Federated Deep AUC Maximization for Heterogeneous Data with a Constant Communication Complexity

Zhuoning Yuan* ZHUONING-YUAN@UIOWA.EDU
The University of Iowa
Zhishuai Guo* ZHISHUAI-GUO@UIOWA.EDU
The University of Iowa
Yi Xu YIXU@ALIBABA-INC.COM
DAMO Academy, Alibaba Group
Yiming Ying YYING@ALBANY.EDU
State University of New York at Albany
Tianbao Yang TIANBAO-YANG@UIOWA.EDU
The University of Iowa

Abstract

Deep AUC (area under the ROC curve) Maximization (DAM) has attracted much attention recently due to its great potential for imbalanced data classification. However, the research on Federated Deep AUC Maximization (FDAM) is still limited. Compared with standard federated learning (FL) approaches that focus on decomposable minimization objectives, FDAM is more complicated due to its minimization objective is non-decomposable over individual examples. In this paper, we propose improved FDAM algorithms for heterogeneous data by solving the popular non-convex strongly-concave min-max formulation of DAM in a distributed fashion, which can also be applied to a class of non-convex strongly-concave min-max problems. A striking result of this paper is that the communication complexity of the proposed algorithm is a constant independent of the number of machines and also independent of the accuracy level, which improves an existing result by orders of magnitude. The experiments have demonstrated the effectiveness of our FDAM algorithm on benchmark datasets, and on medical chest X-ray images from different organizations. Our experiment shows that the performance of FDAM using data from multiple hospitals can improve the AUC score on testing data from a single hospital for detecting life-threatening diseases based on chest radiographs. The proposed method is implemented in our open-sourced library LibAUC (`www.libauc.org`) whose github address is `https://github.com/Optimization-AI/ICML2021_FedDeepAUC_CODASCA`.

1. Introduction

Federated learning (FL) is an emerging paradigm for large-scale learning to deal with data that are (geographically) distributed over multiple clients, e.g., mobile phones, organizations. An important feature of FL is that the data remains at its own clients, allowing

*. Equal Contribution

Table 1: The summary of sample and communication complexities of different algorithms for FDAM under a μ-PL condition in both heterogeneous and homogeneous settings, where K is the number of machines and $\mu \leq 1$. NPA denotes the naive parallel (large mini-batch) version of PPD-SG [22] for DAM, where M denotes the batch size in the NPA. The $*$ indicate the results that are derived by us. $\widetilde{O}(\cdot)$ suppresses a logarithmic factor.

	Heterogeneous Data	Homogeneous Data	Sample Complexity
NPA ($M < \frac{1}{K\mu\epsilon}$)	$\widetilde{O}\left(\frac{1}{KM\mu^2\epsilon} + \frac{1}{\mu\epsilon}\right)$	$\widetilde{O}\left(\frac{1}{KM\mu^2\epsilon} + \frac{1}{\mu\epsilon}\right)$	$\widetilde{O}\left(\frac{M}{\mu\epsilon} + \frac{1}{\mu^2K\epsilon}\right)$
NPA ($M \geq \frac{1}{K\mu\epsilon}$)	$\widetilde{O}\left(\frac{1}{\mu}\right)^*$	$\widetilde{O}\left(\frac{1}{\mu}\right)^*$	$\widetilde{O}\left(\frac{M}{\mu}\right)^*$
CODA+ (CODA)	$\widetilde{O}\left(\frac{K}{\mu} + \frac{1}{\mu\epsilon^{1/2}} + \frac{1}{\mu^{3/2}\epsilon^{1/2}}\right)$	$\widetilde{O}\left(\frac{K}{\mu}\right)^*$	$\widetilde{O}\left(\frac{1}{\mu\epsilon} + \frac{1}{\mu^2K\epsilon}\right)$
CODASCA	$\widetilde{O}\left(\frac{1}{\mu}\right)$	$\widetilde{O}\left(\frac{1}{\mu}\right)$	$\widetilde{O}\left(\frac{1}{\mu\epsilon} + \frac{1}{\mu^2K\epsilon}\right)$

the preservation of data privacy. This feature makes FL attractive not only to internet companies such as Google and Apple but also to conventional industries such as those that provide services to hospitals and banks in the big data era [36, 23]. Data in these industries is usually collected from people who are concerned about data leakage. But in order to provide better services, large-scale machine learning from diverse data sources is important for addressing model bias. For example, most patients in hospitals located in urban areas could have dramatic differences in demographic data, lifestyles, and diseases from patients who are from rural areas. Machine learning models (in particular, deep neural networks) trained based on patients' data from one hospital could dramatically bias towards its major population, which could bring serious ethical concerns [33].

One of the fundamental issues that could cause model bias is data imbalance, where the number of samples from different classes are skewed. Although FL provides an effective framework for leveraging multiple data sources, *most existing FL methods still lack the capability to tackle the model bias caused by data imbalance*. The reason is that most existing FL methods are developed for minimizing the conventional objective function, e.g., the average of a standard loss function on all data, which are not amenable to optimizing more suitable measures such as area under the ROC curve (AUC) for imbalanced data. It has been recently shown that directly maximizing AUC for deep learning can lead to great improvements on real-world difficult classification tasks [49]. For example, Yuan et al. [49] reported the best performance by DAM on the Stanford CheXpert Competition for interpreting chest X-ray images like radiologists [11].

However, the research on FDAM is still limited. To the best of our knowledge, Guo et al. [8] is the only work that was dedicated to FDAM by solving **the non-convex strongly-concave min-max** problem in a distributed manner. Their algorithm (CODA) is similar to the standard FedAvg method [26] except that the periodic averaging is applied both to the primal and the dual variables. Nevertheless, their results on FDAM are not comprehensive. By a deep investigation of their algorithms and analysis, we found that (i) although their FL algorithm CODA was shown to be better than the naive parallel algorithm (NPA) with a small mini-batch for DAM, the NPA using a larger mini-batch at local machines can

enjoy a smaller communication complexity than CODA; (ii) the communication complexity of CODA for homogeneous data becomes better than that was established for the heterogeneous data, but is still worse than that of NPA with a large mini-batch at local clients. These shortcomings of CODA for FDAM motivate us to develop better federated averaging algorithms and analysis with a better communication complexity without sacrificing the sample complexity.

This paper aims to provide more comprehensive results for FDAM, with a focus on improving the communication complexity of CODA for heterogeneous data. In particular, our contributions are summarized below:

- First, we provide a stronger baseline with a simpler algorithm than CODA named CODA+, and establish its complexity in both homogeneous and heterogeneous data settings. Although CODA+ has a slight change from CODA, its analysis is much more involved than that of CODA, which is based on the duality gap analysis instead of the primal objective gap analysis.
- Second, we propose a new variant of CODA+ named CODASCA with a much improved communication complexity than CODA+. The key thrust is to incorporate the idea of stochastic controlled averaging of SCAFFOLD [14] into the framework of CODA+ to correct the client-drift for both local primal updates and local dual updates. A striking result of CODASCA under a PL condition for deep learning is that its communication complexity is independent of the number of machines and the targeted accuracy level, which is even better than CODA+ in the homogeneous data setting. The analysis of CODASCA is also non-trivial that combines the duality gap analysis of CODA+ for a non-convex strongly-concave min-max problem and the variance reduction analysis of SCAFFOLD. The comparison between CODASCA and CODA+ and the NPA for FDAM is shown in Table 1.
- Third, we conduct experiments on benchmark datasets to verify our theory by showing CODASCA can enjoy a larger communication window size than CODA+ without sacrificing the performance. Moreover, we conduct empirical studies on medical chest X-ray images from different hospitals by showing that the performance of CODASCA using data from multiple organizations can improve the performance on testing data from a single hospital.

2. Related Work

Federated Learning (FL). Many empirical studies [34, 38, 27, 5, 18, 13, 48] have shown that FL exhibits good empirical performance for distributed deep learning. For a more thorough survey of FL, we refer the readers to [28]. This paper is closely related to recent studies on the design of distributed stochastic algorithms for FL with provable convergence guarantee.

The most popular FL algorithm is Federated Averaging (FedAvg) [26], also referred to as local SGD [37]. Stich [37] is the first that establishes the convergence of local SGD for strongly convex functions. Yu et al. [47, 46] establishes the convergence of local SGD and their momentum variants for non-convex functions. The analysis in [47] has exhibited the difference of communication complexities of local SGD in homogeneous and heterogeneous data settings, which is also discovered in recent works [15, 42, 41]. These latter studies

provide a tight analysis of local SGD in homogeneous and/or heterogeneous data settings, improving its upper bounds for convex functions and strongly convex functions than some earlier works, which sometimes improve over large mini-batch SGD, e.g., when the level of heterogeneity is sufficiently small.

Haddadpour et al. [10] improve the complexities of local SGD for non-convex optimization by leveraging the Polyak-Łojasiewicz (PL) condition. [14] propose a new FedAvg algorithm SCAFFOLD by introducing control variates (variance reduction) to correct for the 'client-drift' in the local updates for heterogeneous data. The communication complexities of SCAFFOLD are no worse than that of large mini-batch SGD for both strongly convex and non-convex functions. The proposed algorithm CODASCA is inspired by the idea of stochastic controlled averaging of SCAFFOLD. However, the analysis of CODASCA for non-convex min-max optimization under a PL condition of the primal objective function is non-trivial compared to that of SCAFFOLD.

AUC Maximization. This work builds on the foundations of stochastic AUC maximization developed in many previous works. Ying et al. [45] address the scalability issue of optimizing AUC by introducing a min-max reformulation of the AUC square surrogate loss and solving it by a convex-concave stochastic gradient method [30]. Natole et al. [29] improve the convergence rate by adding a strongly convex regularizer into the original formulation. Based on the same min-max formulation as in [45], Liu et al. [21] achieve an improved convergence rate by developing a multi-stage algorithm by leveraging the quadratic growth condition of the problem. However, all of these studies focus on learning a linear model, whose corresponding problem is convex and strongly concave. Yuan et al. [49] propose a more robust margin-based surrogate loss for the AUC score, which can be formulated as a similar min-max problem to the AUC square surrogate loss.

Deep AUC Maximization (DAM). [35] is the first work that develops algorithms and convergence theories for weakly convex and strongly concave min-max problems, which is applicable to DAM. However, their convergence rate is slow for a practical purpose. Liu et al. [22] consider improving the convergence rate for DAM under a practical PL condition of the primal objective function. Guo et al. [9] further develop more generic algorithms for non-convex strongly-concave min-max problems, which can also be applied to DAM. There are also several studies [43, 19, 24, 44] focusing on non-convex strongly concave min-max problems without considering the application to DAM. Based on Liu et al. [22]'s algorithm, Guo et al. [8] propose a communication-efficient FL algorithm (CODA) for DAM. However, its communication cost is still high for heterogeneous data.

DL for Medical Image Analysis. In past decades, machine learning, especially deep learning methods have revolutionized many domains such as machine vision, natural language processing. For medical image analysis, deep learning methods are also showing great potential such as in classification of skin lesions [7, 17], interpretation of chest radiographs [1, 11], and breast cancer screening [3, 25, 39]. Some works have already achieved expert-level performance in different tasks [1, 25, 20]. Recently, Yuan et al. [49] employ DAM for medical image classification and achieve great success on two challenging tasks, namely CheXpert competition for chest X-ray image classification and Kaggle competition for melanoma classification based on skin lesion images. However, to the best of our knowledge, the application of FDAM methods on medical datasets from different hospitals have not be thoroughly investigated.

3. Preliminaries and Notations

We consider federated learning of deep neural networks by maximizing the AUC score. The setting is the same to that was considered as in [8]. Below, we present some preliminaries and notations, which are mostly the same as in [8]. In this paper, we consider the following min-max formulation for distributed problem:

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^d\\(a,b)\in\mathbb{R}^2}} \max_{\alpha\in\mathbb{R}} f(\mathbf{w},a,b,\alpha) = \frac{1}{K}\sum_{k=1}^{K} f_k(\mathbf{w},a,b,\alpha), \tag{1}$$

where K is the total number of machines. This formulation covers a class of non-convex strongly concave min-max problems and specifically for the AUC maximization, $f_k(\mathbf{w},a,b,\alpha)$ is defined below.

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

where $\mathbf{z}^k = (\mathbf{x}^k, y^k) \sim \mathbb{P}_k$, $\mathbb{P}_k$ is the data distribution on machine k, p is the ratio of positive data. When $\phi_k = \phi_l, \forall k \neq l$, this is referred to as the homogeneous data setting; otherwise heterogeneous data setting.

Notations. We define the following notations:

$$\begin{aligned} &\mathbf{v} = (\mathbf{w}^T, a, b)^T, \quad \phi(\mathbf{v}) = \max_{\alpha} f(\mathbf{v},\alpha), \\ &\phi_s(\mathbf{v}) = \phi(\mathbf{v}) + \frac{1}{2\gamma}\|\mathbf{v}-\mathbf{v}_{s-1}\|^2, \\ &f^s(\mathbf{v},\alpha) = f(\mathbf{v},\alpha) + \frac{1}{2\gamma}\|\mathbf{v}-\mathbf{v}_{s-1}\|^2 \\ &F_k^s(\mathbf{v},\alpha;\mathbf{z}_k) = F_k(\mathbf{v},\alpha;\mathbf{z}_k) + \frac{1}{2\gamma}\|\mathbf{v}-\mathbf{v}_{s-1}\|^2 \\ &\mathbf{v}_\phi^* = \arg\min_{\mathbf{v}} \phi(\mathbf{v}), \quad \mathbf{v}_{\phi_s}^* = \arg\min_{\mathbf{v}} \phi_s(\mathbf{v}). \end{aligned}$$

Assumptions. Similar to [8], we make the following assumptions throughout this paper.

Assumption 1

(i) There exist $\mathbf{v}_0, \Delta_0 > 0$ *such that* $\phi(\mathbf{v}_0) - \phi(\mathbf{v}_\phi^*) \leq \Delta_0$.
(ii) PL condition: $\phi(\mathbf{v})$ *satisfies the* μ*-PL condition, i.e.,* $\mu(\phi(\mathbf{v}) - \phi(\mathbf{v}_*)) \leq \frac{1}{2}\|\nabla\phi(\mathbf{v})\|^2$*;*
(iii) Smoothness: For any $\mathbf{z}$*,* $f(\mathbf{v},\alpha;\mathbf{z})$ *is* ℓ*-smooth in* $\mathbf{v}$ *and* α*.* $\phi(\mathbf{v})$ *is* L*-smooth, i.e.,* $\|\nabla\phi(\mathbf{v}_1) - \nabla\phi(\mathbf{v}_2)\| \leq L\|\mathbf{v}_1 - \mathbf{v}_2\|$.
(iv) Bounded variance:

$$\begin{aligned} &\mathbb{E}[\|\nabla_{\mathbf{v}} f_k(\mathbf{v},\alpha) - \nabla_{\mathbf{v}} F_k(\mathbf{v},\alpha;\mathbf{z})\|^2] \leq \sigma^2, \\ &\mathbb{E}[|\nabla_\alpha f_k(\mathbf{v},\alpha) - \nabla_\alpha F_k(\mathbf{v},\alpha;\mathbf{z})|^2] \leq \sigma^2. \end{aligned} \tag{3}$$

Algorithm 1 CODA+

1: Initialization: $(\mathbf{v}_0, \alpha_0, \gamma)$.
2: **for** $s = 1, ..., S$ **do**
3: $\quad \mathbf{v}_s, \alpha_s = \text{DSG+}(\mathbf{v}_{s-1}, \alpha_{s-1}, \eta_s, I_s, \gamma)$;
4: **end for**
5: Return $\mathbf{v}_S, \alpha_S$.

To quantify the drifts between different clients, we introduce the following assumption.

Assumption 2 *Bounded client drift:*

$$\begin{aligned} \frac{1}{K}\sum_{k=1}^{K}\|\nabla_{\mathbf{v}} f_k(\mathbf{v}, \alpha) - \nabla_{\mathbf{v}} f(\mathbf{v}, \alpha)\|^2 \leq D^2, \\ \frac{1}{K}\sum_{k=1}^{K}\|\nabla_{\alpha} f_k(\mathbf{v}, \alpha) - \nabla_{\alpha} f(\mathbf{v}, \alpha)\|^2 \leq D^2. \end{aligned} \tag{4}$$

Remark. D quantifies the drift between the local objectives and the global objective. $D = 0$ denotes the homogeneous data setting that all the local objectives are identical. $D > 0$ corresponds to the heterogeneous data setting.

4. CODA+: A stronger baseline

In this section, we present a stronger baseline than CODA [8]. The motivation is that (i) the CODA algorithm uses a step to compute the dual variable from the primal variable by using sampled data from all clients; but we find this step is unnecessary by an improved analysis; (ii) the complexity of CODA for homogeneous data is not given in its original paper. Hence, CODA+ is a simplified version of CODA but with much refined analysis.

We present the steps of CODA+ in Algorithm 1. It is similar to CODA that uses stage-wise updates. In s-th stage, a strongly convex strongly concave subproblem is constructed:

$$\min_{\mathbf{v}} \max_{\alpha} f(\mathbf{v}, \alpha) + \frac{\gamma}{2}\|\mathbf{v} - \mathbf{v}_0^s\|^2, \tag{5}$$

where $\mathbf{v}_0^s$ is the output of the previous stage.

CODA+ improves upon CODA in two folds. First, CODA+ algorithm is more concise since the output primal and dual variables of each stage can be directly used as input for the next stage, while CODA needs an extra large batch of data after each stage to compute the dual variable. This modification not only reduces the sample complexity, but also makes the algorithm applicable to a boarder family of nonconvex min-max problems. Second, CODA+ has a smaller communication complexity for homogeneous data than that for heterogeneous data while the previous analysis of CODA yields the same communication complexity for homogeneous data and heterogeneous data.

We have the following lemma to bound the convergence for the subproblem in each s-th stage.

Algorithm 2 DSG+$(\mathbf{v}_0, \alpha_0, \eta, T, I, \gamma)$

Each machine does initialization: $\mathbf{v}_0^k = \mathbf{v}_0, \alpha_0^k = \alpha_0$,
for $t = 0, 1, ..., T-1$ **do**
  Each machine k updates its local solution in parallel:
    $\mathbf{v}_{t+1}^k = \mathbf{v}_t^k - \eta(\nabla_{\mathbf{v}} F_k(\mathbf{v}_t^k, \alpha_t^k; \mathbf{z}_t^k) + \gamma(\mathbf{v}_t^k - \mathbf{v}_0))$,
    $\alpha_{t+1}^k = \alpha_t^k + \eta \nabla_\alpha F_k(\mathbf{v}_t^k, \alpha_t^k; \mathbf{z}_t^k)$,
  if $t+1 \bmod I = 0$ **then**
    $\mathbf{v}_{t+1}^k = \frac{1}{K}\sum_{k=1}^K \mathbf{v}_{t+1}^k$, ◇ communicate
    $\alpha_{t+1}^k = \frac{1}{K}\sum_{k=1}^K \alpha_{t+1}^k$, ◇ communicate
  end if
end for
Return $\left(\bar{\mathbf{v}} = \frac{1}{K}\sum_{k=1}^K \frac{1}{T}\sum_{t=1}^T \mathbf{v}_t^k, \bar{\alpha} = \frac{1}{K}\sum_{k=1}^K \frac{1}{T}\sum_{t=1}^T \alpha_t^k\right)$.

Lemma 1 *(One call of Algorithm 2) Let $(\bar{\mathbf{v}}, \bar{\alpha})$ be the output of Algorithm 2. Suppose Assumption 1 and 2 hold. By running Algorithm 2 with given input $\mathbf{v}_0, \alpha_0$ for T iterations, $\gamma = 2\ell$, and $\eta \le \min(\frac{1}{3\ell + 3\ell^2/\mu_2}, \frac{1}{4\ell})$, we have for any $\mathbf{v}$ and α*

$$\mathbb{E}[f^s(\bar{\mathbf{v}}, \alpha) - f^s(\mathbf{v}, \bar{\alpha})] \le \frac{1}{\eta T}\|\mathbf{v}_0 - \mathbf{v}\|^2 + \frac{1}{\eta T}(\alpha_0 - \alpha)^2 + \underbrace{\left(\frac{3\ell^2}{2\mu_2} + \frac{3\ell}{2}\right)(12\eta^2 I\sigma^2 + 36\eta^2 I^2 D^2)\mathbb{I}_{I>1}}_{A_1} + \frac{3\eta\sigma^2}{K},$$

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

Remark. Note that the term A_1 on the RHS is the drift of clients caused by skipping communication. When $D = 0$, i.e., the machines have homogeneous data distribution, we need $\eta I = O\left(\frac{1}{K}\right)$, then A_1 can be merged with the last term. When $D > 0$, we need $\eta I^2 = O\left(\frac{1}{K}\right)$, which means that I has to be smaller in heterogeneous data setting and thus the communication complexity is higher.

Remark. The key difference between the analysis of CODA+ and that of CODA lies at how to handle the term $(\alpha_0 - \alpha)^2$ in Lemma 1. In CODA, the initial dual variable α_0 is computed from the initial primal variable $\mathbf{v}_0$, which reduces the error term $(\alpha_0 - \alpha)^2$ to one similar to $\|\mathbf{v}_0 - \mathbf{v}\|^2$, which is then bounded by the primal objective gap due to the PL condition. However, since we do not conduct the extra computation of α_0 from $\mathbf{v}_0$, our analysis directly deals with such error term by using the duality gap of f^s. This technique is originally developed by [43].

Theorem 1 *Define $\hat{L}$=$L + 2\ell, c = \frac{\mu/\hat{L}}{5+\mu/\hat{L}}$. Set $\gamma = 2\ell$, $\eta_s = \eta_0 \exp(-(s-1)c)$, $T_s = \frac{212}{\eta_0 \min(\ell, \mu_2)}\exp((s-1)c)$. To return $\mathbf{v}_S$ such that $\mathbb{E}[\phi(\mathbf{v}_S) - \phi(\mathbf{v}_\phi^*)] \le \epsilon$, it suffices to*

choose $S \geq O\left(\frac{5\hat{L}+\mu}{\mu}\max\left\{\log\left(\frac{2\Delta_0}{\epsilon}\right), \log S + \log\left[\frac{2\eta_0}{\epsilon}\frac{12(\sigma^2)}{5K}\right]\right\}\right)$. *The iteration complexity is* $\widetilde{O}\left(\max\left(\frac{\Delta_0}{\mu\epsilon\eta_0 K}, \frac{\hat{L}}{\mu^2 K\epsilon}\right)\right)$ *and the communication complexity is* $\widetilde{O}\left(\frac{K}{\mu}\right)$ *by setting* $I_s = \Theta(\frac{1}{K\eta_s})$ *if* $D = 0$, *and is* $\widetilde{O}\left(\max\left(\frac{K}{\mu} + \frac{\Delta_0^{1/2}}{\mu(\eta_0\epsilon)^{1/2}}, \frac{K}{\mu} + \frac{\hat{L}^{1/2}}{\mu^{3/2}\epsilon^{1/2}}\right)\right)$ *by setting* $I_s = \Theta(\frac{1}{\sqrt{K\eta_s}})$ *if* $D > 0$, *where* $\widetilde{O}$ *suppresses logarithmic factors.*

Remark. Due to the PL condition, the step size η decreases geometrically. Accordingly, I increases geometrically due to Lemma 1, and I increases with a faster rate when the data are homogeneous than that when data are heterogeneous. In result, the total number of communications in homogeneous setting is much less than that in heterogeneous setting.

5. CODASCA

Although CODA+ has a highly reduced communication complexity for homogeneous data, it is still suffering from a high communication complexity for heterogeneous data. Even for the homogeneous data, CODA+ has a worse communication complexity with a dependence on the number of clients K than the NPA algorithm with a large batch size.

Can we further reduce the communication complexity for FDAM for both homogeneous and heterogeneous data without using a large batch size?

The main reason for the degeneration in the heterogeneous data setting is the data difference. Even at global optimum $(\mathbf{v}_*, \alpha_*)$, the gradient of local functions in different clients could be different and non-zero. In the homogeneous data setting, different clients still produce different solutions due to stochastic error (cf. the $\eta^2\sigma^2 I$ term of A_1 in Lemma 1). These together contribute to the client drift.

To correct the client drift, we propose to leverage the idea of stochastic controlled averaging due to [14]. The key idea is to maintain and update a control variate to accommodate the client drift, which is taken into account when updating the local solutions. In the proposed algorithm CODASCA, we apply control variates to both primal and dual variables. CODASCA shares the same stagewise framework as CODA+, where a strongly convex strongly concave subproblem is constructed and optimized in a distributed fashion approximately in each stage. The steps of CODASCA are presented in Algorithm 3 and Algorithm 4. Below, we describe the algorithm in each stage.

Each stage has R communication rounds. Between two rounds, there are I local updates, and each machine k does the local updates as

$$\begin{aligned}
\mathbf{v}_{r,t+1}^k &= \mathbf{v}_{r,t}^k - \eta_l(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; \mathbf{z}_{r,t}^t) - c_{\mathbf{v}}^k + c_{\mathbf{v}}),\\
\alpha_{r,t+1}^k &= \alpha_{r,t}^k + \eta_l(\nabla_{\alpha} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; \mathbf{z}_{r,t}^k) - c_{\alpha}^k + c_{\alpha}),
\end{aligned}$$

where $c_{\mathbf{v}}^k, c_{\mathbf{v}}$ are local and global control variates for the primal variable, and c_{α}^k, c_{α} are local and global control variates for the dual variable. Note that $\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; \mathbf{z}_{r,t}^t)$ and $\nabla_{\alpha} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; \mathbf{z}_{r,t}^k)$ are unbiased stochastic gradient on local data. However, they are biased estimate of global gradient when data on different clients are heterogeneous. Intuitively, the term $-c_{\mathbf{v}}^k + c_{\mathbf{v}}$ and $-c_{\alpha}^k + c_{\alpha}$ work to correct the local gradients to get closer to the

global gradient. They also play a role of reducing variance of stochastic gradients, which is helpful as well to reduce the communication complexity in the homogeneous data setting.

At each communication round, the primal and dual variables on all clients get aggregated, averaged and broadcast to all clients. The control variates c at r-th round get updated as

$$\begin{aligned} c_{\mathbf{v}}^k &= c_{\mathbf{v}}^k - c_{\mathbf{v}} + \frac{1}{I\eta_l}(\mathbf{v}_{r-1} - \mathbf{v}_{r,I}^k), \\ c_{\alpha}^k &= c_{\alpha}^k - c_{\alpha} + \frac{1}{I\eta_l}(\alpha_{r,I}^k - \alpha_{r-1}), \end{aligned} \tag{6}$$

which is equivalent to

$$\begin{aligned} c_{\mathbf{v}}^k &= \frac{1}{I}\sum_{t=1}^{I} \nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; \mathbf{z}_{r,t}^k), \\ c_{\alpha}^k &= \frac{1}{I}\sum_{t=1}^{I} \nabla_{\alpha} f_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; \mathbf{z}_{r,t}^k). \end{aligned} \tag{7}$$

Notice that they are simply the average of stochastic gradients used in this round. An alternative way to compute the control variates is by computing the stochastic gradient with a large batch of extra samples at each client, but this would bring extra cost and is unnecessary. $c_{\mathbf{v}}$ and c_{α} are averages of $c_{\mathbf{v}}^k$ and c_{α}^k over all clients. After the local primal and dual variables are averaged, an extrapolation step with $\eta_g > 1$ is performed, which will boost the convergence.

In order to establish the convergence of CODASCA, we first present a key lemma below.

Lemma 2 *(One call of Algorithm 4) Under the same setting as in Theorem 2, with* $\tilde{\eta} = \eta_l\eta_g I \leq \frac{\mu_2}{40\ell^2}$*, for* $\mathbf{v}' = \arg\min_{\mathbf{v}} f^s(\mathbf{v}, \alpha_{\tilde{r}}), \alpha' = \arg\max_{\alpha} f^s(\mathbf{v}_{\tilde{r}}, \alpha)$ *we have*

$$\mathbb{E}[f^s(\mathbf{v}_{\tilde{r}}, \alpha') - f^s(\mathbf{v}', \alpha_{\tilde{r}})] \leq \frac{2}{\eta_l\eta_g T}\|\mathbf{v}_0 - \mathbf{v}'\|^2 + \frac{2}{\eta_l\eta_g T}(\alpha_0 - \alpha')^2 + \underbrace{\frac{10\eta_l\sigma^2}{\eta_g}}_{A_2} + \frac{10\eta_l\eta_g\sigma^2}{K}$$

where $T = I \cdot R$ *is the number of iterations for each stage.*

Remark. Compared the above bound with that in Lemma 1, in particular the term A_2 vs the term A_1, we can see that CODASCA will not be affected by the data heterogeneity $D > 0$, and the stochastic variance is also much reduced. As will seen in the next theorem, the value of $\tilde{\eta}$ and R will keep the same in all stages. Therefore, by decreasing local step size η_l geometrically, the communication window size I_s will increase geometrically to ensure $\tilde{\eta} \leq O(1)$.

The convergence result of CODASCA is presented below.

Theorem 2 *Define* $\hat{L} = L + 2\ell$*,* $c = 4\ell + \frac{248}{53}\hat{L}$*. Set* $\eta_g = \sqrt{K}$*,* $I_s = I_0 \exp\left(\frac{2\mu_1}{c+2\mu_1}(s-1)\right)$*,* $R = \frac{1000}{\tilde{\eta}\mu_2}$*,* $\eta_l^s = \frac{\tilde{\eta}}{\eta_g I_s} = \frac{\tilde{\eta}}{\sqrt{K}I_0}\exp\left(-\frac{2\mu}{c+2\mu}(s-1)\right)$*,* $\tilde{\eta} \leq \min\{\frac{1}{3\ell+3\ell^2/\mu_2}, \frac{\mu_2}{40\ell^2}\}$*. After* $S =$

Algorithm 3 CODASCA

Initialization: $(\mathbf{v}_0, \alpha_0, \gamma)$.
for $s = 1, ..., S$ **do**
  $\mathbf{v}_s, \alpha_s = \text{DSGSCA+}(\mathbf{v}_{s-1}, \alpha_{s-1}, \eta_l, \eta_g, I_s, R_s, \gamma)$;
end for
Return $\mathbf{v}_S, \alpha_S$.

Algorithm 4 DSGSCA+$(\mathbf{v}_0, \alpha_0, \eta_l, \eta_g, I, R, \gamma)$

Each machine does initialization: $\mathbf{v}_{0,0}^k = \mathbf{v}_0, \alpha_{0,0}^k = \alpha_0$, $c_{\mathbf{v}}^k = \mathbf{0}$, $c_\alpha^k = 0$
for $r = 1, ..., R$ **do**
  for $t = 0, 1, ..., I-1$ **do**
    Each machine k updates its local solution in parallel:
    $\mathbf{v}_{r,t+1}^k = \mathbf{v}_{r,t}^k - \eta_l(\nabla_{\mathbf{v}} F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; \mathbf{z}_{r,t}^k) - c_{\mathbf{v}}^k + c_{\mathbf{v}})$,
    $\alpha_{r,t+1}^k = \alpha_{r,t}^k + \eta_l(\nabla_\alpha F_k^s(\mathbf{v}_{r,t}^k, \alpha_{r,t}^k; \mathbf{z}_{r,t}^k) - c_\alpha^k + c_\alpha)$,
  end for
  $c_{\mathbf{v}}^k = c_{\mathbf{v}}^k - c_{\mathbf{v}} + \frac{1}{I\eta_l}(\mathbf{v}_{r-1} - \mathbf{v}_{r,I}^k)$
  $c_\alpha^k = c_\alpha^k - c_\alpha + \frac{1}{I\eta_l}(\alpha_{r,I}^k - \alpha_{r-1})$
  $c_{\mathbf{v}} = \frac{1}{K}\sum_{k=1}^K c_{\mathbf{v}}^k$, $c_\alpha = \frac{1}{K}\sum_{k=1}^K c_\alpha^k$ ◇ communicate
  $\mathbf{v}_r = \frac{1}{K}\sum_{k=1}^K \mathbf{v}_{r,I}^k, \alpha_r = \frac{1}{K}\sum_{k=1}^K \alpha_{r,t}^k$ ◇ communicate
  $\mathbf{v}_r = \mathbf{v}_{r-1} + \eta_g(\mathbf{v}_r - \mathbf{v}_{r-1})$,
  $\alpha_r = \alpha_{r-1} + \eta_g(\alpha_r - \alpha_{r-1})$
  Broadcast $\mathbf{v}_r, \alpha_r, c_{\mathbf{v}}, c_\alpha$ ◇ communicate
end for
Return $\mathbf{v}_{\tilde{r}}, \alpha_{\tilde{r}}$ where $\tilde{r}$ is randomly sampled from $1, ..., R$

$O(\max\left\{\frac{c+2\mu}{2\mu}\log\frac{4\epsilon_0}{\epsilon}, \frac{c+2\mu}{2\mu}\log\frac{160\hat{L}S}{(c+2\mu)\epsilon}\frac{\tilde{\eta}\sigma^2}{KI_0}\right\})$ *stages, the output* $\mathbf{v}_S$ *satisfies* $\mathbb{E}[\phi(\mathbf{v}_S) - \phi(\mathbf{v}_\phi^*)] \leq \epsilon$. *The communication complexity is* $\widetilde{O}\left(\frac{1}{\mu}\right)$. *The iteration complexity is* $\widetilde{O}\left(\max\{\frac{1}{\mu\epsilon}, \frac{1}{\mu^2 K\epsilon}\}\right)$.

Remark. (i) The number of communications is $\widetilde{O}\left(\frac{1}{\mu}\right)$, independent of number of clients K and the accuracy level ϵ. This is a significant improvement over CODA+, which has a communication complexity of $\widetilde{O}\left(K/\mu + 1/(\mu^{3/2}\epsilon^{1/2})\right)$ in heterogeneous setting. Moreover, $\widetilde{O}\left(1/(\mu)\right)$ is a nearly optimal rate up to a logarithmic factor, since $O(1/\mu)$ is the lower bound communication complexity of distributed strongly convex optimization [14, 2] and strongly convexity is a stronger condition than the PL condition.

(ii) Each stage has the same number of communication rounds. However, I_s increases geometrically. Therefore, the number of iterations and samples in a stage increase geometrically. Theoretically, we can also set η_l^s to the same value as the one in the last stage, correspondingly I_s can be set as a fixed large value. But this increases the number of required samples without further speeding up the convergence. Our setting of I_s is a balance between skipping communications and reducing sample complexity. For simplicity, we use the fixed setting of I_s to compare CODASCA and the baseline CODA+ in our experiment to corroborate the theory.

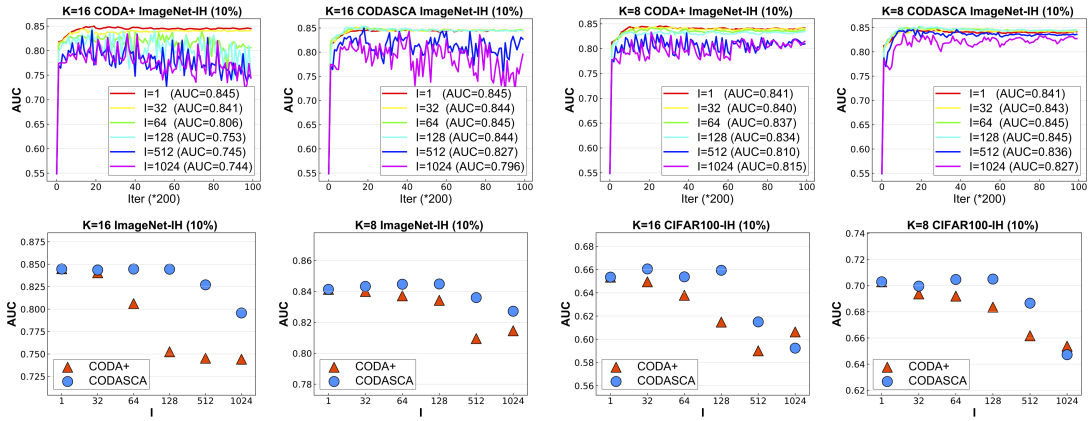

Figure 1: Top row: the testing AUC score of CODASCA vs # of iterations for different values of I on ImageNet-IH and CIFAR100-IH with imratio = 10% and K=16, 8 on Densenet121. Bottom row: the achieved testing AUC vs different values of I for CODASCA and CODA+. The AUC score in the legend in top row figures represent the AUC score at the last iteration.

(iii) The local step size η_l of CODASCA decreases similarly as the step size η in CODA+. But $I_s = O(1/(\sqrt{K}\eta_l^s))$ in CODASCA increases faster than that $I_s = O(1/(\sqrt{K\eta_s}))$ in CODA+ on heterogeneous data. It is noticeable that different from CODA+, we do not need Assumption 2 which bounds the client drift, meaning that CODASCA can be applied to optimize the global objective even if local objectives arbitrarily deviate from the global function.

6. Experiments

In this section, we first verify the effectiveness of CODASCA compared to CODA+ on various datasets, including two benchmark datasets, i.e., ImageNet, CIFAR100 [6, 16] and a constructed large-scale chest X-ray dataset. Then, we demonstrate the effectiveness of FDAM on improving the performance on a single domain (CheXpert) by using data from multiple sources. For notations, K denotes the number of "clients" (# of machines, # of data sources) and I denotes the communication window size. The code used for the experiments are available at `https://github.com/Optimization-AI/ICML2021_FedDeepAUC_CODASCA/`.

Chest X-ray datasets. Five medical chest X-ray datasets, i.e., CheXpert, ChestXray14, MIMIC-CXR, PadChest, ChestXray-AD [11, 40, 12, 4, 32] are collected from different organizations. The statistics of these medical datasets are summarized in Table 2. We construct five binary classification tasks for predicting five popular diseases, Cardiomegaly (C0), Edema (C1), Consolidation (C2), Atelectasis (C3), P. Effusion (C4), as in CheXpert competition [11]. These datasets are naturally imbalanced and heterogeneous due to different patients' populations, different data collection protocols and etc. We refer to the whole medical dataset as ChestXray-IH.

Imbalanced and Heterogeneous (IH) Benchmark Datasets. For benchmark datasets, we manually construct the imbalanced heterogeneous dataset. For ImageNet, we

Table 2: Statistics of Medical Chest X-ray Datasets.

Dataset	Source	Samples
CheXpert	Stanford Hospital (US)	224,316
ChestXray8	NIH Clinical Center (US)	112,120
PadChest	Hospital San Juan (Spain)	110,641
MIMIC-CXR	BIDMC (US)	377,110
ChestXrayAD	H108 and HMUH (Vietnam)	15,000

first randomly select 500 classes as positive class and 500 classes as negative class. To increase data heterogeneity, we further split all positive/negative classes into K groups so that each split only owns samples from unique classes without overlapping with that of other groups. To increase data imbalance level, we randomly remove some samples from positive classes for each machine. Please note that due to this operation, the whole sample set for different K is different. We refer to the proportion of positive samples in all samples as imbalance ratio (*imratio*). For CIFAR100, we follow similar steps to construct imbalanced heterogeneous data. We keep the testing/validation set untouched and keep them balanced. For imbalance ratio (imratio), we explore two ratios: 10% and 30%. We refer to the constructed datasets as ImageNet-IH (10%), ImageNet-IH (30%), CIFAR100-IH (10%), CIFAR100-IH (30%). Due to the limited space, we only report imratio=10% with DenseNet121 and defer the other results to supplement.

Parameters and Settings. We train Desenet121 on all datasets. For the parameters in CODASCA/CODA+, we tune $1/\gamma$ in [500, 700, 1000] and η in [0.1, 0.01, 0.001]. For learning rate schedule, we decay the step size by 3 times every T_0 iterations, where T_0 is tuned in [2000, 3000, 4000]. We experiment with a fixed value of I selected from [1, 32, 64, 128, 512, 1024] and we include experiments with increasing I_s in the supplement. We tune η_g in [1.1, 1, 0.99, 0.999]. The local batch size is set to 32 for each machine. We run a total of 20000 iterations for all experiments.

6.1 Comparison with CODA+

We plot the testing AUC on ImageNet (10%) vs # of iterations for CODASCA and CODA+ in Figure 1 (top row) by varying the value of I for different values of K. Results on CIFAR100 are shown in the Supplement. In the bottom row of Figure 1, we plot the achieved testing AUC score vs different values of I for CODASCA and CODA+. We have the following observations:

• **CODASCA enjoys a larger communication window size**. Comparing CODASCA and CODA+ in the bottom panel of Figure 1, we can see that CODASCA enjoys a larger communication window size without hurting the performance than CODA+, which is consistent with our theory.

• **CODASCA is consistently better for different values of** K. We compare the largest value of I such that the performance does not degenerate too much compared with $I = 1$, which is denoted by $I_{\max}$. From the bottom figures of Figure 1, we can see that the $I_{\max}$ value of CODASCA on ImageNet is 128 (K=16) and 512 (K=8), respectively, and that of CODA+ on ImageNet is 32 (K=16) and 128 (K=8). This demonstrates that CODASCA enjoys consistent advantage over CODA+, i.e., when $K = 16$, $I_{\max}^{\text{CODASCA}}/I_{\max}^{\text{CODA+}} = 4$, and when $K = 8$, $I_{\max}^{\text{CODASCA}}/I_{\max}^{\text{CODA+}} = 4$. The same phenomena occur on CIFAR100 data.

Table 3: Performance on ChestXray-IH testing set when K=16.

Method	I	C0	C1	C2	C3	C4
	1	0.8472	0.8499	0.7406	0.7475	0.8688
CODA+	512	0.8361	**0.8464**	0.7356	0.7449	0.8680
CODASCA	512	**0.8427**	0.8457	**0.7401**	**0.7468**	**0.8680**
CODA+	1024	0.8280	**0.8451**	0.7322	0.7431	0.8660
CODASCA	1024	**0.8363**	0.8444	**0.7346**	**0.7481**	**0.8674**

Table 4: Performance of FDAM on Chexpert validation set for DenseNet121.

#of sources	C0	C1	C2	C3	C4	AVG
K=1	0.9007	0.9536	0.9542	0.9090	0.9571	0.9353
K=2	0.9027	0.9586	0.9542	0.9065	0.9583	0.9361
K=3	0.9021	0.9558	**0.9550**	0.9068	0.9583	0.9356
K=4	0.9055	**0.9603**	0.9542	0.9072	**0.9588**	0.9372
K=5	**0.9066**	0.9583	0.9544	**0.9101**	0.9584	**0.9376**

Next, we compare CODASCA with CODA+ on the ChestXray-IH medical dataset, which is also highly heterogeneous. We split the ChestXray-IH data into $K = 16$ groups according to the patient ID and each machine only owns samples from one organization without overlapping patients. The testing set is the collection of 5% data sampled from each organization. In addition, we use train/val split = 7:3 for the parameter tuning. We run CODASCA and CODA+ with the same number of iterations. The performance on testing set are reported in Table 3. From the results, we can observe that CODASCA performs consistently better than CODA+ on C0, C2, C3, C4.

6.2 FDAM for improving performance on CheXpert

Finally, we show that FDAM can be used to leverage data from multiple hospitals to improve the performance at a single target hospital. For this experiment, we choose CheXpert data from Stanford Hospital as the target data. Its validation data will be used for evaluating the performance of our FDAM method. Note that improving the AUC score on CheXpert is a very challenging task. The top 7 teams on CheXpert leaderboard differ by only 0.1% [1]. Hence, we consider any improvement over 0.1% significant. Our procedure is following: we gradually increase the number of data resources, e.g., $K = 1$ only includes the CheXpert training data, $K = 2$ includes the CheXpert training data and ChestXray8, $K = 3$ includes the CheXpert training data and ChestXray8 and PadChest, and so on.

Parameters and Settings. Due to the limited computing resources, we resize all images to 320x320. We follow the two stage method proposed in [49] and compare with the baseline on a single machine with a single data source (CheXpert training data) (K=1) for learning DenseNet121, DenseNet161. More specifically, we first train a base model by minimizing the Cross-Entropy loss on CheXpert training dataset using Adam with a initial learning rate of 1e-5 and batch size of 32 for 2 epochs. Then, we discard the trained classifier, use the same pretrained model for initializing the local models at all machines and continue training using CODASCA. For the parameter tuning, we try I=[16, 32, 64, 128], learning rate=[0.1, 0.01] and we fix γ=1e-3, T_0=1000 and batch size=32.

Results. We report all results in term of AUC score on the CheXpert validation data in Table 4 and Table 5. We can see that using more data sources from different

1. `https://stanfordmlgroup.github.io/competitions/chexpert/`

Table 5: Performance of FDAM on Chexpert validation set for DenSenet161.

#of sources	C0	C1	C2	C3	C4	AVG
K=1	0.8946	0.9527	0.9544	0.9008	0.9556	0.9316
K=2	0.8938	0.9615	0.9568	0.9109	0.9517	0.9333
K=3	**0.9008**	0.9603	0.9568	0.9127	0.9505	0.9356
K=4	0.8986	**0.9615**	0.9561	0.9128	**0.9564**	0.9367
K=5	0.8986	0.9612	**0.9568**	**0.9130**	0.9552	**0.9370**

organizations can efficiently improve the performance on CheXpert. For DenseNet121, the average improvement across all 5 classification tasks from $K = 1$ to $K = 5$ is over 0.2% which is significant in light of the top CheXpert leaderboard results. Specifically, we can see that CODASCA with K=5 achieves the highest validation AUC score on C0 and C3, and with K=4 achieves the highest on C1 and C4. For DenseNet161, the improvement of average AUC is over 0.5%, which doubles the 0.2% improvement for DenseNet121.

7. Conclusion

In this work, we have conducted comprehensive studies of federated learning for deep AUC maximization. We analyzed a stronger baseline for deep AUC maximization by establishing its convergence for both homogeneous data and heterogeneous data. We also developed an improved variant by adding control variates to the local stochastic gradients for both primal and dual variables, which dramatically reduces the communication complexity. Besides a strong theory guarantee, we exhibit the power of FDAM on real world medical imaging problems. We have shown that our FDAM method can improve the performance on medical imaging classification tasks by leveraging data from different organizations that are kept locally.

Acknowledgements

We are grateful to the anonymous reviewers for their constructive comments and suggestions. This work is partially supported by NSF #1933212 and NSF CAREER Award #1844403.

References

[1] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

[2] Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, pages 1756–1764, 2015.

[3] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning

algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[4] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.

[5] Kai Chen and Qiang Huo. Scalable training of deep learning machines by incremental block training with intra-block parallel optimization and blockwise model-update filtering. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSSP)*, pages 5880–5884, 2016.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE annual Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[7] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[8] Zhishuai Guo, Mingrui Liu, Zhuoning Yuan, Li Shen, Wei Liu, and Tianbao Yang. Communication-efficient distributed stochastic AUC maximization with deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3864–3874, 2020.

[9] Zhishuai Guo, Zhuoning Yuan, Yan Yan, and Tianbao Yang. Fast objective and duality gap convergence for non-convex strongly-concave min-max problems. *arXiv preprint arXiv:2006.06889*, 2020.

[10] Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck R. Cadambe. Local SGD with periodic averaging: Tighter analysis and adaptive synchronization. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 11080–11092, 2019.

[11] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 590–597, 2019.

[12] Alistair E W Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[13] Michael Kamp, Linara Adilova, Joachim Sicking, Fabian Hüger, Peter Schlicht, Tim Wirtz, and Stefan Wrobel. Efficient decentralized deep learning by dynamic model averaging. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 393–409. Springer, 2018.

[14] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[15] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4519–4529, 2020.

[16] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. CIFAR-10 and CIFAR-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6:1, 2009.

[17] Yuexiang Li and Linlin Shen. Skin lesion analysis towards melanoma detection using deep learning network. *Sensors*, 18(2):556, 2018.

[18] Tao Lin, Sebastian U. Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local SGD. In *8th International Conference on Learning Representations (ICLR)*, 2020.

[19] Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 6083–6093, 2020.

[20] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[21] Mingrui Liu, Xiaoxuan Zhang, Zaiyi Chen, Xiaoyu Wang, and Tianbao Yang. Fast stochastic auc maximization with O (1/n)-convergence rate. In *Proceedings of 35th International Conference on Machine Learning (ICML)*, pages 3195–3203, 2018.

[22] Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic AUC maximization with deep neural networks. In *8th International Conference on Learning Representations (ICLR)*, 2020.

[23] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated Learning*, pages 240–254. Springer, 2020.

[24] Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[25] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an AI system for breast cancer screening. *Nature*, 577 (7788):89–94, 2020.

[26] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, 2017.

[27] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

[28] H Brendan McMahan et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1), 2019.

[29] Michael Natole, Yiming Ying, and Siwei Lyu. Stochastic proximal algorithms for auc maximization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3707–3716, 2018.

[30] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[31] Yurii E. Nesterov. *Introductory Lectures on Convex Optimization - A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004.

[32] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *arXiv preprint arXiv:2012.15029*, 2020.

[33] Eduardo HP Pooch, Pedro Ballester, and Rodrigo C Barros. Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In *International Workshop on Thoracic Image Analysis*, pages 74–83. Springer, 2020.

[34] Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur. Parallel training of dnns with natural gradient and parameter averaging. *arXiv preprint arXiv:1410.7455*, 2014.

[35] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

[36] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1): 1–7, 2020.

[37] Sebastian U. Stich. Local SGD converges fast and communicates little. In *7th International Conference on Learning Representations (ICLR)*, 2019.

[38] Hang Su and Haoyu Chen. Experiments on parallel training of deep neural network using model averaging. *arXiv preprint arXiv:1507.01239*, 2015.

[39] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[40] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.

[41] Blake E. Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local SGD for heterogeneous distributed learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[42] Blake E. Woodworth, Kumar Kshitij Patel, Sebastian U. Stich, Zhen Dai, Brian Bullins, H. Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local SGD better than minibatch sgd? In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 10334–10343, 2020.

[43] Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[44] Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[45] Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pages 451–459, 2016.

[46] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 7184–7193, 2019.

[47] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5693–5700, 2019.

[48] Zhuoning Yuan, Zhishuai Guo, Xiaotian Yu, Xiaoyu Wang, and Tianbao Yang. Accelerating deep learning with millions of classes. In *16th European Conference on Computer Vision (ECCV)*, pages 711–726, 2020.

[49] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Appendix 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 [43].

Lemma 3 (Lemma 1 of [43]) *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

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

□

Lemma 4 (Lemma 5 of [43]) *We have the following lower bound for $Gap_s(\mathbf{v}_s, \alpha_s)$*

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

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.

□

Appendix B. Analysis of CODA+

The proof sketch is similar to the proof of CODA in [8]. However, there are two noticeable difference from [8]. First, in Lemma 1, we bound the duality gap instead of the objective gap in [8]. 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 [46], 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 [8], 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* $\mathbf{v}, \alpha$,

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

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,

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

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

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

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

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

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

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

■

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

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

. We have

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

Proof We have

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

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

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

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

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

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

$$\bar{\mathbf{v}}_t - \hat{\mathbf{v}}_t = \eta\left(\nabla_{\mathbf{v}} f^s(\mathbf{v}_{t-1}^k, y_{t-1}^k) - \frac{1}{K}\sum_{k=1}^{K}\nabla_{\mathbf{v}} f_k^s(\mathbf{v}_{t-1}^k, y_{t-1}^k; z_{t-1}^k)\right) \tag{22}$$

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

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

Adding this with (23), we get

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

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 [8].

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*

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

Proof In this proof, we introduce a couple of new notations to make the proof brief: $F_{k,t}^s=F_{k,t}^s(\mathbf{v}_t^k,\alpha_t^k;z_t^k)$ and $f_{k,t}^s=f_{k,t}^s(\mathbf{v}_t^k,\alpha_t^k)$. Similar bounds for minimization problems have been analyzed in [46, 37].

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,

$$\mathbf{v}_t^k=\bar{\mathbf{v}}_{t_0}-\eta\sum_{\tau=t_0}^{t-1}\nabla_{\mathbf{v}}F_{k,\tau}^s. \tag{30}$$

Taking average over all K machines,

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

Therefore,

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

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

$$\begin{aligned}
&\frac{2\eta^2}{K}\sum_{k=1}^{K}\left[\left\|\sum_{\tau=t_0}^{t-1}\left[\left[\nabla_{\mathbf{v}}F_{k,\tau}^s-\nabla_{\mathbf{v}}f_{k,\tau}^s\right]-\frac{1}{K}\sum_{j=1}^{K}\left[\nabla_{\mathbf{v}}F_{j,\tau}^s-\nabla_{\mathbf{v}}f_{j,\tau}^s\right]\right]\right\|^2\right]\\
&\overset{(a)}{\leq}\frac{2\eta^2}{K}\sum_{k=1}^{K}\left[\left\|\sum_{\tau=t_0}^{t-1}\left[\nabla_{\mathbf{v}}F_{k,\tau}^s-\nabla_{\mathbf{v}}f_{k,\tau}^s\right]\right\|^2\right]\\
&\overset{(b)}{=}\frac{2\eta^2}{K}\sum_{k=1}^{K}\sum_{\tau=t_0}^{t-1}\left[\left\|\left[\nabla_{\mathbf{v}}F_{k,\tau}^s-\nabla_{\mathbf{v}}f_{k,\tau}^s\right]\right\|^2\right]\leq2\eta^2I\sigma^2,
\end{aligned} \tag{33}$$

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 $a_k=\sum\limits_{\tau=t_0}^{t-1}[\nabla F_{k,\tau}^s-\nabla_{\mathbf{v}}f_{k,\tau}]$; (b) follows because $\mathbb{E}_{k,\tau-1}[\nabla_{\mathbf{v}}F_{k,\tau}^s-\nabla_{\mathbf{v}}f_{k,\tau}^s]=0$.

Second, we have

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

where

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_{\mathbf{v}} f_{k,\tau}^s-\frac{1}{K}\sum_{j=1}^{K}\nabla_{\mathbf{v}} f_{j,\tau}^s\right\|^2\\
&=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla_{\mathbf{v}} f_{k,\tau}^s-\nabla_{\mathbf{v}} f_k^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)+\nabla_{\mathbf{v}} f_k^s(\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_{j,\tau}^s\right\|^2\\
&\le\frac{1}{K}\sum_{k=1}^{K}\left[3\mathbb{E}\|\nabla_{\mathbf{v}} f_{k,\tau}^s-\nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2+3\mathbb{E}\|\nabla_{\mathbf{v}} f_k^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)-\nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2\right]\\
&\quad+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_{j,\tau}^s\right\|^2\\
&=\frac{1}{K}\sum_{k=1}^{K}\left[3\mathbb{E}\|\nabla_{\mathbf{v}} f_{k,\tau}^s-\nabla_{\mathbf{v}} f_k^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2+3\mathbb{E}\|\nabla_{\mathbf{v}} f_k^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)-\nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2\right]\\
&\quad+3\mathbb{E}\left\|\frac{1}{K}\sum_{j=1}^{K}[\nabla_{\mathbf{v}} f_j^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)-\nabla_{\mathbf{v}} f_{j,\tau}^s]\right\|^2\Big]\\
&\le\frac{1}{K}\sum_{k=1}^{K}\left[3\mathbb{E}\|\nabla_{\mathbf{v}} f_{k,\tau}^s-\nabla_{\mathbf{v}} f_k(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2+3\mathbb{E}\|\nabla_{\mathbf{v}} f_k^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)-\nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2\right]\\
&\quad+3\frac{1}{K}\sum_{j=1}^{K}\mathbb{E}\left\|[\nabla_{\mathbf{v}} f_j^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)-\nabla_{\mathbf{v}} f_{j,\tau}^s]\right\|^2\Big]\\
&\overset{(a)}{\le}\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_k^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)-\nabla_{\mathbf{v}} f^s(\bar{\mathbf{v}}_\tau,\bar{\alpha}_\tau)\|^2\\
&\le\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} \qquad (35)$$

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

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

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

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

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

Similarly for α side, we have

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

Summing up the above two inequalities,

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

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

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

B.2 Proof of Lemma 1

Proof Note $\mathbb{E}\langle \frac{1}{K} \sum_{k=1}^{K} [\nabla_{\mathbf{v}} f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - \nabla_{\mathbf{v}} F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)], \hat{\mathbf{v}}_t - \tilde{\mathbf{v}}_{t-1} \rangle = 0$ and $\mathbb{E}\left\langle -\frac{1}{K} \sum_{k=1}^{K} [\nabla_\alpha f_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k) - F_k(\mathbf{v}_{t-1}^k, \alpha_{t-1}^k; z_{t-1}^k)], \tilde{\alpha}_{t-1} - \hat{\alpha}_t \right\rangle = 0$. And then plugging

Lemma 6 and Lemma 7 into Lemma 5, and taking expectation, we get

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

Since $\eta \le \min(\frac{1}{3\ell+3\ell^2/\mu_2}, \frac{1}{4\ell})$, 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,

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

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

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

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

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

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

Thus, we have

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

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

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

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 $\frac{1}{\eta_sT_s}\le\frac{\min\{\ell,\mu_2\}}{212}$. Set I_s such that $\left(\frac{3\ell^2}{2\mu_2}+\frac{3\ell}{2}\right)(12\eta_s^2I_s+36\eta^2I_s^2D^2)\le\frac{\eta_s\sigma^2}{K}$, where the specific choice of I_s will be made later. Applying Lemma 1 with $\hat{\mathbf{v}}_s(\alpha_s)=\arg\min_{\mathbf{v}'}f^s(\mathbf{v}',\alpha_s)$ and $\hat{\alpha}_s(\mathbf{v}_s)=\arg\max_{\alpha'}f^s(\mathbf{v}_s,\alpha')$, we have

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

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 [31], we have

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

Applying Lemma 4 to (48), we have

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

Combining this with (107), rearranging the terms, and defining a constant $c = 4\ell + \frac{248}{53}\hat{L} \in O(L+\ell)$, we get

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

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

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

Then, we have

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

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

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

Using this inequality recursively, it yields

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

By definition,

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

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

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

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 $\left(\frac{3\ell^2}{2\mu_2}+\frac{3\ell}{2}\right)(12\eta_s^2I_s + 36\eta^2I_s^2D^2) \leq \frac{\eta_s\sigma^2}{K}$ 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

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

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

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

Thus, the communication complexity can be bounded by

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

■

Appendix 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 [22]. 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* $\widetilde{O}\left(\max\left(\frac{\Delta_0}{\mu\epsilon\eta_0 K}, \frac{\hat{L}}{\mu^2 K\epsilon}\right)\right)$ *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 $\widetilde{O}\left(\frac{1}{\mu\epsilon} + \frac{1}{\mu^2 KM\epsilon}\right)$. Thus, the sample complexity of each machines is $\widetilde{O}\left(\frac{M}{\mu\epsilon} + \frac{1}{\mu^2 K\epsilon}\right)$.

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

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

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 $\widetilde{O}\left(\frac{M}{\mu}\right)$.

■

Appendix 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* $\mathbf{v}, \alpha$,

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

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

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

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

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

Adding (66) and (69), we get

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

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

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

■

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

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

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

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

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

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

Adding this with (77), we get

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

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

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

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

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

■

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,

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

Thus,

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

■

Lemma 12

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

Proof

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

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

Then,

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

For α, we have similar results, adding them together

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

Taking average over all machines,

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

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

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

Using (87), we have

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

Rearranging terms,

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

■

D.1 Main Proof of Lemma 2

Proof Plugging Lemma 10 into Lemma 9, we get

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

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,

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

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

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

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

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

where $(\mathbf{v}_{fs}^*, \alpha_{fs}^*)$ denotes a saddle point of f^s and the second inequality uses the strong convexity and strong concavity of f^s. In detail,

$$\begin{aligned} Gap_s(\mathbf{v}_r, \alpha_r) &= \max_\alpha f^s(\mathbf{v}_r, \alpha) - f^s(\mathbf{v}_{fs}^*, \alpha_{fs}^*) + f^s(\mathbf{v}_{fs}^*, \alpha_{fs}^*) - \min_{\mathbf{v}} f^s(\mathbf{v}, \alpha_r) \\ &\geq \frac{\ell}{2}\|\mathbf{v}_r - \mathbf{v}_{fs}^*\|^2 + \frac{\mu_2}{2}(\alpha_r - \alpha_{fs}^*)^2. \end{aligned} \tag{101}$$

Using Lemma 11, we have

$$\begin{aligned} \Xi_r &\leq 4\mathcal{E}_r + 16\tilde{\eta}^2[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r, \alpha_r)\|^2 + (\nabla_\alpha f^s(\mathbf{v}_r, \alpha_r))^2] + \frac{5\tilde{\eta}^2\sigma^2}{KI} \\ &\leq 4\left(\frac{\tilde{\eta}\sigma^2}{2\ell K\eta_g^2} + \tilde{\eta}\ell\Xi_{r-1} + \frac{48\tilde{\eta}^2}{\eta_g^2}[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r, \alpha_r)\|^2 + (\nabla_\alpha f^s(\mathbf{v}_r, \alpha_r))^2]\right) \\ &\quad + 16\tilde{\eta}^2[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r, \alpha_r)\|^2 + (\nabla_\alpha f^s(\mathbf{v}_r, \alpha_r))^2] + \frac{5\tilde{\eta}\sigma^2}{KI} \\ &\leq 4\tilde{\eta}\ell\Xi_{r-1} + 160\tilde{\eta}^2[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r, \alpha_r)\|^2 + (\nabla_\alpha f^s(\mathbf{v}_r, \alpha_r))^2] + \frac{5\tilde{\eta}\sigma^2}{KI}(1 + \frac{K}{\eta_g^2}) \\ &\leq \Xi_{r-1} + 160\tilde{\eta}^2[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r, \alpha_r)\|^2 + (\nabla_\alpha f^s(\mathbf{v}_r, \alpha_r))^2] + \frac{5\tilde{\eta}\sigma^2}{KI}(1 + \frac{K}{\eta_g^2}). \end{aligned} \tag{102}$$

Thus,

$$\begin{aligned} \frac{2\tilde{\eta}\ell^3}{\mu_2 R\eta_g^2}\sum_{r=1}^R \Xi_r \leq& \frac{2\tilde{\eta}\ell^3}{\mu_2 R\eta_g^2}\sum_r \Xi_{r-1} + \frac{320\tilde{\eta}^3\ell^3}{\mu_2 R\eta_g^2}\sum_{r=1}^R[\|\nabla_{\mathbf{v}} f^s(\mathbf{v}_r, \alpha_r)\|^2 + (\nabla_\alpha f^s(\mathbf{v}_r, \alpha_r))^2] \\ &+ \frac{5\tilde{\eta}\sigma^2}{KI}(1 + \frac{K}{\eta_g^2}) \\ \leq& \frac{2\tilde{\eta}\ell^3}{\mu_2 R\eta_g^2}\sum_r \Xi_{r-1} + \frac{1}{2R}\sum_r Gap_r + \frac{5\tilde{\eta}\sigma^2}{KI}(1 + \frac{K}{\eta_g^2}) \end{aligned} \tag{103}$$

Taking $A_0 = 0$,

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

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

$$\mathbb{E}[Gap_{\tilde{r}}^s] \leq \frac{2}{\tilde{\eta}R}\|\mathbf{v}_0 - \mathbf{v}\|^2 + \frac{2}{\tilde{\eta}R}(\alpha_0 - \alpha)^2 + \frac{10\tilde{\eta}\sigma^2}{KI}\left(1 + \frac{K}{\eta_g^2}\right). \tag{104}$$

■

Appendix E. Proof of Theorem 1

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

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

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

Rearranging the terms in (105) yields

$$\begin{aligned}
\phi(\mathbf{v}_s) - \phi(\mathbf{v}_{s-1}) &\leq \frac{1}{8\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 + \frac{1}{4\ell}\langle \partial\phi_s(\mathbf{v}_s), \partial\phi(\mathbf{v}_s)\rangle - \frac{3}{8\ell}\|\partial\phi(\mathbf{v}_s)\|^2 \\
&\overset{(a)}{\leq} \frac{1}{8\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 + \frac{1}{8\ell}(\|\partial\phi_s(\mathbf{v}_s)\|^2 + \|\partial\phi(\mathbf{v}_s)\|^2) - \frac{3}{8\ell}\|\phi(\mathbf{v}_s)\|^2 \\
&= \frac{1}{4\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 - \frac{1}{4\ell}\|\partial\phi(\mathbf{v}_s)\|^2 \\
&\overset{(b)}{\leq} \frac{1}{4\ell}\|\partial\phi_s(\mathbf{v}_s)\|^2 - \frac{\mu}{2\ell}(\phi(\mathbf{v}_s) - \phi(\mathbf{v}_{\phi_s}^*))
\end{aligned} \tag{106}$$

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

$$(4\ell + 2\mu)\left(\phi(\mathbf{v}_s) - \phi(\mathbf{v}_*)\right) - 4\ell(\phi(\mathbf{v}_{s-1}) - \phi(\mathbf{v}_{\phi_s}^*)) \leq \|\partial\phi_s(\mathbf{v}_s)\|^2. \tag{107}$$

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

$$\frac{\ell}{4}\|\hat{\mathbf{v}}_s(\alpha_s) - \mathbf{v}_0^s\|^2 + \frac{\mu_2}{4}(\hat{\alpha}_s(\mathbf{v}_s) - \alpha_0^s)^2 \leq \mathrm{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) + \mathrm{Gap}_s(\mathbf{v}_s, \alpha_s). \tag{108}$$

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

$$\begin{aligned}
&\mathbb{E}[\mathrm{Gap}_s(\mathbf{v}_s, \alpha_s)] \leq \frac{10\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{1}{53}\mathbb{E}\left[\frac{\ell}{4}\|\hat{\mathbf{v}}_s(\alpha_s) - \mathbf{v}_0^s\|^2 + \frac{\mu_2}{4}(\hat{\alpha}_s(\mathbf{v}_s) - \alpha_0^s)^2\right] \\
&\leq \frac{10\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{1}{53}\mathbb{E}\left[\mathrm{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) + \mathrm{Gap}_s(\mathbf{v}_s, \alpha_s)\right].
\end{aligned} \tag{109}$$

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 [31], we have

$$\begin{aligned}
&\mathbb{E}[\|\partial\phi_s(\mathbf{v}_s)\|^2] \leq 2\hat{L}\mathbb{E}(\phi_s(\mathbf{v}_s) - \min_{x\in\mathbb{R}^d}\phi_s(\mathbf{v})) \leq 2\hat{L}\mathbb{E}[\mathrm{Gap}_s(\mathbf{v}_s, \alpha_s)] \\
&= 2\hat{L}\mathbb{E}[4\mathrm{Gap}_s(\mathbf{v}_s, \alpha_s) - 3\mathrm{Gap}_s(\mathbf{v}_s, \alpha_s)] \\
&\leq 2\hat{L}\mathbb{E}\left[4\left(\frac{10\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{1}{53}\left(\mathrm{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) + \mathrm{Gap}_s(\mathbf{v}_s, \alpha_s)\right)\right) - 3\mathrm{Gap}_s(\mathbf{v}_s, \alpha_s)\right] \\
&= 2\hat{L}\mathbb{E}\left[40\frac{\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{4}{53}\mathrm{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) - \frac{155}{53}\mathrm{Gap}_s(\mathbf{v}_s, \alpha_s)\right].
\end{aligned} \tag{110}$$

Applying Lemma 4 to (110), we have

$$\begin{aligned}
\mathbb{E}[\|\partial\phi_s(\mathbf{v}_s)\|^2] \leq 2\hat{L}\mathbb{E}\Bigg[&\frac{40\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{4}{53}\mathrm{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) \\
&- \frac{155}{53}\left(\frac{3}{50}\mathrm{Gap}_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1}) + \frac{4}{5}(\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_0^s))\right)\Bigg] \\
= 2\hat{L}\mathbb{E}\Bigg[&\frac{40\tilde{\eta}\sigma^2}{KI_0 2^s} + \frac{4}{53}\mathrm{Gap}_s(\mathbf{v}_0^s, \alpha_0^s) - \frac{93}{530}\mathrm{Gap}_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1}) - \frac{124}{53}(\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_0^s))\Bigg].
\end{aligned} \tag{111}$$

Combining this with (107), rearranging the terms, and defining a constant $c = 4\ell + \frac{248}{53}\hat{L} \in O(L + \ell)$, we get

$$\begin{aligned}
&(c + 2\mu)\,\mathbb{E}[\phi(\mathbf{v}_0^{s+1}) - \phi(\mathbf{v}_*)] + \frac{93}{265}\hat{L}\mathbb{E}[\mathrm{Gap}_{s+1}(\mathbf{v}_0^{s+1}, \alpha_0^{s+1})] \\
&\leq \left(4\ell + \frac{248}{53}\hat{L}\right)\mathbb{E}[\phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_\phi^*)] + \frac{8\hat{L}}{53}\mathbb{E}[\mathrm{Gap}_s(\mathbf{v}_0^s, \alpha_0^s)] + \frac{80\hat{L}\tilde{\eta}\sigma^2}{KI_0 2^s} \\
&\leq c\mathbb{E}\left[\phi(\mathbf{v}_0^s) - \phi(\mathbf{v}_*) + \frac{8\hat{L}}{53c}\mathrm{Gap}_s(\mathbf{v}_0^s, \alpha_0^s)\right] + \frac{80\hat{L}\tilde{\eta}\sigma^2}{KI_0 2^s}.
\end{aligned} \tag{112}$$

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

$$(c + 2\mu)\frac{8\hat{L}}{53c} = \left(4\ell + \frac{248}{53}\hat{L} + 2\mu\right)\frac{8\hat{L}}{53(4\ell + \frac{248}{53}\hat{L})} \leq \frac{8\hat{L}}{53} + \frac{16\mu_1\hat{L}}{248\hat{L}} \leq \frac{93}{265}\hat{L}. \tag{113}$$

Then, we have

$$\begin{aligned}&(c+2\mu_1)\mathbb{E}\left[\phi(\mathbf{v}_0^{s+1})-\phi(\mathbf{v}_*)+\frac{8\hat{L}}{53c}\mathrm{Gap}_{s+1}(\mathbf{v}_0^{s+1},\alpha_0^{s+1})\right]\\&\leq c\mathbb{E}\left[\phi(\mathbf{v}_0^{s})-\phi(\mathbf{v}_*)+\frac{8\hat{L}}{53c}\mathrm{Gap}_{s}(\mathbf{v}_0^{s},\alpha_0^{s})\right]+\frac{80\hat{L}\tilde{\eta}\sigma^2}{KI_0 2^s}.\end{aligned}\tag{114}$$

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

$$\mathbb{E}[\Delta_{s+1}]\leq\frac{c}{c+2\mu}\mathbb{E}[\Delta_s]+\frac{80\hat{L}}{c+2\mu}\frac{\tilde{\eta}\sigma^2}{KI_0 2^s}\tag{115}$$

Using this inequality recursively, it yields

$$\begin{aligned}E[\Delta_{S+1}]&\leq\left(\frac{c}{c+2\mu}\right)^S E[\Delta_1]+\frac{80\hat{L}}{c+2\mu}\frac{\tilde{\eta}\sigma^2}{KI_0}\sum_{s=1}^{S}\left(\exp\left(-\frac{2\mu}{c+2\mu}(s-1)\right)\left(\frac{c}{c+2\mu}\right)^{S+1-s}\right)\\&\leq 2\epsilon_0\exp\left(\frac{-2\mu S}{c+2\mu}\right)+\frac{80\tilde{\eta}\hat{L}\sigma^2}{(c+2\mu)KI_0}S\exp\left(-\frac{2\mu S}{c+2\mu}\right),\end{aligned}\tag{116}$$

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

$$\begin{aligned}\Delta_1&=\phi(\mathbf{v}_0^1)-\phi(\mathbf{v}^*)+\frac{8\hat{L}}{53c}Gap_1(\mathbf{v}_0^1,\alpha_0^1)\\&=\phi(\mathbf{v}_0)-\phi(\mathbf{v}^*)+\left(f(\mathbf{v}_0,\hat{\alpha}_1(\mathbf{v}_0))+\frac{\gamma}{2}\|\mathbf{v}_0-\mathbf{v}_0\|^2-f(\hat{\mathbf{v}}_1(\alpha_0),\alpha_0)-\frac{\gamma}{2}\|\hat{\mathbf{v}}_1(\alpha_0)-\mathbf{v}_0\|^2\right)\\&\leq\epsilon_0+f(\mathbf{v}_0,\hat{\alpha}_1(\mathbf{v}_0))-f(\hat{\mathbf{v}}(\alpha_0),\alpha_0)\leq 2\epsilon_0.\end{aligned}\tag{117}$$

To make this less than ϵ, it suffices to make

$$\begin{aligned}&2\epsilon_0\exp\left(\frac{-2\mu S}{c+2\mu}\right)\leq\frac{\epsilon}{2},\\&\frac{80\tilde{\eta}\hat{L}\sigma^2}{(c+2\mu)KI_0}S\exp\left(-\frac{2\mu S}{c+2\mu}\right)\leq\frac{\epsilon}{2}.\end{aligned}\tag{118}$$

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

$$
\begin{aligned}
&\sum_{s=1}^{S} T_s = \sum_{s=1}^{S} R_s I_s \\
&= \sum_{s=1}^{S} R_s I_0 \exp(\frac{2\mu}{c+2\mu}(s-1)) = O\left(I_0 \sum_{s} \exp(\frac{2\mu}{c+2\mu}(s-1))\right) \\
&= \widetilde{O}\left(I_0 \frac{\exp(\frac{2\mu}{c+2\mu}S)}{\exp(\frac{2\mu_1}{c+2\mu})}\right) = \widetilde{O}\left(\frac{c}{\mu_2^2 \mu}\left(\frac{\epsilon_0}{\epsilon}, \frac{S\tilde{\eta}\sigma^2}{I_0 K \epsilon}\right)\right) \\
&= \widetilde{O}\left(\max(\frac{1}{\mu\epsilon}, \frac{c^2}{\mu^2}\frac{\tilde{\eta}\sigma^2}{K})\right) = \widetilde{O}\left(\max(\frac{1}{\mu\epsilon}, \frac{1}{K\mu^2\epsilon})\right),
\end{aligned}
\tag{119}
$$

which is also the sample complexity on each single machine.

■

Appendix 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.

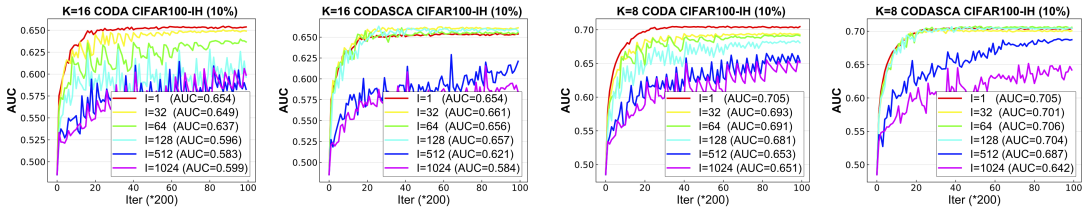

Figure 2: Imbalanced Heterogeneous CIFAR100 with imratio = 10% and K=16,8 on Densenet121.

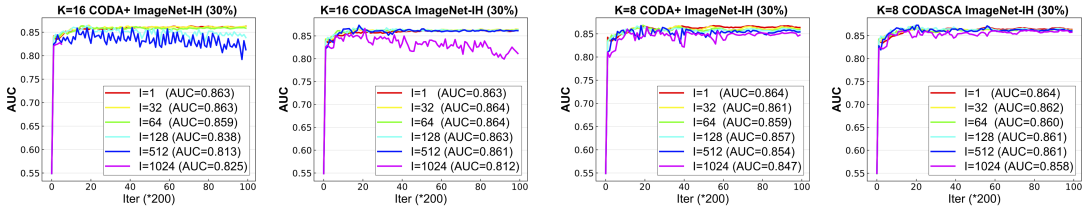

Figure 3: Imbalanced Heterogeneous ImageNet with imratio = 30% and K=16,8 on Densenet121.

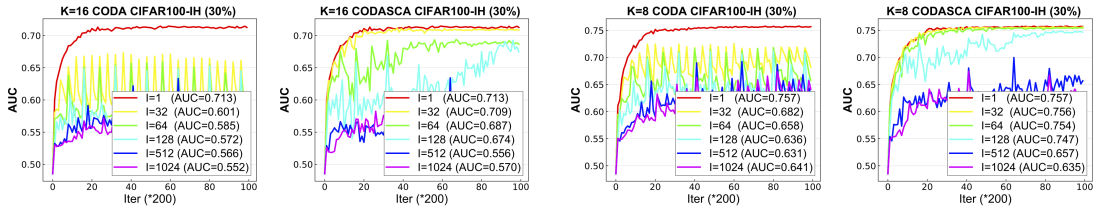

Figure 4: Imbalanced Heterogeneous CIFAR100 with imratio = 30% and K=16,8 on Densenet121.

Appendix G. Descriptions of Datasets

Table 6: Statistics of Medical Chest X-ray Datasets. The numbers for each disease denote 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