left(x_k\right) - \Phi_\mu^* \leqslant \left(L + \frac{\alpha}{\mu}\right) \frac{c}{t^2}.$$

Moreover, by the condition

$$\Phi(x) - \beta\mu \leqslant \Phi_\mu(x) \leqslant \Phi(x) + \beta\mu.$$

This implies

$$\Phi^* \geqslant \Phi_\mu^* - \beta\mu \text{ and } \Phi(x_t) \leqslant \Phi_\mu(x_t) + \beta\mu.$$

Combining inequalities above,

$$\Phi(x_t) - \Phi^* \leqslant \Phi_\mu(x_t) - \Phi_\mu^* + 2\beta\mu \leqslant \frac{Lc}{t^2} + \left(\frac{\alpha c}{t^2}\right)\frac{1}{\mu} + 2\beta\mu. \tag{1}$$

Minimizing the right hand side over μ, we obtain

$$\mu = \sqrt{\frac{\alpha c}{2\beta}}\frac{1}{t}. \tag{2}$$

Plugging it back,

$$\Phi(x_t) - \Phi^* \leqslant \frac{Lc}{t^2} + 2\sqrt{2\alpha\beta c}\frac{1}{t},$$

which is what we want for the first part. Now solve t for

$$\frac{Lc}{t^2} + 2\sqrt{2\alpha\beta c}\frac{1}{t} = \varepsilon.$$

Denote $\nu = \sqrt{c}/t$, then we want to solve

$$L\nu^2 + 2\sqrt{2\alpha\beta}\nu - \varepsilon = 0.$$

This leads to

$$\sqrt{c}/t = \nu = \frac{\varepsilon}{\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\varepsilon}}.$$

We can pick μ according to (2). Therefore, with

$$t = \frac{\sqrt{2\alpha\beta c} + \sqrt{2\alpha\beta c + Lc\varepsilon}}{\varepsilon} \text{ and } \mu = \sqrt{\frac{\alpha}{2\beta}}\frac{\varepsilon}{\sqrt{2\alpha\beta} + \sqrt{2\alpha\beta + L\varepsilon}},$$

we have that

$$\Phi(x_t) - \Phi^* \leqslant \frac{Lc}{t^2} + \left(\frac{\alpha c}{t^2}\right)\frac{1}{\mu} + 2\beta\mu \leqslant \varepsilon.$$

We finish the proof by using $\sqrt{A} + \sqrt{A + B} \leqslant 2\sqrt{A} + \sqrt{B}$ and noting that the right hand side of (1) decreases with t. □

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

$$\mathbb{E}\left[\nabla f(x, \xi)\right] = \nabla f(x), \qquad \mathbb{E}\left[\|\nabla f(x, \xi) - \nabla f(x)\|^2\right] \leqslant \sigma^2. \tag{3}$$

Consider the following method (Proximal Stochastic Gradient Descent)

$$x_{t+1} = \mathrm{prox}_{\eta r}(x_t - \eta \nabla f(x_t, \xi_{t+1})), \tag{4}$$

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

$$\Phi_\rho(x) := \min_{y \in \mathbb{R}^d} \left\{ \Phi(y) + \frac{\rho}{2}\|y - x\|^2 \right\}. \tag{5}$$

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

$$\hat{x} := \mathrm{prox}_{\Phi/\rho}(x) := \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ \Phi(y) + \frac{\rho}{2}\|y - x\|^2 \right\}. \tag{6}$$

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 (4). 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 (4). Prove that for all $t \geqslant 0$

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

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

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

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

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

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 $\hat{x} := \text{prox}_{\eta\Phi}(x)$. 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.

Solution:

(a) By definition of $\hat{x}_t$ and $\Phi(x) := f(x) + r(x)$, we have

$$0 \in \partial\left(f + \frac{\rho}{2}\|\cdot - x_t\|^2 + r\right)(\hat{x}_t) = \nabla f(\hat{x}_t) + \rho(\hat{x}_t - x_t) + \partial r(\hat{x}_t),$$

where the last equality holds, since $f(\cdot) + \frac{\rho}{2}\|\cdot - x_t\|^2$, and $r(\cdot)$ are both convex (due to the conic combination rule, slide 16, Handout 9). Multiplying both sides by $\eta > 0$ and rearranging, we get

$$z_t := \eta\rho x_t - \eta\nabla f(\hat{x}_t) + (1 - \eta\rho)\hat{x}_t \in \hat{x}_t + \eta\partial r(\hat{x}_t)$$

Therefore, by the optimality condition for the proximal sub-problem, we have $\hat{x}_t = \text{prox}_{\eta r}(z_t)$.

(b) Using the result of Question 1, the update rule (4), and non-expansiveness of the proximal operator (Exercise 3 (c) in HW 9), we have

$$\begin{aligned}
&\mathbb{E}\left[\|x_{t+1} - \hat{x}_t\|^2 \mid x_t\right] \\
&= \mathbb{E}\left[\left\|\text{prox}_{\eta r}(x_t - \eta\nabla f(x_t, \xi_{t+1})) - \text{prox}_{\eta r}(\eta\rho x_t - \eta\nabla f(\hat{x}_t) + (1 - \eta\rho)\hat{x}_t)\right\|^2 \mid x_t\right] \\
&\leqslant \mathbb{E}\left[\|x_t - \eta\nabla f(x_t, \xi_{t+1}) - (\eta\rho x_t - \eta\nabla f(\hat{x}_t) + (1 - \eta\rho)\hat{x}_t)\|^2 \mid x_t\right] \\
&= \mathbb{E}\left[\|(1 - \eta\rho)(x_t - \hat{x}_t) - \eta(\nabla f(x_t, \xi_{t+1}) - \nabla f(\hat{x}_t))\|^2 \mid x_t\right] \\
&= \mathbb{E}\left[\|(1 - \eta\rho)(x_t - \hat{x}_t) - \eta(\nabla f(x_t) - \nabla f(\hat{x}_t)) - \eta(\nabla f(x_t, \xi_{t+1}) - \nabla f(x_t))\|^2 \mid x_t\right] \\
&\overset{\text{(i)}}{=} \|(1 - \eta\rho)(x_t - \hat{x}_t) - \eta(\nabla f(x_t) - \nabla f(\hat{x}_t))\|^2 + \eta^2\mathbb{E}\left[\|\nabla f(x_t, \xi_{t+1}) - \nabla f(x_t)\|^2 \mid x_t\right] \\
&\overset{\text{(ii)}}{\leqslant} \|(1 - \eta\rho)(x_t - \hat{x}_t) - \eta(\nabla f(x_t) - \nabla f(\hat{x}_t))\|^2 + \eta^2\sigma^2 \\
&= (1 - \eta\rho)^2\|x_t - \hat{x}_t\|^2 - 2(1 - \eta\rho)\eta\langle x_t - \hat{x}_t, \nabla f(x_t) - \nabla f(\hat{x}_t)\rangle + \eta^2\sigma^2 \\
&\quad + \eta^2\|\nabla f(x_t) - \nabla f(\hat{x}_t)\|^2 \\
&\overset{\text{(iii)}}{\leqslant} (1 - \eta\rho)^2\|x_t - \hat{x}_t\|^2 + 2(1 - \eta\rho)\eta L\|x_t - \hat{x}_t\|^2 + \eta^2L^2\|x_t - \hat{x}_t\|^2 + \eta^2\sigma^2 \\
&= (1 - \eta\rho)\left(1 - \eta\rho + 2\eta L + \frac{\eta^2L^2}{1 - \eta\rho}\right)\|x_t - \hat{x}_t\|^2 + \eta^2\sigma^2 \\
&\leqslant (1 - \eta\rho)\|x_t - \hat{x}_t\|^2 + \eta^2\sigma^2,
\end{aligned}$$

where in (i) and (ii) use unbiasedness of the gradient estimator and bounded variance, i.e., (3). In (iii), we use Cauchy–Schwarz inequality and smoothness of $f(\cdot)$, i.e., $\|\nabla f(\hat{x}_t) - \nabla f(x_t)\| \leqslant L \|\hat{x}_t - x_t\|$. The last inequality holds by the choice of ρ, η and $2\eta L \leqslant \frac{\eta\rho}{2}$, and $\frac{\eta^2 L}{1-\eta\rho} \leqslant \frac{\eta\rho}{2}$.

(c) By optimality of $\hat{x}_{t+1}$, we have

$$\begin{aligned}
\mathbb{E}\left[\Phi_\rho(x_{t+1})\right] &= \mathbb{E}\left[\Phi(\hat{x}_{t+1}) + \frac{\rho}{2}\|\hat{x}_{t+1} - x_{t+1}\|^2\right] \\
&\leqslant \mathbb{E}\left[\Phi(\hat{x}_t) + \frac{\rho}{2}\|\hat{x}_t - x_{t+1}\|^2\right] \\
&\overset{(i)}{\leqslant} \mathbb{E}\left[\Phi(\hat{x}_t) + \frac{\rho}{2}\|\hat{x}_t - x_t\|^2\right] - \frac{\rho^2\eta}{2}\mathbb{E}\left[\|x_t - \hat{x}_t\|^2\right] + \frac{\rho\eta^2\sigma^2}{2} \\
&= \mathbb{E}\left[\Phi_\rho(x_t)\right] - \frac{\rho^2\eta}{2}\mathbb{E}\left[\|x_t - \hat{x}_t\|^2\right] + \frac{\rho\eta^2\sigma^2}{2},
\end{aligned}$$

where in (i) we applied the result of Question 2. The first and the last inequalities hold by definition of $\Phi_\rho(\cdot)$.

By telescoping and rearranging, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\rho(x_t - \hat{x}_t)\|^2\right] \leqslant \frac{2(\Phi_{4L}(x_0) - \inf_x \Phi_{4L}(x))}{\eta T} + 4L\eta\sigma^2.$$

It remains to recognize the expectation with respect to randomization of the index τ on the right hand side to conclude the proof.

□

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:

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

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 $u \in \Omega$, show that

$$\gamma_{t+1} \langle \nabla f(x_{t+1}), z_t - u \rangle \leqslant (\gamma_{t+1} \langle \nabla f(x_{t+1}), z_t - z_{t+1} \rangle) + V_\omega(u, z_t) - V_\omega(u, z_{t+1}).$$

(b) Prove that

$$\gamma_{t+1} \langle \nabla f(x_{t+1}), z_t - z_{t+1} \rangle - \frac{1}{2} \|z_t - z_{t+1}\|^2 \leqslant \gamma_{t+1}^2 L \left(f(x_{t+1}) - f(y_{t+1}) \right).$$

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

$$\gamma_{t+1} \langle \nabla f(x_{t+1}), z_t - u \rangle \leqslant (\gamma_{t+1} \langle \nabla f(x_{t+1}), z_t - z_{t+1} \rangle) - \frac{1}{2} \|z_t - z_{t+1}\|^2 + V_\omega(u, z_t) - V_\omega(u, z_{t+1})$$

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

$$\gamma_{t+1} \langle \nabla f(x_{t+1}), z_t - u \rangle \leqslant \gamma_{t+1}^2 L \left(f(x_{t+1}) - f(y_{t+1}) \right) + V_\omega(u, z_t) - V_\omega(u, z_{t+1}).$$

Hint You might find the relation $z_t - z_{t+1} = \eta_t^{-1}(x_{t+1} - v_t)$ useful in which $v_t = \eta_t z_{t+1} + (1-\eta_t) y_t$. You do not need to prove this relation.

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

$$\gamma_{t+1}^2 L f(y_{t+1}) - (\gamma_{t+1}^2 L - \gamma_{t+1}) f(y_t) + V_\omega(u, z_{t+1}) - V_\omega(u, z_t) \leqslant \gamma_{t+1} f(u).$$

Hint You might find the relation

$$\eta_t (x_{t+1} - z_t) = (1-\eta_t)(y_t - x_{t+1}).$$

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

(d) Assume there exists a minimizer $x^* = \operatorname{argmin}_{x\in\Omega} f(x)$ and for any choice of starting point $x_0 \in \Omega$, we have $V_\omega(x^*, x_0) \leqslant R$, with $R \geqslant 0$. Prove that

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

Solution: The first part of the proof is similar to that of Lemma 10.2 seen in the lecture. Let $u \in \Omega$, then we have

$$\begin{aligned}
\gamma_{t+1}\langle \nabla f(x_{t+1}), z_t - u\rangle &= \gamma_{t+1}\langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle + \gamma_{t+1}\langle \nabla f(x_{t+1}), z_{t+1} - u\rangle \\
&\leqslant \gamma_{t+1}\langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle + \langle \nabla \omega(z_{t+1}) - \nabla\omega(z_t), u - z_{t+1}\rangle \\
&= \gamma_{t+1}\langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle + V_\omega(u, z_t) - V_\omega(u, z_{t+1}) - \underbrace{V_\omega(z_{t+1} - z_t)}_{\geqslant \frac{1}{2}\|z_t - z_{t+1}\|^2} \\
&\leqslant (\gamma_{t+1}\langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle) + V_\omega(u, z_t) - V_\omega(u, z_{t+1}),
\end{aligned}$$

where the first inequality holds by the optimality condition, the second equality holds by the three point identity and the last inequality follows from 1-strong convexity of $\omega(\cdot)$.

Now, let $v = \eta_t z_{t+1} + (1-\eta_t)y_t$. Then $x_{t+1} - v = \eta_t(z_t - z_{t+1})$ and we have

$$\begin{aligned}
&\gamma_{t+1}\langle \nabla f(x_{t+1}), z_t - z_{t+1}\rangle - \frac{1}{2}\|z_t - z_{t+1}\|^2 \\
&\quad= \left\langle \frac{\gamma_{t+1}}{\eta_t}\nabla f(x_t), x_{t+1} - v\right\rangle - \frac{1}{2\eta_t^2}\|x_{t+1} - v\|^2 \\
&\quad= \gamma_{t+1}^2 L\left(\langle \nabla f(x_{t+1}), x_{t+1} - v\rangle - \frac{L}{2}\|x_{t+1} - v\|^2\right) && \text{by } \eta_t = \frac{1}{\gamma_{t+1}L} \\
&\quad= \gamma_{t+1}^2 L\left(-\langle \nabla f(x_{t+1}), v - x_{t+1}\rangle - \frac{L}{2}\|x_{t+1} - v\|^2\right) \\
&\quad\leqslant \gamma_{t+1}^2 L\left(-\langle \nabla f(x_{t+1}), y_{t+1} - x_{t+1}\rangle - \frac{L}{2}\|x_{t+1} - y_{t+1}\|^2\right) && \text{by def. of } y_{t+1} \\
&\quad\leqslant \gamma_{t+1}^2 L\,(f(x_{t+1}) - f(y_{t+1})) && \text{by } f \text{ L-smooth}
\end{aligned}$$

Plugging this into our chain of inequalities above, we get

$$\gamma_{t+1}\langle \nabla f(x_{t+1}), z_t - u\rangle < \gamma_{t+1}^2 L\,(f(x_{t+1}) - f(y_{t+1})) + V_\omega(u, z_t) - V_\omega(u, z_{t+1}),$$

as required.

Question 3: First, observe that if we add $\eta_t x_{t+1}$ on both side of the equation defining x_{t+1}, we have

$$x_{t+1} + \eta_t x_{t+1} = \eta_t z_t + (1-\eta_t)y_t + \eta_t x_{t+1} \quad\Longleftrightarrow\quad \eta_t(x_{t+1} - z_t) = (1-\eta_t)(y_t - x_{t+1}),$$

Now, let $\mathbf{u} \in \Omega$, using the equality above we have

$$
\begin{aligned}
&\gamma_{t+1}\left(f(\mathbf{x}_{t+1}) - f(\mathbf{u})\right) \\
&\leqslant \gamma_{t+1}\left\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{u}\right\rangle && \text{by convexity} \\
&= \gamma_{t+1}\left\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{z}_t\right\rangle + \gamma_{t+1}\left\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{u}\right\rangle \\
&\leqslant \frac{(1-\eta_t)\gamma_{t+1}}{\eta_t}\left\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{y}_t - \mathbf{x}_{t+1}\right\rangle + \gamma_{t+1}\left\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{u}\right\rangle \\
&\leqslant \frac{(1-\eta_t)\gamma_{t+1}}{\eta_t}\left(f(\mathbf{y}_t) - f(\mathbf{x}_{t+1})\right) + \gamma_{t+1}\left\langle \nabla f(\mathbf{x}_{t+1}), \mathbf{z}_t - \mathbf{u}\right\rangle && \text{by convexity} \\
&\leqslant \frac{(1-\eta_t)\gamma_{t+1}}{\eta_t}\left(f(\mathbf{y}_t) - f(\mathbf{x}_{t+1})\right) + \gamma_{t+1}^2 L\left(f(\mathbf{x}_{t+1}) - f(\mathbf{y}_{t+1})\right) \\
&\quad + V_\omega(\mathbf{u}, \mathbf{z}_t) - V_\omega(\mathbf{u}, \mathbf{z}_{t+1}) && \text{by Question 1} \\
&= (\gamma_{t+1}^2 L - \gamma_{t+1}) f(\mathbf{y}_t) - \gamma_{t+1}^2 L f(\mathbf{y}_{t+1}) + \gamma_{t+1} f(\mathbf{x}_{t+1}) + V_\omega(\mathbf{u}, \mathbf{z}_t) - V_\omega(\mathbf{u}, \mathbf{z}_{t+1}) && \text{by } \eta_t = \frac{1}{\gamma_{t+1}L}
\end{aligned}
$$

The terms $\gamma_{t+1} f(\mathbf{x}_{t+1})$ cancel and we obtain

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

as required.

Question 4: First, we observe the following equivalence:

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

Now, using this equivalence and taking the telescoping sum of the result from Question 2 from $t = 0$ to $t = T-1$, we have

$$
\begin{aligned}
& \sum_{t=0}^{T-1} \gamma_{t+1}^2 L f(\mathbf{y}_{t+1}) - (\gamma_{t+1}^2 L - \gamma_{t+1}) f(\mathbf{y}_t) + V_\omega(\mathbf{u}, \mathbf{z}_{t+1}) - V_\omega(\mathbf{u}, \mathbf{z}_t) \leqslant \sum_{t=0}^{T-1} \gamma_{t+1} f(\mathbf{u}) \\
\iff \quad & \sum_{t=0}^{T-1} \gamma_{t+1}^2 L f(\mathbf{y}_{t+1}) - \gamma_t^2 L f(\mathbf{y}_t) + \frac{1}{4L} f(\mathbf{y}_t) + V_\omega(\mathbf{u}, \mathbf{z}_{t+1}) - V_\omega(\mathbf{u}, \mathbf{z}_t) \leqslant \sum_{t=1}^{T} \gamma_t f(\mathbf{u}) \\
\iff \quad & \gamma_T^2 L f(\mathbf{y}_T) - \underbrace{\gamma_0^2 L f(\mathbf{y}_0)}_{= \frac{1}{4L^2} L f(\mathbf{y}_0)} + \frac{1}{4L} f(\mathbf{y}_0) + V_\omega(\mathbf{u}, \mathbf{z}_T) - V_\omega(\mathbf{u}, \mathbf{z}_0) + \sum_{t=1}^{T-1} \frac{1}{4L} f(\mathbf{y}_t) \leqslant \sum_{t=1}^{T} \gamma_t f(\mathbf{u}).
\end{aligned}
$$

Now, we choose $\mathbf{u} = \mathbf{x}^*$ and we observe that

- $\sum_{t=1}^{T} \gamma_t = \sum_{t=1}^{T} \frac{t+1}{2L} = \frac{T(T+3)}{4L}$
- $f(\mathbf{y}_t) \geqslant f(\mathbf{x}^*)$ for $t = 0, 1, \ldots, T-1$ since $\mathbf{x}^*$ is a minimizer
- $V_\omega(\mathbf{u}, \mathbf{z}_T) \geqslant 0$
- $V_\omega(\mathbf{x}^*, \mathbf{z}_0) \leqslant R$

Hence, we have

$$\frac{(T+1)^2}{4L^2} L f(y_T) \leqslant \left(\frac{T(T+3)}{4L} - \frac{T-1}{4L} \right) f(x^*) + R,$$

which, rearranged, is equivalent to

$$f(y_T) - f(x^*) \leqslant \frac{4RL}{(T+1)^2},$$

as required. □