Achieving Near-Optimal Convergence for Distributed Minimax Optimization with Adaptive Stepsizes

Yan Huang
College of Control Science and Engineering
Zhejiang University, China
huangyan5616@zju.edu.cn

Xiang Li
Department of Computer Science
ETH Zurich, Switzerland
xiang.li@inf.ethz.ch

Yipeng Shen
College of Control Science and Engineering
Zhejiang University, China
22332074@zju.edu.cn

Niao He
Department of Computer Science
ETH Zurich, Switzerland
niao.he@inf.ethz.ch

Jinming Xu
College of Control Science and Engineering
Zhejiang University, China
jimmyxu@zju.edu.cn

Abstract

In this paper, we show that applying adaptive methods directly to distributed minimax problems can result in non-convergence due to inconsistency in locally computed adaptive stepsizes. To address this challenge, we propose D-AdaST, a Distributed Adaptive minimax method with Stepsize Tracking. The key strategy is to employ an adaptive stepsize tracking protocol involving the transmission of two extra (scalar) variables. This protocol ensures the consistency among stepsizes of nodes, eliminating the steady-state error due to the lack of coordination of stepsizes among nodes that commonly exists in vanilla distributed adaptive methods, and thus guarantees exact convergence. For nonconvex-strongly-concave distributed minimax problems, we characterize the specific transient times that ensure time-scale separation of stepsizes and quasi-independence of networks, leading to a near-optimal convergence rate of $\tilde{\mathcal{O}}\left(\epsilon^{-(4+\delta)}\right)$ for any small $\delta > 0$, matching that of the centralized counterpart. To our best knowledge, D-AdaST is the *first* distributed adaptive method achieving near-optimal convergence without knowing any problem-dependent parameters for nonconvex minimax problems. Extensive experiments are conducted to validate our theoretical results.

1 Introduction

Distributed optimization has seen significant research progress over the last decade, resulting in numerous algorithms (Nedic and Ozdaglar, 2009; Yuan et al., 2016; Lian et al., 2017; Pu and Nedić, 2021). However, the traditional focus of distributed optimization has primarily been on minimization tasks. With the rapid growth of machine learning research, various applications have emerged that go beyond simple minimization, such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Gulrajani et al., 2017), robust optimization (Mohri et al., 2019; Sinha et al., 2017), adversary training of neural networks (Wang et al., 2021), fair machine learning (Madras et al., 2018), and just

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0622

to name a few. These tasks typically involve a minimax structure as follows:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f\left(x, y\right),$$

where $\mathcal{X} \subseteq \mathbb{R}^p$, $\mathcal{Y} \subseteq \mathbb{R}^d$, and x, y are the primal and dual variables to be learned, respectively. One of the simplest yet effective methods for solving the above minimax problem is Gradient Descent Ascent (GDA) (Dem'yanov and Pevnyi, 1972; Nemirovski et al., 2009) which alternately performs stochastic gradient descent for the primal variable and stochastic gradient ascent for the dual variable. This approach has demonstrated its effectiveness in solving minimax problems, especially for convex-concave objectives (Hsieh et al., 2021; Daskalakis et al., 2021; Antonakopoulos et al., 2021), i.e., the function $f(\cdot, y)$ is convex for any $y \in \mathcal{Y}$, and $f(x, \cdot)$ is concave for any $x \in \mathcal{X}$.

Adaptive gradient methods, such as AdaGrad (Duchi et al., 2011), Adam (Kingma and Ba, 2014), and AMSGrad (Reddi et al., 2018), are often integrated with GDA to effectively solve minimax problems with theoretical guarantees in convex-concave settings (Diakonikolas, 2020; Antonakopoulos et al., 2021; Ene and Lê Nguyen, 2022). These adaptive methods are capable of adjusting stepsizes based on historical gradient information, making it robust to hyper-parameters tuning and can converge without requiring to know problem-dependent parameters (a characteristic often referred to as being "parameter-agnostic"). However, in the nonconvex regime, it has been shown by Lin et al. (2020); Yang et al. (2022b) that it is necessary to have a time-scale separation in stepsizes between the minimization and maximization processes to ensure the convergence of GDA and GDA-based adaptive algorithms. In particular, the stepsize ratio between primal and dual variables needs to be smaller than a threshold depending on the properties of the problem such as the smoothness and strong-concavity parameters (Li et al., 2022; Guo et al., 2021; Huang et al., 2021), which are often unknown or difficult to estimate in real-world tasks, such as training deep neural networks.

Applying GDA-based adaptive methods into decentralized settings poses additional challenges due to the presence of inconsistency in locally computed adaptive stepsizes. In particular, it has been shown that the inconsistency of stepsizes can result in non-convergence in federated learning with heterogeneous computation speeds (Wang et al., 2020; Sharma et al., 2023). This is mainly due to the lack of a central node coordinating the stepsizes of nodes in distributed settings, making it difficult to converge, as observed in minimization problems (Liggett, 2022; Chen et al., 2023b). As a result, the following question arises naturally:

"Can we design an adaptive minimax method that ensures the time-scale separation and consistency of stepsizes with provable convergence in fully distributed settings?"

Contributions. In this paper, we aim to propose a distributed adaptive method for efficiently solving nonconvex-strongly-concave (NC-SC) minimax problems. The contributions are threefold:

- We construct counterexamples showing that directly applying adaptive methods designed for centralized problems will lead to inconsistencies in locally computed adaptive stepsizes, resulting in non-convergence in distributed settings. To tackle this issue, we propose the *first* distributed adaptive minimax method, named D-AdaST, that incorporates an efficient stepsize tracking mechanism to maintain consistency across local stepsizes, which involves transmission of merely two extra (scalar) variables. The proposed algorithm exhibits time-scale separation in stepsizes and parameter-agnostic capability in fully distributed settings.

- Theoretically, we prove that D-AdaST is able to achieve a near-optimal convergence rate of $\tilde{\mathcal{O}}\left(\epsilon^{-(4+\delta)}\right)$ with arbitrarily small $\delta > 0$ to find an ϵ-stationary point for distributed NC-SC minimax problems. In contrast, we also prove the existence of a constant steady-state error in both the lower and upper bounds for GDA-based distributed minimax algorithms when being directly integrated with the adaptive stepsize rule without the stepsize tracking mechanism. Moreover, we explicitly characterize the transient times that ensure time-scale separation and quasi-independence of network, respectively.

- We conduct extensive experiments on real-world datasets to verify our theoretical findings and the effectiveness of D-AdaST on a variety of tasks, including robust training of neural networks and optimizing Wasserstein GANs. In all tasks, we demonstrate the superiority of D-AdaST over several vanilla distributed adaptive methods across various graphs, initial stepsizes and data distributions (see also additional experiments in Appendix A).

https://doi.org/10.52202/079017-0622

1.1 Related Works

Distributed nonconvex minimax methods. In the realm of federated learning, Deng and Mahdavi (2021) introduce Local SGDA algorithm combining FedAvg/Local SGD with stochastic GDA and show an $\tilde{\mathcal{O}}\left(\epsilon^{-6}\right)$ sample complexity for NC-SC objective functions. Sharma et al. (2022) provide improved complexity result of $\tilde{\mathcal{O}}\left(\epsilon^{-4}\right)$ matching that of the lower bound of first-order algorithms for both NC-SC and nonconvex-Polyak-Lojasiewicz (NC-PL) settings (Li et al., 2021; Zhang et al., 2021a) . Yang et al. (2022a) integrate Local SGDA with stochastic gradient estimators to eliminate the data heterogeneity. More recently, Zhang et al. (2023) adopt compressed momentum methods with Local SGD to increase the communication efficiency of the algorithm. For decentralized nonconvex minimax problems, Liu et al. (2020) study the training of GANs using decentralized optimistic stochastic gradient and provide non-asymptotic convergence with fixed stepsizes. Tsaknakis et al. (2020) propose a double-loop decentralized SGDA algorithm with gradient tracking techniques (Pu and Nedić, 2021) and achieve $\tilde{\mathcal{O}}\left(\epsilon^{-4}\right)$ sample complexity. With a stronger assumption of average smoothness, some studies employ variance reduction techniques to accelerate convergence (Zhang et al., 2021b; Chen et al., 2022; Xian et al., 2021; Tarzanagh et al., 2022; Wu et al., 2023; Chen et al., 2024; Zhang et al., 2024), which require more memory and computational resources due to the need for larger batch-sizes or full gradient evaluations. However, all the above-mentioned methods use a fixed or uniformly decaying stepsize, requiring the prior knowledge of smoothness and concavity.

(Distributed) adaptive minimax methods. For centralized nonconvex minimax problems, Yang et al. (2022b) show that, even in deterministic settings, GDA-based methods necessitate the time-scale separation of the stepsizes for primal and dual updates. Many attempts have been made for ensuring the time-scale separation requirement (Lin et al., 2020; Yang et al., 2022c; Boţ and Böhm, 2023; Huang et al., 2023). However, these methods typically come with the prerequisite of having knowledge about problem-dependent parameters, which can be a significant drawback in practical scenarios. To this end, Yang et al. (2022b) introduce a nested adaptive algorithm named NeAda that achieves parameter-agnosticism by incorporating an inner loop to effectively maximize the dual variable, which can obtain an optimal sample complexity of $\tilde{\mathcal{O}}\left(\epsilon^{-4}\right)$ when the strong-concavity parameter is known. More recently, Li et al. (2023) introduce TiAda, a single-loop parameter-agnostic adaptive algorithm for nonconvex minimax optimization which employs separated exponential factors on the adaptive primal and dual stepsizes, improving upon NeAda on the noise-adaptivity. There has been few works dedicated to adaptive minimax optimization in federated learning settings. For instance, Huang et al. (2024) introduces a federated adaptive algorithm that integrates the stepsize rule of Adam with full-client participation, resembling the centralized counterpart. Ju et al. (2023) study a federated Adam algorithm for fair federated learning where the objective function is properly weighted to account for heterogeneous updates among nodes. To the best of our knowledge, it is still unknown how one can design an adaptive minimax method capable of fulfilling the time-scale separation requirement and being parameter-agnostic in *fully distributed settings*.

Notations. Throughout this paper, we denote by $\mathbb{E}\left[\cdot\right]$ the expectation of a random variable, $\left\|\cdot\right\|$ the Frobenius norm, $\langle\cdot,\cdot\rangle$ the inner product of two vectors, $\odot$ the Hadamard product (entry wise), $\otimes$ the Kronecker product. We denote by $\mathbf{1}$ the all-ones vector, $\mathbf{I}$ the identity matrix and $\mathbf{J} = \mathbf{1}\mathbf{1}^T/n$ the averaging matrix with n dimension. For a vector or matrix A and constant α, we denote A^α the entry-wise exponential operations. We denote $\Phi\left(x\right) := f\left(x, y^*\left(x\right)\right)$ as the primal function where $y^*\left(x\right) = \underset{y\in\mathcal{Y}}{\operatorname{argmax}} f\left(x,y\right)$, and $\mathcal{P}_{\mathcal{Y}}\left(\cdot\right)$ as the projection operation onto set $\mathcal{Y}$.

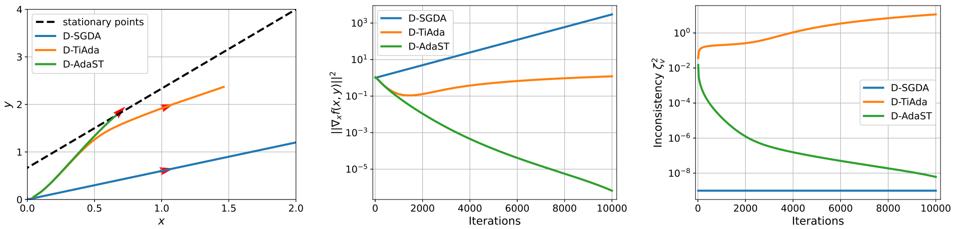
2 Distributed Adaptive Minimax Methods

We consider the distributed minimax problem collaboratively solved by a set of agents over a network. The overall objective of the agents is to solve the following finite-sum problem:

$$\min_{x\in\mathbb{R}^p}\ \max_{y\in\mathcal{Y}} f\left(x,y\right) = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[F_i\left(x,y;\xi_i\right)\right]}_{:=f_i(x,y)}, \tag{1}$$

where $f_i : \mathbb{R}^{p+d} \to \mathbb{R}$ is the local private loss function accessible only by the associated node $i \in \mathcal{N} = \{1, 2, \cdots, n\}$, $\mathcal{Y} \subset \mathbb{R}^d$ is closed and convex, and $\xi_i \sim \mathcal{D}_i$ denotes the data sample locally stored at node $i \in \mathcal{N}$ with distribution $\mathcal{D}_i$. We consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, here, $\mathcal{V} = \{1, 2, ..., n\}$ represents the set of agents, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of edges consisting of ordered pairs (i, j)

https://doi.org/10.52202/079017-0622



(a) trajectory (b) convergence of $\left\|\nabla_x f\left(x,y\right)\right\|^2$ (c) convergence of ζ_v^2

Figure 1: Comparison among D-SGDA, D-TiAda and D-AdaST for NC-SC quadratic objective function (6) with $n = 2$ nodes and $\gamma_x = \gamma_y$. In (a), it shows the trajectories of primal and dual variables of the algorithms, the points on the black dash line are stationary points of f. In (b), it shows the convergence of $\left\|\nabla_x f\left(x_k, y_k\right)\right\|^2$ over the iterations. In (c), it shows the convergence of the inconsistency of stepsizes, ζ_v^2 defined in (8), over the iterations. Notably, ζ_v^2 fails to converge for D-TiAda and $\zeta_v^2 = 0$ for non-adaptive D-SGDA.

representing the communication link from node j to node i. For node i, we define $\mathcal{N}_i = \{j \mid (i,j) \in \mathcal{E}\}$ as the set of its neighboring nodes. Before proceeding to the discussion of distributed algorithms, we first introduce the following notations for brevity:

$$\mathbf{x}_k := \left[x_{1,k}, x_{2,k}, \cdots, x_{n,k}\right]^T \in \mathbb{R}^{n\times p},\ \ \mathbf{y}_k := \left[y_{1,k}, y_{2,k}, \cdots, y_{n,k}\right]^T \in \mathbb{R}^{n\times d},$$

where $x_{i,k} \in \mathbb{R}^p, y_{i,k} \in \mathcal{Y}$ denote the primal and dual variable of node i at each iteration k, and

$$\nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right) := \left[\cdots, \nabla_x F_i\left(x_{i,k}, y_{i,k}; \xi_{i,k}^x\right), \cdots\right]^T,$$
$$\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right) := \left[\cdots, \nabla_y F_i\left(x_{i,k}, y_{i,k}; \xi_{i,k}^y\right), \cdots\right]^T$$

are the corresponding partial stochastic gradients with i.i.d. samples ξ_k^x, ξ_k^y in a compact form.

Next, we will first explain the pitfalls of directly applying centralized adaptive stepsize rules to decentralized settings, and then introduce our newly proposed solution to address the challenge.

2.1 Non-Convergence of Direct Extensions

For the distributed minimax optimization problem as depicted in (1) involving NC-SC objective functions, we will show shortly that the Distributed Stochastic Gradient Descent Ascent (D-SGDA) method may not converge due to the inability of time-scale separation with constant stepsizes (c.f., Figure 1), which is also observed in centralized settings (Lin et al., 2020; Yang et al., 2022b). To address this issue, one can adopt the adaptive stepsize rule used in centralized TiAda (Li et al., 2023) for each individual node, which is renowned for its ability to adaptively fulfill the time-scale separation requirements. As a result, we arrive at the following Distributed TiAda (D-TiAda) algorithm.

$$\mathbf{x}_{k+1} = W\left(\mathbf{x}_k - \gamma_x V_{k+1}^{-\alpha} \nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right), \tag{2a}$$
$$\mathbf{y}_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(W\left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right)\right), \tag{2b}$$

where γ_x and γ_y are the stepsizes, W is a doubly-stochastic weight matrix induced by graph $\mathcal{G}$ (Xiao et al., 2006) (c.f., Assumption 4), and

$$V_{k+1}^{-\alpha} = \operatorname{diag}\left\{v_{i,k+1}^{-\alpha}\right\}_{i=1}^n, \quad U_{k+1}^{-\beta} = \operatorname{diag}\left\{u_{i,k+1}^{-\beta}\right\}_{i=1}^n, \tag{3}$$

with $v_{i,k+1} = \max\left\{m_{i,k+1}^x, m_{i,k+1}^y\right\}, u_{i,k+1} = m_{i,k+1}^y$, and

$$m_{i,k+1}^x = m_{i,k}^x + \left\|\nabla_x F_i\left(x_{i,k}, y_{i,k}; \xi_{i,k}^x\right)\right\|^2,\ \ m_{i,k+1}^y = m_{i,k}^y + \left\|\nabla_y F_i\left(x_{i,k}, y_{i,k}; \xi_{i,k}^y\right)\right\|^2 \tag{4}$$

are the local accumulated gradient norm. Note that we impose a maximum operator in the preconditioner $v_{i,k}$, and employ different stepsize decaying rates, i.e., $0 < \beta < \alpha < 1$, for the primal and

https://doi.org/10.52202/079017-0622

dual variables, respectively. Such design allows to balance the updates of x and y, and achieves the desired time-scale separation without requiring any knowledge of parameters (Li et al., 2023).

However, in the distributed setting, such direct extension may fail to converge to a stationary point because $v_{i,k}$ and $u_{i,k}$ can be inconsistent due to the difference of local objective functions f_i, In particular, we can rewrite the above vanilla distributed optimization algorithm (2) in the sense of average system of primal variables as below,

$$\bar{x}_{k+1} = \underbrace{\bar{x}_k - \gamma_x \bar{v}_k^{-\alpha} \frac{\mathbf{1}^T}{n} \nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)}_{\text{adaptive descent}} - \underbrace{\gamma_x \frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n} \nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)}_{\text{inconsistancy}}, \tag{5}$$

where $\left(\tilde{\boldsymbol{v}}_k^{-\alpha}\right)^T := \left[\cdots, v_{i,k}^{-\alpha} - \bar{v}_k^{-\alpha}, \cdots\right]$, $\bar{x}_k := \mathbf{1}^T \mathbf{x}_k / n$ and $\bar{v}_k := 1/n \sum_{i=1}^n v_{i,k}$.

It is evident that, in comparison to centralized adaptive methods, an unexpected term (i.e., $\tilde{\boldsymbol{v}}_k$) on the right-hand side (RHS) arises due to inconsistencies. This term introduces inaccuracies in the directions of gradient descent, degrading the optimization performance. The theorem presented below reveals a gap near the stationary points in a properly designed counterexample, indicating the non-convergence of D-TiAda. The proof is available in Appendix B.3.

Theorem 1. *There exists a distributed minimax problem in the form of Problem (1) and certain initialization such that after running D-TiAda with any* $0 < \beta < 0.5 < \alpha < 1$ *and* $\gamma_x, \gamma_y > 0$*, it holds that for any* $t = 0, 1, 2, \ldots$*, we have,*

$$\| \nabla_x f(x_t, y_t) \| = \| \nabla_x f(x_0, y_0) \|, \quad \| \nabla_y f(x_t, y_t) \| = \| \nabla_y f(x_0, y_0) \|,$$

where $\|\nabla_x f(x_0, y_0)\|$ *and* $\|\nabla_y f(x_0, y_0)\|$ *can be arbitrarily large depending on the initialization.*

Remark 1. *The counterexample we constructed consists of three nodes, forming a complete graph. Without the stepsize tracking, D-TiAda will remain stationary, and the iterates will not progress if initiated along a specific line. In this counterexample, the only stationary point is at* $(0, 0)$*, but initial points along the line (c.f., Eq. (72)) can be positioned arbitrarily far away from this stationary point, implying the non-convergence of D-TiAda with certain initialization.*

Apart from the counterexample discussed in Theorem 1, we also experimentally observe the divergence of of D-SGDA and D-TiAda even in a simple scenario involving only two connected agents. This phenomenon is illustrated in Figure 1 and the functions are depicted as follows:

$$\begin{aligned} f_1\left(x, y\right) &= -\frac{9}{20} y^2 + \frac{3}{5} y - x + xy - \frac{1}{2} x^2, \\ f_2\left(x, y\right) &= -\frac{9}{20} y^2 + \frac{3}{5} y - x + 2xy - 2x^2. \end{aligned} \tag{6}$$

It is not difficult to verify that the points on the line $3y = 5x + 2$ are stationary points of $f\left(x, y\right) = 1/2\left(f_1\left(x, y\right) + f_2\left(x, y\right)\right)$. It follows from Figure 1(a) and 1(b) that D-SGDA does not converge to a stationary point because of the lack of time-scale separation, and D-TiAda also fails to converge due to stepsize inconsistency, as shown in Figure 1(c). In contrast, the utilization of the stepsize tracking protocol in D-AdaST ensures convergence to a stationary point, with the inconsistency in stepsizes gradually diminishing (c.f., Lemma 9). These two motivating examples effectively highlight the challenges associated with applying adaptive minimax algorithms to distributed settings.

2.2 The Proposed D-AdaST Algorithm

To address the issue of stepsize inconsistency across different nodes, we propose the following Distributed Adaptive minimax optimization algorithm with Stepsize Tracking protocol, termed D-AdaST, which allows us to asymptotically eliminate the stepsize inconsistency in a decentralized manner over networks. The pseudo-code for the algorithm is summarized in Algorithm 1, and can be rewritten in a compact form as follows:

$$\mathbf{m}_{k+1}^x = W\left(\mathbf{m}_k^x + \mathbf{h}_k^x\right), \tag{7a}$$

$$\mathbf{m}_{k+1}^y = W\left(\mathbf{m}_k^y + \mathbf{h}_k^y\right), \tag{7b}$$

$$\mathbf{x}_{k+1} = W\left(\mathbf{x}_k - \gamma_x V_{k+1}^{-\alpha} \nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right), \tag{7c}$$

$$\mathbf{y}_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(W\left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right)\right), \tag{7d}$$

Algorithm 1 **Distributed Adaptive Minimax Method with Stepsize Tracking (D-AdaST)**

```
Initialization: $x_{i,0} \in \mathbb{R}^p$, $y_{i,0} \in \mathcal{Y}$, buffers $m^x_{i,0} = m^y_{i,0} = c > 0$, stepsizes $\gamma_x, \gamma_y > 0$,
    exponential factors $0 < \beta < \alpha < 1$ and weight matrix $W$.
for iteration $k = 0, 1, \cdots$, each node $i \in [n]$, do
    Sample i.i.d. $g^x_{i,k} = \nabla_x F_i\left(x_{i,k}, y_{i,k}; \xi^x_{i,k}\right)$ and $g^y_{i,k} = \nabla_y F_i\left(x_{i,k}, y_{i,k}; \xi^y_{i,k}\right)$.
    Accumulate the gradient norm:
        $m^x_{i,k+1} = m^x_{i,k} + \|g^x_{i,k}\|^2, \; m^y_{i,k+1} = m^y_{i,k} + \|g^y_{i,k}\|^2.$
    Compute the ratio:
        $\psi_{i,k+1} = (m^x_{i,k+1})^\alpha / \max\left\{(m^x_{i,k+1})^\alpha, (m^y_{i,k+1})^\alpha\right\} \leqslant 1.$
    Update primal and dual variables locally:
        $x_{i,k+1} = x_{i,k} - \gamma_x \psi_{i,k+1} \left(m^x_{i,k+1}\right)^{-\alpha} g^x_{i,k}, \; y_{i,k+1} = y_{i,k} + \gamma_y (m^y_{i,k+1})^{-\beta} g^y_{i,k}.$
    Communicate adaptive stepsizes and decision variables with neighbors:
        $\left\{m^x_{i,k+1}, m^y_{i,k+1}, x_{i,k+1}, y_{i,k+1}\right\} \leftarrow \sum_{j \in \mathcal{N}_i} W_{i,j} \left\{m^x_{j,k+1}, m^y_{j,k+1}, x_{j,k+1}, y_{j,k+1}\right\}.$
    Projection of dual variable on the set $\mathcal{Y}$: $y_{i,k+1} \leftarrow \mathcal{P}_\mathcal{Y}\left(y_{i,k+1}\right)$.
end for
```

where $\mathbf{m}^x_k = [\cdots, m^x_{i,k}, \cdots]^T$, $\mathbf{m}^y_k = [\cdots, m^y_{i,k}, \cdots]^T$ denote the tracking variables for the accumulated global gradient norm, i.e., for $z \in \{x, y\}$,

$$\frac{\mathbf{1}^T}{n} \mathbf{m}^z_{k+1} = \frac{1}{n} \sum\nolimits_{i=1}^n \left(\sum\nolimits_{t=0}^k \left\| g^z_{i,t} \right\|^2 + m^z_{i,0} \right)$$

while $\boldsymbol{h}^z_k = [\cdots, \| g^z_{i,k} \|^2, \cdots]^T$, and V_k, U_k are diagonal matrices with $v_{i,k} = \max\left\{m^x_{i,k}, m^y_{i,k}\right\}$ and $u_{i,k} = m^x_{i,k}$. Note that we also provide a variant of D-AdaST with coordinate-wise adaptive stepsizes in Algorithm 2, along with its convergence analysis in Appendix B.5.

3 Convergence Analysis

In this section, we present the main convergence results for the proposed D-AdaST algorithm and compare it with D-TiAda to show the effectiveness of the proposed stepsize tracking protocol.

To this end, letting $\bar{u}_k := 1/n \sum_{i=1}^n u_{i,k}$, we define the following metrics to evaluate the level of inconsistency of stepsizes among nodes, which are ensured to be bounded by Assumption 3.

$$\zeta_v^2 := \sup_{i \in [n], k > 0} \left\{ \left(v_{i,k}^{-\alpha} - \bar{v}_k^{-\alpha} \right)^2 / \left(\bar{v}_k^{-\alpha} \right)^2 \right\}, \; \zeta_u^2 := \sup_{i \in [n], k > 0} \left\{ \left(u_{i,k}^{-\beta} - \bar{u}_k^{-\beta} \right)^2 / \left(\bar{u}_k^{-\beta} \right)^2 \right\}. \quad (8)$$

3.1 Assumptions

We consider the NC-SC setting of Problem (1) with the following assumptions that are commonly used in the existing works (c.f., Remark 2 and Remark 3). Notably, for the function and algorithm class determined by the assumptions of this work, Li et al. (2021) derived a lower complexity bound of $\Omega\left(\epsilon^{-4}\right)$ and proved that such a dependency on ϵ is optimal (c.f., Remark 2).

Assumption 1 (μ-strong concavity in y)**.** *Each objective function $f_i\left(x, y\right)$ is μ-strongly concave in y, i.e., $\forall x \in \mathbb{R}^p$, $\forall y, y' \in \mathcal{Y}$ and $\mu > 0$,*

$$f_i\left(x, y\right) - f_i\left(x, y'\right) \geqslant \left\langle \nabla_y f_i\left(x, y\right), y - y' \right\rangle + \frac{\mu}{2} \left\| y - y' \right\|^2. \quad (9)$$

Assumption 2 (Joint smoothness)**.** *Each objective function $f_i(x, y)$ is L-smooth in x and y, i.e., $\forall x, x' \in \mathbb{R}^p$ and $\forall y, y' \in \mathcal{Y}$, there exists a constant L such that for $z \in \{x, y\}$,*

$$\left\|\nabla_z f_i(x, y) - \nabla_z f_i(x', y')\right\|^2 \leqslant L^2 \left(\|x - x'\|^2 + \|y - y'\|^2\right). \tag{10}$$

Furthermore, f_i is second-order Lipschitz continuous for y, i.e., for $z \in \{x, y\}$,

$$\left\|\nabla_{zy}^2 f_i(x, y) - \nabla_{zy}^2 f_i(x', y')\right\|^2 \leqslant L^2 \left(\|x - x'\|^2 + \|y - y'\|^2\right). \tag{11}$$

Remark 2. *Assumption 1 does not require the convexity in x and the objective function thus can be nonconvex. Assumption 1 and 2 ensure that $y^*(\cdot)$ is smooth (c.f., Lemma 2), which is essential for achieving (near) optimal convergence rate (Chen et al., 2021; Li et al., 2023). Besides, it can be verified that the constructed 'hard' examples for obtaining the lower complexity bound in Li et al. (2021) satisfy the above second-order Lipschitz continuity (11) on y, implying that the achievable optimal complexity for the function and algorithm class considered in this work is $\mathcal{O}\left(\epsilon^{-4}\right)$.*

Assumption 3 (Stochastic gradient)**.** *For i.i.d. sample ξ_i, the stochastic gradient of each i is unbiased, i.e., $\forall x \in \mathbb{R}^p, y \in \mathcal{Y}$, $\mathbb{E}_{\xi_i}\left[\nabla_z F_i(x, y; \xi_i)\right] = \nabla_z f_i(x, y)$, for $z \in \{x, y\}$, and there is a constant $C > 0$ such that $\|\nabla_z F_i(x, y; \xi_i)\| \leqslant C$.*

Remark 3. *Assumption 3 on unbiased stochastic gradient is widely used for establishing convergence rates of both minimization and minimax optimization methods with AdaGrad (Kavis et al., 2022; Li et al., 2023) or Adam (Zou et al., 2019; Chen et al., 2023a; Huang et al., 2024) adaptive stepsize. We note that under Assumption 2, this assumption can be easily satisfied in many real-world tasks by imposing constraints on the compact domain of f, e.g., neural networks with rectified activation (Dinh et al., 2017) and GANs with projections on the critic (Gulrajani et al., 2017).*

Next, we make the following assumption on the underlying graph to ensure its connectivity.

Assumption 4 (Graph connectivity)**.** *The weight matrix W induced by graph $\mathcal{G}$ is doubly stochastic, i.e., $W\mathbf{1} = \mathbf{1}, \mathbf{1}^T W = \mathbf{1}^T$ and $\rho_W := \|W - \mathbf{J}\|_2^2 < 1$.*

Note that one can always find a proper weight matrix W compliant to the graph that satisfies Assumption 4 once the underlying graph is undirected and connected. For instance, the weight matrix can be easily determined based on the Metropolis-Hastings protocol (Xiao et al., 2006). Moreover, this assumption is more general than that in Lian et al. (2017); Borodich et al. (2021) in the sense that W is not required to be symmetric, implying that certain directed graphs can be included in this assumption, e.g., directed ring and exponential graphs (Ying et al., 2021).

3.2 Main Results

We are now ready to present the key convergence results in terms of the primal function $\Phi(x) := f(x, y^*(x))$ with $y^*(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} f(x, y)$, whose proofs can be found in Appendix B.4.

Theorem 2. *Suppose Assumption 1-4 hold. Let $0 < \beta < \alpha < 1$ and the total iteration K satisfy*

$$\Omega\left(\max\left\{\left(\frac{\gamma_x^2 \kappa^4}{\gamma_y^2}\right)^{\frac{1}{\alpha-\beta}}, \left(\frac{1}{(1-\rho_W)^2}\right)^{\max\left\{\frac{1}{\alpha}, \frac{1}{\beta}\right\}}\right\}\right) \tag{12}$$

with $\kappa := L/\mu$ to ensure time-scale separation and quasi-independence of the network. For D-AdaST, we have [1]

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\nabla \Phi(\bar{x}_k)\right\|^2\right] = \tilde{\mathcal{O}}\left(\frac{1}{K^{1-\alpha}} + \frac{1}{(1-\rho_W)^{\alpha} K^{\alpha}}\right) + \tilde{\mathcal{O}}\left(\frac{1}{K^{1-\beta}} + \frac{1}{(1-\rho_W) K^{\beta}}\right). \tag{13}$$

Remark 4 (Near-optimal convergence)**.** *Theorem 2 implies that if the total number of iterations satisfies the conditions (12), the proposed D-AdaST algorithm converges to a stationary point exactly for Problem (1) with an $\tilde{\mathcal{O}}\left(\epsilon^{-(4+\delta)}\right)$ sample complexity for arbitrarily small $\delta > 0$, e.g., letting*

[1] The complete convergence result can be found in (75) in Appendix.

$\alpha = 0.5 + \delta/(8+2\delta)$ and $\beta = 0.5 - \delta/(8+2\delta)$. It is worth noting that this rate is near-optimal compared to the existing lower bound of $\Omega\left(\epsilon^{-4}\right)$ (Li et al., 2021) for a class of smooth NC-SC functions. Moreover, this result recovers the centralized TiAda algorithm (Li et al., 2023) as a special case, i.e., setting $\rho_W = 0$, without assuming the existence of interior optimal point (c.f., Assumption 3.3 Li et al. (2023)). To the best of our knowledge, there is no existing fully parameter-agnostic method that achieves a convergence rate of $\tilde{\mathcal{O}}\left(\epsilon^{-4}\right)$, even in a centralized setting.

Remark 5 (Parameter-agnostic property and transient times)**.** *The above results show that D-AdaST converges without requiring to know any problem-dependent parameters, i.e., L, μ and ρ_W, or tuning the initial stepsize γ_x and γ_y, and is thus parameter-agnostic. Moreover, we explicitly characterize the transient times (c.f., Eq. (12)) that ensure time-scale separation and quasi-independence of the network, respectively. Indeed, we can see that if α and β are close to each other, the time required for time-scale separation to occur increases significantly, which has been observed in (Li et al., 2023). On the other hand, if α and β are relatively large, then $\tilde{\mathcal{O}}\left(1/K^{1-\alpha} + 1/K^{1-\beta}\right)$ dominates the other terms, indicating independence on the network. These observations highlight the trade-offs between the convergence rate and the required duration of the transition phase.*

For proper comparison, we also derive an upper bound for D-TiAda as follows. Together with the lower bound in Theorem 1, we demonstrate that without the stepsize tracking mechanism, the inconsistency among local stepsizes prevents D-TiAda from converging in the distributed setting.

Corollary 1. *Under the same conditions of Theorem 2. For the proposed D-TiAda, we have*

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi\left(\bar{x}_k\right)\right\|^2\right] &= \tilde{\mathcal{O}}\left(\frac{1}{K^{1-\alpha}} + \frac{1}{\left(1-\rho_W\right)^{\alpha}K^{\alpha}}\right)\\ &\quad + \tilde{\mathcal{O}}\left(\frac{1}{K^{1-\beta}} + \frac{1}{\left(1-\rho_W\right)K^{\beta}}\right) + \tilde{\mathcal{O}}\left(\left(\zeta_v^2 + \kappa^2\zeta_u^2\right)C^2\right).\end{aligned} \tag{14}$$

4 Experiments

In this section, we conduct experiments to validate the theoretical findings and demonstrate the effectiveness of the proposed algorithm on real-world machine learning tasks. We compare the proposed D-AdaST with the distributed variants of AdaGrad (Duchi et al., 2011), TiAda (Li et al., 2023) and NeAda (Yang et al., 2022b), namely D-AdaGrad, D-TiAda and D-NeAda, respectively. These experiments run across multiple nodes with different networks, and we consider heterogeneous distributions of local objective functions/datasets. For example, each node can only access samples with a subset of labels on MNIST and CIFAR-10 datasets, which is a common scenario in decentralized and federated learning tasks (Sharma et al., 2023; Huang et al., 2022). The experiments cover three main tasks: synthetic function, robust training of the neural network, and training of Wasserstein GANs (Heusel et al., 2017). For the exponential factors of stepsize, we set $\alpha = 0.6$ and $\beta = 0.4$ for both D-TiAda and D-AdaST. More detailed settings and additional experiments with different initial stepsizes, data distributions and choices of α and β can be found in Appendix A.

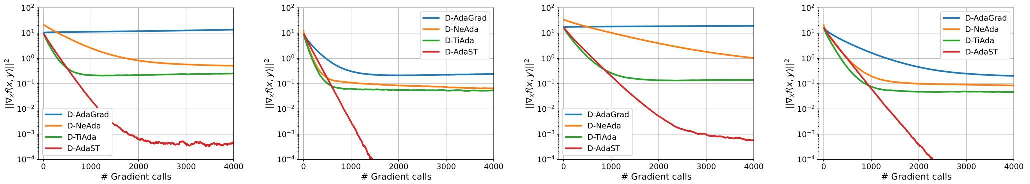
Synthetic example. We consider a distributed minimax problem with the following NC-SC local objective functions over exponential networks with $n = 50$ ($\rho_W = 0.71$) and $n = 100$ ($\rho_W = 0.75$).

$$f_i\left(x,y\right) = -\frac{1}{2}y^2 + L_i xy - \frac{L_i^2}{2}x^2 - 2L_i x + L_i y, \tag{15}$$

where $L_i \sim \mathcal{U}\left(1.5, 2.5\right)$. The local gradient of each node is computed with an additive $\mathcal{N}\left(0, 0.1\right)$ Gaussian noise. It follows from Figure 2 (a) and 2 (b) that the proposed D-AdaST algorithm outperforms other distributed adaptive methods for both initial stepsize settings, especially in cases with a favorable initial stepsize ratio, as illustrated in plots (b) and (d) where $\gamma_x/\gamma_y = 0.2$. Similar observation can be found in Figure 2 (c) and 2 (d), demonstrating the effectiveness of D-AdaST.

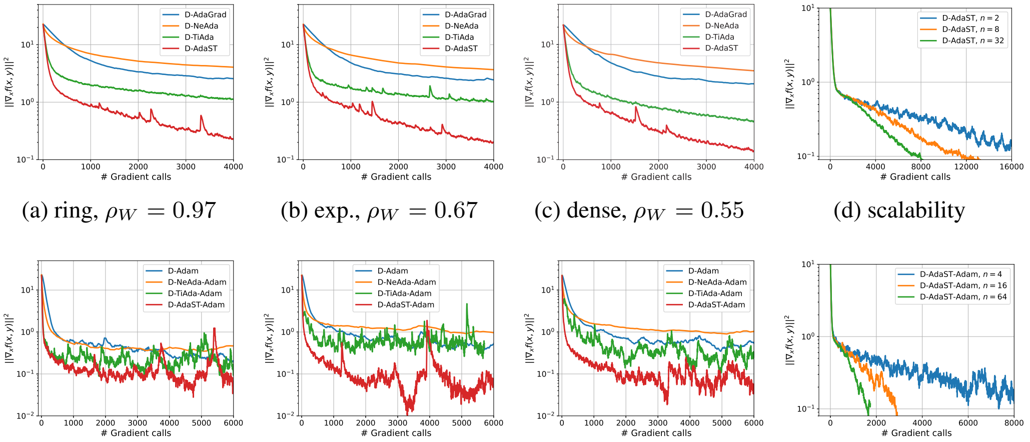
Robust training of neural networks. Next, we consider the task of robust training of neural networks, in the presence of adversarial perturbations on data samples (Sharma et al., 2022; Deng and Mahdavi, 2021). The problem can be formulated as $\min_x \max_y 1/n\sum_{i=1}^{n} f_i\left(x;\xi_i + y\right) - \eta\left\|y\right\|^2$, where x denotes the parameters of the model, y denotes the perturbation and ξ_i denotes the data sample of node i. Note that if η is large enough, the problem is NC-SC. We conduct experiments on

https://doi.org/10.52202/079017-0622



(a) $\gamma_x = 0.1, n = 50$ (b) $\gamma_x = 0.02, n = 50$ (c) $\gamma_x = 0.1, n = 100$ (d) $\gamma_x = 0.02, n = 100$

Figure 2: Performance comparison of algorithms on quadratic functions over exponential graphs with node counts $n = \{50, 100\}$ and *different initial stepsizes* ($\gamma_y = 0.1$).



(a) ring, $\rho_W = 0.97$ (b) exp., $\rho_W = 0.67$ (c) dense, $\rho_W = 0.55$ (d) scalability

(e) ring, $\rho_W = 0.97$ (f) exp., $\rho_W = 0.67$ (g) dense, $\rho_W = 0.55$ (h) scalability

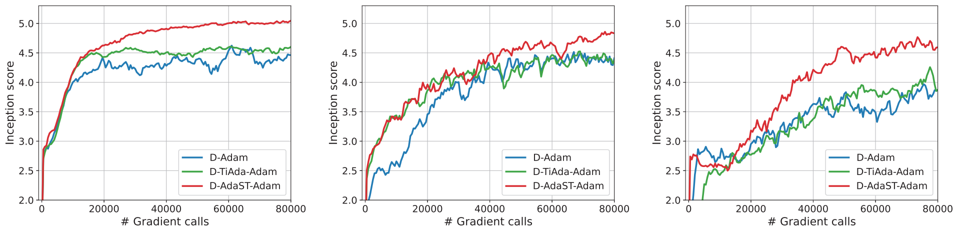
Figure 3: Comparison of the algorithms on training robust CNN on MNIST dataset. The first row shows the results of AdaGrad-like stepsize, and the second row is for Adam-like stepsize. For the first three columns, we compare the algorithms on *different graphs* with $n = 20$. For the last column, we show the scalability of D-AdaST in terms of number of nodes. Initial stepsizes are set as $\gamma_x = 0.01, \gamma_y = 0.1$ for AdaGrad-like stepsize, and $\gamma_x = 0.1, \gamma_y = 0.1$ for Adam-like stepsize.

MNIST dataset over different networks, e.g., ring graph, exponential (exp.) graph (Ying et al., 2021) and dense graph with $n/2$ edges for each node. We consider a heterogeneous scenario in which each node possesses only two distinct classes of labeled samples, resulting in heterogeneity among the local datasets across nodes, while the data is i.i.d within each node.

In Figure 3, we compare D-AdaST with D-AdaGrad, D-TiAda and D-NeAda, using adaptive stepsizes in AdaGrad (first row) and Adam (second row, name suffixed with Adam) respectively, it can be observed from the first three columns that the proposed D-AdaST outperforms the others on three different graphs and it is not very sensitive to the graph connectivity (i.e., ρ_W), demonstrating the quasi-independence of network as indicated in Theorem 2. It should be noted that Adam-like algorithms exhibit more fluctuations in the later stages of optimization as the gradient norm vanishes, leading to an inevitable increase in the Adam stepsize as the optimization process converges (Kingma and Ba, 2014). In plots (d) and (h), we further demonstrate that D-AdaST can scale efficiently with respect to the number of nodes, while keeping a constant batch-size of 64 for each node. This showcases the algorithm's ability to handle large-scale distributed scenarios effectively.

Generative Adversarial Networks. We further illustrate the effectiveness of D-AdaST on another popular task of training GANs, which has a generator and a discriminator used to generate and distinguish samples respectively (Goodfellow et al., 2014). In this experiment, we train Wasserstein GANs (Gulrajani et al., 2017) on CIFAR-10 dataset in a decentralized setting where each discriminator is 1-Lipschitz and has access to only two classes of samples. We compare the inception score of D-AdaST with D-Adam and D-TiAda adopting Adam-like stepsizes in Figure 4. It can be observed from the figure that D-AdaST achieves higher inception scores in three cases with different initial stepsizes, and has a small score loss as the initial step size changes. We believe that this example shows the great potential of D-AdaST in solving real-world problems.

https://doi.org/10.52202/079017-0622



(a) $\gamma_x = \gamma_y = 0.001$ (b) $\gamma_x = \gamma_y = 0.01$ (c) $\gamma_x = \gamma_y = 0.05$

Figure 4: Training GANs on CIFAR-10 dataset over exponential graphs with $n = 10$ nodes.

5 Conclusion

We introduced a new distributed adaptive minimax method, D-AdaST, designed to tackle the issue of non-convergence in nonconvex-strongly-concave minimax problems caused by the inconsistencies among locally computed adaptive stepsizes. Vanilla distributed adaptive methods could suffer from such inconsistencies, as highlighted by the carefully designed counterexamples for demonstrating their potential non-convergence. In contrast, our proposed method employs an efficient adaptive stepsize tracking protocol that not only ensures the time-scale separation, but also guarantees stepsize consistency among nodes and thus effectively eliminates steady-state errors. Theoretically, we showed that D-AdaST can achieve a near-optimal convergence rate of $\tilde{\mathcal{O}}\left(\epsilon^{-(4+\delta)}\right)$ with any arbitrarily small $\delta > 0$. Extensive experiments on both real-world and synthetic datasets have been conducted to validate our theoretical findings across various scenarios.

Acknowledgments

The work of Huang, Shen and Xu has been supported by the National Key R&D Program of China under Grant No. 2022YFB3102100, and in parts by National Natural Science Foundation of China under Grants 62373323, 62088101. The work of Li and He has been supported by the ETH research grant and Swiss National Science Foundation (SNSF) Starting Grant.

References

Antonakopoulos, K., Belmega, V. E., and Mertikopoulos, P. (2021). Adaptive extra-gradient methods for min-max optimization and games. In *ICLR 2021-9th International Conference on Learning Representations*, pages 1–28.

Borodich, E., Beznosikov, A., Sadiev, A., Sushko, V., Savelyev, N., Takáč, M., and Gasnikov, A. (2021). Decentralized personalized federated min-max problems. *arXiv preprint arXiv:2106.07289*.

Boţ, R. I. and Böhm, A. (2023). Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *SIAM Journal on Optimization*, 33(3):1884–1913.

Chen, C., Shen, L., Liu, W., and Luo, Z.-Q. (2023a). Efficient-adam: Communication-efficient distributed adam. *IEEE Transactions on Signal Processing*.

Chen, L., Ye, H., and Luo, L. (2022). A simple and efficient stochastic algorithm for decentralized nonconvex-strongly-concave minimax optimization. *arXiv preprint arXiv:2212.02387*.

Chen, L., Ye, H., and Luo, L. (2024). An efficient stochastic algorithm for decentralized nonconvex-strongly-concave minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1990–1998. PMLR.

Chen, T., Sun, Y., and Yin, W. (2021). Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34:25294–25307.

Chen, X., Karimi, B., Zhao, W., and Li, P. (2023b). On the convergence of decentralized adaptive gradient methods. In *Asian Conference on Machine Learning*, pages 217–232. PMLR.

Daskalakis, C., Skoulakis, S., and Zampetakis, M. (2021). The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1466–1478.

Dem'yanov, V. F. and Pevnyi, A. B. (1972). Numerical methods for finding saddle points. *USSR Computational Mathematics and Mathematical Physics*, 12(5):11–52.

Deng, Y. and Mahdavi, M. (2021). Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1395. PMLR.

Diakonikolas, J. (2020). Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Conference on Learning Theory*, pages 1428–1451. PMLR.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. (2017). Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Ene, A. and Lê Nguyen, H. (2022). Adaptive and universal algorithms for variational inequalities with optimal convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6559–6567.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. (2021). A novel convergence analysis for algorithms of the adam family. *arXiv preprint arXiv:2112.03459*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hsieh, Y.-P., Mertikopoulos, P., and Cevher, V. (2021). The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *International Conference on Machine Learning*, pages 4337–4348. PMLR.

Huang, F., Wang, X., Li, J., and Chen, S. (2024). Adaptive federated minimax optimization with lower complexities. In *International Conference on Artificial Intelligence and Statistics*, pages 4663–4671. PMLR.

Huang, F., Wu, X., and Hu, Z. (2023). Adagda: Faster adaptive gradient descent ascent methods for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2365–2389. PMLR.

Huang, F., Wu, X., and Huang, H. (2021). Efficient mirror descent ascent methods for nonsmooth minimax problems. *Advances in Neural Information Processing Systems*, 34:10431–10443.

Huang, Y., Sun, Y., Zhu, Z., Yan, C., and Xu, J. (2022). Tackling data heterogeneity: A new unified framework for decentralized SGD with sample-induced topology. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9310–9345. PMLR.

Ju, L., Zhang, T., Toor, S., and Hellander, A. (2023). Accelerating fair federated learning: Adaptive federated adam. *arXiv preprint arXiv:2301.09357*.

Kavis, A., Levy, K. Y., and Cevher, V. (2022). High probability bounds for a class of nonconvex algorithms with adagrad stepsize. In *International Conference on Learning Representations*.

https://doi.org/10.52202/079017-0622

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, H., Farnia, F., Das, S., and Jadbabaie, A. (2022). On convergence of gradient descent ascent: A tight local analysis. In *International Conference on Machine Learning*, pages 12717–12740. PMLR.

Li, H., Tian, Y., Zhang, J., and Jadbabaie, A. (2021). Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *Advances in Neural Information Processing Systems*, 34:1792–1804.

Li, X., YANG, J., and He, N. (2023). Tiada: A time-scale adaptive algorithm for nonconvex minimax optimization. In *The Eleventh International Conference on Learning Representations*.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30.

Liggett, B. (2022). Distributed learning with automated stepsizes.

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

Liu, M., Zhang, W., Mroueh, Y., Cui, X., Ross, J., Yang, T., and Das, P. (2020). A decentralized parallel algorithm for training generative adversarial nets. *Advances in Neural Information Processing Systems*, 33:11056–11070.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR.

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

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

Nedic, A., Ozdaglar, A., and Parrilo, P. A. (2010). Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55(4):922–938.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609.

Pu, S. and Nedić, A. (2021). Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1):409–457.

Reddi, S. J., Kale, S., and Kumar, S. (2018). On the convergence of adam and beyond. In *International Conference on Learning Representations*.

Sharma, P., Panda, R., and Joshi, G. (2023). Federated minimax optimization with client heterogeneity. *arXiv preprint arXiv:2302.04249*.

Sharma, P., Panda, R., Joshi, G., and Varshney, P. (2022). Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pages 19683–19730. PMLR.

Sinha, A., Namkoong, H., Volpi, R., and Duchi, J. (2017). Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*.

Tarzanagh, D. A., Li, M., Thrampoulidis, C., and Oymak, S. (2022). Fednest: Federated bilevel, minimax, and compositional optimization. In *International Conference on Machine Learning*, pages 21146–21179. PMLR.

Tsaknakis, I., Hong, M., and Liu, S. (2020). Decentralized min-max optimization: Formulations, algorithms and applications in network poisoning attack. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5755–5759. IEEE.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623.

Wang, J., Zhang, T., Liu, S., Chen, P.-Y., Xu, J., Fardad, M., and Li, B. (2021). Adversarial attack generation empowered by min-max optimization. *Advances in Neural Information Processing Systems*, 34:16020–16033.

Wu, X., Sun, J., Hu, Z., Zhang, A., and Huang, H. (2023). Solving a class of non-convex minimax optimization in federated learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xian, W., Huang, F., Zhang, Y., and Huang, H. (2021). A faster decentralized algorithm for nonconvex minimax problems. *Advances in Neural Information Processing Systems*, 34:25865–25877.

Xiao, L., Boyd, S., and Lall, S. (2006). Distributed average consensus with time-varying metropolis weights. *Automatica*, 1:1–4.

Yang, H., Liu, Z., Zhang, X., and Liu, J. (2022a). Sagda: Achieving $\mathcal{O}\left(\varepsilon^{-2}\right)$ communication complexity in federated min-max learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 7142–7154. Curran Associates, Inc.

Yang, J., Li, X., and He, N. (2022b). Nest your adaptive algorithm for parameter-agnostic nonconvex minimax optimization. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Yang, J., Orvieto, A., Lucchi, A., and He, N. (2022c). Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517. PMLR.

Ying, B., Yuan, K., Chen, Y., Hu, H., Pan, P., and Yin, W. (2021). Exponential graph is provably efficient for decentralized deep training. *Advances in Neural Information Processing Systems*, 34:13975–13987.

Yuan, K., Ling, Q., and Yin, W. (2016). On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854.

Zhang, S., Choudhury, S., Stich, S. U., and Loizou, N. (2023). Communication-efficient gradient descent-accent methods for distributed variational inequalities: Unified analysis and local updates. *arXiv preprint arXiv:2306.05100*.

Zhang, S., Yang, J., Guzmán, C., Kiyavash, N., and He, N. (2021a). The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR.

Zhang, X., Liu, Z., Liu, J., Zhu, Z., and Lu, S. (2021b). Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:18825–18838.

Zhang, X., Mancino-Ball, G., Aybat, N. S., and Xu, Y. (2024). Jointly improving the sample and communication complexities in decentralized stochastic minimax optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20865–20873.

Zhou, D., Chen, J., Cao, Y., Tang, Y., Yang, Z., and Gu, Q. (2018). On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*.

Zou, F., Shen, L., Jie, Z., Zhang, W., and Liu, W. (2019). A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11127–11135.

https://doi.org/10.52202/079017-0622

A Additional Experiments

In this section, we provide detailed experimental settings and perform additional experiments on the task of training robust neural networks with different choices of hyper-parameters. All experiments are deployed in a server with Intel Xeon E5-2680 v4 CPU @ 2.40GHz and 8 Nvidia RTX 3090 GPUs, and implemented using distributed communication package *torch.distributed* in PyTorch 2.0, where each process serves as a node, and we use inter-process communication to mimic communication between nodes. For the AdaGrad-like algorithms considered in the experiments of training neural networks, similar to the Adam-like stepsize, we adopt a coordinate-wise adaptive stepsize rule as commonly used in existing centralized adaptive methods (Yang et al., 2022b; Li et al., 2023). Moreover, since we attempt to develop a parameter-agnostic algorithm that does not need much effort in tuning hyper-parameters, we set $\alpha = 0.6$ and $\beta = 0.4$ for all tasks in the main text, and evaluate the effect of the choices of α and β on the performance of D-AdaST individually in an additional experiment on the synthetic objective function as shown in Appendix A.4.

A.1 Experimental details

Communication topology. For the experiments in the main text, we utilize three commonly used communication topologies: indirect ring, exponential graph and dense graph. An indirect ring is a sparse graph in which each node is sequentially connected to form a ring, with only two neighbors per node. Exponential graph (Ying et al., 2021) is a directed graph where each node is connected to nodes at distances of $2^0, 2^1 ..., 2^{\log n}$. Exponential graphs achieve a good balance between the degree and connectivity of the graph. A dense graph is an indirect graph where each node is connected to nodes at distances of $1, 2, 4, ..., n$. We also consider directed ring and fully connected graphs, which are more sparsely and densely connected, respectively, in the additional experiments.

Robust training of neural network. In this task, we train CNNs with three convolutional layers and one fully connected layer on MNIST dataset containing images of 10 classes. Each layer adopts batch normalization and ELU activation. The total batch-size is 1280, and the batch-size of each node during training is $1280/n$. For Adam-like algorithms, we set the first and second moment parameters as $\beta_1 = 0.9, \beta_2 = 0.999$ respectively. Since NeAda is a double-loop algorithm, for fair comparison, we imply D-AdaGrad and D-Adam using 15 iterations of inner loop in this task.

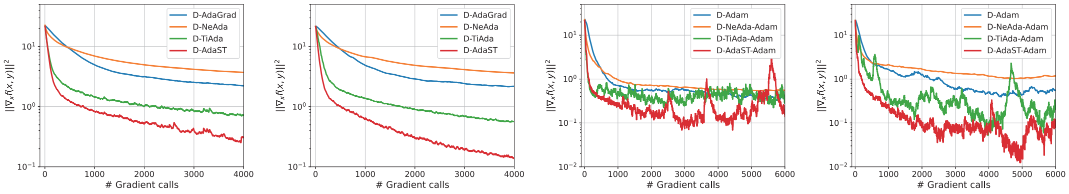
Generative Adversarial Networks. In this task, we train Wasserstein GANs on CIFAR-10 dataset, where the model used for discriminator is a four layer CNNs, and for generator is a four layer CNNs with transpose convolution layers. The total batch-size is 1280, and the batch-size of each node during training is 128 with 10 nodes. For Adam-like algorithms, we use $\beta_1 = 0.5, \beta_2 = 0.9$. To obtain the inception score, we use 8000 artificially generated samples to feed the previously trained inception network.

A.2 Additional experiments on robust training of neural network.

In this part, we conduct additional experiments on robust training of CNNs on MNIST dataset considering a variety of settings. We compare the convergence performance of D-AdaST with D-AdaGrad, D-TiAda and D-NeAda using adaptive stepsizes of AdaGrad and Adam. Unless otherwise specified, the total batch-size is set to 1280; the initial stepsizes for x and y are assigned as $\gamma_x = 0.01, \gamma_y = 0.1$ for AdaGrad-like algorithms, and $\gamma_x = \gamma_y = 0.1$ for Adam-like algorithms. Specifically, we consider two extra graphs that are more sparse and more dense, respectively in Figure 5, e.g., directed ring and fully-connected (fc) graphs. We consider more initial stepsizes settings for x and y respectively in Figure 6. Further, we also consider different data distributions where each node has samples from 4 of the 10 classes in Figure 7. Finally, we perform a comparison experiment with 40 nodes in Figure 8. Under all settings, the proposed D-AdaST outperforms the others, demonstrating the superiority of D-AdaST.

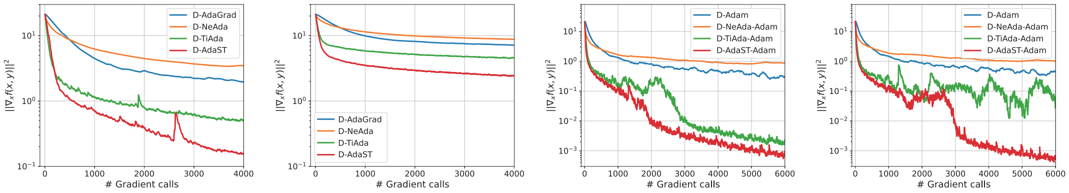
A.3 Additional experiments on training GANs

We provide additional experiments of training GANs on a more complicated dataset CIFAR-100 to further illustrate the effectiveness of the proposed D-AdaST, as shown in Figure 9. We use the entire training set of CIFAR-100 with coarse labels (20 classes) to train GANs over networks, where each node is assigned with four distinct classes of labeled samples. Under the same settings as in



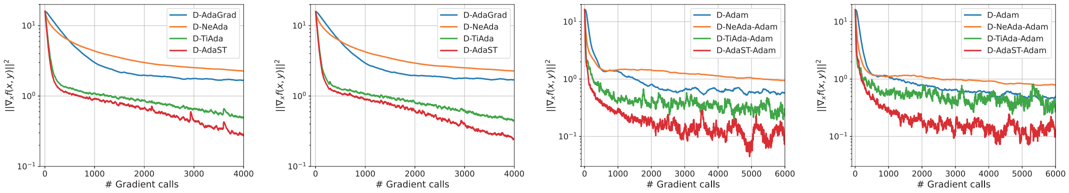
(a) directed-ring (b) fc (c) directed-ring (d) fc

Figure 5: Performance comparison of training CNN on MNIST with $n = 20$ nodes over *directed ring and fully connected graphs*.



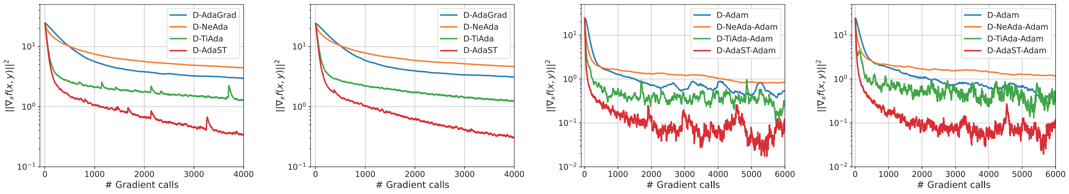
(a) 0.01, 0.01 (b) 0.001, 0.01 (c) 0.01, 0.01 (d) 0.01, 0.1

Figure 6: Performance comparison of training CNN on MNIST with $n = 20$ nodes with *different initial stepsizes* γ_x and γ_y.



(a) exp. (b) dense (c) exp. (d) dense

Figure 7: Performance comparison of training CNN on MNIST with $n = 20$ nodes over exponential and dense graphs where each node has *4 sample classes*.



(a) exp. (b) dense (c) exp. (d) dense

Figure 8: Performance comparison of training CNN on MNIST with $n = 40$ nodes over exponential and dense graphs.

Figure 4 (a), it can be observed that D-AdaST outperforms the others in terms of the inception score. Together with other experimental results in the main text, we believe that we have demonstrated the effectiveness of the proposed D-AdaST method and its potential for further real-world applications.

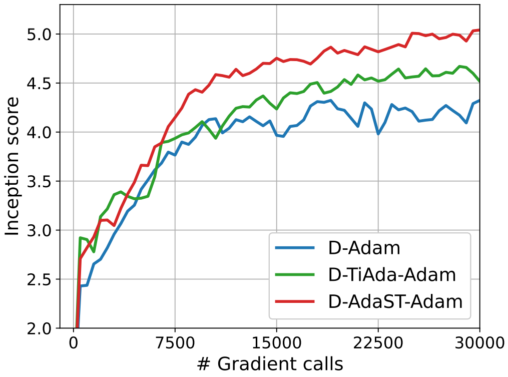


Figure 9: Performance comparison of D-AdaST with D-Adam and D-TiAda adopting Adam-like stepsizes for training GANs on CIFAR-100 with coarse labels over the exponential graph consisting of $n = 10$ nodes under initial stepsizes $\gamma_x = \gamma_y = 0.001$.

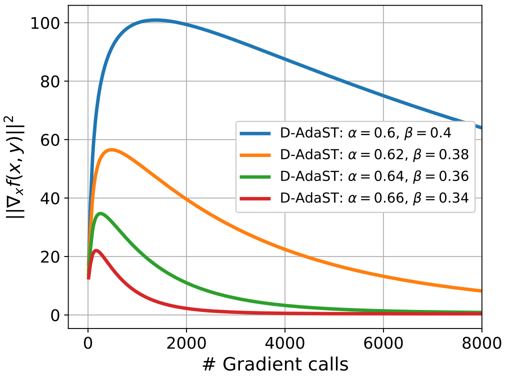


Figure 10: Performance comparison of D-AdaST on quadratic functions over an exponential graph of $n = 50$ nodes with different choices of α and β.

A.4 Additional experiments with different choices of α and β

In this part, we evaluate the effect of the choices of α and β on the performance of D-AdaST. In particular, we provide an additional experimental result on the synthetic quadratic objective functions (15) with a larger ratio of initial stepsizes, i.e., $\gamma_x/\gamma_y = 20$ (indicating faster minimization and slower maximization processes at the beginning). As shown in Figure 10, it can be observed that the transient time (iteration before the inflection point) becomes longer as $\alpha - \beta$ decreases, while the convergence rate is relatively faster, which is consistent with Theorem 2 and the result in the centralized TiAda algorithm (c.f., Figure 5, Li et al., 2023).

B Proof of the main results

We recall here some definitions used in the main text. The averaged variables and the inconsistency are defined as follows:

$$\bar{x}_k := \frac{\mathbf{1}^T}{n}\mathbf{x}_k, \quad \bar{v}_k := \frac{1}{n}\sum_{i=1}^{n} v_{i,k}, \quad \left(\tilde{\boldsymbol{v}}_k^{-\alpha}\right)^T := \left[\cdots, v_{i,k}^{-\alpha} - \bar{v}_k^{-\alpha}, \cdots\right],$$
$$\bar{y}_k := \frac{\mathbf{1}^T}{n}\mathbf{y}_k, \quad \bar{u}_k := \frac{1}{n}\sum_{i=1}^{n} u_{i,k}, \quad \left(\tilde{\boldsymbol{u}}_k^{-\beta}\right)^T := \left[\cdots, u_{i,k}^{-\beta} - \bar{u}_k^{-\beta}, \cdots\right].$$

https://doi.org/10.52202/079017-0622

The inconsistency of stepsizes of the primal and dual variables is defined as follows:

$$\zeta_v^2 := \sup_{i\in[n],k>0}\left\{\left(v_{i,k}^{-\alpha}-\bar{v}_k^{-\alpha}\right)^2/\left(\bar{v}_k^{-\alpha}\right)^2\right\},\ \zeta_u^2 := \sup_{i\in[n],k>0}\left\{\left(u_{i,k}^{-\beta}-\bar{u}_k^{-\beta}\right)^2/\left(\bar{u}_k^{-\beta}\right)^2\right\}.$$

Proof Sketch. The convergence analysis of the main results in Theorem 2 is mainly based on carefully analyzing the average system as shown in (5), and the difference between the distributed system and the averaged system. In general, under Assumption 1-4, we first give a telescoped descent lemma from 0 to $K-1$ iterations in Lemma 3, which is upper bounded by the following key error terms:

- $S_1 := \frac{1}{nK}\sum_{k=0}^{K-1}\mathbb{E}\left[\bar{v}_{k+1}^{-\alpha}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]$: The asymptotically decaying terms by adopting adaptive stepsize;
- $S_2 := \frac{1}{nK}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\right\|^2+\left\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\right\|^2\right]$: The consensus error of x and y between the distributed system and the average system;
- $S_3 := \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f\left(\bar{x}_k,y^*\left(\bar{x}_k\right)\right)-f\left(\bar{x}_k,\bar{y}_k\right)\right]$: The optimality gap in dual variable y;
- $S_4 := \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]$: The inconsistency of stepsize of x.

Next, we prove the contraction properties of these terms in Lemma 4-8 and Lemma 9 respectively. Finally, these results are integrated into the descent lemma to complete the proof. We note that the proof is not trivial in the sense that these terms are coupled and therefore are needed to be carefully analyzed. This proof can also be adapted to analyze the coordinate-wise adaptive stepsize variant of D-AdaST as explained in Appendix B.5, which is of independent interest.

B.1 Supporting lemmas

In this part, we provide several supporting lemmas that have been shown in the existing literature, which are essential to the subsequent convergence analysis.

Lemma 1 (Lemma A.2 in Yang et al. (2022b))**.** *Let $\{x_t\}_{t=0}^{T-1}$ be a sequence of non-negative real numbers, $x_0>0$ and $\alpha\in(0,1)$. Then we have,*

$$\left(\sum_{t=0}^{T-1}x_t\right)^{1-\alpha}\leqslant\sum_{t=0}^{T-1}\frac{x_t}{\left(\sum_{k=0}^{t}x_k\right)^{\alpha}}\leqslant\frac{1}{1-\alpha}\left(\sum_{t=0}^{T-1}x_t\right)^{1-\alpha}. \tag{16}$$

When $\alpha=0$, we have

$$\sum_{t=0}^{T-1}\frac{x_t}{\left(\sum_{k=0}^{t}x_k\right)^{\alpha}}\leqslant 1+\log\left(\frac{\sum_{t=0}^{T-1}x_t}{x_0}\right). \tag{17}$$

Lemma 2. *Suppose Assumption 1 and 2 hold. Define $\Phi(x) := f(x,y^*(x))$ as the envelope function and $y^*(x)=\underset{y\in\mathcal{Y}}{\operatorname{argmax}} f(x,y)$. Then, we have,*

- *$\Phi(\cdot)$ is L_Φ-smooth with $L_\Phi=L(1+\kappa)$, and $\nabla\Phi(x)=\nabla_x f(x,y^*(x))$ (c.f., Lemma 4.3 in Lin et al. (2020));*
- *$y^*(\cdot)$ is κ-Lipschitz and $\hat{L}$-smooth with $\hat{L}=\kappa(1+\kappa)^2$(c.f., Lemma 2 in Chen et al. (2021)).*

B.2 Key Lemmas

In this subsection, we give the key lemmas to help the analysis of the main results. For simplicity, we define $\Delta_k := \left\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\right\|^2+\left\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\right\|^2$ as the consensus error for primal and dual variables. Then, we have the following lemmas.

Lemma 3 (Descent lemma)**.** *Suppose Assumption 1-4 hold. Then, we have*

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi\left(\bar{x}_k\right)\right\|^2\right]\\
&\leqslant \frac{8C^{2\alpha}\left(\Phi^{\max}-\Phi^*\right)}{\gamma_x K^{1-\alpha}}-\frac{4}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla_x f\left(\bar{x}_k,\bar{y}_k\right)\right\|^2\right]\\
&+8\gamma_x L_\Phi\left(1+\zeta_v^2\right)\underbrace{\frac{1}{nK}\sum_{k=0}^{K-1}\mathbb{E}\left[\bar{v}_{k+1}^{-\alpha}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]}_{S_1}+8L^2\underbrace{\frac{1}{nK}\sum_{k=0}^{K-1}\mathbb{E}\left[\Delta_k\right]}_{S_2}\\
&+8\kappa L\underbrace{\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[f\left(\bar{x}_k,y^*\left(\bar{x}_k\right)\right)-f\left(\bar{x}_k,\bar{y}_k\right)\right]}_{S_3}+16\underbrace{\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]}_{S_4},
\end{aligned}
\tag{18}
$$

where $\kappa := L/\mu$ is the condition number of the function in y, $\Phi^{\max}=\max_x \Phi\left(x\right), \Phi^*=\min_x \Phi\left(x\right)$.

Proof. By the smoothness of Φ given in Lemma 2, i.e.,

$$
\Phi\left(\bar{x}_{k+1}\right)-\Phi\left(\bar{x}_k\right)\leqslant\left\langle\nabla\Phi\left(\bar{x}_k\right),\bar{x}_{k+1}-\bar{x}_k\right\rangle+\frac{L_\Phi}{2}\left\|\bar{x}_{k+1}-\bar{x}_k\right\|^2,
$$

and noticing that the scalar $\bar{v}_k, \bar{u}_k$ are random variables, we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\Phi\left(\bar{x}_{k+1}\right)-\Phi\left(\bar{x}_k\right)}{\gamma_x\bar{v}_{k+1}^{-\alpha}}\right]\\
&\leqslant-\mathbb{E}\left[\left\langle\nabla\Phi\left(\bar{x}_k\right),\frac{\mathbf{1}^T}{n}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k\right)\right\rangle\right]-\mathbb{E}\left[\left\langle\nabla\Phi\left(\bar{x}_k\right),\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\rangle\right]\\
&+\frac{\gamma_x L_\Phi}{2}\mathbb{E}\left[\frac{1}{\bar{v}_{k+1}^{-\alpha}}\left\|\left(\frac{\bar{v}_{k+1}^{-\alpha}\mathbf{1}^T}{n}+\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n}\right)\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right],
\end{aligned}
\tag{19}
$$

where we have used the definition of $\bar{x}_{k+1}$ as presented in (5). Then, we bound the inner-product terms on the RHS. Firstly,

$$
\begin{aligned}
&-\mathbb{E}\left[\left\langle\nabla\Phi\left(\bar{x}_k\right),\frac{\mathbf{1}^T}{n}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\rangle\right]\\
&=-\mathbb{E}\left[\left\langle\nabla\Phi\left(\bar{x}_k\right),\frac{\mathbf{1}^T}{n}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k\right)-\frac{\mathbf{1}^T}{n}\nabla_x F\left(\mathbf{1}\bar{x}_k,\mathbf{1}\bar{y}_k\right)+\frac{\mathbf{1}^T}{n}\nabla_x F\left(\mathbf{1}\bar{x}_k,\mathbf{1}\bar{y}_k\right)\right\rangle\right]\\
&\leqslant\frac{1}{4}\mathbb{E}\left[\left\|\nabla\Phi\left(\bar{x}_k\right)\right\|^2\right]+\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k\right)-\frac{\mathbf{1}^T}{n}\nabla_x F\left(\mathbf{1}\bar{x}_k,\mathbf{1}\bar{y}_k\right)\right\|^2\right]\\
&+\frac{1}{2}\left(\mathbb{E}\left[\left\|\nabla\Phi\left(\bar{x}_k\right)-\nabla_x f\left(\bar{x}_k,\bar{y}_k\right)\right\|^2\right]-\mathbb{E}\left[\left\|\nabla\Phi\left(\bar{x}_k\right)\right\|^2\right]-\mathbb{E}\left[\left\|\nabla_x f\left(\bar{x}_k,\bar{y}_k\right)\right\|^2\right]\right)\\
&\leqslant-\frac{1}{4}\mathbb{E}\left[\left\|\nabla\Phi\left(\bar{x}_k\right)\right\|^2\right]+\frac{L^2}{n}\mathbb{E}\left[\Delta_k\right]+\frac{L^2}{2}\mathbb{E}\left[\left\|\bar{y}_k-y^*\left(\bar{x}_k\right)\right\|^2\right]-\frac{1}{2}\mathbb{E}\left[\left\|\nabla_x f\left(\bar{x}_k,\bar{y}_k\right)\right\|^2\right].
\end{aligned}
\tag{20}
$$

wherein the last inequality we have used the smoothness of the objective functions. Then, for the second inner-product in (19), using Young's inequality we get

$$
\begin{aligned}
&-\mathbb{E}\left[\left\langle\nabla\Phi\left(\bar{x}_k\right),\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\rangle\right]\\
&\leqslant\frac{1}{8}\mathbb{E}\left[\left\|\nabla\Phi\left(\bar{x}_k\right)\right\|^2\right]+2\mathbb{E}\left[\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right].
\end{aligned}
\tag{21}
$$

Then, for the last term on the RHS of (18), recalling the definition of stepsize inconsistency in (8), we have

$$
\begin{aligned}
&\frac{\gamma_x L_{\Phi}}{2}\mathbb{E}\left[\frac{1}{\bar{v}_{k+1}^{-\alpha}}\left\|\left(\frac{\bar{v}_{k+1}^{-\alpha}\mathbf{1}^T}{n}+\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n}\right)\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right] \\
&\leqslant \frac{\gamma_x L_{\Phi}\left(1+\zeta_v^2\right)}{n}\mathbb{E}\left[\bar{v}_{k+1}^{-\alpha}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right].
\end{aligned}
\tag{22}
$$

Plugging the obtained inequalities into (18) and telescoping the terms, we get

$$
\begin{aligned}
&\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi\left(\bar{x}_k\right)\right\|^2\right] \\
&\leqslant 8\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\Phi\left(\bar{x}_k\right)-\Phi\left(\bar{x}_{k+1}\right)}{\gamma_x\bar{v}_k^{-\alpha}}\right]-4\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla_x f\left(\bar{x}_k,\bar{y}_k\right)\right\|^2\right] \\
&+4L^2\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\bar{y}_k-\bar{y}^*\right\|^2\right]+\frac{8L^2}{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\Delta_k\right] \\
&+\frac{8\gamma_x L_{\Phi}\left(1+\zeta_v^2\right)}{n}\sum_{k=0}^{K-1}\mathbb{E}\left[\bar{v}_k^{-\alpha}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right] \\
&+16\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right].
\end{aligned}
\tag{23}
$$

Now it remains to bound the first term on the RHS of the above inequality. With the help of Assumption 3, we have

$$
\begin{aligned}
&\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\Phi\left(\bar{x}_k\right)-\Phi\left(\bar{x}_{k+1}\right)}{\gamma_x\bar{v}_{k+1}^{-\alpha}}\right] \\
&=\sum_{k=0}^{K-1}\mathbb{E}\left[\frac{\Phi\left(\bar{x}_k\right)}{\gamma_x\bar{v}_k^{-\alpha}}-\frac{\Phi\left(\bar{x}_{k+1}\right)}{\gamma_x\bar{v}_{k+1}^{-\alpha}}+\Phi\left(\bar{x}_k\right)\left(\frac{1}{\gamma_x\bar{v}_{k+1}^{-\alpha}}-\frac{1}{\gamma_x\bar{v}_k^{-\alpha}}\right)\right] \\
&\leqslant\mathbb{E}\left[\frac{\Phi_{\max}}{\gamma_x\bar{v}_0^{-\alpha}}-\frac{\Phi^*}{\gamma_x\bar{v}_K^{-\alpha}}\right]+\sum_{k=0}^{K-1}\mathbb{E}\left[\Phi_{\max}\left(\frac{1}{\gamma_x\bar{v}_{k+1}^{-\alpha}}-\frac{1}{\gamma_x\bar{v}_k^{-\alpha}}\right)\right] \\
&\leqslant\frac{\left(\Phi_{\max}-\Phi^*\right)}{\gamma_x}\mathbb{E}\left[\bar{v}_K^{\alpha}\right] \\
&\leqslant\frac{\left(\Phi_{\max}-\Phi^*\right)\left(KC^2\right)^{\alpha}}{\gamma_x}.
\end{aligned}
\tag{24}
$$

Noticing that $\mathbb{E}\left[\left\|\bar{y}_k-y^*\left(\bar{x}_k\right)\right\|^2\right]\leqslant\frac{2}{\mu}\mathbb{E}\left[f\left(\bar{x}_k,y^*\left(\bar{x}_k\right)\right)-f\left(\bar{x}_k,\bar{y}_k\right)\right]$, we thus complete the proof.

□

Next, we need to bound the last four terms S_1-S_4 in (18) respectively. For S_1, we have the asymptotic convergence for both primal and dual variables in the following lemma.

Lemma 4. *Suppose Assumption 1-4 hold. Then, we have*

$$
\frac{1}{nK}\sum_{k=0}^{K-1}\mathbb{E}\left[\bar{v}_{k+1}^{-\alpha}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\leqslant\frac{C^{2-2\alpha}}{(1-\alpha)K^{\alpha}},
\tag{25}
$$

and

$$
\frac{1}{nK}\sum_{k=0}^{K-1}\mathbb{E}\left[\bar{u}_{k+1}^{-\beta}\left\|\nabla_y F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^y\right)\right\|^2\right]\leqslant\frac{C^{2-2\beta}}{(1-\beta)K^{\beta}}.
\tag{26}
$$

Proof. With the help of Lemma 1 and Assumption 3, taking the primal variable x as an example, and noticing that $v_{i,0} > 0, i \in [n]$, we have

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\bar{v}_{k+1}^{-\alpha}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\\
&= \frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\frac{\left\|\nabla_x F_i\left(x_{i,k},y_{i,k};\xi_{i,k}^x\right)\right\|^2}{\bar{v}_{k+1}^{\alpha}}\\
&\leqslant \frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\frac{\left\|\nabla_x F_i\left(x_{i,k},y_{i,k};\xi_{i,k}^x\right)\right\|^2}{\left(\sum_{t=0}^{k}\frac{1}{n}\sum_{j=1}^{n}\left\|\nabla_x F_j\left(x_{j,t},y_{j,t};\xi_{j,t}^x\right)\right\|^2\right)^{\alpha}}\\
&\leqslant \frac{1}{1-\alpha}\frac{1}{K}\left(\sum_{k=0}^{K-1}\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_x F_i\left(x_{i,k},y_{i,k};\xi_{i,k}^x\right)\right\|^2\right)^{1-\alpha} \leqslant \frac{C^{2-2\alpha}}{(1-\alpha)K^{\alpha}}.
\end{aligned}
$$

The similar result can be obtained for dual variable y and we thus complete the proof. □

Next, we bound the the consensus error term S_2 in the following lemma.

Lemma 5. *Suppose Assumption 1-4 hold. Then, we have*

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K}\mathbb{E}\left[\mathit{\Delta}_k\right] \leqslant \frac{2\mathbb{E}\left[\mathit{\Delta}_0\right]}{(1-\rho_W)K}\\
&+\frac{8n\rho_W\gamma_x^2\left(1+\zeta_v^2\right)}{(1-\rho_W)^2}\left(\frac{C^{2-4\alpha}}{(1-2\alpha)K^{2\alpha}}\mathbb{I}_{\alpha<1/2}+\frac{1+\log v_K-\log v_1}{K\bar{v}_1^{2\alpha-1}}\mathbb{I}_{\alpha\geqslant 1/2}\right)\\
&+\frac{8n\rho_W\gamma_y^2\left(1+\zeta_u^2\right)}{(1-\rho_W)^2}\left(\frac{C^{2-4\beta}}{(1-2\beta)K^{2\beta}}\mathbb{I}_{\beta<1/2}+\frac{1+\log u_K-\log u_1}{K\bar{u}_1^{2\beta-1}}\mathbb{I}_{\beta\geqslant 1/2}\right),
\end{aligned}
\tag{27}
$$

where $\mathbb{I}_{[\cdot]} \in \{0,1\}$ is the indicator for specific condition, and the initial consensus error $\mathit{\Delta}_0$ can be set to 0 with proper initialization.

Proof. By the updating rule of the primal variable, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{x}_{k+1}-\mathbf{1}\bar{x}_{k+1}\right\|^2\right]\\
&=\mathbb{E}\left[\left\|W\left(\mathbf{x}_k-\gamma_x V_{k+1}^{-\alpha}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right)-\mathbf{J}\left(\mathbf{x}_k-\gamma_x V_{k+1}^{-\alpha}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right)\right\|^2\right]\\
&\leqslant \frac{1+\rho_W}{2}\mathbb{E}\left[\left\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\right\|^2\right]+\frac{2\gamma_x^2\left(1+\rho_W\right)\rho_W}{1-\rho_W}\mathbb{E}\left[\bar{v}_{k+1}^{-2\alpha}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\\
&+\frac{2\gamma_x^2\left(1+\rho_W\right)\rho_W}{1-\rho_W}\mathbb{E}\left[\left\|\left(V_{k+1}^{-\alpha}-\bar{v}_{k+1}^{-\alpha}\mathbf{I}\right)\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right],
\end{aligned}
\tag{28}
$$

where we have used Young's inequality. Then, by the definition of ζ_v in (8), we have

$$
\mathbb{E}\left[\left\|\left(V_{k+1}^{-\alpha}-\bar{v}_{k+1}^{-\alpha}\mathbf{I}\right)\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\leqslant \zeta_v^2\mathbb{E}\left[\bar{v}_{k+1}^{-2\alpha}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right],
\tag{29}
$$

and thus

$$
\begin{aligned}
&\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\mathbf{x}_{k+1}-\mathbf{1}\bar{x}_{k+1}\right\|^2\right]\\
&\leqslant \frac{2}{1-\rho_W}\mathbb{E}\left[\left\|\mathbf{x}_k-\mathbf{1}\bar{x}_k\right\|^2\right]+\frac{8\gamma_x^2\rho_W\left(1+\zeta_v^2\right)}{(1-\rho_W)^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\bar{v}_{k+1}^{-2\alpha}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right].
\end{aligned}
\tag{30}
$$

Then, we bound the last term on the RHS of the above inequality by Lemma 4. For the case $\alpha < 1/2$, by Assumption 3 we have

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}\left[\bar{v}_{k+1}^{-2\alpha}\left\|\nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2\right] \\
&= \sum_{k=0}^{K-1} \sum_{i=1}^{n} \mathbb{E}\left[\frac{\left\|\nabla_x F_i\left(x_{i,k}, y_{i,k}; \xi_{i,k}^x\right)\right\|^2}{\bar{v}_{k+1}^{2\alpha}}\right] \leqslant \frac{n\left(KC^2\right)^{1-2\alpha}}{(1-2\alpha)}.
\end{aligned} \tag{31}$$

For the case $\alpha \geqslant 1/2$, with the help of Lemma 1, we have

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}\left[\bar{v}_{k+1}^{-2\alpha}\left\|\nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2\right] \\
&= \sum_{k=0}^{K-1} \sum_{i=1}^{n} \mathbb{E}\left[\frac{\left\|\nabla_x F_i\left(x_{i,k}, y_{i,k}; \xi_{i,k}^x\right)\right\|^2}{\bar{v}_{k+1} \cdot \bar{v}_{k+1}^{2\alpha-1}}\right] \leqslant \frac{n\left(1+\log v_T - \log v_1\right)}{\bar{v}_1^{2\alpha-1}}.
\end{aligned} \tag{32}$$

For the dual variable, we have

$$\begin{aligned}
\mathbf{y}_{k+1} &= \mathcal{P}_{\mathcal{Y}}\left(W\left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right)\right) \\
&= W\mathbf{y}_k + \gamma_y \nabla_y \hat{G}
\end{aligned}$$

where

$$\nabla_y \hat{G} = \frac{1}{\gamma_y}\left(\mathcal{P}_{\mathcal{Y}}\left(W\left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right)\right) - W\mathbf{y}_k\right).$$

Then, using Young's inequality with parameter λ, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{y}_{k+1} - \mathbf{1}\bar{y}_{k+1}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|W\mathbf{y}_k + \gamma_y \nabla_y \hat{G} - \mathbf{J}\left(W\mathbf{y}_k + \gamma_y \nabla_y \hat{G}\right)\right\|^2\right] \\
&\leqslant (1+\lambda)\,\rho_W \mathbb{E}\left[\left\|\mathbf{y}_k - \mathbf{J}\mathbf{y}_k\right\|^2\right] \\
&+ \left(1+\frac{1}{\lambda}\right) \mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{Y}}\left(W\left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right)\right) - W\mathbf{y}_k\right\|^2\right] \\
&\leqslant \frac{1+\rho_W}{2} \mathbb{E}\left[\left\|\mathbf{y}_k - \mathbf{J}\mathbf{y}_k\right\|^2\right] \\
&+ \frac{1+\rho_W}{1-\rho_W} \mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{Y}}\left(W\left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right)\right) - W\mathbf{y}_k\right\|^2\right].
\end{aligned}$$

Noticing that $W\mathbf{y}_k = \mathcal{P}_{\mathcal{Y}}\left(W\mathbf{y}_k\right)$ holds for convex set $\mathcal{Y}$, we get

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{y}_{k+1} - \mathbf{1}\bar{y}_{k+1}\right\|^2\right] \\
&\leqslant \frac{1+\rho_W}{2} \mathbb{E}\left[\left\|\mathbf{y}_k - \mathbf{J}\mathbf{y}_k\right\|^2\right] \\
&+ \frac{1+\rho_W}{1-\rho_W} \mathbb{E}\left[\left(\left\|\mathcal{P}_{\mathcal{Y}}\left(W\left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right)\right) - \mathcal{P}_{\mathcal{Y}}\left(W\mathbf{y}_k\right)\right\|\right)^2\right] \\
&\leqslant \frac{1+\rho_W}{2} \mathbb{E}\left[\left\|\mathbf{y}_k - \mathbf{J}\mathbf{y}_k\right\|^2\right] + \frac{1+\rho_W}{1-\rho_W} \mathbb{E}\left[\left\|\gamma_y U_{k+1}^{-\beta} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2\right] \\
&\leqslant \frac{1+\rho_W}{2} \mathbb{E}\left[\left\|\mathbf{y}_k - \mathbf{J}\mathbf{y}_k\right\|^2\right] + \frac{4\gamma_y^2\left(1+\zeta_u^2\right)}{\left(1-\rho_W\right)} \mathbb{E}\left[\bar{u}_{k+1}^{-2\beta}\left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2\right],
\end{aligned}$$

where we have used the non-expansiveness of projection operator. Then, we have

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\right\|^2\right] \\
&\leqslant \frac{2}{1-\rho_W}\mathbb{E}\left[\left\|\mathbf{y}_0 - \mathbf{J}\mathbf{y}_0\right\|^2\right] + \frac{8\gamma_y^2\left(1+\zeta_u^2\right)}{\left(1-\rho_W\right)^2}\sum_{k=0}^{K-1}\mathbb{E}\left[\bar{u}_{k+1}^{-2\beta}\left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2\right].
\end{aligned}$$

Similar to the primal variable, we can bound the last term above, which completes the proof. □

Next, we need to bound the term S_3 i.e., the optimality gap in dual variable. The intuition of the proof relies on the adaptive two time-scale protocol, that is, for given α and β, we try to find the threshold of the iterations k_0, after which the inner sub-problem can be well solved (faster) to ensure that the computation of outer sub-problem can be solved accurately (slower). In specific, we suppose that there is a constant G such that $\bar{u}_k \leqslant G$ hold for $k = 0, 1, \cdots, k_0 - 1$, then the analysis is divided into two phases.

Lemma 6 (First phase)**.** *Suppose Assumption 1-4 hold. If $\bar{u}_k \leqslant G, k = 0, 1, \cdots, k_0 - 1$, then we have*

$$\begin{aligned}
&\sum_{k=0}^{k_0-1}\mathbb{E}\left[f\left(\bar{x}_k, y^*\left(\bar{x}_k\right)\right) - f\left(\bar{x}_k, \bar{y}_k\right)\right] \\
&\leqslant \sum_{k=0}^{k_0-1}\mathbb{E}\left[E_{1,k}\right] + \frac{\gamma_x^2\kappa^2\left(1+\zeta_v^2\right)G^{2\beta}}{n\mu\gamma_y^2}\sum_{k=0}^{k_0-1}\mathbb{E}\left[\bar{v}_{k+1}^{-2\alpha}\left\|\nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2\right] \\
&+ \frac{\gamma_y\left(1+\zeta_u^2\right)}{n}\sum_{k=0}^{k_0-1}\mathbb{E}\left[\bar{u}_{k+1}^{-\beta}\left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k\right)\right\|^2\right] + \frac{4\kappa L}{n}\sum_{k=0}^{k_0-1}\mathbb{E}\left[\left\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\right\|^2\right] \\
&+ \frac{4}{\mu}\sum_{k=0}^{k_0-1}\mathbb{E}\left[\left\|\frac{\tilde{\boldsymbol{u}}_{k+1}^{-\beta}}{n\bar{u}_{k+1}^{-\beta}}\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2\right] + C\sum_{k=0}^{k_0-1}\mathbb{E}\left[\sqrt{\frac{1}{n}\left\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\right\|^2}\right],
\end{aligned} \tag{33}$$

where

$$E_{1,k} := \frac{1 - 3\mu\gamma_y\bar{u}_{k+1}^{-\beta}/4}{2\gamma_y\bar{u}_{k+1}^{-\beta}n}\left\|\mathbf{y}_k - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2 - \frac{\left\|\mathbf{y}_{k+1} - \mathbf{1}y^*\left(\bar{x}_{k+1}\right)\right\|^2}{\left(2 + \mu\gamma_y\bar{u}_{k+1}^{-\beta}\right)\gamma_y\bar{u}_{k+1}^{-\beta}n}. \tag{34}$$

Proof. Using Young's inequality with parameter λ_k, we get

$$\begin{aligned}
&\frac{1}{n}\left\|\mathbf{y}_{k+1} - \mathbf{1}\bar{y}^*\left(\bar{x}_{k+1}\right)\right\|^2 \\
&\leqslant \frac{\left(1+\lambda_k\right)}{n}\left\|\mathbf{y}_{k+1} - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2 + \left(1 + \frac{1}{\lambda_k}\right)\left\|y^*\left(\bar{x}_k\right) - y^*\left(\bar{x}_{k+1}\right)\right\|^2.
\end{aligned} \tag{35}$$

Recalling that $\mathbf{y}_{k+1} = \mathcal{P}_{\mathcal{Y}}\left(W\left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta}\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right)\right)$, we further define

$$\hat{\mathbf{y}}_{k+1} = W\left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta}\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right).$$

Then, for the first term on the RHS of (35), by the non-expansiveness property of projection operator $\mathcal{P}_{\mathcal{Y}}(\cdot)$ (c.f., Lemma 1 in (Nedic et al., 2010)), we have

$$\begin{aligned}
&\frac{1}{n}\left\|\mathbf{y}_{k+1} - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2 \\
&\leqslant \frac{1}{n}\left\|\hat{\mathbf{y}}_{k+1} - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2 - \frac{1}{n}\left\|\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1}\right\|^2 \\
&\leqslant \frac{1}{n}\left\|\mathbf{y}_k - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2 + \frac{\gamma_y^2}{n}\left\|U_{k+1}^{-\beta}\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2 \\
&- \frac{1}{n}\sum_{i=1}^{n}2\left\langle\gamma_y\bar{u}_{k+1}^{-\beta}g_{i,k}^y, y_{i,k} - y^*\left(\bar{x}_k\right)\right\rangle - \frac{1}{n}\sum_{i=1}^{n}2\left\langle\gamma_y\left(u_{i,k+1}^{-\beta} - \bar{u}_{k+1}^{-\beta}\right)g_{i,k}^y, y_{i,k} - y^*\left(\bar{x}_k\right)\right\rangle,
\end{aligned} \tag{36}$$

wherein the last inequality we have used the fact $\|W\|_2^2 \leqslant 1$. Then, multiplying by $1/\left(\gamma_y \bar{u}_{k+1}^{-\beta}\right)$ on both sides of (35) we get

$$\begin{aligned}
&\frac{1}{n\gamma_y \bar{u}_{k+1}^{-\beta}} \left\|\mathbf{y}_{k+1} - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2 \\
&\leqslant \frac{1+\lambda_k}{\lambda_k \gamma_y \bar{u}_{k+1}^{-\beta}} \left\|\bar{y}^*\left(\bar{x}_k\right) - \bar{y}^*\left(\bar{x}_{k+1}\right)\right\|^2 \\
&+ \left(1+\lambda_k\right)\left(\frac{1}{n\gamma_y \bar{u}_{k+1}^{-\beta}} \left\|\mathbf{y}_k - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2 + \frac{\gamma_y}{n\bar{u}_{k+1}^{-\beta}} \left\|U_{k+1}^{-\beta} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2\right) \\
&- \left(1+\lambda_k\right)\left(\frac{1}{n}\sum_{i=1}^n 2\left\langle g_{i,k}^y, y_{i,k} - y^*\left(\bar{x}_k\right)\right\rangle - \frac{1}{n}\sum_{i=1}^n 2\left\langle\left(\frac{u_{i,k+1}^{-\beta} - \bar{u}_{k+1}^{-\beta}}{\bar{u}_{k+1}^{-\beta}}\right) g_{i,k}^y, y_{i,k} - y^*\left(\bar{x}_k\right)\right\rangle\right).
\end{aligned} \tag{37}$$

For the inner-product terms on the RHS, taking expectation on both sides, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[-2\left\langle g_{i,k}^y, y_{i,k} - y^*\left(\bar{x}_k\right)\right\rangle\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[-2\left\langle \nabla_y f_i\left(\bar{x}_k, y_{i,k}\right), y_{i,k} - y^*\left(\bar{x}_k\right)\right\rangle\right] \\
&+ \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[-2\left\langle \nabla_y f_i\left(x_{i,k}, y_{i,k}\right) - \nabla_y f_i\left(\bar{x}_k, y_{i,k}\right), y_{i,k} - y^*\left(\bar{x}_k\right)\right\rangle\right] \\
&\leqslant \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[-2\left(f_i\left(\bar{x}_k, y^*\left(\bar{x}_k\right)\right) - f_i\left(\bar{x}_k, y_{i,k}\right)\right) - \mu\left\|y_{i,k} - y^*\left(\bar{x}_k\right)\right\|^2\right] \\
&+ \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{8}{\mu}\left\|\nabla_y f_i\left(x_{i,k}, y_{i,k}\right) - \nabla_y f_i\left(\bar{x}_k, y_{i,k}\right)\right\|^2 + \frac{\mu}{8}\left\|y_{i,k} - \bar{y}^*\left(\bar{x}_k\right)\right\|^2\right] \\
&\leqslant \mathbb{E}\left[-2\left(f\left(\bar{x}_k, y^*\left(\bar{x}_k\right)\right) - f\left(\bar{x}_k, \bar{y}_k\right)\right)\right] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[-2\left(f_i\left(\bar{x}_k, \bar{y}_k\right) - f_i\left(\bar{x}_k, y_{i,k}\right)\right)\right] \\
&+ \frac{8\kappa L}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|x_{i,k} - \bar{x}_k\right\|^2\right] - \frac{7\mu}{8n}\sum_{i=1}^n \mathbb{E}\left[\left\|y_{i,k} - y^*\left(\bar{x}_k\right)\right\|^2\right],
\end{aligned} \tag{38}$$

where we have used Young's inequality and strong-concavity of f_i, and

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[-2\left\langle\left(\frac{u_{i,k+1}^{-\beta} - \bar{u}_{k+1}^{-\beta}}{\bar{u}_{k+1}^{-\beta}}\right) g_{i,k}^y, y_{i,k} - y^*\left(\bar{x}_k\right)\right\rangle\right] \\
&\leqslant \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{8}{\mu}\left\|\left(\frac{u_{i,k+1}^{-\beta} - \bar{u}_{k+1}^{-\beta}}{\bar{u}_{k+1}^{-\beta}}\right) g_{i,k}^y\right\|^2 + \frac{\mu}{8}\left\|y_{i,k} - y^*\left(\bar{x}_k\right)\right\|^2\right].
\end{aligned} \tag{39}$$

For the consensus error of dual variable on the objective function, using strong-concavity of f_i and Jensen's inequality, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n -2\left(f_i\left(\bar{x}_k, \bar{y}_k\right) - f_i\left(\bar{x}_k, y_{i,k}\right)\right) \\
&\leqslant \frac{1}{n}\sum_{i=1}^n 2\left\langle \nabla_y f_i\left(\bar{x}_k, \bar{y}_k\right), y_{i,k} - \bar{y}_k\right\rangle - \frac{\mu}{n}\left\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\right\|^2 \\
&\leqslant 2C\frac{1}{n}\sum_{i=1}^n \left\|y_{i,k} - \bar{y}_k\right\| \leqslant 2C\sqrt{\frac{1}{n}\left\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\right\|^2}.
\end{aligned} \tag{40}$$

Letting $\lambda_k = \mu\gamma_y \bar{u}_{k+1}^{-\beta}/2$, we get

$$\begin{aligned}
&\mathbb{E}\left[f\left(\bar{x}_k, \bar{y}^*\left(\bar{x}_k\right)\right) - f\left(\bar{x}_k, \bar{y}_k\right)\right] \\
&\leqslant \mathbb{E}\left[\frac{1 - 3\mu\gamma_y \bar{u}_{k+1}^{-\beta}/4}{2\gamma_y \bar{u}_{k+1}^{-\beta} n} \left\|\mathbf{y}_k - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2 - \frac{\left\|\mathbf{y}_{k+1} - \mathbf{1}y^*\left(\bar{x}_{k+1}\right)\right\|^2}{\left(2 + \mu\gamma_y \bar{u}_{k+1}^{-\beta}\right)\gamma_y \bar{u}_{k+1}^{-\beta} n}\right] \\
&+ \frac{\gamma_x^2 \kappa^2 \left(1 + \zeta_v^2\right) G^{2\beta}}{n\mu\gamma_y^2} \mathbb{E}\left[\bar{v}_{k+1}^{-2\alpha} \left\|\nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2\right] \\
&+ \frac{\gamma_y \left(1 + \zeta_u^2\right)}{n} \sum_{i=1}^{n} \mathbb{E}\left[\bar{u}_{k+1}^{-\beta} \left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k\right)\right\|^2\right] + \frac{4\kappa L}{n} \mathbb{E}\left[\left\|\mathbf{x}_k - \mathbf{1}\bar{y}_k\right\|^2\right] \\
&+ \frac{4}{\mu} \mathbb{E}\left[\left\|\frac{\tilde{\boldsymbol{u}}_{k+1}^{-\beta}}{n\bar{u}_{k+1}^{-\beta}} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2\right] + C\mathbb{E}\left[\sqrt{\frac{1}{n}\left\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\right\|^2}\right].
\end{aligned} \tag{41}$$

By the κ-smoothness of y^*, we have

$$\begin{aligned}
&\left\|y^*\left(\bar{x}_{k+1}\right) - y^*\left(\bar{x}_k\right)\right\|^2 \\
&\leqslant \kappa^2 \left\|\bar{x}_{k+1} - \bar{x}_k\right\|^2 \\
&= \kappa^2 \left\|\gamma_x \bar{v}_{k+1}^{-\alpha} \frac{\mathbf{1}^T}{n} \nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k\right) - \gamma_x \frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n} \nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2 \\
&\leqslant \frac{2\gamma_x^2 \kappa^2 \left(1 + \zeta_v^2\right) \bar{v}_{k+1}^{-2\alpha}}{n} \left\|\nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2.
\end{aligned} \tag{42}$$

Telescoping the obtained terms from 0 to $k_0 - 1$ and noticing that $\bar{u}_k \leqslant G$ for $k \leqslant k_0 - 1$ we complete the proof. □

For the second phase, i.e., $k \geqslant k_0$, we have the following lemma.

Lemma 7 (Second phase)**.** *Suppose Assumption 1-4 hold. If* $\bar{u}_k \leqslant G, k = 0, 1, \cdots, k_0 - 1$*, then we have*

$$\begin{aligned}
&\sum_{k=k_0}^{K-1} \mathbb{E}\left[f\left(\bar{x}_k, \bar{y}^*\left(\bar{x}_k\right)\right) - f\left(\bar{x}_k, \bar{y}_k\right)\right] \\
&\leqslant \sum_{k=k_0}^{K-1} \mathbb{E}\left[E_{1,k}\right] + \frac{8\gamma_x^2 \kappa^2 \left(1 + \zeta_v^2\right)}{\mu\gamma_y^2 G^{2\alpha - 2\beta}} \sum_{k=k_0}^{K-1} \left\|\nabla_x f\left(\bar{x}_k, \bar{y}_k\right)\right\|^2 \\
&+ \left(\frac{8\gamma_x^2 \kappa^2 L^2 \left(1 + \zeta_v^2\right)}{n\mu\gamma_y^2 G^{2\alpha - 2\beta}} + \frac{4\kappa L}{n}\right) \sum_{k=k_0}^{K-1} \mathbb{E}\left[\mathit{\Delta}_k\right] \\
&+ \frac{\gamma_y \left(1 + \zeta_u^2\right)}{n} \mathbb{E}\left[\bar{u}_{k+1}^{-\beta} \left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k\right)\right\|^2\right] + C \sum_{k=k_0}^{K-1} \mathbb{E}\left[\sqrt{\frac{1}{n}\left\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\right\|^2}\right] \\
&+ \frac{\gamma_x^2 \left(1 + \zeta_v^2\right)}{\gamma_y \bar{v}_1^{\alpha - \beta}} \left(\kappa^2 + \frac{2\gamma_x^2 \left(1 + \zeta_v^2\right) C^2 \hat{L}^2}{\mu\gamma_y \bar{v}_1^{2\alpha - \beta}}\right) \sum_{k=k_0}^{K-1} \mathbb{E}\left[\frac{\bar{v}_{k+1}^{-\alpha}}{n} \left\|\nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2\right] \\
&+ \frac{4\gamma_x \kappa \left(1 + \zeta_v\right) C^2}{\mu\gamma_y \bar{v}_1^{\alpha}} \mathbb{E}\left[\bar{u}_K^{\beta}\right] + \frac{4}{\mu} \sum_{k=k_0}^{K-1} \mathbb{E}\left[\left\|\frac{\tilde{\boldsymbol{u}}_{k+1}^{-\beta}}{n\bar{u}_{k+1}^{-\beta}} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2\right].
\end{aligned} \tag{43}$$

Proof. Firstly, by the non-expansiveness of projection operator, we have

$$\begin{aligned}
&\left\|y_{i,k+1}-y^*\left(\bar{x}_{k+1}\right)\right\|^2\\
&\leqslant\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_{k+1}\right)\right\|^2-\left\|y_{i,k+1}-\hat{y}_{i,k+1}\right\|^2\\
&=\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|^2+\left\|y^*\left(\bar{x}_{k+1}\right)-y^*\left(\bar{x}_k\right)\right\|^2\\
&-2\left\langle\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right),y^*\left(\bar{x}_{k+1}\right)-y^*\left(\bar{x}_k\right)\right\rangle\\
&=\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|^2+\left\|y^*\left(\bar{x}_{k+1}\right)-y^*\left(\bar{x}_k\right)\right\|^2\\
&-2\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\nabla y^*\left(\bar{x}_k\right)\left(\bar{x}_{k+1}-\bar{x}_k\right)^T\\
&-2\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\left(y^*\left(\bar{x}_{k+1}\right)-y^*\left(\bar{x}_k\right)-\nabla y^*\left(\bar{x}_k\right)\left(\bar{x}_{k+1}-\bar{x}_k\right)^T\right).
\end{aligned}\tag{44}$$

Then, for the first inner-product term on the RHS, letting $\nabla_x\tilde{F}_k=\nabla_xF\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k\right)-\nabla_xF\left(\mathbf{x}_k,\mathbf{y}_k\right)$, we get

$$\begin{aligned}
&-2\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\nabla y^*\left(\bar{x}_k\right)\left(\bar{x}_{k+1}-\bar{x}_k\right)^T\\
&=2\gamma_x\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\nabla y^*\left(\bar{x}_k\right)\left(\nabla_xF\left(\mathbf{x}_k,\mathbf{y}_k\right)\right)^T\left(\frac{\mathbf{1}\bar{v}_{k+1}^{-\alpha}}{n}+\frac{\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}}{n}\right)\\
&+2\gamma_x\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\nabla y^*\left(\bar{x}_k\right)\left(\nabla_x\tilde{F}_k\right)^T\left(\frac{\mathbf{1}\bar{v}_{k+1}^{-\alpha}}{n}+\frac{\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}}{n}\right)\\
&\leqslant2\gamma_x\kappa\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|\left\|\left(\nabla_xF\left(\mathbf{x}_k,\mathbf{y}_k\right)\right)^T\left(\frac{\mathbf{1}\bar{v}_{k+1}^{-\alpha}}{n}+\frac{\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}}{n}\right)\right\|\\
&+2\gamma_x\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\nabla y^*\left(\bar{x}_k\right)\left(\nabla_x\tilde{F}_k\right)^T\left(\frac{\mathbf{1}\bar{v}_{k+1}^{-\alpha}}{n}+\frac{\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}}{n}\right).
\end{aligned}\tag{45}$$

wherein the last inequality we have used the fact that y^* is κ-Lipschitz. Then, using Young's inequality with parameter λ_k, we get

$$\begin{aligned}
&-2\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\nabla y^*\left(\bar{x}_k\right)\left(\bar{x}_{k+1}-\bar{x}_k\right)^T\\
&\leqslant\lambda_k\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|^2\\
&+\frac{2\gamma_x^2\bar{v}_{k+1}^{-2\alpha}\kappa^2}{\lambda_k}\left(\left\|\frac{\mathbf{1}^T}{n}\nabla_xF\left(\mathbf{x}_k,\mathbf{y}_k\right)\right\|^2+\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_xF\left(\mathbf{x}_k,\mathbf{y}_k\right)\right\|^2\right)\\
&+2\gamma_x\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\nabla y^*\left(\bar{x}_k\right)\left(\nabla_x\tilde{F}_k\right)^T\left(\frac{\mathbf{1}\bar{v}_{k+1}^{-\alpha}}{n}+\frac{\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}}{n}\right).
\end{aligned}\tag{46}$$

For the second inner-product term on the RHS, noticing that y^* is $\hat{L}=\kappa\left(1+\kappa\right)^2$ smooth given in Lemma 2, we have

$$\begin{aligned}
&2\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\left(y^*\left(\bar{x}_k\right)-y^*\left(\bar{x}_{k+1}\right)+\nabla y^*\left(\bar{x}_k\right)\left(\bar{x}_{k+1}-\bar{x}_k\right)^T\right)\\
&\leqslant2\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|\left\|y^*\left(\bar{x}_k\right)-y^*\left(\bar{x}_{k+1}\right)+\nabla y^*\left(\bar{x}_k\right)\left(\bar{x}_{k+1}-\bar{x}_k\right)\right\|^2\\
&\leqslant2\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|\frac{\hat{L}}{2}\left\|\bar{x}_{k+1}-\bar{x}_k\right\|^2\\
&\leqslant\gamma_x^2\hat{L}\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|\left\|\left(\frac{\bar{v}_{k+1}^{-\alpha}\mathbf{1}^T}{n}+\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n}\right)\nabla_xF\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\\
&\leqslant\gamma_x^2\hat{L}\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|\frac{2\bar{v}_{k+1}^{-2\alpha}\left(1+\zeta_v^2\right)C}{n}\left\|\nabla_xF\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|\\
&\leqslant\tau\gamma_x^2\bar{v}_{k+1}^{-2\alpha}\left(1+\zeta_v^2\right)C^2\hat{L}\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|^2+\frac{\gamma_x^2\bar{v}_{k+1}^{-2\alpha}\left(1+\zeta_v^2\right)\hat{L}}{\tau n}\left\|\nabla_xF\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2,
\end{aligned}\tag{47}$$

wherein the last inequality we have used Young's inequality with parameter τ. Plugging the obtained inequalities into (44), we get

$$\begin{aligned}
&\left\|y_{i,k+1}-y^*\left(\bar{x}_{k+1}\right)\right\|^2\\
&\leqslant\left(1+\lambda_k+\tau\gamma_x^2\bar{v}_{k+1}^{-2\alpha}\left(1+\zeta_v^2\right)C^2\hat{L}\right)\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|^2\\
&+\frac{\gamma_x^2\bar{v}_{k+1}^{-2\alpha}\left(1+\zeta_v^2\right)}{n}\left(2\kappa^2+\frac{\hat{L}}{\tau}\right)\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k\right)\right\|^2\\
&+\frac{2\gamma_x^2\bar{v}_{k+1}^{-2\alpha}\kappa^2}{\lambda_k}\left(\left\|\frac{\mathbf{1}^T}{n}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k\right)\right\|^2+\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k\right)\right\|^2\right)\\
&+2\gamma_x\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\nabla y^*\left(\bar{x}_k\right)\left(\nabla_x\tilde{F}\right)^T\left(\frac{\mathbf{1}\bar{v}_{k+1}^{-\alpha}}{n}+\frac{\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}}{n}\right).
\end{aligned}\tag{48}$$

Setting the parameters for Young's inequalities we used as follows,

$$\lambda_k=\frac{\mu\gamma_y\bar{u}_{k+1}^{-\beta}}{4},\quad\tau=\frac{\mu\gamma_y\bar{v}_0^{2\alpha-\beta}}{4\gamma_x^2\left(1+\zeta_v^2\right)C^2\hat{L}},\tag{49}$$

then we get

$$\begin{aligned}
&\left\|y_{i,k+1}-y^*\left(\bar{x}_{k+1}\right)\right\|^2\\
&\leqslant\left(1+\frac{\mu\gamma_y\bar{u}_{k+1}^{-\beta}}{2}\right)\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|^2\\
&+\frac{\gamma_x^2\left(1+\zeta_v^2\right)}{n}\left(2\kappa^2+\frac{4\gamma_x^2\left(1+\zeta_v^2\right)C^2\hat{L}^2}{\mu\gamma_y\bar{v}_0^{2\alpha-\beta}}\right)\bar{v}_{k+1}^{-2\alpha}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k\right)\right\|^2\\
&+\frac{8\gamma_x^2\bar{v}_{k+1}^{-2\alpha}\kappa^2}{\mu\gamma_y\bar{u}_{k+1}^{-\beta}}\left(\left\|\frac{\mathbf{1}^T}{n}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k\right)\right\|^2+\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k\right)\right\|^2\right)\\
&+2\gamma_x\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\nabla y^*\left(\bar{x}_k\right)\left(\nabla_x\tilde{F}_k\right)^T\left(\frac{\mathbf{1}\bar{v}_{k+1}^{-\alpha}}{n}+\frac{\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}}{n}\right).
\end{aligned}\tag{50}$$

Recalling that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\frac{1}{\gamma_y\bar{u}_{k+1}^{-\beta}}\left\|\hat{y}_{i,k+1}-\bar{y}^*\left(\bar{x}_k\right)\right\|^2\right]\\
&\leqslant\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\frac{1-3\mu\gamma_y\bar{u}_{k+1}^{-\beta}/4}{\gamma_y\bar{u}_{k+1}^{-\beta}}\left\|y_{i,k}-\bar{y}^*\left(\bar{x}_k\right)\right\|^2\right]+\frac{8\kappa L}{n}\mathbb{E}\left[\left\|\mathbf{x}_k-\mathbf{1}\bar{y}_k\right\|^2\right]\\
&+\frac{2\gamma_y\left(1+\zeta_u^2\right)}{n}\mathbb{E}\left[\bar{u}_{k+1}^{-\beta}\left\|\nabla_y F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k\right)\right\|^2\right]-\mathbb{E}\left[2\left(f\left(\bar{x}_k,\bar{y}^*\left(\bar{x}_k\right)\right)-f\left(\bar{x}_k,\bar{y}_k\right)\right)\right]\\
&+\frac{8}{\mu}\mathbb{E}\left[\left\|\frac{\tilde{\boldsymbol{u}}_{k+1}^{-\beta}}{n\bar{u}_{k+1}^{-\beta}}\nabla_y F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^y\right)\right\|^2\right]+2C\mathbb{E}\left[\sqrt{\frac{1}{n}\left\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\right\|^2}\right],
\end{aligned}$$

and multiplying by $\frac{2}{\left(2+\mu\gamma_y\bar{u}_{k+1}^{-\beta}\right)\gamma_y\bar{u}_{k+1}^{-\beta}}$ on both sides of (50), we obtain that

$$
\begin{aligned}
&\mathbb{E}\left[f\left(\bar{x}_k,\bar{y}^*\left(\bar{x}_k\right)\right)-f\left(\bar{x}_k,\bar{y}_k\right)\right]\\
&\leqslant \mathbb{E}\left[E_{1,k}\right]+\frac{\gamma_y\left(1+\zeta_u^2\right)}{n}\mathbb{E}\left[\bar{u}_{k+1}^{-\beta}\left\|\nabla_y F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k\right)\right\|^2\right]+\frac{4\kappa L}{n}\mathbb{E}\left[\left\|\mathbf{x}_k-\mathbf{1}\bar{y}_k\right\|^2\right]\\
&+\frac{4}{\mu}\mathbb{E}\left[\left\|\frac{\tilde{\boldsymbol{u}}_{k+1}^{-\beta}}{n\bar{u}_{k+1}^{-\beta}}\nabla_y F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^y\right)\right\|^2\right]+C\mathbb{E}\left[\sqrt{\frac{1}{n}\left\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\right\|^2}\right]\\
&\underbrace{+\mathbb{E}\left[\frac{4\gamma_x^2\bar{v}_{k+1}^{-2\alpha}\kappa^2}{\mu\gamma_y^2\bar{u}_{k+1}^{-2\beta}}\left(\left\|\frac{\mathbf{1}^T}{n}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k\right)\right\|^2+\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k\right)\right\|^2\right)\right]}_{\mathbb{E}\left[E_{2,k}\right]}\\
&\underbrace{+\frac{\gamma_x^2\left(1+\zeta_v^2\right)}{n}\left(\kappa^2+\frac{2\gamma_x^2\left(1+\zeta_v^2\right)C^2\hat{L}^2}{\mu\gamma_y\bar{v}_1^{2\alpha-\beta}}\right)\mathbb{E}\left[\frac{\bar{v}_{k+1}^{-2\alpha}}{\gamma_y\bar{u}_{k+1}^{-\beta}}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k\right)\right\|^2\right]}_{\mathbb{E}\left[E_{3,k}\right]}\\
&\underbrace{+\frac{1}{n}\sum_{i=1}^n\mathbb{E}\left[\frac{\gamma_x}{\gamma_y\bar{u}_{k+1}^{-\beta}}\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\nabla y^*\left(\bar{x}_k\right)\left(\nabla_x\tilde{F}_k\right)^T\left(\frac{\mathbf{1}\bar{v}_{k+1}^{-\alpha}}{n}+\frac{\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}}{n}\right)\right]}_{\mathbb{E}\left[E_{4,k}\right]}.
\end{aligned}
\tag{51}
$$

Telescoping the terms from t_0 to $K-1$, we get

$$
\begin{aligned}
&\sum_{k=k_0}^{K-1}\mathbb{E}\left[f\left(\bar{x}_k,\bar{y}^*\left(\bar{x}_k\right)\right)-f\left(\bar{x}_k,\bar{y}_k\right)\right]\\
&\leqslant\sum_{k=k_0}^{K-1}\mathbb{E}\left[E_{1,k}\right]+\sum_{k=k_0}^{K-1}\mathbb{E}\left[E_{2,k}\right]+\sum_{k=k_0}^{K-1}\mathbb{E}\left[E_{3,k}\right]+\sum_{k=k_0}^{K-1}\mathbb{E}\left[E_{4,k}\right]\\
&+\frac{\gamma_y\left(1+\zeta_u^2\right)}{n}\mathbb{E}\left[\bar{u}_{k+1}^{-\beta}\left\|\nabla_y F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k\right)\right\|^2\right]+\frac{4\kappa L}{n}\sum_{k=k_0}^{K-1}\mathbb{E}\left[\left\|\mathbf{x}_k-\mathbf{1}\bar{y}_k\right\|^2\right]\\
&+\frac{4}{\mu}\mathbb{E}\left[\left\|\frac{\tilde{\boldsymbol{u}}_{k+1}^{-\beta}}{n\bar{u}_{k+1}^{-\beta}}\nabla_y F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^y\right)\right\|^2\right]+C\sum_{k=k_0}^{K-1}\mathbb{E}\left[\sqrt{\frac{1}{n}\left\|\mathbf{y}_k-\mathbf{1}\bar{y}_k\right\|^2}\right].
\end{aligned}
\tag{52}
$$

Next we need to further bound the running sums of $\mathbb{E}\left[E_{2,k}\right]$, $\mathbb{E}\left[E_{3,k}\right]$ and $\mathbb{E}\left[E_{4,k}\right]$ respectively. For $\mathbb{E}\left[E_{2,k}\right]$, with the help of Assumption 2 and noticing that $\bar{u}_k\leqslant G, k=0,1,\cdots,k_0-1$, we get

$$
\begin{aligned}
&\sum_{k=k_0}^{K-1}\mathbb{E}\left[E_{2,k}\right]\\
&\leqslant\sum_{k=k_0}^{K-1}\mathbb{E}\left[\frac{4\gamma_x^2\bar{v}_{k+1}^{-2\alpha}\kappa^2}{\mu\gamma_y^2\bar{u}_{k+1}^{-2\beta}}\left(\left\|\frac{\mathbf{1}^T}{n}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k\right)\right\|^2+\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k\right)\right\|^2\right)\right]\\
&\leqslant\frac{8\gamma_x^2\kappa^2\left(1+\zeta_v^2\right)}{\mu\gamma_y^2G^{2\alpha-2\beta}}\sum_{k=k_0}^{K-1}\mathbb{E}\left[\left\|\nabla_x f\left(\bar{x}_k,\bar{y}_k\right)\right\|^2+\frac{L^2}{n}\varDelta_k\right].
\end{aligned}
\tag{53}
$$

Then, for the term $\mathbb{E}\left[E_{3,k}\right]$, noticing that $\bar{u}_{k+1} \leqslant \bar{v}_{k+1}$ and $\bar{v}_{k+1} \geqslant \bar{v}_1$, we have

$$\begin{aligned}
&\sum_{k=k_0}^{K-1} \mathbb{E}\left[E_{3,k}\right] \\
&\leqslant \sum_{k=k_0}^{K-1} \mathbb{E}\left[\frac{\gamma_x^2\left(1+\zeta_v^2\right)}{n\gamma_y}\left(\kappa^2+\frac{2\gamma_x^2\left(1+\zeta_v^2\right)C^2\hat{L}^2}{\mu\gamma_y\bar{v}_1^{2\alpha-\beta}}\right)\frac{\bar{v}_{k+1}^{-2\alpha}}{\bar{u}_{k+1}^{-\beta}}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right] \\
&\leqslant \frac{\gamma_x^2\left(1+\zeta_v^2\right)}{\gamma_y\bar{v}_1^{\alpha-\beta}}\left(\kappa^2+\frac{2\gamma_x^2\left(1+\zeta_v^2\right)C^2\hat{L}^2}{\mu\gamma_y\bar{v}_1^{2\alpha-\beta}}\right)\sum_{k=k_0}^{K-1}\mathbb{E}\left[\frac{\bar{v}_{k+1}^{-\alpha}}{n}\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right].
\end{aligned} \tag{54}$$

For the term $E_{4,k}$, we denote

$$e_k := \frac{\gamma_x}{\gamma_y\bar{u}_{k+1}^{-\beta}}\left(\frac{1}{n}\sum_{i=1}^n\left(\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right)^T\right)\nabla y^*\left(\bar{x}_k\right)\left(\nabla_x\tilde{F}_k\right)^T\left(\frac{\mathbf{1}}{n}+\frac{\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}}{n\bar{v}_{k+1}^{-\alpha}}\right),$$

then we have

$$\begin{aligned}
\left|e_k\right| &\leqslant \frac{\gamma_x\kappa}{\gamma_y\bar{u}_{k+1}^{-\beta}}\frac{1}{n}\sum_{i=1}^n\left\|\hat{y}_{i,k+1}-y^*\left(\bar{x}_k\right)\right\|\left\|\left(\nabla_x\tilde{F}_k\right)^T\left(\frac{1}{n}+\frac{\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}}{n\bar{v}_{k+1}^{-\alpha}}\right)\right\| \\
&\leqslant \frac{\gamma_x\kappa\left(1+\zeta_v\right)}{\gamma_y\sqrt{n}\bar{u}_{k+1}^{-\beta}}\left(\frac{1}{n}\sum_{i=1}^n\frac{1}{\mu}\left\|\nabla_y f\left(\bar{x}_k,\hat{y}_{i,k+1}\right)-\nabla_y f\left(\bar{x}_k,y^*\right)\right\|\right)\left\|\nabla_x\tilde{F}\right\| \\
&\leqslant \underbrace{\frac{2\gamma_x\kappa\left(1+\zeta_v\right)C^2\bar{u}_K^\beta}{\mu\gamma_y}}_{M},
\end{aligned} \tag{55}$$

where we have used the Lipschitz continuity of y^* given in Lemma 2 and Assumption 3. Then, noticing that $\mathbb{E}\left[\nabla_x\tilde{F}_k\right]=0$, we obtain

$$\begin{aligned}
\sum_{k=k_0}^{K-1}\mathbb{E}\left[E_{4,k}\right] &= \sum_{k=k_0}^{K-1}\mathbb{E}\left[e_k\bar{v}_{k+1}^{-\alpha}\right] \\
&= \mathbb{E}\left[e_{k_0}\bar{v}_{k_0+1}^{-\alpha}\right]+\underbrace{\sum_{k=k_0+1}^{K-1}\mathbb{E}\left[e_k\bar{v}_k^{-\alpha}\right]}_{0}+\sum_{k=k_0+1}^{K-1}\mathbb{E}\left[-e_k\underbrace{\left(\bar{v}_k^{-\alpha}-\bar{v}_{k+1}^{-\alpha}\right)}_{>0}\right] \\
&\leqslant \mathbb{E}\left[M\bar{v}_{k_0+1}^{-\alpha}\right]+\sum_{k=k_0+1}^{K-1}\mathbb{E}\left[M\left(\bar{v}_k^{-\alpha}-\bar{v}_{k+1}^{-\alpha}\right)\right] \\
&\leqslant 2\mathbb{E}\left[M\bar{v}_{k_0+1}^{-\alpha}\right] \leqslant \frac{4\gamma_x\kappa\left(1+\zeta_v\right)C^2}{\mu\gamma_y\bar{v}_1^\alpha}\mathbb{E}\left[\bar{u}_K^\beta\right].
\end{aligned} \tag{56}$$

Therefore, combining the obtained inequalities, we complete the proof. □

Now, it remains to bound the term $E_{1,k}$.

Lemma 8. *Suppose Assumption 1-4 hold. Then, we have*

$$\sum_{k=0}^{K-1}\mathbb{E}\left[E_{1,k}\right] \leqslant \frac{1}{2\gamma_y\bar{u}_1^{-\beta}n}\left\|\mathbf{y}_0-\mathbf{1}y^*\left(\bar{x}_0\right)\right\|^2+\frac{2\left(4\beta C^2\right)^{2+\frac{1}{1-\beta}}}{\mu^{3+\frac{1}{1-\beta}}\gamma_y^{2+\frac{1}{1-\beta}}\bar{u}_1^{2-2\beta}}. \tag{57}$$

Proof. Recalling the definition of $E_{1,k}$ as given in (34), we have

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}\left[\frac{1-3\mu\gamma_y \bar{u}_{k+1}^{-\beta}/4}{2\gamma_y \bar{u}_{k+1}^{-\beta} n}\left\|\mathbf{y}_k - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2 - \frac{\left\|\mathbf{y}_{k+1} - \mathbf{1}y^*\left(\bar{x}_{k+1}\right)\right\|^2}{\left(2+\mu\gamma_y \bar{u}_{k+1}^{-\beta}\right)\gamma_y \bar{u}_{k+1}^{-\beta} n}\right] \\
&\leqslant \frac{1-3\mu\gamma_y \bar{u}_1^{-\beta}/4}{2\gamma_y \bar{u}_1^{-\beta} n}\left\|\mathbf{y}_0 - \mathbf{1}y^*\left(\bar{x}_0\right)\right\|^2 \\
&+ \sum_{k=1}^{K-1} \mathbb{E}\left[\left(\frac{1-3\mu\gamma_y \bar{u}_{k+1}^{-\beta}/4}{2\gamma_y \bar{u}_{k+1}^{-\beta} n} - \frac{1}{2n\gamma_y \bar{u}_k^{-\beta}\left(2+\mu\gamma_y \bar{u}_k^{-\beta}\right)}\right)\left\|\mathbf{y}_k - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2\right] \\
&\leqslant \frac{1-3\mu\gamma_y \bar{u}_1^{-\beta}/4}{2\gamma_y \bar{u}_1^{-\beta} n}\left\|\mathbf{y}_0 - \mathbf{1}y^*\left(\bar{x}_0\right)\right\|^2 \\
&+ \sum_{k=1}^{K-1} \mathbb{E}\left[\left(\frac{1}{2\gamma_y \bar{u}_{k+1}^{-\beta}} - \frac{1}{4\gamma_y \bar{u}_k^{-\beta}} - \frac{\mu}{8} + \underbrace{\frac{\mu}{2\left(2+\mu\gamma_y \bar{u}_k^{-\beta}\right)} - \frac{\mu}{2}}_{<0}\right)\frac{1}{n}\left\|\mathbf{y}_k - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2\right].
\end{aligned} \tag{58}$$

Next, we show that the term $\frac{1}{2\gamma_y \bar{u}_{k+1}^{-\beta}} - \frac{1}{2\gamma_y \bar{u}_k^{-\beta}} - \frac{\mu}{8}$ is positive for only a constant number of iterations. If the term is positive at iteration k, then we have

$$\begin{aligned}
0 &< \frac{\bar{u}_{k+1}^{\beta}}{2\gamma_y} - \frac{\bar{u}_k^{\beta}}{2\gamma_y} - \frac{\mu}{8} \\
&\leqslant \bar{u}_k^{\beta}\frac{\left(1+\left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2/n\bar{u}_k^{\beta}\right)^{\beta}}{2\gamma_y} - \frac{\bar{u}_k^{\beta}}{2\gamma_y} - \frac{\mu}{8} \\
&\leqslant \bar{u}_k^{\beta}\frac{\left(1+\beta\left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2/n\bar{u}_k\right)}{2\gamma_y} - \frac{\bar{u}_k^{\beta}}{2\gamma_y} - \frac{\mu}{8} \\
&= \frac{\beta\left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2}{2\gamma_y n\bar{u}_k^{1-\beta}} - \frac{\mu}{8},
\end{aligned} \tag{59}$$

wherein the last inequality we used Bernoulli's inequality. Then we have the following two conditions,

$$\begin{cases}
\frac{1}{n}\left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k\right)\right\|^2 \geqslant \frac{\gamma_y \bar{u}_{k+1}^{1-\beta}}{4\beta} \geqslant \frac{\gamma_y \bar{u}_1^{1-\beta}}{4\beta}, \\
\frac{4\beta G^2}{\mu\gamma_y} \geqslant \frac{4\beta\left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2}{\mu\gamma_y n} \geqslant \bar{u}_{k+1}^{1-\beta},
\end{cases} \tag{60}$$

which implies that we have at most

$$\left(\frac{4\beta C^2}{\mu\gamma_y}\right)^{\frac{1}{1-\beta}}\frac{4\beta}{\mu\gamma_y \bar{u}_1^{1-\beta}} \tag{61}$$

constant number of iterations when the term is positive. Furthermore, when the term is positive, by the inequality (59), we have

$$\begin{aligned}
&\left(\frac{1}{2\gamma_y \bar{u}_{k+1}^{-\beta}} - \frac{1}{2\gamma_y \bar{u}_k^{-\beta}} - \frac{\mu}{8}\right)\frac{1}{n}\left\|\mathbf{y}_k - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2 \\
&\leqslant \frac{\beta\left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2}{2\gamma_y n\bar{u}_1^{1-\beta}}\frac{1}{n}\left\|\mathbf{y}_k - \mathbf{1}y^*\left(\bar{x}_k\right)\right\|^2 \\
&\leqslant \frac{\beta C^2}{2\mu^2\gamma_y \bar{u}_1^{1-\beta}}\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla_y f_i\left(\bar{x}_k, y_{i,k}\right) - \nabla_y f_i\left(\bar{x}_k, y^*\right)\right\|^2 \\
&\leqslant \frac{2\beta C^4}{\mu^2\gamma_y \bar{u}_1^{1-\beta}},
\end{aligned} \tag{62}$$

where we have used the concavity of f_i in y and Assumption 3. Then, we have

$$\begin{aligned}
&\sum_{k=1}^{K-1} \mathbb{E}\left[\left(\frac{1}{2\gamma_y \bar{u}_{k+1}^{-\beta}} - \frac{1}{2\gamma_y \bar{u}_{k}^{-\beta}} - \frac{\mu}{8}\right) \frac{1}{n} \left\|\mathbf{y}_k - \mathbf{1} y^*\left(\bar{x}_k\right)\right\|^2\right] \\
&\leqslant \frac{2\beta C^4}{\mu^2 \gamma_y \bar{u}_1^{1-\beta}} \left(\frac{4\beta C^2}{\mu \gamma_y}\right)^{\frac{1}{1-\beta}} \frac{4\beta}{\mu \gamma_y \bar{u}_1^{1-\beta}} \qquad (63)\\
&\leqslant \frac{2\left(4\beta C^2\right)^{2+\frac{1}{1-\beta}}}{\mu^{3+\frac{1}{1-\beta}} \gamma_y^{2+\frac{1}{1-\beta}} \bar{u}_1^{2-2\beta}},
\end{aligned}$$

which completes the proof. □

Next, we show in the following lemma that the inconsistency terms, as described in (5), exhibit asymptotic convergence for the proposed D-AdaST algorithm.

Lemma 9 (Convergence of inconsistency terms)**.** *Suppose Assumption 1-4 hold. For the proposed D-AdaST in Algorithm 1, we have*

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n \bar{v}_{k+1}^{-\alpha}} \nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2\right] \leqslant \sqrt{\frac{1}{n^{1-\alpha}} \left(\frac{4\rho_W}{\left(1-\rho_W\right)^2}\right)^{\alpha} \frac{\left(1+\zeta_v\right) \zeta_v C^{2-\alpha}}{(1-\alpha) K^{\alpha}}}, \qquad (64)$$

and

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\frac{\left(\tilde{\boldsymbol{u}}_{k+1}^{-\beta}\right)^T}{n \bar{u}_{k+1}^{-\beta}} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2\right] \leqslant \sqrt{\frac{1}{n^{1-\beta}} \left(\frac{4\rho_W}{\left(1-\rho_W\right)^2}\right)^{\beta} \frac{\left(1+\zeta_u\right) \zeta_u C^{2-\beta}}{(1-\beta) K^{\beta}}}. \qquad (65)$$

Proof. By the definition of $v_{i,k}$ in (3), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n \bar{v}_{k+1}^{-\alpha}} \nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2\right] \\
&\leqslant \mathbb{E}\left[\frac{1}{n^2} \sum_{i=1}^{n} \left(\bar{v}_{k+1}^{\alpha} - v_{i,k+1}^{\alpha}\right)^2 \frac{\left\|g_{i,k}^x\right\|^2}{v_{i,k+1}^{2\alpha}}\right] \qquad (66)\\
&\leqslant \mathbb{E}\left[\frac{1}{n^2} \sum_{i=1}^{n} \left(\bar{v}_{k+1}^{\alpha} - v_{i,k+1}^{\alpha}\right)^2 \frac{\bar{v}_{k+1}^{\alpha}}{v_{i,k+1}^{2\alpha}} \frac{\left\|g_{i,k}^x\right\|^2}{\bar{v}_{k+1}^{\alpha}}\right].
\end{aligned}$$

Noticing that $\frac{\left|\bar{v}_{k+1}^{\alpha}-v_{i,k+1}^{\alpha}\right|}{v_{i,k+1}^{\alpha}} \leqslant \zeta_v$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right] \\
&\leqslant \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(\bar{v}_{k+1}^{\alpha}-v_{i,k+1}^{\alpha}\right)^2\left(\frac{\bar{v}_{k+1}^{\alpha}-v_{i,k+1}^{\alpha}}{v_{i,k+1}^{2\alpha}}+\frac{1}{v_{i,k+1}^{\alpha}}\right)\frac{\left\|g_{i,k}^x\right\|^2}{\bar{v}_{k+1}^{\alpha}}\right] \\
&\leqslant \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\frac{\left(\bar{v}_{k+1}^{\alpha}-v_{i,k+1}^{\alpha}\right)^2}{v_{i,k+1}^{2\alpha}}\left|\bar{v}_{k+1}^{\alpha}-v_{i,k+1}^{\alpha}\right|\frac{\left\|g_{i,k}^x\right\|^2}{\bar{v}_{k+1}^{\alpha}}\right] \\
&+\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\frac{\left|\bar{v}_{k+1}^{\alpha}-v_{i,k+1}^{\alpha}\right|}{v_{i,k+1}^{\alpha}}\left|\bar{v}_{k+1}^{\alpha}-v_{i,k+1}^{\alpha}\right|\frac{\left\|g_{i,k}^x\right\|^2}{\bar{v}_{k+1}^{\alpha}}\right] \\
&\leqslant \left(1+\zeta_v\right)\zeta_v\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\bar{v}_{k+1}^{\alpha}-v_{i,k+1}^{\alpha}\right|\frac{1}{n}\sum_{i=1}^{n}\frac{\left\|g_{i,k}^x\right\|^2}{\bar{v}_{k+1}^{\alpha}}\right].
\end{aligned}
\tag{67}
$$

By Lemma 4, we get

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right] \\
&\leqslant \left(1+\zeta_v\right)\zeta_v\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\bar{v}_{k+1}^{\alpha}-v_{i,k+1}^{\alpha}\right|\frac{1}{K}\sum_{k=0}^{K-1}\frac{\frac{1}{n}\sum_{i=1}^{n}\left\|g_{i,k}^x\right\|^2}{\bar{v}_{k+1}^{\alpha}}\right] \\
&\leqslant \left(1+\zeta_v\right)\zeta_v\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\bar{v}_{k+1}^{\alpha}-v_{i,k+1}^{\alpha}\right|\right]\frac{C^{2-2\alpha}}{\left(1-\alpha\right)K^{\alpha}} \\
&\leqslant \left(1+\zeta_v\right)\zeta_v\sqrt{\frac{1}{n}\mathbb{E}\left[\left\|\boldsymbol{v}_{k+1}-\mathbf{1}\bar{v}_{k+1}\right\|^{2\alpha}\right]}\frac{C^{2-2\alpha}}{\left(1-\alpha\right)K^{\alpha}}.
\end{aligned}
\tag{68}
$$

Next, for the term of inconsistency of the stepsize $\left\|\boldsymbol{v}_k-\mathbf{1}\bar{v}_k\right\|^2$, we consider two cases due to the max operator we used. At iteration k, for the case $\mathbf{m}_k^x \geqslant \mathbf{m}_k^y$ with $\left\|\mathbf{m}_0^x-\mathbf{1}\bar{m}_0^x\right\|^2=0$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{v}_{k+1}-\mathbf{1}\bar{v}_{k+1}\right\|^2\right]=\mathbb{E}\left[\left\|\mathbf{m}_{k+1}^x-\mathbf{1}\bar{m}_{k+1}^x\right\|^2\right] \\
&=\mathbb{E}\left[\left\|\left(W-\mathbf{J}\right)\left(\mathbf{m}_k^x-\mathbf{1}\bar{m}_k^x\right)+\eta_k\left(W-\mathbf{J}\right)\boldsymbol{h}_k^x\right\|^2\right] \\
&\leqslant \frac{1+\rho_W}{2}\mathbb{E}\left[\left\|\mathbf{m}_k^x-\mathbf{1}\bar{m}_k^x\right\|^2\right]+\frac{\left(1+\rho_W\right)\rho_W}{1-\rho_W}\mathbb{E}\left[\left\|\boldsymbol{h}_k^x\right\|^2\right] \\
&\leqslant \left(\frac{1+\rho_W}{2}\right)^k\mathbb{E}\left[\left\|\mathbf{m}_0^x-\mathbf{1}\bar{m}_0^x\right\|^2\right]+\frac{nC^2\left(1+\rho_W\right)\rho_W}{1-\rho_W}\sum_{t=0}^{k}\left(\frac{1+\rho_W}{2}\right)^{k-t} \\
&\leqslant \frac{2nC^2\left(1+\rho_W\right)\rho_W}{\left(1-\rho_W\right)^2}.
\end{aligned}
\tag{69}
$$

For the case $\mathbf{m}_k^x < \mathbf{m}_k^y$, with $\left\|\mathbf{m}_0^y-\mathbf{1}\bar{m}_0^y\right\|^2=0$,

$$
\mathbb{E}\left[\left\|\boldsymbol{v}_{k+1}-\mathbf{1}\bar{v}_{k+1}\right\|^2\right]=\mathbb{E}\left[\left\|\mathbf{m}_{k+1}^y-\mathbf{1}\bar{m}_{k+1}^y\right\|^2\right]\leqslant\frac{2nC^2\left(1+\rho_W\right)\rho_W}{\left(1-\rho_W\right)^2},
\tag{70}
$$

Combining these two cases, and using Lemma 4 and the fact $\|\boldsymbol{v}_k^{\alpha} - \mathbf{1}\bar{v}_k^{\alpha}\|^2 \leqslant \|\boldsymbol{v}_k - \mathbf{1}\bar{v}_k\|^{2\alpha}$ for $\alpha \in (0, 1)$, we obtain the result for primal decision variable. Following the same proof, we can also derive the result for dual decision variable. We thus complete the proof. □

We further give the following lemma to show that the inconsistency of stepsize remains uniformly bounded for the vanilla D-TiAda algorithm as given in (2).

Lemma 10 (Inconsistency for D-TiAda)**.** *Suppose Assumption 1-4 hold. Then, for D-TiAda, we have*

$$\begin{aligned} &\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\left(\tilde{\boldsymbol{v}}_{k+1}^{-\alpha}\right)^T}{n\bar{v}_{k+1}^{-\alpha}}\nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2\right] \leqslant \zeta_v^2 C^2, \\ &\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\left(\tilde{\boldsymbol{u}}_{k+1}^{-\beta}\right)^T}{n\bar{u}_{k+1}^{-\beta}}\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2\right] \leqslant \zeta_u^2 C^2. \end{aligned} \tag{71}$$

Proof. By the definition of inconsistency of stepsizes in (8) and Assumption 3 on bounded gradient, we immediately get the result. □

B.3 Proof of Theorem 1

Proof of Theorem 1. Consider a complete graph with 3 nodes where the functions corresponding to the nodes are as follows:

$$\begin{aligned} f_1(x, y) &= -\frac{1}{2}y^2 + xy - \frac{1}{2}x^2, \\ f_2(x, y) = f_3(x, y) &= -\frac{1}{2}y^2 - (1 + \frac{1}{a} + \frac{1}{b})xy - \frac{1}{2}x^2, \end{aligned}$$

where $a = 2^{\frac{-1}{2\alpha - 1}}$ and $b = 2^{\frac{-1}{2\beta - 1}}$.

Notice that the only stationary point of $f(x, y) = (f_1(x, y) + f_2(x, y) + f_3(x, y))/3$ is $(0, 0)$. We denote $g_{i,k}^x = \nabla_x f_i(x_k, y_k)$ and $g_{i,k}^y = \nabla_y f_i(x_k, y_k)$.

Now we consider points initialized in line

$$y = -\frac{1 + a}{a + \frac{a}{b}}x, \tag{72}$$

where we have

$$\begin{aligned} g_{1,0}^x &= y_0 - x_0 = -\frac{2ab + a + b}{ab + a}x_0 \\ g_{2,0}^x &= g_{3,0}^x = -\left(1 + \frac{1}{b} + \frac{1}{a}\right)y_0 - x_0 = \frac{2ab + a + b}{a^2(b + 1)}x_0 \\ g_{1,0}^y &= x_0 - y_0 = \frac{2ab + a + b}{ab + a}x_0 \\ g_{2,0}^y &= g_{2,0}^y = -\frac{2ab + a + b}{ab(b + 1)}x_0. \end{aligned}$$

Note that by our assumptions of the range of α and β, we have $a < b$. Thus, we have

$$|g_{1,0}^x| = |g_{1,0}^y| \quad \text{and} \quad |g_{2,0}^x| > |g_{2,0}^y|,$$

which means $g^x_{2,0}$ would be chosen in the maximum operator in the denominator of TiAda stepsize for x. Therefore, after one step, we have

$$
\begin{aligned}
x_1 &= x_0 - \eta^x \underbrace{\left(\frac{g^x_{1,0}}{\left(|g^x_{1,0}|^2\right)^\alpha} + \frac{g^x_{2,0}}{\left(|g^x_{2,0}|^2\right)^\alpha} + \frac{g^x_{3,0}}{\left(|g^x_{3,0}|^2\right)^\alpha} \right)}_{=0} \\
y_1 &= y_0 - \eta^y \underbrace{\left(\frac{g^y_{1,0}}{\left(|g^y_{1,0}|^2\right)^\beta} + \frac{g^y_{2,0}}{\left(|g^y_{2,0}|^2\right)^\beta} + \frac{g^y_{3,0}}{\left(|g^y_{3,0}|^2\right)^\beta} \right)}_{=0}.
\end{aligned}
$$

Next, we will use induction to show that x and y will stay in x_0 and y_0 for any iteration. Assuming for all iterations k in $1, \ldots, t$, $x_k = x_0$ and $y_k = y_0$, then we have in next step

$$
x_{t+1} = x_t - \eta^x \left(\frac{g^x_{1,0}}{\left(t \cdot |g^x_{1,0}|^2\right)^\alpha} + \frac{g^x_{2,0}}{\left(t \cdot |g^x_{2,0}|^2\right)^\alpha} + \frac{g^x_{3,0}}{\left(t \cdot |g^x_{3,0}|^2\right)^\alpha} \right).
$$

Note that $g^x_{1,0} = -a \cdot g^x_{2,0}$. Then, we get

$$
\begin{aligned}
x_{t+1} &= x_t - \eta^x \left(\frac{-p \cdot g^x_{2,0}}{t^\alpha \cdot a^{2\alpha} \cdot |g^x_{2,0}|^{2\alpha}} + \frac{2 g^x_{2,0}}{t^\alpha \cdot |g^x_{2,0}|^{2\alpha}} \right) \\
&= x_t - \frac{g^x_{2,0}}{t^\alpha \cdot |g^x_{2,0}|^{2\alpha}} \underbrace{\left(2 - a^{1-2\alpha}\right)}_{=0 \text{ (by definition of } a)} \\
&= x_t.
\end{aligned}
$$

Similarly, we can show that $y_{t+1} = y_t$. Therefore all iterates will stay at (x_0, y_0) if initialized at line $y = -\frac{ab+b}{ab+a} x$, which implies that the initial gradient norm can be arbitrarily large by picking x_0 to be large. □

B.4 Proof of Theorem 2 and Corollary 1

Proof of Theorem 2. Combining the results obtained in Lemma 6, 7 and 8, we get

$$
\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}\left[f\left(\bar{x}_k, y^*\left(\bar{x}_k\right)\right) - f\left(\bar{x}_k, \bar{y}_k\right)\right] \\
&= \sum_{k=0}^{k_0-1} \mathbb{E}\left[f\left(\bar{x}_k, y^*\left(\bar{x}_k\right)\right) - f\left(\bar{x}_k, \bar{y}_k\right)\right] + \sum_{k=k_0}^{K-1} \mathbb{E}\left[f\left(\bar{x}_k, y^*\left(\bar{x}_k\right)\right) - f\left(\bar{x}_k, \bar{y}_k\right)\right] \\
&\leqslant \frac{1}{2\gamma_y \bar{u}_1^{-\beta} n} \mathbb{E}\left[\left\|\mathbf{y}_0 - \mathbf{1} y^*\left(\bar{x}_0\right)\right\|^2\right] + \frac{2\left(4\beta C^2\right)^{2+\frac{1}{1-\beta}}}{\mu^{3+\frac{1}{1-\beta}} \gamma_y^{2+\frac{1}{1-\beta}} \bar{u}_1^{2-2\beta}} \\
&\quad + \frac{2\gamma_x^2 \kappa^2 \left(1+\zeta_v^2\right) G^{2\beta}}{n\mu\gamma_y^2} \sum_{k=0}^{k_0-1} \mathbb{E}\left[\bar{v}_{k+1}^{-2\alpha} \left\|\nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2\right] \\
&\quad + \frac{\gamma_y\left(1+\zeta_u^2\right)}{n} \sum_{k=0}^{K-1} \mathbb{E}\left[\bar{u}_{k+1}^{-\beta} \left\|\nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k\right)\right\|^2\right] + C \sum_{k=0}^{K-1} \mathbb{E}\left[\sqrt{\frac{1}{n}\left\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\right\|^2}\right] \\
&\quad + \frac{4}{\mu} \sum_{k=0}^{K-1} \mathbb{E}\left[\left\|\frac{\tilde{\boldsymbol{u}}_{k+1}^{-\beta}}{n\bar{u}_{k+1}^{-\beta}} \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y\right)\right\|^2\right] + \frac{8\gamma_x^2\kappa^2\left(1+\zeta_v^2\right)}{\mu\gamma_y^2 G^{2\alpha-2\beta}} \sum_{k=k_0}^{K-1} \left\|\nabla_x f\left(\bar{x}_k, \bar{y}_k\right)\right\|^2 \\
&\quad + \left(\frac{8\gamma_x^2\kappa^2 L^2\left(1+\zeta_v^2\right)}{n\mu\gamma_y^2 G^{2\alpha-2\beta}} + \frac{4\kappa L}{n}\right) \sum_{k=0}^{K-1} \mathbb{E}\left[\varDelta_k\right] + \frac{4\gamma_x\kappa\left(1+\zeta_v\right)C^2}{\mu\gamma_y\bar{v}_1^\alpha} \mathbb{E}\left[\bar{u}_K^\beta\right] \\
&\quad + \frac{\gamma_x^2\left(1+\zeta_v^2\right)}{\gamma_y\bar{v}_1^{\alpha-\beta}} \left(\kappa^2 + \frac{2\gamma_x^2\left(1+\zeta_v^2\right)C^2\hat{L}^2}{\mu\gamma_y\bar{v}_1^{2\alpha-\beta}}\right) \sum_{k=k_0}^{K-1} \mathbb{E}\left[\frac{\bar{v}_{k+1}^{-\alpha}}{n} \left\|\nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x\right)\right\|^2\right].
\end{aligned}
\tag{73}
$$

Letting the separation point between the two phases discussed in Lemma 6 and 7 satisfy

$$G=\left(\frac{16\left(1+\zeta_v^2\right)\gamma_x^2\kappa^4}{\gamma_y^2}\right)^{\frac{1}{2\alpha-2\beta}}, \tag{74}$$

then, plugging above inequality into (18), with the help of Lemma 4-8 and Lemma 9, we get

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi\left(\bar{x}_k\right)\right\|^2\right]\\
&\leqslant E_0+E_G+E_W+\frac{8C^{2\alpha}\left(\Phi^{\max}-\Phi^*\right)}{\gamma_x K^{1-\alpha}}\\
&+\frac{32\gamma_x\kappa^3\left(1+\zeta_v\right)C^{2+2\beta}}{\gamma_y\bar{v}_1^{\alpha}K^{1-\beta}}+\frac{8\kappa L\gamma_y\left(1+\zeta_u^2\right)C^{2-2\beta}}{\left(1-\beta\right)K^{\beta}}\\
&+\left(\gamma_x L_{\Phi}+\frac{\kappa^3 L\gamma_x^2}{\gamma_y\bar{v}_1^{\alpha-\beta}}+\frac{2\gamma_x^4\kappa^2\left(1+\zeta_v^2\right)C^2\hat{L}^2}{\gamma_y^2\bar{v}_1^{3\alpha-2\beta}}\right)\frac{8\left(1+\zeta_v^2\right)C^{2-2\alpha}}{\left(1-\alpha\right)K^{\alpha}}\\
&+\sqrt{\frac{1}{n^{1-\alpha}}\left(\frac{4\rho_W}{\left(1-\rho_W\right)^2}\right)^{\alpha}}\frac{16\left(1+\zeta_v\right)\zeta_v C^{2-\alpha}}{\left(1-\alpha\right)K^{\alpha}}\\
&+\sqrt{\frac{1}{n^{1-\beta}}\left(\frac{4\rho_W}{\left(1-\rho_W\right)^2}\right)^{\beta}}\frac{32\kappa^2\left(1+\zeta_u\right)\zeta_u C^{2-\beta}}{\left(1-\beta\right)K^{\beta}}\\
&+8\kappa LC\sqrt{\frac{8\rho_W\gamma_y^2\left(1+\zeta_u^2\right)}{\left(1-\rho_W\right)^2}\left(\frac{C^{2-4\beta}}{\left(1-2\beta\right)K^{2\beta}}\mathbb{I}_{\beta<1/2}+\frac{1+\log u_K-\log v_1}{K\bar{u}_1^{2\beta-1}}\mathbb{I}_{\beta\geqslant1/2}\right)},
\end{aligned} \tag{75}$$

where $\hat{L}=\kappa\left(1+\kappa\right)^2, L_{\Phi}=L\left(1+\kappa\right)$, and

$$\begin{aligned}
E_0&:=\frac{4\kappa L}{K\gamma_y\bar{u}_1^{-\beta}n}\mathbb{E}\left[\left\|\mathbf{y}_0-\mathbf{1}y^*\left(\bar{x}_0\right)\right\|^2\right]+\frac{16\kappa^2\left(4\beta C^2\right)^{2+\frac{1}{1-\beta}}}{K\mu^{2+\frac{1}{1-\beta}}\gamma_y^{2+\frac{1}{1-\beta}}\bar{u}_1^{2-2\beta}},\\
E_G&:=\frac{16\gamma_x^2\kappa^4\left(1+\zeta_v^2\right)G^{2\beta}}{\gamma_y^2}\left(\frac{C^{2-4\alpha}}{\left(1-2\alpha\right)K^{2\alpha}}\mathbb{I}_{\alpha<1/2}+\frac{1+\log v_K-\log v_1}{K\bar{v}_1^{2\alpha-1}}\mathbb{I}_{\alpha\geqslant1/2}\right),\\
E_W&:=\frac{32\left(8\kappa L+3L^2\right)\rho_W\gamma_x^2\left(1+\zeta_v^2\right)}{\left(1-\rho_W\right)^2}\left(\frac{C^{2-4\alpha}}{\left(1-2\alpha\right)K^{2\alpha}}\mathbb{I}_{\alpha<1/2}+\frac{1+\log v_K-\log v_1}{K\bar{v}_1^{2\alpha-1}}\mathbb{I}_{\alpha\geqslant1/2}\right)\\
&\quad+\frac{32\left(8\kappa L+3L^2\right)\rho_W\gamma_y^2\left(1+\zeta_u^2\right)}{\left(1-\rho_W\right)^2}\left(\frac{C^{2-4\beta}}{\left(1-2\beta\right)K^{2\beta}}\mathbb{I}_{\beta<1/2}+\frac{1+\log u_K-\log v_1}{K\bar{u}_1^{2\beta-1}}\mathbb{I}_{\beta\geqslant1/2}\right).
\end{aligned}$$

Letting the total iteration K satisfy the conditions given in (12) such that the terms E_0, E_G and E_W are dominated by the others, we thus complete the proof. □

Proof of Corollary 1. With the help of Lemma 10, we can directly adapt the proof of Theorem 2 to get the result in (14). □

B.5 Extend the proof to coordinate-wise stepsize

In this subsection, we show how to extend our convergence analysis of D-AdaST to the coordinate-wise adaptive stepsize (Zhou et al., 2018) variant. We first present this variant in Algorithm 2, which can be rewritten in a compact form with the Hadamard product denoted by ⊙.

Algorithm 2 **D-AdaST with coordinate-wise adaptive stepsize**

```
Initialization: $x_{i,0} \in \mathbb{R}^p$, $y_{i,0} \in \mathcal{Y}$, buffers $m^x_{i,0}, m^y_{i,0} > 0$, stepsizes $\gamma_x, \gamma_y > 0$ and $0 < \beta < \alpha < 1$.
for iteration $k = 0, 1, \cdots$, each node $i \in [n]$, do
    Sample i.i.d $\xi^x_{i,k}$ and $\xi^y_{i,k}$, compute:
        $g^x_{i,k} = \nabla_x f_i\left(x_{i,k}, y_{i,k}; \xi^x_{i,k}\right), \; g^y_{i,k} = \nabla_y f_i\left(x_{i,k}, y_{i,k}; \xi^y_{i,k}\right).$
    Accumulate the gradient with Hadamard product:
        $m^x_{i,k+1} = m^x_{i,k} + g^x_{i,k} \odot g^x_{i,k}, \; m^y_{i,k+1} = m^y_{i,k} + g^y_{i,k} \odot g^y_{i,k}$
    Compute the ratio:
        $\psi_{i,k+1} = \left\|m^x_{i,k+1}\right\|^{2\alpha} / \max\left\{\left\|m^x_{i,k+1}\right\|^{2\alpha}, \left\|m^y_{i,k+1}\right\|^{2\alpha}\right\} \leqslant 1.$
    Update primal and dual variables locally:
        $x_{i,k+1} = x_{i,k} - \gamma_x \psi_{i,k+1} \left(m^x_{i,k+1}\right)^{-\alpha} \odot g^x_{i,k},$
        $y_{i,k+1} = y_{i,k} + \gamma_y \left(m^y_{i,k+1}\right)^{-\beta} \odot g^y_{i,k}.$
    Communicate parameters with neighbors:
        $\left\{m^x_{i,k+1}, m^y_{i,k+1}, x_{i,k+1}, y_{i,k+1}\right\} \leftarrow \sum_{j \in \mathcal{N}_i} W_{i,j} \left\{m^x_{j,k+1}, m^y_{j,k+1}, x_{j,k+1}, y_{j,k+1}\right\}.$
    Projection of dual variable on to set $\mathcal{Y}$: $y_{i,k+1} \leftarrow \mathcal{P}_{\mathcal{Y}}\left(y_{i,k+1}\right)$.
end for
```

$$
\begin{aligned}
\mathbf{m}^x_{k+1} &= W\left(\mathbf{m}^x_k + \mathbf{h}^x_k\right), && \text{(76a)}\\
\mathbf{m}^y_{k+1} &= W\left(\mathbf{m}^y_k + \mathbf{h}^y_k\right), && \text{(76b)}\\
\mathbf{x}_{k+1} &= W\left(\mathbf{x}_k - \gamma_x V^{-\alpha}_{k+1} \odot \nabla_x F\left(\mathbf{x}_k, \mathbf{y}_k; \xi^x_k\right)\right), && \text{(76c)}\\
\mathbf{y}_{k+1} &= \mathcal{P}_{\mathcal{Y}}\left(W\left(\mathbf{y}_k + \gamma_y U^{-\beta}_{k+1} \odot \nabla_y F\left(\mathbf{x}_k, \mathbf{y}_k; \xi^y_k\right)\right)\right), && \text{(76d)}
\end{aligned}
$$

where

$$
\boldsymbol{h}^x_k = \left[\cdots, g^x_{i,k} \odot g^x_{i,k}, \cdots\right]^T \in \mathbb{R}^{n \times p}, \; \boldsymbol{h}^y_k = \left[\cdots, g^y_{i,k} \odot g^y_{i,k}, \cdots\right]^T \in \mathbb{R}^{n \times d},
$$

and the matrices U^α_k and V^β_k are redefined as follows:

$$
\begin{aligned}
V^{-\alpha}_k &= \left[\cdots, v^{-\alpha}_{i,k}, \cdots\right]^T, \; \left[v_{i,k}\right]_j = \max\left\{\left[m^x_{i,k}\right]_j, \left[m^y_{i,k}\right]_j\right\}, \; j \in [p],\\
U^{-\beta}_k &= \left[\cdots, u^{-\beta}_{i,k}, \cdots\right]^T, \; \left[u_{i,k}\right]_j = \left[m^y_{i,k}\right]_j, \; j \in [d],
\end{aligned} \tag{77}
$$

where $[\cdot]_j$ denotes the j-th element of a vector.

Recalling the definitions of inconsistency of stepsize in (8), we give the following notations:

$$
\begin{aligned}
\tilde{V}_k &= V_k - \bar{v}_k \mathbf{1}\mathbf{1}^T_p, \; \bar{v}_k = \frac{1}{np} \sum_{i=1}^n \sum_j^p V_{ij}, \; \bar{v}_{i,k} = \frac{1}{p} \sum_j^p V_{ij}, \; \bar{v}_{j,k} = \frac{1}{n} \sum_{i=1}^n V_{ij},\\
\tilde{U}_k &= U_k - \bar{u}_k \mathbf{1}\mathbf{1}^T_p, \; \bar{u}_k = \frac{1}{nd} \sum_{i=1}^n \sum_j^d U_{ij}, \; \bar{u}_{i,k} = \frac{1}{d} \sum_j^d U_{ij}, \; \bar{u}_{j,k} = \frac{1}{n} \sum_{i=1}^n U_{ij},
\end{aligned} \tag{78}
$$

and

$$\zeta_V^2 = \sup_{k\geqslant 0}\left\{\frac{\left\|V_k^{-\alpha} - \bar{v}_k^{-\alpha}\mathbf{1}\mathbf{1}_p^T\right\|^2}{np\left(\bar{v}_k^{-\alpha}\right)^2}\right\},\ \hat{\zeta}_v^2 = \sup_{k\geqslant 0}\left\{\frac{\left\|V_k^{-\alpha} - \left(V_k\mathbf{J}_p\right)^{-\alpha}\right\|^2}{np\left(\bar{v}_k^{-\alpha}\right)^2}\right\},$$

$$\zeta_U^2 = \sup_{k\geqslant 0}\left\{\frac{\left\|U_k^{-\beta} - \bar{u}_k^{-\beta}\mathbf{1}\mathbf{1}_d^T\right\|^2}{nd\left(\bar{u}_k^{-\beta}\right)^2}\right\},\ \hat{\zeta}_u^2 = \sup_{k\geqslant 0}\left\{\frac{\left\|U_k^{-\beta} - \left(U_k\mathbf{J}_d\right)^{-\beta}\right\|^2}{nd\left(\bar{u}_k^{-\beta}\right)^2}\right\}.$$

Building upon the established definitions of coordinate-wise stepsize inconsistency, the subsequent lemma is presented to show the non-convergence of the inconsistency term compared to Lemma 9.

Lemma 11 (Inconsistency, coordinate-wise)**.** *Suppose Assumption 1-4 hold. For the proposed D-AdaST algorithm, we have*

$$\begin{aligned}&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}}\tilde{V}_{k+1}^{-\alpha}\odot\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\\&\leqslant 2\left(1+\zeta_v\right)\zeta_v\sqrt{\frac{1}{n^{1-\alpha}}\left(\frac{4C^2\rho_W}{\left(1-\rho_W\right)^2}\right)^\alpha}\frac{C^{2-2\alpha}}{\left(1-\alpha\right)K^\alpha} + 2np\hat{\zeta}_v^2 C^2\end{aligned}\tag{79}$$

and

$$\begin{aligned}&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n\bar{u}_{k+1}^{-\beta}}\tilde{U}_{k+1}^{-\beta}\odot\nabla_y F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^y\right)\right\|^2\right]\\&\leqslant 2\left(1+\zeta_u\right)\zeta_u\sqrt{\frac{1}{n^{1-\beta}}\left(\frac{4C^2\rho_W}{\left(1-\rho_W\right)^2}\right)^\beta}\frac{C^{2-2\beta}}{\left(1-\beta\right)K^\beta} + 2nd\hat{\zeta}_u^2 C^2.\end{aligned}\tag{80}$$

In contrast, for D-TiAda, we have

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}}\tilde{V}_{k+1}^{-\alpha}\odot\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right] &\leqslant p\zeta_V^2 C^2,\\ \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n\bar{u}_{k+1}^{-\beta}}\tilde{U}_{k+1}^{-\beta}\odot\nabla_y F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^y\right)\right\|^2\right] &\leqslant d\zeta_U^2 C^2.\end{aligned}\tag{81}$$

Proof. For the coordinate-wise adaptive stepsize, with the definitions of Frobenius norm and Hadamard product, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}}\tilde{V}_{k+1}^{-\alpha}\odot\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\\&=\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}}\left(V_{k+1}^{-\alpha} - \left(V_{k+1}\mathbf{J}\right)^{-\alpha} + \left(V_{k+1}\mathbf{J}\right)^{-\alpha} - \bar{v}_{k+1}^{-\alpha}\mathbf{1}\mathbf{1}_p^T\right)\odot\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\\&\leqslant 2\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}}\left(\left(V_{k+1}\mathbf{J}\right)^{-\alpha} - \bar{v}_{k+1}^{-\alpha}\mathbf{1}\mathbf{1}_p^T\right)\odot\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\\&+2\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}}\left(V_{k+1}^{-\alpha} - \left(V_{k+1}\mathbf{J}\right)^{-\alpha}\right)\odot\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right].\end{aligned}\tag{82}$$

For the first term on the RHS, according to the definitions given in (78), we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}}\left(\left(V_{k+1}\mathbf{J}\right)^{-\alpha}-\bar{v}_{k+1}^{-\alpha}\mathbf{1}\mathbf{1}_p^T\right)\odot\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\\&\leqslant\mathbb{E}\left[\frac{1}{n^2\bar{v}_{k+1}^{-2\alpha}}\sum_{i=1}^{n}\left(\bar{v}_{i,k+1}^{\alpha}-\bar{v}_{k+1}^{\alpha}\right)^2\left\|\nabla_x f_i\left(x_{i,k},y_{i,k};\xi_{i,k}^x\right)\right\|^2\right].\end{aligned}\tag{83}$$

Then, for the second part, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}}\left(V_{k+1}^{-\alpha}-\left(V_{k+1}\mathbf{J}\right)^{-\alpha}\right)\odot\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\\&\leqslant\frac{1}{n}\mathbb{E}\left[\left\|\frac{V_{k+1}^{-\alpha}-\left(V_{k+1}\mathbf{J}\right)^{-\alpha}}{\bar{v}_{k+1}^{-\alpha}}\right\|^2\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\\&\leqslant p\hat{\zeta}_v^2\mathbb{E}\left[\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right].\end{aligned}\tag{84}$$

where the term $\hat{\zeta}_v^2$ is not guaranteed to be convergent because the stepsizes between the different dimensions of each node are not consistent. Then, similar to the proof of Lemma 9, we can obtain the result presented in (79).

Next, noticing that for D-TiAda,

$$\mathbb{E}\left[\left\|\frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}}\tilde{V}_{k+1}^{-\alpha}\odot\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\leqslant\frac{1}{n}\mathbb{E}\left[\left\|\frac{\tilde{V}_{k+1}^{-\alpha}}{\bar{v}_{k+1}^{-\alpha}}\right\|^2\left\|\nabla_x F\left(\mathbf{x}_k,\mathbf{y}_k;\xi_k^x\right)\right\|^2\right]\leqslant p\zeta_V^2C^2,\tag{85}$$

and using Lemma 9, we complete the proof. □

Theorem 3. *Suppose Assumption 1-4 hold. Let* $0<\beta<\alpha<1$ *and the total iteration satisfy*

$$K=\Omega\left(\max\left\{\left(\frac{\gamma_x^2\kappa^4}{\gamma_y^2}\right)^{\frac{1}{\alpha-\beta}},\quad\left(\frac{1}{\left(1-\rho_W\right)^2}\right)^{\max\left\{\frac{1}{\alpha},\frac{1}{\beta}\right\}}\right\}\right).$$

to ensure time-scale separation and quasi-independence of network. For D-AdaST with coordinate-wise adaptive stepsize, we have

$$\begin{aligned}&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla\Phi\left(\bar{x}_k\right)\right\|^2\right]\\&=\tilde{\mathcal{O}}\left(\frac{1}{K^{1-\alpha}}+\frac{1}{\left(1-\rho_W\right)^{\alpha}K^{\alpha}}+\frac{1}{K^{1-\beta}}+\frac{1}{\left(1-\rho_W\right)K^{\beta}}\right)+\mathcal{O}\left(n\left(p\hat{\zeta}_v^2+\kappa^2d\hat{\zeta}_u^2\right)C^2\right).\end{aligned}\tag{86}$$

Proof. With the help of Lemma 11 and the obtained result (75) in the proof of Theorem 2, we can derive the convergence results for D-AdaST with coordinate-wise adaptive stepsize. □

Remark 6. *In Theorem 3, we show that the coordinate-wise variant of D-AdaST exhibits a steady-state error in its upper bound. This error depends on the number of nodes and the dimension of the problem, which stems from the stepsize inconsistency in each dimension of the local decision variables for each node (c.f., Line 3 of Algorithm 2).*

https://doi.org/10.52202/079017-0622

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions and scope of this work have been accurately discussed.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] ,

   Justification: We have carefully discussed the limitations of this work in terms of assumptions and main results.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-0622

Justification: We have provided a full set of assumptions and complete proof for the theoretical results. See Section 3 and Appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided detailed experimental settings and reproducibility information for the experiments of this work. See Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-0622

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code of this work is included in the supplementary.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provided detailed experimental settings in Appendix A.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Multiple runs with averaging are used to produce the experimental curves in this work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

https://doi.org/10.52202/079017-0622

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided sufficient information on the computer resources in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel are negative.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-0622

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The license/copyright information of the code and dataset in this paper is clear.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

https://doi.org/10.52202/079017-0622

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code of this paper is included in the supplementary.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: NA

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-0622