

DAS²C: A DISTRIBUTED ADAPTIVE MINIMAX METHOD WITH NEAR-OPTIMAL CONVERGENCE

Anonymous authors

Paper under double-blind review

ABSTRACT

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 DAS²C, a Distributed Adaptive method with time-scale Separated Stepsize Control for minimax optimization. The key strategy is to employ an adaptive stepsize control protocol involving the transmission of two extra (scalar) variables. This protocol ensures the consistency among stepsizes of nodes, eliminating the steady-state errors 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 $\tilde{O}(\epsilon^{-(4+\delta)})$ for any small $\delta > 0$, matching that of the centralized counterpart. To the best of our knowledge, DAS²C is the *first* distributed adaptive method guaranteeing exact convergence without requiring to know any problem-dependent parameters for nonconvex minimax problems.

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), just to name a few. These tasks typically involve a minimax structure as follows

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

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 tackling minimax problems is Stochastic 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 like 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 design of an adaptive minimax method capable of satisfying the time-scale separation requirement and being parameter-agnostic in fully distributed settings remains an open question.*

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

- We construct counterexamples showing that directly applying adaptive methods designed for centralized problems might 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 DAS²C, that incorporates an efficient stepsize control 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.
- Theoretically, we prove that DAS²C is able to achieve a near-optimal convergence rate of $\tilde{O}(\epsilon^{-(4+\delta)})$ with any small $\delta > 0$ to find an ϵ -stationary point for distributed NC-SC problems. For comparison, we also prove the existence of a constant steady-state error in both the lower and upper bounds when directly applying a centralized adaptive algorithm without the stepsize control mechanism. Moreover, we characterize the specific 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 DAS²C on a variety of tasks, including the robust neural network training and optimizing Wasserstein GANs. In all tasks, we show the superiority of DAS²C comparing to several vanilla distributed adaptive methods across various graphs, initial stepsizes and data distributions (see also additional experiments in Appendix A).

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 $\tilde{O}(\epsilon^{-6})$ sample complexity for NC-SC objective functions. Sharma et al. (2022) provide improved complexity result of $\tilde{O}(\epsilon^{-4})$ matching that of the lower bound (Li et al., 2021; Zhang et al., 2021) for both NC-SC and nonconvex-Polyak-Lojasiewicz (NC-PL) settings. Yang et al. (2022a) combine 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 SGDA (D-SGDA) and provide non-asymptotic convergence with fixed stepsizes. Tsaknakis et al. (2020) propose a double-loop D-SGDA algorithm with gradient tracking techniques (Pu and Nedić, 2021) and achieve $\tilde{O}(\epsilon^{-4})$ sample complexity. 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; Boj 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

incorporates an inner loop to effectively maximize the dual variable, yielding parameter-agnosticism and best-known sample complexity of $\tilde{O}(\epsilon^{-4})$. 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 and not requiring mini-batch. There has been few works dedicated to adaptive minimax optimization in federated learning settings. For instance, Huang (2022) introduce a federated adaptive algorithm that integrates the stepsize rule of Adam with full-clients 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 $\mathbb{E}[\cdot]$ the expectation of a stochastic variable, $\|\cdot\|$ the Frobenius norm, $\langle \cdot, \cdot \rangle$ the inner product of two vectors, \odot the Schur 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 $\Phi(x) := f(x, y^*(x))$ as the primal function where $y^*(x) = \arg\max_{y \in \mathcal{Y}} f(x, y)$, and $\mathcal{P}_{\mathcal{Y}}(\cdot)$ 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 communication network. The overall objective of the agents is to solve the following finite-sum problem:

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

where $f_i : \mathbb{R}^{p+d} \rightarrow \mathbb{R}$ is the local private loss function accessible only by the associated node $i \in \mathcal{N} = \{1, 2, \dots, n\}$, $\mathcal{Y} \subset \mathbb{R}^d$ 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, $\mathcal{V} = \{1, 2, \dots, n\}$ 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) 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:

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

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

$$\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) := [\dots, \nabla_x F_i(x_{i,k}, y_{i,k}; \xi_{i,k}^x), \dots]^T \in \mathbb{R}^{n \times p},$$

$$\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y) := [\dots, \nabla_y F_i(x_{i,k}, y_{i,k}; \xi_{i,k}^y), \dots]^T \in \mathbb{R}^{n \times d},$$

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

In what follows, we will first explain the pitfalls of directly applying centralized adaptive algorithms, and then introduce our newly proposed solution to address the challenge.

2.1 NON-CONVERGENCE OF NAIVE DISTRIBUTED ADAPTIVE METHODS

In centralized settings, designing parameter-agnostic adaptive methods for nonconvex-strongly-concave minimax problems is already challenging and demands careful considerations. In fact, simply employing adaptive methods such as AdaGrad and Adam can lead to convergence issues (Yang et al., 2022b). To the best of our knowledge, TiAda (Li et al., 2023) is the SOTA algorithm that achieves near-optimal rates with both parameter and noise adaptivity. Similar to extending SGD into distributed settings (Nedic and Ozdaglar, 2009), TiAda can be adapted for distributed scenarios, which we will refer to as D-TiAda with the following update rules:

$$\mathbf{x}_{k+1} = W(\mathbf{x}_k - \gamma_x V_{k+1}^{-\alpha} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)), \quad (2a)$$

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

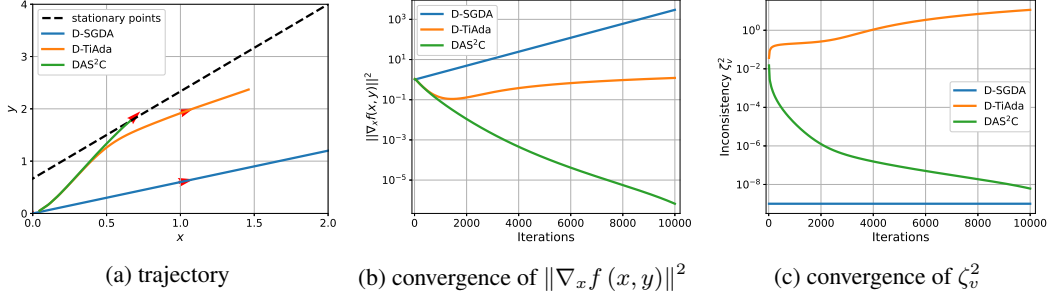


Figure 1: Comparison among D-SGDA, D-TiAda and DAS²C for NC-SC quadratic objective function (5) with $n = 2$ nodes and $\gamma_x = \gamma_y$. In (a), it shows the trajectories of primal and dual variables of the algorithms, the points on the black dash line are stationary points of f . In (b), it shows the convergence of $\|\nabla_x f(x_k, y_k)\|^2$ over the iterations. In (c), it shows the convergence of the inconsistency of stepsizes, ζ_v^2 defined in (7), over the iterations. Notably, ζ_v^2 fails to converge for D-TiAda and $\zeta_v^2 = 0$ for non-adaptive D-SGDA.

where γ_x and γ_y are the stepsizes, W is a doubly-stochastic weight matrix induced by graph \mathcal{G} , and

$$\begin{aligned} V_{k+1}^{-\alpha} &= \text{diag} \left\{ v_{i,k+1}^{-\alpha} \right\}_{i=1}^n, \quad v_{i,k+1} = \max \left\{ m_{i,k+1}^x, m_{i,k+1}^y \right\}, \\ U_{k+1}^{-\beta} &= \text{diag} \left\{ u_{i,k+1}^{-\beta} \right\}_{i=1}^n, \quad u_{i,k+1} = m_{i,k+1}^y, \end{aligned} \quad (3)$$

where $m_{i,k+1}^x = m_{i,k}^x + \left\| \nabla_x F_i(x_{i,k}, y_{i,k}; \xi_{i,k}^x) \right\|^2$, $m_{i,k+1}^y = m_{i,k}^y + \left\| \nabla_y F_i(x_{i,k}, y_{i,k}; \xi_{i,k}^y) \right\|^2$ are the locally computed gradient norm information. TiAda employs a maximum operator in the preconditioner for x , specifically in the definition of $v_{i,k}$, as well as different stepsize decay rates, i.e., $0 < \beta < \alpha < 1$, for the two variables. Such design allows automatic balancing the stepsizes of x and y and achieves the desired time-scale separation without requiring any knowledge of parameters.

However, in the distributed setting, such naive 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 vanilla algorithm (2) above in the sense of average system as below,

$$\begin{aligned} \bar{x}_{k+1} &= \underbrace{\bar{x}_k - \gamma_x \bar{v}_k^{-\alpha} \frac{\mathbf{1}^T}{n} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)}_{\text{adaptive descent}} - \underbrace{\gamma_x \frac{(\tilde{\mathbf{v}}_{k+1}^{-\alpha})^T}{n} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)}_{\text{inconsistency}}, \\ \bar{y}_{k+1} &= \mathcal{P}_y \left(\underbrace{\bar{y}_k + \gamma_y \bar{u}_k^{-\beta} \frac{\mathbf{1}^T}{n} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)}_{\text{adaptive ascent}} + \underbrace{\gamma_y \frac{(\tilde{\mathbf{u}}_k^{-\beta})^T}{n} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)}_{\text{inconsistency}} \right), \end{aligned} \quad (4)$$

where

$$\begin{aligned} \bar{x}_k &:= \frac{\mathbf{1}^T}{n} \mathbf{x}_k, \quad \bar{v}_k := \frac{1}{n} \sum_{i=1}^n v_{i,k}, \quad (\tilde{\mathbf{v}}_k^{-\alpha})^T := [\dots, v_{i,k}^{-\alpha} - \bar{v}_k^{-\alpha}, \dots], \\ \bar{y}_k &:= \frac{\mathbf{1}^T}{n} \mathbf{y}_k, \quad \bar{u}_k := \frac{1}{n} \sum_{i=1}^n u_{i,k}, \quad (\tilde{\mathbf{u}}_k^{-\beta})^T := [\dots, u_{i,k}^{-\beta} - \bar{u}_k^{-\beta}, \dots]. \end{aligned}$$

It is evident that, in comparison to centralized adaptive methods, an unexpected term on the right-hand side (RHS) arises due to inconsistencies, namely, $\tilde{\mathbf{v}}_k$ and $\tilde{\mathbf{u}}_k$. These terms introduce inaccuracies in the directions of gradient descent and ascent, degrading the optimization performance. The following theorem provides an explicit lower bound consisting a constant steady-state-error regarding the non-convergence of D-TiAda, whose proof can be found in Appendix B.4.

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 $0 < \beta < 0.5 < \alpha$ and $\gamma_x, \gamma_y > 0$, it holds*

that for any $t = 0, 1, 2, \dots$, we have

$$\|\nabla_x f(x_t, y_t)\| = \|\nabla_x f(x_0, y_0)\| \quad \text{and} \quad \|\nabla_y f(x_t, y_t)\| = \|\nabla_y f(x_0, y_0)\|,$$

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 control, 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 points along the line (c.f., Eq. (69)) can be positioned arbitrarily far away from this stationary point.

Apart from the counterexample discussed in Theorem 1, where the iterates are stationary, we also experimentally observe the non-convergence for moving iterates of D-TiAda, D-SGDA, and D-AdaGrad (naively applying AdaGrad for each node) even in a simpler scenario involving only two connected agents. This is illustrated in Figure 1 and the functions are as depicted as follows.

$$f_1(x, y) = -\frac{9}{20}y^2 + \frac{3}{5}y - x + xy - \frac{1}{2}x^2, \quad f_2(x, y) = -\frac{9}{20}y^2 + \frac{3}{5}y - x + 2xy - 2x^2. \quad (5)$$

It is not difficult to see that the points on the line $3y = 5x + 2$ are stationary points of $f(x, y) = 1/2(f_1(x, y) + f_2(x, y))$. 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 stepsizes inconsistency, as shown in Figure 1(c). In contrast, the utilization of the stepsize control protocol in DAS²C ensures convergence to a stationary point, with the inconsistency in stepsizes gradually diminishing. These two motivating examples effectively highlight the challenges associated with applying minimax algorithms to distributed scenarios.

2.2 DAS²C: A NEW ALGORITHM DESIGN WITH STEPSIZE CONTROL

To address the issue of inconsistent stepsizes across different nodes, we design the following distributed adaptive minimax optimization algorithm with stepsize control protocol, termed DAS²C, which allows us to asymptotically track the centralized adaptive stepsize 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

$$\mathbf{m}_{k+1}^x = W(\mathbf{m}_k^x + \mathbf{h}_k^x), \quad (6a)$$

$$\mathbf{m}_{k+1}^y = W(\mathbf{m}_k^y + \mathbf{h}_k^y), \quad (6b)$$

$$\mathbf{x}_{k+1} = W(\mathbf{x}_k - \gamma_x V_{k+1}^{-\alpha} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)), \quad (6c)$$

$$\mathbf{y}_{k+1} = \mathcal{P}_Y \left(W \left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y) \right) \right), \quad (6d)$$

where \mathbf{m}_k^x and \mathbf{m}_k^y denote the accumulation of historical gradients with

$$\mathbf{h}_k^x = [\dots, \|g_{i,k}^x\|^2, \dots]^T \in \mathbb{R}^n, \quad \mathbf{h}_k^y = [\dots, \|g_{i,k}^y\|^2, \dots]^T \in \mathbb{R}^n,$$

and V_k and U_k are diagonal matrices with $v_{i,k} = \max\{m_{i,k}^x, m_{i,k}^y\}$, $u_{i,k} = m_{i,k}^x$, where the maximization operator is used to achieve time-scale separation as suggested in TiAda (Li et al., 2023). Note that we also provide a variant of DAS²C with coordinate-wise adaptive stepsizes in Algorithm 2, along with its convergence analysis in Appendix B.6.

3 CONVERGENCE ANALYSIS

In this section, we introduce the main convergence results for the proposed DAS²C algorithm and compare it with D-TiAda to show the effectiveness of the proposed stepsize control protocol. To this end, we define the following metrics to evaluate the level of inconsistency of stepsizes among nodes, which are ensured to be bounded with Assumption 4.

$$\zeta_v^2 := \sup_{k>0} \left\{ \left(v_{i,k}^{-\alpha} - \bar{v}_k^{-\alpha} \right)^2 / \left(\bar{v}_k^{-\alpha} \right)^2 \right\}, \quad \zeta_u^2 := \sup_{k>0} \left\{ \left(u_{i,k}^{-\beta} - \bar{u}_k^{-\beta} \right)^2 / \left(\bar{u}_k^{-\beta} \right)^2 \right\}, \quad i \in [n]. \quad (7)$$

Algorithm 1 Distributed Adaptive Time-Scale Separated Stepsize Control Method (DAS²C)

Initialization: $x_{i,0} \in \mathbb{R}^p$, $y_{i,0} \in \mathcal{Y}$, buffers $m_{i,0}^x = m_{i,0}^y = c > 0$, stepsizes $\gamma_x, \gamma_y > 0$, exponential factors $0 < \beta < \alpha < 1$ and weight matrix W .

1: **for** iteration $k = 0, 1, \dots$, each node $i \in [n]$, **do**

2: Sample i.i.d. $g_{i,k}^x = \nabla_x F_i(x_{i,k}, y_{i,k}; \xi_{i,k}^x)$, $g_{i,k}^y = \nabla_y F_i(x_{i,k}, y_{i,k}; \xi_{i,k}^y)$.

3: Accumulate gradient norm: $m_{i,k+1}^x = m_{i,k}^x + \|g_{i,k}^x\|^2$, $m_{i,k+1}^y = m_{i,k}^y + \|g_{i,k}^y\|^2$.

4: Compute the ratio: $\psi_{i,k+1} = (m_{i,k+1}^x)^\alpha / \max\{(m_{i,k+1}^x)^\alpha, (m_{i,k+1}^y)^\alpha\} \leq 1$.

5: Update primal and dual variables locally:

$$x_{i,k+1} = x_{i,k} - \gamma_x \psi_{i,k+1} (m_{i,k+1}^x)^{-\alpha} g_{i,k}^x, \quad y_{i,k+1} = \mathcal{P}_{\mathcal{Y}}(y_{i,k} - \gamma_y (m_{i,k+1}^y)^{-\beta} g_{i,k}^y).$$

6: Communicate adaptive stepsizes and decision variables with neighbors:

$$\{m_{i,k+1}^x, m_{i,k+1}^y, x_{i,k+1}, y_{i,k+1}\} \leftarrow \sum_{j \in \mathcal{N}_i} W_{i,j} \{m_{j,k+1}^x, m_{j,k+1}^y, x_{j,k+1}, y_{j,k+1}\}.$$

7: **end for**

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).

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

$$f_i(x, y) - f_i(x, y') \geq \langle \nabla_y f_i(x, y), y - y' \rangle + \frac{\mu}{2} \|y - y'\|^2. \quad (8)$$

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

$$\|\nabla_z f_i(x, y) - \nabla_z f_i(x', y')\|^2 \leq L^2 (\|x - x'\|^2 + \|y - y'\|^2), \text{ for } z \in \{x, y\}. \quad (9)$$

Furthermore, f_i is second-order Lipschitz continuous for y , i.e.,

$$\|\nabla_{zy}^2 f_i(x, y) - \nabla_{zy}^2 f_i(x', y')\|^2 \leq L^2 (\|x - x'\|^2 + \|y - y'\|^2), \text{ for } z \in \{x, y\}. \quad (10)$$

Assumption 3 (Interior optimal point). *For all $x \in \mathbb{R}^p$, $y^*(x)$ is in the interior of \mathcal{Y} .*

Assumption 4 (Stochastic gradient). *For i.i.d. sample ξ_i , the stochastic gradient of each i is unbiased, i.e., $\forall x \in \mathbb{R}^p, y \in \mathcal{Y}$, $\mathbb{E}_{\xi_i}[\nabla_z F_i(x, y; \xi_i)] = \nabla_z f_i(x, y)$, for $z \in \{x, y\}$, and there exists a constant $C > 0$ such that $\|\nabla_z F_i(x, y; \xi_i)\| \leq C$.*

Remark 2. Assumption 1 does not require the convexity of primal variable x and the objective function thus can be nonconvex. Assumption 2 and 3 ensure that $y^*(\cdot)$ is smooth (c.f., Lemma 3), which is essential for achieving (near) optimal convergence rate (Chen et al., 2021; Li et al., 2023). Assumption 3 also ensures that $\nabla_y f(x, y^*(x)) = 0$. This is important for AdaGrad-based methods to maintain stepsizes near $y^*(x)$ without being excessively small which otherwise lead to slow convergence. Assumption 4 on bounded stochastic gradient is widely used for establishing convergence rates of adaptive methods (Zou et al., 2019; Kavis et al., 2022; Chen et al., 2023a), and it can be satisfied in many real-world tasks such as neural networks with rectified activation (Dinh et al., 2017) and GANs with projections on the critic (Gulrajani et al., 2017).

Now, we make the following assumption to ensure the connectivity of the graph. Note that the weight matrix is not required to be symmetric thus the graph can be direct, e.g., direct ring and exponential graphs (Ying et al., 2021), which is more general than (Lian et al., 2017; Borodich et al., 2021).

Assumption 5 (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 $\rho_W := \|W - \mathbf{J}\|_2^2 < 1$.*

3.2 MAIN RESULTS

We begin by demonstrating in the following lemma that the inconsistency terms, as described in (4), exhibit asymptotic convergence in the case of the proposed DAS²C algorithm. In contrast, these terms remain non-vanishing for D-TiAda (c.f., Lemma 11).

Lemma 1 (Convergence of inconsistency terms). *Suppose Assumption 1-5 hold. For the proposed DAS²C in Algorithm 1, we have*

$$\begin{aligned} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[\left\| \frac{(\tilde{\mathbf{v}}_{k+1}^{-\alpha})^T}{n\bar{v}_{k+1}^{-\alpha}} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] &\leq \sqrt{\frac{1}{n^{1-\alpha}} \left(\frac{4\rho_W}{(1-\rho_W)^2} \right)^\alpha} \frac{(1+\zeta_v)\zeta_v C^{2-\alpha}}{(1-\alpha)K^\alpha}, \\ \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[\left\| \frac{(\tilde{\mathbf{u}}_{k+1}^{-\alpha})^T}{n\bar{u}_{k+1}^{-\alpha}} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y) \right\|^2 \right] &\leq \sqrt{\frac{1}{n^{1-\beta}} \left(\frac{4\rho_W}{(1-\rho_W)^2} \right)^\beta} \frac{(1+\zeta_u)\zeta_u C^{2-\beta}}{(1-\beta)K^\beta}. \end{aligned} \quad (11)$$

The proof of Lemma 1 can be found Appendix B.3.

We are now ready to present the key convergence results in terms of the primal function $\Phi(x) := f(x, y^*(x))$ with $y^*(x) = \arg\max_{y \in \mathcal{Y}} f(x, y)$, whose proofs can be found in Appendix B.5.

Theorem 2. *Suppose Assumption 1-5 hold. Let $0 < \alpha < \beta < 1$ and the total iteration satisfy*

$$K = \Omega \left(\max \left\{ 1, \left(\gamma_x^2 \kappa^4 / \gamma_y^2 \right)^{1/(\alpha-\beta)}, \left(\rho_W / (1-\rho_W)^2 \right)^{\max\{1/\alpha, 1/\beta\}} \right\} \right), \quad (12)$$

where $\kappa := L/\mu$, to ensure time-scale separation and quasi-independence of network. For DAS²C,

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

Remark 3 (Near-optimal convergence). *Theorem 2 implies that if the total number of iterations satisfies the conditions (12), the proposed DAS²C algorithm converges to a stationary point exactly for Problem 1 with an $\tilde{O}(\epsilon^{-(4+\delta)})$ sample complexity for any small $\delta > 0$ with $\alpha = 0.5 + \delta/(8+2\delta)$ and $\beta = 0.5 - \delta/(8+2\delta)$. It is worth to note that this rate is near-optimal comparing to the best-known result $\tilde{O}(\epsilon^{-4})$ (Li et al., 2021; Yang et al., 2022b) for centralized minimax problems, and recovers the centralized TiAda algorithm (Li et al., 2023) as special case, i.e., letting $\rho_W = 0$.*

Remark 4 (Parameter-agnostic property and transit times). *The above results show that DAS²C is parameter-agnostic without requiring to know any problem-dependent parameters. Furthermore, we characterize the specific transient times (c.f., (12)) that ensure time-scale separation and quasi-independence of network in the sense of $\alpha, \beta < 1$, 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 TiAda. If α or β approaches 0, the transition time for achieving quasi-independence of the network will also increase. These observations highlight the importance of trade-offs between the convergence rate and the required duration of the transition phase.*

For 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 control, the inconsistencies between local stepsizes prevent D-TiAda to converge in the distributed setting.

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

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

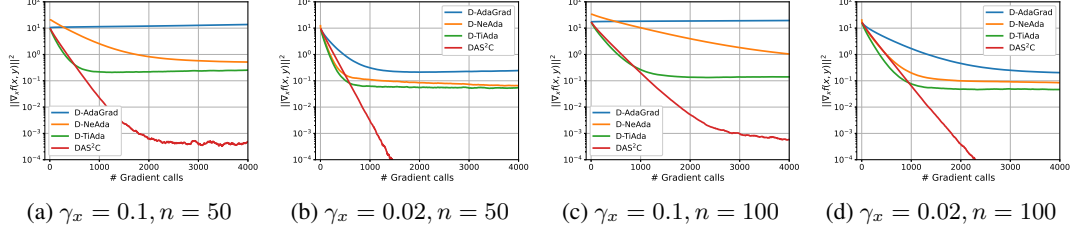


Figure 2: Performance comparison of algorithms on quadratic functions over exponential graphs with node counts $n = \{50, 100\}$ and different initial stepsizes ($\gamma_y = 0.1$).

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 DAS²C 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 communication topologies, 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). More experimental details and additional experiments under other settings 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 $n = 100$ ($\rho_w = 0.75$).

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

where $L_i \sim \mathcal{U}(1.5, 2.5)$. The local gradient of each node is computed with an additive $\mathcal{N}(0, 0.1)$ Gaussian noise. For both D-TiAda and DAS²C, we set the parameters as follows: $\alpha = 0.6$ and $\beta = 0.4$. It follows from Figure 2 (a) and 2 (b) that the proposed DAS²C 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 DAS²C.

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 $\min_x \max_y \frac{1}{n} \sum_{j=1}^n f_i(x; \xi_i + y) - \eta \|y\|^2$, where x denotes the parameters of the model, y denotes the perturbation and ξ_i denotes the data sample of node i . If η is large enough, the problem is NC-SC. We conduct experiments on 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 DAS²C 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 DAS²C 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 fluctuate more in the later stages of optimization as the gradient norm vanishes, leading to the inevitable increase of the Adam stepsize as the optimization process approaches convergence (Kingma and Ba, 2014). In plots (d) and (h), we further demonstrate the efficient scalability of DAS²C 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 larger-scale distributed scenarios effectively.

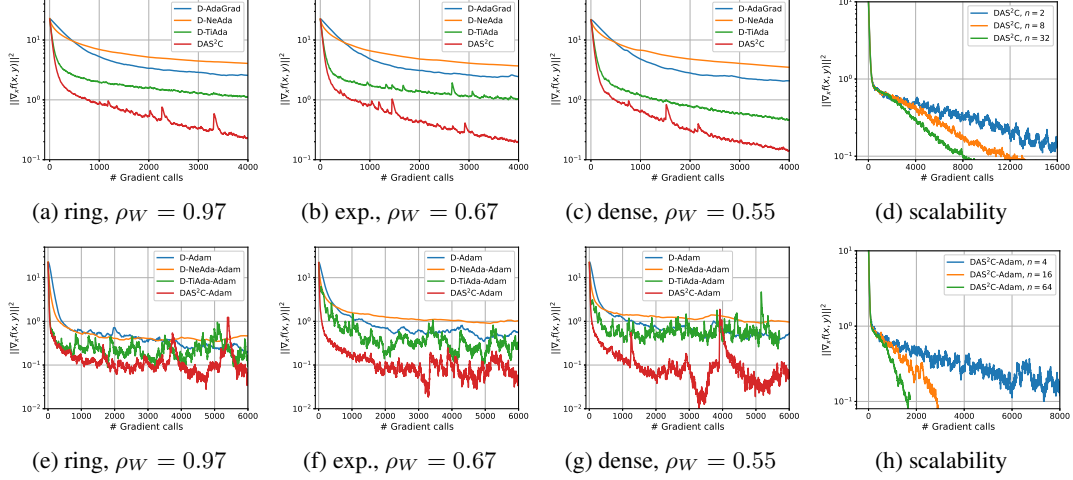


Figure 3: Comparison of the algorithms on training robust CNN on MNIST dataset. The first shows the results of AdaGrad-like stepsize, and the second row is for Adam-like stepsize. For the first three columns, we compare the algorithms on *different graphs* with $n = 20$. For the last column we show the scalability of DAS²C 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.

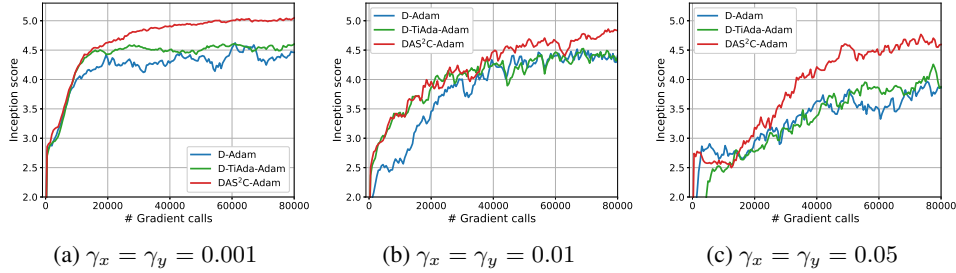


Figure 4: Training GANs on CIFAR-10 dataset over exponential graphs with $n = 10$ nodes using *different initial stepsizes*.

Generative Adversarial Networks. We further illustrate the effectiveness of DAS²C 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 decentralized setting where each discriminator is 1-Lipschitz and has access to only two classes of samples. We compare the inception score of DAS²C with D-Adam and D-TiAda adopting Adam-like stepsizes in Figure 4. It can be observed from the figure that DAS²C 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 DAS²C in solving real-world problems.

5 CONCLUSION

We introduced a new distributed adaptive minimax method, DAS²C, designed to tackle the issue of non-convergence in nonconvex-strongly-concave minimax problems caused by locally computed adaptive stepsize inconsistencies. 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 control protocol that guarantees stepsize consistency among nodes, effectively eliminating steady-state errors. Theoretically, we showed that DAS²C can achieve a near-optimal convergence rate of $\tilde{O}(\epsilon^{-(4+\delta)})$ with any small $\delta > 0$. Extensive experiments on real-world datasets have been conducted to validate our theoretical findings across various scenarios.

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, 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. (2022). Adaptive federated minimax optimization with lower complexities. *arXiv preprint arXiv:2211.07303*.
- 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. *arXiv preprint arXiv:2204.02833*.
- 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.
- 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*.
- 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.
- Yang, H., Liu, Z., Zhang, X., and Liu, J. (2022a). Sagda: Achieving $\mathcal{O}(\varepsilon^{-2})$ 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-ascent 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. (2021). The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR.
- 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.

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. We adapt code from (Yang et al., 2022b; Li et al., 2023) to decentralized settings. We use $\alpha = 0.6$ and $\beta = 0.4$ for all tasks.

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. 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 $2^0, 2^1 \dots 2^{\log(n)}$. Exponential graphs achieve a good balance between the degree and connectivity of the graph. Dense graph is a indirect graph where each node is connected to nodes at distances of $1, 2, 4, \dots, n$. 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 10 class images. 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 implement 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 we use for discriminator is a four layer CNN, and for generator is a four layer CNN 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.

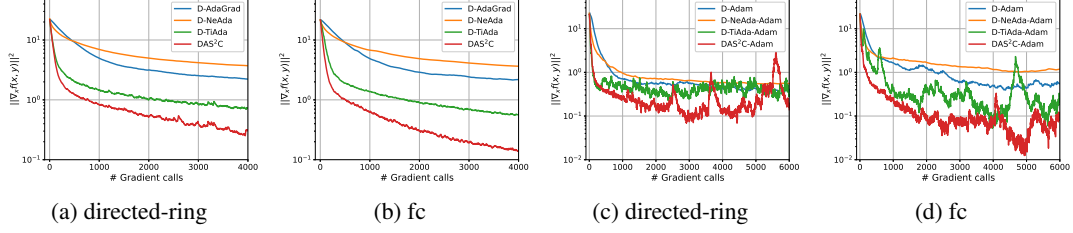
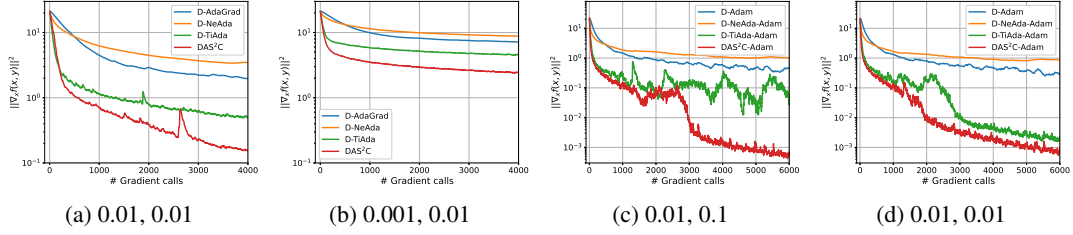
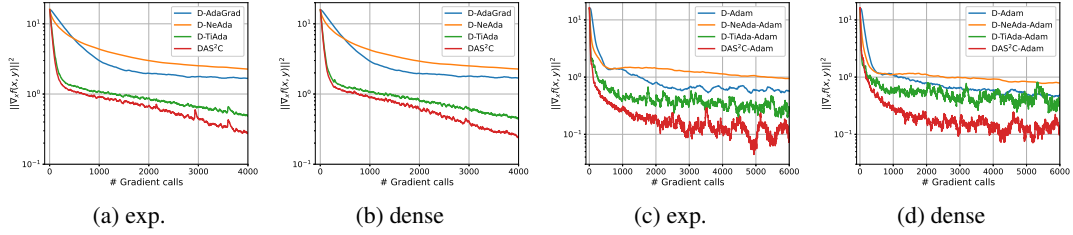
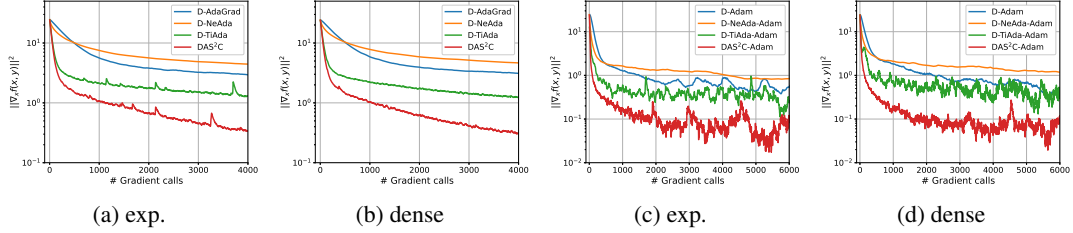
A.2 ADDITIONAL EXPERIMENTS ON ROBUST TRAINING OF NEURAL NETWORK.

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 DAS²C with D-AdaGrad, D-TiAda and D-NeAda using adaptive stepsizes in 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 use more initial stepsizes settings for x and y respectively in Figure 6. Further, we consider another data distribution where each node has data from 4 of the 10 classes in Figure 7. Finally we perform a comparison experiment with 40 nodes. Under all settings, the proposed DAS²C outperforms the others, demonstrating the superiority of DAS²C.

B PROOF OF THE MAIN RESULTS

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

- $S_1 := \frac{1}{nK} \sum_{k=0}^{K-1} \mathbb{E} \left[\bar{v}_{k+1}^{-\alpha} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2 \right]$: The asymptotically decaying terms by adopting adaptive stepsize;

Figure 5: Train CNN on MNIST with $n = 20$ nodes over *directed ring and fully connected graphs*.Figure 6: Train CNN on MNIST with $n = 20$ nodes with *different initial stepsizes γ_x and γ_y* .Figure 7: Train CNN on MNIST with $n = 20$ nodes over *exponential and dense graphs where each node has 4 sample classes*.Figure 8: Train CNN on MNIST with $n = 40$ nodes over *exponential and dense graphs*.

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

Next, we prove that these terms are convergent in Lemma 6-10 and Lemma 1 respectively. Finally, these results are integrated into the decent lemma thus completing the proof. We note that the proof is not trivial in the sense that these terms are coupled and therefore need to be carefully analyzed. This proof can also be adapted to analyze the coordinate-wise adaptive stepsize variant of DAS²C as explained in Appendix B.6, which is of independent interest.

B.1 SUPPORTING LEMMAS

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

Lemma 2 (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,*

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

When $\alpha = 0$, we have

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

Lemma 3. *Under Assumption 1, 2 and 3. Define $\Phi(x) := f(x, y^*(x))$ as the envelope function and $y^*(x) = \operatorname{argmax}_{y \in \mathcal{Y}} f(x, y)$. Then, we have,*

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

Lemma 4. *Let $A, B \in \mathbb{R}^{n \times p}$ be matrices with the same dimension. By the definitions of Frobenius norm and Schur product, we have*

- 1) $\|A \odot B\|^2 \leq \|A\|^2 \|B\|^2$;
- 2) $\left\| \frac{1^T}{n} A \odot B \right\|^2 \leq \frac{1}{n} \|A\|^2 \|B\|^2$;
- 3) For a vector $\mathbf{a} \in \mathbb{R}^n$, $\|\mathbf{a} \mathbf{1}_p^T \odot B\|^2 = \|\operatorname{diag}(\mathbf{a}) B\|^2 \leq \|\mathbf{a}\|^2 \|B\|^2$.

B.2 KEY LEMMAS

In this subsection, we give the key lemmas to help the analysis of the main results. For simplicity, we define $\Delta_k := \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 + \|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2$ as the consensus error for primal and dual variables, and $\tilde{V}_{k+1}^{-\alpha} = V_{k+1}^{-\alpha} - \bar{v}_{k+1}^{-\alpha} \mathbf{1}\mathbf{1}_p^T$ the difference matrix for stepsize. Then, we have the following lemmas.

Lemma 5 (Decent lemma). *Suppose Assumption 1-5 hold. we have*

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

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

Proof. By the smoothness of Φ given in Lemma 3, i.e.,

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

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

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

Then, we bound the inner-product terms on the RHS. Firstly,

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

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

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

Then, for the last term on the RHS of (19), recalling the definition of stepsize inconsistency in (7), we have

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

Plugging the obtained inequalities into (19) and telescoping the terms, we get

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

Now it remains to bound the first term on the RHS of the above inequality. We have

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

wherein the last inequality we have used Assumption 4, and we thus complete the proof. \square

Next, we try to bound the last four terms S_1 - S_4 in (19) respectively. For S_1 , we have the asymptotic convergence for both primal and dual variables in the following lemma.

Lemma 6. *Suppose Assumption 1-5 hold. We have*

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

and

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

Proof. With the help of Lemma 2 and Assumption 4, taking the primal variable x as an example, noticing that $v_{i,0} > 0$, we have

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

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

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

Lemma 7. *Suppose Assumption 1-5 hold. We have*

$$\begin{aligned}
& \frac{1}{K} \sum_{k=0}^K \mathbb{E} [\Delta_k] \leq \frac{2\mathbb{E} [\Delta_0]}{(1-\rho_W) K} \\
&+ \frac{4n\rho_W\gamma_x^2(1+\zeta_v^2)}{(1-\rho_W)^2} \left(\frac{C^{2-4\alpha}}{(1-2\alpha) K^{2\alpha}} \mathbb{I}_{\alpha < 1/2} + \frac{1+\log v_K - \log v_1}{K \bar{v}_1^{2\alpha-1}} \mathbb{I}_{\alpha \geq 1/2} \right) \\
&+ \frac{4n\rho_W\gamma_y^2(1+\zeta_u^2)}{(1-\rho_W)^2} \left(\frac{C^{2-4\beta}}{(1-2\beta) K^{2\beta}} \mathbb{I}_{\beta < 1/2} + \frac{1+\log u_K - \log u_1}{K \bar{u}_1^{2\beta-1}} \mathbb{I}_{\beta \geq 1/2} \right), \tag{27}
\end{aligned}$$

where $\mathbb{I}_{[\cdot]} \in \{0, 1\}$ is the indicator for specific condition, and the initial consensus error Δ_0 can be set to 0 with proper initialization.

Proof. Firstly, for the primal variables, we have

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

Then, by the definition of ζ_v in (7), we have

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

and thus

$$\begin{aligned}
& \sum_{k=0}^{K-1} \mathbb{E} \left[\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 \right] \\
&\leq \frac{1+\rho_W}{2} \sum_{k=0}^{K-1} \mathbb{E} \left[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 \right] \\
&+ \frac{4\gamma_x^2\rho_W(1+\zeta_v^2)}{1-\rho_W} \sum_{k=0}^{K-1} \mathbb{E} \left[\bar{v}_{k+1}^{-2\alpha} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2 \right].
\end{aligned} \tag{30}$$

Then, we bound the last term on the RHS of the above inequality by Lemma 6. For the case $\alpha < 1/2$, by Assumption 4 we have

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

for the case $\alpha \geq 1/2$, with the help of Lemma 2, we have

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

With the similar analysis for the dual variables, we complete the proof. \square

Finally, 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 value 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 $\bar{u}_k \leq G$ hold for $k = 0, 1, \dots, k_0 - 1$, then the analysis is divided into two phases.

Lemma 8 (First phase). *Suppose Assumption 1-5 hold. If $\bar{u}_k \leq G, k = 0, 1, \dots, k_0 - 1$, we have*

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

where

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

Proof. Firstly, we use Young's inequality with parameter λ_k ,

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

Then, for the first term on the RHS, noticing that $\mathcal{P}(\cdot)$ is a linear operator, we have

$$\begin{aligned}
& \|\bar{y}_{k+1} - y^*(\bar{x}_k)\|^2 \\
&= \left\| \mathcal{P}_y \left(\bar{y}_k + \gamma_y \frac{\mathbf{1}^T}{n} U_{k+1}^{-\beta} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y) \right) - y^*(\bar{x}_k) \right\|^2 \\
&\leq \|\bar{y}_k - y^*(\bar{x}_k)\|^2 + \gamma_y^2 \left\| \frac{\mathbf{1}^T}{n} U_{k+1}^{-\beta} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y) \right\|^2 \\
&+ 2\gamma_y \left\langle \frac{\mathbf{1}^T}{n} U_{k+1}^{-\beta} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y), \bar{y}_k - y^*(\bar{x}_k) \right\rangle \\
&\leq \|\bar{y}_k - y^*(\bar{x}_k)\|^2 + \frac{2\gamma_y^2 (1 + \zeta_u^2) \bar{u}_{k+1}^{-2\beta}}{n} \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2 \\
&+ 2\gamma_y \bar{u}_{k+1}^{-\beta} \left\langle \frac{\mathbf{1}^T}{n} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y), \bar{y}_k - y^*(\bar{x}_k) \right\rangle + 2\gamma_y \left\langle \frac{\tilde{\mathbf{u}}_{k+1}^{-\beta}}{n} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y), \bar{y}_k - y^*(\bar{x}_k) \right\rangle.
\end{aligned} \tag{36}$$

Then, multiplying $1/(\gamma_y \bar{u}_{k+1}^{-\beta})$ on both sides we get

$$\begin{aligned}
& \frac{\|\bar{y}_{k+1} - y^*(\bar{x}_{k+1})\|^2}{\gamma_y \bar{u}_{k+1}^{-\beta}} \\
&\leq \left(1 + \frac{1}{\lambda_k}\right) \frac{\|y^*(\bar{x}_{k+1}) - y^*(\bar{x}_k)\|^2}{\gamma_y \bar{u}_{k+1}^{-\beta}} \\
&+ (1 + \lambda_k) \left(\frac{\|\bar{y}_k - y^*(\bar{x}_k)\|^2}{\gamma_y \bar{u}_{k+1}^{-\beta}} + \frac{2\gamma_y \bar{u}_{k+1}^{-\beta} (1 + \zeta_u^2)}{n} \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2 \right) \\
&+ 2(1 + \lambda_k) \left\langle \frac{\mathbf{1}^T}{n} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y), \bar{y}_k - y^*(\bar{x}_k) \right\rangle \\
&+ 2(1 + \lambda_k) \left\langle \frac{\tilde{\mathbf{u}}_{k+1}^{-\beta}}{n \bar{u}_{k+1}^{-\beta}} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y), \bar{y}_k - y^*(\bar{x}_k) \right\rangle.
\end{aligned} \tag{37}$$

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

$$\begin{aligned}
& \mathbb{E} \left[2 \left\langle \frac{\mathbf{1}^T}{n} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y), \bar{y}_k - y^*(\bar{x}_k) \right\rangle \right] \\
&= 2\mathbb{E} \left[\left\langle \frac{\mathbf{1}^T}{n} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k) - \nabla_y f(\bar{x}_k, \bar{y}_k) + \nabla_y f(\bar{x}_k, \bar{y}_k), \bar{y}_k - y^*(\bar{x}_k) \right\rangle \right] \\
&\leq -2\mathbb{E} [f(\bar{x}_k, y^*(\bar{x}_k)) - f(\bar{x}_k, \bar{y}_k)] - \mu \mathbb{E} [\|\bar{y}_k - y^*(\bar{x}_k)\|^2] \\
&+ \frac{8L^2}{\mu n} \mathbb{E} [\Delta_k] + \frac{\mu}{8} \mathbb{E} [\|\bar{y}_k - y^*(\bar{x}_k)\|^2],
\end{aligned} \tag{38}$$

and

$$\begin{aligned}
& \mathbb{E} \left[2 \left\langle \frac{\tilde{\mathbf{u}}_{k+1}^{-\beta}}{n \bar{u}_{k+1}^{-\beta}} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y), \bar{y}_k - y^*(\bar{x}_k) \right\rangle \right] \\
&\leq \frac{8}{\mu} \mathbb{E} \left[\left\| \frac{\tilde{\mathbf{u}}_{k+1}^{-\beta}}{n \bar{u}_{k+1}^{-\beta}} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y) \right\|^2 \right] + \frac{\mu}{8} \mathbb{E} [\|\bar{y}_k - y^*(\bar{x}_k)\|^2].
\end{aligned} \tag{39}$$

Let $\lambda_k = \gamma_y \bar{u}_{k+1}^{-\beta} / 2$. Telescoping the terms from 0 to k_0 , we get

$$\begin{aligned}
& \sum_{k=0}^{k_0-1} \mathbb{E} [f(\bar{x}_k, y^*(\bar{x}_k)) - f(\bar{x}_k, \bar{y}_k)] \\
& \leq \sum_{k=0}^{k_0-1} \mathbb{E} \left[\frac{\|y^*(\bar{x}_{k+1}) - y^*(\bar{x}_k)\|^2}{\gamma_y^2 \bar{u}_{k+1}^{-2\beta}} \right] + \sum_{k=0}^{k_0-1} \mathbb{E} [S_{1,k}] + \frac{4L^2}{\mu n} \sum_{k=0}^{k_0-1} \mathbb{E} [\Delta_k] \\
& + \frac{\gamma_y (1 + \zeta_u^2)}{n} \sum_{k=0}^{k_0-1} \mathbb{E} \left[\bar{u}_{k+1}^{-\beta} \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2 \right] + \frac{4}{\mu} \sum_{k=0}^{k_0-1} \mathbb{E} \left[\left\| \frac{\tilde{\mathbf{u}}_k^{-\beta}}{n \bar{u}_{k+1}^{-\beta}} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y) \right\|^2 \right]. \tag{40}
\end{aligned}$$

By the κ -smoothness of y^* , we have

$$\begin{aligned}
& \|y^*(\bar{x}_{k+1}) - y^*(\bar{x}_k)\|^2 \\
& \leq \kappa^2 \|\bar{x}_{k+1} - \bar{x}_k\|^2 \\
& = \kappa^2 \left\| \gamma_x \bar{v}_{k+1}^{-\alpha} \frac{\mathbf{1}^T}{n} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k) - \gamma_x \frac{(\tilde{\mathbf{v}}_{k+1}^{-\alpha})^T}{n} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \\
& \leq \frac{2\gamma_x^2 \kappa^2 (1 + \zeta_v^2) \bar{v}_{k+1}^{-2\alpha}}{n} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2. \tag{41}
\end{aligned}$$

Noticing that $\bar{u}_k \leq G$ for $k \leq k_0 - 1$, we complete the proof. \square

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

Lemma 9 (Second phase). *Suppose Assumption 1-5 hold. If $\bar{u}_k \leq G, k = 0, 1, \dots, k_0 - 1$, we have*

$$\begin{aligned}
& \sum_{k=k_0}^{K-1} \mathbb{E} [f(\bar{x}_k, y^*(\bar{x}_k)) - f(\bar{x}_k, \bar{y}_k)] \\
& \leq \sum_{k=k_0}^{K-1} \mathbb{E} [S_{1,k}] + \frac{8\gamma_x^2 \kappa^2}{\mu \gamma_y^2 G^{2\alpha-2\beta}} \sum_{k=k_0}^{K-1} \mathbb{E} [\|\nabla_x f(\bar{x}_k, \bar{y}_k)\|^2] \\
& + \left(\frac{4L^2}{\mu n} + \frac{8\gamma_x^2 \kappa^2 L^2}{\mu n \gamma_y^2 G^{2\alpha-2\beta}} \right) \sum_{k=k_0}^{K-1} \mathbb{E} [\Delta_k] \\
& + \left(\frac{4}{\mu} + \frac{4\gamma_x^2 \kappa^2}{\mu \gamma_y^2 G^{2\alpha-2\beta}} \right) \sum_{k=k_0}^{K-1} \mathbb{E} \left[\left\| \frac{\tilde{\mathbf{u}}_{k+1}^{-\beta}}{n \bar{u}_{k+1}^{-\beta}} \right\|^2 \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2 \right] \\
& + \sum_{k=k_0}^{K-1} \mathbb{E} \left[\frac{\gamma_y (1 + \zeta_u^2) \bar{u}_{k+1}^{-\beta}}{n} \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2 \right] \\
& + \frac{\gamma_x^2 (1 + \zeta_v^2)}{\gamma_y \bar{v}_0^{\alpha-\beta}} \left(\kappa^2 + \frac{4\gamma_x^2 (1 + \zeta_v^2) C^2 \hat{L}^2}{\mu \gamma_y \bar{v}_0^{2\alpha-\beta}} \right) \sum_{k=k_0}^{K-1} \mathbb{E} \left[\frac{\bar{v}_k^{-\alpha}}{n} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2 \right] \\
& + \frac{8\gamma_x \kappa (1 + \zeta_v) C^2}{\bar{v}_0^\alpha \gamma_y} \left(\frac{1}{\mu} + \frac{\gamma_y (1 + \zeta_u)}{\bar{u}_0^\beta} \right) \mathbb{E} [\bar{u}_K^\beta]. \tag{42}
\end{aligned}$$

Proof. Firstly, we have

$$\begin{aligned}
& \|\bar{y}_{k+1} - y^*(\bar{x}_{k+1})\|^2 \\
&= \|\bar{y}_{k+1} - y^*(\bar{x}_k)\|^2 + \|y^*(\bar{x}_{k+1}) - y^*(\bar{x}_k)\|^2 \\
&\quad - 2 \langle \bar{y}_{k+1} - y^*(\bar{x}_k), y^*(\bar{x}_{k+1}) - y^*(\bar{x}_k) \rangle \\
&= \|\bar{y}_{k+1} - y^*(\bar{x}_k)\|^2 + \|y^*(\bar{x}_{k+1}) - y^*(\bar{x}_k)\|^2 \\
&\quad - 2 (\bar{y}_{k+1} - y^*(\bar{x}_k))^T \nabla y^*(\bar{x}_k) (\bar{x}_{k+1} - \bar{x}_k)^T \\
&\quad - 2 (\bar{y}_{k+1} - y^*(\bar{x}_k))^T \left(y^*(\bar{x}_{k+1}) - y^*(\bar{x}_k) - \nabla y^*(\bar{x}_k) (\bar{x}_{k+1} - \bar{x}_k)^T \right).
\end{aligned} \tag{43}$$

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

$$\begin{aligned}
& -2 (\bar{y}_{k+1} - y^*(\bar{x}_k))^T \nabla y^*(\bar{x}_k) (\bar{x}_{k+1} - \bar{x}_k)^T \\
&= 2\gamma_x (\bar{y}_{k+1} - y^*(\bar{x}_k))^T \nabla y^*(\bar{x}_k) (\nabla_x F(\mathbf{x}_k, \mathbf{y}_k))^T \left(\frac{\mathbf{1}_{\bar{v}_{k+1}}^{-\alpha}}{n} + \frac{\tilde{\mathbf{v}}_{k+1}^{-\alpha}}{n} \right) \\
&\quad + 2\gamma_x (\bar{y}_{k+1} - y^*(\bar{x}_k))^T \nabla y^*(\bar{x}_k) \left(\nabla_x \tilde{F}_k \right)^T \left(\frac{\mathbf{1}_{\bar{v}_{k+1}}^{-\alpha}}{n} + \frac{\tilde{\mathbf{v}}_{k+1}^{-\alpha}}{n} \right) \\
&\leq 2\gamma_x \kappa \|\bar{y}_{k+1} - y^*(\bar{x}_k)\| \left\| (\nabla_x F(\mathbf{x}_k, \mathbf{y}_k))^T \left(\frac{\mathbf{1}_{\bar{v}_{k+1}}^{-\alpha}}{n} + \frac{\tilde{\mathbf{v}}_{k+1}^{-\alpha}}{n} \right) \right\| \\
&\quad + 2\gamma_x (\bar{y}_{k+1} - y^*(\bar{x}_k))^T \nabla y^*(\bar{x}_k) \left(\nabla_x \tilde{F}_k \right)^T \left(\frac{\mathbf{1}_{\bar{v}_{k+1}}^{-\alpha}}{n} + \frac{\tilde{\mathbf{v}}_{k+1}^{-\alpha}}{n} \right).
\end{aligned} \tag{44}$$

Then, using Young's inequality with parameter λ_k , we get

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

For the second inner-product term on the RHS, noticing that y^* is $\hat{L} = \kappa(1 + \kappa)^2$ smooth given in Lemma 3, we have

$$\begin{aligned}
& 2 (\bar{y}_{k+1} - y^*(\bar{x}_k))^T \left(y^*(\bar{x}_k) - y^*(\bar{x}_{k+1}) + \nabla y^*(\bar{x}_k) (\bar{x}_{k+1} - \bar{x}_k)^T \right) \\
&\leq 2 \|\bar{y}_{k+1} - y^*(\bar{x}_k)\| \|y^*(\bar{x}_k) - y^*(\bar{x}_{k+1}) + \nabla y^*(\bar{x}_k) (\bar{x}_{k+1} - \bar{x}_k)^T\|^2 \\
&\leq 2 \|\bar{y}_{k+1} - y^*(\bar{x}_k)\| \frac{\hat{L}}{2} \|\bar{x}_{k+1} - \bar{x}_k\|^2 \\
&\leq \gamma_x^2 \hat{L} \|\bar{y}_{k+1} - y^*(\bar{x}_k)\| \left\| \left(\frac{\bar{v}_{k+1}^{-\alpha} \mathbf{1}^T}{n} + \frac{(\tilde{\mathbf{v}}_{k+1}^{-\alpha})^T}{n} \right) \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \\
&\leq \gamma_x^2 \hat{L} \|\bar{y}_{k+1} - y^*(\bar{x}_k)\| \frac{2\bar{v}_{k+1}^{-2\alpha} (1 + \zeta_v^2) C}{n} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\| \\
&\leq \tau \gamma_x^2 \bar{v}_{k+1}^{-2\alpha} (1 + \zeta_v^2) C^2 \hat{L} \|\bar{y}_{k+1} - y^*(\bar{x}_k)\|^2 + \frac{\gamma_x^2 \bar{v}_{k+1}^{-2\alpha} (1 + \zeta_v^2) \hat{L}}{\tau n} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2,
\end{aligned} \tag{46}$$

wherein the last inequality we have used Young's inequality with parameter τ . Plugging the obtained inequalities into (43), we get

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

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

$$\lambda_k = \frac{\mu\gamma_y \bar{u}_{k+1}^{-\beta}}{4}, \quad \tau = \frac{\mu\gamma_y \bar{v}_0^{2\alpha-\beta}}{4\gamma_x^2 (1 + \zeta_v^2) C^2 \hat{L}}, \tag{48}$$

we get

$$\begin{aligned}
& \|\bar{y}_{k+1} - y^*(\bar{x}_{k+1})\|^2 \\
& \leq \left(1 + \frac{\mu\gamma_y \bar{u}_{k+1}^{-\beta}}{2}\right) \|\bar{y}_{k+1} - y^*(\bar{x}_k)\|^2 \\
& + \frac{\gamma_x^2 (1 + \zeta_v^2)}{n} \left(2\kappa^2 + \frac{4\gamma_x^2 (1 + \zeta_v^2) C^2 \hat{L}^2}{\mu\gamma_y \bar{v}_0^{2\alpha-\beta}}\right) \bar{v}_{k+1}^{-2\alpha} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k)\|^2 \\
& + \frac{8\gamma_x^2 \bar{v}_{k+1}^{-2\alpha} \kappa^2}{\mu\gamma_y \bar{u}_{k+1}^{-\beta}} \left(\left\| \frac{\mathbf{1}^T}{n} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k) \right\|^2 + \left\| \frac{(\tilde{\mathbf{v}}_{k+1}^{-\alpha})^T}{n \bar{v}_{k+1}^{-\alpha}} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k) \right\|^2 \right) \\
& + 2\gamma_x (\bar{y}_{k+1} - y^*(\bar{x}_k))^T \nabla y^*(\bar{x}_k) \left(\nabla_x \tilde{F}_k \right)^T \left(\frac{\mathbf{1}_{k+1}^{-\alpha}}{n} + \frac{\tilde{\mathbf{v}}_{k+1}^{-\alpha}}{n} \right).
\end{aligned} \tag{49}$$

Recalling that

$$\begin{aligned}
& \mathbb{E} \left[\frac{1}{\gamma_y \bar{u}_{k+1}^{-\beta}} \|\bar{y}_{k+1} - y^*(\bar{x}_k)\|^2 \right] \\
& \leq \mathbb{E} \left[\left(\frac{1 - 3\mu\gamma_y \bar{u}_{k+1}^{-\beta}/4}{\gamma_y \bar{u}_{k+1}^{-\beta}} \right) \|\bar{y}_k - y^*(\bar{x}_k)\|^2 \right] + \mathbb{E} \left[\frac{2\gamma_y (1 + \zeta_u^2) \bar{u}_{k+1}^{-\beta}}{n} \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2 \right] \\
& - 2\mathbb{E} [f(\bar{x}_k, y^*(\bar{x}_k)) - f(\bar{x}_k, \bar{y}_k)] + \frac{8L^2}{\mu n} \mathbb{E} [\Delta_k] + \frac{8}{\mu} \mathbb{E} \left[\left\| \frac{\tilde{\mathbf{u}}_{k+1}^{-\beta}}{n \bar{u}_{k+1}^{-\beta}} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y) \right\|^2 \right].
\end{aligned}$$

Multiplying by $\frac{2}{(2+\mu\gamma_y\bar{u}_{k+1}^{-\beta})\gamma_y\bar{u}_{k+1}^{-\beta}}$ on both sides of (49), we obtain that

$$\begin{aligned}
& \mathbb{E}[f(\bar{x}_k, y^*(\bar{x}_k)) - f(\bar{x}_k, \bar{y}_k)] \\
& \leq \mathbb{E}[S_{1,k}] + \mathbb{E}\left[\frac{\gamma_y(1+\zeta_u^2)\bar{u}_{k+1}^{-\beta}}{n} \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2\right] \\
& + \frac{4}{\mu} \mathbb{E}\left[\left\|\frac{\tilde{\mathbf{u}}_{k+1}^{-\beta}}{n\bar{u}_{k+1}^{-\beta}} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\right\|^2\right] + \frac{4L^2}{\mu n} \mathbb{E}[\Delta_k] \\
& + \underbrace{\frac{4\gamma_x^2\bar{v}_{k+1}^{-2\alpha}\kappa^2}{\mu\gamma_y^2\bar{u}_{k+1}^{-2\beta}} \left(\left\|\frac{\mathbf{1}^T}{n} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k)\right\|^2 + \left\|\frac{(\tilde{\mathbf{v}}_{k+1}^{-\alpha})^T}{n\bar{v}_{k+1}^{-\alpha}} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k)\right\|^2\right)}_{S_{2,k}} \\
& + \underbrace{\frac{\gamma_x^2(1+\zeta_v^2)}{n\gamma_y} \left(\kappa^2 + \frac{2\gamma_x^2(1+\zeta_v^2)C^2\hat{L}^2}{\mu\gamma_y\bar{v}_0^{2\alpha-\beta}}\right) \mathbb{E}\left[\frac{\bar{v}_{k+1}^{-2\alpha}}{\bar{u}_{k+1}^{-\beta}} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2\right]}_{S_{3,k}} \\
& + \underbrace{\frac{2\gamma_x}{\gamma_y\bar{u}_{k+1}^{-\beta}} (\bar{y}_{k+1} - y^*(\bar{x}_k))^T \nabla y^*(\bar{x}_k) (\nabla_x \tilde{F}_k)^T \left(\frac{\mathbf{1}_{k+1}^{-\alpha}}{n} + \frac{\tilde{\mathbf{v}}_{k+1}^{-\alpha}}{n}\right)}_{S_{4,k}}.
\end{aligned} \tag{50}$$

Telescoping the terms, we get

$$\begin{aligned}
& \sum_{k=k_0}^{K-1} \mathbb{E}[f(\bar{x}_k, y^*(\bar{x}_k)) - f(\bar{x}_k, \bar{y}_k)] \\
& \leq \frac{2L^2}{\mu n} \sum_{k=k_0}^{K-1} \mathbb{E}[\Delta_k] + \frac{2}{\mu} \sum_{k=k_0}^{K-1} \mathbb{E}\left[\left\|\frac{\tilde{\mathbf{u}}_{k+1}^{-\beta}}{n\bar{u}_{k+1}^{-\beta}} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\right\|^2\right] \\
& + \sum_{k=k_0}^{K-1} \mathbb{E}\left[\frac{\gamma_y(1+\zeta_u^2)\bar{u}_{k+1}^{-\beta}}{n} \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2\right] \\
& + \sum_{k=k_0}^{K-1} \mathbb{E}[S_{1,k}] + \sum_{k=k_0}^{K-1} \mathbb{E}[S_{2,k}] + \sum_{k=k_0}^{K-1} \mathbb{E}[S_{3,k}] + \sum_{k=k_0}^{K-1} \mathbb{E}[S_{4,k}],
\end{aligned} \tag{51}$$

Next we need to further bound the sums of term $S_{2,k}$, $S_{3,k}$ and $S_{4,k}$ respectively.

$$\begin{aligned}
& \sum_{k=k_0}^{K-1} \mathbb{E}[S_{2,k}] \\
& \leq \sum_{k=k_0}^{K-1} \mathbb{E}\left[\frac{4\gamma_x^2\bar{v}_{k+1}^{-2\alpha}\kappa^2}{\mu\gamma_y^2\bar{u}_{k+1}^{-2\beta}} \left(\left\|\frac{\mathbf{1}^T}{n} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k)\right\|^2 + \left\|\frac{(\tilde{\mathbf{v}}_{k+1}^{-\alpha})^T}{n\bar{v}_{k+1}^{-\alpha}} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k)\right\|^2\right)\right] \\
& \leq \frac{8\gamma_x^2\kappa^2}{\mu\gamma_y^2G^{2\alpha-2\beta}} \sum_{k=k_0}^{K-1} \mathbb{E}\left[\|\nabla_x f(\bar{x}_k, \bar{y}_k)\|^2 + \frac{L^2}{n} \Delta_k\right] \\
& + \sum_{k=k_0}^{K-1} \mathbb{E}\left[\frac{4\gamma_x^2\kappa^2}{\mu\gamma_y^2G^{2\alpha-2\beta}} \left\|\frac{(\tilde{\mathbf{v}}_{k+1}^{-\alpha})^T}{n\bar{v}_{k+1}^{-\alpha}} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k)\right\|^2\right];
\end{aligned} \tag{52}$$

then

$$\begin{aligned}
& \sum_{k=k_0}^{K-1} \mathbb{E} [S_{3,k}] \\
& \leq \sum_{k=k_0}^{K-1} \mathbb{E} \left[\frac{\gamma_x^2 (1 + \zeta_v^2)}{n \gamma_y} \left(\kappa^2 + \frac{2 \gamma_x^2 (1 + \zeta_v^2) C^2 \hat{L}^2}{\mu \gamma_y \bar{v}_0^{2\alpha-\beta}} \right) \frac{\bar{v}_{k+1}^{-2\alpha}}{\bar{u}_{k+1}^{-\beta}} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2 \right] \\
& \leq \frac{\gamma_x^2 (1 + \zeta_v^2)}{\gamma_y \bar{v}_1^{\alpha-\beta}} \left(\kappa^2 + \frac{4 \gamma_x^2 (1 + \zeta_v^2) C^2 \hat{L}^2}{\mu \gamma_y \bar{v}_1^{2\alpha-\beta}} \right) \sum_{k=k_0}^{K-1} \mathbb{E} \left[\frac{\bar{v}_{k+1}^{-\alpha}}{n} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2 \right];
\end{aligned} \tag{53}$$

for the term $S_{4,k}$, we denote $\mathbf{e}_k := \frac{2\gamma_x}{\gamma_y \bar{u}_{k+1}^{-\beta}} (\bar{y}_{k+1} - y^*(\bar{x}_k))^T \nabla y^*(\bar{x}_k) (\nabla_x \tilde{F}_k)^T$, then recalling that

$$\bar{y}_{k+1} = \mathcal{P}_Y \left(\bar{y}_k + \gamma_y \bar{u}_{k+1}^{-\beta} \frac{\mathbf{1}^T}{n} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k) + \gamma_y \frac{(\tilde{\mathbf{u}}_{k+1}^{-\beta})^T}{n} \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k) \right),$$

we have

$$\begin{aligned}
\|\mathbf{e}_k\| & \leq \frac{2\gamma_x \kappa}{\gamma_y \bar{u}_{k+1}^{-\beta}} \|\bar{y}_{k+1} - y^*(\bar{x}_k)\| \|\nabla_x \tilde{F}\| \\
& \leq \frac{2\gamma_x \kappa}{\gamma_y \bar{u}_{k+1}^{-\beta}} \left(\frac{1}{\mu} \|\nabla_y f(\bar{x}_k, \bar{y}_k)\| + \gamma_y \bar{u}_{k+1}^{-\beta} (1 + \zeta_u) C \right) \|\nabla_x \tilde{F}\| \\
& \leq \underbrace{\frac{4\gamma_x \kappa \sqrt{n} C^2}{\gamma_y \bar{u}_1^{-\beta}} \left(\frac{1}{\mu} + \frac{\gamma_y (1 + \zeta_u)}{\bar{u}_1^\beta} \right)}_M
\end{aligned} \tag{54}$$

where we have used the Lipschitz continuity of y^* given in Lemma 3 and Assumption 4. Then, noticing that $\mathbb{E} [\nabla_x \tilde{F}_k] = 0$, we obtain

$$\begin{aligned}
& \sum_{k=k_0}^{K-1} \mathbb{E} [S_{4,k}] = \sum_{k=k_0}^{K-1} \mathbb{E} \left[\mathbf{e}_k \left(\frac{\mathbf{1} \bar{v}_{k+1}^{-\alpha}}{n} + \frac{(\tilde{\mathbf{v}}_{k+1}^{-\alpha})}{n} \right) \right] \\
& = \mathbb{E} \left[\mathbf{e}_{k_0} \left(\frac{\mathbf{1} \bar{v}_{k_0+1}^{-\alpha}}{n} + \frac{(\tilde{\mathbf{v}}_{k_0+1}^{-\alpha})}{n} \right) \right] + \underbrace{\sum_{k=k_0+1}^{K-1} \mathbb{E} \left[\mathbf{e}_k \left(\frac{\mathbf{1} \bar{v}_k^{-\alpha}}{n} + \frac{(\tilde{\mathbf{v}}_k^{-\alpha})}{n} \right) \right]}_0 \\
& + \sum_{k=k_0+1}^{K-1} \mathbb{E} \left[\mathbf{e}_k \left(\frac{\mathbf{1} \bar{v}_{k+1}^{-\alpha}}{n} - \frac{\mathbf{1} \bar{v}_k^{-\alpha}}{n} + \frac{(\tilde{\mathbf{v}}_{k+1}^{-\alpha})}{n} - \frac{(\tilde{\mathbf{v}}_k^{-\alpha})}{n} \right) \right] \\
& \leq \mathbb{E} \left[\frac{M(1 + \zeta_v)}{\bar{v}_1^\alpha \sqrt{n}} \right] + \sum_{k=k_0+1}^{K-1} \mathbb{E} \left[M \left(\underbrace{\frac{(\mathbf{1} \bar{v}_k^{-\alpha} + \tilde{\mathbf{v}}_k^{-\alpha})}{n} - \frac{(\mathbf{1} \bar{v}_{k+1}^{-\alpha} + \tilde{\mathbf{v}}_{k+1}^{-\alpha})}{n}}_{>0} \right) \right] \\
& \leq \frac{8\gamma_x \kappa (1 + \zeta_v) C^2}{\bar{v}_1^\alpha \gamma_y} \left(\frac{1}{\mu} + \frac{\gamma_y (1 + \zeta_u)}{\bar{u}_1^\beta} \right) \mathbb{E} [\bar{u}_K^\beta].
\end{aligned} \tag{55}$$

Therefore combining the obtained inequalities, we complete the proof. \square

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

Lemma 10. Suppose Assumption 1-5 hold. We have

$$\sum_{k=0}^{K-1} \mathbb{E} [S_{1,k}] \leq \mathbb{E} \left[\frac{1}{2\gamma_y \bar{u}_1^{-\beta}} \|\bar{y}_0 - y^*(\bar{x}_0)\|^2 \right] + \frac{(2\beta C^2)^{2+\frac{1}{1-\beta}}}{2\mu^{3+\frac{1}{1-\beta}} \gamma_y^{2+\frac{1}{1-\beta}} \bar{u}_1^{2-2\beta}}. \tag{56}$$

Proof. Firstly, we have

$$\begin{aligned}
& \sum_{k=0}^{K-1} \mathbb{E} \left[\frac{1 - 3\mu\gamma_y \bar{u}_{k+1}^{-\beta}/4}{2\gamma_y \bar{u}_{k+1}^{-\beta}} \|\bar{y}_k - y^*(\bar{x}_k)\|^2 - \frac{1}{\gamma_y \bar{u}_{k+1}^{-\beta} (2 + \mu\gamma_y \bar{u}_{k+1}^{-\beta})} \|\bar{y}_{k+1} - y^*(\bar{x}_{k+1})\|^2 \right] \\
& \leq \frac{1 - 3\mu\gamma_y \bar{u}_1^{-\beta}/4}{2\gamma_y \bar{u}_1^{-\beta}} \|\bar{y}_0 - y^*(\bar{x}_0)\|^2 \\
& + \sum_{k=1}^{K-1} \mathbb{E} \left[\left(\frac{1 - 3\mu\gamma_y \bar{u}_{k+1}^{-\beta}/4}{2\gamma_y \bar{u}_{k+1}^{-\beta}} - \frac{1}{\gamma_y \bar{u}_k^{-\beta} (2 + \mu\gamma_y \bar{u}_k^{-\beta})} \right) \|\bar{y}_k - y^*(\bar{x}_k)\|^2 \right] \\
& \leq \frac{1 - 3\mu\gamma_y \bar{u}_1^{-\beta}/4}{2\gamma_y \bar{u}_1^{-\beta}} \|\bar{y}_0 - y^*(\bar{x}_0)\|^2 \\
& + \sum_{k=1}^{K-1} \mathbb{E} \left[\left(\frac{1}{2\gamma_y \bar{u}_{k+1}^{-\beta}} - \frac{1}{2\gamma_y \bar{u}_k^{-\beta}} - \frac{\mu}{8} + \underbrace{\frac{\mu}{2(2 + \mu\gamma_y \bar{u}_k^{-\beta})} - \frac{\mu}{4}}_{<0} \right) \|\bar{y}_k - y^*(\bar{x}_k)\|^2 \right]. \tag{57}
\end{aligned}$$

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 times.

If the term is positive, noticing that

$$\begin{aligned}
& \frac{\bar{u}_{k+1}^\beta}{2\gamma_y} - \frac{\bar{u}_k^\beta}{2\gamma_y} - \frac{\mu}{8} \\
& \leq \bar{u}_{k+1}^\beta \frac{\left(1 + \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2 / n \bar{u}_{k+1}^\beta\right)^\beta}{2\gamma_y} - \frac{\bar{u}_k^\beta}{2\gamma_y} - \frac{\mu}{8} \\
& \leq \bar{u}_{k+1}^\beta \frac{\left(1 + \beta \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2 / n \bar{u}_{k+1}^\beta\right)}{2\gamma_y} - \frac{\bar{u}_k^\beta}{2\gamma_y} - \frac{\mu}{8} \\
& = \frac{\beta \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2}{2\gamma_y n \bar{u}_{k+1}^{1-\beta}} - \frac{\mu}{8}, \tag{58}
\end{aligned}$$

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

$$\begin{cases} \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k)\|^2 \geq \frac{\gamma_y n \bar{u}_{k+1}^{1-\beta}}{4\beta} \geq \frac{n\mu\gamma_y \bar{u}_1^{1-\beta}}{4\beta}, \\ \frac{4\beta G^2}{2\mu\gamma_y} \geq \frac{4\beta \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2}{2\mu\gamma_y n} \geq \bar{u}_{k+1}^{1-\beta}, \end{cases} \tag{59}$$

which implies that we have at most the following iteration

$$\left(\frac{2\beta CL^2}{\mu\gamma_y} \right)^{\frac{1}{1-\beta}} \frac{4\beta}{\mu\gamma_y \bar{u}_1^{1-\beta}} \tag{60}$$

when the term is positive. Furthermore, when the term is positive, it is also upper bounded,

$$\begin{aligned}
& \left(\frac{1}{2\gamma_y \bar{u}_{k+1}^{-\beta}} - \frac{1}{2\gamma_y \bar{u}_k^{-\beta}} - \frac{\mu}{8} \right) \|\bar{y}_k - y^*(\bar{x}_k)\|^2 \\
& \leq \frac{\beta \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2}{2\gamma_y n \bar{u}_1^{1-\beta}} \|\bar{y}_k - y^*(\bar{x}_k)\|^2 \\
& \leq \frac{\beta C^2}{2\mu^2 \gamma_y \bar{u}_1^{1-\beta}} \|\nabla f(\bar{x}_k, \bar{y}_k)\|^2 \\
& \leq \frac{\beta C^4}{2\mu^2 \gamma_y \bar{u}_1^{1-\beta}}, \tag{61}
\end{aligned}$$

and we have

$$\begin{aligned}
& \sum_{k=1}^{K-1} \mathbb{E} \left[\left(\frac{1}{2\gamma_y \bar{u}_{k+1}^{1-\beta}} - \frac{1}{2\gamma_y \bar{u}_k^{1-\beta}} - \frac{\mu}{8} \right) \|\bar{y}_k - y^*(\bar{x}_k)\|^2 \right] \\
& \leq \frac{\beta C^4}{2\mu^2 \gamma_y \bar{u}_1^{1-\beta}} \left(\frac{2\beta G^2}{\mu \gamma_y} \right)^{\frac{1}{1-\beta}} \frac{4\beta}{\mu \gamma_y \bar{u}_1^{1-\beta}} \\
& = \frac{(2\beta C^2)^{2+\frac{1}{1-\beta}}}{\mu^{3+\frac{1}{1-\beta}} \gamma_y^{2+\frac{1}{1-\beta}} \bar{u}_1^{2-2\beta}},
\end{aligned} \tag{62}$$

which completes the proof. \square

In this part, we bound the term of inconsistency of stepsizes S_3 , which is crucial for the convergence of the proposed algorithm with stepsize control. The formal lemma is given in Lemma 1 in the main text, we give the proof as below.

B.3 PROOF OF LEMMA 1

Proof of Lemma 1. By the definition of $v_{i,k}$ in (3), we have

$$\begin{aligned}
& \mathbb{E} \left[\left\| \frac{(\bar{\mathbf{v}}_{k+1}^{-\alpha})^T}{n \bar{v}_{k+1}^{-\alpha}} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] \\
& \leq \mathbb{E} \left[\frac{1}{n^2} \sum_{i=1}^n (\bar{v}_{k+1}^\alpha - v_{i,k+1}^\alpha)^2 \frac{\|g_{i,k}^x\|^2}{v_{i,k+1}^{2\alpha}} \right] \\
& \leq \mathbb{E} \left[\frac{1}{n^2} \sum_{i=1}^n (\bar{v}_{k+1}^\alpha - v_{i,k+1}^\alpha)^2 \frac{\bar{v}_{k+1}^\alpha}{v_{i,k+1}^{2\alpha}} \frac{\|g_{i,k}^x\|^2}{\bar{v}_{k+1}^\alpha} \right].
\end{aligned} \tag{63}$$

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

$$\begin{aligned}
& \mathbb{E} \left[\left\| \frac{(\bar{\mathbf{v}}_{k+1}^{-\alpha})^T}{n \bar{v}_{k+1}^{-\alpha}} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] \\
& \leq \mathbb{E} \left[\frac{1}{n^2} \sum_{i=1}^n (\bar{v}_{k+1}^\alpha - v_{i,k+1}^\alpha)^2 \left(\frac{\bar{v}_{k+1}^\alpha - v_{i,k+1}^\alpha}{v_{i,k+1}^{2\alpha}} + \frac{1}{v_{i,k+1}^\alpha} \right) \frac{\|g_{i,k}^x\|^2}{\bar{v}_{k+1}^\alpha} \right] \\
& \leq \mathbb{E} \left[\frac{1}{n^2} \sum_{i=1}^n \frac{(\bar{v}_{k+1}^\alpha - v_{i,k+1}^\alpha)^2}{v_{i,k+1}^{2\alpha}} |\bar{v}_{k+1}^\alpha - v_{i,k+1}^\alpha| \frac{\|g_{i,k}^x\|^2}{\bar{v}_{k+1}^\alpha} \right] \\
& + \mathbb{E} \left[\frac{1}{n^2} \sum_{i=1}^n \frac{|\bar{v}_{k+1}^\alpha - v_{i,k+1}^\alpha|}{v_{i,k+1}^\alpha} |\bar{v}_{k+1}^\alpha - v_{i,k+1}^\alpha| \frac{\|g_{i,k}^x\|^2}{\bar{v}_{k+1}^\alpha} \right] \\
& \leq (1 + \zeta_v) \zeta_v \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n |\bar{v}_{k+1}^\alpha - v_{i,k+1}^\alpha| \frac{1}{n} \sum_{i=1}^n \frac{\|g_{i,k}^x\|^2}{\bar{v}_{k+1}^\alpha} \right].
\end{aligned} \tag{64}$$

By Lemma 6, we get

$$\begin{aligned}
& \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[\left\| \frac{(\tilde{\mathbf{v}}_{k+1}^{-\alpha})^T}{n\bar{v}_{k+1}^{-\alpha}} \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] \\
& \leq (1 + \zeta_v) \zeta_v \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n |\bar{v}_{k+1}^\alpha - v_{i,k+1}^\alpha| \frac{1}{K} \sum_{k=0}^{K-1} \frac{\frac{1}{n} \sum_{i=1}^n \|g_{i,k}^x\|^2}{\bar{v}_{k+1}^\alpha} \right] \\
& \leq (1 + \zeta_v) \zeta_v \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n |\bar{v}_{k+1}^\alpha - v_{i,k+1}^\alpha| \right] \frac{C^{2-2\alpha}}{(1-\alpha)K^\alpha} \\
& \leq (1 + \zeta_v) \zeta_v \sqrt{\frac{1}{n} \mathbb{E} [\|\mathbf{v}_{k+1} - \mathbf{1}\bar{v}_{k+1}\|^{2\alpha}]} \frac{C^{2-2\alpha}}{(1-\alpha)K^\alpha}.
\end{aligned} \tag{65}$$

Next, for the term of inconsistency of the stepsize $\|\mathbf{v}_k - \mathbf{1}\bar{v}_k\|^2$, we consider two cases since the max operator we used. At iteration k , for the case $\mathbf{m}_k^x \geq \mathbf{m}_k^y$, we have

$$\begin{aligned}
& \mathbb{E} [\|\mathbf{v}_{k+1} - \mathbf{1}\bar{v}_{k+1}\|^2] = \mathbb{E} [\|\mathbf{m}_{k+1}^x - \mathbf{1}\bar{m}_{k+1}^x\|^2] \\
& = \mathbb{E} [\|(W - \mathbf{J})(\mathbf{m}_k^x - \mathbf{1}\bar{m}_k^x) + \eta_k(W - \mathbf{J})\mathbf{h}_k^x\|^2] \\
& \leq \frac{1 + \rho_W}{2} \mathbb{E} [\|\mathbf{m}_k^x - \mathbf{1}\bar{m}_k^x\|^2] + \frac{(1 + \rho_W)\rho_W}{1 - \rho_W} \mathbb{E} [\|\mathbf{h}_k^x\|^2] \\
& \leq \left(\frac{1 + \rho_W}{2} \right)^k \mathbb{E} [\|\mathbf{m}_0^x - \mathbf{1}\bar{m}_0^x\|^2] + \frac{nC^2(1 + \rho_W)\rho_W}{1 - \rho_W} \sum_{t=0}^k \left(\frac{1 + \rho_W}{2} \right)^{k-t} \\
& \leq \frac{2nC^2(1 + \rho_W)\rho_W}{(1 - \rho_W)^2},
\end{aligned} \tag{66}$$

where we set $\|\mathbf{m}_0^x - \mathbf{1}\bar{m}_0^x\|^2 = 0$; for the case $\mathbf{m}_k^x < \mathbf{m}_k^y$, with $\|\mathbf{m}_0^y - \mathbf{1}\bar{m}_0^y\|^2 = 0$,

$$\mathbb{E} [\|\mathbf{v}_{k+1} - \mathbf{1}\bar{v}_{k+1}\|^2] = \mathbb{E} [\|\mathbf{m}_{k+1}^y - \mathbf{1}\bar{m}_{k+1}^y\|^2] \leq \frac{2nC^2(1 + \rho_W)\rho_W}{(1 - \rho_W)^2}, \tag{67}$$

Combining these two cases, and using Lemma 6 and the fact $\|\mathbf{v}_k^\alpha - \mathbf{1}\bar{v}_k^\alpha\|^2 \leq \|\mathbf{v}_k - \mathbf{1}\bar{v}_k\|^{2\alpha}$ for $\alpha \in (0, 1)$, we complete the proof. \square

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

Lemma 11 (Inconsistency for D-TiAda). *Suppose Assumption 1-5 hold. For D-TiAda, we have*

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

Proof. By the definition of inconsistency of stepsizes in (7) and Assumption 4 on bounded gradient, we immediately get the result. \square

B.4 PROOF OF THEOREM 1

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

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

where

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

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

$$y = -\frac{1+a}{a+\frac{a}{b}}x, \tag{69}$$

where we have

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

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

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

which means gradient of x would be chosen in the maximum operator in the denominator of TiAda stepsize for x . Therefore, after one step, we have

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

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

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

Note that $g_{1,0}^x = -a \cdot g_{2,0}^x$, then we get

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

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

B.5 PROOF OF THEOREM 2

Proof of Theorem 2. Combining the results in Lemma 8, 9 and 10, we get

$$\begin{aligned}
& \sum_{k=0}^{K-1} \mathbb{E} [f(\bar{x}_k, y^*(\bar{x}_k)) - f(\bar{x}_k, \bar{y}_k)] \\
&= \sum_{k=0}^{k_0-1} \mathbb{E} [f(\bar{x}_k, y^*(\bar{x}_k)) - f(\bar{x}_k, \bar{y}_k)] + \sum_{k=k_0}^{K-1} \mathbb{E} [f(\bar{x}_k, y^*(\bar{x}_k)) - f(\bar{x}_k, \bar{y}_k)] \\
&\leq \mathbb{E} \left[\frac{1}{2\gamma_y \bar{u}_1^{-\beta}} \|\bar{y}_0 - y^*(\bar{x}_0)\|^2 \right] + \frac{(2\beta C^2)^{2+\frac{1}{1-\beta}}}{2\mu^{3+\frac{1}{1-\beta}} \gamma_y^{2+\frac{1}{1-\beta}} \bar{u}_1^{2-2\beta}} \\
&+ \left(\frac{4L^2}{\mu n} + \frac{8\gamma_x^2 \kappa^2 L^2}{\mu n \gamma_y^2 G^{2\alpha-2\beta}} \right) \sum_{k=0}^{K-1} \mathbb{E} [\Delta_k] + \frac{\gamma_y (1 + \zeta_u)}{n} \sum_{k=0}^{K-1} \mathbb{E} [\bar{u}_k^{-\beta} \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2] \\
&+ \frac{2\gamma_x^2 \kappa^2 (1 + \zeta_v^2) G^{2\beta}}{\mu \gamma_y^2} \sum_{k=0}^{k_0-1} \mathbb{E} \left[\frac{\bar{v}_{k+1}^{-2\alpha}}{n} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2 \right] \\
&+ \frac{\gamma_x^2 (1 + \zeta_v^2)}{\gamma_y \bar{v}_1^{\alpha-\beta}} \left(\kappa^2 + \frac{4\gamma_x^2 (1 + \zeta_v^2) C^2 \hat{L}^2}{\mu \gamma_y \bar{v}_1^{2\alpha-\beta}} \right) \sum_{k=k_0}^{K-1} \mathbb{E} \left[\frac{\bar{v}_{k+1}^{-\alpha}}{n} \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2 \right] \\
&+ \left(\frac{4}{\mu} + \frac{4\gamma_x^2 \kappa^2}{\mu \gamma_y^2 G^{2\alpha-2\beta}} \right) \sum_{k=0}^{K-1} \mathbb{E} \left[\left\| \frac{\bar{\mathbf{u}}_{k+1}^{-\beta}}{n \bar{u}_{k+1}^{-\beta}} \right\|^2 \|\nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y)\|^2 \right] \\
&+ \frac{8\gamma_x^2 \kappa^2}{\mu \gamma_y^2 G^{2\alpha-2\beta}} \sum_{k=k_0}^{K-1} \mathbb{E} [\|\nabla_x f(\bar{x}_k, \bar{y}_k)\|^2] + \frac{8\gamma_x \kappa (1 + \zeta_v) C^2}{\bar{v}_1^\alpha \gamma_y} \left(\frac{1}{\mu} + \frac{\gamma_y (1 + \zeta_u)}{\bar{u}_1^\beta} \right) \mathbb{E} [\bar{u}_K^\beta].
\end{aligned} \tag{70}$$

Let the dividing point between the two phase in Lemma 8 and 9 be

$$G = \left(\frac{16\gamma_x^2 \kappa^4}{\gamma_y^2} \right)^{\frac{1}{2\alpha-2\beta}}, \tag{71}$$

then, plugging above inequality into (19) in Lemma 5, with the help of Lemma 6-10 and Lemma 1, we can get the following result,

$$\begin{aligned}
& \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} [\|\nabla \Phi(\bar{x}_k)\|^2] \\
&\leq E_0(K) + E_G(K) + E_W(K) + \frac{8C^{2\alpha} (\Phi^{\max} - \Phi^*)}{\gamma_x K^{1-\alpha}} \\
&+ 64(1 + \zeta_v) \zeta_v \sqrt{\frac{1}{n^{1-\alpha}} \left(\frac{\rho_W}{(1 - \rho_W)^2} \right)^\alpha} \frac{C^{2-\alpha}}{(1 - \alpha) K^\alpha} \\
&+ \left(\frac{\gamma_x \kappa^3 L}{\gamma_y \bar{v}_0^{\alpha-\beta}} + \frac{4\gamma_x^3 (1 + \zeta_v^2) C^2 \kappa^2 \hat{L}^2}{\gamma_y^2 \bar{v}_0^{3\alpha-2\beta}} + L_\Phi \right) \frac{8\gamma_x (1 + \zeta_v^2) C^{2-2\alpha}}{(1 - \alpha) K^\alpha} \\
&+ 8(1 + 16\kappa^2) (1 + \zeta_u) \zeta_u \sqrt{\frac{1}{n^{1-\beta}} \left(\frac{\rho_W}{(1 - \rho_W)^2} \right)^\beta} \frac{C^{2-\beta}}{(1 - \beta) K^\beta} \\
&+ \frac{8\gamma_y \kappa L (1 + \zeta_u^2) C^{2-2\beta}}{(1 - \beta) K^\beta} + \frac{64\gamma_x \kappa^2 (1 + \zeta_v) C^2}{\gamma_y \bar{v}_0^\alpha} \left(\kappa + \frac{\gamma_y (1 + \zeta_u) L}{\bar{u}_0^\beta} \right) \frac{C^{2\beta}}{K^{1-\beta}}.
\end{aligned} \tag{72}$$

Algorithm 2 DAS²C with coordinate-wise adaptive stepsize

Initialization: $x_{i,0}, y_{i,0} \in \mathbb{R}^p$, buffers $m_{i,0}^x, m_{i,0}^y > 0$, stepsizes $\gamma_x, \gamma_y > 0$ and $0 < \beta < \alpha < 1$.

- 1: **for** iteration $k = 0, 1, \dots$, each node $i \in [n]$, **do**
- 2: Sample i.i.d $\xi_{i,k}^x$ and $\xi_{i,k}^y$, compute:

$$g_{i,k}^x = \nabla_x f_i(x_{i,k}, y_{i,k}; \xi_{i,k}^x), g_{i,k}^y = \nabla_y f_i(x_{i,k}, y_{i,k}; \xi_{i,k}^y).$$

- 3: Update local gradient sums with Schur product:

$$m_{i,k+1}^x = m_{i,k}^x + g_{i,k}^x \odot g_{i,k}^x, m_{i,k+1}^y = m_{i,k}^y + g_{i,k}^y \odot g_{i,k}^y$$

- 4: Compute the ratio between two gradient sums:

$$\psi_{i,k+1} = \|m_{i,k+1}^x\|^{2\alpha} / \max \left\{ \|m_{i,k+1}^x\|^{2\alpha}, \|m_{i,k+1}^y\|^{2\alpha} \right\} \leq 1.$$

- 5: Update primal and dual variables locally:

$$\begin{aligned} x_{i,k+1} &= x_{i,k} - \gamma_x \psi_{i,k+1} (m_{i,k+1}^x)^{-\alpha} \odot g_{i,k}^x, \\ y_{i,k+1} &= \mathcal{P}_Y \left(y_{i,k} - \gamma_y (m_{i,k+1}^y)^{-\beta} \odot g_{i,k}^y \right). \end{aligned}$$

- 6: Communicate parameters over network:

$$\left\{ m_{i,k+1}^x, m_{i,k+1}^y, x_{i,k+1}, y_{i,k+1} \right\} \leftarrow \sum_{j \in \mathcal{N}_i} W_{i,j} \left\{ m_{j,k+1}^x, m_{j,k+1}^y, x_{j,k+1}, y_{j,k+1} \right\}.$$

7: **end for**

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

$$\begin{aligned} E_0(K) &:= \frac{\mathbb{E} \left[4\kappa L \bar{u}_1^\beta \|\bar{y}_0 - y^*(\bar{x}_0)\|^2 \right]}{\gamma_y K} + \frac{4\kappa^2 (2\beta C^2)^{2+\frac{1}{1-\beta}}}{\mu^{2+\frac{1}{1-\beta}} \gamma_y^{2+\frac{1}{1-\beta}} \bar{u}_1^{2-2\beta} K}, \\ E_G(K) &:= \frac{16\gamma_x^2 \kappa^4 (1 + \zeta_v^2) G^{2\beta}}{\gamma_y^2} \left(\frac{C^{2-4\alpha}}{(1-2\alpha) K^{2\alpha}} \mathbb{I}_{\alpha < 1/2} + \frac{1 + \log v_K - \log v_0}{K \bar{v}_0^{2\alpha-1}} \mathbb{I}_{\alpha \geq 1/2} \right), \\ E_W(K) &:= \frac{16\rho_W \gamma_x^2 (1 + \zeta_v^2) (2 + L^2 + 8\kappa^2) L^2}{(1 - \rho_W)^2} \left(\frac{C^{2-4\alpha}}{(1-2\alpha) K^{2\alpha}} \mathbb{I}_{\alpha < 1/2} + \frac{1 + \log v_K - \log v_0}{K \bar{v}_0^{2\alpha-1}} \mathbb{I}_{\alpha \geq 1/2} \right) \\ &\quad + \frac{16\rho_W \gamma_x^2 (1 + \zeta_u^2) (2 + L^2 + 8\kappa^2) L^2}{(1 - \rho_W)^2} \left(\frac{C^{2-4\beta}}{(1-2\beta) K^{2\beta}} \mathbb{I}_{\beta < 1/2} + \frac{1 + \log u_K - \log v_0}{K \bar{u}_0^{2\beta-1}} \mathbb{I}_{\beta \geq 1/2} \right), \end{aligned}$$

Letting the total iteration K satisfy the conditions in (12) such that the term in E_G and E_W are dominated, we complete the proof. \square

B.5.1 PROOF OF COROLLARY 1

Proof of Corollary 1. With the help of Lemma 11, we can directly adapt the proof of Theorem 2 to get the result in (14). \square

B.6 EXTEND THE PROOF TO COORDINATE-WISE STEPSIZE

In this subsection, we show how to extend our convergence analysis of DAS²C to the coordinate-wise adaptive stepsize (Zhou et al., 2018) variant. We first present this variant in the Algorithm 2, which can be rewritten in a compact form with the Schur product denoted by \odot .

$$\mathbf{m}_{k+1}^x = W(\mathbf{m}_k^x + \mathbf{h}_k^x), \quad (73a)$$

$$\mathbf{m}_{k+1}^y = W(\mathbf{m}_k^y + \mathbf{h}_k^y), \quad (73b)$$

$$\mathbf{x}_{k+1} = W(\mathbf{x}_k - \gamma_x V_{k+1}^{-\alpha} \odot \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k)), \quad (73c)$$

$$\mathbf{y}_{k+1} = \mathcal{P}_y \left(W \left(\mathbf{y}_k + \gamma_y U_{k+1}^{-\beta} \odot \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k) \right) \right), \quad (73d)$$

where

$$\mathbf{h}_k^x = [\cdots, g_{i,k}^x \odot g_{i,k}^x, \cdots]^T \in \mathbb{R}^{n \times p}, \quad \mathbf{h}_k^y = [\cdots, g_{i,k}^y \odot g_{i,k}^y, \cdots]^T \in \mathbb{R}^{n \times d},$$

and the matrices U_k^α and V_k^β are redefined as follows:

$$\begin{aligned} V_k^{-\alpha} &= [\cdots, v_{i,k}^{-\alpha}, \cdots]^T, \quad [v_{i,k}]_j = \max \left\{ [m_{i,k}^x]_j, [m_{i,k}^y]_j \right\}, \quad j \in [p], \\ U_k^{-\beta} &= [\cdots, u_{i,k}^{-\beta}, \cdots]^T, \quad [u_{i,k}]_j = [m_{i,k}^x]_j, \quad j \in [d] \end{aligned} \quad (74)$$

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

Recalling the definitions of inconsistency of stepsize in (7), we give the following notations:

$$\begin{aligned} \tilde{V}_k &= V_k - \bar{v}_k \mathbf{1}\mathbf{1}_p^T, \quad \bar{v}_k = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p V_{ij}, \quad \bar{v}_{i,k} = \frac{1}{p} \sum_{j=1}^p V_{ij}, \quad \bar{v}_{j,k} = \frac{1}{n} \sum_{i=1}^n V_{ij}, \\ \tilde{U}_k &= U_k - \bar{u}_k \mathbf{1}\mathbf{1}_d^T, \quad \bar{u}_k = \frac{1}{nd} \sum_{i=1}^n \sum_{j=1}^d U_{ij}, \quad \bar{u}_{i,k} = \frac{1}{d} \sum_{j=1}^d U_{ij}, \quad \bar{u}_{j,k} = \frac{1}{n} \sum_{i=1}^n U_{ij}, \end{aligned} \quad (75)$$

and

$$\begin{aligned} \zeta_V^2 &= \sup_{k \geq 0} \left\{ \frac{\|V_k^{-\alpha} - \bar{v}_k^{-\alpha} \mathbf{1}\mathbf{1}_p^T\|^2}{np (\bar{v}_k^{-\alpha})^2} \right\}, \quad \hat{\zeta}_v^2 = \sup_{k \geq 0} \left\{ \frac{\|V_k^{-\alpha} - (V_k \mathbf{J}_p)^{-\alpha}\|^2}{np (\bar{v}_k^{-\alpha})^2} \right\}, \\ \zeta_U^2 &= \sup_{k \geq 0} \left\{ \frac{\|U_k^{-\beta} - \bar{u}_k^{-\beta} \mathbf{1}\mathbf{1}_d^T\|^2}{nd (\bar{u}_k^{-\beta})^2} \right\}, \quad \hat{\zeta}_u^2 = \sup_{k \geq 0} \left\{ \frac{\|U_k^{-\beta} - (U_k \mathbf{J}_d)^{-\beta}\|^2}{nd (\bar{u}_k^{-\beta})^2} \right\}. \end{aligned}$$

According to the two definitions of inconsistency of stepsize for Option I and II, we can give the following lemma to show their difference.

Lemma 12 (Inconsistency, coordinate-wise). *Suppose Assumption 1-5 hold. For the proposed DAS²C algorithm, we have*

$$\begin{aligned} & \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[\left\| \frac{\mathbf{1}^T}{n \bar{v}_{k+1}^{-\alpha}} \tilde{V}_{k+1}^{-\alpha} \odot \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] \\ & \leq 2(1 + \zeta_v) \zeta_v \sqrt{\frac{1}{n^{1-\alpha}} \left(\frac{4C^2 \rho_W}{(1 - \rho_W)^2} \right)^\alpha} \frac{C^{2-2\alpha}}{(1 - \alpha) K^\alpha} + 2np \hat{\zeta}_v^2 C^2 \end{aligned} \quad (76)$$

and

$$\begin{aligned} & \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[\left\| \frac{\mathbf{1}^T}{n \bar{u}_{k+1}^{-\beta}} \tilde{U}_{k+1}^{-\beta} \odot \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y) \right\|^2 \right] \\ & \leq 2(1 + \zeta_u) \zeta_u \sqrt{\frac{1}{n^{1-\beta}} \left(\frac{4C^2 \rho_W}{(1 - \rho_W)^2} \right)^\beta} \frac{C^{2-2\beta}}{(1 - \beta) K^\beta} + 2nd \hat{\zeta}_u^2 C^2. \end{aligned} \quad (77)$$

In contrast, for D-TiAda, we have

$$\begin{aligned} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[\left\| \frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}} \tilde{V}_{k+1}^{-\alpha} \odot \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] &\leq p\zeta_V^2 C^2, \\ \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[\left\| \frac{\mathbf{1}^T}{n\bar{u}_{k+1}^{-\alpha}} \tilde{U}_{k+1}^{-\beta} \odot \nabla_y F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^y) \right\|^2 \right] &\leq d\zeta_U^2 C^2. \end{aligned} \quad (78)$$

Proof. For the coordinate-wise stepsize, with the help of Lemma 4, we have

$$\begin{aligned} &\mathbb{E} \left[\left\| \frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}} \tilde{V}_{k+1}^{-\alpha} \odot \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] \\ &= \mathbb{E} \left[\left\| \frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}} \left(V_{k+1}^{-\alpha} - (V_{k+1}\mathbf{J})^{-\alpha} + (V_{k+1}\mathbf{J})^{-\alpha} - \bar{v}_{k+1}^{-\alpha} \mathbf{1}\mathbf{1}_p^T \right) \odot \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] \\ &\leq 2\mathbb{E} \left[\left\| \frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}} \left((V_{k+1}\mathbf{J})^{-\alpha} - \bar{v}_{k+1}^{-\alpha} \mathbf{1}\mathbf{1}_p^T \right) \odot \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] \\ &\quad + 2\mathbb{E} \left[\left\| \frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}} \left(V_{k+1}^{-\alpha} - (V_{k+1}\mathbf{J})^{-\alpha} \right) \odot \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right]. \end{aligned} \quad (79)$$

For the first term in the last line, by Lemma 4, we have

$$\begin{aligned} &\mathbb{E} \left[\left\| \frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}} \left((V_{k+1}\mathbf{J})^{-\alpha} - \bar{v}_{k+1}^{-\alpha} \mathbf{1}\mathbf{1}_p^T \right) \odot \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] \\ &\leq \mathbb{E} \left[\frac{1}{n^2 \bar{v}_{k+1}^{-2\alpha}} \sum_{i=1}^n (\bar{v}_{i,k+1}^{-\alpha} - \bar{v}_{k+1}^{-\alpha})^2 \|\nabla_x f_i(x_{i,k}, y_{i,k}; \xi_{i,k}^x)\|^2 \right]. \end{aligned} \quad (80)$$

which is a similar term with Option I and is convergent. Then, for the second part,

$$\begin{aligned} &\mathbb{E} \left[\left\| \frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}} \left(V_{k+1}^{-\alpha} - (V_{k+1}\mathbf{J})^{-\alpha} \right) \odot \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] \\ &\leq \frac{1}{n} \mathbb{E} \left[\left\| \frac{V_{k+1}^{-\alpha} - (V_{k+1}\mathbf{J})^{-\alpha}}{\bar{v}_{k+1}^{-\alpha}} \right\|^2 \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2 \right] \\ &\leq p\hat{\zeta}_v^2 \mathbb{E} [\|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2]. \end{aligned} \quad (81)$$

where the term $\hat{\zeta}_v^2$ cannot be guaranteed to be convergent because the step size between the different dimensions of each node is inconsistent and uncontrolled. Noticing that for D-TiAda,

$$\mathbb{E} \left[\left\| \frac{\mathbf{1}^T}{n\bar{v}_{k+1}^{-\alpha}} \tilde{V}_{k+1}^{-\alpha} \odot \nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x) \right\|^2 \right] \leq \frac{1}{n} \mathbb{E} \left[\left\| \frac{\tilde{V}_{k+1}^{-\alpha}}{\bar{v}_{k+1}^{-\alpha}} \right\|^2 \|\nabla_x F(\mathbf{x}_k, \mathbf{y}_k; \xi_k^x)\|^2 \right] \leq p\zeta_V^2 C^2, \quad (82)$$

Using Lemma 1, we complete the proof. \square

Theorem 3. Suppose Assumption 1-5 hold. Let $0 < \alpha < \beta < 1$ and the total iteration satisfy

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

to ensure time-scale separation and quasi-independence of network. For DAS²C with coordinate-wise adaptive stepsize, we have

$$\begin{aligned} & \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} \left[\|\nabla \Phi(\bar{x}_k)\|^2 \right] \\ &= \tilde{\mathcal{O}} \left(\frac{1}{K^{1-\alpha}} + \frac{1}{(1-\rho_W)^\alpha K^\alpha} + \frac{1}{K^{1-\beta}} + \frac{1}{(1-\rho_W)^\beta K^\beta} \right) + \mathcal{O} \left(n \left(p\hat{\zeta}_v^2 + \kappa^2 d\hat{\zeta}_u^2 \right) C^2 \right). \end{aligned} \quad (83)$$

Proof. With the help of Lemma 12 and the obtained result (72) in the proof of Theorem 2, we can derive the convergence results for DAS²C with coordinate-wise adaptive stepsize. \square

Remark 5. In Theorem 3, we show that there is a steady-state error in the upper bound of the coordinate-wise variant of DAS²C depending on the number of nodes and the problem’s dimension. However, it’s worth noting that the coordinate-wise scheme exhibits strong performance in numerous real-world experiments, particularly for high-dimensional problems (Li et al., 2023) at the cost of increased communication overhead. The observed gap between the theoretical analysis and experimental results can be attributed to our assumption of bounded gradients (c.f., Assumption 4, i.e., $\|\nabla_z f_i(x, y; \xi_i)\|^2 \leq C$), which hides the information about the dimension of the problem. We believe an interesting direction for future work is to find effective ways to close the gap.