SUPPLEMENT FOR "ACCELERATED GRADIENT-FREE METHOD FOR HEAVILY CONSTRAINED NONCONVEX OPTIMIZATION"

Anonymous authors
Paper under double-blind review

A PROOF OF PROPOSITION 1

Proof 1 *Since the* $(\boldsymbol{w}^*, \boldsymbol{p}^*)$ *is the* ϵ*-stationary point of* $\min_{\boldsymbol{w}} \max_{\boldsymbol{p}\in\Delta^m} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})$*, then we have*

$$\|\nabla_{\boldsymbol{w}} f_0(\boldsymbol{w}^*) + \beta \sum_{j=1}^{m} p_j^* 2\max\{f_j(\boldsymbol{w}^*), 0\}\nabla_{\boldsymbol{w}} f_j(\boldsymbol{w}^*)\|_2^2 \leq \epsilon^2 \tag{1}$$

Let $\alpha_j^* = 2\beta p_j^* \max\{f_j(\boldsymbol{w}^*), 0\}$ *and* $\epsilon \leq \epsilon_1$*, we have*

$$\|\nabla_{\boldsymbol{w}} f_0(\boldsymbol{w}^*) + \sum_{j=1}^{m} \alpha_j^* \nabla_{\boldsymbol{w}} f_j(\boldsymbol{w}^*)\|_2^2 \leq \epsilon_1^2 \tag{2}$$

the first condition in Definition 2 is satisfied.

Using $\|\nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}^*, \boldsymbol{p}^*)\|_2^2 \leq \epsilon$ *and* $0 \leq p_j^2 \leq 1$*, we have*

$$\sum_{j=1}^{m} (\beta\phi_j(\boldsymbol{w}^*) - \lambda p_j^*)^2 \leq \epsilon^2 \tag{3}$$

Using the inequality $\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$*, we have*

$$\begin{aligned}
&\frac{1}{2}\beta^2 \sum_{j=1}^{m} (\max\{f_j(\boldsymbol{w}^*), 0\})^2 \\
\leq &\beta^2 \sum_{j=1}^{m} (\max\{f_j(\boldsymbol{w}^*), 0\} - \lambda p_j^*)^2 + \lambda^2 \sum_{j=1}^{m} (p_j^*)^2 \\
\leq &\epsilon^2 + m\lambda^2
\end{aligned} \tag{4}$$

Rearrange the above inequality and let $\dfrac{2\epsilon^2 + 2m\lambda^2}{\beta^2} \leq \epsilon_2^2$*, we can obtain*

$$\sum_{j=1}^{m} (\max\{f_j(\boldsymbol{w}^*), 0\})^2 \leq \epsilon_2^2 \tag{5}$$

Therefore, the second condition in Definition 2 is satisfied.

Based on the inequality $\|\langle \boldsymbol{a}, \boldsymbol{b}\rangle\|_2^2 \leq \|\boldsymbol{a}\|_2^2 \|\boldsymbol{b}\|_2^2$*, we can multiply* $\sum_{j=1}^m (\alpha_j^*)^2$ *on both sides of the inequality 5, such that we have*

$$\left(\sum_{j=1}^{m} \alpha_j^* \max\{f_j(\boldsymbol{w}^*), 0\}\right)^2 \leq \sum_{j=1}^{m} (\alpha_j^*)^2 \sum_{j=1}^{m} \max\{f_j(\boldsymbol{w}^*), 0\}^2 \leq \epsilon_2^2 \sum_{j=1}^{m} (\alpha_j^*)^2 \tag{6}$$

Since $\alpha_j^* \geq 0$ *and* $\max\{f_j(\boldsymbol{w}^*), 0\} \geq 0$

$$\sum_{j=1}^{m} (\alpha_j^* \max\{f_j(\boldsymbol{w}^*), 0\})^2 \leq \left(\sum_{j=1}^{m} \alpha_j^* \max\{f_j(\boldsymbol{w}^*), 0\}\right)^2 \leq \epsilon_2^2 \sum_{j=1}^{m} (\alpha_j^*)^2 \tag{7}$$

Using inequality 5, we have $(\alpha_j^*)^2 = 4\beta^2(p_j^*)^2(\max\{f_j(\boldsymbol{w}^*), 0\})^2 \leq 4\beta^2\epsilon_2^2$*, Let* $4\beta^2\epsilon_2^2 \leq \epsilon_3^2$*, we have*

$$\sum_{j=1}^{m}(\alpha_j^* \max\{f_j(\boldsymbol{w}^*), 0\})^2 \leq \epsilon_3^2 \tag{8}$$

If $f_j(\boldsymbol{w}^*) \leq 0$*, we have* $\alpha_j^* = 2\beta p_j^* \max\{f_j(\boldsymbol{w}^*), 0\} = 0$*. Therefore, we have*

$$\sum_{j=1}^{m}(\alpha_j^* f_j(\boldsymbol{w}^*))^2 \leq \epsilon_3^2, \tag{9}$$

which means that the third condition in Definition 2 is satisfied.

That completes the proof.

B Proof of Proposition 2

Proof 2 *Assume that a point* $\hat{\boldsymbol{w}}$ *satisfies that* $\|\nabla_{\boldsymbol{w}} g(\hat{\boldsymbol{w}})\|_2 \leq \epsilon$*, the optimization problem* $\max_{\boldsymbol{p}\in\Delta^m} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{y})$ *is strongly concave w.r.t* $\boldsymbol{p}$ *and* $\boldsymbol{p}^*(\hat{\boldsymbol{w}})$ *is uniquely defined. Solving this this strongly concave problem* $\max_{\boldsymbol{p}\in\Delta^m} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{y})$*, we can obtain a point* $\boldsymbol{p}'$ *satisfying that*

$$\|\nabla_{\boldsymbol{p}} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p}')\|_2 \leq \epsilon \; and \; \|\boldsymbol{p}' - \boldsymbol{p}^*(\hat{\boldsymbol{w}})\|_2 \leq \epsilon. \tag{10}$$

If $\|\nabla_{\boldsymbol{w}} g(\hat{\boldsymbol{w}})\|_2 \leq \epsilon$*, we have*

$$\begin{aligned}
&\|\nabla_{\boldsymbol{w}} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p}')\|_2 \\
\leq &\|\nabla_{\boldsymbol{w}} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p}') - \nabla_{\boldsymbol{w}} g(\hat{\boldsymbol{w}})\|_2 + \|\nabla_{\boldsymbol{w}} g(\hat{\boldsymbol{w}})\|_2 \\
= &\|\nabla_{\boldsymbol{w}} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p}') - \nabla_{\boldsymbol{w}} \mathcal{L}(\hat{\boldsymbol{w}}, \boldsymbol{p}^*(\hat{\boldsymbol{w}}))\|_2 + \epsilon \\
\leq &L\|\boldsymbol{p}' - \boldsymbol{p}^*(\hat{\boldsymbol{w}})\|_2 + \epsilon \\
= &\mathcal{O}(\epsilon)
\end{aligned} \tag{11}$$

C Convergence Analysis of DSZOG

Here, we give several lemmas used in our analysis and then give the proof of theorem 1.

Lemma 1 *(Nesterov & Spokoiny (2017) Lemma 3) Under Assumption 1, it holds that*

$$\|\nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}, \boldsymbol{p}) - \nabla_{\boldsymbol{w}} \mathcal{L}(w, \boldsymbol{p})\|_2^2 \leq \frac{\mu^2 L^2 (d+3)^3}{4} \tag{12}$$

Lemma 2 *(Wang et al. (2020) Lemma 12) Under Assumption 1, for any* $\boldsymbol{w} \in \mathbb{R}^d$*,* $\boldsymbol{p}_1, \boldsymbol{p}_2 \in \Delta^m$*, it holds that*

$$\|\nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}, \boldsymbol{p}_1) - \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}, \boldsymbol{p}_2)\|_2 \leq L^2 \|\boldsymbol{p}_1 - \boldsymbol{p}_2\|_2^2 \tag{13}$$

Lemma 3 *(Nesterov & Spokoiny (2017) Lemma 4) Under Assumption 1, it holds that*

$$\|\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})\|_2^2 \leq 2\|\nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}, \boldsymbol{p})\|_2^2 + \frac{\mu^2 L^2 (d+3)^3}{2} \tag{14}$$

Lemma 4 *(Lin et al. (2020) Lemma 3.3)The function* $g(\cdot) := \max_{\boldsymbol{p}\in\Delta^m} \mathcal{L}(\cdot, \boldsymbol{p})$ *is* $L_g = (L + \kappa L)$*-gradient Lipschitz, and* $\nabla g(\boldsymbol{w}) = \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p}^*(\boldsymbol{w}))$*. Moreover,* $\boldsymbol{p}^*(\boldsymbol{w}) = \arg\max_{\boldsymbol{p}\in\Delta^m} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})$ *is* κ*-Lipschitz.*

Lemma 5 *Under Assumption 1, if* $|\mathcal{M}_1| = |\mathcal{M}_2| = M$*, it holds that* $\mathbb{E}_{\boldsymbol{u}_{[q]}, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}}[\|G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]})\|_2^2] \leq 3\|\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})\|_2^2 + \rho(\epsilon, \mu)$*, where* $\rho(\epsilon, \mu) = \frac{\mu^2 L^2 (d+6)^2 \epsilon^2}{8} + \frac{\mu^2 L^2 (d+3)^3}{2} + \frac{\epsilon^2}{2}$.

Proof 3 *Since* $\mathbb{E}[G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]})] = \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}, \boldsymbol{p})$ *and* $|\mathcal{M}_1| = |\mathcal{M}_2| = M$*, we have*

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{u}_{[q]}, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}}[\|G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]})\|_2^2] \\
=&\frac{1}{qM}\mathbb{E}[\|G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_i, f_j, \boldsymbol{u})\|_2^2] \\
=&\frac{1}{qM}\mathbb{E}[\|G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_i, f_j, \boldsymbol{u}) - \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t)\|_2^2] + \|\nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}, \boldsymbol{p})\|_2^2 \\
\leq&\frac{1}{qM}\mathbb{E}[\|G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_i, f_j, \boldsymbol{u})\|_2^2] + 2\|\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)\|_2^2 + 2\|\nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}, \boldsymbol{p}) - \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})\|_2^2 \\
\leq&\frac{1}{qM}\left[2(d+4)\left[\|\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})\|_2^2 + \sigma_1^2\right] + \frac{\mu^2 L^2 (d+6)^3}{2}\right] + 2\|\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})\|_2^2 + \frac{\mu^2 L^2 (d+3)^3}{2} \\
=&\frac{2(d+4)}{qM}\|\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)\|_2^2 + 2\|\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)\|_2^2 + \frac{2(d+4)\sigma_1^2}{qM} + \frac{\mu^2 L^2 (d+6)^3}{2qM} + \frac{\mu^2 L^2 (d+3)^3}{2}
\end{aligned} \tag{15}
$$

The first inequality is due to $\|a+b\|_2^2 \leq 2\|a|_2^2 + \|b\|_2^2$*. The second inequality is due to the Lemma 14 in Nesterov & Spokoiny (2017) and Lemma 1. Let* $qM = 4(d+6)(\sigma_1^2+1)\epsilon^{-2}$*,then we have*

$$
\mathbb{E}_{\boldsymbol{u}_{[q]}, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}}[\|G_\mu(\boldsymbol{w}, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]})\|_2^2] \leq 3\|\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{p})\|_2^2 + \rho(\epsilon, \mu) \tag{16}
$$

where $\rho(\epsilon, \mu) = \dfrac{\mu^2 L^2 (d+6)^2 \epsilon^2}{8} + \dfrac{\mu^2 L^2 (d+3)^3}{2} + \dfrac{\epsilon^2}{2}$.

That completes the proof.

Lemma 6 *Assume* $\{(\boldsymbol{w}_t, \boldsymbol{p}_t)\}$ *is the sequence generated by our Algorithm 1. By setting* $\eta_{\boldsymbol{p}} = \dfrac{1}{2L}$*, the following inequality holds:* $\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_{t-1}) - \boldsymbol{p}_t\|_2^2] \leq (1 - \dfrac{1}{4\kappa})\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_{t-1}) - \boldsymbol{p}_{t-1}\|_2^2] + \dfrac{\sigma^2}{4L^2}$*, where* $\kappa = \dfrac{L}{\tau}$.

Proof 4 *According to the update rules, we have*

$$
\begin{aligned}
&\|\boldsymbol{p}^*(\boldsymbol{w}_{t-1}) - \boldsymbol{p}_t\|_2^2 \\
=&\|\mathrm{Proj}_{\Delta^m}(\boldsymbol{p}_{t-1} + \eta_{\boldsymbol{p}} H(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1}) - \boldsymbol{p}^*(\boldsymbol{w}_{t-1}))\|_2^2 \\
\leq&\|\boldsymbol{p}_{t-1} - \boldsymbol{p}^*(\boldsymbol{w}_{t-1})\|_2^2 + \eta_{\boldsymbol{p}}^2 \|H(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1})\|_2^2 + 2\eta_{\boldsymbol{p}} \langle H(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1}), \boldsymbol{p}_{t-1} - \boldsymbol{p}^*(\boldsymbol{w}_{t-1})\rangle
\end{aligned} \tag{17}
$$

For a given t*, denote by* $\mathbb{E}$ *taking expectation with respect to random random samples conditioned on all previous iterations. By taking expectation to both sides of this inequality, we obtain*

$$
\begin{aligned}
&\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_{t-1}) - \boldsymbol{p}_t\|_2^2] \\
\leq&\mathbb{E}[\|\boldsymbol{p}_{t-1} - \boldsymbol{p}^*(\boldsymbol{w}_{t-1})\|_2^2] + \eta_{\boldsymbol{p}}^2 \mathbb{E}[\|H(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1})\|_2^2] - 2\eta_{\boldsymbol{p}} \langle \mathcal{L}(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1}), \boldsymbol{p}^*(\boldsymbol{w}_{t-1}) - \boldsymbol{p}_{t-1}\rangle \\
\leq&\mathbb{E}[\|\boldsymbol{p}_{t-1} - \boldsymbol{p}^*(\boldsymbol{w}_{t-1})\|_2^2] + \eta_{\boldsymbol{p}}^2 (\|\nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1})\|_2^2 + \sigma_2^2) - 2\eta_{\boldsymbol{p}} \left(\mathcal{L}(\boldsymbol{w}_{t-1}, \boldsymbol{p}^*(\boldsymbol{w}_{t-1})) - \mathcal{L}(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1})\right) \\
\leq&\mathbb{E}[\|\boldsymbol{p}_{t-1} - \boldsymbol{p}^*(\boldsymbol{w}_{t-1})\|_2^2] + \eta_{\boldsymbol{p}}^2 \left(2L(\mathcal{L}(\boldsymbol{w}_{t-1}, \boldsymbol{p}^*(\boldsymbol{w}_{t-1})) - \mathcal{L}(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1})) + \sigma_2^2\right) \\
&- 2\eta_{\boldsymbol{p}} \left(\mathcal{L}(\boldsymbol{w}_{t-1}, \boldsymbol{p}^*(\boldsymbol{w}_{t-1})) - \mathcal{L}(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1})\right) \\
=&\mathbb{E}[\|\boldsymbol{p}_{t-1} - \boldsymbol{p}^*(\boldsymbol{w}_{t-1})\|_2^2] - \frac{1}{2L}\left(\mathcal{L}(\boldsymbol{w}_{t-1}, \boldsymbol{p}^*(\boldsymbol{w}_{t-1})) - \mathcal{L}(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1})\right) + \frac{\sigma_2^2}{4l^2} \\
\leq&\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_{t-1}) - \boldsymbol{p}_{t-1}\|_2^2](1 - \frac{\tau}{4L}) + \frac{\sigma_2^2}{4L^2}
\end{aligned} \tag{18}
$$

The second inequality is due to the cancavity of $\mathcal{L}$ *w.r.t* $\boldsymbol{p}$*. The third inequality is due to the gradient-Lipschitz of* $\mathcal{L}$ *w.r.t* $\boldsymbol{p}$ *and the strongly concave w.r.t* $\boldsymbol{p}$*. The equality is due to* $\eta_{\boldsymbol{p}} = \dfrac{1}{2L}$*. The last is due to strongly concave of* $\mathcal{L}$ *w.r.t* $\boldsymbol{p}$*.*

Lemma 7 *Denote* $\delta_t = \|\boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2$ *and set* $\eta_{\boldsymbol{w}} = \dfrac{1}{4 \times 16^2 \kappa^2 (\kappa+1)^2 (L+1)}$, $\gamma = 1 - \dfrac{1}{4\kappa} + 3\kappa^2 \eta_{\boldsymbol{w}}^2 L^2 \leq 1 - \dfrac{3}{16\kappa} < 1$. *It holds that* $\mathbb{E}[\delta_t] = \gamma^t \mathbb{E}[\delta_0] + 3\kappa^2 \eta_{\boldsymbol{w}}^2 \sum_{i=0}^{t-1} \gamma^{t-1-i} \mathbb{E}[\|\nabla g(\boldsymbol{w}_i)\|_2^2] + \theta_1 \sum_{i=0}^{t-1} \gamma^{t-1-i}$ *where* $\theta_1 = \kappa^2 \eta_{\boldsymbol{w}}^2 \rho(\epsilon, \mu) + \dfrac{\sigma_2^2}{4L^2}$.

Proof 5 *We have*

$$
\begin{aligned}
\mathbb{E}[\delta_t] =& \mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2] \\
\leq& \mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_{t-1}) - \boldsymbol{p}_t\|_2^2] + \mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}^*(\boldsymbol{w}_{t-1})\|_2^2] \\
\leq& (1 - \frac{1}{4\kappa}) \mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_{t-1}) - \boldsymbol{p}_{t-1}\|_2^2] + \frac{\sigma_2^2}{4L^2} + \mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}^*(\boldsymbol{w}_{t-1})\|_2^2] \\
\leq& (1 - \frac{1}{4\kappa}) \mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_{t-1}) - \boldsymbol{p}_{t-1}\|_2^2] + \frac{\sigma_2^2}{4L^2} + \kappa^2 \mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2^2] \\
=& \kappa^2 \eta_{\boldsymbol{w}}^2 \mathbb{E}[\|G_\mu^{\mathcal{L}}(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1}, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]})\|_2^2] + (1 - \frac{1}{4\kappa}) \mathbb{E}[\delta_{t-1}] + \frac{\sigma_2^2}{4L^2}
\end{aligned} \tag{19}
$$

The third inequality is due to Lemma 4. From Lemma 5, we have

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{u}_{[q]}, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}} [\|G_\mu(\boldsymbol{w}_{t-1}, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]})\|_2^2] \\
\leq& 3\mathbb{E}[\|\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_{t-1}, \boldsymbol{p}_{t-1})\|_2^2] + \rho(\epsilon, \mu) \\
\leq& 3\mathbb{E}[\|\nabla g(\boldsymbol{w}_{t-1})\|_2^2] + 3l^2 \mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_{t-1}) - \boldsymbol{p}_{t-1}\|_2^2] + \rho(\epsilon, \mu)
\end{aligned} \tag{20}
$$

Then, combining the above inequalities, we have

$$
\mathbb{E}[\delta_t] = \gamma^t \mathbb{E}[\delta_0] + 3\kappa^2 \eta_{\boldsymbol{w}}^2 \sum_{i=0}^{t-1} \gamma^{t-1-i} \mathbb{E}[\|\nabla g(\boldsymbol{w}_i)\|_2^2] + \theta_1 \sum_{i=0}^{t-1} \gamma^{t-1-i} \tag{21}
$$

where $\gamma = 1 - \dfrac{1}{4\kappa} + 3\kappa^2 \eta_{\boldsymbol{w}}^2 L^2$ *and* $\theta_1 = \kappa^2 \eta_{\boldsymbol{w}}^2 \rho(\epsilon, \mu) + \dfrac{\sigma_2^2}{4L^2}$.

Then, we give the proof of Theorem 1.

Proof 6

$$
\begin{aligned}
& g(\boldsymbol{w}_{t+1}) \\
\leq & g(\boldsymbol{w}_t) - \eta_{\boldsymbol{w}} \langle \nabla g(\boldsymbol{w}_t), G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \rangle + \frac{L_g \eta_{\boldsymbol{w}}^2}{2} \| G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \|_2^2 \\
= & g(\boldsymbol{w}_t) - \eta_{\boldsymbol{w}} \langle \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) - \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) \\
& + \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) - \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t) \\
& + \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t), G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \rangle \\
& + \frac{L_g \eta_{\boldsymbol{w}}^2}{2} \| G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \|_2^2 \\
= & g(\boldsymbol{w}_t) + \frac{L_g \eta_{\boldsymbol{w}}^2}{2} \| G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \|_2^2 \\
& + \eta_{\boldsymbol{w}} \langle \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) - \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)), G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \rangle \\
& + \eta_{\boldsymbol{w}} \langle \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t) - \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)), G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \rangle \\
& - \eta_{\boldsymbol{w}} \langle \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t), G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \rangle \\
\leq & g(\boldsymbol{w}_t) + \frac{L_g \eta_{\boldsymbol{w}}^2}{2} \| G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \|_2^2 \\
& + \frac{1}{L_g} \| \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) - \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) \|_2^2 \\
& + \frac{L_g \eta_{\boldsymbol{w}}^2}{4} \| G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \|_2^2 \\
& + \frac{1}{L_g} \| \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t) - \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) \|_2^2 \\
& + \frac{L_g \eta_{\boldsymbol{w}}^2}{4} \| G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \|_2^2 \\
& - \eta_{\boldsymbol{w}} \langle \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t), G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \rangle \\
\leq & g(\boldsymbol{w}_t) + L_g \eta_{\boldsymbol{w}}^2 \| G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \|_2^2 + \frac{\mu^2 L^2 (d+3)^3}{4 L_g} + \frac{L^2}{L_g} \| \boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t \|_2^2 \\
& - \eta_{\boldsymbol{w}} \langle \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t), G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \rangle
\end{aligned} \tag{22}
$$

where the first inequality is due to Lemma 4. Then second inequality is due to the Young's inequality. The last inequality is due to Lemmas 2 and 1. Then, take expectation w.r.t $\boldsymbol{u}_{[q]}, \xi_{\mathcal{M}_1}, \zeta_{\mathcal{M}_2}$ on both sides of the above inquality and rearrange it, we obtain

$$
\begin{aligned}
& \eta_{\boldsymbol{w}} \mathbb{E}[\| \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t) \|_2^2] \\
\leq & \mathbb{E}[g(\boldsymbol{w}_t)] - \mathbb{E}[g(\boldsymbol{w}_{t+1})] + \frac{L^2}{L_g} \mathbb{E}[\| \boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t \|_2^2] + L_g \eta_{\boldsymbol{w}}^2 \mathbb{E}[\| G_\mu(\boldsymbol{w}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]}) \|_2^2] + \frac{\mu^2 L^2 (d+3)^3}{4 L_g}
\end{aligned} \tag{23}
$$

From Lemma 2, we have

$$
\eta_{\boldsymbol{w}} \mathbb{E}[\| \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}, \boldsymbol{p}^*(\boldsymbol{w})) \|_2^2] \leq \eta_{\boldsymbol{w}} \mathbb{E}[\| \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t) \|_2^2] + \eta_{\boldsymbol{w}} L^2 \| \boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t) \|_2^2 \tag{24}
$$

From Lemma 3, we have

$$
\eta_{\boldsymbol{w}} \| \nabla g(\boldsymbol{w}_t) \|_2^2 \leq 2 \eta_{\boldsymbol{w}} \| \nabla_{\boldsymbol{w}} \mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) \|_2^2 + \frac{\eta_{\boldsymbol{w}} \mu^2 L^2 (d+3)^3}{2} \tag{25}
$$

Combining the above inequalities yields

$$\begin{aligned}
&\eta_{\boldsymbol{w}}\mathbb{E}[\|\nabla g(\boldsymbol{w}_t)\|_2^2] \\
\leq& 2\eta_{\boldsymbol{w}}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t,\boldsymbol{p}_t)\|_2^2] + 2\eta_{\boldsymbol{w}}L^2\|\boldsymbol{p}_t-\boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 + \frac{\eta_{\boldsymbol{w}}\mu^2L^2(d+3)^3}{2} \\
\leq& 2\mathbb{E}[g(\boldsymbol{w}_t)] - 2\mathbb{E}[g(\boldsymbol{w}_{t+1})] + \frac{2L^2}{L_g}\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_t)-\boldsymbol{p}_t\|_2^2] \\
&+ 2L_g\eta_{\boldsymbol{w}}^2\mathbb{E}[\|G_\mu(\boldsymbol{w}_t,\ell_{\mathcal{M}_1},f_{\mathcal{M}_2},\boldsymbol{u}_{[q]})\|_2^2] + \frac{\mu^2L^2(d+3)^3}{2L_g} \\
&+ 2\eta_{\boldsymbol{w}}L^2\|\boldsymbol{p}_t-\boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 + \frac{\eta_{\boldsymbol{w}}\mu^2L^2(d+3)^3}{2} \\
\leq& 2\mathbb{E}[g(\boldsymbol{w}_t)] - 2\mathbb{E}[g(\boldsymbol{w}_{t+1})] + (\frac{2L^2}{L_g}+2\eta_{\boldsymbol{w}}L^2)\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_t)-\boldsymbol{p}_t\|_2^2] \\
&+ \frac{\mu^2L^2(d+3)^3}{2L_g} + \frac{\eta_{\boldsymbol{w}}\mu^2L^2(d+3)^3}{2} \\
&+ 2L_g\eta_{\boldsymbol{w}}^2\left(3\mathbb{E}[\|\nabla g(\boldsymbol{w}_t)\|_2^2] + 3L^2\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_t)-\boldsymbol{p}_t\|_2^2] + \rho(\epsilon,\mu)\right) \\
=& 2\mathbb{E}[g(\boldsymbol{w}_t)] - 2\mathbb{E}[g(\boldsymbol{w}_{t+1})] + 6L_g\eta_{\boldsymbol{w}}^2\mathbb{E}[\|\nabla g(\boldsymbol{w}_t)\|_2^2] + \theta_2\mathbb{E}[\delta_t] + \theta_3
\end{aligned} \tag{26}$$

where

$$\theta_2 = \frac{2L^2}{L_g} + 2\eta_{\boldsymbol{w}}L^2 + 6L_g\eta_{\boldsymbol{w}}^2L^2 \leq 2L + 2\eta_{\boldsymbol{w}}L^2 + 6\eta_{\boldsymbol{w}}L^3(\kappa+1) \tag{27}$$

and

$$\begin{aligned}
&\theta_3 \\
=& \frac{\mu^2L^2(d+3)^3}{2L_g} + \frac{\eta_{\boldsymbol{w}}\mu^2L^2(d+3)^3}{2} + 2L_g\eta_{\boldsymbol{w}}^2\rho(\epsilon,\mu) \\
\leq& \frac{\mu^2L(d+3)^3}{2} + \frac{\eta_{\boldsymbol{w}}\mu^2L^2(d+3)^3}{2} + 2L(\kappa+1)\eta_{\boldsymbol{w}}^2\rho(\epsilon,\mu) \\
\leq& \frac{\mu^2L(d+3)^3}{2} + \frac{\eta_{\boldsymbol{w}}\mu^2L^2(d+3)^3}{2} + L^3(\kappa+1)\eta_{\boldsymbol{w}}^2(\frac{\mu^2(d+6)^2\epsilon^2}{4} + \mu^2(d+3)^3) + L(\kappa+1)\eta_{\boldsymbol{w}}^2\epsilon^2
\end{aligned} \tag{28}$$

Taking sum over $t = 0, \cdots, T$, we get

$$\sum_{t=0}^{T}\mathbb{E}[\delta_t] = \sum_{t=0}^{T}\gamma^t\mathbb{E}[\delta_0] + 3\kappa^2\eta_{\boldsymbol{w}}^2\sum_{t=0}^{T}\sum_{i=0}^{t-1}\gamma^{t-1-i}\mathbb{E}[\|\nabla g(\boldsymbol{w}_i)\|_2^2] + \theta\sum_{t=0}^{T}\sum_{i=0}^{t-1}\gamma^{t-1-i} \tag{29}$$

Moreover, we have

$$\sum_{t=0}^{T}\gamma^t \leq 6\kappa \tag{30}$$

$$\sum_{t=0}^{T}\sum_{i=0}^{t-1}\gamma^{s-1-i} \leq 6\kappa(T+1) \tag{31}$$

$$\sum_{t=0}^{T}\sum_{i=0}^{t-1}\gamma^{s-1-i}\mathbb{E}[\|\nabla g(\boldsymbol{w}_i)\|_2^2] \leq 6\kappa\sum_{t=0}^{T}\mathbb{E}[\|\nabla g(\boldsymbol{w}_t)\|_2^2] \tag{32}$$

Thus, we have

$$\sum_{t=0}^{T}\mathbb{E}[\delta_t] = 6\kappa\mathbb{E}[\delta_0] + 18\kappa^3\eta_{\boldsymbol{w}}^2\sum_{t=0}^{T}\mathbb{E}[\|\nabla g(\boldsymbol{w}_t)\|_2^2] + \theta_1 6\kappa(T+1) \tag{33}$$

Then, summing 26 over $t = 1, \cdots, T$, yields

$$\begin{aligned}
&\eta_{\boldsymbol{w}} \sum_{t=0}^{T} \mathbb{E}[\|\nabla g(\boldsymbol{w}_t)\|_2^2] \\
=&2\mathbb{E}[g(\boldsymbol{w}_0)] - 2\mathbb{E}[g(\boldsymbol{w}_{T+1})] + 6L_g\eta_{\boldsymbol{w}}^2 \sum_{t=0}^{S} \mathbb{E}[\|\nabla g(\boldsymbol{w}_t)\|_2^2] + \theta_2 \sum_{t=0}^{T} \mathbb{E}[\delta_t] + \theta_3(T+1) \\
\leq&2\mathbb{E}[g(\boldsymbol{w}_0)] - 2\mathbb{E}[g(\boldsymbol{w}_{T+1})] + 6L_g\eta_{\boldsymbol{w}}^2 \sum_{t=0}^{S} \mathbb{E}[\|\nabla g(\boldsymbol{w}_t)\|_2^2] \\
&+ \theta_2 \left(6\kappa\mathbb{E}[\delta_0] + 18\kappa^3\eta_{\boldsymbol{w}}^2 \sum_{t=0}^{T} \mathbb{E}[\|\nabla g(\boldsymbol{w}_t)\|_2^2] + \theta_1 6\kappa(T+1) \right) \\
&+ \theta_3(T+1)
\end{aligned} \tag{34}$$

In addition, we have

$$18\kappa^3\eta_{\boldsymbol{w}}^2\theta_2 \leq (\frac{36}{4\times 16^2} + \frac{36}{4^2\times 16^4} + \frac{18\times 6}{4^2\times 16^4})\eta_{\boldsymbol{w}} \leq 0.0616\eta_{\boldsymbol{w}} \tag{35}$$

and

$$6L_g\eta_{\boldsymbol{w}}^2 = 6(\kappa+1)L\eta_{\boldsymbol{w}}^2 \leq 0.0059\eta_{\boldsymbol{w}} \tag{36}$$

Thus, we have

$$0.9325\eta_{\boldsymbol{w}} \sum_{t=0}^{T} \mathbb{E}[\|\nabla g(\boldsymbol{w}_t)\|_2^2] \leq 2\mathbb{E}[g(\boldsymbol{w}_0)] - 2\mathbb{E}[g(\boldsymbol{w}_{T+1})] + \theta_2\left(6\kappa\mathbb{E}[\delta_0] + \theta_1 6\kappa(T+1)\right) + \theta_3(T+1) \tag{37}$$

Dividing both side of the above inequality by $0.9325(T+1)\eta_{\boldsymbol{w}}$ and let $\boldsymbol{p}_0 = \boldsymbol{p}^*(\boldsymbol{w}_0)$, we have

$$\frac{1}{T+1}\sum_{t=0}^{T} \mathbb{E}[\|\nabla g(\boldsymbol{w}_t)\|_2^2] \leq \frac{2(g(\boldsymbol{w}_0) - g(\boldsymbol{w}_{T+1}))}{0.9325(T+1)\eta_{\boldsymbol{w}}} + \frac{6\theta_2\theta_1\kappa}{0.9325\eta_{\boldsymbol{w}}} + \frac{\theta_3}{0.9325\eta_{\boldsymbol{w}}} + \frac{6\theta_2\kappa\sigma_1^2}{0.9325(T+1)\eta_{\boldsymbol{w}}} \tag{38}$$

Then, we o bound the right side by ϵ^2. Then, we have $T > \max\{\frac{2(g(\boldsymbol{w}_0) - g(\boldsymbol{w}_{T+1}))}{0.9325\epsilon^2\eta_{\boldsymbol{w}}}, \frac{\sigma_1^2}{16\kappa\eta_w^2\epsilon^2}\}$.
In additinon We have

$$\theta_3 \leq \frac{\mu^2 L(d+3)^3}{2} + \frac{\eta_{\boldsymbol{w}}\mu^2L^2(d+3)^3}{2} + L^3(\kappa+1)\eta_{\boldsymbol{w}}^2(\frac{\mu^2(d+6)^2\epsilon^2}{4} + \mu^2(d+3)^3) + L(\kappa+1)\eta_{\boldsymbol{w}}^2\epsilon^2 \leq \epsilon^2\eta_w \tag{39}$$

Let $\mu := \mathcal{O}(\epsilon d^{-3/2}L^{-2})$ and the above inequality holds.

D CONVERGENCE ANALYSIS OF ADSGZO

Here, we give several lemmas used in our analysis and then give the proof of theorem 2.

Lemma 8 *Under Assumptions 1 and 3, if $\eta_w L \leq \frac{c_{1,l}}{2c_{1,u}^2}$, we have*

$$\begin{aligned}
&g(\boldsymbol{w}_{t+1}) \\
\leq&g(\boldsymbol{w}_t) - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{4}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \eta_w c_{1,l}\frac{\mu^2L^2(d+3)^3}{2} + \eta_w c_{1,l}L^2\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 \\
&+ \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2
\end{aligned} \tag{40}$$

Proof 7

$$
\begin{aligned}
&g(\boldsymbol{w}_{t+1}) \\
\leq& g(\boldsymbol{w}_t) + \nabla g(\boldsymbol{w}_t)^T(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) + \frac{L}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_2^2 \\
=& g(\boldsymbol{w}_t) - \eta_w \nabla g(\boldsymbol{w}_t)^T \frac{\boldsymbol{z}_{\boldsymbol{w}}^t}{\sqrt{\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2}} + \frac{L}{2}\frac{\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2}{\|\sqrt{\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2}\|_2^2} \\
\leq& g(\boldsymbol{w}_t) - \eta_w c_{1,l} \nabla g(\boldsymbol{w}_t)^T \boldsymbol{z}_{\boldsymbol{w}}^t + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
=& g(\boldsymbol{w}_t) + \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
=& g(\boldsymbol{w}_t) - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
&+ \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t) - \nabla g_\mu(\boldsymbol{w}_t) + \nabla g_\mu(\boldsymbol{w}_t) - \nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t) + \nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t) - \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) + \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
\leq& g(\boldsymbol{w}_t) - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
&+ \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) - \nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t))\|_2^2 + \eta_w c_{1,l}\|\nabla\nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) - \nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t)\|_2^2 \\
&+ \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}_\mu(\boldsymbol{w}_t, \boldsymbol{p}_t) - \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)\|_2^2 + \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
\leq& g(\boldsymbol{w}_t) - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \eta_w c_{1,l}\frac{\mu^2 L^2(d+3)^3}{4} + \eta_w c_{1,l} L^2\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 \\
&+ \eta_w c_{1,l}\frac{\mu^2 L^2(d+3)^3}{4} + \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
\leq& g(\boldsymbol{w}_t) - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \frac{L\eta_w^2 c_{1,u}^2}{2}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \eta_w c_{1,l}\frac{\mu^2 L^2(d+3)^3}{2} + \eta_w c_{1,l} L^2\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 \\
&+ \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 \\
\leq& g(\boldsymbol{w}_t) - \frac{\eta_w c_{1,l}}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 - \frac{\eta_w c_{1,l}}{4}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2 + \eta_w c_{1,l}\frac{\mu^2 L^2(d+3)^3}{2} + \eta_w c_{1,l} L^2\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 \\
&+ \eta_w c_{1,l}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2
\end{aligned}
\tag{41}
$$

The last inequality is due to $\eta_w L \leq \frac{c_{1,l}}{2c_{1,u}^2}$

Lemma 9 *Under Assumptions 1 and 3, if* $a \leq 1$, $\tau \leq L$ *and* $\eta_p \leq \frac{1}{3c_{2.l}L}$, *we have* $\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2 \leq -\frac{1}{4a}\|\boldsymbol{p}_t - \boldsymbol{p}_{t+1}\|_2^2 + \frac{8a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 + (1 - \frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{8L_g^2}{\tau a\eta_p c_{2,l}}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2^2$

Proof 8 *We have*

$$
\begin{aligned}
&\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 \\
\leq& \|(1-a)\boldsymbol{p}_t + a\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 \\
=& \|\boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 + a^2\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 + 2a\langle \boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t), \boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\rangle
\end{aligned}
\tag{42}
$$

Thus, we have

$$
\langle \boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t), \boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\rangle \geq \frac{1}{2a}\left(\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 - \|\boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 - a^2\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2\right)
\tag{43}
$$

Due to the Assumption, we have

$$
\mathcal{L}(\boldsymbol{w}_t, \hat{\boldsymbol{p}}_{t+1}) \geq \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) + \nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T(\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t) - \frac{L}{2}\|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2
\tag{44}
$$

In addition, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}) \\
\leq &\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) + \nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T (\boldsymbol{p} - \boldsymbol{p}_t) - \frac{\tau}{2} \|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2 \\
= &\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) + \nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T (\boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1} + \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t) - \frac{\tau}{2} \|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2 \\
= &\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) + \nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T (\boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}) + \nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T (\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t) - \frac{\tau}{2} \|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2 \\
= &\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) + (\nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t)^T (\boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}) + \langle \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1} \rangle + \nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t)^T (\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t) - \frac{\tau}{2} \|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2
\end{aligned} \tag{45}$$

Then, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}) \\
\leq &\mathcal{L}(\boldsymbol{w}_t, \hat{\boldsymbol{p}}_{t+1}) + (\nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t)^T (\boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1}) + \langle \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1} \rangle - \frac{\tau}{2} \|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2 + \frac{L}{2} \|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2
\end{aligned} \tag{46}$$

Due to the update rule of $\hat{\boldsymbol{p}}$, we have

$$\langle \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t - \eta_{\boldsymbol{p}} \frac{\boldsymbol{z}_{\boldsymbol{p}}^t}{\sqrt{\|\boldsymbol{z}_{\boldsymbol{p}}^t\|_2}}, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1} \rangle \geq 0, \quad \forall \boldsymbol{p} \in \Delta^m \tag{47}$$

Then, we have

$$\begin{aligned}
&\eta_{\boldsymbol{p}} c_{2,l} \langle \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1} \rangle \\
\leq &\langle \eta_{\boldsymbol{p}} \frac{\boldsymbol{z}_{\boldsymbol{p}}^t}{\sqrt{\|\boldsymbol{z}_{\boldsymbol{p}}^t\|_2}}, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1} \rangle \\
\leq &\langle \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t, \boldsymbol{p} - \hat{\boldsymbol{p}}_{t+1} \rangle \\
= &\langle \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t, \boldsymbol{p} - \boldsymbol{p}_t + \boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1} \rangle \\
= &- \|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2 + \langle \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t, \boldsymbol{p} - \boldsymbol{p}_t \rangle
\end{aligned} \tag{48}$$

In addition, we have

$$\begin{aligned}
&\langle \nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p}^*(\boldsymbol{w}_t) - \hat{\boldsymbol{p}}_{t+1} \rangle \\
= &\langle \nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t \rangle + \langle \nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t, \boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1} \rangle \\
\leq &\frac{1}{\tau} \|\nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 + \frac{\tau}{4} \|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{1}{\tau} \|\nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 + \frac{\tau}{4} \|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 \\
\leq &\frac{2}{\tau} \|\nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 + \frac{\tau}{4} \|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{\tau}{4} \|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2
\end{aligned} \tag{49}$$

Then, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}) \\
\leq &\mathcal{L}(\boldsymbol{w}_t, \hat{\boldsymbol{p}}_{t+1}) - \frac{1}{\eta_p c_{2,l}} \|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2 + \frac{1}{\eta_p c_{2,l}} \langle \hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t, \boldsymbol{p} - \boldsymbol{p}_t \rangle - \frac{\tau}{2} \|\boldsymbol{p} - \boldsymbol{p}_t\|_2^2 + \frac{L}{2} \|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2 \\
&+ \frac{2}{\tau} \|\nabla_{\boldsymbol{p}} \mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 + \frac{\tau}{4} \|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{\tau}{4} \|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2
\end{aligned} \tag{50}$$

Let $\boldsymbol{p} = \boldsymbol{p}^*(\boldsymbol{w}_t)$, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}_t, \hat{\boldsymbol{p}}_{t+1}) \\
\leq &\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}^*(\boldsymbol{w}_t)) \\
\leq &\mathcal{L}(\boldsymbol{w}_t, \hat{\boldsymbol{p}}_{t+1}) - \frac{1}{\eta_p c_{2,l}}\|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2 - \frac{\tau}{2}\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{L}{2}\|\hat{\boldsymbol{p}}_{t+1} - \boldsymbol{p}_t\|_2^2 \\
&+ \frac{2}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 + \frac{\tau}{4}\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{\tau}{4}\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 \\
&- \frac{1}{2a\eta_p c_{2,l}}\left(\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 - \|\boldsymbol{p}_t - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 - a^2\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2\right)
\end{aligned} \tag{51}$$

Rearrange the inequality, we have

$$\begin{aligned}
&\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 \\
\leq &-2a\eta_p c_{2,l}\left(\frac{1}{\eta_p c_{2,l}} - \frac{L}{2} - \frac{\tau}{4} - \frac{1}{2b\eta_p c_{2,l}}\right)\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 + \frac{4a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{2})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 \\
\leq &-2a\eta_p c_{2,l}\left(\frac{1}{2\eta_p c_{2,l}} - \frac{3L}{4}\right)\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 + \frac{4a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{2})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2
\end{aligned} \tag{52}$$

where we use $a \leq 1$, $\tau \leq L$ and $\eta_p \leq \frac{1}{3c_{2.l}L}$.

Then, we have

$$\begin{aligned}
&\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2 \\
= &\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t) + \boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2 \\
\leq &(1 + \frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}_{t+1} - \boldsymbol{p}^*(\boldsymbol{w}_t)\|_2^2 + (1 + \frac{4}{\tau a\eta_p c_{2,l}})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2 \\
\leq &-2a\eta_p c_{2,l}(1 + \frac{\tau a\eta_p c_{2,l}}{4})\left(\frac{1}{2\eta_p c_{2,l}} - \frac{3L}{4}\right)\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 + \frac{4a\eta_p c_{2,l}}{\tau}(1 + \frac{\tau a\eta_p c_{2,l}}{4})\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{2})(1 + \frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + (1 + \frac{4}{\tau a\eta_p c_{2,l}})L_g^2\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2^2 \\
\leq &-2a\eta_p c_{2,l}(1 + \frac{\tau a\eta_p c_{2,l}}{4})\left(\frac{1}{2\eta_p c_{2,l}} - \frac{3L}{4}\right)\|\boldsymbol{p}_t - \hat{\boldsymbol{p}}_{t+1}\|_2^2 + \frac{8a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{8L_g^2}{\tau a\eta_p c_{2,l}}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2^2 \\
\leq &-\frac{2\eta_p c_{2,l}}{a}(1 + \frac{\tau a\eta_p c_{2,l}}{4})\left(\frac{1}{2\eta_p c_{2,l}} - \frac{3L}{4}\right)\|\boldsymbol{p}_t - \boldsymbol{p}_{t+1}\|_2^2 + \frac{8a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{8L_g^2}{\tau a\eta_p c_{2,l}}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2^2 \\
\leq &-\frac{1}{4a}\|\boldsymbol{p}_t - \boldsymbol{p}_{t+1}\|_2^2 + \frac{8a\eta_p c_{2,l}}{\tau}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2 \\
&+ (1 - \frac{\tau a\eta_p c_{2,l}}{4})\|\boldsymbol{p}^*(\boldsymbol{w}_t) - \boldsymbol{p}_t\|_2^2 + \frac{8L_g^2}{\tau a\eta_p c_{2,l}}\|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2^2
\end{aligned} \tag{53}$$

Lemma 10 *Under Assumptions 1, 2 and 3, if $b \in (0, 1)$, we have*

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1}, \boldsymbol{p}_{t+1}) - \boldsymbol{z}_{\boldsymbol{p}}^{t+1}\|_2^2] \\
\leq &(1 - b)\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t, \boldsymbol{p}_t) - \boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2] + \frac{1}{b}L^2[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_2^2 + \|\boldsymbol{p}_{t+1} - \boldsymbol{p}_t\|_2^2] + b^2\sigma_2^2
\end{aligned} \tag{54}$$

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\boldsymbol{z}_{\boldsymbol{w}}^{t+1}\|_2^2]\\
\leq&(1-b)\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{w}}^{t}\|_2^2]+\frac{1}{b}L^2[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|_2^2+\|\boldsymbol{p}_{t+1}-\boldsymbol{p}_t\|_2^2]+b^2\sigma_1^2
\end{aligned}\tag{55}$$

Proof 9

$$\boldsymbol{z}_{\boldsymbol{p}}^{t+1}-\boldsymbol{z}_{\boldsymbol{p}}^{t}=-b\boldsymbol{z}_{\boldsymbol{p}}^{t}+bH^{t+1}\tag{56}$$

Then, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\boldsymbol{z}_{\boldsymbol{p}}^{t+1}\|_2^2]\\
=&\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\boldsymbol{z}_{\boldsymbol{p}}^{t}-(\boldsymbol{z}_{\boldsymbol{p}}^{t+1}-\boldsymbol{z}_{\boldsymbol{p}}^{t})\|_2^2]\\
=&\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\boldsymbol{z}_{\boldsymbol{p}}^{t}+b\boldsymbol{z}_{\boldsymbol{p}}^{t}-bH^{t+1}\|_2^2]\\
=&\mathbb{E}[\|(1-b)\left(\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^{t}\right)+(1-b)\left(\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)\right)+b\left(\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-H^{t+1}\right)\|_2^2]\\
=&(1-b)^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^{t}\|_2^2]+(1-b)^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)\|_2^2]\\
&+b^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-H^{t+1}\|_2^2]+(1-b)^2\langle\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^{t},\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)\rangle\\
=&(1-b)^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^{t}\|_2^2]+(1-b)^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)\|_2^2]\\
&+b^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-H^{t+1}\|_2^2]+(1-b)^2\mathbb{E}[\langle\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^{t},\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)\rangle]\\
\leq&(1-b)^2(1+b)\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^{t}\|_2^2]+(1-b)^2(1+\frac{1}{b})\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)\|_2^2\\
&+b^2\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-H^{t+1}\|_2^2]\\
\leq&(1-b)\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^{t}\|_2^2]+\frac{1}{b}L^2\mathbb{E}[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|_2^2+\|\boldsymbol{p}_{t+1}-\boldsymbol{p}_t\|_2^2]+b^2\sigma_2^2
\end{aligned}\tag{57}$$

Similarly, we have

$$\begin{aligned}
&\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\boldsymbol{z}_{\boldsymbol{w}}^{t+1}\|_2^2]\\
\leq&(1-b)\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{w}}^{t}\|_2^2]+\frac{1}{b}L^2\mathbb{E}[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|_2^2+\|\boldsymbol{p}_{t+1}-\boldsymbol{p}_t\|_2^2]+b^2\sigma_1^2
\end{aligned}\tag{58}$$

Then, we give the proof of theorem 2.

Proof 10 *Suming up the above inequality, we have*

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\boldsymbol{z}_{\boldsymbol{w}}^{t+1}\|_2^2]\\
\leq&(1-b)\sum_{t=1}^{T}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{w}}^{t}\|_2^2]+\frac{1}{b}L^2\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|_2^2+\|\boldsymbol{p}_{t+1}-\boldsymbol{p}_t\|_2^2]+b^2\sigma_1^2T
\end{aligned}\tag{59}$$

Then, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\boldsymbol{z}_{\boldsymbol{w}}^{t+1}\|_2^2]\\
\leq&\frac{1}{b}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_1,\boldsymbol{p}_1)-\boldsymbol{z}_{\boldsymbol{w}}^{1}\|_2^2]+\frac{1}{b^2}L^2\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|_2^2+\|\boldsymbol{p}_{t+1}-\boldsymbol{p}_t\|_2^2]+b\sigma_1^2T
\end{aligned}\tag{60}$$

Similarly, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_{t+1},\boldsymbol{p}_{t+1})-\boldsymbol{z}_{\boldsymbol{p}}^{t+1}\|_2^2]\\
\leq&\frac{1}{b}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_1,\boldsymbol{p}_1)-\boldsymbol{z}_{\boldsymbol{w}}^{1}\|_2^2]+\frac{1}{b^2}L^2\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|_2^2+\|\boldsymbol{p}_{t+1}-\boldsymbol{p}_t\|_2^2]+b\sigma_1^2T
\end{aligned}\tag{61}$$

$$
\begin{aligned}
&\sum_{t=1}^{T}\|\boldsymbol{p}_{t+1}-\boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2 \\
\leq&\frac{4}{\tau a\eta_p c_{2,l}}(\|\boldsymbol{p}^*(\boldsymbol{w}_1)-\boldsymbol{p}_1\|_2^2-\frac{1}{4a}\sum_{t=1}^{T}\|\boldsymbol{p}_t-\boldsymbol{p}_{t+1}\|_2^2+\frac{8a\eta_p c_{2,l}}{\tau}\sum_{t=1}^{T}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2+\frac{8L_g^2}{\tau a\eta_p c_{2,l}}\sum_{t=1}^{T}\|\boldsymbol{w}_t-\boldsymbol{w}_{t+1}\|_2^2) \\
\leq&\frac{4}{\tau a\eta_p c_{2,l}}\|\boldsymbol{p}^*(\boldsymbol{w}_1)-\boldsymbol{p}_1\|_2^2-\frac{1}{\tau a^2\eta_p c_{2,l}}\sum_{t=1}^{T}\|\boldsymbol{p}_t-\boldsymbol{p}_{t+1}\|_2^2+\frac{32}{\tau^2}\sum_{t=1}^{T}\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{p}}^t\|_2^2+\frac{32L_g^2\eta_w^2c_{1,u}^2}{\tau^2a^2\eta_p^2c_{2,l}^2}\sum_{t=1}^{T}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2
\end{aligned}
\tag{62}
$$

Thus, we have

$$
\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T}\|\boldsymbol{p}_{t+1}-\boldsymbol{p}^*(\boldsymbol{w}_{t+1})\|_2^2] \\
\leq&\frac{4}{\tau a\eta_p c_{2,l}}\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_1)-\boldsymbol{p}_1\|_2^2]+\mathbb{E}[\frac{32L_g^2\eta_w^2c_{1,u}^2}{\tau^2a^2\eta_p^2c_{2,l}^2}\sum_{t=1}^{T}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2-\frac{1}{\tau a^2\eta_p c_{2,l}}\sum_{t=1}^{T}\|\boldsymbol{p}_t-\boldsymbol{p}_{t+1}\|_2^2] \\
&+\frac{32}{\tau^2 b}\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_1,\boldsymbol{p}_1)-\boldsymbol{z}_{\boldsymbol{p}}^1\|_2^2]+\mathbb{E}\left[\sum_{t=1}^{T}\frac{32}{\tau^2}\left(b\sigma_2^2+\frac{L^2(\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|_2^2+\|\boldsymbol{p}_{t+1}-\boldsymbol{p}_t\|_2^2)}{b^2}\right)\right] \\
\leq&\frac{4}{\tau a\eta_p c_{2,l}}\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_1)-\boldsymbol{p}_1\|_2^2]+\mathbb{E}\left[\left(\frac{32L_g^2\eta_w^2c_{1,u}^2}{\tau^2a^2\eta_p^2c_{2,l}^2}+\frac{32L^2\eta_w^2c_{1,u}^2}{\tau^2b^2}\right)\sum_{t=1}^{T}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2\right] \\
&+\frac{32}{\tau^2 b}\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_1,\boldsymbol{p}_1)-\boldsymbol{z}_{\boldsymbol{p}}^1\|_2^2]+\mathbb{E}\left[\sum_{t=1}^{T}\frac{32}{\tau^2}b\sigma_2^2\right] \\
&+\mathbb{E}\left[\left(\frac{32L^2}{\tau^2b^2}-\frac{1}{\tau a^2\eta_p c_{2,l}}\right)\sum_{t=1}^{T}\|\boldsymbol{p}_t-\boldsymbol{p}_{t+1}\|_2^2]\right]
\end{aligned}
\tag{63}
$$

In addition, we have

$$
\begin{aligned}
&\frac{1}{2}\|\nabla g(\boldsymbol{w}_t)\|_2^2 \\
\leq&\frac{g(\boldsymbol{w}_t)-g(\boldsymbol{w}_{t+1})}{\eta_w c_{1,l}}-\frac{1}{4}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2+\frac{\mu^2L^2(d+3)^3}{2}+L^2\|\boldsymbol{p}^*(\boldsymbol{w}_t)-\boldsymbol{p}_t\|_2^2+\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_t,\boldsymbol{p}_t)-\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2
\end{aligned}
\tag{64}
$$

Summing up from $t=1,\cdots,T$ and taking expectation, we have

$$
\begin{aligned}
&\mathbb{E}[\frac{1}{2}\sum_{t=1}^{T}\|\nabla g(\boldsymbol{w}_t)\|_2^2] \\
\leq&\frac{g(\boldsymbol{w}_1)-g(\boldsymbol{w}_{T+1})}{\eta_w c_{1,l}}-\frac{1}{4}\mathbb{E}[\sum_{t=1}^{T}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2]+\frac{\mu^2TL^2(d+3)^3}{2} \\
&+\frac{4L^2}{\tau a\eta_p c_{2,l}}\mathbb{E}[\|\boldsymbol{p}^*(\boldsymbol{w}_1)-\boldsymbol{p}_1\|_2^2]+\mathbb{E}\left[\left(\frac{L^2\eta_w^2c_{1,u}^2}{b^2}+\frac{32L_g^2L^2\eta_w^2c_{1,u}^2}{\tau^2a^2\eta_p^2c_{2,l}^2}+\frac{32L^4\eta_w^2c_{1,u}^2}{\tau^2b^2}\right)\sum_{t=1}^{T}\|\boldsymbol{z}_{\boldsymbol{w}}^t\|_2^2\right] \\
&+\frac{32L^2}{\tau^2 b}\mathbb{E}[\|\nabla_{\boldsymbol{p}}\mathcal{L}(\boldsymbol{w}_1,\boldsymbol{p}_1)-\boldsymbol{z}_{\boldsymbol{p}}^1\|_2^2]+\mathbb{E}\left[\sum_{t=1}^{T}\frac{32L^2}{\tau^2}b\sigma_2^2\right] \\
&+\mathbb{E}\left[\left(\frac{L^2}{b^2}+\frac{32L^4}{\tau^2b^2}-\frac{L^2}{\tau a^2\eta_p c_{2,l}}\right)\sum_{t=1}^{T}\|\boldsymbol{p}_t-\boldsymbol{p}_{t+1}\|_2^2]\right] \\
&+\frac{1}{b}\mathbb{E}[\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_1,\boldsymbol{p}_1)-\boldsymbol{z}_{\boldsymbol{w}}^1\|_2^2]+b\sigma_1^2T
\end{aligned}
\tag{65}
$$

Let $\boldsymbol{p}^*(w_1) = \boldsymbol{p}_1$, $\boldsymbol{z}_{\boldsymbol{p}}^1 = H(\boldsymbol{w}_t, \boldsymbol{p}_t, f_{\mathcal{M}_3})$, $\boldsymbol{z}_{\boldsymbol{w}}^1 = G_\mu^{\mathcal{L}}(\boldsymbol{w}_t, \boldsymbol{p}_t, \ell_{\mathcal{M}_1}, f_{\mathcal{M}_2}, \boldsymbol{u}_{[q]})$, $\eta_p \leq \min\{\frac{b^2}{\tau a^2 c_{2,l}}, \frac{\tau b^2}{32L^2a^2c_{2,l}}\}$, $\eta_w^2 \leq \min\{\frac{b^2}{4c_{1,u}^2L^2}, \frac{\tau^2a^2\eta_p^2c_{2,l}^2}{128L_g^2L^2c_{1,u}}, \frac{\tau^2b^2}{128L^4c_{1,u}^2}\}$, *we have*

$$\frac{1}{T}\mathbb{E}[\sum_{t=1}^{T}\|\nabla g(\boldsymbol{w}_t)\|_2^2] \leq \frac{2(g(\boldsymbol{w}_1)-g(\boldsymbol{w}_{T+1}))}{T\eta_w c_{1,l}} + \frac{64L^2}{\tau^2 bT}\sigma_2^2 + \frac{2\sigma_1^2}{bT} + \mu^2L^2(d+3)^3 + \frac{64L^2}{\tau^2}b\sigma_2^2 + 2b\sigma_1^2 \tag{66}$$

Bound the left term by ϵ^2, *we have* $\mu \leq \frac{\epsilon}{L(d+3)^{3/2}}$, $b \leq \min\{\frac{\epsilon^2}{2\sigma_1^2}, \frac{\tau^2\epsilon^2}{64\sigma_2^2L^2}\}$ *and* $T \geq \max\{\frac{2(g(\boldsymbol{w}_1)-g(\boldsymbol{w}_T))}{\epsilon^2\eta_w c_{1,l}}, \frac{2\sigma_1^2}{\epsilon^2 b}, \frac{64\sigma_2^2L^2}{\epsilon^2\tau^2 b}\}$

REFERENCES

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pp. 6083–6093. PMLR, 2020.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, and Meisam Razaviyayn. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *stat*, 1050:22, 2020.