Communication Efficient Distributed Learning with Censored, Quantized, and Generalized Group ADMM

Chaouki Ben Issaid  
Centre for Wireless Communications  
University of Oulu, Finland  
CHAOUKI.BENISSAID@OULU.FI

Anis Elgabli  
Centre for Wireless Communications  
University of Oulu, Finland  
ANIS.ELGABLI@OULU.FI

Jihong Park  
School of Information Technology  
Deakin University, Australia  
JIHONG.PARK@DEAKIN.EDU.AU

Mehdi Bennis  
Centre for Wireless Communications  
University of Oulu, Finland  
MEHDI.BENNIS@OULU.FI

Abstract

In this paper, we propose a communication-efficiently decentralized machine learning framework that solves a consensus optimization problem defined over a network of interconnected workers. The proposed algorithm, Censored-and-Quantized Generalized GADMM (CQ-GGADMM), leverages the novel worker grouping and decentralized learning ideas of Group Alternating Direction Method of Multipliers (GADMM), and pushes the frontier in communication efficiency by extending its applicability to a generalized network topologies, while incorporating link censoring for negligible updates after quantization. We theoretically prove that CQ-GGADMM achieves the linear convergence rate when the local objective functions are strongly convex under some mild assumptions. Numerical simulations corroborate that CQ-GGADMM exhibits higher communication efficiency in terms of the number of communication rounds and transmit energy consumption without compromising the accuracy and convergence speed, compared to the benchmark schemes based on decentralized ADMM without censoring, quantization, and/or the worker grouping method of GADMM.

Keywords: communication efficiency, decentralized machine learning, stochastic quantization, decentralized optimization, Alternating Direction Method of Multipliers.

1. Introduction

Machine learning is central to emerging mission-critical applications such as autonomous driving, remote surgery, and the fifth-generation (5G) communication systems and beyond (Park et al. 2019a University of Oulu). These applications commonly require extremely low latency and high reliability while accurately reacting to local environmental dynamics (Park et al. 2020b). To this end, training their machine learning models needs the sheer amount of fresh training data samples that are generated by and dispersed across edge devices (e.g., phones, cars, access points, etc.), hereafter referred to as workers. Collecting these
raw data may not only violate the data privacy, but also incur significant communication overhead under limited bandwidth. This calls for developing communication-efficient and privacy-preserving distributed learning frameworks (Park et al., 2020a; Chen et al., 2019; Singh et al. 2019a). Federated learning is one representative method that ensures learning through periodically exchanging model parameters across workers rather than sending private data samples (McMahan et al., 2017; Kairouz et al., 2019; Park et al., 2019b). Nevertheless, federated learning postulates a parameter server collecting and distributing model parameters, which is not always accessible from faraway workers and is vulnerable to a single point of failure (Kim et al., 2020).

Spurred by this motivation, by generalizing and extending the Group Alternating Direction Method of Multipliers (GADMM, see Fig. 1(a)) and the Quantized GADMM (Q-GADMM) in our prior work (Elgabli et al., 2020c,b), in this article we propose a novel decentralized learning framework, coined Censored-and-Quantized Generalized Group ADMM (CQ-GGADMM, see Fig. 1(c)), which exchanges model parameters in a communication-efficient way without any central entity. Following the same idea of GADMM, workers in CQ-GGADMM are divided into head and tail groups in which the workers in the same group update their models in parallel, whereas the workers in different groups update their models in an alternating way. In essence, CQ-GGADMM exploits three key principles to improve the communication efficiency. First, to reduce the number of communication rounds, it applies a second-order method, i.e., GADMM, which achieves a faster convergence compared to first-order methods such as the decentralized (stochastic) gradient descent (McMahan et al., 2017). Second, to reduce the number of communication links per round, CQ-GGADMM exploits a censoring approach that allows to exchange model parameters only when the updated model is sufficiently changed from the previous model, i.e., skipping small model updates (Sun et al., 2019). Lastly, to reduce the communication payload size per each link, CQ-GGADMM applies a heterogeneous stochastic quantization scheme that decreases the number of bits to represent each model parameter (Elgabli et al., 2020b). These three principles are integrated giving rise to a generalized version of GADMM (GGADMM, see Fig. 1(b)) wherein each worker communicates only with its neighboring workers. Note that in the original GADMM, every worker needs to connect with two neighbors under a chain topology.

Figure 1: A schematic illustration of (a) group ADMM (GADMM) in (Elgabli et al., 2020c), the baseline algorithm under a chain topology, compared to our proposed (b) generalized GADMM (GGADMM) under a generic bipartite graph topology, and (c) censored-and-quantized GGADMM (CQ-GGADMM) that additionally applies link censoring for negligible updates after quantization.
network topology (Elgabli et al., 2020c). By contrast, in CQ-GGADMM, each worker can connect with an arbitrary number of neighbors, as long as the network topology graph is bipartite and connected.

Although the aforementioned principles have been separately studied in preceding works (Elgabli et al., 2020c; Sun et al., 2019; Elgabli et al., 2020b), integrating all of them for maximizing the communication efficiency while guaranteeing fast convergence remains a non-trivial problem. Indeed, first the algorithm convergence rate depends highly on the network topology. Second, both censoring and quantization steps incur model update errors that may propagate over communication rounds due to the lack of central entity. To resolve this problem, we carefully determine the non-increasing target censoring threshold and quantization step size, such that the model updates are more finely tuned as time elapses until convergence. We thereby prove the linear convergence rate of CQ-GGADMM, and show its effectiveness by simulations, in terms of convergence speed, total communication cost, and transmission energy consumption.

2. Related Works and Contributions

Towards improving the communication efficiency of distributed learning, prior works have studied various techniques under centralized and decentralized network architectures, i.e., with and without a parameter server aggregating local model updates, as elaborated next.

**Fast Convergence.** The total communication cost until completing a distributed learning operation can be reduced by accelerating the convergence speed. To this end, departing from the conventional first-order methods such as distributed gradient descent (Boyd et al., 2011), second-order methods are applied under centralized (Konečný et al., 2016; Liu et al., 2019b; Elgabli et al., 2020d) and decentralized architectures (Elgabli et al., 2020c). Furthermore, momentum based training acceleration is utilized under centralized (Yu et al., 2019; Gitman et al., 2019) and decentralized settings (Gao and Huang, 2020).

**Link Sparsification.** In large-scale distributed learning, a large portion of total communication links is often redundant (Mishchenko et al., 2020). In this respect, for each communication round, sparsifying the number of communication links can reduce the communication cost without compromising the accuracy. To this end, link censoring for negligible model updates is applied under centralized (Chen et al., 2018; Sun et al., 2019) and decentralized network topologies (Singh et al., 2019b; Elgabli et al., 2020c).

**Payload Size Reduction.** To reduce the communication payload size per link, model updates are quantized under centralized (Bernstein et al., 2018; Suresh et al., 2017; Sun et al., 2019; Vogels et al., 2019; Alistarh et al., 2017; Horváth et al., 2019) and decentralized network topologies (Sriranga et al., 2019; Zhu et al., 2016; Koloskova et al., 2019; Gao and Huang, 2020; Elgabli et al., 2020b). Alternatively, the entries of model updates can be partially dropped as shown under centralized (Wangni et al., 2018) and decentralized architectures (Stich et al., 2018; Elgabli et al., 2020a). Furthermore, under centralized settings, model parameters can be compressed at the parameter server, with additional training operations, i.e., knowledge distillation (KD) (Hinton et al., 2014) or while training and running KD simultaneously, i.e., federated distillation (Jeong et al., 2018; Ahn et al., 2020; Oh et al., 2020).
Among the aforementioned communication-efficient design principles, this work is closely related to GADMM (Elgabli et al., 2020c), an ADMM-based second-order decentralized learning with neighbor-based communications, which has been extended in various directions. In (Elgabli et al., 2020c), a dynamic version of GADMM (D-GADMM) is considered for coping with a time-varying (chain) network topology. In (Elgabli et al., 2020b), a stochastic quantization is applied for reducing the communication payload size. In (Elgabli et al., 2020a), the payload size is reduced by skipping partial neural network layers at a given interval. All of these works are based on a chain network topology. By contrast, a generic bipartite and connected network topology graph is considered in CQ-GGADMM while additionally incorporating link censoring and payload quantization methods.

**Contributions.** The major contributions of this work are summarized as follows.

- We have proposed CQ-GGADMM, a second-order decentralized learning framework utilizing censoring, quantization, and GADMM for any bipartite and connected network topology graph (Algorithm 2 in Sec. 5).
- We have proven that CQ-GGADMM converges to the optimal solution for convex loss functions (Theorem 1 in Sec. 6).
- We have identified the network topology conditions under which CQ-GGADMM achieves a linear convergence rate (Theorem 2 in Sec. 6) when the loss functions are strongly convex.
- Numerical simulations have corroborated that in linear and logistic regression tasks using synthetic and real datasets, CQ-GGADMM achieves the same convergence speed at significantly less number of communication rounds and several orders of magnitude less transmission energy, compared to the decentralized learning benchmark schemes without censoring and quantization.

**Notations.** Scalars are denoted by non-boldface characters, while vectors and matrices are boldfaced. Throughout this paper, we use the following notations: $\| \cdot \|$, $\| \cdot \|_F$ denote the Euclidean norm of a vector and the Frobenius norm of a matrix, respectively, $\langle \cdot , \cdot \rangle$ is the inner product of two matrices while $(\cdot)^T$ stands for the transpose of a matrix. The notation $| \cdot |$ represents the cardinality of a set, $\nabla f$ stands for the gradient of the function $f$, and $\mathbb{E}[\cdot]$ denotes the expected value.

**Organization.** The remainder of this paper is organized as follows. In section III, we describe the generalized version of GADMM (GGADMM) for a bipartite and connected network topology graph, and formulate the decentralized learning problem. Then, we extend GGADMM to quantized GGADMM (C-GGADMM) by adding a censoring method in Section IV, while Section V further extends C-GGADMM to quantized C-GGADMM (CQ-GGADMM) by applying a stochastic quantization method. In Section VI, we prove the convergence of CQ-GGADMM theoretically, and identify its linear convergence achieving conditions. Finally, Section VII validates the performance of CQ-GGADMM by simulations, followed by concluding remarks in Section VIII. The details of the proofs of our main results are deferred to the Appendices.
3. Problem Formulation

We consider a connected network wherein a set $\mathcal{V}$ of $N$ workers aim to reach a consensus around a solution of a global optimization problem. The problem is solved using only local data and information available for each worker. Moreover, communication is constrained to only take place between neighboring workers. The optimization problem is given by

$$
(P1) \quad \Theta^* := \arg \min_{\Theta} \sum_{n=1}^{N} f_n(\Theta),
$$

where $\Theta \in \mathbb{R}^{d \times 1}$ is the global model parameter and $f_n : \mathbb{R}^d \to \mathbb{R}$ is a local function composed of data stored at worker $n$. Problem (P1) appears in many applications of machine learning, especially when the dataset is very large and the training is carried out using different workers. The connections among workers are represented as an undirected communication graph $G$ having the set $E \subseteq \mathcal{V} \times \mathcal{V}$ of edges. The set of neighbors of worker $n$ is defined as $N_n = \{m|(n,m) \in E\}$ whose cardinality is $|N_n| = d_n$. Before describing our approach, we make the following key assumption.

**Assumption 1.** The communication graph $G$ is bipartite and connected.

Under Assumption 1, following the worker grouping of GADMM (Elgabli et al., 2020c), workers are divided into two groups: a head group $\mathcal{H}$, and a tail group $\mathcal{T}$. Each head worker in $\mathcal{H}$ can only communicate with tail workers in $\mathcal{T}$, and vice versa. In this case, the edge set definition can be re-written as $E = \{(n,m)|n \in \mathcal{H}, m \in \mathcal{T}\}$, and the problem (P1) is equivalent to the following problem

$$
(P2) \quad \theta^* := \arg \min_{\{\theta_n\}_{n=1}^{N}} \sum_{n=1}^{N} f_n(\theta_n)
$$

s.t. $\theta_n = \theta_m, \forall (n,m) \in E,$

where $\theta_n$ is the local copy of the common optimization variable $\Theta$ at worker $n$. Note that, under the formulation (P2), the objective function becomes separable across the workers and as a consequence the problem can be solved in a distributed manner. In this case, the Lagrangian of the optimization problem (P2) can be written as

$$
\mathcal{L}_\rho(\theta, \lambda) = \sum_{n=1}^{N} f_n(\theta_n) + \sum_{(n,m) \in E} \langle \lambda_{n,m}, \theta_n - \theta_m \rangle + \frac{\rho}{2} \sum_{(n,m) \in E} \|\theta_n - \theta_m\|^2, \quad (3)
$$

where $\rho > 0$ is a constant penalty parameter and $\lambda_{n,m}$ is the dual variable between neighboring workers $n$ and $m$, $\forall (n,m) \in E$. At iteration $k+1$, the Generalized Group ADMM (GGADMM) algorithm runs as follows.

1. Every head worker, $n \in \mathcal{H}$, updates its primal variable by solving

$$
\theta_n^{k+1} = \arg \min_{\theta_n} f_n(\theta_n) + \sum_{m \in N_n} \langle \lambda_{n,m}^k, \theta_n - \theta_m^k \rangle + \frac{\rho}{2} \sum_{m \in N_n} \|\theta_n - \theta_m^k\|^2, \quad (4)
$$

and sends its updated model to its neighbors.
(2) The primal variables of tail workers, \( m \in T \), are then updated as

\[
\theta^{k+1}_m = \arg\min_{\theta_m} f_m(\theta_m) + \sum_{n \in N_m} \langle \lambda^{k+1}_{n,m}, \theta^{k+1}_n - \theta_m \rangle + \frac{\rho}{2} \sum_{n \in N_m} \|\theta^{k+1}_n - \theta_m\|^2. \tag{5}
\]

(3) The dual variables are updated locally for every worker, after receiving the model updates from its neighbors, in the following way

\[
\lambda^{k+1}_{n,m} = \lambda^{k}_{n,m} + \rho \left( \theta^{k+1}_n - \theta^{k+1}_m \right), \quad \forall (n, m) \in E. \tag{6}
\]

Note that GGADMM is a generalized version of GADMM algorithm proposed in (Elgabli et al., 2020c). In contrast to GADMM which works for a chain topology, GGADMM considers an arbitrary topology. Introducing the definition of the auxiliary variable \( \alpha \)

\[
\alpha_n = \sum_{m \in N_n} \lambda_{n,m}, \quad \forall n \in V, \tag{7}
\]

we can re-write the above algorithm as follows.

(1) The update of the models of head workers is done in parallel by solving

\[
\theta^{k+1}_n = \arg\min_{\theta_n} f_n(\theta_n) + \langle \theta_n, \alpha^k_n - \rho \sum_{m \in N_n} \theta^k_m \rangle + \frac{\rho}{2} d_n \|\theta_n\|^2. \tag{8}
\]

(2) The models of tail workers are updated in parallel using

\[
\theta^{k+1}_m = \arg\min_{\theta_m} f_m(\theta_m) + \langle \theta_m, \alpha^k_m - \rho \sum_{n \in N_m} \theta^{k+1}_n \rangle + \frac{\rho}{2} d_m \|\theta_m\|^2. \tag{9}
\]

(3) Instead of updating \( \lambda_{n,m} \), each worker will update locally the new auxiliary variable \( \alpha_n \) as follows

\[
\alpha^{k+1}_n = \alpha^k_n + \rho \sum_{m \in N_n} (\theta^{k+1}_n - \theta^{k+1}_m), \quad \forall n \in V. \tag{10}
\]

4. Censored Generalized Group ADMM

In this section, we introduce the communication censoring idea in order to make GGADMM more communication-efficient. In fact, at every iteration, some workers having negligible updates can be censored without compromising accuracy. Accordingly, such workers do not communicate their model updates to their neighbors, based on a “censoring” condition, to be detailed later on. The proposed algorithm will be referred to in the sequel as Censored Generalized Group ADMM (C-GGADMM).

Let \( \{\tau^k\} \) be a decreasing and non-negative sequence that represents the censoring threshold sequence. In our work, we consider the choice to be of the form \( \tau^k = \tau_0 \xi^k \) with \( \tau_0 > 0 \) and \( \xi \in (0, 1) \). At iteration \( k + 1 \), each worker \( n \in V \), computes \( \|\tilde{\theta}^k_n - \theta^{k+1}_n\| \) and compare it to the value of the threshold \( \tau^{k+1} \), where \( \tilde{\theta}^k_n \) is a state variable that stores its most recent (up to time \( k \)) primary variable transmission. Note that the variable \( \tilde{\theta}^k_n \) is...
updated locally for each worker \( n \) and is not shared among workers. If \( \|\tilde{\theta}_n^k - \theta_n^{k+1}\| \geq \tau_{k+1} \), the \( n \)th worker transmits \( \theta_n^{k+1} \) to its neighbors and sets \( \tilde{\theta}_n^{k+1} = \theta_n^{k+1} \). Otherwise, it does not transmit and sets \( \tilde{\theta}_n^{k+1} = \tilde{\theta}_n^k \).

For a given iteration \( k \), note that the censoring condition given by \( \|\tilde{\theta}_n^k - \theta_n^{k+1}\| < \tau_0 \), will be violated as \( \xi \to 0 \), and no communication censoring will take place. In this case, C-GGADMM will reduce to GGADMM. For a fixed \( \xi \), when \( \tau_0 \) is small, more workers will likely transmit their models and the effect of censoring will be less. In the special case \( \tau_0 = 0 \), we get back to GGADMM. However, if \( \tau_0 \) is very large, most workers will be censored from communicating their models, which will slow down the convergence of the algorithm. In this case, the operations of C-GGADMM can be described as follows.

(1) Primal variables for head workers are solved using

\[
\theta_n^{k+1} = \underset{\theta_n}{\arg \min} f_n(\theta_n) + \langle \theta_n, \alpha_n^k - \rho \sum_{m \in \mathcal{N}_n} \tilde{\theta}_m^k \rangle + \frac{\rho}{2} d_n \|\theta_n\|^2. \tag{11}
\]

(2) Primal variables update for tail workers is done as follows

\[
\theta_m^{k+1} = \underset{\theta_m}{\arg \min} f_m(\theta_m) + \langle \theta_m, \alpha_m^k - \rho \sum_{n \in \mathcal{N}_m} \tilde{\theta}_n^{k+1} \rangle + \frac{\rho}{2} d_m \|\theta_m\|^2. \tag{12}
\]

(3) Dual variable of each worker is updated locally

\[
\alpha_n^{k+1} = \alpha_n^k + \rho \sum_{m \in \mathcal{N}_n} (\tilde{\theta}_n^{k+1} - \tilde{\theta}_m^{k+1}), \ \forall n \in \mathcal{V}. \tag{13}
\]

The steps of C-GGADMM are summarized in Algorithm 1. We clearly see that the algorithm is fully decentralized since the updates of the primal and dual variables only depend on local and neighboring information. Moreover, the algorithm allows updating the parameters in parallel for the workers in the same group.

5. Censored Quantized Generalized Group ADMM

Compared to GGADMM, C-GGADMM reduces the communication overhead. However, C-GGADMM still needs to receive the full precision information \( \theta \)'s from the neighbors at each worker \( n \) to update the local model. This creates a communication bottleneck, especially when the dimensions \( d \) of the model \( \theta \) is large. We address this issue by using stochastic quantization in which we use the quantized version of the information \( \hat{Q}_m \), \( \forall m \in \mathcal{N}_n \) to update the primal and dual variables at each worker \( n \).

We follow a similar stochastic quantization scheme to the one described in (Elgabli et al., 2020b) where each worker quantizes the difference between its current model and its previously quantized model before transmission \( (\theta_n^k - \hat{Q}_n^{k-1}) \) as \( \theta_n^k - \hat{Q}_n^{k-1} = Q_n(\theta_n^k, \hat{Q}_n^{k-1}) \). The function \( Q_n(\cdot) \) is a stochastic quantization operator that depends on the quantization probability \( p_{n,i}^k \) for each model vector’s dimension \( i \in \{1, 2, \ldots, d\} \), and on \( b_n^k \) bits used for representing each model vector dimension.

The \( i \)th dimensional element \( [\hat{Q}_n^{k-1}]_i \) of the previously quantized model vector is centred at the quantization range \( 2R_n^k \) that is equally divided into \( 2^{b_n^k} - 1 \) quantization levels, yielding
Algorithm 1 Censored Generalized Group ADMM (C-GGADMM)

1: **Input:** $N, p, \tau_0, \xi, f_n(\theta_n)$ for all $n$
2: $\theta_n^0 = 0, \tilde{\theta}_n^0 = 0, \alpha_n^0 = 0$ for all $n$
3: for $k = 0, 1, 2, \cdots, K$ do
4:   **Head worker** $n \in \mathcal{H}$:
5:     computes its primal variable $\theta_n^{k+1}$ via (11) in parallel
6:     if $\|\tilde{\theta}_n^k - \theta_n^{k+1}\| \geq \tau_0 \xi^{k+1}$ then
7:         worker $n$ sends $\theta_n^{k+1}$ to its neighbor workers $\mathcal{N}_n$ and sets $\tilde{\theta}_n^{k+1} = \theta_n^{k+1}$.
8:     else
9:         worker $n$ **does not transmit** and sets $\tilde{\theta}_n^{k+1} = \tilde{\theta}_n^k$.
10: end if
11: **Tail worker** $m \in \mathcal{T}$:
12:     computes its primal variable $\theta_m^{k+1}$ via (12) in parallel
13:     if $\|\tilde{\theta}_m^k - \theta_m^{k+1}\| \geq \tau_0 \xi^{k+1}$ then
14:         worker $m$ sends $\theta_m^{k+1}$ to its neighbor workers $\mathcal{N}_m$ and sets $\tilde{\theta}_m^{k+1} = \theta_m^{k+1}$.
15:     else
16:         worker $m$ **does not transmit** and sets $\tilde{\theta}_m^{k+1} = \tilde{\theta}_m^k$.
17: end if
18: **Every worker updates** the dual variables $\alpha_n^{k+1}$ via (13) locally.
19: end for

the quantization step size $\Delta_n^k = 2R_n^k/(2^{k+1} - 1)$. In this coordinate, the difference between the $i$th dimensional element $[\theta_n^k]_i$ of the current model vector and $[\hat{Q}_n^{k-1}]_i$ is

$$[c_n(\theta_n^k)]_i = \frac{1}{\Delta_n^k} \left( [\theta_n^k]_i - [\hat{Q}_n^{k-1}]_i + R_n^k \right),$$

(14)

where adding $R_n^k$ ensures the non-negativity of the quantized value. Then, $[c_n(\theta_n^k)]_i$ is mapped to

$$[g_n(\theta_n^k)]_i = \begin{cases} [c_n(\theta_n^k)]_i & \text{with probability } p_{n,i}^k \\ [c_n(\theta_n^k)]_i & \text{with probability } 1 - p_{n,i}^k. \end{cases}$$

(15)

where $[\cdot]$ and $\lfloor \cdot \rfloor$ are the ceiling and floor functions, respectively. Next, the probability $p_{n,i}^k$ in (15) is selected such that the expected quantization error $\mathbb{E} [e_{n,i}^k]$ is zero. Therefore, the probability $p_{n,i}^k$ should satisfy

$$p_{n,i}^k \left( [c_n(\theta_n^k)]_i - [c_n(\theta_n^k)]_i \right) + (1 - p_{n,i}^k) \left( [c_n(\theta_n^k)]_i - [c_n(\theta_n^k)]_i \right) = 0.$$  

(16)

Solving (16) for $p_{n,i}^k$, we obtain

$$p_{n,i}^k = \frac{[c_n(\theta_n^k)]_i - [c_n(\theta_n^k)]_i}{[c_n(\theta_n^k)]_i - [c_n(\theta_n^k)]_i}.$$  

(17)

The choice of $p_{n,i}^k$ in (17) ensures that the quantization in (15) is unbiased and the quantization error variance $\mathbb{E} \left[ \left( e_{n,i}^k \right)^2 \right]$ is less than $(\Delta_n^k)^2$. This implies that $\mathbb{E} \left[ \|e_n^k\|^2 \right] \leq d(\Delta_n^k)^2$. 

8
In addition to the above condition, the convergence of CQ-GGADMM requires non-increasing quantization step sizes over iterations, i.e. \( \Delta^k_n \leq \omega \Delta^{k-1}_n \) for all \( k \) where \( \omega \in (0, 1) \). To satisfy this condition, the parameter \( b^k_n \) is chosen as
\[
b^k_n \geq \left\lceil \log_2 \left( 1 + \frac{2^{b^k_n - 1} - 1}{\omega R^k_n} \right) \right\rceil.
\] (18)
Under this condition, we get that \( \Delta^k_n \leq \omega \Delta^0_n \). Given \( p^k_n, i \) in (17) and \( b^k_n \) in (18), the convergence of CQ-GGADMM is provided in Section 6. With the aforementioned stochastic quantization procedure, \( b^k_n, R^k_n, \) and \( q_n(\theta^k_n) \) suffice to represent \( \hat{Q}^k_n \), where
\[
q_n(\theta^k_n) = (q_n(\theta^k_n)_1, q_n(\theta^k_n)_2, \cdots, q_n(\theta^k_n)_d)^T,
\] (19)
which are transmitted to neighbors. After receiving these values, \( \hat{Q}^k_n \) can be reconstructed as follows:
\[
\hat{Q}^k_n = \hat{Q}^{k-1}_n + \Delta^k_n q_n(\theta^k_n) - R^k_n 1. \tag{20}
\]
Consequently, when the full arithmetic precision uses 32 bits, every transmission payload size of CQ-GGADMM is \( b^k_n d + (b^k_R + b_b) \) bits, where \( b_R \leq 32 \) and \( b_b \leq 32 \) are the required bits to represent \( R^k_n \) and \( b^k_n \), respectively. Compared to GGADMM, whose payload size is \( 32d \) bits, CQ-GGADMM can achieve a huge reduction in communication overhead, particularly for large models, i.e. large \( d \).

Now, we are in a position to explain the censored quantized generalized Group ADMM (CQ-GGADMM). Similarly to Section 4, we introduce a censoring condition to reduce the number of workers communicating at a given iteration by allowing the worker to transmit only when the difference between the current and previously transmitted value is sufficiently different. However, we apply the censoring not on the model itself but on its quantized value, i.e. if the worker is not censored, it transmits its quantized model to its neighbors. According to the communication-censoring strategy, we have that \( \hat{\theta}^k_{n+1} = \hat{Q}^k_n \) provided that \( \|\hat{\theta}^k_n - \hat{Q}^{k+1}_n\| \geq \tau_0 c^{k+1} \) and \( \hat{\theta}^k_{n+1} = \hat{\theta}^k_n \), otherwise. The CQ-GGADMM algorithm can be written in this case as

(1) Primal variables for head workers are found using
\[
\theta^{k+1}_n = \arg \min_{\theta_n} f_n(\theta_n) + \langle \theta_n, \alpha^k_n - \rho \sum_{m \in N_n} \hat{\theta}^k_m \rangle + \frac{\rho}{2} d_n \|\theta_n\|^2. \tag{21}
\]
(2) Primal variables update for tail workers is done as follow
\[
\theta^{k+1}_m = \arg \min_{\theta_m} f_m(\theta_m) + \langle \theta_m, \alpha^k_m - \rho \sum_{n \in N_m} \hat{\theta}^{k+1}_n \rangle + \frac{\rho}{2} d_m \|\theta_m\|^2. \tag{22}
\]
(3) Dual variable of each worker is updated locally
\[
\alpha^{k+1}_n = \alpha^k_n + \rho \sum_{m \in N_n} (\hat{\theta}^{k+1}_n - \hat{\theta}^{k+1}_m), \forall n \in V. \tag{23}
\]
Algorithm 2: Censored Quantized Generalized Group ADMM (CQ-GGADMM)

1: Input: \( \tilde{N}, \rho, \tau_0, \xi, f_n(\theta_n) \) for all \( n \)
2: \( \theta_n^0 = 0, \tilde{\theta}_n^0 = 0, \alpha_n^0 = 0 \) for all \( n \)
3: for \( k = 0, 1, 2, \cdots, K \) do
4:   Head worker \( n \in \mathcal{H} \):
5:     computes its primal variable \( \theta_n^{k+1} \) via (21) in parallel
6:     quantizes its primal variable \( \theta_n^{k+1} \) to \( \tilde{\theta}_n^{k+1} \) as described in section 5
7:     if \( ||\tilde{\theta}_n^k - \tilde{\theta}_n^{k+1}|| \geq \tau_0 \) then
8:       worker \( n \) sends \( q_n(\theta_n^{k+1}), R_n^{k+1}, \) and \( b_n^{k+1} \) to its neighboring workers \( \mathcal{N}_n \) and sets \( \tilde{\theta}_n^{k+1} = \tilde{Q}_n^{k+1} \).
9:     else
10:       worker \( n \) does not transmit and sets \( \tilde{\theta}_n^{k+1} = \tilde{\theta}_n^k \).
11: end if
12: Tail worker \( m \in \mathcal{T} \):
13:     computes its primal variable \( \theta_m^{k+1} \) via (22) in parallel
14:     quantizes its primal variable \( \theta_m^{k+1} \) to \( \tilde{\theta}_m^{k+1} \) as described in section 5
15:     if \( ||\tilde{\theta}_m^k - \tilde{\theta}_m^{k+1}|| \geq \tau_0 \) then
16:       worker \( m \) sends \( q_n(\theta_m^{k+1}), R_m^{k+1}, \) and \( b_m^{k+1} \) to its neighboring workers \( \mathcal{N}_m \) and sets \( \tilde{\theta}_m^{k+1} = \tilde{Q}_m^{k+1} \).
17:     else
18:       worker \( m \) does not transmit and sets \( \tilde{\theta}_m^{k+1} = \tilde{\theta}_m^k \).
19: end if
20: Every worker updates the dual variables \( \alpha_n^{k+1} \) via (23) locally.
21: end for

6. Convergence Analysis

In this section, we prove the optimality and convergence of the CQ-GGADMM algorithm. Before stating the main results of the paper, we further make the following assumptions.

Assumption 2. There exists an optimal solution set to (P1) which has at least one finite element.

Assumption 3. The local cost functions \( f_n \) are convex.

Assumption 4. The local cost functions \( f_n \) are strongly convex with parameter \( \mu_n > 0 \), i.e.
\[
||\nabla f_n(x) - \nabla f_n(y)|| \geq \mu_n ||x - y||^2, \forall x, y \in \mathbb{R}^d.
\] (24)

Assumption 5. The local cost functions \( f_n \) have \( L_n \)-Lipschitz continuous gradient \( (L_n > 0) \), i.e.
\[
||\nabla f_n(x) - \nabla f_n(y)|| \leq L_n ||x - y||, \forall x, y \in \mathbb{R}^d.
\] (25)

Assumptions 1-5 are key assumptions that are often used in the context of distributed optimization (Liu et al., 2019b; Konecný et al., 2016; Chen et al., 2018). While only assumptions...
1-3 are needed to prove the convergence of CQ-GGADMM, assumptions 4 and 5 are further required to show the linear convergence rate of CQ-GGADMM. Note that Assumption 2 ensures that the problem (P2) has at least one optimal solution, denoted by $\theta^*$. Under Assumption 4, the function $f$ is strongly convex with parameter $\mu = \min_{1 \leq n \leq N} \mu_n$, and from Assumption 5, we can see that $f$ has $L$-Lipschitz continuous gradient with $L = \min_{1 \leq n \leq N} L_n$.

To proceed with the analysis, we start by writing the optimality conditions as

$$\theta^* = \theta^*_n, \forall (n, m) \in \mathcal{E},$$
$$\nabla f_n(\theta^*_n) + \alpha^*_n = 0, \forall n \in \mathcal{V},$$
where $\theta^*_n$ and $\alpha^*_n$ are the optimal values of the primal and dual variables, respectively. We define the primal residual $r^{k+1}_{n,m}$, and the dual residual $s^{k+1}_n$ as

$$r^{k+1}_{n,m} = \theta^{k+1}_n - \theta^{k+1}_m, \forall (n, m) \in \mathcal{E},$$
$$s^{k+1}_n = \rho \sum_{m \in \mathcal{N}_n} (\hat{\theta}^{k+1}_m - \hat{\theta}^{k+1}_n), \forall n \in \mathcal{H}.$$  

The total error $e^{k+1}_n$ is given by

$$e^{k+1}_n = \theta^{k+1}_n - \hat{\theta}^{k+1}_n, \forall n = 1, \ldots, N.$$  

The total error can be decomposed as the sum of two errors: (i) a random error coming from the quantization process $e^{k+1}_n = \theta^{k+1}_n - \hat{Q}^{k+1}_n$, and (ii) a deterministic one due to the censoring strategy $\ell^{k+1}_n = \hat{Q}^{k+1}_n - \hat{\theta}^{k+1}_n$. According to the communication-censoring strategy, we have that $\hat{\theta}^{k+1}_n = \hat{Q}^{k+1}_n$ if $\|\hat{\theta}^{k-1}_n - \hat{Q}^{k}_n\| \geq \tau^k$ and $\hat{\theta}^{k+1}_n = \hat{\theta}^{k-1}_n$ if $\|\hat{\theta}^{k-1}_n - \hat{Q}^{k}_n\| < \tau^k$. In both cases, we have

$$\|\ell^{k+1}_n\| = \|\hat{Q}^{k}_n - \hat{\theta}^{k+1}_n\| < \tau^k.$$  

Since the sequence $\{\tau^k\}$ is a decreasing non-negative sequence, then we have that $\|\ell^{k+1}_n\| \leq \tau^k$ and $\|\ell^{k+1}_n\| \leq \tau^k$, $\forall n \in \mathcal{V}$. Since the second moment of the quantization error is bounded by

$$\mathbb{E}[\|\epsilon^k\|^2] \leq d(\Delta^k)^2 \leq d(\Delta^0)^2 \omega^{2k},$$
where $\Delta^0 = \max_{1 \leq n \leq N} \Delta^0_n$, then, the total error can be upper bounded, using (42), by

$$\mathbb{E}[\|\epsilon^k\|^2] \leq 2(\|\ell^k\|^2 + \mathbb{E}[\|\epsilon^k\|^2]) \leq 2 \left(\tau_0^2 \epsilon^{2k} + d(\Delta^0)^2 \omega^{2k}\right) \leq 4C_0^2 \psi^{2k},$$
where $C_0 = \max\{\tau_0, \sqrt{d(\Delta^0)}\}$, and $\psi = \max\{\xi, \omega\} \in (0,1)$.

Note that, at a given iteration $k$, if we have $\sqrt{d(\Delta^0)} \omega^k > \tau_0 \xi^k$, then the quantization error dominates the censoring error; otherwise the censoring error will have more impact than the quantization one. Since both sequences $\{\xi^k\}$ and $\{\omega^k\}$ are decreasing, then the sequence $\{\psi^k\}$ is also decreasing. To prove the convergence of the proposed algorithm, we start by stating and proving the first lemma where we derive upper and lower bounds on the expected value of the optimality gap.
Lemma 1 Under assumptions 1-3, we have the following bounds on the expected value of the optimality gap

(i) Upper bound

\[
\sum_{n=1}^{N} \mathbb{E} \left[ f_n(\theta^{k+1}) - f_n(\theta^*) \right] \leq - \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda^{k+1}_{n,m}, r^{k+1}_{n,m} \rangle \right] + \sum_{n \in \mathcal{H}} \mathbb{E} \left[ \langle s^{k+1}_n, \theta^*_n - \theta^{k+1}_n \rangle \right] + \rho \sum_{n=1}^{N} d_n \mathbb{E} \left[ \langle \epsilon^{k+1}_n, \theta^*_n - \theta^{k+1}_n \rangle \right],
\]

(34)

(ii) Lower bound

\[
\sum_{n=1}^{N} \mathbb{E} \left[ f_n(\theta^{k+1}) - f_n(\theta^*) \right] \geq - \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda^*_n, r^{k+1}_{n,m} \rangle \right].
\]

(35)

Proof The details of the proof are deferred to Appendix B.

Next, we present the first theorem that states the asymptotic convergence of the proposed algorithm. In this theorem, we prove the convergence to zero in the mean square sense of both the primal and dual residuals as well as the convergence to zero in the mean sense of the optimality gap.

Theorem 2 Suppose assumptions 1-3 hold, then the CQ-GGADMM iterates lead to

(i) the convergence of the primal residual to zero in the mean square sense as \( k \to \infty \), i.e.

\[
\lim_{k \to \infty} \mathbb{E} \left[ \| r^{k}_{n,m} \|^2 \right] = 0, \ \forall (n,m) \in \mathcal{E},
\]

(36)

(ii) the convergence of the dual residual to zero in the mean square sense as \( k \to \infty \), i.e.

\[
\lim_{k \to \infty} \mathbb{E} \left[ \| s^{k}_n \|^2 \right] = 0, \ \forall n \in \mathcal{H},
\]

(37)

(iii) the convergence of the optimality gap to zero in the mean sense as \( k \to \infty \), i.e.

\[
\lim_{k \to \infty} \sum_{n=1}^{N} \mathbb{E} \left[ f_n(\theta^{k}_n) - f_n(\theta^*_n) \right] = 0.
\]

(38)

Proof The proof can be found in Appendix C.

The linear convergence of the CQ-GGADMM algorithm is presented next.
Theorem 3 Suppose that assumptions 1, 2, 4 and 5 hold and the dual variable $\alpha$ is initialized such that $\alpha^0$ lies in the column space of the signed incidence matrix $M_-$. Then, provided that $0 < \rho < \bar{\rho}$ where $\bar{\rho}$ is defined in (150), the sequence of iterates of CQ-GGADMM converges at a linear rate, i.e.

$$
\|\theta^{k+1} - \theta^*\|_F^2 \leq \left( \frac{1 + \delta_2}{2} \right)^{k+1} \left( \|\theta^0 - \theta^*\|_F^2 + C_1 \right),
$$

where $\delta_2$ and $C_1$ are defined in (154) and (157), respectively.

Proof The detailed proof is provided in Appendix D. In the proof, we require an extra initialization condition that $\alpha^0$ lies in the column space of $M_-$. This can be simply satisfied by taking $\alpha^0 = 0$. By doing so, we ensure that $\alpha^k$ will always stay in the column space of $M_-$ and therefore, we can write $\alpha^k = M_- \beta^k$, $\forall k \geq 0$. The convergence rate, derived in the proof, depends on the network topology through the values of $\sigma_{\text{max}}(C)$, $\sigma_{\text{max}}(M_-)$ and $\tilde{\sigma}_{\text{min}}(M_-)$, the properties of the local objective functions; more precisely the values of $\mu$ and $L$, the penalty parameter $\rho$ but also on the threshold parameter $\xi$ as well as the parameter $\omega$ used to construct the quantization step sizes.

7. Numerical Results

To validate our theoretical results, we numerically evaluate the performance of CQ-GGADMM compared with GGADMM, C-GGADMM, and C-ADMM (Liu et al., 2019b). Note that C-ADMM performs censoring on top of the Jacobian and decentralized version of the standard ADMM. Note also that, in Jacobian ADMM, all workers update their models in parallel. For the tuning parameters, we choose the values leading to the best performance of all algorithms.

Model and Datasets. All simulations are conducted using synthetic and real datasets. For the synthetic data, we used the datasets that were generated in Chen et al. (2018). We consider two decentralized consensus optimization problems: (i) linear regression, and (ii) logistic regression. Note that the local cost functions are smooth in both cases. The details about the datasets used in our experiments are summarized in Table 1. For each dataset, the number of samples are uniformly distributed across the $N$ workers. The main comparison is based on a network graph that is neither ultra dense nor very sparse. We study the effect of the network graph density later in Section 7.3.

Graph Generation. Similarly to (Shi et al., 2014), we generate randomly a network consisting of $N$ workers with a connectivity ratio $p$. The ratio $p$ is defined as the actual number of edges divided by the number of edges for a fully connected graph, i.e. $N \times (N-1)/2$. Such a random graph is created with $Np \times (N-1)/2$ edges that are uniformly randomly chosen, while ensuring that the generated network is connected. Smaller values of $p$ leads to a sparser graph, while the generated graph becomes denser as $p$ approaches 1.

Communication Energy. We assume that the total system bandwidth 2MHz is equally divided across workers. Therefore, the available bandwidth to the $n$-th worker ($B_n$) at every communication round when utilizing GGADMM is $(4/N)$MHz since only half of the workers are transmitting at each communication round. On the other hand, the available bandwidth
| Dataset                  | Task                  | Data Type | Model Size (d) | Number of Instances |
|-------------------------|-----------------------|-----------|----------------|---------------------|
| synth-linear (Chen et al., 2018) | linear regression | synthetic | 50             | 1200                |
| Body Fat (Dua and Graff, 2017) | linear regression | real      | 14             | 252                 |
| synth-logistic (Chen et al., 2018) | logistic regression | synthetic | 50             | 1200                |
| Derm (Dua and Graff, 2017)    | logistic regression | real      | 34             | 358                 |

Table 1: List of datasets used in the numerical experiments.

to each worker when using C-ADMM is \((2/N)\text{MHz}\). The power spectral density \((N_0)\) is \(10^{-6}\text{W/Hz}\), and each upload/download transmission time \((\tau)\) is 1ms. We assume a free space model, and each worker needs to transmit at a power level that allows transmitting the model vector in one communication round (the rate is bottlenecked by the worst link). For example, using C-ADMM, each worker needs to find the transmission power that achieves the transmission rate \(R = (32d/1\text{ms})\text{ bits/sec}\). Therefore, using Shannon capacity, the corresponding transmission power can be calculated as \(P = \tau D^2 N_0 B_n (2^{R/B_n} - 1)\), and the consumed energy will be \(E = P\tau\).

**Hardware and Software.** To run the experiments, we implemented all algorithms using MATLAB. All methods were evaluated on a MacBook Air computer with 1.8 GHz Intel Core i5 CPU, and a 8 GB 1,600 MHz DDR3 RAM.

### 7.1 Linear Regression

In this case, the local cost function at worker \(n\) is explicitly given by

\[
f_n(\theta) = \frac{1}{2} \| X_n \theta - y_n \|^2,
\]

where \(X_n \in \mathbb{R}^{s \times d}\) and \(y_n \in \mathbb{R}^{s \times 1}\) are private for each worker \(n \in \mathcal{V}\) where \(s\) represents the size of the data at each worker.

Figs. 2-(a) and 3-(a) corroborate that both C-GGADMM and CQ-GGADMM achieve the same convergence speed as GGADMM and significantly outperform C-ADMM, thanks to the the alternation update, censoring, and stochastic quantization. Note that though, C-ADMM allows workers to update their models in parallel, it requires significantly higher number of iterations. Figs. 2-(b) and 3-(b) show that C-GGADMM achieves \(10^{-4}\) objective error with the minimum number of communication rounds outperforming all other algorithms. We also note that introducing quantization on top of censoring has increased the number of communication rounds. However, in terms of the total number of transmitted bits and consumed energy, CQ-GGADMM outperforms all algorithms.

### 7.2 Logistic Regression

In this section, we consider the binary logistic regression problem. We assume that worker \(n\) owns a data matrix \(X_n = (x_{n,1}, \ldots, x_{n,s})^T \in \mathbb{R}^{s \times d}\) along with the corresponding labels \(y_n = (y_{n,1}, \ldots, y_{n,s}) \in \{-1,1\}^s\). The local cost function for worker \(n\) is then given by

\[
f_n(\theta) = \frac{1}{s} \sum_{j=1}^{s} \log \left(1 + \exp \left(-y_{n,j} x_{n,j}^T \theta \right)\right) + \frac{\mu_0}{2} \| \theta \|^2,
\]
where $\mu_0$ is the regularization parameter.

As observed from Figs. 4-(a) and 5-(a), C-GADMM requires more iterations compared to GADMM to achieve the same loss which leads to either no saving in the number of communication rounds (see Fig. 4-(b)) or a small saving in the number of communication rounds (see Fig. 5-(b)). It also appears that the update of each individual worker when not quantizing is important at each iteration and censoring hurts the convergence speed. However, interestingly, when introducing stochastic quantization and performing censoring on top of the quantized models, we overcome this issue, and we show significant savings in the number of communication rounds and the communication overhead per iteration.

To conclude, the combination of quantization and censoring always leads to the most savings in communication overhead for both linear and logistic regression tasks as depicted Figs. 2, and 3, and Figs. 4 and 5, respectively.

### 7.3 Impact of the Network Graph Density

To study how the network graph density (the node degree) affects the performance of the proposed approach, we conduct an experiment using linear regression on real dataset
under different network graphs. In particular, we consider two graphs with different density as shown in Fig. 6 (b) and (c). The first graph, denoted by Graph 1, is a sparse graph (generated with $p = 0.2$), where each worker has a few links (communicating with low number of neighbouring workers). For example, worker 12 communicates only with one neighbour (worker 8). On the other hand, the dense graph (Graph 2) is generated with a connectivity ratio $p = 0.4$ where each worker has at least three links (three neighbours). We clearly see from Fig. 6-(a) that a denser graph leads to faster convergence for all algorithms since each worker uses more information per iteration. However, the ratio in the performance gap in terms of the number of communication rounds remains the same, i.e. C-GGADMM achieves the minimum number of communication rounds followed by CQ-GGADMM which confirms the findings in Fig.3-(b) for more choices of network graph density.

8. Conclusions

In this paper, we have proposed a communication-efficiently decentralized ML algorithm that extends GADMM (Elgabli et al., 2020c) and Q-GADMM (Elgabli et al. 2020b) to arbitrary topologies. Moreover, the proposed algorithm leverages censoring (sparsification) to minimize
Figure 6: *Effect of the graph density on the performance of the algorithms*: loss w.r.t. number of communication rounds (left), Graph 1: Sparse graph (right top); Graph 2: dense graph (right bottom). The number of workers is 18, and the task is linear regression on real dataset.

the number of communication rounds for each worker. Utilizing a decreasing sequence of censoring threshold, stochastic quantization, and adjusting the quantization range at every iteration such that a linear convergence rate is achieved are key features that make CQ-GGADMM robust to errors while ensuring its convergence guarantees. Numerical results in convex linear and logistic regression tasks corroborate the advantages of CQ-GGADMM over GGADMM, and C-ADMM (Liu et al., 2019b).
Appendix A. Basic identities and inequalities

For any two vectors \( x, y \in \mathbb{R}^d \), we have

\[
\|x + y\|^2 \leq 2 (\|x\|^2 + \|y\|^2), \quad \forall x, y \in \mathbb{R}^d,
\]

(42)

\[
2\langle x, y \rangle \leq \frac{1}{\eta} \|x\|^2 + \eta \|y\|^2, \quad \forall x, y \in \mathbb{R}^d, \quad \eta > 0,
\]

(43)

\[
-2\langle x, y \rangle = \|x - y\|^2 - \|x\|^2 - \|y\|^2, \quad \forall x, y \in \mathbb{R}^d.
\]

(44)

\[
\mathbb{E} [\langle x, y \rangle] \leq \left( \mathbb{E} [\|x\|^2] \right)^{\frac{1}{2}} \left( \mathbb{E} [\|y\|^2] \right)^{\frac{1}{2}}, \quad \forall x, y \in \mathbb{R}^d \quad \text{(Cauchy-Schwarz)}.
\]

(45)

For any two matrices \( A \) and \( B \), we have

\[
2\langle A, B \rangle \leq \eta \|A\|^2_F + \frac{1}{\eta} \|B\|^2_F, \quad \forall \eta > 0,
\]

(46)

\[
\|AB\|_F \leq \sigma_{\max}(A) \|B\|_F,
\]

(47)

\[
\|A + B\|^2_F \leq \eta \|A\|^2_F + \frac{\eta}{\eta - 1} \|B\|^2_F, \quad \forall \eta > 1,
\]

(48)

where \( \sigma_{\max}(A) \) denotes the maximum singular value of the matrix \( A \).

Appendix B. Proof of Lemma 1

We start by proving the statement \((i)\). To this end, using (21) the update of the head workers can be written as

\[
\nabla f_n(\theta^{k+1}_n) + \alpha^k_n + \rho \sum_{m \in \mathcal{N}_n} \hat{\theta}^k_m + \rho d_n \theta^{k+1}_n = 0.
\]

(49)

Using the update of \( \alpha^k_n \) as in Eq. (23) and the definition of the dual residual from Eq. (29), we get

\[
\nabla f_n(\theta^{k+1}_n) + \alpha^{k+1}_n + \rho d_n \epsilon^{k+1}_n + s^{k+1}_n = 0.
\]

(50)

Therefore, \( \theta^{k+1}_n \) minimizes the function \( f_n(\theta_n) + \langle \alpha^{k+1}_n + \rho d_n \epsilon^{k+1}_n + s^{k+1}_n, \theta_n \rangle \) and as a consequence

\[
\mathbb{E} \left[ f_n(\theta^{k+1}_n) \right] + \mathbb{E} \left[ \langle \alpha^{k+1}_n + \rho d_n \epsilon^{k+1}_n + s^{k+1}_n, \theta^{k+1}_n \rangle \right] \\
\leq \mathbb{E} \left[ f_n(\theta^*_n) \right] + \mathbb{E} \left[ \langle \alpha^{k+1}_n + \rho d_n \epsilon^{k+1}_n + s^{k+1}_n, \theta^*_n \rangle \right] .
\]

(51)
This concludes the proof of part (ii) of Lemma 1. Now, to prove (ii), we know from the optimality conditions that

$$\nabla f_n(\theta_n^*) + \alpha_n^* = 0, \ \forall n \in \mathcal{V}. \quad (58)$$
Thus, \( \theta_n^* \) minimizes the function \( f_n(\theta_n) + \langle \alpha_n^*, \theta_n \rangle \) and we can write for \( n \in \mathcal{H} \)
\[
E \left[ f_n(\theta_n^*) \right] + E \left[ (\alpha_n^*, \theta_n^*) \right] \leq E \left[ f_n(\theta_n^{k+1}) \right] + E \left[ (\alpha_n^*, \theta_n^{k+1}) \right].
\]
(59)

Similarly, we have, for \( m \in \mathcal{T} \), that
\[
E \left[ f_m(\theta_m^*) \right] + E \left[ (\alpha_m^*, \theta_m^*) \right] \leq E \left[ f_m(\theta_m^{k+1}) \right] + E \left[ (\alpha_m^*, \theta_m^{k+1}) \right].
\]
(60)

Summing over all workers, we get
\[
\sum_{n=1}^{N} E \left[ f_n(\theta_n^{k+1}) - f_n(\theta_n^*) \right] \geq \sum_{n \in \mathcal{H}} E \left[ (\alpha_n^{k+1}, \theta_n^* - \theta_n^{k+1}) \right] + \sum_{m \in \mathcal{T}} E \left[ (\alpha_m^{k+1}, \theta_m^* - \theta_m^{k+1}) \right]
\]
\[
\geq (a) \sum_{(n,m) \in \mathcal{E}} E \left[ (\lambda_{n,m}^*, \theta_n^* - \theta_n^{k+1}) \right] + \sum_{(n,m) \in \mathcal{E}} E \left[ (\lambda_{m,n}^*, \theta_m^* - \theta_m^{k+1}) \right]
\]
\[
\geq (b) - \sum_{(n,m) \in \mathcal{E}} E \left[ (\lambda_{n,m}^*, r_{n,m}^{k+1}) \right].
\]
(61)
(62)

where we used the definition of \( \alpha_n^*, \ n \in \mathcal{V} \) in (a) and that \( \lambda_{n,m}^{k+1} = -\lambda_{m,n}^{k+1} \), and that \( \theta_n^* = \theta_m^*, \ \forall (n,m) \in \mathcal{E} \) in (b).

Appendix C. Proof of Theorem 2

Multiplying Eq. (35) by (-1), adding Eq. (34) and multiplying the sum by 2, we get
\[
2 \sum_{(n,m) \in \mathcal{E}} E \left[ (\lambda_{n,m}^*, \lambda_{n,m}^{k+1}, r_{n,m}^{k+1}) \right] + 2 \sum_{n \in \mathcal{H}} E \left[ (s_n^{k+1}, \theta_n^* - \theta_n^{k+1}) \right]
\]
\[
+ 2 \rho \sum_{n=1}^{N} E \left[ (d_n, e_n^{k+1}, \theta_n^* - \theta_n^{k+1}) \right] \geq 0.
\]
(63)

Since \( \lambda_{n,m}^{k+1} = \lambda_{n,m}^k + \rho r_{n,m}^{k+1} + \rho (e_n^{k+1} - e_n^{k+1}) \), then we can write
\[
2 \sum_{(n,m) \in \mathcal{E}} E \left[ (\lambda_{n,m}^*, \lambda_{n,m}^{k+1}, r_{n,m}^{k+1}) \right]
\]
\[
= 2 \sum_{(n,m) \in \mathcal{E}} E \left[ (\lambda_{n,m}^*, \lambda_{n,m}^{k+1}, r_{n,m}^{k+1}) \right] - 2 \rho \sum_{(n,m) \in \mathcal{E}} E \left[ \| r_{n,m}^{k+1} \|^2 \right]
\]
\[
- 2 \rho \sum_{(n,m) \in \mathcal{E}} E \left[ (e_n^{k+1} - e_n^{k+1}, r_{n,m}^{k+1}) \right].
\]
(64)

Using the identity
\[
r_{n,m}^{k+1} = \frac{1}{\rho} (\lambda_{n,m}^{k+1} - \lambda_{n,m}^*) - \frac{1}{\rho} (\lambda_{n,m}^k - \lambda_{n,m}^*) + e_n^{k+1} - e_m^{k+1},
\]
(65)
we will examine the different terms of (64) starting from the first term

\[
2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^k - \lambda_{n,m}^k, r_{n,m}^{k+1} \rangle \right] \\
= \frac{2}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^k - \lambda_{n,m}^k, \lambda_{n,m}^{k+1} - \lambda_{n,m}^* \rangle \right] + \frac{2}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^k - \lambda_{n,m}^* \|^2 \right] \\
+ 2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^* - \lambda_{n,m}^k, \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^* \rangle \right].
\]  

(66)

The second term can be re-written as

\[
-2\rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| r_{n,m}^{k+1} \|^2 \right] \\
= -\rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| r_{n,m}^{k+1} \|^2 \right] - \frac{1}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^{k+1} - \lambda_{n,m}^* \|^2 \right] - \frac{1}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^k - \lambda_{n,m}^* \|^2 \right] \\
- \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^* \|^2 \right] + \frac{2}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^{k+1} - \lambda_{n,m}^*, \lambda_{n,m}^k - \lambda_{n,m}^* \rangle \right] \\
- 2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^{k+1} - \lambda_{n,m}^*, \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^* \rangle \right] + 2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^k - \lambda_{n,m}^*, \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^* \rangle \right] .
\]  

(67)

The third term can be expanded as

\[
-2\rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^k, r_{n,m}^{k+1} \rangle \right] \\
= -2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^k, \lambda_{n,m}^{k+1} - \lambda_{n,m}^* \rangle \right] + 2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^k, \lambda_{n,m}^k - \lambda_{n,m}^* \rangle \right] \\
+ 2\rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^k \|^2 \right] .
\]  

(68)

From Eqs. (66)-(68), we can re-write Eq. (64) as

\[
2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^* - \lambda_{n,m}^{k+1}, r_{n,m}^{k+1} \rangle \right] \\
= \frac{1}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^k - \lambda_{n,m}^* \|^2 \right] - \frac{1}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^{k+1} - \lambda_{n,m}^* \|^2 \right] - \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| r_{n,m}^{k+1} \|^2 \right] \\
+ 2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^k, \lambda_{n,m}^k - \lambda_{n,m}^* \rangle \right] + \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^k \|^2 \right].
\]  

(69)
The second term of the left hand-side of Eq. (63) can be decomposed as

\begin{align*}
2 \sum_{n \in \mathcal{H}} & \mathbb{E} \left[ \langle s_n^{k+1}, \theta_n^* - \theta_n^{k+1} \rangle \right] \\
= 2 \rho & \sum_{n \in \mathcal{H}} \sum_{m \in \mathcal{N}_n} \mathbb{E} \left[ \langle \hat{\theta}_m^{k+1} - \hat{\theta}_m^k, \theta_n^* - \theta_n^{k+1} \rangle \right] \\
= 2 \rho & \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \hat{\theta}_m^{k+1} - \hat{\theta}_m^k, \theta_n^* - \theta_n^{k+1} \rangle \right] \\
= 2 \rho & \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \hat{\theta}_m^{k+1} - \hat{\theta}_m^k, \theta_n^* - \theta_n^{k+1} \rangle \right] - 2 \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \hat{\theta}_m^{k+1} - \hat{\theta}_m^k, r_{n,m}^{k+1} \rangle \right]. \quad (70)
\end{align*}

Now, we can re-write the first term as

\begin{align*}
-2 \rho & \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \hat{\theta}_m^{k+1} - \hat{\theta}_m^k, r_{n,m}^{k+1} \rangle \right] \\
= -2 \rho & \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \theta_m^{k+1} - \theta_m^k - \epsilon_m^k, r_{n,m}^{k+1} \rangle \right] \\
= -2 \rho & \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \theta_m^{k+1} - \theta_m^k, r_{n,m}^{k+1} \rangle \right] + 2 \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_m^{k+1} - \epsilon_m^k, r_{n,m}^{k+1} \rangle \right]. \quad (71)
\end{align*}

The second term can be expanded as

\begin{align*}
2 \rho & \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \hat{\theta}_m^{k+1} - \hat{\theta}_m^k, \theta_n^* - \theta_n^{k+1} \rangle \right] \\
= 2 \rho & \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \theta_m^{k+1} - \theta_m^k, \theta_n^* - \theta_n^{k+1} \rangle \right] + 2 \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_m^{k+1} - \epsilon_m^k, \theta_n^* - \theta_n^{k+1} \rangle \right]. \quad (72)
\end{align*}

Since \( \theta_n^* = \theta_m^* \), \( \forall (n,m) \in \mathcal{E} \) and \( \theta_n^* - \theta_m^{k+1} = \theta_n^* - \theta_m^k + \theta_n^{k+1} - \theta_m^{k+1} \), we can write

\begin{align*}
2 \rho & \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \theta_m^{k+1} - \theta_m^k, \theta_n^* - \theta_n^{k+1} \rangle \right] \\
= -2 \rho & \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|	heta_m^{k+1} - \theta_m^k\|^2 \right] + 2 \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \theta_m^{k+1} - \theta_m^k, \theta_n^* - \theta_n^{k+1} \rangle \right] \\
= -\rho & \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|	heta_m^{k+1} - \theta_m^k\|^2 \right] - \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|	heta_m^{k+1} - \theta_m^k\|^2 \right] - \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|	heta_m^{k+1} - \theta_m^k\|^2 \right] \\
+ 2 \rho & \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \theta_m^{k+1} - \theta_m^k, \theta_n^* - \theta_n^k \rangle \right]. \quad (73)
\end{align*}
With this expression at hand, we can go back to Eq. (70)

\[ 2 \sum_{n \in \mathcal{H}} \mathbb{E} \left[ \langle s_{n}^{k+1}, \theta_{n}^{*} - \theta_{n}^{k+1} \rangle \right] \]

\[ = \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \theta_{m}^{k} - \theta_{m}^{*} \|^2 \right] - \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \theta_{m}^{k+1} - \theta_{m}^{*} \|^2 \right] + 2 \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \epsilon_{m}^{k+1} - \epsilon_{m}^{k}, \theta_{m}^{k+1} - \theta_{n}^{*} \|^2 \right] \]

\[ - 2 \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \theta_{m}^{k+1} - \theta_{m}^{k}, r_{n,m}^{k+1} \rangle \right] + 2 \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{m}^{k+1} - \epsilon_{m}^{k}, r_{n,m}^{k+1} \rangle \right] - \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \theta_{m}^{k+1} - \theta_{m}^{k} \|^2 \right]. \]  

(74)

Replacing Eq. (69) and (74) in (63), we obtain

\[ \frac{1}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^{k} - \lambda_{n,m}^{*} \|^2 \right] - \frac{1}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^{k+1} - \lambda_{n,m}^{*} \|^2 \right] - \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| r_{n,m}^{k+1} \|^2 \right] \]

\[ + 2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{m}^{k+1} - \epsilon_{m}^{k+1}, \lambda_{n,m}^{k} - \lambda_{n,m}^{*} \rangle \right] + \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \epsilon_{m}^{k+1} - \epsilon_{m}^{k+1} \|^2 \right] \]

\[ + \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \theta_{m}^{k+1} - \theta_{n}^{*} \|^2 \right] - \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \theta_{m}^{k+1} - \theta_{m}^{*} \|^2 \right] - \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \theta_{m}^{k+1} - \theta_{m}^{k} \|^2 \right] \]

\[ - 2 \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \theta_{m}^{k+1} - \theta_{m}^{k}, r_{n,m}^{k+1} \rangle \right] + 2 \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{m}^{k+1} - \epsilon_{m}^{k}, r_{n,m}^{k+1} \rangle \right] \]

\[ + 2 \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{m}^{k+1} - \epsilon_{m}^{k}, \theta_{m}^{k+1} - \theta_{n}^{*} \rangle \right] + 2 \rho \sum_{n=1}^{N} \mathbb{E} \left[ \langle d_{n}, e_{n}^{k+1}, \theta_{n}^{*} - \theta_{n}^{k+1} \rangle \right] \geq 0. \]  

(75)

Using the identity

\[ r_{n,m}^{k+1} = \frac{1}{\rho} (\lambda_{n,m}^{k+1} - \lambda_{n,m}^{k}) + \epsilon_{n}^{k+1} - \epsilon_{m}^{k+1}, \]  

(76)

we can write

\[ - \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| r_{n,m}^{k+1} \|^2 \right] \]

\[ = - \frac{1}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^{k+1} - \lambda_{n,m}^{k} \|^2 \right] - \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \epsilon_{n}^{k+1} - \epsilon_{m}^{k+1} \|^2 \right] \]

\[ + 2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^{k+1} - \lambda_{n,m}^{k}, \epsilon_{m}^{k+1} - \epsilon_{m}^{k+1} \rangle \right]. \]  

(77)
On the other hand, we have

\[
2\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^k, r_{n,m}^{k+1} \rangle \right] + 2\rho \sum_{n=1}^{N} \mathbb{E} \left[ \langle d_n \epsilon_n^{k+1}, \theta^*_n - \theta_m^{k+1} \rangle \right]
= 2\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^k, r_{n,m}^{k+1} \rangle \right] + 2\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1}, \theta^*_n - \theta_m^{k+1} \rangle \right] + 2\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1}, \theta^*_m - \theta_m^{k+1} \rangle \right] - 2\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{m,n}^{k+1}, r_{m,n}^{k+1} \rangle \right]
= 2\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1}, \theta^*_m - \theta_m^{k+1} \rangle \right] + 2\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{m,n}^{k+1}, \theta^*_n - \theta_m^{k+1} \rangle \right] - 2\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1}, \theta^*_m - \theta_m^{k+1} \rangle \right] + 2\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{m,n}^{k+1}, \theta^*_n - \theta_m^{k+1} \rangle \right] - 2\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{m,n}^{k+1}, \theta^*_n - \theta_m^{k+1} \rangle \right] \quad \text{(79)}
\]

Now, recall that \(\theta_m^{k+1}, m \in \mathcal{T}\) minimizes the function \(f_m(\theta_m) + \langle \alpha_m^{k+1} + \rho d_m \epsilon_m^{k+1}, \theta_m \rangle\) and \(\theta_m^k, m \in \mathcal{T}\) minimizes the function \(f_m(\theta_m) + \langle \alpha_m^k + \rho d_m \epsilon_m^k, \theta_m \rangle\), then we could write

\[
\mathbb{E} \left[ f_m(\theta_m^{k+1}) \right] + \mathbb{E} \left[ \langle \alpha_m^{k+1} + \rho d_m \epsilon_m^{k+1}, \theta_m^{k+1} \rangle \right] \leq \mathbb{E} \left[ f_m(\theta_m^k) \right] + \mathbb{E} \left[ \langle \alpha_m^k + \rho d_m \epsilon_m^k, \theta_m^k \rangle \right] \quad \text{(78)}
\]

Adding both equations and re-arranging the terms, we get

\[
\mathbb{E} \left[ \langle \alpha_m^{k+1} - \alpha_m^k, \theta_m^{k+1} - \theta_m^k \rangle \right] \leq -\rho d_m \mathbb{E} \left[ \langle \epsilon_m^{k+1} - \epsilon_m^k, \theta_m^{k+1} - \theta_m^k \rangle \right] \quad \text{(80)}
\]

Using the update of \(\alpha_m^{k+1}\), i.e. \(\alpha_m^{k+1} = \alpha_m^k + \rho \sum_{n \in \mathcal{N}_m, r_{n,m}^{k+1}} \epsilon_{n,m}^{k+1} \), we can re-write Eq. (80) to get

\[
-\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle r_{n,m}^{k+1}, \theta_m^{k+1} - \theta_m^k \rangle \right] \leq -\rho \sum_{(n,m)\in\mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^k, \theta_m^{k+1} - \theta_m^k \rangle \right] \quad \text{(81)}
\]

where we used the fact that \(r_{n,m}^{k+1} = -r_{n,m}^{k+1}\) after summing over \(m \in \mathcal{T}\).
Going back to (75), we can write

\[
\begin{align*}
\frac{1}{\rho} \left\{ \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^{k+1} - \lambda_{n,m}^* - \epsilon_{n,m}^k \|_2^2 \right] - \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^k - \lambda_{n,m}^* \|_2^2 \right] + \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^{k+1} - \lambda_{n,m}^k \|_2^2 \right] \right\} \\
+ \rho \left\{ \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \theta_{m}^{k+1} - \theta_{m}^* \|_2^2 \right] - \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \theta_{m}^k - \theta_{m}^* \|_2^2 \right] + \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \theta_{m}^{k+1} - \theta_{m}^k \|_2^2 \right] \right\} \\
\leq 2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^{k+1} - \lambda_{n,m}^*, \epsilon_{n,m}^{k+1} - \epsilon_{n}^k \rangle \right] + 2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^{k+1} - \lambda_{n,m}^k, \epsilon_{n,m}^{k+1} - \epsilon_{n}^k \rangle \right] \\
- 2\rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \epsilon_{n,m}^{k+1} - \epsilon_{n,m}^k \|_2^2 \right] - 2\rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \theta_{m}^{k+1} - \theta_{m}^k, \epsilon_{n,m}^{k+1} - \epsilon_{n}^k \rangle \right] \\
+ 2\rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1} + \epsilon_{n,m}^k, \theta_{m}^* - \theta_{m}^{k+1} \rangle \right] - 2\rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{m}^{k+1} + \epsilon_{n}^k, \theta_{m}^* - \theta_{m}^{k+1} \rangle \right].
\end{align*}
\]

To upper bound the terms in the right hand side, we will use the identity (43)

\[
\begin{align*}
2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^{k+1} - \lambda_{n,m}^*, \epsilon_{n,m}^{k+1} - \epsilon_{n}^k \rangle \right] \\
\leq \frac{1}{\eta_1} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \epsilon_{n,m}^{k+1} - \epsilon_{n}^k \|_2^2 \right] + \eta_1 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^{k+1} - \lambda_{n,m}^* \|_2^2 \right],
\end{align*}
\]

\[
\begin{align*}
2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \lambda_{n,m}^{k+1} - \lambda_{n,m}^k, \epsilon_{n,m}^{k+1} - \epsilon_{n}^k \rangle \right] \\
\leq \frac{1}{\eta_2} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \epsilon_{n,m}^{k+1} - \epsilon_{n}^k \|_2^2 \right] + \eta_2 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \lambda_{n,m}^{k+1} - \lambda_{n,m}^k \|_2^2 \right],
\end{align*}
\]

\[
\begin{align*}
2\rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \epsilon_{n,m}^{k+1} + \epsilon_{n,m}^k, \theta_{m}^* - \theta_{m}^{k+1} \rangle \right] \\
\leq \frac{\rho}{\eta_3} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \epsilon_{n,m}^{k+1} + \epsilon_{n,m}^k \|_2^2 \right] + \rho\eta_3 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \theta_{m}^* - \theta_{m}^{k+1} \|_2^2 \right],
\end{align*}
\]

\[
\begin{align*}
- 2\rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \langle \theta_{m}^{k+1} - \theta_{m}^k, \epsilon_{n,m}^{k+1} - \epsilon_{n}^k \rangle \right] \\
\leq \frac{\rho}{\eta_4} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \epsilon_{m}^{k+1} - \epsilon_{n}^k \|_2^2 \right] + \rho\eta_4 \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \| \theta_{m}^{k+1} - \theta_{m}^k \|_2^2 \right],
\end{align*}
\]
Finally, we use both identities (42) and (43) to get the following bound

\[
- 2\rho \sum_{(n,m) \in E} \mathbb{E} \left[ \langle \epsilon_{m}^{k}, r_{n,m}^{k+1} \rangle \right] \\
\leq \frac{\rho}{\eta_5} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k}\|^2 \right] + \rho \eta_5 \sum_{(n,m) \in E} \mathbb{E} \left[ \|r_{n,m}^{k+1}\|^2 \right] \\
\leq \frac{\rho}{\eta_5} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k}\|^2 \right] + \frac{2\rho}{\eta_5} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\lambda_{n,m}^{k+1} - \lambda_{n,m}^{k}\|^2 \right] + 2\rho \eta_5 \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k+1} - \epsilon_{m}^{k}\|^2 \right],
\]

(87)

where \( \{\eta_i\}_{i=1}^{5} \) are arbitrary positive constants to be specified later on. Using these bounds and re-arranging the terms in Eq. (82), we can write

\[
2\rho \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k+1} - \epsilon_{m}^{k}\|^2 \right] + \rho (1 - \eta_4) \sum_{(n,m) \in E} \mathbb{E} \left[ \|\theta_{m}^{k+1} - \theta_{m}^{k}\|^2 \right] \\
+ \left( \frac{1}{\rho} - 2\eta_5 \right) \sum_{(n,m) \in E} \mathbb{E} \left[ \|\lambda_{n,m}^{k+1} - \lambda_{n,m}^{k}\|^2 \right] \\
\leq \left( \frac{1}{\eta_1} + \frac{1}{\eta_2} + 2\rho \eta_5 \right) \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k+1} - \epsilon_{m}^{k}\|^2 \right] + \frac{\rho}{\eta_3} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k+1} + \epsilon_{m}^{k}\|^2 \right] \\
+ \frac{1}{\rho} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\lambda_{m}^{k} - \lambda_{m}^{*}\|^2 \right] - \left( \frac{1}{\rho} - \eta_1 \right) \sum_{(n,m) \in E} \mathbb{E} \left[ \|\lambda_{n,m}^{k+1} - \lambda_{n,m}^{*}\|^2 \right] + \rho \sum_{(n,m) \in E} \mathbb{E} \left[ \|\theta_{m}^{k} - \theta_{m}^{*}\|^2 \right] \\
- \rho (1 - \eta_3) \sum_{(n,m) \in E} \mathbb{E} \left[ \|\theta_{m}^{k+1} - \theta_{m}^{*}\|^2 \right] + \frac{\rho}{\eta_5} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k}\|^2 \right] + \frac{\rho}{\eta_4} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k+1} - \epsilon_{m}^{k}\|^2 \right].
\]

(88)

Now, we choose to fix the values of \( \{\eta_i\}_{i=1}^{5} \) to be \((\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = \left( \frac{\psi_k}{2\psi^0}, \frac{1}{4\rho}, \frac{\psi_k}{2\psi^0}, \frac{1}{2}, \frac{1}{4} \right)\).

With these values at hand, we get

\[
\frac{\rho}{2} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\theta_{m}^{k+1} - \theta_{m}^{k}\|^2 \right] + \frac{1}{4\rho} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\lambda_{n,m}^{k+1} - \lambda_{n,m}^{k}\|^2 \right] \\
\leq \left( \frac{5\rho}{2} + \frac{2\rho \psi^0}{\psi_k} \right) \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k+1} - \epsilon_{m}^{k}\|^2 \right] + \frac{2\rho \psi^0}{\psi_k} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k+1} + \epsilon_{m}^{k}\|^2 \right] \\
+ \frac{1}{\rho} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\lambda_{n,m} - \lambda_{n,m}^{*}\|^2 \right] - \frac{1}{\rho} \sum_{(n,m) \in E} \mathbb{E} \left[ \|\lambda_{n,m}^{k+1} - \lambda_{n,m}^{*}\|^2 \right] + \rho \sum_{(n,m) \in E} \mathbb{E} \left[ \|\theta_{m}^{k} - \theta_{m}^{*}\|^2 \right] \\
- \rho \sum_{(n,m) \in E} \mathbb{E} \left[ \|\theta_{m}^{k+1} - \theta_{m}^{*}\|^2 \right] + 4\rho \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k}\|^2 \right] + 2\rho \sum_{(n,m) \in E} \mathbb{E} \left[ \|\epsilon_{m}^{k+1} - \epsilon_{m}^{k}\|^2 \right].
\]

(89)
Re-arranging the terms and upper bounding the terms involving the censoring errors, we can write

\[
\frac{\rho}{2} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\theta_{m}^{k+1} - \theta_{m}^{k}\|^2 \right] + \frac{1}{4\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\lambda_{n,m}^{k+1} - \lambda_{n,m}^{k}\|^2 \right] \\
\leq \frac{1}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\lambda_{n,m}^{k} - \lambda_{n,m}^{*}\|^2 \right] - \frac{1}{\rho} \left(1 - \frac{\psi_{k}}{2\psi_{0}}\right) \mathbb{E} \left[ \|\lambda_{n,m}^{k+1} - \lambda_{n,m}^{*}\|^2 \right] + \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\theta_{m}^{k} - \theta_{m}^{*}\|^2 \right] \\
- \rho \left(1 - \frac{\psi_{k}}{2\psi_{0}}\right) \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\theta_{m}^{k+1} - \theta_{m}^{k}\|^2 \right] + \left(5\rho + \frac{8\rho\psi_{0}}{\psi_{k}}\right) \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\epsilon_{m}^{k+1}\|^2 \right] \\
+ \left(9\rho + \frac{4\rho\psi_{0}}{\psi_{k}}\right) \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\epsilon_{m}^{k}\|^2 \right] + \left(8\rho + \frac{4\rho\psi_{0}}{\psi_{k}}\right) \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\theta_{m}^{k}\|^2 \right].
\]  

(90)

Therefore, using (33), we can write

\[
\frac{\rho}{2} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\theta_{m}^{k+1} - \theta_{m}^{k}\|^2 \right] + \frac{1}{4\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\lambda_{n,m}^{k+1} - \lambda_{n,m}^{k}\|^2 \right] \\
\leq \frac{1}{\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\lambda_{n,m}^{k} - \lambda_{n,m}^{*}\|^2 \right] - \frac{1}{\rho} \left(1 - \frac{\psi_{k}}{2\psi_{0}}\right) \mathbb{E} \left[ \|\lambda_{n,m}^{k+1} - \lambda_{n,m}^{*}\|^2 \right] + \rho \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\theta_{m}^{k} - \theta_{m}^{*}\|^2 \right] \\
- \rho \left(1 - \frac{\psi_{k}}{2\psi_{0}}\right) \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\theta_{m}^{k+1} - \theta_{m}^{k}\|^2 \right] + \gamma_{1}\psi_{k} + \gamma_{2}\psi_{2k},
\]  

(91)

where \(\gamma_{1} = 64\rho C_{0}\psi_{0}|\mathcal{E}|\) and \(\gamma_{2} = 88\rho C_{0}^{2}|\mathcal{E}|\). Now, we define the Lyapunov function

\[
V^{k} = \frac{1}{\rho} \sum_{(n,m) \in \mathcal{E}} \|\lambda_{n,m}^{k} - \lambda_{n,m}^{*}\|^2 + \rho \sum_{(n,m) \in \mathcal{E}} \|\theta_{m}^{k} - \theta_{m}^{*}\|^2.
\]  

(92)

Thus, we get

\[
\frac{\rho}{2} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\theta_{m}^{k+1} - \theta_{m}^{k}\|^2 \right] + \frac{1}{4\rho} \sum_{(n,m) \in \mathcal{E}} \mathbb{E} \left[ \|\lambda_{n,m}^{k+1} - \lambda_{n,m}^{k}\|^2 \right] \\
\leq \mathbb{E} \left[ V^{k} \right] - \left(1 - \frac{\psi_{k}}{2\psi_{0}}\right) \mathbb{E} \left[ V^{k+1} \right] + \gamma_{1}\psi_{k} + \gamma_{2}\psi_{2k}.
\]  

(93)

As a consequence, we can write that

\[
\mathbb{E} \left[ V^{k} \right] - \left(1 - \frac{\psi_{k}}{2\psi_{0}}\right) \mathbb{E} \left[ V^{k+1} \right] + \gamma_{1}\psi_{k} + \gamma_{2}\psi_{2k} \geq 0.
\]  

(94)

Re-arranging the terms, we get

\[
\mathbb{E} \left[ V^{k+1} \right] \leq \left(1 - \frac{\psi_{k}}{2\psi_{0}}\right)^{-1} \left(\mathbb{E} \left[ V^{k} \right] + \gamma_{1}\psi_{k} + \gamma_{2}\psi_{2k}\right).
\]  

(95)

27
Using this equation iteratively, we obtain

\[
E[V^{k+1}] \\
\leq \left( \prod_{j=0}^{k} \left( 1 - \frac{\psi^j}{2\psi^0} \right)^{-1} \right) \mathbb{E}[V_0] + \gamma_1 \sum_{j=0}^{k} \prod_{i=j}^{k} \left( 1 - \frac{\psi^i}{2\psi^0} \right)^{-1} \psi^j + \gamma_2 \sum_{j=0}^{k} \prod_{i=j}^{k} \left( 1 - \frac{\psi^i}{2\psi^0} \right)^{-1} \psi^{2j} \\
\leq \left( \prod_{j=0}^{k} \left( 1 - \frac{\psi^j}{2\psi^0} \right)^{-1} \right) \mathbb{E}[V_0] + \gamma_1 \prod_{j=0}^{k} \left( 1 - \frac{\psi^j}{2\psi^0} \right)^{-1} \sum_{i=0}^{k} \psi^i + \gamma_2 \prod_{j=0}^{k} \left( 1 - \frac{\psi^j}{2\psi^0} \right)^{-1} \sum_{i=0}^{k} \psi^{2i} \\
\leq \prod_{j=0}^{\infty} \left( 1 - \frac{\psi^j}{2\psi^0} \right)^{-1} \left( \mathbb{E}[V^0] + \gamma_1 \sum_{i=0}^{\infty} \psi^i + \gamma_2 \sum_{i=0}^{\infty} \psi^{2i} \right),
\]

(96)

where we have used the fact that \( \left( 1 - \frac{\psi^k}{2\psi^0} \right) \in [\frac{1}{2}, 1] \). Since \( \sum_{i=0}^{\infty} \omega^i < \infty \) and \( \sum_{i=0}^{\infty} \xi^i < \infty \), thus \( \sum_{i=0}^{\infty} \psi^i < \infty \). Furthermore, the sequence \( \{\psi^i\} \) is non-negative, then we get that \( \sum_{i=0}^{\infty} \psi^{2i} < \infty \). To show that \( \prod_{j=0}^{\infty} \left( 1 - \frac{\psi^j}{2\psi^0} \right)^{-1} \) is also finite, we consider its logarithm, i.e.

\[
\log \left( \prod_{j=0}^{\infty} \left( 1 - \frac{\psi^j}{2\psi^0} \right)^{-1} \right) \leq \sum_{j=0}^{\infty} \log \left( \left( 1 - \frac{\psi^j}{2\psi^0} \right)^{-1} \right) \leq \sum_{j=0}^{\infty} \log \left( 1 + \frac{\psi^j}{\psi^0} \right) \leq \frac{1}{\psi^0} \sum_{j=0}^{\infty} \psi^j,
\]

(97)

where we have used that \( \log \left( \left( 1 - \frac{\psi^0}{2\psi^0} \right)^{-1} \right) \leq \log(1+z) \), \( z \geq 1 \) in (a) and \( \log(1+z) \leq z \), \( z \geq 1 \) in (b). Hence, \( \prod_{j=0}^{\infty} \left( 1 - \frac{\psi^j}{2\psi^0} \right)^{-1} \) is also finite and we conclude that the sequence \( \mathbb{E}[V^k] \) is upper bounded by a finite quantity that we denote as \( \tilde{V} \). Going back to Eq. (93) and taking the sum from \( k = 0 \) to \( \infty \) while using the upper bound on \( \mathbb{E}[V^k] \), we can write

\[
\frac{\rho}{2} \sum_{k=0}^{\infty} \sum_{(n,m) \in \mathcal{E}} \mathbb{E}[\| \theta_{m}^{k+1} - \theta_{m}^{k} \|^2] + \frac{1}{4\rho} \sum_{k=0}^{\infty} \sum_{(n,m) \in \mathcal{E}} \mathbb{E}[\| \lambda_{n,m}^{k+1} - \lambda_{n,m}^{k} \|^2] \\
\leq \tilde{V}^0 + \left( \frac{\tilde{V}}{2\psi^0} + \gamma_1 \right) \sum_{k=0}^{\infty} \psi^k + \gamma_2 \sum_{k=0}^{\infty} \psi^{2k}.
\]

(98)

Since the right hand side is finite, we conclude that the left hand side is convergent and as a consequence, we can write that

\[
\lim_{k \to \infty} \mathbb{E}[\| \theta_{m}^{k+1} - \theta_{m}^{k} \|^2] = 0,
\]

(99)

\[
\lim_{k \to \infty} \mathbb{E}[\| \lambda_{n,m}^{k+1} - \lambda_{n,m}^{k} \|^2] = 0.
\]

(100)

We recall the expression of both the primal and dual residuals as

\[
r_{n,m}^{k+1} = \frac{1}{\rho} (\lambda_{n,m}^{k+1} - \lambda_{n,m}^{k}) + \epsilon_{n}^{k+1} - \epsilon_{m}^{k+1},
\]

(101)

\[
s_{n}^{k+1} = \rho \sum_{m \in \mathcal{N}_n} (\theta_{m}^{k+1} - \theta_{m}^{k}) + \rho \sum_{m \in \mathcal{N}_n} (\epsilon_{m}^{k} - \epsilon_{m}^{k+1}).
\]

(102)
Using (42), we can derive the following bounds
\[
\mathbb{E}\left[\|r_{n,m}^{k+1}\|^2\right] \leq 2\left(\frac{1}{\rho^2}\mathbb{E}\left[\|\lambda_{n,m}^{k+1} - \lambda_{n,m}^k\|^2\right] + 2\left(\mathbb{E}\left[\|\epsilon_m^k\|^2\right] + \mathbb{E}\left[\|\epsilon_m^{k+1}\|^2\right]\right)\right), \tag{103}
\]
\[
\mathbb{E}\left[\|s_{n}^{k+1}\|^2\right] \leq 2\rho^2d_n\left(\sum_{m \in N_n} \mathbb{E}\left[\|\theta_{m}^{k+1} - \theta_{m}^k\|^2\right] + \sum_{m \in N_n} \mathbb{E}\left[\|\epsilon_{m}^k - \epsilon_{m}^{k+1}\|^2\right]\right). \tag{104}
\]
Since \(\mathbb{E}\left[\|\epsilon_n^k\|^2\right] \leq 4C_0^2\psi^{2k}, \forall n\), then \(\lim_{k \to \infty} \mathbb{E}\left[\|\epsilon_n^k\|^2\right] = 0\). Using Eqs. (99), (100) and \(\lim_{k \to \infty} \mathbb{E}\left[\|r_{n,m}^{k+1}\|^2\right] = 0\) and \(\lim_{k \to \infty} \mathbb{E}\left[\|s_{n}^{k+1}\|^2\right] = 0\), we conclude that \(\lim_{k \to \infty} \mathbb{E}\left[\|r_{n,m}^{k+1}\|^2\right] = 0\) and \(\lim_{k \to \infty} \mathbb{E}\left[\|s_{n}^{k+1}\|^2\right] = 0\).

Using the Cauchy-Schwarz inequality (45), we can write
\[
\left|\mathbb{E}\left[\langle \lambda_{n,m}^{k+1}, r_{n,m}^{k+1}\rangle\right]\right| \leq \left(\mathbb{E}\left[\|\lambda_{n,m}^{k+1}\|^2\right]\right)^{\frac{1}{2}} \left(\mathbb{E}\left[\|r_{n,m}^{k+1}\|^2\right]\right)^{\frac{1}{2}}, \tag{105}
\]
\[
\left|\mathbb{E}\left[\langle \lambda_{n,m}^*, r_{n,m}^{k+1}\rangle\right]\right| \leq \left(\mathbb{E}\left[\|\lambda_{n,m}^*\|^2\right]\right)^{\frac{1}{2}} \left(\mathbb{E}\left[\|r_{n,m}^{k+1}\|^2\right]\right)^{\frac{1}{2}}, \tag{106}
\]
\[
\left|\mathbb{E}\left[\langle s_{n}^{k+1}, \theta_{n}^* - \theta_{n}^{k+1}\rangle\right]\right| \leq \left(\mathbb{E}\left[\|s_{n}^{k+1}\|^2\right]\right)^{\frac{1}{2}} \left(\mathbb{E}\left[\|\theta_{n}^* - \theta_{n}^{k+1}\|^2\right]\right)^{\frac{1}{2}}, \tag{107}
\]
\[
\left|\mathbb{E}\left[\langle e_{n}^{k+1}, \theta_{n}^* - \theta_{n}^{k+1}\rangle\right]\right| \leq \left(\mathbb{E}\left[\|e_{n}^{k+1}\|^2\right]\right)^{\frac{1}{2}} \left(\mathbb{E}\left[\|\theta_{n}^* - \theta_{n}^{k+1}\|^2\right]\right)^{\frac{1}{2}}. \tag{108}
\]
Since \(\lim_{k \to \infty} \mathbb{E}\left[\|\epsilon_n^k\|^2\right] = 0\), \(\lim_{k \to \infty} \mathbb{E}\left[\|r_{n,m}^{k+1}\|^2\right] = 0\) and \(\lim_{k \to \infty} \mathbb{E}\left[\|s_{n}^{k+1}\|^2\right] = 0\), we get, from (106)-(108), that
\[
\lim_{k \to \infty} \mathbb{E}\left[\langle \lambda_{n,m}^{k+1}, r_{n,m}^{k+1}\rangle\right] = 0, \tag{109}
\]
\[
\lim_{k \to \infty} \mathbb{E}\left[\langle \lambda_{n,m}^*, r_{n,m}^{k+1}\rangle\right] = 0, \tag{110}
\]
\[
\lim_{k \to \infty} \mathbb{E}\left[\langle s_{n}^{k+1}, \theta_{n}^* - \theta_{n}^{k+1}\rangle\right] = 0, \tag{111}
\]
\[
\lim_{k \to \infty} \mathbb{E}\left[\langle e_{n}^{k+1}, \theta_{n}^* - \theta_{n}^{k+1}\rangle\right] = 0. \tag{112}
\]
Furthermore, from (i) and (ii) of Lemma 1, we conclude that
\[
\lim_{k \to \infty} \sum_{n=1}^{N} \mathbb{E}\left[ f_n(\theta_n^k) - f_n(\theta_n^*) \right] = 0. \tag{113}
\]

**Appendix D. Proof of Theorem 3**

The proof of Theorem 3 follows similar steps as the proof of convergence rate of (Liu et al., 2019b) with the additional challenge of the parallel model updates of the head and tail workers. The alternating update nature of our algorithm makes the updates happen in an asymmetric manner, in contrast to the symmetric update in (Liu et al., 2019b), which makes the proof more complex. Recall that for a bipartite graph, the adjacency matrix can be written as
\[
A = \begin{pmatrix} 0_r & B \\ B^T & 0_{ss} \end{pmatrix}, \tag{114}
\]
where \( r = |\mathcal{H}|, s = |\mathcal{T}| \) are the cardinalities of the head \( \mathcal{H} \) and tail \( \mathcal{T} \) groups, respectively. The matrices \( \mathbf{0}_{rr} \), and \( \mathbf{0}_{ss} \) are the null matrices of order \( r \times r \), and \( s \times s \), respectively. The matrix \( \mathbf{B} \in \mathbb{R}^{r\times s} \) is called the bi-adjacency matrix. The adjacency matrix is a boolean matrix where each element is defined as \( A_{i,j} = 1 \) if there exists a link between the nodes \( i \) and \( j \) (i.e. workers), otherwise \( A_{i,j} = 0 \). In our analysis, we introduce the matrix \( \mathbf{C} \) as

\[
\mathbf{C} = \begin{pmatrix}
\mathbf{0}_{rr} & \mathbf{B} \\
\mathbf{0}_{rs} & \mathbf{0}_{ss}
\end{pmatrix}.
\]  

Due to the nature of the updates of the CQ-GGADMM, the matrix \( \mathbf{C} \) is needed to be able to write the updates in a matrix form. For the proof of the convergence rate, we also define the following matrices

\[
\begin{align*}
\mathbf{\theta} &= \begin{pmatrix} \theta_1^T \\ \vdots \\ \theta_N^T \end{pmatrix}, \\
\mathbf{\alpha} &= \begin{pmatrix} \alpha_1^T \\ \vdots \\ \alpha_N^T \end{pmatrix}, \\
\hat{\mathbf{\theta}} &= \begin{pmatrix} \hat{\theta}_1^T \\ \vdots \\ \hat{\theta}_N^T \end{pmatrix}, \\
\mathbf{E} &= \begin{pmatrix} \epsilon_1^T \\ \vdots \\ \epsilon_N^T \end{pmatrix}.
\end{align*}
\]

In this section, we also introduce certain matrices related to the network topology, namely \( \mathbf{D} \) the diagonal degree matrix, \( \mathbf{M}_- \) the signed incidence matrix, and \( \mathbf{M}_+ \) the unsigned incidence matrix. Using Eqs. (49), (52), and (23), the matrix form of the problem can be derived as

\[
\nabla f(\theta^{k+1}) + \mathbf{\alpha}^k - \rho \mathbf{C}\hat{\theta}^k - \rho \mathbf{C}^T\hat{\theta}^{k+1} + \rho \mathbf{D}\theta^{k+1} = 0,
\]

and the optimality conditions are given by

\[
\begin{align*}
\nabla f(\mathbf{\theta}^*) + \mathbf{\alpha}^* &= 0, \\
\mathbf{M}_-^T\mathbf{\alpha}^* &= 0.
\end{align*}
\]

Since \( \mathbf{D} - \mathbf{A} = \frac{1}{2}\mathbf{M}_-\mathbf{M}_+^T \), then we can re-write Eq. (118) as

\[
\mathbf{\alpha}^{k+1} = \mathbf{\alpha}^k + \frac{\rho}{2}\mathbf{M}_-\mathbf{M}_+^T\theta^{k+1} + \frac{\rho}{2}\mathbf{M}_-\mathbf{M}_+^T\mathbf{E}^{k+1}.
\]

Initializing \( \mathbf{\alpha}^0 \) in the column space of \( \mathbf{M}_- \), we get that \( \mathbf{\alpha}^k \) always stays in the column space of \( \mathbf{M}_- \) and thus, we have \( \mathbf{\alpha}^k = \mathbf{M}_-\mathbf{\beta}^k, \forall k \geq 0 \). Therefore, we can further write Eq. (118) as

\[
\mathbf{\beta}^{k+1} = \mathbf{\beta}^k + \frac{\rho}{2}\mathbf{M}_-\mathbf{\theta}^{k+1} + \frac{\rho}{2}\mathbf{M}_-\mathbf{E}^{k+1}.
\]

Using the fact that \( \mathbf{D} = \frac{1}{4}\mathbf{M}_-\mathbf{M}_+^T + \frac{1}{4}\mathbf{M}_+\mathbf{M}_+^T \), \( \mathbf{A} = \frac{1}{4}\mathbf{M}_+\mathbf{M}_+^T - \frac{1}{4}\mathbf{M}_-\mathbf{M}_+^T \) as well as Eq. (122), we can re-write Eq. (117) as

\[
\begin{align*}
\nabla f(\theta^{k+1}) + \mathbf{M}_-\mathbf{\beta}^{k+1} - \rho \mathbf{C}\theta^k + \rho \mathbf{C}\mathbf{E}^{k+1} + \rho \left( \mathbf{C}^T - \frac{1}{2}\mathbf{M}_-\mathbf{M}_+^T \right)\mathbf{E}^{k+1} \\
+ \rho \left( \mathbf{A} - \mathbf{C}^T \right)\theta^{k+1} &= 0.
\end{align*}
\]
Using that $\nabla f(\theta^*) + M_\beta = 0$ and $A = C + C^T$, we can write
\[
\nabla f(\theta^{k+1}) - \nabla f(\theta^*) = M_\beta (\beta^* - \beta^{k+1}) + \rho C \left( \theta^k - \theta^{k+1} \right) - \rho C E^k + \rho \left( \frac{1}{2} M_\beta M_\beta^T - C^T \right) E^{k+1}, \tag{124}
\]
then, multiplying both sides by $\theta^{k+1} - \theta^*$, we get
\[
\mathbb{E} \left[ \langle \nabla f(\theta^{k+1}) - \nabla f(\theta^*), \theta^{k+1} - \theta^* \rangle \right] = \mathbb{E} \left[ \langle M_\beta (\beta^* - \beta^{k+1}), \theta^{k+1} - \theta^* \rangle \right] + \rho \mathbb{E} \left[ \langle C \left( \theta^k - \theta^{k+1} \right), \theta^{k+1} - \theta^* \rangle \right] - \rho \mathbb{E} \left[ \langle C E^k, \theta^{k+1} - \theta^* \rangle \right] - \rho \mathbb{E} \left[ \langle C^T E^{k+1}, \theta^{k+1} - \theta^* \rangle \right] + \frac{\rho}{2} \mathbb{E} \left[ \langle M_\beta M_\beta^T E^{k+1}, \theta^{k+1} - \theta^* \rangle \right]. \tag{125}
\]
The first term of the right hand side can be re-written as
\[
\mathbb{E} \left[ \langle M_\beta (\beta^* - \beta^{k+1}), \theta^{k+1} - \theta^* \rangle \right] = \mathbb{E} \left[ \langle \beta^* - \beta^{k+1}, M_\beta^T (\theta^{k+1} - \theta^*) \rangle \right] \overset{(a)}{=} \mathbb{E} \left[ \langle \beta^* - \beta^{k+1}, M_\beta^T \theta^{k+1} \rangle \right] \overset{(b)}{=} -\frac{2}{\rho} \mathbb{E} \left[ \langle \beta^{k+1} - \beta^*, \beta^{k+1} - \beta^k \rangle \right] - \mathbb{E} \left[ \langle \beta^* - \beta^{k+1}, M_\beta^T E^{k+1} \rangle \right], \tag{126}
\]
where we have used $M_\beta^T \theta^* = 0$ in (a) and $M_\beta^T \theta^{k+1} = \frac{2}{\rho} (\beta^{k+1} - \beta^k) - M_\beta^T E^{k+1}$ in (b). Using the identity (44), we can write
\[
-\frac{2}{\rho} \mathbb{E} \left[ \langle \beta^{k+1} - \beta^*, \beta^{k+1} - \beta^k \rangle \right] = \frac{1}{\rho} \mathbb{E} \left[ \| \beta^k - \beta^* \|^2 \right] - \frac{1}{\rho} \mathbb{E} \left[ \| \beta^{k+1} - \beta^* \|^2 \right] - \frac{1}{\rho} \mathbb{E} \left[ \| \beta^{k+1} - \beta^k \|^2 \right]. \tag{127}
\]
Replacing the terms derived in (126) and (127) by their expressions in Eq. (125), we obtain
\[
\mathbb{E} \left[ \langle \nabla f(\theta^{k+1}) - \nabla f(\theta^*), \theta^{k+1} - \theta^* \rangle \right] = \frac{1}{\rho} \mathbb{E} \left[ \| \beta^k - \beta^* \|^2 \right] - \frac{1}{\rho} \mathbb{E} \left[ \| \beta^{k+1} - \beta^* \|^2 \right] - \frac{1}{\rho} \mathbb{E} \left[ \| \beta^{k+1} - \beta^k \|^2 \right] + \mathbb{E} \left[ \langle \beta^{k+1} - \beta^*, M_\beta^T E^{k+1} \rangle \right] - \rho \mathbb{E} \left[ \langle C E^k, \theta^{k+1} - \theta^* \rangle \right] - \rho \mathbb{E} \left[ \langle C^T E^{k+1}, \theta^{k+1} - \theta^* \rangle \right] + \frac{\rho}{2} \mathbb{E} \left[ \langle M_\beta M_\beta^T E^{k+1}, \theta^{k+1} - \theta^* \rangle \right]. \tag{128}
\]
Using the strong convexity of the function $f$, we can lower bound the left hand side of Eq. (125) as
\[
\mathbb{E} \left[ \langle \nabla f(\theta^{k+1}) - \nabla f(\theta^*), \theta^{k+1} - \theta^* \rangle \right] \geq \mu \mathbb{E} \left[ \| \theta^{k+1} - \theta^* \|^2 \right]. \tag{129}
\]
Hence, we can write
\[
\frac{1}{\rho} \mathbb{E} \left[ \left\| \beta^{k+1} - \beta^* \right\|_F^2 + \mu \left\| \theta^{k+1} - \theta^* \right\|_F^2 \right] \\
\leq \frac{1}{\rho} \mathbb{E} \left[ \left\| \beta^k - \beta^* \right\|_F^2 \right] + \rho \mathbb{E} \left[ \left( \theta^k - \theta^* \right), \theta^{k+1} - \theta^* \right] + \frac{\rho \mathbb{E} \left[ \left( \theta^k - \theta^{k+1} \right), \theta^{k+1} - \theta^* \right]}{\rho} \\
- \frac{1}{\rho} \mathbb{E} \left[ \left\| \beta^{k+1} - \beta^* \right\|_F^2 \right] + \mathbb{E} \left[ \left( \beta^{k+1} - \beta^*, M^T E^{k+1} \right) \right] - \rho \mathbb{E} \left[ \left( \beta^k, \theta^{k+1} - \theta^* \right) \right] \\
- \rho \mathbb{E} \left[ \left( C^T E^{k+1}, \theta^{k+1} - \theta^* \right) \right] + \frac{\rho}{2} \mathbb{E} \left[ \left( M_+ M^T E^{k+1}, \theta^{k+1} - \theta^* \right) \right].
\] (130)

Now, using identities (46) and (47), we get the following bounds
\[
\mathbb{E} \left[ \left( \theta^k - \theta^* \right), \theta^{k+1} - \theta^* \right] \leq \frac{\eta_1}{2} \sigma^2_{\max}(C) \mathbb{E} \left[ \left\| \theta^k - \theta^* \right\|_F^2 \right] + \frac{1}{2 \eta_1} \mathbb{E} \left[ \left\| \theta^{k+1} - \theta^* \right\|_F^2 \right],
\] (131)

\[
\mathbb{E} \left[ \left( \beta^{k+1} - \beta^*, M^T E^{k+1} \right) \right] \leq \frac{\eta_2}{2} \mathbb{E} \left[ \left\| \beta^{k+1} - \beta^* \right\|_F^2 \right] + \frac{\sigma^2_{\max}(M_-)}{2 \eta_2} \mathbb{E} \left[ \left\| E^{k+1} \right\|_F^2 \right],
\] (133)

\[
\mathbb{E} \left[ \left( \beta^{k+1} - \beta^*, M^T E^{k+1} \right) \right] \leq \frac{\eta_3}{2} \mathbb{E} \left[ \left\| \beta^{k+1} - \beta^* \right\|_F^2 \right] + \frac{\sigma^2_{\max}(C)}{2 \eta_3} \mathbb{E} \left[ \left\| E^k \right\|_F^2 \right],
\] (134)

\[
\mathbb{E} \left[ \left( C^T E^{k+1}, \theta^k - \theta^* \right) \right] \leq \frac{\eta_4}{2} \mathbb{E} \left[ \left\| \theta^{k+1} - \theta^* \right\|_F^2 \right] + \frac{\sigma^2_{\max}(C)}{2 \eta_4} \mathbb{E} \left[ \left\| E^{k+1} \right\|_F^2 \right],
\] (135)

Replacing the bounds derived in (131)-(136) in (130) and introducing \( \kappa > 0 \), we get
\[
(1 + \kappa) \frac{1}{\rho} \mathbb{E} \left[ \left\| \beta^{k+1} - \beta^* \right\|_F^2 \right] + \mu \mathbb{E} \left[ \left\| \theta^{k+1} - \theta^* \right\|_F^2 \right]
\leq \frac{1}{\rho} \mathbb{E} \left[ \left\| \beta^k - \beta^* \right\|_F^2 \right] + \frac{\rho}{2 \eta_1} \sigma^2_{\max}(C) \mathbb{E} \left[ \left\| \theta^k - \theta^* \right\|_F^2 \right] + \rho \left( \frac{1}{2 \eta_1} + \frac{\eta_3}{2} + \frac{\eta_4}{2} + \frac{\eta_5}{4} \right) \mathbb{E} \left[ \left\| \theta^{k+1} - \theta^* \right\|_F^2 \right]
+ \left( \frac{\eta_2}{2} + \frac{\kappa}{\rho} \right) \mathbb{E} \left[ \left\| \beta^{k+1} - \beta^* \right\|_F^2 \right] + \rho \left( \frac{\eta_0}{2} \sigma^2_{\max}(C) + \frac{1}{2 \eta_0} \mathbb{E} \left[ \left\| \theta^{k+1} - \theta^* \right\|_F^2 \right]ight)
+ \frac{\rho}{2 \eta_3} \sigma^2_{\max}(C) \mathbb{E} \left[ \left\| E^k \right\|_F^2 \right] + \left( \frac{\sigma^2_{\max}(M_-)}{2 \eta_2} + \frac{\rho}{2 \eta_4} \sigma^2_{\max}(C) + \frac{\rho}{4 \eta_5} \sigma^4_{\max}(M_-) \right) \mathbb{E} \left[ \left\| E^{k+1} \right\|_F^2 \right].
\] (137)
Using that \( \|E^{k+1}\|_F^2 \leq \|E^k\|_F^2 \), and re-arranging the terms, we can further write
\[
\frac{1}{\rho} \mathbb{E} \left[ \|\beta^k - \beta^*\|_F^2 \right] - \frac{1 + \kappa}{\rho} \mathbb{E} \left[ \|\beta^{k+1} - \beta^*\|_F^2 \right] + \frac{\eta_1 \sigma_{\text{max}}^2(C)}{2} \mathbb{E} \left[ \|\theta^k - \theta^*\|_F^2 \right] + \left( \frac{\eta_2}{2} + \frac{\kappa}{\rho} \right) \mathbb{E} \left[ \|\beta^{k+1} - \beta^*\|_F^2 \right] \\
- \left[ \mu - \left( \frac{\eta_0 \sigma_{\text{max}}^2(C)}{2} + \frac{1}{2\eta_0} \right) \mathbb{E} \left[ \|\theta^{k+1} - \theta^*\|_F^2 \right] + \gamma \mathbb{E} \left[ \|E^k\|_F^2 \right] \geq 0, \tag{138} \right.
\]
where \( \gamma = \frac{\sigma_{\text{max}}^2(M_-)}{2\eta_2} + \frac{\kappa}{2} \left( \frac{1}{\eta_4} + \frac{1}{\eta_3} \right) \sigma_{\text{max}}^2(C) + \frac{\eta}{4\eta_0} \sigma_{\text{max}}^2(M_-) > 0 \). Now, we choose to fix \( \eta_2 = \frac{2\kappa}{\rho} \) to get
\[
\frac{1}{\rho} \mathbb{E} \left[ \|\beta^k - \beta^*\|_F^2 \right] - \left( 1 + \kappa \right) \frac{1}{\rho} \mathbb{E} \left[ \|\beta^{k+1} - \beta^*\|_F^2 \right] + \frac{\eta_1 \sigma_{\text{max}}^2(C)}{2} \mathbb{E} \left[ \|\theta^k - \theta^*\|_F^2 \right] + \frac{2\kappa}{\rho} \mathbb{E} \left[ \|\beta^{k+1} - \beta^*\|_F^2 \right] \\
- \left[ \mu - \left( \frac{\eta_0 \sigma_{\text{max}}^2(C)}{2} + \frac{1}{2\eta_0} \right) \mathbb{E} \left[ \|\theta^{k+1} - \theta^*\|_F^2 \right] + \gamma \mathbb{E} \left[ \|E^k\|_F^2 \right] \geq 0. \tag{139} \right.
\]
In order to bound the term \( \mathbb{E} \left[ \|\beta^{k+1} - \beta^*\|_F^2 \right] \) in the left hand side, we use Eq. (124) to write
\[
\mathbb{E} \left[ \|\beta^{k+1} - \beta^*\|_F^2 \right] = \mathbb{E} \left[ \|\nabla f(\theta^{k+1}) - \nabla f(\theta^*) + \rho C \left( \theta^{k+1} - \theta^k \right) + \rho C E^k + \rho \left( C^T - \frac{1}{2} M_- M^T \right) E^{k+1} \|_F^2 \right]. \tag{140} \]
Using identity (42), we can further write
\[
\mathbb{E} \left[ \|\beta^{k+1} - \beta^*\|_F^2 \right] \leq 2 \mathbb{E} \left[ \|\nabla f(\theta^{k+1}) - \nabla f(\theta^*) + \rho C \left( \theta^{k+1} - \theta^k \right) \|_F^2 \right] + 2 \mathbb{E} \left[ \|\rho C E^k + \rho \left( C^T - \frac{1}{2} M_- M^T \right) E^{k+1} \|_F^2 \right]. \tag{141} \]
Using identity (48) for the first term and identity (42) for the second term of the right hand side, we get
\[
\mathbb{E} \left[ \|\beta^{k+1} - \beta^*\|_F^2 \right] \leq 2\eta \mathbb{E} \left[ \|\nabla f(\theta^{k+1}) - \nabla f(\theta^*)\|_F^2 \right] + \frac{2\eta}{\eta - 1} \mathbb{E} \left[ \|\rho C \left( \theta^{k+1} - \theta^k \right) \|_F^2 \right] \\
+ 4 \mathbb{E} \left[ \|\rho C E^k\|_F^2 \right] + 4 \mathbb{E} \left[ \|\rho \left( C^T - \frac{1}{2} M_- M^T \right) E^{k+1}\|_F^2 \right]. \tag{142} \]
On one hand, since both \( \beta^{k+1} \) and \( \beta^* \) belong to the columns space of \( M_- \), we have
\[
\mathbb{E} \left[ \|\beta^{k+1} - \beta^*\|_F^2 \right] \geq \sigma_{\text{min}}^2(M_-) \mathbb{E} \left[ \|\beta^{k+1} - \beta^*\|_F^2 \right], \tag{143} \]
where \( \hat{\sigma}_{\text{min}}(M_-) \) is the minimum non-zero singular value of \( M_- \). On the other hand, from Assumption 5, we have

\[
\mathbb{E} \left[ \| \nabla f(\theta^{k+1}) - \nabla f(\theta^*) \|_F \right] \leq L \mathbb{E} \left[ \| \theta^{k+1} - \theta^* \|_F \right].
\]  

(144)

Therefore, we get the following upper bound

\[
\mathbb{E} \left[ \| \beta^{k+1} - \beta^* \|_F^2 \right] \\ \leq \frac{2\eta}{\hat{\sigma}_{\text{min}}^2(M_-)} \left( L^2 + \frac{2\rho^2}{\eta - 1} \sigma_{\text{max}}^2(C) \right) \mathbb{E} \left[ \| \theta^{k+1} - \theta^* \|_F^2 \right] + \frac{4\eta\rho^2\sigma_{\text{max}}^2(C)}{(\eta - 1)\hat{\sigma}_{\text{min}}^2(M_-)} \mathbb{E} \left[ \| \theta^k - \theta^* \|_F^2 \right] \\
+ \frac{16N\rho^2}{\hat{\sigma}_{\text{min}}^2(M_-)} \left( \sigma_{\text{max}}^2(C) + \sigma_{\text{max}}^2(C^T - \frac{1}{2}M_-M_-^T) \right) \psi^{2k},
\]  

(145)

where we have used that \( \mathbb{E} \left[ \| E^{k+1} \|_F^2 \right] \leq \mathbb{E} \left[ \| E^k \|_F^2 \right] \leq 4C_0^2N\psi^{2k} \). Plugging the bound obtained in Eq. (145) in Eq. (139) we get

\[
\frac{1}{\rho} \mathbb{E} \left[ \| \beta^k - \beta^* \|_F^2 \right] - (1 + \kappa) \frac{1}{\rho} \mathbb{E} \left[ \| \beta^{k+1} - \beta^* \|_F^2 \right] + (b_1 + a\kappa)\rho \mathbb{E} \left[ \| \theta^k - \theta^* \|_F^2 \right] \\
- \left( \mu - \frac{c\kappa}{\rho} - (b_2 + a\kappa)\rho \right) \mathbb{E} \left[ \| \theta^{k+1} - \theta^* \|_F^2 \right] + \nu\psi^{2k} \geq 0,
\]  

(146)

where \( b_1 = \frac{m\sigma_{\text{max}}^2(C)}{2}, b_2 = \frac{m_0\sigma_{\text{max}}^2(C)}{2} + \frac{1}{2m} + \frac{1}{2m} + \frac{y_2}{2} + \frac{y_3}{2} + \frac{y_5}{2}, c = \frac{4\eta L^2}{\hat{\sigma}_{\text{min}}^2(M_-)}, \) \( a = \frac{8\eta\rho\sigma_{\text{max}}^2(C)}{(\eta - 1)\hat{\sigma}_{\text{min}}^2(M_-)}, \) \( \nu = 4N\gamma + \frac{32N\rho\kappa}{\hat{\sigma}_{\text{min}}^2(M_-)} \left( \sigma_{\text{max}}^2(C) + \sigma_{\text{max}}^2(C^T - \frac{1}{2}M_-M_-^T) \right) \).

To ensure that there is a decrease in the optimality gap, we need to determine, for which values of \( \rho \), we have \( c - b_2\rho - a\rho^2 > 0 \). We also want to ensure that

\[
\mu - \frac{c\kappa}{\rho} - (b_2 + a\kappa)\rho \geq (1 + \kappa)(b_1 + a\kappa)\rho > 0.
\]  

(147)

In other words, we need to look for \( \rho \) such that

\[
- [(b_2 + a\kappa) + (1 + \kappa)(b_1 + a\kappa)]\rho^2 + \mu\rho - c\kappa \geq 0.
\]  

(148)

We start by computing the discriminant of the quadratic equation as

\[
\Delta = \mu^2 - 4c\kappa [(b_2 + a\kappa) + (1 + \kappa)(b_1 + a\kappa)].
\]  

(149)

To ensure that we can find \( \rho \) such that Eq. (148) is satisfied, we need to impose that \( \Delta > 0 \). Since Eq. (149) is a third order equation in \( \kappa \), finding for which values of \( \kappa > 0 \) the discriminant \( \Delta \) is positive is not straightforward. However, since when \( \kappa \rightarrow 0, \Delta \rightarrow \mu^2 > 0 \), and knowing that \( \Delta \) is a decreasing function with \( \Delta \rightarrow -\infty \) as \( \kappa \rightarrow \infty \), then we deduce that there exits \( \bar{\kappa} > 0 \) such that for \( 0 < \kappa < \bar{\kappa} \), we have \( \Delta > 0 \). In the rest of the proof, we consider \( \kappa \) such that \( 0 < \kappa < \bar{\kappa} \). Under this condition, we can ensure that for \( 0 < \rho < \bar{\rho} \), Eq. (148) holds where \( \bar{\rho} \) is given by

\[
\bar{\rho} = \frac{\mu + \sqrt{\Delta}}{(b_2 + a\kappa) + (1 + \kappa)(b_1 + a\kappa)}.
\]  

(150)
Therefore, going back to Eq. (146), we can write
\[ \frac{1}{\rho} \mathbb{E} \left[ \| \beta^k - \beta^* \|^2_F \right] - (1 + \kappa) \frac{1}{\rho} \mathbb{E} \left[ \| \beta^{k+1} - \beta^* \|^2_F \right] + \rho (b_1 + a\kappa) \mathbb{E} \left[ \| \theta^k - \theta^* \|^2_F \right] \\
- \rho (1 + \kappa) (b_1 + a\kappa) \mathbb{E} \left[ \| \theta^{k+1} - \theta^* \|^2_F \right] + \nu \psi^{2k} \geq 0. \] (151)

Re-arranging the terms, we get
\[ \frac{1}{\rho} \mathbb{E} \left[ \| \beta^{k+1} - \beta^* \|^2_F \right] + \rho (b_1 + a\kappa) \mathbb{E} \left[ \| \theta^{k+1} - \theta^* \|^2_F \right] \\
\leq \frac{1}{1 + \kappa} \left( \frac{1}{\rho} \mathbb{E} \left[ \| \beta^k - \beta^* \|^2_F \right] + \rho (b_1 + a\kappa) \mathbb{E} \left[ \| \theta^k - \theta^* \|^2_F \right] \right) + \frac{\nu}{1 + \kappa} \psi^{2k}. \] (152)

Using this equation iteratively, we can write
\[ \frac{1}{\rho} \mathbb{E} \left[ \| \beta^{k+1} - \beta^* \|^2_F \right] + \rho (b_1 + a\kappa) \mathbb{E} \left[ \| \theta^{k+1} - \theta^* \|^2_F \right] \\
\leq \left( \frac{1}{1 + \kappa} \right)^{k+1} \left( \frac{1}{\rho} \mathbb{E} \left[ \| \beta^0 - \beta^* \|^2_F \right] + \rho (b_1 + a\kappa) \mathbb{E} \left[ \| \theta^0 - \theta^* \|^2_F \right] \right) + \nu \sum_{j=0}^k \left( \frac{1}{1 + \kappa} \right)^{k-j+1} \psi^{2j}. \] (153)

Introducing the two constants
\[ \delta_1 = \min\{(1 + \kappa)^{-1}, \psi^2\}, \quad \delta_2 = \max\{(1 + \kappa)^{-1}, \psi^2\} \] (154)
we can further write
\[ \frac{1}{\rho} \mathbb{E} \left[ \| \beta^{k+1} - \beta^* \|^2_F \right] + \rho (b_1 + a\kappa) \mathbb{E} \left[ \| \theta^{k+1} - \theta^* \|^2_F \right] \]
\[ \leq (1 + \delta_2)^{k+1} \left( \frac{1}{\rho} \mathbb{E} \left[ \| \beta^0 - \beta^* \|^2_F \right] + \rho (b_1 + a\kappa) \mathbb{E} \left[ \| \theta^0 - \theta^* \|^2_F \right] \right) + \nu \sum_{j=0}^k \left( \frac{1 + \delta_2}{2} \right)^{k-j+1} \delta_1^j \] (a)
\[ \leq \left( \frac{1 + \delta_2}{2} \right)^{k+1} \left( \frac{1}{\rho} \mathbb{E} \left[ \| \beta^0 - \beta^* \|^2_F \right] + \rho (b_1 + a\kappa) \mathbb{E} \left[ \| \theta^0 - \theta^* \|^2_F \right] \right) + \nu \left( \frac{1 + \delta_2}{2} \right)^{k+1} \sum_{j=0}^k \left( \frac{2\delta_1}{1 + \delta_2} \right)^j \] (b)
\[ \leq \left( \frac{1 + \delta_2}{2} \right)^{k+1} \left( \frac{1}{\rho} \mathbb{E} \left[ \| \beta^0 - \beta^* \|^2_F \right] + \rho (b_1 + a\kappa) \mathbb{E} \left[ \| \theta^0 - \theta^* \|^2_F \right] + \frac{\nu(1 + \delta_2)}{1 + \delta_2 - 2\delta_1} \right), \] (155)

where we have used in (a) the fact that \( \delta_2 \leq (1 + \delta_2)/2 \) since \( \kappa > 0 \) and \( \psi \in (0, 1) \) and \((2\delta_1)/(1 + \delta_2) \in (0, 1) \) in (b). Since \((1 + \delta_2)/2 \in (0, 1)\), then we deduce that the sequence \((\theta^k, \beta^k)\) converges to \((\theta^*, \beta^*)\) at a linear rate. Equivalently, we can write
\[ \| \theta^{k+1} - \theta^* \|^2_F \leq \left( \frac{1 + \delta_2}{2} \right)^{k+1} \left( \| \theta^0 - \theta^* \|^2_F + C_1 \right), \] (156)

where the constant \( C_1 \) is given by
\[ C_1 = \frac{\| \beta^0 - \beta^* \|^2_F}{\rho^2 (b_1 + a\kappa)} + \frac{\nu (1 + \delta_2)}{\rho (b_1 + a\kappa) (1 + \delta_2 - 2\delta_1)}. \] (157)
References

Jin-Hyun Ahn, Osvaldo Simeone, and Joonhyuk Kang. Cooperative learning via federated distillation over fading channels. *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain*, 2020.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 560–569, Stockholm Sweden, 2018.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Mingzhe Chen, Ursula Challita, Walid Saad, Changchuan Yin, and Mérouane Debbah. Machine learning for wireless networks with artificial intelligence: A tutorial on neural networks. *IEEE Communications Surveys & Tutorials*, 21(4):3039–3071, 2019.

Tianyi Chen, Georgios Giannakis, Tao Sun, and Wotao Yin. LAG: Lazily aggregated gradient for communication-efficient distributed learning. In *Advances in Neural Information Processing Systems 31*, pages 5050–5060, 2018.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Anis Elgabli, Jihong Park, Sabbir Ahmed, and Mehdi Bennis. L-FGADMM: Layer-wise federated group ADMM for communication efficient decentralized deep learning. *Proc. IEEE WCNC, Seoul, Korea*, 2020a.

Anis Elgabli, Jihong Park, Amrit S. Bedi, Chaouki Ben Issaid, Mehdi Bennis, and Vaneet Aggarwal. Q-GADMM: Quantized group ADMM for communication efficient decentralized machine learning. to appear in IEEE Transactions on Communications, 2020b.

Anis Elgabli, Jihong Park, Amrit S. Bedi, Mehdi Bennis, and Vaneet Aggarwal. GADMM: Fast and communication efficient framework for distributed machine learning. *Journal of Machine Learning Research*, 21(76):1–39, 2020c.

Anis Elgabli, Jihong Park, Chaouki Ben Issaid, and Mehdi Bennis. Harnessing wireless channels for scalable and privacy-preserving federated learning. ArXiv preprint, arXiv:2007.01790, 2020d.

Hongchang Gao and Heng Huang. Adaptive serverless learning. ArXiv preprint, arXiv:2008.10422, 2020.
Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. In Advances in Neural Information Processing Systems, pages 9633–9643, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In Proceedings of NeurIPS Workshop on Deep Learning, pages 1–9, Montréal, Canada, December 2014.

Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Sebastian Stich, and Peter Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. ArXiv preprint, arXiv:1904.05115, 2019.

Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-IID private data. Neural Information Processing Systems Workshop on Machine Learning on the Phone and other Consumer Devices (MLPCD), Montréal, Canada, 2018.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. ArXiv preprint, arXiv:1912.04977, 2019.

H. Kim, J. Park, M. Bennis, and S. Kim. Blockchained on-device federated learning. IEEE Communications Letters, 24(6):1279–1283, 2020.

Anastasia Koloskova, Tao Lin, Sebastian U. Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. ArXiv preprint, arXiv:1907.09356, 2019.

Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. ArXiv preprint, arXiv:1610.02527, 2016.

Wei Liu, Li Chen, Yunfei Chen, and Wenyi Zhang. Accelerating federated learning via momentum gradient descent. ArXiv preprint, arXiv:1910.03197, 2019a.

Yaohua Liu, Wei Xu, Gang Wu, Zhi Tian, and Qing Ling. Communication-censored ADMM for decentralized consensus optimization. IEEE Transactions on Signal Processing, 67 (10):2565–2579, 2019b.

H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In Proc. of AISTATS, Fort Lauderdale, FL, USA, April 2017.

Konstantin Mishchenko, Filip Hanzely, and Peter Richtarik. 99% of worker-master communication in distributed optimization is not needed. volume 124 of Proceedings of Machine Learning Research, pages 979–988, Virtual, Aug 2020.
Seungeun Oh, Jihong Park, Eunjeong Jeong, Hyesung Kim, Mehdi Bennis, and Seong-Lyun Kim. Mix2FLD: Downlink federated learning after uplink federated distillation with two-way mixup. to appear in IEEE Communications Letters., 2020.

Jihong Park, Sumudu Samarakoon, Mehdi Bennis, and Mérouane Debbah. Wireless network intelligence at the edge. Proceedings of the IEEE, 107(11):2204–2239, October 2019a.

Jihong Park, Shiqiang Wang, Anis Elgabli, Seungeun Oh, Eunjeong Jeong, Han Cha, Hyesung Kim, Seong-Lyun Kim, and Mehdi Bennis. Distilling on-device intelligence at the network edge. ArXiv preprint, arXiv:1908.05895, 2019b.

Jihong Park, Sumudu Samarakoon, Anis Elgabli, Joongheon Kim, Mehdi Bennis, Seong-Lyun Kim, and Mérouane Debbah. Communication-efficient and distributed learning over wireless networks: Principles and applications. ArXiv preprint, arXiv:2008.02608, 2020a.

Jihong Park, Sumudu Samarakoon, Hamid Shiri, Mohamed K Abdel-Aziz, Takayuki Nishio, Anis Elgabli, and Mehdi Bennis. Extreme URLLC: Vision, challenges, and key enablers. ArXiv preprint, arXiv:2001.09683, 2020b.

Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the admm in decentralized consensus optimization. IEEE Transactions on Signal Processing, 62(7):1750–1761, 2014.

Abhishek Singh, Praneeth Vepakomma, Otkrist Gupta, and Ramesh Raskar. Detailed comparison of communication efficiency of split learning and federated learning. ArXiv preprint, arXiv:1909.09145, 2019a.

Navjot Singh, Deepesh Data, Jemin George, and Suhas Diggavi. SPARQ-SGD: Event-triggered and compressed communication in decentralized stochastic optimization. ArXiv preprint, arXiv:1910.14280, 2019b.

Nandan Sriranga, Chandra R. Murthy, and Vaneet Aggarwal. A method to improve consensus averaging using quantized ADMM. In 2019 IEEE International Symposium on Information Theory (ISIT). IEEE, 2019.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In Advances in Neural Information Processing Systems 31, pages 4447–4458. 2018.

Jun Sun, Tianyi Chen, Georgios B. Giannakis, and Zaiyue Yang. Communication-efficient distributed learning via lazily aggregated quantized gradients. Proc. NeurIPS, Vancouver, Canada, December 2019.

Ananda Theertha Suresh, Felix X Yu, Sanjiv Kumar, and H Brendan McMahan. Distributed mean estimation with limited communication. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3329–3337, 2017.

University of Oulu. 6G Flagship. [online, Accessed: 2018-12-04]. http://www.oulu.fi/6gflagship.
Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems*, pages 14259–14268, 2019.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pages 1299–1309, 2018.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. ArXiv preprint, arXiv:1905.03817, 2019.

Shengyu Zhu, Mingyi Hong, and Biao Chen. Quantized consensus ADMM for multi-agent distributed optimization. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Shanghai, China*, March 2016.