A. Proof of Theorem 1

Proof. Define $\phi_s(\mathbf{v}) = \phi(\mathbf{v}) + \frac{1}{2\gamma}\|\mathbf{v}-\mathbf{v}_{s-1}\|^2$. We can see that $\phi_s(\mathbf{v})$ is convex and smooth since $\gamma \leq 1/L_{\mathbf{v}}$. The smooth coefficient of ϕ_s is $\hat{L}_{\mathbf{v}} = L_{\mathbf{v}} + 1/\gamma$. According to Theorem 2.1.5 of (Nesterov, 2004), we have

$$\|\nabla\phi_s(\mathbf{v}_s)\|^2 \leq 2\hat{L}_{\mathbf{v}}(\phi_s(\mathbf{v}_s) - \phi_s(\mathbf{v}^*_{\phi_s})). \tag{5}$$

Applying Lemma 2, we have

$$E_{s-1}[\phi_s(\mathbf{v}_s) - \phi_s(\mathbf{v}^*_{\phi_s})] \leq \frac{2}{\eta_s T_s}\|\mathbf{v}_{s-1} - \mathbf{v}^*_{\phi_s}\|^2 + \frac{1}{\eta_s T_s}(\alpha_{s-1} - \alpha^*(\mathbf{v}_s))^2 + H\eta_s^2 I_s^2 B^2 \mathbb{I}_{I_s>1} + \frac{\eta_s(2\sigma_{\mathbf{v}}^2 + 3\sigma_\alpha^2)}{2K}.$$

Denote $\mathbf{x}^k_{1:m_s} = (\mathbf{x}^k_1, ..., \mathbf{x}^k_{m_s})$, $y^k_{1:m_s} = (y^k_1, ..., y^k_{m_s})$, and $\tilde{f}_k(\mathbf{x}^k_{1:m_s}, y^k_{1:m_s}) = \frac{\sum_{i=1}^{m_s} h(\mathbf{w}_s;\mathbf{x}^k_i)\mathbb{I}_{y^k_i=y}}{\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y}} - E_{\mathbf{x}^k}[h(\mathbf{w}_s;\mathbf{x}^k)|y]$. If $\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y} > 0$, then $\frac{\sum_{i=1}^{m_s} h(\mathbf{w}_s;\mathbf{x}^k_i)\mathbb{I}_{y^k_i=y}}{\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y}}$ is an unbiased estimation of $E_{\mathbf{x}^k}[h(\mathbf{w}_s;\mathbf{x}^k)|y]$. Noting $0 \leq h(\mathbf{w};\mathbf{x}) \leq 1$, we have $Var(h(\mathbf{w};\mathbf{x}^k)|y) \leq \tilde{\sigma}^2 \leq 1$. Then we know that

$$\begin{aligned}
E_{\mathbf{x}^k_{1:m_s}}[(\tilde{f}(\mathbf{x}^k_{1:m_s}, y^k_{1:m_s}))^2|y^k_{1:m_s}] &\leq \frac{\tilde{\sigma}^2}{\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y}}\mathbb{I}_{(\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y}>0)} + 1\cdot\mathbb{I}_{(\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y}=0)} \\
&\leq \frac{\mathbb{I}_{(\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y}>0)}}{\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y}} + \mathbb{I}_{(\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y}=0)}.
\end{aligned} \tag{6}$$

Hence,

$$\begin{aligned}
&E_{s-1}[\tilde{f}_k(\mathbf{x}^k_{1:m_s}, y^k_{1:m_s})] = E_{y^k_{1:m_s}}\left[E_{\mathbf{x}^k_{1:m_s}}[(\tilde{f}_k(\mathbf{x}^k_{1:m_s}, y^k_{1:m_s}))^2|y^k_{1:m_s}]\right] \\
&\leq E_{y^k_{1:m_s}}\left[\frac{\mathbb{I}_{(\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y}>0)}}{\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y}} + \mathbb{I}_{\sum_{i=1}^{m_s}\mathbb{I}_{y^k_i=y}=0}\right] \leq \frac{1}{m_s\text{Pr}(y^k_i=y)} + (1-\text{Pr}(y^k_i=y))^{m_s}.
\end{aligned} \tag{7}$$

Denote

$$\begin{aligned}
\alpha^*(\mathbf{v}_s) &= \arg\max_\alpha f(\mathbf{v}_s,\alpha) = \frac{1}{K}\sum_{k=1}^K E\left[\frac{h(\mathbf{w}_s;\mathbf{x}^k)\mathbb{I}_{y^k=-1}}{1-p} - \frac{h(\mathbf{w}_s;\mathbf{x}^k)\mathbb{I}_{y^k=1}}{p}\right] \\
&= \frac{1}{K}\sum_{k=1}^K\left[E\left[h(\mathbf{w}_s;\mathbf{x}^k)|y^k=-1\right] - E\left[h(\mathbf{w}_s;\mathbf{x}^k)|y^k=1\right]\right].
\end{aligned} \tag{8}$$

Therefore,

$$\begin{aligned}
E_{s-1}[(\alpha_{s-1} - \alpha^*(\mathbf{v}_{s-1}))^2] = E_{s-1}&\left[\frac{1}{K}\sum_{k=1}^K\frac{\sum_{i=1}^{m_{s-1}} h(\mathbf{w}_{s-1};\mathbf{x}^k_i)\mathbb{I}_{y^k_i=-1}}{\sum_{i=1}^{m_{s-1}}\mathbb{I}_{y^k_i=-1}} - E\left[\frac{1}{K}\sum_{k=1}^K h(\mathbf{w}_{s-1};\mathbf{x}^k_i)|y=-1\right]\right. \\
&\left.+E\left[\frac{1}{K}\sum_{k=1}^K h(\mathbf{w}_{s-1};\mathbf{x}^k_i)|y=1\right] - \frac{1}{K}\sum_{k=1}^K\frac{\sum_{i=1}^{m_{s-1}} h(\mathbf{w}_{s-1};\mathbf{x}^k_i)\mathbb{I}_{y^k_i=1}}{\sum_{i=1}^{m_{s-1}}\mathbb{I}_{y^k_i=1}}\right]^2 \\
&\leq \frac{2}{Km_{s-1}\text{Pr}(y^k_i=-1)} + \frac{2(1-\text{Pr}(y^k_i=-1))^{m_{s-1}}}{K} + (1-\text{Pr}(y^k_i=-1))^{2m_{s-1}} \\
&\quad+\frac{2}{Km_{s-1}\text{Pr}(y^k_i=1)} + \frac{2(1-\text{Pr}(y_i=1))^{m_{s-1}}}{K} + (1-\text{Pr}(y^k_i=1))^{2m_{s-1}} \\
&\leq \frac{2}{Km_{s-1}p(1-p)} + \frac{3p^{m_{s-1}}}{K} + \frac{3(1-p)^{m_{s-1}}}{K} \leq 2\left(\frac{1}{Km_{s-1}p(1-p)} + \frac{3\tilde{p}^{m_{s-1}}}{K}\right) \\
&\leq 2\left(\frac{1}{Km_{s-1}p(1-p)} + \frac{C}{Km_{s-1}}\right) \leq \frac{2(1+C)}{Km_{s-1}p(1-p)},
\end{aligned} \tag{9}$$

where $C = \frac{3\tilde{p}^{\frac{1}{\ln(1/\tilde{p})}}}{2\ln(1/\tilde{p})}$ and $\tilde{p} = \max(p, 1-p)$.

Since $h(\mathbf{w};\mathbf{x})$ is G_h-Lipschitz, $E[h(\mathbf{w},\mathbf{x})|y=-1] - E[h(\mathbf{w},\mathbf{x})|y=1]$ is $2G_h$-Lipschitz. It follows that

$$\begin{aligned}
&E_{s-1}[(\alpha_{s-1} - \alpha^*(\mathbf{v}_s))^2] = E_{s-1}[(\alpha_{s-1} - \alpha^*(\mathbf{v}_{s-1}) + \alpha^*(\mathbf{v}_{s-1}) - \alpha^*(\mathbf{v}_s))^2] \\
&\leq E_{s-1}[2(\alpha_{s-1} - \alpha^*(\mathbf{v}_{s-1}))^2 + 2(\alpha^*(\mathbf{v}_{s-1}) - \alpha^*(\mathbf{v}_s))^2] \\
&= E_{s-1}[2(\alpha_{s-1} - \alpha^*(\mathbf{v}_s))^2] \\
&+ 2\left\|\frac{1}{K}\sum_{k=1}^{K}\left[E_{s-1}[h(\mathbf{w}_{s-1};\mathbf{x}^k)|y^k=-1] - E_{s-1}[h(\mathbf{w}_{s-1};\mathbf{x}^k)|y^k=1]] - [E_{s-1}[h(\mathbf{w}_s;\mathbf{x})|y^k=-1] - E_{s-1}[h(\mathbf{w}_s;\mathbf{x}^k)|y^k=1]\right]\right\|^2 \\
&\leq \frac{2(1+C)}{m_{s-1}K4p^2(1-p)^2} + 8G_h^2 E_{s-1}[\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2].
\end{aligned} \tag{10}$$

Since $m_{s-1} \geq \frac{1+C}{\eta_s^2 T_s \sigma_\alpha^2 p^2(1-p)^2}$, then we have

$$\begin{aligned}
E[\phi_s(\mathbf{v}_s) - \phi_s(\mathbf{v}^*_{\phi_s})] &\leq \frac{2\|\mathbf{v}_{s-1} - \mathbf{v}^*_{\phi_s}\|^2 + 8G_h^2 E[\|\mathbf{v}_{s-1} - \mathbf{v}_s\|]}{\eta_s T_s} + \frac{\eta_s\sigma_\alpha^2}{2K} + H\eta_s^2 I_s^2 B^2 \mathbb{I}_{I_s>1} + \frac{\eta_s(2\sigma_\mathbf{v}^2 + 3\sigma_\alpha^2)}{2K} \\
&\leq \frac{2\|\mathbf{v}_{s-1} - \mathbf{v}^*_{\phi_s}\|^2 + 8G_h^2 E[\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2]}{\eta_s T_s} + H\eta_s^2 I_s^2 B^2 \mathbb{I}_{I_s>1} + \frac{2\eta_s(\sigma_\mathbf{v}^2 + \sigma_\alpha^2)}{K}.
\end{aligned} \tag{11}$$

We define $I'_s = 1/\sqrt{K\eta_s} = \frac{1}{K\sqrt{\eta_0}}\exp(\frac{c(s-1)}{2})$. Applying this and (11) to (5), we get

$$\begin{aligned}
E[\|\nabla\phi_s(\mathbf{v}_s)\|^2] &\leq 2\hat{L}_\mathbf{v}\left[\frac{2\|\mathbf{v}_{s-1} - \mathbf{v}^*_{\phi_s}\|^2 + 8G_h^2 E[\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2]}{\eta_s T_s} + H\eta_s^2 {I'}_s^2 B^2 + \frac{2\eta_s(\sigma_\mathbf{v}^2 + \sigma_\alpha^2)}{K}\right] \\
&\leq 2\hat{L}_\mathbf{v}\left[\frac{2\|\mathbf{v}_{s-1} - \mathbf{v}^*_{\phi_s}\|^2 + 8G_h^2 E[\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2]}{\eta_s T_s} + H\eta_s^2 {I'}_s^2 B^2 + \frac{2\eta_s(\sigma_\mathbf{v}^2 + \sigma_\alpha^2)}{K}\right].
\end{aligned} \tag{12}$$

Taking $\gamma = \frac{1}{2L_\mathbf{v}}$, then $\hat{L}_\mathbf{v} = 3L_\mathbf{v}$. Note that $\phi_s(\mathbf{v})$ is $(\gamma^{-1} - L_\mathbf{v})$-strongly convex, we have

$$\phi_s(\mathbf{v}_{s-1}) \geq \phi_s(\mathbf{v}^*_{\phi_s}) + \frac{L_\mathbf{v}}{2}\|\mathbf{v}_{s-1} - \mathbf{v}^*_{\phi_s}\|^2. \tag{13}$$

Plugging (13) into (11), we get

$$\begin{aligned}
&E_{s-1}[\phi(\mathbf{v}_s) + L_\mathbf{v}\|\mathbf{v}_s - \mathbf{v}_{s-1}\|^2] \\
&\leq \phi_s(\mathbf{v}^*_{\phi_s}) + \frac{2\|\mathbf{v}_{s-1} - \mathbf{v}^*_{\phi_s}\|^2 + 8G_h^2 E_{s-1}[\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2]}{\eta_s T_s} + H\eta_s^2 {I'}_s^2 B^2 + \frac{2\eta_s(\sigma_\mathbf{v}^2 + \sigma_\alpha^2)}{K} \\
&\leq \phi_s(\mathbf{v}_{s-1}) - \frac{L_\mathbf{v}}{2}\|\mathbf{v}_{s-1} - \mathbf{v}^*_{\phi_s}\|^2 + \\
&\qquad \frac{2\|\mathbf{v}_{s-1} - \mathbf{v}^*_{\phi_s}\|^2 + 8G_h^2 E_{s-1}[\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2]}{\eta_s T_s} + H\eta_s^2 {I'}_s^2 B^2 + \frac{2\eta_s(\sigma_\mathbf{v}^2 + \sigma_\alpha^2)}{K}.
\end{aligned} \tag{14}$$

Noting $\eta_s T_s L_\mathbf{v} = \max(8, 8G_h^2)$ and $\phi_s(\mathbf{v}_{s-1}) = \phi(\mathbf{v}_{s-1})$, we rearrange terms and get

$$\frac{2\|\mathbf{v}_{s-1} - \mathbf{v}^*_{\phi_s}\|^2 + 8G_h^2 E_{s-1}[\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2]}{\eta_s T_s} \leq \phi(\mathbf{v}_{s-1}) - E_{s-1}[\phi(\mathbf{v}_s)] + H\eta_s^2 {I'}_s^2 B^2 + \frac{2\eta_s(\sigma_\mathbf{v}^2 + \sigma_\alpha^2)}{K}. \tag{15}$$

Combining (12) and (15), we get

$$\begin{aligned}
E_{s-1}\|\nabla\phi_s(\mathbf{v}_s)\|^2 &\leq 2\hat{L}_\mathbf{v}\left[\phi(\mathbf{v}_{s-1}) - E_{s-1}[\phi(\mathbf{v}_s)] + 2H\eta_s^2 {I'}_s^2 B^2 + \frac{4\eta_s(\sigma_\mathbf{v}^2 + \sigma_\alpha^2)}{K}\right] \\
&= 6L_\mathbf{v}\left[\phi(\mathbf{v}_{s-1}) - E_{s-1}[\phi(\mathbf{v}_s)] + 2H\eta_s^2 {I'}_s^2 B^2 + \frac{4\eta_s(\sigma_\mathbf{v}^2 + \sigma_\alpha^2)}{K}\right].
\end{aligned} \tag{16}$$

Taking expectation on both sides over all randomness until $\mathbf{v}_{s-1}$ is generated and by tower property, we have

$$E\|\nabla\phi_s(\mathbf{v}_s)\|^2 \leq 6L_{\mathbf{v}}\left(E[\phi(\mathbf{v}_{s-1})-\phi(\mathbf{v}_\phi^*)] - E[\phi(\mathbf{v}_s)-\phi(\mathbf{v}_\phi^*)] + 2H\eta_s^2 {I'}_s^2 B^2 + \frac{4\eta_s(\sigma_{\mathbf{v}}^2+\sigma_\alpha^2)}{K}\right) \tag{17}$$

Since $\phi(\mathbf{v})$ is $L_{\mathbf{v}}$-smooth and hence is $L_{\mathbf{v}}$-weakly convex, we have

$$\begin{aligned}
&\phi(\mathbf{v}_{s-1}) \geq \phi(\mathbf{v}_s) + \langle \nabla\phi(\mathbf{v}_s), \mathbf{v}_{s-1}-\mathbf{v}_s\rangle - \frac{L_{\mathbf{v}}}{2}\|\mathbf{v}_{s-1}-\mathbf{v}_s\|^2 \\
&= \phi(\mathbf{v}_s) + \langle \nabla\phi(\mathbf{v}_s) + 2L_{\mathbf{v}}(\mathbf{v}_s - \mathbf{v}_{s-1}), \mathbf{v}_{s-1}-\mathbf{v}_s\rangle + \frac{3}{2}L_{\mathbf{v}}\|\mathbf{v}_{s-1}-\mathbf{v}_s\|^2 \\
&= \phi(\mathbf{v}_s) + \langle \nabla\phi_s(\mathbf{v}_s), \mathbf{v}_{s-1}-\mathbf{v}_s\rangle + \frac{3}{2}L_{\mathbf{v}}\|\mathbf{v}_{s-1}-\mathbf{v}_s\|^2 \\
&= \phi(\mathbf{v}_s) - \frac{1}{2L_{\mathbf{v}}}\langle \nabla\phi_s(\mathbf{v}_s), \nabla\phi_s(\mathbf{v}_s) - \nabla\phi(\mathbf{v}_s)\rangle + \frac{3}{8L_{\mathbf{v}}}\|\nabla\phi_s(\mathbf{v}_s) - \nabla\phi(\mathbf{v}_s)\|^2 \\
&= \phi(\mathbf{v}_s) - \frac{1}{8L_{\mathbf{v}}}\|\nabla\phi_s(\mathbf{v}_s\|^2 - \frac{1}{4L_{\mathbf{v}}}\langle \nabla\phi_s(\mathbf{v}_s), \nabla\phi(\mathbf{v}_s)\rangle + \frac{3}{8L_{\mathbf{v}}}\|\nabla\phi(\mathbf{v}_s)\|^2
\end{aligned} \tag{18}$$

Rearranging terms, it yields

$$\begin{aligned}
&\phi(\mathbf{v}_s) - \phi(\mathbf{v}_{s-1}) \leq \frac{1}{8L_{\mathbf{v}}}\|\nabla\phi_s(\mathbf{v}_s)\|^2 + \frac{1}{4L_{\mathbf{v}}}\langle \nabla\phi_s(\mathbf{v}_s), \nabla\phi(\mathbf{v}_s)\rangle - \frac{3}{8L_{\mathbf{v}}}\|\nabla\phi(\mathbf{V}_s)\|^2 \\
&\leq \frac{1}{8L_{\mathbf{v}}}\|\nabla\phi_s(\mathbf{v}_s)\|^2 + \frac{1}{8L_{\mathbf{v}}}(\|\nabla\phi_s(\mathbf{v}_s)\|^2 + \|\nabla\phi(\mathbf{v}_s)\|^2) - \frac{3}{8L_{\mathbf{v}}}\|\nabla\phi(\mathbf{V}_s)\|^2 \\
&= \frac{1}{4L_{\mathbf{v}}}\|\nabla\phi_s(\mathbf{v}_s)\|^2 - \frac{1}{4L_{\mathbf{v}}}\|\nabla\phi(\mathbf{v}_s)\|^2 \\
&\leq \frac{1}{4L_{\mathbf{v}}}\|\nabla\phi_s(\mathbf{v}_s)\|^2 - \frac{\mu}{2L_{\mathbf{v}}}(\phi(\mathbf{v}_s) - \phi(\mathbf{v}_\phi^*))
\end{aligned} \tag{19}$$

Define $\Delta_s = \phi(\mathbf{v}_s) - \phi(\mathbf{v}_\phi^*)$. Combining (17) and (19), we get

$$E[\Delta_s - \Delta_{s-1}] \leq \frac{3}{2}E(\Delta_{s-1} - \Delta_s) + 3H\eta_s^2 {I'}_s^2 B^2 + \frac{6\eta_s(\sigma_{\mathbf{v}}^2+\sigma_\alpha^2)}{K} - \frac{\mu}{2L_{\mathbf{v}}}E[\Delta_s] \tag{20}$$

Therefore,

$$\left(\frac{5}{2} + \frac{\mu}{2L_{\mathbf{v}}}\right)E[\Delta_s] \leq \frac{5}{2}E[\Delta_{s-1}] + 3H\eta_s^2 {I'}_s^2 B^2 + \frac{6\eta_s(\sigma_{\mathbf{v}}^2+\sigma_\alpha^2)}{K} \tag{21}$$

Using $c = \frac{\mu/L_{\mathbf{v}}}{5+\mu/L_{\mathbf{v}}}$ as defined in the theorem,

$$\begin{aligned}
&E[\Delta_S] \leq \frac{5L_{\mathbf{v}}}{5L_{\mathbf{v}}+\mu}E[\Delta_{S-1}] + \frac{2L_{\mathbf{v}}}{5L_{\mathbf{v}}+\mu}\left[3H\eta_S^2 {I'}_S^2 B^2 + \frac{6\eta_S(\sigma_{\mathbf{v}}^2+\sigma_\alpha^2)}{K}\right] \\
&= (1-c)\left[E[\Delta_{S-1}] + \frac{2}{5}\left(3H\eta_S^2 {I'}_S^2 B^2 + \frac{6\eta_S(\sigma_{\mathbf{v}}^2+\sigma_\alpha^2)}{K}\right)\right] \\
&\leq (1-c)^S E[\Delta_0] + \frac{6HB^2}{5}\sum_{j=1}^{S}\eta_j^2 {I'}_j^2 (1-c)^{S+1-j} + \frac{12(\sigma_{\mathbf{v}}^2+\sigma_\alpha^2)}{5K}\sum_{j=1}^{S}\eta_j(1-c)^{S+1-j} \\
&= (1-c)^S E[\Delta_0] + \frac{6HB^2}{5}\sum_{j=1}^{S}\eta_j^2 {I'}_j^2 (1-c)^{S+1-j} + \frac{12(\sigma_{\mathbf{v}}^2+\sigma_\alpha^2)}{5K}\sum_{j=1}^{S}\eta_j(1-c)^{S+1-j}
\end{aligned} \tag{22}$$

We then have

$$\begin{aligned}
E[\Delta_S] &\leq (1-c)^S E[\Delta_0] + \left(\frac{6HB^2}{5K} + \frac{12(\sigma_{\mathbf{v}}^2 + \sigma_\alpha^2)}{5K}\right)\sum_{j=1}^{S}\eta_j(1-c)^{S+1-j} \\
&\leq \exp(-cS)\Delta_0 + \left(\frac{6HB^2}{5K} + \frac{12(\sigma_{\mathbf{v}}^2 + \sigma_\alpha^2)}{5K}\right)\sum_{j=1}^{S}\eta_j \exp(-c(S+1-j)) \\
&= \exp(-cS)\Delta_0 + \left(\frac{6HB^2}{5} + \frac{12(\sigma_{\mathbf{v}}^2 + \sigma_\alpha^2)}{5}\right)\eta_0 S\exp(-cS).
\end{aligned} \tag{23}$$

To achieve $E[\Delta_S] \leq \epsilon$, it suffices to make

$$\exp(-cS)\Delta_0 \leq \epsilon/2 \tag{24}$$

and

$$\left(\frac{6HB^2}{5} + \frac{12(\sigma_{\mathbf{v}}^2 + \sigma_\alpha^2)}{5}\right)\eta_0 S\exp(-cS) \leq \epsilon/2. \tag{25}$$

So, it suffices to make

$$S \geq c^{-1}\max\left\{\log\left(\frac{2\Delta_0}{\epsilon}\right), \log S + \log\left[\frac{2\eta_0}{\epsilon}\frac{6HB^2 + 12(\sigma_{\mathbf{v}}^2 + \sigma_\alpha^2)}{5}\right]\right\}. \tag{26}$$

Taking summation of iteration over $s = 1, ..., S$, we have the total iteration complexity as

$$\begin{aligned}
T &= \sum_{s=1}^{S} T_s \leq \frac{\max\{8, 8G_h^2\}}{L_{\mathbf{v}}\eta_0 K}\frac{\exp(cS)-1}{\exp(c)-1} \leq \frac{\max\{8, 8G_h^2\}}{L_{\mathbf{v}}\eta_0 K}\frac{5L_{\mathbf{v}}+\mu}{\mu}\exp(cS) \\
&= \tilde{O}\left(\max\left(\frac{\Delta_0}{\mu\epsilon\eta_0 K}, \frac{S(6HB^2 + 12(\sigma_{\mathbf{v}}^2 + \sigma_\alpha^2))}{\mu\epsilon K}\right)\right) = \tilde{O}\left(\max\left(\frac{\Delta_0}{\mu\epsilon\eta_0 K}, \frac{L_{\mathbf{v}}}{\mu^2 K\epsilon}\right)\right).
\end{aligned} \tag{27}$$

To analyze the total communication complexity, we will analyze two cases: (1) $\frac{1}{K\sqrt{\eta_0}} > 1$; (2) $\frac{1}{K\sqrt{\eta_0}} \leq 1$.

(1) If $\frac{1}{K\sqrt{\eta_0}} > 1$, then $I_s = \max(1, \frac{1}{K\sqrt{\eta_0}}\exp(\frac{c(s-1)}{2})) = \frac{1}{K\sqrt{\eta_0}}\exp(\frac{c(s-1)}{2})$ for any $s \geq 1$.

The total number of communications:

$$\begin{aligned}
\sum_{s=1}^{S}\frac{T_s}{I_s} &= \sum_{s=1}^{S}\frac{\max(8, 8G_h^2)}{L_{\mathbf{v}}\eta_0^{1/2}}\exp\left(\frac{c(s-1)}{2}\right) = \frac{\max(8, 8G_h^2)}{L_{\mathbf{v}}\eta_0^{1/2}}\frac{\exp(cS/2)-1}{\exp(c/2)-1} \\
&= \tilde{O}\left(\max\left(\frac{(2\Delta_0/\epsilon)^{1/2}}{\mu\eta_0^{1/2}}, \frac{(S(6HB^2 + 12(\sigma_{\mathbf{v}}^2 + \sigma_\alpha^2))^{1/2}}{\mu\epsilon^{1/2}}\right)\right) = \tilde{O}\left(\frac{\Delta_0^{1/2}}{\mu(\eta_0\epsilon)^{1/2}}, \frac{L_{\mathbf{v}}^{1/2}}{\mu^{3/2}\epsilon^{1/2}}\right).
\end{aligned} \tag{28}$$

(2) If $\frac{1}{K\sqrt{\eta_0}} \leq 1$, then $I_s = 1$ for $s \leq \lceil 2c^{-1}\log(K\sqrt{\eta_0}) + 1\rceil := S_1$ and $I_s = \frac{1}{K\sqrt{\eta_0}}\exp\left(\frac{s-1}{2}\right)$ for $s > \frac{2(5+\mu/L_{\mathbf{v}})}{\mu/L_{\mathbf{v}}}\log(K\sqrt{\eta_0}) + 1$.

Obviously, $S_1 \leq \frac{2(5+\mu/L_{\mathbf{v}})}{\mu/L_{\mathbf{v}}} \log(K\sqrt{\eta_0}) + 2$. The number of iterations from $s = 1$ to S_1 is

$$\begin{aligned}
\sum_{s=1}^{S_1} T_s &= \sum_{s=1}^{S_1} \frac{\max\{8, 8G_h^2\}}{\eta_0 L_{\mathbf{v}} K} \exp(c(s-1)) \\
&= \frac{\max\{8, 8G_h^2\}}{\eta_0 L_{\mathbf{v}} K} \frac{\exp(cS_1) - 1}{\exp(c) - 1} \\
&\leq c^{-1} \frac{\max\{8, 8G_h^2\}}{\eta_0 L_{\mathbf{v}} K} \exp\left(2\log(K\sqrt{\eta_0}) + 2c\right) \\
&= c^{-1} \frac{\max\{8, 8G_h^2\}}{\eta_0 L_{\mathbf{v}} K} K^2 \eta_0 \exp\left(\frac{2\mu/L_{\mathbf{v}}}{5+\mu/L_{\mathbf{v}}}\right) \\
&\leq c^{-1} \max\{8, 8G_h^2\} K \exp(2).
\end{aligned} \tag{29}$$

Thus, the total number of communications is

$$\begin{aligned}
&\sum_{s=1}^{S_1} T_s + \sum_{s=S_1+1}^{S} \frac{T_s}{I_s} \\
&= c^{-1} \max\{8, 8G_h^2\} K \exp(2) + \sum_{s=S_1+1}^{S} \frac{\max(8, 8G_h^2)}{L_{\mathbf{v}} \eta_0^{1/2}} \exp\left(\frac{s-1}{2} \frac{\mu/L_{\mathbf{v}}}{5+\mu/L_{\mathbf{v}}}\right) \\
&\leq c^{-1} \max\{8, 8G_h^2\} K \exp(2) + \sum_{s=1}^{S} \frac{\max(8, 8G_h^2)}{L_{\mathbf{v}} \eta_0^{1/2}} \exp\left(\frac{s-1}{2} \frac{\mu/L_{\mathbf{v}}}{5+\mu/L_{\mathbf{v}}}\right) \\
&\leq c^{-1} \max\{8, 8G_h^2\} K \exp(2) + \frac{\max(8, 8G_h^2)}{L_{\mathbf{v}} \eta_0^{1/2}} \frac{\exp(\frac{S}{2} \frac{\mu/L_{\mathbf{v}}}{5+\mu/L_{\mathbf{v}}}) - 1}{\exp(\frac{\mu/L_{\mathbf{v}}}{2(5+\mu/L_{\mathbf{v}})}) - 1} \\
&\in O\left(\max\left(\frac{K}{\mu} + \frac{\Delta_0}{\mu \eta_0^{1/2} \epsilon^{1/2}}, \frac{K}{\mu} + \frac{L_{\mathbf{v}}^{1/2}}{\mu^{3/2} \epsilon^{1/2}}\right)\right).
\end{aligned} \tag{30}$$

B. Proof of Lemma 1

To prove Lemma 1, we need the following Lemma 7 and Lemma 8 to show that the trajectories of α, a and b are constrained in closed sets in Algorithm 2.

Lemma 7 *Suppose Assumption (1) holds and $\eta \leq \frac{1}{2p(1-p)}$. Running Algorithm 2 with the input given by Algorithm 1, we have $|\alpha_t^k| \leq \frac{\max\{p,(1-p)\}}{p(1-p)}$ for any iteration t and any machine k.*

Proof. Firstly, we need to show that the input for any call of Algorithm (2) satisfies $|\alpha_0| \leq \frac{\max\{p,(1-p)\}}{p(1-p)}$. If Algorithm 2 is called by Algorithm 1 for the first time, we know $|\alpha_0| = 0 \leq \frac{\max\{p,(1-p)\}}{p(1-p)}$. Otherwise, by the update of $alpha_s$ in Algorithm (1) (lines 4-7), we know that the input for Algorithm (2) satisfies $|\alpha_0| \leq 2 \leq \frac{\max\{p,(1-p)\}}{p(1-p)}$ since $h(\mathbf{w}; \mathbf{x}^k) \in [0, 1]$ by Assumption 1(iv).

Next, we will show by induction that $|\alpha_t^k| \leq \frac{\max\{p,(1-p)\}}{p(1-p)}$ for any iteration t and any machine k in Algorithm 2. Obviously, $|a_0^k| \leq 2 \leq \frac{\max\{p,(1-p)\}}{p(1-p)}$ for any k.

Assume $|a_t^k| \leq \frac{\max\{p,(1-p)\}}{p(1-p)}$ for any k.

(1) If $t+1 \mod I \neq 0$, then we have

$$\begin{aligned}|\alpha_{t+1}^k| &= \left|\alpha_t^k + \eta(2(ph(\mathbf{w}_t^k;\mathbf{x})\mathbb{I}_{[y=-1]} - (1-p)h(\mathbf{w}_t^k;\mathbf{x})\mathbb{I}_{[y=1]}) - 2p(1-p)\alpha_t)\right| \\ &\leq \left|(1-2\eta p(1-p))\alpha_t^k\right| + \left|2\eta(ph(\mathbf{w}_t^k;\mathbf{x})\mathbb{I}_{[y=-1]} - (1-p)h(\mathbf{w}_t^k;\mathbf{x})\mathbb{I}_{[y=1]})\right| \\ &\leq (1-2\eta p(1-p))\frac{\max\{p,(1-p)\}}{p(1-p)} + 2\eta\max\{p,(1-p)\} \\ &= (1-2\eta p(1-p) + 2\eta p(1-p))\frac{\max\{p,(1-p)\}}{p(1-p)} \\ &= \frac{\max\{p,(1-p)\}}{p(1-p)}.\end{aligned} \tag{31}$$

(2) If $t+1 \mod I = 0$, then by same analysis as above, we know that $|\alpha_{t+1}^k| \leq \frac{\max\{p,(1-p)\}}{p(1-p)}$ before being averaged across machines. Therefore, after being averaged across machines, it still holds that $|\alpha_{t+1}^k| \leq \frac{\max\{p,(1-p)\}}{p(1-p)}$.

Therefore, $|\alpha_t^k| \leq \frac{\max\{p,(1-p)\}}{p(1-p)}$ holds for any iteration t and any machine k at any call of Algorithm (2). □

Lemma 8 *Suppose Assumption (1) (1) holds and $\eta \leq \min(\frac{1}{2(1-p)}, \frac{1}{2p})$. Running Algorithm 2 with the input given by Algorithm (1), we have that $|a_t^k| \leq 1$ and $|b_t^k| \leq 1$ for any iteration t and any machine k.*

Proof. At the first call of Algorithm (2), the input satisfies $|a_0| \leq 1$ and $|b_0| \leq 1$. Thus $|a_0^k| \leq 1$ and $|b_0^k| \leq 1$ for any machine k.

Assume $|a_t^k| \leq 1$ and $|b_t^k| \leq 1$. Then:

(1) $t+1 \mod I \neq 0$, then we have

$$\begin{aligned}|a_t^k| &= \left|\frac{\gamma}{\eta+\gamma}a_{t-1}^k + \frac{\eta}{\eta+\gamma}a_0 - \frac{\eta\gamma}{\eta+\gamma}\nabla_a F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k, \mathbf{z}_{t-1}^k)\right| \\ &= \left|\frac{\gamma}{\eta+\gamma}a_{t-1}^k + \frac{\eta}{\eta+\gamma}a_0 + \frac{\eta\gamma}{\eta+\gamma}(2(1-p)(h(\mathbf{w}_{t-1}^k;\mathbf{x}_{t-1}^k) - a_{t-1}^k))\mathbb{I}_{y^k=1}\right| \\ &= \left|\frac{\eta}{\eta+\gamma}a_0 + \frac{\gamma}{\eta+\gamma}a_{t-1}^k(1-2\eta(1-p))\mathbb{I}_{y^k=1} + \frac{\eta\gamma}{\eta+\gamma}2(1-p)h(\mathbf{w}_{t-1}^k;\mathbf{x}_{t-1}^k)\mathbb{I}_{y^k=1}\right| \\ &\leq \left|\frac{\eta}{\eta+\gamma}a_0\right| + \left|\frac{\gamma}{\eta+\gamma}a_{t-1}^k(1-2\eta(1-p))\mathbb{I}_{y^k=1}\right| + \left|\frac{\eta\gamma}{\eta+\gamma}2(1-p)h(\mathbf{w}_{t-1}^k;\mathbf{x}_{t-1}^k)\mathbb{I}_{y^k=1}\right| \\ &\leq \frac{\eta}{\eta+\gamma} + \frac{\gamma}{\eta+\gamma}(1-2\eta(1-p)) + \frac{\eta\gamma}{\eta+\gamma}2(1-p) \\ &= 1.\end{aligned} \tag{32}$$

(2) If $t+1 \mod I = 0$, then by the same analysis as above, we have that $|a_{t+1}^k| \leq 1$ before being averaged across machines. Therefore, after being averaged across machines, it still holds that $|a_{t+1}^k| \leq 1$.

Thus, we can see that $|a_t^k| \leq 1$ holds for any iteration t and any machine k in this call of Algorithm 2. Therefore, the output of the stage also has $|\tilde{a}| \leq 1$.

Then we know that in the next call of Algorithm (2), the input satisfies $|a_0| \leq 1$, by the same proof, we can see that $|a_t^k| \leq 1$ holds for any iteration t and any machine k in any call of Algorithm (2). With the same techniques, we can prove that $|b_t^k|$ holds for any iteration t and any machine k in any call of Algorithm (2). □

With the above lemmas, we are ready to prove Lemma 1 and derive the claimed constants.

By definition of $F(\mathbf{v},\alpha;\mathbf{z})$ and noting that $\mathbf{v} = (\mathbf{w}, a, b)$, we have

$$\nabla_{\mathbf{v}} F_k(\mathbf{v},\alpha;\mathbf{z}) = [\nabla_{\mathbf{w}} F_k(\mathbf{v},\alpha;\mathbf{z})^T, \nabla_a F_k(\mathbf{v},\alpha;\mathbf{z}), \nabla_b F_k(\mathbf{v},\alpha;\mathbf{z})]^T. \tag{33}$$

Addressing each of the three terms on RHS, it follows that

$$\begin{aligned}\nabla_{\mathbf{w}} F_k(\mathbf{v}, \alpha; \mathbf{z}) &= \Big[2(1-p)(h(\mathbf{w}; \mathbf{x}^k) - a) - 2(1+\alpha)(1-p)\Big] \nabla h(\mathbf{w}; \mathbf{x}^k) \mathbb{I}_{[y^k=1]} \\ &\quad + \Big[2p(h(\mathbf{w}; \mathbf{x}^k) - b) + 2(1+\alpha)p\Big] \nabla h(\mathbf{w}; \mathbf{x}^k) \mathbb{I}_{[y^k=-1]}, \\ \nabla_a F_k(\mathbf{v}, \alpha; \mathbf{z}) &= -2(1-p)(h(\mathbf{w}; x^k) - a)\mathbb{I}_{[y^k=1]}, \\ \nabla_b F_k(\mathbf{v}, \alpha; \mathbf{z}) &= -2p(h(\mathbf{w}; \mathbf{x}^k) - b).\end{aligned} \tag{34}$$

Since $|h(\mathbf{w}; \mathbf{x}^k)| \in [0, 1]$, $\|\nabla h(\mathbf{w}; \mathbf{x}^k)\| \le G_h$, $|\alpha| \le \frac{\max\{p, 1-p\}}{p(1-p)}$, $|a| \le 1$ and $b \le 1$, we have

$$\begin{aligned}\|\nabla_{\mathbf{w}} F_k(\mathbf{v}, \alpha; \mathbf{z})\| &\le \|2(1-p)(h(\mathbf{w}; \mathbf{x}^k) - a) - 2(1+\alpha)(1-p)\| G_h + \|2p(h(\mathbf{w}; \mathbf{x}^k) - b) + 2(1+\alpha)p\| G_h \\ &\le |6 + 2\alpha|(1-p)G_h + |6 + 2\alpha| p G_h \\ &\le \left(6 + 2\frac{\max\{p, 1-p\}}{p(1-p)}\right) G_h,\end{aligned} \tag{35}$$

$$\|\nabla_a F_k(\mathbf{v}, \alpha; \mathbf{z})\| \le 4(1-p), \tag{36}$$

$$\|\nabla_b F_k(\mathbf{v}, \alpha; \mathbf{z})\| \le 4p. \tag{37}$$

Thus,

$$\begin{aligned}\|\nabla_{\mathbf{v}} F_k(\mathbf{v}, \alpha; \mathbf{z})\|^2 &= \|\nabla_{\mathbf{w}} F_k(\mathbf{v}, \alpha; \mathbf{z})\|^2 + \|\nabla_a F_k(\mathbf{v}, \alpha; \mathbf{z})\|^2 + \|\nabla_b F_k(\mathbf{v}, \alpha; \mathbf{z})\|^2 \\ &\le \left(6 + \frac{2\max\{p, 1-p\}}{p(1-p)}\right)^2 G_h^2 + 16(1-p)^2 + 16p^2.\end{aligned} \tag{38}$$

$$\begin{aligned}\|\nabla_\alpha F_k(\mathbf{v}, \alpha; \mathbf{z})\|^2 &= \|2ph(\mathbf{w}; \mathbf{x}^k)\mathbb{I}_{y^k=-1} - 2(1-p)h(\mathbf{w}; \mathbf{x}^k)\mathbb{I}_{y^k=1} - 2p(1-p)\alpha\|^2 \\ &\le (2p + 2(1-p) + 4\max\{p, 1-p\})^2 = (2 + 4\max\{p, 1-p\})^2.\end{aligned} \tag{39}$$

Thus, $B_{\mathbf{v}}^2 = \left(6 + \frac{2\max\{p,1-p\}}{p(1-p)}\right)^2 G_h^2 + 16(1-p)^2 + 16p^2$ and $B_\alpha^2 = (2 + 4\max\{p, 1-p\})^2$.

It follow that

$$|\nabla_{\mathbf{v}} f_k(\mathbf{v}, \alpha)| = |E[\nabla_\alpha F_k(\mathbf{v}, \alpha; \mathbf{z}^k)]| \le B_{\mathbf{v}}. \tag{40}$$

Therefore,

$$E[\|\nabla_{\mathbf{v}} f_k(\mathbf{v}, \alpha) - \nabla_{\mathbf{v}} F_k(\mathbf{v}, \alpha; \mathbf{z}^k)\|^2] \le [2|\nabla_{\mathbf{v}} f_k(\mathbf{v}, \alpha)|^2 + 2|E[\nabla_{\mathbf{v}} F_k(\mathbf{v}, \alpha; \mathbf{z}^k)]|^2] \le 4B_{\mathbf{v}}^2. \tag{41}$$

Similarly,

$$|\nabla_\alpha f_k(\mathbf{w}, a, b, \alpha)| = |E[\nabla_\alpha F_k(\mathbf{w}, a, b, \alpha; \mathbf{z}^k)]| \le B_\alpha. \tag{42}$$

Therefore,

$$E[\|\nabla_\alpha f_k(\mathbf{v}, \alpha) - \nabla_\alpha F_k(\mathbf{v}, \alpha; \mathbf{z}^k)\|^2] \le 2|\nabla_\alpha f_k(\mathbf{v}, \alpha)|^2 + 2E[F_k(\mathbf{v}, \alpha; \mathbf{z}^k)]|^2 \le 4B_\alpha^2. \tag{43}$$

Thus, $\sigma_{\mathbf{v}}^2 = 4B_{\mathbf{v}}^2$ and $\sigma_\alpha^2 = 4B_\alpha^2$.

Now, it remains to derive the constant L_2 such that $\|\nabla_{\mathbf{v}} F_k(\mathbf{v}_1, \alpha; \mathbf{z}) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_2, \alpha; \mathbf{z})\| \leq L_2 \|\mathbf{v}_1 - \mathbf{v}_2\|$.

By (34), we get

$$
\begin{aligned}
&\|\nabla_{\mathbf{w}} F_k(\mathbf{v}_1, \alpha; \mathbf{z}) - \nabla_{\mathbf{w}} F_k(\mathbf{v}_2, \alpha; \mathbf{z})\| \\
&= \Bigg\| \Big[2(1-p)(h(\mathbf{w}_1; \mathbf{x}^k) - a_1) - 2(1+\alpha)(1-p)\Big] \nabla h(\mathbf{w}_1; \mathbf{x}^k) \mathbb{I}_{[y^k=1]} + \Big[2p(h(\mathbf{w}_1; \mathbf{x}^k) - b_1) + 2(1+\alpha)p\Big] \nabla h(\mathbf{w}_1; x^k) \mathbb{I}_{[y^k=-1]} \\
&\quad - \Big[2(1-p)(h(\mathbf{w}_2; \mathbf{x}^k) - a_2) - 2(1+\alpha)(1-p)\Big] \nabla h(\mathbf{w}_2; \mathbf{x}^k) \mathbb{I}_{[y^k=1]} - \Big[2p(h(\mathbf{w}_2; \mathbf{x}^k) - b_2) + 2(1+\alpha)p\Big] \nabla h(\mathbf{w}_2; \mathbf{x}^k) \mathbb{I}_{[y^k=-1]} \Big] \Bigg\| \\
&= \Bigg\| 2(1-p) \Big[h(\mathbf{w}_1; \mathbf{x}^k) \nabla h(\mathbf{w}_1; \mathbf{x}^k) - h(\mathbf{w}_2; \mathbf{x}^k) \nabla h(\mathbf{w}_2; \mathbf{x}^k)\Big] \mathbb{I}_{[y^k=1]} + 2p \Big[h(\mathbf{w}_1; \mathbf{x}^k) \nabla h(\mathbf{w}_1; \mathbf{x}^k) - h(\mathbf{w}_2; \mathbf{x}^k) \nabla h(\mathbf{w}_2; \mathbf{x}^k)\Big] \mathbb{I}_{[y^k=-1]} \\
&- (2(1+\alpha))(1-p)(\nabla h(\mathbf{w}_1; \mathbf{x}^k) - \nabla h(\mathbf{w}_2; \mathbf{x}^k)) \mathbb{I}_{[y^k=1]} + (2(1+\alpha)p)(\nabla h(\mathbf{w}_1; \mathbf{x}^k) - \nabla h(\mathbf{w}_2; \mathbf{x}^k)) \mathbb{I}_{[y^k=-1]} \\
&- 2(1-p)(a_1 \nabla h(\mathbf{w}_1; \mathbf{x}^k) - a_2 \nabla h(\mathbf{w}_2; \mathbf{x}^k)) \mathbb{I}_{y^k=1} - 2p(b_1 \nabla h(\mathbf{w}_1; \mathbf{x}^k) - b_2 \nabla h(\mathbf{w}_2; \mathbf{x}^k)) \mathbb{I}_{[y^k=-1]} \Bigg\| \\
&\leq 2(1-p) \|h(\mathbf{w}_1; \mathbf{x}^k) \nabla h(\mathbf{w}_1; \mathbf{x}^k) - h(\mathbf{w}_2; \mathbf{x}^k) \nabla h(\mathbf{w}_2; \mathbf{x}^k)\| + 2p \|h(\mathbf{w}_1; \mathbf{x}^k) \nabla h(\mathbf{w}_1; \mathbf{x}^k) - h(\mathbf{w}_2; \mathbf{x}^k) \nabla h(\mathbf{w}_2; \mathbf{x}^k)\| \\
&\quad + \|2(1+\alpha)(1-p)\| \|\nabla h(\mathbf{w}_1; \mathbf{x}^k) - \nabla h(\mathbf{w}_2; \mathbf{x}^k)\| + \|2(1+\alpha)p\| \|\nabla h(\mathbf{w}_1; \mathbf{x}^k) - \nabla h(\mathbf{w}_2; \mathbf{x}^k)\| \\
&\quad + 2(1-p) \|a_1 \nabla h(\mathbf{w}_1; \mathbf{x}^k) - a_2 \nabla h(\mathbf{w}_2; \mathbf{x}^k)\| + 2p \|b_1 \nabla h(\mathbf{w}_1; \mathbf{x}^k) - b_2 \nabla h(\mathbf{w}_2; \mathbf{x}^k)\|.
\end{aligned} \tag{44}
$$

Denoting $\Gamma_1(\mathbf{w}; \mathbf{x}^k) = h(\mathbf{w}; \mathbf{x}^k) \nabla h(\mathbf{w}; \mathbf{x}^k)$,

$$
\begin{aligned}
\|\nabla \Gamma_1(\mathbf{w}; \mathbf{x}^k)\| &= \|\nabla h(\mathbf{w}; \mathbf{x}^k) \nabla h(\mathbf{w}; \mathbf{x}^k)^T + h(\mathbf{w}; \mathbf{x}^k) \nabla^2 h(\mathbf{w}; \mathbf{x}^k)\| \\
&\leq \|\nabla h(\mathbf{w}; \mathbf{x}^k) \nabla h(\mathbf{w}; \mathbf{x}^k)^T\| + \|h(\mathbf{w}; \mathbf{x}^k) \nabla^2 h(\mathbf{w}; \mathbf{x}^k)\| \\
&\leq G_h^2 + L_h.
\end{aligned} \tag{45}
$$

Thus, $\|\Gamma_1(\mathbf{w}_1; \mathbf{x}^k) - \Gamma_1(\mathbf{w}_2; \mathbf{x}^k)\| = \|h(\mathbf{w}_1; \mathbf{x}^k) h'(\mathbf{w}_1; \mathbf{x}^k) - h(\mathbf{w}_2; \mathbf{x}^k) h'(\mathbf{w}_2; \mathbf{x}^k)\| \leq (G_h^2 + L_h) \|\mathbf{w}_1 - \mathbf{w}_2\|$. Define $\Gamma_2(\mathbf{w}, \alpha; \mathbf{x}^k) = a \nabla h(\mathbf{w}; \mathbf{x}^k)$. By Lemma 8 and Assumption 1, we have

$$
\nabla_{\mathbf{w},a} \Gamma_2(\mathbf{w}, a; \mathbf{x}^k) \leq \|\nabla_{\mathbf{w}} \Gamma_2(\mathbf{w}, a; \mathbf{z}^k)\| + \|\nabla_a \Gamma_2(\mathbf{w}, a; \mathbf{z}^k)\| = \|a \nabla^2 h(\mathbf{w}; \mathbf{x}^k)\| + \|\nabla h(\mathbf{w}; \mathbf{x}^k)\| \leq L_h + G_h. \tag{46}
$$

Therefore,

$$
\|\Gamma_2(\mathbf{w}_1, a_1; \mathbf{x}^k) - \Gamma_2(\mathbf{w}_2, a_2; \mathbf{x}^k)\| = \|a_1 \nabla h(\mathbf{w}_1; \mathbf{x}^k) - a_2 \nabla h(\mathbf{w}_2; \mathbf{x}^k)\| \leq (L_h + G_h) \sqrt{\|\mathbf{w}_1 - \mathbf{w}_2\|^2 + \|a_1 - a_2\|^2}. \tag{47}
$$

Similarly, we can prove that

$$
\|b_1 \nabla h(\mathbf{w}_1; \mathbf{x}^k) - b_2 \nabla h(\mathbf{w}_2; \mathbf{x}^k)\| \leq (L_h + G_h) \sqrt{\|\mathbf{w}_1 - \mathbf{w}_2\|^2 + \|b_1 - b_2\|^2}. \tag{48}
$$

Then plugging (47), (48) and Assumption 1 into (44), we have

$$
\begin{aligned}
&\|\nabla_{\mathbf{w}} F_k(\mathbf{v}_1, \alpha; \mathbf{z}) - \nabla_{\mathbf{w}} F_k(\mathbf{v}_2, \alpha; \mathbf{z})\| \\
&\leq 2(G_h^2 + L_h) \|\mathbf{w}_1 - \mathbf{w}_2\| + 2|1+\alpha| G_h \|\mathbf{w}_1 - \mathbf{w}_2\| \\
&\quad + (L_h + G_h) \sqrt{\|\mathbf{w}_1 - \mathbf{w}_2\|^2 + \|a_1 - a_2\|^2} + (L_h + G_h) \sqrt{\|\mathbf{w}_1 - \mathbf{w}_2\|^2 + \|b_1 - b_2\|^2} \\
&\leq (2(G_h^2 + L_h) + |2(1+\alpha)| G_h + 2L_h + 2G_h) \sqrt{\|\mathbf{w}_1 - \mathbf{w}_2\|^2 + \|a_1 - a_2\|^2 + \|b_1 - b_2\|^2} \\
&\leq \left(2G_h^2 + 4L_h + \left(4 + \frac{2\max\{p, 1-p\}}{p(1-p)} \right) G_h \right) \|\mathbf{v}_1 - \mathbf{v}_2\|.
\end{aligned} \tag{49}
$$

By (34), we also have

$$
\begin{aligned}
&\|\nabla_a F_k(\mathbf{v}_1, \alpha; \mathbf{z}) - \nabla_a F_k(\mathbf{v}_2, \alpha; \mathbf{z})\|^2 \leq 4(1-p)^2 (\|h(\mathbf{w}_1; \mathbf{x}^k) - h(\mathbf{w}_2; \mathbf{x}^k)\|^2 + \|a_1 - a_2\|^2) \\
&\leq 4(1-p)^2 (G_h^2 \|\mathbf{w}_1 - \mathbf{w}_2\|^2 + \|a_1 - a_2\|^2 + \|b_1 - b_2\|^2) \leq 4(1-p)^2 (G_h^2 + 1) \|\mathbf{v}_1 - \mathbf{v}_2\|^2,
\end{aligned} \tag{50}
$$

and

$$\|\nabla_b F_k(\mathbf{v}_1,\alpha;\mathbf{z}) - \nabla_b F_k(\mathbf{v}_2,\alpha;\mathbf{z})\|^2 \leq 4(1-p)^2(\|h(\mathbf{w}_1;\mathbf{x}^k) - h(\mathbf{w}_2;\mathbf{x}^k)\|^2 + \|b_1-b_2\|^2) \\ \leq 4(1-p)^2(G_h^2\|\mathbf{w}_1-\mathbf{w}_2\|^2 + \|a_1-a_2\|^2 + \|b_1-b_2\|^2) \leq 4(1-p)^2(G_h^2+1)\|\mathbf{v}_1-\mathbf{v}_2\|^2. \tag{51}$$

$$\begin{aligned}\|\nabla_{\mathbf{v}} F_k(\mathbf{v}_1,\alpha;\mathbf{z}) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_2,\alpha;\mathbf{z})\|^2 &= \|\nabla_{\mathbf{w}} F_k(\mathbf{v}_1,\alpha;\mathbf{z}) - \nabla_{\mathbf{w}} F_k(\mathbf{v}_2,\alpha;\mathbf{z})\|^2 \\ &\quad + \|\nabla_a F_k(\mathbf{v}_1,\alpha;\mathbf{z}) - \nabla_b F_k(\mathbf{v}_2,\alpha;\mathbf{z})\|^2 + \|\nabla_b F_k(\mathbf{v}_1,\alpha;\mathbf{z}) - \nabla_b F_k(\mathbf{v}_1,\alpha;\mathbf{z})\|^2 \\ &\leq \left(G_h^2 + L_h + 4 + \frac{2\max\{p,1-p\}}{p(1-p)} 8(1-p)^2(G_h^2+1)\right)\|\mathbf{v}_1-\mathbf{v}_2\|^2.\end{aligned} \tag{52}$$

Thus, we get $L_2 = \left(G_h^2 + L_h + 4 + \frac{2\max\{p,1-p\}}{p(1-p)} 8(1-p)^2(G_h^2+1)\right)^{1/2}$.

C. Proof of Lemma 2

Proof. Plugging Lemma 4 and Lemma 5 into Lemma 3, we get

$$\begin{aligned}&\psi(\tilde{\mathbf{v}}) - \psi(\mathbf{v}_\psi^*) \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\Bigg[\underbrace{\left(\frac{L_{\mathbf{v}} + 3G_\alpha^2/\mu_\alpha}{2} - \frac{1}{2\eta}\right)\|\bar{\mathbf{v}}_{t-1} - \bar{\mathbf{v}}_t\|^2 + \left(\frac{L_\alpha + 3G_{\mathbf{v}}^2/L_{\mathbf{v}}}{2} - \frac{1}{2\eta}\right)(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2}_{C_1} \\ &+ \underbrace{\left(\frac{1}{2\eta} - \frac{\mu_\alpha}{3}\right)(\bar{\alpha}_{t-1} - \alpha^*(\tilde{\mathbf{v}}))^2 - \left(\frac{1}{2\eta} - \frac{\mu_\alpha}{3}\right)(\bar{\alpha}_t - \alpha^*(\tilde{\mathbf{v}}))^2}_{C_2} + \underbrace{\left(\frac{2L_{\mathbf{v}}}{3} + \frac{1}{2\eta}\right)\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_\psi^*\|^2 - \left(\frac{1}{2\eta} + \frac{2L_{\mathbf{v}}}{3}\right)\|\bar{\mathbf{v}}_t - \mathbf{v}_\psi^*\|^2}_{C_3} \\ &+ \underbrace{\frac{1}{2\eta}((\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_{t-1})^2 - (\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_t)^2)}_{C_4} + \underbrace{\left(\frac{3G_{\mathbf{v}}^2}{2\mu_\alpha} + \frac{3L_{\mathbf{v}}}{2}\right)\frac{1}{K}\sum_{k=1}^{K}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{t-1}^k\|^2 + \left(\frac{3G_\alpha^2}{2L_{\mathbf{v}}} + \frac{3L_\alpha^2}{2\mu_\alpha}\right)\frac{1}{K}\sum_{k=1}^{K}(\bar{\alpha}_{t-1} - \alpha_{t-1}^k)^2}_{C_5} \\ &+ \underbrace{\eta\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)]\right\|^2}_{C_6} + \underbrace{\frac{3\eta}{2}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)\right\|^2}_{C_7} \\ &+ \underbrace{\left\langle\frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \hat{\mathbf{v}}_t - \mathbf{v}_\psi^*\right\rangle}_{C_8} + \underbrace{\left\langle\frac{1}{K}\sum_{k=1}^{K}[\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \tilde{\alpha}_{t-1} - \hat{\alpha}_t\right\rangle}_{C_9}\Bigg].\end{aligned} \tag{53}$$

Since $\eta \leq \min(\frac{1}{L_{\mathbf{v}}+3G_\alpha^2/\mu_\alpha}, \frac{1}{L_\alpha+3G_{\mathbf{v}}^2/L_{\mathbf{v}}})$, thus in the RHS of (53), C_1 can be cancelled. C_2, C_3 and C_4 will be handled by telescoping sum. C_5 can be bounded by Lemma 6.

Taking expectation over C_6,

$$
\begin{aligned}
&E\left[\eta\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}}f_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k)-\nabla_{\mathbf{v}}F_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k;\mathbf{z}_{t-1}^k)]\right\|^2\right]\\
&=E\left[\frac{\eta}{K^2}\left\|\sum_{k=1}^{K}[\nabla_{\mathbf{v}}f_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k)-\nabla_{\mathbf{v}}F_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k;\mathbf{z}_{t-1}^k)]\right\|^2\right]\\
&=E\left[\frac{\eta}{K^2}\left(\sum_{k=1}^{K}\|\nabla_{\mathbf{v}}f_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k)-\nabla_{\mathbf{v}}F_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k;\mathbf{z}_{t-1}^k)\|^2\right.\right.\\
&\quad\left.\left.+2\sum_{i=1}^{K}\sum_{j=i+1}^{K}\left\langle\nabla_{\mathbf{v}}f_i(\mathbf{v}_{t-1}^i,\alpha_{t-1}^i)-\nabla_{\mathbf{v}}F_i(\mathbf{v}_{t-1}^i,\alpha_{t-1}^i;\mathbf{z}_{t-1}^i),\nabla_{\mathbf{v}}f_j(\mathbf{v}_{t-1}^j,\alpha_{t-1}^j)-\nabla_{\mathbf{v}}F_j(\mathbf{v}_{t-1}^j,\alpha_{t-1}^j;\mathbf{z}_{t-1}^j)\right\rangle\right)\right]\\
&\leq\frac{\eta\sigma_{\mathbf{v}}^2}{K}.
\end{aligned}
\tag{54}
$$

The last inequality holds because $\|\nabla_{\mathbf{v}}f_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k)-\nabla_{\mathbf{v}}F_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k;\mathbf{z}_{t-1}^k)\|^2 \leq \sigma_{\mathbf{v}}^2$ for any k and $E\left[\left\langle\nabla_{\mathbf{v}}f_i(\mathbf{v}_{t-1}^i,\alpha_{t-1}^i)-\nabla_{\mathbf{v}}F_i(\mathbf{v}_{t-1}^i,\alpha_{t-1}^i;\mathbf{z}_{t-1}^i),\nabla_{\mathbf{v}}f_j(\mathbf{v}_{t-1}^j,\alpha_{t-1}^j)-\nabla_{\mathbf{v}}F_j(\mathbf{v}_{t-1}^j,\alpha_{t-1}^j;\mathbf{z}_{t-1}^j)\right\rangle\right]=0$ for any $i\neq j$ as each machine draws data independently. Similarly, we take expectation over C_7 and have

$$
E\left[\frac{3\eta}{2}\left(\frac{1}{K}\sum_{k=1}^{K}[\nabla_{\alpha}f_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k)-\nabla_{\alpha}F_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k;\mathbf{z}_{t-1}^k)]\right)^2\right]\leq\frac{3\eta\sigma_{\alpha}^2}{2K}. \tag{55}
$$

Note $E\left[\left\langle\frac{1}{K}\sum\limits_{k=1}^{K}[\nabla_{\mathbf{v}}f_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k)-\nabla_{\mathbf{v}}F_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k;\mathbf{z}_{t-1}^k)],\hat{\mathbf{v}}_t-\mathbf{v}_{\psi}^*\right\rangle\right]=0$ and $E\left[\left\langle-\frac{1}{K}\sum\limits_{k=1}^{K}[\nabla_{\alpha}f_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k)-F_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k;\mathbf{z}_{t-1}^k)],\tilde{\alpha}_{t-1}-\hat{\alpha}_t\right\rangle\right]=0$. Therefore, C_8 and C_9 will diminish after taking expectation.

As $\eta\leq\frac{1}{L_{\mathbf{v}}+3G_{\alpha}^2/\mu_{\alpha}}$, we have $L_{\mathbf{v}}\leq\frac{1}{\eta}$. Plugging (54) and (55) into (53), and taking expectation, it yields

$$
\begin{aligned}
E[\psi(\tilde{\mathbf{v}})-\psi(\mathbf{v}_{\psi}^*)]\leq E\bigg\{&\frac{1}{T}\left(\frac{2L_{\mathbf{v}}}{3}+\frac{1}{2\eta}\right)\|\bar{\mathbf{v}}_0-\mathbf{v}_{\psi}^*\|^2+\frac{1}{T}\left(\frac{1}{2\eta}-\frac{\mu_{\alpha}}{3}\right)(\bar{\alpha}_0-\alpha^*(\tilde{\mathbf{v}}))^2+\frac{1}{2\eta T}(\tilde{\alpha}_0-\alpha^*(\tilde{\mathbf{v}}))^2\\
&+\frac{1}{T}\sum_{t=1}^{T}\left(\frac{3G_{\mathbf{v}}^2}{2\mu_{\alpha}}+\frac{3L_{\mathbf{v}}}{2}\right)\frac{1}{K}\sum_{k=1}^{K}\|\bar{\mathbf{v}}_{t-1}-\mathbf{v}_{t-1}^k\|^2+\frac{1}{T}\sum_{t=1}^{T}\left(\frac{3G_{\alpha}^2}{2L_{\mathbf{v}}}+\frac{3L_{\alpha}^2}{2\mu_{\alpha}}\right)\frac{1}{K}\sum_{k=1}^{K}\|\bar{\alpha}_{t-1}-\alpha_{t-1}^k\|^2\\
&+\frac{1}{T}\sum_{t=1}^{T}\frac{\eta\sigma_{\mathbf{v}}^2}{K}+\frac{1}{T}\sum_{t=1}^{T}\frac{3\eta\sigma_{\alpha}^2}{2K}\bigg\}\\
\leq\ &\frac{2}{\eta T}\|\mathbf{v}_0-\mathbf{v}_{\psi}^*\|^2+\frac{1}{\eta T}(\alpha_0-\alpha^*(\tilde{\mathbf{v}}))^2+\left(\frac{6G_{\mathbf{v}}^2}{\mu_{\alpha}}+6L_{\mathbf{v}}+\frac{6G_{\alpha}^2}{L_{\mathbf{v}}}+\frac{6L_{\alpha}^2}{\mu_{\alpha}}\right)\eta^2I^2B^2\mathbb{I}_{I>1}+\frac{\eta(2\sigma_{\mathbf{v}}^2+3\sigma_{\alpha}^2)}{2K},
\end{aligned}
$$

where we use Lemma 6, $\mathbf{v}_0=\bar{\mathbf{v}}_0$, $\alpha_0=\bar{\alpha}_0$ and $B^2=\max\{B_{\mathbf{v}}^2,B_{\alpha}^2\}$ in the last inequality. □

D. Proof of Lemma 3

Proof. Define $\alpha^*(\tilde{\mathbf{v}}) = \arg\max_\alpha f(\tilde{\mathbf{v}}, \alpha)$ and $\tilde{\alpha} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{T}\sum_{t=1}^{T}\alpha_t^k$.

$$\begin{aligned}
\psi(\tilde{\mathbf{v}}) - \min_{\mathbf{v}}\psi(\mathbf{v}) &= \max_\alpha\left[f(\tilde{\mathbf{v}}, \alpha) + \frac{1}{2\gamma}\|\tilde{\mathbf{v}} - \mathbf{v}_0\|^2\right] - \min_{\mathbf{v}}\max_\alpha\left[f(\mathbf{v}, \alpha) + \frac{1}{2\gamma}\|\mathbf{v} - \mathbf{v}_0\|^2\right] \\
&= \left[f(\tilde{\mathbf{v}}, \alpha^*(\tilde{\mathbf{v}})) + \frac{1}{2\gamma}\|\tilde{\mathbf{v}} - \mathbf{v}_0\|^2\right] - \max_\alpha\left[f(\mathbf{v}_\psi^*, \alpha) + \frac{1}{2\gamma}\|\mathbf{v}_\psi^* - \mathbf{v}_0\|^2\right] \\
&\leq \left[f(\tilde{\mathbf{v}}, \alpha^*(\tilde{\mathbf{v}})) + \frac{1}{2\gamma}\|\tilde{\mathbf{v}} - \mathbf{v}_0\|^2\right] - \left[f(\mathbf{v}_\psi^*, \tilde{\alpha}) + \frac{1}{2\gamma}\|\mathbf{v}_\psi^* - \mathbf{v}_0\|^2\right] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left[\left(f(\bar{\mathbf{v}}_t, \alpha^*(\tilde{\mathbf{v}})) + \frac{1}{2\gamma}\|\bar{\mathbf{v}}_t - \mathbf{v}_0\|^2\right) - \left(f(\mathbf{v}_\psi^*, \bar{\alpha}_t) + \frac{1}{2\gamma}\|\mathbf{v}_\psi^* - \mathbf{v}_0\|^2\right)\right],
\end{aligned} \tag{56}$$

where the last inequality uses Jensen's inequality and the fact that $f(\mathbf{v}, \alpha) + \frac{1}{2\gamma}\|\mathbf{v} - \mathbf{v}_0\|^2$ is convex w.r.t. $\mathbf{v}$ and concave w.r.t. α.

By $L_{\mathbf{v}}$-weakly convexity of $f(\cdot)$ w.r.t. $\mathbf{v}$, we have

$$f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) + \langle\nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \mathbf{v}_\psi^* - \bar{\mathbf{v}}_{t-1}\rangle - \frac{L_{\mathbf{v}}}{2}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_\psi^*\|^2 \leq f(\mathbf{v}_\psi^*, \bar{\alpha}_{t-1}), \tag{57}$$

and by $L_{\mathbf{v}}$-smoothness of $f(\cdot)$ w.r.t. $\mathbf{v}$, we have

$$\begin{aligned}
f(\bar{\mathbf{v}}_t, \alpha^*(\tilde{\mathbf{v}})) &\leq f(\bar{\mathbf{v}}_{t-1}, \alpha^*(\tilde{\mathbf{v}})) + \langle\nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \alpha^*(\tilde{\mathbf{v}})), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle + \frac{L_{\mathbf{v}}}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 \\
&= f(\bar{\mathbf{v}}_{t-1}, \alpha^*(\tilde{\mathbf{v}})) + \langle\nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \alpha^*(\tilde{\mathbf{v}})), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle + \frac{L_{\mathbf{v}}}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 \\
&\quad + \langle\nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle - \langle\nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle \\
&= f(\bar{\mathbf{v}}_{t-1}, \alpha^*(\tilde{\mathbf{v}})) + \langle\nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle + \frac{L_{\mathbf{v}}}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 \\
&\quad + \langle\nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \alpha^*(\tilde{\mathbf{v}})) - \nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle \\
&\overset{(a)}{\leq} f(\bar{\mathbf{v}}_{t-1}, \alpha^*(\tilde{\mathbf{v}})) + \langle\nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle + \frac{L_{\mathbf{v}}}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 \\
&\quad + G_\alpha|\bar{\alpha}_{t-1} - \alpha^*(\tilde{\mathbf{v}})|\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\| \\
&\overset{(b)}{\leq} f(\bar{\mathbf{v}}_{t-1}, \alpha^*(\tilde{\mathbf{v}})) + \langle\nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle + \frac{L_{\mathbf{v}}}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 \\
&\quad + \frac{\mu_\alpha}{6}|\bar{\alpha}_{t-1} - \alpha^*(\tilde{\mathbf{v}})|^2 + \frac{3G_\alpha^2}{2\mu_\alpha}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2,
\end{aligned} \tag{58}$$

where (a) holds because we know that $\nabla_{\mathbf{v}} f(\cdot)$ is $G_\alpha = 2\max\{p, 1-p\}$-Lipshitz w.r.t. α by the definition of $f(\cdot)$, and (b) holds by Young's inequality.

By $\frac{1}{\gamma}$-strong convexity of $\frac{1}{2\gamma}\|\mathbf{v} - \mathbf{v}_0\|^2$ w.r.t. $\mathbf{v}$, we have

$$\frac{1}{2\gamma}\|\bar{\mathbf{v}}_t - \mathbf{v}_0\|^2 + \frac{1}{\gamma}\langle\bar{\mathbf{v}}_t - \mathbf{v}_0, \mathbf{v}_\psi^* - \mathbf{v}_t\rangle + \frac{1}{2\gamma}\|\mathbf{v}_\psi^* - \mathbf{v}_t\|^2 \leq \frac{1}{2\gamma}\|\mathbf{v}_\psi^* - \mathbf{v}_0\|^2. \tag{59}$$

Adding (57), (58), (59), and rearranging terms, we have

$$\begin{aligned}
&f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) + f(\bar{\mathbf{v}}_t, \alpha^*(\tilde{\mathbf{v}})) + \frac{1}{2\gamma}\|\bar{\mathbf{v}}_t - \mathbf{v}_0\|^2 - \frac{1}{2\gamma}\|\mathbf{v}_\psi^* - \mathbf{v}_0\|^2 \\
\leq& f(\mathbf{v}_\psi^*, \bar{\alpha}_{t-1}) + f(\bar{\mathbf{v}}_{t-1}, \alpha^*(\tilde{\mathbf{v}})) + \langle\nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_\psi^*\rangle + \frac{L_{\mathbf{v}} + 3G_\alpha^2/\mu_\alpha}{2}\eta^2\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 + \frac{L_{\mathbf{v}}}{2}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_\psi^*\|^2 \\
&+ \frac{\mu_\alpha}{6}(\bar{\alpha}_{t-1} - \alpha^*(\tilde{\mathbf{v}})) - \frac{1}{2\gamma}\|\mathbf{v}_\psi^* - \mathbf{v}_t\|^2 + \frac{1}{\gamma}\langle\bar{\mathbf{v}}_t - \mathbf{v}_0, \mathbf{v}_t - \mathbf{v}_\psi^*\rangle.
\end{aligned} \tag{60}$$

By definition, we know that $f(\cdot)$ is $\mu_\alpha := 2p(1-p)$-strong concavity w.r.t. α ($-f(\cdot)$ is μ_α-strong convexity w.r.t. α). Thus, we have

$$-f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - \nabla_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1})^T(\alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_{t-1}) + \frac{\mu_\alpha}{2}(\alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_{t-1})^2 \leq -f(\bar{\mathbf{v}}_{t-1}, \alpha^*(\tilde{\mathbf{v}})) \tag{61}$$

By definition, we know that $f(\cdot)$ is smooth in α (with coefficient $L_\alpha := 2p(1-p)$), we get

$$\begin{aligned}
&- f(\mathbf{v}^*_\psi, \bar{\alpha}_t) \leq -f(\mathbf{v}^*_\psi, \bar{\alpha}_{t-1}) - \langle \nabla_\alpha f(\mathbf{v}^*_\psi, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \bar{\alpha}_{t-1} \rangle + \frac{L_\alpha}{2}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 \\
&= -f(\mathbf{v}^*_\psi, \bar{\alpha}_{t-1}) - \langle \nabla_\alpha f(\mathbf{v}^*_\psi, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \bar{\alpha}_{t-1} \rangle + \frac{L_\alpha}{2}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 \\
&\quad - \langle \nabla_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \bar{\alpha}_{t-1} \rangle + \langle \nabla_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \bar{\alpha}_{t-1} \rangle \\
&\overset{(a)}{\leq} -f(\mathbf{v}^*_\psi, \bar{\alpha}_{t-1}) - \langle \nabla_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \bar{\alpha}_{t-1} \rangle + \frac{L_\alpha}{2}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 + G_{\mathbf{v}} |\langle \mathbf{v}^*_\psi - \bar{\mathbf{v}}_{t-1}, \bar{\alpha}_t - \bar{\alpha}_{t-1} \rangle| \\
&\leq -f(\mathbf{v}^*_\psi, \bar{\alpha}_{t-1}) - \langle \nabla_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \bar{\alpha}_{t-1} \rangle + \frac{L_\alpha}{2}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 + \frac{L_{\mathbf{v}}}{6}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}^*_\psi\|^2 + \frac{3G_{\mathbf{v}}^2}{2L_{\mathbf{v}}}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2,
\end{aligned} \tag{62}$$

where (a) holds because $\nabla_\alpha f(\cdot)$ is Lipshitz in α with coefficient $G_{\mathbf{v}} = 2\max\{p, 1-p\}G_h$ by definition of $f(\cdot)$.

Adding (61), (62) and arranging terms, we have

$$\begin{aligned}
-f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - f(\mathbf{v}^*_\psi, \bar{\alpha}_t) \leq -f(\bar{\mathbf{v}}_{t-1}, \alpha^*(\tilde{\mathbf{v}})) - f(\mathbf{v}^*_\psi, \bar{\alpha}_{t-1}) - \langle \nabla_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \alpha^*(\tilde{\mathbf{v}}) \rangle + \frac{L_\alpha}{2}\|\bar{\alpha}_t - \bar{\alpha}_{t-1}\|^2 \\
+ \frac{L_{\mathbf{v}}}{6}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}^*_\psi\|^2 + \frac{3G_{\mathbf{v}}^2}{2L_{\mathbf{v}}}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 - \frac{\mu_\alpha}{2}(\alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_{t-1})^2.
\end{aligned} \tag{63}$$

Adding (60) and (63), we get

$$\begin{aligned}
&\left[f(\bar{\mathbf{v}}_t, \alpha^*(\tilde{\mathbf{v}})) + \frac{1}{2\gamma}\|\bar{\mathbf{v}}_t - \mathbf{v}_0\|^2 \right] - \left[f(\mathbf{v}^*_\psi, \bar{\alpha}_t) + \frac{1}{2\gamma}\|\mathbf{v}^*_\psi - \mathbf{v}_0\|^2 \right] \leq \\
&\langle \nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \mathbf{v}^*_\psi \rangle - \langle \nabla_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \alpha^*(\tilde{\mathbf{v}}) \rangle \\
&+ \frac{L_{\mathbf{v}} + 3G_\alpha^2/\mu_\alpha}{2}\eta^2\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 + \left(\frac{L_{\mathbf{v}}}{6} + \frac{L_{\mathbf{v}}}{2}\right)\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}^*_\psi\|^2 - \frac{1}{2\gamma}\|\mathbf{v}^*_\psi - \mathbf{v}_t\|^2 \\
&+ \frac{L_\alpha + 3G_{\mathbf{v}}^2/L_{\mathbf{v}}}{2}\eta^2\|\bar{\alpha}_t - \bar{\alpha}_{t-1}\|^2 - \frac{\mu_\alpha}{3}(\bar{\alpha}_{t-1} - \alpha^*(\tilde{\mathbf{v}}))^2 \\
&+ \frac{1}{\gamma}\langle \bar{\mathbf{v}}_t - \mathbf{v}_0, \bar{\mathbf{v}}_t - \mathbf{v}^*_\psi \rangle.
\end{aligned} \tag{64}$$

Applying $\gamma = \frac{1}{2L_{\mathbf{v}}}$ to (64) and then plugging it into (56), we get

$$\begin{aligned}
\psi(\tilde{\mathbf{v}}) - \min_{\mathbf{v}} \psi(\mathbf{v}) \leq \frac{1}{T}\sum_{t=1}^{T} \Bigg[&\langle \nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \mathbf{v}^*_\psi \rangle + 2L_{\mathbf{v}}\langle \bar{\mathbf{v}}_t - \mathbf{v}_0, \bar{\mathbf{v}}_t - \mathbf{v}^*_\psi \rangle + \langle \nabla_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t \rangle \\
&+ \frac{L_{\mathbf{v}} + 3G_\alpha^2/\mu_\alpha}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 + \frac{L_\alpha + 3G_{\mathbf{v}}^2/L_{\mathbf{v}}}{2}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 \\
&+ \frac{2L_{\mathbf{v}}}{3}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}^*_\psi\|^2 - L_{\mathbf{v}}\|\bar{\mathbf{v}}_t - \mathbf{v}^*_\psi\|^2 - \frac{\mu_\alpha}{3}(\bar{\alpha}_{t-1} - \alpha^*(\tilde{\mathbf{v}}))^2 \Bigg]. \ \square
\end{aligned}$$

E. Proof of Lemma 4

Proof. According to the update rule of $\mathbf{v}$ and taking $\gamma = \frac{1}{2L_{\mathbf{v}}}$, we have

$$2L_{\mathbf{v}}(\mathbf{v}^k_t - \mathbf{v}_0) = -\nabla_{\mathbf{v}} F_k(\mathbf{v}^k_{t-1}, \alpha^k_{t-1}; \mathbf{z}^k_{t-1}) - \frac{1}{\eta}(\mathbf{v}^k_t - \mathbf{v}^k_{t-1}). \tag{65}$$

Taking average over K machines, we have

$$2L_{\mathbf{v}}(\bar{\mathbf{v}}_t - \mathbf{v}_0) = -\frac{1}{K}\sum_{k=1}^{K} \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) - \frac{1}{\eta}(\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}). \tag{66}$$

It follows that

$$\begin{aligned}
&\langle \nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \mathbf{v}_\psi^* \rangle + 2L_{\mathbf{v}} \langle \bar{\mathbf{v}}_t - \mathbf{v}_0, \bar{\mathbf{v}}_t - \mathbf{v}_\psi^* \rangle \\
&= \left\langle \frac{1}{K}\sum_{k=1}^{K} \nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \mathbf{v}_\psi^* \right\rangle - \left\langle \frac{1}{K}\sum_{k=1}^{K} \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k), \bar{\mathbf{v}}_t - \mathbf{v}_\psi^* \right\rangle + \frac{1}{\eta}\langle \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}, \bar{\mathbf{v}}_t - \mathbf{v}_\psi^* \rangle \\
&\leq \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - \nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_{t-1}, \alpha_{t-1}^k)], \bar{\mathbf{v}}_t - \mathbf{v}_\psi^* \right\rangle \qquad ① \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_{t-1}, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)], \bar{\mathbf{v}}_t - \mathbf{v}_\psi^* \right\rangle \qquad ② \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \bar{\mathbf{v}}_t - \mathbf{v}_\psi^* \right\rangle \qquad ③ \\
&\quad + \frac{1}{2\eta}(\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_\psi^*\|^2 - \|\bar{\mathbf{v}}_{t-1} - \bar{\mathbf{v}}_t\|^2 - \|\bar{\mathbf{v}}_t - \mathbf{v}_\psi^*\|^2).
\end{aligned} \tag{67}$$

Then we will bound ①, ② and ③ separately,

$$\begin{aligned}
① &\overset{(a)}{\leq} \frac{3}{2L_{\mathbf{v}}} \left\| \frac{1}{K}\sum_{k=1}^{K} [\nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - \nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_{t-1}, \alpha_{t-1}^k)] \right\|^2 + \frac{L_{\mathbf{v}}}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}_\psi^*\|^2 \\
&\overset{(b)}{\leq} \frac{3}{2L_{\mathbf{v}}} \frac{1}{K}\sum_{k=1}^{K} \|\nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - \nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_{t-1}, \alpha_{t-1}^k)\|^2 + \frac{L_{\mathbf{v}}}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}_\psi^*\|^2 \\
&\overset{(c)}{\leq} \frac{3G_\alpha^2}{2L_{\mathbf{v}}} \frac{1}{K}\sum_{k=1}^{K} \|\bar{\alpha}_{t-1} - \alpha_{t-1}^k\|^2 + \frac{L_{\mathbf{v}}}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}_\psi^*\|^2,
\end{aligned} \tag{68}$$

where (a) follows from Young's inequality and (b) follows from Jensen's inequality. (c) holds because $\nabla_{\mathbf{v}} f_k(\mathbf{v}, \alpha)$ is Lipschitz in α with coefficient $G_\alpha = 2\max(p, 1-p)$ for any $\mathbf{v}$ by definition of $f_k(\cdot)$. By similar techniques, we have

$$\begin{aligned}
② &\leq \frac{3}{2L_{\mathbf{v}}} \frac{1}{K}\sum_{k=1}^{K} \|\nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_{t-1}, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)\|^2 + \frac{L_{\mathbf{v}}}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}_\psi^*\|^2 \\
&\leq \frac{3L_{\mathbf{v}}}{2} \frac{1}{K}\sum_{k=1}^{K} \|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{t-1}^k\|^2 + \frac{L_{\mathbf{v}}}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}_\psi^*\|^2.
\end{aligned} \tag{69}$$

Let $\hat{\mathbf{v}}_t = \arg\min_{\mathbf{v}} \left(\frac{1}{K}\sum_{k=1}^{K} \nabla_{\mathbf{v}} f(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) \right)^T \mathbf{v} + \frac{1}{2\eta}\|\mathbf{v} - \bar{\mathbf{v}}_{t-1}\|^2 + \frac{1}{2\gamma}\|\mathbf{v} - \mathbf{v}_0\|^2$. Then we have

$$\bar{\mathbf{v}}_t - \hat{\mathbf{v}}_t = \frac{\eta\gamma}{\eta+\gamma}\left(\nabla_{\mathbf{v}} f(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \frac{1}{K}\sum_{k=1}^{K} \nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) \right). \tag{70}$$

Hence we get

$$
\begin{aligned}
③ &= \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \bar{\mathbf{v}}_t - \hat{\mathbf{v}}_t \right\rangle \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \hat{\mathbf{v}}_t - \mathbf{v}_{\psi}^* \right\rangle \\
&= \frac{\eta\gamma}{\eta+\gamma}\left\| \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)] \right\|^2 \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \hat{\mathbf{v}}_t - \mathbf{v}_{\psi}^* \right\rangle \\
&\leq \eta\left\| \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)] \right\|^2 \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \hat{\mathbf{v}}_t - \mathbf{v}_{\psi}^* \right\rangle
\end{aligned} \tag{71}
$$

Plugging (68), (69) and (71) into (67), we get

$$
\begin{aligned}
&\left\langle \nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \mathbf{v}_{\psi}^* \right\rangle + \frac{1}{\gamma}\langle \bar{\mathbf{v}}_t - \mathbf{v}_0, \bar{\mathbf{v}}_t - \mathbf{v}_{\psi}^* \rangle \\
&\leq \frac{3G_\alpha^2}{2L_{\mathbf{v}}}\frac{1}{K}\sum_{k=1}^{K}\|\bar{\alpha}_{t-1} - \alpha_{t-1}^k\|^2 + \frac{L_{\mathbf{v}}}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}_{\psi}^*\|^2 + \frac{3L_{\mathbf{v}}}{2}\frac{1}{K}\sum_{k=1}^{K}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{t-1}^k\|^2 + \frac{L_{\mathbf{v}}}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}_{\psi}^*\|^2 \\
&\quad + \eta\left\| \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)] \right\|^2 \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \hat{\mathbf{v}}_t - \mathbf{v}_{\psi}^* \right\rangle \\
&\quad + \frac{1}{2\eta}(\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{\psi}^*\|^2 - \|\bar{\mathbf{v}}_{t-1} - \bar{\mathbf{v}}_t\|^2 - \|\bar{\mathbf{v}}_t - \mathbf{v}_{\psi}^*\|^2). \square
\end{aligned} \tag{72}
$$

F. Proof of Lemma 5

Proof.

$$
\begin{aligned}
\langle \nabla_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t \rangle &= \left\langle \frac{1}{K}\sum_{k=1}^{K}\nabla_\alpha f_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t \right\rangle \\
&= \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_\alpha f_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - \nabla_\alpha f_k(\bar{\mathbf{v}}_{t-1}, \alpha_{t-1}^k)], \alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t \right\rangle \quad ④ \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_\alpha f_k(\bar{\mathbf{v}}_{t-1}, \alpha_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)], \alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t \right\rangle \quad ⑤ \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t \right\rangle \quad ⑥ \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K}\nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k), \alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t \right\rangle \quad ⑦
\end{aligned} \tag{73}
$$

$$\begin{aligned}
④ &\overset{(a)}{\leq} \frac{3}{2\mu_\alpha}\left(\frac{1}{K}\sum_{k=1}^{K}[\nabla_\alpha f_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - \nabla_\alpha f_k(\bar{\mathbf{v}}_{t-1}, \alpha_{t-1}^k)]\right)^2 + \frac{\mu_\alpha}{6}(\bar{\alpha}_t - \alpha^*(\tilde{\mathbf{v}}))^2 \\
&\overset{(b)}{\leq} \frac{3}{2\mu_\alpha}\frac{1}{K}\sum_{k=1}^{K}(\nabla_\alpha f_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - \nabla_\alpha f_k(\bar{\mathbf{v}}_{t-1}, \alpha_{t-1}^k))^2 + \frac{\mu_\alpha}{6}(\bar{\alpha}_t - \alpha^*(\tilde{\mathbf{v}}))^2 \\
&\overset{(c)}{\leq} \frac{3L_\alpha^2}{2\mu_\alpha}\frac{1}{K}\sum_{k=1}^{K}(\bar{\alpha}_{t-1} - \alpha_{t-1}^k)^2 + \frac{\mu_\alpha}{6}(\bar{\alpha}_t - \alpha^*(\tilde{\mathbf{v}}))^2,
\end{aligned} \tag{74}$$

where (a) follows from Young's inequality, (b) follows from Jensen's inequality, and (c) holds because $f_k(\mathbf{v}, \alpha)$ is smooth in α with coefficient $L_\alpha = 2p(1-p)$ for any $\mathbf{v}$ by definition of $f_k(\cdot)$.

$$\begin{aligned}
⑤ &\overset{(a)}{\leq} \frac{3}{2\mu_\alpha}\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla_\alpha f_k(\bar{\mathbf{v}}_{t-1}, \alpha_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)]\right\|^2 + \frac{\mu_\alpha}{6}(\alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t)^2 \\
&\overset{(b)}{\leq} \frac{3}{2\mu_\alpha}\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla_\alpha f_k(\bar{\mathbf{v}}_{t-1}, \alpha_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)\right\|^2 + \frac{\mu_\alpha}{6}(\alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t)^2 \\
&\overset{(c)}{\leq} \frac{3G_{\mathbf{v}}^2}{2\mu_\alpha}\frac{1}{K}\sum_{k=1}^{K}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{t-1}^k\|^2 + \frac{\mu_\alpha}{6}(\alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t)^2,
\end{aligned} \tag{75}$$

where (a) follows from Young's inequality, (b) follows from Jensen's inequality. (c) holds because $\nabla_\alpha f_k(\mathbf{v}, \alpha)$ is Lipschitz in $\mathbf{v}$ with coefficient $G_{\mathbf{v}} = 2\max(p, 1-p)G_h$ by definition of $f_k(\cdot)$.

Let $\hat{\alpha}_t = \bar{\alpha}_{t-1} + \frac{\eta}{K}\sum\limits_{k=1}^{K}\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)$. Then we have

$$\bar{\alpha}_t - \hat{\alpha}_t = \eta\left(\frac{1}{K}\sum_{k=1}^{K}\nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)\right). \tag{76}$$

And for the auxiliary sequence $\tilde{\alpha}_t$, we can verify that

$$\tilde{\alpha}_t = \arg\min_\alpha \left(\frac{1}{K}\sum_{k=1}^{K}(\nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\right)^T \alpha + \frac{1}{2\eta}(\alpha - \tilde{\alpha}_{t-1})^2 := \lambda_{t-1}(\alpha). \tag{77}$$

Since $\lambda_{t-1}(\alpha)$ is $\frac{1}{\eta}$-strongly convex, we have

$$\begin{aligned}
&\frac{1}{2}(\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_t)^2 \leq \lambda_{t-1}(\alpha^*(\tilde{\mathbf{v}})) - \lambda_{t-1}(\tilde{\alpha}_t) \\
&= \left(\frac{1}{K}\sum_{k=1}^K (\nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\right)^T \alpha^*(\tilde{\mathbf{v}}) + \frac{1}{2\eta}(\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_{t-1})^2 \\
&\quad - \left(\frac{1}{K}\sum_{k=1}^K (\nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\right)^T \tilde{\alpha}_t - \frac{1}{2\eta}(\tilde{\alpha}_t - \tilde{\alpha}_{t-1})^2 \\
&= \left(\frac{1}{K}\sum_{k=1}^K (\nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\right)^T (\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_{t-1}) + \frac{1}{2\eta}(\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_{t-1})^2 \\
&\quad - \left(\frac{1}{K}\sum_{k=1}^K (\nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\right)^T (\tilde{\alpha}_t - \tilde{\alpha}_{t-1}) - \frac{1}{2\eta}(\tilde{\alpha}_t - \tilde{\alpha}_{t-1})^2 \\
&\leq \left(\frac{1}{K}\sum_{k=1}^K (\nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\right)^T (\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_{t-1}) + \frac{1}{2\eta}(\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_{t-1})^2 \\
&\quad + \frac{\eta}{2}\left(\frac{1}{K}\sum_{k=1}^K (\nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) - \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\right)^2.
\end{aligned} \tag{78}$$

Hence we get

$$\begin{aligned}
⑥ &= \left\langle \frac{1}{K}\sum_{k=1}^K [\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \hat{\alpha}_t - \bar{\alpha}_t \right\rangle \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^K [\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \alpha^*(\tilde{\mathbf{v}}) - \hat{\alpha}_t \right\rangle \\
&= \eta\left(\frac{1}{K}\sum_{k=1}^K [\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)]\right)^2 \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^K [\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \alpha^*(\tilde{\mathbf{v}}) - \hat{\alpha}_t \right\rangle.
\end{aligned} \tag{79}$$

Combining (78) and (79), we get

$$\begin{aligned}
⑥ \leq &\frac{3\eta}{2}\left(\frac{1}{K}\sum_{k=1}^K [\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)]\right)^2 \\
&+ \left\langle \frac{1}{K}\sum_{k=1}^K [\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k)], \tilde{\alpha}_{t-1} - \hat{\alpha}_t \right\rangle \\
&+ \frac{1}{2\eta}(\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_{t-1})^2 - \frac{1}{2\eta}(\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_t)^2.
\end{aligned} \tag{80}$$

⑦ can be bounded as

$$⑦ = \langle \bar{\alpha}_t - \bar{\alpha}_{t-1}, \alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t \rangle = \frac{1}{2\eta}((\bar{\alpha}_{t-1} - \alpha^*(\tilde{\mathbf{v}}))^2 - (\bar{\alpha}_{t-1} - \bar{\alpha}_t)^2 - (\bar{\alpha}_t - \alpha^*(\tilde{\mathbf{v}}))^2). \tag{81}$$

Adding (74), (75), (80) and (81), we get

$$\begin{aligned}
\langle \nabla_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \alpha^*(\tilde{\mathbf{v}}) - \bar{\alpha}_t \rangle \leq\ & \frac{3G_{\mathbf{v}}^2}{2\mu_\alpha}\frac{1}{K}\sum_{k=1}^K \|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{t-1}^k\|^2 + \frac{3L_\alpha^2}{2\mu_\alpha}\frac{1}{K}\sum_{k=1}^K (\bar{\alpha}_{t-1} - \alpha_{t-1}^k)^2 \\
& + \frac{3\eta}{2}\left(\frac{1}{K}\sum_{k=1}^K [\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1})]\right)^2 \\
& + \frac{1}{K}\sum_{k=1}^K \langle \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k), \tilde{\alpha}_{t-1} - \hat{\alpha}_t \rangle \\
& + \frac{1}{2\eta}((\bar{\alpha}_{t-1} - \alpha^*(\tilde{\mathbf{v}}))^2 - (\bar{\alpha}_{t-1} - \bar{\alpha}_t)^2 - (\bar{\alpha}_t - \alpha^*(\tilde{\mathbf{v}})))^2) + \frac{\mu_\alpha}{3}(\bar{\alpha}_t - \alpha^*(\tilde{\mathbf{v}})))^2 \\
& + \frac{1}{2\eta}((\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_{t-1}) - (\alpha^*(\tilde{\mathbf{v}}) - \tilde{\alpha}_t)). \ \square
\end{aligned}$$

G. Proof of Lemma 6

Proof. If $I = 1$, $\|\mathbf{v}_t^k - \bar{\mathbf{v}}_t^k\| = 0$ and $|\alpha_t^k - \bar{\alpha}_t^k| = 0$ for any iteration t and any machine k since $\mathbf{v}$ and α are averaged across machines at each iteration.

We prove the case when $I > 1$ in the following. For any iteration t, there must be an iteration with index t_0 before t such that $t \bmod I = 0$ and $t - t_0 \leq I$. Since $\mathbf{v}$ and α are averaged across machines at t_0, we have $\bar{\mathbf{v}}_{t_0} = \mathbf{v}_{t_0}^k$.

(1) For $\mathbf{v}$, according to the update rule,

$$\mathbf{v}_t^k = -\frac{\eta\gamma}{\eta+\gamma}\nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) + \frac{\gamma}{\eta+\gamma}\mathbf{v}_{t-1}^k + \frac{\eta}{\eta+\gamma}\mathbf{v}_0, \tag{82}$$

and hence

$$\bar{\mathbf{v}}_t = -\frac{\eta\gamma}{\eta+\gamma}\frac{1}{K}\sum_{k=1}^K \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) + \frac{\gamma}{\eta+\gamma}\bar{\mathbf{v}}_{t-1} + \frac{\eta}{\eta+\gamma}\mathbf{v}_0. \tag{83}$$

Thus,

$$\begin{aligned}
\|\bar{\mathbf{v}}_t - \mathbf{v}_t^k\| &\leq \frac{\eta\gamma}{\eta+\gamma}\left\|\nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_t^k; \mathbf{z}_t^k) - \frac{1}{K}\sum_{i=1}^K \nabla_{\mathbf{v}} F_i(\mathbf{v}_{t-1}^i, \alpha_{t-1}^i; \mathbf{z}_{t-1}^i)\right\| + \frac{\gamma}{\eta+\gamma}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{t-1}^k\| \\
&\leq 2B_{\mathbf{v}}\frac{\eta\gamma}{\eta+\gamma} + \frac{\gamma}{\eta+\gamma}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{t-1}^k\|.
\end{aligned} \tag{84}$$

Since $\bar{\mathbf{v}}_{t_0} = \mathbf{v}_{t_0}^k$ (for any k), we can see $\|\bar{\mathbf{v}}_{t_0+1} - \mathbf{v}_{t_0+1}^k\| \leq 2\frac{\eta\gamma}{\gamma+\eta}B_{\mathbf{v}} \leq 2\eta B_{\mathbf{v}}$, Assuming $\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{t-1}^k\| \leq 2(t-1-t_0)\eta B_{\mathbf{v}}$, then $\|\bar{\mathbf{v}}_t - \mathbf{v}_t^k\| \leq 2(t-t_0)\eta B_{\mathbf{v}}$ by (84). Thus, by induction, we know that for any t, $\|\bar{\mathbf{v}}_t - \mathbf{v}_t^k\| \leq 2(t-t_0)\eta B_{\mathbf{v}} \leq 2\eta I B_{\mathbf{v}}$. Hence proved.

(ii)

$$\alpha_t^k = \alpha_{t-1}^k + \eta\nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k), \tag{85}$$

and

$$\bar{\alpha}_t = \bar{\alpha}_{t-1} + \eta\frac{1}{K}\sum_{k=1}^K \nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k). \tag{86}$$

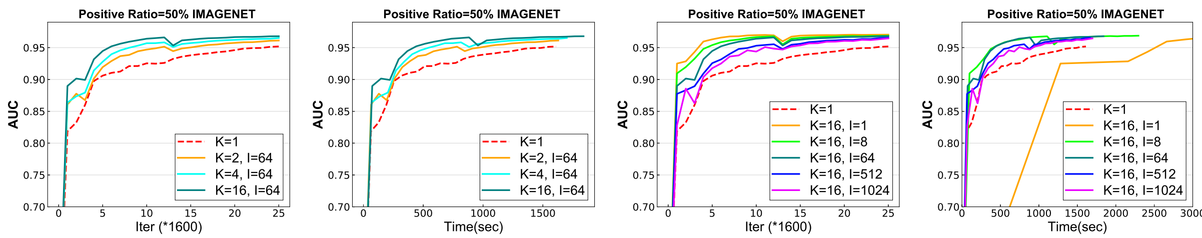
Thus,

$$\begin{aligned}
|\bar{\alpha}_t - \alpha_t^k| &\leq |\bar{\alpha}_{t-1} - \alpha_{t-1}^k| + \eta\left|\nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; \mathbf{z}_{t-1}^k) - \frac{1}{K}\sum_{i=1}^K \nabla_\alpha F_i(\mathbf{v}_{t-1}^i, \alpha_{t-1}^i; \mathbf{z}_{t-1}^i)\right| \\
&\leq |\bar{\alpha}_{t-1} - \alpha_{t-1}^k| + 2\eta B_\alpha.
\end{aligned} \tag{87}$$

Since $\bar{\alpha}_{t_0} = \alpha^k_{t_0}$ (for any k), we can see that $\|\bar{\alpha}_{t_0+1} - \alpha^k_{t_0+1}\| \leq 2\eta B_\alpha$. Assuming $|\bar{\alpha}_{t-1} - \alpha^k_{t-1}| \leq 2(t-1-t_0)\eta B_\alpha$, then $|\bar{\alpha}_t - \alpha^k_t| \leq 2(t-t_0)\eta B_\alpha$. Thus, by induction, we know that for any t, $\|\bar{\alpha}_t - \alpha^k_t\| \leq 2(t-t_0)\eta B_\alpha \leq 2\eta I B_\alpha$. Hence proved. $\square$

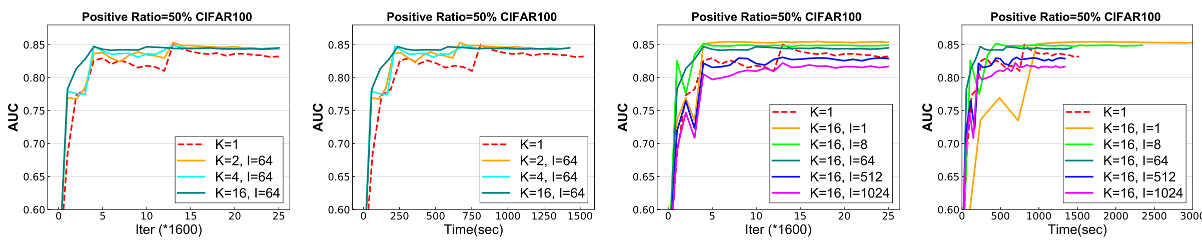
H. More Experiments

In this section, we include more experimental results. Most of the settings are the same as in the Experiments section in the main paper, except that in Figure 10, we set $I = I_0 * 3^{(s-1)}$, other than set I to be a constant. This means that a later stage will communicate less frequently since the step size is decreased after each stage (see the first remark of Theorem 1).



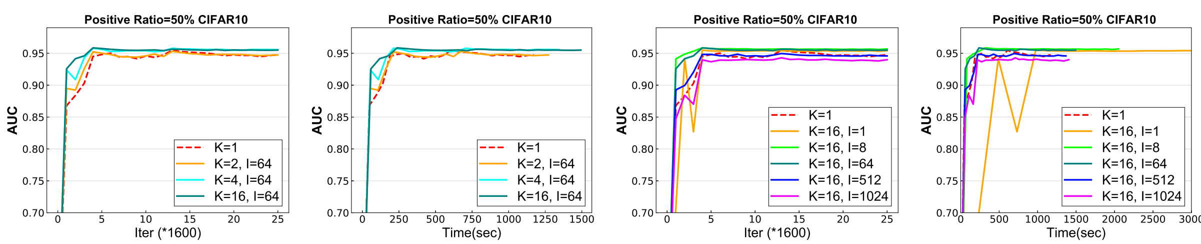
(a) Fix I, vary K (b) Fix K, vary I

Figure 6. ImageNet, positive ratio = 50%.



(a) Fix I, vary K (b) Fix K, vary I

Figure 7. Cifar100, positive ratio = 50%.



(a) Fix I, vary K (b) Fix K, vary I

Figure 8. Cifar10, positive ratio = 50%.

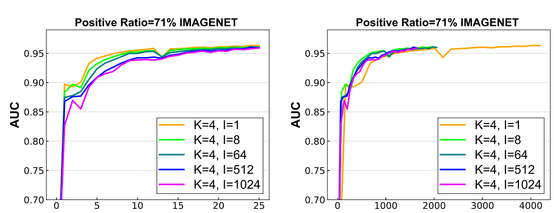
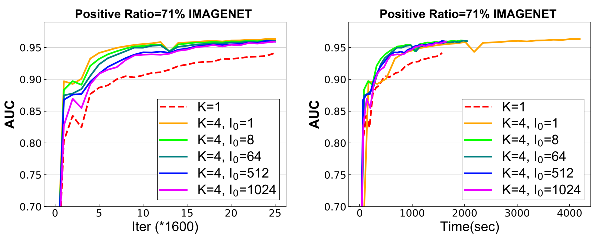
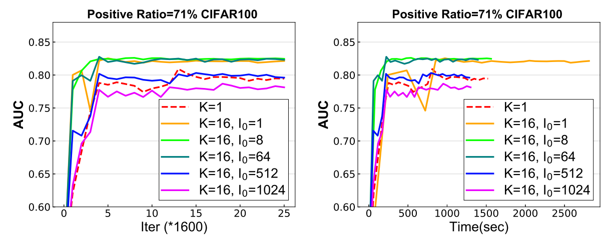


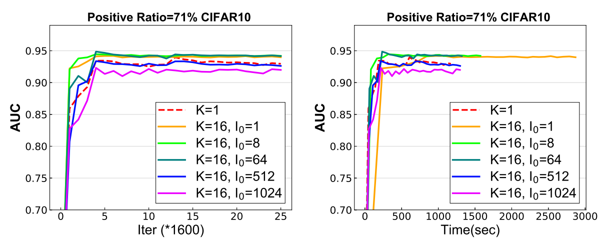
Figure 9. ImageNet, postive ratio=71%, K=4.



(a) ImageNet



(b) Cifar100



(c) Cifar10

Figure 10. $I_s = I_0 3^{(s-1)}$, positive ratio = 71%.