A Single-Loop Algorithm for Decentralized Bilevel Optimization

Youran Dong* Shiqian Ma† Junfeng Yang‡ Chao Yin§

Abstract

Bilevel optimization has received more and more attention recently due to its wide applications in machine learning. In this paper, we consider bilevel optimization in decentralized networks. In particular, we propose a novel single-loop algorithm for solving decentralized bilevel optimization with strongly convex lower level problem. Our algorithm is fully single-loop and does not require heavy matrix-vector multiplications when approximating the hypergradient. Moreover, unlike existing methods for decentralized bilevel optimization and federated bilevel optimization, our algorithm does not require any gradient heterogeneity assumption. Our analysis shows that the proposed algorithm achieves the best known convergence rate for bilevel optimization algorithms.

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 [25, 10], meta learning [10, 28, 14] and adversarial training [2, 31, 33]. A generic BO has the following form:

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

Throughout this paper, we assume that the lower-level (LL) function f is 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 $\nabla\Phi(x)$ has the following form:

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

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; (ii) how to deal with the

*Department of Mathematics, Nanjing University, Nanjing, P. R. China. Email: yrdong@smail.nju.edu.cn.

†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.

‡Department of Mathematics, Nanjing University, Nanjing, P. R. China. Email: jfyang@nju.edu.cn. Research supported by NSFC (12371301).

§Department of Mathematics, Nanjing University, Nanjing, P. R. China. Email: yinchao@smail.nju.edu.cn.

matrix inversion, or equivalently, solve the linear system in (2). Different approaches addressing these two questions lead to different algorithms for solving (1). A basic algorithm along this line for solving (1) updates the iterates as follows:

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

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

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

In practice, exactly calculating the Hessian inverse or solving the linear system 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, 30, 14, 12, 16, 15], leverage automatic differentiation to approximate the hypergradient using backpropagation. Approaches related to AID, including those proposed in [25, 11, 12, 16, 15, 4, 13, 8], use various methods to solve the linear system (4). Some of these methods employ gradient descent or conjugate gradient methods, while others use Neumann series to approximate the inverse matrix. 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, 13] 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 involve heavy Hessian- and Jacobian-vector multiplications. Recently, Dagréou et al. [8] proposed a fully single-loop framework (named SOBA) for solving (1) that does not need heavy matrix-vector multiplications to approximately solve the linear system in (4). 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 α_k, β_k and η_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

$$\begin{aligned}
d_y(x, y) &= \nabla_2 f(x, y), \\
d_v(x, y, v) &= \nabla_2 F(x, y) - \nabla^2_{22} f(x, y) v, \\
d_x(x, y, v) &= \nabla_1 F(x, y) - \nabla^2_{12} f(x, y) v.
\end{aligned}$$

The SOBA framework [8] was later extended to the case where LL problem is merely convex by Liu et al. [20], and the authors named their algorithm sl-BAMM. However, the sl-BAMM algorithm needs to assume that the sequence $\{x^k, y^k, v^k\}$ is uniformly bounded in order to prove convergence. This is a very strong assumption and not verifiable in practice. Other notable BO algorithms include [17, 3] which first convert (1) to an equivalent constrained single-level problem, and then approximately solve this reformulated problem. This avoids calculating the approximated hypergradient (4), but special care is needed for handling the constraints.

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 distributed to n agents, and each agent can only communicate with its neighbors in the network. The n agents cooperatively solve the BO problem through local updates and communications. Decentralized optimization has many benefits such as speeding up the convergence rate and protecting data privacy [19]. The DBO problem is given below:

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

where the i-th agent only has access to the data related to F_i and f_i. The main challenge in designing a decentralized gradient method for solving the DBO (6) is how to compute the hypergradient information. Note that, the hypergradient $\nabla\Phi(x)$ of (6) is given by

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

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 Chen et al. [5], where the authors proposed the DSBO algorithm that incorporates a decentralized algorithm to solve the linear system 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 [22], Lu et al. 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. [32] 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., [26]). We list below the heterogeneity assumptions in these papers.

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

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

- (SLAM, Theorem 1 in [22]) There exists a constant $L \geq 0$ such that

$$\left\|\nabla^2_{22} f_i(x_i,y_i) - \frac{1}{n}\sum_{i=1}^n \nabla^2_{22} f_i(x_i,y_i')\right\| \leq L\|y_i - y_i'\|, \quad \forall x_i, y_i, y_i'.$$

  It should be noted that if $y_i = y_i'$, this assumption becomes $\nabla^2_{22} f_i(x_i,y_i) = \frac{1}{n}\sum_{i=1}^n \nabla^2_{22} f_i(x_i,y_i)$ for all i.

- (SimFBO, Assumption 4 in [32]) 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 function and the global function. Our algorithm does not need any heterogeneity assumptions like these.

Main contributions. Our contributions in this work lie in several folds. First, we propose a single-loop algorithm for DBO. Our algorithm has two main features: (i) it is of single-loop structure; (ii) it does not involve heavy computation of matrix-vector multiplications. Second, we provide a convergence rate analysis for our proposed algorithm without requiring 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 $\nabla^2_{12} f(x,y)$ and $\nabla^2_{22} f(x,y)$ respectively. If there is no further specification, it is assumed that $\|\cdot\|$ denotes the ℓ_2 norm for vectors and the Frobenius norm for matrices. The operator norm of matrix Z is denoted as $\|Z\|_{\text{op}}$.

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.

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 [11, 16, 5, 6, 15], and Assumption 2.2 has been utilized in [27, 23, 7, 5, 6, 22]. These assumptions are established as standard in the literature and are taken as the basis for this paper.

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

(a) For any fixed x, $f_i(x,\cdot)$ is σ-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 ∇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 ∇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 nodes i and j are not connected. Moreover, we assume that W is symmetric and doubly stochastic, i.e. $W = W^{\mathsf{T}}$ 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 $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n$ and $\rho := \max\{|\lambda_2|, |\lambda_n|\} < 1$.*

2.2 The Proposed SLDBO Algorithm

Our aim is to extend the idea of SOBA [8] to deal with the DBO (6), which poses a significant challenge, particularly without imposing any heterogeneity assumptions. To address this, we propose to project y^k and v^k onto Euclidean balls with controllable radii to manage the consensus error. Remarkably, these radii can be pre-defined constants, ensuring the convergence of our algorithm. Before introducing our SLDBO algorithm, we define the following constants:

$$r_v := L_{F,0}/\sigma, \quad B_1 := L_{F,0} + L_{f,1} r_v, \quad \rho_1 := \frac{1-\rho^2}{\rho} + \frac{4}{1-\rho} + \frac{16}{(1-\rho)^3},$$
$$r_x := \sqrt{2\frac{\rho_1}{1-\rho} B_1^2 n \bar{\alpha}^2 + 2\sum_{i=1}^{n} \|x_i^0\|^2}, \quad r_y := \frac{L_{f,1} r_x}{\sigma} + \frac{1}{\sigma}\|\nabla_2 f(0,0)\|, \tag{8}$$

where x_i^0, $i = 1, \ldots, n$, are the initial points of our algorithm, $\bar{\alpha} > 0$ is an arbitrary constant, and other constants such as σ, ρ, $L_{F,0}$, $L_{f,1}$ are defined in Assumptions 2.1 and 2.2. We also define the projection operator $\mathcal{P}_r$, which projects a given point onto a Euclidean ball with radius $r \geq 0$, as

$$\mathcal{P}_r[z] := \arg\min_{\|z'\| \leq r} \|z' - z\| = \min\{1, r/\|z\|\}\, z.$$

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

- Initial points $d_{x,i}^{-1} = d_{y,i}^{-1} = d_{v,i}^{-1} = t_{x,i}^{-1} = t_{y,i}^{-1} = t_{v,i}^{-1} = 0$ $(i = 1, 2, \ldots, n)$.
- Initial points $x_i^{-1} = x_j^{-1}$, $y_i^{-1} = y_j^{-1}$, $v_i^{-1} = v_j^{-1}$, $i, j \in \{1, 2, \ldots, n\}$, satisfying $\|y_i^{-1}\| \leq r_y$ and $\|v_i^{-1}\| \leq r_v$, where r_v and r_y are defined in (8). Note that both inequalities can easily be satisfied by choosing $y_i^{-1} = 0$ and $v_i^{-1} = 0$, for instance.
- Compute initial points

$$x_i^0 = \sum_{j=1}^{n} w_{ij}(x_j^{-1} - \alpha t_{x,j}^{-1}) = x_i^{-1},$$
$$y_i^0 = \mathcal{P}_{r_y}\left[\textstyle\sum_{j=1}^{n} w_{ij}(y_j^{-1} - \alpha t_{y,j}^{-1})\right] = y_i^{-1},$$
$$v_i^0 = \mathcal{P}_{r_v}\left[\textstyle\sum_{j=1}^{n} w_{ij}(v_j^{-1} - \alpha t_{v,j}^{-1})\right] = v_i^{-1}.$$

BOX 1: Initial points of Algorithm 1.

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 $W = (w_{ij})$. In this regard, we employ the adapt-then-combine diffusion strategy, which has been demonstrated to perform better in practical numerical experiments [29], rather than the combine-then-adapt diffusion strategy.

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

Input: Define r_y and r_v as in (8) and let K be the maximum iteration number. Set initial points as in BOX 1, as well as step sizes $\alpha = \frac{\bar{\alpha}}{K+1}$, $\beta = \frac{\bar{\beta}}{\sqrt{K+1}}$ and $\eta = \frac{\bar{\eta}}{\sqrt{K+1}}$ with $\bar{\alpha} > 0$, $0 < \bar{\beta} < \frac{2}{\sigma + L_{f,1}}$ and $0 < \bar{\eta} < \frac{1}{L_{f,1}}$.

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

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

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

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

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

$$t_{y,i}^k = \sum_{j=1}^n w_{ij} t_{y,j}^{k-1} + d_{y,i}^k - d_{y,i}^{k-1}, \quad y_i^{k+1} = \mathcal{P}_{r_y}\left[\sum_{j=1}^n w_{ij}(y_j^k - \beta t_{y,j}^k)\right]; \tag{12}$$

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

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

end for

end for

(iii) The projection steps outlined in (12) and (13) are essential for ensuring that the sequences $\{y_i^k\}$ and $\{v_i^k\}$ remain bounded. In contrast, we can omit a projection step for the update of x_i^{k+1} since it can be shown that $\{x_i^k\}$ remains bounded when $\{y_i^k\}$ and $\{v_i^k\}$ are bounded.

(iv) 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 method in distributed optimization literature [21, 27, 23]. The tracking sequences $t_{x,i}^k$, $t_{y,i}^k$, and $t_{v,i}^k$ have been introduced to reduce the consensus error and to accelerate the convergence rate.

3 Convergence Rate Results for SLDBO

In this section, we provide the convergence results for Algorithm 1. First, we show that the boundedness of $\{y_i^k\}$ and $\{v_i^k\}$ results in the boundedness of $\{x_i^k\}$, which subsequently implies the boundedness of $\{d_{x,i}^k\}$, $\{d_{y,i}^k\}$, and $\{d_{v,i}^k\}$. These outcomes are pivotal for proving the convergence rate of SLDBO (Algorithm 1). We then construct a Lyapunov function, and 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, and the proof is postponed to the Appendix.

Theorem 3.1. *For any integer $K \geq 0$, 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) Consensus Error. *For any $0 \leq k \leq K$, we have*

$$\frac{1}{n}\sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 = O\left(\frac{1}{K^3}\right), \ \frac{1}{n}\sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 = O\left(\frac{1}{K^2}\right), \ \frac{1}{n}\sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 = O\left(\frac{1}{K^2}\right).$$

(b) Stationarity. *There exists an absolute constant K_0 such that for all $K \geq K_0$, we have*

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

According to Part (b) of Theorem 3.1, the convergence rate on stationarity is sublinear, with a rate of $O(1/K)$. This finding is consistent with the results reported in [15]. At present, this is the best known convergence rate result that has been achieved for both BO and DBO algorithms, as far as we know.

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 $i, j = 1, \ldots, n$, $w_{ij} = w$ 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 first conduct logistic regression with ℓ_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 node i, $i = 1, \ldots, 8$, we have

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

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

where $\mathcal{D}_i$ and $\mathcal{D}_i'$ denote the training and testing datasets on node i, respectively. We aim to identify the optimal hyperparameter λ such that $\omega^*(\lambda)$ represents the optimal model parameter corresponding to λ. To achieve this, we utilize synthetic heterogeneous data, generated in the same manner as in [6]. Specifically, the data distribution of x_e on node i follows a normal distribution with mean 0 and variance $i^2 \cdot r^2$, where r is the heterogeneity rate. In our experiments, we set r to 1. For the response variable, we let $y_e = x_e^\top w + 0.1z$, where z is sampled from the standard normal distribution.

In our experiments, we compare SLDBO to MA-DSBO [6]. 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 and r_y to be 20, and $\bar{\alpha}$ and $\bar{\eta}$ to be 0.5. Additionally, we use a value of 1.2 for $\bar{\beta}$. 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, particularly when $T = N = 2$. Note that MA-DSBO requires sufficient inner-loop iterations to accurately estimate the LL solution and the hypergradient. In addition, as the dimension of the data increases, SLDBO exhibits a more significant improvement in convergence rate. Its single-loop structure proves advantageous in terms

of savings in the number of required matrix-vector products, which becomes particularly beneficial when dealing with high-dimensional data.

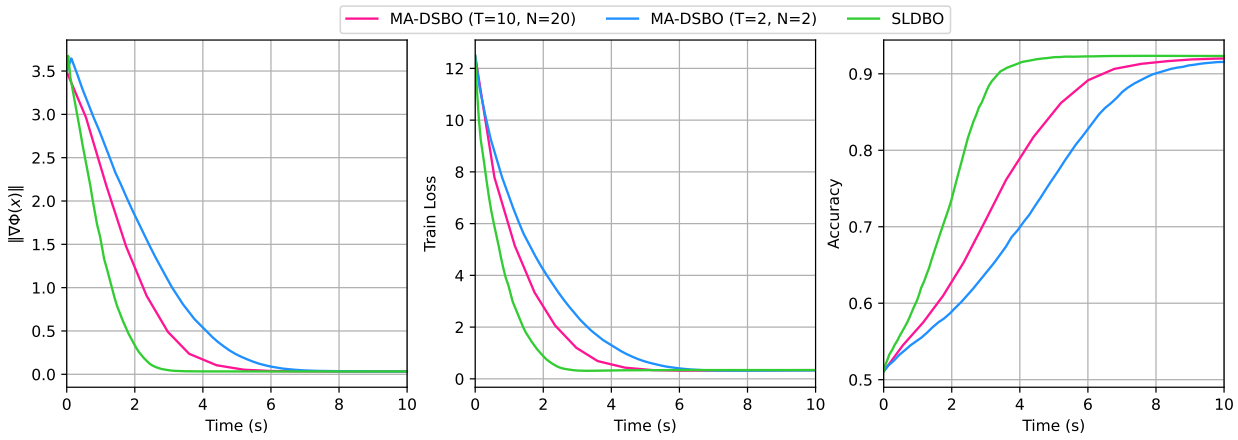

Figure 1: Comparison between MA-DSBO and SLDBO on synthetic data ($p = 50$).

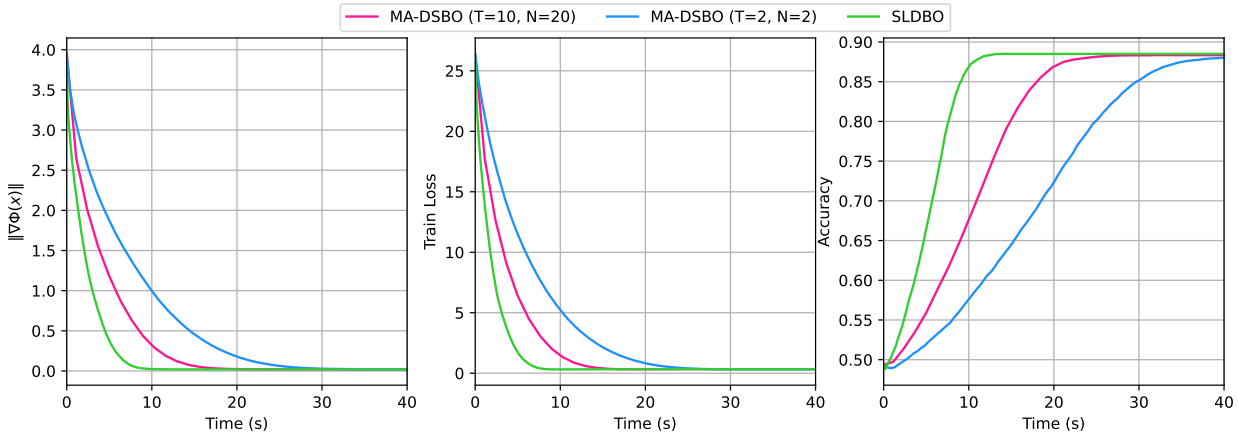

Figure 2: Comparison between MA-DSBO and SLDBO on synthetic data ($p = 200$).

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 node i. We next apply SLDBO to solve the following hyperparameter problem using the MNIST database [18]:

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

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

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, where 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 node to 1,000. In the case of SLDBO, the hyperparameters were set as follows: $r_v = r_y = 100$, $\bar{\alpha} = \bar{\eta} = 6$, and $\bar{\beta} = 15$. For MA-DSBO, we set $T = N = 5$. The comparison of test loss, train loss, and classification accuracy between the SLDBO and MA-DSBO algorithms is illustrated in Figure 3. These results demonstrate that our proposed algorithm SLDBO can efficiently solve this problem with improved convergence rate and classification accuracy.

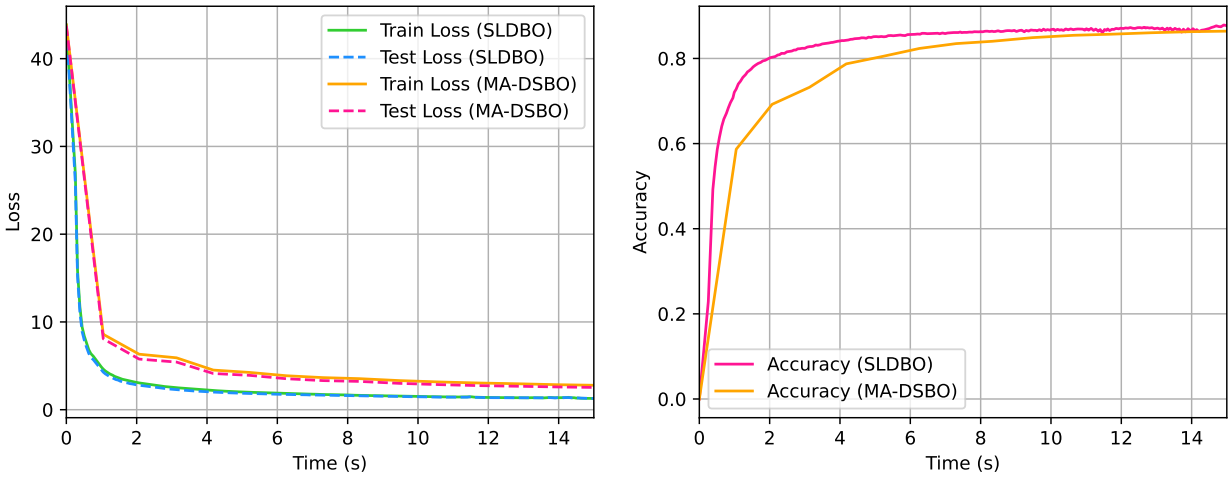

Figure 3: Comparison of test loss, train loss, and classificaiton accuracy between MA-DSBO and SLDBO on real-world MNIST dataset.

5 Conclusion

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 and does not make any assumptions related to heterogeneity. Our numerical experiments confirm the effectiveness of SLDBO. Nevertheless, we would like to acknowledge that computing the full gradient during practical applications can be time-consuming. It is therefore worth investigating the extension of our algorithm to the stochastic setting. Furthermore, it is important to emphasize that our algorithm is unable to achieve linear speedup in terms of the number of agents, and we will address this in a future work.

References

[1] A. Beck. *First-order methods in optimization*, volume 25. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2017. 24

[2] N. Bishop, L. Tran-Thanh, and E. Gerding. Optimal learning from verified training data. In *Advances in Neural Information Processing Systems*, volume 33, pages 9520–9529. Curran Associates, Inc., 2020. 1

[3] L. Chen, Y. Ma, and J. Zhang. Near-optimal fully first-order algorithms for finding stationary points in bilevel optimization. *arXiv preprint arXiv:2306.14853*, 2023. 2

[4] T. Chen, Y. Sun, and W. Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *Advances in Neural Information Processing Systems*, volume 34, pages 25294–25307. Curran Associates, Inc., 2021. 2

[5] X. Chen, M. Huang, and S. Ma. Decentralized bilevel optimization. *arXiv preprint arXiv:2206.05670*, 2022. 3, 4

[6] X. Chen, M. Huang, S. Ma, and K. Balasubramanian. Decentralized stochastic bilevel optimization with improved per-iteration complexity. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 4641–4671. PMLR, 2023. 3, 4, 7

[7] W. Choi and J. Kim. On the convergence analysis of the decentralized projected gradient descent. *arXiv preprint arXiv:2303.08412*, 2023. 4, 15

[8] M. Dagréou, P. Ablin, S. Vaiter, and T. Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. In *Advances in Neural Information Processing Systems*, volume 35, pages 26698–26710. Curran Associates, Inc., 2022. 2, 5

[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. 7

[10] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1568–1577. PMLR, 2018. 1, 2

[11] S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018. 2, 4, 24

[12] R. Grazzi, L. Franceschi, M. Pontil, and S. Salzo. On the iteration complexity of hypergradient computation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 3748–3758. PMLR, 2020. 2

[13] 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. 2

[14] K. Ji, J. D. Lee, Y. Liang, and H. V. Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. In *Advances in Neural Information Processing Systems*, volume 33, pages 11490–11500. Curran Associates, Inc., 2020. 1, 2

[15] K. Ji, M. Liu, Y. Liang, and L. Ying. Will bilevel optimizers benefit from loops. In *Advances in Neural Information Processing Systems*, volume 35, pages 3011–3023. Curran Associates, Inc., 2022. 2, 4, 7

[16] K. Ji, J. Yang, and Y. Liang. Bilevel optimization: Convergence analysis and enhanced design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 4882–4892. PMLR, 2021. 2, 4

[17] J. Kwon, D. Kwon, S. Wright, and R. D. Nowak. A fully first-order method for stochastic bilevel optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 18083–18113. PMLR, 2023. 2

[18] 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. 8

[19] 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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3

[20] 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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 21839–21866. PMLR, 2023. 2, 14

[21] 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. 6

[22] 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 *Advances in Neural Information Processing Systems*, volume 35, pages 30638–30650. Curran Associates, Inc., 2022. 3, 4

[23] 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. 4, 6

[24] Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, Cham, second edition, 2018. 26

[25] F. Pedregosa. Hyperparameter optimization with approximate gradient. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 737–746, New York, New York, USA, 2016. PMLR. 1, 2

[26] S. Pu and A. Nedić. Distributed stochastic gradient tracking methods. *Math. Program.*, 187(1-2):409–457, 2021. 3, 14

[27] G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Trans. Control Netw. Syst.*, 5(3):1245–1260, 2018. 4, 6

[28] A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1

[29] A. H. Sayed. Adaptation, learning, and optimization over networks. *Foundations and Trends® in Machine Learning*, 7(4-5):311–801, 2014. 5

[30] A. Shaban, C.-A. Cheng, N. Hatch, and B. Boots. Truncated back-propagation for bilevel optimization. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 1723–1732. PMLR, 2019. 2

[31] J. Wang, H. Chen, R. Jiang, X. Li, and Z. Li. Fast algorithms for Stackelberg prediction game with least squares loss. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 10708–10716. PMLR, 2021. 1

[32] Y. Yang, P. Xiao, and K. Ji. SimFBO: Towards simple, flexible and communication-efficient federated bilevel learning. *arXiv preprint arXiv:2305.19442*, 2023. 3, 4

[33] 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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 26693–26712. PMLR, 2022. 1

A Proof of the Convergence Results

In this section we provide the proof of convergence results. Assumption 2.1 and 2.2 are used throughout the proofs.

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

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

- $\bar{x}^k = \frac{1}{n}\sum_{i=1}^n x_i^k$, $\bar{y}^k = \frac{1}{n}\sum_{i=1}^n y_i^k$, $\bar{v}^k = \frac{1}{n}\sum_{i=1}^n v_i^k$.
- $\bar{d}_x^k = \frac{1}{n}\sum_{i=1}^n d_{x,i}^k$, $\bar{d}_y^k = \frac{1}{n}\sum_{i=1}^n d_{y,i}^k$, $\bar{d}_v^k = \frac{1}{n}\sum_{i=1}^n d_{v,i}^k$.
- $\bar{t}_x^k = \frac{1}{n}\sum_{i=1}^n t_{x,i}^k$, $\bar{t}_y^k = \frac{1}{n}\sum_{i=1}^n t_{y,i}^k$, $\bar{t}_v^k = \frac{1}{n}\sum_{i=1}^n t_{v,i}^k$.

We now give a list of constants and their definitions that we will use in our proof. Note that all these constants are absolute constants – they do not depend on the maximum iteration number K of the Algorithm 1.

$$
\begin{aligned}
r_v &:= L_{F,0}/\sigma, \\
B_1 &:= L_{F,0} + L_{f,1} r_v, \\
\rho_1 &:= \frac{1-\rho^2}{\rho} + \frac{4}{1-\rho} + \frac{16}{(1-\rho)^3}, \\
r_x &:= \sqrt{2\frac{\rho_1}{1-\rho} B_1^2 n \bar{\alpha}^2 + 2\sum_{i=1}^n \|x_i^0\|^2}, \\
\rho_2 &:= \left(\frac{2(1-\rho^2)}{\rho} + \frac{4+8\rho^3+8\rho^4}{1-\rho}\right)\left(1 + \frac{4}{(1-\rho)^2}\right), \\
r_y &:= \frac{L_{f,1} r_x}{\sigma} + \frac{1}{\sigma}\|\nabla_2 f(0,0)\|, \\
B_2 &:= L_{f,1}(r_x + r_y) + \max_{i=1,\ldots,n}\{\|\nabla_2 f_i(0,0)\|\}, \\
C_1 &:= 3\max\{(L_{F,1} + r_v L_{f,2})^2, L_{f,1}^2\}, \\
C_2 &:= L_{f,1}^2.
\end{aligned}
\tag{15}
$$

The Roadmap of the Proof. Here we briefly discuss of the roadmap of the proof of Theorem 3.1. First, we define the Lyapunov function

$$
V_k = a\left[F(\bar{x}^k, y^*(\bar{x}^k)) - F^*\right] + b\|\bar{y}^k - y^*(\bar{x}^k)\|^2 + c\|\bar{v}^k - v^*(\bar{x}^k)\|^2,
$$

with appropriately chosen positive coefficients a, b, c. We will manage to show the following inequality:

$$
\begin{aligned}
&V_{k+1} - V_k \\
\leq &-\frac{a\alpha}{2}\|\nabla\Phi(\bar{x}^k)\|^2 - \widehat{\alpha}\|\bar{x}^{k+1} - \bar{x}^k\|^2 - \widehat{\beta}\|\bar{y}^k - y^*(\bar{x}^k)\|^2 - \widehat{\eta}\|\bar{v}^k - v^*(\bar{x}^k)\|^2 + E_k,
\end{aligned}
$$

where we can prove that $E_k = O(1/K)$, $\forall 0 \leq k \leq K$, and constants $\hat{\alpha}$, $\hat{\beta}$, $\hat{\eta}$ are all positive. By taking the telescoping sum of this inequality, we can prove

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

Remark A.1. *We point out that the construction of this Lyapunov function and parts of our proof were inspired by [20].*

We begin the proof by presenting a few useful lemmas. Lemma A.1 is a common result in decentralized optimization (e.g., [26, Lemma 1]).

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

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

(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).** $\sum_{i=1}^n \left\|\sum_{j=1}^n w_{ij}x_j\right\|^2 \leq \sum_{i=1}^n \sum_{j=1}^n w_{ij}\|x_j\|^2 = \sum_{j=1}^n \sum_{i=1}^n w_{ij}\|x_j\|^2 = \sum_{i=1}^n \|x_i\|^2$, where the inequality follows from the convexity of $\|\cdot\|^2$, the last equality follows from $\sum_{i=1}^n w_{ij} = 1$.
(b). Let

$$X = \begin{pmatrix} - & x_1^\top & - \\ - & x_2^\top & - \\ & \vdots & \\ - & x_n^\top & - \end{pmatrix} \in \mathbb{R}^{n\times d},$$

we have

$$\left\|WX - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^\top X\right\|^2 \leq \left\|W - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^\top\right\|_{\text{op}}^2 \left\|X - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^\top X\right\|^2, \tag{16}$$

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

$$W = P\text{diag}(\lambda_i)P^{-1},\ \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^\top = P\text{diag}(1, 0, \cdots, 0)P^{-1},$$

where λ_i $(i = 1, \ldots, n)$ are eigenvalues of W, satisfying $1 = \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n$ and $\rho := \max\{|\lambda_2|, |\lambda_n|\} < 1$, thus

$$\left\|W - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^\top\right\|_{\text{op}} = \left\|P\left(\text{diag}(\lambda_i) - \text{diag}(1, 0, \ldots, 0)\right)P^{-1}\right\|_{\text{op}} \leq \rho. \tag{17}$$

Moreover, we have

$$\left\|WX - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^\top X\right\|^2 = \left\|WX - \mathbf{1_n}\bar{x}^\top\right\|^2 = \sum_{i=1}^n \left\|\sum_{j=1}^n w_{ij}\,(x_j - \bar{x})\right\|^2 \tag{18}$$

and

$$\left\|X - \frac{1}{n}\mathbf{1_n}\mathbf{1_n}^\top X\right\|^2 = \left\|X - \mathbf{1_n}\bar{x}^\top\right\|^2 = \sum_{i=1}^n \|x_i - \bar{x}\|^2 . \tag{19}$$

We can get the desired result by combining (16), (17), (18) and (19). □

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 $\mathcal{P}_r[z] := \operatorname{argmin}_{\{z':\|z'\|\in[0,r]\}}\|z' - z\|$.

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

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

Proof. Consider the function

$$G(x) = \sum_{i=1}^n \left\|\mathcal{P}_r\left[x_i\right] - x\right\|^2 ,$$

which is minimized by $x = \frac{1}{n}\sum_{i=1}^n \mathcal{P}_r\left[x_i\right]$. Therefore, we have

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

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

A.2 Consensus Error of Algorithm 1 (part (a) of Theorem 3.1)

The main result in this subsection is Theorem A.1, which is the same as part (a) of Theorem 3.1. We first prove a few useful lemmas.

Lemma A.3. *The sequences $\{t_{x,i}^k\}$ and $\{d_{x,i}^k\}$ generated by Algorithm 1 satisfy*

$$\sum_{i=1}^n \|t_{x,i}^{k+1} - \bar{t}_x^{k+1}\|^2 \le \frac{1}{1-\rho}\sum_{l=0}^{k+1}\left(\rho^{k+1-l}\sum_{i=1}^n \|d_{x,i}^l - d_{x,i}^{l-1}\|^2\right), \tag{20}$$

and

$$\sum_{i=1}^n \|t_{x,i}^{k+1}\|^2 \le \frac{1}{1-\rho}\sum_{l=0}^{k+1}\left(\rho^{k+1-l}\sum_{i=1}^n \|d_{x,i}^l - d_{x,i}^{l-1}\|^2\right) + n\|\bar{d}_x^{k+1}\|^2. \tag{21}$$

Proof. From the update of x_i^k and $t_{x,i}^k$ in (14), we have

$$\bar{t}_x^k = \bar{t}_x^{k-1} + \bar{d}_x^k - \bar{d}_x^{k-1}, \quad \bar{t}_x^{-1} = \bar{d}_x^{-1} = 0, \quad \bar{x}^{k+1} = \bar{x}^k - \alpha\bar{t}_x^k,$$

which implies (by induction)

$$\bar{t}_x^k = \bar{d}_x^k, \quad \bar{x}^{k+1} = \bar{x}^k - \alpha\bar{d}_x^k. \tag{22}$$

Therefore, we have

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

where the first inequality is due to Cauchy-Schwarz inequality, and the second inequality follows from Lemma A.1. By $t_{x,i}^{-1}=0$, we can deduce that

$$\sum_{i=1}^{n}\|t_{x,i}^{k+1}-\bar{t}_x^{k+1}\|^2\le\frac{1}{1-\rho}\sum_{l=0}^{k+1}\left(\rho^{k+1-l}\sum_{i=1}^{n}\|d_{x,i}^{l}-d_{x,i}^{l-1}\|^2\right),$$

which proves (20). The inequality (21) follows immediately by noticing

$$\sum_{i=1}^{n}\|t_{x,i}^{k+1}\|^2=\sum_{i=1}^{n}\|t_{x,i}^{k+1}-\bar{t}_x^{k+1}\|^2+n\|\bar{t}_x^{k+1}\|^2.$$

□

Remark A.2. *Following similar steps, we have*

$$\sum_{i=1}^{n}\|t_{v,i}^{k+1}-\bar{t}_v^{k+1}\|^2\le\frac{1}{1-\rho}\sum_{l=0}^{k+1}\left(\rho^{k+1-l}\sum_{i=1}^{n}\|d_{v,i}^{l}-d_{v,i}^{l-1}\|^2\right),\tag{23}$$

$$\sum_{i=1}^{n}\|t_{y,i}^{k+1}-\bar{t}_y^{k+1}\|^2\le\frac{1}{1-\rho}\sum_{l=0}^{k+1}\left(\rho^{k+1-l}\sum_{i=1}^{n}\|d_{y,i}^{l}-d_{y,i}^{l-1}\|^2\right).\tag{24}$$

The proof is omitted for brevity.

The following lemma proves the boundedness of $\|x_i^k\|$.

Lemma A.4. *The sequence of $\{x_i^k\}$ generated by Algorithm 1 satisfies*

$$\sum_{i=1}^{n}\|x_i^{k+1}-x_i^{k}\|^2=O(\alpha^2).\tag{25}$$

Moreover, x_i^k is uniformly bounded by r_x, i.e., $\|x_i^k\|\le r_x$, $\forall i,k$, where r_x is defined in (15).

Proof. By Assumption 2.1, we have

$$\begin{aligned}\|d_{x,i}^k\| &= \|\nabla_1 F_i(x_i^k, y_i^k) - \nabla_{12}^2 f_i(x_i^k, y_i^k) v_i^k\| \\ &\le \|\nabla_1 F_i(x_i^k, y_i^k)\| + \|\nabla_{12}^2 f_i(x_i^k, y_i^k)\| \|v_i^k\| \le L_{F,0} + L_{f,1} r_v = B_1,\end{aligned} \tag{26}$$

and

$$\|\bar{d}_x^k\| \le \frac{1}{n} \sum_{i=1}^n \|d_{x,i}^k\| \le B_1. \tag{27}$$

By (21), (26) and (27), we can deduce that

$$\sum_{i=1}^n \|t_{x,i}^{k+1}\|^2 \le \frac{4nB_1^2}{1-\rho} \sum_{l=0}^{k+1} \left(\rho^{k+1-l}\right) + nB_1^2 \le \left(1 + \frac{4}{(1-\rho)^2}\right) B_1^2 n. \tag{28}$$

From (14), we have

$$\begin{aligned}\sum_{i=1}^n \|x_i^{k+1} - x_i^k\|^2 &\le \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij}(x_j^k - \alpha t_{x,j}^k) - \sum_{j=1}^n w_{ij}(x_j^{k-1} - \alpha t_{x,j}^{k-1}) \right\|^2 \\ &\le \frac{1}{\rho} \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij}(x_j^k - x_j^{k-1}) \right\|^2 + \frac{1}{1-\rho} \alpha^2 \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij}(t_{x,j}^k - t_{x,j}^{k-1}) \right\|^2 \\ &\le \frac{1}{\rho} \sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij}(x_j^k - x_j^{k-1}) \right\|^2 + \frac{1}{1-\rho} \alpha^2 \sum_{i=1}^n \|t_{x,i}^k - t_{x,i}^{k-1}\|^2,\end{aligned} \tag{29}$$

where the second inequality is due to the Cauchy-Schwarz inequality, and the last inequality follows from Lemma A.1 (a). By Lemma A.1 (b), we have

$$\sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} x_j^k - \sum_{j=1}^n w_{ij} x_j^{k-1} - \left(\bar{x}^k - \bar{x}^{k-1}\right) \right\|^2 \le \rho^2 \sum_{i=1}^n \|x_i^k - x_i^{k-1} - \left(\bar{x}^k - \bar{x}^{k-1}\right)\|^2. \tag{30}$$

Moreover, we have

$$\begin{aligned}&\sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} x_j^k - \sum_{j=1}^n w_{ij} x_j^{k-1} - \left(\bar{x}^k - \bar{x}^{k-1}\right) \right\|^2 \\ =&\sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} x_j^k - \sum_{j=1}^n w_{ij} x_j^{k-1} \right\|^2 + n\|\bar{x}^k - \bar{x}^{k-1}\|^2 - 2\sum_{i=1}^n \left\langle \sum_{j=1}^n w_{ij} x_j^k - \sum_{j=1}^n w_{ij} x_j^{k-1}, \bar{x}^k - \bar{x}^{k-1} \right\rangle \\ =&\sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} x_j^k - \sum_{j=1}^n w_{ij} x_j^{k-1} \right\|^2 - n\|\bar{x}^k - \bar{x}^{k-1}\|^2,\end{aligned} \tag{31}$$

and

$$\begin{aligned}&\sum_{i=1}^n \|x_i^k - x_i^{k-1} - \left(\bar{x}^k - \bar{x}^{k-1}\right)\|^2 = \sum_{i=1}^n \|x_i^k - x_i^{k-1}\|^2 + n\|\bar{x}^k - \bar{x}^{k-1}\|^2 - 2\sum_{i=1}^n \left\langle x_i^k - x_i^{k-1}, \bar{x}^k - \bar{x}^{k-1} \right\rangle \\ =&\sum_{i=1}^n \|x_i^k - x_i^{k-1}\|^2 - n\|\bar{x}^k - \bar{x}^{k-1}\|^2.\end{aligned} \tag{32}$$

By combining (30), (31), (32), we obtain

$$\sum_{i=1}^n \left\| \sum_{j=1}^n w_{ij} x_j^k - \sum_{j=1}^n w_{ij} x_j^{k-1} \right\|^2 \le \rho^2 \sum_{i=1}^n \|x_i^k - x_i^{k-1}\|^2 + (1-\rho^2) n \|\bar{x}^k - \bar{x}^{k-1}\|^2 \tag{33}$$
$$= \rho^2 \sum_{i=1}^n \|x_i^k - x_i^{k-1}\|^2 + n(1-\rho^2)\alpha^2 \|\bar{d}_x^{k-1}\|^2,$$

where the equality follows from (22). Together with (29), we have

$$\sum_{i=1}^n \|x_i^{k+1} - x_i^k\|^2 \le \rho \sum_{i=1}^n \|x_i^k - x_i^{k-1}\|^2 + n\frac{(1-\rho^2)}{\rho}\alpha^2 \|\bar{d}_x^{k-1}\|^2 + \frac{1}{1-\rho}\alpha^2 \sum_{i=1}^n \|t_{x,i}^k - t_{x,i}^{k-1}\|^2$$
$$\le \rho \sum_{i=1}^n \|x_i^k - x_i^{k-1}\|^2 + n\frac{(1-\rho^2)}{\rho}\alpha^2 \|\bar{d}_x^{k-1}\|^2 + \frac{2}{1-\rho}\alpha^2 \left(\sum_{i=1}^n \|t_{x,i}^k\|^2 + \sum_{i=1}^n \|t_{x,i}^{k-1}\|^2 \right),$$

which, combining with (28) and (27), yields that (recall $x_i^{-1} = x_i^0$, and ρ_1 is defined in (15))

$$\begin{aligned} \sum_{i=1}^n \|x_i^{k+1} - x_i^k\|^2 \le & \rho \sum_{i=1}^n \|x_i^k - x_i^{k-1}\|^2 + \rho_1 B_1^2 n \alpha^2 \\ \le & \rho^k \sum_{i=1}^n \|x_i^0 - x_i^{-1}\|^2 + \rho_1 B_1^2 n \alpha^2 \sum_{l=0}^k \rho^l \\ \le & \frac{\rho_1}{1-\rho} B_1^2 n \alpha^2 = O(\alpha^2). \end{aligned} \tag{34}$$

This proves (25). Since $\alpha = \frac{\bar{\alpha}}{K+1}$, we have

$$\sum_{i=1}^n \|x_i^k\|^2 \le 2\sum_{i=1}^n \|x_i^k - x_i^0\|^2 + 2\sum_{i=1}^n \|x_i^0\|^2 \le 2k \sum_{l=0}^{k-1} \sum_{i=1}^n \|x_i^{l+1} - x_i^l\|^2 + 2\sum_{i=1}^n \|x_i^0\|^2$$
$$\le \frac{\rho_1}{1-\rho} B_1^2 n \bar{\alpha}^2 \frac{2k^2}{(K+1)^2} + 2\sum_{i=1}^n \|x_i^0\|^2 \le 2\frac{\rho_1}{1-\rho} B_1^2 n \bar{\alpha}^2 + 2\sum_{i=1}^n \|x_i^0\|^2,$$

which implies $\|x_i^k\|^2 \le \sum_{i=1}^n \|x_i^k\|^2 \le r_x^2$, where r_x is defined in (15). □

Lemma A.5. *The sequence $\{v_i^k\}$ generated by Algorithm 1 satisfy*

$$\frac{1}{n}\sum_{i=1}^n \|v_i^{k+1} - \bar{v}^{k+1}\|^2 \le \frac{\rho^2}{(1-\rho)} \eta^2 \sum_{l=0}^k \left(\rho^{k-l} \frac{1}{n} \sum_{i=1}^n \|t_{v,i}^l - \bar{t}_v^l\|^2 \right). \tag{35}$$

Proof. First, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\|v_i^{k+1}-\bar{v}^{k+1}\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\left\|\mathcal{P}_{r_v}\left[\sum_{j=1}^{n}w_{ij}(v_j^k+\eta t_{v,j}^k)\right]-\frac{1}{n}\sum_{s=1}^{n}\mathcal{P}_{r_v}\left[\sum_{j=1}^{n}w_{sj}(v_j^k+\eta t_{v,j}^k)\right]\right\|^2\\
&\le \frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}w_{ij}(v_j^k+\eta t_{v,j}^k)-\frac{1}{n}\sum_{s=1}^{n}\sum_{j=1}^{n}w_{sj}(v_j^k+\eta t_{v,j}^k)\right\|^2\\
&\le \left(1+\frac{1-\rho}{\rho}\right)\frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}w_{ij}v_j^k-\bar{v}^k\right\|^2+(1+\frac{\rho}{1-\rho})\eta^2\frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}w_{ij}t_{v,j}^k-\bar{t}_v^k\right\|^2\\
&\le \rho\frac{1}{n}\sum_{i=1}^{n}\|v_i^k-\bar{v}^k\|^2+\frac{\rho^2}{(1-\rho)}\eta^2\frac{1}{n}\sum_{i=1}^{n}\|t_{v,i}^k-\bar{t}_v^k\|^2,\\
&\le \frac{\rho^2}{(1-\rho)}\eta^2\sum_{l=0}^{k}\left(\rho^{k-l}\frac{1}{n}\sum_{i=1}^{n}\|t_{v,i}^l-\bar{t}_v^l\|^2\right),
\end{aligned}$$

where the first inequality follows from Lemma A.2, the third inequality follows from Lemma A.1, the forth inequality follows from $v_i^0=\bar{v}^0, i=1,\dots,n$. □

Remark A.3. *Following similar steps, we have*

$$\frac{1}{n}\sum_{i=1}^{n}\|x_i^{k+1}-\bar{x}^{k+1}\|^2 \le \frac{\rho^2}{(1-\rho)}\alpha^2\sum_{l=0}^{k}\left(\rho^{k-l}\frac{1}{n}\sum_{i=1}^{n}\|t_{x,i}^l-\bar{t}_x^l\|^2\right), \tag{36}$$

$$\frac{1}{n}\sum_{i=1}^{n}\|y_i^{k+1}-\bar{y}^{k+1}\|^2 \le \frac{\rho^2}{(1-\rho)}\beta^2\sum_{l=0}^{k}\left(\rho^{k-l}\frac{1}{n}\sum_{i=1}^{n}\|t_{y,i}^l-\bar{t}_y^l\|^2\right). \tag{37}$$

The proof is omitted for brevity.

Lemma A.6. *The sequences of $\{v_i^k\}$ and $\{y_i^k\}$ generated by Algorithm 1 satisfy*

$$\sum_{i=1}^{n}\|v_i^{k+1}-v_i^k\|^2 = O(\eta^2), \tag{38}$$

$$\sum_{i=1}^{n}\|y_i^{k+1}-y_i^k\|^2 = O(\beta^2). \tag{39}$$

Proof. By Assumption 2.1, we have

$$\begin{aligned}
\|d_{v,i}^k\| &= \|\nabla_2 F_i(x_i^k,y_i^k)-\nabla_{22}^2 f_i(x_i^k,y_i^k)v_i^k\|\\
&\le \|\nabla_2 F_i(x_i^k,y_i^k)\|+\|\nabla_{22}^2 f_i(x_i^k,y_i^k)\|\|v_i^k\| \le L_{F,0}+L_{f,1}r_v = B_1.
\end{aligned}$$

Since the updates of $t_{x,i}^k=\sum_{j=1}^n w_{ij}t_{x,j}^{k-1}+d_{x,i}^k-d_{x,i}^{k-1}$ and $t_{v,i}^k=\sum_{j=1}^n w_{ij}t_{v,j}^{k-1}+d_{v,i}^k-d_{v,i}^{k-1}$ are of the same form, by replacing $t_{x,i}^k,\bar{d}_x^k,B_1$ with $t_{v,i}^k,\bar{d}_v^k,B_1$ respectively in (27) and (28), we can deduce that

$$\|\bar{d}_v^k\| \le B_1, \tag{40}$$

and

$$n\|\bar{t}_v^{k+1}\|^2 \le \sum_{i=1}^{n}\|t_{v,i}^{k+1}\|^2 \le \left(1+\frac{4}{(1-\rho)^2}\right)B_1^2 n. \tag{41}$$

From the updates in Algorithm 1, we have

$$\begin{aligned}\sum_{i=1}^{n}\|v_i^{k+1}-v_i^k\|^2 &\le \sum_{i=1}^{n}\left\|\mathcal{P}_{r_v}\left[\sum_{j=1}^{n}w_{ij}(v_j^k+\eta t_{v,j}^k)\right]-\mathcal{P}_{r_v}\left[\sum_{j=1}^{n}w_{ij}(v_j^{k-1}+\eta t_{v,j}^{k-1})\right]\right\|^2\\ &\le \sum_{i=1}^{n}\left\|\sum_{j=1}^{n}w_{ij}(v_j^k+\eta t_{v,j}^k)-\sum_{j=1}^{n}w_{ij}(v_j^{k-1}+\eta t_{v,j}^{k-1})\right\|^2\\ &\le \frac{1}{\rho}\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}w_{ij}(v_j^k-v_j^{k-1})\right\|^2+\frac{1}{1-\rho}\eta^2\sum_{i=1}^{n}\|t_{v,i}^k-t_{v,i}^{k-1}\|^2, \end{aligned} \tag{42}$$

where the third inequality follows from Cauchy-Schwarz inequality and Lemma A.1 (a), similar to (29).

Next, following the similar steps in (30), (31), (32) and (33), we can deduce that

$$\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}w_{ij}v_j^k-\sum_{j=1}^{n}w_{ij}v_j^{k-1}\right\|^2\le\rho^2\sum_{i=1}^{n}\|v_i^k-v_i^{k-1}\|^2+(1-\rho^2)n\|\bar{v}^k-\bar{v}^{k-1}\|^2. \tag{43}$$

Moreover,

$$\begin{aligned}\|\bar{v}^k-\bar{v}^{k-1}\|^2 &\le \left\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{P}_{r_v}\left[\sum_{j=1}^{n}w_{ij}(v_j^{k-1}+\eta t_{v,j}^{k-1})\right]-\bar{v}^{k-1}\right\|^2\\ &\le \frac{1}{n}\sum_{i=1}^{n}\left\|\mathcal{P}_{r_v}\left[\sum_{j=1}^{n}w_{ij}(v_j^{k-1}+\eta t_{v,j}^{k-1})\right]-\bar{v}^{k-1}\right\|^2\le\frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}w_{ij}(v_j^{k-1}+\eta t_{v,j}^{k-1})-\bar{v}^{k-1}\right\|^2\\ &\le \frac{2}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}w_{ij}v_j^{k-1}-\bar{v}^{k-1}\right\|^2+\eta^2\frac{2}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{n}w_{ij}t_{v,j}^{k-1}\right\|^2\\ &\le \frac{2\rho^2}{n}\sum_{i=1}^{n}\|v_i^{k-1}-\bar{v}^{k-1}\|^2+\eta^2\frac{2}{n}\sum_{i=1}^{n}\|t_{v,i}^{k-1}\|^2. \end{aligned} \tag{44}$$

Hence, combining (42), (43) and (44), we have

$$\begin{aligned}\sum_{i=1}^{n}\|v_i^{k+1}-v_i^k\|^2 &\le \rho\sum_{i=1}^{n}\|v_i^k-v_i^{k-1}\|^2+2\rho(1-\rho^2)\sum_{i=1}^{n}\|v_i^{k-1}-\bar{v}^{k-1}\|^2\\ &\quad+\frac{2(1-\rho^2)}{\rho}\eta^2\sum_{i=1}^{n}\|t_{v,i}^{k-1}\|^2+\frac{1}{1-\rho}\eta^2\sum_{i=1}^{n}\|t_{v,i}^k-t_{v,i}^{k-1}\|^2\\ &\le \rho\sum_{i=1}^{n}\|v_i^k-v_i^{k-1}\|^2+2\rho(1-\rho^2)\sum_{i=1}^{n}\|v_i^{k-1}-\bar{v}^{k-1}\|^2\\ &\quad+\eta^2\left(\frac{2(1-\rho^2)}{\rho}+\frac{2}{1-\rho}\right)\sum_{i=1}^{n}\|t_{v,i}^{k-1}\|^2+\eta^2\frac{2}{1-\rho}\sum_{i=1}^{n}\|t_{v,i}^k\|^2,\\ &\le \rho\sum_{i=1}^{n}\|v_i^k-v_i^{k-1}\|^2+2\rho(1-\rho^2)\sum_{i=1}^{n}\|v_i^{k-1}-\bar{v}^{k-1}\|^2+\left(\frac{2(1-\rho^2)}{\rho}+\frac{4}{1-\rho}\right)\left(1+\frac{4}{(1-\rho)^2}\right)nB_1^2\eta^2\\ &\le \rho\sum_{i=1}^{n}\|v_i^k-v_i^{k-1}\|^2+\rho_2 nB_1^2\eta^2, \end{aligned}$$

where ρ_2 is defined in (15), the third inequality uses (41), the fourth inequality follows from the fact that by (35) and (41), we have

$$\begin{aligned}\sum_{i=1}^n \|v_i^{k-1} - \bar{v}^{k-1}\|^2 &\le n\frac{\rho^2}{(1-\rho)}\eta^2 \sum_{l=0}^k \left(\rho^{k-l}\frac{1}{n}\sum_{i=1}^n \|t_{v,i}^l - \bar{t}_v^l\|^2\right) \\ &\le \frac{\rho^2}{(1-\rho)}\eta^2 \sum_{l=0}^k \left(\rho^{k-l}\left(2\sum_{i=1}^n \|t_{v,i}^l\|^2 + 2\sum_{i=1}^n \|\bar{t}_v^l\|^2\right)\right) \\ &\le \frac{4\rho^2}{(1-\rho)^2}\left(1 + \frac{4}{(1-\rho)^2}\right) nB_1^2\eta^2.\end{aligned}$$

Because $v_i^{-1} = v_i^0$, we have

$$\begin{aligned}\sum_{i=1}^n \|v_i^{k+1} - v_i^k\|^2 \le \rho\sum_{i=1}^n \|v_i^k - v_i^{k-1}\|^2 + \rho_2 nB_1^2\eta^2 &\le \rho^{k+1}\sum_{i=1}^n \|v_i^0 - v_i^{-1}\|^2 + \rho_2 nB_1^2\eta^2 \sum_{l=0}^k \rho^l \\ &\le \frac{\rho_2}{1-\rho} nB_1^2\eta^2, \end{aligned} \tag{45}$$

where the last inequality uses the initialization in Algorithm 1. This proves (38).

Since $\alpha = \frac{\bar{\alpha}}{K+1}$, by Assumption 2.1 and Lemma A.4, we have

$$\|\nabla_2 f_i(x_i^k, y_i^k) - \nabla_2 f_i(0,0)\|^2 \le L_{f,1}^2\left(\|x_i^k\|^2 + \|y_i^k\|^2\right) \le L_{f,1}^2\left(r_x^2 + r_y^2\right),$$

which implies

$$\begin{aligned}\|d_{y,i}^k\| &= \|\nabla_2 f_i(x_i^k, y_i^k)\| \\ &\le \|\nabla_2 f_i(x_i^k, y_i^k) - \nabla_2 f_i(0,0)\| + \|\nabla_2 f_i(0,0)\| \le L_{f,1}(r_x + r_y) + \|\nabla_2 f_i(0,0)\| \le B_2,\end{aligned}$$

where B_2 is defined in (15). Therefore, following similar steps in the above proof and replacing v_i^k, B_1, η with y_i^k, B_2, β in (45), we can deduce that

$$\sum_{i=1}^n \|y_i^{k+1} - y_i^k\|^2 \le \frac{\rho_2}{1-\rho} nB_2^2\beta^2, \tag{46}$$

which proves (39). □

Lemma A.7. *The sequences of $d_{x,i}^k$, $d_{v,i}^k$ and $d_{y,i}^k$ satisfy*

$$\sum_{i=1}^n \|d_{x,i}^k - d_{x,i}^{k-1}\|^2 \le C_1\sum_{i=1}^n \left(\|x_i^k - x_i^{k-1}\|^2 + \|y_i^k - y_i^{k-1}\|^2 + \|v_i^k - v_i^{k-1}\|^2\right), \tag{47}$$

$$\sum_{i=1}^n \|d_{v,i}^k - d_{v,i}^{k-1}\|^2 \le C_1\sum_{i=1}^n \left(\|x_i^k - x_i^{k-1}\|^2 + \|y_i^k - y_i^{k-1}\|^2 + \|v_i^k - v_i^{k-1}\|^2\right), \tag{48}$$

$$\sum_{i=1}^n \|d_{y,i}^k - d_{y,i}^{k-1}\|^2 \le C_2\sum_{i=1}^n \left(\|x_i^k - x_i^{k-1}\|^2 + \|y_i^k - y_i^{k-1}\|^2\right), \tag{49}$$

where C_1 and C_2 are defined in (15).

Proof. Note that

$$\begin{aligned}
&\|d_{x,i}^k - d_{x,i}^{k-1}\| \\
=&\|\nabla_1 F(x_i^k, y_i^k) - \nabla_{12}^2 f_i(x_i^k, y_i^k) v_i^k - \nabla_1 F(x_i^{k-1}, y_i^{k-1}) + \nabla_{12}^2 f_i(x_i^{k-1}, y_i^{k-1}) v_i^{k-1}\| \\
\le&\|\nabla_1 F(x_i^k, y_i^k) - \nabla_1 F(x_i^{k-1}, y_i^{k-1})\| + \|(\nabla_{12}^2 f_i(x_i^k, y_i^k) - \nabla_{12}^2 f_i(x_i^{k-1}, y_i^{k-1})) v_i^k\| \\
&+ \|\nabla_{12}^2 f_i(x_i^{k-1}, y_i^{k-1})(v_i^k - v_i^{k-1})\| \\
\le&(L_{F,1} + r_v L_{f,2})\|x_i^k - x_i^{k-1}\| + (L_{F,1} + r_v L_{f,2})\|y_i^k - y_i^{k-1}\| + L_{f,1}\|v_i^k - v_i^{k-1}\|,
\end{aligned}$$

so we have

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

Similarly, by the definition of $d_{v,i}^k$, we also have

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

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

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

□

Now we are ready to prove part (a) of Theorem 3.1, which gives the rate of consensus error.

Theorem A.1. *The sequence of $\{x_i^k, y_i^k, v_i^k\}$ generated by Algorithm 1 satisfies*

$$\frac{1}{n}\sum_{i=1}^n \|x_i^{k+1} - \bar{x}^{k+1}\|^2 = O\left(\alpha^2 \max\{\alpha^2, \beta^2, \eta^2\}\right) = O\left(\frac{1}{K^3}\right), \tag{50}$$

$$\frac{1}{n}\sum_{i=1}^n \|y_i^{k+1} - \bar{y}^{k+1}\|^2 = O\left(\beta^2 \max\{\alpha^2, \beta^2\}\right) = O\left(\frac{1}{K^2}\right), \tag{51}$$

$$\frac{1}{n}\sum_{i=1}^n \|v_i^{k+1} - \bar{v}^{k+1}\|^2 = O\left(\eta^2 \max\{\alpha^2, \beta^2, \eta^2\}\right) = O\left(\frac{1}{K^2}\right). \tag{52}$$

Proof. We know that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|x_i^{k+1}-\bar{x}^{k+1}\|^2 \le \frac{\rho^2}{(1-\rho)}\alpha^2\sum_{l=0}^{k}\left(\rho^{k-l}\frac{1}{n}\sum_{i=1}^{n}\|t_{x,i}^{l}-\bar{t}_x^{l}\|^2\right)\\
\le&\frac{\rho^2}{(1-\rho)}\alpha^2\sum_{l=0}^{k}\left(\rho^{k-l}\frac{1}{n}\frac{1}{1-\rho}\sum_{s=0}^{l}\left(\rho^{l-s}\sum_{i=1}^{n}\|d_{x,i}^{s}-d_{x,i}^{s-1}\|^2\right)\right)\\
\le&\frac{\rho^2}{(1-\rho)}\alpha^2\sum_{l=0}^{k}\left(\rho^{k-l}\frac{1}{n}\frac{C_1}{1-\rho}\sum_{s=0}^{l}\left(\rho^{l-s}\sum_{i=1}^{n}\left(\|x_i^{s}-x_i^{s-1}\|^2+\|y_i^{s}-y_i^{s-1}\|^2+\|v_i^{s}-v_i^{s-1}\|^2\right)\right)\right)\\
\le&\frac{\rho^2}{(1-\rho)}\alpha^2\sum_{l=0}^{k}\left[\rho^{k-l}\frac{C_1}{1-\rho}\sum_{s=0}^{l}\left(\rho^{l-s}\left(\frac{\rho_1}{1-\rho}B_1^2\alpha^2+\frac{\rho_2}{1-\rho}B_2^2\beta^2+\frac{\rho_2}{1-\rho}B_1^2\eta^2\right)\right)\right]\\
\le&\frac{\rho^2 C_1}{(1-\rho)^4}\alpha^2\left[\frac{\rho_1}{1-\rho}B_1^2\alpha^2+\frac{\rho_2}{1-\rho}B_2^2\beta^2+\frac{\rho_2}{1-\rho}B_1^2\eta^2\right]\\
=&O\left(\alpha^2\max\{\alpha^2,\beta^2,\eta^2\}\right),
\end{aligned}$$

where the first inequality uses (36), the second inequality uses (20), the third inequality uses (47), the fourth inequality uses (34), (46) and (45), the last inequality is by summing the geometric series. This proves (50).

Similarly, using (37), (24), (49), (34) and (46), we have

$$\frac{1}{n}\sum_{i=1}^{n}\|y_i^{k+1}-\bar{y}^{k+1}\|^2 = O\left(\beta^2\max\{\alpha^2,\beta^2\}\right),$$

using (35), (23), (48), (34), (46) and (45), we have

$$\frac{1}{n}\sum_{i=1}^{n}\|v_i^{k+1}-\bar{v}^{k+1}\|^2 = O\left(\eta^2\max\{\alpha^2,\beta^2,\eta^2\}\right).$$

This completes the proof. □

A.3 Convergence Rate of Algorithm 1 (part (b) of Theorem 3.1)

Recall that $\Phi(x)=F(x,y^*(x))$, we first establish the descent of $\Phi(\bar{x}^k)$.

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

$$\begin{aligned}
&\Phi(\bar{x}^{k+1})-\Phi(\bar{x}^k)\\
\le&-\frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k)\|^2-\frac{1}{2}\left(\frac{1}{\alpha}-L_\Phi\right)\|\bar{x}^{k+1}-\bar{x}^k\|^2\\
&+\frac{5}{2}\alpha L_{f,1}^2\|\bar{v}^k-v^*(\bar{x}^k)\|^2+\frac{5}{2}\alpha\left(L_{F,1}+L_{f,2}r_v\right)^2\|\bar{y}^k-y^*(\bar{x}^k)\|^2\\
&+\frac{5}{2}\alpha\left(L_{F,1}+r_vL_{f,2}\right)^2\frac{1}{n}\left(\sum_{i=1}^{n}\|x_i^k-\bar{x}^k\|^2+\sum_{i=1}^{n}\|y_i^k-\bar{y}^k\|^2\right)+\frac{5}{2}\alpha L_{f,1}^2\frac{1}{n}\sum_{i=1}^{n}\|v_i^k-\bar{v}^k\|^2, \qquad (53)
\end{aligned}$$

where L_Φ is a constant given by

$$L_\Phi := L_{F,1}+\frac{2L_{F,1}L_{f,1}+L_{f,2}L_{F,0}^2}{\sigma}+\frac{2L_{f,1}L_{F,0}L_{f,2}+L_{f,1}^2L_{F,1}}{\sigma^2}+\frac{L_{f,2}L_{f,1}^2L_{F,0}}{\sigma^3}.$$

Proof. From Lemma 2.2 in [11] we know that $\nabla\Phi(x)$ is L_Φ-Lipschitz continuous. By Lemma 5.7 in [1], we have

$$\begin{aligned}
&F(\bar{x}^{k+1}, y^*(\bar{x}^{k+1})) - F(\bar{x}^k, y^*(\bar{x}^k)) = \Phi(\bar{x}^{k+1}) - \Phi(\bar{x}^k) \\
\leq& \left\langle \nabla\Phi(\bar{x}^k), \bar{x}^{k+1} - \bar{x}^k \right\rangle + \frac{L_\Phi}{2}\|\bar{x}^{k+1} - \bar{x}^k\|^2 \\
=& -\alpha \left\langle \nabla\Phi(\bar{x}^k), \bar{d}_x^k \right\rangle + \frac{L_\Phi}{2}\|\bar{x}^{k+1} - \bar{x}^k\|^2 \\
=& -\frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k)\|^2 - \frac{\alpha}{2}\|\bar{d}_x^k\|^2 + \frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k) - \bar{d}_x^k\|^2 + \frac{L_\Phi}{2}\|\bar{x}^{k+1} - \bar{x}^k\|^2 \\
=& -\frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k)\|^2 - \frac{1}{2\alpha}\|\bar{x}^{k+1} - \bar{x}^k\|^2 + \frac{\alpha}{2}\|\nabla\Phi(\bar{x}^k) - \bar{d}_x^k\|^2 + \frac{L_\Phi}{2}\|\bar{x}^{k+1} - \bar{x}^k\|^2, \qquad (54)
\end{aligned}$$

where the second and the fourth equality use the update of $\bar{x}^{k+1}$ from (22). By the definition we get

$$\begin{aligned}
\nabla\Phi(\bar{x}^k) &= \nabla_1 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_{12}^2 f(\bar{x}^k, y^*(\bar{x}^k)) v^*(\bar{x}^k), \\
\bar{d}_x^k &= \frac{1}{n}\sum_{i=1}^n \left(\nabla_1 F_i(x_i^k, y_i^k) - \nabla_{12}^2 f_i(x_i^k, y_i^k) v_i^k\right),
\end{aligned}$$

and we have

$$\begin{aligned}
&\left\|\nabla\Phi(\bar{x}^k) - \bar{d}_x^k\right\| \\
\leq& \left\|\nabla_1 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_{12}^2 f(\bar{x}^k, y^*(\bar{x}^k)) v^*(\bar{x}^k) - \nabla_1 F(\bar{x}^k, \bar{y}^k) + \nabla_{12}^2 f(\bar{x}^k, \bar{y}^k)\bar{v}^k\right\| \\
&+ \left\|\nabla_1 F(\bar{x}^k, \bar{y}^k) - \nabla_{12}^2 f(\bar{x}^k, \bar{y}^k)\bar{v}^k - \frac{1}{n}\sum_{i=1}^n (\nabla_1 F_i(x_i^k, y_i^k) - \nabla_{12}^2 f_i(x_i^k, y_i^k) v_i^k)\right\|. \qquad (55)
\end{aligned}$$

For the first term on the right hand side of (55), we have

$$\begin{aligned}
&\left\|\nabla_1 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_{12}^2 f(\bar{x}^k, y^*(\bar{x}^k)) v^*(\bar{x}^k) - \nabla_1 F(\bar{x}^k, \bar{y}^k) + \nabla_{12}^2 f(\bar{x}^k, \bar{y}^k)\bar{v}^k\right\| \\
\leq& \|\nabla_1 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_1 F(\bar{x}^k, \bar{y}^k)\| + \left\|\left[\nabla_{12}^2 f(\bar{x}^k, \bar{y}^k) - \nabla_{12}^2 f(\bar{x}^k, y^*(\bar{x}^k))\right]\bar{v}^k\right\| \\
&+ \left\|\nabla_{12}^2 f(\bar{x}^k, y^*(\bar{x}^k))[\bar{v}^k - v^*(\bar{x}^k)]\right\| \\
\leq& (L_{F,1} + L_{f,2} r_v)\left\|\bar{y}^k - y^*(\bar{x}^k)\right\| + L_{f,1}\|\bar{v}^k - v^*(\bar{x}^k)\|, \qquad (56)
\end{aligned}$$

and for the second term on the right hand side of (55), we have

$$\begin{aligned}
&\left\|\nabla_1 F(\bar{x}^k, \bar{y}^k) - \nabla_{12}^2 f(\bar{x}^k, \bar{y}^k)\bar{v}^k - \frac{1}{n}\sum_{i=1}^n (\nabla_1 F_i(x_i^k, y_i^k) - \nabla_{12}^2 f_i(x_i^k, y_i^k) v_i^k)\right\| \\
\leq& \frac{1}{n}\sum_{i=1}^n \left\|\nabla_1 F_i(\bar{x}^k, \bar{y}^k) - \nabla_{12}^2 f_i(\bar{x}^k, \bar{y}^k)\bar{v}^k - \nabla_1 F_i(x_i^k, y_i^k) + \nabla_{12}^2 f_i(x_i^k, y_i^k) v_i^k)\right\| \\
\leq& \frac{1}{n}\sum_{i=1}^n \left\|\nabla_1 F_i(\bar{x}^k, \bar{y}^k) - \nabla_1 F_i(x_i^k, y_i^k)\right\| + \frac{1}{n}\sum_{i=1}^n \left\|\left[\nabla_{12}^2 f_i(\bar{x}^k, \bar{y}^k) - \nabla_{12}^2 f_i(x_i^k, y_i^k)\right] v_i^k\right\| \\
&+ \frac{1}{n}\sum_{i=1}^n \|[\nabla_{12}^2 f_i(\bar{x}^k, \bar{y}^k)](\bar{v}^k - v_i^k)\| \\
\leq& (L_{F,1} + r_v L_{f,2})\left(\frac{1}{n}\sum_{i=1}^n \|x_i^k - \bar{x}^k\| + \frac{1}{n}\sum_{i=1}^n \|y_i^k - \bar{y}^k\|\right) + L_{f,1}\frac{1}{n}\sum_{i=1}^n \|v_i^k - \bar{v}^k\|. \qquad (57)
\end{aligned}$$

Combining (55), (56), (57) and using the inequality $(\sum_{l=1}^s a_l)^2 \le s\sum_{l=1}^s a_l^2$, we have

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

which, together with (54), yields the desired result. □

The next lemma proves the boundedness of $v^*(x)$ and $y^*(x)$.

Lemma A.9. *For $y^*(x)$ and $v^*(x)$, we have (recall that r_v and r_y are defined in (15)):*

(a) $\|v^*(x)\| \le r_v$,

(b) $\|y^*(x)\| \le r_y$, *for any x satisfying $\|x\| \le r_x$.*

Proof. **(a).** By the $L_{F,0}$-Lipschitz continuity of F and σ-strong convexity of $f(x,\cdot)$ in Assumption 2.1, we can deduce that

$$\|\nabla_2 F(x, y^*(x))\| \le L_{F,0} \quad \text{and} \quad \left\|\left[\nabla_{22}^2 f(x, y^*(x))\right]^{-1}\right\|_{\text{op}} \le \frac{1}{\sigma}.$$

For $v^*(x) = \left[\nabla_{22}^2 f(x, y^*(x))\right]^{-1}\nabla_2 F(x, y^*(x))$, we have

$$\|v^*(x)\| \le \left\|\left[\nabla_{22}^2 f(x, y^*(x))\right]^{-1}\right\|_{\text{op}} \|\nabla_2 F(x, y^*(x))\| \le \frac{L_{F,0}}{\sigma}.$$

(b). Assume $\|x\| \le r_x$. By the optimality of $y^*(x)$, we know that $\nabla_2 f(x, y^*(x)) = 0$, and thus

$$\left(\nabla_2 f(x, y^*(x)) - \nabla_2 f(x, y^*(x'))\right) + \left(\nabla_2 f(x, y^*(x')) - \nabla_2 f(x', y^*(x'))\right) = 0.$$

Then by the strong convexity of $f(x,\cdot)$ and Lipschitz continuity of ∇f, we have

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

That is,

$$\|y^*(x) - y^*(x')\| \le \frac{L_{f,1}}{\sigma}\|x - x'\|. \tag{58}$$

Take $x' = 0$, we have

$$\begin{aligned}
\|y^*(x)\| &\le \frac{L_{f,1}}{\sigma}\|x - 0\| + \|y^*(0)\| = \frac{L_{f,1}}{\sigma}\|x\| + \|y^*(0) - 0\| \\
&\le \frac{L_{f,1}}{\sigma}\|x\| + \frac{1}{\sigma}\|\nabla_2 f(0, y^*(0)) - \nabla_2 f(0,0)\| \\
&\le \frac{L_{f,1}}{\sigma}r_x + \frac{1}{\sigma}\|\nabla_2 f(0,0)\|,
\end{aligned}$$

where the second inequality follows from the strong convexity of $f(x,\cdot)$, the last inequality follows from the optimality of $y^*(0)$. □

Next, we analyze the errors of $\bar{y}^k$ and $\bar{v}^k$.

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

$$\left\|\bar{y}^{k+1} - y^*(\bar{x}^k)\right\|^2 \leq \left(1 - \frac{1}{2}\beta\sigma\right)\|\bar{y}^k - y^*(\bar{x}^k)\|^2 + \frac{6\beta}{\sigma}L_{f,1}^2\left(\frac{1}{n}\sum_{i=1}^n\|\bar{x}^k - x_i^k\|^2 + \frac{1}{n}\sum_{i=1}^n\|\bar{y}^k - y_i^k\|^2\right) + 2\rho^2\frac{1}{n}\sum_{i=1}^n\|y_i^k - \bar{y}^k\|^2 + 2\rho^2\beta^2\frac{1}{n}\sum_{i=1}^n\|t_{y,i}^k - \bar{t}_y^k\|^2. \tag{59}$$

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

$$\begin{aligned}&\left\|\bar{y}^k - \beta\bar{d}_y^k - y^*(\bar{x}^k)\right\|^2 \\ \leq&(1+\xi)\left\|\bar{y}^k - \beta\nabla_2 f(\bar{x}^k, \bar{y}^k) - y^*(\bar{x}^k)\right\|^2 + (1 + 1/\xi)\beta^2\left\|\nabla_2 f(\bar{x}^k, \bar{y}^k) - \bar{d}_y^k\right\|^2.\end{aligned} \tag{60}$$

The first term on the right hand side of (60) is equal to

$$\begin{aligned}&\left\|\bar{y}^k - \beta\nabla_2 f(\bar{x}^k, \bar{y}^k) - y^*(\bar{x}^k)\right\|^2 \\ =&\|\bar{y}^k - y^*(\bar{x}^k)\|^2 - 2\beta\left\langle\bar{y}^k - y^*(\bar{x}^k), \nabla_2 f(\bar{x}^k, \bar{y}^k)\right\rangle + \beta^2\|\nabla_2 f(\bar{x}^k, \bar{y}^k)\|^2.\end{aligned}$$

Since $\nabla_2 f(\bar{x}^k, y^*(\bar{x}^k)) = 0$, by the σ-strong convexity of $f(\bar{x}^k, \cdot)$, and $L_{f,1}$-smoothness of f, Theorem 2.1.12 in [24] implies that

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

Since $\bar{\beta} \leq \frac{2}{\sigma + L_{f,1}}$ in Algorithm 1, we have

$$\left\|\bar{y}^k - \beta\nabla_2 f(\bar{x}^k, \bar{y}^k) - y^*(\bar{x}^k)\right\|^2 \leq \left(1 - 2\beta\frac{\sigma L_{f,1}}{\sigma + L_{f,1}}\right)\|\bar{y}^k - y^*(\bar{x}^k)\|^2 \leq (1 - \beta\sigma)\|\bar{y}^k - y^*(\bar{x}^k)\|^2. \tag{61}$$

Then we consider the second term on the right hand side of (60):

$$\begin{aligned}\left\|\nabla_2 f(\bar{x}^k, \bar{y}^k) - \bar{d}_y^k\right\| &\leq \frac{1}{n}\sum_{i=1}^n\left\|\nabla_2 f_i(\bar{x}^k, \bar{y}^k) - \nabla_2 f_i(x_i^k, y_i^k)\right\| \\ &\leq L_{f,1}\left(\frac{1}{n}\sum_{i=1}^n\|\bar{x}^k - x_i^k\| + \frac{1}{n}\sum_{i=1}^n\|\bar{y}^k - y_i^k\|\right).\end{aligned} \tag{62}$$

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

$$\begin{aligned}\sum_{i=1}^n\|y_i^{k+1} - y^*(\bar{x}^k)\|^2 &= \sum_{i=1}^n\left\|\mathcal{P}_{r_y}\left[\sum_{j=1}^n w_{ij}(y_j^k - \beta t_{y,j}^k)\right] - y^*(\bar{x}^k)\right\|^2 \leq \sum_{i=1}^n\left\|\sum_{j=1}^n w_{ij}(y_j^k - \beta t_{y,j}^k) - y^*(\bar{x}^k)\right\|^2 \\ &= \sum_{i=1}^n\left\|\sum_{j=1}^n w_{ij}y_j^k - \bar{y}^k - \beta\left(\sum_{j=1}^n w_{ij}t_{y,j}^k - \bar{t}_y^k\right)\right\|^2 + \sum_{i=1}^n\|\bar{y}^k - y^*(\bar{x}^k) - \beta\bar{t}_y^k\|^2,\end{aligned}$$

where the inequality holds because $\|y^*(\bar{x}^k)\| \le r_y$ by Lemma A.9, and the second equality holds because

$$\sum_{i=1}^n \left[\sum_{j=1}^n w_{ij} y_j^k - \bar{y}^k - \beta\left(\sum_{j=1}^n w_{ij} t_{y,j}^k - \bar{t}_y^k\right)\right] = 0.$$

Using Lemma A.1, we can deduce that

$$\begin{aligned}\sum_{i=1}^n \left\|\sum_{j=1}^n w_{ij} y_j^k - \bar{y}^k - \beta\left(\sum_{j=1}^n w_{ij} t_{y,j}^k - \bar{t}_y^k\right)\right\|^2 &\le 2\sum_{i=1}^n \left\|\sum_{j=1}^n w_{ij} y_j^k - \bar{y}^k\right\|^2 + 2\beta^2 \sum_{i=1}^n \left\|\sum_{j=1}^n w_{ij} t_{y,j}^k - \bar{t}_y^k\right\|^2 \\ &\le 2\rho^2 \sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 + 2\beta^2\rho^2 \sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2,\end{aligned}$$

which, together with (60), (61) and (62), yields that (recall $\bar{d}_y^k = \bar{t}_y^k$)

$$\begin{aligned}&\|\bar{y}^{k+1} - y^*(\bar{x}^k)\|^2 \le \frac{1}{n}\sum_{i=1}^n \|y_i^{k+1} - \bar{y}^{k+1}\|^2 + \|\bar{y}^{k+1} - y^*(\bar{x}^k)\|^2 = \frac{1}{n}\sum_{i=1}^n \|y_i^{k+1} - y^*(\bar{x}^k)\|^2 \\ &\le \|\bar{y}^k - y^*(\bar{x}^k) - \beta\bar{t}_y^k\|^2 + 2\rho^2\frac{1}{n}\sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 + 2\rho^2\beta^2\frac{1}{n}\sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2 \\ &\le (1+\xi)(1-\beta\sigma)\|\bar{y}^k - y^*(\bar{x}^k)\|^2 + (1+1/\xi)\beta^2 2L_{f,1}^2\left(\frac{1}{n}\sum_{i=1}^n \|\bar{x}^k - x_i^k\|^2 + \frac{1}{n}\sum_{i=1}^n \|\bar{y}^k - y_i^k\|^2\right) \\ &\quad + 2\rho^2\frac{1}{n}\sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 + 2\rho^2\beta^2\frac{1}{n}\sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2.\end{aligned}$$

Then we get the desired result by taking $\xi = \frac{1}{2}\beta\sigma$ and using $\beta\sigma \le 1$. □

Lemma A.11. *The sequence of (x_i^k, y_i^k, v_i^k) generated by Algorithm 1 satisfies*

$$\begin{aligned}&\|\bar{v}^{k+1} - v^*(\bar{x}^k)\|^2 \\ &\le \left(1 - \frac{1}{2}\eta\sigma\right)\left\|\bar{v}^k - v^*(\bar{x}^k)\right\|^2 + \frac{3\eta}{\sigma}\left(L_{F,1} + L_{f,2}r_v\right)^2\left\|\bar{y}^k - y^*(\bar{x}^k)\right\|^2 \\ &\quad + \frac{9\eta}{\sigma}\left[\left(L_{F,1} + L_{f,2}r_v\right)^2\left(\frac{1}{n}\sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 + \frac{1}{n}\sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2\right) + L_{f,1}^2\frac{1}{n}\sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2\right] \\ &\quad + 2\rho^2\frac{1}{n}\sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 + 2\eta^2\rho^2\frac{1}{n}\sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2. \qquad (63)\end{aligned}$$

Proof. By the Cauchy-Schwarz inequality, for any $\delta > 0$, we have

$$\begin{aligned}&\left\|\bar{v}^k + \eta\bar{d}_v^k - v^*(\bar{x}^k)\right\|^2 \\ \le&(1+\delta)\left\|\bar{v}^k + \eta\left(\nabla_2 F(\bar{x}^k, \bar{y}^k) - [\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k)]\bar{v}^k\right) - v^*(\bar{x}^k)\right\|^2 \\ &+ (1 + 1/\delta)\eta^2\left\|[\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k)]\bar{v}^k - \nabla_2 F(\bar{x}^k, \bar{y}^k) + \bar{d}_v^k\right\|^2. \qquad (64)\end{aligned}$$

Firstly, we evaluate the first term on the right hand of (64). Since $\left[\nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k))\right] v^*(\bar{x}^k) = \nabla_2 F(\bar{x}^k, y^*(\bar{x}^k))$, we have

$$\begin{aligned}
&\bar{v}^k + \eta\left(\nabla_2 F(\bar{x}^k, \bar{y}^k) - [\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k)]\bar{v}^k\right) - v^*(\bar{x}^k) \\
&\quad = \bar{v}^k - v^*(\bar{x}^k) - \eta\left[\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k)\right]\left[\bar{v}^k - v^*(\bar{x}^k)\right] \\
&\quad\quad - \eta\left[\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k) - \nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k))\right] v^*(\bar{x}^k) - \eta\left[\nabla_2 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_2 F(\bar{x}^k, \bar{y}^k)\right].
\end{aligned}$$

Then, by Cauchy-Schwarz inequality, for any $\delta_1 > 0$, we have

$$\begin{aligned}
&\left\|\bar{v}^k + \eta\left(\nabla_2 F(\bar{x}^k, \bar{y}^k) - [\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k)]\bar{v}^k\right) - v^*(\bar{x}^k)\right\|^2 \\
&\le (1+\delta_1)\left\|\left[I - \eta\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k)\right]\left[\bar{v}^k - v^*(\bar{x}^k)\right]\right\|^2 \\
&\quad + (1 + 1/\delta_1)\,\eta^2\left\|\left[\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k) - \nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k))\right] v^*(\bar{x}^k) + \nabla_2 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_2 F(\bar{x}^k, \bar{y}^k)\right\|^2.
\end{aligned}$$

Since $\bar{\eta} \le 1/L_{f,1}$, by the σ-strong convexity of $f(x,\cdot)$, we have

$$\left\|\left[I - \eta\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k)\right]\left[\bar{v}^k - v^*(\bar{x}^k)\right]\right\| \le \left\|I - \eta\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k)\right\|_{\mathrm{op}}\left\|\bar{v}^k - v^*(\bar{x}^k)\right\| \le (1-\eta\sigma)\left\|\bar{v}^k - v^*(\bar{x}^k)\right\|.$$

Next, by Assumption 2.1, we get

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

Taking $\delta_1 = \eta\sigma$, we have

$$\begin{aligned}
&\left\|\bar{v}^k + \eta\left(\nabla_2 F(\bar{x}^k, \bar{y}^k) - [\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k)]\bar{v}^k\right) - v^*(\bar{x}^k)\right\|^2 \\
&\le (1+\eta\sigma)(1-\eta\sigma)^2\left\|\bar{v}^k - v^*(\bar{x}^k)\right\|^2 + (1+1/\eta\sigma)\,\eta^2\left(L_{f,2} r_v + L_{F,1}\right)^2\left\|\bar{y}^k - y^*(\bar{x}^k)\right\|^2 \\
&\le (1-\eta\sigma)\left\|\bar{v}^k - v^*(\bar{x}^k)\right\|^2 + \frac{2\eta}{\sigma}\left(L_{f,2} r_v + L_{F,1}\right)^2\left\|\bar{y}^k - y^*(\bar{x}^k)\right\|^2, \qquad (65)
\end{aligned}$$

where the second inequality uses $\eta\sigma \le \eta L_{f,1} \le 1$.

Then we evaluate the second term on the right hand side of (64). The definition of $\bar{d}_v^k$ implies that

$$\begin{aligned}
&\left\|[\nabla_{22}^2 f(\bar{x}^k, \bar{y}^k)]\bar{v}^k - \nabla_2 F(\bar{x}^k, \bar{y}^k) + \bar{d}_v^k\right\| \\
&\le \frac{1}{n}\sum_{i=1}^n\left\|[\nabla_{22}^2 f_i(\bar{x}^k, \bar{y}^k)]\bar{v}^k - \nabla_2 F_i(\bar{x}^k, \bar{y}^k) - [\nabla_{22}^2 f_i(x_i^k, y_i^k)]v_i^k + \nabla_2 F_i(x_i^k, y_i^k)\right\| \\
&\le \frac{1}{n}\sum_{i=1}^n\|[\nabla_{22}^2 f_i(\bar{x}^k, \bar{y}^k)](\bar{v}^k - v_i^k)\| + \frac{1}{n}\sum_{i=1}^n\|[\nabla_{22}^2 f_i(\bar{x}^k, \bar{y}^k) - \nabla_{22}^2 f_i(x_i^k, y_i^k)]v_i^k\| \\
&\quad + \frac{1}{n}\sum_{i=1}^n\|\nabla_2 F_i(\bar{x}^k, \bar{y}^k) - \nabla_2 F_i(x_i^k, y_i^k)\| \\
&\le (L_{f,2} r_v + L_{F,1})\left(\frac{1}{n}\sum_{i=1}^n\|x_i^k - \bar{x}^k\| + \frac{1}{n}\sum_{i=1}^n\|y_i^k - \bar{y}^k\|\right) + L_{f,1}\frac{1}{n}\sum_{i=1}^n\|v_i^k - \bar{v}^k\|. \qquad (66)
\end{aligned}$$

Next, similar to Lemma A.10, we evaluate $\sum_{i=1}^n \|v_i^{k+1} - v^*(\bar{x}^k)\|^2$ in order to bound $\|\bar{v}^{k+1} - v^*(\bar{x}^k)\|^2$. By the update of v_i^k in Algorithm 1, we have

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

where the second equality holds because $\sum_{i=1}^n \left[\sum_{j=1}^n w_{ij} v_j^k - \bar{v}^k + \eta(\sum_{j=1}^n w_{ij} t_{v,j}^k - \bar{t}_v^k)\right] = 0$, and the last inequality uses Lemma A.1. Then, together with (64), (65) and (66), we have (note that $\bar{d}_v^k = \bar{t}_v^k$)

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

We can get the desired results by taking $\delta = \frac{1}{2}\eta\sigma$ and using $\eta\sigma \le 1$. □

Lemma A.12. *For $y^*(\bar{x}^k)$ and $v^*(\bar{x}^k)$, we have*

$$\|y^*(\bar{x}^{k+1}) - y^*(\bar{x}^k)\| \le \frac{L_{f,1}}{\sigma}\|\bar{x}^{k+1} - \bar{x}^k\|, \tag{67}$$

and

$$\|v^*(\bar{x}^{k+1}) - v^*(\bar{x}^k)\| \le \frac{L_v}{\sigma}\|\bar{x}^{k+1} - \bar{x}^k\|, \tag{68}$$

where $L_v = (L_{F,1} + L_{f,2} r_v)\left(1 + \frac{L_{f,1}}{\sigma}\right)$.

Proof. By (58), we have

$$\|y^*(\bar{x}^{k+1}) - y^*(\bar{x}^k)\| \le \frac{L_{f,1}}{\sigma}\|\bar{x}^{k+1} - \bar{x}^k\|.$$

Since $\left[\nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k))\right] v^*(\bar{x}^k) = \nabla_2 F(\bar{x}^k, y^*(\bar{x}^k))$, we have

$$\begin{aligned}
&\nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k))(v^*(\bar{x}^k) - v^*(\bar{x}^{k+1})) \\
=&\left[\nabla_2 F(\bar{x}^k, y^*(\bar{x}^k)) - \nabla_2 F(\bar{x}^{k+1}, y^*(\bar{x}^{k+1}))\right] + \left[\nabla_{22}^2 f(\bar{x}^{k+1}, y^*(\bar{x}^{k+1})) - \nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k))\right] v^*(\bar{x}^{k+1}).
\end{aligned}$$

Using Assumption 2.1, we know that

$$\begin{aligned}
&\sigma\|v^*(\bar{x}^k) - v^*(\bar{x}^{k+1})\| \le \|\nabla_{22}^2 f(\bar{x}^k, y^*(\bar{x}^k))(v^*(\bar{x}^k) - v^*(\bar{x}^{k+1}))\| \\
\le& L_{F,1}\left(\|\bar{x}^k - \bar{x}^{k+1}\| + \|y^*(\bar{x}^k) - y^*(\bar{x}^{k+1})\|\right) + L_{f,2} r_v \left(\|\bar{x}^k - \bar{x}^{k+1}\| + \|y^*(\bar{x}^k) - y^*(\bar{x}^{k+1})\|\right) \\
\le& (L_{F,1} + L_{f,2} r_v)\left(1 + \frac{L_{f,1}}{\sigma}\right)\|\bar{x}^k - \bar{x}^{k+1}\|,
\end{aligned}$$

which implies the desired result. □

Now we are ready to prove the following main result, which is the same as part (b) of Theorem 3.1.

Theorem A.2. *For the sequence generated by Algorithm 1, we have*

$$\min_{0\le k\le K-1}\{\|\nabla\Phi(\bar{x}^k)\|^2\} = O\left(\frac{1}{K}\right),$$

where $\bar{x}^k = \frac{1}{n}\sum_{i=1}^n x_i^k$.

Proof. Using Cauchy-Schwarz inequality, we know that

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

By combining (59), (67) and $\beta\sigma \le 1$, we have

$$\begin{aligned}
&\left\|\bar{y}^{k+1} - y^*(\bar{x}^{k+1})\right\|^2 \\
\le&(1 - \frac{1}{4}\beta\sigma)\|\bar{y}^k - y^*(\bar{x}^k)\|^2 + \frac{5L_{f,1}^2}{\beta\sigma^3}\|\bar{x}^{k+1} - \bar{x}^k\|^2 + \frac{15}{2}\frac{\beta}{\sigma}L_{f,1}^2\left(\frac{1}{n}\sum_{i=1}^n\|\bar{x}^k - x_i^k\|^2 + \frac{1}{n}\sum_{i=1}^n\|\bar{y}^k - y_i^k\|^2\right) \\
&+ \frac{5}{2}\rho^2\frac{1}{n}\sum_{i=1}^n\|y_i^k - \bar{y}^k\|^2 + \frac{5}{2}\rho^2\beta^2\frac{1}{n}\sum_{i=1}^n\|t_{y,i}^k - \bar{t}_y^k\|^2. \qquad (69)
\end{aligned}$$

Also, using Cauchy-Schwarz inequality, we know that

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

By combining (63), (68) and $\eta\sigma \le 1$, we have

$$\begin{aligned}
&\|\bar{v}^{k+1} - v^*(\bar{x}^{k+1})\|^2 \\
\le&\left(1 - \frac{1}{4}\eta\sigma\right)\left\|\bar{v}^k - v^*(\bar{x}^k)\right\|^2 + \frac{15}{4}\frac{\eta}{\sigma}\left(L_{F,1} + L_{f,2}r_v\right)^2\left\|\bar{y}^k - y^*(\bar{x}^k)\right\|^2 + \frac{5L_v^2}{\eta\sigma^3}\|\bar{x}^{k+1} - \bar{x}^k\|^2 \\
&+ \frac{45}{4}\frac{\eta}{\sigma}\left[(L_{F,1} + L_{f,2}r_v)^2\left(\frac{1}{n}\sum_{i=1}^n\|x_i^k - \bar{x}^k\|^2 + \frac{1}{n}\sum_{i=1}^n\|y_i^k - \bar{y}^k\|^2\right) + L_{f,1}^2\frac{1}{n}\sum_{i=1}^n\|v_i^k - \bar{v}^k\|^2\right] \\
&+ \frac{5}{2}\rho^2\frac{1}{n}\sum_{i=1}^n\|v_i^k - \bar{v}^k\|^2 + \frac{5}{2}\eta^2\rho^2\frac{1}{n}\sum_{i=1}^n\|t_{v,i}^k - \bar{t}_v^k\|^2. \qquad (70)
\end{aligned}$$

Next, we define the Lyapunov function

$$V_k = a\left[F(\bar{x}^k, y^*(\bar{x}^k)) - F^*\right] + b\|\bar{y}^k - y^*(\bar{x}^k)\|^2 + c\|\bar{v}^k - v^*(\bar{x}^k)\|^2,$$

where

$$a := K+1,\ b := \bar{b}(K+1),\ c := K+1,\ \bar{b} := \frac{15L_1^2\bar{\eta}}{\sigma^2\bar{\beta}} + 1.$$

With (53), (69) and (70), we can get

$$\begin{aligned}
&V_{k+1} - V_k \\
\leq& a\left[F(\bar{x}^{k+1}, y^*(\bar{x}^{k+1})) - F(\bar{x}^k, y^*(\bar{x}^k))\right] + b\left[\|\bar{y}^{k+1} - y^*(\bar{x}^{k+1})\|^2 - \|\bar{y}^k - y^*(\bar{x}^k)\|^2\right] \\
&+ c\left[\|\bar{v}^{k+1} - v^*(\bar{x}^{k+1})\|^2 - \|\bar{v}^k - v^*(\bar{x}^k)\|^2\right] \\
\leq& -\frac{a\alpha}{2}\|\nabla\Phi(\bar{x}^k)\|^2 - \widehat{\alpha}\|\bar{x}^{k+1} - \bar{x}^k\|^2 - \widehat{\beta}\|\bar{y}^k - y^*(\bar{x}^k)\|^2 - \widehat{\eta}\|\bar{v}^k - v^*(\bar{x}^k)\|^2 \\
&+ \tilde{\alpha}\frac{1}{n}\sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 + \tilde{\beta}\frac{1}{n}\sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 + \tilde{\eta}\frac{1}{n}\sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 \\
&+ \frac{5}{2}\rho^2 b\beta^2\frac{1}{n}\sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2 + \frac{5}{2}\rho^2 c\eta^2\frac{1}{n}\sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2,
\end{aligned} \tag{71}$$

where the coefficients are given as

$$\begin{aligned}
\widehat{\alpha} :=& \frac{1}{2}\frac{a}{\alpha} - \frac{L_\Phi}{2}a - \frac{5L_{f,1}^2}{\sigma^3}\frac{b}{\beta} - \frac{5L_v^2}{\sigma^3}\frac{c}{\eta},\ \widehat{\beta} := \frac{\sigma}{4}b\beta - \frac{5}{2}L_1^2 a\alpha - \frac{15L_1^2}{4\sigma}c\eta, \\
\widehat{\eta} :=& \frac{\sigma}{4}c\eta - \frac{5}{2}L_{f,1}^2 a\alpha, \\
\tilde{\alpha} :=& \frac{5}{2}L_1^2 a\alpha + \frac{15L_{f,1}^2}{2\sigma}b\beta + \frac{45L_1^2}{4\sigma}c\eta,\ \tilde{\beta} := \frac{5}{2}L_1^2 a\alpha + \frac{15L_{f,1}^2}{2\sigma}b\beta + \frac{45L_1^2}{4\sigma}c\eta + \frac{5\rho^2}{2}b, \\
\tilde{\eta} :=& \frac{5}{2}L_{f,1}^2 a\alpha + \frac{45L_1^2}{4\sigma}c\eta + \frac{5\rho^2}{2}c,
\end{aligned}$$

where we defined $L_1 := L_{F,1} + L_{f,2}r_v$. Denote

$$\begin{aligned}
&K_0 := \\
&\max\left\{\lceil L_\Phi\bar{\alpha}\rceil, \left\lceil\left(\frac{10L_{f,1}^2}{\sigma^3}\frac{\bar{\alpha}\bar{b}}{\bar{\beta}}\right)^2\right\rceil, \left\lceil\left(\frac{10L_v^2}{\sigma^3}\frac{\bar{\alpha}}{\bar{\eta}}\right)^2\right\rceil, \left\lceil\left(\frac{10L_1^2}{\sigma}\frac{\bar{\alpha}}{\bar{\beta}\bar{b}}\right)^2\right\rceil, \left\lceil\left(\frac{10L_{f,1}^2}{\sigma}\frac{\bar{\alpha}}{\bar{\eta}}\right)^2\right\rceil\right\}.
\end{aligned}$$

Then we can deduce that when $K \geq K_0$, $\widehat{\alpha}, \widehat{\beta}, \widehat{\eta}$ are positive for all $0 \leq k \leq K-1$.

Denote

$$\begin{aligned}
E_k :=& \tilde{\alpha}\sum_{i=1}^n \|x_i^k - \bar{x}^k\|^2 + \tilde{\beta}\sum_{i=1}^n \|y_i^k - \bar{y}^k\|^2 + \tilde{\eta}\sum_{i=1}^n \|v_i^k - \bar{v}^k\|^2 \\
&+ \frac{5}{2}\rho^2 b\beta^2\frac{1}{n}\sum_{i=1}^n \|t_{y,i}^k - \bar{t}_y^k\|^2 + \frac{5}{2}\rho^2 c\eta^2\frac{1}{n}\sum_{i=1}^n \|t_{v,i}^k - \bar{t}_v^k\|^2.
\end{aligned}$$

By (24), (49), (25) and (39), we have

$$\begin{aligned}\sum_{i=1}^{n}\|t_{y,i}^{k}-\bar{t}_{y}^{k}\|^{2} &\leq \frac{1}{1-\rho}\sum_{l=0}^{k}\left(\rho^{k-l}\sum_{i=1}^{n}\|d_{y,i}^{l}-d_{y,i}^{l-1}\|^{2}\right)\\ &\leq \frac{C_2}{1-\rho}\sum_{l=0}^{k}\left(\rho^{k-l}\sum_{i=1}^{n}\left(\|x_i^l-x_i^{l-1}\|^2+\|y_i^l-y_i^{l-1}\|^2\right)\right)\\ &\leq \frac{C_2}{1-\rho}\sum_{l=0}^{k}\rho^{k-l}\left(\frac{\rho_1}{1-\rho}nB_2^2\alpha^2+\frac{\rho_2}{1-\rho}nB_2^2\beta^2\right)\\ &\leq \frac{C_2}{(1-\rho)^2}\left(\frac{\rho_1}{1-\rho}nB_2^2\alpha^2+\frac{\rho_2}{1-\rho}nB_2^2\beta^2\right)=O\left(\max\{\alpha^2,\beta^2\}\right).\end{aligned}$$

Similarly, by (23), (48), (25), (39) and (38), we have

$$\sum_{i=1}^{n}\|t_{y,i}^{k}-\bar{t}_{y}^{k}\|^{2}=O\left(\max\{\alpha^2,\beta^2,\eta^2\}\right),$$

which, combining with Theorem A.1, yields

$$E_k=O\left(\frac{1}{K}\right),\forall k.$$

Then by (71), we have

$$\frac{a\alpha}{2}\sum_{k=0}^{K-1}\|\nabla\Phi(\bar{x}^k)\|^2\leq V_0+\sum_{k=0}^{K-1}E_k.$$

Therefore,

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

which completes the proof. □