Distributed Distributionally Robust Optimization with Non-Convex Objectives

Yang Jiao
Tongji University
yangjiao@tongji.edu.cn

Kai Yang*
Tongji University
kaiyang@tongji.edu.cn

Dongjin Song
University of Connecticut
dongjin.song@uconn.edu

Abstract

Distributionally Robust Optimization (DRO), which aims to find an optimal decision that minimizes the worst case cost over the ambiguity set of probability distribution, has been widely applied in diverse applications, *e.g.*, network behavior analysis, risk management, *etc.* However, existing DRO techniques face three key challenges: 1) how to deal with the asynchronous updating in a distributed environment; 2) how to leverage the prior distribution effectively; 3) how to properly adjust the degree of robustness according to different scenarios. To this end, we propose an asynchronous distributed algorithm, named **A**synchronous **S**ingle-loo**P** alternat**I**ve g**R**adient proj**E**ction (ASPIRE) algorithm with the it**E**rative **A**ctive **SE**t method (EASE) to tackle the distributed distributionally robust optimization (DDRO) problem. Furthermore, a new uncertainty set, *i.e.*, constrained D-norm uncertainty set, is developed to effectively leverage the prior distribution and flexibly control the degree of robustness. Finally, our theoretical analysis elucidates that the proposed algorithm is guaranteed to converge and the iteration complexity is also analyzed. Extensive empirical studies on real-world datasets demonstrate that the proposed method can not only achieve fast convergence, and remain robust against data heterogeneity as well as malicious attacks, but also tradeoff robustness with performance.

1 Introduction

The past decade has witnessed the proliferation of smartphones and Internet of Things (IoT) devices, which generate a plethora of data everyday. Centralized machine learning requires gathering the data to a particular server to train models which incurs high communication overhead [46] and suffers privacy risks [43]. As a remedy, distributed machine learning methods have been proposed. Considering a distributed system composed of N workers (devices), we denote the dataset of these workers as $\{D_1, \cdots, D_N\}$. For the j^{th} $(1 \leq j \leq N)$ worker, the labeled dataset is given as $D_j = \{\mathbf{x}_j^i, y_j^i\}$, where $\mathbf{x}_j^i \in \mathbb{R}^d$ and $y_j^i \in \{1, \cdots, c\}$ denote the i^{th} data sample and the corresponding label, respectively. The distributed learning tasks can be formulated as the following optimization problem,

$$\min_{\boldsymbol{w} \in \boldsymbol{\mathcal{W}}} F(\boldsymbol{w}) \quad \text{with} \quad F(\boldsymbol{w}) := \sum\nolimits_j f_j(\boldsymbol{w}), \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^p$ is the model parameter to be learned and $\boldsymbol{\mathcal{W}} \subseteq \mathbb{R}^p$ is a nonempty closed convex set, $f_j(\cdot)$ is the empirical risk over the j^{th} worker involving only the local data:

$$f_j(\boldsymbol{w}) = \sum\nolimits_{i:\mathbf{x}_j^i \in D_j} \frac{1}{|D_j|} \mathcal{L}_j(\mathbf{x}_j^i, y_j^i; \boldsymbol{w}), \tag{2}$$

*Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

where $\mathcal{L}_j$ is the local objective function over the j^{th} worker. Problem in Eq. (1) arises in numerous areas, such as distributed signal processing [19], multi-agent optimization [36], *etc*. However, such problem does not consider the data heterogeneity [57, 40, 39, 30] among different workers (*i.e.*, data distribution of workers could be substantially different from each other [44]). Indeed, it has been shown that traditional federated approaches, such as FedAvg [33], built for independent and identically distributed (IID) data may perform poorly when applied to Non-IID data [27]. This issue can be mitigated via learning a robust model that aims to achieve uniformly good performance over all workers by solving the following distributionally robust optimization (DRO) problem in a distributed manner:

$$\min_{\boldsymbol{w}\in\boldsymbol{\mathcal{W}}} \max_{\mathbf{p}\in\boldsymbol{\Omega}\subseteq\Delta_N} F(\boldsymbol{w},\mathbf{p}) := \sum\nolimits_j p_j f_j(\boldsymbol{w}), \tag{3}$$

where $\mathbf{p} = [p_1, \cdots, p_N] \in \mathbb{R}^N$ is the adversarial distribution in N workers, the j^{th} entry in this vector, *i.e.*, p_j represents the adversarial distribution value for the j^{th} worker. $\Delta_N = \{\mathbf{p} \in \mathbb{R}_+^N : \mathbf{1}^\top \mathbf{p} = 1\}$ and $\boldsymbol{\Omega}$ is a subset of Δ_N. Agnostic federated learning (AFL) [35] firstly introduces the distributionally robust (agnostic) loss in federated learning and provides the convergence rate for (strongly) convex functions. However, AFL does not discuss the setting of $\boldsymbol{\Omega}$. DRFA-Prox [16] considers $\boldsymbol{\Omega} = \Delta_N$ and imposes a regularizer on adversarial distribution to leverage the prior distribution. Nevertheless, three key challenges have not yet been addressed by prior works. First, whether it is possible to construct an uncertainty framework that can not only flexibly maintain the trade-off between the model robustness and performance but also effectively leverage the prior distribution? Second, how to design asynchronous algorithms with guaranteed convergence? Compared to synchronous algorithms, the master in asynchronous algorithms can update its parameters after receiving updates from only a small subset of workers [58, 10]. Asynchronous algorithms are particularly desirable in practice since they can relax strict data dependencies and ensure convergence even in the presence of device failures [58]. Finally, whether it is possible to flexibly adjust the degree of robustness? Moreover, it is necessary to provide convergence guarantee when the objectives (*i.e.*, $f_j(\boldsymbol{w}_j), \forall j$) are non-convex.

To this end, we propose ASPIRE-EASE to effectively address the aforementioned challenges. Firstly, different from existing works, the prior distribution is incorporated within the constraint in our formulation, which can not only leverage the prior distribution more effectively but also achieve guaranteed feasibility for any adversarial distribution within the uncertainty set. The prior distribution can be obtained from side information or uniform distribution [41], which is necessary to construct the uncertainty (ambiguity) set and obtain a more robust model [16]. Specifically, we formulate the prior distribution informed distributionally robust optimization (PD-DRO) problem as:

$$\begin{aligned} &\min_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}},\{\boldsymbol{w}_j\in\boldsymbol{\mathcal{W}}\}} \max_{\mathbf{p}\in\boldsymbol{\mathcal{P}}} \sum\nolimits_j p_j f_j(\boldsymbol{w}_j) \\ &\quad\text{s.t.} \quad \boldsymbol{z} = \boldsymbol{w}_j,\ j = 1,\cdots,N, \\ &\quad\text{var.} \quad \boldsymbol{z}, \boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_N, \end{aligned} \tag{4}$$

where $\boldsymbol{z} \in \mathbb{R}^p$ is the global consensus variable, $\boldsymbol{w}_j \in \mathbb{R}^p$ is the local variable (local model parameter) of j^{th} worker and $\boldsymbol{\mathcal{Z}} \subseteq \mathbb{R}^p$ is a nonempty closed convex set. $\boldsymbol{\mathcal{P}} \subseteq \mathbb{R}_+^N$ is the uncertainty (ambiguity) set of adversarial distribution $\mathbf{p}$, which is set based on the prior distribution. To solve the PD-DRO problem in an asynchronous distributed manner, we first propose **A**synchronous **S**ingle-loo**P** alternat**I**ve g**R**adient proj**E**ction (ASPIRE), which employs *simple* gradient projection steps for the update of primal and dual variables at every iteration, thus is computationally *efficient*. Next, the it**E**rative **A**ctive **SE**t method (EASE) is employed to replace the traditional cutting plane method to improve the computational efficiency and speed up the convergence. We further provide the convergence guarantee for the proposed algorithm. Furthermore, a new uncertainty set, *i.e.*, constrained D-norm (CD-norm), is proposed in this paper and its advantages include: 1) it can flexibly control the degree of robustness; 2) the resulting subproblem is computationally simple; 3) it can effectively leverage the prior distribution and flexibly set the bounds for every p_j.

Contributions. Our contributions can be summarized as follows:

1. We formulate a PD-DRO problem with CD-norm uncertainty set. PD-DRO incorporates the prior distribution as constraints which can leverage prior distribution more effectively and guarantee robustness. In addition, CD-norm is developed to model the ambiguity set around the prior distribution and it provides a flexible way to control the trade-off between model robustness and performance.

2. We develop a *single-loop asynchronous* algorithm, namely ASPIRE-EASE, to optimize PD-DRO in an asynchronous distributed manner. ASPIRE employs simple gradient projection steps to update the variables at every iteration, which is computationally efficient. And EASE is proposed to replace cutting plane method to enhance the computational efficiency and speed up the convergence. We demonstrate that even if the objectives $f_j(w_j), \forall j$ are non-convex, the proposed algorithm is guaranteed to converge. We also theoretically derive the iteration complexity of ASPIRE-EASE.

3. Extensive empirical studies on four different real world datasets demonstrate the superior performance of the proposed algorithm. It is seen that ASPIRE-EASE can not only ensure the model's robustness against data heterogeneity but also mitigate malicious attacks.

2 Preliminaries

2.1 Distributionally Robust Optimization

Optimization problems often contain uncertain parameters. A small perturbation of the parameters could render the optimal solution of the original optimization problem infeasible or completely meaningless [5]. Distributionally robust optimization (DRO) [28, 17, 7] assumes that the probability distributions of uncertain parameters are unknown but remain in an ambiguity (uncertainty) set and aims to find a decision that minimizes the worst case expected cost over the ambiguity set, whose general form can be expressed as,

$$\min_{x \in X} \max_{P \in \mathcal{P}} \mathbb{E}_P[r(x; \xi)], \tag{5}$$

where $x \in X$ represents the decision variable, $\mathcal{P}$ is the ambiguity set of probability distributions P of uncertain parameters ξ. Existing methods for solving DRO can be broadly grouped into two widely-used categories [42]: 1) Dual methods [15, 50, 18] reformulate the primal DRO problems as deterministic optimization problems through duality theory. Ben-Tal et al. [2] reformulate the robust linear optimization (RLO) problem with an ellipsoidal uncertainty set as a second-order cone optimization problem (SOCP). 2) Cutting plane methods [34, 6] (also called adversarial approaches [21]) continuously solve an approximate problem with a finite number of constraints of the primal DRO problem, and subsequently check whether new constraints are needed to refine the feasible set. Recently, several new methods [41, 29, 23] have been developed to solve DRO, which need to solve the inner maximization problem at every iteration.

2.2 Cutting Plane Method for PD-DRO

In this section, we introduce the cutting plane method for PD-DRO in Eq. (4). We first reformulate PD-DRO by introducing an additional variable $h \in H$ ($H \subseteq R^1$ is a nonempty closed convex set) and protection function $g(\{w_j\})$ [55]. Introducing additional variable h is an epigraph reformulation [3, 56]. In this case, Eq. (4) can be reformulated as the form with uncertainty in the constraints:

$$\begin{aligned} &\min_{z \in Z, \{w_j \in W\}, h \in H} \quad h \\ &\text{s.t.} \quad \sum_j \bar{p}_j f_j(w_j) + g(\{w_j\}) - h \le 0, \\ &\qquad z = w_j, \; j = 1, \cdots, N, \\ &\text{var.} \quad z, w_1, w_2, \cdots, w_N, h, \end{aligned} \tag{6}$$

where $\bar{p}_j$ is the nominal value of the adversarial distribution for every worker and $g(\{w_j\}) = \max_{p \in \mathcal{P}} \sum_j (p_j - \bar{p}_j) f_j(w_j)$ is the protection function. Eq. (6) is a semi-infinite program (SIP) which contains infinite constraints and cannot be solved directly [42]. Denoting the set of cutting plane parameters in $(t+1)^{th}$ iteration as $\mathcal{A}^t \subset R^N$, the following function is used to approximate $g(\{w_j\})$:

$$\bar{g}(\{w_j\}) = \max_{a_l \in \mathcal{A}^t} a_l^\top f(w) = \max_{a_l \in \mathcal{A}^t} \sum_j a_{l,j} f_j(w_j), \tag{7}$$

where $a_l = [a_{l,1}, \cdots, a_{l,N}] \in R^N$ denotes the parameters of l^{th} cutting plane in $\mathcal{A}^t$ and $f(w) = [f_1(w_1), \cdots, f_N(w_N)] \in R^N$. Substituting the protection function $g(\{w_j\})$ with $\bar{g}(\{w_j\})$, we can obtain the following approximate problem:

$$\begin{aligned} &\min_{z \in Z, \{w_j \in W\}, h \in H} \quad h \\ &\text{s.t.} \quad \sum_j (\bar{p}_j + a_{l,j}) f_j(w_j) - h \le 0, \forall a_l \in \mathcal{A}^t, \\ &\qquad z = w_j, \; j = 1, \cdots, N, \\ &\text{var.} \quad z, w_1, w_2, \cdots, w_N, h. \end{aligned} \tag{8}$$

3 ASPIRE

Distributed optimization is an attractive approach for large-scale learning tasks [54, 8] since it does not require data aggregation, which protects data privacy while also reducing bandwidth requirements [45]. When the neural network models (i.e., $f_j(w_j), \forall j$ are non-convex functions) are used, solving problem in Eq. (8) in a distributed manner facing two challenges: 1) Computing the optimal solution to a non-convex subproblem requires a large number of iterations and therefore is highly computationally intensive if not impossible. Thus, the traditional Alternating Direction Method of Multipliers (ADMM) is ineffective. 2) The communication delays of workers may differ significantly [11], thus, asynchronous algorithms are strongly preferred.

To this end, we propose the Asynchronous Single-looP alternatIve gRadient projEction (ASPIRE). The advantages of the proposed algorithm include: 1) ASPIRE uses simple gradient projection steps to update variables in each iteration and therefore it is computationally more efficient than the traditional ADMM method, which seeks to find the optimal solution in non-convex (for $w_j, \forall j$) and convex (for z and h) optimization subproblems every iteration, 2) the proposed asynchronous algorithm does not need strict synchronization among different workers. Therefore, ASPIRE remains resilient against communication delays and potential hardware failures from workers. Details of the algorithm are given below. Firstly, we define the node as master which is responsible for updating the global variable z, and we define the node which is responsible for updating the local variable w_j as worker j. In each iteration, the master updates its variables once it receives updates from at least S workers, i.e., active workers, satisfying $1 \le S \le N$. Q^{t+1} denotes the index subset of workers from which the master receives updates during $(t+1)^{th}$ iteration. We also assume the master will receive updated variables from every worker at least once for each τ iterations. The augmented Lagrangian function of Eq. (8) can be written as:

$$L_p = h + \sum_l \lambda_l \Big(\sum_j (\bar{p} + a_{l,j}) f_j(w_j) - h \Big) + \sum_j \phi_j^\top (z - w_j) + \sum_j \frac{\kappa_1}{2} \| z - w_j \|^2, \tag{9}$$

where $L_p = L_p(\{w_j\}, z, h, \{\lambda_l\}, \{\phi_j\})$, $\lambda_l \in \Lambda, \forall l$ and $\phi_j \in \Phi, \forall j$ represent the dual variables of inequality and equality constraints in Eq. (8), respectively. $\Lambda \subseteq R^1$ and $\Phi \subseteq R^p$ are nonempty closed convex sets, constant $\kappa_1 > 0$ is a penalty parameter. Note that Eq. (9) does not consider the second-order penalty term for inequality constraint since it will invalidate the distributed optimization. Following [52], the regularized version of Eq. (9) is employed to update all variables as follows,

$$\widetilde{L}_p(\{w_j\}, z, h, \{\lambda_l\}, \{\phi_j\}) = L_p - \sum_l \frac{c_1^t}{2} \|\lambda_l\|^2 - \sum_j \frac{c_2^t}{2} \|\phi_j\|^2, \tag{10}$$

where c_1^t and c_2^t denote the regularization terms in $(t+1)^{th}$ iteration. To avoid enumerating the whole dataset, the mini-batch loss could be used. A batch of instances with size m can be randomly sampled from each worker during each iteration. The loss function of these instances from j^{th} worker is given by $\hat{f}_j(w_j) = \sum_{i=1}^{m} \frac{1}{m} L_j(x_j^i, y_j^i, w_j)$. It is evident that $E[\hat{f}_j(w_j)] = f_j(w_j)$ and $E[\nabla \hat{f}_j(w_j)] = \nabla f_j(w_j)$. In $(t+1)^{th}$ master iteration, the proposed algorithm proceeds as follows.

1) Active workers update the local variables w_j as follows,

$$w_j^{t+1} = \begin{cases} P_W\big(w_j^t - \eta_w^{\tilde{t}_j} \nabla_{w_j} \widetilde{L}_p(\{w_j^{\tilde{t}_j}\}, z^{\tilde{t}_j}, h^{\tilde{t}_j}, \{\lambda_l^{\tilde{t}_j}\}, \{\phi_j^{\tilde{t}_j}\})\big), \forall j \in Q^{t+1}, \\ w_j^t, \forall j \notin Q^{t+1}, \end{cases} \tag{11}$$

where $\tilde{t}_j$ is the last iteration during which worker j was active. It is seen that $\forall j \in Q^{t+1}, w_j^t = w_j^{\tilde{t}_j}$ and $\phi_j^t = \phi_j^{\tilde{t}_j}$. $\eta_w^{\tilde{t}_j}$ represents the step-size and let $\tilde{\eta}_w^t = \eta_w^t$ when $t < T_1$ and $\tilde{\eta}_w^t = \underline{\eta_w}$ when $t \ge T_1$, where η_w^t and constant $\underline{\eta_w}$ will be introduced below. P_W represents the projection onto the closed convex set W and we set $W = \{w_j \mid \|w_j\|_1 \le \alpha_1\}$, α_1 is a positive constant. And then, the active workers ($j \in Q^{t+1}$) transmit their local model parameters w_j^{t+1} and loss $f_j(w_j)$ to the master.

2) After receiving the updates from active workers, the master updates the global consensus variable z, additional variable h and dual variables λ_l as follows,

$$z^{t+1} = P_Z\big(z^t - \eta_z^t \nabla_z \widetilde{L}_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\})\big); \tag{12}$$

$$h^{t+1} = P_H\big(h^t - \eta_h^t \nabla_h \widetilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^t, \{\lambda_l^t\}, \{\phi_j^t\})\big); \tag{13}$$

$$\lambda_l^{t+1} = P_{\Lambda}\left(\lambda_l^t + \eta_1 \nabla_{\lambda_l} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\lambda_l^t\}, \{\varphi_j^t\})\right),\ l = 1, \cdots, |A^t|, \tag{14}$$

where η_z^t, η_h^t and η_1 represent the step-sizes. P_Z, P_H and P_Λ respectively represent the projection onto the closed convex sets Z, H and Λ. We set $Z = \{z \mid \|z\|_1 \le \alpha_1\}$, $H = \{h \mid 0 \le h \le \alpha_2\}$ and $\Lambda = \{\lambda_l \mid 0 \le \lambda_l \le \alpha_3\}$, where α_2 and α_3 are positive constants. $|A^t|$ denotes the number of cutting planes. Then, master broadcasts z^{t+1}, h^{t+1}, $\{\lambda_l^{t+1}\}$ to the active workers.

3) Active workers update the local dual variables φ_j as follows,

$$\varphi_j^{t+1} = \begin{cases} P_{\Phi}\left(\varphi_j^t + \eta_2 \nabla_{\varphi_j} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\lambda_l^{t+1}\}, \{\varphi_j^t\})\right), \forall j \in Q^{t+1}, \\ \varphi_j^t, \forall j \notin Q^{t+1}, \end{cases} \tag{15}$$

where η_2 represents the step-size and P_Φ represents the projection onto the closed convex set Φ and we set $\Phi = \{\varphi_j \mid \|\varphi_j\|_1 \le \alpha_4\}$, α_4 is a positive constant. And master can also obtain $\{\varphi_j^{t+1}\}$ according to Eq. (15). It is seen that the projection operation in each step is computationally simple since the closed convex sets have simple structures [4].

4 Iterative Active Set Method

Cutting plane methods may give rise to numerous linear constraints and lots of extra message passing [55]. Moreover, more iterations are required to obtain the ϵ-stationary point when the size of a set containing cutting planes increases (which corresponds to a larger M), which can be seen in Theorem 1. To improve the computational efficiency and speed up the convergence, we consider removing the inactive cutting planes. The proposed Efficient Active SEt method (EASE) can be divided into the two steps: during T_1 iterations, 1) solving the cutting plane generation subproblem to generate cutting plane, and 2) removing the inactive cutting plane every k iterations, where $k > 0$ is a pre-set constant and can be controlled flexibly.

The cutting planes are generated according to the uncertainty set. For example, if we employ ellipsoid uncertainty set, the cutting plane is generated via solving a SOCP. In this paper, we propose CD-norm uncertainty set, which can be expressed as follows,

$$P = \{p : -\hat{p}_j \le p_j - q_j \le \hat{p}_j;\ \sum_j \left|\frac{p_j - q_j}{\hat{p}_j}\right| \le \Gamma;\ 1^\top p = 1\}, \tag{16}$$

where $\Gamma \in R^1$ can flexibly control the level of robustness, $q = [q_1, \cdots, q_N] \in R^N$ represents the prior distribution, $-\hat{p}_j$ and $\hat{p}_j$ ($\hat{p}_j \ge 0$) represent the lower and upper bounds for $p_j - q_j$, respectively. The setting of q and $\hat{p}_j, \forall j$ are based on the prior knowledge. D-norm is a classical uncertainty set (which is also called as budget uncertainty set) [5]. We call Eq. (16) CD-norm uncertainty set since p is a probability vector so all the entries of this vector are non-negative and add up to exactly one, i.e., $1^\top p = 1$. Due to the special structure of CD-norm, the cutting plane generation subproblem is easy to solve and the level of robustness in terms of the outage probability, i.e., probabilistic bounds of the violations of constraints can be flexibly adjusted via a single parameter Γ. We claim that l_1-norm (or twice total variation distance) uncertainty set is closely related to CD-norm uncertainty set. Nevertheless, there are two differences: 1) CD-norm uncertainty set could be regarded as a weighted l_1-norm with additional constraints. 2) CD-norm uncertainty set can flexibly set the lower and upper bounds for every p_j (i.e., $q_j - \hat{p}_j \le p_j \le p_j + \hat{p}_j$), while $0 \le p_j \le 1, \forall j$ in l_1-norm uncertainty set. Based on the CD-norm uncertainty set, the cutting plane can be derived as follows,

1) Solve the following problem,

$$\begin{aligned} p^{t+1} = \arg\max_{p_1, \cdots, p_N} & \sum_j (p_j - \bar{p}) f_j(w_j) \\ \text{s.t.:} & \sum_j \left|\frac{p_j - q_j}{\hat{p}_j}\right| \le \Gamma;\ -\hat{p}_j \le p_j - q_j \le \hat{p}_j, \forall j;\ \sum_j p_j = 1 \\ \text{var:} & \quad p_1, \cdots, p_N, \end{aligned} \tag{17}$$

where $p^{t+1} = [p_1^{t+1}, \cdots, p_N^{t+1}] \in R^N$. Let $a^{t+1} = p^{t+1} - \bar{p}$, where $\bar{p} = [\bar{p}, \cdots, \bar{p}] \in R^N$. This first step aims to obtain the distribution a^{t+1} by solving problem in Eq. (17). This problem can be effectively solved through combining merge sort [13] (for sorting $\hat{p}_j f_j(w_j), j = 1, \cdots, N$) with few basic arithmetic operations (for obtaining $p_j^{t+1}, j = 1, \cdots, N$). Since N is relatively large in

---

Algorithm 1 ASPIRE-EASE

Initialization: iteration $t = 0$, variables $\{w_j^0\}$, z^0, h^0, $\{\lambda_l^0\}$, $\{\theta_j^0\}$ and set A^0.
repeat
  for active worker **do**
    updates local w_j^{t+1} according to Eq. (11);
  end for
  active workers transmit local model parameters and loss to *master*;
  master receives updates from *active workers* **do**
    updates z^{t+1}, h^{t+1}, $\{\lambda_l^{t+1}\}$, $\{\theta_j^{t+1}\}$ in master according to Eq. (12), (13), (14), (15);
  master broadcasts z^{t+1}, h^{t+1}, $\{\lambda_l^{t+1}\}$ to *active workers*;
  for active worker **do**
    updates local θ_j^{t+1} according to Eq. (15);
  end for
  if $(t + 1)$ mod $k == 0$ and $t < T_1$ **then**
    master updates A^{t+1} according to Eq. (19) and (20), and broadcast parameters to all workers;
  end if
  $t = t + 1$;
until convergence

---

distributed system, the arithmetic complexity of solving problem in Eq. (17) is dominated by merge sort, which can be regarded as $O(N \log(N))$.

2) Let $f(w) = [f_1(w_1), \cdots, f_N(w_N)] \in R^N$, check the feasibility of the following constraints:

$$a^{t+1\top} f(w) - \max_{a_l \in A^t} a_l^\top f(w). \tag{18}$$

3) If Eq. (18) is violated, a^{t+1} will be added into A^t:

$$A^{t+1} = \begin{cases} A^t \cup \{a^{t+1}\}, & \text{if Eq.(18) is violated,} \\ A^t, & \text{otherwise,} \end{cases} \tag{19}$$

when a new cutting plane is added, its corresponding dual variable $\lambda^{t+1}_{|A^t|+1} = 0$ will be generated. After the cutting plane subproblem is solved, the inactive cutting plane will be removed, that is:

$$A^{t+1} = \begin{cases} \complement_{A^{t+1}}\{a_l\}, & \text{if } \lambda_l^{t+1} = 0 \text{ and } \lambda_l^t = 0, 1 \le l \le |A^t|, \\ A^{t+1}, & \text{otherwise,} \end{cases} \tag{20}$$

where $\complement_{A^{t+1}}\{a_l\}$ is the complement of $\{a_l\}$ in A^{t+1}, and the dual variable will be removed. Then master broadcasts A^{t+1}, $\{\lambda_l^{t+1}\}$ to all workers. Details of algorithm are summarized in Algorithm 1.

5 Convergence Analysis

Definition 1 (Stationarity gap) Following [52, 32, 53], the stationarity gap of our problem at t^{th} iteration is defined as:

$$\nabla G^t = \begin{bmatrix} \left\{\frac{1}{\eta_w^t}\left(w_j^t - P_{\mathcal{W}}\left(w_j^t - \eta_w^t \nabla_{w_j} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\})\right)\right)\right\} \\ \frac{1}{\eta_z^t}\left(z^t - P_{\mathcal{Z}}\left(z^t - \eta_z^t \nabla_z L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\})\right)\right) \\ \frac{1}{\eta_h^t}\left(h^t - P_{\mathcal{H}}\left(h^t - \eta_h^t \nabla_h L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\})\right)\right) \\ \left\{\frac{1}{\eta_1}\left(\lambda_l^t - P_{\Lambda}\left(\lambda_l^t + \eta_1 \nabla_{\lambda_l} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\})\right)\right)\right\} \\ \left\{\frac{1}{\eta_2}\left(\theta_j^t - P_{\Theta}\left(\theta_j^t + \eta_2 \nabla_{\theta_j} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\})\right)\right)\right\} \end{bmatrix}, \tag{21}$$

where ∇G^t is the simplified form of $\nabla G(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\})$.

Definition 2 (ε-stationary point) $(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\})$ is an ε-stationary point ($\varepsilon \ge 0$) of a differentiable function L_p, if $||\nabla G^t|| \le \varepsilon$. $T(\varepsilon)$ is the first iteration index such that $||\nabla G^t|| \le \varepsilon$, i.e., $T(\varepsilon) = \min\{t \mid ||\nabla G^t|| \le \varepsilon\}$.

Assumption 1 (Smoothness/Gradient Lipschitz) L_p has Lipschitz continuous gradients. We assume that there exists $L > 0$ satisfying

$$||\nabla L_p(\{w_j\}, z, h, \{\lambda_l\}, \{\phi_j\}) - \nabla L_p(\{\hat{w}_j\}, \hat{z}, \hat{h}, \{\hat{\lambda}_l\}, \{\hat{\phi}_j\})|| \le L||[w_{cat} - \hat{w}_{cat}; z - \hat{z}; h - \hat{h}; \lambda_{cat} - \hat{\lambda}_{cat}; \phi_{cat} - \hat{\phi}_{cat}]||,$$

where $\in \{\{w_j\}, z, h, \{\lambda_l\}, \{\phi_j\}\}$ and $[;]$ represents the concatenation. $w_{cat} - \hat{w}_{cat} = [w_1 - \hat{w}_1; \cdots; w_N - \hat{w}_N] \in R^{pN}$, $\lambda_{cat} - \hat{\lambda}_{cat} = [\lambda_1 - \hat{\lambda}_1; \cdots; \lambda_{|A^t|} - \hat{\lambda}_{|A^t|}] \in R^{|A^t|}$, $\phi_{cat} - \hat{\phi}_{cat} = [\phi_1 - \hat{\phi}_1; \cdots; \phi_N - \hat{\phi}_N] \in R^{pN}$.

Assumption 2 (Boundedness) Before obtaining the ε-stationary point (i.e., $t \le T(\varepsilon) - 1$), we assume variables in master satisfy that $||z^{t+1} - z^t||^2 + ||h^{t+1} - h^t||^2 + \sum_l ||\lambda_l^{t+1} - \lambda_l^t||^2 \ge \vartheta$, where $\vartheta > 0$ is a relative small constant. The change of the variables in master is upper bounded within iterations:

$$||z^t - z^{t-k}||^2 \le k_1\vartheta;\ ||h^t - h^{t-k}||^2 \le k_1\vartheta;\ \sum_l ||\lambda_l^t - \lambda_l^{t-k}||^2 \le k_1\vartheta;\ \forall 1 \le k \le \tau,$$

where $k_1 > 0$ is a constant.

Setting 1 (Bounded $|A^t|$) $|A^t| \le M$, $\forall t$, i.e., an upper bound is set for the number of cutting planes.

Setting 2 (Setting of c_1^t, c_2^t) $c_1^t = \frac{1}{\eta_1 (t+1)^{\frac{1}{6}}} \ge \underline{c}_1$ and $c_2^t = \frac{1}{\eta_2 (t+1)^{\frac{1}{6}}} \ge \underline{c}_2$ are nonnegative non-increasing sequences, where $\underline{c}_1$ and $\underline{c}_2$ are positive constants and meet $M\underline{c}_1{}^2 + N\underline{c}_2{}^2 \le \frac{\varepsilon^2}{4}$.

Theorem 1 (Iteration complexity) Suppose Assumption 1 and 2 hold. We set $\eta_w^t = \eta_z^t = \eta_h^t = \frac{2}{L + \eta_1|A^t|L^2 + \eta_2 NL^2 + 8(\frac{|A^t|L^2}{\eta_1 (c_1^t)^2} + \frac{NL^2}{\eta_2 (c_2^t)^2})}$ and $\underline{\eta}_w = \frac{2}{L + \eta_1 ML^2 + \eta_2 NL^2 + 8(\frac{ML^2}{\eta_1 \underline{c}_1{}^2} + \frac{NL^2}{\eta_2 \underline{c}_2{}^2})}$. And we set constants $\eta_1 < \min\{\frac{2}{L + 2c_1^0}, \frac{1}{15k_1 NL^2}\}$ and $\eta_2 \le \frac{2}{L + 2c_2^0}$, respectively. For a given ε, we have:

$$T(\varepsilon) \sim O(\max\{(\frac{4M\sigma_1{}^2}{\eta_1{}^2} + \frac{4N\sigma_2{}^2}{\eta_2{}^2})^3\frac{1}{\varepsilon^6}, (\frac{4(d_6 + \frac{\eta_2 (N-S)L^2}{2}\sigma^2)(d + k_d(\tau - 1))d_5}{\varepsilon^2} + (T_1 + \tau)^{\frac{1}{3}})^3\}), \tag{22}$$

where σ_1, σ_2, σ, τ, k_d, d, d_5, d_6 and T_1 are constants. The detailed proof is given in Appendix A.

There exists a wide array of works regarding the convergence analysis of various algorithms for nonconvex/convex optimization problems involved in machine learning [25, 53]. Our analysis, however, differs from existing works in two aspects. First, we solve the non-convex PD-DRO in an asynchronous distributed manner. To our best knowledge, there are few works focusing on solving the DRO in a distributed manner. Compared to solving the non-convex PD-DRO in a centralized manner, solving it in an asynchronous distributed manner poses significant challenges in algorithm design and convergence analysis. Secondly, we do not assume the inner problem can be solved nearly optimally for each outer iteration, which is numerically difficult to achieve in practice [4]. Instead, ASPIRE-EASE is single loop and involves simple gradient projection operation at each step.

6 Experiment

In this section, we conduct experiments on four real-world datasets to assess the performance of the proposed method. Specifically, we evaluate the robustness against data heterogeneity, robustness against malicious attacks and efficiency of the proposed method. Ablation study is also carried out to demonstrate the excellent performance of ASPIRE-EASE.

6.1 Datasets and Baseline Methods

We compare the proposed ASPIRE-EASE with baseline methods based on SHL [29], Person Activity [26], Single Chest-Mounted Accelerometer (SM-AC) [9] and Fashion MNIST [51] datasets. The baseline methods include Ind_j (learning the model from an individual worker j), Mix_{Even} (learning the model from all workers with even weights using ASPIRE), FedAvg [39], AFL [35] and DRFA-Prox [16]. The detailed descriptions of datasets and baselines are given in Appendix C.

In our empirical studies, since the downstream tasks are multi-class classification, the cross entropy loss is used on each worker (i.e., $L_j(\cdot), \forall j$). For SHL, Person Activity, and SM-AC datasets, we adopt the deep multilayer perceptron [49] as the base model. And we use the same logistic regression model as in [35, 16] for Fashion MNIST dataset. The base models are trained with SGD. More details are given in Appendix C. Following related works in this direction [41, 35, 16], worst case performance are reported for the comparison of robustness. Specifically, we use Acc_w and $Loss_w$ to represent the worst case test accuracy and training loss (i.e., the test accuracy and training loss on the worker with worst performance), respectively. We also report the standard deviation of

Table 1: Performance comparisons based on Acc_w (%) ↑, $Loss_w$ ↓ and Std ↓ (↑ and ↓ respectively denote higher scores represent better performance and lower scores represent better performance). The boldfaced digits represent the best results, "–" represents not available.

Model	SHL			Person Activity			SC-MA			Fashion MNIST		
	Acc_w ↑	$Loss_w$ ↓	Std ↓	Acc_w ↑	$Loss_w$ ↓	Std ↓	Acc_w ↑	$Loss_w$ ↓	Std ↓	Acc_w ↑	$Loss_w$ ↓	Std ↓
$\max\{Ind_j\}$	19.06±0.65		29.1	49.38±0.08		8.32	22.56±0.78		17.5			
Mix_{Even}	69.87±3.10	0.806±0.018	4.81	56.31±0.69	1.165±0.017	3.00	49.81±0.21	1.424±0.024	6.99	66.80±0.18	0.784±0.003	10.1
FedAvg [33]	69.96±3.07	0.802±0.023	5.21	56.28±0.63	1.154±0.019	3.13	49.53±0.96	1.441±0.015	7.17	66.58±0.39	0.781±0.002	10.2
AFL [35]	78.11±1.99	0.582±0.021	1.87	58.39±0.37	1.081±0.014	0.99	54.56±0.79	1.172±0.018	3.50	77.32±0.15	0.703±0.001	1.86
DRFA-Prox [16]	78.34±1.46	0.532±0.034	1.85	58.62±0.16	1.096±0.037	1.26	54.61±0.76	1.151±0.039	4.69	77.95±0.51	0.702±0.007	1.34
ASPIRE-EASE	79.16±1.13	0.515±0.019	1.02	59.43±0.44	1.053±0.010	0.82	56.31±0.29	1.127±0.021	3.16	78.82±0.07	0.696±0.004	1.01
ASPIRE-EASE_{per}	78.94±1.27	0.521±0.023	1.36	59.54±0.21	1.051±0.016	0.79	56.71±0.16	1.119±0.023	2.48	78.73±0.06	0.698±0.006	1.09

$[Acc_1; \ldots; Acc_N]$ (the test accuracy on every worker). In the experiments, S is set as 1, that means the master will make an update once it receives a message. Each experiment is repeated 10 times, both mean and standard deviations are reported. We implement our model with PyTorch and conduct all the experiments on a server with two TITAN V GPUs.

6.2 Results

Robustness against Data Heterogeneity. We first assess the robustness of the proposed ASPIRE-EASE by comparing it with baseline methods when data are heterogeneously distributed across different workers. Specifically, we compare the Acc_w, $Loss_w$ and Std of different methods on all datasets. The performance comparison results are shown in Table 1. In this table, we can observe that $\max\{Ind_j\}$, which represents the best performance of individual training over all workers, exhibits the worst robustness on SHL, Person Activity, and SC-MA. This is because individual training ($\max\{Ind_j\}$) only learns from the data in its local worker and cannot generalize to other workers due to different data distributions. Note that $\max\{Ind_j\}$ is unavailable for Fashion MNIST since each worker only contains one class of data and cross entropy loss cannot be used in this case. $\max\{Ind_j\}$ also does not have $Loss_w$, since Ind_j is trained only on individual worker j. The FedAvg and Mix_{Even} exhibit better performance than $\max\{Ind_j\}$ since they consider the data from all workers. Nevertheless, FedAvg and Mix_{Even} only assign the fixed weight for each worker. AFL is more robust than FedAvg and Mix_{Even} since it not only utilizes the data from all workers but also considers optimizing the weight of each worker. DRFA-Prox outperforms AFL since it also considers the prior distribution and regards it as a regularizer in the objective function. Finally, we can observe that the proposed ASPIRE-EASE shows excellent robustness, which can be attributed to two factors: 1) ASPIRE-EASE considers data from all workers and can optimize the weight of each worker; 2) compared with DRFA-Prox which uses prior distribution as a regularizer, the prior distribution is incorporated within the constraint in our formulation (Eq. 4), which can be leveraged more effectively. And it is seen that ASPIRE-EASE can perform periodic communication since ASPIRE-EASE_{per}, which represents ASPIRE-EASE with periodic communication, also has excellent performance.

Within ASPIRE-EASE, the level of robustness can be controlled by adjusting η. Specially, when $\eta = 0$, we obtain a nominal optimization problem in which no adversarial distribution is considered. The size of the uncertainty set will increase with η (when $\eta \leq N$), which enhances the adversarial robustness of the model. As shown in Figure 1, the robustness of ASPIRE-EASE can be gradually enhanced when η increases. More results are available in Figure C2 of Appendix C.

Robustness against Malicious Attacks. To assess the model robustness against malicious attacks, malicious workers with backdoor attacks [1, 48], which attempt to mislead the model training process, are added to the distributed system. Following [14], we report the success attack rate of backdoor attacks for comparison. It can be calculated by checking how many instances in the backdoor dataset can be misled and categorized into the target labels. Lower success attack rates indicate more robustness against backdoor attacks. The comparison results are summarized in Table 2 and more detailed settings of backdoor attacks are available in Appendix C. In Table 2, we observe that AFL can be attacked easily since it could assign higher weights to malicious workers. Compared to AFL, FedAvg and Mix_{Even} achieve relatively lower success attack rates since they assign equal weights to the malicious workers and other workers. DRFA-Prox can achieve even lower success attack rates since it can leverage the prior distribution to assign lower weights for malicious workers. The proposed ASPIRE-EASE achieves the lowest success attack rates since it can leverage the prior distribution more effectively. Specifically, it will assign lower weights to malicious workers with tight theoretical guarantees.

(a) Person Activity (b) SC-MA

Figure 1: control the degree of robustness (worst case performance in the problem) on (a) Person Activity, (b) SC-MA datasets.

(a) Person Activity (b) SC-MA

Figure 2: Comparison of the convergence time on worst case worker on (a) Person Activity, (b) SC-MA datasets.

(a) Person Activity (b) SC-MA

Figure 3: Comparison of ASPIRE-CP and ASPIRE-EASE regarding the number of cutting planes on (a) Person Activity, (b) SC-MA datasets.

Table 2: Performance comparisons about the success attack rate(%) #. The boldfaced digits represent the best results.

Model	SHL	Person Activity	SC-MA	Fashion MNIST
Mix $_{Even}$	36.21±2.23	34.32±2.18	52.14±2.89	83.18±2.07
FedAvg [33]	38.15±3.02	33.25±2.49	55.39±3.13	82.04±1.84
AFL [35]	68.63±4.24	43.66±3.87	75.81±4.03	90.04±2.52
DRFA-Prox [16]	21.23±3.63	27.27±3.31	30.79±3.65	63.24±2.47
ASPIRE-EASE	9.17±1.65	22.36±2.33	14.51±3.21	45.10±1.64

Ef ciency. In Figure 2, we compare the convergence speed of the proposed ASPIRE-EASE with AFL and DRFA-Prox by considering different communication and computation delays for each worker. The proposed ASPIRE-EASE has two variants, ASPIRE-CP (ASPIRE with cutting plane method), ASPIRE-EASE(-)(ASPIRE-EASE without asynchronous setting). More results are available in Figure C3 of Appendix C. Based on the comparison, we can observe that the proposed ASPIRE-EASE generally converges faster than baseline methods and its two variants. This is because 1) compared with AFL, DRFA-Prox, and ASPIRE-EASE(-), ASPIRE-EASE is an asynchronous algorithm in which the master updates its parameters only after receiving the updates from active workers instead of all workers; 2) unlike DRFA-Prox, the master in ASPIRE-EASE only needs to communicate with active workers once per iteration; 3) compared with ASPIRE-CP, ASPIRE-EASE utilizes active set method instead of cutting plane method, which is more ef cient. It is seen from Figure 2 that, the convergence speed of ASPIRE-EASE mainly bene ts from the asynchronous setting.

Ablation Study. For ASPIRE, compared with cutting plane method, EASE is more ef cient since it considers removing the inactive cutting planes. To demonstrate the ef ciency of EASE, we rstly compare ASPIRE-EASE with ASPIRE-CP concerning the number of cutting planes used during the training. In Figure 3, we can observe that ASPIRE-EASE uses fewer cutting planes than ASPIRE-CP, thus is more ef cient. The convergence speed of ASPIRE-EASE and ASPIRE-CP in Figure 2 also suggests that ASPIRE-EASE converges much faster than ASPIRE-CP. More results are available in Figure C3 and C4, Appendix C.

7 Conclusion

In this paper, we present ASPIRE-EASE method to effectively solve the distributed distributionally robust optimization problem with non-convex objectives. In addition, CD-norm uncertainty set has been proposed to effectively incorporate the prior distribution into the problem formulation, which allows for exible adjustment of the degree of robustness of DRO. Theoretical analysis has also been conducted to analyze the convergence properties and the iteration complexity of ASPIRE-EASE. ASPIRE-EASE exhibits strong empirical performance on multiple real-world datasets and is effective in tackling DRO problems in a fully distributed and asynchronous manner. In the future work, more uncertainty sets could be designed for our framework and more update rule for variables in ASPIRE could be considered.

Acknowledgments and Disclosure of Funding

The work of Yang Jiao and Kai Yang was supported in part by the Fundamental Research Funds for the Central Universities of China, in part by the Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), in part by the National Natural Science Foundation of China under Grant 61771013, and in part by the Fundamental Research Funds of Shanghai Jiading District.

References

[1] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2938–2948. PMLR, 2020.

[2] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations research letters*, 25(1):1–13, 1999.

[3] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton university press, 2009.

[4] D. P. Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

[5] D. Bertsimas and M. Sim. The price of robustness. *Operations research*, 52(1):35–53, 2004.

[6] D. Bertsimas, I. Dunning, and M. Lubin. Reformulation versus cutting-planes for robust optimization. *Computational Management Science*, 13(2):195–217, 2016.

[7] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[8] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[9] P. Casale, O. Pujol, and P. Radeva. Personalization and user verification in wearable systems using biometric walking patterns. *Personal and Ubiquitous Computing*, 16(5):563–580, 2012.

[10] T.-H. Chang, M. Hong, W.-C. Liao, and X. Wang. Asynchronous distributed ADMM for large-scale optimization—Part I: Algorithm and convergence analysis. *IEEE Transactions on Signal Processing*, 64(12):3118–3130, 2016.

[11] Y. Chen, Y. Ning, M. Slawski, and H. Rangwala. Asynchronous online federated learning for edge devices with Non-IID data. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 15–24. IEEE, 2020.

[12] A. Cohen, A. Daniely, Y. Drori, T. Koren, and M. Schain. Asynchronous stochastic optimization robust to arbitrary delays. *Advances in Neural Information Processing Systems*, 34:9024–9035, 2021.

[13] R. Cole. Parallel merge sort. *SIAM Journal on Computing*, 17(4):770–785, 1988.

[14] J. Dai, C. Chen, and Y. Li. A backdoor attack against LSTM-based text classification systems. *IEEE Access*, 7:138872–138878, 2019.

[15] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[16] Y. Deng, M. M. Kamani, and M. Mahdavi. Distributionally robust federated averaging. *arXiv preprint arXiv:2102.12660*, 2021.

[17] J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[18] R. Gao and A. J. Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[19] G. Geraci, M. Wildemeersch, and T. Q. Quek. Energy efficiency of distributed signal processing in wireless networks: A cross-layer analysis. IEEE Transactions on Signal Processing, 64(4): 1034–1047, 2015.

[20] H. Gjoreski, M. Ciliberto, L. Wang, F. J. O. Morales, S. Mekki, S. Valentin, and D. Roggen. The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices. IEEE Access, 6:42592–42604, 2018.

[21] B. L. Gorissen, I. Yanıkoğlu, and D. den Hertog. A practical guide to robust optimization. Omega, 53:124–137, 2015.

[22] F. Haddadpour, M. M. Kamani, M. Mahdavi, and A. Karbasi. Learning distributionally robust models at scale via composite optimization. arXiv preprint arXiv:2203.09607, 2022.

[23] Y. Hu, X. Chen, and N. He. On the bias-variance-cost tradeoff of stochastic optimization. Advances in Neural Information Processing Systems, 34, 2021.

[24] J. Jiang, W. Zhang, J. Gu, and W. Zhu. Asynchronous decentralized online learning. Advances in Neural Information Processing Systems, 34:20185–20196, 2021.

[25] C. Jin, P. Netrapalli, and M. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In International Conference on Machine Learning, pages 4880–4889. PMLR, 2020.

[26] B. Kaluža, V. Mirchevska, E. Dovgan, M. Luštrek, and M. Gams. An agent-based approach to care in independent living. In International joint conference on ambient intelligence, pages 177–186. Springer, 2010.

[27] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh. SCAFFOLD: Stochastic Controlled Averaging for On-Device Federated Learning. 2019.

[28] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In Operations Research & Management Science in the Age of Analytics, pages 130–166. INFORMS, 2019.

[29] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. Advances in Neural Information Processing Systems, 33:8847–8860, 2020.

[30] W.-H. Liao and Y.-T. Huang. Investigation of DNN model robustness using heterogeneous datasets. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 4393–4397. IEEE, 2021.

[31] T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In International Conference on Machine Learning, pages 6083–6093. PMLR, 2020.

[32] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. IEEE Transactions on Signal Processing, 68:3676–3691, 2020.

[33] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics, pages 1273–1282. PMLR, 2017.

[34] S. Mehrotra and D. Papp. A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. SIAM Journal on Optimization, 24(4): 1670–1697, 2014.

[35] M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In International Conference on Machine Learning, pages 4615–4625. PMLR, 2019.

[36] A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. IEEE Transactions on Automatic Control, 54(1):48–61, 2009.

[37] Y. Nesterov. Introductory lectures on convex optimization: A basic course, volume 87. Springer Science & Business Media, 2003.

[38] Q. Qi, Z. Guo, Y. Xu, R. Jin, and T. Yang. An online method for a class of distributionally robust optimization with non-convex objectives. Advances in Neural Information Processing Systems, 34:10067–10080, 2021.

[39] J. Qian, X. Fafoutis, and L. K. Hansen. Towards federated learning: Robustness analytics to data heterogeneity. arXiv preprint arXiv:2002.05038, 2020.

[40] J. Qian, L. K. Hansen, X. Fafoutis, P. Tiwari, and H. M. Pandey. Robustness analytics to data heterogeneity in edge computing. Computer Communications, 164:229–239, 2020.

[41] Q. Qian, S. Zhu, J. Tang, R. Jin, B. Sun, and H. Li. Robust optimization over multiple domains. In Proceedings of the AAAI Conference on Arti cial Intelligence, volume 33, pages 4739–4746, 2019.

[42] H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. arXiv preprint arXiv:1908.05659, 2019.

[43] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini. Security, privacy and trust in Internet of Things: The road ahead. Computer networks, 76:146–164, 2015.

[44] K. Singhal, H. Sidahmed, Z. Garrett, S. Wu, J. Rush, and S. Prakash. Federated reconstruction: Partially local federated learning. Advances in Neural Information Processing Systems, 34, 2021.

[45] T. Subramanya and R. Riggio. Centralized and federated learning for predictive VNF autoscaling in multi-domain 5G networks and beyond. IEEE Transactions on Network and Service Management, 18(1):63–78, 2021.

[46] J. Sun, T. Chen, G. B. Giannakis, and Z. Yang. Communication-ef cient distributed learning via lazily aggregated quantized gradients. arXiv preprint arXiv:1909.07588, 2019.

[47] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Ef cient algorithms for smooth minimax optimization. Advances in Neural Information Processing Systems, 32, 2019.

[48] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In 2019 IEEE Symposium on Security and Privacy (SP), pages 707–723. IEEE, 2019.

[49] Z. Wang, W. Yan, and T. Oates. Time series classi cation from scratch with deep neural networks: A strong baseline. In 2017 International joint conference on neural networks (IJCNN), pages 1578–1585. IEEE, 2017.

[50] W. Wiesemann, D. Kuhn, and B. Rustem. Robust Markov decision processes. Mathematics of Operations Research, 38(1):153–183, 2013.

[51] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.

[52] Z. Xu, H. Zhang, Y. Xu, and G. Lan. A uni ed single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. arXiv preprint arXiv:2006.02032, 2020.

[53] Z. Xu, J. Shen, Z. Wang, and Y. Dai. Zeroth-order alternating randomized gradient projection algorithms for general nonconvex-concave minimax problems. arXiv preprint arXiv:2108.00473, 2021.

[54] K. Yang, Y. Wu, J. Huang, X. Wang, and S. Verdú. Distributed robust optimization for communication networks. In IEEE INFOCOM 2008-The 27th Conference on Computer Communications, pages 1157–1165. IEEE, 2008.

[55] K. Yang, J. Huang, Y. Wu, X. Wang, and M. Chiang. Distributed robust optimization (DRO), part I: Framework and example. Optimization and Engineering, 15(1):35–67, 2014.

[56] I. Yan koglu, B. L. Gorissen, and D. den Hertog. A survey of adjustable robust optimization. European Journal of Operational Research, 277(3):799–813, 2019.

[57] S. Zawad, A. Ali, P.-Y. Chen, A. Anwar, Y. Zhou, N. Baracaldo, Y. Tian, and F. Yan. Curse or redemption? how data heterogeneity affects the robustness of federated learning. *arXiv preprint arXiv:2102.00655*, 2021.

[58] R. Zhang and J. Kwok. Asynchronous distributed ADMM for consensus optimization. In *International conference on machine learning*, pages 1701–1709. PMLR, 2014.

[59] X. Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] There is no potential negative societal impact of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 5.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A and B.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The references of the data used in this paper are added in Section 6.1.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section C.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 6.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 6.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Proof of Theorem 1

Before proceeding to the detailed proofs, we provide some notations for the clarity in presentation. We use notation $\langle \cdot, \cdot \rangle$ to denote the inner product and we use $||\cdot||$ to denote the l_2-norm. $|A^t|$ and $|Q^{t+1}|$ respectively denote the number of cutting planes and active workers in $(t+1)^{th}$ iteration.

Then, we cover some Lemmas which are useful for the deduction of Theorem 1.

Lemma 1 Suppose Assumption 1 and 2 hold, $\forall t \geq T_1 + \tau$, we have,

$$L_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \leq \sum_{j=1}^{N} \left(\frac{L+1}{2} - \frac{1}{\eta_w^t}\right)||w_j^{t+1} - w_j^t||^2 + \frac{3\tau k_1 NL^2}{2}\left(||z^{t+1} - z^t||^2 + ||h^{t+1} - h^t||^2 + \sum_{l=1}^{|A^t|}||\lambda_l^{t+1} - \lambda_l^t||^2\right); \tag{A.1}$$

$$L_p(\{w_j^{t+1}\}, z^{t+1}, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \leq \left(\frac{L}{2} - \frac{1}{\eta_z^t}\right)||z^{t+1} - z^t||^2; \tag{A.2}$$

$$L_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_j^{t+1}\}, z^{t+1}, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \leq \left(\frac{L}{2} - \frac{1}{\eta_h^t}\right)||h^{t+1} - h^t||^2. \tag{A.3}$$

Proof of Lemma 1:

According to Assumption 1, we have,

$$\begin{aligned}
&L_p(\{w_1^{t+1}, w_2^t, \cdots, w_N^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \\
&\leq \langle \nabla_{w_1} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}), w_1^{t+1} - w_1^t \rangle + \frac{L}{2}||w_1^{t+1} - w_1^t||^2; \\
&L_p(\{w_1^{t+1}, w_2^{t+1}, w_3^t, \cdots, w_N^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_1^{t+1}, w_2^t, \cdots, w_N^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \\
&\leq \langle \nabla_{w_2} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}), w_2^{t+1} - w_2^t \rangle + \frac{L}{2}||w_2^{t+1} - w_2^t||^2; \\
&\vdots \\
&L_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_1^{t+1}, \cdots, w_{N-1}^{t+1}, w_N^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \\
&\leq \langle \nabla_{w_N} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}), w_N^{t+1} - w_N^t \rangle + \frac{L}{2}||w_N^{t+1} - w_N^t||^2.
\end{aligned} \tag{A.4}$$

Summing up the above inequalities in Eq. (A.4), we have,

$$L_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \leq \sum_{j=1}^{N}\left(\langle \nabla_{w_j} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}), w_j^{t+1} - w_j^t \rangle + \frac{L}{2}||w_j^{t+1} - w_j^t||^2\right). \tag{A.5}$$

According to $\nabla_{w_j} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) = \nabla_{w_j} \tilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\})$ and the optimal condition for Eq. (11), for active nodes, i.e., $\forall j \in Q^{t+1}, \forall t \geq T_1 + \tau$, we have,

$$\left\langle w_j^t - w_j^{t+1}, w_j^{t+1} - w_j^t + \eta_w^t \nabla_{w_j} L_p(\{w_j^{\hat{t}_j}\}, z^{\hat{t}_j}, h^{\hat{t}_j}, \{\lambda_l^{\hat{t}_j}\}, \{\theta_j^{\hat{t}_j}\}) \right\rangle \geq 0. \tag{A.6}$$

According to Eq. (A.6), $\forall t \geq T_1 + \tau$, we have,

$$\left\langle w_j^{t+1} - w_j^t, \nabla_{w_j} L_p(\{w_j^{\hat{t}_j}\}, z^{\hat{t}_j}, h^{\hat{t}_j}, \{\lambda_l^{\hat{t}_j}\}, \{\theta_j^{\hat{t}_j}\}) \right\rangle \leq -\frac{1}{\eta_w^t}||w_j^{t+1} - w_j^t||^2 \leq -\frac{1}{\eta_w^t}||w_j^{t+1} - w_j^t||^2. \tag{A.7}$$

And according to the Cauchy-Schwarz inequality, Assumption 1 and 2, we can get,

$$\begin{aligned}
&\left\langle w_j^{t+1} - w_j^t, \nabla_{w_j} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) - \nabla_{w_j} L_p(\{w_j^{\tilde{t}_j}\}, z^{\tilde{t}_j}, h^{\tilde{t}_j}, \{\lambda_l^{\tilde{t}_j}\}, \{\theta_j^{\tilde{t}_j}\})\right\rangle \\
&\le \tfrac{1}{2}\|w_j^{t+1} - w_j^t\|^2 + \tfrac{L^2}{2}\Big(\|z^t - z^{\tilde{t}_j}\|^2 + \|h^t - h^{\tilde{t}_j}\|^2 + \sum_{l=1}^{|P^t|}\|\lambda_l^t - \lambda_l^{\tilde{t}_j}\|^2\Big) \\
&\le \tfrac{1}{2}\|w_j^{t+1} - w_j^t\|^2 + \tfrac{3k_1L^2}{2}\Big(\|z^{t+1} - z^t\|^2 + \|h^{t+1} - h^t\|^2 + \sum_{l=1}^{|P^t|}\|\lambda_l^{t+1} - \lambda_l^t\|^2\Big).
\end{aligned} \tag{A.8}$$

Combining the above Eq. (A.5), (A.7) with Eq. (A.8), we can obtain Eq. (A.1), that is,

$$\begin{aligned}
&L_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \\
&\le \sum_{j=1}^{N}\Big(\tfrac{L+1}{2} - \tfrac{1}{\eta_w^t}\Big)\|w_j^{t+1} - w_j^t\|^2 + \tfrac{3k_1NL^2}{2}\Big(\|z^{t+1} - z^t\|^2 + \|h^{t+1} - h^t\|^2 + \sum_{l=1}^{|P^t|}\|\lambda_l^{t+1} - \lambda_l^t\|^2\Big).
\end{aligned}$$

Following Assumption 1, we have,

$$\begin{aligned}
&L_p(\{w_j^{t+1}\}, z^{t+1}, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \\
&\le \left\langle \nabla_z L_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}), z^{t+1} - z^t\right\rangle + \tfrac{L}{2}\|z^{t+1} - z^t\|^2.
\end{aligned} \tag{A.9}$$

According to $\nabla_z L_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) = \nabla_z \tilde{L}_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\})$ and the optimal condition for Eq. (12), we have,

$$\left\langle z^t - z^{t+1}, z^{t+1} - z^t + \eta_z^t \nabla_z L_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\})\right\rangle \ge 0. \tag{A.10}$$

Combining Eq. (A.9) with Eq. (A.10), we can obtain the Eq. (A.2), that is,

$$L_p(\{w_j^{t+1}\}, z^{t+1}, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_j^{t+1}\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \le \Big(\tfrac{L}{2} - \tfrac{1}{\eta_z^t}\Big)\|z^{t+1} - z^t\|^2.$$

According to Assumption 1, we have:

$$\begin{aligned}
&L_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_j^{t+1}\}, z^{t+1}, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \\
&\le \left\langle \nabla_h L_p(\{w_j^{t+1}\}, z^{t+1}, h^t, \{\lambda_l^t\}, \{\theta_j^t\}), h^{t+1} - h^t\right\rangle + \tfrac{L}{2}\|h^{t+1} - h^t\|^2.
\end{aligned} \tag{A.11}$$

According to $\nabla_h L_p(\{w_j^{t+1}\}, z^{t+1}, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) = \nabla_h \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^t, \{\lambda_l^t\}, \{\theta_j^t\})$ and the optimal condition for Eq. (13), we have:

$$\left\langle h^t - h^{t+1}, h^{t+1} - h^t + \eta_h^t \nabla_h L_p(\{w_j^{t+1}\}, z^{t+1}, h^t, \{\lambda_l^t\}, \{\theta_j^t\})\right\rangle \ge 0. \tag{A.12}$$

Combining Eq. (A.11) with Eq. (A.12), we can show that,

$$L_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\lambda_l^t\}, \{\theta_j^t\}) - L_p(\{w_j^{t+1}\}, z^{t+1}, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \le \Big(\tfrac{L}{2} - \tfrac{1}{\eta_h^t}\Big)\|h^{t+1} - h^t\|^2.$$

Lemma 2 Suppose Assumption 1 and 2 hold, $\forall t \ge T_1 + \tau$, we have:

$$\begin{aligned}
&L_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\lambda_l^{t+1}\}, \{\theta_j^{t+1}\}) - L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\theta_j^t\}) \\
&\le \Big(\tfrac{L+1}{2} - \tfrac{1}{\eta_w^t} + \tfrac{|A^t|L^2}{2a_1} + \tfrac{|Q^{t+1}|L^2}{2a_3}\Big)\sum_{j=1}^{N}\|w_j^{t+1} - w_j^t\|^2 + \Big(\tfrac{L+3k_1NL^2}{2} - \tfrac{1}{\eta_z^t} + \tfrac{|A^t|L^2}{2a_1} + \tfrac{|Q^{t+1}|L^2}{2a_3}\Big)\|z^{t+1} - z^t\|^2 \\
&+ \Big(\tfrac{L+3k_1NL^2}{2} - \tfrac{1}{\eta_h^t} + \tfrac{|A^t|L^2}{2a_1} + \tfrac{|Q^{t+1}|L^2}{2a_3}\Big)\|h^{t+1} - h^t\|^2 + \Big(\tfrac{a_1+3k_1NL^2}{2} - \tfrac{c_1^{t-1} - c_1^t}{2} + \tfrac{1}{2\rho_1}\Big)\sum_{l=1}^{|P^t|}\|\lambda_l^{t+1} - \lambda_l^t\|^2 \\
&+ \tfrac{c_1^{t-1}}{2}\sum_{l=1}^{|P^t|}\big(\|\lambda_l^{t+1}\|^2 - \|\lambda_l^t\|^2\big) + \tfrac{1}{2\rho_1}\sum_{l=1}^{|P^t|}\|\lambda_l^t - \lambda_l^{t-1}\|^2 + \Big(\tfrac{a_3}{2} - \tfrac{c_2^{t-1} - c_2^t}{2} + \tfrac{1}{2\rho_2}\Big)\sum_{j=1}^{N}\|\theta_j^{t+1} - \theta_j^t\|^2 \\
&+ \tfrac{c_2^{t-1}}{2}\sum_{j=1}^{N}\big(\|\theta_j^{t+1}\|^2 - \|\theta_j^t\|^2\big) + \tfrac{1}{2\rho_2}\sum_{j=1}^{N}\|\theta_j^t - \theta_j^{t-1}\|^2;
\end{aligned} \tag{A.13}$$

where $a_1 > 0$ and $a_3 > 0$ are constants.

Proof of Lemma 2

First of all, at $(t+1)^{th}$ iteration, the following equations hold and will be used in the derivation:

$$\sum_{j=1}^{N} \| {}_j^{t+1} - {}_j^{t} \|^2 = \sum_{j \in Q^{t+1}} \| {}_j^{t+1} - {}_j^{t} \|^2; \quad \sum_{j=1}^{N} (\| {}_j^{t+1} \|^2 - \| {}_j^{t} \|^2) = \sum_{j \in Q^{t+1}} (\| {}_j^{t+1} \|^2 - \| {}_j^{t} \|^2).$$

According to Eq. (14), in $(t+1)^{th}$ iteration, $\forall \in$, it follows that:

$$\left\langle {}_l^{t+1} - {}_l^{t} - {}_1 \nabla_{l} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ {}_l^{t}\}, \{ {}_j^{t}\}), \; {}_l^{t+1} \right\rangle \geq 0. \tag{A.14}$$

Let $= {}_l^t$, we can obtain:

$$\left\langle \nabla_{l} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ {}_l^{t}\}, \{ {}_j^{t}\}) - \frac{1}{{}_1}({}_l^{t+1} - {}_l^{t}), \; {}_l^{t} - {}_l^{t+1} \right\rangle \geq 0. \tag{A.15}$$

Likewise, in t^{th} iteration, we can obtain:

$$\left\langle \nabla_{l} \tilde{L}_p(\{w_j^{t}\}, z^{t}, h^{t}, \{ {}_l^{t-1}\}, \{ {}_j^{t-1}\}) - \frac{1}{{}_1}({}_l^{t} - {}_l^{t-1}), \; {}_l^{t+1} - {}_l^{t} \right\rangle \geq 0. \tag{A.16}$$

$\forall t \geq T_1$, since $\tilde{L}_p(\{w_j\}, z, h, \{ {}_l\}, \{ {}_j\})$ is concave with respect to ${}_l$, we have,

$$\begin{aligned}
&\tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ {}_l^{t+1}\}, \{ {}_j^{t}\}) - \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ {}_l^{t}\}, \{ {}_j^{t}\}) \\
&\leq \sum_{l=1}^{|P^t|} \left\langle \nabla_{l} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ {}_l^{t}\}, \{ {}_j^{t}\}), \; {}_l^{t+1} - {}_l^{t} \right\rangle \\
&\leq \sum_{l=1}^{|P^t|} \Big\langle \nabla_{l} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ {}_l^{t}\}, \{ {}_j^{t}\}) - \nabla_{l} \tilde{L}_p(\{w_j^{t}\}, z^{t}, h^{t}, \{ {}_l^{t-1}\}, \{ {}_j^{t-1}\}), \; {}_l^{t+1} - {}_l^{t} \\
&\quad + \frac{1}{{}_1} \left\langle {}_l^{t} - {}_l^{t-1}, \; {}_l^{t+1} - {}_l^{t} \right\rangle \Big).
\end{aligned} \tag{A.17}$$

Denoting $v_{1,l}^{t+1} = {}_l^{t+1} - {}_l^{t} - ({}_l^{t} - {}_l^{t-1})$, we have,

$$\begin{aligned}
&\sum_{l=1}^{|P^t|} \left\langle \nabla_{l} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ {}_l^{t}\}, \{ {}_j^{t}\}) - \nabla_{l} \tilde{L}_p(\{w_j^{t}\}, z^{t}, h^{t}, \{ {}_l^{t-1}\}, \{ {}_j^{t-1}\}), \; {}_l^{t+1} - {}_l^{t} \right\rangle \\
&= \sum_{l=1}^{|P^t|} \left\langle \nabla_{l} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ {}_l^{t}\}, \{ {}_j^{t}\}) - \nabla_{l} \tilde{L}_p(\{w_j^{t}\}, z^{t}, h^{t}, \{ {}_l^{t}\}, \{ {}_j^{t}\}), \; {}_l^{t+1} - {}_l^{t} \right\rangle \quad (1a) \\
&+ \sum_{l=1}^{|P^t|} \left\langle \nabla_{l} \tilde{L}_p(\{w_j^{t}\}, z^{t}, h^{t}, \{ {}_l^{t}\}, \{ {}_j^{t}\}) - \nabla_{l} \tilde{L}_p(\{w_j^{t}\}, z^{t}, h^{t}, \{ {}_l^{t-1}\}, \{ {}_j^{t-1}\}), \; v_{1,l}^{t+1} \right\rangle \quad (1b) \\
&+ \sum_{l=1}^{|P^t|} \left\langle \nabla_{l} \tilde{L}_p(\{w_j^{t}\}, z^{t}, h^{t}, \{ {}_l^{t}\}, \{ {}_j^{t}\}) - \nabla_{l} \tilde{L}_p(\{w_j^{t}\}, z^{t}, h^{t}, \{ {}_l^{t-1}\}, \{ {}_j^{t-1}\}), \; {}_l^{t} - {}_l^{t-1} \right\rangle \quad (1c).
\end{aligned} \tag{A.18}$$

Firstly, we focus on the $(1a)$ in Eq. (A.18), we can write $(1a)$ as:

$$\begin{aligned}
&\left\langle \nabla_{l} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ {}_l^{t}\}, \{ {}_j^{t}\}) - \nabla_{l} \tilde{L}_p(\{w_j^{t}\}, z^{t}, h^{t}, \{ {}_l^{t}\}, \{ {}_j^{t}\}), \; {}_l^{t+1} - {}_l^{t} \right\rangle \\
&= \left\langle \nabla_{l} L_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ {}_l^{t}\}, \{ {}_j^{t}\}) - \nabla_{l} L_p(\{w_j^{t}\}, z^{t}, h^{t}, \{ {}_l^{t}\}, \{ {}_j^{t}\}), \; {}_l^{t+1} - {}_l^{t} \right\rangle \\
&+ (c_1^{t-1} - c_1^{t}) \left\langle {}_l^{t}, \; {}_l^{t+1} - {}_l^{t} \right\rangle \\
&= \left\langle \nabla_{l} L_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ {}_l^{t}\}, \{ {}_j^{t}\}) - \nabla_{l} L_p(\{w_j^{t}\}, z^{t}, h^{t}, \{ {}_l^{t}\}, \{ {}_j^{t}\}), \; {}_l^{t+1} - {}_l^{t} \right\rangle \\
&+ \frac{c_1^{t-1} - c_1^{t}}{2} (\| {}_l^{t+1} \|^2 - \| {}_l^{t} \|^2) - \frac{c_1^{t-1} - c_1^{t}}{2} \| {}_l^{t+1} - {}_l^{t} \|^2.
\end{aligned} \tag{A.19}$$

And according to Cauchy-Schwarz inequality and Assumption 1, we can obtain,

$$
\begin{aligned}
&\langle \nabla_{\lambda_l} L_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\lambda_l^t\}, \{\phi_j^t\}) - \nabla_{\lambda_l} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}), \lambda_l^{t+1} - \lambda_l^t \rangle \\
&\le \frac{L^2}{2a_1}\Big(\sum_{j=1}^{N} \|w_j^{t+1} - w_j^t\|^2 + \|z^{t+1} - z^t\|^2 + \|h^{t+1} - h^t\|^2\Big) + \frac{a_1}{2}\|\lambda_l^{t+1} - \lambda_l^t\|^2,
\end{aligned} \tag{A.20}
$$

where $a_1 > 0$ is a constant. Combining Eq. (A.19) with Eq. (A.20), we can obtain the upper bound of (1a), that is,

$$
\begin{aligned}
&\sum_{l=1}^{|P^t|} \Big\langle \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\lambda_l^t\}, \{\phi_j^t\}) - \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}), \lambda_l^{t+1} - \lambda_l^t \Big\rangle \\
&\le \sum_{l=1}^{|P^t|} \Big(\frac{L^2}{2a_1}\Big(\sum_{j=1}^{N} \|w_j^{t+1} - w_j^t\|^2 + \|z^{t+1} - z^t\|^2 + \|h^{t+1} - h^t\|^2\Big) + \frac{a_1}{2}\|\lambda_l^{t+1} - \lambda_l^t\|^2 \\
&\quad + \frac{c_1^{t-1} - c_1^t}{2}\big(\|\lambda_l^{t+1}\|^2 - \|\lambda_l^t\|^2\big) - \frac{c_1^{t-1} - c_1^t}{2}\|\lambda_l^{t+1} - \lambda_l^t\|^2 \Big).
\end{aligned} \tag{A.21}
$$

Secondly, we focus on the (1b) in Eq. (A.18). According to Cauchy-Schwarz inequality we can write the (1b) as,

$$
\begin{aligned}
&\sum_{l=1}^{|P^t|} \Big\langle \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}) - \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^{t-1}\}, \{\phi_j^{t-1}\}), v_{1,l}^{t+1} \Big\rangle \\
&\le \sum_{l=1}^{|P^t|} \Big(\frac{a_2}{2}\|\nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}) - \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^{t-1}\}, \{\phi_j^{t-1}\})\|^2 + \frac{1}{2a_2}\|v_{1,l}^{t+1}\|^2\Big),
\end{aligned} \tag{A.22}
$$

where $a_2 > 0$ is a constant. Then, we focus on the (1c) in Eq. (A.18). Firstly, $\forall \lambda_l$, we have,

$$
\begin{aligned}
&\|\nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}) - \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^{t-1}\}, \{\phi_j^{t-1}\})\| \\
&= \|\nabla_{\lambda_l} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}) - \nabla_{\lambda_l} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^{t-1}\}, \{\phi_j^t\}) - c_1^{t-1}(\lambda_l^t - \lambda_l^{t-1})\| \\
&\le (L + c_1^{t-1})\|\lambda_l^t - \lambda_l^{t-1}\|,
\end{aligned} \tag{A.23}
$$

where the last inequality comes from Assumption 1 and the trigonometric inequality. Denoting $L_1{}' = L + c_1^0$, we can obtain,

$$
\|\nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}) - \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^{t-1}\}, \{\phi_j^{t-1}\})\| \le L_1{}'\|\lambda_l^t - \lambda_l^{t-1}\|. \tag{A.24}
$$

Following from Eq. (A.24) and the strong concavity of $\widetilde{L}_p(\{w_j\}, z, h, \{\lambda_l\}, \{\phi_j\})$ w.r.t λ_l [37, 52], we can obtain the upper bound of (1c):

$$
\begin{aligned}
&\sum_{l=1}^{|P^t|} \Big\langle \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}) - \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^{t-1}\}, \{\phi_j^{t-1}\}), \lambda_l^t - \lambda_l^{t-1} \Big\rangle \\
&\le \sum_{l=1}^{|P^t|} \Big(-\frac{1}{L_1{}' + c_1^{t-1}}\|\nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}) - \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^{t-1}\}, \{\phi_j^{t-1}\})\|^2 \\
&\quad - \frac{c_1^{t-1} L_1{}'}{L_1{}' + c_1^{t-1}}\|\lambda_l^t - \lambda_l^{t-1}\|^2 \Big).
\end{aligned} \tag{A.25}
$$

In addition, the following inequality can be obtained,

$$
\frac{1}{\eta_\lambda}\langle \lambda_l^t - \lambda_l^{t-1}, \lambda_l^{t+1} - \lambda_l^t \rangle \le \frac{1}{2\eta_\lambda}\|\lambda_l^{t+1} - \lambda_l^t\|^2 - \frac{1}{2\eta_\lambda}\|v_{1,l}^{t+1}\|^2 + \frac{1}{2\eta_\lambda}\|\lambda_l^t - \lambda_l^{t-1}\|^2. \tag{A.26}
$$

Combining Eq. (A.17), (A.18), (A.21), (A.22), (A.25), (A.26), $\frac{1}{2} \; \frac{1}{L_1{}^0 + c_1^0}$, and setting $a_2 = \; _1$, we have:

$$L_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\ _l^{t+1}\}, \{\ _j^t\}) \quad L_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\ _l^t\}, \{\ _j^t\})$$

$$\sum_{l=1}^{|P^t|} \Big(\big\langle r_{\ l} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\ _l^t\}, \{\ _j^t\}) \;\; r_{\ l} \tilde{L}_p(\{w_j^t\}, z^t, h^t, \{\ _l^{t-1}\}, \{\ _j^{t-1}\}), \ _l^{t+1} \;\; _l^t \big\rangle$$

$$+ \frac{1}{\ _1} \langle \ _l^t \;\; _l^{t-1}, \ _l^{t+1} \;\; _l^t \rangle + \frac{c_1^t}{2}(||\ _l^{t+1}||^2 \;\; ||\ _l^t||^2)\Big)$$

$$\sum_{l=1}^{|P^t|} \Big(\frac{L^2}{2a_1}\big(\sum_{j=1}^{N} ||w_j^{t+1} \;\; w_j^t||^2 + ||z^{t+1} \;\; z^t||^2 + ||h^{t+1} \;\; h^t||^2\big)$$

$$+ \big(\frac{a_1}{2} \;\; \frac{c_1^{t-1} \;\; c_1^t}{2} + \frac{1}{2\ _1}\big)||\ _l^{t+1} \;\; _l^t||^2 + \frac{c_1^{t-1}}{2}(||\ _l^{t+1}||^2 \;\; ||\ _l^t||^2) + \frac{1}{2\ _1}||\ _l^t \;\; _l^{t-1}||^2\Big)$$

$$= \frac{|A^t| L^2}{2a_1}\big(\sum_{j=1}^{N} ||w_j^{t+1} \;\; w_j^t||^2 + ||z^{t+1} \;\; z^t||^2 + ||h^{t+1} \;\; h^t||^2\big)$$

$$+ \big(\frac{a_1}{2} \;\; \frac{c_1^{t-1} \;\; c_1^t}{2} + \frac{1}{2\ _1}\big)\sum_{l=1}^{|P^t|} ||\ _l^{t+1} \;\; _l^t||^2 + \frac{c_1^{t-1}}{2}\sum_{l=1}^{|P^t|}(||\ _l^{t+1}||^2 \;\; ||\ _l^t||^2) + \frac{1}{2\ _1}\sum_{l=1}^{|P^t|} ||\ _l^t \;\; _l^{t-1}||^2. \tag{A.27}$$

According to Eq. (15), $8 \;\; 2$, it follows that,

$$\big\langle \ _j^{t+1} \;\; _j^t \;\; _2 r_{\ j} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\ _l^{t+1}\}, \{\ _j^t\}), \;\; _j^{t+1} \big\rangle \;\; 0. \tag{A.28}$$

Choosing $= \ _j^t$, we can obtain,

$$\langle r_{\ j} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\ _l^{t+1}\}, \{\ _j^t\}) \;\; \frac{1}{\ _2}(\ _j^{t+1} \;\; _j^t), \ _j^t \;\; _j^{t+1} \rangle \;\; 0. \tag{A.29}$$

Likewise, we have,

$$\langle r_{\ j} \tilde{L}_p(\{w_j^t\}, z^t, h^t, \{\ _l^t\}, \{\ _j^{t-1}\}) \;\; \frac{1}{\ _2}(\ _j^t \;\; _j^{t-1}), \ _j^{t+1} \;\; _j^t \rangle \;\; 0. \tag{A.30}$$

Since $\tilde{L}_p(\{w_j\}, z, h, \{\ _l\}, \{\ _j\})$ is concave with respect to $_j$ and follows from Eq. (A.30):

$$\tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\ _l^{t+1}\}, \{\ _j^{t+1}\}) \quad \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\ _l^{t+1}\}, \{\ _j^t\})$$

$$\sum_{j=1}^{N} \big\langle r_{\ j} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\ _l^{t+1}\}, \{\ _j^t\}), \ _j^{t+1} \;\; _j^t \big\rangle$$

$$\sum_{j=1}^{N} \Big(\big\langle r_{\ j} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\ _l^{t+1}\}, \{\ _j^t\}) \;\; r_{\ j} \tilde{L}_p(\{w_j^t\}, z^t, h^t, \{\ _l^t\}, \{\ _j^{t-1}\}), \ _j^{t+1} \;\; _j^t \big\rangle$$

$$+ \frac{1}{\ _2}\langle \ _j^t \;\; _j^{t-1}, \ _j^{t+1} \;\; _j^t \rangle\Big). \tag{A.31}$$

Denoting $v_{2,l}^{t+1} = \ _j^{t+1} \;\; _j^t \;\; (\ _j^t \;\; _j^{t-1})$, we can write the first term in the last inequality of Eq. (A.31) as

$$\sum_{j=1}^{N} \big\langle r_{\ j} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\ _l^{t+1}\}, \{\ _j^t\}) \;\; r_{\ j} \tilde{L}_p(\{w_j^t\}, z^t, h^t, \{\ _l^t\}, \{\ _j^{t-1}\}), \ _j^{t+1} \;\; _j^t \big\rangle$$

$$= \sum_{j=1}^{N} \big\langle r_{\ j} \tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{\ _l^{t+1}\}, \{\ _j^t\}) \;\; r_{\ j} \tilde{L}_p(\{w_j^t\}, z^t, h^t, \{\ _l^t\}, \{\ _j^t\}), \ _j^{t+1} \;\; _j^t \big\rangle \quad (2a)$$

$$+ \sum_{j=1}^{N} \big\langle r_{\ j} \tilde{L}_p(\{w_j^t\}, z^t, h^t, \{\ _l^t\}, \{\ _j^t\}) \;\; r_{\ j} \tilde{L}_p(\{w_j^t\}, z^t, h^t, \{\ _l^t\}, \{\ _j^{t-1}\}), v_{2,l}^{t+1} \big\rangle \quad (2b)$$

$$+ \sum_{j=1}^{N} \big\langle r_{\ j} \tilde{L}_p(\{w_j^t\}, z^t, h^t, \{\ _l^t\}, \{\ _j^t\}) \;\; r_{\ j} \tilde{L}_p(\{w_j^t\}, z^t, h^t, \{\ _l^t\}, \{\ _j^{t-1}\}), \ _j^t \;\; _j^{t-1} \big\rangle \quad (2c). \tag{A.32}$$

We firstly focus on the (2a) in Eq. (A.32), we can write the (2a) as,

$$
\begin{aligned}
&\left\langle \nabla_{\theta_j} \tilde{L}_p(\{w_j^{t+1}\},z^{t+1},h^{t+1},\{\lambda_l^{t+1}\},\{\theta_j^t\}) - \nabla_{\theta_j} \tilde{L}_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}), \theta_j^{t+1}-\theta_j^t \right\rangle \\
&= \left\langle \nabla_{\theta_j} L_p(\{w_j^{t+1}\},z^{t+1},h^{t+1},\{\lambda_l^{t+1}\},\{\theta_j^t\}) - \nabla_{\theta_j} L_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}), \theta_j^{t+1}-\theta_j^t \right\rangle \\
&+(c_2^{t-1}-c_2^t)\left\langle \theta_j^t, \theta_j^{t+1}-\theta_j^t \right\rangle \\
&= \left\langle \nabla_{\theta_j} L_p(\{w_j^{t+1}\},z^{t+1},h^{t+1},\{\lambda_l^{t+1}\},\{\theta_j^t\}) - \nabla_{\theta_j} L_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}), \theta_j^{t+1}-\theta_j^t \right\rangle \\
&+\frac{c_2^{t-1}-c_2^t}{2}(||\theta_j^{t+1}||^2-||\theta_j^t||^2)-\frac{c_2^{t-1}-c_2^t}{2}||\theta_j^{t+1}-\theta_j^t||^2).
\end{aligned}
\tag{A.33}
$$

And according to Cauchy-Schwarz inequality and Assumption 1, we can obtain,

$$
\begin{aligned}
&\left\langle \nabla_{\theta_j} L_p(\{w_j^{t+1}\},z^{t+1},h^{t+1},\{\lambda_l^{t+1}\},\{\theta_j^t\}) - \nabla_{\theta_j} L_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}), \theta_j^{t+1}-\theta_j^t \right\rangle \\
&= \left\langle \nabla_{\theta_j} L_p(\{w_j^{t+1}\},z^{t+1},h^{t+1},\{\lambda_l^{t}\},\{\theta_j^t\}) - \nabla_{\theta_j} L_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}), \theta_j^{t+1}-\theta_j^t \right\rangle \\
&\le \frac{L^2}{2a_3}\Big(\sum_{j=1}^{N}||w_j^{t+1}-w_j^t||^2+||z^{t+1}-z^t||^2+||h^{t+1}-h^t||^2\Big)+\frac{a_3}{2}||\theta_j^{t+1}-\theta_j^t||^2,
\end{aligned}
\tag{A.34}
$$

where $a_3>0$ is a constant. Thus, we can obtain the upper bound of (2a) by combining the above Eq. (A.33) and Eq. (A.34),

$$
\begin{aligned}
&\sum_{j=1}^{N}\left\langle \nabla_{\theta_j} \tilde{L}_p(\{w_j^{t+1}\},z^{t+1},h^{t+1},\{\lambda_l^{t+1}\},\{\theta_j^t\}) - \nabla_{\theta_j} \tilde{L}_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}), \theta_j^{t+1}-\theta_j^t \right\rangle \\
&= \sum_{j\in Q^{t+1}}\left\langle \nabla_{\theta_j} \tilde{L}_p(\{w_j^{t+1}\},z^{t+1},h^{t+1},\{\lambda_l^{t+1}\},\{\theta_j^t\}) - \nabla_{\theta_j} \tilde{L}_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}), \theta_j^{t+1}-\theta_j^t \right\rangle \\
&\le \sum_{j\in Q^{t+1}}\Big(\frac{L^2}{2a_3}\Big(\sum_{j=1}^{N}||w_j^{t+1}-w_j^t||^2+||z^{t+1}-z^t||^2+||h^{t+1}-h^t||^2\Big)+\frac{a_3}{2}||\theta_j^{t+1}-\theta_j^t||^2 \\
&\qquad +\frac{c_2^{t-1}-c_2^t}{2}(||\theta_j^{t+1}||^2-||\theta_j^t||^2)-\frac{c_2^{t-1}-c_2^t}{2}||\theta_j^{t+1}-\theta_j^t||^2).
\end{aligned}
\tag{A.35}
$$

Next we focus on the (2b) in Eq. (A.32). According to Cauchy-Schwarz inequality we can write the (2b) as

$$
\begin{aligned}
&\sum_{j=1}^{N}\left\langle \nabla_{\theta_j} \tilde{L}_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}) - \nabla_{\theta_j} \tilde{L}_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^{t-1}\}), v_{2,l}^{t+1} \right\rangle \\
&\le \sum_{j=1}^{N}\Big(\frac{a_4}{2}||\nabla_{\theta_j} \tilde{L}_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}) - \nabla_{\theta_j} \tilde{L}_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^{t-1}\})||^2+\frac{1}{2a_4}||v_{2,l}^{t+1}||^2\Big),
\end{aligned}
\tag{A.36}
$$

where $a_4>0$ is a constant. Then, we focus on the (2c) in Eq. (A.32), we have,

$$
\begin{aligned}
&||\nabla_{\theta_j} \tilde{L}_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}) - \nabla_{\theta_j} \tilde{L}_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^{t-1}\})|| \\
&\le ||\nabla_{\theta_j} L_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}) - \nabla_{\theta_j} L_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^{t-1}\})||+c_2^{t-1}||\theta_j^t-\theta_j^{t-1}|| \\
&\le (L+c_2^{t-1})||\theta_j^t-\theta_j^{t-1}||,
\end{aligned}
\tag{A.37}
$$

where the last inequality comes from Assumption 1 and the trigonometric inequality. Denoting $L_2'=L+c_2^0$, we can obtain,

$$
||\nabla_{\theta_j} \tilde{L}_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^t\}) - \nabla_{\theta_j} \tilde{L}_p(\{w_j^{t}\},z^{t},h^{t},\{\lambda_l^{t}\},\{\theta_j^{t-1}\})|| \le L_2'||\theta_j^t-\theta_j^{t-1}||.
\tag{A.38}
$$

Following Eq. (A.38) and the strong concavity of $\tilde{\mathcal{L}}_p(\{w_j\}, z, h, \{ _l\}, \{ _j\})$ w.r.t $_j$, we can obtain the upper bound of 2(c),

$$\sum_{j=1}^{N} \left\langle \nabla_{ _j} \tilde{\mathcal{L}}_p(\{w_j^t\}, z^t, h^t, \{ _l^t\}, \{ _j^t\}) - \nabla_{ _j} \tilde{\mathcal{L}}_p(\{w_j^t\}, z^t, h^t, \{ _l^t\}, \{ _j^{t-1}\}), _j^t - _j^{t-1} \right\rangle$$
$$\leq \sum_{j=1}^{N} \Big(-\frac{1}{L_2 {}^0 + c_2^{t-1}} \|\nabla_{ _j} \tilde{\mathcal{L}}_p(\{w_j^t\}, z^t, h^t, \{ _l^t\}, \{ _j^t\}) - \nabla_{ _j} \tilde{\mathcal{L}}_p(\{w_j^t\}, z^t, h^t, \{ _l^t\}, \{ _j^{t-1}\})\|^2$$
$$- \frac{c_2^{t-1} L_2 {}^0}{L_2 {}^0 + c_2^{t-1}} \| _j^t - _j^{t-1}\|^2 \Big). \tag{A.39}$$

In addition, the following inequality can also be obtained,

$$\sum_{j=1}^{N} \frac{1}{ _2} \langle _j^t - _j^{t-1}, _j^{t+1} - _j^t \rangle \leq \sum_{j=1}^{N} \left(\frac{1}{2 _2} \| _j^{t+1} - _j^t\|^2 - \frac{1}{2 _2} \|v_{2,l}^{t+1}\|^2 + \frac{1}{2 _2} \| _j^t - _j^{t-1}\|^2 \right). \tag{A.40}$$

Combining Eq. (A.31), (A.32), (A.35), (A.36), (A.39), (A.40), $\frac{2}{2} \leq \frac{1}{L_2 {}^0 + c_2^0}$, and setting $a_4 = _2$, we have,

$$\mathcal{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ _l^{t+1}\}, \{ _j^{t+1}\}) - \mathcal{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ _l^{t+1}\}, \{ _j^t\})$$
$$\leq \sum_{j=1}^{N} \Big(\left\langle \nabla_{ _j} \tilde{\mathcal{L}}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ _l^{t+1}\}, \{ _j^t\}) - \nabla_{ _j} \tilde{\mathcal{L}}_p(\{w_j^t\}, z^t, h^t, \{ _l^t\}, \{ _j^{t-1}\}), _j^{t+1} - _j^t \right\rangle$$
$$+ \frac{1}{ _2} \langle _j^t - _j^{t-1}, _j^{t+1} - _j^t \rangle + \frac{c_2^t}{2} (\| _j^{t+1}\|^2 - \| _j^t\|^2) \Big)$$
$$\leq \frac{|Q^{t+1}| L^2}{2a_3} \Big(\sum_{j=1}^{N} \|w_j^{t+1} - w_j^t\|^2 + \|z^{t+1} - z^t\|^2 + \|h^{t+1} - h^t\|^2 \Big)$$
$$+ \Big(\frac{a_3}{2} - \frac{c_2^{t-1} - c_2^t}{2} + \frac{1}{2 _2} \Big) \sum_{j=1}^{N} \| _j^{t+1} - _j^t\|^2 + \frac{c_2^{t-1}}{2} \sum_{j=1}^{N} (\| _j^{t+1}\|^2 - \| _j^t\|^2) + \frac{1}{2 _2} \sum_{j=1}^{N} \| _j^t - _j^{t-1}\|^2. \tag{A.41}$$

By combining Lemma 1 with Eq. (A.27) and Eq. (A.41), we conclude the proof of Lemma 2.

Lemma 3 Firstly, we denote S_1^{t+1}, S_2^{t+1} and F^{t+1} as,

$$S_1^{t+1} = \frac{4}{ _1^2 c_1^{t+1}} \sum_{l=1}^{|\mathcal{A}^t|} \| _l^{t+1} - _l^t\|^2 - \frac{4}{ _1} \Big(\frac{c_1^{t-1}}{c_1^t} - 1 \Big) \sum_{l=1}^{|\mathcal{A}^t|} \| _l^{t+1}\|^2; \tag{A.42}$$

$$S_2^{t+1} = \frac{4}{ _2^2 c_2^{t+1}} \sum_{j=1}^{N} \| _j^{t+1} - _j^t\|^2 - \frac{4}{ _2} \Big(\frac{c_2^{t-1}}{c_2^t} - 1 \Big) \sum_{j=1}^{N} \| _j^{t+1}\|^2; \tag{A.43}$$

$$F^{t+1} = \mathcal{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{ _l^{t+1}\}, \{ _j^{t+1}\}) + S_1^{t+1} + S_2^{t+1}$$
$$- \frac{7}{2 _1} \sum_{l=1}^{|\mathcal{P}^t|} \| _l^{t+1} - _l^t\|^2 - \frac{c_1^t}{2} \sum_{l=1}^{|\mathcal{P}^t|} \| _l^{t+1}\|^2 - \frac{7}{2 _2} \sum_{j=1}^{N} \| _j^{t+1} - _j^t\|^2 - \frac{c_2^t}{2} \sum_{j=1}^{N} \| _j^{t+1}\|^2; \tag{A.44}$$

then $\forall t \geq T_1 +$, we have,

$$F^{t+1} - F^{t} \le \Big(\frac{L+1}{2} - \frac{1}{{}^{t}_{w}} + \frac{{}_1\|A^t\|L^2}{2} + \frac{{}_2\|Q^{t+1}\|L^2}{2} + \frac{8\|A^t\|L^2}{{}_1(c_1^t)^2} + \frac{8NL^2}{{}_2(c_2^t)^2}\Big)\sum_{j=1}^{N}\|w_j^{t+1} - w_j^t\|^2$$
$$+\Big(\frac{L+3k_1NL^2}{2} - \frac{1}{{}^{t}_{z}} + \frac{{}_1\|A^t\|L^2}{2} + \frac{{}_2\|Q^{t+1}\|L^2}{2} + \frac{8\|A^t\|L^2}{{}_1(c_1^t)^2} + \frac{8NL^2}{{}_2(c_2^t)^2}\Big)\|z^{t+1} - z^t\|^2$$
$$+\Big(\frac{L+3k_1NL^2}{2} - \frac{1}{{}^{t}_{h}} + \frac{{}_1\|A^t\|L^2}{2} + \frac{{}_2\|Q^{t+1}\|L^2}{2} + \frac{8\|A^t\|L^2}{{}_1(c_1^t)^2} + \frac{8NL^2}{{}_2(c_2^t)^2}\Big)\|h^{t+1} - h^t\|^2$$
$$-\Big(\frac{1}{10\,{}_1} - \frac{3k_1NL^2}{2}\Big)\sum_{l=1}^{|\mathcal{P}^t|}\|{}_l^{t+1} - {}_l^t\|^2 - \frac{1}{10\,{}_2}\sum_{j=1}^{N}\|{}_j^{t+1} - {}_j^t\|^2 + \frac{c_1^{t-1} - c_1^t}{2}\sum_{l=1}^{|\mathcal{P}^t|}\|{}_l^{t+1}\|^2$$
$$+\frac{c_2^{t-1} - c_2^t}{2}\sum_{j=1}^{N}\|{}_j^{t+1}\|^2 + \frac{4}{{}_1}\Big(\frac{c_1^{t-2}}{c_1^{t-1}} - \frac{c_1^{t-1}}{c_1^t}\Big)\sum_{l=1}^{|\mathcal{P}^t|}\|{}_l^t\|^2 + \frac{4}{{}_2}\Big(\frac{c_2^{t-2}}{c_2^{t-1}} - \frac{c_2^{t-1}}{c_2^t}\Big)\sum_{j=1}^{N}\|{}_j^t\|^2. \tag{A.45}$$

Proof of Lemma 3:

Let $a_1 = \frac{1}{{}_1}$, $a_3 = \frac{1}{{}_2}$ and substitute them into Lemma 2, $\forall t \geq T_1 +$, we have,

$$L_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{{}_l^{t+1}\}, \{{}_j^{t+1}\}) - L_p(\{w_j^t\}, z^t, h^t, \{{}_l^t\}, \{{}_j^t\})$$
$$\le \Big(\frac{L+1}{2} - \frac{1}{{}^{t}_{w}} + \frac{{}_1\|A^t\|L^2}{2} + \frac{{}_2\|Q^{t+1}\|L^2}{2}\Big)\sum_{j=1}^{N}\|w_j^{t+1} - w_j^t\|^2$$
$$+\Big(\frac{L+3k_1NL^2}{2} - \frac{1}{{}^{t}_{z}} + \frac{{}_1\|A^t\|L^2}{2} + \frac{{}_2\|Q^{t+1}\|L^2}{2}\Big)\|z^{t+1} - z^t\|^2$$
$$+\Big(\frac{L+3k_1NL^2}{2} - \frac{1}{{}^{t}_{h}} + \frac{{}_1\|A^t\|L^2}{2} + \frac{{}_2\|Q^{t+1}\|L^2}{2}\Big)\|h^{t+1} - h^t\|^2 \tag{A.46}$$
$$+\Big(\frac{3k_1NL^2}{2} + \frac{1}{{}_1} - \frac{c_1^{t-1} - c_1^t}{2}\Big)\sum_{l=1}^{|\mathcal{P}^t|}\|{}_l^{t+1} - {}_l^t\|^2$$
$$+\frac{c_1^{t-1}}{2}\sum_{l=1}^{|\mathcal{P}^t|}(\|{}_l^{t+1}\|^2 - \|{}_l^t\|^2) + \frac{1}{2\,{}_1}\sum_{l=1}^{|\mathcal{P}^t|}\|{}_l^t - {}_l^{t-1}\|^2 + \Big(\frac{1}{{}_2} - \frac{c_2^{t-1} - c_2^t}{2}\Big)\sum_{j=1}^{N}\|{}_j^{t+1} - {}_j^t\|^2$$
$$+\frac{c_2^{t-1}}{2}\sum_{j=1}^{N}(\|{}_j^{t+1}\|^2 - \|{}_j^t\|^2) + \frac{1}{2\,{}_2}\sum_{j=1}^{N}\|{}_j^t - {}_j^{t-1}\|^2.$$

According to Eq. (14), in $(t+1)^{th}$ iteration, it follows that:

$$\Big\langle {}_l^{t+1} - {}_l^t - {}_1\nabla_{{}_l}\tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{{}_l^t\}, \{{}_j^t\}),\ {}_l - {}_l^{t+1}\Big\rangle \le 0. \tag{A.47}$$

Similar to Eq. (A.47), in t^{th} iteration, we have,

$$\Big\langle {}_l^t - {}_l^{t-1} - {}_1\nabla_{{}_l}\tilde{L}_p(\{w_j^t\}, z^t, h^t, \{{}_l^{t-1}\}, \{{}_j^{t-1}\}),\ {}_l^{t+1} - {}_l^t\Big\rangle \le 0. \tag{A.48}$$

$\forall t \geq T_1$, we can obtain the following inequality,

$$\sum_{l=1}^{|\mathcal{P}^t|}\frac{1}{{}_1}\Big\langle v_{1,l}^{t+1},\ {}_l^{t+1} - {}_l^t\Big\rangle$$
$$\le \sum_{l=1}^{|\mathcal{P}^t|}\Big(\Big\langle \nabla_{{}_l}\tilde{L}_p(\{w_j^{t+1}\}, z^{t+1}, h^{t+1}, \{{}_l^t\}, \{{}_j^t\}) - \nabla_{{}_l}\tilde{L}_p(\{w_j^t\}, z^t, h^t, \{{}_l^t\}, \{{}_j^t\}),\ {}_l^{t+1} - {}_l^t\Big\rangle$$
$$+\Big\langle \nabla_{{}_l}\tilde{L}_p(\{w_j^t\}, z^t, h^t, \{{}_l^t\}, \{{}_j^t\}) - \nabla_{{}_l}\tilde{L}_p(\{w_j^t\}, z^t, h^t, \{{}_l^{t-1}\}, \{{}_j^{t-1}\}),\ v_{1,l}^{t+1}\Big\rangle$$
$$+\Big\langle \nabla_{{}_l}\tilde{L}_p(\{w_j^t\}, z^t, h^t, \{{}_l^t\}, \{{}_j^t\}) - \nabla_{{}_l}\tilde{L}_p(\{w_j^t\}, z^t, h^t, \{{}_l^{t-1}\}, \{{}_j^{t-1}\}),\ {}_l^t - {}_l^{t-1}\Big\rangle\Big). \tag{A.49}$$

Since we have the following equality,

$$\frac{1}{{}_1}\Big\langle v_{1,l}^{t+1},\ {}_l^{t+1} - {}_l^t\Big\rangle = \frac{1}{2\,{}_1}\|{}_l^{t+1} - {}_l^t\|^2 + \frac{1}{2\,{}_1}\|v_{1,l}^{t+1}\|^2 - \frac{1}{2\,{}_1}\|{}_l^t - {}_l^{t-1}\|^2, \tag{A.50}$$

it follows that,

$$\sum_{l=1}^{|P^t|}\Big(\frac{1}{2\eta_1}\|\lambda_l^{t+1}-\lambda_l^t\|^2+\frac{1}{2\eta_1}\|v_{1;l}^{t+1}\|^2-\frac{1}{2\eta_1}\|\lambda_l^t-\lambda_l^{t-1}\|^2\Big)$$
$$\le\sum_{l=1}^{|P^t|}\Big(\frac{L^2}{2b_1^t}\big(\sum_{j=1}^{N}\|w_j^{t+1}-w_j^t\|^2+\|z^{t+1}-z^t\|^2+\|h^{t+1}-h^t\|^2\big)+\frac{b_1^t}{2}\|\lambda_l^{t+1}-\lambda_l^t\|^2$$
$$+\frac{c_1^{t-1}-c_1^t}{2}\big(\|\lambda_l^{t+1}\|^2-\|\lambda_l^t\|^2\big)-\frac{c_1^{t-1}-c_1^t}{2}\|\lambda_l^{t+1}-\lambda_l^t\|^2$$
$$+\frac{1}{2}\|\nabla_{\lambda_l}\tilde{L}_p(\{w_j^t\},z^t,h^t,\{\lambda_l^t\},\{\phi_j^t\})-\nabla_{\lambda_l}\tilde{L}_p(\{w_j^t\},z^t,h^t,\{\lambda_l^{t-1}\},\{\phi_j^{t-1}\})\|^2+\frac{1}{2\eta_1}\|v_{1;l}^{t+1}\|^2$$
$$-\frac{1}{L_1^0+c_1^{t-1}}\|\nabla_{\lambda_l}\tilde{L}_p(\{w_j^t\},z^t,h^t,\{\lambda_l^t\},\{\phi_j^t\})-\nabla_{\lambda_l}\tilde{L}_p(\{w_j^t\},z^t,h^t,\{\lambda_l^{t-1}\},\{\phi_j^{t-1}\})\|^2$$
$$-\frac{c_1^{t-1}L_1^0}{L_1^0+c_1^{t-1}}\|\lambda_l^t-\lambda_l^{t-1}\|^2\Big);\tag{A.51}$$

where $b_1^t>0$. According to the setting that $c_1^0\le L_1^0$, we have $-\frac{c_1^{t-1}L_1^0}{L_1^0+c_1^{t-1}}\le-\frac{c_1^{t-1}L_1^0}{2L_1^0}=-\frac{c_1^{t-1}}{2}\le-\frac{c_1^t}{2}$. Multiplying both sides of the inequality Eq. (A.51) by $\frac{8}{\eta_1 c_1^t}$, we have,

$$\sum_{l=1}^{|P^t|}\Big(\frac{4}{\eta_1^2c_1^t}\|\lambda_l^{t+1}-\lambda_l^t\|^2-\frac{4}{\eta_1}\big(\frac{c_1^{t-1}-c_1^t}{c_1^t}\big)\|\lambda_l^{t+1}\|^2\Big)$$
$$\le\sum_{l=1}^{|P^t|}\Big(\frac{4}{\eta_1^2c_1^t}\|\lambda_l^t-\lambda_l^{t-1}\|^2-\frac{4}{\eta_1}\big(\frac{c_1^{t-1}-c_1^t}{c_1^t}\big)\|\lambda_l^t\|^2+\frac{4b_1^t}{\eta_1c_1^t}\|\lambda_l^{t+1}-\lambda_l^t\|^2-\frac{4}{\eta_1}\|\lambda_l^t-\lambda_l^{t-1}\|^2$$
$$+\frac{4L^2}{\eta_1c_1^tb_1^t}\big(\sum_{j=1}^{N}\|w_j^{t+1}-w_j^t\|^2+\|z^{t+1}-z^t\|^2+\|h^{t+1}-h^t\|^2\big)\Big).\tag{A.52}$$

Setting $b_1^t=\frac{c_1^t}{2}$ in Eq. (A.52) and using the de nition of S_1^t, we have,

$$S_1^{t+1}-S_1^t$$
$$\le\sum_{l=1}^{|P^t|}\Big(\frac{4}{\eta_1}\big(\frac{c_1^{t-2}}{c_1^{t-1}}-\frac{c_1^{t-1}}{c_1^t}\big)\|\lambda_l^t\|^2+\frac{8L^2}{\eta_1(c_1^t)^2}\big(\sum_{j=1}^{N}\|w_j^{t+1}-w_j^t\|^2+\|z^{t+1}-z^t\|^2+\|h^{t+1}-h^t\|^2\big)$$
$$+\big(\frac{2}{\eta_1}+\frac{4}{\eta_1^2}\big(\frac{1}{c_1^{t+1}}-\frac{1}{c_1^t}\big)\big)\|\lambda_l^{t+1}-\lambda_l^t\|^2-\frac{4}{\eta_1}\|\lambda_l^t-\lambda_l^{t-1}\|^2\Big)$$
$$=\sum_{l=1}^{|P^t|}\frac{4}{\eta_1}\big(\frac{c_1^{t-2}}{c_1^{t-1}}-\frac{c_1^{t-1}}{c_1^t}\big)\|\lambda_l^t\|^2+\sum_{l=1}^{|P^t|}\big(\frac{2}{\eta_1}+\frac{4}{\eta_1^2}\big(\frac{1}{c_1^{t+1}}-\frac{1}{c_1^t}\big)\big)\|\lambda_l^{t+1}-\lambda_l^t\|^2$$
$$-\sum_{l=1}^{|P^t|}\frac{4}{\eta_1}\|\lambda_l^t-\lambda_l^{t-1}\|^2+\frac{8|A^t|L^2}{\eta_1(c_1^t)^2}\big(\sum_{j=1}^{N}\|w_j^{t+1}-w_j^t\|^2+\|z^{t+1}-z^t\|^2+\|h^{t+1}-h^t\|^2\big).\tag{A.53}$$

Likewise, according to Eq. (15), we have that,

$$\frac{1}{\eta_2}\big\langle v_{2;l}^{t+1},\phi_j^{t+1}-\phi_j^t\big\rangle$$
$$\le\big\langle\nabla_{\phi_j}\tilde{L}_p(\{w_j^{t+1}\},z^{t+1},h^{t+1},\{\lambda_l^{t+1}\},\{\phi_j^t\})-\nabla_{\phi_j}\tilde{L}_p(\{w_j^t\},z^t,h^t,\{\lambda_l^t\},\{\phi_j^{t-1}\}),\phi_j^{t+1}-\phi_j^t\big\rangle$$
$$=\big\langle\nabla_{\phi_j}\tilde{L}_p(\{w_j^{t+1}\},z^{t+1},h^{t+1},\{\lambda_l^{t+1}\},\{\phi_j^t\})-\nabla_{\phi_j}\tilde{L}_p(\{w_j^t\},z^t,h^t,\{\lambda_l^t\},\{\phi_j^t\}),\phi_j^{t+1}-\phi_j^t\big\rangle$$
$$+\big\langle\nabla_{\phi_j}\tilde{L}_p(\{w_j^t\},z^t,h^t,\{\lambda_l^t\},\{\phi_j^t\})-\nabla_{\phi_j}\tilde{L}_p(\{w_j^t\},z^t,h^t,\{\lambda_l^t\},\{\phi_j^{t-1}\}),v_{2;l}^{t+1}\big\rangle$$
$$+\big\langle\nabla_{\phi_j}\tilde{L}_p(\{w_j^t\},z^t,h^t,\{\lambda_l^t\},\{\phi_j^t\})-\nabla_{\phi_j}\tilde{L}_p(\{w_j^t\},z^t,h^t,\{\lambda_l^t\},\{\phi_j^{t-1}\}),\phi_j^t-\phi_j^{t-1}\big\rangle.\tag{A.54}$$

In addition, since

$$\frac{1}{\eta_2}\big\langle v_{2;l}^{t+1},\phi_j^{t+1}-\phi_j^t\big\rangle=\frac{1}{2\eta_2}\|\phi_j^{t+1}-\phi_j^t\|^2+\frac{1}{2\eta_2}\|v_{2;l}^{t+1}\|^2-\frac{1}{2\eta_2}\|\phi_j^t-\phi_j^{t-1}\|^2,\tag{A.55}$$

it follows that,

$$
\begin{aligned}
&\tfrac{1}{2\ \ 2}\|\ \ {}_j^{t+1}\ \ {}_j^{t}\|^2+\tfrac{1}{2\ \ 2}\|v_{2;l}^{t+1}\|^2\ \ \tfrac{1}{2\ \ 2}\|\ \ {}_j^{t}\ \ {}_j^{t\ \ 1}\|^2\\
&\ \ \tfrac{L^2}{2b_2^t}(\sum_{j=1}\|w_j^{t+1}\ \ w_j^t\|^2+\|z^{t+1}\ \ z^t\|^2+\|h^{t+1}\ \ h^t\|^2)+\tfrac{b_2^t}{2}\|\ \ {}_j^{t+1}\ \ {}_j^{t}\|^2\\
&+\tfrac{c_2^{t\ \ 1}\ \ c_2^t}{2}(\|\ \ {}_j^{t+1}\|^2\ \ \|\ \ {}_j^{t}\|^2)\ \ \tfrac{c_2^{t\ \ 1}\ \ c_2^t}{2}\|\ \ {}_j^{t+1}\ \ {}_j^{t}\|^2\ \ \tfrac{c_2^{t\ \ 1}L_2^0}{L_2^0+c_2^{t\ \ 1}}\|\ \ {}_j^{t}\ \ {}_j^{t\ \ 1}\|^2\\
&+\tfrac{2}{2}\|r\ \ {}_j\,\mathcal{L}_p(\{w_j^t\},z^t,h^t,\{\ \ {}_l^t\},\{\ \ {}_j^t\})\ \ r\ \ {}_j\,\mathcal{L}_p(\{w_j^t\},z^t,h^t,\{\ \ {}_l^t\},\{\ \ {}_j^{t\ \ 1}\})\|^2+\tfrac{1}{2\ \ 2}\|v_{2;l}^{t+1}\|^2\\
&\ \ \tfrac{1}{L_2^0+c_2^{t\ \ 1}}\|r\ \ {}_j\,\mathcal{L}_p(\{w_j^t\},z^t,h^t,\{\ \ {}_l^t\},\{\ \ {}_j^t\})\ \ r\ \ {}_j\,\mathcal{L}_p(\{w_j^t\},z^t,h^t,\{\ \ {}_l^t\},\{\ \ {}_j^{t\ \ 1}\})\|^2:
\end{aligned}
\tag{A.56}
$$

According to the setting $c_2^0\ \ L_2^0$, we have $\frac{c_2^{t\ \ 1}L_2^0}{L_2^0+c_2^{t\ \ 1}}\ \ \frac{c_2^{t\ \ 1}L_2^0}{2L_2^0}=\frac{c_2^{t\ \ 1}}{2}\ \ \frac{c_2^t}{2}$. Multiplying both sides of the inequality Eq. (A.56) by $\frac{8}{2c_2^t}$, we have,

$$
\begin{aligned}
&\tfrac{4}{2^2c_2^t}\|\ \ {}_j^{t+1}\ \ {}_j^{t}\|^2\ \ \tfrac{4}{2}(\tfrac{c_2^{t\ \ 1}\ \ c_2^t}{c_2^t})\|\ \ {}_j^{t+1}\|^2\\
&\ \ \tfrac{4}{2^2c_2^t}\|\ \ {}_j^{t}\ \ {}_j^{t\ \ 1}\|^2\ \ \tfrac{4}{2}(\tfrac{c_2^{t\ \ 1}\ \ c_2^t}{c_2^t})\|\ \ {}_j^{t}\|^2+\tfrac{4b_2^t}{2c_2^t}\|\ \ {}_j^{t+1}\ \ {}_j^{t}\|^2\ \ \tfrac{4}{2}\|\ \ {}_j^{t}\ \ {}_j^{t\ \ 1}\|^2\\
&+\tfrac{4L^2}{2c_2^tb_2^t}(\sum_{j=1}\|w_j^{t+1}\ \ w_j^t\|^2+\|z^{t+1}\ \ z^t\|^2+\|h^{t+1}\ \ h^t\|^2):
\end{aligned}
\tag{A.57}
$$

Setting $b_2^t=\frac{c_2^t}{2}$ in Eq. (A.57) and using the definition of S_2^t, we can obtain,

$$
\begin{aligned}
&S_2^{t+1}\ \ S_2^t\\
&\ \ \sum_{j=1}(\tfrac{4}{2}(\tfrac{c_2^{t\ \ 2}}{c_2^{t\ \ 1}}\ \ \tfrac{c_2^{t\ \ 1}}{c_2^t})\|\ \ {}_j^{t}\|^2+\tfrac{8L^2}{2(c_2^t)^2}(\sum_{j=1}\|w_j^{t+1}\ \ w_j^t\|^2+\|z^{t+1}\ \ z^t\|^2+\sum_{l=1}^{|Q^t|}\|\ \ {}_l^{t+1}\ \ {}_l^{t}\|^2)\\
&+(\tfrac{2}{2}+\tfrac{2}{2^2}(\tfrac{1}{c_2^{t+1}}\ \ \tfrac{1}{c_2^t}))\|\ \ {}_j^{t+1}\ \ {}_j^{t}\|^2\ \ \tfrac{4}{2}\|\ \ {}_j^{t}\ \ {}_j^{t\ \ 1}\|^2)\\
&=\sum_{j=1}\tfrac{4}{2}(\tfrac{c_2^{t\ \ 2}}{c_2^{t\ \ 1}}\ \ \tfrac{c_2^{t\ \ 1}}{c_2^t})\|\ \ {}_j^{t}\|^2+\sum_{j=1}(\tfrac{2}{2}+\tfrac{4}{2^2}(\tfrac{1}{c_2^{t+1}}\ \ \tfrac{1}{c_2^t}))\|\ \ {}_j^{t+1}\ \ {}_j^{t}\|^2\\
&\ \ \sum_{j=1}\tfrac{4}{2}\|\ \ {}_j^{t}\ \ {}_j^{t\ \ 1}\|^2+\tfrac{8NL^2}{2(c_2^t)^2}(\sum_{j=1}\|w_j^{t+1}\ \ w_j^t\|^2+\|z^{t+1}\ \ z^t\|^2+\|h^{t+1}\ \ h^t\|^2):
\end{aligned}
\tag{A.58}
$$

According to the setting about c_1^t and c_2^t, we have $\frac{1}{10}\ \ \frac{1}{c_1^{t+1}}\ \ \frac{1}{c_1^t};\ \frac{2}{10}\ \ \frac{1}{c_2^{t+1}}\ \ \frac{1}{c_2^t};\ 8t\ \ T_1$. Using the definition of F^{t+1} and combining it with Eq. (A.53) and Eq. (A.58), $8t\ \ T_1+\ \ $, we have,

$$
\begin{aligned}
&F^{t+1}\ \ F^t\\
&\ \ (\tfrac{L+1}{2}\ \ \tfrac{1}{{}_w^t}+\tfrac{{}_1|A^t|L^2}{2}+\tfrac{{}_2|Q^{t+1}|L^2}{2}+\tfrac{8|A^t|L^2}{{}_1(c_1^t)^2}+\tfrac{8NL^2}{{}_2(c_2^t)^2})\sum_{j=1}\|w_j^{t+1}\ \ w_j^t\|^2\\
&+(\tfrac{L+3\ \ k_1NL^2}{2}\ \ \tfrac{1}{{}_z^t}+\tfrac{{}_1|A^t|L^2}{2}+\tfrac{{}_2|Q^{t+1}|L^2}{2}+\tfrac{8|A^t|L^2}{{}_1(c_1^t)^2}+\tfrac{8NL^2}{{}_2(c_2^t)^2})\|z^{t+1}\ \ z^t\|^2\\
&+(\tfrac{L+3\ \ k_1NL^2}{2}\ \ \tfrac{1}{{}_h^t}+\tfrac{{}_1|A^t|L^2}{2}+\tfrac{{}_2|Q^{t+1}|L^2}{2}+\tfrac{8|A^t|L^2}{{}_1(c_1^t)^2}+\tfrac{8NL^2}{{}_2(c_2^t)^2})\|h^{t+1}\ \ h^t\|^2\\
&\ \ (\tfrac{1}{10\ \ {}_1}\ \ \tfrac{3\ \ k_1NL^2}{2})\sum_{l=1}^{|P^t|}\|\ \ {}_l^{t+1}\ \ {}_l^{t}\|^2\ \ \tfrac{1}{10\ \ {}_2}\sum_{j=1}\|\ \ {}_j^{t+1}\ \ {}_j^{t}\|^2+\tfrac{c_1^{t\ \ 1}\ \ c_1^t}{2}\sum_{l=1}^{|P^t|}\|\ \ {}_l^{t+1}\|^2\\
&+\tfrac{c_2^{t\ \ 1}\ \ c_2^t}{2}\sum_{j=1}\|\ \ {}_j^{t+1}\|^2+\tfrac{4}{1}(\tfrac{c_1^{t\ \ 2}}{c_1^{t\ \ 1}}\ \ \tfrac{c_1^{t\ \ 1}}{c_1^t})\sum_{l=1}^{|P^t|}\|\ \ {}_l^{t}\|^2+\tfrac{4}{2}(\tfrac{c_2^{t\ \ 2}}{c_2^{t\ \ 1}}\ \ \tfrac{c_2^{t\ \ 1}}{c_2^t})\sum_{j=1}\|\ \ {}_j^{t}\|^2:
\end{aligned}
\tag{A.59}
$$

Next, we will combine Lemma 1, Lemma 2 with Lemma 3 to derive Theorem 1. Firstly, we make some definitions about our problem.

Definition A.3 The stationarity gap at t^{th} iteration is defined as:

$$\nabla G^t = \begin{bmatrix} \left\{ \frac{1}{\eta_w^t}(w_j^t - P_{\mathcal{W}}(w_j^t - \eta_w^t \nabla_{w_j} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))) \right\} \\ \frac{1}{\eta_z^t}(z^t - P_{\mathcal{Z}}(z^t - \eta_z^t \nabla_z L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))) \\ \frac{1}{\eta_h^t}(h^t - P_{\mathcal{H}}(h^t - \eta_h^t \nabla_h L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))) \\ \left\{ \frac{1}{\rho_1}(\lambda_l^t - P_{\Lambda}(\lambda_l^t + \rho_1 \nabla_{\lambda_l} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))) \right\} \\ \left\{ \frac{1}{\rho_2}(\phi_j^t - P_{\Phi}(\phi_j^t + \rho_2 \nabla_{\phi_j} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))) \right\} \end{bmatrix}. \tag{A.60}$$

And we also define:

$$\begin{aligned} (\nabla G^t)_{w_j} &= \tfrac{1}{\eta_w^t}(w_j^t - P_{\mathcal{W}}(w_j^t - \eta_w^t \nabla_{w_j} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))); \\ (\nabla G^t)_{z} &= \tfrac{1}{\eta_z^t}(z^t - P_{\mathcal{Z}}(z^t - \eta_z^t \nabla_z L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))); \\ (\nabla G^t)_{h} &= \tfrac{1}{\eta_h^t}(h^t - P_{\mathcal{H}}(h^t - \eta_h^t \nabla_h L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))); \\ (\nabla G^t)_{\lambda_l} &= \tfrac{1}{\rho_1}(\lambda_l^t - P_{\Lambda}(\lambda_l^t + \rho_1 \nabla_{\lambda_l} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))); \\ (\nabla G^t)_{\phi_j} &= \tfrac{1}{\rho_2}(\phi_j^t - P_{\Phi}(\phi_j^t + \rho_2 \nabla_{\phi_j} L_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))). \end{aligned} \tag{A.61}$$

It follows that,

$$||\nabla G^t||^2 = \sum_{j=1}^{N} ||(\nabla G^t)_{w_j}||^2 + ||(\nabla G^t)_z||^2 + ||(\nabla G^t)_h||^2 + \sum_{l=1}^{|\mathcal{A}^t|} ||(\nabla G^t)_{\lambda_l}||^2 + \sum_{j=1}^{N} ||(\nabla G^t)_{\phi_j}||^2. \tag{A.62}$$

Definition A.4 At t^{th} iteration, the stationarity gap w.r.t. $\widetilde{L}_p(\{w_j\}, z, h, \{\lambda_l\}, \{\phi_j\})$ is defined as:

$$\nabla \widetilde{G}^t = \begin{bmatrix} \left\{ \frac{1}{\eta_w^t}(w_j^t - P_{\mathcal{W}}(w_j^t - \eta_w^t \nabla_{w_j} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))) \right\} \\ \frac{1}{\eta_z^t}(z^t - P_{\mathcal{Z}}(z^t - \eta_z^t \nabla_z \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))) \\ \frac{1}{\eta_h^t}(h^t - P_{\mathcal{H}}(h^t - \eta_h^t \nabla_h \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))) \\ \left\{ \frac{1}{\rho_1}(\lambda_l^t - P_{\Lambda}(\lambda_l^t + \rho_1 \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))) \right\} \\ \left\{ \frac{1}{\rho_2}(\phi_j^t - P_{\Phi}(\phi_j^t + \rho_2 \nabla_{\phi_j} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))) \right\} \end{bmatrix}. \tag{A.63}$$

We further define:

$$\begin{aligned} (\nabla \widetilde{G}^t)_{w_j} &= \tfrac{1}{\eta_w^t}(w_j^t - P_{\mathcal{W}}(w_j^t - \eta_w^t \nabla_{w_j} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))); \\ (\nabla \widetilde{G}^t)_{z} &= \tfrac{1}{\eta_z^t}(z^t - P_{\mathcal{Z}}(z^t - \eta_z^t \nabla_z \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))); \\ (\nabla \widetilde{G}^t)_{h} &= \tfrac{1}{\eta_h^t}(h^t - P_{\mathcal{H}}(h^t - \eta_h^t \nabla_h \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))); \\ (\nabla \widetilde{G}^t)_{\lambda_l} &= \tfrac{1}{\rho_1}(\lambda_l^t - P_{\Lambda}(\lambda_l^t + \rho_1 \nabla_{\lambda_l} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))); \\ (\nabla \widetilde{G}^t)_{\phi_j} &= \tfrac{1}{\rho_2}(\phi_j^t - P_{\Phi}(\phi_j^t + \rho_2 \nabla_{\phi_j} \widetilde{L}_p(\{w_j^t\}, z^t, h^t, \{\lambda_l^t\}, \{\phi_j^t\}))). \end{aligned} \tag{A.64}$$

It follows that,

$$||\nabla \widetilde{G}^t||^2 = \sum_{j=1}^{N} ||(\nabla \widetilde{G}^t)_{w_j}||^2 + ||(\nabla \widetilde{G}^t)_z||^2 + ||(\nabla \widetilde{G}^t)_h||^2 + \sum_{l=1}^{|\mathcal{A}^t|} ||(\nabla \widetilde{G}^t)_{\lambda_l}||^2 + \sum_{j=1}^{N} ||(\nabla \widetilde{G}^t)_{\phi_j}||^2. \tag{A.65}$$

Definition A.5 In our asynchronous algorithm, for the worker j in t^{th} iteration, we define the last iteration where worker j was active as $\hat{t}_j$. And we define the next iteration that worker j will be active as $\overline{t_j}$. For the iteration index set that worker j is active from T_1^{th} to $(T_1+T+\tau)^{th}$ iteration, we define it as $\mathcal{V}_j(T)$. And the i^{th} element in $\mathcal{V}_j(T)$ is defined as $\hat{v}_j(i)$.

Proof of Theorem 1:

Firstly, setting:

$$a_5^t = \frac{4|A^t|(\gamma-2)L^2}{\rho_1(c_1^t)^2} + \frac{4N(\gamma-2)L^2}{\rho_2(c_2^t)^2} + \frac{\rho_2(N-|Q^{t+1}|)L^2}{2} - \frac{1}{2}; \tag{A.66}$$

$$a_6^t = \frac{4|A^t|(\gamma-2)L^2}{\rho_1(c_1^t)^2} + \frac{4N(\gamma-2)L^2}{\rho_2(c_2^t)^2} + \frac{\rho_2(N-|Q^{t+1}|)L^2}{2} - \frac{3k_1NL^2}{2}; \tag{A.67}$$

where γ is a constant which satisfies $\gamma > 2$ and $\frac{4(\gamma-2)L^2}{\rho_1(c_1^0)^2} + \frac{4N(\gamma-2)L^2}{\rho_2(c_2^0)^2} + \frac{\rho_2(N-S)L^2}{2} \geq \max\{\frac{1}{2}; \frac{3k_1NL^2}{2}\}$. It is seen that the $a_5^t; a_6^t$ are nonnegative sequences. Since $\forall t \geq 0$, $|A^0| \geq |A^t|$, $(c_1^0)^2 \geq (c_1^t)^2$, $(c_2^0)^2 \geq (c_2^t)^2$, and we assume that $|Q^{t+1}| = S; \forall t$, thus we have $a_5^0 \geq a_5^t; a_6^0 \geq a_6^t; \forall t$. According to the setting of η_w^t, η_z^t, η_h^t and c_1^t, c_2^t, we have,

$$\frac{L+1}{2} - \frac{1}{\eta_w^t} + \frac{\rho_1|A^t|L^2}{2} + \frac{\rho_2|Q^{t+1}|L^2}{2} + \frac{8|A^t|L^2}{\rho_1(c_1^t)^2} + \frac{8NL^2}{\rho_2(c_2^t)^2} = -a_5^t; \tag{A.68}$$

$$\frac{L+3k_1NL^2}{2} - \frac{1}{\eta_z^t} + \frac{\rho_1|A^t|L^2}{2} + \frac{\rho_2|Q^{t+1}|L^2}{2} + \frac{8|A^t|L^2}{\rho_1(c_1^t)^2} + \frac{8NL^2}{\rho_2(c_2^t)^2} = -a_6^t; \tag{A.69}$$

$$\frac{L+3k_1NL^2}{2} - \frac{1}{\eta_h^t} + \frac{\rho_1|A^t|L^2}{2} + \frac{\rho_2|Q^{t+1}|L^2}{2} + \frac{8|A^t|L^2}{\rho_1(c_1^t)^2} + \frac{8NL^2}{\rho_2(c_2^t)^2} = -a_6^t: \tag{A.70}$$

Combining Eq. (A.68), (A.69), (A.70) with Lemma 8, $\forall t \geq T_1+\tau$, it follows that,

$$\begin{aligned}
&a_5^t\sum_{j=1}^{N}||w_j^{t+1}-w_j^t||^2 + a_6^t||z^{t+1}-z^t||^2 + a_6^t||h^{t+1}-h^t||^2 \\
&+\left(\frac{1}{10\eta_1}-\frac{3k_1NL^2}{2}\right)\sum_{l=1}^{|P^t|}||\lambda_l^{t+1}-\lambda_l^t||^2 + \frac{1}{10\eta_2}\sum_{j=1}^{N}||\phi_j^{t+1}-\phi_j^t||^2 \\
&\leq F^t - F^{t+1} + \frac{c_1^{t-1}-c_1^t}{2}\sum_{l=1}^{|P^t|}||\lambda_l^{t+1}||^2 + \frac{c_2^{t-1}-c_2^t}{2}\sum_{j=1}^{N}||\phi_j^{t+1}||^2 \\
&+\frac{4}{\eta_1}\left(\frac{c_1^{t-2}}{c_1^{t-1}}-\frac{c_1^{t-1}}{c_1^t}\right)\sum_{l=1}^{|P^t|}||\lambda_l^t||^2 + \frac{4}{\eta_2}\left(\frac{c_2^{t-2}}{c_2^{t-1}}-\frac{c_2^{t-1}}{c_2^t}\right)\sum_{j=1}^{N}||\phi_j^t||^2:
\end{aligned} \tag{A.71}$$

Combining the definition of $(\nabla G^t)_{w_j}$ with trigonometric inequality, Cauchy-Schwarz inequality and Assumption 1 and 2, $\forall t \geq T_1+\tau$, we have,

$$||(\nabla G^t)_{w_j}||^2 \leq \frac{2}{\underline{\eta_w}^2}||w_j^{\overline{t_j}}-w_j^t||^2 + 6\tau k_1L^2\left(||z^{t+1}-z^t||^2 + ||h^{t+1}-h^t||^2 + \sum_{l=1}^{|P^t|}||\lambda_l^{t+1}-\lambda_l^t||^2\right): \tag{A.72}$$

Combining the definition of $(\nabla G^t)_z$ with trigonometric inequality and Cauchy-Schwarz inequality, we can obtain the following inequality,

$$||(\nabla G^t)_z||^2 \leq 2L^2\sum_{j=1}^{N}||w_j^{t+1}-w_j^t||^2 + \frac{2}{(\eta_z^t)^2}||z^{t+1}-z^t||^2: \tag{A.73}$$

Likewise, combining the definition of $(\nabla G^t)_h$ with trigonometric inequality and Cauchy-Schwarz inequality, we have that,

$$||(\nabla G^t)_h||^2 \leq 2L^2\left(\sum_{j=1}^{N}||w_j^{t+1}-w_j^t||^2 + ||z^{t+1}-z^t||^2\right) + \frac{2}{(\eta_h^t)^2}||h^{t+1}-h^t||^2: \tag{A.74}$$

Combining the definition of $(\nabla \tilde{L}^t)_{\lambda_l}$ with trigonometric inequality and Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\|(\nabla \tilde{L}^t)_{\lambda_l}\|^2 \\
&\le \frac{3}{\eta_1^2}\|\lambda_l^{t+1}-\lambda_l^t\|^2+3L^2\Big(\sum_{j=1}^{N}\|w_j^{t+1}-w_j^t\|^2+\|z^{t+1}-z^t\|^2+\|h^{t+1}-h^t\|^2\Big)+3(c_1^{t-1}-c_1^t)^2\|\lambda_l^t\|^2 \\
&\le \frac{3}{\eta_1^2}\|\lambda_l^{t+1}-\lambda_l^t\|^2+3L^2\Big(\sum_{j=1}^{N}\|w_j^{t+1}-w_j^t\|^2+\|z^{t+1}-z^t\|^2+\|h^{t+1}-h^t\|^2\Big)+3((c_1^{t-1})^2-(c_1^t)^2)\|\lambda_l^t\|^2.
\end{aligned}
\tag{A.75}
$$

Combining the definition of $(\nabla \tilde{L}^t)_{\phi_j}$ with Cauchy-Schwarz inequality and Assumption 2, we have,

$$
\begin{aligned}
&\|(\nabla \tilde{L}^t)_{\phi_j}\|^2 \\
&\le \frac{3}{\eta_2^2}\|\phi_j^{\overline{t_j}}-\phi_j^t\|^2+3L^2\Big(\sum_{j=1}^{N}\|w_j^{\overline{t_j}}-w_j^t\|^2+\|z^{\overline{t_j}}-z^t\|^2\Big)+3(c_2^{\tilde{t}_j-1}-c_2^{\overline{t_j}-1})^2\|\phi_j^t\|^2 \\
&\le \frac{3}{\eta_2^2}\|\phi_j^{\overline{t_j}}-\phi_j^t\|^2+3L^2\Big(\sum_{j=1}^{N}\|w_j^{\overline{t_j}}-w_j^t\|^2+k_\tau\big(\|z^{t+1}-z^t\|^2+\|h^{t+1}-h^t\|^2+\sum_{l=1}^{|P^t|}\|\lambda_l^{t+1}-\lambda_l^t\|^2\big)\Big) \\
&\quad+3((c_2^{\tilde{t}_j-1})^2-(c_2^{\overline{t_j}-1})^2)\|\phi_j^t\|^2.
\end{aligned}
\tag{A.76}
$$

According to the Definition A.4 as well as Eq. (A.72), (A.73), (A.74), (A.75) and Eq. (A.76), $\forall t \ge T_1+\tau$, we have that,

$$
\begin{aligned}
\|\nabla \tilde{L}^t\|^2 &= \sum_{j=1}^{N}\|(\nabla \tilde{L}^t)_{w_j}\|^2+\|(\nabla \tilde{L}^t)_z\|^2+\|(\nabla \tilde{L}^t)_h\|^2+\sum_{l=1}^{|P^t|}\|(\nabla \tilde{L}^t)_{\lambda_l}\|^2+\sum_{j=1}^{N}\|(\nabla \tilde{L}^t)_{\phi_j}\|^2 \\
&\le \Big(\frac{2}{\eta_w^2}+3NL^2\Big)\sum_{j=1}^{N}\|w_j^{\overline{t_j}}-w_j^t\|^2+(4+3|A^t|)L^2\sum_{j=1}^{N}\|w_j^{t+1}-w_j^t\|^2 \\
&\quad+\Big(\frac{2}{(\eta_z^t)^2}+(2+9k_\tau N+3|A^t|)L^2\Big)\|z^{t+1}-z^t\|^2+\Big(\frac{2}{(\eta_h^t)^2}+(9k_\tau N+3|A^t|)L^2\Big)\|h^{t+1}-h^t\|^2 \\
&\quad+\sum_{l=1}^{|P^t|}\Big(\frac{3}{\eta_1^2}+9k_\tau NL^2\Big)\|\lambda_l^{t+1}-\lambda_l^t\|^2+\sum_{l=1}^{|P^t|}3((c_1^{t-1})^2-(c_1^t)^2)\|\lambda_l^t\|^2 \\
&\quad+\sum_{j=1}^{N}\frac{3}{\eta_2^2}\|\phi_j^{\overline{t_j}}-\phi_j^t\|^2+\sum_{j=1}^{N}3((c_2^{\tilde{t}_j-1})^2-(c_2^{\overline{t_j}-1})^2)\|\phi_j^t\|^2.
\end{aligned}
\tag{A.77}
$$

We set constants d_1, d_2, d_3 as,

$$
d_1=\frac{2k_\tau+(4+3M+3k_\tau N)L^2\eta_w^2}{\eta_w^2(a_5^0)^2}\ge\frac{2k_\tau+(4+3|A^t|+3k_\tau N)L^2\eta_w^2}{\eta_w^2(a_5^t)^2}; \tag{A.78}
$$

$$
d_2=\frac{2+(2+9k_\tau N+3M)L^2\eta_z^2}{\eta_z^2(a_6^0)^2}\ge\frac{2+(2+9k_\tau N+3|A^t|)L^2(\eta_z^t)^2}{(\eta_z^t)^2(a_6^t)^2}; \tag{A.79}
$$

$$
d_3=\frac{2+(9k_\tau N+3M)L^2\eta_h^2}{\eta_h^2(a_6^0)^2}\ge\frac{2+(9k_\tau N+3|A^t|)L^2(\eta_h^t)^2}{(\eta_h^t)^2(a_6^t)^2}; \tag{A.80}
$$

where k_τ, η_z and η_h are positive constants. $\eta_z=\frac{2}{L+\rho_1 ML^2+\rho_2 NL^2+8(\frac{M\gamma^2}{\eta_1 c_1^2}+\frac{N\gamma^2}{\eta_2 c_2^2})}\le\eta_z^t$ and $\eta_h=\frac{2}{L+\rho_1 ML^2+\rho_2 NL^2+8(\frac{M\gamma^2}{\eta_1 c_1^2}+\frac{N\gamma^2}{\eta_2 c_2^2})}\le\eta_h^t$, $\forall t$. Thus, combining Eq. (A.77) with Eq. (A.78), (A.79), (A.80), $\forall t \ge T_1+\tau$, we have,

$$
\begin{aligned}
\|\nabla \tilde{L}^t\|^2 &\le \sum_{j=1}^{N}d_1(a_5^t)^2\|w_j^{t+1}-w_j^t\|^2+d_2(a_6^t)^2\|z^{t+1}-z^t\|^2+d_3(a_6^t)^2\|h^{t+1}-h^t\|^2 \\
&\quad+\sum_{l=1}^{|P^t|}\Big(\frac{3}{\eta_1^2}+9k_\tau NL^2\Big)\|\lambda_l^{t+1}-\lambda_l^t\|^2+\sum_{l=1}^{|P^t|}3((c_1^{t-1})^2-(c_1^t)^2)\|\lambda_l^t\|^2+\sum_{j=1}^{N}\frac{3}{\eta_2^2}\|\phi_j^{\overline{t_j}}-\phi_j^t\|^2 \\
&\quad+\sum_{j=1}^{N}3((c_2^{\tilde{t}_j-1})^2-(c_2^{\overline{t_j}-1})^2)\|\phi_j^t\|^2+\Big(\frac{2}{\eta_w^2}+3NL^2\Big)\sum_{j=1}^{N}\|w_j^{\overline{t_j}}-w_j^t\|^2 \\
&\quad-\Big(\frac{2k_\tau}{\eta_w^2}+3k_\tau NL^2\Big)\sum_{j=1}^{N}\|w_j^{t+1}-w_j^t\|^2.
\end{aligned}
\tag{A.81}
$$

Let d_4^t denote a nonnegative sequence:

$$d_4^t = \frac{1}{\max\{d_1 a_5^t; d_2 a_6^t; d_3 a_6^t; \frac{\frac{30}{\eta_1}+90\eta_1 k\tau_1 N L^2}{1-15\eta_1 k\tau_1 N L^2}; \frac{30}{\eta_2}\}}: \tag{A.82}$$

It is seen that $d_4^0 \geq d_4^t, \forall t \geq 0$. And we denote the lower bound of d_4^t as $\underline{d_4}$, it appears that $d_4^t \geq \underline{d_4} > 0, \forall t \geq 0$. And we set the constant k satisfies $k \geq \frac{d_4^0(\frac{2}{\eta_w^2}+3\tau NL^2)}{\underline{d_4}(\frac{2}{\eta_w^2}+3\tau NL^2)}$, where $\underline{\eta_w}$ is the step-size in terms of w_j in the first iteration (it is seen that $\underline{\eta_w} \geq \eta_w^t, \forall t$). Then, $\forall t \geq T_1 + \tau$, we can obtain the following inequality from Eq. (A.81) and Eq. (A.82):

$$\begin{aligned} d_4^t\|\nabla \tilde{G}^t\|^2 \leq\ & a_5^t \sum_{j=1}^{N} \|w_j^{t+1} - w_j^t\|^2 + a_6^t\|z^{t+1} - z^t\|^2 + a_6^t\|h^{t+1} - h^t\|^2 \\ & +(\frac{1}{10\eta_1} - \frac{3k\tau_1 NL^2}{2})\sum_{l=1}^{|P^t|}\|\lambda_l^{t+1} - \lambda_l^t\|^2 + \frac{1}{10\eta_2}\sum_{j=1}^{N}\|\theta_j^{\overline{t_j}} - \theta_j^t\|^2 \\ & +3d_4^t((c_1^{t-1})^2 - (c_1^t)^2)\sum_{l=1}^{|P^t|}\|\lambda_l^t\|^2 + 3d_4^t\sum_{j=1}^{N}((c_2^{\tilde{t}_j-1})^2 - (c_2^{\overline{t_j}-1})^2)\|\theta_j^t\|^2 \\ & + d_4^t(\frac{2}{\eta_w^2}+3\tau NL^2)\sum_{j=1}^{N}\|w_j^{\overline{t_j}} - w_j^t\|^2 - d_4^t(\frac{2k}{\eta_w^2}+3k\tau NL^2)\sum_{j=1}^{N}\|w_j^{t+1} - w_j^t\|^2: \end{aligned} \tag{A.83}$$

Combining Eq. (A.83) with Eq. (A.71) and according to the setting $\|\lambda\|^2 \leq \sigma_1^2$, $\|\theta_j^t\|^2 \leq \sigma_2^2$ (where $\sigma_1^2 = \sigma_3^2$, $\sigma_2^2 = p\sigma_4^2$) and $d_4^0 \geq d_4^t \geq \underline{d_4}$, thus, $\forall t \geq T_1 + \tau$, we have,

$$\begin{aligned} & d_4^t\|\nabla \tilde{G}^t\|^2 \\ & \leq F^t - F^{t+1} + \frac{c_1^{t-1} - c_1^t}{2}M\sigma_1^2 + \frac{c_2^{t-1} - c_2^t}{2}N\sigma_2^2 + \frac{4}{\eta_1}(\frac{c_1^{t-2}}{c_1^{t-1}} - \frac{c_1^{t-1}}{c_1^t})M\sigma_1^2 \\ & + \frac{4}{\eta_2}(\frac{c_2^{t-2}}{c_2^{t-1}} - \frac{c_2^{t-1}}{c_2^t})N\sigma_2^2 + 3d_4^0((c_1^{t-1})^2 - (c_1^t)^2)M\sigma_1^2 + 3d_4^0\sum_{j=1}^{N}((c_2^{\tilde{t}_j-1})^2 - (c_2^{\overline{t_j}-1})^2)\sigma_2^2 \\ & + \frac{1}{10\eta_2}\sum_{j=1}^{N}\|\theta_j^{\overline{t_j}} - \theta_j^t\|^2 - \frac{1}{10\eta_2}\sum_{j=1}^{N}\|\theta_j^{t+1} - \theta_j^t\|^2 \\ & + d_4^0(\frac{2}{\eta_w^2} + 3\tau NL^2)\sum_{j=1}^{N}\|w_j^{\overline{t_j}} - w_j^t\|^2 - \underline{d_4}(\frac{2k}{\eta_w^2} + 3k\tau NL^2)\sum_{j=1}^{N}\|w_j^{t+1} - w_j^t\|^2: \end{aligned} \tag{A.84}$$

Denoting $\tilde{T}(\varepsilon)$ as $\tilde{T}(\varepsilon) = \min\{t \mid \|\nabla \tilde{G}^{T_1+\tau+t}\| \leq \frac{\varepsilon}{2}, t \geq 0\}$. Summing up Eq. (A.84) from $t = T_1 + \tau$ to $t = T_1 + \tau + \tilde{T}(\varepsilon)$, we have,

$$\begin{aligned} & \sum_{t=T_1+\tau}^{T_1+\tau+\tilde{T}(\varepsilon)} d_4^t\|\nabla \tilde{G}^t\|^2 \\ & \leq F^{T_1+\tau} - \underline{L} + \frac{4}{\eta_1}(\frac{c_1^{T_1+\tau-2}}{c_1^{T_1+\tau-1}} + \frac{c_1^{T_1+\tau-1}}{c_1^{T_1+\tau}})M\sigma_1^2 + \frac{c_1^{T_1+\tau-1}}{2}M\sigma_1^2 + \frac{7}{2\eta_1}M\sigma_3^2 + 3d_4^0(c_1^0)^2M\sigma_1^2 \\ & + \frac{4}{\eta_2}(\frac{c_1^{T_1+\tau-2}}{c_1^{T_1+\tau-1}} + \frac{c_1^{T_1+\tau-1}}{c_1^{T_1+\tau}})N\sigma_2^2 + \frac{c_2^{T_1+\tau-1}}{2}N\sigma_2^2 + \frac{7}{2\eta_2}N\sigma_4^2 + \sum_{j=1}^{N}\sum_{t=T_1+\tau}^{T_1+\tau+\tilde{T}(\varepsilon)} 3d_4^0((c_2^{\tilde{t}_j-1})^2 - (c_2^{\overline{t_j}-1})^2)\sigma_2^2 \\ & + \frac{c_1^{T_1+\tau}}{2}M\sigma_1^2 + \frac{c_2^{T_1+\tau}}{2}N\sigma_2^2 + \frac{1}{10\eta_2}\sum_{j=1}^{N}\sum_{t=T_1+\tau}^{T_1+\tau+\tilde{T}(\varepsilon)}\|\theta_j^{\overline{t_j}} - \theta_j^t\|^2 - \frac{1}{10\eta_2}\sum_{j=1}^{N}\sum_{t=T_1+\tau}^{T_1+\tau+\tilde{T}(\varepsilon)}\|\theta_j^{t+1} - \theta_j^t\|^2 \\ & + d_4^0(\frac{2}{\eta_w^2} + 3\tau NL^2)\sum_{j=1}^{N}\sum_{t=T_1+\tau}^{T_1+\tau+\tilde{T}(\varepsilon)}\|w_j^{\overline{t_j}} - w_j^t\|^2 - \underline{d_4}(\frac{2k}{\eta_w^2} + 3k\tau NL^2)\sum_{j=1}^{N}\sum_{t=T_1+\tau}^{T_1+\tau+\tilde{T}(\varepsilon)}\|w_j^{t+1} - w_j^t\|^2; \end{aligned} \tag{A.85}$$

where $_3 = \max \{ \| _1 \quad _2 \| \| _1; \ _2 \in \ \}$, $_4 = \max \{ \| _1 \quad _2 \| \| _1; \ _2 \in \ \}$ and $L =$ $\min_{\{w_j \in W\};z \in Z;h \in H;\{_l \in \};\{_j \in \}} L_p(\{w_j\};z;h;\{_l\};\{_j\})$, which satisfy that $\forall t \quad T_1 + \ ,$

$$F^{t+1} \quad L \quad \frac{4}{_1} \frac{c_1^{T_1+} \quad 1}{c_1^{T_1+}} M \quad _1^2 \quad \frac{4}{_2} \frac{c_2^{T_1+} \quad 1}{c_2^{T_1+}} N \quad _2^2 \quad \frac{7}{2 \ _1} M \quad _3^2 \quad \frac{7}{2 \ _2} N \quad _4^2 \quad \frac{c_1^{T_1+}}{2} M \quad _1^2 \quad \frac{c_2^{T_1+}}{2} N \quad _2^2 : \tag{A.86}$$

For each worker j, the iterations between the last iteration and the next iteration where it is active is no more than , i.e., $\overline{t_j} \quad \tilde{t}_j$, we have,

$$\begin{aligned} &\sum_{t=T_1+}^{T_1+\tilde{T}(")} 3d_4^0((c_2^{\tilde{t}_j} \quad 1)^2 \quad (c_2^{\overline{t_j}} \quad 1)^2) \quad _2^2 \\ & \qquad \sum_{\substack{\hat{v}_j(i) \in V_j(\tilde{T}(")); \\ T_1+ \quad \hat{v}_j(i) \quad T_1+\tilde{T}(")}} 3d_4^0((c_2^{\hat{v}_j(i)} \quad 1)^2 \quad (c_2^{\hat{v}_j(i+1)} \quad 1)^2) \quad _2^2 \\ & \quad 3 \ d_4^0(c_2^0)^2 \quad _2^2 : \end{aligned} \tag{A.87}$$

Since the idle workers do not update their variables in each iteration, for any t that satisfies $\hat{v}_j(i \quad 1)$ $t < \hat{v}_j(i)$, we have $_j^t = \ _j^{\hat{v}_j(i) \quad 1}$. And for $t \notin V_j(T)$, we have $\| _j^t \quad _j^{t \quad 1}\|^2 = 0$. Combing with $\hat{v}_j(i) \quad \hat{v}_j(i \quad 1)$, we can obtain that,

$$\begin{aligned} \sum_{j=1}^{N} \sum_{t=T_1+}^{T_1+\tilde{T}(")} \| _j^{\overline{t_j}} \quad _j^t \|^2 \quad & \sum_{j=1}^{N} \sum_{\substack{\hat{v}_j(i) \in V_j(\tilde{T}(")); \\ T_1+ \ +1 \quad \hat{v}_j(i)}} \| _j^{\hat{v}_j(i)} \quad _j^{\hat{v}_j(i) \quad 1}\|^2 \\ &= \sum_{j=1}^{N} \sum_{t=T_1+}^{T_1+\tilde{T}(")} \| _j^{t+1} \quad _j^t \|^2 + \sum_{j=1}^{N} \sum_{t=T_1+\tilde{T}(")+1}^{T_1+\tilde{T}(")+ \quad 1} \| _j^{t+1} \quad _j^t \|^2 \\ & \sum_{j=1}^{N} \sum_{t=T_1+}^{T_1+\tilde{T}(")} \| _j^{t+1} \quad _j^t \|^2 + 4 \quad (\quad 1) N \quad _2^2 : \end{aligned} \tag{A.88}$$

Similarly, for any t that satisfies $\hat{v}_j(i \quad 1) \quad t < \hat{v}_j(i)$, we have $w_j^t = w_j^{\hat{v}_j(i) \quad 1}$. And for $t \notin V_j(T)$, we have $\|w_j^t \quad w_j^{t \quad 1}\|^2 = 0$. Combing with $\hat{v}_j(i) \quad \hat{v}_j(i \quad 1)$, we can obtain,

$$\begin{aligned} \sum_{j=1}^{N} \sum_{t=T_1+}^{T_1+\tilde{T}(")} \|w_j^{\overline{t_j}} \quad w_j^t\|^2 \quad & \sum_{j=1}^{N} \sum_{\substack{\hat{v}_j(i) \in V_j(\tilde{T}(")); \\ T_1+ \ +1 \quad \hat{v}_j(i)}} \|w_j^{\hat{v}_j(i)} \quad w_j^{\hat{v}_j(i) \quad 1}\|^2 \\ &= \sum_{j=1}^{N} \sum_{t=T_1+}^{T_1+\tilde{T}(")} \|w_j^{t+1} \quad w_j^t\|^2 + \sum_{j=1}^{N} \sum_{t=T_1+\tilde{T}(")+1}^{T_1+\tilde{T}(")+ \quad 1} \|w_j^{t+1} \quad w_j^t\|^2 \\ & \sum_{j=1}^{N} \sum_{t=T_1+}^{T_1+\tilde{T}(")} \|w_j^{t+1} \quad w_j^t\|^2 + 4 \quad (\quad 1) pN \quad _1^2 : \end{aligned} \tag{A.89}$$

It follows from Eq. (A.85), (A.87), (A.88), (A.89) that,

$$\begin{aligned} &\sum_{t=T_1+}^{T_1+\tilde{T}(")} d_4^t \|r \ \tilde{G}^t\|^2 \\ & \quad F^{T_1+} \quad L + \tfrac{4}{_1}(\tfrac{c_1^{T_1+} \quad 2}{c_1^{T_1+} \quad 1} + \tfrac{c_1^{T_1+} \quad 1}{c_1^{T_1+}}) M \quad _1^2 + \tfrac{c_1^{T_1+} \quad 1}{2} M \quad _1^2 + \tfrac{7}{2 \ _1} M \quad _3^2 + 3 \, d_4^0 (c_1^0)^2 M \quad _1^2 \\ & + \tfrac{4}{_2}(\tfrac{c_1^{T_1+} \quad 2}{c_1^{T_1+} \quad 1} + \tfrac{c_1^{T_1+} \quad 1}{c_1^{T_1+}}) N \quad _2^2 + \tfrac{c_2^{T_1+} \quad 1}{2} N \quad _2^2 + \tfrac{7}{2 \ _2} N \quad _4^2 + 3 \ d_4^0 (c_2^0)^2 N \quad _2^2 \\ & + \tfrac{c_1^{T_1+}}{2} M \quad _1^2 + \tfrac{c_2^{T_1+}}{2} N \quad _2^2 + (\tfrac{2N \quad _2^2}{5 \ _2} + 4 d_4^0 (\tfrac{2}{\ _w^2} + 3 N L^2) pN \quad _1^2 \quad)(\quad 1) \\ & = \ d + k_d(\quad 1); \end{aligned} \tag{A.90}$$

where d and k_d are constants. And constant d_5^t is given by,

$$
\begin{aligned}
d_5 &= \max\{\frac{d_1}{a_6^0}; \frac{d_2}{a_5^0}; \frac{d_3}{a_5^0}; \frac{\frac{30}{1}+90 \quad {}_1\, k\, {}_1 NL^2}{(1 \quad 15\ {}_1\, k\, {}_1 NL^2)a_5^0 a_6^0}; \frac{30}{{}_2 a_5^0 a_6^0}\} \\
&\max\{\frac{d_1}{a_6^t}; \frac{d_2}{a_5^t}; \frac{d_3}{a_5^t}; \frac{\frac{30}{1}+90 \quad {}_1\, k\, {}_1 NL^2}{(1 \quad 15\ {}_1\, k\, {}_1 NL^2)a_5^t a_6^t}; \frac{30}{{}_2 a_5^t a_6^t}\} \\
&= \frac{1}{d_4^t a_5^t a_6^t}:
\end{aligned}
\tag{A.91}
$$

Thus, we can obtain that,

$$
\sum_{t=T_1+}^{T_1+\hat{T}('')} \frac{1}{d_5 a_5^t a_6^t} ||r \quad \mathcal{G}^{T_1+\hat{T}('')}||^2 \quad \sum_{t=T_1+}^{T_1+\hat{T}('')} \frac{1}{d_5 a_5^t a_6^t} ||r \quad \mathcal{G}^t||^2 \quad \sum_{t=T_1+}^{T_1+\hat{T}('')} d_4^t ||r \quad \mathcal{G}^t||^2 \quad d + k_d(\quad 1):
\tag{A.92}
$$

And it follows from Eq. (A.92) that,

$$
||r \quad \mathcal{G}^{T_1+\hat{T}('')}||^2 \quad \frac{(d + k_d(\quad 1))d_5}{\sum_{t=T_1+}^{T_1+\hat{T}('')} \frac{1}{a_5^t a_6^t}}:
\tag{A.93}
$$

According to the setting of c_1^t, c_2^t and Eq. (A.66), (A.67), we have,

$$
\frac{1}{a_5^t a_6^t} \quad \frac{1}{(4(\quad 2)L^2(M \quad {}_1 + N \quad {}_2)(t+1)^{\frac{1}{3}} + \frac{{}_2(N \quad S)L^2}{2})^2}:
\tag{A.94}
$$

Summing up $\frac{1}{a_5^t a_6^t}$ from $t = T_1 +$ to $t = T_1 + \hat{T}('')$, it follows that,

$$
\begin{aligned}
\sum_{t=T_1+}^{T_1+\hat{T}('')} \frac{1}{a_5^t a_6^t} &\quad \sum_{t=T_1+}^{T_1+\hat{T}('')} \frac{1}{(4(\quad 2)L^2(M \quad {}_1 + N \quad {}_2)(t+1)^{\frac{1}{3}} + \frac{{}_2(N \quad S)L^2}{2})^2} \\
&\quad \sum_{t=T_1+}^{T_1+\hat{T}('')} \frac{1}{(4(\quad 2)L^2(M \quad {}_1 + N \quad {}_2)(t+1)^{\frac{1}{3}} + \frac{{}_2(N \quad S)L^2}{2}(t+1)^{\frac{1}{3}})^2} \\
&\quad \frac{(T_1+\hat{T}(''))^{\frac{1}{3}} \quad (T_1 + \quad)^{\frac{1}{3}}}{(4(\quad 2)L^2(M \quad {}_1 + N \quad {}_2) + \frac{{}_2(N \quad S)L^2}{2})^2}:
\end{aligned}
\tag{A.95}
$$

The second inequality in Eq. (A.95) is due to that $8t \quad T_1 +$, we have,

$$
4(\quad 2)L^2(M \quad {}_1 + N \quad {}_2)(t+1)^{\frac{1}{3}} + \frac{{}_2(N \quad S)L^2}{2} \quad (4(\quad 2)L^2(M \quad {}_1 + N \quad {}_2) + \frac{{}_2(N \quad S)L^2}{2})(t+1)^{\frac{1}{3}}:
\tag{A.96}
$$

The last inequality in Eq. (A.95) follows from the fact that $\sum_{t=T_1+}^{T_1+\hat{T}('')} \frac{1}{(t+1)^{\frac{2}{3}}} \quad (T_1+\hat{T}(''))^{\frac{1}{3}} \quad (T_1 + \quad)^{\frac{1}{3}}$.

Thus, plugging Eq. (A.95) into Eq. (A.93), we can obtain:

$$
||r \quad \mathcal{G}^{T_1+\hat{T}('')}||^2 \quad \frac{(d + k_d(\quad 1))d_5}{\sum_{t=T_1+}^{T_1+\hat{T}('')} \frac{1}{a_5^t a_6^t}} \quad \frac{(4(\quad 2)L^2(M \quad {}_1 + N \quad {}_2) + \frac{{}_2(N \quad S)L^2}{2})^2 (d + k_d(\quad 1))d_5}{(T_1 + \hat{T}(''))^{\frac{1}{3}} \quad (T_1 + \quad)^{\frac{1}{3}}}:
\tag{A.97}
$$

According to the definition of $\hat{T}('')$, we have:

$$
T_1 + \hat{T}('') \quad (\frac{4(4(\quad 2)L^2(M \quad {}_1 + N \quad {}_2) + \frac{{}_2(N \quad S)L^2}{2})^2 (d + k_d(\quad 1))d_5}{''^2} + (T_1 + \quad)^{\frac{1}{3}})^3:
\tag{A.98}
$$

Combining the definition of ∇G^t and $\nabla \widetilde{G}^t$ with trigonometric inequality, we then get:

$$||\nabla G^t|| - ||\nabla \widetilde{G}^t|| \leq ||\nabla G^t - \nabla \widetilde{G}^t|| \leq \sqrt{\sum_{l=1}^{|A^t|} ||c_1^{t-1} - \lambda_l^t||^2 + \sum_{j=1}^{N} ||c_2^{t-1} - \lambda_j^t||^2}. \quad \text{(A.99)}$$

Denoting constant d_6 as $d_6 = 4(\gamma - 2)L^2(M\alpha_1 + N\alpha_2)$. If $t > (\frac{4M\alpha_1^2}{\eta_1^2} + \frac{4N\alpha_2^2}{\eta_2^2})^3 \frac{1}{\varepsilon^6}$, then we have $\sqrt{\sum_{l=1}^{|A^t|} ||c_1^{t-1} - \lambda_l^t||^2 + \sum_{j=1}^{N} ||c_2^{t-1} - \lambda_j^t||^2} \leq \frac{\varepsilon}{2}$. Combining it with Eq. (A.98), we can conclude that there exists a

$$T(\varepsilon) \sim O(\max\{(\frac{4M\alpha_1^2}{\eta_1^2} + \frac{4N\alpha_2^2}{\eta_2^2})^3 \frac{1}{\varepsilon^6}, (\frac{4(d_6 + \frac{\rho_2(N-S)L^2}{\eta_2^2})^2 (d + k_d(\tau - 1)) d_5}{\varepsilon^2} + (T_1 + \tau)^{\frac{1}{3}})^3\}), \quad \text{(A.100)}$$

such that $||\nabla G^t|| - ||\nabla \widetilde{G}^t|| + \sqrt{\sum_{l=1}^{|A^t|} ||c_1^{t-1} - \lambda_l^t||^2 + \sum_{j=1}^{N} ||c_2^{t-1} - \lambda_j^t||^2} \leq \varepsilon$, which concludes our proof.

B Time Efficiency Comparison

In a distributed communication network, the communication and computation delays of workers are inevitable. Due to differences in system configuration, communication and computation delays vary across different workers, the existence of lagging workers (*i.e.*, stragglers and stale workers) is inevitable. For synchronous algorithm, it will lead to idling and wastage of computing resources since the master only updates the variables after receiving updates from all workers (as illustrated in Figure B1). Different from the synchronous algorithm, the asynchronous algorithm allows the master updates the variables whenever it receives updates from a subset of workers, which is more efficient. In this section, we compare the time for our asynchronous and synchronous algorithms to return an ε-stationary point.

Fact 1 Let T_1 and $T_1(\varepsilon)$ denote the convergence time and iterations for the proposed asynchronous algorithm. Let T_2 and $T_2(\varepsilon)$ denote the convergence time and iterations for the synchronous algorithm, we have,

$$\frac{T_1}{T_2} = \frac{T_1(\varepsilon) S}{T_2(\varepsilon)(\frac{\hat{d}}{\hat{d}_1} + \frac{\hat{d}}{\hat{d}_2} + \cdots + \frac{\hat{d}}{\hat{d}_N})}; \quad \text{(B.101)}$$

where $\hat{d}$ is the maximum (computation + communication) delay of all workers.

Proof of Fact 1:

In this part, we do not consider the delays of master. Suppose that there are N workers in a distributed system and the number of active workers is S. For brevity, we assume the delay for each work remains the same during the iteration. Let $[\hat{d}_1, \hat{d}_2, \cdots, \hat{d}_N] \in R^N$ denote the (computation + communication) delay for N workers. And we define the maximum delay of all workers as $\hat{d}$. For the j^{th} worker, it has communicated with the master $\frac{T_1}{\hat{d}_j}$ times during time T_1. Thus, for time T_1, it satisfies that,

$$\frac{T_1}{\hat{d}_1} + \frac{T_1}{\hat{d}_2} + \cdots + \frac{T_1}{\hat{d}_N} = T_1(\varepsilon) S. \quad \text{(B.102)}$$

For the synchronous algorithm, the time T_2 needs to satisfy that:

$$T_2 = T_2(\varepsilon) \hat{d}. \quad \text{(B.103)}$$

Thus, we have that,

$$\frac{T_1}{T_2} = \frac{T_1(\varepsilon) S}{T_2(\varepsilon)(\frac{\hat{d}}{\hat{d}_1} + \frac{\hat{d}}{\hat{d}_2} + \cdots + \frac{\hat{d}}{\hat{d}_N})}. \quad \text{(B.104)}$$

For the special case that the asynchronous algorithm degrades to the synchronous algorithm, *i.e.*, the master is required to update its parameters only after receiving the updates from all workers, we have $S = N$, $T_1(\varepsilon) = T_2(\varepsilon)$ and $\hat{d}_j = \hat{d}, \forall j = 1, \cdots, N$ since all workers are required to wait the slowest worker. Back to Eq. (B.104), we can obtain that $\frac{T_1}{T_2} = 1$.

Figure B1: The illustration of synchronous and asynchronous algorithms. represents the number of iterations. In the asynchronous algorithm (at the bottom), the master begins to update its parameters after receiving the update from one worker.

(a) Clean images whose labels are T-shirt.

(b) Attacked images whose target labels are Pullover.

Figure C1: Backdoor attacks on Fashion MNIST dataset. Through adding triggers on local patch of clean images, the attacked images are misclassi ed as the target labels.

C Experiments

In this section, we present the detailed results of our experiments. We rst give a detailed description of the datasets and baseline methods used in our experiments.

C.1 Datasets and Baseline Methods

In this section, we provide a detailed introduction to datasets and baseline methods. The number of workers and categories of every dataset are summarized in Table C1.

Datasets

1. SHL dataset The SHL dataset was collected using four cellphones on four body locations where people usually carry cellphones. The SHL dataset provides multimodal locomotion and transportation data collected in real-world settings using eight various modes of transportation. We separated the data into six workers with varied proportions based on the four body locations of smartphones to imitate the different tendencies of workers (users) in positioning cellphones.
2. Person Activity dataset Data contains recordings of ve participants performing eleven different activities. Each participant wears four sensors in four different body locations (ankle left, ankle right, belt, and chest) while performing the activities. Each participant corresponds to one worker in the experiment.

Table C1: The number of workers and categories of datasets

	SHL	Person Activity	SC-MA	Fashion MNIST
Number of workers	6	5	15	3
Number of categories	8	11	7	3

Table C2: Model structure that used for SHL dataset.

No.	Layer type	Number of neurons	Activation
1	Fully-connected	96	ReLU
2	Fully-connected	48	ReLU
3	Fully-connected	24	ReLU
4	Output	8	Softmax

3. Single Chest-Mounted Accelerometer dataset: Data was collected from fteen participants engaged in seven distinct activities. Each participant (worker) wears an accelerometer mounted on the chest.
4. Fashion MNIST: Similar to MNIST, Fashion MNIST is a dataset where images are grouped into ten categories of clothing. The subset of the data labeled with Pullover, Shirt and T-shirt are extracted as three workers and each worker consists of one class of clothing.

Baseline Methods

1. Ind_j : It learns the model from an individual worker j.
2. Mix_{Even} : It learns the model from all workers with even weights using the proposed distributed algorithm.
3. FedAvg: It learns the model from all workers with even weights. It aggregates the local model parameters from workers through using model averaging.
4. AFL : It aims to address the fairness issues in federated learning. AFL adopts the strategy that alternately update the model parameters and the weight of each worker through alternating projected gradient descent/ascent.
5. DRFA-Prox: It aims to mitigate the data heterogeneity issue in federated learning. Compared with AFL, it is communication-ef cient which requires fewer communication rounds. Moreover, it leverages the prior distribution and introduces it as a regularizer in the objective function.
6. ASPIRE-EASE(-): The proposed ASPIRE-EASE without asynchronous setting.
7. ASPIRE-CP: The proposed ASPIRE with cutting plane method.
8. $ASPIRE\text{-}EASE_{per}$: The proposed ASPIRE-EASE with periodic communication.

C.2 Training Details

In our empirical studies, since the downstream tasks are multi-class classi cation, the cross entropy loss is used on each worker (i.e., $L_j( )$; 8j). For SHL, Person Activity and SM-AC datasets, we adopt the deep multilayer perceptron [49] as the base model. Speci cally, we exhibit the model structures that are used for SHL, Person Activity and SM-AC datasets in Table C2, Table C3 and Table C4. And we use the same logistic regression model as in [35, 16] for the Fashion MNIST dataset. In the experiments, we use the SGD optimizer for model training, and we implement our model with PyTorch and conduct all the experiments on a server with two TITAN V GPUs.

C.3 Additional Results

We rst show the detailed experiment settings about robustness against malicious attacks. We conduct experiments in the setting where there are malicious workers which attempt to mislead the model

Table C3: Model structure that used for Peson Activity dataset.

No.	Layer type	Number of neurons	Activation
1	Fully-connected	64	ReLU
2	Fully-connected	32	ReLU
3	Fully-connected	16	ReLU
4	Output	11	Softmax

Table C4: Model structure that used for SC-MA dataset.

No.	Layer type	Number of neurons	Activation
1	Fully-connected	32	ReLU
2	Fully-connected	16	ReLU
3	Output	7	Softmax

training process. The backdoor attack [1, 48] is adopted in the experiment which aims to bury the backdoor during the training phase of the model. The buried backdoor will be activated by the preset trigger. When the backdoor is not activated, the attacked model performs normally to other local models. When the backdoor is activated, the output of the attack model is misled as the target label which is pre-specified by the attacker. In the experiment, one worker is chosen as the malicious worker. We add triggers to a small part of the data and change their primal labels to target labels (e.g, triggers are added on the local patch of clean images on the Fashion MNIST dataset, which are shown in Figure C1). Furthermore, the malicious worker can purposefully raise the training loss to mislead the master. To evaluate the model's robustness against malicious attacks, following [14], we calculate the success attack rate of the backdoor attacks. The success attack rate can be calculated by checking how many instances in the backdoor dataset can be misled into the target labels. The lower success attack rate indicates better robustness against backdoor attacks. The success attack rates of different models on three datasets are reported in Table 2. In Table 2, we observe that AFL can be attacked easily since it could assign higher weights to malicious workers. Compared to AFL, FedAvg and Mix_{Even} achieve relatively lower success attack rates since they assign equal weights to the malicious workers and other workers. DRFA-Prox can achieve even lower success attack rates since it can leverage the prior distribution to assign lower weights for malicious workers. The proposed ASPIRE-EASE achieves the lowest success attack rates since it can leverage the prior distribution more effectively. Specifically, it will assign lower weights to malicious workers with tight theoretical guarantees.

We also report additional experiment results on SHL and Fashion MNIST datasets. We first show that the proposed ASPIRE-EASE can flexibly control the level of robustness by adjusting λ, which is presented in Figure C2. It is seen that the robustness of ASPIRE-EASE can be gradually enhanced when ρ increases. Next, the comparison of convergence speed by considering different communication and computation delays for each worker is exhibited in Figure C3. We can observe that the proposed ASPIRE-EASE is generally the most efficient since the ASPIRE is an asynchronous algorithm and the proposed EASE is effective. Finally, to further demonstrate the efficiency of EASE, we compare ASPIRE-EASE with ASPIRE-CP concerning the number of cutting planes used during the training. As shown in Theorem 1, a smaller number of cutting planes (which corresponds to a smaller M) will need fewer iterations to achieve convergence. In Figure C4, we can see that ASPIRE-EASE uses fewer cutting planes and thus is more efficient.

D Solve PD-DRO in Centralized Manner

Considering to solve the PD-DRO problem in Eq. (4) in centralized manner, we can rewrite the problem in Eq. (4) as:

$$\min_{w \in W} \max_{p \in P} \sum_{j=1}^{N} p_j f_j(w) \tag{D.105}$$

(a) SHL (b) Fashion MNIST

Figure C2: control the degree of robustness (worst case performance in the problem) on (a) SHL, (b) Fashion MNIST datasets.

(a) SHL (b) Fashion MNIST

Figure C3: Comparison of the convergence time on worst case worker on (a) SHL, (b) Fashion MNIST datasets.

where $w \in R^p$ is the model parameter. Utilizing the cutting plane method, we can obtain the approximate problem of Eq. (D.105),

$$\min_{w \in W, h \in H} \quad h \tag{D.106}$$

$$s.t. \sum_{j=1}^{N} (p + a_{l,j}) f_j(w) - h \le 0, \forall a_l \in A^t;$$

$$\text{var:} \quad w, h.$$

Thus, the Lagrangian function of Eq. (D.106) can be written as:

$$L_p(w, h, \{\lambda_l\}) = h + \sum_{l=1}^{|A^t|} \lambda_l \Big(\sum_{j=1}^{N} (p + a_{l,j}) f_j(w) - h \Big). \tag{D.107}$$

Following [52], the regularized version of (D.107) is employed to update all variables as follows,

$$\tilde{L}_p(w, h, \{\lambda_l\}) = h + \sum_{l=1}^{|A^t|} \lambda_l \Big(\sum_{j=1}^{N} (p + a_{l,j}) f_j(w) - h \Big) - \sum_{l=1}^{|A^t|} \frac{c_1^t}{2} ||\lambda_l||^2, \tag{D.108}$$

where c_1^t denotes the regularization term in $(t+1)^{th}$ iteration. To avoid enumerating the whole dataset, the mini-batch loss $\hat{f}_j(w) = \sum_{i=1}^{m} \frac{1}{m} L_j(x_j^i, y_j^i; w)$ can be used, where m is the mini-batch size. It is evident that $E[\hat{f}_j(w)] = f_j(w)$ and $E[\nabla \hat{f}_j(w)] = \nabla f_j(w)$. The centralized algorithm, which aims to solve problem in Eq. (4) in centralized manner, proceeds as follows in $(t+1)^{th}$ iteration:

(a) SHL (b) Fashion MNIST

Figure C4: Comparison of ASPIRE-CP and ASPIRE-EASE regarding the number of cutting planes on (a) SHL, (b) Fashion MNIST datasets. ASPIRE-CP represents ASPIRE with cutting plane method.

1. Updating the model parameter w as follows,

$$w^{t+1} = \mathcal{P}_{\mathcal{W}}(w^t - \eta_w^t \nabla_w \tilde{L}_p(w^t; h^t; \{\lambda_l^t\})); \tag{D.109}$$

   where η_w^t represents the step-size and $\mathcal{P}_{\mathcal{W}}$ represents the projection onto the convex set $\mathcal{W}$.

2. Updating the additional variable h as follows,

$$h^{t+1} = \mathcal{P}_{\mathcal{H}}(h^t - \eta_h^t \nabla_h \tilde{L}_p(w^{t+1}; h^t; \{\lambda_l^t\})); \tag{D.110}$$

   where η_h^t represents the step-size and $\mathcal{P}_{\mathcal{H}}$ represents the projection onto the convex set $\mathcal{H}$.

3. Updating the dual variable λ_l as follows,

$$\lambda_l^{t+1} = \mathcal{P}_{\Lambda}(\lambda_l^t + \eta_\lambda \nabla_{\lambda_l} \tilde{L}_p(w^{t+1}; h^{t+1}; \{\lambda_l^t\})); \; l = 1, \cdots, |A^t|; \tag{D.111}$$

   where η_λ represents the step-size and $\mathcal{P}_{\Lambda}$ represents the projection onto the convex set Λ.

Then, during T_1 iterations, EASE is utilized to update the set A^{t+1} every k iterations.

Definition D.1 Following [52, 32, 53], the stationarity gap at t^{th} iteration is defined as,

$$\nabla G^t = \begin{bmatrix} \frac{1}{\eta_w^t}(w^t - \mathcal{P}_{\mathcal{W}}(w^t - \eta_w^t \nabla_w L_p(w^t; h^t; \{\lambda_l^t\}))) \\ \frac{1}{\eta_h^t}(h^t - \mathcal{P}_{\mathcal{H}}(h^t - \eta_h^t \nabla_h L_p(w^t; h^t; \{\lambda_l^t\}))) \\ \{\frac{1}{\eta_\lambda}(\lambda_l^t - \mathcal{P}_{\Lambda}(\lambda_l^t + \eta_\lambda \nabla_{\lambda_l} L_p(w^t; h^t; \{\lambda_l^t\})))\} \end{bmatrix}. \tag{D.112}$$

And we also define:

$$\begin{aligned} (\nabla G^t)_w &= \frac{1}{\eta_w^t}(w^t - \mathcal{P}_{\mathcal{W}}(w^t - \eta_w^t \nabla_w L_p(w^t; h^t; \{\lambda_l^t\}))); \\ (\nabla G^t)_h &= \frac{1}{\eta_h^t}(h^t - \mathcal{P}_{\mathcal{H}}(h^t - \eta_h^t \nabla_h L_p(w^t; h^t; \{\lambda_l^t\}))); \\ (\nabla G^t)_{\lambda_l} &= \frac{1}{\eta_\lambda}(\lambda_l^t - \mathcal{P}_{\Lambda}(\lambda_l^t + \eta_\lambda \nabla_{\lambda_l} L_p(w^t; h^t; \{\lambda_l^t\}))). \end{aligned} \tag{D.113}$$

It follows that:

$$||\nabla G^t||^2 = ||(\nabla G^t)_w||^2 + ||(\nabla G^t)_h||^2 + \sum_{l=1}^{|A^t|} ||(\nabla G^t)_{\lambda_l}||^2. \tag{D.114}$$

Definition D.2 At t^{th} iteration, the stationarity gap w.r.t. $\tilde{L}_p(w; h; \{\lambda_l\})$ is defined as:

$$\nabla \tilde{G}^t = \begin{bmatrix} \frac{1}{\eta_w^t}(w^t - \mathcal{P}_{\mathcal{W}}(w^t - \eta_w^t \nabla_w \tilde{L}_p(w^t; h^t; \{\lambda_l^t\}))) \\ \frac{1}{\eta_h^t}(h^t - \mathcal{P}_{\mathcal{H}}(h^t - \eta_h^t \nabla_h \tilde{L}_p(w^t; h^t; \{\lambda_l^t\}))) \\ \{\frac{1}{\eta_\lambda}(\lambda_l^t - \mathcal{P}_{\Lambda}(\lambda_l^t + \eta_\lambda \nabla_{\lambda_l} \tilde{L}_p(w^t; h^t; \{\lambda_l^t\})))\} \end{bmatrix}. \tag{D.115}$$

And we also define:

$$(\nabla \tilde{G}^t)_w = \frac{1}{\eta_w^t}(w^t - P_W(w^t - \eta_w^t \nabla_w \tilde{L}_p(w^t; h^t; \{\lambda_l^t\})));$$
$$(\nabla \tilde{G}^t)_h = \frac{1}{\eta_h^t}(h^t - P_H(h^t - \eta_h^t \nabla_h \tilde{L}_p(w^t; h^t; \{\lambda_l^t\})));\tag{D.116}$$
$$(\nabla \tilde{G}^t)_{\lambda_l} = \frac{1}{\rho_1}(\lambda_l^t - P_\Lambda(\lambda_l^t + \rho_1 \nabla_{\lambda_l} \tilde{L}_p(w^t; h^t; \{\lambda_l^t\}))).$$

It follows that:

$$||\nabla \tilde{G}^t||^2 = ||(\nabla \tilde{G}^t)_w||^2 + ||(\nabla \tilde{G}^t)_h||^2 + \sum_{l=1}^{|A^t|} ||(\nabla \tilde{G}^t)_{\lambda_l}||^2.\tag{D.117}$$

Assumption D.1 L_p has Lipschitz continuous gradients. We assume that there exists $L > 0$ satisfying that,

$$||\nabla L_p(w; h; \{\lambda_l\}) - \nabla L_p(\hat{w}; \hat{h}; \{\hat{\lambda}_l\})|| \le L||[w - \hat{w}; h - \hat{h}; \lambda_{cat} - \hat{\lambda}_{cat}]||;$$

where $\nabla \in \{\nabla_w; \nabla_h; \nabla_{\lambda_l}\}$, $[;]$ represents the concatenation and $\lambda_{cat} - \hat{\lambda}_{cat} = [\lambda_1 - \hat{\lambda}_1; \cdots; \lambda_{|A^t|} - \hat{\lambda}_{|A^t|}] \in R^{|A^t|}$.

Setting D.1 $|A^t| \le M; \forall t$, i.e., an upper bound is set for the number of cutting planes.

Setting D.2 $c_1^t = \frac{1}{\rho_1 (t+1)^{\frac{1}{4}}} \ge \underline{c}_1$ is nonnegative non-increasing sequence, where $\underline{c}_1 > 0$ meets $\underline{c}_1^2 \le \frac{\varepsilon^2}{4M}$.

Theorem D.1 Suppose Assumption D.1 holds. We set $\eta_w^t = \eta_h^t = \frac{2}{L + \rho_1 |A^t| L^2 + 8\frac{|A^t| L^2}{\rho_1 (c_1^t)^2}}$, and we set constant $\rho_1 \le \frac{2}{L + 2c_1^0}$. For a given ε, we have:

$$T(\varepsilon) \le O(\max\{(\frac{16(\ldots - 2)L^2 M \rho_1 d\, d_5}{\varepsilon^2} + (T_1 + 2)^{\frac{1}{2}})^2; \frac{16 M^2 \rho_1^4}{\rho_1^4}\frac{1}{\varepsilon^4}\});\tag{D.118}$$

where ρ_1, $\ldots$, d, d_5 and T_1 are constants.

Lemma D.1 Suppose Assumption D.1 holds, we have:

$$L_p(w^{t+1}; h^t; \{\lambda_l^t\}) - L_p(w^t; h^t; \{\lambda_l^t\}) \le (\frac{L}{2} - \frac{1}{\eta_w^t})||w^{t+1} - w^t||^2;\tag{D.119}$$

$$L_p(w^{t+1}; h^{t+1}; \{\lambda_l^t\}) - L_p(w^{t+1}; h^t; \{\lambda_l^t\}) \le (\frac{L}{2} - \frac{1}{\eta_h^t})||h^{t+1} - h^t||^2.\tag{D.120}$$

Proof:

According to Assumption D.1, we have,

$$\begin{aligned} &L_p(w^{t+1}; h^t; \{\lambda_l^t\}) - L_p(w^t; h^t; \{\lambda_l^t\}) \\ &\le \langle \nabla_w L_p(w^t; h^t; \{\lambda_l^t\}), w^{t+1} - w^t \rangle + \frac{L}{2}||w^{t+1} - w^t||^2. \end{aligned}\tag{D.121}$$

According to the optimal condition for Eq. (D.109) and $\nabla_w \tilde{L}_p(w^t; h^t; \{\lambda_l^t\}) = \nabla_w L_p(w^t; h^t; \{\lambda_l^t\})$, we have,

$$\langle w^{t+1} - w^t, \nabla_w L_p(w^t; h^t; \{\lambda_l^t\}) \rangle \le -\frac{1}{\eta_w^t}||w^{t+1} - w^t||^2.\tag{D.122}$$

Combining Eq. (D.121) with Eq. (D.122), we have that,

$$L_p(w^{t+1}; h^t; \{\lambda_l^t\}) - L_p(w^t; h^t; \{\lambda_l^t\}) \le (\frac{L}{2} - \frac{1}{\eta_w^t})||w^{t+1} - w^t||^2.$$

Similar to Eq. (D.119), we can easily have Eq. (D.120).

Lemma D.2 Suppose Assumption D.1 holds, 8t, T_1, we have:

$$L_p(w^{t+1};h^{t+1};f \; {}_l^{t+1} g) \quad L_p(w^t;h^t;f \; {}_l^t g)$$
$$(\frac{L}{2} \quad \frac{1}{{}_w^t} + \frac{|A^t|L^2}{2a_1})||w^{t+1} \quad w^t||^2 + (\frac{L}{2} \quad \frac{1}{{}_h^t} + \frac{|A^t|L^2}{2a_1})||h^{t+1} \quad h^t||^2$$
$$+(\frac{a_1}{2} \quad \frac{c_1^{t\;1} \; c_1^t}{2} + \frac{1}{2\;_1})\sum_{l=1}^{|P^t|}|| \;{}_l^{t+1} \quad {}_l^t||^2 + \frac{c_1^{t\;1}}{2}\sum_{l=1}^{|P^t|}(|| \;{}_l^{t+1}||^2 \; || \;{}_l^t||^2) + \frac{1}{2\;_1}\sum_{l=1}^{|P^t|}|| \;{}_l^t \quad {}_l^{t\;1}||^2: \tag{D.123}$$

Proof:

According to Eq. (D.111), in $(t+1)^{th}$ iteration, 8 2 , it follows that,

$$\left\langle {}_l^{t+1} \quad {}_l^t \quad {}_1 r \; {}_l \, \mathcal{E}_p(w^{t+1};h^{t+1};f \; {}_l^t g); \quad {}_l^{t+1} \right\rangle \quad 0: \tag{D.124}$$

Let $= {}_l^t$, we can obtain,

$$r \; {}_l \, \mathcal{E}_p(w^{t+1};h^{t+1};f \; {}_l^t g) \quad \frac{1}{_1}(\; {}_l^{t+1} \quad {}_l^t); \; {}_l^t \quad {}_l^{t+1} \quad 0: \tag{D.125}$$

Likewise, in t^{th} iteration, we have that,

$$r \; {}_l \, \mathcal{E}_p(w^t;h^t;f \; {}_l^{t\;1} g) \quad \frac{1}{_1}(\; {}_l^t \quad {}_l^{t\;1}); \; {}_l^{t+1} \quad {}_l^t \quad 0: \tag{D.126}$$

8t T_1, since $\mathcal{E}_p(w;h;f \; {}_l g)$ is concave with respect to $_l$, we have,

$$\begin{aligned}
&\mathcal{E}_p(w^{t+1};h^{t+1};f \; {}_l^{t+1} g) \quad \mathcal{E}_p(w^{t+1};h^{t+1};f \; {}_l^t g) \\
&\sum_{l=1}^{|P^t|}\left\langle r \; {}_l \, \mathcal{E}_p(w^{t+1};h^{t+1};f \; {}_l^t g); \; {}_l^{t+1} \quad {}_l^t \right\rangle \\
&\sum_{l=1}^{|P^t|}\left\langle (r \; {}_l \, \mathcal{E}_p(w^{t+1};h^{t+1};f \; {}_l^t g) \; r \; {}_l \, \mathcal{E}_p(w^t;h^t;f \; {}_l^{t\;1} g); \; {}_l^{t+1} \quad {}_l^t \right\rangle \\
&\quad + \frac{1}{_1} \; {}_l^t \quad {}_l^{t\;1}; \; {}_l^{t+1} \quad {}_l^t \;):
\end{aligned} \tag{D.127}$$

Denoting $v_{1;l}^{t+1} = {}_l^{t+1} \quad {}_l^t \quad (\; {}_l^t \quad {}_l^{t\;1})$, we have,

$$\begin{aligned}
&\sum_{l=1}^{|P^t|}\left\langle (r \; {}_l \, \mathcal{E}_p(w^{t+1};h^{t+1};f \; {}_l^t g) \; r \; {}_l \, \mathcal{E}_p(w^t;h^t;f \; {}_l^{t\;1} g); \; {}_l^{t+1} \quad {}_l^t \right\rangle \\
= &\sum_{l=1}^{|P^t|}\left\langle r \; {}_l \, \mathcal{E}_p(w^{t+1};h^{t+1};f \; {}_l^t g) \; r \; {}_l \, \mathcal{E}_p(w^t;h^t;f \; {}_l^t g); \; {}_l^{t+1} \quad {}_l^t \right\rangle \; (1a) \\
+ &\sum_{l=1}^{|P^t|}\left\langle r \; {}_l \, \mathcal{E}_p(w^t;h^t;f \; {}_l^t g) \; r \; {}_l \, \mathcal{E}_p(w^t;h^t;f \; {}_l^{t\;1} g); v_{1;l}^{t+1} \right\rangle \; (1b) \\
+ &\sum_{l=1}^{|P^t|}\left\langle r \; {}_l \, \mathcal{E}_p(w^t;h^t;f \; {}_l^t g) \; r \; {}_l \, \mathcal{E}_p(w^t;h^t;f \; {}_l^{t\;1} g); \; {}_l^t \quad {}_l^{t\;1} \right\rangle \; (1c):
\end{aligned} \tag{D.128}$$

We firstly focus on (1a) in Eq. (D.128). According to the Cauchy-Schwarz inequality and Assumption D.1, we have,

$$\begin{aligned}
&\sum_{l=1}^{|P^t|}\left\langle r \; {}_l \, \mathcal{E}_p(w^{t+1};h^{t+1};f \; {}_l^t g) \; r \; {}_l \, \mathcal{E}_p(w^t;h^t;f \; {}_l^t g); \; {}_l^{t+1} \quad {}_l^t \right\rangle \\
&\sum_{l=1}^{|P^t|}(\frac{L^2}{2a_1}(||w^{t+1} \quad w^t||^2 + ||h^{t+1} \quad h^t||^2) + \frac{a_1}{2}|| \; {}_l^{t+1} \quad {}_l^t||^2 \\
&\quad + \frac{c_1^{t\;1} \; c_1^t}{2}(|| \; {}_l^{t+1}||^2 \; || \; {}_l^t||^2) \quad \frac{c_1^{t\;1} \; c_1^t}{2}|| \; {}_l^{t+1} \quad {}_l^t||^2);
\end{aligned} \tag{D.129}$$

where $a_1 > 0$ is a constant. Secondly, according to Cauchy-Schwarz inequality we write (1b) in Eq. (D.128) as,

$$\begin{aligned}&\sum_{l=1}^{|P^t|}\left\langle \nabla_{\theta_l}\tilde{L}_p(w^t;h^t;\{\theta_l^t\}) - \nabla_{\theta_l}\tilde{L}_p(w^t;h^t;\{\theta_l^{t-1}\}), v_{1;l}^{t+1}\right\rangle \\ \le &\sum_{l=1}^{|P^t|}\left(\frac{a_2}{2}||\nabla_{\theta_l}\tilde{L}_p(w^t;h^t;\{\theta_l^t\}) - \nabla_{\theta_l}\tilde{L}_p(w^t;h^t;\{\theta_l^{t-1}\})||^2 + \frac{1}{2a_2}||v_{1;l}^{t+1}||^2\right);\end{aligned} \tag{D.130}$$

where $a_2 > 0$ is a constant. Then, we focus on the (1c) in Eq. (D.128). Denoting $L_1{}' = L + c_1^0$, according to Assumption D.1, trigonometric inequality and the strong concavity of $\tilde{L}_p(w;h;\{\theta_l\})$ w.r.t θ_l [37, 52], we have,

$$\begin{aligned}&\sum_{l=1}^{|P^t|}\left\langle \nabla_{\theta_l}\tilde{L}_p(w^t;h^t;\{\theta_l^t\}) - \nabla_{\theta_l}\tilde{L}_p(w^t;h^t;\{\theta_l^{t-1}\}), \theta_l^t - \theta_l^{t-1}\right\rangle \\ \le &\sum_{l=1}^{|P^t|}\left(-\frac{1}{L_1{}'+c_1^{t-1}}||\nabla_{\theta_l}\tilde{L}_p(w^t;h^t;\{\theta_l^t\}) - \nabla_{\theta_l}\tilde{L}_p(w^t;h^t;\{\theta_l^{t-1}\}); v_{1;l}^{t+1}||^2 - \frac{c_1^{t-1}L_1{}'}{L_1{}'+c_1^{t-1}}||\theta_l^t - \theta_l^{t-1}||^2\right).\end{aligned} \tag{D.131}$$

In addition, we can obtain the following inequality,

$$\frac{1}{\eta_1}\langle \theta_l^t - \theta_l^{t-1}, \theta_l^{t+1} - \theta_l^t\rangle \le \frac{1}{2\eta_1}||\theta_l^{t+1} - \theta_l^t||^2 - \frac{1}{2\eta_1}||v_{1;l}^{t+1}||^2 + \frac{1}{2\eta_1}||\theta_l^t - \theta_l^{t-1}||^2. \tag{D.132}$$

Combining Eq. (D.127), (D.129), (D.130), (D.131), (D.132) with $\frac{\eta_1}{2} \ge \frac{1}{L_1{}'+c_1^t}$, and setting $a_2 = \eta_1$, $\forall t \ge T_1$, we have,

$$\begin{aligned}&L_p(w^{t+1};h^{t+1};\{\theta_l^{t+1}\}) - L_p(w^{t+1};h^{t+1};\{\theta_l^t\}) \\ \le &\frac{|A^t|L^2}{2a_1}(||w^{t+1}-w^t||^2 + ||h^{t+1}-h^t||^2) + \left(\frac{a_1}{2} - \frac{c_1^{t-1}-c_1^t}{2} + \frac{1}{2\eta_1}\right)\sum_{l=1}^{|P^t|}||\theta_l^{t+1}-\theta_l^t||^2 \\ &+ \frac{c_1^{t-1}}{2}\sum_{l=1}^{|P^t|}(||\theta_l^{t+1}||^2 - ||\theta_l^t||^2) + \frac{1}{2\eta_1}\sum_{l=1}^{|P^t|}||\theta_l^t - \theta_l^{t-1}||^2.\end{aligned} \tag{D.133}$$

By combining Lemma D.1 with Eq. (D.133), we conclude the proof of Lemma D.2.

Lemma D.3 Denote:

$$S_1^{t+1} = \frac{4}{\eta_1^2 c_1^{t+1}}\sum_{l=1}^{|A^t|}||\theta_l^{t+1} - \theta_l^t||^2 - \frac{4}{\eta_1}\left(\frac{c_1^{t-1}}{c_1^t} - 1\right)\sum_{l=1}^{|A^t|}||\theta_l^{t+1}||^2; \tag{D.134}$$

$$F^{t+1} = L_p(w^{t+1};h^{t+1};\{\theta_l^{t+1}\}) + S_1^{t+1} - \frac{7}{2\eta_1}\sum_{l=1}^{|P^t|}||\theta_l^{t+1} - \theta_l^t||^2 - \frac{c_1^t}{2}\sum_{l=1}^{|P^t|}||\theta_l^{t+1}||^2; \tag{D.135}$$

then $\forall t \ge T_1$, we have:

$$\begin{aligned}&F^{t+1} - F^t \\ \le &\left(\frac{L}{2} - \frac{1}{\eta_w^t} + \frac{\eta_1|A^t|L^2}{2} + \frac{8|A^t|L^2}{\eta_1(c_1^t)^2}\right)||w^{t+1} - w^t||^2 + \left(\frac{L}{2} - \frac{1}{\eta_h^t} + \frac{\eta_1|A^t|L^2}{2} + \frac{8|A^t|L^2}{\eta_1(c_1^t)^2}\right)||h^{t+1} - h^t||^2 \\ &- \frac{1}{10\eta_1}\sum_{l=1}^{|P^t|}||\theta_l^{t+1} - \theta_l^t||^2 + \frac{c_1^{t-1} - c_1^t}{2}\sum_{l=1}^{|P^t|}||\theta_l^{t+1}||^2 + \frac{4}{\eta_1}\left(\frac{c_1^{t-2}}{c_1^{t-1}} - \frac{c_1^{t-1}}{c_1^t}\right)\sum_{l=1}^{|P^t|}||\theta_l^t||^2.\end{aligned} \tag{D.136}$$

Proof:

Let $a_1 = \frac{1}{\eta_1}$ and substitute it into Lemma D.2, $\forall t \ge T_1$, we have,

$$\begin{aligned}&L_p(w^{t+1};h^{t+1};\{\theta_l^{t+1}\}) - L_p(w^t;h^t;\{\theta_l^t\}) \\ \le &\left(\frac{L}{2} - \frac{1}{\eta_w^t} + \frac{\eta_1|A^t|L^2}{2}\right)||w^{t+1} - w^t||^2 + \left(\frac{L}{2} - \frac{1}{\eta_h^t} + \frac{\eta_1|A^t|L^2}{2}\right)||h^{t+1} - h^t||^2 \\ &+ \left(\frac{c_1^{t-1} - c_1^t}{2} + \frac{1}{\eta_1}\right)\sum_{l=1}^{|P^t|}||\theta_l^{t+1} - \theta_l^t||^2 + \frac{c_1^{t-1}}{2}\sum_{l=1}^{|P^t|}(||\theta_l^{t+1}||^2 - ||\theta_l^t||^2) + \frac{1}{2\eta_1}\sum_{l=1}^{|P^t|}||\theta_l^t - \theta_l^{t-1}||^2.\end{aligned} \tag{D.137}$$

Firstly, $\forall t \geq T_1$, we can obtain the following inequality,

$$\begin{aligned}
&\sum_{l=1}^{|P^t|} \frac{1}{\gamma_1}\left\langle v_{1;l}^{t+1}, \theta_l^{t+1} - \theta_l^t \right\rangle \\
\leq &\sum_{l=1}^{|P^t|} \Big(\left\langle \nabla_{\theta_l} \tilde{L}_p(w^{t+1}; h^{t+1}; \{\theta_l^t\}) - \nabla_{\theta_l} \tilde{L}_p(w^t; h^t; \{\theta_l^t\}), \theta_l^{t+1} - \theta_l^t \right\rangle \\
&+ \left\langle \nabla_{\theta_l} \tilde{L}_p(w^t; h^t; \{\theta_l^t\}) - \nabla_{\theta_l} \tilde{L}_p(w^t; h^t; \{\theta_l^{t-1}\}), v_{1;l}^{t+1} \right\rangle \\
&+ \left\langle \nabla_{\theta_l} \tilde{L}_p(w^t; h^t; \{\theta_l^t\}) - \nabla_{\theta_l} \tilde{L}_p(w^t; h^t; \{\theta_l^{t-1}\}), \theta_l^t - \theta_l^{t-1} \right\rangle \Big).
\end{aligned} \tag{D.138}$$

Since

$$\frac{1}{\gamma_1}\left\langle v_{1;l}^{t+1}, \theta_l^{t+1} - \theta_l^t \right\rangle = \frac{1}{2\gamma_1}||\theta_l^{t+1} - \theta_l^t||^2 + \frac{1}{2\gamma_1}||v_{1;l}^{t+1}||^2 - \frac{1}{2\gamma_1}||\theta_l^t - \theta_l^{t-1}||^2; \tag{D.139}$$

it follows that,

$$\begin{aligned}
&\sum_{l=1}^{|P^t|} \left(\frac{1}{2\gamma_1}||\theta_l^{t+1} - \theta_l^t||^2 + \frac{1}{2\gamma_1}||v_{1;l}^{t+1}||^2 - \frac{1}{2\gamma_1}||\theta_l^t - \theta_l^{t-1}||^2\right) \\
\leq &\sum_{l=1}^{|P^t|} \Big(\frac{L^2}{2b_1^t}(||w^{t+1} - w^t||^2 + ||h^{t+1} - h^t||^2) + \frac{b_1^t}{2}||\theta_l^{t+1} - \theta_l^t||^2 \\
&+ \frac{c_1^{t-1} - c_1^t}{2}(||\theta_l^{t+1}||^2 - ||\theta_l^t||^2) - \frac{c_1^{t-1} - c_1^t}{2}||\theta_l^{t+1} - \theta_l^t||^2 \\
&+ \frac{1}{2}||\nabla_{\theta_l} \tilde{L}_p(w^t; h^t; \{\theta_l^t\}) - \nabla_{\theta_l} \tilde{L}_p(w^t; h^t; \{\theta_l^{t-1}\}); v_{1;l}^{t+1}||^2 + \frac{1}{2\gamma_1}||v_{1;l}^{t+1}||^2 \\
&- \frac{1}{L\lambda' + c_1^{t-1}}||\nabla_{\theta_l} \tilde{L}_p(w^t; h^t; \{\theta_l^t\}) - \nabla_{\theta_l} \tilde{L}_p(w^t; h^t; \{\theta_l^{t-1}\}); v_{1;l}^{t+1}||^2 - \frac{c_1^{t-1}L\lambda'}{L\lambda' + c_1^{t-1}}||\theta_l^t - \theta_l^{t-1}||^2 \Big);
\end{aligned} \tag{D.140}$$

where $b_1^t > 0$. According to the setting that $c_1^0 \geq L\lambda'$, we have $\frac{c_1^{t-1}L\lambda'}{L\lambda' + c_1^{t-1}} \geq \frac{c_1^{t-1}L\lambda'}{2L\lambda'} = \frac{c_1^{t-1}}{2} \geq \frac{c_1^t}{2}$. Multiplying both sides of the inequality Eq. (D.140) by $\frac{8}{\gamma_1 c_1^t}$ and setting $b_1^t = \frac{c_1^t}{2}$, $\forall t \geq T_1$, we have,

$$\begin{aligned}
S_1^{t+1} - S_1^t \leq &\sum_{l=1}^{|P^t|} \frac{4}{\gamma_1}\left(\frac{c_1^{t-2}}{c_1^{t-1}} - \frac{c_1^{t-1}}{c_1^t}\right)||\theta_l^t||^2 + \sum_{l=1}^{|P^t|}\left(\frac{2}{\gamma_1} + \frac{4}{\gamma_1^2}\left(\frac{1}{c_1^{t+1}} - \frac{1}{c_1^t}\right)\right)||\theta_l^{t+1} - \theta_l^t||^2 \\
&- \sum_{l=1}^{|P^t|} \frac{4}{\gamma_1}||\theta_l^t - \theta_l^{t-1}||^2 + \frac{8|A^t|L^2}{\gamma_1(c_1^t)^2}(||w^{t+1} - w^t||^2 + ||h^{t+1} - h^t||^2).
\end{aligned} \tag{D.141}$$

According to the setting about c_1^t, we have $\frac{\gamma_1}{10} \geq \frac{1}{c_1^{t+1}} - \frac{1}{c_1^t}$; $\forall t \geq T_1$. Using the definition of F^{t+1} and combining it with Eq. (D.141) and Eq. (D.138), $\forall t \geq T_1$, we have,

$$\begin{aligned}
&F^{t+1} - F^t \\
\leq &\left(\frac{L}{2} - \frac{1}{\eta_w^t} + \frac{\gamma_1|A^t|L^2}{2} + \frac{8|A^t|L^2}{\gamma_1(c_1^t)^2}\right)||w^{t+1} - w^t||^2 + \left(\frac{L}{2} - \frac{1}{\eta_h^t} + \frac{\gamma_1|A^t|L^2}{2} + \frac{8|A^t|L^2}{\gamma_1(c_1^t)^2}\right)||h^{t+1} - h^t||^2 \\
&- \frac{1}{10\gamma_1}\sum_{l=1}^{|P^t|}||\theta_l^{t+1} - \theta_l^t||^2 + \frac{c_1^{t-1} - c_1^t}{2}\sum_{l=1}^{|P^t|}||\theta_l^{t+1}||^2 + \frac{4}{\gamma_1}\left(\frac{c_1^{t-2}}{c_1^{t-1}} - \frac{c_1^{t-1}}{c_1^t}\right)\sum_{l=1}^{|P^t|}||\theta_l^t||^2.
\end{aligned}$$

Proof of Theorem D.1:

Firstly, we set that $a_5^t = \frac{4|A^t|(\gamma - 2)L^2}{\gamma_1(c_1^t)^2}$, where $\gamma > 2$ is a constant. According to the setting of η_w^t, η_h^t and c_1^t, we have,

$$\frac{L}{2} - \frac{1}{\eta_w^t} + \frac{\gamma_1|A^t|L^2}{2} + \frac{8|A^t|L^2}{\gamma_1(c_1^t)^2} \leq -a_5^t; \quad \frac{L}{2} - \frac{1}{\eta_h^t} + \frac{\gamma_1|A^t|L^2}{2} + \frac{8|A^t|L^2}{\gamma_1(c_1^t)^2} \leq -a_5^t. \tag{D.142}$$

Combining with Lemma D.3, $\forall t \geq T_1$, it follows that,

$$\begin{aligned}
&a_5^t||w^{t+1} - w^t||^2 + a_5^t||h^{t+1} - h^t||^2 + \frac{1}{10\gamma_1}\sum_{l=1}^{|P^t|}||\theta_l^{t+1} - \theta_l^t||^2 \\
\leq \; &F^t - F^{t+1} + \frac{c_1^{t-1} - c_1^t}{2}\sum_{l=1}^{|P^t|}||\theta_l^{t+1}||^2 + \frac{4}{\gamma_1}\left(\frac{c_1^{t-2}}{c_1^{t-1}} - \frac{c_1^{t-1}}{c_1^t}\right)\sum_{l=1}^{|P^t|}||\theta_l^t||^2.
\end{aligned} \tag{D.143}$$

According to the definition of $(\nabla \widetilde{G}(t))_{\boldsymbol{w}}$, we have,

$$||(\nabla \widetilde{G}^t)_{\boldsymbol{w}}||^2 \leq \frac{1}{(\eta_w^t)^2}||\boldsymbol{w}^{t+1}-\boldsymbol{w}^t||^2. \tag{D.144}$$

Combining the definition of $(\nabla \widetilde{G}(t))_h$ with trigonometric, Cauchy-Schwarz inequality and Assumption D.1, we have,

$$||(\nabla \widetilde{G}^t)_h||^2 \leq 2L^2||\boldsymbol{w}^{t+1}-\boldsymbol{w}^t||^2+\frac{2}{(\eta_h^t)^2}||h^{t+1}-h^t||^2. \tag{D.145}$$

Combining the definition of $(\nabla \widetilde{G}^t)_{\lambda_l}$ with trigonometric inequality and Cauchy-Schwarz inequality,

$$||(\nabla \widetilde{G}^t)_{\lambda_l}||^2 \leq \frac{3}{{\rho_1}^2}||\lambda_l^{t+1}-\lambda_l^t||^2+3L^2(||\boldsymbol{w}^{t+1}-\boldsymbol{w}^t||^2+||h^{t+1}-h^t||^2)+3((c_1^{t-1})^2-(c_1^t)^2)||\lambda_l^t||^2. \tag{D.146}$$

According to the Definition D.2 as well as Eq. (D.144), (D.145) and Eq. (D.146), we can obtain,

$$\begin{aligned}||\nabla \widetilde{G}^t||^2 \leq\ & (\tfrac{1}{(\eta_w^t)^2}+2L^2+3|\mathbf{A}^t|L^2)||\boldsymbol{w}^{t+1}-\boldsymbol{w}^t||^2+(\tfrac{2}{(\eta_h^t)^2}+3|\mathbf{A}^t|L^2)||h^{t+1}-h^t||^2 \\ & +\sum_{l=1}^{|\mathbf{A}^t|}\tfrac{3}{{\rho_1}^2}||\lambda_l^{t+1}-\lambda_l^t||^2+\sum_{l=1}^{|\mathbf{A}^t|}3((c_1^{t-1})^2-(c_1^t)^2)||\lambda_l^t||^2.\end{aligned} \tag{D.147}$$

We set constants d_1, d_2 as

$$d_1 = \frac{1+(2+3M)L^2\underline{\eta_{\boldsymbol{w}}}^2}{\underline{\eta_{\boldsymbol{w}}}^2(a_5^0)^2} \geq \frac{1+(2+3|\mathbf{A}^t|)L^2(\eta_{\boldsymbol{w}}^t)^2}{(\eta_{\boldsymbol{w}}^t)^2(a_5^t)^2}, \tag{D.148}$$

$$d_2 = \frac{2+3ML^2\underline{\eta_h}^2}{\underline{\eta_h}^2(a_5^0)^2} \geq \frac{2+3|\mathbf{A}^t|L^2(\eta_h^t)^2}{(\eta_h^t)^2(a_5^t)^2}, \tag{D.149}$$

where $\underline{\eta_{\boldsymbol{w}}} = \frac{2}{L+\rho_1 ML^2+8\frac{M\gamma L^2}{\rho_1 \underline{c}_1^2}} \leq \eta_{\boldsymbol{w}}^t$ and $\underline{\eta_h} = \frac{2}{L+\rho_1 ML^2+8\frac{M\gamma L^2}{\rho_1 \underline{c}_1^2}} \leq \eta_h^t, \forall t$ are positive constants. Thus, combining Eq. (D.147) with Eq. (D.148) and Eq. (D.149), we can obtain,

$$\begin{aligned}||\nabla \widetilde{G}^t||^2 \leq\ & d_1(a_5^t)^2||\boldsymbol{w}^{t+1}-\boldsymbol{w}^t||^2+d_2(a_5^t)^2||h^{t+1}-h^t||^2 \\ & +\sum_{l=1}^{|\mathbf{A}^t|}\tfrac{3}{{\rho_1}^2}||\lambda_l^{t+1}-\lambda_l^t||^2+\sum_{l=1}^{|\mathbf{A}^t|}3((c_1^{t-1})^2-(c_1^t)^2)||\lambda_l^t||^2.\end{aligned} \tag{D.150}$$

Let d_3^t denote a nonnegative sequence, $d_3^t = \frac{1}{\max\{d_1 a_5^t, d_2 a_5^t, \frac{30}{\rho_1}\}}$, and we have,

$$\begin{aligned}d_3^t||\nabla \widetilde{G}^t||^2 \leq\ & a_5^t||\boldsymbol{w}^{t+1}-\boldsymbol{w}^t||^2+a_5^t||h^{t+1}-h^t||^2 \\ & +\tfrac{1}{10\rho_1}\sum_{l=1}^{|\mathbf{A}^t|}||\lambda_l^{t+1}-\lambda_l^t||^2+3d_3^t((c_1^{t-1})^2-(c_1^t)^2)\sum_{l=1}^{|\mathbf{A}^t|}||\lambda_l^t||^2.\end{aligned} \tag{D.151}$$

Combining Eq. (D.151) with Eq. (D.143) and according to the setting $||\lambda_l^t||^2 \leq {\sigma_1}^2$ (where ${\sigma_1}^2 = {\alpha_3}^2$) and $d_3^0 \geq d_3^t, \forall t \geq T_1$, we have that,

$$d_3^t||\nabla \widetilde{G}^t||^2 \leq F^t-F^{t+1}+\frac{c_1^{t-1}-c_1^t}{2}M{\sigma_1}^2+\frac{4}{\rho_1}(\frac{c_1^{t-2}}{c_1^{t-1}}-\frac{c_1^{t-1}}{c_1^t})M{\sigma_1}^2+3d_3^0((c_1^{t-1})^2-(c_1^t)^2)M{\sigma_1}^2. \tag{D.152}$$

Denoting $\widetilde{T}(\varepsilon)$ as $\widetilde{T}(\varepsilon) = \min\{t \mid \ ||\nabla \widetilde{G}^{T_1+t}|| \leq \frac{\varepsilon}{2}, t \geq 2\}$. Summing up Eq. (D.152) from $t = T_1+2$ to $t = T_1+\widetilde{T}(\varepsilon)$, we have,

$$\begin{aligned}\sum_{t=T_1+2}^{T_1+\widetilde{T}(\varepsilon)} d_3^t||\nabla \widetilde{G}^t||^2 \leq\ & F^{T_1+2}-\underline{L}+\tfrac{4}{\rho_1}(\tfrac{c_1^{T_1}}{c_1^{T_1+1}}+\tfrac{c_1^{T_1+1}}{c_1^{T_1+2}})M{\sigma_1}^2+\tfrac{c_1^{T_1+1}}{2}M{\sigma_1}^2 \\ & +\tfrac{7}{2\rho_1}M{\sigma_3}^2+\tfrac{c_1^{T_1+2}}{2}M{\sigma_1}^2+3d_3^0(c_1^0)^2M{\sigma_1}^2 \\ =\ & \bar{d},\end{aligned} \tag{D.153}$$

where $\sigma_3 = \max\{||\lambda_1 - \lambda_2|| \,|\lambda_1, \lambda_2 \in \mathbf{\Lambda}\}$ and $\underline{L} = \min_{\boldsymbol{w}\in\boldsymbol{\mathcal{W}}, h\in\boldsymbol{\mathcal{H}}, \{\lambda_l\in\mathbf{\Lambda}\}} L_p(\boldsymbol{w}, h, \{\lambda_l\})$, which satisfy that,

$$F^{t+1} \geq \underline{L} - \frac{4}{\rho_1}\frac{c_1^{T_1+1}}{c_1^{T_1+2}} M{\sigma_1}^2 - \frac{7}{2\rho_1} M{\sigma_3}^2 - \frac{c_1^{T_1+2}}{2} M{\sigma_1}^2, \ \forall t \geq T_1 + 2, \tag{D.154}$$

and $\bar{d}$ is a constant. Constant d_5 is given by,

$$d_5 = \max\{d_1, d_2, \tfrac{30}{\rho_1 a_5^0}\} \geq \max\{d_1, d_2, \tfrac{30}{\rho_1 a_5^t}\} = \tfrac{1}{d_3^t a_5^t} \ . \tag{D.155}$$

Thus, we can obtain that,

$$\sum_{t=T_1+2}^{T_1+\widetilde{T}(\varepsilon)} \frac{1}{d_5 a_5^t} ||\nabla \widetilde{G}^{T_1+\widetilde{T}(\varepsilon)}||^2 \leq \sum_{t=T_1+2}^{T_1+\widetilde{T}(\varepsilon)} d_3^t ||\nabla \widetilde{G}^t||^2 \leq \bar{d}\,. \tag{D.156}$$

Summing up $\frac{1}{a_5^t}$ from $t = T_1 + 2$ to $t = T_1 + \widetilde{T}(\varepsilon)$, it follows that,

$$\sum_{t=T_1+2}^{T_1+\widetilde{T}(\varepsilon)} \tfrac{1}{a_5^t} \geq \sum_{t=T_1+2}^{T_1+\widetilde{T}(\varepsilon)} \tfrac{1}{4(\gamma-2)L^2 M\rho_1 (t+1)^{\frac{1}{2}}} \geq \tfrac{(T_1+\widetilde{T}(\varepsilon))^{\frac{1}{2}} - (T_1+2)^{\frac{1}{2}}}{4(\gamma-2)L^2 M\rho_1}. \tag{D.157}$$

Combining Eq. (D.156), (D.157) with the definition of $\widetilde{T}(\varepsilon)$, we have that,

$$T_1 + \widetilde{T}(\varepsilon) \geq (\frac{16(\gamma-2)L^2 M\rho_1 \bar{d}\, d_5}{\varepsilon^2} + (T_1+2)^{\frac{1}{2}})^2. \tag{D.158}$$

According to trigonometric inequality, we then get $||\nabla G^t|| - ||\nabla \widetilde{G}^t|| \leq ||\nabla G^t - \nabla \widetilde{G}^t|| \leq \sqrt{\sum_{l=1}^{|\mathbf{A}^t|} ||c_1^{t-1}\lambda_l^t||^2}$. If $t > \frac{16M^2{\sigma_1}^4}{{\rho_1}^4}\frac{1}{\varepsilon^4}$, we have $\sqrt{\sum_{l=1}^{|\mathbf{A}^t|} ||c_1^{t-1}\lambda_l^t||^2} \leq \frac{\varepsilon}{2}$. Combining it with Eq. (D.158), we can conclude that there exists a

$$T(\varepsilon) \sim \mathcal{O}(\max\{(\frac{16(\gamma-2)L^2 M\rho_1 \bar{d}\, d_5}{\varepsilon^2} + (T_1+2)^{\frac{1}{2}})^2, \frac{16M^2{\sigma_1}^4}{{\rho_1}^4}\frac{1}{\varepsilon^4}\}), \tag{D.159}$$

such that $||\nabla G^t|| \leq \varepsilon$, which concludes our proof.

E Convergence Rate Analysis

In this section, we compare the convergence results of the proposed method against the existing methods in the literature (with centralized and distributed setting). GDmax [25] is proposed recently, which can be utilized to solve the nonconvex-concave minimax problems (related to the setting of our problem) with iteration complexity $\mathcal{O}(\frac{1}{\varepsilon^6})$ to obtain the ε-stationary point (*i.e.*, $||\Phi(\cdot)||^2 \leq \varepsilon^2$, where $\Phi(\cdot) = \max_y f(\cdot, y)$). However, GDmax is nested-loop which has to solve the inner subproblem every iteration [52]. Gradient descent-ascent (GDA) method [31] is proposed, which performs alternating gradient descent-ascent every iteration. The iteration complexity of GDA to obtain the ε-stationary point (*i.e.*, $||\Phi(\cdot)||^2 \leq \varepsilon^2$) for nonconvex-concave minimax problems is upper bounded by $\mathcal{O}(\frac{1}{\varepsilon^6})$. COVER [38] is proposed to solve the distributionally robust optimization with nonconvex objectives, which can obtain the ε-stationary point (*i.e.*, $||\mathcal{G}_\eta(\cdot)||^2 \leq \varepsilon^2$, where $\mathcal{G}_\eta$ is a proximal gradient measure) with the complexity $\mathcal{O}(\frac{1}{\varepsilon^3})$. Nevertheless, all the algorithms mentioned above do not discuss about the distributed algorithms. Recently, GCIVR [22] is proposed to solve the distributionally robust optimization problem in centralized and distributed manners. GCIVR is effective, which can respectively obtain the ε-stationary point (*i.e.*, $||\mathcal{G}_\eta(\cdot)||^2 \leq \varepsilon^2$) with the complexity $\mathcal{O}(\min\{\frac{\sqrt{N}}{\varepsilon^2}, \frac{1}{\varepsilon^3}\})$ and $\mathcal{O}(\min\{\frac{\sqrt{N}}{p\varepsilon^2} + \frac{\sqrt{N}}{\varepsilon^2}, \frac{1}{p\varepsilon^3} + \frac{1}{\varepsilon^3}\})$ (p is the number of workers, in this problem $p = N$) in centralized and distributed manners when the objective is nonconvex.

Table E1: Convergence rate of algorithms related to our work (with centralized and distributed setting).

Method	Centralized	Synchronous (Distributed)	Asynchronous (Distributed)
GDmax [25]	$\mathcal{O}(\frac{1}{\varepsilon^6})^{1,3}$	NA[5]	NA[5]
GDA [31]	$\mathcal{O}(\frac{1}{\varepsilon^6})^{1}$	NA[5]	NA[5]
COVER [38]	$\mathcal{O}(\frac{1}{\varepsilon^3})^{2}$	NA[5]	NA[5]
GCIVR [22]	$\mathcal{O}(\min\{\frac{\sqrt{N}}{\varepsilon^2}, \frac{1}{\varepsilon^3}\})^{2}$	$\mathcal{O}(\min\{\frac{\sqrt{N}}{p\varepsilon^2}+\frac{\sqrt{N}}{\varepsilon^2}, \frac{1}{p\varepsilon^3}+\frac{1}{\varepsilon^3}\})^{2,4}$	NA[5]
ASPIRE-EASE	$\mathcal{O}(\frac{1}{\varepsilon^4})$	NA[5]	$\mathcal{O}(\frac{1}{\varepsilon^6})$

[1] This complexity is to find an ε-stationary point of $\Phi(\cdot) = \max_y f(\cdot, y)$, that is $||\Phi(\cdot)||^2 \leq \varepsilon^2$.
[2] This complexity is to find an ε-stationary point such that $||\mathcal{G}_\eta(\cdot)||^2 \leq \varepsilon^2$, where $\mathcal{G}_\eta$ is a proximal gradient measure.
[3] This complexity corresponds to the number of iterations to solve the inner subproblem. It does not consider the complexity of solving the inner subproblem.
[4] p is the number of workers, in this problem $p = N$.
[5] NA represents not applicable.

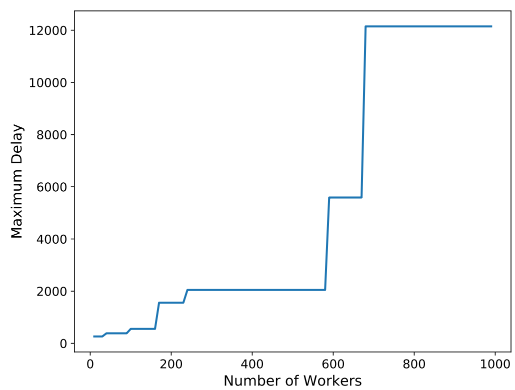

Figure E1: With the increase of the number of workers in the distributed system, the maximum delay would increase dramatically. The delay follows log-normal distribution LN(1, 0.4) in the experiment.

The proposed algorithm differs significantly from the aforementioned methods because it is designed for solving the PD-DRO problem in Eq. (4) in an *asynchronous distributed manner*. The asynchronous distributed algorithm does not suffer from the straggler problem [24] and therefore is critical for large scale distributed optimization in practice. On the contrary, synchronous distributed algorithm suffers from the straggler problem, *i.e.*, its speed is limited by the worker with maximum delay [10] and may not scale well with the size of a distributed system. For instance, we assume that the delays of workers follow a heavy-tailed distribution as given in [12]. With the increase of the number of workers in the distributed system, the maximum delay may increase dramatically as shown in Figure E1. Hence, the synchronous algorithm may incur huge delays and become practically infeasible for a large-scale distributed systems with tens of thousands of workers. Moreover, if a few workers fail to respond, which is very common in real-world large-scale data centers, the synchronous algorithm will come to an immediate halt [58]. Therefore, the asynchronous algorithm is strongly preferred in practice.

The asynchronous setting is considered when we design the distributed algorithm. Compared with centralized algorithm, the asynchronous distributed algorithm is more complicated, which pose the major challenge against the theoretical analysis. In the future work, how to improve the iteration complexity will be taken into consideration. And we summarize the convergence results of different methods in Table E1.

F Explanation about Assumption

The gradient Lipschitz (or smoothness) is a common assumption that has been widely used [16, 38, 31]. In some other works [47, 59], if the function L_p is $\widetilde{L}$-smooth, it has to satisfy,

$$\begin{aligned}&||\nabla_\theta L_p(\{\boldsymbol{w}_j\},\boldsymbol{z},h,\{\lambda_l\},\{\boldsymbol{\phi}_j\})-\nabla_\theta L_p(\{\hat{\boldsymbol{w}}_j\},\hat{\boldsymbol{z}},\hat{h},\{\hat{\lambda}_l\},\{\hat{\boldsymbol{\phi}}_j\})||\\&\leq\widetilde{L}(\sum_{j=1}^{N}||\boldsymbol{w}_j-\hat{\boldsymbol{w}}_j||+||\boldsymbol{z}-\hat{\boldsymbol{z}}||+||h-\hat{h}||+\sum_{l=1}^{M}||\lambda_l-\hat{\lambda}_l||+\sum_{j=1}^{N}||\boldsymbol{\phi}_j-\hat{\boldsymbol{\phi}}_j||),\end{aligned}\tag{F.160}$$

where $\theta\in\{\{\boldsymbol{w}_j\},\boldsymbol{z},h,\{\lambda_l\},\{\boldsymbol{\phi}_j\}\}$ and we demonstrate L_p that satisfies Eq. (F. 160) is also satisfied with our Assumption 1.

From Eq. (F. 160) and according to Cauchy-Schwarz inequality, we can obtain,

$$\begin{aligned}&||\nabla_\theta L_p(\{\boldsymbol{w}_j\},\boldsymbol{z},h,\{\lambda_l\},\{\boldsymbol{\phi}_j\})-\nabla_\theta L_p(\{\hat{\boldsymbol{w}}_j\},\hat{\boldsymbol{z}},\hat{h},\{\hat{\lambda}_l\},\{\hat{\boldsymbol{\phi}}_j\})||^2\\&\leq(2N+M+2)\widetilde{L}^2(\sum_{j=1}^{N}||\boldsymbol{w}_j-\hat{\boldsymbol{w}}_j||^2+||\boldsymbol{z}-\hat{\boldsymbol{z}}||^2+||h-\hat{h}||^2+\sum_{l=1}^{M}||\lambda_l-\hat{\lambda}_l||^2+\sum_{j=1}^{N}||\boldsymbol{\phi}_j-\hat{\boldsymbol{\phi}}_j||^2).\end{aligned}\tag{F.161}$$

Let $L=\sqrt{(2N+M+2)}\widetilde{L}$, we can obtain,

$$\begin{aligned}&||\nabla_\theta L_p(\{\boldsymbol{w}_j\},\boldsymbol{z},h,\{\lambda_l\},\{\boldsymbol{\phi}_j\})-\nabla_\theta L_p(\{\hat{\boldsymbol{w}}_j\},\hat{\boldsymbol{z}},\hat{h},\{\hat{\lambda}_l\},\{\hat{\boldsymbol{\phi}}_j\})||\\&\leq L||[\boldsymbol{w}_{\rm cat}-\hat{\boldsymbol{w}}_{\rm cat};\boldsymbol{z}-\hat{\boldsymbol{z}};h-\hat{h};\boldsymbol{\lambda}_{\rm cat}-\hat{\boldsymbol{\lambda}}_{\rm cat};\boldsymbol{\phi}_{\rm cat}-\hat{\boldsymbol{\phi}}_{\rm cat}]||.\end{aligned}\tag{F.162}$$

G Discussion about CD-norm Uncertainty Set

In this paper, we utilize the CD-norm uncertainty set in our framework. Compared with ellipsoid and KL-divergence uncertainty sets, whose cutting plane generation subproblems are respectively a second-order cone optimization (SOCP) problem and a relative entropy programming (REP) problem, the cutting plane generation subproblem (Eq. (17)) is an LP-type problem when utilizing CD-norm uncertainty set. Please note that the LP-type problem in Eq. (17) can be efficiently solved by merge sort. Therefore, the cutting plane generation subproblem with CD-norm uncertainty set is much easier to solve than those with the ellipsoid and KL-divergence uncertainty sets.