Training Neural Networks for and by Interpolation: Supplementary

Leonard Berrada [1] **Andrew Zisserman** [2] **M. Pawan Kumar** [2]

1. Local Interpretation of the Polyak Step-Size

In this section, we provide two results that shed light on a geometrical interpretation of the Polyak step-size. First, proposition 1 provides a proximal interpretation for the standard Polyak step-size. Second, proposition 2 gives a similar result when using a maximal learning-rate, which corresponds to the update used by ALI-G.

Proposition 1. *Suppose that the problem is unconstrained: $\Omega = \mathbb{R}^p$. Let $\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{f(\mathbf{w}_t) - f_\star}{\|\nabla f(\mathbf{w}_t)\|^2} \nabla f(\mathbf{w}_t)$. Then $\mathbf{w}_{t+1}$ verifies:*

$$\mathbf{w}_{t+1} = \underset{\mathbf{w} \in \mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{w} - \mathbf{w}_t\| \text{ subject to: } f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top (\mathbf{w} - \mathbf{w}_t) = f_\star, \tag{1}$$

where we remind that $f_\star$ is the minimum of f, and $\mathbf{w} \mapsto f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top (\mathbf{w} - \mathbf{w}_t)$ is the linearization of f at $\mathbf{w}_t$. In other words, $\mathbf{w}_{t+1}$ is the closest point to $\mathbf{w}_t$ that lies on the hyper-plane $f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top (\mathbf{w} - \mathbf{w}_t) = f_\star$.

Proof : First we show that $\mathbf{w}_{t+1}$ satisfies the linear equality constraint:

$$\begin{aligned} & f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top (\mathbf{w}_{t+1} - \mathbf{w}_t) \\ & \quad = f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top \left(- \frac{f(\mathbf{w}_t) - f_\star}{\|\nabla f(\mathbf{w}_t)\|^2} \nabla f(\mathbf{w}_t) \right), \\ & \quad = f(\mathbf{w}_t) - f(\mathbf{w}_t) + f_\star, \\ & \quad = f_\star. \end{aligned} \tag{2}$$

Now let us show that it has a minimal distance to $\mathbf{w}_t$.

We take $\hat{\boldsymbol{w}} \in \mathbb{R}^p$ a solution of the linear equality constraint, and we will show that $\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \leq \|\hat{\boldsymbol{w}} - \mathbf{w}_t\|$. By definition, we have that $\hat{\boldsymbol{w}}$ satisfies:

$$f(\mathbf{w}_t) + \nabla f(\mathbf{w}_t)^\top (\hat{\boldsymbol{w}} - \mathbf{w}_t) = f_\star. \tag{3}$$

Now we can write:

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}_t\| &= \| \frac{f(\mathbf{w}_t) - f_\star}{\|\nabla f(\mathbf{w}_t)\|^2} \nabla f(\mathbf{w}_t) \|, \\ &= \frac{f(\mathbf{w}_t) - f_\star}{\|\nabla f(\mathbf{w}_t)\|}, \\ &= \frac{|\nabla f(\mathbf{w}_t)^\top (\hat{\boldsymbol{w}} - \mathbf{w}_t)|}{\|\nabla f(\mathbf{w}_t)\|}, \\ &\leq \frac{\|\nabla f(\mathbf{w}_t)\| \|\hat{\boldsymbol{w}} - \mathbf{w}_t\|}{\|\nabla f(\mathbf{w}_t)\|}, \quad \text{(Cauchy-Schwarz)} \\ &= \|\hat{\boldsymbol{w}} - \mathbf{w}_t\|. \end{aligned} \tag{4}$$

■

[1]DeepMind, London, United Kingdom. Work performed while at University of Oxford. [2]Department of Engineering Science, University of Oxford, Oxford, United Kingdom. Correspondence to: Leonard Berrada <lberrada@google.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

Proposition 2. *[Proximal Interpretation] Suppose that $\Omega = \mathbb{R}^p$ and let $\delta = 0$. We consider the update performed by SGD: $\boldsymbol{w}_{t+1}^{SGD} = \boldsymbol{w}_t - \eta_t \nabla \ell_{z_t}(\boldsymbol{w}_t)$; and the update performed by ALI-G: $\boldsymbol{w}_{t+1}^{ALI\text{-}G} = \boldsymbol{w}_t - \gamma_t \nabla \ell_{z_t}(\boldsymbol{w}_t)$, where $\gamma_t = \min\left\{\frac{\ell_{z_t}(\boldsymbol{w}_t)}{\|\nabla \ell_{z_t}(\boldsymbol{w}_t)\|^2 + \delta}, \eta\right\}$. Then we have:*

$$\boldsymbol{w}_{t+1}^{SGD} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^p} \left\{ \frac{1}{2\eta_t}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2 + \ell_{z_t}(\boldsymbol{w}_t) + \nabla \ell_{z_t}(\boldsymbol{w}_t)^\top (\boldsymbol{w} - \boldsymbol{w}_t) \right\}, \tag{5}$$

$$\boldsymbol{w}_{t+1}^{ALI\text{-}G} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^p} \left\{ \frac{1}{2\eta}\|\boldsymbol{w} - \boldsymbol{w}_t\|^2 + \max\left\{ \ell_{z_t}(\boldsymbol{w}_t) + \nabla \ell_{z_t}(\boldsymbol{w}_t)^\top (\boldsymbol{w} - \boldsymbol{w}_t), 0 \right\} \right\}. \tag{6}$$

Proof : In order to make the notation simpler, we use $\boldsymbol{d}_t \triangleq \nabla \ell_{z_t}(\mathbf{w}_t)$ and $l_t \triangleq \ell_{z_t}(\mathbf{w}_t)$.
First, let us consider $\boldsymbol{d}_t = \mathbf{0}$.

Then we choose $\gamma_t = 0$ and it is clear that $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\gamma_t \boldsymbol{d}_t = \mathbf{w}_t$ is the optimal solution of problem (6).

We now assume $\boldsymbol{d}_t \neq \mathbf{0}$.

We can successively re-write the proximal problem (6) as :

$$\begin{aligned}
&\min_{\mathbf{w}\in\mathbb{R}^p} \left\{ \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_t\|^2 + \max\left\{ \ell_{z_t}(\mathbf{w}_t) + \nabla \ell_{z_t}(\mathbf{w}_t)^\top (\mathbf{w} - \mathbf{w}_t), 0 \right\} \right\}, \\
&\min_{\mathbf{w}\in\mathbb{R}^p} \left\{ \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_t\|^2 + \max\left\{ l_t + \boldsymbol{d}_t^\top (\mathbf{w} - \mathbf{w}_t), 0 \right\} \right\}, \\
&\min_{\mathbf{w}\in\mathbb{R}^p, \upsilon} \left\{ \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_t\|^2 + \upsilon \right\} \text{ subject to: } \upsilon \geq 0,\ \upsilon \geq l_t + \boldsymbol{d}_t^\top (\mathbf{w} - \mathbf{w}_t) \\
&\min_{\mathbf{w}\in\mathbb{R}^p, \upsilon} \sup_{\mu,\nu\geq 0} \left\{ \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_t\|^2 + \upsilon - \mu\upsilon - \nu(\upsilon - l_t - \boldsymbol{d}_t^\top (\mathbf{w} - \mathbf{w}_t)) \right\} \\
&\sup_{\mu,\nu\geq 0} \min_{\mathbf{w}\in\mathbb{R}^p, \upsilon} \left\{ \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_t\|^2 + \upsilon - \mu\upsilon - \nu(\upsilon - l_t - \boldsymbol{d}_t^\top (\mathbf{w} - \mathbf{w}_t)) \right\},
\end{aligned} \tag{7}$$

where the last equation uses strong duality. The inner problem is now smooth in $\mathbf{w}$ and υ. We write its KKT conditions:

$$\frac{\partial \cdot}{\partial \upsilon} = 0 : \quad 1 - \mu - \nu = 0 \tag{8}$$

$$\frac{\partial \cdot}{\partial \mathbf{w}} = 0 : \quad \frac{1}{\eta}(\mathbf{w} - \mathbf{w}_t) + \nu \boldsymbol{d}_t = \mathbf{0} \tag{9}$$

We plug in these results and obtain:

$$\begin{aligned}
&\sup_{\mu,\nu\geq 0} \left\{ \frac{1}{2\eta}\|\eta\nu\boldsymbol{d}_t\|^2 + \nu(l_t + \boldsymbol{d}_t^\top(-\eta\nu\boldsymbol{d}_t)) \right\} \\
&\text{st:} \quad \mu + \nu = 1 \\
&\sup_{\nu\in[0,1]} \left\{ \frac{\eta}{2}\nu^2\|\boldsymbol{d}_t\|^2 + \nu l_t - \eta\nu^2\|\boldsymbol{d}_t^\top\|^2 \right\} \\
&\sup_{\nu\in[0,1]} \left\{ -\frac{\eta}{2}\nu^2\|\boldsymbol{d}_t\|^2 + \nu l_t \right\}
\end{aligned} \tag{10}$$

This is a one-dimensional quadratic problem in ν. It can be solved in closed-form by finding the global maximum of the quadratic objective, and projecting the solution on $[0, 1]$. We have:

$$\frac{\partial \cdot}{\partial \nu} = 0 : -\eta\nu\|\boldsymbol{d}_t\|^2 + l_t = 0 \tag{11}$$

Since $\boldsymbol{d}_t \neq \mathbf{0}$ and $\eta \neq 0$, this gives the optimal solution:

$$\nu = \min\left\{ \max\left\{ \frac{l_t}{\eta\|\boldsymbol{d}_t\|^2}, 0 \right\}, 1 \right\} = \min\left\{ \frac{l_t}{\eta\|\boldsymbol{d}_t\|^2}, 1 \right\}, \tag{12}$$

since $l_t, \eta, \|\boldsymbol{d}_t\|^2 \geq 0$.
Plugging this back in the KKT conditions, we obtain that the solution $\mathbf{w}_{t+1}$ of the primal problem can be written as:

$$\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t - \eta\nu\boldsymbol{d}_t, \\
&= \mathbf{w}_t - \eta \min\left\{\frac{l_t}{\eta\|\boldsymbol{d}_t\|^2}, 1\right\}\boldsymbol{d}_t, \\
&= \mathbf{w}_t - \eta \min\left\{\frac{\ell_{z_t}(\mathbf{w}_t)}{\eta\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2}, 1\right\}\nabla\ell_{z_t}(\mathbf{w}_t), \\
&= \mathbf{w}_t - \min\left\{\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2}, \eta\right\}\nabla\ell_{z_t}(\mathbf{w}_t).
\end{aligned} \tag{13}$$

■

2. Summary of Convergence Results

Problem Formulation. We remind the problem setting as follows. The learning task can be expressed as the problem $(\mathcal{P})$ of finding a feasible vector of parameters $\mathbf{w}_\star \in \Omega$ that minimizes f:

$$\mathbf{w}_\star \in \operatorname*{argmin}_{\boldsymbol{w}\in\Omega} f(\boldsymbol{w}). \tag{$\mathcal{P}$}$$

Also note that $f_\star$ refers to the minimum value of f over Ω: $f_\star \triangleq \min_{\boldsymbol{w}\in\Omega} f(\boldsymbol{w})$.

In the remainder of this section, we give an overview of convergence results of ALI-G in various stochastic settings. First, we summarize convergence results in the convex setting in section 2.1. Notably, these results show convergence for any maximal learning-rate η, including $\eta = \infty$, which is equivalent to not using any clipping to a maximal value. Second, we give results for a class of non-convex problems. These results show that a maximal learning-rate is necessary and sufficient for convergence of the Polyak step-size. Indeed we show that the Polyak step-size can oscillate indefinitely without a maximal learning-rate, and that using a maximal learning-rate provably leads to (exponentially fast) convergence.

2.1. Convex Setting

For simplicity purposes, we assume that we are in the perfect interpolation setting: $\forall z,\ \ell_z(\mathbf{w}_\star) = 0$. Detailed results with an interpolation tolerance $\varepsilon > 0$ are given in section 3. Since we are in the perfect interpolation setting, note that we can safely set the small constant for numerical stability to zero: $\delta = 0$. The summary of the results is presented in table 1.

Assumption on Loss Functions	Distance Considered	Convergence Rate	
		Small η	Large η (potentially ∞)
Convex and C-Lipschitz	$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star$	$\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{\eta(T+1)} + \sqrt{\frac{C^2\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1}}$	$\sqrt{\frac{C^2\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1}}$
Convex and β-Smooth	$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star$	$\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{\eta(T+1)}$	$\frac{2\beta\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1}$
α-Strongly Convex and β-Smooth	$\mathbb{E}[f(\mathbf{w}_{T+1})] - f_\star$	$\frac{\beta}{2}\exp\left(\frac{-\alpha\eta T}{2}\right)\|\mathbf{w}_0-\mathbf{w}_\star\|^2$	$\frac{\beta}{2}\exp\left(-\frac{\alpha t}{4\beta}\right)\|\mathbf{w}_0-\mathbf{w}_\star\|^2$

Table 1. Summary of convergence rates for convex problems in the perfect interpolation setting. We remind that η denotes the hyper-parameter used by ALI-G to clip its learning-rate to a maximal value. Our convergence results yield different results when η has a small value (middle column), and when η has a large, possibly even infinite, value (right column). The formal statements of these results are available in section 3, along with their proofs.

The overall convergence speed is similar to that of *non-stochastic* Polyak step-size, which is itself the same as the optimal rate of *non-stochastic* gradient descent: $\mathcal{O}(1/\sqrt{T})$ for convex Lipschitz functions, $\mathcal{O}(1/T)$ for convex and smooth functions, and $\mathcal{O}(\exp(-kT))$ (for some constant k) for smooth and strongly convex functions (Hazan & Kakade, 2019).

2.2. Non-Convex Setting

We also assume that we are in the perfect interpolation setting and thus we set the constant for numerical stability δ to zero. We further assume that the problem is unconstrained. The summary of the results is presented in table 2.

Convergence Result	
$0 < \eta \leq \frac{2\alpha}{\beta^2}$	$\eta = \infty$
$f(\mathbf{w}_{T+1}) - f_\star \leq \frac{\beta}{2} \exp\left(-\kappa T\right) \left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2$	Can Fail to Converge (Proved)

Table 2. Summary of convergence results for α-RSI and β-smooth loss functions in the perfect interpolation setting. We remind that η denotes the hyper-parameter used by ALI-G to clip its learning-rate to a maximal value. The constant κ depends on α, β and η. These results show that using a maximal learning-rate is necessary and sufficient for convergence. The formal statements of these results are available in section 4, along with their proofs.

3. Detailed Convex Results

3.1. Lipschitz Convex Functions

Theorem 1. *We assume that Ω is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and C-Lipschitz. Let $\mathbf{w}_\star$ be a solution of ($\mathcal{P}$) such that $\forall z \in \mathcal{Z}$, $\ell_z(\mathbf{w}_\star) \leq \varepsilon$. We further assume that $\eta > \frac{\varepsilon}{\delta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star \leq \frac{\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{(\eta - \frac{\varepsilon}{\delta})(T+1)} + \frac{\varepsilon^2}{\delta(\eta - \frac{\varepsilon}{\delta})} + \sqrt{\frac{(C^2+\delta)\left\|\mathbf{w}_0 - \mathbf{w}_\star\right\|^2}{T+1}} + \varepsilon\sqrt{\frac{C^2}{\delta} + 1}. \tag{14}$$

Proof :

We consider the update at time t, which we condition on the draw of $z_t \in \mathcal{Z}$:

$$\begin{aligned}
&\left\|\mathbf{w}_{t+1} - \mathbf{w}_\star\right\|^2 \\
&\quad= \left\|\Pi_\Omega(\mathbf{w}_t - \gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)) - \mathbf{w}_\star\right\|^2 \\
&\quad\leq \left\|\mathbf{w}_t - \gamma_t \nabla \ell_{z_t}(\mathbf{w}_t) - \mathbf{w}_\star\right\|^2 \quad (\Pi_\Omega \text{ projection}) \\
&\quad= \left\|\mathbf{w}_t - \mathbf{w}_\star\right\|^2 - 2\gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}_\star) + \gamma_t^2 \left\|\nabla \ell_{z_t}(\mathbf{w}_t)\right\|^2 \\
&\quad\leq \left\|\mathbf{w}_t - \mathbf{w}_\star\right\|^2 - 2\gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}_\star) + \gamma_t \frac{\ell_{z_t}(\mathbf{w}_t)}{\left\|\nabla \ell_{z_t}(\mathbf{w}_t)\right\|^2 + \delta} \left\|\nabla \ell_{z_t}(\mathbf{w}_t)\right\|^2 \\
&\qquad\text{(because } \gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t)}{\left\|\nabla \ell_{z_t}(\mathbf{w}_t)\right\|^2 + \delta}\text{)} \\
&\quad\leq \left\|\mathbf{w}_t - \mathbf{w}_\star\right\|^2 - 2\gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)^\top(\mathbf{w}_t - \mathbf{w}_\star) + \gamma_t \frac{\ell_{z_t}(\mathbf{w}_t)}{\left\|\nabla \ell_{z_t}(\mathbf{w}_t)\right\|^2} \left\|\nabla \ell_{z_t}(\mathbf{w}_t)\right\|^2 \\
&\qquad\text{(because } \ell_{z_t}(\mathbf{w}_t) \geq 0 \text{ and } \delta \geq 0\text{)} \\
&\quad\leq \left\|\mathbf{w}_t - \mathbf{w}_\star\right\|^2 - 2\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \ell_{z_t}(\mathbf{w}_t) \quad (\text{convexity of } \ell_{z_t}) \\
&\quad= \left\|\mathbf{w}_t - \mathbf{w}_\star\right\|^2 - 2\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \ell_{z_t}(\mathbf{w}_\star) \\
&\quad= \left\|\mathbf{w}_t - \mathbf{w}_\star\right\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \ell_{z_t}(\mathbf{w}_\star)
\end{aligned} \tag{15}$$

We now consider different cases, according to the value that γ_t takes: $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t)}{\left\|\nabla \ell_{z_t}(\mathbf{w}_t)\right\|^2 + \delta}$ or $\gamma_t = \eta$.

First, suppose that $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$. Then we have:

$$
\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t\Big(\ell_{z_t}(\mathbf{w}_t) - 2\ell_{z_t}(\mathbf{w}_\star)\Big) \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}\Big(\ell_{z_t}(\mathbf{w}_t)^2 - 2\ell_{z_t}(\mathbf{w}_t)\ell_{z_t}(\mathbf{w}_\star)\Big) \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}\Big((\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2 - \ell_{z_t}(\mathbf{w}_\star)^2\Big) \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} + \frac{\ell_{z_t}(\mathbf{w}_\star)^2}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{C^2 + \delta} + \frac{\ell_{z_t}(\mathbf{w}_\star)^2}{\delta} \\
&\qquad \text{(because we have } 0 \leq \|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 \leq C^2\text{)} \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{C^2 + \delta} + \frac{\varepsilon^2}{\delta} \quad \text{(definition of } \varepsilon\text{)}
\end{aligned} \tag{16}
$$

Now suppose $\gamma_t = \eta$ and $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0$. We can use $\gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$ to write:

$$
\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \ell_{z_t}(\mathbf{w}_\star), \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}\ell_{z_t}(\mathbf{w}_\star),
\end{aligned} \tag{17}
$$

where the last inequality has used $\gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$, $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0$ and $\ell_{z_t}(\mathbf{w}_\star) \geq 0$. Therefore we are exactly in the same situation as the first case (where we used $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$), and thus we have again:

$$
\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{C^2 + \delta} + \frac{\varepsilon^2}{\delta}. \tag{18}
$$

Now suppose that $\gamma_t = \eta$ and $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$. The inequality (15) gives:

$$
\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \ell_{z_t}(\mathbf{w}_\star), \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \ell_{z_t}(\mathbf{w}_\star), \quad (\gamma_t = \eta) \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \varepsilon, \quad \text{(definition of } \varepsilon, \gamma_t \geq 0\text{)} \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \varepsilon\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}, \\
&\qquad \text{(because } \gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}, \varepsilon \geq 0\text{)} \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \varepsilon\frac{\ell_{z_t}(\mathbf{w}_t)}{\delta}, \\
&\qquad \text{(because } \|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 \geq 0\text{)} \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \varepsilon\frac{\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) + \ell_{z_t}(\mathbf{w}_\star)}{\delta}, \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \varepsilon\frac{\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) + \varepsilon}{\delta}, \\
&\qquad \text{(because } \ell_{z_t}(\mathbf{w}_\star) \leq \varepsilon\text{)} \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \Big(\eta - \frac{\varepsilon}{\delta}\Big)(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\varepsilon^2}{\delta}.
\end{aligned} \tag{19}
$$

We now introduce $\mathcal{I}_T$ and $\mathcal{J}_T$ as follows:

$$\begin{aligned}\mathcal{I}_T &\triangleq \{t \in \{0, ..., T\} : \gamma_t = \eta \text{ and } \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0\} \\ \mathcal{J}_T &\triangleq \{0, ..., T\} \setminus \mathcal{I}_T\end{aligned} \tag{20}$$

Then, by combining inequalities (16), (18) and (19), and using a telescopic sum, we obtain:

$$\begin{aligned}\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 &+ \sum_{t \in \mathcal{J}_T} \left(-\frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{C^2 + \delta} + \frac{\varepsilon^2}{\delta} \right) \\ &+ \sum_{t \in \mathcal{I}_T} \left(-\left(\eta - \frac{\varepsilon}{\delta}\right)(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\varepsilon^2}{\delta} \right)\end{aligned} \tag{21}$$

Using $\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2 \geq 0$, we obtain:

$$\begin{aligned}\frac{1}{C^2 + \delta} \sum_{t \in \mathcal{J}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2 + \left(\eta - \frac{\varepsilon}{\delta}\right) \sum_{t \in \mathcal{I}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \\ \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta}\end{aligned} \tag{22}$$

In particular, the inequality (22) gives that:

$$\left(\eta - \frac{\varepsilon}{\delta}\right) \sum_{t \in \mathcal{I}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta}. \tag{23}$$

Furthermore, for every $t \in \mathcal{I}_T$, we have $(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq 0$, which yields $\left(\eta - \frac{\varepsilon}{\delta}\right) \sum_{t \in \mathcal{I}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq 0$ since $\eta > \frac{\epsilon}{\delta}$. Thus the inequality (22) also gives:

$$\frac{1}{C^2 + \delta} \sum_{t \in \mathcal{J}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2 \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta}. \tag{24}$$

Using the Cauchy-Schwarz inequality, we can further write:

$$\left(\sum_{t \in \mathcal{J}_T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \right)^2 \leq |\mathcal{J}_T| \sum_{t \in \mathcal{J}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2. \tag{25}$$

Therefore we have:

$$\begin{aligned}\sum_{t \in \mathcal{J}_T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) &\leq \sqrt{|\mathcal{J}_T| \sum_{t \in \mathcal{J}_T} (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}, \\ &\leq \sqrt{|\mathcal{J}_T|(C^2 + \delta)\left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta}\right)}.\end{aligned} \tag{26}$$

We can now put together inequalities (23) and (26) by writing:

$$\begin{aligned}
&\sum_{t=0}^{T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \\
&\quad = \sum_{t\in\mathcal{I}_T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) + \sum_{t\in\mathcal{J}_T} \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \\
&\quad \leq \frac{1}{\eta - \frac{\varepsilon}{\delta}} \left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta} \right) \\
&\qquad + \sqrt{|\mathcal{J}_T|(C^2+\delta)\left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta} \right)} \\
&\quad \leq \frac{1}{\eta - \frac{\varepsilon}{\delta}} \left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta} \right) \\
&\qquad + \sqrt{(T+1)(C^2+\delta)\left(\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + (T+1)\frac{\varepsilon^2}{\delta} \right)}
\end{aligned} \tag{27}$$

Dividing by $T+1$ and taking the expectation, we obtain:

$$\begin{aligned}
&f\left(\frac{1}{T+1}\sum_{t=0}^{T} \mathbf{w}_t \right) - f_\star \\
&\quad \leq \frac{1}{T+1}\sum_{t=0}^{T} f(\mathbf{w}_t) - f_\star, \quad (f \text{ is convex}) \\
&\quad \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{(\eta - \frac{\varepsilon}{\delta})(T+1)} + \frac{\varepsilon^2}{\delta(\eta - \frac{\varepsilon}{\delta})} + \sqrt{(C^2+\delta)\left(\frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1} + \frac{\varepsilon^2}{\delta} \right)}, \\
&\quad \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{(\eta - \frac{\varepsilon}{\delta})(T+1)} + \frac{\varepsilon^2}{\delta(\eta - \frac{\varepsilon}{\delta})} + \sqrt{\frac{(C^2+\delta)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}} + \varepsilon\sqrt{\frac{C^2}{\delta} + 1}.
\end{aligned} \tag{28}$$

■

When η is small, the convergence error of Theorem 1 is large. This is corrected in the following result which is informative in the regime where η is small:

Theorem 2. *We assume that Ω is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is convex and C-Lipschitz. Let $\mathbf{w}_\star$ be a solution of ($\mathcal{P}$) such that $\forall z \in \mathcal{Z},\ \ell_z(\mathbf{w}_\star) \leq \varepsilon$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T} \mathbf{w}_t \right) - f_\star \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\eta(T+1)} + \varepsilon + \sqrt{\frac{(C^2+\delta)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}} + \eta\varepsilon\sqrt{C^2+\delta}. \tag{29}$$

Proof :

We consider the update at time t, which we condition on the draw of $z_t \in \mathcal{Z}$. We re-use the inequality (15) from the proof of Theorem 1:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \ell_{z_t}(\mathbf{w}_\star) \tag{30}$$

We consider again different cases, according to the value of γ_t and the sign of $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)$.

Suppose that $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0$. Then the inequality (30) gives:

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \\
&\quad\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \ell_{z_t}(\mathbf{w}_\star), \\
&\quad\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \ell_{z_t}(\mathbf{w}_\star), \\
&\qquad\text{(because } \gamma_t \leq \eta, \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0) \\
&\quad\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \varepsilon, \quad \text{(definition of } \varepsilon, \gamma_t \geq 0) \\
&\quad\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \eta \varepsilon, \quad (\gamma_t \leq \eta, \varepsilon \geq 0)
\end{aligned} \tag{31}$$

Now suppose $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$ and $\gamma_t = \eta$. Then the inequality (30) gives:

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \ell_{z_t}(\mathbf{w}_\star), \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \eta \ell_{z_t}(\mathbf{w}_\star), \\
&\qquad\text{(because } \gamma_t = \eta) \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \eta \varepsilon, \\
&\qquad\text{(definition of } \varepsilon, \eta \geq 0)
\end{aligned} \tag{32}$$

Finally, suppose that $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$ and $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}$. Then the inequality (30) gives:

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \\
&\quad\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \gamma_t \ell_{z_t}(\mathbf{w}_\star), \\
&\quad\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \eta \ell_{z_t}(\mathbf{w}_\star), \\
&\qquad\text{(because } \gamma_t \leq \eta, \ell_{z_t}(\mathbf{w}_\star) \geq 0) \\
&\quad\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \eta \varepsilon, \quad \text{(definition of } \varepsilon, \eta \geq 0) \\
&\quad= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \eta \varepsilon, \\
&\qquad\text{(because } \gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}) \\
&\quad\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} + \eta \varepsilon, \\
&\qquad\text{(because } \ell_{z_t}(\mathbf{w}_t) \geq \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0) \\
&\quad\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{C^2 + \delta} + \eta \varepsilon, \quad (\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 \leq C^2)
\end{aligned} \tag{33}$$

We now introduce $\mathcal{I}_T$ and $\mathcal{J}_T$ as follows:

$$\begin{aligned}
\mathcal{J}_T &\triangleq \left\{ t \in \{0, ..., T\} : \gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \text{ and } \ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0 \right\} \\
\mathcal{I}_T &\triangleq \{0, ..., T\} \setminus \mathcal{I}_T
\end{aligned} \tag{34}$$

Then, by combining inequalities (31), (32) and (33), and using a telescopic sum, we obtain:

$$\begin{aligned}
\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 &+ \sum_{t \in \mathcal{J}_T} \left(-\frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star))^2}{C^2 + \delta} + \eta \varepsilon \right) \\
&+ \sum_{t \in \mathcal{I}_T} (-\eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \eta \varepsilon)
\end{aligned} \tag{35}$$

Using $\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2 \geq 0$, we obtain:

$$\begin{aligned}\frac{1}{C^2+\delta}\sum_{t\in\mathcal{J}_T}\left(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\right)^2+\eta\sum_{t\in\mathcal{I}_T}\left(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\right)\\ \leq \|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\end{aligned} \tag{36}$$

We now take the expectation and obtain:

$$\begin{aligned}\frac{1}{C^2+\delta}\sum_{t\in\mathcal{J}_T}\mathbb{E}\left[\left(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\right)^2\right]+\eta\sum_{t\in\mathcal{I}_T}\left(f(\mathbf{w}_t)-f_\star\right)\\ \leq \|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\end{aligned} \tag{37}$$

Since $\mathbb{E}[U]^2 \leq \mathbb{E}[U^2]$ for any real-valued random variable, we can write:

$$\frac{1}{C^2+\delta}\sum_{t\in\mathcal{J}_T}\left(f(\mathbf{w}_t)-f_\star\right)^2+\eta\sum_{t\in\mathcal{I}_T}\left(f(\mathbf{w}_t)-f_\star\right)\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon \tag{38}$$

Since each $f(\mathbf{w}_t)-f_\star\geq 0$, the inequality (38) gives that:

$$\eta\sum_{t\in\mathcal{I}_T}\left(f(\mathbf{w}_t)-f_\star\right)\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon, \tag{39}$$

and:

$$\frac{1}{C^2+\delta}\sum_{t\in\mathcal{J}_T}\left(f(\mathbf{w}_t)-f_\star\right)^2\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon. \tag{40}$$

Using the Cauchy-Schwarz inequality, we can further write:

$$\left(\sum_{t\in\mathcal{J}_T}f(\mathbf{w}_t)-f_\star\right)^2\leq|\mathcal{J}_T|\sum_{t\in\mathcal{J}_T}\left(f(\mathbf{w}_t)-f_\star\right)^2. \tag{41}$$

Therefore we have:

$$\begin{aligned}\sum_{t\in\mathcal{J}_T}f(\mathbf{w}_t)-f_\star&\leq\sqrt{|\mathcal{J}_T|\sum_{t\in\mathcal{J}_T}\left(f(\mathbf{w}_t)-f_\star\right)^2},\\ &\leq\sqrt{|\mathcal{J}_T|(C^2+\delta)\left(\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\right)}.\end{aligned} \tag{42}$$

We can now put together inequalities (39) and (42) by writing:

$$\begin{aligned}\sum_{t=0}^{T}f(\mathbf{w}_t)-f_\star&=\sum_{t\in\mathcal{I}_T}f(\mathbf{w}_t)-f_\star+\sum_{t\in\mathcal{J}_T}f(\mathbf{w}_t)-f_\star\\ &\leq\frac{1}{\eta}\left(\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\right)+\sqrt{|\mathcal{J}_T|(C^2+\delta)\left(\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\right)}\\ &\leq\frac{1}{\eta}\left(\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\right)+\sqrt{(T+1)(C^2+\delta)\left(\|\mathbf{w}_0-\mathbf{w}_\star\|^2+(T+1)\eta\varepsilon\right)}\end{aligned} \tag{43}$$

Dividing by $T+1$, we obtain:

$$\begin{aligned} & f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right) - f_\star \\ & \leq \frac{1}{T+1}\sum_{t=0}^{T} f(\mathbf{w}_t) - f_\star, \quad (f \text{ is convex}) \\ & \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\eta(T+1)} + \varepsilon + \sqrt{(C^2+\delta)\left(\frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1} + \eta\varepsilon\right)}, \\ & \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\eta(T+1)} + \varepsilon + \sqrt{\frac{(C^2+\delta)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{T+1}} + \eta\varepsilon\sqrt{C^2+\delta}. \end{aligned} \tag{44}$$

■

3.2. Smooth Convex Functions

We now tackle the convex and β-smooth case. Our proof techniques naturally produce the separation $\eta \geq \frac{1}{2\beta}$ and $\eta \leq \frac{1}{2\beta}$.

Lemma 1. *Let $z \in \mathcal{Z}$. Assume that ℓ_z is β-smooth and non-negative on $\mathbb{R}^p$. Then we have:*

$$\forall\, \mathbf{w} \in \mathbb{R}^p,\ \ell_z(\mathbf{w}) \geq \frac{1}{2\beta}\|\nabla \ell_z(\mathbf{w})\|^2 \tag{45}$$

Note that we do not assume that ℓ_z is convex.

Proof :

Let $\boldsymbol{w} \in \mathbb{R}^p$. By Lemma 3.4 of (Bubeck, 2015), we have:

$$\forall\, \boldsymbol{u} \in \mathbb{R}^p,\ |\ell_z(\boldsymbol{u}) - \ell_z(\boldsymbol{w}) - \nabla \ell_z(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w})| \leq \frac{\beta}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2. \tag{46}$$

Therefore we can write:

$$\forall\, \boldsymbol{u} \in \mathbb{R}^p,\ \ell_z(\boldsymbol{u}) \leq \ell_z(\boldsymbol{w}) + \nabla \ell_z(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w}) + \frac{\beta}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2. \tag{47}$$

And since $\forall\, \boldsymbol{u} \in \mathbb{R}^p,\ \ell_z(\boldsymbol{u}) \geq 0$, we have:

$$\forall\, \boldsymbol{u} \in \mathbb{R}^p,\ 0 \leq \ell_z(\boldsymbol{w}) + \nabla \ell_z(\boldsymbol{w})^\top(\boldsymbol{u} - \boldsymbol{w}) + \frac{\beta}{2}\|\boldsymbol{u} - \boldsymbol{w}\|^2. \tag{48}$$

We now choose $\boldsymbol{u} = \boldsymbol{w} - \frac{1}{\beta}\nabla \ell_z(\boldsymbol{w})$, which yields:

$$0 \leq \ell_z(\boldsymbol{w}) - \frac{1}{\beta}\|\nabla \ell_z(\boldsymbol{w})\|^2 + \frac{1}{2\beta}\|\nabla \ell_z(\boldsymbol{w})\|^2, \tag{49}$$

which gives the desired result.

■

Lemma 2. *Let $z \in \mathcal{Z}$. Assume that ℓ_z is β-smooth and non-negative on $\mathbb{R}^p$. Then we have:*

$$\forall\, \mathbf{w} \in \mathbb{R}^p,\ \frac{\ell_z(\mathbf{w})}{\|\nabla \ell_z(\mathbf{w})\|^2 + \delta} \geq \frac{1}{2\beta} - \frac{\delta}{4\beta^2 \ell_z(\mathbf{w})} \tag{50}$$

Proof :

Let $\mathbf{w} \in \mathbb{R}^p$. We apply Lemma 1 and we write successively:

$$\begin{aligned}\frac{\ell_z(\mathbf{w})}{\|\nabla\ell_z(\mathbf{w})\|^2+\delta} &\geq \frac{\ell_z(\mathbf{w})}{2\beta\ell_z(\mathbf{w})+\delta}, \quad \text{(Lemma 1)}\\ &= \frac{\ell_z(\mathbf{w})+\frac{\delta}{2\beta}-\frac{\delta}{2\beta}}{2\beta(\ell_z(\mathbf{w})+\frac{\delta}{2\beta})},\\ &= \frac{1}{2\beta}-\frac{\frac{\delta}{2\beta}}{2\beta(\ell_z(\mathbf{w})+\frac{\delta}{2\beta})},\\ &\geq \frac{1}{2\beta}-\frac{\delta}{4\beta^2\ell_z(\mathbf{w})}. \quad (\delta\geq 0)\end{aligned} \tag{51}$$

■

Theorem 3. *We assume that Ω is a convex set, and that for every $z\in\mathcal{Z}$, ℓ_z is convex and β-smooth. Let $\mathbf{w}_\star$ be a solution of ($\mathcal{P}$) such that $\forall z\in\mathcal{Z},\ \ell_z(\mathbf{w}_\star)\leq\varepsilon$, and suppose that $\delta>2\beta\varepsilon$. Further assume that $\eta\geq\frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right)-f_\star\leq\frac{\delta}{\beta(1-\frac{2\beta\varepsilon}{\delta})}+\frac{2\beta}{1-\frac{2\beta\varepsilon}{\delta}}\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1}. \tag{52}$$

Proof :

We re-use the inequality (15) from the proof of Theorem 1:

$$\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t\ell_{z_t}(\mathbf{w}_\star) \tag{53}$$

As previously, we lower bound $\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))$ and upper bound $\gamma_t\ell_{z_t}(\mathbf{w}_\star)$ individually.

We begin with $\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))$. We remark that either $\gamma_t=\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$ or $\gamma_t=\eta$.

Suppose $\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\geq 0$ and $\gamma_t=\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$. Then we can write:

$$\begin{aligned}&\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))\\ &\quad=\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)), \quad \text{(definition of } \gamma_t)\\ &\quad\geq\left(\frac{1}{2\beta}-\frac{\delta}{4\beta^2\ell_z(\mathbf{w}_t)}\right)(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))\\ &\qquad\text{(using Lemma 2, } \ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\geq 0)\\ &\quad=\frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))-\frac{\delta}{4\beta^2}\frac{\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)}{\ell_{z_t}(\mathbf{w}_t)}\\ &\quad\geq\frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))-\frac{\delta}{4\beta^2} \quad (\ell_{z_t}(\mathbf{w}_\star)\geq 0,\ell_{z_t}(\mathbf{w}_t)\geq 0)\end{aligned} \tag{54}$$

Now suppose $\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\geq 0$ and $\gamma_t=\eta$. Then we have:

$$\begin{aligned}\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))&=\eta(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))\\ &\geq\eta(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))-\frac{\delta}{4\beta^2}\\ &\geq\frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))-\frac{\delta}{4\beta^2}\\ &\qquad\text{(because } \eta\geq\frac{1}{2\beta},\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star)\geq 0).\end{aligned} \tag{55}$$

Now suppose $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0$. We have:

$$\begin{aligned}
\gamma_t &\leq \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&\leq \frac{\ell_{z_t}(\mathbf{w}_\star)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \quad (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0) \\
&\leq \frac{\varepsilon}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \quad \text{(definition of } \varepsilon) \\
&\leq \frac{\varepsilon}{\delta} \quad (\|\nabla \ell_{z_t}(\mathbf{w}_t)\| \geq 0) \\
&\leq \frac{1}{2\beta} \quad (\delta \geq 2\beta\varepsilon)
\end{aligned} \tag{56}$$

We now write:

$$\begin{aligned}
\gamma_t \left(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)\right) &\geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \quad (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0, \gamma_t \leq \frac{1}{2\beta}) \\
&\geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{4\beta^2}
\end{aligned} \tag{57}$$

In conclusion, in all cases, it holds true that:

$$\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{4\beta^2} \tag{58}$$

We now upper bound $\gamma_t \ell_{z_t}(\mathbf{w}_\star)$:

$$\begin{aligned}
\gamma_t \ell_{z_t}(\mathbf{w}_\star) &\leq \frac{\ell_{z_t}(\mathbf{w}_t)\ell_{z_t}(\mathbf{w}_\star)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}, \quad \text{(definition of } \gamma_t \text{ and } \ell_{z_t}(\mathbf{w}_\star) \geq 0) \\
&\leq \frac{\ell_{z_t}(\mathbf{w}_t)\ell_{z_t}(\mathbf{w}_\star)}{\delta}, \quad (\|\nabla \ell_{z_t}(\mathbf{w}_t)\| \geq 0) \\
&\leq \frac{(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) + \varepsilon)\varepsilon}{\delta}, \quad \text{(definition of } \varepsilon \text{ twice)} \\
&= \frac{\varepsilon}{\delta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\varepsilon^2}{\delta}.
\end{aligned} \tag{59}$$

We now put together inequalities (53), (58) and (59):

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \\
&\quad \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \frac{\varepsilon}{\delta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\varepsilon^2}{\delta}, \\
&\quad = \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \left(\frac{1}{2\beta} - \frac{\varepsilon}{\delta}\right)(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \frac{\varepsilon^2}{\delta}.
\end{aligned} \tag{60}$$

Therefore we have:

$$\left(\frac{1}{2\beta} - \frac{\varepsilon}{\delta}\right)(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \left(\frac{\delta}{4\beta^2} + \frac{\varepsilon^2}{\delta}\right) \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2. \tag{61}$$

By summing over t and taking the expectation over the z_t, we obtain:

$$\begin{aligned}
&\sum_{t=0}^{T}\left(\frac{\delta-2\beta\varepsilon}{2\beta\delta}\left(f(\mathbf{w}_t)-f(\mathbf{w}_\star)\right)-\frac{\delta^2+4\beta^2\varepsilon^2}{4\beta^2\delta}\right)\\
&\quad\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2-\mathbb{E}\left[\|\mathbf{w}_{T+1}-\mathbf{w}_\star\|^2\right],\\
&\quad\leq\|\mathbf{w}_0-\mathbf{w}_\star\|^2.
\end{aligned}\tag{62}$$

By assumption, we have that $\delta-2\beta\varepsilon>0$. Dividing by $T+1$ and using the convexity of f, we finally obtain:

$$\begin{aligned}
f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right)-f_\star&\leq\frac{1}{T+1}\sum_{t=0}^{T}f(\mathbf{w}_t)-f_\star\quad\text{(convexity of }f\text{)},\\
&=\frac{2\beta\delta}{\delta-2\beta\varepsilon}\frac{\delta^2+4\beta^2\varepsilon^2}{4\beta^2\delta}+\frac{2\beta\delta}{\delta-2\beta\varepsilon}\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1},\\
&=\frac{\delta^2+4\beta^2\varepsilon^2}{2\beta(\delta-2\beta\varepsilon)}+\frac{2\beta\delta}{\delta-2\beta\varepsilon}\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1},\\
&\leq\frac{\delta^2}{\beta(\delta-2\beta\varepsilon)}+\frac{2\beta\delta}{\delta-2\beta\varepsilon}\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1},\quad(\delta-2\beta\varepsilon\geq0)\\
&=\frac{\delta}{\beta(1-\frac{2\beta\varepsilon}{\delta})}+\frac{2\beta}{1-\frac{2\beta\varepsilon}{\delta}}\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{T+1}.
\end{aligned}\tag{63}$$

■

Theorem 4. *We assume that Ω is a convex set, and that for every $z\in\mathcal{Z}$, ℓ_z is convex and β-smooth. Let $\mathbf{w}_\star$ be a solution of ($\mathcal{P}$) such that $\forall z\in\mathcal{Z},\ \ell_z(\mathbf{w}_\star)\leq\varepsilon$, and suppose that $\delta>2\beta\varepsilon$. Further assume that $\eta\leq\frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t\right)-f_\star\leq\frac{\|\mathbf{w}_0-\mathbf{w}_\star\|^2}{\eta(T+1)}+\frac{\delta}{2\beta}+\varepsilon.\tag{64}$$

Proof :

Similarly to the beginning of previous proofs, we have that:

$$\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2\leq\|\mathbf{w}_t-\mathbf{w}_\star\|^2-\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))+\gamma_t\ell_{z_t}(\mathbf{w}_\star)\tag{65}$$

As previously, we lower bound $\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))$ and upper bound $\gamma_t\ell_{z_t}(\mathbf{w}_\star)$ individually.

We begin with $\gamma_t(\ell_{z_t}(\mathbf{w}_t)-\ell_{z_t}(\mathbf{w}_\star))$. We remark that either $\gamma_t=\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla\ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$ or $\gamma_t=\eta$.

Suppose $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$ and $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$. First we write:

$$\begin{aligned}
\gamma_t &= \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&= \frac{\ell_{z_t}(\mathbf{w}_t) + \frac{\delta}{2\beta}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} - \frac{\frac{\delta}{2\beta}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&\geq \frac{\frac{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2}{2\beta} + \frac{\delta}{2\beta}}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} - \frac{\delta}{2\beta}\frac{1}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \quad \text{(Lemma 1)} \\
&= \frac{1}{2\beta} - \frac{\delta}{2\beta}\frac{1}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&\geq \eta - \frac{\delta}{2\beta}\frac{1}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \quad (\eta \leq \frac{1}{2\beta})
\end{aligned} \tag{66}$$

Since $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$, this yields:

$$\begin{aligned}
\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) &\geq \left(\eta - \frac{\delta}{2\beta}\frac{1}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta}\right)(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \\
&= \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{2\beta}\frac{\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&\geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\delta}{2\beta}\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2 + \delta} \\
&\qquad \text{(because } \ell_{z_t}(\mathbf{w}_\star) \geq 0\text{)}
\end{aligned} \tag{67}$$

We now notice that since $\gamma_t = \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2+\delta}$, and $\gamma_t \leq \eta$, then necessarily $\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2+\delta} \leq \eta$. This gives:

$$\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\eta\delta}{2\beta} \tag{68}$$

Now suppose $\gamma_t = \eta$ and $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \geq 0$. Then we have:

$$\begin{aligned}
\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) &= \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \\
&\geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\eta\delta}{2\beta}.
\end{aligned} \tag{69}$$

Now suppose $\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0$. Since $\gamma_t \leq \eta$ by definition, we have that:

$$\begin{aligned}
\gamma_t\left(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)\right) &\geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \quad (\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star) \leq 0) \\
&\geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\eta\delta}{2\beta}.
\end{aligned} \tag{70}$$

In conclusion, in all cases, it holds true that:

$$\gamma_t(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) \geq \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) - \frac{\eta\delta}{2\beta} \tag{71}$$

We upper bound $\gamma_t \ell_{z_t}(\mathbf{w}_\star)$ as follows:

$$\begin{aligned}
\gamma_t \ell_{z_t}(\mathbf{w}_\star) &\leq \eta \ell_{z_t}(\mathbf{w}_\star) \quad (\ell_{z_t}(\mathbf{w}_\star) \geq 0) \\
&\leq \eta\varepsilon \quad \text{(definition of } \varepsilon\text{)}
\end{aligned} \tag{72}$$

We combine inequalities (65), (71) and (72) and obtain:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\eta\delta}{2\beta} + \eta\varepsilon. \tag{73}$$

By taking the expectation and using a telescopic sum, we obtain:

$$0 \leq \|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2 \leq \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 - \sum_{t=0}^{T} \left(\eta(f(\mathbf{w}_t) - f_\star) + \frac{\eta\delta}{2\beta} + \eta\varepsilon \right). \tag{74}$$

Re-arranging and using the convexity of f, we finally obtain:

$$f\left(\frac{1}{T+1} \sum_{t=0}^{T} \mathbf{w}_t \right) \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_\star\|^2}{\eta(T+1)} + \frac{\delta}{2\beta} + \varepsilon. \tag{75}$$

■

3.3. Smooth and Strongly Convex Functions

Finally, we consider the α-strongly convex and β-smooth case. Again, our proof yields a natural separation between $\eta \geq \frac{1}{2\beta}$ and $\eta \leq \frac{1}{2\beta}$.

Lemma 3. *Let $z \in \mathcal{Z}$. Assume that ℓ_z is α-strongly convex, non-negative on $\mathbb{R}^p$, and such that $\inf \ell_z \leq \varepsilon$. In addition, suppose that $\delta \geq 2\alpha\varepsilon$. Then we have:*

$$\forall\, \mathbf{w} \in \mathbb{R}^p,\ \frac{\ell_z(\mathbf{w})}{\|\nabla \ell_z(\mathbf{w})\|^2 + \delta} \leq \frac{1}{2\alpha}. \tag{76}$$

Proof :

Let $\mathbf{w} \in \mathbb{R}^p$ and suppose that ℓ_z reaches its minimum at $\underline{\mathbf{w}} \in \mathbb{R}^p$ (this minimum exists because of strong convexity). By definition of strong convexity, we have that:

$$\forall\, \hat{\mathbf{w}} \in \mathbb{R}^p,\ \ell_z(\hat{\mathbf{w}}) \geq \ell_z(\mathbf{w}) + \nabla \ell_z(\mathbf{w})^\top (\hat{\mathbf{w}} - \mathbf{w}) + \frac{\alpha}{2}\|\hat{\mathbf{w}} - \mathbf{w}\|^2 \tag{77}$$

We minimize the right hand-side over $\hat{\mathbf{w}}$, which gives:

$$\begin{aligned} \forall \hat{\mathbf{w}} \in \mathbb{R}^p,\ \ell_z(\hat{\mathbf{w}}) &\geq \ell_z(\mathbf{w}) + \nabla \ell_z(\mathbf{w})^\top (\hat{\mathbf{w}} - \mathbf{w}) + \frac{\alpha}{2}\|\hat{\mathbf{w}} - \mathbf{w}\|^2 \\ &\geq \ell_z(\mathbf{w}) - \frac{1}{2\alpha}\|\nabla \ell_z(\mathbf{w})\|^2 \end{aligned} \tag{78}$$

Thus by choosing $\hat{\mathbf{w}} = \underline{\mathbf{w}}$ and re-ordering, we obtain the following result (a.k.a. the Polyak-Lojasiewicz inequality):

$$\ell_z(\mathbf{w}) - \ell_z(\underline{\mathbf{w}}) \leq \frac{1}{2\alpha}\|\nabla \ell_z(\mathbf{w})\|^2 \tag{79}$$

Therefore we can write:

$$\frac{\ell_z(\mathbf{w})}{\|\nabla \ell_z(\mathbf{w})\|^2 + \delta} \leq \frac{\ell_z(\mathbf{w}) - \ell_z(\underline{\mathbf{w}}) + \varepsilon}{\|\nabla \ell_z(\mathbf{w})\|^2 + \delta} \leq \frac{\frac{1}{2\alpha}\|\nabla \ell_z(\mathbf{w})\|^2 + \varepsilon}{\|\nabla \ell_z(\mathbf{w})\|^2 + \delta}. \tag{80}$$

We introduce the function $\psi : x \in \mathbb{R}^+ \mapsto \dfrac{\frac{1}{2\alpha}x + \varepsilon}{x + \delta}$, and we compute its derivative:

$$\begin{aligned} \psi'(x) &= \frac{\frac{1}{2\alpha}(x+\delta) - \frac{1}{2\alpha}x - \varepsilon}{(x+\delta)^2}, \\ &= \frac{\frac{\delta}{2\alpha} - \varepsilon}{(x+\delta)^2} \geq 0. \quad (\delta \geq 2\alpha\varepsilon) \end{aligned} \tag{81}$$

Therefore ψ is monotonically increasing. As a result, we have:

$$\forall\, x \in \mathbb{R}^+,\ \psi(x) \leq \lim_{x\to\infty} \psi(x) = \frac{1}{2\alpha}. \tag{82}$$

Therefore we have that:

$$\frac{\frac{1}{2\alpha}\|\nabla \ell_z(\mathbf{w})\|^2 + \varepsilon}{\|\nabla \ell_z(\mathbf{w})\|^2 + \delta} = \psi\left(\|\nabla \ell_z(\mathbf{w})\|^2\right) \leq \frac{1}{2\alpha}, \tag{83}$$

which concludes the proof. ∎

Lemma 4. *For any $a, b \in \mathbb{R}^p$, we have that:*

$$\|a\|^2 + \|b\|^2 \geq \frac{1}{2}\|a-b\|^2 \tag{84}$$

Proof : This is a simple application of the parallelogram law, but we give the proof here for completeness.

$$\begin{aligned}
\|a\|^2 + \|b\|^2 - \frac{1}{2}\|a-b\|^2 &= \|a\|^2 + \|b\|^2 - \frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2 + a^\top b \\
&= \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2 + a^\top b \\
&= \frac{1}{2}\|a+b\|^2 \\
&\geq 0
\end{aligned}$$

∎

Lemma 5. *Let $z \in \mathcal{Z}$. Assume that ℓ_z is α-strongly convex and achieves its (possibly constrained) minimum at $\mathbf{w}_\star \in \Omega$. Then we have:*

$$\forall\, \mathbf{w} \in \Omega,\ \ell_z(\mathbf{w}) - \ell_z(\mathbf{w}_\star) \geq \frac{\alpha}{2}\|\mathbf{w} - \mathbf{w}_\star\|^2 \tag{85}$$

Proof : By definition of strong-convexity (Bubeck, 2015), we have:

$$\forall\, \mathbf{w} \in \Omega,\ \ell_z(\mathbf{w}) - \ell_z(\mathbf{w}_\star) - \nabla \ell_z(\mathbf{w}_\star)^\top(\mathbf{w} - \mathbf{w}_\star) \geq \frac{\alpha}{2}\|\mathbf{w} - \mathbf{w}_\star\|^2. \tag{86}$$

In addition, since $\mathbf{w}_\star$ minimizes ℓ_z, then necessarily:

$$\forall\, \mathbf{w} \in \Omega,\ \nabla \ell_z(\mathbf{w}_\star)^\top(\mathbf{w} - \mathbf{w}_\star) \geq 0. \tag{87}$$

Combining the two equations gives the desired result. ∎

Theorem 5. *We assume that Ω is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is α-strongly convex and β-smooth. Let $\mathbf{w}_\star$ be a solution of ($\mathcal{P}$) such that $\forall z \in \mathcal{Z},\ \ell_z(\mathbf{w}_\star) \leq \varepsilon$, and suppose that $\delta > 2\beta\varepsilon$. Further assume that $\eta \geq \frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$\mathbb{E}[f(\mathbf{w}_{T+1})] - f_\star \leq \beta \exp\left(-\frac{\alpha t}{4\beta}\right)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{\delta}{\alpha} + 2\frac{\beta}{\alpha}\varepsilon + 2\frac{\beta^2}{\alpha^2}\varepsilon. \tag{88}$$

Proof :

We condition the update on z_t drawn at random. The beginning of the proof is identical to that of Theorem 3 (and in particular requires $\delta > 2\beta\varepsilon$). In addition, we remark that $\delta > 2\beta\varepsilon \geq 2\alpha\varepsilon$, because it always holds true that $\beta \geq \alpha$. Combining inequalities (15) and (58), we obtain:

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \gamma_t \ell_{z_t}(\mathbf{w}_\star), \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \gamma_t \varepsilon, \quad \text{(definition of } \varepsilon\text{)} \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{1}{2\beta}(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2} + \frac{\varepsilon}{2\alpha}. \quad \text{(Lemma 3)}
\end{aligned} \tag{89}$$

Taking the expectation over z_t, we obtain:

$$\begin{aligned}\mathbb{E}_{z_t}[\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2] &\leq \|\mathbf{w}_t-\mathbf{w}_\star\|^2 - \frac{1}{2\beta}(f(\mathbf{w}_t)-f(\mathbf{w}_\star)) + \frac{\delta}{4\beta^2}+\frac{\varepsilon}{2\alpha},\\ &\leq \|\mathbf{w}_t-\mathbf{w}_\star\|^2 - \frac{\alpha}{4\beta}\|\mathbf{w}_t-\mathbf{w}_\star\|^2 + \frac{\delta}{4\beta^2}+\frac{\varepsilon}{2\alpha}. \quad \text{(by lemma 5)}\end{aligned}$$

We use a trivial induction over t and write:

$$\begin{aligned}&\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2]\\ &\quad\leq \left(1-\frac{\alpha}{4\beta}\right)\mathbb{E}[\|\mathbf{w}_t-\mathbf{w}_\star\|^2] + \frac{\delta}{4\beta^2}+\frac{\varepsilon}{2\alpha},\\ &\quad\leq \left(1-\frac{\alpha}{4\beta}\right)^t\|\mathbf{w}_0-\mathbf{w}_\star\|^2 + \sum_{k=0}^{t}\left(1-\frac{\alpha}{4\beta}\right)^{t-k}\left(\frac{\delta}{4\beta^2}+\frac{\varepsilon}{2\alpha}\right),\\ &\quad\leq \left(1-\frac{\alpha}{4\beta}\right)^t\|\mathbf{w}_0-\mathbf{w}_\star\|^2 + \sum_{k=0}^{\infty}\left(1-\frac{\alpha}{4\beta}\right)^{k}\left(\frac{\delta}{4\beta^2}+\frac{\varepsilon}{2\alpha}\right),\\ &\quad= \left(1-\frac{\alpha}{4\beta}\right)^t\|\mathbf{w}_0-\mathbf{w}_\star\|^2 + \frac{1}{\frac{\alpha}{4\beta}}\left(\frac{\delta}{4\beta^2}+\frac{\varepsilon}{2\alpha}\right),\\ &\quad= \left(1-\frac{\alpha}{4\beta}\right)^t\|\mathbf{w}_0-\mathbf{w}_\star\|^2 + \frac{4\beta}{\alpha}\left(\frac{\delta}{4\beta^2}+\frac{\varepsilon}{2\alpha}\right).\end{aligned} \tag{90}$$

Given an arbitrary $\mathbf{w}\in\mathbb{R}^p$, we now wish to relate the distance $\|\mathbf{w}-\mathbf{w}_\star\|^2$ to the function values $f(\mathbf{w})-f(\mathbf{w}_\star)$.

Since each ℓ_z is α-strongly convex and β-smooth, so is $f=\mathbb{E}_z[\ell_z]$. We introduce $\underline{\mathbf{w}}$ the minimizer of f on its unconstrained domain $\mathbb{R}^p$. Then we can write that for any $\mathbf{w}\in\mathbb{R}^p$:

$$\begin{aligned}&f(\mathbf{w})-f(\mathbf{w}_\star)\\ &\quad\leq f(\mathbf{w})-f(\underline{\mathbf{w}}), \quad (f(\underline{\mathbf{w}})\leq f(\mathbf{w}_\star))\\ &\quad\leq \nabla f(\underline{\mathbf{w}})^\top(\mathbf{w}-\underline{\mathbf{w}}) + \frac{\beta}{2}\|\mathbf{w}-\underline{\mathbf{w}}\|^2, \quad (f \text{ is } \beta\text{-smooth})\\ &\quad= \frac{\beta}{2}\|\mathbf{w}-\underline{\mathbf{w}}\|^2, \quad (\nabla f(\underline{\mathbf{w}})=\mathbf{0})\\ &\quad\leq \beta(\|\mathbf{w}-\mathbf{w}_\star\|^2+\|\mathbf{w}_\star-\underline{\mathbf{w}}\|^2), \quad \text{(Lemma 4)}\\ &\quad\leq \beta\|\mathbf{w}-\mathbf{w}_\star\|^2 + \frac{2\beta}{\alpha}\left(f(\mathbf{w}_\star)-f(\underline{\mathbf{w}})\right), \quad (f \text{ is } \alpha\text{-strongly convex})\\ &\quad\leq \beta\|\mathbf{w}-\mathbf{w}_\star\|^2 + \frac{2\beta}{\alpha}f(\mathbf{w}_\star), \quad (0\leq f(\underline{\mathbf{w}}))\\ &\quad\leq \beta\|\mathbf{w}-\mathbf{w}_\star\|^2 + 2\frac{\beta\varepsilon}{\alpha}, \quad \text{(definition of } \varepsilon)\end{aligned} \tag{91}$$

Taking the expectation, we can combine the results to obtain the final result:

$$\begin{aligned}\mathbb{E}[f(\mathbf{w}_{t+1})]-f(\mathbf{w}_\star) &\leq \beta\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_\star\|^2] + 2\frac{\beta\varepsilon}{\alpha},\\ &\leq \beta\left(\left(1-\frac{\alpha}{4\beta}\right)^t\|\mathbf{w}_0-\mathbf{w}_\star\|^2 + \frac{4\beta}{\alpha}\left(\frac{\delta}{4\beta^2}+\frac{\varepsilon}{2\alpha}\right)\right) + 2\frac{\beta\varepsilon}{\alpha},\\ &= \beta\left(1-\frac{\alpha}{4\beta}\right)^t\|\mathbf{w}_0-\mathbf{w}_\star\|^2 + \frac{4\beta}{\alpha}\left(\frac{\delta}{4\beta}+\frac{\varepsilon\beta}{2\alpha}\right) + 2\frac{\beta\varepsilon}{\alpha},\\ &= \beta\left(1-\frac{\alpha}{4\beta}\right)^t\|\mathbf{w}_0-\mathbf{w}_\star\|^2 + \frac{\delta}{\alpha} + 2\frac{\beta}{\alpha}\varepsilon + 2\frac{\beta^2}{\alpha^2}\varepsilon,\end{aligned}$$

$$\leq \beta \exp\left(-\frac{\alpha t}{4\beta}\right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{\delta}{\alpha} + 2\frac{\beta}{\alpha}\varepsilon + 2\frac{\beta^2}{\alpha^2}\varepsilon.$$

■

Theorem 6. *We assume that Ω is a convex set, and that for every $z \in \mathcal{Z}$, ℓ_z is α-strongly convex and β-smooth. Let $\mathbf{w}_\star$ be a solution of ($\mathcal{P}$) such that $\forall z \in \mathcal{Z},\ \ell_z(\mathbf{w}_\star) \leq \varepsilon$, and suppose that $\delta > 2\beta\varepsilon$. Further assume that $\eta \leq \frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$\mathbb{E}[f(\mathbf{w}_{T+1})] - f_\star \leq \beta \exp\left(\frac{-\alpha\eta T}{2}\right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{\delta}{\alpha} + \frac{4\varepsilon\beta}{\alpha}. \tag{92}$$

Proof : Re-using inequalities (65) and (71) from the proof of Theorem 4, we can write:

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\eta\delta}{2\beta} + \gamma_t \ell_{z_t}(\mathbf{w}_\star), \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(\ell_{z_t}(\mathbf{w}_t) - \ell_{z_t}(\mathbf{w}_\star)) + \frac{\eta\delta}{2\beta} + \eta\varepsilon \\
&\qquad \text{(using } \gamma_t \leq \eta,\ 0 \leq \ell_{z_t}(\mathbf{w}_\star) \leq \varepsilon\text{)}.
\end{aligned} \tag{93}$$

Taking the expectation over z_t, we obtain:

$$\mathbb{E}_{z_t}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2] \leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \eta(f(\mathbf{w}_t) - f(\mathbf{w}_\star)) + \frac{\eta\delta}{2\beta} + \eta\varepsilon. \tag{94}$$

Therefore, we can write:

$$\begin{aligned}
\mathbb{E}_{z_t}[\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2] &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \frac{\alpha\eta}{2}\|\mathbf{w}_t - \mathbf{w}_\star\|^2 + \frac{\eta\delta}{2\beta} + \eta\varepsilon, \quad \text{(Lemma 5)} \\
&= \left(1 - \frac{\alpha\eta}{2}\right) \|\mathbf{w}_t - \mathbf{w}_\star\|^2 + \frac{\eta\delta}{2\beta} + \eta\varepsilon.
\end{aligned} \tag{95}$$

Then a trivial induction gives that:

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2] &\leq \left(1 - \frac{\alpha\eta}{2}\right)^T \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \left(\frac{\eta\delta}{2\beta} + \eta\varepsilon\right) \sum_{t=0}^{T} \left(1 - \frac{\alpha\eta}{2}\right)^t, \\
&\leq \left(1 - \frac{\alpha\eta}{2}\right)^T \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \left(\frac{\eta\delta}{2\beta} + \eta\varepsilon\right) \sum_{t=0}^{\infty} \left(1 - \frac{\alpha\eta}{2}\right)^t, \\
&= \left(1 - \frac{\alpha\eta}{2}\right)^T \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \left(\frac{\eta\delta}{2\beta} + \eta\varepsilon\right) \frac{1}{1 - \left(1 - \frac{\alpha\eta}{2}\right)}, \\
&= \left(1 - \frac{\alpha\eta}{2}\right)^T \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{\delta}{\alpha\beta} + \frac{2\varepsilon}{\alpha}.
\end{aligned} \tag{96}$$

We now re-use the inequality (91) in expectation to write:

$$\begin{aligned}
\mathbb{E}[f(\mathbf{w}_{T+1})] - f_\star &\leq \beta\mathbb{E}[\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2] + \frac{2\beta\varepsilon}{\alpha}, \\
&\leq \beta \left(1 - \frac{\alpha\eta}{2}\right)^T \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{\delta}{\alpha} + \frac{4\varepsilon\beta}{\alpha}, \\
&\leq \beta \exp\left(\frac{-\alpha\eta T}{2}\right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2 + \frac{\delta}{\alpha} + \frac{4\varepsilon\beta}{\alpha}.
\end{aligned} \tag{97}$$

■

4. Detailed Non-Convex Results

The Restricted Secant Inequality (RSI) is a milder assumption than convexity. It can be defined as follows:

Definition 1. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a lower-bounded differentiable function achieving its minimum at $\mathbf{w}_\star$. We say that f satisfies the RSI if there exists $\alpha > 0$ such that:*

$$\forall \boldsymbol{w} \in \mathbb{R}^p,\ \nabla f(\mathbf{w})^\top (\boldsymbol{w} - \mathbf{w}_\star) \geq \alpha \|\boldsymbol{w} - \mathbf{w}_\star\|^2. \tag{98}$$

The RSI is sometimes used to prove convergence of optimization algorithms without assuming convexity (Vaswani et al., 2019).

As we prove below, the Polyak step-size may fail to converge under the RSI assumption, even in a non-stochastic setting with the exact minimum known.

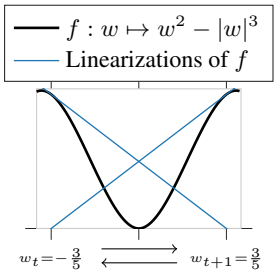

Figure 1. Illustration of the function f, which satisfies the RSI. When starting at $w = -3/5$, gradient descent with the Polyak step-size oscillates between $w = -3/5$ and $w = 3/5$.

Proposition 3. *Let $f : w \in [\frac{-3}{5}; \frac{3}{5}] \mapsto w^2 - |w|^3$. Then f satisfies the RSI with $\alpha = \frac{1}{5}$.*

Proof : First we note that f achieves its minimum at $w_\star = 0$, and that $f(w_\star) = 0$. In addition, we introduce the sign function $\sigma(w)$, which is equal to 1 if $w \geq 0$, and -1 otherwise. Now let $w \in [\frac{-3}{5}; \frac{3}{5}]$. Then we have that:

$$\begin{aligned}
\nabla f(w)(w - w_\star) - \frac{1}{5}(w - w_\star)^2, &= (2w - 3\sigma(w)w^2)(w - 0) - \frac{1}{5}(w - 0)^2, \\
&= \frac{9}{5}w^2 - 3\sigma(w)w^3, \\
&= 3w^2(\frac{3}{5} - \sigma(w)w), \\
&\geq 0.
\end{aligned} \tag{99}$$

■

Proposition 4. *Assume that we apply the Polyak step-size to $f : w \in [\frac{-3}{5}; \frac{3}{5}] \mapsto w^2 - |w|^3$, starting from the initial point $w_0 = -3/5$. Then the iterates oscillate between $-3/5$ and $3/5$.*

Proof : We show that, starting with $w_0 = -\frac{3}{5}$, we obtain $w_1 = \frac{3}{5}$. This will prove oscillation of the iterates by symmetry of the problem. Since $w_0 = \frac{-3}{5}$, we have $f(w_0) = \frac{9}{25} - \frac{27}{125} = \frac{18}{125}$. Furthermore, $\nabla f(w_0) = 2(\frac{-3}{5}) + 3(\frac{9}{25}) = \frac{-3}{25}$. Therefore:

$$\begin{aligned}
w_1 &= w_0 - \frac{f(w_0)}{(\nabla f(w_0))^2}\nabla f(w_0), \\
&= w_0 - \frac{f(w_0)}{\nabla f(w_0)}, \\
&= \frac{-3}{5} + \frac{\frac{18}{125}}{\frac{3}{25}}, \\
&= \frac{-3}{5} + \frac{6}{5}, \\
&= \frac{3}{5}.
\end{aligned} \tag{100}$$

■

Theorem 7. *We assume that $\Omega = \mathbb{R}^p$, and that for every $z \in \mathcal{Z}$, ℓ_z is β-smooth and satisfies the RSI with constant α. We further assume that there exists $\mathbf{w}_\star$ a solution of ($\mathcal{P}$) such that $\forall z \in \mathcal{Z}$, $\ell_z(\mathbf{w}_\star) = 0$. Let η be such that $\frac{1}{2\beta} \leq \eta \leq \frac{2\alpha}{\beta^2}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f(\mathbf{w}_{T+1}) - f_\star \leq \frac{\beta}{2}\exp\left(\left(-\frac{\alpha}{\beta} + \frac{\eta\beta}{2}\right)T\right)\|\mathbf{w}_0 - \mathbf{w}_\star\|^2. \tag{101}$$

Note: this result assumes perfect interpolation, and thus we set $\delta = 0$ (no small constant for numerical stability).

Proof : We consider the update at time t, which we condition on the draw of $z_t \in \mathcal{Z}$. Since we consider $\delta = 0$, we have $\gamma_t = \min\left\{\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2}, \eta\right\}$. We suppose $\nabla \ell_{z_t}(\mathbf{w}_t) \neq \mathbf{0}$.

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &= \|\Pi_\Omega(\mathbf{w}_t - \gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)) - \mathbf{w}_\star\|^2, \\
&\leq \|\mathbf{w}_t - \gamma_t \nabla \ell_{z_t}(\mathbf{w}_t) - \mathbf{w}_\star\|^2, \quad (\Pi_\Omega \text{ projection}) \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{w}_\star) + \gamma_t^2 \|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2, \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{w}_\star) + \gamma_t \ell_{z_t}(\mathbf{w}_t), \quad (\text{since } \gamma_t \leq \frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2}) \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\gamma_t \nabla \ell_{z_t}(\mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{w}_\star) + \gamma_t \frac{\beta}{2} \|\mathbf{w}_t - \mathbf{w}_\star\|^2, \quad (\text{Lemma 3.4 of (Bubeck, 2015)}) \\
&\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\gamma_t \alpha \|\mathbf{w}_t - \mathbf{w}_\star\|^2 + \gamma_t \frac{\beta}{2} \|\mathbf{w}_t - \mathbf{w}_\star\|^2, \quad (\text{RSI inequality}) \\
&= \left(1 - 2\gamma_t \alpha + \gamma_t \frac{\beta}{2}\right) \|\mathbf{w}_t - \mathbf{w}_\star\|^2.
\end{aligned} \tag{102}$$

Since we know that $\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2} \geq \frac{1}{2\beta}$ (Lemma 1) and $\eta \geq \frac{1}{2\beta}$, we have that $\gamma_t \geq \frac{1}{2\beta}$. Then, using both $\gamma_t \geq \frac{1}{2\beta}$ and $\gamma_t \leq \eta$, we can write:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \left(1 - \frac{\alpha}{\beta} + \frac{\eta\beta}{2}\right) \|\mathbf{w}_t - \mathbf{w}_\star\|^2. \tag{103}$$

With a trivial induction we obtain:

$$\begin{aligned}
\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2 &\leq \left(1 - \frac{\alpha}{\beta} + \frac{\eta\beta}{2}\right)^T \|\mathbf{w}_0 - \mathbf{w}_\star\|^2, \\
&\leq \exp\left(\left(-\frac{\alpha}{\beta} + \frac{\eta\beta}{2}\right) T\right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2.
\end{aligned} \tag{104}$$

Since f is β-smooth and the problem is unconstrained by assumption, we have $f(\mathbf{w}_{T+1}) \leq \frac{\beta}{2}\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2$ (by Lemma 3.4 of (Bubeck, 2015)), and we obtain the desired result. ∎

Theorem 8. *We assume that $\Omega = \mathbb{R}^p$, and that for every $z \in \mathcal{Z}$, ℓ_z is β-smooth and satisfies the RSI with constant α. We further assume that there exists $\mathbf{w}_\star$ a solution of ($\mathcal{P}$) such that $\forall z \in \mathcal{Z}$, $\ell_z(\mathbf{w}_\star) = 0$. Let η be such that $0 < \eta \leq \frac{1}{2\beta}$. Then if we apply ALI-G with a maximal learning-rate of η to f, we have:*

$$f(\mathbf{w}_{T+1}) - f_\star \leq \frac{\beta}{2} \exp\left(\left(-\eta\left(2\alpha - \frac{\beta}{2}\right)\right) T\right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2. \tag{105}$$

Note: this result assumes perfect interpolation, and thus we set $\delta = 0$ (no small constant for numerical stability).

Proof : We consider the update at time t, which we condition on the draw of $z_t \in \mathcal{Z}$. Since we consider $\delta = 0$, we have $\gamma_t = \min\left\{\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2}, \eta\right\}$. We suppose $\nabla \ell_{z_t}(\mathbf{w}_t) \neq \mathbf{0}$. We re-use equation (102) to write:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \left(1 - 2\gamma_t \alpha + \gamma_t \frac{\beta}{2}\right) \|\mathbf{w}_t - \mathbf{w}_\star\|^2. \tag{106}$$

Since we know that $\frac{\ell_{z_t}(\mathbf{w}_t)}{\|\nabla \ell_{z_t}(\mathbf{w}_t)\|^2} \geq \frac{1}{2\beta}$ (Lemma 1) and $\eta \leq \frac{1}{2\beta}$, we have that $\gamma_t = \eta$ necessarily. Thus we obtain:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 \leq \left(1 - 2\eta\alpha + \eta \frac{\beta}{2}\right) \|\mathbf{w}_t - \mathbf{w}_\star\|^2.$$

With a trivial induction we obtain:

$$\begin{aligned}
\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2 &\leq \left(1 - \eta\left(2\alpha - \frac{\beta}{2}\right)\right)^T \|\mathbf{w}_0 - \mathbf{w}_\star\|^2, \\
&\leq \exp\left(\left(-\eta\left(2\alpha - \frac{\beta}{2}\right)\right) T\right) \|\mathbf{w}_0 - \mathbf{w}_\star\|^2.
\end{aligned}$$

Since f is β-smooth and the problem is unconstrained by assumption, we have $f(\mathbf{w}_{T+1}) \leq \frac{\beta}{2}\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2$ (by Lemma 3.4 of (Bubeck, 2015)), and we obtain the desired result. ∎

5. Additional Experimental Details

5.1. Standard Deviation of CIFAR Results

Task	Optimizer	Avg	Std
DN10	ADAMW	92.6	0.08
DN10	ALIG	95.0	0.16
DN10	AMSGRAD	91.7	0.25
DN10	DFW	94.6	0.22
DN10	L4ADAM	90.8	0.09
DN10	L4MOM	91.9	0.17
DN10	SGD	95.1	0.21
DN10	YOGI	92.1	0.38
DN100	ADAMW	69.5	0.54
DN100	ALIG	76.3	0.14
DN100	AMSGRAD	69.4	0.41
DN100	DFW	73.2	0.29
DN100	L4ADAM	60.5	0.64
DN100	L4MOM	62.6	1.98
DN100	SGD	76.3	0.22
DN100	YOGI	69.6	0.34
WRN10	ADAMW	92.1	0.34
WRN10	ALIG	95.2	0.09
WRN10	AMSGRAD	90.8	0.31
WRN10	DFW	94.2	0.19
WRN10	L4ADAM	90.5	0.09
WRN10	L4MOM	91.6	0.24
WRN10	SGD	95.3	0.31
WRN10	YOGI	91.2	0.27
WRN100	ADAMW	69.6	0.51
WRN100	ALIG	75.8	0.29
WRN100	AMSGRAD	68.7	0.70
WRN100	DFW	76.0	0.24
WRN100	L4ADAM	61.7	2.17
WRN100	L4MOM	61.4	0.86
WRN100	SGD	77.8	0.13
WRN100	YOGI	68.7	0.47

Table 3. Test Accuracy (%) on CIFAR including standard deviations. Each experiment was run three times.

5.2. Additional Details About Training Protocol on ImageNet

Data Processing. We use 1.23M images for training. As mentioned in the paper, we do not use any data augmentation on this task. Our data processing can be described as follows. Each training image is resized so that its smaller dimension is of 224 pixels, after which we take a centered square crop of 224 by 224. The cropped image is then centered and normalized per channel (for this, the mean and standard deviation per channel is computed across all training images), before being fed to the neural network.

Loss Function. We use the top-k truncated cross-entropy (Lapin et al., 2016) as our loss function for training the model on ImageNet. In particular, we use $k = 5$ so that we optimize for the commonly used top-5 error, and we use the default temperature parameter $\tau = 1$.

Our PyTorch code re-uses the implementation from `https://github.com/locuslab/lml`.

References

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 2015.

Hazan, E. and Kakade, S. Revisiting the polyak step size. *arXiv preprint*, 2019.

Lapin, M., Hein, M., and Schiele, B. Loss functions for top-k error: Analysis and insights. *Conference on Computer Vision and Pattern Recognition*, 2016.

Vaswani, S., Mishkin, A., Laradji, I., Schmidt, M., Gidel, G., and Lacoste-Julien, S. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *arXiv preprint*, 2019.