# Covariance Selection over Networks

**Wenfu Xia**
ShanghaiTech

**Fengpei Li**
ShanghaiTech

**Ying Sun**
Penn State

**Ziping Zhao**
ShanghaiTech

## Abstract

Covariance matrix estimation is a fundamental problem in multivariate data analysis, which becomes particularly challenging in high-dimensional settings due to the curse of dimensionality. To enhance estimation accuracy, structural regularization is often imposed on the precision matrix (the inverse covariance matrix) for covariance selection. In this paper, we study covariance selection in a distributed setting, where data is spread across a network of agents. We formulate the problem as a Gaussian maximum likelihood estimation problem with structural penalties and propose a novel algorithmic framework called NetGGM. Unlike existing methods that rely on a central coordinator, NetGGM operates in a fully decentralized manner with low computational complexity. We provide theoretical guarantees showing that NetGGM converges linearly to the global optimum while ensuring consensus among agents. Numerical experiments validate its convergence properties and demonstrate that it outperforms state-of-the-art methods in precision matrix estimation.

## 1 INTRODUCTION

Estimating the covariance matrix is a fundamental task across various fields related to data analysis, including machine learning (Jolliffe, 2002), finance (Markowitz, 1952), and biology (Schäfer and Strimmer, 2005). In high-dimensional settings, where the number of samples is comparable to or smaller than the number of dimensions, accurate estimation of the covariance matrix becomes particularly challenging. A widely adopted approach to address this issue is covariance selection (Dempster, 1972), which imposes structural constraints on the precision matrix (the inverse covariance matrix) to improve estimation accuracy.

Under the Gaussian assumption, the off-diagonal elements of the precision matrix represent conditional dependencies between variables, aligning precision estimation with the construction of Gaussian graphical models (GGM) (Lauritzen, 1996; Li and Zhao, 2024). A commonly used structural assumption in GGM is sparsity, where many off-diagonal elements in the precision matrix are zero, leading to sparse GGM. This assumption reduces the number of parameters to estimate, thereby improving accuracy. A well-established approach for estimating sparse GGM is the $\ell_1$-penalized Gaussian maximum likelihood estimation (Banerjee et al., 2008; Friedman et al., 2008). Beyond the lasso penalty (Tibshirani, 1996), various regularization techniques, including group lasso (Marlin and Murphy, 2009), smoothly clipped absolute deviation (Fan et al., 2009), ridge penalty (Kuismin et al., 2017), elastic net (Ryali et al., 2012), multivariate total positivity of order 2 ($\text{MTP}_2$) (Karlin and Rinott, 1983), and sorted $\ell_1$-penalized estimation (SLOPE) (Mazza-Anthony et al., 2020) have also been applied within the maximum likelihood estimation framework.

In modern applications, data is often distributed across multiple agents, such as geographically dispersed sensors, satellites in different orbits, or institutions spanning multiple continents. Due to constraints like communication and storage overhead, privacy concerns, and regulatory policies, transferring distributed data to a central processor can be inefficient or even infeasible. For example, in disease detection studies, hospitals may seek to collaborate to increase the sample size, but privacy regulations prevent the direct sharing of the electronic health records of patients (Warnat-Herresthal et al., 2021). This highlights the urgent need for decentralized estimation methods (Maros and Scutari, 2022; Ji et al., 2023; Xia et al., 2024) that enable network-wide analysis while preserving data privacy. This paper investigates the problem of precision matrix estimation over a network, where samples are distributed across multiple agents, each observing all variables. Several studies

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. 

have explored precision matrix estimation over distributed samples (Arroyo and Hou, 2016; Nezakati and Pircalabelu, 2023; Wang and Cui, 2021; Dong and Liu, 2024). However, these methods either rely on divide-and-conquer strategies or require a central node to iteratively aggregate information, preventing them from operating in a fully decentralized manner.

In this paper, we address the problem of high-dimensional decentralized GGM estimation, i.e., structured precision matrix estimation. This problem presents several challenges. A primary difficulty in distributed estimation is that each agent lacks access to data from other agents. Additionally, although the problem is convex, the likelihood function is neither Lipschitz smooth nor strongly convex over the set of positive definite matrices, further complicating the estimation process. The main contributions of this paper are summarized as follows:

- We propose NetGGM, a decentralized single-loop algorithm based on proximal gradient descent (Beck, 2017) and gradient tracking techniques (Di Lorenzo and Scutari, 2016), to solve the structured GGM estimation problem. In each iteration, each agent updates its local estimate via a lightweight soft-thresholding operation while integrating information from neighboring agents. Compared to existing distributed precision matrix estimation methods (Arroyo and Hou, 2016; Wang and Cui, 2021; Nezakati and Pircalabelu, 2023; Dong and Liu, 2024), NetGGM eliminates the reliance on a central node and provides theoretical guarantees of positive definiteness.
- We prove that NetGGM achieves both linear convergence and consensus starting from any positive definite initialization. This resolves a key challenge in decentralized settings, where agents lack global information to set an initialization within a specific range, as required in centralized proximal gradient methods for GGM (Rolfs et al., 2012). Furthermore, by carefully selecting the algorithm parameters, we ensure that the local estimates remain positive definite at each iteration. This is achieved without the need to explicitly enforce a positive definiteness constraint, which would otherwise lead to proximal steps that lack closed-form solutions.
- We discuss the applicability of NetGGM to a broader class of GGM problems and demonstrate that it maintains linear convergence and consensus, even when the proximal steps only have inexact solutions.
- Numerical experiments confirm the linear convergence and consensus of NetGGM and demonstrate its superior performance compared to state-of-the-art methods in precision matrix estimation.

## 2 RELATED WORK

**GGM estimation.** Estimation methods for sparse GGMs have been developed for both primal and dual formulations. For the primal problems, several algorithms have been proposed, including interior point methods (Yuan and Lin, 2007), alternating-direction methods (Yuan, 2012), and block coordinate descent methods (Mazumder and Agarwal, 2011; Mazumder and Hastie, 2012). However, many of these approaches lack theoretical convergence guarantees. Scheinberg et al. (2010) introduced a method based on an alternating linearization technique and established sublinear convergence. The QUIC algorithm, a second-order proximal point method, was proposed by Hsieh et al. (2011) and exhibits local superlinear convergence. The most relevant approach to our work is the graphical iterative shrinkage thresholding algorithm (G-ISTA) (Rolfs et al., 2012), which is based on the proximal gradient descent framework (Beck, 2017). It has been proven to converge linearly to the global optimum, provided that the initialization lies within a specified set. For dual methods, block coordinate descent algorithms (Banerjee et al., 2008; Friedman et al., 2008), Nesterov's smooth approximation framework (d'Aspremont et al., 2008; Lu, 2009, 2010), and proximal gradient descent (Dalal and Rajaratnam, 2017) have been employed. Additionally, from a primal-dual perspective, the alternating direction method of multipliers (ADMM) has also been applied (Boyd et al., 2011). For a comprehensive review, see (Chen, 2024). Furthermore, many studies have extended these algorithms to other GGM models. For instance, block coordinate descent methods (Slawski and Hein, 2015; Lauritzen et al., 2019) and quasi-Newton methods (Cai et al., 2024) have been employed to estimate GGMs under $\text{MTP}_2$ constraints, while ADMM has been adapted to GGMs with SLOPE regularization (Defazio and Caetano, 2012; Mazza-Anthony et al., 2020).

**Distributed GGM estimation.** In cases where samples are distributed across multiple agents, research on decentralized methods for precision matrix estimation remains limited. In contrast, when a network includes a centralized node, a common approach is the divide-and-conquer strategy, where each agent computes a local estimate from its data and sends it to the central node for aggregation. In Arroyo and Hou (2016), each agent applies the debiased $\ell_1$-penalized maximum likelihood estimation (Janková and van de Geer, 2015) to obtain local estimates, which are then

averaged by the central node. Nezakati and Pircalabelu (2023) extended this approach by introducing an element-wise weighted average of local estimates to account for varying sample sizes across nodes. However, these divide-and-conquer methods do not guarantee the positive definiteness of the estimates. Additionally, some studies focus on cases where variables, rather than samples, are distributed across agents (Wiesel and Hero, 2011; Meng et al., 2014; Tavassolipour et al., 2019), which falls beyond the scope of this paper.

**GGM estimation with other loss functions.** Several other centralized methods have been proposed for GGM estimation, including the $D$-trace method (Zhang and Zou, 2014), column-by-column estimation (Meinshausen and Bühlmann, 2006), and linear programming-based approaches (Yuan, 2010; Cai et al., 2011; Liu and Luo, 2015). Wang and Cui (2021) extended the $D$-trace loss to distributed settings using a divide-and-conquer approach, while Dong and Liu (2024) introduced an alternating block-based gradient descent method to iteratively solve the $\ell_1$-penalized $D$-trace loss over networks. While the latter method moves beyond the divide-and-conquer framework, it still lacks theoretical guarantees for convergence and positive definiteness.

# 3 PROBLEM FORMULATION

In this section, we present the assumptions regarding the network and the problem setup for GGM estimation. We first discuss high-dimensional sparse GGM estimation and then extend the framework to other high-dimensional structured GGM estimation problems.

## 3.1 Network Topology

The network is modeled as a time-invariant undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices $\mathcal{V} = \{1, \ldots, m\}$ corresponding to the $m$ agents, and $\mathcal{E}$ represents the communication links among agents. We make the following standard assumption regarding the connectivity of the graph.

**Assumption 1.** *The graph $\mathcal{G}$ is connected.*

The neighborhood of agent $i$ is defined as $\mathcal{N}_i = \{j \mid (i,j) \in \mathcal{E}\} \cup \{i\}$. Define the weight matrix associated with $\mathcal{G}$ as $\mathbf{W} \in \mathbb{R}^{m \times m}$, where its $(i,j)$-th entry is

$$\begin{cases} W_{ij} \in [\kappa, 1], & \text{if } j \in \mathcal{N}_i, \\ W_{ij} = 0, & \text{otherwise,} \end{cases}$$

with $\kappa \in (0,1)$ a constant. Throughout this paper, we assume that $\mathbf{W}$ fulfills the following assumption.

**Assumption 2.** $\mathbf{W}$ *is doubly stochastic, i.e.,* $\mathbf{W1} = \mathbf{1}$ *and* $\mathbf{1}^\top \mathbf{W} = \mathbf{1}^\top$.

## 3.2 The Optimization Problem

For a zero-mean random vector $\mathbf{x} \in \mathbb{R}^d$, we assume that agent $i$, for $i = 1, 2, \ldots, m$, possesses $n_i$ independent and identically distributed samples $\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i}$. We are interested in estimating the precision matrix $\mathbf{\Theta}$ given all the samples in the network, defined as the solution to the following problem:

$$\underset{\mathbf{\Theta} \succ \mathbf{0}}{\text{minimize}} \quad U(\mathbf{\Theta}) = \underbrace{\sum_{i=1}^{m} f_i(\mathbf{\Theta})}_{=F(\mathbf{\Theta})} + \lambda \|\mathbf{\Theta}\|_1, \tag{1}$$

where the local likelihood function is

$$f_i(\mathbf{\Theta}) = \frac{n_i}{N}\left(-\log\det(\mathbf{\Theta}) + \langle \mathbf{S}_i, \mathbf{\Theta} \rangle\right)$$

with $N = \sum_{i=1}^m n_i$ the total sample size, and $\mathbf{S}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_{ij}\mathbf{x}_{ij}^\top$ the sample covariance matrix of agent $i$. Problem (1) has been proven to have a unique solution (Banerjee et al., 2008; Lu, 2009). Furthermore, since the objective function is convex, we can establish the following lemma.

**Lemma 1.** *The objective function $U$ is coercive over $\mathbb{S}_{++}^d$, i.e., $\lim_{\lambda_{\min}(\mathbf{\Theta}) \to 0} U(\mathbf{\Theta}) = +\infty$ and $\lim_{\|\mathbf{\Theta}\|_F \to +\infty} U(\mathbf{\Theta}) = +\infty$.*

Since each data set is local, agent $i$ can only access $f_i$ and cannot solve (1) independently. Our goal is to design an algorithm that computes the optimal $\mathbf{\Theta}$ without exchanging data or local covariance matrices.

# 4 ALGORITHM DESIGN

We first briefly introduce a single-loop proximal gradient-type algorithm (Rolfs et al., 2012) for solving the sparse GGM problem in (1) in a centralized manner. Then, we propose a decentralized algorithm termed Network Gaussian graphical models (NetGGM) to solve the problem over networks.

## 4.1 Proximal Gradient Algorithm for Sparse GGM

In the centralized setting, a standard approach to solving (1) is via the proximal gradient method, as summarized in Algorithm 1. At iteration $k$, it updates the estimate by minimizing a proximal regularized lin-

earization of $F$:

$$\boldsymbol{\Theta}^{(k+1)} = \arg\min_{\boldsymbol{\Theta}\succ\mathbf{0}}\Big\{F\left(\boldsymbol{\Theta}^{(k)}\right)+\left\langle\nabla F\left(\boldsymbol{\Theta}^{(k)}\right),\boldsymbol{\Theta}-\boldsymbol{\Theta}^{(k)}\right\rangle + \frac{\gamma}{2}\left\|\boldsymbol{\Theta}-\boldsymbol{\Theta}^{(k)}\right\|_F^2+\lambda\|\boldsymbol{\Theta}\|_1\Big\}, \tag{2}$$

where $\gamma>0$ is the step size. It is proven that when the step size is sufficiently small, solving (2) is equivalent to a soft thresholding operation (Rolfs et al., 2012):

$$\boldsymbol{\Theta}^{(k+1)} = \mathrm{ST}_{\frac{\lambda}{\gamma}}\left(\boldsymbol{\Theta}^{(k)}-\frac{1}{\gamma}\left(\mathbf{S}-\left(\boldsymbol{\Theta}^{(k)}\right)^{-1}\right)\right),$$

where $\mathbf{S}=\sum_{i=1}^m\frac{n_i}{N}\mathbf{S}_i$ is the global sample covariance matrix. For any $\mathbf{X}\in\mathbb{R}^{d\times d}$ and $\omega>0$, the $(i,j)$-th entry of $\mathrm{ST}_\omega(\mathbf{X})$ is defined as

$$\mathrm{ST}_\omega(\mathbf{X})_{ij}=\mathrm{sign}(X_{ij})\max\{|X_{ij}|-\omega,0\}.$$

### 4.2 Proposed Algorithm: NetGGM

However, in the distributed setting, due to the lack of global information, agents cannot locally solve problem (2). To address this, we propose NetGGM to solve problem (1) over the network. NetGGM extends the proximal gradient algorithm to the distributed network setting using gradient tracking techniques (Di Lorenzo and Scutari, 2016). Gradient tracking is a standard technique for solving decentralized optimization problems, enabling each agent to iteratively update a local auxiliary variable by integrating information from its neighbors, thereby tracking the global gradient of the problem. NetGGM consists of two main steps: local optimization and information mixing.

**Local optimization.** For agent $i$, we define $\boldsymbol{\Theta}_i$ as a local estimator for the optimization variable $\boldsymbol{\Theta}$ and $\mathbf{Y}_i$ as a local auxiliary variable that aims to asymptotically track $\frac{1}{m}\sum_{i=1}^m\nabla f_i(\boldsymbol{\Theta}_i)$. Denote $\boldsymbol{\Theta}^{(k)}=\left[\boldsymbol{\Theta}_1^{(k)};\boldsymbol{\Theta}_2^{(k)};\ldots;\boldsymbol{\Theta}_m^{(k)}\right]$ and $\mathbf{Y}^{(k)}=\left[\mathbf{Y}_1^{(k)};\mathbf{Y}_2^{(k)};\ldots;\mathbf{Y}_m^{(k)}\right]$ as the matrices of local variables of all agents. We decompose $F$ into two parts: $f_i$ and $\sum_{j\neq i}f_j$, and separately approximate these two terms. At iteration $k$, the surrogate function of $F$ at agent $i$ is constructed as

$$\tilde{F}_i^{(k)}(\boldsymbol{\Theta}_i)=\underbrace{f_i\left(\boldsymbol{\Theta}_i^{(k)}\right)+\left\langle\nabla f_i\left(\boldsymbol{\Theta}_i^{(k)}\right),\boldsymbol{\Theta}_i-\boldsymbol{\Theta}_i^{(k)}\right\rangle}_{(a)} + \underbrace{\left\langle m\mathbf{Y}_i^{(k)}-\nabla f_i\left(\boldsymbol{\Theta}_i^{(k)}\right),\boldsymbol{\Theta}_i-\boldsymbol{\Theta}_i^{(k)}\right\rangle}_{(b)} + \frac{\tau}{2}\left\|\boldsymbol{\Theta}_i-\boldsymbol{\Theta}_i^{(k)}\right\|_F^2, \tag{3}$$

**Algorithm 1** Sparse GGM via proximal gradient

**given** $\boldsymbol{\Theta}^{(0)}\succeq\mathbf{0}$, $\gamma>0$, $k=0$
**while** not converge, **do**
  $\boldsymbol{\Theta}^{(k+1)}=\mathrm{ST}_{\frac{\lambda}{\gamma}}\left(\boldsymbol{\Theta}^{(k)}-\frac{1}{\gamma}\left(\mathbf{S}-\left(\boldsymbol{\Theta}^{(k)}\right)^{-1}\right)\right)$,
  $k=k+1$
**end while**
**return** $\boldsymbol{\Theta}^{(k)}$.

**Algorithm 2** NetGGM

**given** $\boldsymbol{\Theta}_i^{(0)}\succeq\mathbf{0}$, $\mathbf{Y}_i^{(0)}=\frac{n_1}{N}\left(\mathbf{S}_i-\left(\boldsymbol{\Theta}_i^{(0)}\right)^{-1}\right)$, $\mathbf{W}$, $\tau>0$, $\alpha\in(0,1]$, $k=0$
**while** not converge, each agent $i$ **do**
  **Local optimization:**
    $\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}=\mathrm{ST}_{\frac{\lambda}{\tau}}\left(\boldsymbol{\Theta}_i^{(k)}-\frac{m}{\tau}\mathbf{Y}_i^{(k)}\right)$,
  **Information mixing:**
    $\tilde{\boldsymbol{\Theta}}_i^{(k)}=\boldsymbol{\Theta}_i^{(k)}+\alpha\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}_i^{(k)}\right)$,
    $\boldsymbol{\Theta}_i^{(k+1)}=\sum_{j=1}^m W_{ij}\tilde{\boldsymbol{\Theta}}_j^{(k)}$,
    $\tilde{\mathbf{Y}}_i^{(k)}=\mathbf{Y}_i^{(k)}+\left(\boldsymbol{\Theta}_i^{(k)}\right)^{-1}-\left(\boldsymbol{\Theta}_i^{(k+1)}\right)^{-1}$,
    $\mathbf{Y}_i^{(k+1)}=\sum_{j=1}^m W_{ij}\tilde{\mathbf{Y}}_j^{(k)}$,
  $k=k+1$
**end while**
**return** $\boldsymbol{\Theta}^{(k)}$.

where $\tau>0$ is a parameter. Function $\tilde{F}_i$ is a strongly convex function with constant $\tau$ where terms $(a)$ and $(b)$ represent the linearization of $f_i$ and $\sum_{j\neq i}f_j$ around $\boldsymbol{\Theta}_i^{(k)}$, respectively. Then, the local proximal step for agent $i$ is given as

$$\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}=\arg\min_{\boldsymbol{\Theta}_i\succ\mathbf{0}}\left\{\tilde{F}_i^{(k)}(\boldsymbol{\Theta}_i)+\lambda\|\boldsymbol{\Theta}_i\|_1\right\}. \tag{4}$$

Due to the positive definiteness constraint and the $\ell_1$ penalty, the subproblem (4) lacks a closed-form solution. To address this challenge, we propose to discard the positive definite constraint and consequently, updating $\boldsymbol{\Theta}_i$ can be done using a single soft-thresholding operation

$$\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}=\mathrm{ST}_{\frac{\lambda}{\tau}}\left(\boldsymbol{\Theta}_i^{(k)}-\frac{m}{\tau}\mathbf{Y}_i^{(k)}\right).$$

As to be elaborated in the next section, we theoretically prove that as long as the parameter $\tau$ is appropriately chosen, $\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\succ\mathbf{0}$ is automatically satisfied for any $k\in\mathbb{N}$. In this way, NetGGM avoids an extra computation loop for solving (4) with an iterative algorithm, which is typically computationally demanding.

**Information mixing.** After the local optimization, each agent $i$ collects information from its neighbors

and updates both $\mathbf{\Theta}_i$ and $\mathbf{Y}_i$. It first computes

$$\tilde{\mathbf{\Theta}}_i^{(k)} = \mathbf{\Theta}_i^{(k)} + \alpha\left(\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \mathbf{\Theta}_i^{(k)}\right),$$

where $\alpha \in (0,1]$ is a constant step size. It then receives $\left\{\tilde{\mathbf{\Theta}}_j^{(k)}\right\}_{j\in\mathcal{N}_i}$ from its neighbors and applies the consensus-based step to enforce asymptotic agreement among $\mathbf{\Theta}$, given as follows:

$$\mathbf{\Theta}_i^{(k+1)} = \sum_{j=1}^{m} W_{ij}\tilde{\mathbf{\Theta}}_j^{(k)}.$$

Next, agent $i$ updates its local gradient estimator according to

$$\tilde{\mathbf{Y}}_i^{(k)} = \mathbf{Y}_i^{(k)} + \left(\mathbf{\Theta}_i^{(k)}\right)^{-1} - \left(\mathbf{\Theta}_i^{(k+1)}\right)^{-1},$$

exchanges it with its neighbors, and updates $\mathbf{Y}_i$ with

$$\mathbf{Y}_i^{(k+1)} = \sum_{j=1}^{m} W_{ij}\tilde{\mathbf{Y}}_j^{(k)}.$$

NetGGM iteratively performs the above operations until convergence. We summarize it as Algorithm 2.

## 5 CONVERGENCE ANALYSIS

In this section, we prove the linear convergence of NetGGM. One of the challenges is that function $F$ does not exhibit global Lipschitz smoothness and strong convexity over $\mathbf{\Theta} \succ \mathbf{0}$. This can be seen from the Hessian $\nabla^2 F(\mathbf{\Theta}) = \mathbf{\Theta}^{-1} \otimes \mathbf{\Theta}^{-1}$, whose eigenvalues lack positive upper and lower bounds. Moreover, the local optimization updates in NetGGM do not explicitly enforce positive definiteness constraints. To address these difficulties, our analysis proceeds in two steps: (1) choosing appropriate parameters and utilizing the coercivity of $U$ to guarantee local Lipschitz smoothness and strong convexity; (2) using this local smoothness and strong convexity to establish the linear convergence of NetGGM.

**Local Properties.** Our analysis hinges on the potential function $V$ that combines the objective function with the consensus error as follows:

$$V\left(\mathbf{\Theta}^{(k)}, \mathbf{Y}^{(k)}\right) = \sum_{i=1}^{m} U\left(\mathbf{\Theta}_i^{(k)}\right) + \underbrace{a\sum_{i=1}^{m}\left\|\mathbf{\Theta}_i^{(k)} - \bar{\mathbf{\Theta}}^{(k)}\right\|_F^2 + b\sum_{i=1}^{m}\left\|\mathbf{Y}_i^{(k)} - \bar{\mathbf{Y}}^{(k)}\right\|_F^2}_{\text{consensus error}},$$

where $\bar{\mathbf{\Theta}}^{(k)} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{\Theta}_i^{(k)}$, $\bar{\mathbf{Y}}^{(k)} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{Y}_i^{(k)}$, and $a, b > 0$. The function $V$ is coercive, which follows from Lemma 1 and the fact that the consensus error is non-negative. For any initialization $\mathbf{\Theta}^{(0)} \succ \mathbf{0}$, we can define a level set

$$\mathcal{A} = \left\{(\mathbf{\Theta}, \mathbf{Y}) \mid V(\mathbf{\Theta}, \mathbf{Y}) \le V\left(\mathbf{\Theta}^{(0)}, \mathbf{Y}^{(0)}\right)\right\}.$$

Since $V$ is coercive, there exist constants $\overline{R}_\Theta > \underline{R}_\Theta > 0$ and $R_Y > 0$ such that for every $(\mathbf{\Theta}, \mathbf{Y}) \in \mathcal{A}$ we have $\underline{R}_\Theta \mathbf{I} \preceq \mathbf{\Theta}_i \preceq \overline{R}_\Theta \mathbf{I}$ and $-R_Y\mathbf{I} \preceq \mathbf{Y}_i - \bar{\mathbf{Y}} \preceq R_Y\mathbf{I}$ for $i = 1, 2, \ldots, m$.

Then, by carefully selecting parameters $\tau$ and $\alpha$, we ensure that every $\left(\mathbf{\Theta}^{(k)}, \mathbf{Y}^{(k)}\right)$, $k \in \mathbb{N}$ obtained by NetGGM remain within $\mathcal{A}$, thereby achieving local Lipschitz smoothness and strong convexity while maintaining the positive definiteness of the estimate. We prove this conclusion by induction, as shown in Theorem 3.

**Theorem 3.** *Assume that Assumptions 1 and 2 are satisfied. Based on Lemma 1, when $\tau \ge \underline{\tau}_1$ and $\alpha \in (0, \bar{\alpha}_1]$, for $\left\{\mathbf{\Theta}^{(k)}, \mathbf{Y}^{(k)}\right\}_{k\in\mathbb{N}}$ obtained by NetGGM, we have*

$$V\left(\mathbf{\Theta}^{(k+1)}, \mathbf{Y}^{(k+1)}\right) \le V\left(\mathbf{\Theta}^{(k)}, \mathbf{Y}^{(k)}\right) - \beta\sum_{i=1}^{m}\left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \mathbf{\Theta}_i^{(k)}\right\|_F^2,$$

*where $\underline{\tau}_1$, $\bar{\alpha}_1$, and $\beta$ are positive universal constants.*

Utilizing Theorem 3 and iterating to $k = 0$, we have

$$V\left(\mathbf{\Theta}^{(k)}, \mathbf{Y}^{(k)}\right) \le \ldots \le V\left(\mathbf{\Theta}^{(0)}, \mathbf{Y}^{(0)}\right).$$

Therefore, $\left(\mathbf{\Theta}^{(k)}, \mathbf{Y}^{(k)}\right) \in \mathcal{A}$ and all the $\mathbf{\Theta}^{(k)}$ are positive definite for every iteration $k \in \mathbb{N}$. This conclusion leads to the local Lipschitz smoothness and strong convexity at each iteration.

Note that Lu (2009) proves that the unique solution to problem (1) lies in a convex subset of $\mathbb{S}_{++}^d$, which is defined based on the global sample covariance matrix $\mathbf{S}$, the dimension $d$, and the thresholding parameter $\lambda$. In the centralized case, the proximal gradient algorithm has been shown to ensure that, as long as the initialization is within this subset and the step size is sufficiently small, the resulting sequence remains in the subset (Rolfs et al., 2012). However, in the network setting, since each agent lacks the knowledge of $\mathbf{S}$, we cannot guarantee that all $\mathbf{\Theta}^{(0)}$ lie within this subset. In contrast, in our proof, the set $\mathcal{A}$ is constructed based on $\mathbf{\Theta}^{(0)}$, ensuring that for any positive definite initialization, each update of NetGGM remains within set $\mathcal{A}$, thereby ensuring local Lipschitz smoothness and strong convexity throughout the iterations.

**Linear Convergence.** Based on the local Lipschitz smoothness and strong convexity, we establish the following result.

**Theorem 4.** *In the same setting as Theorem 3, if the proximal parameter is sufficiently large satisfying* $\tau \geq \max\left\{\underline{\tau}_1, \frac{4}{\underline{R}_\Theta^2}\right\}$ *and the step size satisfies* $\alpha \in (0, \bar{\alpha}_1]$ *for some constant* $\underline{\tau}_1$ *and* $\bar{\alpha}_1$*, we have*

$$\sum_{i=1}^m \left\|\boldsymbol{\Theta}_i^{(k+1)} - \hat{\boldsymbol{\Theta}}\right\|_F^2 \leq \gamma \sum_{i=1}^m \left\|\boldsymbol{\Theta}_i^{(k)} - \hat{\boldsymbol{\Theta}}\right\|_F^2 + a' \sum_{i=1}^m \left\|\boldsymbol{\Theta}_i^{(k)} - \bar{\boldsymbol{\Theta}}^{(k)}\right\|_F^2 + b' \sum_{i=1}^m \left\|\mathbf{Y}_i^{(k)} - \bar{\mathbf{Y}}^{(k)}\right\|_F^2,$$

*where* $a', b' > 0$*, and* $\gamma \in (0,1)$ *are universal constants.*

Theorem 4 illustrates the linear convergence of NetGGM for the consensus error. Then, we bound the consensus error and establish the R-linear convergence of NetGGM.

**Theorem 5.** *In the same setting as Theorem 4, if the proximal parameter satisfies* $\tau \geq \max\left\{\underline{\tau}_1, \frac{4}{\underline{R}_\Theta^2}\right\}$ *and the step size* $\alpha$ *satisfies* $\alpha \in (0, \bar{\alpha}_2)$*, we have*

$$\sum_{i=1}^m \left\|\boldsymbol{\Theta}_i^{(k)} - \hat{\boldsymbol{\Theta}}\right\|_F^2 \leq A\underline{z}^k$$

*for all* $k \in \mathbb{N}$*, where* $\bar{\alpha}_2, A > 0$*, and* $\underline{z} \in (0,1)$ *are constants depend on the data, the initialization of* $\boldsymbol{\Theta}^{(0)}$*, and the network connectivity.*

Theorem 5 shows that NetGGM converges linearly to the global optimum of problem (1), which indicates that NetGGM can achieve the same estimation accuracy as the centralized estimator (Rothman et al., 2008; Ravikumar et al., 2011).

**Complexity.** According to Theorem 5, NetGGM converges within $\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$ iterations to reach an error less than $\epsilon > 0$. In each iteration, the complexity bottleneck is the computation of $\boldsymbol{\Theta}^{-1}$, which requires $\mathcal{O}(d^3)$ operations. With $m$ agents performing this computation in parallel, the overall time complexity for NetGGM to converge is $\mathcal{O}(md^3 \log(\frac{1}{\epsilon}))$. Each iteration involves exchanging $\mathcal{O}(|\mathcal{E}|d^2)$ units of information, where $|\mathcal{E}|$ stands for the number of edges. Thus, the total communication complexity for NetGGM to converge is $\mathcal{O}(|\mathcal{E}|d^2 \log(\frac{1}{\epsilon}))$.

## 6 OTHER REGULARIZERS

In addition to the $\ell_1$ penalty considered in Section 3.2, various other regularization terms can be integrated into problem (1) to induce various structures. In this section, we illustrate how the NetGGM algorithm can be applied to more general GGM estimation problems by considering the following formulation:

$$\underset{\boldsymbol{\Theta} \succ \mathbf{0}}{\text{minimize}} \quad F(\boldsymbol{\Theta}) + G(\boldsymbol{\Theta}), \tag{5}$$

where $F$ is defined in (1), and $G$ is a generic convex regularization term. To solve (5) using NetGGM, the local optimization step in (4) is replaced by the following one:

$$\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)} = \arg\min_{\boldsymbol{\Theta}_i} \left\{\tilde{F}_i^{(k)}(\boldsymbol{\Theta}_i) + G(\boldsymbol{\Theta})\right\}, \tag{6}$$

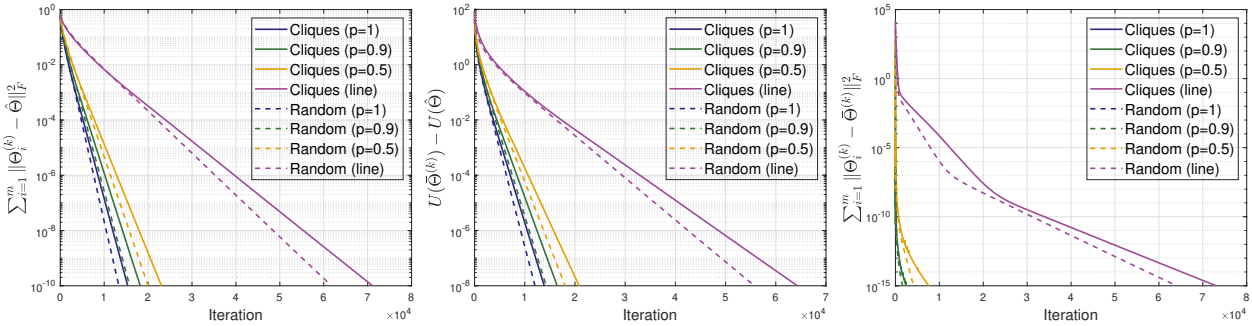
where $\tilde{F}_i^{(k)}$ is defined in (3).

In practice, $G$ can take many different possible forms. Below, we provide some examples. When prior knowledge about the sparsity pattern is available, a weighted $\ell_1$-penalty can be used to encourage certain elements to approach zero (Li and Jackson, 2015). If the sparsity pattern is fully known, constraints can be directly imposed to ensure the estimator matches this pattern (Egilmez et al., 2017). When the dimensionality exceeds the sample size, the $\ell_1$-penalty can estimate at most as many edges as there are available samples (Zou and Hastie, 2005). In such cases, the elastic net penalty can help overcome this limitation by allowing for denser solutions (Ryali et al., 2012). The SLOPE penalty can be employed to achieve graph estimation with a controlled false discovery rate (Mazza-Anthony et al., 2020). Additionally, in fields such as finance, where non-negative correlations exist, the $\text{MTP}_2$ constraint (Karlin and Rinott, 1983) can be applied. With the above possible penalty function, the proximal step (6) has a closed-form solution like the $\ell_1$-norm case.

It is worth noting that there exist regularization terms $G$ that prevent subproblems (6) from having closed-form solutions. For example, in some graphical model estimation problems, it may be desirable to select edges as a group. Given a set of such groups $\mathcal{G}_1, ..., \mathcal{G}_K \subset \{1, \ldots, d\}^2$, the group lasso penalty (Marlin and Murphy, 2009)

$$G(\boldsymbol{\Theta}) = \lambda \sum_{i=1}^K |\mathcal{G}_i|^{\frac{1}{2}} \left\|(\boldsymbol{\Theta})_{\mathcal{G}_i}\right\|_2$$

can achieve this goal, where $|\mathcal{G}_i|$ is the number of elements in $\mathcal{G}_i$, and $(\boldsymbol{\Theta})_{\mathcal{G}_i}$ represents the vector of elements corresponding to $\mathcal{G}_i$. However, when these groups overlap with each other, i.e., when elements appear in multiple groups, the subproblem (6) no longer admits a closed-form solution. In this case, we may only obtain an $\varepsilon^{(k)}$-optimal solution $\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}$ to (6)




(a) $\sum_{i=1}^{m} \left\| \boldsymbol{\Theta}_i^{(k)} - \hat{\boldsymbol{\Theta}} \right\|_F^2$ (b) $U\left(\bar{\boldsymbol{\Theta}}^{(k)}\right) - U\left(\hat{\boldsymbol{\Theta}}\right)$ (c) $\sum_{i=1}^{m} \left\| \boldsymbol{\Theta}_i^{(k)} - \bar{\boldsymbol{\Theta}}^{(k)} \right\|_F^2$

Figure 1: Linear convergence of NetGGM over different networks for two models. ($m = 20$ and $N = 25$)

from an iterative algorithm, which satisfies

$$\tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) + G\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) \leq \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}^{\left(k+\frac{1}{2}\right)\star}\right) + G\left(\boldsymbol{\Theta}^{\left(k+\frac{1}{2}\right)\star}\right) + \varepsilon^{(k)},$$

where $\varepsilon^{(k)} > 0$, $k \in \mathbb{N}$.

In this case, we can prove that NetGGM still exhibits R-linear convergence, as described in Theorem 6.

**Theorem 6.** *Assume that Assumptions 1 and 2 are satisfied. Assume that the error $\varepsilon^{(k)}$ decay at least at an exponential rate (i.e., there exist constants $c > 0$ and $\bar{\varepsilon} \in (0,1)$ such that $\varepsilon^{(k)} \leq c\bar{\varepsilon}^k$) and satisfies $\varepsilon^{(k)} < \frac{\left(\sqrt{2}-1\right)^2 \underline{R}_\Theta^2}{8}$. For $\tau \geq \underline{\tau}_2$ and $\alpha \in (0, \bar{\alpha}_3)$, we have*

$$\sum_{i=1}^{m} \left\| \boldsymbol{\Theta}_i^{(k)} - \hat{\boldsymbol{\Theta}} \right\|_F^2 \leq (A' + Bc)\, \underline{z}'^k$$

*for all $k \in \mathbb{N}$, where $\underline{\tau}_2, \bar{\alpha}_3, A', B > 0$, and $\underline{z}' \in (0,1)$ are constants depend on the data, the initialization of $\boldsymbol{\Theta}^{(0)}$, and the network connectivity.*

# 7 NUMERICAL SIMULATIONS

In this section, we demonstrate the convergence of NetGGM and assess its estimation performance on both synthetic data and real data from Parkinson's disease. We use NetGGM to solve sparse GGM over networks and select the shrinkage parameter $\lambda$ of for all cases by five-fold cross-validation. The parameters $\tau$ and $\alpha$ were tuned to ensure the convergence of the algorithm. Empirically, we choose $\alpha = 0.1$ and choose $\tau$ as $\eta^t$ where $\eta > 1$ and $t \in \mathbb{N}$ is the smallest number ensuring the convergence of NetGGM. We compare the performance of NetGGM with G-ISTA (Rolfs et al., 2012) and two existing divide-and-conquer based estimators D&C I (Arroyo and Hou, 2016) and D&C II (Nezakati and Pircalabelu, 2023). Note that the baselines are not decentralized methods, since G-ISTA solves sparse GGM in centralized settings while D&C I and D&C II need a central node to aggregate information. The parameters for these baselines are adjusted as described in the original references. All approaches are implemented in MATLAB without any code in compiled languages on a 3.3 GHz 12-core Intel Xeon W processor. The weight matrix $\mathbf{W}$ for each simulation is set according to the Metropolis weights (Xiao et al., 2005)

$$W_{ij} = \begin{cases} \frac{1}{\max(d_i, d_j)+1}, & i \neq j \text{ and } (i,j) \in \mathcal{E}, \\ 0, & i \neq j \text{ and } (i,j) \notin \mathcal{E}, \\ 1 - \sum_{l \neq i} W_{il}, & i = j, \end{cases}$$

where $d_i$ is the degree of agent $i$.

## 7.1 Synthetic Data

In the synthetic data simulation, we consider a $d$-variate Gaussian distribution $\mathcal{N}\left(\mathbf{0}, \left(\boldsymbol{\Theta}^\star\right)^{-1}\right)$ with $d = 50$. $N$ samples are drawn from the Gaussian distribution and randomly distributed to the $m$ agents. We consider two precision matrix models (Bien and Tibshirani, 2011):

- Cliques model: $\boldsymbol{\Theta}^\star = \text{blkdiag}\left(\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(5)}\right)$, where the off-diagonal elements of the submatrices $\boldsymbol{\Theta}_{(1)}, \ldots, \boldsymbol{\Theta}_{(5)}$ are set to $\pm 1$;
- Random model: Each off-diagonal element is set to $\pm 1$ with a probability of 0.05.

For these two models, the diagonal elements are set to a constant which makes the condition number of $\boldsymbol{\Theta}^\star$ equal to $d$ as in (Rothman et al., 2008). The performance of NetGGM is tested over four different connected time-invariant undirected networks: three Erdős–Rényi models (Erdős and Rényi, 1959) (each

Table 1: Average and variance (between parentheses) of performance of NetGGM and baselines.

| **Cliques model** | Methods | NMSE | | | Positive definiteness | | |
|---|---|---|---|---|---|---|---|
| | G-ISTA | **0.2585 (0.0001)** | | | **100/100** | | |
| | Number of agents $m$ | 5 | 10 | 20 | 5 | 10 | 20 |
| $N = 25$ | NetGGM ($p = 0.9$) | **0.2585 (0.0001)** | **0.2585 (0.0001)** | **0.2585 (0.0001)** | **100/100** | **100/100** | **100/100** |
| | NetGGM ($p = 0.5$) | **0.2585 (0.0001)** | **0.2585 (0.0001)** | **0.2585 (0.0001)** | **100/100** | **100/100** | **100/100** |
| | NetGGM (line) | **0.2585 (0.0001)** | **0.2585 (0.0001)** | **0.2585 (0.0001)** | **100/100** | **100/100** | **100/100** |
| | D&C I | 24.1020 (2024.0296) | 135.0798 (32190.5248) | 1337.6720 (643780.6406) | 8/100 | 8/100 | 8/100 |
| | D&C II | 0.7847 (0.0033) | 0.8367 (0.0025) | 1.0055 (0.0041) | 0/100 | 0/100 | 0/100 |
| | G-ISTA | **0.1313 (0.00002)** | | | **100/100** | | |
| | Number of agents $m$ | 5 | 10 | 20 | 5 | 10 | 20 |
| $N = 100$ | NetGGM ($p = 0.9$) | **0.1313 (0.00002)** | **0.1313 (0.00002)** | **0.1313 (0.00002)** | **100/100** | **100/100** | **100/100** |
| | NetGGM ($p = 0.5$) | **0.1313 (0.00002)** | **0.1313 (0.00002)** | **0.1313 (0.00002)** | **100/100** | **100/100** | **100/100** |
| | NetGGM (line) | **0.1313 (0.00002)** | **0.1313 (0.00002)** | **0.1313 (0.00002)** | **100/100** | **100/100** | **100/100** |
| | D&C I | 0.7086 (0.7272) | 0.8279 (2.9621) | 11.0683 (164.1900) | 9/100 | 9/100 | 9/100 |
| | D&C II | 0.3496 (0.0001) | 0.3847 (0.0001) | 0.4306 (0.0001) | 0/100 | 0/100 | 0/100 |
| **Random model** | Methods | NMSE | | | Positive definiteness | | |
| | G-ISTA | **0.1955 (0.0001)** | | | **100/100** | | |
| | Number of agents $m$ | 5 | 10 | 20 | 5 | 10 | 20 |
| $N = 25$ | NetGGM ($p = 0.9$) | **0.1955 (0.0001)** | **0.1955 (0.0001)** | **0.1955 (0.0001)** | **100/100** | **100/100** | **100/100** |
| | NetGGM ($p = 0.5$) | **0.1955 (0.0001)** | **0.1955 (0.0001)** | **0.1955 (0.0001)** | **100/100** | **100/100** | **100/100** |
| | NetGGM (line) | **0.1955 (0.0001)** | **0.1955 (0.0001)** | **0.1955 (0.0001)** | **100/100** | **100/100** | **100/100** |
| | D&C I | 21.8069 938.1516 | 89.01776 3063.7182 | 525.9566 28055.0031 | 5/100 | 5/100 | 8/100 |
| | D&C II | 0.8754 0.0067 | 0.8843 0.0081 | 13.4111 370.7767 | 0/100 | 0/100 | 0/100 |
| | G-ISTA | **0.0667 (0.00004)** | | | **100/100** | | |
| | Number of agents $m$ | 5 | 10 | 20 | 5 | 10 | 20 |
| $N = 100$ | NetGGM ($p = 0.9$) | **0.0667 (0.00004)** | **0.0667 (0.00004)** | **0.0667 (0.00004)** | **100/100** | **100/100** | **100/100** |
| | NetGGM ($p = 0.5$) | **0.0667 (0.00004)** | **0.0667 (0.00004)** | **0.0667 (0.00004)** | **100/100** | **100/100** | **100/100** |
| | NetGGM (line) | **0.0667 (0.00004)** | **0.0667 (0.00004)** | **0.0667 (0.00004)** | **100/100** | **100/100** | **100/100** |
| | D&C I | 0.2009 (0.0017) | 2.4042 (2.3302) | 6.9944 (38.3535) | 5/100 | 5/100 | 5/100 |
| | D&C II | 0.3056 (0.0002) | 0.3453 (0.0004) | 0.3761 (0.0005) | 0/100 | 0/100 | 0/100 |

pair of agents is connected randomly with probability $p = 1$, $p = 0.9$, and $p = 0.5$) and a line-structured model (agent $i$ is only connected to agent $i-1$ and agent $i+1$, $i = 2, \ldots, m-1$).

We first demonstrate the convergence of NetGGM. Figure 1 shows the decrease in the average distance between $\mathbf{\Theta}_i^{(k)}$ obtained by NetGGM and $\hat{\mathbf{\Theta}}$ obtained by G-ISTA in the centralized setting $\sum_{i=1}^m \left\|\mathbf{\Theta}_i^{(k)} - \hat{\mathbf{\Theta}}\right\|_F^2$, along with the reduction in their optimality gap $U\left(\bar{\mathbf{\Theta}}^{(k)}\right) - U\left(\hat{\mathbf{\Theta}}\right)$ and consensus error $\sum_{i=1}^m \left\|\mathbf{\Theta}_i^{(k)} - \bar{\mathbf{\Theta}}^{(k)}\right\|_F^2$ under two models over the line model network. This confirms that NetGGM achieves consensus while converging to the global optimum linearly, with accuracy almost identical to G-ISTA. Note that when $p = 1$, the graph becomes a fully connected graph, and NetGGM essentially becomes G-ISTA. As the connectivity of the the network increases, the convergence rate of NetGGM improves accordingly.

Then, we compared the estimation performance of NetGGM with the baselines. The estimation performance evaluated at $\bar{\mathbf{\Theta}}^{(k)}$ is measured by the normalized mean square error (NMSE) defined as

$$\mathsf{NMSE}\left(\bar{\mathbf{\Theta}}^{(k)}\right) = \frac{\left\|\bar{\mathbf{\Theta}}^{(k)} - \mathbf{\Theta}^\star\right\|^2}{\left\|\mathbf{\Theta}^\star\right\|^2}.$$

We simulated the cases with sample sizes $N = \{25, 100\}$ and number of agents $m = \{5, 10, 20\}$ on two models, and the results averaging over 100

Monte Carlo trials are presented in Table 1. It can be observed that across all networks and scenarios, NetGGM achieves the same estimation performance as G-ISTA and outperforms the two divide-and-conquer methods. This advantage is especially evident when the sample size is small and the number of agents is large. Among these methods, the most naive approach, divide-and-conquer I, performs significantly worse than others and has a very low probability of producing positive definite estimates. Divide-and-conquer II, which employs an element-wise weighting strategy to handle the unbalanced sample, achieves better estimation accuracy and stability compared to divide-and-conquer I, but it is even less likely to yield a positive definite inverse covariance matrix. In contrast to these two methods, NetGGM ensures positive definiteness of the estimated inverse covariance matrix.

### 7.2 Real Data

To evaluate the performance of our proposed methods, we conduct experiments using the Leukemia dataset from Golub et al. (1999). This dataset consists of $N = 72$ gene expression profiles, with 47 samples from acute lymphoblastic leukemia (ALL) patients and 25 samples from acute myeloid leukemia (AML) patients. Each sample is represented by 7129 gene expression levels. Following Rothman et al. (2009); Cui et al. (2016), we first calculate the $F$ statistic

$$F(x_j) = \frac{\frac{1}{K-1}\sum_{l=1}^{K} N_{(l)}\left(\bar{x}_{j(l)} - \bar{x}_j\right)}{\frac{1}{N-K}\sum_{l=1}^{K} N_{(l)}\left(N_{(l)} - 1\right)\hat{\sigma}^2_{(l)}},$$

for each gene $j$, where $K = 2$ is the number of classes, $N_{(l)}$ is the sample size of class $l$, $\bar{x}_j$ and $\bar{x}_{j(l)}$ are the overall mean and mean of class $l$, and $\hat{\sigma}^2_{(l)}$ is the sample variance of class $l$. Then we select the top $d = 50$ genes with the highest $F$ statistic for experiments. We randomly partition the dataset into 100 different subsets, each containing 35 training samples (23 ALL and 12 AML) and 37 test samples (24 ALL and 13 AML). In practice, these data may be held by separate hospitals and cannot be shared directly due to privacy regulations. To mimic this scenario, we distribute the training samples randomly among $m = \{3, 5\}$ agents and estimate the precision matrices. We then incorporate the estimated precision matrices into a quadratic discriminant analysis (QDA) (Hastie et al., 2009) model and evaluate the classification performance based on the misclassification rate.

As the estimates obtained by the two divide-and-conquer methods do not always guarantee positive definiteness, we only present results for cases where the estimates are positive definite. As shown in Figure 2, NetGGM consistently achieves classification accuracy

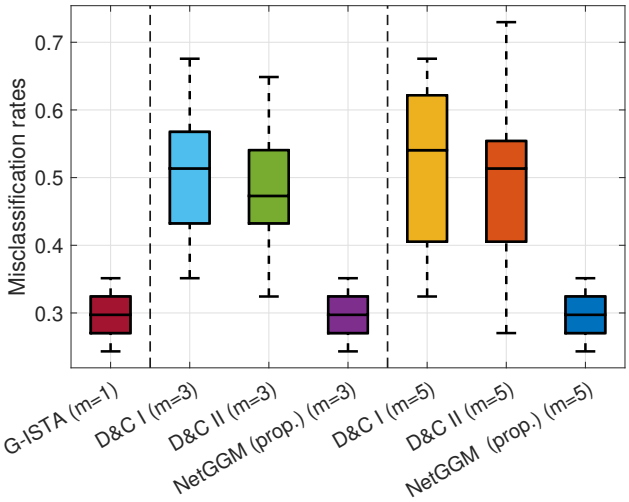



Figure 2: Misclassification rate of QDA with G-ISTA, NetGGM, D&C I, and D&C II on Leukemia dataset. Here we report the results of NetGGM on the line-structured model network.

comparable to that of G-ISTA across all cases and outperforms both divide-and-conquer methods. This result demonstrates the superior estimation performance of NetGGM. For divide-and-conquer methods, as the number of agents increases, each agent has fewer samples, leading to lower estimation accuracy and higher variability. Among them, D&C II utilizes element-wise weighted averaging, resulting in better performance compared to D&C I.

## 8 CONCLUSIONS AND FUTURE WORK

This paper has investigated covariance selection in the context of data distributed across a network of agents. The problem has been formulated as a Gaussian maximum likelihood estimation with structural penalties, and a novel lightweight algorithmic framework, called NetGGM, has been introduced. Unlike existing methods that have relied on a central node, NetGGM has operated in a fully decentralized manner with low computational complexity. We have provided theoretical proof showing that NetGGM has converged linearly to the global optimum from any positive definite initialization, achieving consensus among agents and preserving positive definiteness throughout. Numerical simulations have validated the convergence properties of NetGGM and have demonstrated that it has consistently outperformed state-of-the-art methods in estimating precision matrices. Future research could focus on reducing the number of iterations required for convergence and minimizing communication overhead during each exchange.

**Acknowledgements**

This work was supported by the National Nature Science Foundation of China (NSFC) under Grant 62001295.

**References**

Jesus Arroyo and Elizabeth Hou. Efficient distributed estimation of inverse covariance matrices. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5. IEEE, 2016.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

Amir Beck. *First-Order Methods in Optimization*, volume 25. SIAM, 2017.

Jacob Bien and Robert J Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4): 807–820, 2011.

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.

Jian-Feng Cai, José Vinícius de Miranda Cardoso, Daniel Palomar, and Jiaxi Ying. Fast projected Newton-like method for precision matrix estimation under total positivity. *Advances in Neural Information Processing Systems*, 36, 2024.

Tony Cai, Weidong Liu, and Xi Luo. A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Li-Pang Chen. Estimation of graphical models: An overview of selected topics. *International Statistical Review*, 92(2):194–245, 2024.

Shixiang Chen, Alfredo Garcia, and Shahin Shahrampour. On distributed nonconvex optimization: Projected subgradient method for weakly convex problems in networks. *IEEE Transactions on Automatic Control*, 67(2):662–675, 2021.

Ying Cui, Chenlei Leng, and Defeng Sun. Sparse estimation of high-dimensional correlation matrices. *Computational Statistics & Data Analysis*, 93:390–403, 2016.

Onkar Dalal and Bala Rajaratnam. Sparse Gaussian graphical model estimation via alternating minimization. *Biometrika*, 104(2):379–395, 2017.

Alexandre d'Aspremont, Onureena Banerjee, and Laurent El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.

Aaron Defazio and Tibério Caetano. A convex formulation for learning scale-free networks via submodular relaxation. *Advances in neural information processing systems*, 25, 2012.

Arthur P Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

Paolo Di Lorenzo and Gesualdo Scutari. NEXT: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

Wei Dong and Hongzhen Liu. Distributed sparse precision matrix estimation via alternating block-based gradient descent. *Mathematics*, 12(5):646, 2024.

Hilmi E Egilmez, Eduardo Pavez, and Antonio Ortega. Graph learning from data under Laplacian and structural constraints. *IEEE Journal of Selected Topics in Signal Processing*, 11(6):825–841, 2017.

Paul Erdős and Alfréd Rényi. On random graphs. *Publicationes Mathematicaes*, 6:290–297, 1959.

Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521, 2009.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439): 531–537, 1999.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Cho-Jui Hsieh, Inderjit Dhillon, Pradeep Ravikumar, and Mátyás Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. *Advances in neural information processing systems*, 24, 2011.

Jana Janková and Sara van de Geer. Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9:1205–1229, 2015.

Yao Ji, Gesualdo Scutari, Ying Sun, and Harsha Honnappa. Distributed sparse regression via penalization. *Journal of Machine Learning Research*, 24(272):1–62, 2023.

Ian T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, NY, USA, 2nd edition, 2002.

Samuel Karlin and Yosef Rinott. M-matrices as covariance matrices of multinormal distributions. *Linear Algebra and Its Applications*, 52:419–438, 1983.

Markku O Kuismin, Jukka T Kemppainen, and Mikko J Sillanpää. Precision matrix estimation with ROPE. *Journal of Computational and Graphical Statistics*, 26(3):682–694, 2017.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Steffen L Lauritzen, Caroline Uhler, and Piotr Zwiernik. Maximum likelihood estimation in Gaussian models under total positivity. *The Annals of Statistics*, 47(4):1835–1863, 2019.

Fengpei Li and Ziping Zhao. Guaranteed robust large precision matrix estimation under t-distribution. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 3414–3419. IEEE, 2024.

Yupeng Li and Scott A Jackson. Gene network reconstruction by integration of prior biological knowledge. *G3: Genes, Genomes, Genetics*, 5(6):1075–1079, 2015.

Weidong Liu and Xi Luo. Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of Multivariate Analysis*, 135:153–162, 2015.

Zhaosong Lu. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19(4):1807–1827, 2009.

Zhaosong Lu. Adaptive first-order methods for general sparse inverse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2000–2016, 2010.

Harry M Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–79, 1952.

Benjamin M Marlin and Kevin P Murphy. Sparse Gaussian graphical models with unknown block structure. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 705–712, 2009.

Marie Maros and Gesualdo Scutari. DGD$^2$: A linearly convergent distributed algorithm for high-dimensional statistical recovery. *Advances in Neural Information Processing Systems*, 35:3475–3487, 2022.

Rahul Mazumder and Deepak K Agarwal. A flexible, scalable and efficient algorithmic framework for primal graphical lasso. *arXiv preprint arXiv:1110.5508*, 2011.

Rahul Mazumder and Trevor Hastie. The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125, 2012.

Cody Mazza-Anthony, Bogdan Mazoure, and Mark Coates. Learning Gaussian graphical models with ordered weighted $\ell_1$ regularization. *IEEE Transactions on Signal Processing*, 69:489–499, 2020.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

Zhaoshi Meng, Dennis Wei, Ami Wiesel, and Alfred O Hero. Marginal likelihoods for distributed parameter estimation of Gaussian graphical models. *IEEE Transactions on Signal Processing*, 62(20): 5425–5438, 2014.

Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Ensiyeh Nezakati and Eugen Pircalabelu. Unbalanced distributed estimation and inference for the precision matrix in Gaussian graphical models. *Statistics and Computing*, 33(2):47, 2023.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Benjamin Rolfs, Bala Rajaratnam, Dominique Guillot, Ian Wong, and Arian Maleki. Iterative thresholding algorithm for sparse inverse covariance estimation. *Advances in Neural Information Processing Systems*, 25, 2012.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Adam J Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

Srikanth Ryali, Tianwen Chen, Kaustubh Supekar, and Vinod Menon. Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage*, 59(4):3852–3861, 2012.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

Katya Scheinberg, Shiqian Ma, and Donald Goldfarb. Sparse inverse covariance selection via alternating linearization methods. *Advances in Nneural Information Processing Systems*, 23, 2010.

Martin Slawski and Matthias Hein. Estimation of positive definite $M$-matrices and structure learning for attractive Gaussian Markov random fields. *Linear Algebra and its Applications*, 473:145–179, 2015.

Ying Sun, Gesualdo Scutari, and Amir Daneshmand. Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation. *SIAM Journal on Optimization*, 32(2):354–385, 2022.

Mostafa Tavassolipour, Armin Karamzade, Reza Mirzaeifard, Seyed Abolfazl Motahari, and Mohammad-Taghi Manzuri Shalmani. Structure learning of sparse GGMs over multiple access networks. *IEEE Transactions on Communications*, 68(2):987–997, 2019.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Guan Peng Wang and Heng Jian Cui. Efficient distributed estimation of high-dimensional sparse precision matrix for transelliptical graphical models. *Acta Mathematica Sinica, English Series*, 37(5):689–706, 2021.

Stefanie Warnat-Herresthal, Hartmut Schultze, Krishnaprasad Lingadahalli Shastry, Sathyanarayanan Manamohan, Saikat Mukherjee, Vishesh Garg, Ravi Sarveswara, Kristian Händler, Peter Pickkers, N Ahmad Aziz, et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862):265–270, 2021.

Ami Wiesel and Alfred O Hero. Distributed covariance estimation in Gaussian graphical models. *IEEE Transactions on Signal Processing*, 60(1):211–220, 2011.

Wenfu Xia, Ziping Zhao, and Ying Sun. Distributed sparse covariance matrix estimation. In *2024 IEEE 13rd Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 1–5. IEEE, 2024.

Lin Xiao, Stephen Boyd, and Sanjay Lall. A scheme for robust distributed sensor fusion based on average consensus. In *IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005.*, pages 63–70. IEEE, 2005.

Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *The Journal of Machine Learning Research*, 11:2261–2286, 2010.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94 (1):19–35, 2007.

Xiaoming Yuan. Alternating direction method for covariance selection models. *Journal of Scientific Computing*, 51:261–273, 2012.

Teng Zhang and Hui Zou. Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika*, 101(1):103–120, 2014.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# Supplementary Materials for "Covariance Selection over Networks"

## Contents

# A NOTATIONS

The notation in this paper is mostly standard. $\mathbb{N}$ stands for the set of all natural numbers. $\mathbb{R}^n$ stands for $n$-dimensional real-valued vector space. $\mathbb{S}^n$ stands for the set of all $n \times n$ real symmetric matrices and $\mathbb{S}^n_{++}$ stands for the set of all $n \times n$ real symmetric positive definite matrices. For any matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \preceq \mathbf{B}$ stands for $\mathbf{B} - \mathbf{A}$ is positive semidefinite. $\mathbf{0}$ stands for all zero matrix and $\mathbf{1}$ stands for vector with all elements equal to one. $\mathbf{I}$ stands for the identity matrix. $\det(\cdot)$ stands for matrix determinant. $\|\cdot\|_F$ stands for Frobenius norm and $\|\cdot\|_1$ stands for the $\ell_1$-norm. $\langle\cdot\rangle$ stands for inner product and $\otimes$ stands for Kronecker product. $\nabla$ stands for the Jacobian matrix and $\nabla^2$ stands for the Hessian matrix. $A_{ij}$ stands for the $(i,j)$th entry of $\mathbf{A} \in \mathbb{R}^{n\times n}$. For any convex function $f$, $\partial f$ stands for the set of all subgradients. For simplicity, we define the following compact notations:

$$
\begin{aligned}
\mathbf{\Theta}^{(k)} &= \left[\mathbf{\Theta}_1^{(k)}; \mathbf{\Theta}_2^{(k)}; \ldots; \mathbf{\Theta}_m^{(k)}\right], & \mathbf{E}_\Theta^{(k)} &= \mathbf{\Theta}^{(k)} - \mathbf{1} \otimes \bar{\mathbf{\Theta}}^{(k)}, \\
\mathbf{Y}^{(k)} &= \left[\mathbf{Y}_1^{(k)}; \mathbf{Y}_2^{(k)}; \ldots; \mathbf{Y}_m^{(k)}\right], & \mathbf{E}_Y^{(k)} &= \mathbf{Y}^{(k)} - \mathbf{1} \otimes \bar{\mathbf{Y}}^{(k)}, \\
\mathbf{D}_i^{(k)} &= \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \mathbf{\Theta}_i^{(k)}, & \mathbf{D}^{(k)} &= \left[\mathbf{D}_1^{(k)}; \mathbf{D}_2^{(k)}; \ldots; \mathbf{D}_m^{(k)}\right], \\
\tilde{\mathbf{\Theta}}_i^{(k)} &= \mathbf{\Theta}_i^{(k)} + \alpha \mathbf{D}_i^{(k)}, & \mathbf{\Delta}_i^{(k)} &= \nabla F\left(\mathbf{\Theta}_i^{(k)}\right) - m\mathbf{Y}_i^{(k)}, \\
\mathbf{\Delta}^{(k)} &= \left[\mathbf{\Delta}_1^{(k)}; \mathbf{\Delta}_2^{(k)}; \ldots; \mathbf{\Delta}_m^{(k)}\right].
\end{aligned}
$$

# B PROOF OF THEOREM 3

We start by proving that $\left\|\mathbf{Y}_i^{(k)}\right\|_F$ is bounded for all $k \in \mathbb{N}$. First, we introduce the following lemma.

**Lemma 2.** *Assume that $f$ is $L$-smooth, we have*

$$
f\left(\sum_{i=1}^n a_i \mathbf{X}_i\right) \geq \sum_{i=1}^n a_i f\left(\mathbf{X}_i\right) - \frac{L}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_i a_j \left\|\mathbf{X}_i - \mathbf{X}_j\right\|_F^2,
$$

*where $\sum_{i=1}^n a_i = 1$ and $a_i \geq 0$ for all $i$.*

The proof for Lemma 2 follows a similar approach to that of Lemma II.1 (a result of weakly convex functions) in (Chen et al., 2021) and is therefore omitted. Based on Lemma 2, we can prove that $\left\|\mathbf{Y}_i^{(k)}\right\|_F$ are bounded.

**Lemma 3.** *For every $\left(\mathbf{\Theta}^{(k)}, \mathbf{E}_Y^{(k)}\right) \in \mathcal{A}$, $k \in \mathbb{N}$, we have*

$$
\left\|\mathbf{Y}_i^{(k)}\right\|_F \leq \sqrt{\sum_{i=1}^m u_{\nabla f_i} + dR_Y^2},
$$

*where $u_{\nabla f_i} = \frac{n_i^2}{N^2}\left(\left\|\mathbf{S}_i\right\|_F + \frac{\sqrt{d}}{R_\Theta}\right)^2$.*

*Proof.* When we set $\mathbf{Y}^{(0)} = \nabla f^{(0)}$, the update rule for $\bar{\mathbf{Y}}$ can be expressed as

$$
\bar{\mathbf{Y}}^{(k+1)} = \frac{1}{m} \sum_{i=1}^m \nabla f_i\left(\mathbf{\Theta}_i^{(k+1)}\right). \tag{7}
$$

Applying Lemma 2, equation (7), and the 2-Lipschitz smoothness of $\|\cdot\|_F^2$, we obtain

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\left\|\mathbf{Y}_i^{(k)}\right\|_F^2 &\le \left\|\frac{1}{m}\sum_{i=1}^{m}\mathbf{Y}_i^{(k)}\right\|_F^2+\frac{1}{m^2}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\left\|\mathbf{Y}_i^{(k)}-\mathbf{Y}_j^{(k)}\right\|_F^2\\ &=\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i\left(\mathbf{\Theta}_i^{(k)}\right)\right\|_F^2+\frac{1}{m^2}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\left\|\mathbf{Y}_i^{(k)}-\mathbf{Y}_j^{(k)}\right\|_F^2\\ &\le \frac{1}{m}\sum_{i=1}^{m}\left\|\nabla f_i\left(\mathbf{\Theta}_i^{(k)}\right)\right\|_F^2+\frac{1}{m}\left\|\mathbf{E}_Y^{(k)}\right\|_F^2.\end{aligned}$$

Since $U$ is coercive and $\left(\mathbf{\Theta}^{(k)},\mathbf{E}_Y^{(k)}\right)\in\mathcal{A}$, we have

$$\left\|\nabla f_i\left(\mathbf{\Theta}_i^{(k)}\right)\right\|_F\le\max_{(\mathbf{\Theta},\mathbf{E}_Y)\in\mathcal{A}}\|\nabla f_i\left(\mathbf{\Theta}_i\right)\|_F\le\frac{n_i}{N}\|\mathbf{S}_i\|_F+\frac{n_i}{N}\max_{(\mathbf{\Theta},\mathbf{E}_Y)\in\mathcal{A}}\left\|\mathbf{\Theta}_i^{-1}\right\|_F\le\frac{n_i}{N}\left(\|\mathbf{S}_i\|_F+\frac{\sqrt{d}}{\underline{R}_\Theta}\right).$$

Moreover, $\left\|\mathbf{E}_Y^{(k)}\right\|_F^2\le\|R_Y\mathbf{I}\|_F^2=dR_Y^2$ due to $\left(\mathbf{\Theta}^{(k)},\mathbf{E}_Y^{(k)}\right)\in\mathcal{A}$. Therefore, we obtain

$$\left\|\mathbf{Y}_i^{(k)}\right\|_F\le\sqrt{\sum_{i=1}^{m}\left\|\mathbf{Y}_i^{(k)}\right\|_F^2}\le\sqrt{\sum_{i=1}^{m}\frac{n_i^2}{N^2}\left(\|\mathbf{S}_i\|_F+\frac{\sqrt{d}}{\underline{R}_\Theta}\right)^2+dR_Y^2}.$$

□

Based on Lemma 3, we show that if $\left(\mathbf{\Theta}^{(k)},\mathbf{E}_Y^{(k)}\right)\in\mathcal{A}$ and $\tau$ is chosen sufficiently large, $\mathbf{\Theta}^{(k+1)}$ generated by NetGGM must be positive definite.

**Lemma 4.** *If* $\left(\mathbf{\Theta}^{(k)},\mathbf{E}_Y^{(k)}\right)\in\mathcal{A}$, *and*

$$\tau\ge\frac{2\left(m\sqrt{\sum_{i=1}^{m}u_{\nabla f_i}+dR_Y^2}+\lambda d\right)}{\underline{R}_\Theta},\tag{8}$$

*then* $\left\{\mathbf{\Theta}^{\left(k+\frac{1}{2}\right)},\mathbf{\Theta}^{(k+1)}\right\}\in\mathcal{B}$, *where*

$$\mathcal{B}=\left\{\mathbf{\Theta}\mid\frac{\underline{R}_\Theta}{2}\mathbf{I}\preceq\mathbf{\Theta}_i\preceq\left(\overline{R}_\Theta+\frac{\underline{R}_\Theta}{2}\right)\mathbf{I},i=1,2,\ldots,m\right\}.$$

*Proof.* Since $\tilde{F}_i^{(k)}$ is $\tau$-strongly convex and $\lambda\|\mathbf{\Theta}\|_1$ is convex, for any $\boldsymbol{\zeta}\in\lambda\partial\left\|\mathbf{\Theta}_i^{(k)}\right\|_1$ and any $\mathbf{\Phi}\succ\mathbf{0}$, we have

$$\begin{aligned}&\tilde{F}_i^{(k)}\left(\mathbf{\Phi}\right)+\lambda\|\mathbf{\Phi}\|_1\\ \ge&\tilde{F}_i^{(k)}\left(\mathbf{\Theta}_i^{(k)}\right)+\lambda\left\|\mathbf{\Theta}_i^{(k)}\right\|_1+\left\langle\nabla\tilde{F}_i^{(k)}\left(\mathbf{\Theta}_i^{(k)}\right)+\boldsymbol{\zeta},\mathbf{\Phi}-\mathbf{\Theta}_i^{(k)}\right\rangle+\frac{\tau}{2}\left\|\mathbf{\Phi}-\mathbf{\Theta}_i^{(k)}\right\|_F^2\\ =&\tilde{F}_i^{(k)}\left(\mathbf{\Theta}_i^{(k)}\right)+\lambda\left\|\mathbf{\Theta}_i^{(k)}\right\|_1-\frac{1}{2\tau}\left\|\nabla\tilde{F}_i^{(k)}\left(\mathbf{\Theta}_i^{(k)}\right)+\boldsymbol{\zeta}\right\|_F^2+\frac{\tau}{2}\left\|\frac{1}{\tau}\left(\nabla\tilde{F}_i^{(k)}\left(\mathbf{\Theta}_i^{(k)}\right)+\boldsymbol{\zeta}\right)+\mathbf{\Phi}-\mathbf{\Theta}_i^{(k)}\right\|_F^2\\ \ge&\tilde{F}_i^{(k)}\left(\mathbf{\Theta}_i^{(k)}\right)+\lambda\left\|\mathbf{\Theta}_i^{(k)}\right\|_1-\frac{1}{2\tau}\left\|\nabla\tilde{F}_i^{(k)}\left(\mathbf{\Theta}_i^{(k)}\right)+\boldsymbol{\zeta}\right\|_F^2.\end{aligned}\tag{9}$$

When choosing $\mathbf{\Phi}=\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}$, (9) becomes

$$\frac{1}{2\tau}\left\|\nabla\tilde{F}_i^{(k)}\left(\mathbf{\Theta}_i^{(k)}\right)+\boldsymbol{\zeta}\right\|_F^2\ge\tilde{F}_i^{(k)}\left(\mathbf{\Theta}_i^{(k)}\right)+\lambda\left\|\mathbf{\Theta}_i^{(k)}\right\|_1-\tilde{F}_i^{(k)}\left(\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)-\lambda\left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_1.\tag{10}$$

On the other hand, according to the first-order optimality condition, there exists $\boldsymbol{\xi} \in \lambda\partial \left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_1$ such that

$$-\left\langle \nabla \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)+\boldsymbol{\xi}, \mathbf{D}_i^{(k)}\right\rangle \geq 0. \tag{11}$$

In addition, because $\tilde{F}_i^{(k)}$ is $\tau$-strongly convex and $\lambda\|\cdot\|_1$ is convex, we have

$$\begin{aligned}\tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{(k)}\right)+\lambda\left\|\boldsymbol{\Theta}_i^{(k)}\right\|_1 \geq & \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)+\lambda\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_1 \\ & -\left\langle \nabla \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)+\boldsymbol{\xi}, \mathbf{D}_i^{(k)}\right\rangle+\frac{\tau}{2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2 .\end{aligned} \tag{12}$$

Plugging (11) into (12), we can get

$$\tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{(k)}\right)+\lambda\left\|\boldsymbol{\Theta}_i^{(k)}\right\|_1-\tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)-\lambda\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_1 \geq \frac{\tau}{2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2 . \tag{13}$$

Combining (10) and (13) and transforming the inequality leads to

$$\begin{aligned}\left\|\mathbf{D}_i^{(k)}\right\|_F & \leq \frac{1}{\tau}\left\|\nabla \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{(k)}\right)+\boldsymbol{\zeta}\right\|_F \\ & \leq \frac{1}{\tau}\left\|\nabla \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{(k)}\right)\right\|_F+\frac{1}{\tau} \max_{\boldsymbol{\eta} \in \lambda\partial\left\|\boldsymbol{\Theta}_i^{(k)}\right\|_1}\|\boldsymbol{\eta}\|_F .\end{aligned}$$

Since each entry of $\partial\lambda\left\|\boldsymbol{\Theta}_i^{(k)}\right\|_1$ falls within $[-\lambda, \lambda]$ and there are $d^2$ entries in total, it follows that $\max_{\boldsymbol{\eta} \in \lambda\partial\left\|\boldsymbol{\Theta}_i^{(k)}\right\|_1}\|\boldsymbol{\eta}\|_F \leq \lambda d$. Consequently, we have

$$\left\|\mathbf{D}_i^{(k)}\right\|_F \leq \frac{m}{\tau}\left\|\mathbf{Y}_i^{(k)}\right\|_F+\frac{1}{\tau}\lambda d. \tag{14}$$

Based on Lemma 3 and (8), we further derive

$$\frac{m}{\tau}\left\|\mathbf{Y}_i^{(k)}\right\|_F+\frac{1}{\tau}\lambda d \leq \frac{\underline{R}_\Theta}{2} \tag{15}$$

for any $k \in \mathbb{N}$. Combining (14) and (15), we conclude

$$\left\|\mathbf{D}_i^{(k)}\right\|_F \leq \frac{\underline{R}_\Theta}{2}.$$

Therefore, we have $-\frac{\underline{R}_\Theta}{2}\mathbf{I} \preceq \mathbf{D}_i^{(k)} \preceq \frac{\underline{R}_\Theta}{2}\mathbf{I}$. Recall that $\boldsymbol{\Theta}_i^{(k)} \in \left\{\boldsymbol{\Theta} \mid \underline{R}_\Theta\mathbf{I} \preceq \boldsymbol{\Theta} \preceq \overline{R}_\Theta\mathbf{I}\right\}$, so we have $\boldsymbol{\Theta}_i^{(k)} \in \mathcal{B}$ and $\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)} \in \mathcal{B}$. In addition, due to Step 2 of NetGGM, $w_{ij} \in [0,1], i, j=1,2,\ldots,m$, $\sum_{j=1}^m w_{ij}=1$, and $\alpha \leq 1$, we have $\boldsymbol{\Theta}_i^{(k+1)} \in \mathcal{B}$. □

Lemma 4 reveals that if $\boldsymbol{\Theta}^{(k)} \in \mathcal{A}$ and $\tau$ satisfying (8) for all $i=1,2,\ldots,m$, then both $\boldsymbol{\Theta}_i^{(k)}$ and $\boldsymbol{\Theta}_i^{(k+1)} \in \mathcal{B}$. Since $\mathcal{B}$ is a compact set, we have $f_i$ is $L_i$-smooth and $F$ is $\mu$-strongly convex on set $\mathcal{B}$, i.e., for any $\boldsymbol{\Theta}, \boldsymbol{\Phi} \in \mathcal{B}$ we have

$$\|\nabla f_i(\boldsymbol{\Theta})-\nabla f_i(\boldsymbol{\Phi})\|_F \leq L_i\|\boldsymbol{\Theta}-\boldsymbol{\Phi}\|_F ,$$

and

$$\|\nabla F(\boldsymbol{\Theta})-\nabla F(\boldsymbol{\Phi})\|_F \geq \mu\|\boldsymbol{\Theta}-\boldsymbol{\Phi}\|_F .$$

Then we construct a coercive function $V$ and prove, given two iterates in $\mathcal{B}$, the decrease in the value of function $V$ is bounded. We first bound $p$ at the $(k+1)$-th iteration by the result of the $k$-th iteration. We have the following upper bound of the optimality gap with respect to consensus errors.

**Lemma 5.** *Based on Lemma 4, there holds*

$$\sum_{i=1}^{m} U\left(\boldsymbol{\Theta}_i^{(k+1)}\right) \leq \sum_{i=1}^{m} U\left(\boldsymbol{\Theta}_i^{(k)}\right) + \frac{\alpha}{2}\epsilon_p^{-1} m^2 \left(4L_{\max}^2 \left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2 + 2\left\|\mathbf{E}_Y^{(k)}\right\|_F^2\right) - \alpha\left(\tau - \frac{2\alpha}{\underline{R}_\Theta^2} - \frac{1}{2}\epsilon_p\right)\left\|\mathbf{D}^{(k)}\right\|_F^2, \tag{16}$$

*where* $\epsilon_p > 0$ *and* $\alpha < \frac{\underline{R}_\Theta^2}{2}\tau$.

*Proof.* Consider the Taylor expansion of $F$

$$\begin{aligned} F\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right) &= F\left(\boldsymbol{\Theta}_i^{(k)}\right) + \left\langle \nabla F\left(\boldsymbol{\Theta}_i^{(k)}\right), \alpha\mathbf{D}_i^{(k)}\right\rangle + \frac{1}{2}\mathrm{vec}\left(\alpha\mathbf{D}_i^{(k)}\right)^\top \nabla^2 F\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right)\mathrm{vec}\left(\alpha\mathbf{D}_i^{(k)}\right) \\ &= F\left(\boldsymbol{\Theta}_i^{(k)}\right) + \left\langle \boldsymbol{\Delta}_i^{(k)}, \alpha\mathbf{D}_i^{(k)}\right\rangle + \left\langle m\mathbf{Y}_i^{(k)}, \alpha\mathbf{D}_i^{(k)}\right\rangle + \frac{1}{2}\mathrm{vec}\left(\alpha\mathbf{D}_i^{(k)}\right)^\top \nabla^2 F\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right)\mathrm{vec}\left(\alpha\mathbf{D}_i^{(k)}\right). \end{aligned} \tag{17}$$

Because $\widetilde{F}_i$ is strongly convex with $\tau$ and $\lambda\|\cdot\|_1$ is convex, according to the first-order optimality condition, we have

$$\lambda\left\|\boldsymbol{\Theta}_i^{(k)}\right\|_1 - \lambda\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_1 \geq \left\langle m\mathbf{Y}_i^{(k)}, \mathbf{D}_i^{(k)}\right\rangle + \tau\left\|\mathbf{D}_i^{(k)}\right\|_F^2. \tag{18}$$

Using the convexity of $G$, we have

$$\lambda\left\|\tilde{\boldsymbol{\Theta}}_i^{(k)}\right\|_1 = \lambda\left\|\alpha\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)} + (1-\alpha)\boldsymbol{\Theta}_i^{(k)}\right\|_1 \leq \alpha\lambda\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_1 + (1-\alpha)\lambda\left\|\boldsymbol{\Theta}_i^{(k)}\right\|_1. \tag{19}$$

Substituting (18) and (19) into (17), we have

$$\begin{aligned} F\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right) \leq\ & F\left(\boldsymbol{\Theta}_i^{(k)}\right) + \left\langle \boldsymbol{\Delta}_i^{(k)}, \alpha\mathbf{D}_i^{(k)}\right\rangle + \lambda\left\|\boldsymbol{\Theta}_i^{(k)}\right\|_1 - \lambda\left\|\tilde{\boldsymbol{\Theta}}_i^{(k)}\right\|_1 - \alpha\tau\left\|\mathbf{D}_i^{(k)}\right\|_F^2 \\ &+ \frac{1}{2}\mathrm{vec}\left(\alpha\mathbf{D}_i^{(k)}\right)^\top \nabla^2 F\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right)\mathrm{vec}\left(\alpha\mathbf{D}_i^{(k)}\right). \end{aligned} \tag{20}$$

According to Lemma 4, we have $\tilde{\boldsymbol{\Theta}}_i^{(k)} \succeq \frac{\underline{R}_\Theta}{2}\mathbf{I}$, and hence

$$\nabla^2 F\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right) = \left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right)^{-1} \otimes \left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right)^{-1} \preceq \frac{4}{\underline{R}_\Theta^2}\mathbf{I}. \tag{21}$$

Substituting (21) into (20), we have

$$\begin{aligned} F\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right) &\leq F\left(\boldsymbol{\Theta}_i^{(k)}\right) + \left\langle \boldsymbol{\Delta}_i^{(k)}, \alpha\mathbf{D}_i^{(k)}\right\rangle + \lambda\left\|\boldsymbol{\Theta}_i^{(k)}\right\|_1 - \lambda\left\|\tilde{\boldsymbol{\Theta}}_i^{(k)}\right\|_1 - \alpha\tau\left\|\mathbf{D}_i^{(k)}\right\|_F^2 + \frac{2\alpha^2}{\underline{R}_\Theta^2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2 \\ &\leq F\left(\boldsymbol{\Theta}_i^{(k)}\right) + \lambda\left\|\boldsymbol{\Theta}_i^{(k)}\right\|_1 - \lambda\left\|\tilde{\boldsymbol{\Theta}}_i^{(k)}\right\|_1 - \alpha\left(\tau - \frac{2\alpha}{\underline{R}_\Theta^2}\right)\left\|\mathbf{D}_i^{(k)}\right\|_F^2 + \alpha\left\|\mathbf{D}_i^{(k)}\right\|\left\|\boldsymbol{\Delta}_i^{(k)}\right\|_F, \end{aligned}$$

which equals to

$$U\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right) \leq U\left(\boldsymbol{\Theta}_i^{(k)}\right) - \alpha\left(\tau - \frac{2\alpha}{\underline{R}_\Theta^2}\right)\left\|\mathbf{D}_i^{(k)}\right\|_F^2 + \alpha\left\|\mathbf{D}_i^{(k)}\right\|_F\left\|\boldsymbol{\Delta}_i^{(k)}\right\|_F. \tag{22}$$

Invoking the convexity of $U$ and the doubly stochasticity of $\mathbf{W}$, we can bound $\sum_{i=1}^m U\left(\boldsymbol{\Theta}_i^{(k)}\right)$ as

$$\sum_{i=1}^{m} U\left(\boldsymbol{\Theta}_i^{(k+1)}\right) = \sum_{i=1}^{m} U\left(\sum_{j=1}^{m} W_{ij}\tilde{\boldsymbol{\Theta}}_j^{(k)}\right) \leq \sum_{i=1}^{m}\sum_{j=1}^{m} W_{ij} U\left(\tilde{\boldsymbol{\Theta}}_j^{(k)}\right) = \sum_{i=1}^{m} U\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right). \tag{23}$$

We can now substitute (22) into (23) and get

$$\sum_{i=1}^{m} U\left(\boldsymbol{\Theta}_i^{(k+1)}\right) \leq \sum_{i=1}^{m} U\left(\boldsymbol{\Theta}_i^{(k)}\right) + \sum_{i=1}^{m}\left(\alpha\left\|\mathbf{D}_i^{(k)}\right\|_F\left\|\boldsymbol{\Delta}_i^{(k)}\right\|_F - \alpha\left(\tau - \frac{2\alpha}{\underline{R}_\Theta^2}\right)\left\|\mathbf{D}_i^{(k)}\right\|_F^2\right).$$

Using Young's inequality, we have

$$\alpha \left\|\mathbf{D}_i^{(k)}\right\|_F \left\|\mathbf{\Delta}_i^{(k)}\right\|_F \leq \frac{\alpha}{2}\epsilon_p \left\|\mathbf{D}_i^{(k)}\right\|_F^2 + \frac{\alpha}{2}\epsilon_p^{-1} \left\|\mathbf{\Delta}_i^{(k)}\right\|_F^2, \tag{24}$$

where $\epsilon_p > 0$. Therefore, we have

$$\begin{aligned}\sum_{i=1}^m U\left(\mathbf{\Theta}_i^{(k+1)}\right) &\leq \sum_{i=1}^m U\left(\mathbf{\Theta}_i^{(k)}\right) + \sum_{i=1}^m \left(\frac{\alpha}{2}\epsilon_p^{-1}\left\|\mathbf{\Delta}_i^{(k)}\right\|_F^2 - \alpha\left(\tau - \frac{2\alpha}{\underline{R}_\Theta^2} - \frac{1}{2}\epsilon_p\right)\left\|\mathbf{D}_i^{(k)}\right\|_F^2\right) \\ &= \sum_{i=1}^m U\left(\mathbf{\Theta}_i^{(k)}\right) + \frac{\alpha}{2}\epsilon_p^{-1}\left\|\mathbf{\Delta}^{(k)}\right\|_F^2 - \alpha\left(\tau - \frac{2\alpha}{\underline{R}_\Theta^2} - \frac{1}{2}\epsilon_p\right)\left\|\mathbf{D}^{(k)}\right\|_F^2, \end{aligned} \tag{25}$$

and we choose $\epsilon_p$ such that

$$\tau - \frac{2\alpha}{\underline{R}_\Theta^2} - \frac{1}{2}\epsilon_p > 0.$$

Then, we bound $\left\|\mathbf{\Delta}^{(k)}\right\|_F^2$ in terms of the consensus errors $\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2$ and $\left\|\mathbf{E}_Y^{(k)}\right\|_F^2$. Recall the definition of the tracking error $\left\|\mathbf{\Delta}^{(k)}\right\|_F^2$ and (7), we have

$$\begin{aligned}\left\|\mathbf{\Delta}^{(k)}\right\|_F^2 &= \sum_{i=1}^m \left\|\nabla F\left(\mathbf{\Theta}_i^{(k)}\right) - m\bar{\mathbf{Y}}^{(k)} + m\bar{\mathbf{Y}}^{(k)} - m\mathbf{Y}_i^{(k)}\right\|_F^2 \\ &= \sum_{i=1}^m \left\|\sum_{j=1}^m \nabla f_j\left(\mathbf{\Theta}_i^{(k)}\right) - \sum_{j=1}^m \nabla f_j\left(\mathbf{\Theta}_j^{(k)}\right) + m\bar{\mathbf{Y}}^{(k)} - m\mathbf{Y}_i^{(k)}\right\|_F^2 \\ &\leq 2m\sum_{i=1}^m\sum_{j=1}^m \left\|\nabla f_j\left(\mathbf{\Theta}_i^{(k)}\right) - \nabla f_j\left(\mathbf{\Theta}_j^{(k)}\right)\right\|_F^2 + 2m^2\left\|\mathbf{E}_Y^{(k)}\right\|_F^2. \end{aligned}$$

Recall the Lipschitz smoothness of $f_i$, $i = 1, 2, \ldots, m$, we have

$$\begin{aligned}\left\|\mathbf{\Delta}^{(k)}\right\|_F^2 &\leq 2m\sum_{i=1}^m\sum_{j=1}^m L_j^2\left\|\mathbf{\Theta}_i^{(k)} - \mathbf{\Theta}_j^{(k)}\right\|_F^2 + 2m^2\left\|\mathbf{E}_Y^{(k)}\right\|_F^2 \\ &= 4m^2L_{\max}^2\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2 + 2m^2\left\|\mathbf{E}_Y^{(k)}\right\|_F^2. \end{aligned} \tag{26}$$

Substituting (26) into (25), we obtain the desired result (16). □

Subsequently, we bound the consensus errors with the following lemma from (Sun et al., 2022).

**Lemma 6.** *(Sun et al., 2022) The disagreements* $\left\|\mathbf{E}_\Theta^{(k)}\right\|_F$ *and* $\left\|\mathbf{E}_Y^{(k)}\right\|_F$ *are bounded by*

$$\left\|\mathbf{E}_\Theta^{(k+1)}\right\|_F \leq \rho\left\|\mathbf{E}_\Theta^{(k)}\right\|_F + \alpha\rho\left\|\mathbf{D}^{(k)}\right\|_F, \tag{27}$$

*and*

$$\left\|\mathbf{E}_Y^{(k+1)}\right\|_F \leq \rho\left\|\mathbf{E}_Y^{(k)}\right\|_F + 2L_{\max}\rho\left\|\mathbf{E}_\Theta^{(k)}\right\|_F + \alpha L_{\max}\rho\left\|\mathbf{D}^{(k)}\right\|_F, \tag{28}$$

*where* $\rho = \left\|\mathbf{W}\otimes\mathbf{I} - \frac{1}{m}\left(\mathbf{1}\mathbf{1}^\top\otimes\mathbf{I}\right)\right\|_2$.

Based on Lemmas 5 and 6, we can prove Theorem 3 and give the specific expression of the parameters.

**Theorem 7.** *Assume that Assumptions 1 and 2 are satisfied. Based on Lemma 1, When* $\tau \geq \underline{\tau}_1$ *and* $\alpha \in (0, \bar{\alpha}_1)$, *where*

$$\underline{\tau}_1 = \frac{2\left(m\sqrt{\sum_{i=1}^m u_{\nabla f_i} + dR_Y^2} + \lambda d\right)}{\underline{R}_\Theta},$$

$$\bar{\alpha}_1 = \min\left\{\frac{\underline{R}_\Theta^2}{2}\tau, \frac{\tau\left(\sqrt{\frac{1}{\underline{R}_\Theta^4}+2\left(\frac{4m^2L_{\max}^2\rho^2\left(1+\epsilon_d^{-1}\right)}{1-\rho^2(1+\epsilon_d)}+\frac{8m^2L_{\max}^2\rho^4\left(1+\epsilon_d^{-1}\right)^2}{(1-\rho^2(1+\epsilon_d))^2}\right)}-\frac{1}{\underline{R}_\Theta^2}\right)}{2\left(\frac{4m^2L_{\max}^2\rho^2\left(1+\epsilon_d^{-1}\right)}{1-\rho^2(1+\epsilon_d)}+\frac{8m^2L_{\max}^2\rho^4\left(1+\epsilon_d^{-1}\right)^2}{(1-\rho^2(1+\epsilon_d))^2}\right)}, 1\right\}, \tag{29}$$

*$\epsilon_d > 0$, and $1-\rho^2(1+\epsilon_d) > 0$, then for $\left\{\mathbf{\Theta}^{(k)}, \mathbf{Y}^{(k)}\right\}_{k\in\mathbb{N}}$ obtained by NetGGM, we have*

$$V\left(\mathbf{\Theta}^{(k+1)}, \mathbf{Y}^{(k+1)}\right) \le V\left(\mathbf{\Theta}^{(k)}, \mathbf{Y}^{(k)}\right) - \beta\sum_{i=1}^{m}\left\|\mathbf{D}_i^{(k)}\right\|_F^2, \tag{30}$$

*where*

$$\begin{aligned} V\left(\mathbf{\Theta}^{(k)}, \mathbf{E}_Y^{(k)}\right) =& \sum_{i=1}^{m} U\left(\mathbf{\Theta}_i^{(k)}\right) + \frac{\alpha\left(\tau-\frac{2\alpha}{\underline{R}_\Theta^2}\right)^{-1}m^2}{1-\rho^2(1+\epsilon_d)}\left\|\mathbf{E}_Y^{(k)}\right\|_F^2 \\ &+ \frac{2\alpha\left(\tau-\frac{2\alpha}{\underline{R}_\Theta^2}\right)^{-1}m^2L_{\max}^2\left(4\rho^2\left(1+\epsilon_d^{-1}\right)+1-\rho^2(1+\epsilon_d)\right)}{\left(1-\rho^2(1+\epsilon_d)\right)^2}\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2, \end{aligned}$$

*and*

$$\beta = \alpha\left(\frac{\tau}{2}-\frac{\alpha}{\underline{R}_\Theta^2}-\frac{4\alpha^2\left(\tau-\frac{2\alpha}{\underline{R}_\Theta^2}\right)^{-1}m^2L_{\max}^2\rho^2\left(1+\epsilon_d^{-1}\right)}{1-\rho^2(1+\epsilon_d)}-\frac{8\alpha^2\left(\tau-\frac{2\alpha}{\underline{R}_\Theta^2}\right)^{-1}m^2L_{\max}^2\rho^4\left(1+\epsilon_d^{-1}\right)^2}{\left(1-\rho^2(1+\epsilon_d)\right)^2}\right) \ge 0.$$

*Proof.* Squaring both sides of inequality (27) and utilizing Young's inequality provides

$$\begin{aligned} \left\|\mathbf{E}_\Theta^{(k+1)}\right\|_F^2 &\le \rho^2\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2 + \alpha^2\rho^2\left\|\mathbf{D}^{(k)}\right\|_F^2 + 2\alpha\rho^2\left\|\mathbf{E}_\Theta^{(k)}\right\|_F\left\|\mathbf{D}^{(k)}\right\|_F \\ &\le \rho^2(1+\epsilon_d)\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2 + \alpha^2\rho^2\left(1+\epsilon_d^{-1}\right)\left\|\mathbf{D}^{(k)}\right\|_F^2. \end{aligned} \tag{31}$$

Similarly, we have

$$\begin{aligned} \left\|\mathbf{E}_Y^{(k+1)}\right\|_F^2 &\le \rho^2\left\|\mathbf{E}_Y^{(k)}\right\|_F^2 + \left(2L_{\max}\rho\left\|\mathbf{E}_\Theta^{(k)}\right\|_F + \alpha L_{\max}\rho\left\|\mathbf{D}^{(k)}\right\|_F\right)^2 \\ &\quad + 2\rho\left\|\mathbf{E}_Y^{(k)}\right\|_F\left(2L_{\max}\rho\left\|\mathbf{E}_\Theta^{(k)}\right\|_F + \alpha L_{\max}\rho\left\|\mathbf{D}^{(k)}\right\|_F\right) \\ &\le \rho^2(1+\epsilon_d)\left\|\mathbf{E}_Y^{(k)}\right\|_F^2 + L_{\max}^2\rho^2\left(1+\epsilon_d^{-1}\right)\left(2\left\|\mathbf{E}_\Theta^{(k)}\right\|_F + \alpha\left\|\mathbf{D}^{(k)}\right\|_F\right)^2 \\ &\le \rho^2(1+\epsilon_d)\left\|\mathbf{E}_Y^{(k)}\right\|_F^2 + 2L_{\max}^2\rho^2\left(1+\epsilon_d^{-1}\right)\left(4\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2 + \alpha^2\left\|\mathbf{D}^{(k)}\right\|_F^2\right) \\ &= \rho^2(1+\epsilon_d)\left\|\mathbf{E}_Y^{(k)}\right\|_F^2 + 8L_{\max}^2\rho^2\left(1+\epsilon_d^{-1}\right)\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2 + 2L_{\max}^2\rho^2\left(1+\epsilon_d^{-1}\right)\alpha^2\left\|\mathbf{D}^{(k)}\right\|_F^2. \end{aligned} \tag{32}$$

Multiplying $\frac{2\alpha\epsilon_p^{-1}L_{\max}^2\left(4\rho^2\left(1+\epsilon_d^{-1}\right)+1-\rho^2(1+\epsilon_d)\right)}{\left(1-\rho^2(1+\epsilon_d)\right)^2}$ and add $\frac{2\alpha\epsilon_p^{-1}L_{\max}^2\left(4\rho^2\left(1+\epsilon_d^{-1}\right)+1-\rho^2(1+\epsilon_d)\right)}{1-\rho^2(1+\epsilon_d)}\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2$ on both sides of (31) leads to

$$\begin{aligned} &\frac{2\alpha\epsilon_p^{-1}m^2L_{\max}^2\left(4\rho^2\left(1+\epsilon_d^{-1}\right)+1-\rho^2(1+\epsilon_d)\right)}{\left(1-\rho^2(1+\epsilon_d)\right)^2}\left\|\mathbf{E}_\Theta^{(k+1)}\right\|_F^2 \\ &+\frac{2\alpha\epsilon_p^{-1}m^2L_{\max}^2\left(4\rho^2\left(1+\epsilon_d^{-1}\right)+1-\rho^2(1+\epsilon_d)\right)}{1-\rho^2(1+\epsilon_d)}\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2 \\ \le&\frac{2\alpha\epsilon_p^{-1}m^2L_{\max}^2\left(4\rho^2\left(1+\epsilon_d^{-1}\right)+1-\rho^2(1+\epsilon_d)\right)}{\left(1-\rho^2(1+\epsilon_d)\right)^2}\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2 \\ &+\frac{2\alpha^3\epsilon_p^{-1}m^2L_{\max}^2\left(4\rho^2\left(1+\epsilon_d^{-1}\right)+1-\rho^2(1+\epsilon_d)\right)\rho^2\left(1+\epsilon_d^{-1}\right)}{\left(1-\rho^2(1+\epsilon_d)\right)^2}\left\|\mathbf{D}^{(k)}\right\|_F^2, \end{aligned} \tag{33}$$

while multiplying $\frac{\alpha\epsilon_p^{-1}m^2}{1-\rho^2(1+\epsilon_d)}$ and add $\alpha\epsilon_p^{-1}m^2\left\|\mathbf{E}_Y^{(k)}\right\|_F^2$ on both sides of (32) leads to

$$\begin{aligned}&\frac{\alpha\epsilon_p^{-1}m^2}{1-\rho^2\left(1+\epsilon_d\right)}\left\|\mathbf{E}_Y^{(k+1)}\right\|_F^2+\alpha\epsilon_p^{-1}m^2\left\|\mathbf{E}_Y^{(k)}\right\|_F^2 \\ \leq&\frac{\alpha\epsilon_p^{-1}m^2}{1-\rho^2\left(1+\epsilon_d\right)}\left\|\mathbf{E}_Y^{(k)}\right\|_F^2+\frac{8\alpha\epsilon_p^{-1}m^2L_{\max}^2\rho^2\left(1+\epsilon_d^{-1}\right)}{1-\rho^2\left(1+\epsilon_d\right)}\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2+\frac{2\alpha\epsilon_p^{-1}m^2L_{\max}^2\rho^2\left(1+\epsilon_d^{-1}\right)\alpha^2}{1-\rho^2\left(1+\epsilon_d\right)}\left\|\mathbf{D}^{(k)}\right\|_F^2.\end{aligned} \tag{34}$$

Summing (16), (33), and (34), and choosing $\epsilon_p=\tau-\frac{2\alpha}{\underline{R}_\Theta^2}$ leads to the desired result (30). □

## C PROOF OF THEOREM 4

**Theorem 8.** *In the same setting as Theorem 7, if the proximal parameter is sufficiently large satisfying* $\tau\geq\max\left\{\underline{\tau}_1,\frac{4}{\underline{R}_\Theta^2}\right\}$ *and the step size satisfies* $\alpha\in(0,\bar{\alpha}_1]$*, we have*

$$\left\|\mathbf{\Theta}^{(k+1)}-\mathbf{1}\otimes\hat{\mathbf{\Theta}}\right\|_F^2\leq\left(1-\frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\alpha\right)\left\|\mathbf{\Theta}^{(k)}-\mathbf{1}\otimes\hat{\mathbf{\Theta}}\right\|_F^2+\frac{\alpha m^2}{\epsilon_\Theta\left(\tau-\epsilon_\Theta\right)}\left(4L_{\max}^2\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2+2\left\|\mathbf{E}_Y^{(k)}\right\|_F^2\right). \tag{35}$$

*where* $\epsilon_\Theta\in(0,\mu)$.

*Proof.* Recall that $f_i$ is $L_i$-smooth and $F$ is $\mu$-strongly convex on set $\mathcal{B}$, we have

$$F\left(\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)-F\left(\mathbf{\Theta}_i^{(k)}\right)-\left\langle\nabla F\left(\mathbf{\Theta}_i^{(k)}\right),\mathbf{D}_i^{(k)}\right\rangle\leq\frac{L}{2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2$$

and

$$F\left(\hat{\mathbf{\Theta}}\right)-F\left(\mathbf{\Theta}_i^{(k)}\right)-\left\langle\nabla F\left(\mathbf{\Theta}_i^{(k)}\right),\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{(k)}\right\rangle\geq\frac{\mu}{2}\left\|\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{(k)}\right\|_F^2,$$

where $L=\sum_{i=1}^m L_i$. Summing these two inequalities, we have

$$0\leq F\left(\hat{\mathbf{\Theta}}\right)-F\left(\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)+\frac{L}{2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2-\frac{\mu}{2}\left\|\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{(k)}\right\|_F^2-\left\langle\nabla F\left(\mathbf{\Theta}_i^{(k)}\right),\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\rangle. \tag{36}$$

Since $\tilde{F}_i$ is strongly convex with $\tau$ and $\lambda\|\cdot\|_1$ is convex, according to the first order optimality condition of $\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}$, we have

$$\lambda\left\|\hat{\mathbf{\Theta}}\right\|_1-\lambda\left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_1\geq\left\langle m\mathbf{Y}_i^{(k)}+\tau\left(\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{(k)}\right),\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\mathbf{\Theta}}\right\rangle+\tau\left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\mathbf{\Theta}}\right\|_F^2. \tag{37}$$

Summing (36) and (37), we have

$$\begin{aligned}\tau\left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\mathbf{\Theta}}\right\|_F^2\leq&U\left(\hat{\mathbf{\Theta}}\right)-U\left(\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)+\frac{L}{2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2-\frac{\mu}{2}\left\|\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{(k)}\right\|_F^2\\&+\tau\left\langle\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{(k)},\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\rangle+\left\langle\mathbf{\Delta}_i^{(k)},\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\mathbf{\Theta}}\right\rangle.\end{aligned}$$

Using the fact that $U\left(\hat{\mathbf{\Theta}}\right)-U\left(\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)\leq 0$, we have

$$\begin{aligned}\left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\mathbf{\Theta}}\right\|_F^2\leq&\frac{L}{2\tau}\left\|\mathbf{D}_i^{(k)}\right\|_F^2-\frac{\mu}{2\tau}\left\|\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{(k)}\right\|_F^2\\&+\left\langle\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{(k)},\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\rangle+\frac{1}{\tau}\left\langle\mathbf{\Delta}_i^{(k)},\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\mathbf{\Theta}}\right\rangle.\end{aligned}$$

When $\tau \geq L$, we have

$$\begin{aligned}\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 \leq & \frac{1}{2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2-\frac{\mu}{2\tau}\left\|\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{(k)}\right\|_F^2+\left\langle\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{(k)}, \hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\rangle+\frac{1}{\tau}\left\|\boldsymbol{\Delta}_i^{(k)}\right\|_F\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F \\ = & \left(\frac{1}{2}-\frac{\mu}{2\tau}\right)\left\|\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{(k)}\right\|_F^2+\frac{1}{2}\left\|\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_F^2+\frac{1}{\tau}\left\|\boldsymbol{\Delta}_i^{(k)}\right\|_F\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F .\end{aligned}$$

Rearranging the inequality above and using Young's inequality, we have

$$\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 \leq\left(1-\frac{\mu}{\tau}\right)\left\|\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{(k)}\right\|_F^2+\frac{1}{\tau \epsilon_\Theta}\left\|\boldsymbol{\Delta}_i^{(k)}\right\|_F^2+\frac{\epsilon_\Theta}{\tau}\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F^2,$$

which can be written as

$$\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 \leq\left(1-\frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\right)\left\|\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{(k)}\right\|_F^2+\frac{1}{\epsilon_\Theta\left(\tau-\epsilon_\Theta\right)}\left\|\boldsymbol{\Delta}_i^{(k)}\right\|_F^2, \tag{38}$$

where $\mu>\epsilon_\Theta$. Summing (38) for all agents leads to

$$\begin{aligned}\left\|\boldsymbol{\Theta}^{(k+1)}-\mathbf{1} \otimes \hat{\boldsymbol{\Theta}}\right\|_F^2 & =\sum_{i=1}^m\left\|\sum_{j=1}^m w_{ij} \tilde{\boldsymbol{\Theta}}_j^{(k)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 \leq \sum_{j=1}^m\left\|\tilde{\boldsymbol{\Theta}}_j^{(k)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 \\ & \leq \alpha \sum_{j=1}^m\left\|\boldsymbol{\Theta}_j^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F^2+(1-\alpha) \sum_{j=1}^m\left\|\boldsymbol{\Theta}_j^{(k)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 \\ & \leq\left(1-\frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta} \alpha\right)\left\|\boldsymbol{\Theta}^{(k)}-\mathbf{1} \otimes \hat{\boldsymbol{\Theta}}\right\|_F^2+\frac{\alpha}{\epsilon_\Theta\left(\tau-\epsilon_\Theta\right)}\left\|\boldsymbol{\Delta}^{(k)}\right\|_F^2 .\end{aligned}$$

Recalling (26), we obtain the desired result (35).

Note that since $\nabla^2 F(\boldsymbol{\Theta})=\boldsymbol{\Theta}^{-1} \otimes \boldsymbol{\Theta}^{-1} \preceq \frac{4}{\underline{R}_\Theta^2} \mathbf{I}$ for $\boldsymbol{\Theta} \in \mathcal{B}$ , we can choose $L=\frac{4}{\underline{R}_\Theta^2}$, and hence $\tau \geq \frac{4}{\underline{R}_\Theta^2}=L$. □

## D PROOF OF THEOREM 5

We first bound $\left\|\mathbf{D}^{(k)}\right\|_F$, which is detailed in the following proposition.

**Proposition 1.** *The following upper bound holds for* $\left\|\mathbf{D}^{(k)}\right\|_F$*:*

$$\left\|\mathbf{D}^{(k)}\right\|_F^2 \leq\left(\frac{3 m^2 L_{\max }^2}{\tau^2}+12\right)\left\|\mathbf{1} \otimes \hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^{(k)}\right\|_F^2+\frac{3 m^2}{\tau^2}\left\|\mathbf{E}_Y^{(k)}\right\|_F^2 . \tag{39}$$

*Proof.* According to the first order optimality condition of $\hat{\boldsymbol{\Theta}}$ along with the convexity of $\lambda\|\cdot\|_1$, we have

$$0 \leq\left\langle\nabla F\left(\hat{\boldsymbol{\Theta}}\right), \boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\rangle+\lambda\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_1-\lambda\left\|\hat{\boldsymbol{\Theta}}\right\|_1 . \tag{40}$$

Summing (37) and (40) yields

$$\begin{aligned}\tau\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 & \leq\left\langle\nabla F\left(\hat{\boldsymbol{\Theta}}\right)-m \mathbf{Y}_i^{(k)}-\tau\left(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{(k)}\right), \boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\rangle \\ & \leq\left\|\nabla F\left(\hat{\boldsymbol{\Theta}}\right)-m \mathbf{Y}_i^{(k)}-\tau\left(\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{(k)}\right)\right\|_F\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F .\end{aligned}$$

Dividing $\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F$ on both sides, we have

$$\tau\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F \leq\left\|\nabla F\left(\hat{\boldsymbol{\Theta}}\right)-m \bar{\mathbf{Y}}^{(k)}\right\|_F+m\left\|\bar{\mathbf{Y}}^{(k)}-\mathbf{Y}_i^{(k)}\right\|_F+\tau\left\|\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{(k)}\right\|_F .$$

Since

$$\tau \left\|\mathbf{D}_i^{(k)}\right\|_F - \tau \left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}\right\|_F \leq \tau \left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \hat{\mathbf{\Theta}}\right\|_F,$$

we obtain

$$\left\|\mathbf{D}_i^{(k)}\right\|_F \leq \frac{1}{\tau}\left\|\nabla F\left(\hat{\mathbf{\Theta}}\right) - m\bar{\mathbf{Y}}^{(k)}\right\|_F + \frac{m}{\tau}\left\|\bar{\mathbf{Y}}^{(k)} - \mathbf{Y}_i^{(k)}\right\|_F + 2\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}\right\|_F.$$

Taking the square of both sides and recalling (7) and the Lipschitz Smoothness of $f_i$, we have

$$\begin{aligned}
\left\|\mathbf{D}_i^{(k)}\right\|_F^2 &\leq \frac{3}{\tau^2}\left\|\nabla F\left(\hat{\mathbf{\Theta}}\right) - m\bar{\mathbf{Y}}^{(k)}\right\|_F^2 + \frac{3m^2}{\tau^2}\left\|\bar{\mathbf{Y}}^{(k)} - \mathbf{Y}_i^{(k)}\right\|_F^2 + 12\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}\right\|_F^2 \\
&\leq \frac{3m}{\tau^2}\sum_{j=1}^{m}\left\|\nabla f_j\left(\hat{\mathbf{\Theta}}\right) - \nabla f_j\left(\mathbf{\Theta}_j^{(k)}\right)\right\|_F^2 + \frac{3m^2}{\tau^2}\left\|\bar{\mathbf{Y}}^{(k)} - \mathbf{Y}_i^{(k)}\right\|_F^2 + 12\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}\right\|_F^2 \\
&\leq \frac{3m}{\tau^2}\sum_{j=1}^{m} L_j^2\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_j^{(k)}\right\|_F^2 + \frac{3m^2}{\tau^2}\left\|\bar{\mathbf{Y}}^{(k)} - \mathbf{Y}_i^{(k)}\right\|_F^2 + 12\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}\right\|_F^2.
\end{aligned}$$

Summing over $i = 1, \ldots, m$, we have the desired result (39). □

We are now ready to prove the linear rate of NetGGM. We build on the following intermediate result, introduced in (Nedic et al., 2017).

**Lemma 7.** *(Nedic et al., 2017) Given the sequence $\left\{s^{(k)}\right\}$, define the transformations*

$$S^{(K)}(z) = \max_{k=0,\ldots,K}\left|s^{(k)}\right| z^{-k} \text{ and } S(z) = \sup_{k\in\mathbb{N}}\left|s^{(k)}\right| z^{-k} \tag{41}$$

*for $z \in (0,1)$. If $S(z) \leq A < +\infty$, then $\left|s^{(k)}\right| \leq Az^k$.*

We show next how to chain the inequalities (35), (31), (32), and (39) so that Lemma 7 can be applied to the sequences $\left\{\left\|\mathbf{\Theta}^{(k)} - \mathbf{1}\otimes\hat{\mathbf{\Theta}}\right\|_F^2\right\}$, $\left\{\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2\right\}$, $\left\{\left\|\mathbf{E}_Y^{(k)}\right\|_F^2\right\}$, and $\left\{\left\|\mathbf{D}^{(k)}\right\|_F^2\right\}$, establishing thus their linear convergence.

**Proposition 2.** *Let $O^{(K)}(z)$, $E_\Theta^{(K)}(z)$, $E_Y^{(K)}(z)$, and $D^{(K)}(z)$ denote the transformation (41) applied to the sequences $\left\{\left\|\mathbf{\Theta}^{(k)} - \mathbf{1}\otimes\hat{\mathbf{\Theta}}\right\|_F^2\right\}$, $\left\{\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2\right\}$, $\left\{\left\|\mathbf{E}_Y^{(k)}\right\|_F^2\right\}$, and $\left\{\left\|\mathbf{D}^{(k)}\right\|_F^2\right\}$, respectively. Given the free parameter $\epsilon_d > 0$, the following holds:*

$$\begin{aligned}
O^{(K)}(z) &\leq a_O(\alpha, z)\left(4L_{\max}^2 E_\Theta^{(K)}(z) + 2E_Y^{(K)}(z)\right) + b_O(\alpha, z), \\
E_\Theta^{(K)}(z) &\leq a_E(z)\,\alpha^2 D^{(K)}(z) + b_\Theta(z), \\
E_Y^{(K)}(z) &\leq a_E(z)\,L_{\max}^2\left(8E_\Theta^{(K)}(z) + 2\alpha^2 D^{(K)}(z)\right) + b_Y(z), \\
D^{(K)}(z) &\leq a_{DO}O^{(K)}(z) + a_{DY}E_Y^{(K)}(z),
\end{aligned} \tag{42}$$

for all $z \in \left(\max\left\{1 - \frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\alpha, \rho^2(1+\epsilon_d)\right\}, 1\right)$, where

$$\begin{aligned}
a_O(\alpha, z) &= \frac{\frac{\alpha m^2}{\epsilon_\Theta(\tau-\epsilon_\Theta)}}{z - \left(1 - \frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\alpha\right)}, \quad b_O(\alpha, z) = \frac{z\left\|\mathbf{\Theta}^{(0)} - \mathbf{1}\otimes\hat{\mathbf{\Theta}}\right\|_F^2}{z - \left(1 - \frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\alpha\right)}, \\
a_E(z) &= \frac{\rho^2\left(1+\epsilon_d^{-1}\right)}{z - \rho^2(1+\epsilon_d)}, \qquad b_\Theta(z) = \frac{z}{z - \rho^2(1+\epsilon_d)}\left\|\mathbf{E}_\Theta^{(0)}\right\|_F^2, \quad b_Y(z) = \frac{z}{z - \rho^2(1+\epsilon_d)}\left\|\mathbf{E}_Y^{(0)}\right\|_F^2, \\
a_{DO} &= \frac{3m^2L_{\max}^2}{\tau^2} + 12, \qquad a_{DY} = \frac{3m^2}{\tau^2}.
\end{aligned}$$

*Proof.* Multiplying $z^{-k}$ on both sides of (35) and taking the maximum over $k = 0, \ldots, K$, we have

$$\begin{aligned}
&\max_{k=0,\ldots,K}\left\{\left\|\boldsymbol{\Theta}^{(k+1)}-\mathbf{1}\otimes\hat{\boldsymbol{\Theta}}\right\|_F^2 z^{-k}\right\}\\
\leq&\max_{k=0,\ldots,K}\left\{\left(1-\frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\alpha\right)\left\|\boldsymbol{\Theta}^{(k)}-\mathbf{1}\otimes\hat{\boldsymbol{\Theta}}\right\|_F^2 z^{-k}+\frac{\alpha m^2}{\epsilon_\Theta\left(\tau-\epsilon_\Theta\right)}\left(4L_{\max}^2\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2 z^{-k}+2\left\|\mathbf{E}_Y^{(k)}\right\|_F^2 z^{-k}\right)\right\}\\
\leq&\left(1-\frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\alpha\right)O^{(K)}\left(z\right)+\frac{\alpha m^2}{\epsilon_\Theta\left(\tau-\epsilon_\Theta\right)}\left(4L_{\max}^2E_\Theta^{(K)}\left(z\right)+2E_Y^{(K)}\left(z\right)\right).
\end{aligned}$$

Moreover, according to the fact that

$$\max_{k=0,\ldots,K}\left\{\left\|\boldsymbol{\Theta}^{(k+1)}-\mathbf{1}\otimes\hat{\boldsymbol{\Theta}}\right\|_F^2 z^{-k}\right\}\geq\max_{k=1,\ldots,K}\left\{\left\|\boldsymbol{\Theta}^{(k)}-\mathbf{1}\otimes\hat{\boldsymbol{\Theta}}\right\|_F^2 z^{-(k-1)}\right\}\geq zO^{(K)}\left(z\right)-z\left\|\boldsymbol{\Theta}^{(0)}-\mathbf{1}\otimes\hat{\boldsymbol{\Theta}}\right\|_F^2,$$

we have the desired result (42).

Then, applying a similar procedure as the one used to obtain (42) to (31), (32), and (39), we have the rest of the results. □

Chaining the inequalities in 2, we can finally prove Theorem 5.

**Theorem 9.** *In the same setting as Theorem 8, if the proximal parameter satisfies* $\tau \geq \max\left\{\underline{\tau}_1, \frac{4}{\underline{R}_\Theta^2}\right\}$ *and the step size* $\alpha$ *satisfies* $\alpha \in (0, \bar{\alpha}_2)$*, where* $\bar{\alpha}_2 = \min\left\{\bar{\alpha}_1, \frac{(1-\rho)^2}{A_{\frac{1}{2}}}\right\}$*, for* $\alpha \in (0, \bar{\alpha}_2)$*, we have*

$$\sum_{i=1}^m\left\|\boldsymbol{\Theta}_i^{(k)}-\hat{\boldsymbol{\Theta}}\right\|_F^2\leq A\underline{z}^k$$

*for all* $k \in \mathbb{N}$*, where* $A_{\frac{1}{2}}, A > 0$*, and* $\underline{z} \in (0,1)$ *are defined in* (52)*,* (48)*, and* (55)*.*

*Proof.* Chaining the inequalities in Proposition 2, we have

$$D^{(K)}\left(z\right)\leq a_D\left(\alpha,z\right)D^{(K)}\left(z\right)+b_D\left(\alpha,z\right),$$

where

$$\begin{aligned}
a_D\left(\alpha,z\right)=&\,4a_{DO}a_O\left(\alpha,z\right)L_{\max}^2a_E\left(z\right)\alpha^2\\
&+8\left(2a_{DO}a_O\left(\alpha,z\right)+a_{DY}\right)a_E\left(z\right)L_{\max}^2a_E\left(z\right)\alpha^2\\
&+2\left(2a_{DO}a_O\left(\alpha,z\right)+a_{DY}\right)a_E\left(z\right)L_{\max}^2\alpha^2, \qquad (43)\\
b_D\left(\alpha,z\right)=&\,a_{DO}b_O\left(\alpha,z\right)\\
&+4a_{DO}a_O\left(\alpha,z\right)L_{\max}^2b_\Theta\left(z\right)\\
&+8\left(2a_{DO}a_O\left(\alpha,z\right)+a_{DY}\right)a_E\left(z\right)L_{\max}^2b_\Theta\left(z\right)\\
&+\left(2a_{DO}a_O\left(\alpha,z\right)+a_{DY}\right)b_Y\left(z\right). \qquad (44)
\end{aligned}$$

Therefore, for $a\left(\alpha,z\right)<1$, we have

$$D^{(K)}\left(z\right)\leq\frac{b_D\left(\alpha,z\right)}{1-a_D\left(\alpha,z\right)}<+\infty, \qquad (45)$$

Plugging (45) into (42), we have

$$O^{(K)}\left(z\right)\leq a\left(\alpha,z\right)\frac{b_D\left(\alpha,z\right)}{1-a_D\left(\alpha,z\right)}+b\left(\alpha,z\right)\leq+\infty,$$

where

$$\begin{aligned}
a\left(\alpha,z\right)=&\left(a_O\left(\alpha,z\right)4L_{\max}^2+a_O\left(\alpha,z\right)2a_E\left(z\right)L_{\max}^28\right)a_E\left(z\right)\alpha^2+a_O\left(\alpha,z\right)2a_E\left(z\right)L_{\max}^22\alpha^2, \qquad (46)\\
b\left(\alpha,z\right)=&\left(a_O\left(\alpha,z\right)4L_{\max}^2+a_O\left(\alpha,z\right)2a_E\left(z\right)L_{\max}^28\right)b_\Theta\left(z\right)+a_O\left(\alpha,z\right)2b_Y\left(z\right)+b_O\left(\alpha,z\right). \qquad (47)
\end{aligned}$$

According to Lemma 7, we have

$$\left\|\boldsymbol{\Theta}^{(k)}-\mathbf{1}\otimes\hat{\boldsymbol{\Theta}}\right\|_F^2\leq Az^k,$$

where

$$A=a\left(\alpha,z\right)\frac{b_D\left(\alpha,z\right)}{1-a_D\left(\alpha,z\right)}+b\left(\alpha,z\right), \tag{48}$$

This means that $\left\{\left\|\boldsymbol{\Theta}^{(k)}-\mathbf{1}\otimes\hat{\boldsymbol{\Theta}}\right\|^2\right\}$ vanishes R-linearly at a rate of at least $z$.

Finally, we prove that there exist $\alpha\in(0,1]$ and $z\in\left(\max\left\{1-\frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\alpha,\rho^2\left(1+\epsilon_d\right)\right\},1\right)$ such that $a_D\left(\alpha,z\right)<1$. We first set $z=1$ and find $\bar{\alpha}\in(0,1)$ such that $a_D\left(\alpha,1\right)<1$ for all $\alpha\in(0,\bar{\alpha})$. For parameter $\epsilon_d$, $a_D\left(\alpha,1\right)$ is minimized when

$$\epsilon_d=\arg\min_{\epsilon>0}\frac{1+\epsilon^{-1}}{1-\rho^2\left(1+\epsilon\right)}=\frac{1-\rho}{\rho}.$$

For $\epsilon_\Theta$, $a_D\left(\alpha,1\right)$ is minimized by solving the following problem:

$$\begin{aligned}&\underset{\epsilon_\Theta}{\text{minimize}} \quad a_O\left(\alpha,1\right)=\frac{m^2}{\epsilon_\Theta\left(\mu-\epsilon_\Theta\right)}\\&\text{subject to} \quad \epsilon_\Theta\in(0,\mu)\end{aligned} \tag{49}$$

The solution of (49) is obtained when $\epsilon_\Theta=\frac{\mu}{2}$, and hence

$$a_O^\star\left(\alpha,1\right)=\frac{4m^2}{\mu^2}.$$

Then since $a_D\left(\alpha,1\right)$ is continuous on $[0,+\infty)$, $a_D\left(\alpha,1\right)\geq 0$ for all $\alpha\in(0,1]$, and $a_D\left(0,1\right)=0$ when $\alpha=0$, there exists some $\bar{\alpha}\in(0,1]$ such that $a_D\left(\alpha,1\right)<1$ for all $\alpha\in(0,\bar{\alpha})$. Then since $a_D\left(\bar{\alpha},z\right)$ is continuous at $z=1$, there exists $z\left(\bar{\alpha}\right)\in(0,1)$ such that $a\left(\alpha,z\left(\bar{\alpha}\right)\right)<1$.

Then, we find the lower bound for $z$ and the upper bound for $\alpha$. Recall that $z\in\left(\max\left\{1-\frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\alpha,\rho^2\left(1+\epsilon_d\right)\right\},1\right)$. For $z>1-\frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\alpha$, we impose the stronger version

$$z\geq 1-\frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\alpha+\frac{\theta\alpha\left(\mu-\epsilon_\Theta\right)}{\tau-\epsilon_\Theta}, \tag{50}$$

where $\theta\in(0,1)$. Hence, $a_O\left(\alpha,z\right)$ can be bounded by

$$a_O\left(\alpha,z\right)=\frac{\frac{\alpha m^2}{\epsilon_\Theta\left(\tau-\epsilon_\Theta\right)}}{z-1+\frac{\mu-\epsilon_\Theta}{\tau-\epsilon_\Theta}\alpha}\leq\inf_{\epsilon_\Theta\in(0,\mu)}\frac{m^2}{\theta\epsilon_\Theta\left(\mu-\epsilon_\Theta\right)}=\frac{4m^2}{\mu^2}\theta^{-1}.$$

Substituting the upper bound into $a_D\left(\alpha,z\right)$, we have

$$\begin{aligned}a_D\left(\alpha,z\right)\leq\ &4a_{DO}\frac{4m^2}{\mu^2}\theta^{-1}L_{\max}^2a_E\left(z\right)\alpha^2\\&+8\left(2a_{DO}\frac{4m^2}{\mu^2}\theta^{-1}+a_{DY}\right)a_E\left(z\right)L_{\max}^2a_E\left(z\right)\alpha^2\\&+2\left(2a_{DO}\frac{4m^2}{\mu^2}\theta^{-1}+a_{DY}\right)a_E\left(z\right)L_{\max}^2\alpha^2.\end{aligned}$$

The right-hand side is minimized at $\epsilon_d=\frac{\sqrt{z}-\rho}{\rho}$. Then, the sufficient condition of $a_D\left(\alpha,z\right)<1$ is

$$\alpha\leq\left(\left(A_{\theta,1}+A_{\theta,3}\right)\frac{1}{\left(\sqrt{z}-\rho\right)^2}+A_{\theta,2}\frac{1}{\left(\sqrt{z}-\rho\right)^4}\right)^{-\frac{1}{2}},$$

where

$$A_{\theta,1} = \frac{4m^2}{\mu^2}\theta^{-1}a_{DO}4\rho^2 L_{\max}^2,$$
$$A_{\theta,2} = \left(\frac{8m^2}{\mu^2}\theta^{-1}a_{DO} + a_{DY}\right)8\rho^4 L_{\max}^2,$$
$$A_{\theta,3} = \left(\frac{8m^2}{\mu^2}\theta^{-1}a_{DO} + a_{DY}\right)2\rho^2 L_{\max}^2.$$

This implies that

$$z \geq \left(\rho + \sqrt{\alpha A_\theta}\right)^2, \tag{51}$$

where

$$A_\theta = \sqrt{A_{\theta,1} + A_{\theta,2} + A_{\theta,3}}. \tag{52}$$

Defining $\epsilon_\Theta = \frac{\mu}{2}$, $\theta = \frac{1}{2}$, and

$$J = \frac{\mu}{2\left(2\tau - \mu\right)},$$

inequalities (50) and (51) lead to

$$\underline{z} = \max\left\{\left(\rho + \sqrt{\alpha A_{\frac{1}{2}}}\right)^2, 1 - J\alpha\right\}. \tag{53}$$

Recall that $\epsilon_d = \frac{\sqrt{z}-\rho}{\rho}$, it follows from (51) that

$$z \geq \left(\rho + \sqrt{\alpha A_\theta}\right)^2 > \rho^2 \Rightarrow z > \rho\sqrt{z} = \rho^2\left(1 + \epsilon_d\right).$$

For the upper bound of $\alpha$, $z < 1$ implies that

$$\alpha < \frac{\left(1-\rho\right)^2}{A_{\frac{1}{2}}}. \tag{54}$$

Combining (29) and (54), we have

$$\bar{\alpha}_2 = \min\left\{\bar{\alpha}_1, \frac{\left(1-\rho\right)^2}{A_{\frac{1}{2}}}\right\}.$$

Note that the function $\left(\rho + \sqrt{\alpha A_{\frac{1}{2}}}\right)^2$ increases monotonically as $\alpha$ increases, while $1 - J\alpha$ decreases monotonically as $\alpha$ increases. Then since $\rho < 1$, inequality (53) is equivalent to

$$\underline{z} = \begin{cases} 1 - J\alpha, & \alpha \in \left(0, \alpha^\star\right), \\ \left(\rho + \sqrt{\alpha A_{\frac{1}{2}}}\right)^2, & \alpha \in \left[\alpha^\star, \bar{\alpha}_2\right), \end{cases} \tag{55}$$

where

$$\alpha^\star = \left(\frac{-\rho\sqrt{A_{\frac{1}{2}}} + \sqrt{A_{\frac{1}{2}} + J\left(1-\rho^2\right)}}{A_{\frac{1}{2}} + J}\right)^2.$$

□

# E PROOF OF THEOREM 6

To prove Theorem 6, we first revisit the definition of the $\varepsilon$-subgradient and its properties.

**Definition 1.** For convex function $f$, $\boldsymbol{\xi}$ is an $\varepsilon$-subgradient of $f$ at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \boldsymbol{\xi}, \mathbf{y} - \mathbf{x} \rangle - \varepsilon, \quad \forall y. \tag{56}$$

Then, the following properties hold for subdifferential:

1. $\varepsilon$-subdifferential of a convex function is nonempty, convex and compact;
2. $\mathbf{x}$ is $\varepsilon$-optimal of $G$ if and only if $0 \in \partial_\varepsilon G(\mathbf{x})$;
3. If $G = G_1 + G_2$ and $G_1$ and $G_2$ are convex, then $\partial_\varepsilon G(\mathbf{x}) \subset \partial_\varepsilon G_1(\mathbf{x}) + \partial_\varepsilon G_2(\mathbf{x})$.

**Lemma 8.** *If $f(\mathbf{x}) := \mathbb{R}^n \to \mathbb{R}$ is $\mu$-strongly convex, we have $\nabla f(\mathbf{x}) + \boldsymbol{\delta} \in \partial_\varepsilon f(\mathbf{x})$ for all $\|\boldsymbol{\delta}\|_F^2 \leq 2\mu\varepsilon$, where $\varepsilon > 0$ and $\partial_\varepsilon$ is the set of $\varepsilon$-subgradient.*

*Proof.* Because $\|\boldsymbol{\delta}\|_F^2 \leq 2\mu\varepsilon$, we have

$$\begin{aligned}
\boldsymbol{\delta}^\top(\mathbf{y} - \mathbf{x}) - \varepsilon &= \left\| \boldsymbol{\delta}^\top(\mathbf{y} - \mathbf{x}) \right\|_F - \varepsilon \\
&\leq \|\boldsymbol{\delta}\|_F \|(\mathbf{y} - \mathbf{x})\|_F - \varepsilon \\
&\leq \sqrt{2\mu\varepsilon} \|(\mathbf{y} - \mathbf{x})\|_F - \varepsilon \\
&\leq \frac{\mu}{2} \|(\mathbf{y} - \mathbf{x})\|_F^2 .
\end{aligned}$$

Then, due to strong convexity, it follows that

$$\begin{aligned}
f(\mathbf{y}) \geq & f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_F^2 \\
\geq & f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \boldsymbol{\delta}^\top(\mathbf{y} - \mathbf{x}) - \varepsilon \\
= & f(\mathbf{x}) + (\nabla f(\mathbf{x}) + \boldsymbol{\delta})^\top(\mathbf{y} - \mathbf{x}) - \varepsilon,
\end{aligned}$$

which completes the proof. □

## E.1 Local strong convexity and Lipschitz smoothness

Assuming the inexact solutions of local optimizations are $\varepsilon^{(k)}$-optimal and $\varepsilon^{(k)}$ decay at least at an exponential rate, we assume that there exist constants $c > 0$ and $\bar{\varepsilon} \in (0, 1)$ such that $\varepsilon^{(k)} \leq c\bar{\varepsilon}^k$. We redefine set $\mathcal{A}$ as

$$\mathcal{A} = \left\{ (\boldsymbol{\Theta}, \mathbf{Y}) \mid V(\boldsymbol{\Theta}, \mathbf{Y}) \leq V\left(\boldsymbol{\Theta}^{(0)}, \mathbf{Y}^{(0)}\right) + \frac{2mc}{1 - \bar{\varepsilon}} \right\}.$$

Compared to the case where subproblems have closed-form solutions, inexact solutions result in a smaller step size required to ensure convergence, as stated in Lemma 9.

**Lemma 9.** *On the basis of Lemma 3, if $\left(\boldsymbol{\Theta}^{(k)}, \mathbf{E}_Y^{(k)}\right) \in \mathcal{A}, i = 1, 2, \ldots, m$ and*

$$\tau \geq \max \left\{ \frac{m\sqrt{\sum_{i=1}^m u_{\nabla f_i} + dR_Y^2} + \max_{\boldsymbol{\eta} \in \partial G\left(\boldsymbol{\Theta}_i^{(k)}\right)} \|\boldsymbol{\eta}\|_F}{\sqrt{\left(\frac{R_\Theta}{2} - \sqrt{2\varepsilon^{(k)}}\right)^2 - 4\varepsilon^{(k)}}}, 1 \right\}, \tag{57}$$

*where $\varepsilon^{(k)} < \frac{\left(\sqrt{2}-1\right)^2 R_\Theta^2}{8}$, then $\boldsymbol{\Theta}^{\left(k+\frac{1}{2}\right)}, \boldsymbol{\Theta}^{(k+1)} \in \mathcal{B}$, where*

$$\mathcal{B} = \left\{ \boldsymbol{\Theta} \mid \frac{R_\Theta}{2}\mathbf{I} \preceq \boldsymbol{\Theta}_i \preceq \left(\overline{R}_\Theta + \frac{R_\Theta}{2}\right)\mathbf{I}, i = 1, 2, \ldots, m \right\}. \tag{58}$$

*Proof.* Since $0 \in \partial_{\varepsilon^{(k)}} \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) + G\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)$, there exists some $\boldsymbol{\xi} \in \partial_{\varepsilon^{(k)}} \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)$ and $-\boldsymbol{\xi} \in \partial_{\varepsilon^{(k)}} G\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right)$. Because $\tilde{F}_i^{(k)}$ is $\tau$-strongly convex, according to Lemma 8, we can find some $\|\boldsymbol{\delta}\|_F^2 \leq 2\tau\varepsilon^{(k)}$ so that $\boldsymbol{\xi} = \nabla \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) + \boldsymbol{\delta}$. Recall the definition of $\varepsilon^{(k)}$-subgradient, we have

$$G\left(\boldsymbol{\Theta}_i^{(k)}\right) \geq G\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) + \left\langle \nabla \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) + \boldsymbol{\delta}, \mathbf{D}_i^{(k)} \right\rangle - \varepsilon^{(k)}. \tag{59}$$

In addition, because $\tilde{F}_i^{(k)}$ is $\tau$-strongly convex, we have

$$\tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{(k)}\right) \geq \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) - \left\langle \nabla \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right), \mathbf{D}_i^{(k)} \right\rangle + \frac{\tau}{2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2. \tag{60}$$

Plugging (59) into (60), we can get

$$\begin{aligned} \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{(k)}\right) + G\left(\boldsymbol{\Theta}_i^{(k)}\right) - \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) - G\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) \geq & \frac{\tau}{2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2 + \left\langle \boldsymbol{\delta}, \mathbf{D}_i^{(k)} \right\rangle - \varepsilon^{(k)} \\ \geq & \frac{\tau}{2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2 - \sqrt{2\tau\varepsilon^{(k)}}\left\|\mathbf{D}_i^{(k)}\right\|_F - \varepsilon^{(k)}. \end{aligned} \tag{61}$$

Combining (10) and (61) and replacing $\lambda\|\cdot\|_1$ with $G$ leads to

$$\frac{1}{2\tau}\left\|\nabla \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{(k)}\right) + \boldsymbol{\zeta}\right\|_F^2 \geq \frac{\tau}{2}\left(\left\|\mathbf{D}_i^{(k)}\right\|_F - \sqrt{\frac{2}{\tau}\varepsilon^{(k)}}\right)^2 - 2\varepsilon^{(k)},$$

and hence

$$\begin{aligned} \left\|\mathbf{D}_i^{(k)}\right\|_F \leq & \sqrt{\frac{1}{\tau^2}\left\|\nabla \tilde{F}_i^{(k)}\left(\boldsymbol{\Theta}_i^{(k)}\right) + \boldsymbol{\zeta}\right\|_F^2 + \frac{4}{\tau}\varepsilon^{(k)}} + \sqrt{\frac{2}{\tau}\varepsilon^{(k)}} \\ \leq & \sqrt{\frac{1}{\tau^2}\left(m\left\|\mathbf{Y}_i^{(k)}\right\|_F + \max_{\boldsymbol{\eta} \in \partial G\left(\boldsymbol{\Theta}_i^{(k)}\right)} \|\boldsymbol{\eta}\|_F\right)^2 + \frac{4}{\tau}\varepsilon^{(k)}} + \sqrt{\frac{2}{\tau}\varepsilon^{(k)}}. \end{aligned}$$

When $\tau \geq 1$, we have

$$\left\|\mathbf{D}_i^{(k)}\right\|_F \leq \sqrt{\frac{1}{\tau^2}\left(m\left\|\mathbf{Y}_i^{(k)}\right\|_F + \max_{\boldsymbol{\eta} \in \partial G\left(\boldsymbol{\Theta}_i^{(k)}\right)} \|\boldsymbol{\eta}\|_F\right)^2 + 4\varepsilon^{(k)}} + \sqrt{2\varepsilon^{(k)}}.$$

Then due to Proposition 3 and (57), we have

$$\left\|\mathbf{D}_i^{(k)}\right\|_F \leq \frac{R_\Theta}{2} \tag{62}$$

for any $k \in \mathbb{N}$, and hence $\boldsymbol{\Theta}_i^{(k+1)} \in \mathcal{B}$. □

We have the following upper bound of the optimality gap with respect to consensus errors.

**Lemma 10.** *Based on Lemma 9, there holds*

$$\begin{aligned} \sum_{i=1}^m U\left(\boldsymbol{\Theta}_i^{(k+1)}\right) \leq & \sum_{i=1}^m U\left(\boldsymbol{\Theta}_i^{(k)}\right) + \frac{\alpha}{2}\epsilon_p^{-1}m^2\left(4L_{\max}^2\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2 + 2\left\|\mathbf{E}_Y^{(k)}\right\|_F^2\right) - \alpha\left(\frac{\tau}{2} - \frac{2\alpha}{\underline{R}_\Theta^2} - \frac{1}{2}\epsilon_p\right)\left\|\mathbf{D}^{(k)}\right\|_F^2 \\ & + 2m\alpha\varepsilon^{(k)}, \end{aligned} \tag{63}$$

*where* $\epsilon_p > 0$ *and* $\alpha < \frac{\underline{R}_\Theta^2}{4}\tau$.

*Proof.* Because $\widetilde{F}_i$ is strongly convex with $\tau$ and $G$ is convex, according to the $\varepsilon$optimality condition, we have

$$G\left(\boldsymbol{\Theta}_i^{(k)}\right) - G\left(\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) \geq \left\langle m\mathbf{Y}_i^{(k)} + \boldsymbol{\delta}_i, \mathbf{D}_i^{(k)} \right\rangle + \tau \left\|\mathbf{D}_i^{(k)}\right\|_F^2 - \varepsilon^{(k)}. \tag{64}$$

Substituting (64) and (19) into (17), we have

$$\begin{aligned} F\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right) &\leq F\left(\boldsymbol{\Theta}_i^{(k)}\right) + \left\langle \boldsymbol{\Delta}_i^{(k)} - \boldsymbol{\delta}_i, \alpha\mathbf{D}_i^{(k)} \right\rangle + G\left(\boldsymbol{\Theta}_i^{(k)}\right) - G\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right) - \alpha\tau \left\|\mathbf{D}_i^{(k)}\right\|_F^2 + \frac{2\alpha^2}{\underline{R}_\Theta^2} \left\|\mathbf{D}_i^{(k)}\right\|_F^2 + \alpha\varepsilon^{(k)} \\ &\leq F\left(\boldsymbol{\Theta}_i^{(k)}\right) + G\left(\boldsymbol{\Theta}_i^{(k)}\right) - G\left(\tilde{\boldsymbol{\Theta}}_i^{(k)}\right) - \alpha\left(\tau - \frac{2\alpha}{\underline{R}_\Theta^2}\right) \left\|\mathbf{D}_i^{(k)}\right\|_F^2 \\ &\quad + \alpha \left\|\mathbf{D}_i^{(k)}\right\|_F \left\|\boldsymbol{\Delta}_i^{(k)}\right\|_F + \alpha\sqrt{2\tau\varepsilon^{(k)}} \left\|\mathbf{D}_i^{(k)}\right\|_F + \alpha\varepsilon^{(k)}. \end{aligned} \tag{65}$$

We can now substitute (65) into (23) and get

$$\begin{aligned} \sum_{i=1}^m U\left(\boldsymbol{\Theta}_i^{(k+1)}\right) \leq \sum_{i=1}^m U\left(\boldsymbol{\Theta}_i^{(k)}\right) \\ + \sum_{i=1}^m \left( \alpha \left\|\mathbf{D}_i^{(k)}\right\|_F \left\|\boldsymbol{\Delta}_i^{(k)}\right\|_F - \alpha\left(\tau - \frac{2\alpha}{\underline{R}_\Theta^2}\right) \left\|\mathbf{D}_i^{(k)}\right\|_F^2 + \alpha\sqrt{2\tau\varepsilon^{(k)}} \left\|\mathbf{D}_i^{(k)}\right\|_F + \alpha\varepsilon^{(k)} \right). \end{aligned}$$

Using Young's inequality, we have

$$\alpha\sqrt{2\tau\varepsilon^{(k)}} \left\|\mathbf{D}_i^{(k)}\right\|_F \leq \alpha\varepsilon^{(k)} + \frac{\alpha\tau}{2} \left\|\mathbf{D}_i^{(k)}\right\|_F^2.$$

Recall (24) and (26), we obtain the desired result (63). □

Following similar proof as in Theorem (7) and choosing $\epsilon_p = \frac{\tau}{2} - \frac{2\alpha}{\underline{R}_\Theta^2}$, we have the following theorem.

**Theorem 10.** *When $\tau \geq \underline{\tau}_2$ and $\alpha \in (0, \bar{\alpha}_1')$, where*

$$\underline{\tau}_2 = \max\left\{ \frac{m\sqrt{\sum_{i=1}^m u_{\nabla f_i} + dR_Y^2} + \max_{\boldsymbol{\eta}\in\partial G\left(\boldsymbol{\Theta}_i^{(k)}\right)} \|\boldsymbol{\eta}\|_F}{\sqrt{\left(\frac{R_\Theta}{2} - \sqrt{2\varepsilon^{(k)}}\right)^2 - 4\varepsilon^{(k)}}}, 1 \right\},$$

$\varepsilon^{(k)} < \frac{\left(\sqrt{2}-1\right)^2 \underline{R}_\Theta^2}{8}$, *and*

$$\bar{\alpha}_1' = \min\left\{ \underline{R}_\Theta^2\tau, \frac{\tau\left(\sqrt{\frac{1}{\underline{R}_\Theta^4} + 2\left(\frac{4m^2L_{\max}^2\rho^2\left(1+\epsilon_d^{-1}\right)}{1-\rho^2(1+\epsilon_d)} + \frac{8m^2L_{\max}^2\rho^4\left(1+\epsilon_d^{-1}\right)^2}{(1-\rho^2(1+\epsilon_d))^2}\right)} - \frac{1}{\underline{R}_\Theta^2}\right)}{2\left(\frac{4m^2L_{\max}^2\rho^2\left(1+\epsilon_d^{-1}\right)}{1-\rho^2(1+\epsilon_d)} + \frac{8m^2L_{\max}^2\rho^4\left(1+\epsilon_d^{-1}\right)^2}{(1-\rho^2(1+\epsilon_d))^2}\right)}, 1 \right\}, \tag{66}$$

*then for $\left\{\boldsymbol{\Theta}^{(k)}, \mathbf{Y}^{(k)}\right\}_{k\in\mathbb{N}}$ obtained by NetGGM, we have*

$$V\left(\boldsymbol{\Theta}^{(k+1)}, \mathbf{Y}^{(k+1)}\right) \leq V\left(\boldsymbol{\Theta}^{(k)}, \mathbf{Y}^{(k)}\right) - \beta\sum_{i=1}^m \left\|\mathbf{D}_i^{(k)}\right\|_F^2 + 2m\alpha\varepsilon^{(k)}, \tag{67}$$

*where*

$$\begin{aligned} V\left(\boldsymbol{\Theta}^{(k)}, \mathbf{E}_Y^{(k)}\right) = &\sum_{i=1}^m U\left(\boldsymbol{\Theta}_i^{(k)}\right) + \frac{\alpha\left(\frac{\tau}{2} - \frac{2\alpha}{\underline{R}_\Theta^2}\right)^{-1} m^2}{1-\rho^2\left(1+\epsilon_d\right)} \left\|\mathbf{E}_Y^{(k)}\right\|_F^2 \\ &+ \frac{2\alpha\left(\frac{\tau}{2} - \frac{2\alpha}{\underline{R}_\Theta^2}\right)^{-1} m^2L_{\max}^2\left(4\rho^2\left(1+\epsilon_d^{-1}\right) + 1 - \rho^2\left(1+\epsilon_d\right)\right)}{\left(1-\rho^2\left(1+\epsilon_d\right)\right)^2} \left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2, \end{aligned}$$

*and*

$$\beta = \alpha\left(\frac{\tau}{4} - \frac{\alpha}{\underline{R}_\Theta^2} - \frac{4\alpha^2\left(\frac{\tau}{2} - \frac{2\alpha}{\underline{R}_\Theta^2}\right)^{-1} m^2 L_{\max}^2 \rho^2 \left(1+\epsilon_d^{-1}\right)}{1-\rho^2\left(1+\epsilon_d\right)} - \frac{8\alpha^2\left(\frac{\tau}{2} - \frac{2\alpha}{\underline{R}_\Theta^2}\right)^{-1} m^2 L_{\max}^2 \rho^4 \left(1+\epsilon_d^{-1}\right)^2}{\left(1-\rho^2\left(1+\epsilon_d\right)\right)^2}\right) \geq 0.$$

Utilizing the fact that $\varepsilon^{(k)} \leq c\bar{\varepsilon}^k$, $\alpha \leq 1$, and Theorem 10, and iterating to $k=0$, we have

$$\begin{aligned} V\left(\mathbf{\Theta}^{(k+1)}, \mathbf{Y}^{(k+1)}\right) &\leq \ldots \leq V\left(\mathbf{\Theta}^{(0)}, \mathbf{Y}^{(0)}\right) + 2m\alpha\sum_{l=1}^{k}\varepsilon^{(l)} \leq V\left(\mathbf{\Theta}^{(0)}, \mathbf{Y}^{(0)}\right) + 2mc\sum_{l=1}^{k}\bar{\varepsilon}^k \\ &\leq V\left(\mathbf{\Theta}^{(0)}, \mathbf{Y}^{(0)}\right) + \frac{2mc}{1-\bar{\varepsilon}}. \end{aligned}$$

Therefore, $\left(\mathbf{\Theta}^{(k)}, \mathbf{Y}^{(k)}\right) \in \mathcal{A}$ and all the $\mathbf{\Theta}_i^{(k)}$ are positive definite for every iteration $k \in \mathbb{N}$.

### E.2 Linear convergence

**Theorem 11.** *Based on Theorem 10, when $\tau \geq \frac{4}{\underline{R}_\Theta^2}$, we have*

$$\begin{aligned} \left\|\mathbf{\Theta}^{(k+1)} - \mathbf{1}\otimes\hat{\mathbf{\Theta}}\right\|_F^2 \leq& \left(1 - \frac{\mu - 2\epsilon_\Theta}{\tau - 2\epsilon_\Theta}\alpha\right)\left\|\mathbf{\Theta}^{(k)} - \mathbf{1}\otimes\hat{\mathbf{\Theta}}\right\|_F^2 + \frac{\alpha m^2}{\epsilon_\Theta\left(\tau - 2\epsilon_\Theta\right)}\left(4L_{\max}^2\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2 + 2\left\|\mathbf{E}_Y^{(k)}\right\|_F^2\right) \\ &+ \frac{2\left(\epsilon_\Theta^{-1}\tau + 1\right)}{\tau - 2\epsilon_\Theta}\varepsilon^{(k)}. \end{aligned} \tag{68}$$

*where $\epsilon_\Theta \in \left(0, \frac{\mu}{2}\right)$.*

*Proof.* Since $\widetilde{F}_i$ is strongly convex with $\tau$ and $G$ is convex, according to the first order optimality condition of $\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}$, we have

$$G\left(\hat{\mathbf{\Theta}}\right) - G\left(\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) \geq \left\langle m\mathbf{Y}_i^{(k)} + \tau\left(\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}\right) + \boldsymbol{\delta}_i, \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \hat{\mathbf{\Theta}}\right\rangle + \tau\left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \hat{\mathbf{\Theta}}\right\|_F^2 - \varepsilon^{(k)}. \tag{69}$$

Summing (36) and (69), we have

$$\begin{aligned} \tau\left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \hat{\mathbf{\Theta}}\right\|_F^2 \leq& U\left(\hat{\mathbf{\Theta}}\right) - U\left(\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right) + \frac{L}{2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2 - \frac{\mu}{2}\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}\right\|_F^2 \\ &+ \tau\left\langle\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}, \hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\rangle + \left\langle\mathbf{\Delta}_i^{(k)}, \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \hat{\mathbf{\Theta}}\right\rangle + \left\langle\boldsymbol{\delta}_i, \hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\rangle + \varepsilon^{(k)} \\ \leq& \frac{L}{2\tau}\left\|\mathbf{D}_i^{(k)}\right\|_F^2 - \frac{\mu}{2\tau}\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}\right\|_F^2 \\ &+ \left\langle\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}, \hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\rangle + \frac{1}{\tau}\left\langle\mathbf{\Delta}_i^{(k)}, \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \hat{\mathbf{\Theta}}\right\rangle + \frac{1}{\tau}\left\langle\boldsymbol{\delta}_i, \hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\rangle + \frac{1}{\tau}\varepsilon^{(k)}. \end{aligned}$$

When $\tau \geq L$, we have

$$\begin{aligned} \left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \hat{\mathbf{\Theta}}\right\|_F^2 \leq& \frac{1}{2}\left\|\mathbf{D}_i^{(k)}\right\|_F^2 - \frac{\mu}{2\tau}\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}\right\|_F^2 + \left\langle\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}, \hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\rangle \\ &+ \frac{1}{\tau}\left\|\mathbf{\Delta}_i^{(k)}\right\|_F\left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \hat{\mathbf{\Theta}}\right\|_F + \frac{\sqrt{2\tau\varepsilon^{(k)}}}{\tau}\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_F + \frac{1}{\tau}\varepsilon^{(k)} \\ =& \left(\frac{1}{2} - \frac{\mu}{2\tau}\right)\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{(k)}\right\|_F^2 + \frac{1}{2}\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_F^2 \\ &+ \frac{1}{\tau}\left\|\mathbf{\Delta}_i^{(k)}\right\|_F\left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)} - \hat{\mathbf{\Theta}}\right\|_F + \frac{\sqrt{2\tau\varepsilon^{(k)}}}{\tau}\left\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}\right\|_F + \frac{1}{\tau}\varepsilon^{(k)}. \end{aligned}$$

Rearranging the inequality above and using Young's inequality, we have

$$\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 \leq \left(1-\frac{\mu}{\tau}\right)\left\|\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{(k)}\right\|_F^2+\frac{1}{\tau\epsilon_\Theta}\left\|\boldsymbol{\Delta}_i^{(k)}\right\|_F^2+\frac{2\epsilon_\Theta}{\tau}\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F^2+\frac{2}{\tau}\left(\epsilon_\Theta^{-1}\tau+1\right)\varepsilon^{(k)},$$

which can be written as

$$\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 \leq \left(1-\frac{\mu-2\epsilon_\Theta}{\tau-2\epsilon_\Theta}\right)\left\|\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}_i^{(k)}\right\|_F^2+\frac{1}{\epsilon_\Theta\left(\tau-2\epsilon_\Theta\right)}\left\|\boldsymbol{\Delta}_i^{(k)}\right\|_F^2+\frac{2\left(\epsilon_\Theta^{-1}\tau+1\right)}{\tau-2\epsilon_\Theta}\varepsilon^{(k)}, \tag{70}$$

where $\mu>2\epsilon_\Theta$. Summing (70) for all agents leads to

$$\begin{aligned}
\left\|\boldsymbol{\Theta}^{(k+1)}-\mathbf{1}\otimes\hat{\boldsymbol{\Theta}}\right\|_F^2 &= \sum_{i=1}^m\left\|\sum_{j=1}^m w_{ij}\tilde{\boldsymbol{\Theta}}_j^{(k)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 \leq \sum_{j=1}^m\left\|\tilde{\boldsymbol{\Theta}}_j^{(k)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 \\
&\leq \alpha\sum_{j=1}^m\left\|\boldsymbol{\Theta}_j^{\left(k+\frac{1}{2}\right)}-\hat{\boldsymbol{\Theta}}\right\|_F^2+(1-\alpha)\sum_{j=1}^m\left\|\boldsymbol{\Theta}_j^{(k)}-\hat{\boldsymbol{\Theta}}\right\|_F^2 \\
&\leq \left(1-\frac{\mu-2\epsilon_\Theta}{\tau-2\epsilon_\Theta}\alpha\right)\left\|\boldsymbol{\Theta}^{(k)}-\mathbf{1}\otimes\hat{\boldsymbol{\Theta}}\right\|_F^2+\frac{\alpha}{\epsilon_\Theta\left(\tau-2\epsilon_\Theta\right)}\left\|\boldsymbol{\Delta}^{(k)}\right\|_F^2+\frac{2m\alpha\left(\epsilon_\Theta^{-1}\tau+1\right)}{\tau-2\epsilon_\Theta}\varepsilon^{(k)}.
\end{aligned}$$

Recalling (26), we obtain the desired result (68). □

We then bound $\left\|\mathbf{D}^{(k)}\right\|_F$, which is detailed in the following proposition.

**Proposition 3.** *The following upper bound holds for* $\left\|\mathbf{D}^{(k)}\right\|_F$*:*

$$\left\|\mathbf{D}^{(k)}\right\|_F^2 \leq \left(\frac{8mL_{\max}^2}{\tau^2}+10\right)\left\|\mathbf{1}\otimes\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}^{(k)}\right\|_F^2+\frac{8m^2}{\tau^2}\left\|\mathbf{E}_Y^{(k)}\right\|_F^2+\frac{20m}{\tau}\varepsilon^{(k)}. \tag{71}$$

*Proof.* Summing (69) and (40) and replacing $\lambda\|\cdot\|_1$ with $G$ yields

$$\begin{aligned}
\tau\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\|_F^2 &\leq \left\langle\nabla F\left(\boldsymbol{\Theta}^\star\right)-m\bar{\mathbf{Y}}^{(k)}+m\bar{\mathbf{Y}}^{(k)}-m\mathbf{Y}_i^{(k)}-\tau\left(\boldsymbol{\Theta}^\star-\boldsymbol{\Theta}_i^{(k)}\right)-\boldsymbol{\delta}_i,\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\rangle+\varepsilon^{(k)} \\
&\leq \left\langle\nabla F\left(\boldsymbol{\Theta}^\star\right)-m\bar{\mathbf{Y}}^{(k)},\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\rangle+m\left\|\bar{\mathbf{Y}}^{(k)}-\mathbf{Y}_i^{(k)}\right\|_F\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\|_F \\
&\quad+\tau\left\|\boldsymbol{\Theta}^\star-\boldsymbol{\Theta}_i^{(k)}\right\|_F\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\|_F+\sqrt{2\tau\varepsilon^{(k)}}\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\|_F+\varepsilon^{(k)}.
\end{aligned}$$

Recalling (7) and the Lipschitz Smoothness of $f_i$ and using Young's inequality, we have

$$\begin{aligned}
&\tau\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\|_F^2 \\
\leq&\sum_{j=1}^m\left\langle\nabla f_j\left(\boldsymbol{\Theta}^\star\right)-\nabla f_j\left(\boldsymbol{\Theta}_j^{(k)}\right),\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\rangle+\frac{m}{2\epsilon_D}\left\|\bar{\mathbf{Y}}^{(k)}-\mathbf{Y}_i^{(k)}\right\|_F^2+\frac{\epsilon_D}{2}\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\|_F^2 \\
&+\frac{\tau^2}{2\epsilon_D}\left\|\boldsymbol{\Theta}^\star-\boldsymbol{\Theta}_i^{(k)}\right\|_F^2+\frac{\epsilon_D}{2}\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\|_F^2+\tau\epsilon_D^{-1}\varepsilon^{(k)}+\frac{\epsilon_D}{2}\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\|_F^2+\varepsilon^{(k)} \\
\leq&\sum_{j=1}^m L_j\left\|\boldsymbol{\Theta}^\star-\boldsymbol{\Theta}_j^{(k)}\right\|_F\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\|_F+\frac{m}{2\epsilon_D}\left\|\bar{\mathbf{Y}}^{(k)}-\mathbf{Y}_i^{(k)}\right\|_F^2+\frac{\tau^2}{2\epsilon_D}\left\|\boldsymbol{\Theta}^\star-\boldsymbol{\Theta}_i^{(k)}\right\|_F^2 \\
&+\frac{3\epsilon_D}{2}\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\|_F^2+\left(\tau\epsilon_D^{-1}+1\right)\varepsilon^{(k)} \\
\leq&\frac{L_{\max}^2}{2\epsilon_D}\sum_{j=1}^m\left\|\boldsymbol{\Theta}^\star-\boldsymbol{\Theta}_j^{(k)}\right\|_F^2+\frac{m}{2\epsilon_D}\left\|\bar{\mathbf{Y}}^{(k)}-\mathbf{Y}_i^{(k)}\right\|_F^2+\frac{\tau^2}{2\epsilon_D}\left\|\boldsymbol{\Theta}^\star-\boldsymbol{\Theta}_i^{(k)}\right\|_F^2+2\epsilon_D\left\|\boldsymbol{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\boldsymbol{\Theta}^\star\right\|_F^2+\left(\tau\epsilon_D^{-1}+1\right)\varepsilon^{(k)}.
\end{aligned}$$

Since

$$\frac{1}{2}\left\|\mathbf{D}_i^{(k)}\right\|^2-\left\|\hat{\mathbf{\Theta}}-\mathbf{\Theta}_i^{(k)}\right\|^2\leq\left\|\mathbf{\Theta}_i^{\left(k+\frac{1}{2}\right)}-\hat{\mathbf{\Theta}}\right\|^2,$$

and let $\epsilon_D=\frac{\tau}{4}$, we obtain

$$\left\|\mathbf{D}_i^{(k)}\right\|^2\leq\frac{8L_{\max}^2}{\tau^2}\sum_{j=1}^m\left\|\mathbf{\Theta}^\star-\mathbf{\Theta}_j^{(k)}\right\|^2+\frac{8m}{\tau^2}\left\|\bar{\mathbf{Y}}^{(k)}-\mathbf{Y}_i^{(k)}\right\|^2+10\left\|\mathbf{\Theta}^\star-\mathbf{\Theta}_i^{(k)}\right\|^2+\frac{20}{\tau}\varepsilon^{(k)}.$$

Summing over $i=1,\ldots,m$, we have the desired result (71). □

We are now ready to prove the linear rate of NetGGM. Following similar steps in Proposition 2, we have the following proposition.

**Proposition 4.** *Let $O^{(K)}(z)$, $E_\Theta^{(K)}(z)$, $E_Y^{(K)}(z)$, $D^{(K)}(z)$, and $\Xi^{(K)}(z)$ denote the transformation (41) applied to the sequences $\left\{\left\|\mathbf{\Theta}^{(k)}-\mathbf{1}\otimes\hat{\mathbf{\Theta}}\right\|_F^2\right\}$, $\left\{\left\|\mathbf{E}_\Theta^{(k)}\right\|_F^2\right\}$, $\left\{\left\|\mathbf{E}_Y^{(k)}\right\|_F^2\right\}$, $\left\{\|\mathbf{D}^{(k)}\|_F^2\right\}$, and $\{\varepsilon^{(k)}\}$, respectively. Given the free parameter $\epsilon_d>0$, the following holds:*

$$\begin{aligned}
O^{(K)}(z)&\leq a_O(\alpha,z)\left(4L_{\max}^2E_\Theta^{(K)}(z)+2E_Y^{(K)}(z)\right)+b_O(\alpha,z)+c_O(\alpha,z)\,\Xi^{(K)}(z),\\
E_\Theta^{(K)}(z)&\leq a_E(z)\,\alpha^2D^{(K)}(z)+b_\Theta(z),\\
E_Y^{(K)}(z)&\leq a_E(z)\,L_{\max}^2\left(8E_\Theta^{(K)}(z)+2\alpha^2D^{(K)}(z)\right)+b_Y(z),\\
D^{(K)}(z)&\leq a_{DO}O^{(K)}(z)+a_{DY}E_Y^{(K)}(z)+c_{DO}\Xi^{(K)}(z),
\end{aligned}\tag{72}$$

for all $z\in\left(\max\left\{1-\frac{\mu-2\epsilon_\Theta}{\tau-2\epsilon_\Theta}\alpha,\rho^2(1+\epsilon_d)\right\},1\right)$, where

$$\begin{aligned}
a_O(\alpha,z)&=\frac{\frac{\alpha m^2}{\epsilon_\Theta(\tau-2\epsilon_\Theta)}}{z-\left(1-\frac{\mu-2\epsilon_\Theta}{\tau-2\epsilon_\Theta}\alpha\right)},\quad
b_O(\alpha,z)=\frac{z\left\|\mathbf{\Theta}^{(0)}-\mathbf{1}\otimes\hat{\mathbf{\Theta}}\right\|_F^2}{z-\left(1-\frac{\mu-2\epsilon_\Theta}{\tau-2\epsilon_\Theta}\alpha\right)},\quad
c_O(\alpha,z)=\frac{\frac{2\left(\epsilon_\Theta^{-1}\tau+1\right)}{\tau-2\epsilon_\Theta}}{z-\left(1-\frac{\mu-2\epsilon_\Theta}{\tau-2\epsilon_\Theta}\alpha\right)}\\
a_E(z)&=\frac{\rho^2\left(1+\epsilon_d^{-1}\right)}{z-\rho^2(1+\epsilon_d)},\quad
b_\Theta(z)=\frac{z}{z-\rho^2(1+\epsilon_d)}\left\|\mathbf{E}_\Theta^{(0)}\right\|_F^2,\quad
b_Y(z)=\frac{z}{z-\rho^2(1+\epsilon_d)}\left\|\mathbf{E}_Y^{(0)}\right\|_F^2,\\
a_{DO}&=\frac{8mL_{\max}^2}{\tau^2}+10,\quad
a_{DY}=\frac{8m^2}{\tau^2}.\quad
c_{DO}=\frac{20m}{\tau}.
\end{aligned}$$

Chaining the inequalities in Proposition 4, we can finally prove Theorem 6.

**Theorem 12.** *Assume that Assumptions 1 and 2 are satisfied. Assume that the error $\varepsilon^{(k)}$ decays at least at an exponential rate (i.e., there exist constants $c>0$ and $\bar{\varepsilon}\in(0,1)$ such that $\varepsilon^{(k)}\leq c\bar{\varepsilon}^k$) and satisfies $\varepsilon^{(k)}<\frac{\left(\sqrt{2}-1\right)^2\underline{R}_\Theta^2}{8}$. For $\tau\geq\underline{\tau}_2$ and $\alpha\in(0,\bar{\alpha}_3)$, where*

$$\bar{\alpha}_3=\min\left\{\bar{\alpha}_1,\frac{(1-\rho)^2}{A'_{\frac{1}{2}}}\right\},$$

$$\begin{aligned}
A_{\theta,1}&=\frac{8m^2}{\mu^2}\theta^{-1}a_{DO}4\rho^2L_{\max}^2,\\
A_{\theta,2}&=\left(\frac{16m^2}{\mu^2}\theta^{-1}a_{DO}+a_{DY}\right)8\rho^4L_{\max}^2,\\
A_{\theta,3}&=\left(\frac{16m^2}{\mu^2}\theta^{-1}a_{DO}+a_{DY}\right)2\rho^2L_{\max}^2,
\end{aligned}$$

*and*

$$A_\theta = \sqrt{A_{\theta,1} + A_{\theta,2} + A_{\theta,3}},$$

*we have*

$$\sum_{i=1}^{m} \left\| \mathbf{\Theta}_i^{(k)} - \hat{\mathbf{\Theta}} \right\|_F^2 \leq (A' + Bc)\, \underline{z}'^k$$

*for all* $k \in \mathbb{N}$, *where* $A', B > 0$ *are constants defined in* (74) *and* (75), $\underline{z}'$ *is defined as*

$$\underline{z}' = \max\{\underline{z}, \bar{\varepsilon}\},$$

*and* $\underline{z}$ *is defined in* (55).

*Proof.* Chaining the inequalities in 4, we have

$$D^{(K)}(z) \leq a_D(\alpha, z)\, D^{(K)}(z) + b_D(\alpha, z) + c_D(\alpha, z)\, \Xi^{(K)}(z),$$

where $a_D(\alpha, z)$ and $b_D(\alpha, z)$ are defined in (43) and (44), and

$$c_D(\alpha, z) = a_{DO} c_O(\alpha, z) + c_{DO}.$$

Therefore, for $a(\alpha, z) < 1$, we have

$$D^{(K)}(z) \leq \frac{b_D(\alpha, z)}{1 - a_D(\alpha, z)} + \frac{c_D(\alpha, z)}{1 - a_D(\alpha, z)} \Xi^{(K)}(z). \tag{73}$$

Plugging (73) into (72), we have

$$O^{(K)}(z) \leq a(\alpha, z) \frac{b_D(\alpha, z)}{1 - a_D(\alpha, z)} + b(\alpha, z) + c(\alpha, z)\, \Xi^{(K)}(z),$$

where $a(\alpha, z)$ and $b(\alpha, z)$ are defined in (46) and (47), and

$$c(\alpha, z) = a(\alpha, z) \frac{c_D(\alpha, z)}{1 - a_D(\alpha, z)} + c_O(\alpha, z).$$

According to Lemma 7, we have

$$\left\| \mathbf{\Theta}^{(k)} - \mathbf{1} \otimes \hat{\mathbf{\Theta}} \right\|_F^2 \leq \left( A' + B \Xi^{(K)}(z) \right) z^k,$$

where

$$A' = a(\alpha, z) \frac{b_D(\alpha, z)}{1 - a_D(\alpha, z)} + b(\alpha, z), \tag{74}$$

$$B = c(\alpha, z). \tag{75}$$

This means that if $\Xi^{(K)}(z)$ has an upper bound, $\left\{ \left\| \mathbf{\Theta}^{(k)} - \mathbf{1} \otimes \hat{\mathbf{\Theta}} \right\|^2 \right\}$ vanishes R-linearly at a rate of at least $z$.

Following the same steps in the proof of Theorem 9 and choosing $\epsilon_\Theta = \frac{\mu}{4}$, we can prove that $a_D(\alpha, z) < 1$ when $\alpha \in (0, \bar{\alpha}_3)$ and $z \in (\underline{z}, 1)$. In addition, since $\varepsilon^{(k)} \leq c\bar{\varepsilon}^k$ and $\underline{z}' \geq \bar{\varepsilon}$, we have $\Xi^{(K)}(z) \leq c$, and hence

$$\left\| \mathbf{\Theta}^{(k)} - \mathbf{1} \otimes \hat{\mathbf{\Theta}} \right\|_F^2 \leq (A' + Bc)\, \underline{z}'^k,$$

which proves Theorem 12. □

# F ADDITIONAL EXPERIMENTS

We also compare the execution times of all methods as shown in Table 2. NetGGM requires sacrificing some computation time to achieve the same level of accuracy as the centralized algorithm. The convergence time of NetGGM is influenced by various factors. As shown in the table, the larger the number of nodes in the network and the lower the network connectivity, the longer the convergence time. In the worst-case scenario, NetGGM can estimate a 50-dimensional precision matrix using 25 samples on a linear network with 20 agents within approximately 300 seconds. As for the longer runtime of the D&C methods, this is because we use G-ISTA to obtain local estimates, and the limited number of local samples at each agent leads to slower convergence of G-ISTA.

Table 2: Average and variance (between parentheses) of CPU times of NetGGM and baselines on synthetic data.

| **Cliques model** | Methods | | CPU time (s) | |
|---|---|---|---|---|
| | G-ISTA | | 0.6592 (0.0518) | |
| | Number of agents $m$ | 5 | 10 | 20 |
| $N = 25$ | NetGGM ($p = 0.9$) | 8.7564 (74.0955) | 17.2320 (305.1681) | 51.9478 (2173.9923) |
| | NetGGM ($p = 0.5$) | 10.9004 (101.2909) | 23.5953 (718.8357) | 60.4709 (3062.2584) |
| | NetGGM (line) | 17.3498 (319.7095) | 50.6574 (1210.2413) | 308.0480 (11031.5613) |
| | D&C I | 10.4738 (206.9460) | 282.5218 (8634.6486) | 541.5241 (26244.6776) |
| | D&C II | 10.3181 (205.4229) | 281.9859 (8363.1977) | 539.7789 (22503.2595) |
| | G-ISTA | | 0.9815 (0.0513) | |
| | Number of agents $m$ | 5 | 10 | 20 |
| $N = 100$ | NetGGM ($p = 0.9$) | 10.2981 (43.7608) | 20.1145 (130.9594) | 61.7357 (1736.4533) |
| | NetGGM ($p = 0.5$) | 12.5014 (59.6681) | 30.7408 (475.2664) | 71.0667 (1993.2196) |
| | NetGGM (line) | 18.1808 (163.9079) | 57.6531 (775.9505) | 312.7071 (7117.7892) |
| | D&C I | 0.2916 (0.0347) | 39.5273 (4372.2945) | 262.1904 (17539.3305) |
| | D&C II | 0.3012 (0.0333) | 39.6439 (4390.4094) | 261.6074 (16182.8145) |
| **Random model** | Methods | | CPU time (s) | |
| | G-ISTA | | 0.6803 (0.0488) | |
| | Number of agents $m$ | 5 | 10 | 20 |
| $N = 25$ | NetGGM ($p = 0.9$) | 6.9368 (54.5397) | 14.5805 (155.1912) | 58.6180 (2418.4794) |
| | NetGGM ($p = 0.5$) | 9.3587 (77.3082) | 18.7734 (177.8163) | 61.8095 (2293.8800) |
| | NetGGM (line) | 13.7527 (178.8802) | 49.5722 (1119.4247) | 274.0014 (10676.5477) |
| | D&C I | 32.6488 (563.2589) | 94.4452 (1159.9948) | 573.9397 (1822.7583) |
| | D&C II | 32.6147 (571.9648) | 95.6593 (1087.2280) | 575.3351 (2492.2283) |
| | G-ISTA | | 0.6937 (0.0267) | |
| | Number of agents $m$ | 5 | 10 | 20 |
| $N = 100$ | NetGGM ($p = 0.9$) | 5.9506 (11.7812) | 14.0879 (73.8083) | 46.6213 (1219.8495) |
| | NetGGM ($p = 0.5$) | 8.3510 (19.4260) | 22.4274 (108.4904) | 56.1714 (1543.4996) |
| | NetGGM (line) | 12.0197 (38.4656) | 36.6198 (302.9742) | 220.6651 (4541.1462) |
| | D&C I | 12.4499 (313.0276) | 31.9262 (2657.2532) | 275.9662 (7110.5214) |
| | D&C II | 12.4423 (307.0222) | 31.1597 (2532.3115) | 277.2202 (7081.5730) |