A Two-timescale Primal-dual Algorithm for Decentralized Optimization with Compression

Haoming Liu, Chung-Yiu Yau, Hoi-To Wai
Dept. of Systems Engineering & Engineering Management, The Chinese University of Hong Kong
Emails: haomingliu@cuhk.edu.hk, cyyau@se.cuhk.edu.hk, htwai@se.cuhk.edu.hk.

***Abstract*—This paper proposes a two-timescale compressed primal-dual (TiCoPD) algorithm for decentralized optimization with improved communication efficiency over prior works on primal-dual decentralized optimization. The algorithm is built upon the primal-dual optimization framework and utilizes a majorization-minimization procedure. The latter naturally suggests the agents to share a compressed difference term during the iteration. Furthermore, the TiCoPD algorithm incorporates a fast timescale mirror sequence for agent consensus on nonlinearly compressed terms, together with a slow timescale primal-dual recursion for optimizing the objective function. We show that the TiCoPD algorithm converges with a constant step size. It also finds an $\mathcal{O}(1/T)$ stationary solution after T iterations. Numerical experiments on decentralized training of a neural network validate the efficacy of TiCoPD algorithm.**

***Index Terms*—decentralized optimization, nonlinear compression, two-timescale iteration, majorization-minimization.**

I. Introduction

Let $\mathcal{G} = (V, E)$ be an undirected and connected graph of n agents, with the node set given by $V = [n] := \{1, ..., n\}$ and the edge set $E \subseteq V \times V$. Consider the following distributed optimization problem:

$$\min_{\mathbf{X} \in \mathbb{R}^{nd}} \quad \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{X}_i) \quad \text{s.t.} \quad \mathbf{X}_i = \mathbf{X}_j, \ \forall (i,j) \in E, \tag{1}$$

where $\mathbf{X} = [\mathbf{X}_1^\top \ \mathbf{X}_2^\top \ \cdots \ \mathbf{X}_n^\top]^\top$. For each $i \in V$, $\mathbf{X}_i \in \mathbb{R}^d$ stands for the local decision variable of the ith agent, and the local objective function $f_i : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable (possibly non-convex). Distributed solution methods for (1) has found applications in machine learning [1], [2], signal processing [3]–[5], etc.

In scenarios where the local objective function (and the associated training data) has to be stored or processed in a distributed network, and/or there is no central server, a decentralized algorithm is preferred for tackling (1) where agents do not share their local objective functions with other parties. The pioneering work by Nedić and Ozdaglar [6] proposed the decentralized gradient descent (DGD) method. The latter is shown to converge in the non-convex setting in a follow-up publication [7]. A number of subsequent works have been proposed to improve DGD, e.g., EXTRA in [8], gradient tracking in [9], etc. The (proximal) primal-dual algorithm [10]–[12] yields a general framework for decentralized optimization with good convergence property where it encompasses the previous algorithms as special cases.

For high dimensional instances of (1) such that $d \gg 1$, the communication overhead incurred with decentralized optimization can be a bottleneck impeding the convergence speed of these algorithms. Reducing the bandwidth usage through lossy communication compression has thus become an important issue in recent works. Note that directly applying compression scheme such as quantization in distributed optimization may result in a non-converging algorithm [13]. Thus, many works have proposed to combine the decentralized algorithms with an error feedback subroutine to achieve exact convergence. For example, [14]–[17] focused on the deterministic gradient setting and extended algorithms such as DGD and gradient tracking with communication compression, [18]–[21] considered similar extensions but have considered using stochastic gradients. Alternatively, [22]–[24] considered reducing the communication frequency by local updates, [25] considered event triggered communication.

Meanwhile, existing results entail various limitations in guaranteeing convergence for the general non-convex setting — the algorithms in [14], [19] require a bounded heterogeneity assumption such that $\|\nabla f_i(\mathbf{X}) - \nabla f_j(\mathbf{X})\|$ is upper bounded for any $\mathbf{X}$ and may require a diminishing stepsize for convergence, the algorithms in [17], [20] requires applying two separate communication compression operations that may increase storage complexity, and other algorithms [15], [16], [18] have not been analyzed for the non-convex optimization setting.

This paper aims to address the above shortcomings with communication efficient decentralized optimization. Our key idea is to develop the algorithm using the primal-dual framework [10], [11] and to incorporate compressed communication updates under the classical majorization-minimization framework [26]. Our contributions are:

- We propose the **T**wo-**ti**mescale **Co**mpressed **P**rimal-**D**ual (TiCoPD) algorithm as a nonlinearly compressed primal-dual algorithm for decentralized optimization. The TiCoPD algorithm follows a two-timescale update rule which separates the *communication* and *optimization* steps, handled using a large stepsize and a small stepsize, respectively.
- To incorporate nonlinear compression into the primal-dual algorithm, we develop a majorization-minimization (MM) procedure which suggests agents to transmit the compressed difference terms – a scheme that coincides with the popular error feedback mechanism. This offers a new perspective for extending the error feedback mechanism popularized by [18] and establishing a connection to the nonlinear gossiping algorithm in [27]. We believe that this observation will be of independent interest.
- For optimization problems with continuously differentiable (possibly non-convex) objective functions, we show that the TiCoPD algorithm converges at a rate of $\mathcal{O}(1/T)$ towards a stationary solution of (1). Furthermore, the convergence is guaranteed with a constant stepsize and without relying on additional assumptions such as bounded heterogeneity.

Finally, we present numerical experiments to demonstrate the efficacy of TiCoPD algorithm against state-of-the-art algorithms in tackling a toy example of training neural networks over network for (1).

II. Problem Statement

This section introduces the basic ideas for the proposed TiCoPD algorithm to tackling (1). We first introduce a few extra notations to

This research is supported in part by project #MMT-p5-23 of the Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong.

facilitate the development: define

$$f(\bar{\mathbf{X}}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\bar{\mathbf{X}}), \quad \nabla f(\bar{\mathbf{X}}) := \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\bar{\mathbf{X}}), \tag{2}$$

as the global objective function and gradient evaluated on a common decision variable $\bar{\mathbf{X}} \in \mathbb{R}^d$, respectively. Moreover, the consensus constraint $\mathbf{X}_i = \mathbf{X}_j,\ \forall (i,j) \in E$ can be replaced by the equality $\widetilde{\mathbf{A}}\mathbf{X} = \mathbf{0}$, where $\widetilde{\mathbf{A}} = \mathbf{A} \otimes \mathbf{I}_d$ and $\widetilde{\mathbf{A}} \in \{0, 1, -1\}^{|E|\times n}$ denotes the incidence matrix of the graph G.

Using the above notations and introducing the Lagrange multiplier variable $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_i)_{i\in E} \in \mathbb{R}^{|E|d}$ for the equality constraint $\widetilde{\mathbf{A}}\mathbf{X} = \mathbf{0}$, we consider the following augmented Lagrangian function:

$$\mathcal{L}(\mathbf{X}, \boldsymbol{\lambda}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{X}_i) + \boldsymbol{\lambda}^\top \widetilde{\mathbf{A}}\mathbf{X} + \frac{\theta}{2}\|\widetilde{\mathbf{A}}\mathbf{X}\|^2, \tag{3}$$

where $\theta > 0$ is a regularization parameter. It can be shown that any stationary point to $\mathcal{L}(\cdot)$ satisfying $\nabla_{\mathbf{X}}\mathcal{L}(\mathbf{X}, \boldsymbol{\lambda}) = \mathbf{0}$, $\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{X}, \boldsymbol{\lambda}) = \mathbf{0}$ is a KKT point of (1). Importantly, applying the standard gradient descent-ascent (GDA) algorithm on $\mathcal{L}(\cdot)$ yields the decentralized algorithm proposed in [12] which finds a stationary solution to (1) for a class of non-convex problems.

However, a key drawback for [12] and similar algorithms proposed using the primal-dual optimization framework (e.g., [11]) lies on their high bandwidth usage for high-dimensional problems when $d \gg 1$. In particular, at each iteration, such algorithms require the agents to share their current local decision variable with neighboring agents, which demands transmitting an $\mathbb{R}^d$ vector. Such step introduces a considerable communication overhead for their implementation.

III. Proposed TiCoPD Algorithm

Our idea is to develop a **T**wo-t**i**mescale **Co**mpressed **P**rimal-**D**ual (TiCoPD) algorithm that supports general *nonlinear compression* in primal-dual decentralized updates. The key idea of TiCoPD is to separately treat *compressed communication* and *optimization* as lower and upper level updates, respectively, to be updated at different speed. The algorithm depends on two ingredients: (i) a majorization-minimization step that introduces a *surrogate* variable to separate the communication step from the optimization step, (ii) a *two-timescale* update that incorporates the nonlinearly compressed update of the surrogate variable. The above ideas will be introduced in the sequel.

Majorization Minimization. We begin by inspecting the augmented Lagrangian function $\mathcal{L}(\cdot)$ again. Here, computing the gradient for the last term in (3) at a fixed primal variable $\mathbf{X}^t$ leads to

$$\nabla_{\mathbf{X}_i}\|\widetilde{\mathbf{A}}\mathbf{X}^t\|^2 = 2\textstyle\sum_{j\in\mathcal{N}_i}(\mathbf{X}_j^t - \mathbf{X}_i^t), \tag{4}$$

where $\mathcal{N}_i$ is the neighbor set of agent i. Consequently, the primal update necessitates the communication of the neighbors' decision variables $\mathbf{X}_j$, $j \in \mathcal{N}_i$ and leads to a communication bottleneck.

Our idea is to sidestep this term through a majorization-minimization (MM) procedure with a surrogate variable. Let there be a sequence of surrogate variables $\{\hat{\mathbf{X}}^t\}_{t\geq 0}$ such that (i) $\hat{\mathbf{X}}^t \approx \mathbf{X}^t$, and (ii) it is possible for agent i to acquire the neighbors' surrogate variables $(\hat{\mathbf{X}}_j^t)_{j\in\mathcal{N}_i^t}$ with *compressed communication*. We will illustrate how to construct such sequence later. With $M := \|\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\|_2$, the following *majorization* holds for any $\mathbf{X}$,

$$\|\widetilde{\mathbf{A}}\mathbf{X}\|^2 \leq \|\widetilde{\mathbf{A}}\hat{\mathbf{X}}^t\|^2 + 2(\mathbf{X} - \hat{\mathbf{X}}^t)^\top\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\hat{\mathbf{X}}^t + M\|\mathbf{X} - \hat{\mathbf{X}}^t\|^2. \tag{5}$$

Unlike the original term $\|\widetilde{\mathbf{A}}\mathbf{X}\|^2$, evaluating the gradient w.r.t. $\mathbf{X}_i$ on the upper bound only requires aggregating the surrogate variables $\hat{\mathbf{X}}_j^t, j \in \mathcal{N}_i$, which can be obtained from compressed communication.

We further upper bounding the first term in (3) using standard truncated Taylor approximation. Now, the $\mathbf{X}$-update can be computed using the following *minimization* step:

$$\begin{aligned}\mathbf{X}^{t+1} &= \arg\min_{\mathbf{X}\in\mathbb{R}^{nd}} \nabla\mathbf{f}(\mathbf{X}^t)^\top(\mathbf{X} - \mathbf{X}^t) + \mathbf{X}^\top\widetilde{\mathbf{A}}^\top\boldsymbol{\lambda}^t + \frac{\theta}{2}\|\widetilde{\mathbf{A}}\hat{\mathbf{X}}^t\|^2 \\ &\qquad + \theta\mathbf{X}^\top\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\hat{\mathbf{X}}^t + \frac{\theta M}{2}\|\mathbf{X} - \hat{\mathbf{X}}^t\|^2 + \frac{1}{2\widetilde{\alpha}}\|\mathbf{X} - \mathbf{X}^t\|^2 \\ &= \beta\mathbf{X}^t + (1-\beta)\hat{\mathbf{X}}^t - \alpha\big(\nabla\mathbf{f}(\mathbf{X}^t) + \widetilde{\mathbf{A}}^\top\boldsymbol{\lambda}^t + \theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\hat{\mathbf{X}}^t\big),\end{aligned} \tag{6}$$

where $\nabla\mathbf{f}(\mathbf{X}^t) = [\nabla f_1(\mathbf{X}_1^t)^\top \cdots \nabla f_n(\mathbf{X}_n^t)^\top]^\top$, $\alpha = \frac{1}{\frac{1}{\widetilde{\alpha}} + \theta M}$, and $\beta = \frac{\alpha}{\widetilde{\alpha}}$. For the $\boldsymbol{\lambda}$-subproblem, similarly we replace $\mathbf{X}^t$ with the surrogate variable $\hat{\mathbf{X}}^t$ to obtain:

$$\begin{aligned}\boldsymbol{\lambda}^{t+1} &= \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^{|E|d}} \left\{-\boldsymbol{\lambda}^\top\widetilde{\mathbf{A}}\hat{\mathbf{X}}^t + \frac{1}{2\eta}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^t\|^2\right\} \\ &= \boldsymbol{\lambda}^t + \eta\widetilde{\mathbf{A}}\hat{\mathbf{X}}^t.\end{aligned} \tag{7}$$

Substituting the variable $\widetilde{\boldsymbol{\lambda}}^t = \widetilde{\mathbf{A}}^\top\boldsymbol{\lambda}^t \in \mathbb{R}^{nd}$ yields a decentralized primal-dual algorithm that only requires aggregating $\hat{\mathbf{X}}_i^t$ at each step. Our remaining task is to study how to effectively construct the sequence $\{\hat{\mathbf{X}}^t\}_{t\geq 0}$ in a compressed communication friendly fashion.

Two-timescale Updates. Recall that a key requirement for our construction of $\{\hat{\mathbf{X}}^t\}_{t\geq 0}$ is that the surrogate variable should *track* the original decision variable $\hat{\mathbf{X}}^t \approx \mathbf{X}^t$.

To this end, our construction consists of applying a (randomized) compression operator $\hat{Q} : \mathbb{R}^d \times \Omega_i \to \mathbb{R}^d$ satisfying: there exists $\delta \in (0, 1]$ such that

$$\mathbb{E}\left[\|\hat{Q}(\mathbf{x};\xi_i) - \mathbf{x}\|^2\right] \leq (1-\delta)^2\mathbb{E}\left[\|\mathbf{x}\|^2\right],\ \forall\, \mathbf{x} \in \mathbb{R}^d, i \in [n], \tag{8}$$

where $\xi_i \in \Omega_i$ is a random variable. For example, the above property can be satisfied with the randomized quantization operator:

$$\mathrm{qsgd}_s(\mathbf{x};\xi_i) = \frac{\mathrm{sign}(\mathbf{x})\cdot\|\mathbf{x}\|}{s\tau}\cdot\left\lfloor s\frac{|\mathbf{x}|}{\|\mathbf{x}\|} + \xi_i\right\rfloor, \tag{9}$$

where $s > 0$ is the number of precision levels, $\tau = 1 + \min\{d/s^2, \sqrt{d}/s\}$ and $\xi_i \sim \mathcal{U}[0,1]^d$ is an additive noise [28]. The operator can be implemented with an *encoder* (denoted $\mathrm{ENC}(\cdot)$) which turns the d-dimensional input vector into a $d\log_2 s$ bits string to represent the quantized levels, a d bits string to represent $\mathrm{sign}(x)$ and a floating-point value for $\|\mathbf{x}\|$. The latter can be transmitted on bandwidth limited channels. On the receiver's side, a *decoder* (denoted $\mathrm{DEC}(\cdot)$) may convert the received symbols into the quantized real vector in (9). Particularly, the randomized quantization operator satisfies (8) with $\delta = \frac{1}{2\tau}$; see [18], [28] for discussions on contractive compressors satisfying (8). Finally, we denote $Q : \mathbb{R}^{nd} \times \Omega_1 \times \ldots \times \Omega_n \to \mathbb{R}^{nd}$ as the compression operator for the column stacked variables such that $[Q(x;\xi)]_i = \hat{Q}(x_i;\xi_i)$, where $\xi = (\xi_1, \ldots, \xi_n)$.

The contraction property in (8) suggests a compressed communication friendly procedure to achieve $\hat{\mathbf{X}}^t \approx \mathbf{X}^t$. In particular, we notice that the mean field $\mathbb{E}_\xi[Q(\mathbf{X}^t - \mathbf{X};\xi)]$ has a unique fixed point at $\mathbf{X} = \mathbf{X}^t$. In particular, we observe that at the tth primal-dual iteration, if we let k denotes the contraction iteration index and $\gamma \in (0, 1]$ be a stepsize parameter, the recursion

$$\hat{\mathbf{X}}^{t,k+1} = \hat{\mathbf{X}}^{t,k} + \gamma\, Q(\mathbf{X}^t - \hat{\mathbf{X}}^{t,k};\xi^{t,k+1}),\ \forall\ k \geq 0. \tag{10}$$

finds $\hat{\mathbf{X}}^{t,k} \overset{k\to\infty}{\to} \mathbf{X}^t$. The insight behind the above update is that the compression error can be gradually reduced by compressing only the error itself $\mathbf{X}^t - \hat{\mathbf{X}}^{t,k}$. For implementation on a decentralized system, for each (t,k), the agents only need to encode and transmit

Algorithm 1 TiCoPD

1: **Input:** Parameter $\alpha, \theta, \beta, \eta$, initialize $\mathbf{X}^0, \hat{\mathbf{X}}^0, \widetilde{\lambda}^0$.
2: **for** $t = 1, \ldots, T$ **do**
3: *Surrogate variable update:* for any $i \in V$,
$$\hat{\mathbf{X}}_i^t = \hat{\mathbf{X}}_i^{t-1} + \hat{Q}(\mathbf{X}_i^t - \hat{\mathbf{X}}_i^{t-1}; \xi_i^t),$$
and transmit the compressed message $\text{ENC}(\mathbf{X}_i^t - \hat{\mathbf{X}}_i^{t-1}; \xi_i^t)$ to agent $j \in \mathcal{N}_i$.
4: *Aggregate received messages:* for any $i \in V$,
$$\hat{\mathbf{X}}_{i,-i}^t = \hat{\mathbf{X}}_{i,-i}^{t-1} + \textstyle\sum_{j \in \mathcal{N}_i} \text{DEC}(\text{ENC}(\mathbf{X}_j^t - \hat{\mathbf{X}}_j^{t-1}; \xi_j^t)),$$
where $\hat{Q}(\cdot) = \text{DEC} \circ \text{ENC}(\cdot)$ and $\text{DEC}(\cdot)$ is a decoder.
5: *Primal-dual update:* for any $i \in V$,
$$\begin{aligned}\mathbf{X}_i^{t+1} &= \beta \mathbf{X}_i^t + (1-\beta)\hat{\mathbf{X}}_i^t \\ &\quad - \alpha\big[\nabla f_i(\mathbf{X}_i^t) + \widetilde{\boldsymbol{\lambda}}_i^t + \theta(|\mathcal{N}_i|\hat{\mathbf{X}}_i^t - \hat{\mathbf{X}}_{i,-i}^t)\big], \\ \widetilde{\boldsymbol{\lambda}}_i^{t+1} &= \widetilde{\boldsymbol{\lambda}}_i^t + \eta(|\mathcal{N}_i|\hat{\mathbf{X}}_i^t - \hat{\mathbf{X}}_{i,-i}^t).\end{aligned}$$
6: **end for**

the differences $\mathbf{X}^t - \hat{\mathbf{X}}^{t,k}$. In particular, it holds $\mathbb{E}[\|\hat{\mathbf{X}}^{t,k} - \mathbf{X}^t\|^2] \leq (1-\gamma\delta)^k \|\hat{\mathbf{X}}^{t,0} - \mathbf{X}^t\|^2$. For the rest of this paper, we shall take $\gamma = 1$ for simplicity. However, we remark that in cases when (8) is not satisfied, the stepsize parameter $\gamma \in (0,1]$ can be used to control the process for the construction of $\hat{\mathbf{X}}^t \approx \mathbf{X}^t$.

The convergence of the procedure (10) relies on holding $\mathbf{X}^t$ fixed as k increases. While repeating the recursion (10) until $k \to \infty$ guarantees $\hat{\mathbf{X}}^{t,k} = \mathbf{X}^t$, this may incur a prohibitive communication cost for each of the primal-dual update (6), (7) indexed by t. Fortunately, we observe that each recursion in (10) can actually reduces the error by a factor of $(1-\delta)$, i.e., it exhibits a geometric convergence rate. Suppose the primal-dual update stepsizes are sufficiently small compared to γ, it is possible to satisfy the *tracking condition* $\hat{\mathbf{X}}^t \approx \mathbf{X}^t$ with only one step of the update in (10) per primal-dual iteration.

Overall, the above discussion suggests a *two-timescale* simultaneous update of (6), (7), (10). We call this algorithm the TiCoPD algorithm which is summarized in Algorithm 1. Note that the algorithm is fully decentralized. For each iteration, the agents always communicate with one round of *compressed* message exchanges.

Remark III.1. *The recently proposed CP-SGD algorithm [21] uses a similar idea of contractive compressor on primal-dual algorithm as TiCoPD. There are several key differences: (i) CP-SGD uses an extra auxiliary variable $\mathbf{X}^c$ and entails an additional stepsize parameter α_x, (ii) in practice we observe that CP-SGD yields slower convergence in the consensus error for certain problems. We highlight that TiCoPD was developed directly from the MM procedure and two-timescale updates. Our algorithm also belongs to a general framework that can incorporate scenarios with less restrictions on the communication architecture, e.g., noisy communication, time varying graphs, etc. Such features are missing in the CP-SGD algorithm [21].*

IV. Convergence analysis

This section establishes the convergence of the TiCoPD algorithm towards a stationary point of (1) at a sublinear rate. To facilitate our discussions, we define the consensus error operator $\widetilde{\mathbf{K}} := (\mathbf{I}_n - \mathbf{1}\mathbf{1}^\top/n) \otimes \mathbf{I}_d$ and $\widetilde{\mathbf{Q}} := (\widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}})^\dagger$, where $(\cdot)^\dagger$ denotes the Moore-Penrose inverse. From the definitions, we observe that $\widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} \widetilde{\mathbf{K}} = \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} = \widetilde{\mathbf{K}} \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{Q}} \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} = \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} \widetilde{\mathbf{Q}} = \widetilde{\mathbf{K}}$. We first state the assumptions about the objective function, graph, and the compression operator.

Assumption IV.1. *For any $i \in [n]$, the function f_i is L-smooth, i.e.,*
$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \ \forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{11}$$

Note that (11) implies the global objective function is L-smooth. It is also a standard assumption in the optimization literature.

Assumption IV.2. *There exists $\tilde{\rho}_1 \geq \tilde{\rho}_2 > 0$ such that*
$$\tilde{\rho}_2 \widetilde{\mathbf{K}} \preceq \widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}} \preceq \tilde{\rho}_1 \widetilde{\mathbf{K}}. \tag{12}$$

Recall that as $\widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}}$ is the graph Laplacian matrix, the assumption above is satisfied when G is a connected graph. Furthermore, we have $\tilde{\rho}_1 = \|\widetilde{\mathbf{A}}^\top \widetilde{\mathbf{A}}\|_2$. It also holds $\tilde{\rho}_1^{-1} \widetilde{\mathbf{K}} \preceq \widetilde{\mathbf{Q}} \preceq \tilde{\rho}_2^{-1} \widetilde{\mathbf{K}}$.

Assumption IV.3. *For any fixed $\mathbf{X} \in \mathbb{R}^d$, there exist $0 < \delta \leq 1$ such that the compression operator $\hat{Q}(\cdot; \xi_q)$ satisfies*
$$\mathbb{E}\left[\|\hat{Q}(\mathbf{X}; \xi_q) - \mathbf{X}\|^2\right] \leq (1-\delta)^2 \|\mathbf{X}\|^2. \tag{13}$$

As discussed before, the randomized quantizer (9) satisfies the above with $\delta = s/\sqrt{d}$ for sufficiently large d.

To describe our main results, we need to define a few extra notations. The network-average decision variable is denoted as
$$\bar{\mathbf{X}}^t = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^t = \frac{1}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)\mathbf{X}^t,$$
and the consensus error is given by
$$\sum_{i=1}^n \|\mathbf{X}_i^t - \bar{\mathbf{X}}^t\|^2 = \|\mathbf{X}^t - \mathbf{1}_n \otimes \bar{\mathbf{X}}^t\|^2 = \|\widetilde{\mathbf{K}}\mathbf{X}^t\|^2 = \|\mathbf{X}^t\|_{\widetilde{\mathbf{K}}}^2.$$
We also set $\mathbf{v}^t = \alpha(\widetilde{\boldsymbol{\lambda}}^t + \nabla \mathbf{f}((\mathbf{1} \otimes \mathbf{I}_d)\bar{\mathbf{X}}^t))$ as an auxiliary variable to measures the violation of tracking the average deterministic gradient. The convergence result follows:

Theorem IV.4. *Under Assumptions IV.1–IV.3, suppose the step sizes satisfy $\eta > 0, \theta \geq \theta_{lb}, \alpha \leq \alpha_{ub}$, where*
$$\begin{aligned}\theta_{lb} &= \max\Big\{\frac{4L^2}{n\tilde{\rho}_2 \mathtt{a}}, \frac{2}{\tilde{\rho}_2}(1 + 2L + 8\eta\tilde{\rho}_1^2\tilde{\rho}_2^{-1} \\ &\qquad + \delta_1(\frac{3}{2} + 3L^2 + \eta\tilde{\rho}_1 + \frac{\tilde{\rho}_1^2}{2}))\Big\}, \\ \alpha_{ub} &= \frac{\max\{16\eta, \delta\}}{320\theta^2} \cdot \min\Big\{\frac{1}{M^2}, \frac{n\mathtt{a}}{M^2}, \frac{1}{\tilde{\rho}_1^2}\Big\},\end{aligned} \tag{14}$$
where $\delta_2 = \max\{\frac{16\eta}{\delta}, 1\}$, $\delta_1 = 12\max\{2, 2\tilde{\rho}_2^{-1}\eta^{-1}, \delta_2\tilde{\delta}\}$, $\tilde{\delta} = \max\{\frac{(1-\delta)^2(1-\frac{\delta}{2})^2}{(1-\frac{\delta}{2})^2 - (1-\delta)^2}, 1\}$. Then, for any $T \geq 1$, it holds
$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(\bar{\mathbf{X}}^t)\|^2\right] &\leq \frac{\mathbb{E}[F_0] - \mathbb{E}[F_T]}{\alpha T/16}, \\ \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\mathbf{X}^t\|_{\widetilde{\mathbf{K}}}^2\right] &\leq \frac{\mathbb{E}[F_0] - \mathbb{E}[F_T]}{\alpha\theta\tilde{\rho}_2 \mathtt{a} T/4}.\end{aligned} \tag{15}$$
where the expectation is taken w.r.t. randomness in the compression operator, $\mathtt{a} > 0$ is a free quantity, and
$$\begin{aligned}F_t &= f(\bar{\mathbf{X}}^t) - f^\star + \frac{\mathtt{a}}{\eta\alpha}\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{Q}} + \alpha\left((\frac{\delta_1}{2}(\theta+\eta) - \theta) - \delta_2\theta\right)\widetilde{\mathbf{K}}} \\ &\quad + \mathtt{a}\|\mathbf{X}^t\|_{\widetilde{\mathbf{K}}}^2 + \delta_1 \mathtt{a}\left\langle \mathbf{X}^t \mid \mathbf{v}^t \right\rangle_{\widetilde{\mathbf{K}}} + \delta_2 \mathtt{a}\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2,\end{aligned} \tag{16}$$
can be shown to be non-negative.

Our analysis involves constructing a Lyapunov function to track the joint convergence of the coupled variables. However, the proof is rather technical and are relegated to an online appendix; see https://www1.se.cuhk.edu.hk/~htwai/pdf/icassp25-ticopd.pdf.

The above theorem shows that upon fixing the dual step size at $\eta = \delta$, there exists a sufficiently small constant primal step size $\alpha \leq \alpha_{ub}$[1] (independent of T) such that for sufficiently large T,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{X}}^t)\|^2\right] = \mathcal{O}(1/T), \tag{17}$$

where it implies that there exists $\hat{t} \in \{0, \ldots, T-1\}$ such that $\mathbb{E}\left[\|\nabla f(\bar{\mathbf{X}}^{\hat{t}})\|^2\right] = \mathcal{O}(1/T)$. Notice that this is the same rate (w.r.t. $1/T$) as a centralized gradient algorithm on the smooth optimization problem (1). Moreover, compared to the results for decentralized algorithms such as DGD [7] and CHOCO-SGD [19], our results do not require using a diminishing step size nor bounded gradient heterogeneity.

We also comment on the impact of compression on the convergence rate and step size selection. Recall that $\delta \in (0, 1]$ of Assumption IV.3 is affected by quality of the compressor, where $\delta \approx 0$ with an aggressive compression scheme, e.g., the number of quantization levels, s, is small. In this case, we have $\tilde{\delta} \asymp \delta^{-1}, \theta_{lb} \asymp \delta_1 \asymp \delta^{-1}$, and thus $\alpha_{ub} = \mathcal{O}(\delta^3)$. Consequently, we observe that the convergence rate will be $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(\bar{\mathbf{X}}^t)\|^2\right] = \mathcal{O}(\delta^3 T^{-1})$ and the upper bound evaluates to $\mathcal{O}(d^{1.5}s^{-3}T^{-1})$ for the case of randomized quantization. As such, we conclude that the convergence of TiCoPD is sensitive to the quality of compressor. A future direction is to improve such dependence on δ.

V. Numerical Experiments

We compare the performance of the proposed algorithm on a typical machine learning model training task. Here, $f_i(\mathbf{X}_i)$ is taken as the cross-entropy classification loss where $\mathbf{X}_i \in \mathbb{R}^{79510}$ denotes the weights of a 2 layer feed-forward neural network with 100 neurons and sigmoid activation. We consider training the neural network on the MNIST dataset. We benchmark the performance among algorithms that support compressed communication including CHOCO-SGD [18] and CP-SGD [21] (both implemented with exact gradients), while also comparing against the classical DGD method, and a heuristic variant of DGD that only aggregates quantized parameters. The stepsizes for each of the above algorithms are fine tuned so that they achieve the best performance after 10^5 iterations. The experiments are performed on a 40-core Intel Xeon server with 64GB memory using the PyTorch package.

We simulate the algorithms on a ring network of 10 nodes. The MNIST dataset (of $M = 60,000$ images, each with $p = 784$ pixels) is distributed to each node according to the class label $\{0, ..., 9\}$, i.e., each node only access one class of images. We consider unshuffled dataset to maximize data heterogeneity across agents. Note that the convergence of primal-dual algorithms (TiCoPD, CP-SGD) are unaffected by the data heterogeneity issue.

Fig. 1 compares the training loss ($\max_{i\in[n]} f(\mathbf{X}_i^t)$), gradient norm ($\max_{i\in[n]} \|\nabla f(\mathbf{X}_i^t)\|^2$), consensus error ($\|\mathbf{X}^t\|^2_{\tilde{\mathbf{K}}}$), and test accuracy of the worst local model $\mathbf{X}_i^t$ against the iteration number and communication cost (in bits transmitted). Notice that for the uncompressed DGD method, we assume that a 32 bits full precision representation for a real number. We observe that both the primal-dual algorithms, TiCoPD, CP-SGD, achieve substantially better performance than the other algorithms that are primal-only. It illustrates

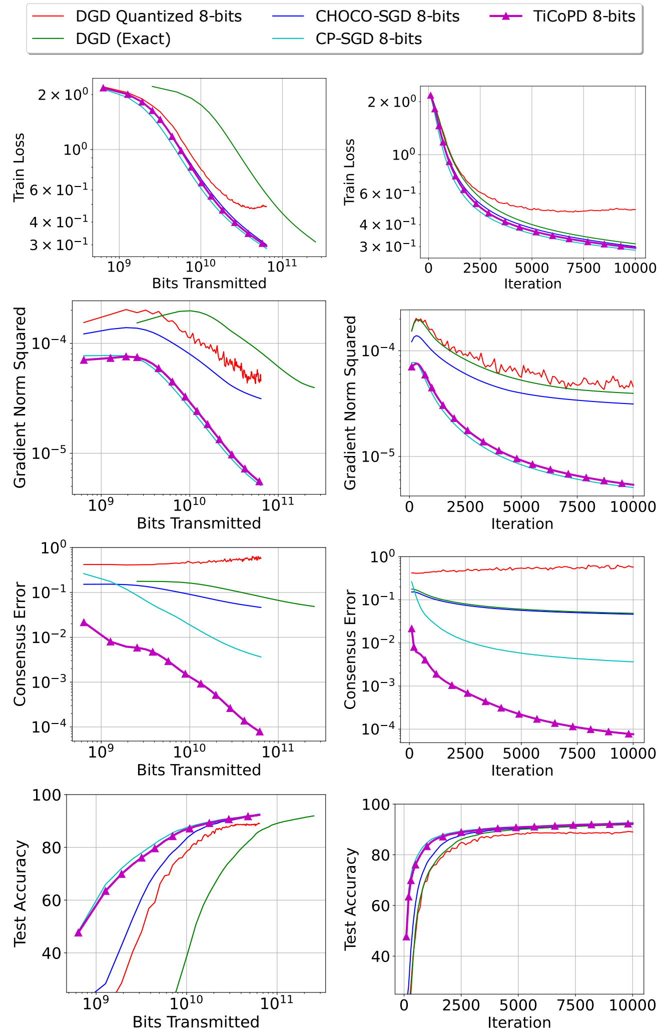

Fig. 1. Training a 2-layer feedforward network using the MNIST data. The bit-rates for communication quantization are displayed in the legend.

how data heterogeneity will dampen the convergence of primal-only algorithms. Furthermore, we note that the heuristic modification of DGD with direct quantization results in a non-converging algorithm, indicating the necessity of developing better approaches such as error feedback. When comparing between TiCoPD and CP-SGD, we can see that TiCoPD achieves lower consensus error by two orders of magnitude as the number of iterations grows. Meanwhile, the two algorithms share similar convergence rates on other metrics.

VI. Conclusions

This paper studies a communication efficient primal-dual algorithm for decentralized optimization with support for compression schemes such as quantized message exchanges. Unlike prior works, our key idea is to develop the algorithm from the augmented Lagrangian framework, and to incorporate classical designs such as majorization-minimization and two-timescale updates. The resultant algorithm is an algorithm that converges at the rate of $\mathcal{O}(T^{-1})$ for smooth (possibly non-convex) problems. We envisage that the proposed algorithmic framework can be extended for numerous tasks in signal processing and machine learning.

[1]Since it holds $\alpha = \frac{1}{\frac{1}{\tilde{\alpha}}+\theta L}$, once θ is fixed, the upper bound of α can be achieved by selecting a sufficiently small $\tilde{\alpha}$.

REFERENCES

[1] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5262–5276, 2010.

[2] 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," *Advances in neural information processing systems*, vol. 30, 2017.

[3] A. G. Dimakis, S. Kar, J. M. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.

[4] L. Xiao, S. Boyd, and S.-J. Kim, "Distributed average consensus with least-mean-square deviation," *Journal of parallel and distributed computing*, vol. 67, no. 1, pp. 33–46, 2007.

[5] S. Kar and J. M. Moura, "Consensus+ innovations distributed inference over networks: cooperation and sensing in networked systems," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 99–109, 2013.

[6] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[7] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Transactions on signal processing*, vol. 66, no. 11, pp. 2834–2848, 2018.

[8] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[9] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1245–1260, 2017.

[10] T.-H. Chang, M. Hong, H.-T. Wai, X. Zhang, and S. Lu, "Distributed learning in the nonconvex world: From batch data to streaming and beyond," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 26–38, 2020.

[11] D. Hajinezhad and M. Hong, "Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization," *Mathematical Programming*, vol. 176, no. 1, pp. 207–245, 2019.

[12] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Linear convergence of first-and zeroth-order primal–dual algorithms for distributed nonconvex optimization," *IEEE Transactions on Automatic Control*, vol. 67, no. 8, pp. 4194–4201, 2021.

[13] M. G. Rabbat and R. D. Nowak, "Quantized incremental algorithms for distributed optimization," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 798–808, 2005.

[14] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "An exact quantized decentralized gradient descent algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4934–4947, 2019.

[15] S. Magnússon, H. Shokri-Ghadikolaei, and N. Li, "On maintaining linear convergence of distributed learning and optimization under limited communication," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6101–6116, 2020.

[16] X. Liu and Y. Li, "Linear convergent decentralized optimization with compression," in *International Conference on Learning Representations*, 2021.

[17] H. Zhao, B. Li, Z. Li, P. Richtárik, and Y. Chi, "Beer: Fast $o(1/t)$ rate for decentralized nonconvex optimization with communication compression," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31653–31667, 2022.

[18] A. Koloskova, S. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in *International Conference on Machine Learning*, pp. 3478–3487, PMLR, 2019.

[19] A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi, "Decentralized deep learning with arbitrary communication compression," in *Proceedings of the 8th International Conference on Learning Representations*, 2019.

[20] C.-Y. Yau and H.-T. Wai, "Docom: Compressed decentralized optimization with near-optimal sample complexity," *arXiv preprint arXiv:2202.00255*, 2022.

[21] A. Xie, X. Yi, X. Wang, M. Cao, and X. Ren, "A communication-efficient stochastic gradient descent algorithm for distributed nonconvex optimization," *arXiv preprint arXiv:2403.01322*, 2024.

[22] S. U. Stich, "Local sgd converges fast and communicates little," *arXiv preprint arXiv:1805.09767*, 2018.

[23] H. Yu, S. Yang, and S. Zhu, "Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 5693–5700, 2019.

[24] D. Basu, D. Data, C. Karakus, and S. N. Diggavi, "Qsparse-local-sgd: Distributed sgd with quantization, sparsification, and local computations," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 217–226, 2020.

[25] N. Singh, D. Data, J. George, and S. Diggavi, "Sparq-sgd: Event-triggered and compressed communication in decentralized stochastic optimization," *arXiv preprint arXiv:1910.14280*, 2019.

[26] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2016.

[27] A. S. Mathkar and V. S. Borkar, "Nonlinear gossip," *SIAM Journal on Control and Optimization*, vol. 54, no. 3, pp. 1535–1557, 2016.

[28] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," *Advances in neural information processing systems*, vol. 30, 2017.

Appendix A
Proof of Theorem IV.4

For some constants $\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e} > 0$, we define the potential function F_t as

$$\begin{aligned} F_t = & f(\bar{\mathbf{X}}^t) + \mathtt{a}\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}} + \mathtt{b}\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{Q}}+\mathtt{c}\widetilde{\mathbf{K}}} \\ & + \mathtt{d}\left\langle \mathbf{X}^t \mid \mathbf{v}^t \right\rangle_{\widetilde{\mathbf{K}}} + \mathtt{e}\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2, \end{aligned} \tag{18}$$

Based on the update of algorithm, it follows:

$$\begin{aligned} \mathbb{E}\left[F_{t+1}\right] \le & \mathbb{E}\left[F_t\right] + \omega_f \mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{X}}^t)\right\|^2\right] + \alpha^2\omega_\sigma \sum_{i=1}^n \sigma_i^2 \\ & + \mathbb{E}\left[\omega_x\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}} + \omega_v\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}} + \omega_{\hat{x}}\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2 + \left\langle \mathbf{X}^t \mid \mathbf{v}^t \right\rangle_{\mathbf{W}_{xv}}\right], \end{aligned} \tag{19}$$

where we denote

$$\mathbf{W}_{xv} = \mathtt{a}\cdot -2(\widetilde{\mathbf{K}} - \alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}) + \mathtt{b}\cdot 2\alpha\eta(\widetilde{\mathbf{K}} + \mathtt{c}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}) - \mathtt{d}\cdot(\alpha\theta + \alpha\eta)\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} + \mathtt{e}\cdot 2\tilde{\delta}\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} \tag{20}$$

$$\omega_x = \frac{\alpha L^2}{n} + \mathtt{a}\cdot\left(\alpha - 2\alpha\theta\tilde{\rho}_2 + \alpha^2\theta^2\tilde{\rho}_1^2 + (3 + \frac{2}{\alpha})\alpha^2 L^2\right) \tag{21}$$

$$+ \mathtt{b}\cdot\left(4\alpha^2\eta^2\tilde{\rho}_1^2(\mathtt{c} + \tilde{\rho}_2^{-1}) + 9\alpha^2(\tilde{\rho}_2^{-1} + \mathtt{c})L^4\right) \tag{22}$$

$$+ \mathtt{d}\cdot\left(\frac{3}{2}\alpha + 3L^2\alpha + (\alpha\eta - \alpha^2\theta - 2\alpha^2\theta)\tilde{\rho}_1 + (\frac{\alpha}{2} - \alpha^2\eta\theta + \alpha^3\theta^2 + \frac{\alpha^3\eta^2}{2})\tilde{\rho}_1^2 + (\alpha + \frac{1}{2}\alpha^2)8\alpha^2 L^4\right) \tag{23}$$

$$+ \mathtt{e}\cdot\tilde{\delta}\left(2\alpha^2\theta^2\tilde{\rho}_1^2 + 3\alpha^2 L^2\right), \tag{24}$$

$$\omega_v = \mathtt{a}\cdot 3 + \mathtt{b}\cdot 2\alpha(\tilde{\rho}_2^{-1} + \mathtt{c}) + \mathtt{d}\cdot -\frac{1}{2} + \mathtt{e}\cdot 2\tilde{\delta}, \tag{25}$$

$$\omega_{\hat{x}} = \theta^2 M^2\frac{L\alpha + 2}{n}\alpha + \mathtt{a}\cdot\left((2\alpha + 3\alpha^2)\theta^2(\tilde{\rho}_1 + M)^2\right) \tag{26}$$

$$+ \mathtt{b}\cdot\left(4\alpha^2\eta^2\tilde{\rho}_1^2(\tilde{\rho}_2^{-1} + \mathtt{c}) + \alpha\eta^2 + 6\alpha^3(\tilde{\rho}_2^{-1} + \mathtt{c})\theta^2 M^2 L^2\right) \tag{27}$$

$$+ \mathtt{d}\cdot\left(3\alpha^2\theta^2(\tilde{\rho}_1 + M)^2) + \frac{\alpha\eta^2}{2}\tilde{\rho}_1^2 + 2(\alpha + \frac{1}{2}\alpha^2)\theta^2 M^2 L^2\alpha^2\right) \tag{28}$$

$$+ \mathtt{e}\cdot\left(2\tilde{\delta}\alpha^2\theta^2(\tilde{\rho}_1 + M)^2 + (1 - \frac{\delta}{2})^2 - 1\right) \tag{29}$$

$$\omega_f = -\frac{\alpha}{4} + \mathtt{b}\cdot 9\alpha^3(\tilde{\rho}_2^{-1} + \mathtt{c})nL^2 + \mathtt{d}\cdot 4(\alpha + \frac{1}{2}\alpha^2)\alpha^2 nL^2 + \mathtt{e}\cdot 2\tilde{\delta}\alpha^2 \tag{30}$$

Lemma A.1. *Under Assumption IV.1, when* $\alpha \le \frac{1}{4L}$,

$$\begin{aligned} \mathbb{E}\left[f(\bar{\mathbf{X}}^{t+1})\right] \le & \mathbb{E}\Bigg[f(\bar{\mathbf{X}}^t) - \frac{\alpha}{4}\left\|\nabla f(\bar{\mathbf{X}}^t)\right\|^2 \\ & + (\frac{\alpha L^2}{2n} + \frac{2\alpha^2 L^3}{n})\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\Bigg] + \frac{\alpha^2 L}{n^2}\sum_{i=1}^n \sigma_i^2 \\ & + \theta^2 M^2\frac{L\alpha + 2}{n}\alpha\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right], \end{aligned} \tag{31}$$

See Appendix B for the proof. Therefore, next we need to develop the bound of consensus error $\mathbb{E}\left[\|\mathbf{X}\|^2_{\widetilde{\mathbf{K}}}\right]$ and surrogate variable error $\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right]$.

Lemma A.2. *Under Assumptions IV.1–IV.2 and step size* $\alpha \le \frac{1}{\theta\tilde{\rho}_1}$, *the consensus error follows the inequality*

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{X}^{t+1}\|^2_{\widetilde{\mathbf{K}}}\right] \le & \big(1 + \alpha - 2\alpha\theta\tilde{\rho}_2 + \alpha^2\theta^2\tilde{\rho}_1^2 \\ & + (2\alpha + 3\alpha^2)L^2\big)\mathbb{E}\left[\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\right] - 2\mathbb{E}\left[\left\langle \mathbf{X}^t \mid \mathbf{v}^t \right\rangle_{(\mathbf{I} - \alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})\widetilde{\mathbf{K}}}\right] \\ & + 3\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}}\right] + (2\alpha + 3\alpha^2)\theta^2(\tilde{\rho}_1 + M)^2\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right], \end{aligned} \tag{32}$$

See Appendix C for the proof. The above lemma provides the updates of consensus error $\mathbb{E}\left[\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\right]$ depend on $\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}}\right]$ and the weighted inner product of $\mathbf{X}^t, \mathbf{v}^t$. Therefore, next we will consider the bound of these two terms.

Lemma A.3. *Under Assumption IV.1–IV.2 and step size* $\alpha \leq 1$*, for any constant* $\mathtt{c} > 0$*, it follows the inequality*

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{v}^{t+1}\|^2_{\widetilde{\mathbf{Q}}+\mathtt{c}\widetilde{\mathbf{K}}}\right] \overset{(67)}{\leq} (1+2\alpha)\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{Q}}+\mathtt{c}\widetilde{\mathbf{K}}}\right] \\
&+ \left(4\alpha^2\eta^2\tilde{\rho}_1^2(\mathtt{c}+\tilde{\rho}_2^{-1}) + 9\alpha^3(\tilde{\rho}_2^{-1}+\mathtt{c})L^4\right)\mathbb{E}\left[\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\right] \\
&+ 2\alpha\eta\mathbb{E}\left[\langle \mathbf{v}^t \mid \mathbf{X}^t\rangle_{\widetilde{\mathbf{K}}+\mathtt{c}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}}\right] + \left(4\alpha^2\eta^2\tilde{\rho}_1^2(\tilde{\rho}_2^{-1}+\mathtt{c})\right. \\
&\left.+ \alpha\eta^2 + 9\alpha^3(\tilde{\rho}_2^{-1}+\mathtt{c})\theta^2M^2L^2\right)\mathbb{E}\left[\left\|\hat{\mathbf{X}}^t - \mathbf{X}^t\right\|^2\right] \\
&+ 9\alpha^3(\tilde{\rho}_2^{-1}+\mathtt{c})nL^2\mathbb{E}\left[\left\|\nabla\mathbf{f}(\bar{\mathbf{X}}^t)\right\|^2\right].
\end{aligned} \tag{33}$$

See Appendix D for the proof.

Lemma A.4. *Under Assumption IV.1–IV.2, for any constant* $\mathtt{c} > 0, \delta_1 > 0$*, it follows the inequality*

$$\begin{aligned}
&\mathbb{E}\left[\langle \mathbf{X}^{t+1} \mid \mathbf{v}^{t+1}\rangle_{\widetilde{\mathbf{K}}}\right] \leq \mathbb{E}\left[\langle \mathbf{X}^t \mid \mathbf{v}^t\rangle_{\widetilde{\mathbf{K}}-(\alpha\theta+\alpha\eta)\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}}\right] \\
&- \frac{1}{2}\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}}\right] + \mathbf{N}\mathbb{E}\left[\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\right] + \left(3\alpha^2\theta^2(\tilde{\rho}_1+M)^2\right) \\
&+ \frac{\alpha\eta^2}{2}\tilde{\rho}_1^2 + 2(\alpha+\frac{1}{2}\alpha^2)\theta^2M^2L^2\alpha^2)\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right] \\
&+ (\alpha+\frac{1}{2}\alpha^2)4\alpha^2nL^2\mathbb{E}\left[\left\|\nabla\mathbf{f}(\bar{\mathbf{X}}^t)\right\|^2\right],
\end{aligned} \tag{34}$$

where $\mathbf{N} = \frac{3}{2}\alpha + 3L^2\alpha + (\alpha\eta - \alpha^2\theta - 2\alpha^2\theta)\tilde{\rho}_1 + (\frac{\alpha}{2} - \alpha^2\eta\theta + \alpha^3\theta^2 + \frac{\alpha^3\eta^2}{2})\tilde{\rho}_1^2 + (\alpha + \frac{1}{2}\alpha^2)8\alpha^2L^4$.

See Appendix E for the proof.

Lemma A.5. *Under Assumptions IV.1–IV.3, the mirror sequence error follows the inequality*

$$\begin{aligned}
&\mathbb{E}\left[\|\hat{\mathbf{X}}^{t+1} - \mathbf{X}^{t+1}\|^2\right] \\
&\leq \tilde{\delta}\left(2\alpha^2\theta^2\tilde{\rho}_1^2 + 3\alpha^2L^2\right)\mathbb{E}\left[\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\right] \\
&+ 2\tilde{\delta}\mathbb{E}\left[\langle \mathbf{X}^t \mid \mathbf{v}^t\rangle_{\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}}\right] + 2\tilde{\delta}\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}}\right] \\
&+ \left(2\tilde{\delta}\alpha^2\theta^2(\tilde{\rho}_1+M)^2 + (1-\frac{\delta}{2})^2\right)\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right] \\
&+ 2\tilde{\delta}\alpha^2\left\|\nabla\mathbf{f}(\bar{\mathbf{X}}^t)\right\|^2.
\end{aligned} \tag{35}$$

See Appendix F for the proof.

Lemma A.6. *Suppose that*

$$\begin{aligned}
&\mathtt{b} = \mathtt{a}\cdot\frac{1}{\alpha\eta}, \quad \mathtt{c} = \frac{(\alpha\theta+\alpha\eta)\mathtt{d} - 2\alpha\theta\mathtt{a} - \mathtt{e}2\tilde{\delta}\alpha\theta}{2\alpha\eta\mathtt{b}}, \\
&\mathtt{d} = \max\{24, 24\tilde{\rho}_2^{-1}\eta^{-1}, 192\frac{\tilde{\delta}\eta}{\delta}, 12\tilde{\delta}\}\mathtt{a}, \quad \mathtt{e} = \max\{\frac{16\eta}{\delta}, 1\}\mathtt{a},
\end{aligned} \tag{36}$$

then for $\eta > 0, \theta \geq \theta_{lb}, \alpha \leq \alpha_{ub}$*, it holds that* $F_t \geq f(\bar{\mathbf{X}}^t) \geq f^* > -\infty$*, and*

$$\begin{aligned}
&\mathbf{W}_{xv} = \mathbf{0}, \quad \omega_x \leq -\frac{1}{4}\alpha\theta\tilde{\rho}_2\mathtt{a}, \quad \omega_v \leq -2\mathtt{a}, \\
&\omega_{\hat{x}} \leq -\frac{\delta}{16}\mathtt{a}, \quad \omega_f \leq -\frac{1}{16}\alpha.
\end{aligned} \tag{37}$$

Summing up the inequality (18) from $t = 0$ to $t = T-1$ and divide both sides by T concludes the proof of the theorem.

Appendix B
Proof of Lemma A.1

Firstly, by (14) and (6), the updates of $\bar{\mathbf{X}}$ follows:

$$\bar{\mathbf{X}}^{t+1} = \frac{1}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)\left(\mathbf{X}^t + (1-\beta)(\hat{\mathbf{X}}^t - \mathbf{X}^t) - \alpha\left(\nabla\mathbf{f}(\mathbf{X}^t) + \widetilde{\mathbf{A}}^\top\lambda^t + \theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\hat{\mathbf{X}}^t\right)\right) \tag{38}$$

$$= \bar{\mathbf{X}}^t + \frac{1-\beta}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)(\hat{\mathbf{X}}^t - \mathbf{X}^t) - \frac{\alpha}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)\nabla\mathbf{f}(\mathbf{X}^t), \tag{39}$$

where the last equation utilize the fact $\mathbf{1}^\top\widetilde{\mathbf{A}}^\top = \mathbf{0}$.

By Assumption IV.1 and (39),

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{f}(\bar{\mathbf{X}}^{t+1})\right] && (40)\\
\le&\mathbb{E}\left[\mathbf{f}(\bar{\mathbf{X}}^{t}) + \left\langle \nabla\mathbf{f}(\bar{\mathbf{X}}^{t}) \;\middle|\; \frac{1-\beta}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)(\hat{\mathbf{X}}^t - \mathbf{X}^t) - \frac{\alpha}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)\nabla\mathbf{f}(\mathbf{X}^t)\right\rangle\right]\\
&+\mathbb{E}\left[\frac{L}{2}\left\|\frac{1-\beta}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)(\hat{\mathbf{X}}^t - \mathbf{X}^t) - \frac{\alpha}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)\nabla\mathbf{f}(\mathbf{X}^t)\right\|^2\right] && (41)
\end{aligned}
$$

The second term of (41) can be bounded as

$$
\begin{aligned}
&\mathbb{E}\left[\left\langle \nabla\mathbf{f}(\bar{\mathbf{X}}^{t}) \;\middle|\; \frac{1-\beta}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)(\hat{\mathbf{X}}^t - \mathbf{X}^t) - \frac{\alpha}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)\nabla\mathbf{f}(\mathbf{X}^t)\right\rangle\right] && (42)\\
\le&\frac{\alpha}{8}\mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{X}}^{t})\|^2\right] + \frac{2(1-\beta)^2}{n\alpha}\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right] - \frac{\alpha}{2}\mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{X}}^{t})\|^2\right] + \frac{\alpha}{2n}\mathbb{E}\left[\left\|\nabla\mathbf{f}(\bar{\mathbf{X}}^{t}) - (\mathbf{1}^\top \otimes \mathbf{I}_d)\nabla\mathbf{f}(\mathbf{X}^t)\right\|^2\right] && (43)\\
\le&-\frac{3\alpha}{8}\mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{X}}^{t})\|^2\right] + \frac{2\theta^2 M^2}{n}\alpha\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right] + \frac{\alpha L^2}{2n}\mathbb{E}\left[\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\right], && (44)
\end{aligned}
$$

where we use (IV.1) and the following relationship between α and $1-\beta$:

$$
1-\beta = 1 - \frac{1}{1+\tilde{\alpha}\theta M} = \frac{\theta M}{\frac{1}{\tilde{\alpha}} + \theta M} = \theta M \alpha. \tag{45}
$$

The third term of (41) can be bounded as

$$
\begin{aligned}
&\mathbb{E}\left[\frac{L}{2}\left\|\frac{1-\beta}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)(\hat{\mathbf{X}}^t - \mathbf{X}^t) - \frac{\alpha}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)\nabla\mathbf{f}(\mathbf{X}^t)\right\|^2\right] && (46)\\
\le&\frac{\alpha^2 L}{n^2}\mathbb{E}\left[\left\|(\mathbf{1}^\top \otimes \mathbf{I}_d)\nabla\mathbf{f}(\mathbf{X}^t)\right\|^2\right] + \frac{(1-\beta)^2 L}{n^2}\mathbb{E}\left[\left\|\frac{1}{n}(\mathbf{1}^\top \otimes \mathbf{I}_d)(\hat{\mathbf{X}}^t - \mathbf{X}^t)\right\|^2\right] && (47)\\
\le&\frac{2\alpha^2 L^3}{n}\mathbb{E}\left[\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\right] + 2\alpha^2 L\mathbb{E}\left[\|\nabla\mathbf{f}(\bar{\mathbf{X}}^{t})\|^2\right] + \frac{\theta^2 M^2 L\alpha^2}{n}\mathbb{E}\left[\left\|\hat{\mathbf{X}}^t - \mathbf{X}^t\right\|^2\right]. && (48)
\end{aligned}
$$

Combining the above and setting the step size $\alpha \le \frac{1}{4L}$ concludes the proof of the lemma. □

APPENDIX C
PROOF OF LEMMA A.2

To facilitate our analysis, we introduce the following quantities:

$$
\begin{aligned}
\mathbf{e}_g^t &:= \alpha\left(\nabla\mathbf{f}((\mathbf{1}_n \otimes \mathbf{I}_d)\bar{\mathbf{X}}^t) - \nabla\mathbf{f}(\mathbf{X}^t)\right), && (49)\\
\mathbf{v}^t &:= \alpha\widetilde{\boldsymbol{\lambda}}^t + \alpha\nabla\mathbf{f}((\mathbf{1}_n \otimes \mathbf{I}_d)\bar{\mathbf{X}}^t), && (50)\\
\mathbf{e}_{\hat{\mathbf{X}}}^t &:= \left(\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} + (1-\beta)\mathbf{I}_{nd}\right)(\hat{\mathbf{X}}^t - \mathbf{X}^t). && (51)
\end{aligned}
$$

By (49), (50) and (51), we can rewrite the primal update as

$$
\begin{aligned}
\mathbf{X}^{t+1} &= \beta\mathbf{X}^t + (1-\beta)\hat{\mathbf{X}}^t - \alpha(\nabla\mathbf{f}(\mathbf{X}^t) + \widetilde{\lambda}^t + \theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\hat{\mathbf{X}}^t) && (52)\\
&= \left(\mathbf{I} - \alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\mathbf{X}^t - \mathbf{v}^t + \mathbf{e}_g^t + \mathbf{e}_{\hat{\mathbf{X}}}^t && (53)
\end{aligned}
$$

By (53), the consensus error can be measured by

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{X}^{t+1}\|^2_{\widetilde{\mathbf{K}}}\right] = \mathbb{E}\left[\left\|\left(\mathbf{I} - \alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\mathbf{X}^t - \mathbf{v}^t + \mathbf{e}_s^t + \mathbf{e}_g^t + \mathbf{e}_{\hat{\mathbf{X}}}^t\right\|^2_{\widetilde{\mathbf{K}}}\right] && (54)\\
&\le \mathbb{E}\left[\left\|\left(\mathbf{I} - \alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\mathbf{X}^t\right\|^2_{\widetilde{\mathbf{K}}}\right] + 3\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}}\right] + 3\mathbb{E}\left[\|\mathbf{e}_g^t\|^2_{\widetilde{\mathbf{K}}}\right] + 3\mathbb{E}\left[\|\mathbf{e}_{\hat{\mathbf{X}}}^t\|^2_{\widetilde{\mathbf{K}}}\right]\\
&\quad - 2\mathbb{E}\left[\left\langle\left(\mathbf{I} - \alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\mathbf{X}^t \;\middle|\; \mathbf{v}^t - \mathbf{e}_g^t - \mathbf{e}_{\hat{\mathbf{X}}}^t\right\rangle_{\widetilde{\mathbf{K}}}\right] && (55)\\
&= \mathbb{E}\left[\left\|\left(\mathbf{I} - \alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\mathbf{X}^t\right\|^2_{\widetilde{\mathbf{K}}}\right] + 3\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}}\right] + 3\mathbb{E}\left[\|\mathbf{e}_g^t\|^2_{\widetilde{\mathbf{K}}}\right] + 3\mathbb{E}\left[\|\mathbf{e}_{\hat{\mathbf{X}}}^t\|^2_{\widetilde{\mathbf{K}}}\right]\\
&\quad - 2\mathbb{E}\left[\left\langle\left(\mathbf{I} - \alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\mathbf{X}^t \;\middle|\; \mathbf{v}^t\right\rangle_{\widetilde{\mathbf{K}}}\right] + 2\alpha\mathbb{E}\left[\left\langle\left(\mathbf{I} - \alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\mathbf{X}^t \;\middle|\; \frac{1}{\alpha}(\mathbf{e}_g^t + \mathbf{e}_{\hat{\mathbf{X}}}^t)\right\rangle_{\widetilde{\mathbf{K}}}\right] && (56)\\
&\le (1+\alpha)(1 - 2\alpha\theta\tilde{\rho}_2 + \alpha^2\theta^2\tilde{\rho}_1^2)\mathbb{E}\left[\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\right] - 2\mathbb{E}\left[\left\langle\mathbf{X}^t \;\middle|\; \mathbf{v}^t\right\rangle_{(\mathbf{I} - \alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})\widetilde{\mathbf{K}}}\right] && (57)\\
&\quad + 3\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}}\right] + (3 + \frac{2}{\alpha})\mathbb{E}\left[\|\mathbf{e}_g^t\|^2_{\widetilde{\mathbf{K}}}\right] + (3 + \frac{2}{\alpha})\mathbb{E}\left[\|\mathbf{e}_{\hat{\mathbf{X}}}^t\|^2_{\widetilde{\mathbf{K}}}\right]. && (58)
\end{aligned}
$$

The $,\mathbb{E}\left[\|\mathbf{e}_g^t\|_{\widetilde{\mathbf{K}}}^2\right], \mathbb{E}\left[\|\mathbf{e}_{\hat{\mathbf{X}}}^t\|_{\widetilde{\mathbf{K}}}^2\right]$ term can be simplified as follow. Using the fact that each difference term in $\mathbf{e}_s^t$ has mean zero and are independent when conditioned on $(\mathbf{X}^t, \boldsymbol{\lambda}^t)$, we can obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{e}_g^t\|_{\widetilde{\mathbf{K}}}^2\right] = \alpha^2\mathbb{E}\left[\|\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t) - \nabla\mathbf{f}(\mathbf{X}^t)\|_{\widetilde{\mathbf{K}}}^2\right] \leq \mathbb{E}\left[\|\mathbf{e}_g^t\|^2\right] = \alpha^2\mathbb{E}\left[\|\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t) - \nabla\mathbf{f}(\mathbf{X}^t)\|^2\right] \\
&\overset{(11)}{\leq}\alpha^2L^2\mathbb{E}\left[\|(\mathbf{1}\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t - \mathbf{X}^t\|^2\right] = \alpha^2L^2\mathbb{E}\left[\|\widetilde{\mathbf{K}}\mathbf{X}^t\|^2\right] = \alpha^2L^2\mathbb{E}\left[\|\mathbf{X}^t\|_{\widetilde{\mathbf{K}}}^2\right],
\end{aligned} \tag{59}$$

$$\mathbb{E}\left[\|\mathbf{e}_{\hat{\mathbf{X}}}^t\|_{\widetilde{\mathbf{K}}}^2\right] \leq \mathbb{E}\left[\|\mathbf{e}_{\hat{\mathbf{X}}}^t\|^2\right] = \mathbb{E}\left[\left\|\left(\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} + (1-\beta)\mathbf{I}\right)(\hat{\mathbf{X}}^t - \mathbf{X}^t)\right\|^2\right] \tag{60}$$

$$\leq\alpha^2\theta^2(\tilde{\rho}_1 + M)^2\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right], \tag{61}$$

Combining the upper bounds of $,\mathbb{E}\left[\|\mathbf{e}_g^t\|_{\widetilde{\mathbf{K}}}^2\right], \mathbb{E}\left[\|\mathbf{e}_{\hat{\mathbf{X}}}^t\|_{\widetilde{\mathbf{K}}}^2\right]$ and step size $\alpha \leq \frac{1}{\theta\tilde{\rho}_1}$, the proof of Lemma A.2 has been completed. □

Appendix D
Proof of Lemma A.3

Consider expanding $\mathbf{v}^{t+1}$ as

$$\mathbf{v}^{t+1} \overset{(50)}{=} \alpha\widetilde{\lambda}^{t+1} + \alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) \tag{62}$$

$$= \mathbf{v}^t + \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{X}^t + \eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}(\hat{\mathbf{X}}^t - \mathbf{X}^t) + \alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t) \tag{63}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{v}^{t+1}\|_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}^2\right] \leq \mathbb{E}\left[\|\mathbf{v}^t\|_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}^2\right] \\
&+ 2\mathbb{E}\left[\left\langle\mathbf{v}^t \,\middle|\, \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{X}^t + \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}(\hat{\mathbf{X}}^t - \mathbf{X}^t) + \alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t)\right\rangle_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}\right] \\
&+ 2\mathbb{E}\left[\|\alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{X}^t + \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}(\hat{\mathbf{X}}^t - \mathbf{X}^t)\|_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}^2\right] + 2\mathbb{E}\left[\|\alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t)\|_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}^2\right]
\end{aligned} \tag{64}$$

Due to Assumption IV.2, it holds:

$$\mathbb{E}\left[\|\cdot\|_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}^2\right] \leq (\tilde{\rho}_2^{-1} + \mathsf{c})\mathbb{E}\left[\|\cdot\|_{\widetilde{\mathbf{K}}}^2\right]. \tag{65}$$

The second term of (64) can be simplified as

$$2\mathbb{E}\left[\left\langle\mathbf{v}^t \,\middle|\, \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{X}^t + \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}(\hat{\mathbf{X}}^t - \mathbf{X}^t) + \alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t)\right\rangle_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}\right] \tag{66}$$

$$\begin{aligned}
&\leq 2\alpha\eta\mathbb{E}\left[\left\langle\mathbf{v}^t \,\middle|\, \mathbf{X}^t\right\rangle_{(\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}})\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}}\right] + 2\alpha\mathbb{E}\left[\|\mathbf{v}^t\|_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}^2\right] + \alpha\eta^2\tilde{\rho}_1^2\mathbb{E}\left[\left\|\hat{\mathbf{X}}^t - \mathbf{X}^t\right\|_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}^2\right] \\
&+ \alpha(\tilde{\rho}_2^{-1} + \mathsf{c})\mathbb{E}\left[\left\|\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t)\right\|^2\right]
\end{aligned} \tag{67}$$

The third term of (64) can be simplified as

$$2\mathbb{E}\left[\|\alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{X}^t + \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}(\hat{\mathbf{X}}^t - \mathbf{X}^t)\|_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}^2\right] \tag{68}$$

$$\leq 4\alpha^2\eta^2\tilde{\rho}_1^2\mathbb{E}\left[\|\mathbf{X}^t\|_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}^2\right] + 4\alpha^2\eta^2\tilde{\rho}_1^2(\tilde{\rho}_2^{-1} + \mathsf{c})\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right] \tag{69}$$

The fourth term of (64) can be simplified as

$$2\mathbb{E}\left[\left\|\alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t)\right\|_{\widetilde{\mathbf{Q}}+\mathsf{c}\widetilde{\mathbf{K}}}^2\right] \leq 2\alpha^2\cdot(\tilde{\rho}_2^{-1} + \mathsf{c})\mathbb{E}\left[\left\|\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t)\right\|^2\right] \tag{70}$$

The proof is concluded by combining the above inequalities and simplifing $\widetilde{\mathbf{Q}}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} = \widetilde{\mathbf{K}}$ and $\widetilde{\mathbf{K}}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} = \widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}$. □

Appendix E
Proof of Lemma A.4

By (53) and (63),

$$\mathbb{E}\left[\left\langle\mathbf{X}^{t+1} \,\middle|\, \mathbf{v}^{t+1}\right\rangle_{\widetilde{\mathbf{K}}}\right] \tag{71}$$

$$= \mathbb{E}\left[\left\langle\mathbf{X}^t \,\middle|\, \mathbf{v}^t\right\rangle_{(\mathbf{I}-\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})\widetilde{\mathbf{K}}-\alpha\eta\widetilde{\mathbf{K}}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}}\right] + \mathbb{E}\left[\|\mathbf{X}^t\|_{\alpha\eta(\mathbf{I}-\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})\widetilde{\mathbf{K}}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}}^2\right] \tag{72}$$

$$+ \alpha\mathbb{E}\left[\left\langle\mathbf{X}^t \,\middle|\, \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t)\right\rangle_{(\mathbf{I}-\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})\widetilde{\mathbf{K}}}\right] \tag{73}$$

$$- \mathbb{E}\left[\|\mathbf{v}^t\|_{\widetilde{\mathbf{K}}}^2\right] - \alpha\mathbb{E}\left[\left\langle\mathbf{v}^t \,\middle|\, \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t)\right\rangle_{\widetilde{\mathbf{K}}}\right] \tag{74}$$

$$+ \mathbb{E}\left[\left\langle\left(\mathbf{I} - \alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\mathbf{X}^t - \mathbf{v}^t \,\middle|\, \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}(\hat{\mathbf{X}}^t - \mathbf{X}^t)\right\rangle_{\widetilde{\mathbf{K}}}\right] \tag{75}$$

$$+ \mathbb{E}\left[\left\langle\mathbf{e}_g^t + \mathbf{e}_{\hat{\mathbf{X}}}^t \,\middle|\, \mathbf{v}^t + \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{X}^t + \alpha(\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t))\right\rangle_{\widetilde{\mathbf{K}}}\right] \tag{76}$$

Now notice that by applying Young's inequality on the third, fifth, sixth and seventh terms of the above, we get

$$\alpha\mathbb{E}\left[\left\langle \mathbf{X}^t \mid \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t})\right\rangle_{(\mathbf{I}-\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})\widetilde{\mathbf{K}}}\right] \tag{77}$$

$$\leq \frac{\alpha}{2}\cdot\mathbb{E}\left[\|\mathbf{X}^t\|^2_{((\mathbf{I}-\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}})\widetilde{\mathbf{K}})^2}\right] + \frac{\alpha}{2}\cdot\mathbb{E}\left[\|\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t})\|^2\right], \tag{78}$$

$$-\alpha\mathbb{E}\left[\left\langle \mathbf{v}^t \mid \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t})\right\rangle_{\widetilde{\mathbf{K}}}\right] \tag{79}$$

$$\leq \frac{\alpha}{2}\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}}\right] + \frac{\alpha}{2}\mathbb{E}\left[\|\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t})\|^2\right] \tag{80}$$

$$\mathbb{E}\left[\left\langle \left(\mathbf{I}-\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\mathbf{X}^t - \mathbf{v}^t \;\middle|\; \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}(\hat{\mathbf{X}}^t - \mathbf{X}^t)\right\rangle_{\widetilde{\mathbf{K}}}\right] \tag{81}$$

$$\leq \alpha\mathbb{E}\left[\|\left(\mathbf{I}-\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\right)\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\right] + \alpha\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}}\right] + \frac{\alpha\eta^2}{2}\tilde{\rho}_1^2\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\right] \tag{82}$$

$$\mathbb{E}\left[\left\langle \mathbf{e}_g^t + \mathbf{e}_{\hat{\mathbf{X}}}^t \;\middle|\; \mathbf{v}^t + \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\hat{\mathbf{X}}^t + \alpha(\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t}))\right\rangle_{\widetilde{\mathbf{K}}}\right] \tag{83}$$

$$\leq \frac{3}{2}\mathbb{E}\left[\|\mathbf{e}_g^t + \mathbf{e}_{\hat{\mathbf{X}}}^t\|^2_{\widetilde{\mathbf{K}}}\right] + \frac{1}{6}\mathbb{E}\left[\|\mathbf{v}^t + \alpha\eta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\hat{\mathbf{X}}^t + \alpha(\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t}))\|^2_{\widetilde{\mathbf{K}}}\right] \tag{84}$$

$$\leq 3\mathbb{E}\left[\|\mathbf{e}_g^t\|^2_{\widetilde{\mathbf{K}}}\right] + 3\alpha^2\theta^2(\tilde{\rho}_1+M)^2\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right] + \frac{1}{2}\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}}\right] + \frac{\alpha^2\eta^2}{2}\mathbb{E}\left[\|\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}}\right]$$
$$+\frac{1}{2}\alpha^2\mathbb{E}\left[\|\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t})\|^2_{\widetilde{\mathbf{K}}}\right] \tag{85}$$

The proof is concluded by combining Lemma H.1 and simplifying $\widetilde{\mathbf{A}}\widetilde{\mathbf{K}} = \widetilde{\mathbf{A}}$. □

Appendix F
Proof of Lemma A.5

By Lemma IV.3, it holds that

$$\begin{aligned}\mathbb{E}(\|\hat{\mathbf{X}}^{t+1} - \mathbf{X}^{t+1}\|^2 &= \mathbb{E}(\|\hat{\mathbf{X}}^{t} - \mathbf{X}^{t+1} + Q(\mathbf{X}^{t+1} - \hat{\mathbf{X}}^t;\xi_q^{t+1})\|^2\\ &\leq (1-\delta)^2\mathbb{E}\left[(1+\tau)\|\mathbf{X}^{t+1}-\mathbf{X}^t\|^2 + (1+\frac{1}{\tau})\|\mathbf{X}^t - \hat{\mathbf{X}}^t\|^2\right]\\ &\leq \tilde{\delta}\mathbb{E}\left[\|\mathbf{X}^{t+1}-\mathbf{X}^t\|^2\right] + (1-\frac{\delta}{2})^2\mathbb{E}\left[\|\mathbf{X}^t - \hat{\mathbf{X}}^t\|^2\right],\end{aligned} \tag{86}$$

where we denote $\tau = \frac{(1-\delta)^2}{\delta - \frac{3}{4}\delta^2}$, which satisfied $(1-\delta)^2(1+\frac{1}{\tau}) = (1-\frac{\delta}{2})^2$. The last term $\mathbb{E}\left[\|\mathbf{X}^t - \hat{\mathbf{X}}^t\|^2\right]$ can be simplified as

$$\mathbb{E}\left[\|\mathbf{X}^{t+1}-\mathbf{X}^t\|^2\right] = \mathbb{E}\left[\left\|-\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{X}^t - \mathbf{v}^t + \mathbf{e}_s^t + \mathbf{e}_g^t + \mathbf{e}_{\hat{\mathbf{X}}}^t\right\|^2\right] \tag{87}$$

$$\begin{aligned}\leq& 2\mathbb{E}\left[\left\|-\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}\mathbf{X}^t\right\|^2\right] - 2\mathbb{E}\left[\left\langle \mathbf{X}^t \mid \mathbf{v}^t\right\rangle_{-\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}}\right]\\ &+2\mathbb{E}\left[\|\mathbf{v}^t\|^2\right] + 3\mathbb{E}\left[\|\mathbf{e}_g^t\|^2\right] + 2\mathbb{E}\left[\|\mathbf{e}_{\hat{\mathbf{X}}}^t\|^2\right] + 2\mathbb{E}\left[\|\mathbf{e}_s^t\|^2\right]\end{aligned} \tag{88}$$

$$\begin{aligned}\overset{(i)}{\leq}& \left(2\alpha^2\theta^2\tilde{\rho}_1^2 + 3\alpha^2L^2 + 2\alpha^2\theta^2\sigma_A^2\tilde{\rho}_1\right)\mathbb{E}\left[\left\|\mathbf{X}^t\right\|^2_{\widetilde{\mathbf{K}}}\right] + 2\mathbb{E}\left[\left\langle \mathbf{X}^t \mid \mathbf{v}^t\right\rangle_{\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}}\right]\\ &+2\mathbb{E}\left[\|\mathbf{v}^t\|^2\right] + 2\alpha^2\theta^2(\tilde{\rho}_1+M)^2\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right] + 2\alpha^2\sum_{i=1}^n\sigma_i^2,\end{aligned} \tag{89}$$

where (i) uses the bound of , $\mathbb{E}\left[\|\mathbf{e}_g^t\|^2_{\widetilde{\mathbf{K}}}\right]$, $\mathbb{E}\left[\|\mathbf{e}_{\hat{\mathbf{X}}}^t\|^2_{\widetilde{\mathbf{K}}}\right]$. Notice that $\widetilde{\lambda}^t = \widetilde{\mathbf{A}}^\top\lambda^t \Rightarrow (\mathbf{1}\otimes\mathbf{I_d})^\top\widetilde{\lambda}^t = 0$, it follows

$$\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}} = \|\alpha\widetilde{\lambda}^t + \alpha\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t)\|^2_{(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\top)\otimes\mathbf{I}_d} = \|\mathbf{v}^t\|^2 - \alpha^2\|\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^t)\|^2_{\frac{1}{n}\mathbf{1}\mathbf{1}^\top\otimes\mathbf{I}_d} \tag{90}$$

$$= \|\mathbf{v}^t\|^2 - \alpha^2\left\|\nabla\mathbf{f}(\bar{\mathbf{X}}^t)\right\|^2. \tag{91}$$

It follows

$$\mathbb{E}\left[\|\mathbf{X}^{t+1}-\mathbf{X}^t\|^2\right] \tag{92}$$

$$\begin{aligned}\leq& \left(2\alpha^2\theta^2\tilde{\rho}_1^2 + 3\alpha^2L^2 + 2\alpha^2\theta^2\sigma_A^2\tilde{\rho}_1\right)\mathbb{E}\left[\left\|\mathbf{X}^t\right\|^2_{\widetilde{\mathbf{K}}}\right] + 2\mathbb{E}\left[\left\langle \mathbf{X}^t \mid \mathbf{v}^t\right\rangle_{\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}}\right]\\ &+2\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}}\right] + 2\alpha^2\theta^2(\tilde{\rho}_1+M)^2\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right] + 2\alpha^2\sum_{i=1}^n\sigma_i^2 + 2\alpha^2\left\|\nabla\mathbf{f}(\bar{\mathbf{X}}^t)\right\|^2\end{aligned} \tag{93}$$

Combing the above, the proof is concluded. □

APPENDIX G
PROOF OF LEMMA A.6

Firstly, we show the lower bound of F_t for any $t \geq 0$. By the inequality $|\langle \mathbf{x} \mid \mathbf{y}\rangle| \leq \frac{1}{2\delta_0}\|\mathbf{x}\|^2 + \frac{\delta_0}{2}\|\mathbf{y}\|^2$ for any $\delta_0 > 0$,

$$F_t \geq \mathbf{f}(\bar{\mathbf{X}}^t) + \mathtt{a}\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}} + \|\mathbf{v}^t\|^2_{\mathtt{b}\widetilde{\mathbf{Q}}+\mathtt{bc}\widetilde{\mathbf{K}}} - \frac{\mathtt{d}}{2\delta_0}\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}} - \frac{\mathtt{d}\delta_0}{2}\|\mathbf{v}^t\|^2_{\widetilde{\mathbf{K}}} + \mathtt{e}\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2 \tag{94}$$

$$= \mathbf{f}(\bar{\mathbf{X}}^t) + (\mathtt{a} - \frac{\mathtt{d}}{2\delta_0})\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}} + \|\mathbf{v}^t\|^2_{\mathtt{b}\widetilde{\mathbf{Q}}+(\mathtt{bc}-\mathtt{d}\delta_0/2)\widetilde{\mathbf{K}}} + \mathtt{e}\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2 \tag{95}$$

$$\overset{(\delta_0=\frac{\mathtt{d}}{2\mathtt{a}})}{=} \mathbf{f}(\bar{\mathbf{X}}^t) + 0\cdot\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}} + \|\mathbf{v}^t\|^2_{\mathtt{b}\widetilde{\mathbf{Q}}+(\mathtt{bc}-\frac{\mathtt{d}^2}{4\mathtt{a}})\widetilde{\mathbf{K}}} + \mathtt{e}\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2 \tag{96}$$

$$\overset{(12)}{\geq} \mathbf{f}(\bar{\mathbf{X}}^t) + \|\mathbf{v}^t\|^2_{(\mathtt{b}\cdot\tilde{\rho}_1^{-1}+\mathtt{bc}-\frac{\mathtt{d}^2}{4\mathtt{a}})\widetilde{\mathbf{K}}} + \mathtt{e}\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2 \tag{97}$$

Assume that $\mathtt{d} = \delta_1\mathtt{a}, \mathtt{e} = \delta_2\mathtt{a},\ \delta_1, \delta_2 \geq 0$. By choosing $\delta_1 \geq 2\delta_2 + 2, \alpha \leq \frac{4}{\eta\tilde{\rho}_1\delta_1^2}$, it holds

$$\mathtt{b}\cdot\tilde{\rho}_1^{-1} + \mathtt{bc} - \frac{\mathtt{d}^2}{4\mathtt{a}} = \mathtt{a}\left(\frac{1}{\alpha\eta\tilde{\rho}_1} + \alpha(\frac{\delta_1}{2}(\theta+\eta) - \theta) - \delta_2\alpha\theta - \frac{\delta_1^2}{4}\right) \geq 0. \tag{98}$$

Therefore, we can obtain

$$F_t \geq \mathbf{f}(\bar{\mathbf{X}}^t) \geq f^* > -\infty. \tag{99}$$

If the parameters are choosed according to (36), then it holds

$$\mathbf{W}_{xv} = -\mathtt{a}\cdot 2\widetilde{\mathbf{K}} + \mathtt{a}\cdot 2\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} + \mathtt{b}\cdot 2\alpha\eta(\widetilde{\mathbf{K}} + \mathtt{c}\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}}) - \mathtt{d}\cdot(\alpha\theta+\alpha\eta)\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} + \mathtt{e}2\tilde{\delta}\alpha\theta\widetilde{\mathbf{A}}^\top\widetilde{\mathbf{A}} = 0, \tag{100}$$

and the inner product term in (18) vanishes. Combining (36), we can obtain

$$\mathtt{c} = \alpha(\frac{\delta_1}{2}(\theta+\eta) - \theta) - \delta_2\alpha\theta. \tag{101}$$

By choosing $\alpha \leq \frac{1}{\tilde{\rho}_2(\frac{\delta_1}{2}(\theta+\eta)-\theta)}$ and (12), we can simplify $c \leq \tilde{\rho}_2^{-1}$. Now we argue that $\omega_x, \omega_v, \omega_{\hat{x}}, \omega_f, \omega_\sigma$ are negative.

To upper bound ω_x, by choosing $\alpha \leq 1, \eta \geq 1$, we have

$$\begin{aligned}\omega_x \leq& \frac{\alpha L^2}{n} + \mathtt{a}\cdot\alpha\left(1 + 2L + 8\eta\tilde{\rho}_1^2\tilde{\rho}_2^{-1} + \delta_1(\frac{3}{2} + 3L^2 + \eta\tilde{\rho}_1 + \frac{\tilde{\rho}_1^2}{2}) - 2\theta\tilde{\rho}_2\right) &(102)\\ &+ \mathtt{a}\cdot\left(\alpha^2\theta^2\tilde{\rho}_1^2 + 3\alpha^2L^2 + 18\alpha^2\tilde{\rho}_2^{-1}L^4\eta^{-1}\right) &(103)\\ &+ \mathtt{a}\cdot\delta_1\left((-\alpha^2\theta - 2\alpha^2\theta)\tilde{\rho}_1 - \alpha^2\eta\theta\tilde{\rho}_1^2 + \alpha^2\theta^2\tilde{\rho}_1^2 + \alpha^2\frac{\eta^2}{2}\tilde{\rho}_1^2 + 12\alpha^2L^4\right) &(104)\\ &+ \mathtt{a}\cdot\tilde{\delta}\delta_2\left(2\alpha^2\theta^2\tilde{\rho}_1^2 + 3\alpha^2L^2\right). &(105)\end{aligned}$$

It holds that $\omega_x \leq -\frac{1}{4}\alpha\theta\tilde{\rho}_2\mathtt{a} < 0$, with the following step size condition:

$$\begin{cases} \frac{\alpha L^2}{n} \leq \frac{1}{4}\alpha\theta\tilde{\rho}_2\mathtt{a} & \Leftrightarrow \theta \geq \frac{4L^2}{n\tilde{\rho}_2\mathtt{a}},\\ \theta \geq 2(1 + 2L + 8\eta\tilde{\rho}_1^2\tilde{\rho}_2^{-1} + \delta_1(\frac{3}{2} + 3L^2 + \eta\tilde{\rho}_1 + \frac{\tilde{\rho}_1^2}{2}))\tilde{\rho}_2^{-1}, & \\ \alpha^2\theta^2\tilde{\rho}_1^2 \leq \frac{1}{10}\alpha\theta\tilde{\rho}_2 & \Leftrightarrow \alpha \leq \frac{\tilde{\rho}_2}{10\theta\tilde{\rho}_1^2},\\ 3\alpha^2L^2 \leq \frac{1}{10}\alpha\theta\tilde{\rho}_2 & \Leftrightarrow \alpha \leq \frac{\theta\tilde{\rho}_2}{30L^2},\\ 18\alpha^2\tilde{\rho}_2^{-1}L^4\eta^{-1} \leq \frac{1}{10}\alpha\theta\tilde{\rho}_2 & \Leftrightarrow \alpha \leq \frac{\theta\tilde{\rho}_2^2\eta}{180L^4},\\ \delta_1\alpha^2\theta^2\tilde{\rho}_1^2 \leq \frac{1}{10}\alpha\theta\tilde{\rho}_2 & \Leftrightarrow \alpha \leq \frac{\tilde{\rho}_2}{10\delta_1\theta\tilde{\rho}_1^2},\\ \delta_1\alpha^2\frac{\eta^2}{2}\tilde{\rho}_1^2 \leq \frac{1}{10}\alpha\theta\tilde{\rho}_2 & \Leftrightarrow \alpha \leq \frac{\theta\tilde{\rho}_2}{5\delta_1\eta^2\tilde{\rho}_1^2},\\ \delta_1 12\alpha^2L^4 \leq \frac{1}{10}\alpha\theta\tilde{\rho}_2 & \Leftrightarrow \alpha \leq \frac{\theta\tilde{\rho}_2}{120\delta_1L^4},\\ 2\tilde{\delta}\delta_2\alpha^2\theta^2\tilde{\rho}_1^2 \leq \frac{1}{10}\alpha\theta\tilde{\rho}_2 & \Leftrightarrow \alpha \leq \frac{\tilde{\rho}_2}{20\tilde{\delta}\delta_2\theta\tilde{\rho}_1^2},\\ 3\tilde{\delta}\delta_2\alpha^2L^2 \leq \frac{1}{10}\alpha\theta\tilde{\rho}_2 & \Leftrightarrow \alpha \leq \frac{\theta\tilde{\rho}_2}{30\tilde{\delta}\delta_2L^2}, \end{cases} \tag{106}$$

To upper bound ω_v,

$$\omega_v \leq \mathtt{a}\cdot 3 + \mathtt{a}\cdot 4\tilde{\rho}_2^{-1}\eta^{-1} + \mathtt{a}\cdot -\frac{1}{2}\delta_1 + \mathtt{a}\cdot 2\tilde{\delta}\delta_2 \tag{107}$$

It holds that $\omega_v \leq -\mathtt{a}\frac{1}{12}\delta_1 \leq -2\mathtt{a} < 0$, with the following step size condition:

$$\begin{cases} 3 \leq \frac{1}{8}\delta_1 & \Leftrightarrow \delta_1 \geq 24,\\ 4\tilde{\rho}_2^{-1}\eta^{-1} \leq \frac{1}{6}\delta_1 & \Leftrightarrow \delta_1 \geq 24\tilde{\rho}_2^{-1}\eta^{-1},\\ 2\tilde{\delta}\delta_2 \leq \frac{1}{6}\delta_1 & \Leftrightarrow \delta_1 \geq 12\tilde{\delta}\delta_2, \end{cases} \tag{108}$$

To upper bound $\omega_{\hat{x}}$,

$$\omega_{\hat{x}} \leq \theta^2 M^2 \frac{3}{n}\alpha + \mathtt{a} \cdot \left(5\alpha\theta^2(\tilde{\rho}_1 + M)^2\right) \tag{109}$$
$$+ \mathtt{a} \cdot \left(\eta + 8\alpha\eta\tilde{\rho}_1^2\tilde{\rho}_2^{-1} + 18\alpha^2\eta^{-1}\tilde{\rho}_2^{-1}\theta^2 M^2 L^2\right) \tag{110}$$
$$+ \mathtt{a} \cdot \delta_1 \left(3\alpha^2\theta^2(\tilde{\rho}_1 + M)^2 + \frac{\alpha\eta^2}{2}\tilde{\rho}_1^2 + 3\alpha^2\theta^2 M^2 L^2\right) \tag{111}$$
$$+ \mathtt{a} \cdot \delta_2 \left(2\tilde{\delta}\alpha^2\theta^2(\tilde{\rho}_1 + M)^2 - \frac{3\delta}{4}\right), \tag{112}$$

Utilizing $\alpha \leq \min\{1, \eta, \frac{1}{L}\}$ and (45), it holds that $\omega_{\hat{x}} \leq -\frac{\delta}{16}\mathtt{a} < 0$, with the following step size condition:

$$\begin{cases} 4\tilde{\delta}\alpha^2\theta^2\tilde{\rho}_1^2 \leq \frac{\delta}{16} & \Leftrightarrow \alpha \leq \frac{\sqrt{\delta/\tilde{\delta}}}{8\theta\tilde{\rho}_1}, \\ 4\tilde{\delta}\alpha^2\theta^2 M^2 \leq \frac{\delta}{16} & \Leftrightarrow \alpha \leq \frac{\sqrt{\delta/\tilde{\delta}}}{8\theta M}, \\ \theta^2 M^2 \frac{3}{n}\alpha \leq \frac{\delta}{16}\delta_2\mathtt{a} & \Leftarrow \alpha \leq \frac{\delta_2\mathtt{a}\delta n}{48\theta^2 M^2}, \\ 10\mathtt{a}\alpha\theta^2\tilde{\rho}_1^2 \leq \frac{\delta}{32}\delta_2\mathtt{a} & \Leftrightarrow \alpha \leq \frac{\delta\delta_2}{320\theta^2\tilde{\rho}_1^2}, \\ 10\mathtt{a}\alpha\theta^2 M^2 \leq \frac{\delta}{32}\delta_2\mathtt{a} & \Leftrightarrow \alpha \leq \frac{\delta\delta_2}{320\theta^2 M^2}, \\ \mathtt{a}\eta \leq \frac{\delta}{16}\delta_2\mathtt{a} & \Leftrightarrow \delta_2 \geq \frac{16\eta}{\delta}, \\ \mathtt{a}(8\alpha\eta\tilde{\rho}_1^2\tilde{\rho}_2^{-1}) \leq \frac{\delta}{16}\delta_2\mathtt{a} & \Leftrightarrow \alpha \leq \frac{\delta\delta_2}{128(\eta\tilde{\rho}_1^2\tilde{\rho}_2^{-1})} \\ \mathtt{a}(18\alpha^2\tilde{\rho}_2^{-1}\theta^2 M^2 L^2) \leq \frac{\delta}{8}\delta_2\mathtt{a} & \Leftrightarrow \alpha \leq \frac{\sqrt{\delta\delta_2\tilde{\rho}_2}}{12ML\theta}, \\ 6\mathtt{a}\delta_1\alpha^2\theta^2\tilde{\rho}_1^2 \leq \frac{\delta}{32}\delta_2\mathtt{a} & \Leftrightarrow \alpha \leq \frac{\sqrt{\delta\delta_2/192\delta_1}}{\theta\tilde{\rho}_1}, \\ 6\mathtt{a}\delta_1\alpha^2\theta^2 M^2 \leq \frac{\delta}{32}\delta_2\mathtt{a} & \Leftrightarrow \alpha \leq \frac{\sqrt{\delta\delta_2/192\delta_1}}{\theta M}, \\ \mathtt{a}\delta_1\frac{\alpha\eta^2}{2}\tilde{\rho}_1^2 \leq \frac{\delta}{32}\delta_2\mathtt{a} & \Leftrightarrow \alpha \leq \frac{\delta\delta_2}{16\delta_1\eta^2\tilde{\rho}_1^2}, \\ \mathtt{a}\delta_1 3\alpha^2\theta^2 M^2 L^2 \leq \frac{\delta}{32}\delta_2\mathtt{a} & \Leftrightarrow \alpha \leq \frac{\sqrt{\delta\delta_2/96\delta_1}}{\theta ML}, \end{cases} \tag{113}$$

To upper bound ω_f,

$$\omega_f \leq -\frac{1}{4}\alpha + 18\mathtt{a}\alpha^2\eta^{-1}\tilde{\rho}_2^{-1}nL^2 + 6\mathtt{a}\delta_1\alpha^2 nL^2 + 2\mathtt{a}\tilde{\delta}\delta_2\alpha^2 \tag{114}$$

It holds that $\omega_f \leq -\frac{1}{16}\alpha$, with the following step size condition:

$$\begin{cases} 18\mathtt{a}\alpha^2\eta^{-1}\tilde{\rho}_2^{-1}nL^2 \leq \frac{1}{16}\alpha & \Leftrightarrow \alpha \leq \frac{\eta\tilde{\rho}_2}{288nL^2\mathtt{a}}, \\ 6\mathtt{a}\delta_1\alpha^2 nL^2 \leq \frac{1}{16}\alpha & \Leftrightarrow \alpha \leq \frac{1}{96\delta_1 nL^2\mathtt{a}}, \\ 2\mathtt{a}\tilde{\delta}\delta_2\alpha^2 \leq \frac{1}{16}\alpha & \Leftrightarrow \alpha \leq \frac{1}{32\tilde{\delta}\delta_2\mathtt{a}}, \end{cases} \tag{115}$$

Without loss of general, assume $L \geq 1$. Based on the above conditions, the upper and lower bounds of the parameters and step size can be obtained. $\eta \geq 1, \theta \geq \theta_{lb}, \alpha \leq \alpha_{ub}$, where

$$\begin{aligned} \delta_1 &= \max\{24, 24\tilde{\rho}_2^{-1}\eta^{-1}, 12\delta_2\tilde{\delta}\}, \quad \delta_2 = \max\{\frac{16\eta}{\delta}, 1\}, \\ \theta_{lb} &= \max\{\frac{4L^2}{n\tilde{\rho}_2\mathtt{a}}, 2\tilde{\rho}_2^{-1}(1 + 2L + 8\eta\tilde{\rho}_1^2\tilde{\rho}_2^{-1} + \delta_1(\frac{3}{2} + 3L^2 + \eta\tilde{\rho}_1 + \frac{\tilde{\rho}_1^2}{2}))\}, \\ \alpha_{ub} &= \frac{\delta\delta_2}{320\theta^2} \cdot \min\{\frac{1}{M^2}, \frac{n\mathtt{a}}{M^2}, \frac{1}{\tilde{\rho}_1^2}, \frac{1}{n\mathtt{a}}\}. \end{aligned} \tag{116}$$

□

Appendix H
Auxiliary Lemmas

Lemma H.1. *Under Assumption IV.1,*

$$\begin{aligned} &\mathbb{E}\left[\left\|\nabla\mathbf{f}((\mathbf{1}_n \otimes \mathbf{I}_d)\bar{\mathbf{X}}^{t+1}) - \nabla\mathbf{f}((\mathbf{1}_n \otimes \mathbf{I}_d)\bar{\mathbf{X}}^t)\right\|^2\right] \\ &\leq 3\alpha^2 nL^2\mathbb{E}\left[\left\|\nabla\mathbf{f}(\bar{\mathbf{X}}^t)\right\|^2\right] + 3\alpha^2 L^4\|\mathbf{X}^t\|^2_{\widetilde{\mathbf{K}}} + 3\theta^2 M^2 L^2\alpha^2\mathbb{E}\left[\|\hat{\mathbf{X}}^t - \mathbf{X}^t\|^2\right]. \end{aligned} \tag{117}$$

Proof of Lemma H.1. By the Lipschitz gradient assumption on each local objective function f_i,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t+1})-\nabla\mathbf{f}((\mathbf{1}_n\otimes\mathbf{I}_d)\bar{\mathbf{X}}^{t})\right\|^2\right] &(118)\\
\le& nL^2\mathbb{E}\left[\|\bar{\mathbf{X}}^{t+1}-\bar{\mathbf{X}}^{t}\|^2\right] &(119)\\
\le& 3\alpha^2nL^2\mathbb{E}\left[\left\|\nabla\mathbf{f}(\bar{\mathbf{X}}^{t})\right\|^2+\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(\mathbf{X}^t)-\nabla f_i(\bar{\mathbf{X}}^{t})\right\|^2\right]\\
&+3L^2(1-\beta)^2\mathbb{E}\left[\|\hat{\mathbf{X}}^t-\mathbf{X}^t\|^2\right] &(120)\\
\le& 3\alpha^2nL^2\mathbb{E}\left[\left\|\nabla\mathbf{f}(\bar{\mathbf{X}}^{t})\right\|^2\right]+3\alpha^2L^4\|\mathbf{X}^t\|^2_{\tilde{\mathbf{K}}}+3\theta^2M^2L^2\alpha^2\mathbb{E}\left[\|\hat{\mathbf{X}}^t-\mathbf{X}^t\|^2\right]. &(121)
\end{aligned}$$

□