A. Auxiliary Lemmas

Noting all algorithms discussed in thpaper including the baselines implement a stagewise framework, we define the duality gap of s-th stage at a point $(\mathbf{v}, \alpha)$ as

$$Gap_s(\mathbf{v}, \alpha) = \max_{\alpha'} f^s(\mathbf{v}, \alpha') - \min_{\mathbf{v}'} f^s(\mathbf{v}', \alpha). \tag{8}$$

Before we show the proofs, we first present the lemmas from (Yan et al., 2020).

Lemma 3 (Lemma 1 of (Yan et al., 2020))**.** *Suppose a function $h(\mathbf{v}, \alpha)$ is λ_1-strongly convex in $\mathbf{v}$ and λ_2-strongly concave in α. Consider the following problem*

$$\min_{\mathbf{v}\in X} \max_{\alpha\in Y} h(\mathbf{v}, \alpha),$$

where X and Y are convex compact sets. Denote $\hat{\mathbf{v}}_h(y) = \arg\min_{\mathbf{v}'\in X} h(\mathbf{v}', \alpha)$ and $\hat{\alpha}_h(\mathbf{v}) = \arg\max_{\alpha'\in Y} h(\mathbf{v}, \alpha')$. Suppose we have two solutions $(\mathbf{v}_0, \alpha_0)$ and $(\mathbf{v}_1, \alpha_1)$. Then the following relation between variable distance and duality gap holds

$$\begin{aligned}\frac{\lambda_1}{4}\|\hat{\mathbf{v}}_h(\alpha_1) - \mathbf{v}_0\|^2 + \frac{\lambda_2}{4}\|\hat{\alpha}_h(\mathbf{v}_1) - \alpha_0\|^2 \le & \max_{\alpha'\in Y} h(\mathbf{v}_0, \alpha') - \min_{\mathbf{v}'\in X} h(\mathbf{v}', \alpha_0) \\ & + \max_{\alpha'\in Y} h(\mathbf{v}_1, \alpha') - \min_{\mathbf{v}'\in X} h(\mathbf{v}', \alpha_1).\end{aligned} \tag{9}$$

□

Lemma 4 (Lemma 5 of (Yan et al., 2020))**.** *We have the following lower bound for $Gap_s(\mathbf{v}_s, \alpha_s)$*

$$Gap_s(\mathbf{v}_s, \alpha_s) \ge \frac{3}{50} Gap_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1}) + \frac{4}{5}(\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_0^s)),$$

where $\mathbf{v}_0^{s+1} = \mathbf{v}_s$ and $\alpha_0^{s+1} = \alpha_s$, i.e., the initialization of $(s+1)$-th stage is the output of the s-th stage.

□

B. Analysis of CODA+

The proof sketch is similar to the proof of CODA in (Guo et al., 2020a). However, there are two noticeable difference from (Guo et al., 2020a). First, in Lemma 1, we bound the duality gap instead of the objective gap in (Guo et al., 2020a). This is because the analysis later in this proof requires the bound of the duality gap.

Second, in Lemma 1, where the bound for homogeneous data is better than that of heterogeneous data. The better analysis for homogeneous data is inspired by the analysis in (Yu et al., 2019a), which tackles a minimization problem. Note that f^s denotes the subproblem for stage s, we omit the index s in variables when the context is clear.

B.1. Lemmas

We need following lemmas for the proof. The Lemma 5, Lemma 6 and Lemma 7 are similar to Lemma 3, Lemma 4 and Lemma 5 of (Guo et al., 2020a), respectively. For the sake of completeness, we will include the proof of Lemma 5 and Lemma 6 since a change in the update of the primal variable.

Lemma 5. *Define $\bar{\mathbf{v}}_t = \frac{1}{K}\sum_{k=1}^N \mathbf{v}_t^k, \bar{\alpha}_t = \frac{1}{K}\sum_{k=1}^N y_t^k$. Suppose Assumption 1 holds and by running Algorithm 2, we have for any $\mathbf{v}, \alpha$,*

$$\begin{aligned} f^s(\bar{\mathbf{v}}, \alpha) - f^s(\mathbf{v}, \bar{\alpha}) \le \frac{1}{T}\sum_{t=1}^T \Bigg[& \underbrace{\langle \nabla_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - x\rangle}_{B_1} + \underbrace{\langle \nabla_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), y - \bar{\alpha}_t\rangle}_{B_2} \\ & + \underbrace{\frac{3\ell + 3\ell^2/\mu_2}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 + 2\ell(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2}_{B_3} - \frac{\ell}{3}\|\bar{\mathbf{v}}_t - \mathbf{v}\|^2 - \frac{\mu_2}{3}(\bar{\alpha}_{t-1} - \alpha)^2 \Bigg], \end{aligned}$$

where $\mu_2 = 2p(1-p)$ is the strong concavity coefficient of $f(\mathbf{v}, \alpha)$ in α.

Proof. For any $\mathbf{v}$ and α, using Jensen's inequality and the fact that $f^s(\mathbf{v}, \alpha)$ is convex in $\mathbf{v}$ and concave in α,

$$f^s(\bar{\mathbf{v}}, \alpha) - f^s(\mathbf{v}, \bar{\alpha}) \leq \frac{1}{T}\sum_{t=1}^{T}(f^s(\bar{\mathbf{v}}_t, \alpha) - f^s(\mathbf{v}, \bar{\alpha}_t)) \tag{10}$$

By ℓ-strongly convexity of $f^s(\mathbf{v}, \alpha)$ in $\mathbf{v}$, we have

$$f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) + \langle \partial_{\mathbf{v}} f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \mathbf{v} - \bar{\mathbf{v}}_{t-1}\rangle + \frac{\ell}{2}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}\|^2 \leq f(\mathbf{v}, \bar{\alpha}_{t-1}). \tag{11}$$

By 3ℓ-smoothness of $f^s(\mathbf{v}, \alpha)$ in $\mathbf{v}$, we have

$$\begin{aligned} f^s(\bar{\mathbf{v}}_t, \alpha) &\leq f^s(\bar{\mathbf{v}}_{t-1}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\bar{\mathbf{v}}_{t-1}, \alpha), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle + \frac{3\ell}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 \\ &= f^s(\bar{\mathbf{v}}_{t-1}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle + \frac{3\ell}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 \\ &\quad + \langle \partial_{\mathbf{v}} f^s(\bar{\mathbf{v}}_{t-1}, \alpha) - \partial_{\mathbf{v}} f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle \\ &\overset{(a)}{\leq} f^s(\bar{\mathbf{v}}_{t-1}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle + \frac{3\ell}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 \\ &\quad + \ell|\bar{\alpha}_{t-1} - \alpha|\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\| \\ &\overset{(b)}{\leq} f^s(\bar{\mathbf{v}}_{t-1}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\rangle + \frac{3\ell}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 \\ &\quad + \frac{\mu_2}{6}(\bar{\alpha}_{t-1} - \alpha)^2 + \frac{3\ell^2}{2\mu_2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2, \end{aligned} \tag{12}$$

where (a) holds because that we know $\partial_{\mathbf{v}} f(\mathbf{v}, \alpha)$ is ℓ-Lipschitz in α since $f(\mathbf{v}, \alpha)$ is ℓ-smooth, (b) holds by Young's inequality, and $\mu_2 = 2p(1-p)$ is the strong concavity coefficient of f^s in α.

Adding (11) and (12), rearranging terms, we have

$$\begin{aligned} & f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) + f^s(\bar{\mathbf{v}}_t, \alpha) \\ & \leq f(\mathbf{v}, \bar{\alpha}_{t-1}) + f(\bar{\mathbf{v}}_{t-1}, \alpha) + \langle \partial_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \mathbf{v}\rangle + \frac{3\ell + 3\ell^2/\mu_2}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 \\ & \quad - \frac{\ell}{2}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}\|^2 + \frac{\mu_2}{6}(\bar{\alpha}_{t-1} - \alpha)^2. \end{aligned} \tag{13}$$

We know $f^s(\mathbf{v}, \alpha)$ is μ_2-strong concavity in α ($-f(\mathbf{v}, \alpha)$ is μ_2-strong convexity of in α). Thus, we have

$$-f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - \partial_\alpha f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1})^\top(\alpha - \bar{\alpha}_{t-1}) + \frac{\mu_2}{2}(\alpha - \bar{\alpha}_{t-1})^2 \leq -f^s(\bar{\mathbf{v}}_{t-1}, \alpha). \tag{14}$$

Since $f(\mathbf{v}, \alpha)$ is ℓ-smooth in α, we get

$$\begin{aligned} -f^s(\mathbf{v}, \bar{\alpha}_t) &\leq -f^s(\mathbf{v}, \bar{\alpha}_{t-1}) - \langle \partial_\alpha f^s(\mathbf{v}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \bar{\alpha}_{t-1}\rangle + \frac{\ell}{2}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 \\ &= -f^s(\mathbf{v}, \bar{\alpha}_{t-1}) - \langle \partial_\alpha f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \bar{\alpha}_{t-1}\rangle + \frac{\ell}{2}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 \\ &\quad - \langle \partial_\alpha (f^s(\mathbf{v}, \bar{\alpha}_{t-1}) - f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1})), \bar{\alpha}_t - \bar{\alpha}_{t-1}\rangle \\ &\overset{(a)}{\leq} -f^s(\mathbf{v}, \bar{\alpha}_{t-1}) - \langle \partial_\alpha f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \bar{\alpha}_{t-1}\rangle + \frac{\ell}{2}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 \\ &\quad + \ell\|\mathbf{v} - \bar{\mathbf{v}}_{t-1}\|(\bar{\alpha}_t - \bar{\alpha}_{t-1}) \\ &\leq -f^s(\mathbf{v}, \bar{\alpha}_{t-1}) - \langle \partial_\alpha f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \bar{\alpha}_{t-1}\rangle + \frac{\ell}{2}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 \\ &\quad + \frac{\ell}{6}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}\|^2 + \frac{3\ell}{2}(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 \end{aligned} \tag{15}$$

where (a) holds because that $\partial_\alpha f^s(\mathbf{v}, \alpha)$ is ℓ-Lipschitz in $\mathbf{v}$.

Adding (14), (15) and arranging terms, we have

$$\begin{aligned}
&- f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - f^s(\mathbf{v}, \bar{\alpha}_t) \leq -f^s(\bar{\mathbf{v}}_{t-1}, \alpha) - f^s(\mathbf{v}, \bar{\alpha}_{t-1}) - \langle \partial_\alpha f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \alpha \rangle \\
&\quad + 2\ell(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 + \frac{\ell}{6}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}\|^2 - \frac{\mu_2}{2}(\alpha - \bar{\alpha}_{t-1})^2.
\end{aligned} \tag{16}$$

Adding (13) and (16), we get

$$\begin{aligned}
&f^s(\bar{\mathbf{v}}_t, \alpha) - f^s(\mathbf{v}, \bar{\alpha}_t) \\
&\leq \langle \partial_{\mathbf{v}} f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \mathbf{v} \rangle - \langle \partial_\alpha f(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\alpha}_t - \alpha \rangle \\
&\quad + \frac{3\ell + 3\ell^2/\mu_2}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 + 2\ell(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2 \\
&\quad - \frac{\ell}{3}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}\|^2 - \frac{\mu_2}{3}(\bar{\alpha}_{t-1} - \alpha)^2
\end{aligned} \tag{17}$$

Taking average over $t = 1, ..., T$, we get

$$\begin{aligned}
&f^s(\bar{\mathbf{v}}, \alpha) - f^s(\mathbf{v}, \bar{\alpha}) \\
&\leq \frac{1}{T}\sum_{t=1}^{T}[f^s(\bar{\mathbf{v}}_t, \alpha) - f^s(\mathbf{v}, \bar{\alpha}_t)] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\Bigg[\underbrace{\langle \partial_{\mathbf{v}} f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \mathbf{v} \rangle}_{B_1} + \underbrace{\langle \partial_\alpha f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \alpha - \bar{\alpha}_t \rangle}_{B_2} \\
&\quad \underbrace{+ \frac{3\ell + 3\ell^2/\mu_2}{2}\|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}\|^2 + 2\ell(\bar{\alpha}_t - \bar{\alpha}_{t-1})^2}_{B_3} \\
&\quad - \frac{\ell}{3}\|\mathbf{v} - \bar{\mathbf{v}}_t\|^2 - \frac{\mu_2}{3}(\bar{\alpha}_{t-1} - \alpha)^2\Bigg]
\end{aligned}$$

□

In the following, we will bound the term B_1 by Lemma 6, B_2 by Lemma 7 and B_3 by Lemma 8.

Lemma 6. *Define* $\hat{\mathbf{v}}_t = \bar{\mathbf{v}}_{t-1} - \frac{\eta}{K}\sum\limits_{k=1}^{K}\nabla_{\mathbf{v}} f^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)$ *and*

$$\tilde{\mathbf{v}}_t = \tilde{\mathbf{v}}_{t-1} - \frac{\eta}{K}\sum_{k=1}^{K}\left(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, y_{t-1}^k; z_{t-1}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)\right), \textit{ for } t > 0\textit{; } \tilde{\mathbf{v}}_0 = \mathbf{v}_0. \tag{18}$$

. We have

$$\begin{aligned}
B_1 &\leq \frac{3\ell}{2}\frac{1}{K}\sum_{k=1}^{K}(\bar{\alpha}_{t-1} - \alpha_{t-1}^k)^2 + \frac{3\ell}{2}\frac{1}{K}\sum_{k=1}^{K}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{t-1}^k\|^2 \\
&\quad + \frac{3\eta}{2}\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)]\right\|^2 \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)], \hat{\mathbf{v}}_t - \tilde{\mathbf{v}}_{t-1} \right\rangle \\
&\quad + \frac{1}{2\eta}(\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}\|^2 - \|\bar{\mathbf{v}}_{t-1} - \bar{\mathbf{v}}_t\|^2 - \|\bar{\mathbf{v}}_t - \mathbf{v}\|^2) \\
&\quad + \frac{\ell}{3}\|\bar{\mathbf{v}}_t - \mathbf{v}\|^2 + \frac{1}{2\eta}(\|\mathbf{v} - \tilde{\mathbf{v}}_{t-1}\|^2 - \|\mathbf{v} - \tilde{\mathbf{v}}_t\|^2)
\end{aligned}$$

Proof. We have

$$
\begin{aligned}
&\langle \nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \mathbf{v}\rangle = \left\langle \frac{1}{K}\sum_{k=1}^{K} \nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}), \bar{\mathbf{v}}_t - \mathbf{v}\right\rangle \\
&\leq \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - \nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_{t-1}, \alpha^k_{t-1})], \bar{\mathbf{v}}_t - \mathbf{v}\right\rangle \qquad ① \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_{t-1}, \alpha^k_{t-1}) - \nabla_{\mathbf{v}} f^s_k(\mathbf{v}^k_{t-1}, \alpha^k_{t-1})], \bar{\mathbf{v}}_t - \mathbf{v}\right\rangle \qquad ② \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K} [\nabla_{\mathbf{v}} f^s_k(\mathbf{v}^k_{t-1}, \alpha^k_{t-1}) - \nabla_{\mathbf{v}} F^s_k(\bar{\mathbf{v}}_{t-1}, \alpha^k_{t-1}; z^k_{t-1})], \bar{\mathbf{v}}_t - \mathbf{v}\right\rangle \qquad ③ \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K} \nabla_{\mathbf{v}} F^s_k(\bar{\mathbf{v}}_{t-1}, \alpha^k_{t-1}; z^k_{t-1}), \bar{\mathbf{v}}_t - \mathbf{v}\right\rangle \qquad ④
\end{aligned} \tag{19}
$$

Then we will bound ①, ②, ③ and ④, respectively,

$$
\begin{aligned}
① &\overset{(a)}{\leq} \frac{3}{2\ell}\left\|\frac{1}{K}\sum_{k=1}^{K} [\nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - \nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_{t-1}, \alpha^k_{t-1})]\right\|^2 + \frac{\ell}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}\|^2 \\
&\overset{(b)}{\leq} \frac{3}{2\ell}\frac{1}{K}\sum_{k=1}^{K} \|\nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_{t-1}, \bar{\alpha}_{t-1}) - \nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_{t-1}, \alpha^k_{t-1})\|^2 + \frac{\ell}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}\|^2 \\
&\overset{(c)}{\leq} \frac{3\ell}{2}\frac{1}{K}\sum_{k=1}^{K} (\bar{\alpha}_{t-1} - \alpha^k_{t-1})^2 + \frac{\ell}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}\|^2,
\end{aligned} \tag{20}
$$

where (a) follows from Young's inequality, (b) follows from Jensen's inequality. and (c) holds because $\nabla_{\mathbf{v}} f^s_k(\mathbf{v}, \alpha)$ is ℓ-Lipschitz in α. Using similar techniques, we have

$$
\begin{aligned}
② &\leq \frac{3}{2\ell}\frac{1}{K}\sum_{k=1}^{K} \|\nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_{t-1}, \alpha^k_{t-1}) - \nabla_{\mathbf{v}} f^s_k(\mathbf{v}^k_{t-1}, \alpha^k_{t-1})\|^2 + \frac{\ell}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}\|^2 \\
&\leq \frac{3\ell}{2}\frac{1}{K}\sum_{k=1}^{K} \|\bar{\mathbf{v}}_{t-1} - \mathbf{v}^k_{t-1}\|^2 + \frac{\ell}{6}\|\bar{\mathbf{v}}_t - \mathbf{v}\|^2.
\end{aligned} \tag{21}
$$

Let $\hat{\mathbf{v}}_t = \arg\min_{\mathbf{v}} \left(\frac{1}{K}\sum\limits_{k=1}^{K} \nabla_{\mathbf{v}} f^s(\mathbf{v}^k_{t-1}, \alpha^k_{t-1})\right)^\top x + \frac{1}{2\eta}\|\mathbf{v} - \bar{\mathbf{v}}_{t-1}\|^2$, then we have

$$
\bar{\mathbf{v}}_t - \hat{\mathbf{v}}_t = \eta\Bigg(\nabla_{\mathbf{v}} f^s(\mathbf{v}^k_{t-1}, y^k_{t-1}) - \frac{1}{K}\sum_{k=1}^{K} \nabla_{\mathbf{v}} f^s_k(\mathbf{v}^k_{t-1}, y^k_{t-1}; z^k_{t-1})\Bigg) \tag{22}
$$

Hence we get

$$
\begin{aligned}
③ &= \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)], \bar{\mathbf{v}}_t - \hat{\mathbf{v}}_t \right\rangle \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)], \hat{\mathbf{v}}_t - \mathbf{v} \right\rangle \\
&= \eta \left\| \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)] \right\|^2 \\
&\quad + \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)], \hat{\mathbf{v}}_t - \mathbf{v} \right\rangle
\end{aligned} \tag{23}
$$

Define another auxiliary sequence as

$$
\tilde{\mathbf{v}}_t = \tilde{\mathbf{v}}_{t-1} - \frac{\eta}{K}\sum_{k=1}^{K}\left(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, y_{t-1}^k; z_{t-1}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)\right), \text{ for } t > 0; \tilde{\mathbf{v}}_0 = \mathbf{v}_0. \tag{24}
$$

Denote

$$
\Theta_{t-1}(\mathbf{v}) = \left(-\frac{1}{K}\sum_{k=1}^{K}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, y_{t-1}^k; z_{t-1}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\right)^{\top} x + \frac{1}{2\eta}\|\mathbf{v} - \tilde{\mathbf{v}}_{t-1}\|^2. \tag{25}
$$

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

$$
\tilde{\mathbf{v}}_t = \arg\min_{\mathbf{v}} \Theta_{t-1}(\mathbf{v}). \tag{26}
$$

Since $\Theta_{t-1}(\mathbf{v})$ is $\frac{1}{\eta}$-strongly convex, we have

$$
\begin{aligned}
&\frac{1}{2}\|\mathbf{v} - \tilde{\mathbf{v}}_t\|^2 \le \Theta_{t-1}(\mathbf{v}) - \Theta_{t-1}(\tilde{\mathbf{v}}_t) \\
&= \Big(-\frac{1}{K}\sum_{k=1}^{K}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\Big)^{\top} x + \frac{1}{2\eta}\|\mathbf{v} - \tilde{\mathbf{v}}_{t-1}\|^2 \\
&\quad - \Big(-\frac{1}{K}\sum_{k=1}^{K}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\Big)^{\top} \tilde{\mathbf{v}}_t - \frac{1}{2\eta}\|\tilde{\mathbf{v}}_t - \tilde{\mathbf{v}}_{t-1}\|^2 \\
&= \Big(-\frac{1}{K}\sum_{k=1}^{K}(\nabla_{\alpha} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k) - \nabla_{\alpha} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\Big)^{\top} (\mathbf{v} - \tilde{\mathbf{v}}_{t-1}) + \frac{1}{2\eta}\|\mathbf{v} - \tilde{\mathbf{v}}_{t-1}\|^2 \\
&\quad - \Big(-\frac{1}{K}\sum_{k=1}^{K}(\nabla_{\alpha} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k) - \nabla_{\alpha} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\Big)^{\top} (\tilde{\mathbf{v}}_t - \tilde{\mathbf{v}}_{t-1}) - \frac{1}{2\eta}\|\tilde{\mathbf{v}}_t - \tilde{\mathbf{v}}_{t-1}\|^2 \\
&\le \Big(-\frac{1}{K}\sum_{k=1}^{K}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k))\Big)^{\top} (\mathbf{v} - \tilde{\mathbf{v}}_{t-1}) + \frac{1}{2\eta}\|\mathbf{v} - \tilde{\mathbf{v}}_{t-1}\|^2 \\
&\quad + \frac{\eta}{2}\left\| \frac{1}{K}\sum_{k=1}^{K}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)) \right\|^2
\end{aligned} \tag{27}
$$

Adding this with (23), we get

$$
\begin{aligned}
③ \le& \frac{3\eta}{2}\left\| \frac{1}{K}\sum_{k=1}^{K}(\nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k) - \nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)) \right\|^2 + \frac{1}{2\eta}\|\mathbf{v} - \tilde{\mathbf{v}}_{t-1}\|^2 - \frac{1}{2}\|\mathbf{v} - \tilde{\mathbf{v}}_t\|^2 \\
&+ \left\langle \frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)], \hat{\mathbf{v}}_t - \tilde{\mathbf{v}}_{t-1} \right\rangle
\end{aligned} \tag{28}
$$

④ can be bounded as

$$④ = -\frac{1}{\eta}\langle \bar{\mathbf{v}}_t - \bar{\mathbf{v}}_{t-1}, \bar{\mathbf{v}}_t - \mathbf{v}\rangle = \frac{1}{2\eta}(\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}\|^2 - \|\bar{\mathbf{v}}_{t-1} - \bar{\mathbf{v}}_t\|^2 - \|\bar{\mathbf{v}}_t - \bar{\mathbf{v}}\|^2) \tag{29}$$

Plug (20), (21), (28) and (29) into (19), we get

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

□

B_2 can be bounded by the following lemma, whose proof is identical to that of Lemma 5 in (Guo et al., 2020a).

Lemma 7. *Define* $\hat{\alpha}_t = \bar{\alpha}_{t-1} + \frac{\eta}{K}\sum_{k=1}^{K}\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k)$*, and*

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

We have,

$$\begin{aligned}
B_2 &\leq \frac{3\ell^2}{2\mu_2}\frac{1}{K}\sum_{k=1}^{K}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{t-1}^k\|^2 + \frac{3\ell^2}{2\mu_2}\frac{1}{K}\sum_{k=1}^{K}(\bar{\alpha}_{t-1} - \alpha_{t-1}^k)^2 \\
&\quad + \frac{3\eta}{2}\left(\frac{1}{K}\sum_{k=1}^{K}[\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1})]\right)^2 \\
&\quad + \frac{1}{K}\sum_{k=1}^{K}\langle \nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_i(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k), \tilde{\alpha}_{t-1} - \hat{\alpha}_t\rangle \\
&\quad + \frac{1}{2\eta}((\bar{\alpha}_{t-1} - \alpha)^2 - (\bar{\alpha}_{t-1} - \bar{\alpha}_t)^2 - (\bar{\alpha}_t - \alpha)^2) \\
&\quad + \frac{\mu_2}{3}(\bar{\alpha}_t - \alpha)^2 + \frac{1}{2\eta}(\alpha - \tilde{\alpha}_{t-1})^2 - \frac{1}{2\eta}(\alpha - \tilde{\alpha}_t)^2.
\end{aligned}$$

□

B_3 can be bounded by the following lemma.

Lemma 8. *If K machines communicate every I iterations, where $I \leq \frac{1}{18\sqrt{2}\eta\ell}$, then*

$$\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\bar{\mathbf{v}}_t - \mathbf{v}_t^k\|^2 + \|\bar{\alpha}_t - \alpha_t^k\|^2\right] \leq \left(12\eta^2 I\sigma^2 T + 36\eta^2 I^2 D^2 T\right)\mathbb{I}_{I>1}$$

Proof. In this proof, we introduce a couple of new notations to make the proof brief: $F^s_{k,t} = F^s_{k,t}(\mathbf{v}^k_t, \alpha^k_t; z^k_t)$ and $f^s_{k,t} = f^s_{k,t}(\mathbf{v}^k_t, \alpha^k_t)$. Similar bounds for minimization problems have been analyzed in (Yu et al., 2019a; Stich, 2019).

Denote t_0 as the nearest communication round before t, i.e., $t - t_0 \leq I$. By the update rule of $\mathbf{v}$, we have that on each machine k,

$$\mathbf{v}^k_t = \bar{\mathbf{v}}_{t_0} - \eta \sum_{\tau=t_0}^{t-1} \nabla_{\mathbf{v}} F^s_{k,\tau}. \tag{30}$$

Taking average over all K machines,

$$\bar{\mathbf{v}}_t = \bar{\mathbf{v}}_{t_0} - \eta \sum_{\tau=t_0}^{t-1} \frac{1}{K} \sum_{k=1}^{K} \nabla_{\mathbf{v}} F^s_{k,\tau}. \tag{31}$$

Therefore,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K} \|\bar{\mathbf{v}}_t - \mathbf{v}^k_t\|^2 = \frac{\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\sum_{\tau=t_0}^{t-1}\left[\nabla_{\mathbf{v}} F^s_{k,\tau} - \frac{1}{K}\sum_{j=1}^{K}\nabla_{\mathbf{v}} F^s_{j,\tau}\right]\right\|^2\right] \\
&\leq \frac{2\eta^2}{K}\sum_{k=1}^{K}\left[\left\|\sum_{\tau=t_0}^{t-1}\left[[\nabla_{\mathbf{v}} F^s_{k,\tau} - \nabla_{\mathbf{v}} f^s_{k,\tau}] - \frac{1}{K}\sum_{j=1}^{K}[\nabla_{\mathbf{v}} F^s_{j,\tau} - \nabla_{\mathbf{v}} f^s_{j,\tau}]\right]\right\|^2\right] \\
&\quad + \frac{2\eta^2}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\sum_{\tau=t_0}^{t-1}\left[\nabla_{\mathbf{v}} f^s_{k,\tau} - \frac{1}{K}\sum_{j=1}^{K}\nabla_{\mathbf{v}} f^s_{j,\tau}\right]\right\|^2\right]
\end{aligned} \tag{32}$$

In the following, we will address these two terms on the right hand side separately. First, we have

$$\begin{aligned}
&\frac{2\eta^2}{K}\sum_{k=1}^{K}\left[\left\|\sum_{\tau=t_0}^{t-1}\left[[\nabla_{\mathbf{v}} F^s_{k,\tau} - \nabla_{\mathbf{v}} f^s_{k,\tau}] - \frac{1}{K}\sum_{j=1}^{K}[\nabla_{\mathbf{v}} F^s_{j,\tau} - \nabla_{\mathbf{v}} f^s_{j,\tau}]\right]\right\|^2\right] \\
&\overset{(a)}{\leq} \frac{2\eta^2}{K}\sum_{k=1}^{K}\left[\left\|\sum_{\tau=t_0}^{t-1}[\nabla_{\mathbf{v}} F^s_{k,\tau} - \nabla_{\mathbf{v}} f^s_{k,\tau}]\right\|^2\right] \\
&\overset{(b)}{=} \frac{2\eta^2}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t-1}\left[\left\|[\nabla_{\mathbf{v}} F^s_{k,\tau} - \nabla_{\mathbf{v}} f^s_{k,\tau}]\right\|^2\right] \\
&\leq 2\eta^2 I \sigma^2,
\end{aligned} \tag{33}$$

where (a) holds by $\frac{1}{K}\sum\limits_{k=1}^{K}\|a_k - \left[\frac{1}{K}\sum\limits_{j=1}^{K} a_j\right]\|^2 = \frac{1}{K}\sum\limits_{k=1}^{K}\|a_k\|^2 - \|\frac{1}{K}\sum\limits_{k=1}^{K} a_k\|^2 \leq \frac{1}{K}\sum\limits_{k=1}^{K}\|a_k\|^2$, where $a_k = \sum\limits_{\tau=t_0}^{t-1}[\nabla F^s_{k,\tau} - \nabla_{\mathbf{v}} f_{k,\tau}]$; (b) follows because $\mathbb{E}_{k,\tau-1}[\nabla_{\mathbf{v}} F^s_{k,\tau} - \nabla_{\mathbf{v}} f^s_{k,\tau}] = 0$.

Second, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\sum_{\tau=t_0}^{t-1}\left[\nabla_{\mathbf{v}} f^s_{i,\tau} - \frac{1}{K}\sum_{j=1}^{K}\nabla_{\mathbf{v}} f^s_{j,\tau}\right]\right\|^2\right] \\
&\leq \frac{1}{K}\sum_{k=1}^{K}(t-t_0)\sum_{\tau=t_0}^{t-1}\mathbb{E}\left[\left\|\nabla_{\mathbf{v}} f^s_{i,\tau} - \frac{1}{K}\sum_{j=1}^{K}\nabla_{\mathbf{v}} f^s_{j,\tau}\right\|^2\right] \\
&\leq I\sum_{\tau=t_0}^{t-1}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla_{\mathbf{v}} f^s_{k,\tau} - \frac{1}{K}\sum_{j=1}^{K}\nabla_{\mathbf{v}} f^s_{j,\tau}\right\|^2\right],
\end{aligned} \tag{34}$$

where

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_{\mathbf{v}} f^s_{k,\tau} - \frac{1}{K}\sum_{j=1}^{K}\nabla_{\mathbf{v}} f^s_{j,\tau}\right\|^2 \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_{\mathbf{v}} f^s_{k,\tau} - \nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau) + \nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau) - \nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau) + \nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau) - \frac{1}{K}\sum_{j=1}^{K}\nabla_{\mathbf{v}} f^s_{j,\tau}\right\|^2 \\
&\leq \frac{1}{K}\sum_{k=1}^{K}\left[3\mathbb{E}\|\nabla_{\mathbf{v}} f^s_{k,\tau} - \nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2 + 3\mathbb{E}\|\nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau) - \nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2\right. \\
&\quad \left. + 3\mathbb{E}\left\|\nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau) - \frac{1}{K}\sum_{j=1}^{K}\nabla_{\mathbf{v}} f^s_{j,\tau}\right\|^2\right. \\
&= \frac{1}{K}\sum_{k=1}^{K}\left[3\mathbb{E}\|\nabla_{\mathbf{v}} f^s_{k,\tau} - \nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2 + 3\mathbb{E}\|\nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau) - \nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2\right. \\
&\quad \left. + 3\mathbb{E}\left\|\frac{1}{K}\sum_{j=1}^{K}[\nabla_{\mathbf{v}} f^s_j(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau) - \nabla_{\mathbf{v}} f^s_{j,\tau}]\right\|^2\right] \\
&\leq \frac{1}{K}\sum_{k=1}^{K}\left[3\mathbb{E}\|\nabla_{\mathbf{v}} f^s_{k,\tau} - \nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2 + 3\mathbb{E}\|\nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau) - \nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2\right. \\
&\quad \left. + 3\frac{1}{K}\sum_{j=1}^{K}\mathbb{E}\left\|[\nabla_{\mathbf{v}} f^s_j(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau) - \nabla_{\mathbf{v}} f^s_{j,\tau}]\right\|^2\right] \\
&\overset{(a)}{\leq} \frac{54\ell^2}{K}\sum_{k=1}^{K}\left[\|\mathbf{v}_{k,\tau} - \bar{\mathbf{v}}_\tau\|^2 + |\alpha_{k,\tau} - \bar{\alpha}_\tau|^2\right] + \frac{3}{K}\sum_{k=1}^{K}\|\nabla_{\mathbf{v}} f^s_k(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau) - \nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2 \\
&\leq \frac{54\ell^2}{K}\sum_{k=1}^{K}\left[\|\mathbf{v}_{k,\tau} - \bar{\mathbf{v}}_\tau\|^2 + |\alpha_{k,\tau} - \bar{\alpha}_\tau|^2\right] + 3D^2,
\end{aligned}
\tag{35}
$$

where (a) holds because f is ℓ-smooth, i.e., f^s is 3ℓ-smooth.

Combining (32), (33), (34) and (35),

$$
\frac{1}{K}\sum_{k=1}^{K}\|\bar{\mathbf{v}}_t - \mathbf{v}^k_t\|^2 \leq 2\eta^2 I\sigma^2 + 2\eta^2\left(I\sum_{\tau=t_0}^{t-1}\left[\frac{54\ell^2}{K}\sum_{k=1}^{K}\left[\|\mathbf{v}^k_\tau - \bar{\mathbf{v}}_\tau\|^2 + \|\alpha_{k,\tau} - \bar{\alpha}_\tau\|^2\right] + 3D^2\right]\right) \tag{36}
$$

Summing over $t = \{0, ..., T-1\}$,

$$
\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\|\bar{\mathbf{v}}_t - \mathbf{v}^k_t\|^2 \leq 2\eta^2 I\sigma^2 T + 108\eta^2 I^2\ell^2\sum_{t=0}^{T-1}\frac{1}{K}\left(\|\mathbf{v}^k_t - \bar{\mathbf{v}}_t\|^2 + \|\alpha^k_t - \bar{\alpha}_\tau\|^2\right) + 6\eta^2 I^2 D^2 T. \tag{37}
$$

Similarly for α side, we have

$$
\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}\|\bar{\alpha}_t - \alpha^k_t\|^2 \leq 2\eta^2 I\sigma^2 T + 108\eta^2 I^2\ell^2\sum_{t=0}^{T-1}\frac{1}{K}\left(\|\mathbf{v}^k_t - \bar{\mathbf{v}}_t\|^2 + \|\alpha^k_t - \bar{\alpha}_t\|^2\right) + 6\eta^2 I^2 D^2 T. \tag{38}
$$

Summing up the above two inequalities,

$$
\begin{aligned}
\sum_{t=0}^{T-1}\frac{1}{K}\sum_{k=1}^{K}[\|\bar{\mathbf{v}}_t - \mathbf{v}^k_t\|^2 + \mathbb{E}[\|\bar{\alpha}_t - \alpha^k_t\|^2] &\leq \frac{4\eta^2 I\sigma^2}{1-216\eta^2 I^2\ell^2}T + \frac{12\eta^2 I^2 D^2}{1-216\eta^2 I^2\ell^2}T \\
&\leq 12\eta^2 I\sigma^2 T + 36\eta^2 I^2 D^2 T
\end{aligned}
\tag{39}
$$

where the second inequality is due to $I \leq \frac{1}{18\sqrt{2}\eta\ell}$, i.e., $1 - 216\eta^2 I^2\ell^2 \geq \frac{2}{3}$. $\square$

Based on above lemmas, we are ready to give the convergence of duality gap in one stage of CODA+.

B.2. Proof of Lemma 1

Proof. Noting $\mathbb{E}\langle \frac{1}{K}\sum\limits_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)], \hat{\mathbf{v}}_t - \tilde{\mathbf{v}}_{t-1}\rangle = 0$ and $\mathbb{E}\left\langle -\frac{1}{K}\sum\limits_{k=1}^{K}[\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)], \tilde{\alpha}_{t-1} - \hat{\alpha}_t \right\rangle = 0.$ and then plugging Lemma 6 and Lemma 7 into Lemma 5, and taking expectation, we get

$$\begin{aligned}
&\mathbb{E}[f^s(\bar{\mathbf{v}}, \alpha) - f^s(\mathbf{v}, \bar{\alpha})] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Bigg[\underbrace{\left(\frac{3\ell + 3\ell^2/\mu_2}{2} - \frac{1}{2\eta}\right)\|\bar{\mathbf{v}}_{t-1} - \bar{\mathbf{v}}_t\|^2 + \left(2\ell - \frac{1}{2\eta}\right)\|\bar{\alpha}_t - \bar{\alpha}_{t-1}\|^2}_{C_1} \\
&+ \underbrace{\left(\frac{1}{2\eta} - \frac{\mu_2}{3}\right)\|\bar{\alpha}_{t-1} - \alpha\|^2 - \left(\frac{1}{2\eta} - \frac{\mu_2}{3}\right)(\bar{\alpha}_t - \alpha)^2}_{C_2} + \underbrace{\left(\frac{1}{2\eta} - \frac{\ell}{3}\right)\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}\|^2 - \left(\frac{1}{2\eta} - \frac{\ell}{3}\right)\|\bar{\mathbf{v}}_t - \mathbf{v}\|^2}_{C_3} \\
&+ \underbrace{\frac{1}{2\eta}((\alpha - \tilde{\alpha}_{t-1})^2 - (\alpha - \tilde{\alpha}_t)^2)}_{C_4} + \underbrace{\frac{1}{2\eta}(\|\mathbf{v} - \tilde{\mathbf{v}}_{t-1}\|^2 - \|\mathbf{v} - \tilde{\mathbf{v}}_t\|^2)}_{C_5} \\
&+ \underbrace{\left(\frac{3\ell^2}{2\mu_2} + \frac{3\ell}{2}\right)\frac{1}{K}\sum_{k=1}^{K}\|\bar{\mathbf{v}}_{t-1} - \mathbf{v}_{t-1}^k\|^2 + \left(\frac{3\ell}{2} + \frac{3\ell^2}{2\mu_2}\right)\frac{1}{K}\sum_{k=1}^{K}(\bar{\alpha}_{t-1} - \alpha_{t-1}^k)^2}_{C_6} \\
&+ \underbrace{\frac{3\eta}{2}\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)]\right\|^2}_{C_7} \\
&+ \underbrace{\frac{3\eta}{2}\left\|\frac{1}{K}\sum_{k=1}^{K}\nabla_\alpha f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_\alpha F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)\right\|^2}_{C_8}\Bigg]
\end{aligned} \tag{40}$$

Since $\eta \leq \min(\frac{1}{3\ell+3\ell^2/\mu_2}, \frac{1}{4\ell})$, thus in the RHS of (40), C_1 can be cancelled. C_2, C_3, C_4 and C_5 will be handled by telescoping sum. C_6 can be bounded by Lemma 8.

Taking expectation over C_7,

$$\begin{aligned}
&\mathbb{E}\left[\frac{3\eta}{2}\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)]\right\|^2\right] \\
&= \mathbb{E}\left[\frac{3\eta}{2K^2}\left\|\sum_{k=1}^{K}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)]\right\|^2\right] \\
&= \mathbb{E}\Bigg[\frac{3\eta}{2K^2}\Bigg(\sum_{k=1}^{K}\|\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)\|^2 \\
&\quad +2\sum_{k=1}^{K}\sum_{j=i+1}^{K}\left\langle \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k), \nabla_{\mathbf{v}} f_j(\mathbf{v}_{t-1}^j, \alpha_{t-1}^j) - \nabla_{\mathbf{v}} F_j^s(\mathbf{v}_{t-1}^j, \alpha_{t-1}^j; z_{t-1}^j)\right\rangle\Bigg)\Bigg] \\
&\leq \frac{3\eta\sigma^2}{2K}.
\end{aligned} \tag{41}$$

The last inequality holds because $\|\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)\|^2 \leq \sigma^2$ and $\mathbb{E}\langle \nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k), \nabla_{\mathbf{v}} f_j(\mathbf{v}_{t-1}^j, \alpha_{t-1}^j) - \nabla_{\mathbf{v}} F_j(\mathbf{v}_{t-1}^j, \alpha_{t-1}^j; z_{t-1}^j)\rangle = 0$ for any $k \neq j$ as each machine draws data independently. Similarly, we take expectation over C_8 and have

$$\mathbb{E}\left[\frac{3\eta}{2}\left\|\frac{1}{K}\sum_{k=1}^{K}[\nabla_\alpha f_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k)-\nabla_\alpha F_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k;\mathbf{z}_{t-1}^k)]\right\|^2\right]\leq\frac{3\eta\sigma^2}{2K}. \tag{42}$$

Plugging (41) and (42) into (97), and taking expectation, it yields

$$\begin{aligned}
&\mathbb{E}[f^s(\bar{\mathbf{v}},\alpha)-f^s(\mathbf{v},\bar{\alpha})\\
&\leq\mathbb{E}\bigg\{\frac{1}{T}\left(\frac{1}{2\eta}-\frac{\ell}{3}\right)\|\bar{\mathbf{v}}_0-\mathbf{v}\|^2+\frac{1}{2\eta T}\|\tilde{\mathbf{v}}_0-\mathbf{v}\|^2+\frac{1}{T}\left(\frac{1}{2\eta}-\frac{\mu_2}{3}\right)\|\bar{\alpha}_0-\alpha\|^2+\frac{1}{2\eta T}\|\tilde{\alpha}_0-\alpha\|^2\\
&\quad+\frac{1}{T}\sum_{t=1}^{T}\left(\frac{3\ell^2}{2\mu_2}+\frac{3\ell}{2}\right)\frac{1}{K}\sum_{k=1}^{K}\|\bar{\mathbf{v}}_{t-1}-\mathbf{v}_{t-1}^k\|^2+\frac{1}{T}\sum_{t=1}^{T}\left(\frac{3\ell}{2}+\frac{3\ell^2}{2\mu_2}\right)\frac{1}{K}\sum_{k=1}^{K}(\bar{\alpha}_{t-1}-\alpha_{t-1}^k)^2\\
&\quad+\frac{1}{T}\sum_{t=1}^{T}\frac{3\eta\sigma^2}{K}\bigg\}\\
&\leq\frac{1}{\eta T}\|\mathbf{v}_0-\mathbf{v}\|^2+\frac{1}{\eta T}\|\alpha_0-\alpha\|^2+\left(\frac{3\ell^2}{2\mu_2}+\frac{3\ell}{2}\right)(12\eta^2 I\sigma^2+36\eta^2I^2D^2)\mathbb{I}_{I>1}+\frac{3\eta\sigma^2}{K},
\end{aligned}$$

where we use Lemma 8, $\mathbf{v}_0=\bar{\mathbf{v}}_0$, and $\alpha_0=\bar{\alpha}_0$ in the last inequality. □

B.3. Main Proof of Theorem 1

Proof. Since $f(\mathbf{v},\alpha)$ is ℓ-smooth (thus ℓ-weakly convex) in $\mathbf{v}$ for any α, $\phi(\mathbf{v})=\max_{\alpha'}f(\mathbf{v},\alpha')$ is also ℓ-weakly convex. Taking $\gamma=2\ell$, we have

$$\begin{aligned}
\phi(\mathbf{v}_{s-1})&\geq\phi(\mathbf{v}_s)+\langle\partial\phi(\mathbf{v}_s),\mathbf{v}_{s-1}-\mathbf{v}_s\rangle-\frac{\ell}{2}\|\mathbf{v}_{s-1}-\mathbf{v}_s\|^2\\
&=\phi(\mathbf{v}_s)+\langle\partial\phi(\mathbf{v}_s)+2\ell(\mathbf{v}_s-\mathbf{v}_{s-1}),\mathbf{v}_{s-1}-\mathbf{v}_s\rangle+\frac{3\ell}{2}\|\mathbf{v}_{s-1}-\mathbf{v}_s\|^2\\
&\overset{(a)}{=}\phi(\mathbf{v}_s)+\langle\partial\phi_s(\mathbf{v}_s),\mathbf{v}_{s-1}-\mathbf{v}_s\rangle+\frac{3\ell}{2}\|\mathbf{v}_{s-1}-\mathbf{v}_s\|^2\\
&\overset{(b)}{=}\phi(\mathbf{v}_s)-\frac{1}{2\ell}\langle\partial\phi_s(\mathbf{v}_s),\partial\phi_s(\mathbf{v}_s)-\partial\phi(\mathbf{v}_s)\rangle+\frac{3}{8\ell}\|\partial\phi_s(\mathbf{v}_s)-\partial\phi(\mathbf{v}_s)\|^2\\
&=\phi(\mathbf{v}_s)-\frac{1}{8\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2-\frac{1}{4\ell}\langle\partial\phi_s(\mathbf{v}_s),\partial\phi(\mathbf{v}_s)\rangle+\frac{3}{8\ell}\|\partial\phi(\mathbf{v}_s)\|^2,
\end{aligned}\tag{43}$$

where (a) and (b) hold by the definition of $\phi_s(\mathbf{v})$.

Rearranging the terms in (43) yields

$$\begin{aligned}
\phi(\mathbf{v}_s)-\phi(\mathbf{v}_{s-1})&\leq\frac{1}{8\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2+\frac{1}{4\ell}\langle\partial\phi_s(\mathbf{v}_s),\partial\phi(\mathbf{v}_s)\rangle-\frac{3}{8\ell}\|\partial\phi(\mathbf{v}_s)\|^2\\
&\overset{(a)}{\leq}\frac{1}{8\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2+\frac{1}{8\ell}(\|\partial\phi_s(\mathbf{v}_s)\|^2+\|\partial\phi(\mathbf{v}_s)\|^2)-\frac{3}{8\ell}\|\phi(\mathbf{v}_s)\|^2\\
&=\frac{1}{4\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2-\frac{1}{4\ell}\|\partial\phi(\mathbf{v}_s)\|^2\\
&\overset{(b)}{\leq}\frac{1}{4\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2-\frac{\mu}{2\ell}(\phi(\mathbf{v}_s)-\phi(\mathbf{v}_*))
\end{aligned}\tag{44}$$

where (a) holds by using $\langle\mathbf{a},\mathbf{b}\rangle\leq\frac{1}{2}(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2)$, and (b) holds by the μ-PL property of $\phi(\mathbf{v})$.

Thus, we have

$$(4\ell+2\mu)\left(\phi(\mathbf{v}_s)-\phi(\mathbf{v}_*)\right)-4\ell(\phi(\mathbf{v}_{s-1})-\phi(\mathbf{v}_*))\leq\|\partial\phi_s(\mathbf{v}_s)\|^2. \tag{45}$$

Since $\gamma = 2\ell$, $f^s(\mathbf{v}, \alpha)$ is ℓ-strongly convex in $\mathbf{v}$ and $\mu_2 = 2p(1-p)$ strong concave in α. Apply Lemma 3 to f^s, we know that

$$\frac{\ell}{4}\|\hat{\mathbf{v}}_s(\alpha_s) - \mathbf{v}_0^s\|^2 + \frac{\mu_2}{4}\|\hat{\alpha}_s(\mathbf{v}_s) - \alpha_0^s\|^2 \leq \text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) + \text{Gap}_s(\mathbf{v}_s, \alpha_s). \tag{46}$$

By the setting of $\eta_s = \eta_0 \exp\left(-(s-1)\frac{2\mu}{c+2\mu}\right)$, and $T_s = \frac{212}{\eta_0 \min\{\ell, \mu_2\}} \exp\left((s-1)\frac{2\mu}{c+2\mu}\right)$, we note that $\frac{1}{\eta_s T_s} \leq \frac{\min\{\ell,\mu_2\}}{212}$. Set I_s such that $\left(\frac{3\ell^2}{2\mu_2} + \frac{3\ell}{2}\right)(12\eta_s^2 I_s + 36\eta^2 I_s^2 D^2) \leq \frac{\eta_s \sigma^2}{K}$, where the specific choice of I_s will be made later. Applying Lemma 1 with $\hat{\mathbf{v}}_s(\alpha_s) = \arg\min\limits_{\mathbf{v}'} f^s(\mathbf{v}', \alpha_s)$ and $\hat{\alpha}_s(\mathbf{v}_s) = \arg\max\limits_{\alpha'} f^s(\mathbf{v}_s, \alpha')$, we have

$$\begin{aligned}
&\mathbb{E}[\text{Gap}_s(\mathbf{v}_s, \alpha_s)] \leq \frac{4\eta_s\sigma^2}{K} + \frac{1}{53}\mathbb{E}\left[\frac{\ell}{4}\|\hat{\mathbf{v}}_s(\alpha_s) - \mathbf{v}_0^s\|^2 + \frac{\mu_2}{4}\|\hat{\alpha}_s(\mathbf{v}_s) - \alpha_0^s\|^2\right] \\
&\leq \frac{4\eta_s\sigma^2}{K} + \frac{1}{53}\mathbb{E}\left[\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) + \text{Gap}_s(\mathbf{v}_s, \alpha_s)\right].
\end{aligned} \tag{47}$$

Since $\phi(\mathbf{v})$ is L-smooth and $\gamma = 2\ell$, then $\phi_s(\mathbf{v})$ is $\hat{L} = (L + 2\ell)$-smooth. According to Theorem 2.1.5 of (Nesterov, 2004), we have

$$\begin{aligned}
&\mathbb{E}[\|\partial\phi_s(\mathbf{v}_s)\|^2] \leq 2\hat{L}\mathbb{E}(\phi_s(\mathbf{v}_s) - \min_{x\in\mathbb{R}^d}\phi_s(\mathbf{v})) \leq 2\hat{L}\mathbb{E}[\text{Gap}_s(\mathbf{v}_s, \alpha_s)] \\
&= 2\hat{L}\mathbb{E}[4\text{Gap}_s(\mathbf{v}_s, \alpha_s) - 3\text{Gap}_s(\mathbf{v}_s, \alpha_s)] \\
&\leq 2\hat{L}\mathbb{E}\left[4\left(\frac{4\eta_s\sigma^2}{K} + \frac{1}{53}\left(\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) + \text{Gap}_s(\mathbf{v}_s, \alpha_s)\right)\right) - 3\text{Gap}_s(\mathbf{v}_s, \alpha_s)\right] \\
&= 2\hat{L}\mathbb{E}\left[\frac{16\eta_s\sigma^2}{K} + \frac{4}{53}\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) - \frac{155}{53}\text{Gap}_s(\mathbf{v}_s, \alpha_s)\right]
\end{aligned} \tag{48}$$

Applying Lemma 4 to (48), we have

$$\begin{aligned}
\mathbb{E}[\|\partial\phi_s(\mathbf{v}_s)\|^2] \leq 2\hat{L}\mathbb{E}\Bigg[&\frac{16\eta_s\sigma^2}{K} + \frac{4}{53}\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) \\
&- \frac{155}{53}\left(\frac{3}{50}\text{Gap}_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1}) + \frac{4}{5}(\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_0^s))\right)\Bigg] \\
= 2\hat{L}\mathbb{E}\Bigg[\frac{16\eta_s\sigma^2}{K} + \frac{4}{53}\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) &- \frac{93}{530}\text{Gap}_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1}) - \frac{124}{53}(\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_0^s))\Bigg].
\end{aligned} \tag{49}$$

Combining this with (45), rearranging the terms, and defining a constant $c = 4\ell + \frac{248}{53}\hat{L} \in O(L + \ell)$, we get

$$\begin{aligned}
&(c + 2\mu)\,\mathbb{E}[\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_*)] + \frac{93}{265}\hat{L}\mathbb{E}[\text{Gap}_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1})] \\
&\leq \left(4\ell + \frac{248}{53}\hat{L}\right)\mathbb{E}[\phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_*)] + \frac{8\hat{L}}{53}\mathbb{E}[\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s)] + \frac{32\eta_s\hat{L}\sigma^2}{K} \\
&\leq c\mathbb{E}\left[\phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_*) + \frac{8\hat{L}}{53c}\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s)\right] + \frac{32\eta_s\hat{L}\sigma^2}{K}
\end{aligned} \tag{50}$$

Using the fact that $\hat{L} \geq \mu$,

$$(c + 2\mu)\frac{8\hat{L}}{53c} = \left(4\ell + \frac{248}{53}\hat{L} + 2\mu\right)\frac{8\hat{L}}{53(4\ell + \frac{248}{53}\hat{L})} \leq \frac{8\hat{L}}{53} + \frac{16\mu\hat{L}}{248\hat{L}} \leq \frac{93}{265}\hat{L}. \tag{51}$$

Then, we have

$$\begin{aligned}&(c+2\mu)\mathbb{E}\left[\phi(\mathbf{v}_0^{s+1})-\phi(\mathbf{v}_*)+\frac{8\hat{L}}{53c}\text{Gap}_{s+1}(\mathbf{v}_0^{s+1},\alpha_0^{s+1})\right]\\&\leq c\mathbb{E}\left[\phi(\mathbf{v}_0^{s})-\phi(\mathbf{v}_*)+\frac{8\hat{L}}{53c}\text{Gap}_{s}(\mathbf{v}_0^{s},\alpha_0^{s})\right]+\frac{32\eta_s\hat{L}\sigma^2}{K}.\end{aligned} \tag{52}$$

Defining $\Delta_s=\phi(\mathbf{v}_0^s)-\phi(\mathbf{v}_*)+\frac{8\hat{L}}{53c}\text{Gap}_s(\mathbf{v}_0^s,\alpha_0^s)$, then

$$\mathbb{E}[\Delta_{s+1}]\leq\frac{c}{c+2\mu}\mathbb{E}[\Delta_s]+\frac{32\eta_s\hat{L}\sigma^2}{(c+2\mu)K} \tag{53}$$

Using this inequality recursively, it yields

$$E[\Delta_{S+1}]\leq\left(\frac{c}{c+2\mu}\right)^S E[\Delta_1]+\frac{32\hat{L}\sigma^2}{(c+2\mu)K}\sum_{s=1}^{S}\left(\eta_s\left(\frac{c}{c+2\mu}\right)^{S+1-s}\right) \tag{54}$$

By definition,

$$\begin{aligned}\Delta_1&=\phi(\mathbf{v}_0^1)-\phi(\mathbf{v}^*)+\frac{8\hat{L}}{53c}\widehat{Gap}_1(\mathbf{v}_0^1,\alpha_0^1)\\&=\phi(\mathbf{v}_0)-\phi(\mathbf{v}^*)+\left(f(\mathbf{v}_0,\hat{\alpha}_1(\mathbf{v}_0))+\frac{\gamma}{2}\|\mathbf{v}_0-\mathbf{v}_0\|^2-f(\hat{\mathbf{v}}_1(\alpha_0),\alpha_0)-\frac{\gamma}{2}\|\hat{\mathbf{v}}_1(\alpha_0)-\mathbf{v}_0\|^2\right)\\&\leq\epsilon_0+f(\mathbf{v}_0,\hat{\alpha}_1(\mathbf{v}_0))-f(\hat{\mathbf{v}}(\alpha_0),\alpha_0)\leq 2\epsilon_0.\end{aligned} \tag{55}$$

Using inequality $1-x\leq\exp(-x)$, we have

$$\begin{aligned}&\mathbb{E}[\Delta_{S+1}]\leq\exp\left(\frac{-2\mu S}{c+2\mu}\right)\mathbb{E}[\Delta_1]+\frac{32\eta_0\hat{L}\sigma^2}{(c+2\mu)K}\sum_{s=1}^{S}\exp\left(-\frac{2\mu S}{c+2\mu}\right)\\&\leq 2\epsilon_0\exp\left(\frac{-2\mu S}{c+2\mu}\right)+\frac{32\eta_0\hat{L}\sigma^2}{(c+2\mu)K}S\exp\left(-\frac{2\mu S}{(c+2\mu)}\right).\end{aligned}$$

To make this less than ϵ, it suffices to make

$$\begin{aligned}&2\epsilon_0\exp\left(\frac{-2\mu S}{c+2\mu}\right)\leq\frac{\epsilon}{2}\\&\frac{32\eta_0\hat{L}\sigma^2}{(c+2\mu)K}S\exp\left(-\frac{2\mu S}{c+2\mu}\right)\leq\frac{\epsilon}{2}\end{aligned} \tag{56}$$

Let S be the smallest value such that $\exp\left(\frac{-2\mu S}{c+2\mu}\right)\leq\min\{\frac{\epsilon}{4\epsilon_0},\frac{(c+2\mu)K\epsilon}{64\eta_0\hat{L}S\sigma^2}\}$. We can set $S=\max\left\{\frac{c+2\mu}{2\mu}\log\frac{4\epsilon_0}{\epsilon},\frac{c+2\mu}{2\mu}\log\frac{64\eta_0\hat{L}S\sigma^2}{(c+2\mu)K\epsilon}\right\}$.

Then, the total iteration complexity is

$$
\begin{aligned}
\sum_{s=1}^{S} T_s &\leq O\left(\frac{424}{\eta_0 \min\{\ell, \mu_2\}} \sum_{s=1}^{S} \exp\left((s-1)\frac{2\mu}{c+2\mu}\right)\right) \\
&\leq O\bigg(\frac{1}{\eta_0 \min\{\ell, \mu_2\}} \frac{\exp(S\frac{2\mu}{c+2\mu}) - 1}{\exp(\frac{2\mu}{c+2\mu}) - 1}\bigg) \\
&\overset{(a)}{\leq} \widetilde{O}\left(\frac{c}{\eta_0 \mu \min\{\ell, \mu_2\}} \max\left\{\frac{\epsilon_0}{\epsilon}, \frac{\eta_0 \hat{L} S \sigma^2}{(c+2\mu)K\epsilon}\right\}\right) \\
&\leq \widetilde{O}\left(\max\left\{\frac{(L+\ell)\epsilon_0}{\eta_0 \mu \min\{\ell, \mu_2\}\epsilon}, \frac{(L+\ell)^2 \sigma^2}{\mu^2 \min\{\ell, \mu_2\} K\epsilon}\right\}\right) \\
&\leq \widetilde{O}\left(\max\left\{\frac{1}{\mu_1 \mu_2^2 \epsilon}, \frac{1}{\mu_1^2 \mu_2^3 K\epsilon}\right\}\right),
\end{aligned}
\tag{57}
$$

where (a) uses the setting of S and $\exp(x) - 1 \geq x$, and $\widetilde{O}$ suppresses logarithmic factors.

$\eta_s = \eta_0 \exp(-(s-1)\frac{2\mu}{c+2\mu}), T_s = \frac{212}{\eta_0 \mu_2} \exp\left((s-1)\frac{2\mu}{c+2\mu}\right)$.

Next, we will analyze the communication cost. We investigate both $D = 0$ and $D > 0$ cases.

(i) Homogeneous Data (D = 0): To assure $\left(\frac{3\ell^2}{2\mu_2} + \frac{3\ell}{2}\right)(12\eta_s^2 I_s + 36\eta^2 I_s^2 D^2) \leq \frac{\eta_s \sigma^2}{K}$ which we used in above proof, we take $I_s = \frac{1}{MK\eta_s} = \frac{\exp((s-1)\frac{2\mu}{c+2\mu})}{MK\eta_0}$, where M is a proper constant.

If $\frac{1}{MK\eta_0} > 1$, then $I_s = \max(1, \frac{\exp((s-1)\frac{2\mu}{c+2\mu})}{MK\eta_0}) = \frac{\exp((s-1)\frac{2\mu}{c+2\mu})}{MK\eta_0}$.

Otherwise, $\frac{1}{MK\eta_0} \leq 1$, then $K_s = 1$ for $s \leq S_1 := \frac{c+2\mu}{2\mu}\log(MK\eta_0) + 1$ and $K_s = \frac{\exp((s-1)\frac{2\mu}{c+2\mu})}{MK\eta_0}$ for $s > S_1$.

$$
\begin{aligned}
\sum_{s=1}^{S_1} T_s &= \sum_{s=1}^{S_1} O\left(\frac{212}{\eta_0} \exp\left((s-1)\frac{2\mu}{c+2\mu}\right)\right) \\
&= \widetilde{O}\left(\frac{212}{\eta_0} \frac{\exp\left(\frac{2\mu}{c+2\mu}S_1\right) - 1}{\exp\left(\exp(\frac{2\mu}{c+2\mu}) - 1\right)}\right) \\
&= \widetilde{O}\left(\frac{K}{\mu}\right)
\end{aligned}
\tag{58}
$$

Thus, for both above cases, the total communication complexity can be bounded by

$$
\begin{aligned}
&\sum_{s=1}^{S_1} T_s + \sum_{s=S_1+1}^{S} \frac{T_s}{I_s} \\
&= \widetilde{O}\left(\frac{K}{\mu} + KS\right) \leq \widetilde{O}\left(\frac{K}{\mu}\right).
\end{aligned}
\tag{59}
$$

(ii) Heterogeneous Data ($D > 0$):

To assure $\left(\frac{3\ell^2}{2\mu_2} + \frac{3\ell}{2}\right)(12\eta_s^2 I_s + 36\eta^2 I_s^2 D^2) \leq \frac{\eta_s \sigma^2}{K}$ which we used in above proof, we take $I_s = \frac{1}{M\sqrt{K\eta_s}}$, where M is proper constant.

If $\frac{1}{M\sqrt{N\eta_0}} \leq 1$, then $I_s = 1$ for $s \leq S_2 := \frac{c+2\mu}{2\mu}\log(M^2K\eta_0) + 1$ and $I_s = \frac{\exp((s-1)\frac{2\mu}{c+2\mu})}{N\eta_0}$ for $s > S_2$.

$$\begin{aligned} \sum_{s=1}^{S_2} T_s &= \sum_{s=1}^{S_2} O\left(\frac{212}{\eta_0}\exp\left((s-1)\frac{2\mu}{c+2\mu}\right)\right) \\ &= \widetilde{O}\left(\frac{K}{\mu}\right) \end{aligned} \tag{60}$$

Thus, the communication complexity can be bounded by

$$\begin{aligned} \sum_{s=1}^{S_2} T_s + \sum_{s=S_2+1}^{S} \frac{T_s}{I_s} &= \widetilde{O}\left(\frac{K}{\mu} + \sqrt{K}\exp\left(\frac{(s-1)\frac{2\mu}{c+2\mu}}{2}\right)\right) \\ &\leq \widetilde{O}(\frac{K}{\mu} + \sqrt{K}\frac{\exp\left(\frac{S}{2}\frac{2\mu}{c+2\mu}\right)-1}{\exp\frac{\mu}{c+2\mu}-1}) \\ &\leq O\left(\frac{K}{\mu} + \frac{1}{\mu^{3/2}\epsilon^{1/2}}\right). \end{aligned} \tag{61}$$

$\square$

C. Baseline: Naive Parallel Algorithm

Note that if we set $I_s = 1$ for all s, CODA+ will be reduced to a naive parallel version of PPD-SG (Liu et al., 2020). We analyze this naive parallel algorithm in the following theorem.

Theorem 3. *Consider Algorithm 1 with $I_s = 1$. Set $\gamma = 2\ell$, $\hat{L} = L + 2\ell$, $c = \frac{\mu/\hat{L}}{5+\mu/\hat{L}}$.*

(1) If $M < \frac{1}{K\mu\epsilon}$, set $\eta_s = \eta_0\exp(-(s-1)c) \leq O(1)$ and $T_s = \frac{212}{\eta_0\min(\ell,\mu_2)}\exp((s-1)c)$, then the communication/iteration complexity is $\widetilde{O}\left(\max\left(\frac{\Delta_0}{\mu\epsilon\eta_0 K}, \frac{\hat{L}}{\mu^2K\epsilon}\right)\right)$ to return $\mathbf{v}_S$ such that $\mathbb{E}[\phi(\mathbf{v}_S) - \phi(\mathbf{v}_\phi^*)] \leq \epsilon$.

(2) If $M \geq \frac{1}{K\mu\epsilon}$, set $\eta_s = \min(\frac{1}{3\ell+3\ell^2/\mu_2}, \frac{1}{4\ell})$ and $T_s = \frac{212}{\eta_s\min\{\ell,\mu_2\}}$, then the communication/iteration complexity is $\widetilde{O}\left(\frac{1}{\mu}\right)$ to return $\mathbf{v}_S$ such that $\mathbb{E}[\phi(\mathbf{v}_S) - \phi(\mathbf{v}_\phi^*)] \leq \epsilon$.

Proof. (1) If $M < \frac{1}{K\mu\epsilon}$, note that the setting of η_s and T_s are identical to that in CODA+ (Theorem 1). However, as a batch of M is used on each machine at each iteration, the variance at each iteration is reduced to $\frac{\sigma^2}{KM}$. Therefore, by similar analysis of Theorem 1 (specifically (57)), we see that the iteration complexity of NPA is $\widetilde{O}\left(\frac{1}{\mu\epsilon} + \frac{1}{\mu^2KM\epsilon}\right)$. Thus, the sample complexity of each machines is $\widetilde{O}\left(\frac{M}{\mu\epsilon} + \frac{1}{\mu^2K\epsilon}\right)$.

(2) If $M \geq \frac{1}{K\mu\epsilon}$, . Note $\frac{1}{\eta_sT_s} \leq \frac{\min\{\ell,\mu_2\}}{212}$, we can follow the proof of Theorem 1 and derive

$$\begin{aligned} \Delta_{s+1} &\leq \frac{c}{c+2\mu}\mathbb{E}[\Delta_s] + \frac{32\eta_s\hat{L}\sigma^2}{KM} \\ &\leq \frac{c}{c+2\mu}\mathbb{E}[\Delta_s] + 32\eta_s\hat{L}\sigma^2\mu\epsilon \end{aligned} \tag{62}$$

where the first inequality is similar to (53) and the Δ is defined as that in Theorem 1. Thus,

$$\begin{aligned}\Delta_{S+1} &\leq \left(\frac{c}{c+2\mu}\right)^S + \mu\epsilon O\left(\sum_{s=1}^{S}\left(\frac{c}{c+2\mu}\right)^{s-1}\right) \\ &\leq \left(\frac{c}{c+2\mu}\right)^S + O(\epsilon) \\ &\leq \exp\left(\frac{-2\mu S}{c+2\mu}\right) + O(\epsilon)\end{aligned} \tag{63}$$

Therefore, it suffices to take $S = \widetilde{O}\left(\frac{1}{\mu}\right)$. Hence, the total number of communication is $S \cdot T_s = \widetilde{O}\left(\frac{1}{\mu}\right)$ and the sample complexity on each machine is $\widetilde{O}\left(\frac{M}{\mu}\right)$.

□

D. Proof of Lemma 2

In this section, we will prove Lemma 2, which is the convergence analysis of one stage in CODASCA.

First, the duality gap in stage s can be bounded as

Lemma 9. *For any* $\mathbf{v}, \alpha$*,*

$$\begin{aligned}&\frac{1}{R}\sum_{r=1}^{R}[f^s(\mathbf{v}_r, \alpha) - f^s(\mathbf{v}, \alpha_r)] \\ &\leq \frac{1}{R}\sum_{r=1}^{R}\Bigg[\underbrace{\langle \partial_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v}\rangle}_{B4} + \underbrace{\langle \partial_{\alpha} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \alpha - \alpha_r\rangle}_{B5} \\ &\quad + \frac{3\ell + 3\ell^2/\mu_2}{2}\|\mathbf{v}_r - \mathbf{v}_{r-1}\|^2 + 2\ell(\alpha_r - \alpha_{r-1})^2 - \frac{\ell}{3}\|\mathbf{v}_{r-1} - \mathbf{v}\|^2 - \frac{\mu_2}{3}(\alpha_{r-1} - \alpha)^2\Bigg]\end{aligned}$$

Proof. By ℓ-strongly convexity of $f^s(\mathbf{v}, \alpha)$ in $\mathbf{v}$, we have

$$f^s(\mathbf{v}_{r-1}, \alpha_{r-1}) + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v} - \mathbf{v}_{r-1}\rangle + \frac{\ell}{2}\|\mathbf{v}_{r-1} - \mathbf{v}\|^2 \leq f^s(\mathbf{v}, \alpha_{r-1}). \tag{64}$$

By 3ℓ-smoothness of $f^s(\mathbf{v}, \alpha)$ in $\mathbf{v}$, we have

$$\begin{aligned}f^s(\mathbf{v}_r, \alpha) &\leq f^s(\mathbf{v}_{r-1}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha), \mathbf{v}_r - \mathbf{v}_{r-1}\rangle + \frac{3\ell}{2}\|\mathbf{v}_r - \mathbf{v}_{r-1}\|^2 \\ &= f^s(\mathbf{v}_{r-1}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v}_{r-1}\rangle + \frac{3\ell}{2}\|\mathbf{v}_r - \mathbf{v}_{r-1}\|^2 \\ &\quad + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha) - \partial_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v}_{r-1}\rangle \\ &\overset{(a)}{\leq} f^s(\mathbf{v}_{r-1}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v}_{r-1}\rangle + \frac{3\ell}{2}\|\mathbf{v}_r - \mathbf{v}_{r-1}\|^2 \\ &\quad + \ell|\alpha_{r-1} - \alpha|\|\mathbf{v}_r - \mathbf{v}_{r-1}\| \\ &\overset{(b)}{\leq} f^s(\mathbf{v}_{r-1}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v}_{r-1}\rangle + \frac{3\ell}{2}\|\mathbf{v}_r - \mathbf{v}_{r-1}\|^2 \\ &\quad + \frac{\mu_2}{6}(\alpha_{r-1} - \alpha)^2 + \frac{3\ell^2}{2\mu_2}\|\mathbf{v}_r - \mathbf{v}_{r-1}\|^2,\end{aligned} \tag{65}$$

where (a) holds because that we know $\partial_{\mathbf{v}} f^s(\mathbf{v}, \alpha)$ is ℓ-Lipschitz in α since $f(\mathbf{v}, \alpha)$ is ℓ-smooth and (b) holds by Young's inequality.

Adding (64) and (65), by rearranging terms, we have

$$\begin{aligned} & f^s(\mathbf{v}_{r-1}, \alpha_{r-1}) + f^s(\mathbf{v}_r, \alpha) \\ & \leq f^s(\mathbf{v}, \alpha_{r-1}) + f^s(\mathbf{v}_{r-1}, \alpha) + \langle \partial_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v} \rangle \\ & \quad + \frac{3\ell + 3\ell^2/\mu_2}{2} \|\mathbf{v}_r - \mathbf{v}_{r-1}\|^2 - \frac{\ell}{2} \|\mathbf{v}_{r-1} - \mathbf{v}\|^2 + \frac{\mu_2}{6} (\alpha_{r-1} - \alpha)^2. \end{aligned} \tag{66}$$

We know $f^s(\mathbf{v}, \alpha)$ is μ_2-strong concave in α ($-f^s(\mathbf{v}, \alpha)$ is μ_2-strong convexity of in α). Thus, we have

$$-f^s(\mathbf{v}_{r-1}, \alpha_{r-1}) - \langle \partial_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \alpha - \alpha_{r-1} \rangle + \frac{\mu_2}{2} (\alpha - \alpha_{r-1})^2 \leq -f^s(\mathbf{v}_{r-1}, \alpha). \tag{67}$$

Since $f^s(\mathbf{v}, \alpha)$ is ℓ-smooth in α, we get

$$\begin{aligned} -f^s(\mathbf{v}, \alpha_r) & \leq -f^s(\mathbf{v}, \alpha_{r-1}) - \langle \partial_\alpha f^s(\mathbf{v}, \alpha_{r-1}), \alpha_r - \alpha_{r-1} \rangle + \frac{\ell}{2} (\alpha_r - \alpha_{r-1})^2 \\ & = -f^s(\mathbf{v}, \alpha_{r-1}) - \langle \partial_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \alpha_r - \alpha_{r-1} \rangle + \frac{\ell}{2} (\alpha_r - \alpha_{r-1})^2 \\ & \quad - \langle \partial_\alpha (f^s(\mathbf{v}, \alpha_{r-1}) - f^s(\mathbf{v}_{r-1}, \alpha_{r-1})), \alpha_r - \alpha_{r-1} \rangle \\ & \overset{(a)}{\leq} -f^s(\mathbf{v}, \alpha_{r-1}) - \langle \partial_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \alpha_r - \alpha_{r-1} \rangle + \frac{\ell}{2} (\alpha_r - \alpha_{r-1})^2 \\ & \quad + \ell \|\mathbf{v} - \mathbf{v}_{r-1}\| |\alpha_r - \alpha_{r-1}| \\ & \leq -f^s(\mathbf{v}, \alpha_{r-1}) - \langle \partial_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \alpha_r - \alpha_{r-1} \rangle + \frac{\ell}{2} (\alpha_r - \alpha_{r-1})^2 \\ & \quad + \frac{\ell}{6} \|\mathbf{v}_{r-1} - \mathbf{v}\|^2 + \frac{3\ell}{2} (\alpha_r - \alpha_{r-1})^2 \end{aligned} \tag{68}$$

where (a) holds because that $\partial_\alpha f^s(\mathbf{v}, \alpha)$ is ℓ-Lipschitz in α.
Adding (67), (68) and arranging terms, we have

$$\begin{aligned} & - f^s(\mathbf{v}_{r-1}, \alpha_{r-1}) - f^s(\mathbf{v}, \alpha_r) \leq -f^s(\mathbf{v}_{r-1}, \alpha) - f^s(\mathbf{v}, \alpha_{r-1}) - \langle \partial_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \alpha_r - \alpha \rangle \\ & \quad + 2\ell (\alpha_r - \alpha_{r-1})^2 + \frac{\ell}{6} \|\mathbf{v}_{r-1} - \mathbf{v}\|^2 - \frac{\mu_2}{2} (\alpha - \alpha_{r-1})^2. \end{aligned} \tag{69}$$

Adding (66) and (69), we get

$$\begin{aligned} & f^s(\mathbf{v}_r, \alpha) - f^s(\mathbf{v}, \alpha_r) \\ & \leq \langle \partial_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v} \rangle - \langle \partial_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \alpha_r - \alpha \rangle \\ & \quad + \frac{3\ell + 3\ell^2/\mu_2}{2} \|\mathbf{v}_r - \mathbf{v}_{r-1}\|^2 + 2\ell (\alpha_r - \alpha_{r-1})^2 \\ & \quad - \frac{\ell}{3} \|\mathbf{v}_{r-1} - \mathbf{v}\|^2 - \frac{\mu_2}{3} (\alpha_{r-1} - \alpha)^2 \end{aligned} \tag{70}$$

Taking average over $r = 1, ..., R$, we get

$$\begin{aligned} & \frac{1}{R} \sum_{r=1}^{R} [f^s(\mathbf{v}_r, \alpha) - f^s(\mathbf{v}, \alpha_r)] \\ & \leq \frac{1}{R} \sum_{r=1}^{R} \Bigg[\underbrace{\langle \partial_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v} \rangle}_{B_4} + \underbrace{\langle \partial_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \alpha - \alpha_r \rangle}_{B_5} \\ & \quad + \frac{3\ell + 3\ell^2/\mu_2}{2} \|\mathbf{v}_r - \mathbf{v}_{r-1}\|^2 + 2\ell (\alpha_r - \alpha_{r-1})^2 - \frac{\ell}{3} \|\mathbf{v}_{r-1} - \mathbf{v}\|^2 - \frac{\mu_2}{3} (\alpha_{r-1} - \alpha)^2 \Bigg] \end{aligned}$$

□

B_4 and B_5 can be bounded by the following lemma. For simplicity of notation, we define

$$\Xi_r = \frac{1}{KI}\sum_{k,t}\mathbb{E}[\|\mathbf{v}_{r,t}^k - \mathbf{v}_r\|^2 + (\alpha_{r,t}^k - \alpha_r)^2], \tag{71}$$

which is the drift of the variables between te sequence in r-th round and the ending point, and

$$\mathcal{E}_r = \frac{1}{KI}\sum_{k,t}\mathbb{E}[\|\mathbf{v}_{r,t}^k - \mathbf{v}_{r-1}\|^2 + (\alpha_{r,t}^k - \alpha_{r-1})^2], \tag{72}$$

which is the drift of the variables between te sequence in r-th round and the starting point.

B_4 can be bounded as

Lemma 10.

$$\begin{aligned}
&\mathbb{E}\left\langle \nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v}\right\rangle \\
&\leq \frac{3\ell}{2}\mathcal{E}_r + \frac{\ell}{3}\mathbb{E}\|\bar{\mathbf{v}}_r - \mathbf{v}\|^2 + \frac{3\tilde{\eta}}{2}\mathbb{E}\left\|\frac{1}{NK}\sum_{i,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k)]\right\|^2 \\
&\quad + \frac{1}{2\tilde{\eta}}\mathbb{E}(\|\mathbf{v}_{r-1} - \mathbf{v}\|^2 - \|\mathbf{v}_{r-1} - \mathbf{v}_r\|^2 - \|\mathbf{v}_r - \mathbf{v}\|^2) + \frac{1}{2\tilde{\eta}}\mathbb{E}(\|\tilde{\mathbf{v}}_{r-1} - \mathbf{v}\|^2 - \|\tilde{\mathbf{v}}_r - \mathbf{v}\|^2),
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\langle \nabla_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), y - \alpha_r\rangle \leq \frac{3\ell^2}{2\mu_2}\mathcal{E}_r + \frac{\mu_2}{3}\mathbb{E}(\bar{\alpha}_r - \alpha)^2 \\
&\quad + \frac{3\tilde{\eta}}{2}\mathbb{E}\left(\frac{1}{NK}\sum_{i,t}[\nabla_\alpha f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_\alpha F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k)]\right)^2 \\
&\quad + \frac{1}{2\tilde{\eta}}\mathbb{E}((\bar{\alpha}_{r-1} - \alpha)^2 - (\bar{\alpha}_{r-1} - \bar{\alpha}_r)^2 - (\bar{\alpha}_r - \alpha)^2) + \frac{1}{2\tilde{\eta}}\mathbb{E}((\alpha - \tilde{\alpha}_{r-1})^2 - (\alpha - \tilde{\alpha}_r)^2).
\end{aligned}$$

Proof.

$$\begin{aligned}
&\langle \nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v}\rangle \\
&= \left\langle \frac{1}{KI}\sum_{k,t}\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v}\right\rangle \\
&\leq \left\langle \frac{1}{KI}\sum_{k,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r-1}) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r,t}^k)], \mathbf{v}_r - \mathbf{v}\right\rangle \qquad \text{①} \\
&\quad + \left\langle \frac{1}{KI}\sum_{i,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k)], \mathbf{v}_r - \mathbf{v}\right\rangle \qquad \text{②} \\
&\quad + \left\langle \frac{1}{KI}\sum_{k,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k)], \mathbf{v}_r - \mathbf{v}\right\rangle \qquad \text{③} \\
&\quad + \left\langle \frac{1}{KI}\sum_{k,t}\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k), \mathbf{v}_r - \mathbf{v}\right\rangle \qquad \text{④}
\end{aligned} \tag{73}$$

Then we will bound ①, ② and ③, respectively,

$$\begin{aligned}
① &\overset{(a)}{\leq} \frac{3}{2\ell}\left\|\frac{1}{KI}\sum_{k,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r-1}) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r,t}^k)]\right\|^2 + \frac{\ell}{6}\|\mathbf{v}_r - \mathbf{v}\|^2 \\
&\overset{(b)}{\leq} \frac{3}{2\ell}\frac{1}{KI}\sum_{k,t}\|\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r-1}) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r,t}^k)\|^2 + \frac{\ell}{6}\|\mathbf{v}_r - \mathbf{v}\|^2 \\
&\overset{(c)}{\leq} \frac{3\ell}{2}\frac{1}{KI}\sum_{k,t}\|\alpha_{r-1} - \alpha_{r,t}^k\|^2 + \frac{\ell}{6}\|\mathbf{v}_r - \mathbf{v}\|^2,
\end{aligned} \tag{74}$$

where (a) follows from Young's inequality, (b) follows from Jensen's inequality. and (c) holds because $\nabla_{\mathbf{v}} f_k^s(\mathbf{v}, \alpha)$ is ℓ-smooth in α. Using similar techniques, we have

$$\begin{aligned}
② &\leq \frac{3}{2\ell}\frac{1}{KI}\sum_{k,t}\|\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k)\|^2 + \frac{\ell}{6}\|\mathbf{v}_r - \mathbf{v}\|^2 \\
&\leq \frac{3\ell}{2}\frac{1}{KI}\sum_{k,t}\|\mathbf{v}_{r-1} - \mathbf{v}_{r,t}^i\|^2 + \frac{\ell}{6}\|\mathbf{v}_r - \mathbf{v}\|^2.
\end{aligned} \tag{75}$$

Let $\hat{\mathbf{v}}_r = \arg\min_{\mathbf{v}} \left(\frac{1}{KI}\sum_{k,t}\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, y_{r,t}^k)\right)^\top \mathbf{v} + \frac{1}{2\tilde{\eta}}\|\mathbf{v} - \mathbf{v}_{r-1}\|^2$, then we have

$$\bar{\mathbf{v}}_r - \hat{\mathbf{v}}_r = \frac{\tilde{\eta}}{KI}\sum_{k,t}\left(\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, y_{r,t}^k; z_{r,t}^k)\right). \tag{76}$$

Hence we get

$$\begin{aligned}
③ &= \left\langle \frac{1}{KI}\sum_{k,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k)], \mathbf{v}_r - \hat{\mathbf{v}}_r \right\rangle \\
&\quad + \left\langle \frac{1}{KI}\sum_{k,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k)], \hat{\mathbf{v}}_r - \mathbf{v} \right\rangle \\
&= \tilde{\eta}\left\|\frac{1}{KI}\sum_{k,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k)]\right\|^2 \\
&\quad + \left\langle \frac{1}{KI}\sum_{k,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k)], \hat{\mathbf{v}}_r - \mathbf{v} \right\rangle.
\end{aligned} \tag{77}$$

Define another auxiliary sequence as

$$\tilde{\mathbf{v}}_r = \tilde{\mathbf{v}}_{r-1} - \frac{\tilde{\eta}}{KI}\sum_{k,t}\left(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, y_{r,t}^k; z_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k)\right), \text{ for } r > 0; \tilde{\mathbf{v}}_0 = \mathbf{v}_0. \tag{78}$$

Denote

$$\Theta_r(\mathbf{v}) = \left(\frac{1}{KI}\sum_{k,t}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, y_{r,t}^k; z_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k))\right)^\top \mathbf{v} + \frac{1}{2\tilde{\eta}}\|\mathbf{v} - \tilde{\mathbf{v}}_{r-1}\|^2. \tag{79}$$

Hence, for the auxiliary sequence $\tilde{\alpha}_r$, we can verify that

$$\tilde{\mathbf{v}}_r = \arg\min_{\mathbf{v}} \Theta_r(\mathbf{v}). \tag{80}$$

Since $\Theta_r(\mathbf{v})$ is $\frac{1}{\tilde{\eta}}$-strongly convex, we have

$$
\begin{aligned}
&\frac{1}{2\tilde{\eta}}\|\mathbf{v}-\tilde{\mathbf{v}}_r\|^2 \leq \Theta_r(\mathbf{v}) - \Theta_r(\tilde{\mathbf{v}}_r) \\
&= \left(\frac{1}{KI}\sum_{k,t}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k))\right)^\top \mathbf{v} + \frac{1}{2\tilde{\eta}}\|\mathbf{v}-\tilde{\mathbf{v}}_{r-1}\|^2 \\
&\quad - \left(\frac{1}{KI}\sum_{k,t}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^i, \alpha_{r,t}^k; z_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^i, \alpha_{r,t}^k))\right)^\top \tilde{\mathbf{v}}_r - \frac{1}{2\tilde{\eta}}\|\tilde{\mathbf{v}}_r-\tilde{\mathbf{v}}_{r-1}\|^2 \\
&= \left(\frac{1}{KI}\sum_{k,t}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k))\right)^\top (\mathbf{v}-\tilde{\mathbf{v}}_{r-1}) + \frac{1}{2\tilde{\eta}}\|\mathbf{v}-\tilde{\mathbf{v}}_{r-1}\|^2 \\
&\quad - \left(\frac{1}{KI}\sum_{k,t}(\nabla_{\alpha} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k) - \nabla_{\alpha} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k))\right)^\top (\tilde{\mathbf{v}}_r-\tilde{\mathbf{v}}_{r-1}) - \frac{1}{2\tilde{\eta}}\|\tilde{\mathbf{v}}_r-\tilde{\mathbf{v}}_{r-1}\|^2 \\
&\leq \left(\frac{1}{KI}\sum_{k,t}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k))\right)^\top (\mathbf{v}-\tilde{\mathbf{v}}_{r-1}) + \frac{1}{2\tilde{\eta}}\|\mathbf{v}-\tilde{\mathbf{v}}_{r-1}\|^2 \\
&\quad + \frac{\tilde{\eta}}{2}\left\|\frac{1}{KI}\sum_{k,t}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k))\right\|^2
\end{aligned} \tag{81}
$$

Adding this with (77), we get

$$
\begin{aligned}
③ \leq &\frac{3\tilde{\eta}}{2}\left\|\frac{1}{KI}\sum_{k,t}(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k))\right\|^2 + \frac{1}{2\tilde{\eta}}\|\mathbf{v}-\tilde{\mathbf{v}}_{r-1}\|^2 - \frac{1}{2\tilde{\eta}}\|\mathbf{v}-\tilde{\mathbf{v}}_r\|^2 \\
&+ \left\langle \frac{1}{KI}\sum_{k,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k)], \hat{\mathbf{v}}_r - \tilde{\mathbf{v}}_{r-1}\right\rangle
\end{aligned} \tag{82}
$$

④ can be bounded as

$$
④ = \frac{1}{\tilde{\eta}}\langle \mathbf{v}_r - \mathbf{v}_{r-1}, \mathbf{v}-\mathbf{v}_r\rangle = \frac{1}{2\tilde{\eta}}(\|\mathbf{v}_{r-1}-\mathbf{v}\|^2 - \|\mathbf{v}_{r-1}-\mathbf{v}_r\|^2 - \|\mathbf{v}_r-\mathbf{v}\|^2) \tag{83}
$$

Plug (74), (75), (82) and (83) into (73), we get

$$
\begin{aligned}
&\mathbb{E}\left\langle \nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), \mathbf{v}_r - \mathbf{v}\right\rangle \\
&\leq \frac{3\ell}{2}\mathcal{E}_r + \frac{\ell}{3}\mathbb{E}\|\bar{\mathbf{v}}_r - \mathbf{v}\|^2 + \frac{3\tilde{\eta}}{2}\mathbb{E}\left\|\frac{1}{KI}\sum_{k,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k)]\right\|^2 \\
&\quad + \frac{1}{2\tilde{\eta}}\mathbb{E}(\|\mathbf{v}_{r-1}-\mathbf{v}\|^2 - \|\mathbf{v}_{r-1}-\mathbf{v}_r\|^2 - \|\mathbf{v}_r-\mathbf{v}\|^2) + \frac{1}{2\tilde{\eta}}\mathbb{E}(\|\tilde{\mathbf{v}}_{r-1}-\mathbf{v}\|^2 - \|\tilde{\mathbf{v}}_r-\mathbf{v}\|^2)
\end{aligned}
$$

Similarly for α, noting f_k^s is ℓ-smooth and μ_2-strongly concave in α,

$$
\begin{aligned}
&\mathbb{E}\langle \nabla_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}), y - \alpha_r\rangle \leq \frac{3\ell^2}{2\mu_2}\mathcal{E}_r + \frac{\mu_2}{3}\mathbb{E}(\bar{\alpha}_r - \alpha)^2 \\
&\quad + \frac{3\tilde{\eta}}{2}\mathbb{E}\left(\frac{1}{KI}\sum_{k,t}[\nabla_\alpha f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_\alpha F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k)]\right)^2 \\
&\quad + \frac{1}{2\tilde{\eta}}\mathbb{E}((\bar{\alpha}_{r-1}-\alpha)^2 - (\bar{\alpha}_{r-1}-\bar{\alpha}_r)^2 - (\bar{\alpha}_r-\alpha)^2) + \frac{1}{2\tilde{\eta}}\mathbb{E}((\alpha-\tilde{\alpha}_{r-1})^2 - (\alpha-\tilde{\alpha}_r)^2)
\end{aligned}
$$

□

We show the following lemmas where Ξ and $\mathcal{E}$ are coupled.

Lemma 11.

$$\Xi_r \leq 4\mathcal{E}_r + 8\tilde{\eta}^2[\|\nabla_{\mathbf{v}} f(\mathbf{v}_r, \alpha_r)\|^2 + \|\nabla_\alpha f(\mathbf{v}_r, \alpha_r)\|^2] + \frac{5\tilde{\eta}^2\sigma^2}{KI}. \tag{84}$$

Proof.

$$\begin{aligned}
\mathbb{E}[\|\mathbf{v}_r - \mathbf{v}_{r-1}\|^2] &= \mathbb{E}\left\|-\frac{\tilde{\eta}}{KI}\sum_{k,t}(\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k) - c_{\mathbf{v}}^k + c_{\mathbf{v}})\right\|^2 \\
&= \mathbb{E}\left\|-\frac{\tilde{\eta}}{KI}\sum_{k,t}\left[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; z_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) + \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k)\right]\right\|^2 \\
&\leq \mathbb{E}\left\|-\frac{\tilde{\eta}}{KI}\sum_{k,t}\left[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k)\right]\right\|^2 + \frac{\tilde{\eta}^2\sigma^2}{KI} \\
&= \mathbb{E}\left\|-\frac{\tilde{\eta}}{KI}\sum_{k,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r-1})] + \tilde{\eta}\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}))\right\|^2 + \frac{\tilde{\eta}^2\sigma^2}{KI} \\
&\leq 2\mathbb{E}\left\|-\frac{\tilde{\eta}}{KI}\sum_{k,t}[\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r-1})]\right\|^2 + 2\tilde{\eta}^2\mathbb{E}\left\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\right\|^2 + \frac{\tilde{\eta}^2\sigma^2}{KI} \\
&\leq \frac{2\tilde{\eta}^2\ell^2}{KI}\sum_{k,t}\mathbb{E}[\|\mathbf{v}_{r,t}^k - \mathbf{v}_{r-1}\|^2 + (\alpha_{r,t}^k - \alpha_{r-1})^2] + 2\tilde{\eta}^2\mathbb{E}\left\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\right\|^2 + \frac{\tilde{\eta}^2\sigma^2}{KI} \\
&\leq 2\tilde{\eta}^2\ell^2\mathcal{E}_r + 2\tilde{\eta}^2\mathbb{E}\left\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\right\|^2 + \frac{\tilde{\eta}^2\sigma^2}{KI}
\end{aligned} \tag{85}$$

Similarly,

$$\mathbb{E}[(\alpha_r - \alpha_{r-1})^2] \leq 2\tilde{\eta}^2\ell^2\mathcal{E}_r + 2\tilde{\eta}^2\mathbb{E}\left(\nabla_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\right)^2 + \frac{\tilde{\eta}^2\sigma^2}{KI}. \tag{86}$$

Using the 3ℓ-smoothness of f^s and combining with above results,

$$\begin{aligned}
&\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2 + (\nabla_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}))^2 \\
&= \|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}) - \nabla_{\mathbf{v}} f^s(\mathbf{v}_r, \alpha_r) + \nabla_{\mathbf{v}} f^s(\mathbf{v}_r, \alpha_r)\|^2 + (\nabla_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}) - \nabla_\alpha f^s(\mathbf{v}_r, \alpha_r) + \nabla_\alpha f^s(\mathbf{v}_r, \alpha_r))^2 \\
&\leq 2[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r, \alpha_r)\|^2 + \|\nabla_\alpha f^s(\mathbf{v}_r, \alpha_r)\|^2] + 18\ell^2(\|\mathbf{v}_{r-1} - \mathbf{v}_r\|^2 + (\alpha_{r-1} - \alpha_r)^2) \\
&\leq 2[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r, \alpha_r)\|^2 + \|\nabla_\alpha f^s(\mathbf{v}_r, \alpha_r)\|^2] + 60\ell^4\tilde{\eta}^2\mathcal{E}_r + \frac{40\tilde{\eta}^2\ell^2\sigma^2}{KI} \\
&\leq 2[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r, \alpha_r)\|^2 + \|\nabla_\alpha f^s(\mathbf{v}_r, \alpha_r)\|^2] + \frac{\ell^2}{24}\mathcal{E}_r + \frac{\sigma^2}{144KI}.
\end{aligned} \tag{87}$$

$$
\begin{aligned}
\Xi_r &= \frac{1}{KI}\sum_{k,t}\mathbb{E}[\|\mathbf{v}_{r,t}^k - \mathbf{v}_r\|^2 + (\alpha_{r,t}^k - \alpha_r)^2] \\
&\leq \frac{2}{KI}\sum_{k,t}\mathbb{E}[\|\mathbf{v}_{r,t}^k - \mathbf{v}_{r-1}\|^2 + \|\mathbf{v}_{r-1} - \mathbf{v}_r\|^2 + (\alpha_{r,t}^k - \alpha_{r-1})^2 + (\alpha_{r-1} - \alpha_r)^2] \\
&\leq 2\mathcal{E}_r + 2\mathbb{E}[\|\mathbf{v}_{r-1} - \mathbf{v}_r\|^2 + (\alpha_{r-1} - \alpha_r)^2] \\
&\leq 2\mathcal{E}_r + 8\tilde{\eta}^2\ell^2\mathcal{E}_r + 4\tilde{\eta}^2\mathbb{E}[(\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1},\alpha_{r-1}))^2 + (\nabla_\alpha f^s(\mathbf{v}_{r-1},\alpha_{r-1}))^2] + \frac{4\tilde{\eta}^2\sigma^2}{KI} \\
&\leq 3\mathcal{E}_r + 4\tilde{\eta}^2\left(2[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r,\alpha_r)\|^2 + (\nabla_\alpha f^s(\mathbf{v}_r,\alpha_r))^2] + \frac{\ell^2}{24}\mathcal{E}_r + \frac{\sigma^2}{144KI}\right) + \frac{4\tilde{\eta}^2\sigma^2}{KI} \\
&\leq 4\mathcal{E}_r + 8\tilde{\eta}^2[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r,\alpha_r)\|^2 + (\nabla_\alpha f^s(\mathbf{v}_r,\alpha_r))^2] + \frac{5\tilde{\eta}^2\sigma^2}{KI}.
\end{aligned} \tag{88}
$$

□

Lemma 12.

$$
\mathcal{E}_r \leq \frac{\tilde{\eta}\sigma^2}{2\ell K\eta_g^2} + \tilde{\eta}\ell\Xi_{r-1} + \frac{48\tilde{\eta}^2}{\eta_g^2}[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r,\alpha_r)\|^2 + \|\nabla_\alpha f^s(\mathbf{v}_r,\alpha_r)\|^2]. \tag{89}
$$

Proof.

$$
\begin{aligned}
\mathbb{E}\|\mathbf{v}_{r,t}^k - \mathbf{v}_{r-1}\|^2 &= \mathbb{E}\|\mathbf{v}_{r,t-1}^k - \eta_l(\nabla_{\mathbf{v}} f_k(\mathbf{v}_{r,t-1}^k, y_{r,t-1}^k; z_{r,t-1}^k) - c_{\mathbf{v}}^k + c_{\mathbf{v}}) - \mathbf{v}_{r-1}\|^2 \\
&\leq \mathbb{E}\|\mathbf{v}_{r,t-1}^k - \eta_l(\nabla_{\mathbf{v}} f_k(\mathbf{v}_{r,t-1}^k, y_{r,t-1}^k) - \mathbb{E}[c_{\mathbf{v}}^k] + \mathbb{E}[c_{\mathbf{v}}]) - \mathbf{v}_{r-1}\|^2 + 2\eta_l^2\sigma^2 \\
&\leq \left(1 + \frac{1}{I-1}\right)\mathbb{E}\|\mathbf{v}_{r,t-1}^k - \mathbf{v}_{r-1}\|^2 + I\eta_l^2\mathbb{E}\|\nabla_{\mathbf{v}} f_k(\mathbf{v}_{r,t-1}^k, \alpha_{r,t-1}^k) - \mathbb{E}[c_{\mathbf{v}}^k] + \mathbb{E}[c_{\mathbf{v}}]\|^2 + 2\eta_l^2\sigma^2,
\end{aligned} \tag{90}
$$

where $\mathbb{E}[c_{\mathbf{v}}^k] = \frac{1}{I}\sum\limits_{t=1}^{I} f^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k)$ and $\mathbb{E}[c_{\mathbf{v}}] = \frac{1}{K}\sum\limits_{k=1}^{K}\frac{1}{I}\sum\limits_{t=1}^{I} f^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k)$.

Then,

$$
\begin{aligned}
&I\eta_l^2\mathbb{E}\|\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t-1}^k, \alpha_{r,t-1}^k) - \mathbb{E}[c_{\mathbf{v}}^k] + \mathbb{E}[c_{\mathbf{v}}]\|^2 \\
&\leq I\eta_l^2\mathbb{E}\|\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t-1}^k, \alpha_{r,t-1}^k) - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r-1}) + (\mathbb{E}[c_{\mathbf{v}}] - \nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})) \\
&\qquad + \nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}) - (\mathbb{E}[c_{\mathbf{v}}^k] - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r-1}))\|^2 \\
&\leq 4I\eta_l^2\ell^2\left(\mathbb{E}[\|\mathbf{v}_{r,t-1}^k - \mathbf{v}_{r-1}\|^2] + \mathbb{E}[\|\alpha_{r,t-1}^k - \alpha_{r-1}\|^2]\right) + 4I\eta_l^2\mathbb{E}[\|\mathbb{E}[c_{\mathbf{v}}^k] - \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2] \\
&\quad + 4I\eta_l^2\mathbb{E}[\|\mathbb{E}[c_{\mathbf{v}}] - \nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1}\|^2] + 4I\eta_l^2\mathbb{E}[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2 \\
&\leq 4I\eta_l^2\ell^2\left(\mathbb{E}[\|\mathbf{v}_{k-1,r}^k - \mathbf{v}_{r-1}\|^2] + \mathbb{E}[\|\alpha_{k-1,r}^k - \alpha_{r-1}\|^2]\right) + 4I\eta_l^2\ell^2\frac{1}{I}\sum_{\tau=1}^{I}\mathbb{E}[\|\mathbf{v}_{r-1,\tau}^k - \mathbf{v}_{r-1}\|^2 + \|\alpha_{r-1,\tau}^k - \alpha_{r-1}\|^2] \\
&\quad + 4I\eta_l^2\ell^2\frac{1}{KI}\sum_{j=1}^{K}\sum_{t=1}^{I}\mathbb{E}[\|\mathbf{v}_{r-1,t}^j - \mathbf{v}_{r-1}\|^2 + \|\alpha_{r-1,k}^j - \alpha_{r-1}\|^2] + 4I\eta_l^2\mathbb{E}[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2
\end{aligned} \tag{91}
$$

For α, we have similar results, adding them together

$$\begin{aligned}
&\mathbb{E}\|\mathbf{v}^k_{k,r} - \mathbf{v}_{r-1}\|^2 + \mathbb{E}\|\alpha^k_{k,r} - \alpha_{r-1}\|^2 \leq \left(1 + \frac{1}{K-1} + 8K\eta_l^2\ell^2\right)(\mathbb{E}\|\mathbf{v}^k_{k-1,r} - \mathbf{v}_{r-1}\|^2 + \mathbb{E}\|\alpha^k_{k-1,r} - \alpha_{r-1}\|^2) \\
&\quad + 2\eta_l^2\sigma^2 + 4I\eta_l^2\ell^2\Xi_{r-1} + 4I\eta_l^2\frac{1}{I}\sum_{\tau=1}^{I}\mathbb{E}[\|\mathbf{v}^k_{r-1,\tau} - \mathbf{v}_{r-1}\|^2 + \|\alpha^k_{r-1,\tau} - \alpha_{r-1}\|^2] \\
&\quad + 4I\eta_l^2\mathbb{E}[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2 + \|\nabla_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2]
\end{aligned} \tag{92}$$

Taking average over all machines,

$$\begin{aligned}
&\frac{1}{K}\sum_k \mathbb{E}\|\mathbf{v}^k_{r,t} - \mathbf{v}_{r-1}\|^2 + \mathbb{E}(\alpha^k_{r,t} - \alpha_{r-1})^2 \\
&\leq \left(1 + \frac{1}{I-1} + 8I\eta_l^2\ell^2\right)\frac{1}{K}\sum_k(\mathbb{E}\|\mathbf{v}^k_{r,t-1} - \mathbf{v}_{r-1}\|^2 + \mathbb{E}(\alpha^k_{r,t-1} - \alpha_{r-1})^2) + 2\eta_l^2\sigma^2 \\
&\quad + 8I\eta_l^2\ell^2\Xi_{r-1} + 4I\eta_l^2\mathbb{E}[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2 + \|\nabla_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2]] \\
&\leq \left(2\eta_l^2\sigma^2 + 8I\eta_l^2\ell^2\Xi_{r-1} + 4I\eta_l^2\mathbb{E}[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2 + (\nabla_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}))^2\right)\left(\sum_{\tau=0}^{t-1}(1 + \frac{1}{I-1} + 8I\eta_l^2\ell^2)^\tau\right) \\
&\leq \left(\frac{2\tilde{\eta}^2\sigma^2}{I^2\eta_g^2} + \frac{8\tilde{\eta}^2\ell^2}{I\eta_g^2}\Xi_{r-1} + \frac{4\tilde{\eta}^2}{I\eta_g^2}\mathbb{E}[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2 + (\nabla_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}))^2]\right)3I \\
&\leq \left(\frac{\tilde{\eta}\sigma^2}{24\ell I^2\eta_g^2} + \frac{\tilde{\eta}\ell}{3I\eta_g^2}\Xi_{r-1} + \frac{4\tilde{\eta}^2}{I\eta_g^2}\mathbb{E}[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2 + (\nabla_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1}))^2]\right)3I
\end{aligned} \tag{93}$$

Taking average over $t = 1, ..., I$,

$$\mathcal{E}_r \leq \frac{\tilde{\eta}\sigma^2}{8\ell I\eta_g^2} + \tilde{\eta}\ell\Xi_{r-1} + \frac{12\tilde{\eta}^2}{\eta_g^2}\mathbb{E}[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2 + \|\nabla_\alpha f^s(\mathbf{v}_{r-1}, \alpha_{r-1})\|^2] \tag{94}$$

Using (87), we have

$$\mathcal{E}_r \leq \frac{\tilde{\eta}\sigma^2}{8\ell I\eta_g^2} + \tilde{\eta}\ell\Xi_{r-1} + \frac{12\tilde{\eta}^2}{\eta_g^2}\left(4[\|\nabla_{\mathbf{v}} f(\mathbf{v}_r, \alpha_r)\|^2 + \|\nabla_\alpha f^s(\mathbf{v}_r, \alpha_r)\|^2] + \frac{\ell^2}{24}\mathcal{E}_r + \frac{\sigma^2}{144KI}\right). \tag{95}$$

Rearranging terms,

$$\mathcal{E}_r \leq \frac{\tilde{\eta}\sigma^2}{2\ell I\eta_g^2} + \tilde{\eta}\ell\Xi_{r-1} + \frac{48\tilde{\eta}^2}{\eta_g^2}[\|\nabla_x f^s(\mathbf{v}_r, \alpha_r)\|^2 + \|\nabla_\alpha f^s(\mathbf{v}_r, \alpha_r)\|^2] \tag{96}$$

□

D.1. Main Proof of Lemma 2

Proof. Plugging Lemma 10 into Lemma 9, we get

$$
\begin{aligned}
&\frac{1}{R}\sum_{r=1}^{R}[f^s(\mathbf{v}_r,\alpha)-f^s(\mathbf{v},\alpha_r)]\\
&\leq \frac{1}{R}\sum_{r=1}^{R}\Bigg[\underbrace{\left(\frac{3\ell+3\ell^2/\mu_2}{2}-\frac{1}{2\tilde\eta}\right)\|\mathbf{v}_{r-1}-\mathbf{v}_r\|^2+\left(2\ell-\frac{1}{2\tilde\eta}\right)\|\alpha_r-\alpha_{r-1}\|^2}_{C_1}\\
&+\underbrace{\left(\frac{1}{2\tilde\eta}-\frac{\mu_2}{3}\right)\|\alpha_{r-1}-\alpha\|^2-\left(\frac{1}{2\tilde\eta}-\frac{\mu_2}{3}\right)(\alpha_r-\alpha)^2}_{C_2}+\underbrace{\left(\frac{1}{2\tilde\eta}-\frac{\ell}{3}\right)\|\mathbf{v}_{r-1}-\mathbf{v}\|^2-\left(\frac{1}{2\tilde\eta}-\frac{\ell}{3}\right)\|\mathbf{v}_r-\mathbf{v}\|^2}_{C_3}\\
&+\underbrace{\frac{1}{2\tilde\eta}((\alpha-\tilde\alpha_{r-1})^2-(\alpha-\tilde\alpha_r)^2)}_{C_4}+\underbrace{\left(\frac{3\ell}{2}+\frac{3\ell^2}{2\mu_2}\right)\mathcal{E}_r}_{C_5}\\
&+\underbrace{\frac{3\tilde\eta}{2}\left\|\frac{1}{KI}\sum_{i,t}[\nabla_{\mathbf{v}}f_k^s(\mathbf{v}_{r,t}^k,\alpha_{r,t}^k)-\nabla_{\mathbf{v}}F_k^s(\mathbf{v}_{r,t}^k,\alpha_{r,t}^k;z_{r,t}^k)]\right\|^2}_{C_6}+\underbrace{\frac{3\tilde\eta}{2}\left(\frac{1}{KI}\sum_{i,t}\nabla_\alpha f_k^s(\mathbf{v}_{r,t}^k,\alpha_{r,t}^k)-\nabla_\alpha F_k^s(\mathbf{v}_{r,t}^k,\alpha_{r,t}^k;z_{r,t}^k)\right)^2}_{C_7}
\end{aligned}
\tag{97}
$$

Since $\tilde\eta\leq\min(\frac{1}{3\ell+3\ell^2/\mu_2},\frac{1}{4\ell},\frac{3}{2\mu_2})$, thus in the RHS of (97), 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 12.

Taking expectation over C_6,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{3\tilde\eta}{2}\left\|\frac{1}{KI}\sum_{i,t}[\nabla_{\mathbf{v}}f_k^s(\mathbf{v}_{r,t}^k,\alpha_{r,t}^k)-\nabla_{\mathbf{v}}F_k^s(\mathbf{v}_{r,t}^k,\alpha_{r,t}^k;z_{r,t}^k)]\right\|^2\right]\\
&=\mathbb{E}\left[\frac{3\tilde\eta}{2K^2I^2}\sum_{i,t}\|\nabla_{\mathbf{v}}f_k^s(\mathbf{v}_{r,t}^k,\alpha_{r,t}^k)-\nabla_{\mathbf{v}}F_k^s(\mathbf{v}_{r,t}^k,\alpha_{r,t}^k;z_{r,t}^k)\|^2\right]\\
&\leq\frac{3\tilde\eta\sigma^2}{2KI}.
\end{aligned}
\tag{98}
$$

The equality is due to $\mathbb{E}_{r,t}\left\langle\nabla_{\mathbf{v}}f_k^s(\mathbf{v}_{r,t}^k,\alpha_{r,t}^k)-\nabla_{\mathbf{v}}F_k^s(\mathbf{v}_{r,t}^i,\alpha_{r,t}^i;z_{r,t}^k),\nabla_{\mathbf{v}}f_j^s(\mathbf{v}_{r,t}^j,\alpha_{r,t}^j)-\nabla_{\mathbf{v}}F_j^s(\mathbf{v}_{r,t}^j,\alpha_{r,t}^j;z_{r,t}^j)\right\rangle=0$ for any $i\neq j$ as each machine draws data independently, where $\mathbb{E}_{r,t}$ denotes an expectation in round r conditioned on events until k. The last inequality holds because $\|\nabla_{\mathbf{v}}f_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k)-\nabla_{\mathbf{v}}F_k(\mathbf{v}_{t-1}^k,\alpha_{t-1}^k;z_{t-1}^k)\|^2\leq\sigma^2$ for any i. Similarly, we take expectation over C_7 and have

$$
\mathbb{E}\left[\frac{3\tilde\eta}{2}\left(\frac{1}{NK}\sum_{i,t}^{N}[\nabla_\alpha f_k(\mathbf{v}_{r,t}^k,\alpha_{r,t}^k)-\nabla_\alpha F_k(\mathbf{v}_{r,t}^k,\alpha_{r,t}^k;\mathbf{z}_{r,t}^k)]\right)^2\right]\leq\frac{3\tilde\eta\sigma^2}{2KI}.
\tag{99}
$$

Plugging (98) and (99) into (97), and taking expectation, it yields

$$\begin{aligned}
&\frac{1}{R}\sum_r \mathbb{E}[f^s(\mathbf{v}_r,\alpha)-f^s(\mathbf{v},\alpha_r)]\\
&\leq \mathbb{E}\bigg\{\frac{1}{R}\left(\frac{1}{2\tilde{\eta}}-\frac{\ell_2}{3}\right)\|\mathbf{v}_0-\mathbf{v}\|^2+\frac{1}{R}\left(\frac{1}{2\tilde{\eta}}-\frac{\mu_2}{3}\right)\|\alpha_0-\alpha\|^2+\frac{1}{2\tilde{\eta}R}\|\mathbf{v}_0-\mathbf{v}\|^2+\frac{1}{2\tilde{\eta}R}\|\alpha_0-\alpha\|^2\\
&\quad+\frac{1}{R}\sum_{r=1}^{R}\left(\frac{3\ell^2}{2\mu_2}+\frac{3\ell}{2}\right)\mathcal{E}_r+\frac{3\tilde{\eta}\sigma^2}{KI}\bigg\}\\
&\leq \frac{1}{\tilde{\eta}R}\|\mathbf{v}_0-\mathbf{v}\|^2+\frac{1}{\tilde{\eta}R}\|\alpha_0-\alpha\|^2+\frac{3\ell^2}{\mu_2}\frac{1}{R}\sum_{r=1}^{R}\mathcal{E}_r+\frac{3\tilde{\eta}\sigma^2}{KI},
\end{aligned}$$

where we use $\mathbf{v}_0=\bar{\mathbf{v}}_0$, and $\alpha_0=\bar{\alpha}_0$ in the last inequality.

Using Lemma 12,

$$\begin{aligned}
&\frac{1}{R}\sum_r \mathbb{E}[f^s(\mathbf{v}_r,\alpha)-f^s(\mathbf{v},\alpha_r)]\\
&\leq \frac{1}{\tilde{\eta}R}\|\mathbf{v}_0-\mathbf{v}\|^2+\frac{1}{\tilde{\eta}R}\|\alpha_0-\alpha\|^2+\frac{3\ell^2}{\mu_2}\frac{1}{R}\sum_{r=1}^{R}\mathcal{E}_r+\frac{3\tilde{\eta}\sigma^2}{KI}\\
&\leq \frac{1}{\tilde{\eta}R}\|\mathbf{v}_0-\mathbf{v}\|^2+\frac{1}{\tilde{\eta}R}\|\alpha_0-\alpha\|^2\\
&\quad+\frac{3\ell^2}{\mu_2}\frac{1}{R}\sum_{r=1}^{R}\left[\left(\frac{\tilde{\eta}\sigma^2}{2\ell I\eta_g^2}+\tilde{\eta}\ell\Xi_{r-1}+\frac{48\tilde{\eta}^2}{\eta_g^2}\mathbb{E}[\|\nabla_{\mathbf{v}}f^s(\mathbf{v}_r,\alpha_r)\|^2+\|\nabla_\alpha f^s(\mathbf{v}_r,\alpha_r)\|^2]\right)\right]+\frac{3\tilde{\eta}\sigma^2}{KI}\\
&\leq \frac{1}{\tilde{\eta}R}\|\mathbf{v}_0-\mathbf{v}\|^2+\frac{1}{\tilde{\eta}R}\|\alpha_0-\alpha\|^2+\frac{3\tilde{\eta}\ell^3}{\mu_2R\eta_g^2}\sum_r\Xi_{r-1}+\frac{5\ell}{\mu_2I\eta_g^2}\tilde{\eta}\sigma^2+\frac{3000\tilde{\eta}^2\ell^4}{\mu_2^2\eta_g^2}\frac{1}{R}\sum_{r=1}^{R}Gap_r
\end{aligned}$$

where the last inequality holds because

$$\|\nabla_{\mathbf{v}}f^s(\mathbf{v}_r,\alpha_r)\|^2+\|\nabla_\alpha f^s(\mathbf{v}_r,\alpha_r)\|^2\leq 9\ell^2(\|\mathbf{v}_r-\mathbf{v}^*_{\phi_s}\|^2+\|\alpha_r-\alpha^*_{\phi_s}\|^2)\leq\frac{18\ell^2}{\mu_2}Gap_s(\mathbf{v}_r,\alpha_r). \tag{100}$$

Using Lemma 11,

$$\begin{aligned}
\Xi_r&\leq 4\mathcal{E}_r+16\tilde{\eta}^2[\|\nabla_{\mathbf{v}}f^s(\mathbf{v}_r,\alpha_r)\|^2+\|\nabla_\alpha f^s(\mathbf{v}_r,\alpha_r)\|^2]+\frac{5\tilde{\eta}^2\sigma^2}{KI}\\
&\leq 4\left(\frac{\tilde{\eta}\sigma^2}{2\ell K\eta_g^2}+\tilde{\eta}\ell\Xi_{r-1}+\frac{48\tilde{\eta}^2}{\eta_g^2}[\|\nabla_{\mathbf{v}}f^s(\mathbf{v}_r,\alpha_r)\|^2+\|\nabla_\alpha f^s(\mathbf{v}_r,\alpha_r)\|^2]\right)\\
&\quad+16\tilde{\eta}^2[\|\nabla_{\mathbf{v}}f^s(\mathbf{v}_r,\alpha_r)\|^2+\|\nabla_\alpha f^s(\mathbf{v}_r,\alpha_r)\|^2]+\frac{5\tilde{\eta}\sigma^2}{KI}\\
&\leq 4\tilde{\eta}\ell\Xi_{r-1}+160\tilde{\eta}^2[\|\nabla_{\mathbf{v}}f^s(\mathbf{v}_r,\alpha_r)\|^2+\|\nabla_\alpha f^s(\mathbf{v}_r,\alpha_r)\|^2]+\frac{5\tilde{\eta}\sigma^2}{KI}(1+\frac{K}{\eta_g^2})\\
&\leq \Xi_{r-1}+160\tilde{\eta}^2[\|\nabla_{\mathbf{v}}f^s(\mathbf{v}_r,\alpha_r)\|^2+\|\nabla_\alpha f^s(\mathbf{v}_r,\alpha_r)\|^2]+\frac{5\tilde{\eta}\sigma^2}{KI}(1+\frac{K}{\eta_g^2}).
\end{aligned} \tag{101}$$

Thus,

$$\begin{aligned}
\frac{2\tilde{\eta}\ell^3}{\mu_2R\eta_g^2}\sum_{r=1}^{R}\Xi_r&\leq\frac{2\tilde{\eta}\ell^3}{\mu_2R\eta_g^2}\sum_r\Xi_{r-1}+\frac{320\tilde{\eta}^3\ell^3}{\mu_2R\eta_g^2}\sum_{r=1}^{R}[\|\nabla_{\mathbf{v}}f^s(\mathbf{v}_r,\alpha_r)\|^2+\|\nabla_\alpha f^s(\mathbf{v}_r,\alpha_r)\|^2]+\frac{5\tilde{\eta}\sigma^2}{KI}(1+\frac{K}{\eta_g^2})\\
&\leq\frac{2\tilde{\eta}\ell^3}{\mu_2R\eta_g^2}\sum_r\Xi_{r-1}+\frac{1}{2R}\sum_r Gap_r+\frac{5\tilde{\eta}\sigma^2}{KI}(1+\frac{K}{\eta_g^2})
\end{aligned} \tag{102}$$

Taking $A_0 = 0$,

$$\begin{aligned}
&\frac{1}{R}\sum_r \mathbb{E}[f^s(\mathbf{v}_r,\alpha) - f^s(\mathbf{v},\alpha_r)] \\
&\leq \frac{1}{\tilde{\eta}R}\|\mathbf{v}_0 - \mathbf{v}\|^2 + \frac{1}{\tilde{\eta}R}\|\alpha_0 - \alpha\|^2 + \frac{1}{2R}\sum_r Gap_r + \frac{5\tilde{\eta}\sigma^2}{NK}(1 + \frac{N}{\eta_g^2})
\end{aligned}$$

It follows that

$$\begin{aligned}
&\frac{1}{R}\sum_r \mathbb{E}[f^s(\mathbf{v}_r,\alpha) - f^s(\mathbf{v},\alpha_r)] - \frac{1}{2R}\sum_r Gap_r \\
&\leq \frac{1}{\tilde{\eta}R}\|\mathbf{v}_0 - \mathbf{v}\|^2 + \frac{1}{\tilde{\eta}R}\|\alpha_0 - \alpha\|^2 + \frac{5\tilde{\eta}\sigma^2}{KI}(1 + \frac{K}{\eta_g^2}).
\end{aligned}$$

Sample a $\tilde{r}$ from $1, ..., R$, we have

$$\mathbb{E}[Gap_{\tilde{r}}^s] \leq \frac{2}{\tilde{\eta}R}\|\mathbf{v}_0 - \mathbf{v}\|^2 + \frac{2}{\tilde{\eta}R}\|\alpha_0 - \alpha\|^2 + \frac{10\tilde{\eta}\sigma^2}{KI}\left(1 + \frac{K}{\eta_g^2}\right). \tag{103}$$

□

E. Proof of Theorem 2

Proof. Since $f(\mathbf{v},\alpha)$ is ℓ-weakly convex in $\mathbf{v}$ for any α, $\phi(\mathbf{v}) = \max_{\alpha'} f(\mathbf{v},\alpha')$ is also ℓ-weakly convex. Taking $\gamma = 2\ell$, we have

$$\begin{aligned}
\phi(\mathbf{v}_{s-1}) &\geq \phi(\mathbf{v}_s) + \langle \partial\phi(\mathbf{v}_s), \mathbf{v}_{s-1} - \mathbf{v}_s\rangle - \frac{\ell}{2}\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2 \\
&= \phi(\mathbf{v}_s) + \langle \partial\phi(\mathbf{v}_s) + 2\ell(\mathbf{v}_s - \mathbf{v}_{s-1}), \mathbf{v}_{s-1} - \mathbf{v}_s\rangle + \frac{3\ell}{2}\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2 \\
&\overset{(a)}{=} \phi(\mathbf{v}_s) + \langle \partial\phi_s(\mathbf{v}_s), \mathbf{v}_{s-1} - \mathbf{v}_s\rangle + \frac{3\ell}{2}\|\mathbf{v}_{s-1} - \mathbf{v}_s\|^2 \\
&\overset{(b)}{=} \phi(\mathbf{v}_s) - \frac{1}{2\ell}\langle \partial\phi_s(\mathbf{v}_s), \partial\phi_s(\mathbf{v}_s) - \partial\phi(\mathbf{v}_s)\rangle + \frac{3}{8\ell}\|\partial\phi_s(\mathbf{v}_s) - \partial\phi(\mathbf{v}_s)\|^2 \\
&= \phi(\mathbf{v}_s) - \frac{1}{8\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 - \frac{1}{4\ell}\langle \partial\phi_s(\mathbf{v}_s), \partial\phi(\mathbf{v}_s)\rangle + \frac{3}{8\ell}\|\partial\phi(\mathbf{v}_s)\|^2,
\end{aligned} \tag{104}$$

where (a) and (b) hold by the definition of $\phi_s(\mathbf{v})$.

Rearranging the terms in (104) yields

$$\begin{aligned}
\phi(\mathbf{v}_s) - \phi(\mathbf{v}_{s-1}) &\leq \frac{1}{8\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 + \frac{1}{4\ell}\langle \partial\phi_s(\mathbf{v}_s), \partial\phi(\mathbf{v}_s)\rangle - \frac{3}{8\ell}\|\partial\phi(\mathbf{v}_s)\|^2 \\
&\overset{(a)}{\leq} \frac{1}{8\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 + \frac{1}{8\ell}(\|\partial\phi_s(\mathbf{v}_s)\|^2 + \|\partial\phi(\mathbf{v}_s)\|^2) - \frac{3}{8\ell}\|\phi(\mathbf{v}_s)\|^2 \\
&= \frac{1}{4\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 - \frac{1}{4\ell}\|\partial\phi(\mathbf{v}_s)\|^2 \\
&\overset{(b)}{\leq} \frac{1}{4\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 - \frac{\mu}{2\ell}(\phi(\mathbf{v}_s) - \phi(\mathbf{v}_{\phi_s}^*))
\end{aligned} \tag{105}$$

where (a) holds by using $\langle \mathbf{a}, \mathbf{b}\rangle \leq \frac{1}{2}(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$, and (b) holds by the μ-PL property of $\phi(\mathbf{v})$.

Thus, we have

$$(4\ell + 2\mu)\left(\phi(\mathbf{v}_s) - \phi(\mathbf{v}_*)\right) - 4\ell(\phi(\mathbf{v}_{s-1}) - \phi(\mathbf{v}_{\phi_s}^*)) \leq \|\partial\phi_s(\mathbf{v}_s)\|^2. \tag{106}$$

Since $\gamma = 2\ell$, $f_s(\mathbf{v}, \alpha)$ is ℓ-strongly convex in $\mathbf{v}$ and μ_2 strong concave in α. Apply Lemma 3 to f_s, we know that

$$\frac{\ell}{4}\|\hat{\mathbf{v}}_s(\alpha_s) - \mathbf{v}_0^s\|^2 + \frac{\mu_2}{4}\|\hat{\alpha}_s(\mathbf{v}_s) - \alpha_0^s\|^2 \leq \text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) + \text{Gap}_s(\mathbf{v}_s, \alpha_s). \tag{107}$$

By the setting of $\tilde{\eta}_s$, $I_s = I_0 * 2^s$, and $R_s = \frac{1000}{\tilde{\eta}\min(\ell,\mu_2)}$, we note that $\frac{4}{\tilde{\eta}R_s} \leq \frac{\min\{\ell,\mu_2\}}{212}$. Applying Lemma (2), we have

$$\begin{aligned}
&\mathbb{E}[\text{Gap}_s(\mathbf{v}_s, \alpha_s)] \leq \frac{10\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{1}{53}\mathbb{E}\left[\frac{\ell}{4}\|\hat{\mathbf{v}}_s(\alpha_s) - \mathbf{v}_0^s\|^2 + \frac{\mu_2}{4}\|\hat{\alpha}_s(\mathbf{v}_s) - \alpha_0^s\|^2\right] \\
&\leq \frac{10\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{1}{53}\mathbb{E}\left[\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) + \text{Gap}_s(\mathbf{v}_s, \alpha_s)\right].
\end{aligned} \tag{108}$$

Since $\phi(\mathbf{v})$ is L-smooth and $\gamma = 2\ell$, then $\phi_k(\mathbf{v})$ is $\hat{L} = (L + 2\ell)$-smooth. According to Theorem 2.1.5 of (Nesterov, 2004), we have

$$\begin{aligned}
&\mathbb{E}[\|\partial\phi_s(\mathbf{v}_s)\|^2] \leq 2\hat{L}\mathbb{E}(\phi_s(\mathbf{v}_s) - \min_{x\in\mathbb{R}^d}\phi_s(\mathbf{v})) \leq 2\hat{L}\mathbb{E}[\text{Gap}_s(\mathbf{v}_s, \alpha_s)] \\
&= 2\hat{L}\mathbb{E}[4\text{Gap}_s(\mathbf{v}_s, \alpha_s) - 3\text{Gap}_s(\mathbf{v}_s, \alpha_s)] \\
&\leq 2\hat{L}\mathbb{E}\left[4\left(\frac{10\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{1}{53}\left(\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) + \text{Gap}_s(\mathbf{v}_s, \alpha_s)\right)\right) - 3\text{Gap}_s(\mathbf{v}_s, \alpha_s)\right] \\
&= 2\hat{L}\mathbb{E}\left[40\frac{\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{4}{53}\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) - \frac{155}{53}\text{Gap}_s(\mathbf{v}_s, \alpha_s)\right].
\end{aligned} \tag{109}$$

Applying Lemma 4 to (109), we have

$$\begin{aligned}
\mathbb{E}[\|\partial\phi_s(\mathbf{v}_s)\|^2] \leq 2\hat{L}\mathbb{E}\Bigg[&\frac{40\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{4}{53}\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) \\
&- \frac{155}{53}\left(\frac{3}{50}\text{Gap}_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1}) + \frac{4}{5}(\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_0^s))\right)\Bigg] \\
= 2\hat{L}\mathbb{E}\Bigg[40\frac{\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{4}{53}&\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) - \frac{93}{530}\text{Gap}_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1}) - \frac{124}{53}(\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_0^s))\Bigg].
\end{aligned} \tag{110}$$

Combining this with (106), rearranging the terms, and defining a constant $c = 4\ell + \frac{248}{53}\hat{L} \in O(L + \ell)$, we get

$$\begin{aligned}
&(c + 2\mu)\,\mathbb{E}[\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_*)] + \frac{93}{265}\hat{L}\mathbb{E}[\text{Gap}_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1})] \\
&\leq \left(4\ell + \frac{248}{53}\hat{L}\right)\mathbb{E}[\phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_\phi^*)] + \frac{8\hat{L}}{53}\mathbb{E}[\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s)] + \frac{80\hat{L}\tilde{\eta}\sigma^2}{KI_0 2^s} \\
&\leq c\mathbb{E}\left[\phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_*) + \frac{8\hat{L}}{53c}\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s)\right] + \frac{80\hat{L}\tilde{\eta}\sigma^2}{KI_0 2^s}.
\end{aligned} \tag{111}$$

Using the fact that $\hat{L} \geq \mu$,

$$(c + 2\mu)\frac{8\hat{L}}{53c} = \left(4\ell + \frac{248}{53}\hat{L} + 2\mu\right)\frac{8\hat{L}}{53(4\ell + \frac{248}{53}\hat{L})} \leq \frac{8\hat{L}}{53} + \frac{16\mu_1\hat{L}}{248\hat{L}} \leq \frac{93}{265}\hat{L}. \tag{112}$$

Then, we have

$$\begin{aligned}
&(c + 2\mu_1)\mathbb{E}\left[\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_*) + \frac{8\hat{L}}{53c}\text{Gap}_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1})\right] \\
&\leq c\mathbb{E}\left[\phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_*) + \frac{8\hat{L}}{53c}\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s)\right] + \frac{80\hat{L}\tilde{\eta}\sigma^2}{KI_0 2^s}.
\end{aligned} \tag{113}$$

Defining $\Delta_s = \phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_*) + \frac{8\hat{L}}{53c}\text{Gap}_s(\mathbf{v}_0^s, \alpha_0^s)$, then

$$\mathbb{E}[\Delta_{s+1}] \leq \frac{c}{c+2\mu}\mathbb{E}[\Delta_s] + \frac{80\hat{L}}{c+2\mu}\frac{\tilde{\eta}\sigma^2}{KI_0 2^s} \tag{114}$$

Using this inequality recursively, it yields

$$\begin{aligned} E[\Delta_{S+1}] &\leq \left(\frac{c}{c+2\mu}\right)^S E[\Delta_1] + \frac{80\hat{L}}{c+2\mu}\frac{\tilde{\eta}\sigma^2}{KI_0}\sum_{s=1}^{S}\left(\exp\left(-\frac{2\mu}{c+2\mu}(s-1)\right)\left(\frac{c}{c+2\mu}\right)^{S+1-s}\right) \\ &\leq 2\epsilon_0 \exp\left(\frac{-2\mu S}{c+2\mu}\right) + \frac{80\tilde{\eta}\hat{L}\sigma^2}{(c+2\mu)KI_0} S\exp\left(-\frac{2\mu S}{c+2\mu}\right), \end{aligned} \tag{115}$$

where the second inequality uses the fact $1 - x \leq \exp(-x)$, and

$$\begin{aligned} \Delta_1 &= \phi(\mathbf{v}_0^1) - \phi(\mathbf{v}^*) + \frac{8\hat{L}}{53c}Gap_1(\mathbf{v}_0^1, \alpha_0^1) \\ &= \phi(\mathbf{v}_0) - \phi(\mathbf{v}^*) + \left(f(\mathbf{v}_0, \hat{\alpha}_1(\mathbf{v}_0)) + \frac{\gamma}{2}\|\mathbf{v}_0 - \mathbf{v}_0\|^2 - f(\hat{\mathbf{v}}_1(\alpha_0), \alpha_0) - \frac{\gamma}{2}\|\hat{\mathbf{v}}_1(\alpha_0) - \mathbf{v}_0\|^2\right) \\ &\leq \epsilon_0 + f(\mathbf{v}_0, \hat{\alpha}_1(\mathbf{v}_0)) - f(\hat{\mathbf{v}}(\alpha_0), \alpha_0) \leq 2\epsilon_0. \end{aligned} \tag{116}$$

To make this less than ϵ, it suffices to make

$$\begin{aligned} 2\epsilon_0 \exp\left(\frac{-2\mu S}{c+2\mu}\right) &\leq \frac{\epsilon}{2} \\ \frac{80\tilde{\eta}\hat{L}\sigma^2}{(c+2\mu)KI_0} S\exp\left(-\frac{2\mu S}{c+2\mu}\right) &\leq \frac{\epsilon}{2} \end{aligned} \tag{117}$$

Let S be the smallest value such that $\exp\left(\frac{-2\mu S}{c+2\mu}\right) \leq \min\{\frac{\epsilon}{4\epsilon_0}, \frac{(c+2\mu)\epsilon}{160\hat{L}S}\frac{KI_0}{\tilde{\eta}\sigma^2}\}$. We can set S to be the smallest value such that $S > \max\left\{\frac{c+2\mu}{2\mu}\log\frac{4\epsilon_0}{\epsilon}, \frac{c+2\mu}{2\mu}\log\frac{160\hat{L}S}{(c+2\mu)\epsilon}\frac{\tilde{\eta}\sigma^2}{KI_0}\right\}$.

Then, the total communication complexity is

$$\sum_{s=1}^{S} R_s \leq O\left(\frac{1000}{\tilde{\eta}\mu_2}S\right) \leq \widetilde{O}\left(\frac{1}{\tilde{\eta}\mu_2}\frac{c}{\mu}\right) \leq \widetilde{O}\left(\frac{1}{\mu}\right).$$

Total iteration complexity is

$$\begin{aligned} &\sum_{s=1}^{S} T_s = \sum_{s=1}^{S} R_s I_s \\ &= \sum_{s=1}^{S} R_s I_0 \exp(\frac{2\mu}{c+2\mu}(s-1)) = O\left(I_0 \sum_s \exp(\frac{2\mu}{c+2\mu}(s-1))\right) \\ &= \widetilde{O}\left(I_0 \frac{\exp(\frac{2\mu}{c+2\mu}S)}{\exp(\frac{2\mu_1}{c+2\mu})}\right) \\ &= \widetilde{O}\left(\frac{c}{\mu_2^2\mu}\left(\frac{\epsilon_0}{\epsilon}, \frac{S\tilde{\eta}\sigma^2}{I_0 K\epsilon}\right)\right) \\ &= \widetilde{O}\left(\max(\frac{1}{\mu\epsilon}, \frac{c^2}{\mu^2}\frac{\tilde{\eta}\sigma^2}{K})\right) = \widetilde{O}\left(\max(\frac{1}{\mu\epsilon}, \frac{1}{K\mu^2\epsilon})\right), \end{aligned} \tag{118}$$

which is also the sample complexity on each single machine.

□

F. More Results

In this section, we report more experiment results for imratio=30% with DenseNet121 on ImageNet-IH, and CIFAR100-IH in Figure 2,3 and 4. We also verify the proposed CODASCA using stagewise $I = I_0 \times 3^{(s-1)}$, where s is the stage number, indicating that all machines will communicate less frequently at later stages during training. The results for imratio=10% and K=16 with DenseNet121 on ImageNet-IH, and CIFAR100-IH are included in Figure 5. In addition, we conduct experiments on imbalanced heterogeneous CIFAR100 from the same sample set for K=16 and K=8 and the results are included in Figure 6 and Figure 7.

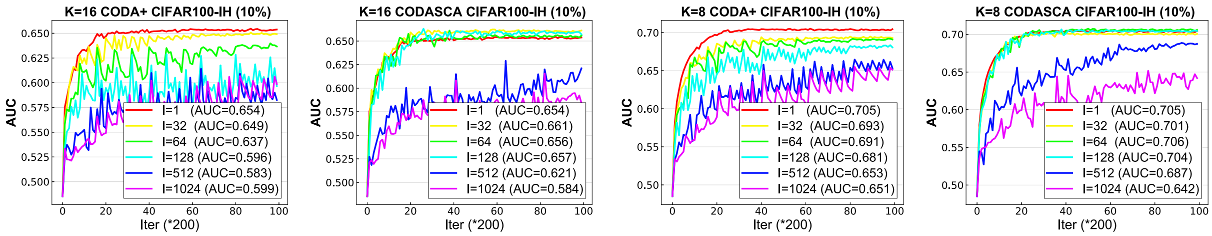

Figure 2. Imbalanced Heterogeneous CIFAR100 with imratio = 10% and K=16,8 on Densenet121.

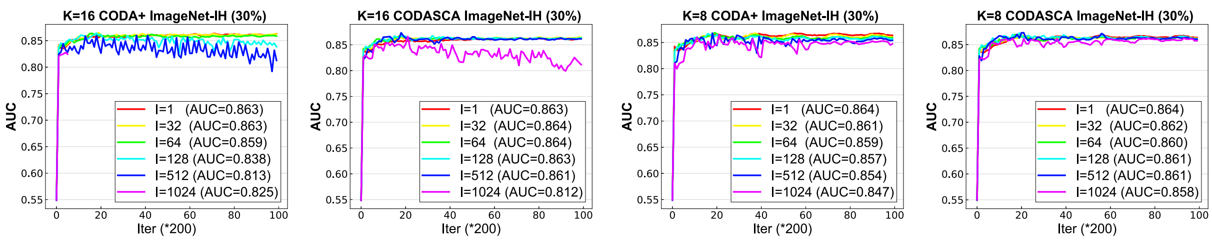

Figure 3. Imbalanced Heterogeneous ImageNet with imratio = 30% and K=16,8 on Densenet121.

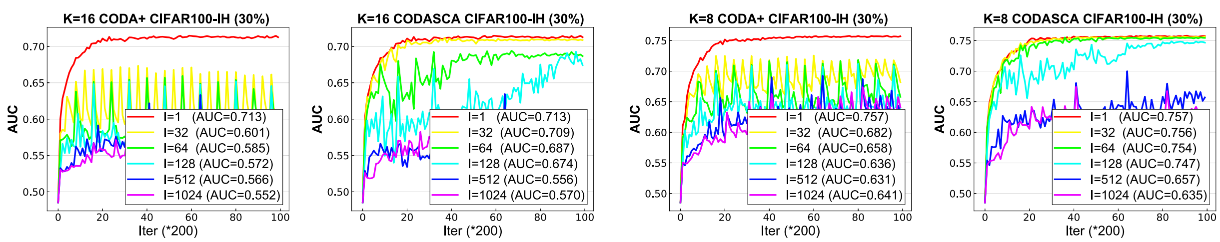

Figure 4. Imbalanced Heterogeneous CIFAR100 with imratio = 30% and K=16,8 on Densenet121.

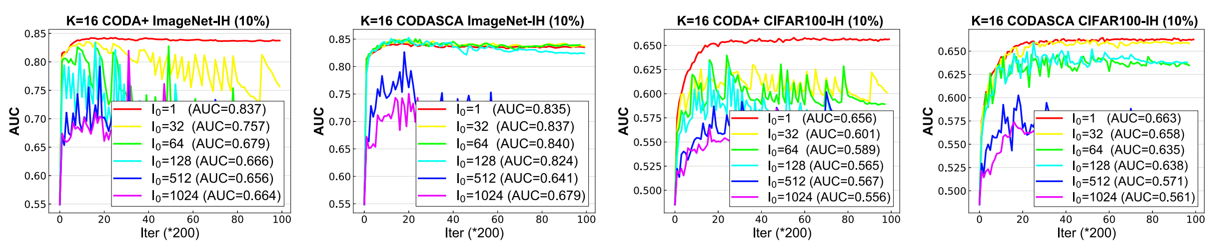

Figure 5. Imbalanced Heterogeneous ImageNet, CIFAR100 with imratio = 10%, K=16 and increasing I on Densenet121.

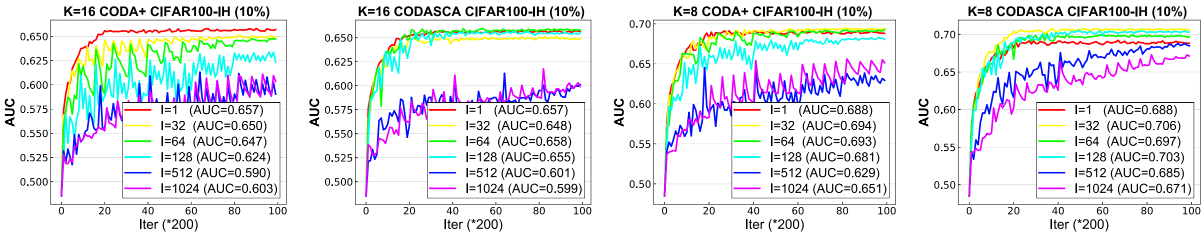

Figure 6. Imbalanced Heterogeneous CIFAR100 with imratio = 10%, K=16,8 on Densenet121.

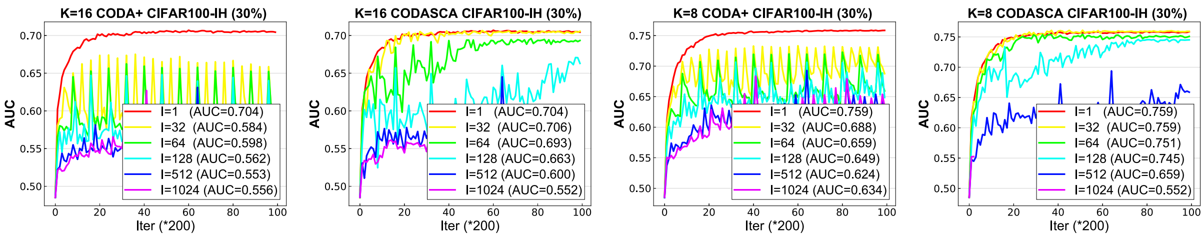

Figure 7. Imbalanced Heterogeneous CIFAR100 with imratio = 30%, K=16,8 on Densenet121.

G. Descriptions of Datasets

Table 6. Statistics of Medical Chest X-ray Datasets. The numbers for each disease indicate the imbalance ratio (imratio).

Dataset	Source	Samples	Cardiomegaly	Edema	Consolidation	Atelectasis	Effusion
CheXpert	Stanford Hospital (US)	224,316	0.211	0.342	0.120	0.310	0.414
ChestXray8	NIH Clinical Center (US)	112,120	0.025	0.021	0.042	0.103	0.119
PadChest	Hospital San Juan (Spain)	110,641	0.089	0.012	0.015	0.056	0.064
MIMIC-CXR	BIDMC (US)	377,110	0.196	0.179	0.047	0.246	0.237
ChestXrayAD	H108 and HMUH (Vietnam)	15,000	0.153	0.000	0.024	0.012	0.069