

# A Single-Loop Algorithm for Decentralized Bilevel Optimization

Youran Dong\*, Shiqian Ma<sup>†</sup>, Junfeng Yang\*, Chao Yin\*

## Abstract

Bilevel optimization has gained significant attention in recent years due to its broad applications in machine learning. This paper focuses on bilevel optimization in decentralized networks and proposes a novel single-loop algorithm for solving decentralized bilevel optimization with a strongly convex lower-level problem. Our approach is a fully single-loop method that approximates the hypergradient using only two matrix-vector multiplications per iteration. Importantly, our algorithm does not require any gradient heterogeneity assumption, distinguishing it from existing methods for decentralized bilevel optimization and federated bilevel optimization. Our analysis demonstrates that the proposed algorithm achieves the best-known convergence rate for bilevel optimization algorithms. We also present experimental results on hyperparameter optimization problems using both synthetic and MNIST datasets, which demonstrate the efficiency of our proposed algorithm.

## 1 Introduction

Bilevel optimization (BO) has received increasing attention in recent studies due to its wide applications in machine learning, including but not limited to hyperparameter optimization [27, 10], meta learning [10, 30, 16] and adversarial training [2, 33, 37]. A generic BO takes the form

$$\min_{x \in \mathbb{R}^p} \Phi(x) = F(x, y^*(x)), \quad \text{s.t. } y^*(x) = \arg \min_{y \in \mathbb{R}^q} f(x, y). \quad (1)$$

Throughout this paper, we assume that the lower-level (LL) objective function  $f$  is twice continuously differentiable and strongly convex with respect to  $y$  for any fixed  $x$ . Problem (1) aims at minimizing the upper-level (UL) function  $F$  with respect to  $x$  with  $y$  being the optimal solution of the LL problem for fixed  $x$ . Algorithms for solving BO (1) have been studied extensively. When the LL problem is strongly convex with respect to  $y$  so that it admits unique solution for fixed  $x$ , a natural idea to solve (1) is to apply gradient descent for the UL problem. Under the assumption that  $F$  is smooth, the gradient descent method for solving (1) updates the iterate as follows:

$$x^{k+1} := x^k - \tau_{x,k} \nabla \Phi(x^k),$$

where  $\tau_{x,k} > 0$  is a step size, and the hypergradient  $\nabla \Phi(x)$  is given by

$$\nabla \Phi(x) := \nabla_1 F(x, y^*(x)) - \nabla_{12}^2 f(x, y^*(x)) [\nabla_{22}^2 f(x, y^*(x))]^{-1} \nabla_2 F(x, y^*(x)). \quad (2)$$

---

\*Department of Mathematics, Nanjing University, Nanjing, P. R. China. Email: yrdong@smail.nju.edu.cn, jfyang@nju.edu.cn, yinchao@smail.nju.edu.cn. Research supported by the National Natural Science Foundation of China (NSFC-12371301).

<sup>†</sup>Department of Computational Applied Mathematics and Operations Research, Rice University, Houston, USA. Email: sqma@rice.edu. Research supported in part by NSF grants DMS-2243650, CCF-2308597, CCF-2311275 and ECCS-2326591, and a startup fund from Rice University.

Two challenges arise from computing the hypergradient in (2): (i) how to efficiently (approximately) compute  $y^*(x^k)$ , which requires solving the LL problem for given  $x^k$ ; and (ii) how to deal with the matrix inversion, or equivalently, solve the linear system in (2). Different approaches addressing these two issues lead to different algorithms for solving (1). Let  $K$  and  $T$  be given positive integers. A basic algorithm along this line for solving (1) updates the iterates as follows:

$$\begin{aligned}
& \text{for } k = 0, 1, \dots, K - 1 \\
& \quad y^{k,0} = y^{k-1,T} \\
& \quad \text{for } t = 0, 1, \dots, T - 1 \\
& \quad \quad y^{k,t+1} = y^{k,t} - \tau_{y,k} \nabla_2 f(x^k, y^{k,t}) \\
& \quad \text{end for} \\
& \quad x^{k+1} = x^k - \tau_{x,k} \widetilde{\nabla} \Phi(x^k) \\
& \text{end for}
\end{aligned} \tag{3}$$

where  $\widetilde{\nabla} \Phi(x^k)$  is an approximation of the hypergradient  $\nabla \Phi(x^k)$  and is defined as

$$\widetilde{\nabla} \Phi(x^k) = \nabla_1 F(x^k, y^{k,T}) - \nabla_{12}^2 f(x^k, y^{k,T}) [\nabla_{22}^2 f(x^k, y^{k,T})]^{-1} \nabla_2 F(x^k, y^{k,T}). \tag{4}$$

In practice, exactly calculating the Hessian inverse or solving the linear system of equations in (4) is computationally inefficient, and hence two representative approaches to estimate (4) have been proposed in the literature: iterative differentiation (ITD) and approximate implicit differentiation (AID). Approaches related to ITD, such as those proposed in [10, 32, 16, 13, 18, 17], leverage automatic differentiation to approximate the hypergradient using backpropagation. Approaches related to AID, including those proposed in [27, 12, 13, 18, 17, 4, 14, 8], use various methods to approximately evaluate  $[\nabla_{22}^2 f(x^k, y^{k,T})]^{-1} \nabla_2 F(x^k, y^{k,T})$  in (4). Some of these methods employ gradient descent or conjugate gradient methods, while others use Neumann series to approximate the Hessian inverse. Additionally, it should be noted that the update scheme (3) involves a double-loop structure, where updating  $x$  constitutes the outer loop while updating  $y$  represents the inner loop. However, this structure is not preferable in practical settings. Some works [4, 14] eliminate this double-loop structure by taking  $T = 1$  in (3), yet still require the use of the AID approach to estimate (4). Furthermore, both AID and ITD approaches need another sub-loop to estimate (4), which may involve  $\Theta(\log K)$  Hessian- and Jacobian-vector multiplications per iteration (see, e.g., [18, Theorem 2]).

Recently, Dagr eou et al. [8] proposed a fully single-loop framework (named SOBA) for solving (1) that only needs one matrix-vector multiplication to approximately solve the linear system of equations in (4) in each iteration. The SOBA algorithm maintains three sequences and updates them as

$$y^{k+1} = y^k - \beta_k D_y^k, \quad v^{k+1} = v^k + \eta_k D_v^k, \quad x^{k+1} = x^k - \alpha_k D_x^k, \tag{5}$$

where  $\alpha_k$ ,  $\beta_k$  and  $\eta_k$  are stepsizes,  $D_y^k$ ,  $D_v^k$  and  $D_x^k$  are respectively unbiased stochastic estimators of  $d_y(x^k, y^k)$ ,  $d_v(x^k, y^k, v^k)$  and  $d_x(x^k, y^k, v^k)$  defined as  $d_y(x, y) = \nabla_2 f(x, y)$ ,

$$d_v(x, y, v) = \nabla_2 F(x, y) - \nabla_{22}^2 f(x, y) v \quad \text{and} \quad d_x(x, y, v) = \nabla_1 F(x, y) - \nabla_{12}^2 f(x, y) v.$$

The SOBA framework [8] has been extended by [22] to handle the case where the LL problem is merely convex. In this case, it is assumed that the sequence  $\{v^k\}$  is bounded to ensure convergence. Other notable BO algorithms include [19, 3] which first convert (1) to an equivalent constrained single-level

problem, and then approximately solve this reformulated problem. While this approach circumvents the need for computing the approximate hypergradient (4), it requires careful consideration when handling the constraints. In [15], the authors suggest transforming BO into an unconstrained constraint dissolving problem, enabling the direct application of efficient unconstrained optimization methods to BO problems.

The main focus of this paper is to design a single-loop algorithm for decentralized bilevel optimization (DBO). DBO considers BO in a decentralized network, where the data are naturally distributed to  $n$  agents, each with access to their own local data and communication limited to their immediate neighbors in the network. The  $n$  agents cooperatively solve the BO problem through local updates and communications. Decentralized optimization has many benefits, such as enhancing computational efficiency and sharing data while protecting data privacy [21]. The general form of DBO is given below:

$$\begin{aligned} \min_{x \in \mathbb{R}^p} \Phi(x) = F(x, y^*(x)) &:= \frac{1}{n} \sum_{i=1}^n F_i(x, y^*(x)) \\ \text{s.t. } y^*(x) = \arg \min_{y \in \mathbb{R}^q} f(x, y) &:= \frac{1}{n} \sum_{i=1}^n f_i(x, y), \end{aligned} \quad (6)$$

where the  $i$ -th agent only has access to the data related to  $F_i$  and  $f_i$ . To illustrate DBO, let us consider the example of training a binary classification model using medical data from  $n$  different hospitals. Suppose we want to train a logistic regression model that can predict whether a tumor is benign or malignant based on medical data (such as MRI). The hospitals are local agents that cannot share their data with other hospitals to preserve patients' privacy. In this case, the hyperparameter tuning problem can be formulated as follows, and we need to design a decentralized algorithm to solve it in a distributed network:

$$\min_{\lambda} \frac{1}{n} \sum_{i=1}^n \sum_{(x_e, y_e) \in D'_i} \psi(y_e x_e^\top \omega^*(\lambda)), \quad \text{s.t. } \omega^*(\lambda) = \arg \min_{\omega} \frac{1}{n} \sum_{i=1}^n \sum_{(x_e, y_e) \in D_i} \psi(y_e x_e^\top \omega) + \lambda \|\omega\|_2^2.$$

In this problem,  $\psi$  is a loss function used to train the logistic regression model. A commonly used loss function is the logistic loss, given by  $\psi(t) = \log(1 + e^{-t})$ . The parameter vector  $\omega$  comprises the regression parameters, while  $\lambda$  represents the hyperparameter of the  $\ell_2$ -norm-squared regularizer, which is used to prevent overfitting. The training and testing datasets from hospital  $i$  are denoted by  $D_i$  and  $D'_i$ , respectively.

The main challenge in designing a decentralized gradient method for solving the DBO (6) is how to compute the hypergradient. Note that, the hypergradient  $\nabla \Phi(x)$  of (6) is given by

$$\begin{aligned} \nabla \Phi(x) &= \frac{1}{n} \sum_{i=1}^n \nabla_1 F_i(x, y^*(x)) \\ &\quad - \left[ \frac{1}{n} \sum_{i=1}^n \nabla_{12}^2 f_i(x, y^*(x)) \right] \left[ \frac{1}{n} \sum_{i=1}^n \nabla_{22}^2 f_i(x, y^*(x)) \right]^{-1} \frac{1}{n} \sum_{i=1}^n \nabla_2 F_i(x, y^*(x)). \end{aligned} \quad (7)$$

Calculation of the hypergradient (7) is not possible through a single agent, and instead requires cooperative computation among all agents through communication. The first algorithm for solving DBO (6) was due to [5], where the authors proposed the DSBO algorithm that incorporates a decentralized algorithm to solve the linear system of equations in (7). The per-iteration complexity of the DSBO algorithm was later improved by the same authors by employing the moving average technique [6]. In [24], the authors proposed a stochastic linearized augmented Lagrangian method

(SLAM) for solving DBO (6). Another type of distributed BO, Federated BO, has also been studied in the literature. For example, Yang et al. [35] proposed the SimFBO algorithm for solving (6) but in a federated network. All these algorithms for decentralized BO and federated BO require certain gradient heterogeneity in order to guarantee the convergence or lower per-iteration complexity. However, this kind of assumption is very strong and may not hold in certain scenarios (see, e.g., [28]). We list below the heterogeneity assumptions in these papers.

- (DSBO, [5, Assumption 2.4]) Assume the data associated with  $f_i$  is independent and identically distributed,  $i = 1, \dots, n$ .
- (MA-DSBO, [6, Assumption 2.3]) For all  $i$ , there exists a constant  $\delta \geq 0$  such that

$$\left\| \nabla_2 f_i(x, y) - \frac{1}{n} \sum_{j=1}^n \nabla_2 f_j(x, y) \right\| \leq \delta, \quad \forall x, y.$$

- (SLAM, [24, Theorem 1]) For all  $i$ , there exists a constant  $L \geq 0$  such that

$$\left\| \nabla_{22}^2 f_i(x, y) - \frac{1}{n} \sum_{j=1}^n \nabla_{22}^2 f_j(x, y') \right\| \leq L \|y - y'\|, \quad \forall x, y, y'.$$

It should be noted that if  $y = y'$ , this assumption becomes  $\nabla_{22}^2 f_i(x, y) = \frac{1}{n} \sum_{j=1}^n \nabla_{22}^2 f_j(x, y)$ .

- (SimFBO, [35, Assumption 4]) There exist constants  $\delta_1 \geq 1$  and  $\delta_2 \geq 0$  such that

$$\frac{1}{n} \sum_{i=1}^n \|\nabla_2 f_i(x, y)\|^2 \leq \delta_1^2 \left\| \frac{1}{n} \sum_{i=1}^n \nabla_2 f_i(x, y) \right\|^2 + \delta_2^2, \quad \forall x, y.$$

These assumptions indicate the level of similarity between the local and the global objective functions. Our proposed algorithm, however, with the help of gradient tracking and projection technique, does not need any heterogeneity assumptions like these. Here we emphasize that in this paper we focus on deterministic DBO, because it is already very challenging. Extending the algorithms in this paper to stochastic DBO is definitely another very important task, and we leave it to a future work.

**Main contributions.** Our contributions in this work lie in several folds. First, we propose a single-loop algorithm for DBO, which has two main features: (i) it is of a single-loop structure; and (ii) it only needs two matrix-vector multiplications in each iteration. Second, we provide a convergence rate analysis for the proposed algorithm in the absence of any heterogeneity assumptions. This is in sharp contrast to existing works on decentralized BO and federated BO. Third, we demonstrate the great potential of our algorithm through numerical experiments on hyperparameter optimization.

**Notation.** We denote the optimal value of (6) as  $F^*$ . The gradients of  $f$  with respect to  $x$  and  $y$  are denoted as  $\nabla_1 f(x, y)$  and  $\nabla_2 f(x, y)$  respectively, while the Jacobian matrix of  $\nabla_1 f$  and Hessian matrix of  $f$  with respect to  $y$  are denoted as  $\nabla_{12}^2 f(x, y)$  and  $\nabla_{22}^2 f(x, y)$  respectively. If there is no further specification, it is assumed that  $\|\cdot\|$  denotes the  $\ell_2$  norm for vectors and the Frobenius norm for matrices. The operator norm of a matrix  $Z$  is denoted by  $\|Z\|_{\text{op}}$ .

## 2 A Single-Loop Algorithm for Decentralized Bilevel Optimization

In this section, we propose our single-loop algorithm for DBO (SLDBO). Its convergence results are given in Section 3. Before presenting the algorithm, we specify our assumptions.

## 2.1 Assumptions

Throughout this paper, we adopt the following standard assumptions, which are commonly used in existing literature on bilevel optimization and decentralized optimization. For instance, Assumption 2.1 has been employed in previous works such as [12, 18, 5, 6, 17], and Assumption 2.2 has been utilized in [29, 25, 7, 5, 6, 24].

**Assumption 2.1.** *The following assumptions hold for functions  $F_i$  and  $f_i$ ,  $i = 1, \dots, n$ , in (6).*

- (a) *For any fixed  $x$ ,  $f_i(x, \cdot)$  is  $\sigma$ -strongly convex, with  $\sigma > 0$  being a constant.*
- (b) *The function  $F_i$  is Lipschitz continuous with a Lipschitz constant of  $L_{F,0}$ , and its gradient  $\nabla F_i$  is also Lipschitz continuous with a Lipschitz constant of  $L_{F,1}$ .*
- (c) *The function  $f_i$  is twice differentiable, with its gradient  $\nabla f_i$  being Lipschitz continuous and having a Lipschitz constant of  $L_{f,1}$ . Moreover, the Hessian of  $f_i$ , denoted by  $\nabla^2 f_i$ , is also Lipschitz continuous with a Lipschitz constant of  $L_{f,2}$ .*

**Assumption 2.2** (Network topology). *Suppose the communication network is represented by a nonnegative weight matrix  $W = (w_{ij}) \in \mathbb{R}^{n \times n}$ , where  $w_{ij} = 0$  if  $i \neq j$  and agents  $i$  and  $j$  are not connected. Moreover, we assume that  $W$  is symmetric and doubly stochastic, i.e.  $W = W^\top$  and  $W\mathbf{1}_n = \mathbf{1}_n$ , where  $\mathbf{1}_n$  is the all-one vector in  $\mathbb{R}^n$ . Furthermore, the eigenvalues of  $W$  satisfy  $1 = \lambda_1 > \lambda_2 \geq \dots \geq \lambda_n$  and  $\rho := \max\{|\lambda_2|, |\lambda_n|\} < 1$ .*

## 2.2 The Proposed SLDBO Algorithm

Our goal is to extend the idea of SOBA [8] to handle the DBO (6), which presents a significant challenge, especially without any heterogeneity assumptions. To tackle this, we propose to project  $v^k$  onto a Euclidean ball with a pre-defined radius. By combining this with the gradient tracking technique, we successfully eliminate all heterogeneity assumptions.\* Before introducing our SLDBO algorithm, we define the following constants:

$$\begin{cases} r_v := L_{F,0}/\sigma, & L_v := (L_{F,1} + L_{f,2}r_v)(1 + L_{f,1}/\sigma), & L_1 := L_{F,1} + L_{f,2}r_v, \\ L_\Phi := L_{F,1} + \frac{2L_{F,1}L_{f,1} + L_{f,2}L_{F,0}^2}{\sigma} + \frac{2L_{f,1}L_{F,0}L_{f,2} + L_{f,1}^2L_{F,1}}{\sigma^2} + \frac{L_{f,2}L_{f,1}^2L_{F,0}}{\sigma^3}, \end{cases} \quad (8)$$

where constants such as  $\sigma$ ,  $L_{F,0}$ ,  $L_{F,1}$ ,  $L_{f,1}$ ,  $L_{f,2}$  are defined in Assumptions 2.1. We also define the projection operator  $\mathcal{P}_r$  as  $\mathcal{P}_r[z] := \arg \min_{\|z'\| \leq r} \|z' - z\|$ , which projects a given point onto the Euclidean ball with radius  $r \geq 0$ .

The details of our proposed algorithm, SLDBO, are presented in Algorithm 1, while the setup of its initial points is shown in BOX 1.

---

\*Here we point out that shortly after we released the first version of our paper on arXiv, another work [36] was uploaded to arxiv. The first version of [36] extended SOBA to solving the stochastic DBO, but it did not provide any proof for the theoretical results. The second version of [36] appeared on arxiv a few months later where the authors changed their algorithm in the first version to a different algorithm and a convergence analysis was provided for this new algorithm. We note that this new algorithm incorporated a projection step similar to the one used in our Algorithm 1, and this projection step was missing in the algorithm in the first version of [36]. This confirms the crucial role of our projection technique for a provably convergent algorithm for DBO.

- Initial points  $d_{x,i}^{-1} = d_{y,i}^{-1} = d_{v,i}^{-1} = t_{x,i}^{-1} = t_{y,i}^{-1} = t_{v,i}^{-1} = 0$  ( $i = 1, 2, \dots, n$ ).
- Initial points  $x_i^{-1} = x_j^{-1}$ ,  $y_i^{-1} = y_j^{-1}$ ,  $v_i^{-1} = v_j^{-1}$ ,  $i, j \in \{1, 2, \dots, n\}$ , satisfying  $\|v_i^{-1}\| \leq r_v$ , where  $r_v$  is defined in (8). Note that the inequality can easily be satisfied by choosing  $v_i^{-1} = 0$ , for instance.
- Compute initial points  $x_i^0 = \sum_{j=1}^n w_{ij}(x_j^{-1} - \alpha t_{x,j}^{-1}) = x_i^{-1}$ ,

$$y_i^0 = \sum_{j=1}^n w_{ij}(y_j^{-1} - \beta t_{y,j}^{-1}) = y_i^{-1} \quad \text{and} \quad v_i^0 = \mathcal{P}_{r_v} \left[ \sum_{j=1}^n w_{ij}(v_j^{-1} + \eta t_{v,j}^{-1}) \right] = v_i^{-1},$$

for  $i = 1, 2, \dots, n$ , where  $\alpha$ ,  $\beta$  and  $\eta$  are constant stepsizes.

BOX 1: Initialization of Algorithm 1.

---

**Algorithm 1** A Single-Loop Algorithm for DBO (SLDBO)

---

**Input:** Let  $K$  be the maximum iteration number and  $r_v$  be defined in (8). Set initial points as in BOX 1, as well as constant step sizes  $\alpha, \beta, \eta > 0$  satisfying the upper bounds in (62).

**for**  $k = 0, 1, \dots, K - 1$  **do**

**for**  $i = 1, \dots, n$  **do**

$$d_{y,i}^k = \nabla_2 f_i(x_i^k, y_i^k); \tag{9}$$

$$d_{v,i}^k = \nabla_2 F_i(x_i^k, y_i^k) - \nabla_{22}^2 f_i(x_i^k, y_i^k) v_i^k; \tag{10}$$

$$d_{x,i}^k = \nabla_1 F_i(x_i^k, y_i^k) - \nabla_{12}^2 f_i(x_i^k, y_i^k) v_i^k; \tag{11}$$

$$t_{y,i}^k = \sum_{j=1}^n w_{ij} t_{y,j}^{k-1} + d_{y,i}^k - d_{y,i}^{k-1}, \quad y_i^{k+1} = \sum_{j=1}^n w_{ij} (y_j^k - \beta t_{y,j}^k); \tag{12}$$

$$t_{v,i}^k = \sum_{j=1}^n w_{ij} t_{v,j}^{k-1} + d_{v,i}^k - d_{v,i}^{k-1}, \quad v_i^{k+1} = \mathcal{P}_{r_v} \left[ \sum_{j=1}^n w_{ij} (v_j^k + \eta t_{v,j}^k) \right]; \tag{13}$$

$$t_{x,i}^k = \sum_{j=1}^n w_{ij} t_{x,j}^{k-1} + d_{x,i}^k - d_{x,i}^{k-1}, \quad x_i^{k+1} = \sum_{j=1}^n w_{ij} (x_j^k - \alpha t_{x,j}^k). \tag{14}$$

**end for**

**end for**

---

**Remark 2.1.** Some remarks on the SLDBO (Algorithm 1) are in demand.

(i) The  $d_{y,i}^k$ ,  $d_{v,i}^k$  and  $d_{x,i}^k$  in (9)-(11) are the decentralized counterparts of  $D_y^k$ ,  $D_v^k$  and  $D_x^k$  in SOBA (5).

(ii) The updates for  $y_i^{k+1}$ ,  $v_i^{k+1}$ , and  $x_i^{k+1}$  outlined in (12)-(14) are based on similar ideas as in SOBA (5). However, since we are now addressing the decentralized problem (6), we require communication steps using the communication matrix  $W = (w_{ij})$ . As it is introduced in Assumption 2.2, if  $w_{ij} > 0$ , agent  $j$  communicates with agent  $i$ , and the data communicated is multiplied by the weight  $w_{ij}$ . If  $w_{ij} = 0$ , then there is no communication between the two agents. In this regard, we employ the adapt-then-combine diffusion strategy, which has been demonstrated to perform better in practical numerical experiments [31], rather than the combine-then-adapt diffusion strategy.

(iii) The updates for  $t_{y,i}^k$ ,  $t_{v,i}^k$ , and  $t_{x,i}^k$  described in (12)-(14) rely on the gradient tracking technique, a commonly employed strategy in distributed optimization literature [23, 29, 25] and DBO literature [5, 6, 11]. Tracking sequences have been employed to eliminate heterogeneity assumptions for single-level problems. However, for DBO problems, an additional projection step is required in (13) to accomplish this task. Specifically, the projection step ensures that the sequence  $\{v_i^k\}$  remains bounded, which is crucial for bounding the terms  $d_{x,i}^{k+1} - d_{x,i}^k$  and  $d_{v,i}^{k+1} - d_{v,i}^k$  in gradient tracking steps. We refer to Lemma A.5 in the Appendix for more discussions on this point.

### 3 Convergence Rate Results for SLDBO

In this section, we provide the convergence results for Algorithm 1. Firstly, we construct a Lyapunov function, with the terms related to the consensus error. Then, by suitably selecting the parameters, we establish a descent property of the Lyapunov function, which further results in the  $O(1/K)$  convergence rate for some stationarity measure. Our primary convergence rate results for Algorithm 1 are summarized in Theorem 3.1.

**Theorem 3.1.** *For any integer  $K \geq 1$ , when  $0 \leq k \leq K$ , define  $\bar{x}^k = \frac{1}{n} \sum_{i=1}^n x_i^k$ ,  $\bar{y}^k = \frac{1}{n} \sum_{i=1}^n y_i^k$  and  $\bar{v}^k = \frac{1}{n} \sum_{i=1}^n v_i^k$ . The following convergence rate results hold for Algorithm 1.*

- (a) **Stationarity.** *For any integer  $K \geq 1$ , there holds  $\min_{0 \leq k \leq K-1} \|\nabla \Phi(\bar{x}^k)\|^2 = O(1/K)$ .*
- (b) **Consensus Error.** *For any integer  $K \geq 1$ , we have  $\min_{0 \leq k \leq K-1} \frac{1}{n} \sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 = O(1/K)$ ,  $\min_{0 \leq k \leq K-1} \frac{1}{n} \sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 = O(1/K)$  and  $\min_{0 \leq k \leq K-1} \frac{1}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 = O(1/K)$ .*

As per Part (a) of Theorem 3.1, the convergence rate for stationarity is sublinear, with a rate of  $O(1/K)$ . This result is in line with the findings presented in [17]. As far as we know, this is the currently best-known convergence rate result for both BO and DBO algorithms. Furthermore, in each iteration, our algorithm only requires  $\Theta(1)$  communication rounds and computational complexity, while other algorithms, such as those in [6, 24], involve computation of  $\Theta(\log K)$  matrix-vector multiplications per iteration. This makes our algorithm computationally efficient and well-suited for large-scale DBO problems.

**Roadmap of the Proof.** Here we briefly describe the roadmap of the proof of Theorem 3.1 and the details are postponed to the Appendix. First, we define the Lyapunov function

$$\begin{aligned} V_k := & F(\bar{x}^k, y^*(\bar{x}^k)) - F^* + a_1 \|\bar{y}^k - y^*(\bar{x}^k)\|^2 + a_2 \|\bar{v}^k - v^*(\bar{x}^k)\|^2 \\ & + \frac{a_3}{n} \sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 + \frac{a_4}{n} \sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 + \frac{a_5}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 \\ & + \frac{a_6 \alpha^2}{n} \sum_{i=1}^n \|t_{x,i}^k - \bar{t}_x^k\|^2 + \frac{a_7 \beta^2}{n} \sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2 + \frac{a_8 \eta^2}{n} \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2, \end{aligned} \quad (15)$$

where we define  $v^*(x) := [\nabla_{22}^2 f(x, y^*(x))]^{-1} \nabla_2 F(x, y^*(x))$  and  $a_1, a_2, \dots, a_8$  are constants. We will show that, for another set of positive constants  $A_1, A_2, \dots, A_9$ , the following inequality holds:

$$\begin{aligned} V_{k+1} - V_k \leq & -\frac{\alpha}{2} \|\nabla \Phi(\bar{x}^k)\|^2 - \frac{A_1}{\alpha^2} \|\bar{x}^{k+1} - \bar{x}^k\|^2 - A_2 \|\bar{y}^k - y^*(\bar{x}^k)\|^2 - A_3 \|\bar{v}^k - v^*(\bar{x}^k)\|^2 \\ & - \frac{A_4}{n} \sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 - \frac{A_5}{n} \sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 - \frac{A_6}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 \\ & - \frac{A_7}{n} \sum_{i=1}^n \|t_{x,i}^k - \bar{t}_x^k\|^2 - \frac{A_8}{n} \sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2 - \frac{A_9}{n} \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2. \end{aligned}$$



By taking the telescoping sum of this inequality, we can complete the proof. The construction of the Lyapunov function (15) and part of our proof were inspired by [22].

## 4 Numerical Experiments

In this section, we conduct experiments on hyperparameter optimization to evaluate the effectiveness of our proposed SLDBO. To test the decentralized setting, we used a local device equipped with 8 cores, i.e.,  $n = 8$ , and employed mpi4py [9] for parallel computing.

We adopted a ring topology to model the network for distributed computation, represented by a weight matrix  $W = (w_{ij}) \in \mathbb{R}^{n \times n}$  given by: for  $i, j = 1, \dots, n$ ,  $w_{ij} = w$  if  $i = j$ ,  $w_{ij} = (1 - w)/2$  if  $i = j \pm 1$  or  $(i, j) \in \{(1, n), (n, 1)\}$ , and  $w_{ij} = 0$  otherwise, where  $w \in (0, 1)$  is a constant. We take  $w = 0.4$  in our experiments. In this ring topology, each agent has exactly two neighbours. Our experiments involve both synthetic and real-world data.

### 4.1 Synthetic Data

We conduct logistic regression with  $\ell_2$  regularization. Let  $\psi(t) = \log(1 + e^{-t})$  for  $t \in \mathbb{R}$  and  $p$  be the dimension of the data. Following [6], on agent  $i$ ,  $i = 1, \dots, 8$ , we have

$$F_i(\lambda, \omega) = \sum_{(x_e, y_e) \in \mathcal{D}'_i} \psi(y_e x_e^\top \omega),$$

$$f_i(\lambda, \omega) = \sum_{(x_e, y_e) \in \mathcal{D}_i} \psi(y_e x_e^\top \omega) + \frac{1}{2} \sum_{j=1}^p e^{\lambda_j} \omega_j^2,$$

where  $\mathcal{D}_i$  and  $\mathcal{D}'_i$  denote the training and testing datasets on agent  $i$ , respectively. We aim to identify the optimal hyperparameter  $\lambda$  such that  $\omega^*(\lambda)$  represents the optimal model parameter corresponding to  $\lambda$ . To achieve this, we utilize synthetic heterogeneous data, generated in the same manner as in [6]. Specifically, the data distribution of  $x_e$  on agent  $i$  follows a normal distribution with mean 0 and variance  $i^2 \cdot r^2$ , where  $r$  is the heterogeneity rate. For the response variable, we let  $y_e = x_e^\top \omega + 0.1z$ , where  $z$  is sampled from the standard normal distribution.

Note that the parameter  $r$  controls the data heterogeneity rate. We first compare SLDBO with MA-DSBO (Moving Average-DSBO) [6] under conditions of low data heterogeneity with  $r = 1$  to investigate the advantages of the single-loop structure of SLDBO. Full gradients are calculated in both algorithms, and we use a training dataset and a testing dataset consisting of 20,000 samples. In SLDBO, we set  $r_v = 2$ ,  $\alpha = \eta = 0.025$  and  $\beta = 0.06$ . The results under different data dimensions are shown in Figures 1-2. MA-DSBO employs two key parameters, where  $T$  represents the number of iterations performed in the inner loop, and  $N$  represents the number of Hessian-inverse-gradient product iterations. Comparing SLDBO and MA-DSBO, it is evident that SLDBO is faster. In particular, SLDBO is much faster than MA-DSBO with  $T = N = 2$ . Note that MA-DSBO requires sufficient inner-loop iterations to accurately estimate the LL solution and the hypergradient. By comparing the results in Figure 1 ( $p = 50$ ) with those in Figure 2 ( $p = 200$ ), we can observe that SLDBO demonstrates a greater improvement in convergence rate as the dimension of the data increases. SLDBO's single-loop structure offers an advantage in terms of reducing the number of required matrix-vector products, which proves to be particularly beneficial when dealing with high-dimensional data.

Subsequently, we set different heterogeneity rates to highlight the exceptional performance of SLDBO under conditions of high data heterogeneity, as well as the necessity of the projection



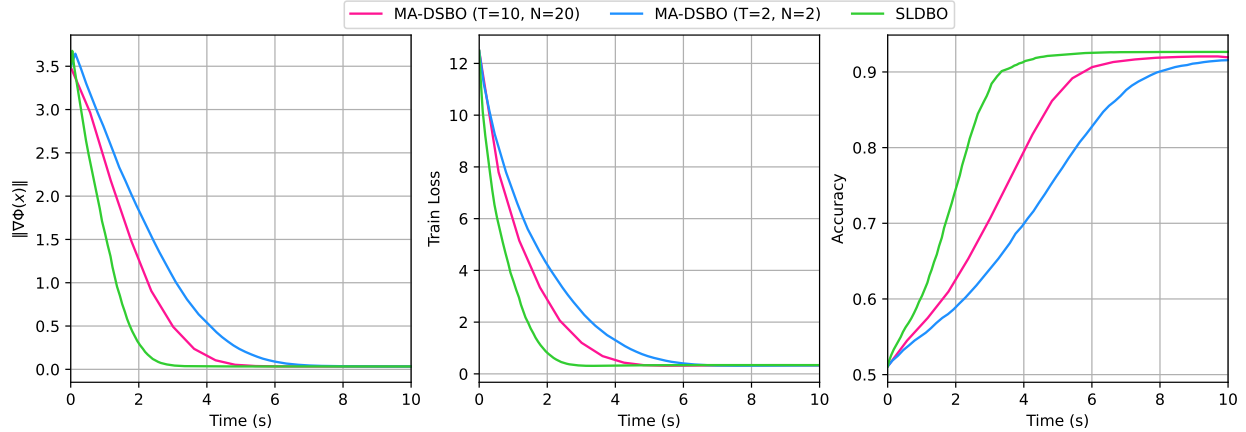


Figure 1: Comparison between MA-DSBO and SLDBO on synthetic data ( $p = 50$ ).

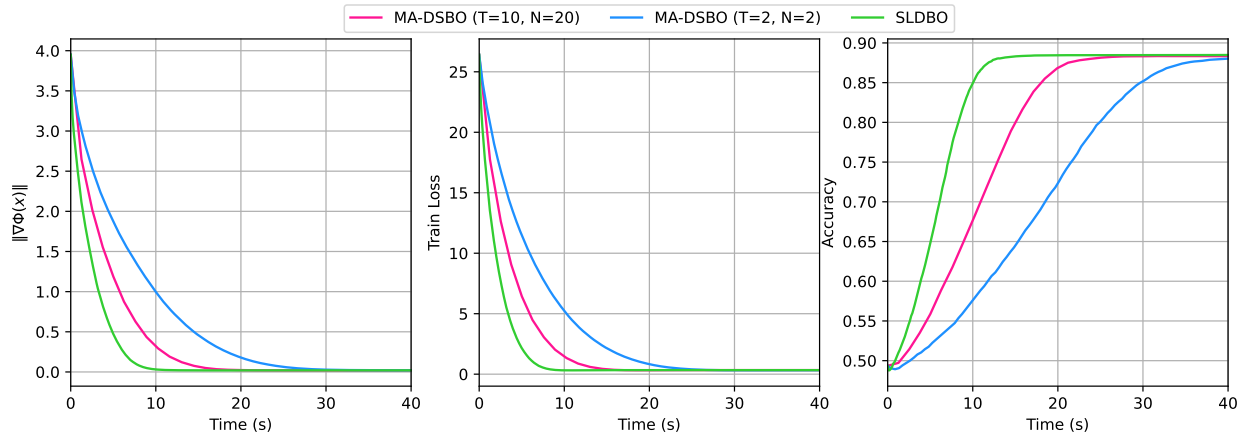


Figure 2: Comparison between MA-DSBO and SLDBO on synthetic data ( $p = 200$ ).

technique in SLDBO. The experimental results are illustrated in Figure 3. SLDBO (w/o proj.) represents SLDBO without the projection step, this means replacing (13) with

$$t_{v,i}^k = \sum_{j=1}^n w_{ij} t_{v,j}^{k-1} + d_{v,i}^k - d_{v,i}^{k-1}, \quad v_i^{k+1} = \sum_{j=1}^n w_{ij} (v_j^k + \eta t_{v,j}^k).$$

As shown in the left plot of Figure 3, when data heterogeneity is low ( $r = 1$ ), SLDBO with or without projection performs similarly, indicating that the projection step is almost ineffective. However, it is noteworthy that at a high level of data heterogeneity ( $r = 40$ , the right plot of Figure 3), SLDBO without projection and MA-DSBO only achieve an accuracy of 0.5, while SLDBO (with projection) achieves an accuracy of about 0.96. This demonstrates the critical role of the projection technique in our proposed algorithm.

## 4.2 Real-World Data

Similar to the case of synthetic data, we define  $\mathcal{D}_i$  and  $\mathcal{D}'_i$  as the training and testing datasets, respectively, for agent  $i$ . We next apply SLDBO to solve the following hyperparameter problem

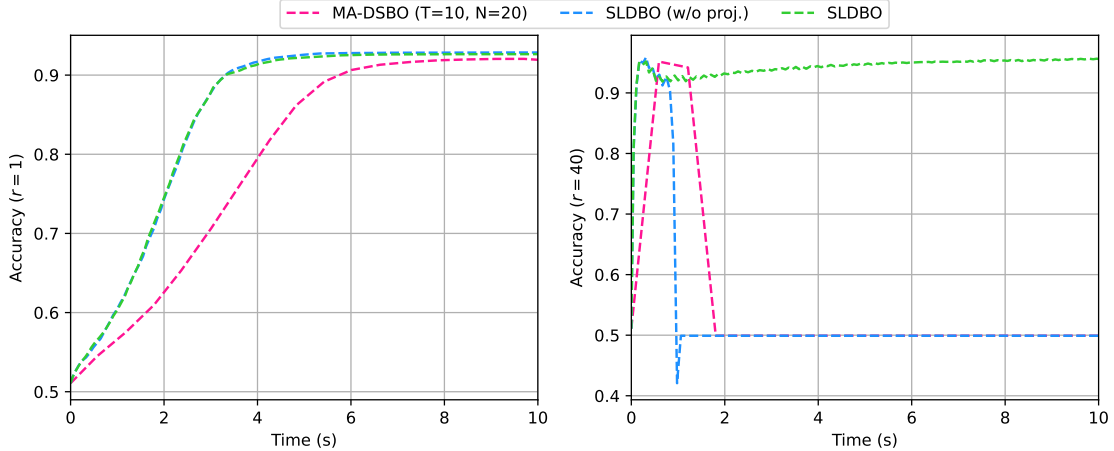


Figure 3: Comparison between MA-DSBO, SLDBO (w/o proj.) and SLDBO on synthetic data. Dimension:  $p = 50$ . Heterogeneity rate:  $r = 1$  (left),  $r = 40$  (right).

using the MNIST database [20]:

$$F_i(\lambda, \omega) = \frac{1}{|\mathcal{D}'_i|} \sum_{(x_e, y_e) \in \mathcal{D}'_i} L(x_e^\top \omega, y_e),$$

$$f_i(\lambda, \omega) = \frac{1}{|\mathcal{D}_i|} \sum_{(x_e, y_e) \in \mathcal{D}_i} L(x_e^\top \omega, y_e) + \frac{1}{cp} \sum_{i=1}^c \sum_{j=1}^p e^{\lambda_j} \omega_{ij}^2,$$

where  $\omega \in \mathbb{R}^{c \times p}$  denotes the model parameter,  $L$  denotes the cross entropy loss, and  $|S|$  denotes the cardinality of a set  $S$ . In our experiments, we set  $c$  and  $p$  to be 10 and 784, respectively. Here  $c$  represents the number of classes and  $p$  represents the number of features. The training and testing sets comprise 60,000 samples each, with balanced representation across all classes. To reduce the computational overhead associated with estimating gradients from a large dataset in our SLDBO algorithm, we adopt a technique inspired by stochastic gradient descent. Specifically, we extract a representative subset of samples to estimate the gradients instead of using the entire dataset. For both the SLDBO and MA-DSBO algorithms, we set the batch size on each computing agent to 1,000. In the case of SLDBO, the hyperparameters were set as follows:  $r_v = 10$ ,  $\alpha = \eta = 0.024$ , and  $\beta = 0.06$ . For MA-DSBO, we set  $T = N = 5$ . Figure 4 presents a comparison of the test loss, train loss, and classification accuracy between the SLDBO and MA-DSBO algorithms. These results demonstrate that our proposed algorithm SLDBO can efficiently solve this problem with improved convergence rate and classification accuracy.

## 5 Concluding Remarks

This paper presents a novel single-loop algorithm, called SLDBO, for efficiently solving DBO problems with a guaranteed sublinear convergence rate. Notably, SLDBO is the first single-loop algorithm for DBO that operates without nested matrix-vector products (only two matrix-vector products are required at each iteration) and does not make any assumptions related to data heterogeneity. Our numerical experiments confirm the effectiveness of SLDBO. Nevertheless, we acknowledge that computing the full gradient during practical applications can be time-consuming. Therefore, it is worth investigating the extension of our algorithm to the stochastic setting. Additionally,

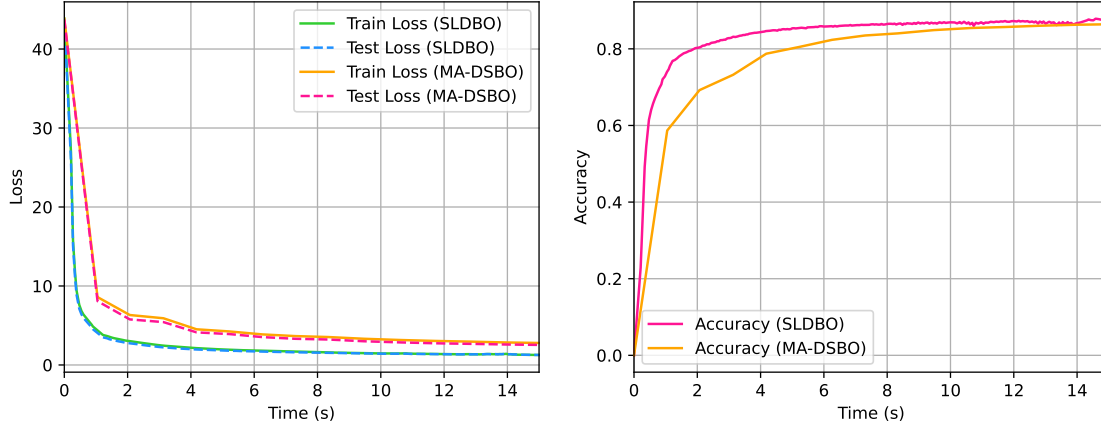


Figure 4: Comparison of test loss, train loss, and classification accuracy between MA-DSBO and SLDBO on real-world MNIST dataset.

techniques for saving communications, such as those used in E-AiPOD [34], could be integrated into our algorithm for decentralized settings. These topics are left for further investigation.

## References

- [1] A. Beck. *First-order methods in optimization*, volume 25. SIAM, 2017.
- [2] N. Bishop, L. Tran-Thanh, and E. Gerding. Optimal learning from verified training data. In *NeurIPS*, volume 33, pages 9520–9529, 2020.
- [3] L. Chen, Y. Ma, and J. Zhang. Near-optimal fully first-order algorithms for finding stationary points in bilevel optimization. *arXiv:2306.14853*, 2023.
- [4] T. Chen, Y. Sun, and W. Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *NeurIPS*, volume 34, pages 25294–25307, 2021.
- [5] X. Chen, M. Huang, and S. Ma. Decentralized bilevel optimization. *Optimization Letters*, 2024.
- [6] X. Chen, M. Huang, S. Ma, and K. Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *ICML*, volume 202, pages 4641–4671, 2023.
- [7] W. Choi and J. Kim. On the convergence analysis of the decentralized projected gradient descent. *arXiv:2303.08412*, 2023.
- [8] M. Dagr eou, P. Ablin, S. Vaiter, and T. Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *NeurIPS*, volume 35, pages 26698–26710, 2022.
- [9] L. Dalcin and Y.-L. L. Fang. mpi4py: Status update after 12 years of development. *Computing in Science & Engineering*, 23(4):47–54, 2021.
- [10] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *ICML*, volume 80, pages 1568–1577, 2018.
- [11] H. Gao, B. Gu, and M. T. Thai. On the convergence of distributed stochastic bilevel optimization algorithms over a network. In *AISTATS*, volume 206, pages 9238–9281, 2023.
- [12] S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *arXiv:1802.02246*, 2018.
- [13] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. In *ICML*, volume 119, pages 3748–3758, 2020.

- [14] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.
- [15] X. Hu, N. Xiao, X. Liu, and K.-C. Toh. An improved unconstrained approach for bilevel optimization. *SIAM Journal on Optimization*, 33(4):2801–2829, 2023.
- [16] K. Ji, J. D. Lee, Y. Liang, and H. V. Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. In *NeurIPS*, volume 33, pages 11490–11500, 2020.
- [17] K. Ji, M. Liu, Y. Liang, and L. Ying. Will bilevel optimizers benefit from loops. In *NeurIPS*, volume 35, pages 3011–3023, 2022.
- [18] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *ICML*, volume 139, pages 4882–4892, 2021.
- [19] J. Kwon, D. Kwon, S. Wright, and R. D. Nowak. A fully first-order method for stochastic bilevel optimization. In *ICML*, volume 202, pages 18083–18113, 2023.
- [20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [21] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *NeurIPS*, volume 30, page 5336–5346, 2017.
- [22] R. Liu, Y. Liu, W. Yao, S. Zeng, and J. Zhang. Averaged method of multipliers for bi-level optimization without lower-level strong convexity. In *ICML*, volume 202, pages 21839–21866, 2023.
- [23] P. D. Lorenzo and G. Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.
- [24] S. Lu, S. Zeng, X. Cui, M. Squillante, L. Horesh, B. Kingsbury, J. Liu, and M. Hong. A stochastic linearized augmented lagrangian method for decentralized bilevel optimization. In *NeurIPS*, volume 35, pages 30638–30650, 2022.
- [25] A. Nedić, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM J. Optim.*, 27(4):2597–2633, 2017.
- [26] Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, Cham, 2018.
- [27] F. Pedregosa. Hyperparameter optimization with approximate gradient. In *ICML*, volume 48, pages 737–746, 2016.
- [28] S. Pu and A. Nedić. Distributed stochastic gradient tracking methods. *Math. Program.*, 187(1-2):409–457, 2021.
- [29] G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Trans. Control Netw. Syst.*, 5(3):1245–1260, 2018.
- [30] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. In *NeurIPS*, volume 32, page 113–124, 2019.
- [31] A. H. Sayed. Adaptation, learning, and optimization over networks. *Foundations and Trends® in Machine Learning*, 7(4-5):311–801, 2014.
- [32] A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots. Truncated back-propagation for bilevel optimization. In *AISTATS*, volume 89, pages 1723–1732, 2019.
- [33] J. Wang, H. Chen, R. Jiang, X. Li, and Z. Li. Fast algorithms for Stackelberg prediction game with least squares loss. In *ICML*, volume 139, pages 10708–10716, 2021.
- [34] Q. Xiao, H. Shen, W. Yin, and T. Chen. Alternating projected SGD for equality-constrained bilevel optimization. In *AISTATS*, volume 206, pages 987–1023, 2023.

- [35] Y. Yang, P. Xiao, and K. Ji. SimFBO: Towards simple, flexible and communication-efficient federated bilevel learning. *arXiv:2305.19442*, 2023.
- [36] Y. Zhang, M. T. Thai, J. Wu, and H. Gao. On the communication complexity of decentralized bilevel optimization. *arXiv:2311.11342*, 2023.
- [37] Y. Zhang, G. Zhang, P. Khanduri, M. Hong, S. Chang, and S. Liu. Revisiting and advancing fast adversarial training through the lens of bi-level optimization. In *ICML*, volume 162, pages 26693–26712, 2022.

## A Proof of the Convergence Results

In this appendix, we provide the proof of our convergence results, i.e., Theorem 3.1. Assumption 2.1 and 2.2 are used throughout the proof.

### A.1 Notation, Constants and Basic Lemmas

For the ease of presentation, we define some notation below.

$$\begin{cases} \bar{x}^k := \frac{1}{n} \sum_{i=1}^n x_i^k, & \bar{y}^k := \frac{1}{n} \sum_{i=1}^n y_i^k, & \bar{v}^k := \frac{1}{n} \sum_{i=1}^n v_i^k, \\ \bar{d}_x^k := \frac{1}{n} \sum_{i=1}^n d_{x,i}^k, & \bar{d}_y^k := \frac{1}{n} \sum_{i=1}^n d_{y,i}^k, & \bar{d}_v^k := \frac{1}{n} \sum_{i=1}^n d_{v,i}^k, \\ \bar{t}_x^k := \frac{1}{n} \sum_{i=1}^n t_{x,i}^k, & \bar{t}_y^k := \frac{1}{n} \sum_{i=1}^n t_{y,i}^k, & \bar{t}_v^k := \frac{1}{n} \sum_{i=1}^n t_{v,i}^k. \end{cases} \quad (16)$$

Recall that  $r_v$  is defined in (8). To assist in our proof, we now define some additional constants as follows:

$$C_1 := 3 \max\{(L_{F,1} + r_v L_{f,2})^2, L_{f,1}^2\}, \quad C_2 := L_{f,1}^2. \quad (17)$$

Before the proof, we present some useful lemmas. Among others, Lemma A.1 is a common result in decentralized optimization, see, e.g., [28, Lemma 1]. For completeness, we provide the proof.

**Lemma A.1.** *Consider the mixing matrix  $W = (w_{ij}) \in \mathbb{R}^{n \times n}$  defined in Assumption 2.2, for any  $x_1, \dots, x_n \in \mathbb{R}^d$ , let  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , we have*

- (a)  $\sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} x_j \right\|^2 \leq \sum_{j=1}^n \|x_j\|^2$ .
- (b)  $\sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} (x_j - \bar{x}) \right\|^2 \leq \rho^2 \sum_{i=1}^n \|x_i - \bar{x}\|^2$ .

*Proof.* (a).  $\sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} x_j \right\|^2 \leq \sum_{i=1}^n \sum_{j=1}^n w_{ij} \|x_j\|^2 = \sum_{j=1}^n \sum_{i=1}^n w_{ij} \|x_j\|^2 = \sum_{j=1}^n \|x_j\|^2$ , where the inequality follows from the convexity of  $\|\cdot\|^2$  and the last equality is due to  $\sum_{i=1}^n w_{ij} = 1$ .

(b). Define  $X = [x_1^\top; x_2^\top; \dots; x_n^\top] \in \mathbb{R}^{n \times d}$ . Then, we have

$$\left\| WX - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top X \right\| \leq \left\| W - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \right\|_{\text{op}} \left\| X - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top X \right\|, \quad (18)$$

where the inequality holds because  $(W - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top) \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top X = 0$  by Assumption 2.2. Since  $W$  and  $\frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top$  are both symmetric and commute with each other, they are simultaneously diagonalizable, i.e., there exists an orthonormal matrix  $P$  such that

$$W = P \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) P^{-1} \quad \text{and} \quad \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top = P \text{diag}(1, 0, \dots, 0) P^{-1},$$

where the eigenvalues of  $W$  satisfy  $1 = \lambda_1 > \lambda_2 \geq \dots \geq \lambda_n$  and  $\rho = \max\{|\lambda_2|, |\lambda_n|\} < 1$ , as stated in Assumption 2.2. Hence, it is easy to verify that  $\|W - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\|_{\text{op}} = \|\text{diag}(0, \lambda_2, \dots, \lambda_n)\|_{\text{op}} = \rho$ . Moreover, we have

$$\left\|WX - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top X\right\|^2 = \left\|WX - \mathbf{1}_n\bar{x}^\top\right\|^2 = \sum_{i=1}^n \left\|\sum_{j=1}^n w_{ij}(x_j - \bar{x})\right\|^2 \quad (19)$$

and

$$\left\|X - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top X\right\|^2 = \left\|X - \mathbf{1}_n\bar{x}^\top\right\|^2 = \sum_{i=1}^n \|x_i - \bar{x}\|^2. \quad (20)$$

The desired result follows by combining (18)-(20) and  $\|W - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\|_{\text{op}} = \rho$ .  $\square$

Lemma A.2 is adopted from Lemma 3.2 in [7], and it describes the relationship between consensus error before and after projection. Recall that  $\mathcal{P}_r[z] := \text{argmin}_{\{z': \|z' - z\| \in [0, r]\}} \|z' - z\|$ .

**Lemma A.2.** *For any  $x_1, \dots, x_n \in \mathbb{R}^d$ , we have*

$$\sum_{i=1}^n \left\|\mathcal{P}_r[x_i] - \frac{1}{n}\sum_{j=1}^n \mathcal{P}_r[x_j]\right\|^2 \leq \sum_{i=1}^n \left\|x_i - \frac{1}{n}\sum_{j=1}^n x_j\right\|^2.$$

Consider the function  $G(x) = \sum_{i=1}^n \|\mathcal{P}_r[x_i] - x\|^2$ , which is minimized by  $x = \frac{1}{n}\sum_{i=1}^n \mathcal{P}_r[x_i]$ . Therefore, we have

$$\sum_{i=1}^n \left\|\mathcal{P}_r[x_i] - \frac{1}{n}\sum_{j=1}^n \mathcal{P}_r[x_j]\right\|^2 \leq \sum_{i=1}^n \left\|\mathcal{P}_r[x_i] - \mathcal{P}_r\left[\frac{1}{n}\sum_{j=1}^n x_j\right]\right\|^2 \leq \sum_{i=1}^n \left\|x_i - \frac{1}{n}\sum_{j=1}^n x_j\right\|^2,$$

where the last inequality holds because the projection operator is non-expansive.

## A.2 Consensus Error of Algorithm 1

Recall that  $\bar{x}^k, \bar{y}^k, \bar{v}^k, \bar{d}_x^k, \bar{d}_y^k, \bar{d}_v^k, \bar{t}_x^k, \bar{t}_y^k$  and  $\bar{t}_v^k$  are defined in (16). In this section, we bound the terms related to the consensus error in (15). We first prove some useful lemmas.

**Lemma A.3.** *The sequence  $\{v_i^k\}$  generated by Algorithm 1 satisfies*

$$\sum_{i=1}^n \|v_i^{k+1} - \bar{v}^{k+1}\|^2 \leq \rho \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 + \frac{\rho^2 \eta^2}{1 - \rho} \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2. \quad (21)$$

*Proof.* In fact, we have

$$\begin{aligned} \sum_{i=1}^n \|v_i^{k+1} - \bar{v}^{k+1}\|^2 &= \sum_{i=1}^n \left\| \mathcal{P}_{r_v} \left[ \sum_{j=1}^n w_{ij}(v_j^k + \eta t_{v,j}^k) \right] - \frac{1}{n} \sum_{s=1}^n \mathcal{P}_{r_v} \left[ \sum_{j=1}^n w_{sj}(v_j^k + \eta t_{v,j}^k) \right] \right\|^2 \\ &\leq \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij}(v_j^k + \eta t_{v,j}^k) - \frac{1}{n} \sum_{s=1}^n \sum_{j=1}^n w_{sj}(v_j^k + \eta t_{v,j}^k) \right\|^2 \\ &\leq \left(1 + \frac{1 - \rho}{\rho}\right) \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij}v_j^k - \bar{v}^k \right\|^2 + \left(1 + \frac{\rho}{1 - \rho}\right) \eta^2 \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij}t_{v,j}^k - \bar{t}_v^k \right\|^2 \\ &\leq \rho \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 + \frac{\rho^2 \eta^2}{1 - \rho} \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2, \end{aligned}$$

where the first inequality follows from Lemma A.2, the second follows from Cauchy-Schwartz inequality and the notation defined in (16), the third follows from Lemma A.1.  $\square$

**Remark A.1.** By following similar deductions as in Lemma A.3, we can derive (details are omitted to avoid redundancy)

$$\sum_{i=1}^n \|x_i^{k+1} - \bar{x}^{k+1}\|^2 \leq \rho \sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 + \frac{\rho^2 \alpha^2}{1-\rho} \sum_{i=1}^n \|t_{x,i}^k - \bar{t}_x^k\|^2, \quad (22)$$

$$\sum_{i=1}^n \|y_i^{k+1} - \bar{y}^{k+1}\|^2 \leq \rho \sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 + \frac{\rho^2 \beta^2}{1-\rho} \sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2. \quad (23)$$

**Lemma A.4.** The sequences  $\{t_{x,i}^k\}$  and  $\{d_{x,i}^k\}$  generated by Algorithm 1 satisfy

$$\sum_{i=1}^n \|t_{x,i}^{k+1} - \bar{t}_x^{k+1}\|^2 \leq \rho \sum_{i=1}^n \|t_{x,i}^k - \bar{t}_x^k\|^2 + \frac{1}{1-\rho} \sum_{i=1}^n \|d_{x,i}^{k+1} - d_{x,i}^k\|^2. \quad (24)$$

*Proof.* From the update of  $x_i^k$  and  $t_{x,i}^k$  in (14), we have  $\bar{t}_x^k = \bar{t}_x^{k-1} + \bar{d}_x^k - \bar{d}_x^{k-1}$ ,  $\bar{t}_x^{-1} = \bar{d}_x^{-1} = 0$  and  $\bar{x}^{k+1} = \bar{x}^k - \alpha \bar{t}_x^k$ , which implies (by induction)

$$\bar{t}_x^k = \bar{d}_x^k \quad \text{and} \quad \bar{x}^{k+1} = \bar{x}^k - \alpha \bar{d}_x^k. \quad (25)$$

Therefore, we have

$$\begin{aligned} \sum_{i=1}^n \|t_{x,i}^{k+1} - \bar{t}_x^{k+1}\|^2 &= \sum_{i=1}^n \left\| \left( \sum_{j=1}^n w_{ij} t_{x,j}^k + d_{x,i}^{k+1} - d_{x,i}^k - \bar{t}_x^k \right) + (\bar{t}_x^k - \bar{t}_x^{k+1}) \right\|^2 \\ &= \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} t_{x,j}^k + d_{x,i}^{k+1} - d_{x,i}^k - \bar{t}_x^k \right\|^2 + 2 \sum_{i=1}^n \langle \bar{t}_x^{k+1} - \bar{t}_x^k, \bar{t}_x^k - \bar{t}_x^{k+1} \rangle + \sum_{i=1}^n \|\bar{t}_x^k - \bar{t}_x^{k+1}\|^2 \\ &= \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} t_{x,j}^k + d_{x,i}^{k+1} - d_{x,i}^k - \bar{t}_x^k \right\|^2 - n \|\bar{d}_x^k - \bar{d}_x^{k+1}\|^2 \\ &\leq \left(1 + \frac{1-\rho}{\rho}\right) \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} t_{x,j}^k - \bar{t}_x^k \right\|^2 + \left(1 + \frac{\rho}{1-\rho}\right) \sum_{i=1}^n \|d_{x,i}^{k+1} - d_{x,i}^k\|^2 - n \|\bar{d}_x^k - \bar{d}_x^{k+1}\|^2 \\ &\leq \rho \sum_{i=1}^n \|t_{x,i}^k - \bar{t}_x^k\|^2 + \frac{1}{1-\rho} \sum_{i=1}^n \|d_{x,i}^{k+1} - d_{x,i}^k\|^2, \end{aligned}$$

where the first inequality is due to Cauchy-Schwartz inequality, and the second inequality follows from Lemma A.1.  $\square$

**Remark A.2.** The following inequalities can be derived by following similar steps as in Lemma A.4:

$$\sum_{i=1}^n \|t_{v,i}^{k+1} - \bar{t}_v^{k+1}\|^2 \leq \rho \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2 + \frac{1}{1-\rho} \sum_{i=1}^n \|d_{v,i}^{k+1} - d_{v,i}^k\|^2, \quad (26)$$

$$\sum_{i=1}^n \|t_{y,i}^{k+1} - \bar{t}_y^{k+1}\|^2 \leq \rho \sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2 + \frac{1}{1-\rho} \sum_{i=1}^n \|d_{y,i}^{k+1} - d_{y,i}^k\|^2. \quad (27)$$

The details are omitted.



**Lemma A.5.** *The sequences  $\{d_{x,i}^k\}$ ,  $\{d_{v,i}^k\}$  and  $\{d_{y,i}^k\}$  generated by Algorithm 1 satisfy*

$$\sum_{i=1}^n \|d_{x,i}^k - d_{x,i}^{k-1}\|^2 \leq C_1 \sum_{i=1}^n \left( \|x_i^k - x_i^{k-1}\|^2 + \|y_i^k - y_i^{k-1}\|^2 + \|v_i^k - v_i^{k-1}\|^2 \right), \quad (28)$$

$$\sum_{i=1}^n \|d_{v,i}^k - d_{v,i}^{k-1}\|^2 \leq C_1 \sum_{i=1}^n \left( \|x_i^k - x_i^{k-1}\|^2 + \|y_i^k - y_i^{k-1}\|^2 + \|v_i^k - v_i^{k-1}\|^2 \right), \quad (29)$$

$$\sum_{i=1}^n \|d_{y,i}^k - d_{y,i}^{k-1}\|^2 \leq C_2 \sum_{i=1}^n \left( \|x_i^k - x_i^{k-1}\|^2 + \|y_i^k - y_i^{k-1}\|^2 \right), \quad (30)$$

where  $C_1$  and  $C_2$  are defined in (17).

*Proof.* By the definition of  $d_{x,i}^k$  in (11), it follows that

$$\begin{aligned} \sum_{i=1}^n \|d_{x,i}^k - d_{x,i}^{k-1}\|^2 &\leq 3 \sum_{i=1}^n \|\nabla_1 F_i(x_i^k, y_i^k) - \nabla_1 F_i(x_i^{k-1}, y_i^{k-1})\|^2 \\ &\quad + 3 \sum_{i=1}^n \|(\nabla_{12}^2 f_i(x_i^k, y_i^k) - \nabla_{12}^2 f_i(x_i^{k-1}, y_i^{k-1}))v_i^k\|^2 + 3 \sum_{i=1}^n \|\nabla_{12}^2 f_i(x_i^{k-1}, y_i^{k-1})(v_i^k - v_i^{k-1})\|^2 \\ &\leq 3(L_{F,1} + r_v L_{f,2})^2 \sum_{i=1}^n \|x_i^k - x_i^{k-1}\|^2 + 3(L_{F,1} + r_v L_{f,2})^2 \sum_{i=1}^n \|y_i^k - y_i^{k-1}\|^2 + 3L_{f,1}^2 \sum_{i=1}^n \|v_i^k - v_i^{k-1}\|^2 \\ &\leq C_1 \sum_{i=1}^n \left( \|x_i^k - x_i^{k-1}\|^2 + \|y_i^k - y_i^{k-1}\|^2 + \|v_i^k - v_i^{k-1}\|^2 \right). \end{aligned}$$

Similarly, by the definition of  $d_{v,i}^k$  in (10), we can also derive

$$\sum_{i=1}^n \|d_{v,i}^k - d_{v,i}^{k-1}\|^2 \leq C_1 \sum_{i=1}^n \left( \|x_i^k - x_i^{k-1}\|^2 + \|y_i^k - y_i^{k-1}\|^2 + \|v_i^k - v_i^{k-1}\|^2 \right).$$

Then by the definition of  $d_{y,i}^k$  in (9), we obtain

$$\begin{aligned} \sum_{i=1}^n \|d_{y,i}^k - d_{y,i}^{k-1}\|^2 &\leq \sum_{i=1}^n \|\nabla_2 f(x_i^k, y_i^k) - \nabla_2 f(x_i^{k-1}, y_i^{k-1})\|^2 \\ &\leq L_{f,1}^2 \sum_{i=1}^n \|x_i^k - x_i^{k-1}\|^2 + L_{f,1}^2 \sum_{i=1}^n \|y_i^k - y_i^{k-1}\|^2 = C_2 \sum_{i=1}^n \left( \|x_i^k - x_i^{k-1}\|^2 + \|y_i^k - y_i^{k-1}\|^2 \right), \end{aligned}$$

which completes the proof.  $\square$

**Lemma A.6.** *The sequence  $\{v_i^k\}$  generated by Algorithm 1 satisfies*

$$\sum_{i=1}^n \|v_i^{k+1} - v_i^k\|^2 \leq 8 \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 + 4\eta^2 \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2 + 4n\eta^2 \|\bar{d}_v^k\|^2. \quad (31)$$

*Proof.* Follow the update of  $v_i^k$  in (13), we have

$$\begin{aligned} \sum_{i=1}^n \|v_i^{k+1} - v_i^k\|^2 &= \sum_{i=1}^n \|\mathcal{P}_{r_v} \left[ \sum_{j=1}^n w_{ij}(v_j^k + \eta t_{v,j}^k) \right] - v_i^k\|^2 \leq \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij}(v_j^k + \eta t_{v,j}^k) - v_i^k \right\|^2 \\ &\leq 4 \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij}v_j^k - \bar{v}^k \right\|^2 + 4 \sum_{i=1}^n \|\bar{v}^k - v_i^k\|^2 + 4\eta^2 \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij}t_{v,j}^k - \bar{t}_v^k \right\|^2 + 4\eta^2 \sum_{i=1}^n \|\bar{d}_v^k\|^2 \\ &\leq 8 \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 + 4\eta^2 \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2 + 4\eta^2 \sum_{i=1}^n \|\bar{d}_v^k\|^2, \end{aligned}$$

where the first inequality holds because the projection operator is non-expansive, and the third inequality is due to Lemma A.1.  $\square$

**Remark A.3.** *Similarly, we can deduce the following inequalities, details are omitted to avoid repetition.*

$$\sum_{i=1}^n \|x_i^{k+1} - x_i^k\|^2 \leq 8 \sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 + 4\alpha^2 \sum_{i=1}^n \|t_{x,i}^k - \bar{t}_x^k\|^2 + 4n\alpha^2 \|\bar{d}_x^k\|^2, \quad (32)$$

$$\sum_{i=1}^n \|y_i^{k+1} - y_i^k\|^2 \leq 8 \sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 + 4\beta^2 \sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2 + 4n\beta^2 \|\bar{d}_y^k\|^2. \quad (33)$$

**Lemma A.7.** *The sequences  $\{d_{y,i}^k\}$ ,  $\{d_{v,i}^k\}$  generated by Algorithm 1 satisfy*

$$\|\bar{d}_y^k\|^2 \leq \frac{2L_{f,1}^2}{n} \sum_{i=1}^n \left( \|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2 \right) + 2L_{f,1}^2 \|\bar{y}^k - y^*(\bar{x}^k)\|^2, \quad (34)$$

$$\begin{aligned} \|\bar{d}_v^k\|^2 &\leq \frac{5(L_{f,2r_v} + L_{F,1})^2}{n} \sum_{i=1}^n \left( \|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2 \right) + \frac{5L_{f,1}^2}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 \\ &\quad + 5(L_{F,1} + r_v L_{f,2})^2 \|\bar{y}^k - y^*(\bar{x}^k)\|^2 + 5L_{f,1}^2 \|\bar{v}^k - v^*(\bar{x}^k)\|^2. \end{aligned} \quad (35)$$

*Proof.* For  $\nabla_2 f(\bar{x}^k, y^*(\bar{x}^k)) = 0$ , by the definition of  $d_{y,i}^k$  in (9), we obtain

$$\begin{aligned} \|\bar{d}_y^k\|^2 &\leq \frac{2}{n} \sum_{i=1}^n \|\nabla_2 f_i(x_i^k, y_i^k) - \nabla_2 f_i(\bar{x}^k, \bar{y}^k)\|^2 + 2\|\nabla_2 f(\bar{x}^k, \bar{y}^k) - \nabla_2 f(\bar{x}^k, y^*(\bar{x}^k))\|^2 \\ &\leq \frac{2L_{f,1}^2}{n} \sum_{i=1}^n \left( \|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2 \right) + 2L_{f,1}^2 \|\bar{y}^k - y^*(\bar{x}^k)\|^2. \end{aligned}$$

For convenience, we define

$$\begin{cases} \Delta_1 := \left\| \nabla_2 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k)) v^*(\bar{x}^k) - \nabla_2 F(\bar{x}^k, \bar{y}^k) + \nabla_{22}^2 f(\bar{x}^k, \bar{y}^k) \bar{v}^k \right\|, \\ \Delta_2 := \left\| \nabla_2 F(\bar{x}^k, \bar{y}^k) - \nabla_{22}^2 f(\bar{x}^k, \bar{y}^k) \bar{v}^k - \frac{1}{n} \sum_{i=1}^n \left( \nabla_2 F_i(x_i^k, y_i^k) - \nabla_{22}^2 f_i(x_i^k, y_i^k) v_i^k \right) \right\|. \end{cases}$$

From the triangle inequality, it follows that

$$\begin{aligned} \Delta_1 &\leq \left\| \nabla_2 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_2 F(\bar{x}^k, \bar{y}^k) \right\| + \left\| \left( \nabla_{22}^2 f(\bar{x}^k, \bar{y}^k) - \nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k)) \right) \bar{v}^k \right\| \\ &\quad + \left\| \nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k)) (\bar{v}^k - v^*(\bar{x}^k)) \right\| \\ &\leq (L_{F,1} + L_{f,2r_v}) \|\bar{y}^k - y^*(\bar{x}^k)\| + L_{f,1} \|\bar{v}^k - v^*(\bar{x}^k)\|. \end{aligned} \quad (36)$$

By using the triangle inequality again and considering  $F = \frac{1}{n} \sum_{i=1}^n F_i$  and  $f = \frac{1}{n} \sum_{i=1}^n f_i$ , we derive

$$\begin{aligned} \Delta_2 &\leq \frac{1}{n} \sum_{i=1}^n \left\| \nabla_2 F_i(\bar{x}^k, \bar{y}^k) - \nabla_{22}^2 f_i(\bar{x}^k, \bar{y}^k) \bar{v}^k - \nabla_2 F_i(x_i^k, y_i^k) + \nabla_{22}^2 f_i(x_i^k, y_i^k) v_i^k \right\| \\ &\leq \frac{1}{n} \sum_{i=1}^n \left\| \nabla_2 F_i(\bar{x}^k, \bar{y}^k) - \nabla_2 F_i(x_i^k, y_i^k) \right\| + \frac{1}{n} \sum_{i=1}^n \left\| \left( \nabla_{22}^2 f_i(\bar{x}^k, \bar{y}^k) - \nabla_{22}^2 f_i(x_i^k, y_i^k) \right) v_i^k \right\| \\ &\quad + \frac{1}{n} \sum_{i=1}^n \left\| \nabla_{22}^2 f_i(\bar{x}^k, \bar{y}^k) (\bar{v}^k - v_i^k) \right\| \\ &\leq \frac{L_{F,1} + r_v L_{f,2}}{n} \sum_{i=1}^n (\|x_i^k - \bar{x}^k\| + \|y_i^k - \bar{y}^k\|) + \frac{L_{f,1}}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|. \end{aligned} \quad (37)$$

Then, for  $\nabla_2 F(\bar{x}^k, y^*(\bar{x}^k)) = \nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k))v^*(\bar{x}^k)$ , combining (36) and (37), using the inequality  $(\sum_{l=1}^s a_l)^2 \leq s \sum_{l=1}^s a_l^2$ , we can deduce that

$$\begin{aligned} \|\bar{d}_v^k\|^2 &\leq (\Delta_1 + \Delta_2)^2 \leq \frac{5(L_{f,2r_v} + L_{F,1})^2}{n} \sum_{i=1}^n \left( \|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2 \right) \\ &\quad + \frac{5L_{f,1}^2}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 + 5(L_{f,2r_v} + L_{F,1})^2 \|\bar{y}^k - y^*(\bar{x}^k)\|^2 + 5L_{f,1}^2 \|\bar{v}^k - v^*(\bar{x}^k)\|^2, \end{aligned}$$

which completes the proof.  $\square$

Now, combining (28)-(30), (31), (32)-(33), (34)-(35) with Lemma A.4 and Remark A.2, we can establish the boundness of  $\frac{1}{n} \sum_{i=1}^n \|t_{v,i}^{k+1} - \bar{t}_v^{k+1}\|^2$ ,  $\frac{1}{n} \sum_{i=1}^n \|t_{y,i}^{k+1} - \bar{t}_y^{k+1}\|^2$  and  $\frac{1}{n} \sum_{i=1}^n \|t_{x,i}^{k+1} - \bar{t}_x^{k+1}\|^2$ :

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|t_{y,i}^{k+1} - \bar{t}_y^{k+1}\|^2 &\leq (\rho + \frac{4C\beta^2}{1-\rho}) \frac{1}{n} \sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2 + \frac{4C\alpha^2}{(1-\rho)n} \sum_{i=1}^n \|t_{x,i}^k - \bar{t}_x^k\|^2 \\ &\quad + \frac{C(8 + 8L_{f,1}^2\beta^2)}{(1-\rho)n} \sum_{i=1}^n (\|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2) + \frac{4C\alpha^2}{1-\rho} \|\bar{d}_x^k\|^2 + \frac{8CL_{f,1}^2\beta^2}{1-\rho} \|\bar{y}^k - y^*(\bar{x}^k)\|^2, \end{aligned} \quad (38)$$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|t_{x,i}^{k+1} - \bar{t}_x^{k+1}\|^2 &\leq (\rho + \frac{4C\alpha^2}{1-\rho}) \frac{1}{n} \sum_{i=1}^n \|t_{x,i}^k - \bar{t}_x^k\|^2 + \frac{4C}{(1-\rho)n} \sum_{i=1}^n (\beta^2 \|t_{y,i}^k - \bar{t}_y^k\|^2 + \eta^2 \|t_{v,i}^k - \bar{t}_v^k\|^2) \\ &\quad + \frac{C(8 + 8L_{f,1}^2\beta^2 + 20L_1^2\eta^2)}{(1-\rho)n} \sum_{i=1}^n (\|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2) + \frac{C(8 + 20L_{f,1}^2\eta^2)}{(1-\rho)n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 \\ &\quad + \frac{4C\alpha^2}{1-\rho} \|\bar{d}_x^k\|^2 + \frac{C(8L_{f,1}^2\beta^2 + 20L_1^2\eta^2)}{1-\rho} \|\bar{y}^k - y^*(\bar{x}^k)\|^2 + \frac{20CL_{f,1}^2\eta^2}{1-\rho} \|\bar{v}^k - v^*(\bar{x}^k)\|^2, \end{aligned} \quad (39)$$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|t_{v,i}^{k+1} - \bar{t}_v^{k+1}\|^2 &\leq (\rho + \frac{4C\eta^2}{1-\rho}) \frac{1}{n} \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2 + \frac{4C}{(1-\rho)n} \sum_{i=1}^n (\beta^2 \|t_{y,i}^k - \bar{t}_y^k\|^2 + \alpha^2 \|t_{x,i}^k - \bar{t}_x^k\|^2) \\ &\quad + \frac{C(8 + 8L_{f,1}^2\beta^2 + 20L_1^2\eta^2)}{(1-\rho)n} \sum_{i=1}^n (\|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2) + \frac{C(8 + 20L_{f,1}^2\eta^2)}{(1-\rho)n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 \\ &\quad + \frac{4C\alpha^2}{1-\rho} \|\bar{d}_x^k\|^2 + \frac{C(8L_{f,1}^2\beta^2 + 20L_1^2\eta^2)}{1-\rho} \|\bar{y}^k - y^*(\bar{x}^k)\|^2 + \frac{20CL_{f,1}^2\eta^2}{1-\rho} \|\bar{v}^k - v^*(\bar{x}^k)\|^2, \end{aligned} \quad (40)$$

where  $C = \max\{C_1, C_2\}$  and  $L_1$  is defined in (8).

### A.3 Convergence Rate of Algorithm 1

Before the final proof, we first establish several useful lemmas. Recall that  $\Phi(x) = F(x, y^*(x))$  denotes the overall objective function.

**Lemma A.8.** *The sequence  $\{(x_i^k, y_i^k, v_i^k)\}$  generated by Algorithm 1 satisfies*

$$\begin{aligned} \Phi(\bar{x}^{k+1}) - \Phi(\bar{x}^k) &\leq -\frac{\alpha}{2} \|\nabla\Phi(\bar{x}^k)\|^2 - \frac{1}{2} \left( \frac{1}{\alpha} - L_\Phi \right) \|\bar{x}^{k+1} - \bar{x}^k\|^2 + \frac{5\alpha L_{f,1}^2}{2} \|\bar{v}^k - v^*(\bar{x}^k)\|^2 \\ &\quad + \frac{5\alpha(L_{F,1} + L_{f,2r_v})^2}{2} \|\bar{y}^k - y^*(\bar{x}^k)\|^2 + \frac{5\alpha L_{f,1}^2}{2n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 \\ &\quad + \frac{5\alpha(L_{F,1} + r_v L_{f,2})^2}{2n} \sum_{i=1}^n \left( \|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2 \right), \end{aligned} \quad (41)$$

where  $L_\Phi$  is defined in (8).

*Proof.* It follows from [12, Lemma 2.2] that  $\nabla\Phi(x)$  is  $L_\Phi$ -Lipschitz continuous. On the other hand, from [1, Lemma 5.7] we derive

$$\begin{aligned}
\Phi(\bar{x}^{k+1}) - \Phi(\bar{x}^k) &\leq \langle \nabla\Phi(\bar{x}^k), \bar{x}^{k+1} - \bar{x}^k \rangle + \frac{L_\Phi}{2} \|\bar{x}^{k+1} - \bar{x}^k\|^2 \\
&\stackrel{(25)}{=} -\alpha \langle \nabla\Phi(\bar{x}^k), \bar{d}_x^k \rangle + \frac{L_\Phi}{2} \|\bar{x}^{k+1} - \bar{x}^k\|^2 \\
&= -\frac{\alpha}{2} \|\nabla\Phi(\bar{x}^k)\|^2 - \frac{\alpha}{2} \|\bar{d}_x^k\|^2 + \frac{\alpha}{2} \|\nabla\Phi(\bar{x}^k) - \bar{d}_x^k\|^2 + \frac{L_\Phi}{2} \|\bar{x}^{k+1} - \bar{x}^k\|^2 \\
&\stackrel{(25)}{=} -\frac{\alpha}{2} \|\nabla\Phi(\bar{x}^k)\|^2 - \frac{1}{2\alpha} \|\bar{x}^{k+1} - \bar{x}^k\|^2 + \frac{\alpha}{2} \|\nabla\Phi(\bar{x}^k) - \bar{d}_x^k\|^2 + \frac{L_\Phi}{2} \|\bar{x}^{k+1} - \bar{x}^k\|^2.
\end{aligned} \tag{42}$$

By definition, we have  $\nabla\Phi(\bar{x}^k) = \nabla_1 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_{12}^2 f(\bar{x}^k, y^*(\bar{x}^k)) v^*(\bar{x}^k)$  and

$$\bar{d}_x^k = \frac{1}{n} \sum_{i=1}^n \left( \nabla_1 F_i(x_i^k, y_i^k) - \nabla_{12}^2 f_i(x_i^k, y_i^k) v_i^k \right).$$

For convenience, we define

$$\begin{cases} \Delta_3 := \left\| \nabla_1 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_{12}^2 f(\bar{x}^k, y^*(\bar{x}^k)) v^*(\bar{x}^k) - \nabla_1 F(\bar{x}^k, \bar{y}^k) + \nabla_{12}^2 f(\bar{x}^k, \bar{y}^k) \bar{v}^k \right\|, \\ \Delta_4 := \left\| \nabla_1 F(\bar{x}^k, \bar{y}^k) - \nabla_{12}^2 f(\bar{x}^k, \bar{y}^k) \bar{v}^k - \frac{1}{n} \sum_{i=1}^n \left( \nabla_1 F_i(x_i^k, y_i^k) - \nabla_{12}^2 f_i(x_i^k, y_i^k) v_i^k \right) \right\|. \end{cases}$$

From the triangle inequality, following the similar steps in (36) and (37) respectively, we have

$$\Delta_3 \leq (L_{F,1} + L_{f,2} r_v) \|\bar{y}^k - y^*(\bar{x}^k)\| + L_{f,1} \|\bar{v}^k - v^*(\bar{x}^k)\|, \tag{43}$$

and

$$\Delta_4 \leq \frac{L_{F,1} + r_v L_{f,2}}{n} \sum_{i=1}^n (\|x_i^k - \bar{x}^k\| + \|y_i^k - \bar{y}^k\|) + \frac{L_{f,1}}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|. \tag{44}$$

It is easy to observe from the triangle inequality that  $\|\nabla\Phi(\bar{x}^k) - \bar{d}_x^k\| \leq \Delta_3 + \Delta_4$ . Then, by using (43)-(44), it is easy to derive

$$\begin{aligned}
\left\| \nabla\Phi(\bar{x}^k) - \bar{d}_x^k \right\|^2 &\leq 5(L_{F,1} + L_{f,2} r_v)^2 \|\bar{y}^k - y^*(\bar{x}^k)\|^2 + 5L_{f,1}^2 \|\bar{v}^k - v^*(\bar{x}^k)\|^2 \\
&\quad + \frac{5(L_{F,1} + r_v L_{f,2})^2}{n} \sum_{i=1}^n (\|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2) + \frac{5L_{f,1}^2}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2,
\end{aligned}$$

which, together with (42), yields the desired result.  $\square$

The next two lemmas bound  $\|\bar{y}^{k+1} - y^*(\bar{x}^k)\|$  and  $\|\bar{v}^{k+1} - v^*(\bar{x}^k)\|$ , respectively.

**Lemma A.9.** *The sequence  $\{(x_i^k, y_i^k, v_i^k)\}$  generated by Algorithm 1 satisfies*

$$\|\bar{y}^{k+1} - y^*(\bar{x}^k)\|^2 \leq \left(1 - \frac{\beta\sigma}{2}\right) \|\bar{y}^k - y^*(\bar{x}^k)\|^2 + \frac{3\beta L_{f,1}^2}{n\sigma} \sum_{i=1}^n (\|\bar{x}^k - x_i^k\|^2 + \|\bar{y}^k - y_i^k\|^2). \tag{45}$$

*Proof.* First, by Cauchy-Schwartz inequality, for any  $\xi > 0$ , we have

$$\begin{aligned}
\|\bar{y}^{k+1} - y^*(\bar{x}^k)\|^2 &= \|\bar{y}^k - \beta \nabla_2 f(\bar{x}^k, \bar{y}^k) - y^*(\bar{x}^k)\|^2 + \beta^2 \|\nabla_2 f(\bar{x}^k, \bar{y}^k) - \bar{d}_y^k\|^2 \\
&\leq (1 + \xi) \|\bar{y}^k - \beta \nabla_2 f(\bar{x}^k, \bar{y}^k) - y^*(\bar{x}^k)\|^2 + (1 + 1/\xi) \beta^2 \|\nabla_2 f(\bar{x}^k, \bar{y}^k) - \bar{d}_y^k\|^2.
\end{aligned} \tag{46}$$

Note that  $\nabla_2 f(\bar{x}^k, y^*(\bar{x}^k)) = 0$ . It follows from the  $\sigma$ -strong convexity of  $f(\bar{x}^k, \cdot)$ ,  $L_{f,1}$ -smoothness of  $f$  and [26, Theorem 2.1.12] that

$$\begin{aligned} \langle \bar{y}^k - y^*(\bar{x}^k), \nabla_2 f(\bar{x}^k, \bar{y}^k) \rangle &= \langle \bar{y}^k - y^*(\bar{x}^k), \nabla_2 f(\bar{x}^k, \bar{y}^k) - \nabla_2 f(\bar{x}^k, y^*(\bar{x}^k)) \rangle \\ &\geq \frac{\sigma L_{f,1}}{\sigma + L_{f,1}} \|\bar{y}^k - y^*(\bar{x}^k)\|^2 + \frac{1}{\sigma + L_{f,1}} \|\nabla_2 f(\bar{x}^k, \bar{y}^k)\|^2. \end{aligned} \quad (47)$$

By expanding  $\|\bar{y}^k - \beta \nabla_2 f(\bar{x}^k, \bar{y}^k) - y^*(\bar{x}^k)\|^2$ , plugging in (47), and noting that  $\beta \leq \frac{2}{\sigma + L_{f,1}}$  in Algorithm 1, we obtain

$$\|\bar{y}^k - \beta \nabla_2 f(\bar{x}^k, \bar{y}^k) - y^*(\bar{x}^k)\|^2 \leq \left(1 - \frac{2\beta\sigma L_{f,1}}{\sigma + L_{f,1}}\right) \|\bar{y}^k - y^*(\bar{x}^k)\|^2 \leq (1 - \beta\sigma) \|\bar{y}^k - y^*(\bar{x}^k)\|^2. \quad (48)$$

It is elementary to show from  $f = \frac{1}{n} \sum_{i=1}^n f_i$ , the definition of  $\bar{d}_y^k$  in (16), the triangle inequality and Assumption 2.1 (c) that

$$\|\nabla_2 f(\bar{x}^k, \bar{y}^k) - \bar{d}_y^k\|^2 \leq \frac{L_{f,1}^2}{n} \sum_{i=1}^n \left( \|\bar{x}^k - x_i^k\|^2 + \|\bar{y}^k - y_i^k\|^2 \right), \quad (49)$$

which, together with (46), (48) and the relation  $\bar{d}_y^k = \bar{t}_y^k$ , yields

$$\|\bar{y}^{k+1} - y^*(\bar{x}^k)\|^2 \leq (1 + \xi)(1 - \beta\sigma) \|\bar{y}^k - y^*(\bar{x}^k)\|^2 + (1 + 1/\xi) \frac{\beta^2 L_{f,1}^2}{n} \sum_{i=1}^n \left( \|\bar{x}^k - x_i^k\|^2 + \|\bar{y}^k - y_i^k\|^2 \right).$$

Finally, we arrive at the desired result (45) by taking  $\xi = \beta\sigma/2$  and using  $\beta\sigma \leq 1$ .  $\square$

**Lemma A.10.** *The sequence  $\{(x_i^k, y_i^k, v_i^k)\}$  generated by Algorithm 1 satisfies*

$$\begin{aligned} \|\bar{v}^{k+1} - v^*(\bar{x}^k)\|^2 &\leq (1 - \eta\sigma/2) \|\bar{v}^k - v^*(\bar{x}^k)\|^2 + \frac{3\eta(L_{F,1} + L_{f,2}r_v)^2}{\sigma} \|\bar{y}^k - y^*(\bar{x}^k)\|^2 \\ &+ \frac{9\eta(L_{F,1} + L_{f,2}r_v)^2}{n\sigma} \sum_{i=1}^n \left( \|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2 \right) + \frac{2\eta^2\rho^2}{n} \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2 \\ &+ \left( 9\eta L_{f,1}^2/\sigma + 2\rho^2 \right) \frac{1}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2. \end{aligned} \quad (50)$$

*Proof.* For convenience, we define  $\Delta_5 := \bar{v}^k + \eta \left[ \nabla_2 F(\bar{x}^k, \bar{y}^k) - \nabla_{22}^2 f(\bar{x}^k, \bar{y}^k) \bar{v}^k \right] - v^*(\bar{x}^k)$ . Similar to (46), for any  $\delta > 0$ , we have

$$\|\bar{v}^k + \eta \bar{d}_v^k - v^*(\bar{x}^k)\|^2 \leq (1 + \delta) \|\Delta_5\|^2 + (1 + 1/\delta) \eta^2 \Delta_2^2. \quad (51)$$

Recall that the term  $\Delta_2$  is treated in (37). Next, we treat the term  $\|\Delta_5\|^2$  in (51). Since  $\nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k)) v^*(\bar{x}^k) = \nabla_2 F(\bar{x}^k, y^*(\bar{x}^k))$ , we have the following reformulation:  $\Delta_5 = \Delta_{5,1} - \eta(\Delta_{5,2} + \Delta_{5,3})$ , where

$$\begin{cases} \Delta_{5,1} := \left[ I - \eta \nabla_{22}^2 f(\bar{x}^k, \bar{y}^k) \right] \left( \bar{v}^k - v^*(\bar{x}^k) \right), \\ \Delta_{5,2} := \left[ \nabla_{22}^2 f(\bar{x}^k, \bar{y}^k) - \nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k)) \right] v^*(\bar{x}^k), \\ \Delta_{5,3} := \nabla_2 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_2 F(\bar{x}^k, \bar{y}^k). \end{cases}$$

By Cauchy-Schwartz inequality, for any  $\delta_1 > 0$ , we have

$$\|\Delta_5\|^2 \leq (1 + \delta_1) \|\Delta_{5,1}\|^2 + (1 + 1/\delta_1) \eta^2 \|\Delta_{5,2} + \Delta_{5,3}\|^2.$$

Since  $\eta \leq \bar{\eta} \leq 1/L_{f,1}$  and  $f(x, \cdot)$  is  $\sigma$ -strongly convex, there holds

$$\|\Delta_{5,1}\| \leq \left\| I - \eta \nabla_{22}^2 f(\bar{x}^k, \bar{y}^k) \right\|_{\text{op}} \left\| \bar{v}^k - v^*(\bar{x}^k) \right\| \leq (1 - \eta\sigma) \left\| \bar{v}^k - v^*(\bar{x}^k) \right\|.$$

Furthermore, it is apparent from Assumption 2.1 that  $\|\Delta_{5,2} + \Delta_{5,3}\| \leq (L_{f,2r_v} + L_{F,1}) \|\bar{y}^k - y^*(\bar{x}^k)\|$ . Taking  $\delta_1 = \eta\sigma$  and noting  $\eta\sigma \leq \eta L_{f,1} \leq 1$ , we obtain

$$\begin{aligned} \|\Delta_5\|^2 &\leq (1 + \eta\sigma) (1 - \eta\sigma)^2 \left\| \bar{v}^k - v^*(\bar{x}^k) \right\|^2 + (1 + 1/\eta\sigma) \eta^2 (L_{f,2r_v} + L_{F,1})^2 \left\| \bar{y}^k - y^*(\bar{x}^k) \right\|^2 \\ &\leq (1 - \eta\sigma) \left\| \bar{v}^k - v^*(\bar{x}^k) \right\|^2 + \frac{2\eta (L_{f,2r_v} + L_{F,1})^2}{\sigma} \left\| \bar{y}^k - y^*(\bar{x}^k) \right\|^2. \end{aligned} \quad (52)$$

We next evaluate  $\sum_{i=1}^n \|v_i^{k+1} - v^*(\bar{x}^k)\|^2$  to bound  $\|\bar{v}^{k+1} - v^*(\bar{x}^k)\|^2$ . By the update of  $v_i^k$  in Algorithm 1, we have

$$\begin{aligned} \sum_{i=1}^n \|v_i^{k+1} - v^*(\bar{x}^k)\|^2 &= \sum_{i=1}^n \left\| \mathcal{P}_{r_v} \left[ \sum_{j=1}^n w_{ij} (v_j^k + \eta t_{v,j}^k) \right] - v^*(\bar{x}^k) \right\|^2 \\ &\leq \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} (v_j^k + \eta t_{v,j}^k) - v^*(\bar{x}^k) \right\|^2 \\ &= \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} v_j^k - \bar{v}^k + \eta \left( \sum_{j=1}^n w_{ij} t_{v,j}^k - \bar{t}_v^k \right) \right\|^2 + \sum_{i=1}^n \|\bar{v}^k - v^*(\bar{x}^k) + \eta \bar{t}_v^k\|^2 \\ &\leq \sum_{i=1}^n \|\bar{v}^k - v^*(\bar{x}^k) + \eta \bar{t}_v^k\|^2 + 2 \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} v_j^k - \bar{v}^k \right\|^2 + 2\eta^2 \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} t_{v,j}^k - \bar{t}_v^k \right\|^2 \\ &\leq \sum_{i=1}^n \|\bar{v}^k - v^*(\bar{x}^k) + \eta \bar{t}_v^k\|^2 + 2\rho^2 \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 + 2\eta^2 \rho^2 \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2, \end{aligned}$$

where the second equality holds because  $\sum_{i=1}^n \left[ \sum_{j=1}^n w_{ij} v_j^k - \bar{v}^k + \eta \left( \sum_{j=1}^n w_{ij} t_{v,j}^k - \bar{t}_v^k \right) \right] = 0$ , and the last inequality follows from Lemma A.1. Note that  $\bar{d}_v^k = \bar{t}_v^k$ . Then, by combining (51), (52) and (37), it can be concluded that

$$\begin{aligned} \|\bar{v}^{k+1} - v^*(\bar{x}^k)\|^2 &\leq \frac{1}{n} \sum_{i=1}^n \|v_i^{k+1} - v^*(\bar{x}^k)\|^2 \\ &\leq \|\bar{v}^k - v^*(\bar{x}^k) + \eta \bar{t}_v^k\|^2 + \frac{2\rho^2}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 + \frac{2\eta^2 \rho^2}{n} \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2 \\ &\leq (1 + \delta) (1 - \eta\sigma) \left\| \bar{v}^k - v^*(\bar{x}^k) \right\|^2 + \frac{2(1 + \delta)\eta (L_{F,1} + L_{f,2r_v})^2}{\sigma} \left\| \bar{y}^k - y^*(\bar{x}^k) \right\|^2 \\ &\quad + \frac{3(1 + 1/\delta)\eta^2}{n} \left[ (L_{F,1} + L_{f,2r_v})^2 \sum_{i=1}^n \left( \|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2 \right) + L_{f,1}^2 \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 \right] \\ &\quad + \frac{2\rho^2}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 + \frac{2\eta^2 \rho^2}{n} \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2. \end{aligned}$$

Finally, the desired result (50) follows by setting  $\delta = \eta\sigma/2$  and using  $\eta\sigma \leq 1$ .  $\square$

**Lemma A.11.** *Let  $L_v$  be defined in (8). There hold*

$$\|y^*(\bar{x}^{k+1}) - y^*(\bar{x}^k)\| \leq \frac{L_{f,1}}{\sigma} \|\bar{x}^{k+1} - \bar{x}^k\| \quad \text{and} \quad \|v^*(\bar{x}^{k+1}) - v^*(\bar{x}^k)\| \leq \frac{L_v}{\sigma} \|\bar{x}^{k+1} - \bar{x}^k\|. \quad (53)$$

*Proof.* Due to the optimality of  $y^*(x)$ , we have  $\nabla_2 f(x, y^*(x)) = 0$  for any  $x$ . Now, let  $x$  and  $x'$  be arbitrarily fixed. Then, it follows from the  $\sigma$ -strong convexity of  $f(x, \cdot)$  and  $L_{f,1}$ -Lipschitz continuity of  $\nabla f$  that

$$\begin{aligned} \sigma \|y^*(x) - y^*(x')\| &\leq \|\nabla_2 f(x, y^*(x)) - \nabla_2 f(x, y^*(x'))\| \\ &= \|\nabla_2 f(x, y^*(x')) - \nabla_2 f(x', y^*(x'))\| \leq L_{f,1} \|x - x'\| \end{aligned}$$

Hence, we have

$$\|y^*(x) - y^*(x')\| \leq (L_{f,1}/\sigma) \|x - x'\|. \quad (54)$$

Then we can get first inequality immediately by taking  $x = \bar{x}^{k+1}$  and  $x' = \bar{x}^k$ . Next, to obtain the second inequality, we define

$$\begin{cases} \Delta_6 := \nabla_2 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_2 F(\bar{x}^{k+1}, y^*(\bar{x}^{k+1})), \\ \Delta_7 := [\nabla_{22}^2 f(\bar{x}^{k+1}, y^*(\bar{x}^{k+1})) - \nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k))] v^*(\bar{x}^{k+1}). \end{cases}$$

By using Assumption 2.1, we have

$$\begin{cases} \|\Delta_6\| \leq L_{F,1} (\|\bar{x}^{k+1} - \bar{x}^k\| + \|y^*(\bar{x}^{k+1}) - y^*(\bar{x}^k)\|), \\ \|\Delta_7\| \leq L_{f,2} r_v (\|\bar{x}^{k+1} - \bar{x}^k\| + \|y^*(\bar{x}^{k+1}) - y^*(\bar{x}^k)\|). \end{cases} \quad (55)$$

It follows from  $\nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k)) v^*(\bar{x}^k) = \nabla_2 F(\bar{x}^k, y^*(\bar{x}^k))$  that

$$\nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k)) (v^*(\bar{x}^k) - v^*(\bar{x}^{k+1})) = \Delta_6 + \Delta_7. \quad (56)$$

The  $\sigma$ -strong convexity of  $f(\bar{x}^k, \cdot)$  implies that

$$\begin{aligned} \sigma \|v^*(\bar{x}^k) - v^*(\bar{x}^{k+1})\| &\leq \|\nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k)) (v^*(\bar{x}^k) - v^*(\bar{x}^{k+1}))\| \\ &\stackrel{(56)}{=} \|\Delta_6 + \Delta_7\| \leq \|\Delta_6\| + \|\Delta_7\| \stackrel{(55,54)}{\leq} (L_{F,1} + L_{f,2} r_v) (1 + L_{f,1}/\sigma) \|\bar{x}^{k+1} - \bar{x}^k\|, \end{aligned}$$

which implies the second inequality in (53) by noting the definition of  $L_v$  in (8).  $\square$

Now, we are ready to prove Theorem 3.1.

*Proof.* By using Cauchy-Schwartz inequality again, we derive

$$\|\bar{y}^{k+1} - y^*(\bar{x}^{k+1})\|^2 \leq \left(1 + \frac{\beta\sigma}{4}\right) \|\bar{y}^{k+1} - y^*(\bar{x}^k)\|^2 + \left(1 + \frac{4}{\beta\sigma}\right) \|y^*(\bar{x}^{k+1}) - y^*(\bar{x}^k)\|^2.$$

Taking into account (45), the first inequality in (53), and  $\beta\sigma \leq 1$ , we obtain

$$\begin{aligned} \|\bar{y}^{k+1} - y^*(\bar{x}^{k+1})\|^2 &\leq \left(1 - \frac{\beta\sigma}{4}\right) \|\bar{y}^k - y^*(\bar{x}^k)\|^2 \\ &\quad + \frac{15\beta L_{f,1}^2}{4n\sigma} \sum_{i=1}^n (\|\bar{x}^k - x_i^k\|^2 + \|\bar{y}^k - y_i^k\|^2) + \frac{5L_{f,1}^2}{\beta\sigma^3} \|\bar{x}^{k+1} - \bar{x}^k\|^2. \end{aligned} \quad (57)$$



Using Cauchy-Schwartz inequality again, we derive

$$\|\bar{v}^{k+1} - v^*(\bar{x}^{k+1})\|^2 \leq \left(1 + \frac{\eta\sigma}{4}\right) \|\bar{v}^{k+1} - v^*(\bar{x}^k)\|^2 + \left(1 + \frac{4}{\eta\sigma}\right) \|v^*(\bar{x}^{k+1}) - v^*(\bar{x}^k)\|^2.$$

Similarly, taking into account (50), the second inequality in (53), and  $\eta\sigma \leq 1$ , we obtain

$$\begin{aligned} \|\bar{v}^{k+1} - v^*(\bar{x}^{k+1})\|^2 &\leq \left(1 - \frac{\eta\sigma}{4}\right) \|\bar{v}^k - v^*(\bar{x}^k)\|^2 + \frac{15\eta(L_{F,1} + L_{f,2}r_v)^2}{4\sigma} \|\bar{y}^k - y^*(\bar{x}^k)\|^2 \\ &\quad + \frac{5L_v^2}{\eta\sigma^3} \|\bar{x}^{k+1} - \bar{x}^k\|^2 + \frac{45\eta(L_{F,1} + L_{f,2}r_v)^2}{4n\sigma} \sum_{i=1}^n (\|x_i^k - \bar{x}^k\|^2 + \|y_i^k - \bar{y}^k\|^2) \\ &\quad + \left(\frac{45\eta L_{f,1}^2}{4\sigma} + \frac{5\rho^2}{2}\right) \frac{1}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 + \frac{5\rho^2\eta^2}{2n} \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2. \end{aligned} \quad (58)$$

Define the Lyapunov function as in (15). By combining (39)-(40), (22)-(21), (41), (57), (58), it is straightforward to derive

$$V_{k+1} - V_k \leq -\frac{\alpha}{2} \|\nabla\Phi(\bar{x}^k)\|^2 - A_1 \|\bar{d}_x^k\|^2 - A_2 \|\bar{y}^k - y^*(\bar{x}^k)\|^2 - A_3 \|\bar{v}^k - v^*(\bar{x}^k)\|^2 \quad (59)$$

$$- \frac{A_4}{n} \sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 - \frac{A_5}{n} \sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 - \frac{A_6}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 \quad (60)$$

$$- \frac{A_7}{n} \sum_{i=1}^n \|t_{x,i}^k - \bar{t}_x^k\|^2 - \frac{A_8}{n} \sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2 - \frac{A_9}{n} \sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2, \quad (61)$$

where the coefficients are given by

$$\begin{aligned} A_1 &= \frac{\alpha}{2} - \frac{L_\Phi\alpha^2}{2} - \frac{5L_{f,1}^2 a_1\alpha^2}{\sigma^3\beta} - \frac{5L_v^2 a_2\alpha^2}{\sigma^3\eta} - \frac{4C(a_6\alpha^2 + a_7\beta^2 + a_8\eta^2)\alpha^2}{1-\rho}, \\ A_2 &= \frac{\sigma a_1\beta}{4} - \frac{5L_1^2\alpha}{2} - \frac{15L_1^2 a_2\eta}{4\sigma} - \frac{C(a_6\alpha^2 + a_8\eta^2)(8L_{f,1}^2\beta^2 + 20L_1^2\eta^2)}{1-\rho} - \frac{8CL_{f,1}^2 a_7\beta^4}{1-\rho}, \\ A_3 &= \frac{\sigma a_2\eta}{4} - \frac{5L_{f,1}^2\alpha}{2} - \frac{20CL_{f,1}^2\eta^2(a_6\alpha^2 + a_8\eta^2)}{1-\rho}, \\ A_4 &= a_3(1-\rho) - \frac{5L_1^2\alpha}{2} - \frac{15L_{f,1}^2 a_1\beta}{4\sigma} - \frac{45L_1^2 a_2\eta}{4\sigma} - \frac{C(a_6\alpha^2 + a_8\eta^2)(8 + 8L_{f,1}^2\beta^2 + 20L_1^2\eta^2)}{1-\rho} - \frac{Ca_7\beta^2(8 + 8L_{f,1}^2\beta^2)}{1-\rho}, \\ A_5 &= a_4(1-\rho) - \frac{5L_1^2\alpha}{2} - \frac{15L_{f,1}^2 a_1\beta}{4\sigma} - \frac{45L_1^2 a_2\eta}{4\sigma} - \frac{C(a_6\alpha^2 + a_8\eta^2)(8 + 8L_{f,1}^2\beta^2 + 20L_1^2\eta^2)}{1-\rho} - \frac{Ca_7\beta^2(8 + 8L_{f,1}^2\beta^2)}{1-\rho}, \\ A_6 &= a_5(1-\rho) - \frac{5L_{f,1}^2\alpha}{2} - \frac{45L_1^2 a_2\eta}{4\sigma} - \frac{5\rho^2 a_2}{2} - \frac{C(a_6\alpha^2 + a_8\eta^2)(8 + 20L_{f,1}^2\eta^2)}{1-\rho}, \\ A_7 &= a_6\alpha^2(1-\rho) - \frac{\rho^2 a_3\alpha^2}{1-\rho} - \frac{4C(a_6\alpha^2 + a_8\eta^2 + a_7\beta^2)\alpha^2}{1-\rho}, \\ A_8 &= a_7\beta^2(1-\rho) - \frac{\rho^2 a_4\beta^2}{1-\rho} - \frac{4C(a_6\alpha^2 + a_8\eta^2 + a_7\beta^2)\beta^2}{1-\rho}, \\ A_9 &= a_8\eta^2(1-\rho) - \frac{\rho^2 a_5\eta^2}{1-\rho} - \frac{5\rho^2 a_2\eta^2}{2} - \frac{4C(a_6\alpha^2 + a_8\eta^2)\eta^2}{1-\rho}. \end{aligned}$$

Here, constants such as  $L_\Phi$ ,  $L_v$  and  $L_1$  are defined in (8), take

$$a_1 = a_2 = a_3 = a_4 = 1, a_5 = \frac{10\rho^2}{1-\rho}, a_6 = a_7 = \frac{2\rho^2}{(1-\rho)^2}, a_8 = \max\left\{\frac{30\rho^4}{(1-\rho)^3}, \frac{15\rho^2}{2(1-\rho)}\right\},$$

and choose stepsizes satisfying

$$\begin{aligned}\beta &< \min \left\{ \frac{4\sigma(1-\rho)}{75L_{f,1}^2}, \frac{1}{L_{f,1}}, \frac{1-\rho}{\sqrt{96Ca_7}}, \frac{\sigma(1-\rho)}{128Ca_7}, \frac{1-\rho}{\sqrt{24C}} \right\}, \\ \eta &< \min \left\{ \frac{4\sigma(1-\rho)}{225L_1^2}, \frac{2\sigma\rho^2}{9L_1^2}, \frac{\sigma\beta}{60L_1^2}, \frac{1}{L_{f,1}}, \frac{1}{L_1}, \frac{\rho(1-\rho)}{\sqrt{360Ca_8}}, \sqrt{\frac{\sigma(1-\rho)\beta}{896Ca_8}}, \frac{\sigma(1-\rho)}{320Ca_8}, \frac{1-\rho}{\sqrt{24Ca_8/a_6}}, \frac{1-\rho}{\sqrt{24C}} \right\}, \\ \alpha &< \min \left\{ \frac{1}{4L_\Phi}, \frac{\sigma^3\beta}{40L_{f,1}^2}, \frac{\sigma^3\eta}{40L_v^2}, \frac{\sigma\beta}{40L_1^2}, \frac{\sigma\eta}{20L_{f,1}^2}, \frac{2(1-\rho)}{25L_1^2}, \frac{\rho^2}{L_{f,1}^2}, 1, \frac{\rho(1-\rho)}{\sqrt{360Ca_6}}, \sqrt{\frac{\sigma(1-\rho)\beta}{896Ca_6}}, \sqrt{\frac{\sigma(1-\rho)\eta}{320Ca_6}}, \frac{(1-\rho)^{3/2}}{\sqrt{24C}} \right\},\end{aligned}\tag{62}$$

it is elementary to show that  $A_1, A_2, \dots, A_9$  are all nonnegative. Then it follows that

$$V_{k+1} - V_k \leq -\frac{\alpha}{2} \|\nabla\Phi(\bar{x}^k)\|^2.$$

By telescoping,

$$\min_{0 \leq k \leq K-1} \|\nabla\Phi(\bar{x}^k)\|^2 \leq \frac{1}{K} \sum_{k=0}^{K-1} \|\nabla\Phi(\bar{x}^k)\|^2 \leq \frac{2V_0}{\alpha K} = O\left(\frac{1}{K}\right).$$

The consensus error can also be established by

$$V_{k+1} - V_k \leq -\frac{A_4}{n} \sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2,$$

which yields

$$\min_{0 \leq k \leq K-1} \frac{1}{n} \sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 \leq \frac{1}{nK} \sum_{k=0}^{K-1} \sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 \leq \frac{V_0}{A_4 K} = O\left(\frac{1}{K}\right).$$

Similarly, we can derive

$$\min_{0 \leq k \leq K-1} \frac{1}{n} \sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 = O\left(\frac{1}{K}\right) \quad \text{and} \quad \min_{0 \leq k \leq K-1} \frac{1}{n} \sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 = O\left(\frac{1}{K}\right),$$

which completes the proof of Theorem 3.1.  $\square$