
Implicit Regularization of Decentralized Gradient Descent for Decentralized Sparse Regression

Tongle Wu

The Pennsylvania State University
tfw5381@psu.edu

Ying Sun

The Pennsylvania State University
ybs5190@psu.edu

Abstract

We consider learning a sparse model from linear measurements taken by a network of agents. Different from existing decentralized methods designed based on the LASSO regression with explicit ℓ_1 norm regularization, we exploit the implicit regularization of the decentralized optimization method applied to an over-parameterized nonconvex least squares formulation without sparse penalization. Our first result shows that despite nonconvexity, if the network connectivity is good, the well-known decentralized gradient descent algorithm (DGD) with small initialization and early stopping can compute the statistically optimal solution. Sufficient conditions on the initialization scale, choice of step size, network connectivity, and stopping time are further provided to achieve convergence. Our result recovers the convergence rate of gradient descent in the centralized setting, showing its tightness. Based on the analysis of DGD, we further propose a communication-efficient version, termed T-DGD, by truncating the iterates before transmission. In the high signal-to-noise ratio (SNR) regime, we show that T-DGD achieves comparable statistical accuracy to DGD, while the communication cost is logarithmic in the number of parameters. Numerical results are provided to validate the effectiveness of DGD and T-DGD for sparse learning through implicit regularization.

1 Introduction

Modern deep learning is generally in the over-parameterized regime where the models have significantly more parameters than available training examples [6, 41]. Although deep learning models exhibit remarkable performance in multiple domains, the theoretical understanding of optimization and generalization for deep learning is still limited. Recent studies show that despite being over-parameterized, gradient-based methods applied to minimize the empirical loss exhibit the implicit regularization phenomenon. For example, a line of works [3, 5, 11, 16] shows that with certain initialization, networks trained with gradient descent (GD) land in the “kernel regime” and share similar behaviors to the kernel method. However, the literature [2, 9, 24] suggests that kernel regime analyses fall short in explaining the success of deep learning because neural networks analyzed in the kernel regime are almost linearized, thus hindering feature learning from data. Further, many works [35, 38, 40] start investigating the “rich regime”, showing that GD with small initialization induces structures on the solution, such as sparsity and low-rankness, that better explains the generalization capability of NNs. However, all aforementioned results are limited to the centralized setting, where data are stored on a single machine. Practical constraints such as limited computing and storage resources, data privacy and security, and regulation rules make the centralized learning framework increasingly inadequate for contemporary applications. Although a variety of decentralized learning algorithms can be applied to NN training, the questions of which solution they can converge, along with its generalization performance, are largely unclear.

In this paper, we study the sparse learning problem [34] in the overparameterized regime, which shares many key characteristics with deep learning models but is more tractable to analyze, as a prototype for understanding the computation and statistical guarantees of decentralized learning algorithms. Specifically, we consider learning a sparse model $\mathbf{w}^*$ from its noisy linear measurements over m agents. These agents communicate over an undirected connected mesh network without a central coordinator, and each agent can only communicate with its one-hop neighbors. The i -th agent has its own n samples $\{(\mathbf{x}_{i,j}, y_{i,j})\}_{j=1}^n$. Each j -th data pair $(\mathbf{x}_{i,j}, y_{i,j})$ is generated according to the noisy linear model

$$y_{i,j} = \mathbf{x}_{i,j}^T \mathbf{w}^* + \xi_{i,j}, \quad \forall i \in [m] \text{ and } \forall j \in [n], \quad (1)$$

where $\mathbf{x}_{i,j} \in \mathbb{R}^d$ and $y_{i,j} \in \mathbb{R}$ denotes respectively the j -th feature and its corresponding response at i -th agent, $\xi_{i,j}$ is the observation noise, and $\mathbf{w}^* \in \mathbb{R}^d$ is the sparse model parameter to be learned, common to all agents, and has only s ($s \ll d$) non-zero elements. We are interested in the high-dimensional setting where the ambient dimension d is substantially larger than the total sample size $N := mn$, i.e., $d \gg N$. By re-parameterizing $\mathbf{w} = \mathbf{u} \odot \mathbf{u} - \mathbf{v} \odot \mathbf{v}$, the loss function can be formulated as minimizing the following regularization-free nonlinear least square problem:

$$F(\mathbf{u}, \mathbf{v}) := \frac{1}{m} \sum_{i=1}^m f_i(\mathbf{u}, \mathbf{v}), \quad \text{with} \quad f_i(\mathbf{u}, \mathbf{v}) := \frac{1}{n} \|\mathbf{X}_i(\mathbf{u} \odot \mathbf{u} - \mathbf{v} \odot \mathbf{v}) - \mathbf{y}_i\|^2; \quad \forall i \in [m], \quad (2)$$

where $f_i(\mathbf{u}, \mathbf{v})$ corresponds to the loss function of i -th agent. Problem (2) can also be regarded as the supervised learning problem on the diagonal linear network of degree-2 [38].

Problem (2) is highly non-convex with $\mathbf{u}$ and $\mathbf{v}$, however, recent works [12, 36, 42] demonstrate that if the design matrix satisfies the restricted isometric property (RIP) condition, the centralized GD without any regularization can yield the statistically optimal estimator with properly chosen initialization scale, step size, and early stopping time. This intriguing phenomenon is derived from the implicit regularization of GD. Roughly speaking, with small initialization, the gap $|u_i^2 - v_i^2|$ would increase with iteration for coordinate i such that $w_i^* \neq 0$, where the remaining ones stay small enough before early stopping. As a result, GD identifies the support of $\mathbf{w}^*$ as the algorithm processes. In the decentralized setting, general results from the pure optimization perspective can only certify convergence to the stationary points, implying neither global optimality nor generalizability. Given the encouraging result of GD achieved in the centralized setting, it is natural to ask if statistically optimal solutions are also computable by decentralized gradient-type algorithms, which algorithm can achieve the goal, and what are the regularity conditions.

This paper aims to analyse the renowned decentralized gradient algorithm (DGD) for minimizing (2) over the undirected mesh networks. The main contributions of this paper are detailed as follows.

- **Statistical guarantee.** It is well established that even for convex objectives, DGD cannot compute an exact minimizer. It only converges to the neighborhood of the solution whose radius depends on the step size. However, we show that under specific conditions—namely, if the global design matrix satisfies the RIP condition, the initialization scale is sufficiently small, and the network is sufficiently connected—the solution computed by DGD with early stopping is statistically optimal.
- **Computational complexity.** Our convergence analysis reveals that the early stopping time increases logarithmically with the ambient dimension d . While network connectivity does not affect statistical error when it satisfies mild conditions, it does influence the stopping time of DGD to find the optimal estimator. Networks with poor connectivity will delay the early stopping time, and thus increase the iteration complexity.
- **Technical analysis.** Compared to the techniques used for analyzing the centralized GD [33, 38], proving the convergence of DGD faces the following challenge. Because the consensus error terms induced from the mesh network result in a perturbed version of the multiplicative update. Compared with the exact multiplicative updates, the challenge is that the additional error term outside of multiplication prevents applying the centralized analysis directly. In addition, the error terms within the multiplication have more complicated consensus error terms than that of the centralized setting, which requires bounding the consensus error terms carefully to control these error terms that can achieve the same order statistical error. To achieve this goal, we separately control the consensus errors on support $\mathcal{S}$ and non-support $\mathcal{S}^c$ by the magnitudes of parameters on support $\mathcal{S}$ and non-support $\mathcal{S}^c$, respectively. Our fine-grained analysis for consensus errors is distinct

from existing decentralized optimization analyses that bound the consensus errors uniformly. The additional error term also complicates the transfer of proof from the simplified non-negative w^* case to the general w^* setting as the centralized setting, we conduct a comprehensive induction process to both u and v simultaneously for general w^* .

- **Truncated DGD.** We propose a communication-efficient truncated DGD (T-DGD) method that at each iteration, vectors being transmitted are truncated, keeping only s elements with the largest magnitudes nonzero. We prove that if each agent has sufficient samples and the signal-to-noise ratio is high enough, T-DGD can perform as well as the vanilla DGD while reducing communication complexity to logarithmic dependence on ambient dimension d .

2 Related works

We categorize the existing works most relevant to our study into three main groups.

- **Implicit regularizations for sparse regression.** The recent study in [15] reparameterized the model parameter through overparameterized Hadamard product and discovered encouraged empirical performance by the first-order optimization algorithms. The statistical and convergence guarantees for this phenomenon are established in [36, 42] under mild conditions. Woodworth et al. [38] studied the impact of initialization scale on solutions. Scott et al. [27] demonstrated the benefit of stochasticity of SGD in sparse regression and explored the impact of momentum in [25]. The more recent process in understanding the linear diagonal networks can be found in references [7, 10, 12, 19, 23, 26, 43]. To the best of our knowledge, existing works have only discussed implicit regularizations induced by centralized optimization methods in linear diagonal networks. However, the question of whether decentralized algorithms induce implicit regularizations, and what type of implicit regularization they may induce has not been studied so far.
- **Decentralized sparse regression with explicit regularization.** For estimating ground truth w^* in high-dimensional sparse linear regression under the decentralized setting, Ji et al. [18] proposed DGD-CTA algorithm for tackling LASSO objective with consensus penalty and proved linear convergence rate to the neighbor of the statistical optimal estimator, but the convergence rate has polynomial dependence on ambient dimension d . Further Sun et al. [31] proposed the NetLASSO based on the gradient tracking method and obtained d -independent convergence rate and optimal statistical accuracy. To complement work [18], Ji et al. proposed DGD-ATC by mixing the local gradients along iterations and achieved logarithmic dependence on d in [17]. Maros et al. [21] proposed DGD² method based on a double mixing for solving decentralized LASSO and obtained similar theoretical guarantees in [17, 31]. To improve the computation efficiency, Maros et al. [20] integrated accelerated proximal gradient descent with gradient tracking to solve decentralized LASSO. Despite these developments, it remains unclear whether leveraging the unregularized overparameterization and implicit regularization of decentralized optimization methods can achieve the optimal statistical guarantee over mesh networks.
- **Implicit regularizations of decentralized optimization.** Implicit bias or regularizations of centralized optimization methods for overparameterized models have been extensively studied [13, 30], but only a few works have investigated the implicit regularization of decentralized optimization methods. Richards et al. [29] studied the implicit regularization for decentralized stochastic gradient descent for solving general unregularized convex problems. Zhu et al. [44] demonstrated that decentralized stochastic gradient descent implicitly executes the sharpness-aware minimization algorithm for general non-convex problems. Taheri et al. [33] studied the implicit regularization of DGD in overparameterized classification for separable data. Recent work [22] demonstrated the implicit regularization of the DGD² in solving the overparameterized matrix sensing problem. Different from these works, we establish the statistical and computational results for specific non-convex sparse regression problem.

3 Preliminaries

In this section, we will introduce the basic notations used in this paper, and then formulate the DGD for solving the problem (2). Finally, we provide the necessary assumptions and definitions for the decentralized sparse regression problem.

3.1 Notations

Throughout this paper, we use $[m]$ to denote set $\{1, \dots, m\}$ for given positive integer m , $\mathbf{1}_d$ denotes d -dimensional vector that all elements are one and $\mathbf{I}_d$ denotes d -dimensional identity matrix. For ground truth parameter $\mathbf{w}^*$, the relevant notations are support set $\mathcal{S} := \{j | w_j^* \neq 0\}$, positive support set $\mathcal{S}^+ := \{j | w_j^* > 0\}$, negative support set $\mathcal{S}^- := \{j | w_j^* < 0\}$ and non-support set $\mathcal{S}^c := \{j | w_j^* = 0\}$, $w_{\max}^* := \max_{j \in \mathcal{S}} |w_j^*|$ and $w_{\min}^* := \min_{j \in \mathcal{S}} |w_j^*|$. $\forall \mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}_{\mathcal{S}} := \mathbf{1}_{\mathcal{S}} \odot \mathbf{x}$ where $\mathbf{1}_{\mathcal{S}}$ denotes a vector equal to one for all coordinates $j \in \mathcal{S}$ and equal to zero everywhere else. Symbol “ $\odot$ ” denotes Hadamard product that $(\mathbf{a} \odot \mathbf{b})_j = a_j b_j, \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. The averaged signal is defined as $\bar{\mathbf{w}}^t := \frac{1}{m} \sum_{i=1}^m \mathbf{w}^{t,i}$ and similar notations can be extended to $\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t$.

$\mathbf{X} := [\mathbf{X}_1; \dots; \mathbf{X}_m]$ denotes the concatenated sample matrix, where each row of $\mathbf{X}_i$ represents one feature vector in agent i . $\|\cdot\|$ denotes the Frobenius norm for vector and the spectral norm (maximum singular value) for matrix. $\|\mathbf{A}\|_{\infty} := \max_{i,j} |A_{ij}|$ denotes infinity norm. We use $a = \mathcal{O}(b)$ to denote that inequality $a \leq Cb$ holds with some absolute constants C that do not depend on any parameters of the problem. The notation $a \lesssim b$ shares the same meaning as $a = \mathcal{O}(b)$. Finally, we use $a \gtrsim b$ if there exists a universal constant c such that $a \geq cb$.

3.2 Method and Assumptions

We focus on DGD solving problem (2) over mesh network, modeled as an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where nodes $\mathcal{V} = \{1, \dots, m\}$ represent the set of agents and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ represent the communication links. An unordered pair $\{i, j\}$ is included in $\mathcal{E}$ if and only if there is a bidirectional communication link between agent i and j . The set of one-hop neighbors for agent i is denoted by $\mathcal{N}_i := \{j \in \mathcal{V} | (i, j) \in \mathcal{E}\} \cup \{i\}$.

DGD allows each agent to independently update its parameters based on local gradient descent and then synchronize with neighboring agents by weighted averaging these updates. The recursive iteration of DGD for each agent is described as follows.

$$\mathbf{u}^{t+1,i} = \sum_{j=1}^m W_{ij} \left(\mathbf{u}^{t,j} - \eta \frac{4}{n} \mathbf{u}^{t,j} \odot \left(\mathbf{X}_j^T (\mathbf{X}_j (\mathbf{u}^{t,j} \odot \mathbf{u}^{t,j} - \mathbf{v}^{t,j} \odot \mathbf{v}^{t,j}) - \mathbf{y}_j) \right) \right), \forall i \in [m]; \quad (3)$$

$$\mathbf{v}^{t+1,i} = \sum_{j=1}^m W_{ij} \left(\mathbf{v}^{t,j} + \eta \frac{4}{n} \mathbf{v}^{t,j} \odot \left(\mathbf{X}_j^T (\mathbf{X}_j (\mathbf{u}^{t,j} \odot \mathbf{u}^{t,j} - \mathbf{v}^{t,j} \odot \mathbf{v}^{t,j}) - \mathbf{y}_j) \right) \right), \forall i \in [m], \quad (4)$$

where the $\mathbf{w}^{t,i} := \mathbf{u}^{t,i} \odot \mathbf{u}^{t,i} - \mathbf{v}^{t,i} \odot \mathbf{v}^{t,i}$ denotes the local estimator in agent i at t^{th} iteration, the initialization is $\mathbf{u}^{0,i} = \mathbf{v}^{0,i} = \alpha \mathbf{1}_d, \forall i \in [m]$ and η is constant step size. $\mathbf{W}$ is the nonnegative weight mixing matrix for the undirected mesh network, where $W_{ij} > 0$ if there is a link between agents i and j , and $W_{ij} = 0$ otherwise. The mixing matrix $\mathbf{W}$ related to the undirected graph satisfies the following assumption.

Assumption 1. The communication network $\mathcal{G}$ is connected. The weight matrix $\mathbf{W} = [w_{ij}]_{i,j=1}^m$ for this graph has the following properties: (i) $w_{ij} = 0$ for all pairs (i, j) that are not in $\mathcal{E}$; (ii) it is double stochastic that $\mathbf{1}_m^T \mathbf{W} = \mathbf{1}_m^T$ and $\mathbf{W} \mathbf{1}_m = \mathbf{1}_m$; (iii) the spectral gap $\rho := \|\mathbf{W} - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T\| \leq 1$.

This assumption is common in decentralized optimization literature [17, 22]. We need the following RIP condition which is a key condition to obtain the optimal estimator for sparse regression.

Definition 1. The global design matrix $\mathbf{X}/\sqrt{N} \in \mathbb{R}^{N \times d}$ satisfies the (δ, s) -Restricted Isometry Property (RIP) if for any s -sparse vector $\mathbf{w} \in \mathbb{R}^d$, there is $(1 - \delta) \|\mathbf{w}\|^2 \leq \left\| \mathbf{X} \mathbf{w} / \sqrt{N} \right\|^2 \leq (1 + \delta) \|\mathbf{w}\|^2$.

The RIP condition was first introduced in the compressed sensing literature in [8] which is a little more restrictive condition to achieve optimal statistical rate than the restricted eigenvalue condition in [1]. We inherit this assumption in the centralized setting [36, 42] to achieve optimal estimator error under the condition that parameter δ is upper bounded. Besides the global RIP condition, we have the following local RIP condition for local design matrices $\{\mathbf{X}_i\}_{i=1}^m$.

Definition 2. The local design matrices $\mathbf{X}_1/\sqrt{n}, \dots, \mathbf{X}_m/\sqrt{n} \in \mathbb{R}^{n \times d}$ satisfy the local $(\delta_{\max}, s)$ -(RIP) condition, if for any s -sparse vector $\mathbf{w} \in \mathbb{R}^d$ and any local design matrix $\mathbf{X}_i/\sqrt{n}$, there is $(1 - \delta_{\max}) \|\mathbf{w}\|^2 \leq \|\mathbf{X}_i \mathbf{w}/\sqrt{n}\|^2 \leq (1 + \delta_{\max}) \|\mathbf{w}\|^2$.

The definition of the local RIP condition is just for ease of proof presentation, as we do not necessitate any upper bound on the local RIP parameter $\delta_{\max}$.

4 Main Result

Based on the above method and assumptions, we now give theoretical guarantees of DGD in solving problem (2) for sparse regression problem (1) as follows.

Theorem 1. Considering the sequence generated by (3) and (4) based on DGD for solving problem (2) and $\forall \epsilon > 0$, if the global design matrix $\mathbf{X}/\sqrt{N}$ satisfies $(\delta, s+1)$ -RIP condition with bounded RIP parameter $\delta \lesssim \frac{1}{\sqrt{s}}$, the local design matrices $\{\mathbf{X}_i/\sqrt{n}\}_{i=1}^m$ satisfy local $(\delta_{\max}, s+1)$ -RIP condition, and the mesh network satisfies assumption 1, the initialization satisfies $\alpha \lesssim \min \left\{ 1, \frac{\epsilon^2}{(12d+1)^2}, \frac{\epsilon}{w_{\max}^*}, \frac{\zeta}{6(w_{\max}^*)^2}, \frac{w_{\min}^*}{4} \right\}$, the constant stepsize η satisfies

$$\eta \lesssim \min \left\{ \frac{1 - \sqrt{\rho}}{64\sqrt{\rho}w_{\max}^*}, \frac{\log \frac{1}{\alpha^4} \left(1 - \sqrt{\frac{1+\sqrt{\rho}}{2}} \right)}{4w_{\max}^*}, \frac{1 - \left(\frac{1+\sqrt{\rho}}{2} \right)^{\frac{1}{4}}}{w_{\max}^*} \right\}, \quad (5)$$

and the spectral gap ρ satisfies

$$\rho^{\frac{1}{4}} \lesssim \min \left\{ \frac{1}{\sqrt{s}\delta_{\max} + 1}, \frac{\delta}{8\delta_{\max}}, \frac{\left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_{\infty}}{8 \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty}} \right\}, \quad (6)$$

then after running $t = \mathcal{O} \left(\frac{1}{\eta\zeta} \log \frac{1}{\alpha} \right)$ iterations. There would be

$$|\bar{w}_j^t - w_j^*| \leq \begin{cases} \mathcal{O}(\varsigma) & \text{if } j \in \mathcal{S} \text{ and } w_{\min}^* \leq \mathcal{O}(\varsigma) \\ \mathcal{O} \left(\max \left\{ \left| \left(\frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right)_j \right|, \delta \sqrt{s} \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \odot \mathbf{1}_{\mathcal{S}} \right\|_{\infty}, \epsilon \right\} \right) & \text{if } j \in \mathcal{S} \text{ and } w_{\min}^* \geq \mathcal{O}(\varsigma) \\ \mathcal{O}(\sqrt{\alpha}) & \text{if } j \notin \mathcal{S}, \end{cases} \quad (7)$$

where $\varsigma := \max \left\{ \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_{\infty}, \epsilon \right\}$, $\zeta := \max \left\{ \frac{w_{\min}^*}{5}, 960\varsigma \right\}$.

- **Mechanism to promote sparsity.** The consensus errors induced from decentralized network complicate the multiplicative updates, which becomes inexact multiplicative updates as $\bar{\mathbf{u}}^{t+1} = \bar{\mathbf{u}}^t \odot \left(1 - 4\eta \left(\bar{\mathbf{u}}^t \odot \bar{\mathbf{u}}^t - \mathbf{w}^* + \hat{\mathbf{p}}^t + \hat{\mathbf{b}}^t \right) \right) + \mathbf{e}^t$. Compared with the exact multiplicative updates of GD in [36], the challenge is that the extra error term $\mathbf{e}^t$ outside of the multiplication prevents applying the centralized analysis trivially. In addition, the perturbation error terms $\hat{\mathbf{p}}^t, \hat{\mathbf{b}}^t$ within the multiplication are much more complicated than that of the centralized setting due to additional multiple consensus errors. This requires bounding the consensus error terms carefully, which should control the complicated perturbation errors $\hat{\mathbf{p}}^t, \hat{\mathbf{b}}^t, \mathbf{e}^t$ not to be large. Thus, we can use network connectivity to control the consensus errors to bound these three perturbation errors small enough to make the distance between two trajectories obtained by inexact and exact multiplicative updates within statistical accuracy, which can promote sparsity in the decentralized setting. The detailed theoretical mechanism of promoting sparsity has been demystified in Proposition 3.
- **Statistical Guarantee.** Based on the result in (7) and conditions in (5) and (6), we can observe that if the initialization α is small enough and network connectivity is sufficiently well, the DGD with early stopping can obtain the desired estimator for sparse ground truth parameter $\mathbf{w}^*$ that achieves the same order of statistical error as the centralized setting in [36]. The formula in (7) not only

illustrates that we establish the network-independent estimator error bound but also inherits the benefit of implicit regularization, which indicates that if the signal-to-noise is high enough, the statistical error is independent of ambient dimension d . In contrast, existing results in decentralized LASSO methods [17, 18, 21, 31], have consistent dependence on d in any case.

- *Computational Complexity.* The iteration complexity of early stopping is network-dependent that is because the $t = \mathcal{O}\left(\frac{1}{\eta\zeta} \log \frac{1}{\alpha}\right)$ has the dependence on stepsize, which should satisfy the condition in (5). This suggests that poorer network connectivity leads to higher computational complexity. Although the initialization has no dependence on network connectivity, $\frac{1}{\alpha}$ has polynomial dependence on d , and the dependence of complexity on d is just logarithmic, which is similar to DGD in solving LASSO in [17, 21] and improves the polynomial dependence on d in [18].
- *Dependence on network connectivity.* For accurate estimation, it is essential that the network should be well-connected, as specified in condition (6). When this condition is not satisfied, we can run multiple rounds of communication per iteration. It is observable that the smaller ratio between the global RIP parameter δ and the local RIP parameter $\delta_{\max}$, and smaller ratio between the local noise and the global noise magnitude, necessitate a higher degree of network connectivity. This can be understood from the perspective of heterogeneity, where smaller ratios indicate a significant disparity between local and global design matrices. Consequently, condition (6) is reasonable as it suggests that higher levels of heterogeneity necessitate improved network connectivity. In numerical experiments, we can observe that if ρ does not satisfy the condition as (6), obtaining optimal statistical error is not achievable, which indicates the optimal statistical error undergoes a phase transition with the network connectivity.

Our results demonstrate the benefit of overparameterization for DGD. Theorem 1 shows that standard DGD is sufficient to provide a satisfied statistical estimator with efficient computation without gradient correction techniques. This finding challenges the widely held belief in decentralized optimization literature that extra techniques like gradient tracking and other gradient-correction-based methods are necessary for heterogeneous scenarios [39, 32]. The following corollary considers the well-known instance where the design matrix and noise are generated from sub-Gaussian distribution, which indicates that DGD with early stopping can achieve the minimax optimal statistical rate under the ℓ_2 metric.

Corollary 1. Suppose that entries of global design matrix $\mathbf{X}$ generated from i.i.d 1-sub-Gaussian distribution, and the total sample size satisfies $N \gtrsim s \left(s \log \frac{ed}{s} + \log \frac{dN}{m} \right)$. The noise vector $\boldsymbol{\xi}$ is generated from independent σ^2 -sub-Gaussian entries, and the initialization is set as Theorem 1 with

$$\epsilon = \mathcal{O}\left(\sigma \sqrt{\frac{\log d}{N}}\right). \text{ If the spectral gap satisfies } \rho \lesssim \frac{1}{m^4} \text{ and stepsize is set as } \eta = \mathcal{O}\left(\frac{1 - \left(\frac{1+\sqrt{\rho}}{2}\right)^{\frac{1}{4}}}{w_{\max}^*}\right),$$

then after running $t = \mathcal{O}\left(\frac{w_{\max}^* \sqrt{N}}{\sigma \sqrt{\log d} \left(1 - \left(\frac{1+\sqrt{\rho}}{2}\right)^{\frac{1}{4}}\right)} \log \frac{1}{\alpha}\right)$ iterations, the sequence generated by (3)

and (4) based on DGD for solving problem (2) would obtain estimator that $\|\bar{\mathbf{w}}^t - \mathbf{w}^*\| \lesssim \sigma \sqrt{\frac{s \log d}{N}}$ with probability at least $1 - \frac{3}{8d^3}$.

Corollary 1 indicates that in the sub-Gaussian setting, network-independent statistical error obtained by DGD matches optimal rate $\mathcal{O}\left(\sigma \sqrt{\frac{s \log d}{N}}\right)$ under ℓ_2 metric in the centralized setting [28]. In this context, the condition for network connectivity implies that the smaller ρ is required as the number of agents m increases. This is reasonable because when the total sample size N is fixed, an increase in the number of agents results in fewer samples assigned to each agent. Consequently, better network connectivity is necessary to achieve optimal estimation.

5 Communication Efficient DGD via Truncation

It is apparent that iterations in (3) and (4) of DGD, each agent has to transmit two d -dimensional vectors $\mathbf{u}^{t,i}$ and $\mathbf{v}^{t,i}$ to its neighboring agents per iteration. Because we are considering the high-dimensional regime where the feature has ultra-high dimension, which leads to the $\mathcal{O}\left(d \cdot \frac{1}{\eta\zeta} \log \frac{1}{\alpha}\right)$ high communication complexity (in terms of the bits transmitted) for DGD. The primary idea is

whether it is possible to transmit fewer partial elements instead of the entire d -dimensional vectors for $\mathbf{u}^{t,i}, \mathbf{v}^{t,i}$ in all rounds of communication. Since all elements of $\mathbf{u}^{t,i}$ and $\mathbf{v}^{t,i}$ equal to α at initialization, we can utilize the one step of local gradient descent step in each agent to distinguish the support set and non-support based on changes of magnitudes for each element. The intuition is that the elements on the support would grow more rapidly than those on the non-support. Thus, we propose the **Truncated Decentralized Gradient Descent (T-DGD)** as

$$\begin{aligned}\mathbf{u}^{t+1,i} &= \sum_{j=1}^m W_{ij} \cdot \text{Trun}_s \left(\left(\mathbf{u}^{t,j} - \eta \frac{4}{n} \mathbf{u}^{t,j} \odot \left(\mathbf{X}_j^T (\mathbf{X}_j (\mathbf{u}^{t,j} \odot \mathbf{u}^{t,j} - \mathbf{v}^{t,j} \odot \mathbf{v}^{t,j}) - \mathbf{y}_j) \right) \right) \right); \\ \mathbf{v}^{t+1,i} &= \sum_{j=1}^m W_{ij} \cdot \text{Trun}_s \left(\left(\mathbf{v}^{t,j} + \eta \frac{4}{n} \mathbf{v}^{t,j} \odot \left(\mathbf{X}_j^T (\mathbf{X}_j (\mathbf{u}^{t,j} \odot \mathbf{u}^{t,j} - \mathbf{v}^{t,j} \odot \mathbf{v}^{t,j}) - \mathbf{y}_j) \right) \right) \right),\end{aligned}\quad (8)$$

for $\forall i \in [m]$, where $\text{Trun}_s(\mathbf{x})$ is the operator that preserves only the s largest magnitude elements of the vector $\mathbf{x}$ while setting all other elements to zero. The following proposition shows the benefit of T-DGD in sparse regression under proper conditions.

Proposition 1. *With the same setup in Corollary 1, if the ground truth $\mathbf{w}^*$ satisfies $\frac{w_{\min}^*}{2} \gtrsim \sqrt{s} \delta_{\max} w_{\max}^* + \sigma \sqrt{\frac{\log d}{n}}$, then the sequence generated by T-DGD as (8) for solving problem (2) would obtain estimator that $\|\bar{\mathbf{w}}^t - \mathbf{w}^*\| \lesssim \sigma \sqrt{\frac{s \log d}{N}}$ with probability at least $1 - \frac{3}{8d^3}$. However, the communication complexity in terms of transmitted bits would be at most $\mathcal{O} \left(s \cdot (w_{\max}^* \sqrt{N}) / \left(\sigma \sqrt{\log d} \left(1 - \left(\frac{1+\sqrt{\rho}}{2} \right)^{\frac{1}{4}} \right) \right) \log \frac{1}{\alpha} \right)$.*

To ensure condition $\frac{w_{\min}^*}{2} \gtrsim \sqrt{s} \delta_{\max} w_{\max}^* + \sigma \sqrt{\frac{\log d}{n}}$ satisfied, it is necessary to require that each agent has sufficient samples and SNR is high enough such that $\delta_{\max} \lesssim \frac{w_{\min}^*}{\sqrt{s w_{\max}^*}}$ and $\sigma \sqrt{\frac{\log d}{n}} \lesssim w_{\min}^*$. This proposition enables each agent to transmit only s elements of d -dimensional vector per communication round, which can achieve optimal statistical rate and eliminate the d -linear increasing communication complexity. The result in Proposition 1 validates the usefulness of the Hadamard product over-parameterization in decentralized gradient-based optimization.

6 Numerical Results

This section conducts the experimental studies to evaluate the theoretical findings of DGD and T-DGD for solving problem (2) in Subsection 6.1, Subsection 6.2, respectively. In Subsection 6.3, we compare the effectiveness of implicit regularization of DGD with explicit regularization based decentralized methods. The communication networks $\mathcal{G}$ are generated from Erdős Rényi (ER) graphs with link activation under given probabilities. By default, unless stated otherwise, all the design matrices $\mathbf{X}$ have i.i.d. standard Gaussian elements, noise $\boldsymbol{\xi}$ follows i.i.d. $\mathcal{N}(0, 0.5^2)$ distribution, and the magnitudes of elements on support $\mathcal{S}$ are 1. All experiments are conducted on 12th Gen Intel(R) Core(TM) i7-12700@2.10GHz processor and 16.0GB RAM under Windows 11 system.

6.1 Simulations on DGD

We organize the experiments as follows: 1) We visualize the dynamics of averaged variables and consensus errors that allow us to evaluate the implicit regularization of DGD and the soundness of our technical analysis. 2) We check whether DGD can achieve optimal statistical error, the impacts of ambient dimension d and initialization scale α on statistical and computational properties. 3) We evaluate the condition of (6) that reveals the relationship between network connectivity and network scale for achieving the statistical accuracy of centralized setting.

- **Dynamics of $\bar{\mathbf{w}}^t, \bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t$ and $\mathbf{u}^{t,i} - \bar{\mathbf{u}}^t, \mathbf{v}^{t,i} - \bar{\mathbf{v}}^t$.** In this case, we set $d = 2000, s = 10, m = 10, N = 400, \rho = 0.1778, \alpha = 10^{-6}$. Fig. 1(a) demonstrates the convergence of averaged $\bar{\mathbf{w}}^t$ in DGD, showing successful convergence of elements on support $\mathcal{S}$ and maintenance of small

magnitudes for elements on non-support $\mathcal{S}^c$. Fig. 1(b) and Fig. 1(c) further illustrate how DGD utilizes $\bar{\mathbf{u}}^t$ and $\bar{\mathbf{v}}^t$ to fit parameters on positive and negative support, respectively. Additionally, the magnitudes of $\bar{\mathbf{u}}^t$ and $\bar{\mathbf{v}}^t$ on non-positive and non-negative support remain small enough as the initialization. Consensus errors $\mathbf{u}^{t,i} - \bar{\mathbf{u}}^t$ and $\mathbf{v}^{t,i} - \bar{\mathbf{v}}^t$ are depicted in Fig. 1(d) and Fig. 1(e), respectively. The trends in these curves correspond to the magnitudes of the model parameter, affirming the validity of our analysis.

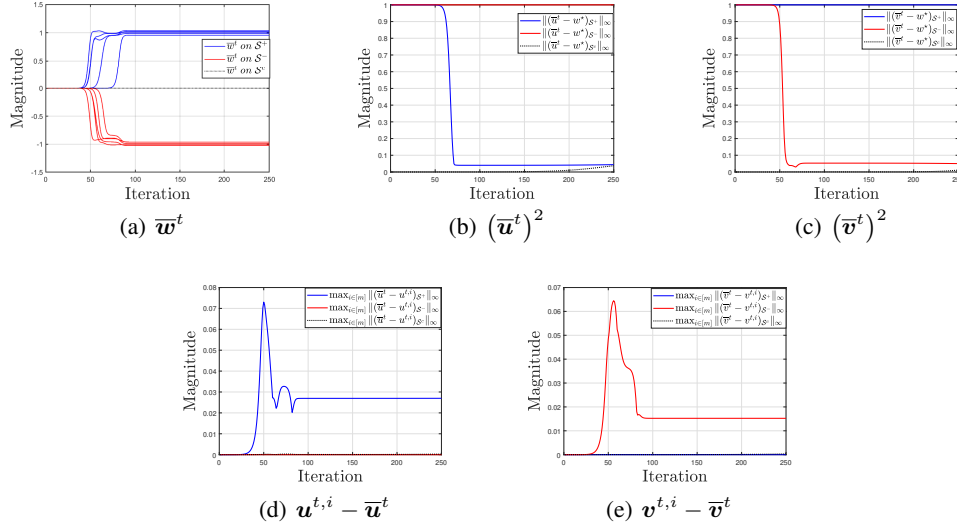


Figure 1: Dynamics of averaged variables and consensus errors.

- Impact of d and α on optimal estimation.** We vary the dimension of d ($4 \times 10^2, 4 \times 10^3, 4 \times 10^4$) to access effect of d on both statistical and computational properties. With $s = \lceil \log d \rceil$ and N chosen to satisfy $s \log d/N \approx 0.25$, we aim to maintain the same order optimal statistical error $\mathcal{O}\left(\sigma \sqrt{\frac{s \log d}{N}}\right)$. Testing is conducted on two networks with $m = 20$ but different ρ . For each d , we select the maximum initialization α that achieves optimal statistical error, resulting in $\alpha = 10^{-8}$ for $d = 4 \times 10^2$, $\alpha = 10^{-8.5}$ for $d = 4 \times 10^3$ and $\alpha = 10^{-9}$ for $d = 4 \times 10^4$. The results for $\rho = 0.1778$ and $\rho = 0.7519$ are displayed in Fig. 2(a) and Fig. 2(b), respectively. It is observable that DGD obtains estimators with statistical error matching that of the centralized setting, with computational complexity remaining largely unaffected by ambient dimension d across different network conditions. To assess the influence of α , we set $d = 2000$, $s = 10$, $m = 20$, $N = 400$, and different values for α on network with $m = 20$, $\rho = 0.1778$. The results in Fig. 2(c) illustrate that it is necessary to use small enough initialization to obtain optimal estimator.

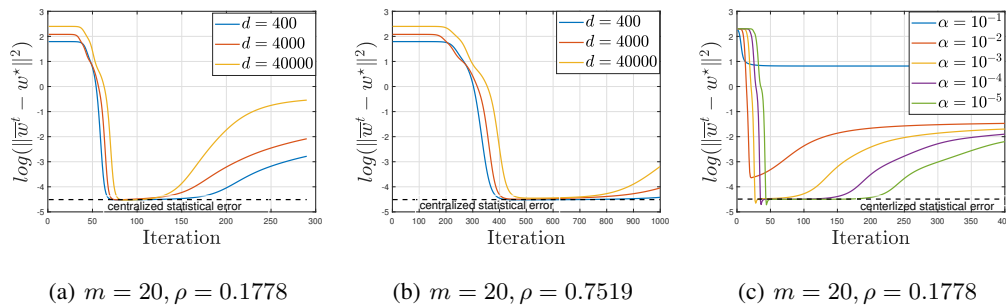


Figure 2: Impact of ambient dimension d and initialization α .

- Dependence on ρ and m .** We set $d = 2000$, $s = 10$, $N = 200$, $\alpha = 10^{-6}$ and test on networks with different numbers of agents. The results are shown in Fig. 3 where Fig. 3(a) and Fig.

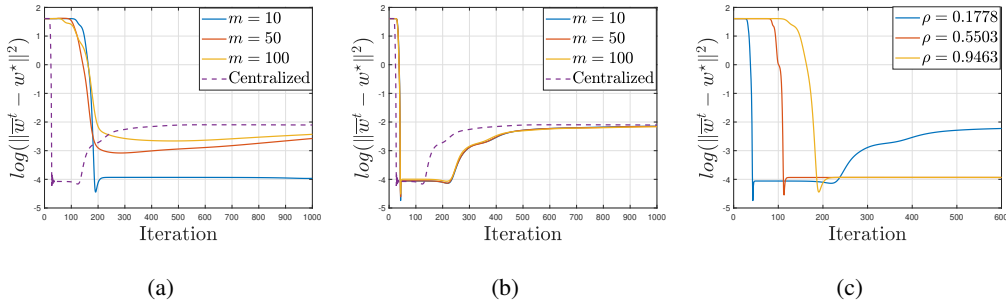


Figure 3: (a) $\rho = 0.9400$; (b) $\rho = 0.1778$; (c) $m = 10$.

3(b) display the performance with varied numbers of agents under fixed $\rho = 0.9400$ and fixed $\rho = 0.1778$, respectively. Fig. 3(a) indicates that DGD would not obtain the optimal estimator as the centralized setting when the number of agents is large which violates the condition in (6). When network connectivity is sufficiently connected as $\rho = 0.1778$, Fig. 3(b) conveys that this can allow a larger scale of agents to attain optimal statistical error. In Fig. 3(c), we fix $m = 10$ and observe the phenomenon under varied ρ by choosing proper stepsizes to achieve the best statistical error. Fig. 3(c) illustrates that ρ would influence the stopping time when DGD can obtain the optimal estimator. The worse the network, the more iterations it takes to find the optimal estimator.

6.2 Simulations on T-DGD

In this section, we evaluate the effectiveness of T-DGD. Initially, we vary the values of N , keeping the other parameters consistent with the simulations in Fig. 1. Fig. 4(a) illustrates that when each agent has inadequate samples ($N = 100, n = 10$), T-DGD would fail in achieving optimal estimation. However, with increasing local samples ($N = 400, n = 10$), T-DGD matches DGD in both statistical accuracy and convergence performance. Subsequently, we set the magnitudes of the ground truth on support as 100 and $N = 300$. The performance is depicted as dashed lines in Fig. 4(c), indicating failure of T-DGD under higher noise level ($\sigma = 0.5$). We further reduce the noise magnitude to $\sigma = 0.1$, and solid lines in Fig. 4(c) demonstrate the usefulness of T-DGD in sparse regression. These observations validate the statement in Proposition 1.

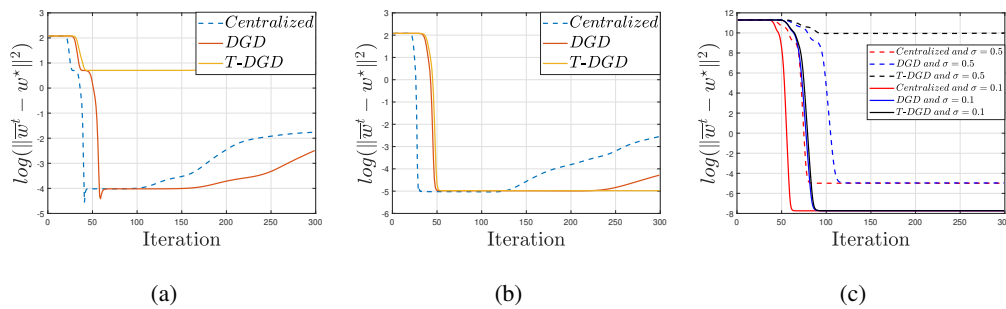


Figure 4: (a) $N = 100$; (b) $N = 400$; (c) Different noise intensities.

6.3 Comparison with explicit regularization

We have compared our proposed method with three existing decentralized methods, namely: CTA-DGD (LASSO) [18], ATC-DGD (LASSO) [17], and DGT (NetLASSO) [31]. These methods are all derived based on the LASSO formulation with explicit regularization. The numerical results presented in Fig. 5 compare all four methods under three different network connectivity settings. For each method, we tuned the step size to achieve the best performance. Our proposed method demonstrated the best recovery performance in all network settings with minimal iterations.

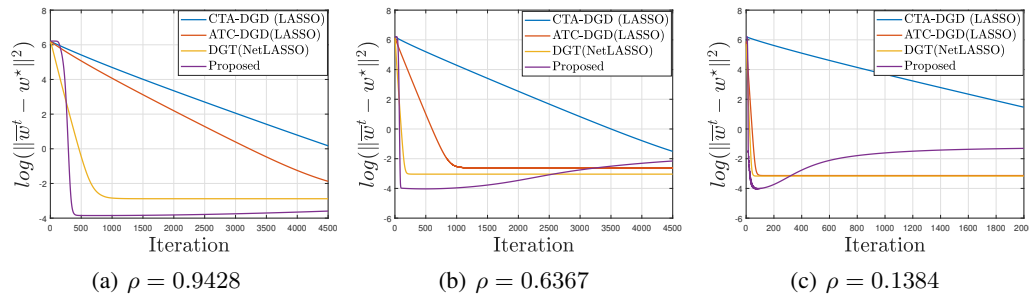


Figure 5: Comparison with decentralized sparse solvers under varying communication network. The setting is $d = 1000, k = 5, m = 50, N = 280, \sigma = 0.5$ and magnitude of sparse signal is 10.

We further compared T-DGD with existing methods with truncated versions of existing methods: Trun-CTA-DGD (LASSO), Trun-ATC-DGD (LASSO), and Trun-DGT (NetLASSO) which use the same Top- s truncation operator. As shown in Fig. 6, our proposed method is the only one to achieve successful recovery, while all other truncated decentralized methods failed. The numerical evidence demonstrates that naively combining sparsification with decentralized algorithms is *not granted to converge*. This is precisely one of the motivations of this work: to provide communication-efficient algorithms with both provably statistical and computational guarantees. This result also demonstrates the unique benefit of overparameterization and implicit regularization for decentralized learning setting, which has not been explored in the literature of learning theory.

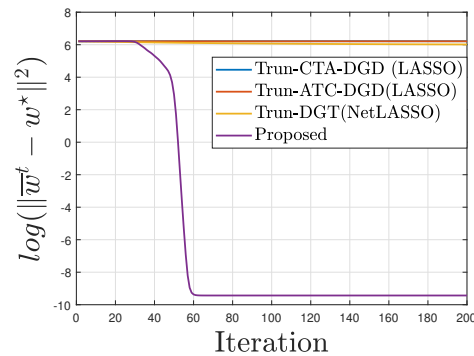


Figure 6: Truncated version: comparison with truncated decentralized sparse solvers. The setting is $d = 1000, s = 5, m = 50, N = 550, \sigma = 0.1, \rho = 0.2458$ and magnitude of sparse signal is 10.

7 Conclusion

In this paper, we study the implicit regularization of decentralized gradient descent for decentralized sparse regression in the unpenalized and overparameterized regimes. We establish both statistical and computational guarantees for the decentralized estimator under mild conditions of network connectivity, underscoring the utility of DGD in addressing overparameterized models. Furthermore, the proposed truncated DGD (T-DGD) offers a promising idea to reduce communication complexity while maintaining performance. In future work, exploring the possibility of relaxing the RIP condition in our assumption and leveraging the restricted eigenvalue condition to achieve optimal estimator in the decentralized setting is an interesting topic. Additionally, investigating alternative forms of implicit regularizations in decentralized optimization algorithms for more complicated overparameterized models is another intriguing direction.

References

- [1] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. *Advances in Neural Information Processing Systems*, 23, 2010.
- [2] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2022.
- [3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019.
- [4] Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Sgd with large step sizes learns sparse features. In *International Conference on Machine Learning*, pages 903–925. PMLR, 2023.
- [5] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.
- [6] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.
- [7] Raphaël Berthier. Incremental learning in diagonal linear networks. *Journal of Machine Learning Research*, 24(171):1–26, 2023.
- [8] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.
- [9] Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35:25237–25250, 2022.
- [10] Hung-Hsu Chou, Johannes Maly, and Holger Rauhut. More is less: inducing sparsity via overparameterization. *Information and Inference: A Journal of the IMA*, 12(3):1437–1460, 2023.
- [11] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.
- [12] Jianqing Fan, Zhuoran Yang, and Mengxin Yu. Understanding implicit regularization in over-parameterized single index model. *Journal of the American Statistical Association*, 118(544):2315–2328, 2023.
- [13] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [15] Peter D Hoff. Lasso, fractional norm and structured sparse estimation using a hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017.
- [16] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.
- [17] Yao Ji, Gesualdo Scutari, Ying Sun, and Harsha Honnappa. Distributed (atc) gradient descent for high dimension sparse regression. *IEEE Transactions on Information Theory*, 2023.

- [18] Yao Ji, Gesualdo Scutari, Ying Sun, and Harsha Honnappa. Distributed sparse regression via penalization. *Journal of Machine Learning Research*, 24(272):1–62, 2023.
- [19] Jiangyuan Li, Thanh Nguyen, Chinmay Hegde, and Ka Wai Wong. Implicit sparse regularization: The impact of depth and early stopping. *Advances in Neural Information Processing Systems*, 34:28298–28309, 2021.
- [20] Marie Maros and Gesualdo Scutari. Acceleration in distributed sparse regression. *Advances in Neural Information Processing Systems*, 35:36832–36844, 2022.
- [21] Marie Maros and Gesualdo Scutari. Dgd²: A linearly convergent distributed algorithm for high-dimensional statistical recovery. *Advances in Neural Information Processing Systems*, 35:3475–3487, 2022.
- [22] Marie Maros and Gesualdo Scutari. Decentralized matrix sensing: Statistical guarantees and fast convergence. *Advances in Neural Information Processing Systems*, 36, 2024.
- [23] Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pages 16270–16295. PMLR, 2022.
- [24] Eshaan Nichani, Yu Bai, and Jason D Lee. Identifying good directions to escape the ntk regime and efficiently learn low-degree plus sparse polynomials. *Advances in Neural Information Processing Systems*, 35:14568–14581, 2022.
- [25] Hristo Georgiev Papazov, Scott Pesme, and Nicolas Flammarion. Leveraging continuous time to understand momentum when training diagonal linear networks. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AIS-TATS) 2024*, 2024.
- [26] Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. *Advances in Neural Information Processing Systems*, 36, 2024.
- [27] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.
- [28] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.
- [29] Dominic Richards and Patrick Rebeschini. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 21(34):1–44, 2020.
- [30] Ohad Shamir. The implicit bias of benign overfitting. In *Conference on Learning Theory*, pages 448–478. PMLR, 2022.
- [31] Ying Sun, Marie Maros, Gesualdo Scutari, and Guang Cheng. High-dimensional inference over networks: Linear convergence and statistical guarantees. *arXiv preprint arXiv:2201.08507*, 2022.
- [32] 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.
- [33] Hossein Taheri and Christos Thrampoulidis. On generalization of decentralized learning with separable data. In *International Conference on Artificial Intelligence and Statistics*, pages 4917–4945. PMLR, 2023.
- [34] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.
- [35] Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6):86–93, 2023.

- [36] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.
- [37] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.
- [38] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.
- [39] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.
- [40] Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2020.
- [41] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2016.
- [42] Peng Zhao, Yun Yang, and Qiao-Chu He. Implicit regularization via hadamard product overparametrization in high-dimensional linear regression. *arXiv preprint arXiv:1903.09367*, 2(4):8, 2019.
- [43] Mo Zhou and Rong Ge. Implicit regularization leads to benign overfitting for sparse linear regression. In *International Conference on Machine Learning*, pages 42543–42573. PMLR, 2023.
- [44] Tongtian Zhu, Fengxiang He, Kaixuan Chen, Mingli Song, and Dacheng Tao. Decentralized sgd and average-direction sam are asymptotically equivalent. In *International Conference on Machine Learning*, pages 43005–43036. PMLR, 2023.

A Appendix.A

Section A.1, A.2 give the additional notations and useful basic lemmas, respectively. Section A.3 provides the properties for the simplified setting where all the elements of ground truth parameter $\mathbf{w}^*$ are non-negative. Section A.4 gives the key proposition for the general case where $\mathbf{w}^*$ contains both positive and negative elements on support $\mathcal{S}$. Final Section A.5 concludes the proofs for the Theorem 1, Corollary 1, and Proposition 1 in the main paper based on Proposition 3 in Section A.4.

A.1 Full Notations

In addition to introduced notations from Section 3.1 in the main paper, we need additional notations to present the proof. The consensus error is denoted as $\Delta_{\mathbf{u}}^{t,i} := \mathbf{u}^{t,i} - \bar{\mathbf{u}}^t$. $\Delta_{\mathbf{u},\mathcal{S}^+}^{t,i} = \Delta_{\mathbf{u}}^{t,i} \odot \mathbf{1}_{\mathcal{S}^+}$ and notations $\Delta_{\mathbf{u},\mathcal{S}^-}^{t,i}, \Delta_{\mathbf{u},\mathcal{S}^c}^{t,i}$ can be defined similarly. The additional notations are defined as

$$\begin{aligned} \mathbf{U}^t &:= [\mathbf{u}^{t,i}, \dots, \mathbf{u}^{t,m}], \quad \Delta_{\mathbf{u}}^t := [\Delta_{\mathbf{u}}^{t,1}, \dots, \Delta_{\mathbf{u}}^{t,m}], \quad \bar{\mathbf{U}}^t := \bar{\mathbf{u}}^t \mathbf{1}_m^T; \\ \nabla_{\mathbf{u}} \mathbf{f}(\bar{\mathbf{U}}^t, \bar{\mathbf{V}}^t) &:= [\nabla_{\mathbf{u}} f_1(\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t), \dots, \nabla_{\mathbf{u}} f_m(\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t)], \quad \nabla_{\mathbf{u}} \mathbf{F}(\bar{\mathbf{U}}^t, \bar{\mathbf{V}}^t) := \nabla_{\mathbf{u}} F(\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t) \mathbf{1}_m; \\ \nabla_{\mathbf{u}} \mathbf{f}(\mathbf{U}^t, \mathbf{V}^t) &:= [\nabla_{\mathbf{u}} f_1(\mathbf{u}^{t,1}, \mathbf{v}^{t,1}), \dots, \nabla_{\mathbf{u}} f_m(\mathbf{u}^{t,m}, \mathbf{v}^{t,m})]; \\ \nabla_{\mathbf{u}} \mathbf{f}(\bar{\mathbf{U}}^t, \mathbf{V}^t) &:= [\nabla_{\mathbf{u}} f_1(\bar{\mathbf{u}}^t, \mathbf{v}^{t,1}), \dots, \nabla_{\mathbf{u}} f_m(\bar{\mathbf{u}}^t, \mathbf{v}^{t,m})]. \end{aligned} \quad (9)$$

Above definitions can be also extended to variable $\mathbf{v}$ similarly.

A.2 Preliminary Lemmas

Lemma 1. (Theorem 6.5 in [37]) Consider the random feature vector $\mathbf{x} \in \mathbb{R}^d$ that all entries obey i.i.d. 1-sub-Gaussian distribution, if the sample size satisfies $n \gtrsim \delta^{-2} (s \log \frac{ed}{s} + \log \frac{2}{\epsilon})$, then the sample covariance matrix $\hat{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$ satisfies (δ, s) -RIP condition with probability $1 - \epsilon$.

Lemma 2. (Lemma A.3 in [36]) Suppose that $\frac{\mathbf{X}}{\sqrt{n}} \in \mathbb{R}^{n \times d}$ satisfies the $(\delta, s+1)$ -RIP, if $\mathbf{w} \in \mathbb{R}^d$ is a s -sparse vector, then $\left\| \left(\frac{\mathbf{X}^T \mathbf{X}}{n} - \mathbf{I} \right) \mathbf{w} \right\|_{\infty} \leq \sqrt{s} \delta \|\mathbf{w}\|_{\infty}$.

Lemma 3. (Lemma A.4 in [36]) Suppose that $\frac{1}{\sqrt{n}} \mathbf{X} \in \mathbb{R}^{n \times d}$ that satisfies $(\delta, 1)$ -RIP with $0 \leq \delta \leq 1$, then we have $\left\| \frac{\mathbf{X}^T \mathbf{X}}{n} \mathbf{w} \right\|_{\infty} \leq 2d \|\mathbf{w}\|_{\infty}, \forall \mathbf{w} \in \mathbb{R}^d$.

Lemma 4. (Lemma B.5 in [42]) Let $\xi \in \mathcal{R}^n$ is a vector of independent σ -sub-Gaussian random variables and all ℓ_2 norm of column vectors of $\mathbf{X} \in \mathbb{R}^{n \times d}$ are bounded, then with high probability $1 - \frac{1}{8d^3}$ such that $\left\| \frac{\mathbf{X}^T \xi}{n} \right\|_{\infty} \lesssim \sigma \sqrt{\frac{\log d}{n}}$.

A.3 Non-negative Case

We consider the simplified setting where all the elements on support are positive for ground truth $\mathbf{w}^*$. The following lemma shows the recursion of average variable $\bar{\mathbf{u}}^t$ on support $\mathcal{S}$ and non-support $\mathcal{S}^c$.

Lemma 5. Consider the sequence $\{\mathbf{u}^{t,i}\}$ generated according to (3) and (4) by DGD for solving loss function in (2), the average signal $\bar{\mathbf{u}}^t$ on support $\mathcal{S}$ and non-support $\mathcal{S}^c$ are updated according to the following formulas

$$\begin{aligned} \bar{\mathbf{u}}_{\mathcal{S}}^{t+1} &= \bar{\mathbf{u}}_{\mathcal{S}}^t \odot \left(\mathbf{1}_d - 4\eta (\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*) - 4\eta \frac{\mathbf{X}^T \mathbf{X}}{N} (\bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \bar{\mathbf{u}}_{\mathcal{S}^c}^t) + 4\eta \frac{\mathbf{X}^T \xi}{N} \right. \\ &\quad \left. - 4\eta \left(\frac{\mathbf{X} \mathbf{X}^T}{N} - \mathbf{I} \right) (\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*) - 4\eta \mathbf{p}^t \right) - 4\eta \mathbf{q}^t; \\ \bar{\mathbf{u}}_{\mathcal{S}^c}^{t+1} &= \bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \left(\mathbf{1}_d - 4\eta \left(\frac{\mathbf{X}^T \mathbf{X}}{N} (\bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \bar{\mathbf{u}}_{\mathcal{S}^c}^t) - \frac{\mathbf{X}^T \xi}{N} + \left(\frac{\mathbf{X} \mathbf{X}^T}{N} - \mathbf{I} \right) (\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*) \right) \right) \end{aligned} \quad (10)$$

$$-4\eta\mathbf{g}^t) - 4\eta\mathbf{f}^t, \quad (11)$$

where the perturbed error terms $\mathbf{p}^t, \mathbf{q}^t, \mathbf{g}^t, \mathbf{f}^t$ induced from decentralized network are defined as

$$\begin{aligned} \mathbf{p}^t = & \frac{1}{m} \sum_{i=1}^m \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left(2\bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i} + \Delta_S^{t,i} \odot \Delta_S^{t,i} \right) + 3\Delta_S^{t,i} \odot \Delta_S^{t,i} \\ & + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left(2\bar{\mathbf{u}}_{S^c}^t \odot \Delta_{S^c}^{t,i} + \Delta_{S^c}^{t,i} \odot \Delta_{S^c}^{t,i} \right); \end{aligned} \quad (12)$$

$$\begin{aligned} \mathbf{q}^t = & \frac{1}{m} \sum_{i=1}^m \Delta_S^{t,i} \odot \left(\left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left(\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{u}}_S^t - \mathbf{w}^* + 2\bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i} + \Delta_S^{t,i} \odot \Delta_S^{t,i} \right) \right. \\ & \left. + \Delta_S^{t,i} \odot \Delta_S^{t,i} - \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left(\bar{\mathbf{u}}_{S^c}^t + \Delta_{S^c}^{t,i} \right)^2 \right); \end{aligned} \quad (13)$$

$$\mathbf{g}^t = \frac{1}{m} \sum_{i=1}^m \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left(2\bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i} + \Delta_S^{t,i} \odot \Delta_S^{t,i} \right) + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left(2\bar{\mathbf{u}}_{S^c}^t \odot \Delta_{S^c}^{t,i} + \Delta_{S^c}^{t,i} \odot \Delta_{S^c}^{t,i} \right); \quad (14)$$

$$\begin{aligned} \mathbf{f}^t = & \frac{1}{m} \sum_{i=1}^m \Delta_{S^c}^{t,i} \odot \left(\left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left(\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{u}}_S^t - \mathbf{w}^* + 2\bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i} + \Delta_S^{t,i} \odot \Delta_S^{t,i} \right) \right. \\ & \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left(\bar{\mathbf{u}}_{S^c}^t + \Delta_{S^c}^{t,i} \right)^2 - \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right). \end{aligned} \quad (15)$$

Proof. Based on the updating of DGD, one-step iteration of the averaged parameter is

$$\bar{\mathbf{u}}^{t+1} = \bar{\mathbf{u}}^t - \eta \nabla F(\bar{\mathbf{u}}^t) + \frac{\eta}{m} \sum_{i=1}^m (\nabla f_i(\bar{\mathbf{u}}^t) - \nabla f_i(\mathbf{u}^{t,i})). \quad (16)$$

The gradient difference has the formula as

$$\begin{aligned} \nabla f_i(\bar{\mathbf{u}}^t) - \nabla f_i(\mathbf{u}^{t,i}) = & \bar{\mathbf{u}}^t \odot \left(\frac{4}{n} \mathbf{X}_i^T \mathbf{X}_i (\bar{\mathbf{u}}^t \odot \bar{\mathbf{u}}^t - \mathbf{w}^*) - \frac{4}{n} \mathbf{X}_i^T \boldsymbol{\xi}_i \right) \\ & - \mathbf{u}^{t,i} \odot \left(\frac{4}{n} \mathbf{X}_i^T \mathbf{X}_i (\mathbf{u}^{t,i} \odot \mathbf{u}^{t,i} - \mathbf{w}^*) - \frac{4}{n} \mathbf{X}_i^T \boldsymbol{\xi}_i \right) \\ = & (\bar{\mathbf{u}}^t - \mathbf{u}^{t,i}) \odot \left(\frac{4}{n} \mathbf{X}_i^T \mathbf{X}_i (\mathbf{u}^{t,i} \odot \mathbf{u}^{t,i} - \mathbf{w}^*) - \frac{4}{n} \mathbf{X}_i^T \boldsymbol{\xi}_i \right) \\ & + \bar{\mathbf{u}}^t \odot \left(\frac{4}{n} \mathbf{X}_i^T \mathbf{X}_i (\bar{\mathbf{u}}^t \odot \bar{\mathbf{u}}^t - \mathbf{u}^{t,i} \odot \mathbf{u}^{t,i}) \right) \\ = & -\Delta^{t,i} \odot \left(\frac{4}{n} \mathbf{X}_i^T \mathbf{X}_i (\bar{\mathbf{u}}^t \odot \bar{\mathbf{u}}^t - \mathbf{w}^* + 2\bar{\mathbf{u}}^t \odot \Delta^{t,i} + \Delta^{t,i} \odot \Delta^{t,i}) - \frac{4}{n} \mathbf{X}_i^T \boldsymbol{\xi}_i \right) \\ & - \bar{\mathbf{u}}^t \odot \left(\frac{4}{n} \mathbf{X}_i^T \mathbf{X}_i (2\bar{\mathbf{u}}^t \odot \Delta^{t,i} + \Delta^{t,i} \odot \Delta^{t,i}) \right), \end{aligned} \quad (17)$$

where the last inequality is due to the definition of $\Delta^{t,i}$. Substituting the above equality into (16) would have

$$\begin{aligned} \bar{\mathbf{u}}^{t+1} = & \bar{\mathbf{u}}^t - 4\eta\bar{\mathbf{u}}^t \odot \left(\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{u}}_S^t - \mathbf{w}^* + \frac{\mathbf{X}^T \mathbf{X}}{N} (\bar{\mathbf{u}}_{S^c}^t \odot \bar{\mathbf{u}}_{S^c}^t) - \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right. \\ & \left. + \left(\frac{\mathbf{X} \mathbf{X}^T}{N} - \mathbf{I} \right) (\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{u}}_S^t - \mathbf{w}^*) \right) \\ & - 4\eta\bar{\mathbf{u}}^t \odot \frac{1}{m} \sum_{i=1}^m \left(2 \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i}) + 2\bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i} \right. \end{aligned}$$

$$\begin{aligned}
& +2\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left(\bar{\mathbf{u}}_{S^c}^t \odot \Delta_{S^c}^{t,i} \right) + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left(\Delta_S^{t,i} \odot \Delta_S^{t,i} \right) + \Delta_S^{t,i} \odot \Delta_S^{t,i} \\
& + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left(\Delta_{S^c}^{t,i} \odot \Delta_{S^c}^{t,i} \right) \\
& - 4\eta \frac{1}{m} \sum_{i=1}^m \Delta^{t,i} \odot \left(\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{u}}_S^t - \mathbf{w}^* + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{u}}_S^t - \mathbf{w}^*) \right. \\
& \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\bar{\mathbf{u}}_{S^c}^t \odot \bar{\mathbf{u}}_{S^c}^t) - \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right) \\
& - 4\eta \frac{1}{m} \sum_{i=1}^m \Delta^{t,i} \odot \left(2 \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i}) + 2\bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i} \right. \\
& \left. + 2\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\bar{\mathbf{u}}_{S^c}^t \odot \Delta_{S^c}^{t,i}) + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\Delta_S^{t,i} \odot \Delta_S^{t,i}) + \Delta_S^{t,i} \odot \Delta_S^{t,i} \right. \\
& \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\Delta_{S^c}^{t,i} \odot \Delta_{S^c}^{t,i}) \right)
\end{aligned} \tag{18}$$

Because there are $\frac{1}{m} \sum_{i=1}^m \bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i} = \mathbf{0}$ and $\frac{1}{m} \sum_{i=1}^m \Delta_S^{t,i} \odot (\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{u}}_S^t - \mathbf{w}^*) = \mathbf{0}$, the above formula can be simplified as

$$\begin{aligned}
\bar{\mathbf{u}}^{t+1} = & \bar{\mathbf{u}}^t - 4\eta \bar{\mathbf{u}}^t \odot \left(\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{u}}_S^t - \mathbf{w}^* + \frac{\mathbf{X}^T \mathbf{X}}{N} (\bar{\mathbf{u}}_{S^c}^t \odot \bar{\mathbf{u}}_{S^c}^t) - \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right. \\
& \left. + \left(\frac{\mathbf{X} \mathbf{X}^T}{N} - \mathbf{I} \right) (\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{u}}_S^t - \mathbf{w}^*) \right) \\
& - 4\eta \bar{\mathbf{u}}^t \odot \frac{1}{m} \sum_{i=1}^m \left(2 \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i}) \right. \\
& \left. + 2\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\bar{\mathbf{u}}_{S^c}^t \odot \Delta_{S^c}^{t,i}) + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\Delta_S^{t,i} \odot \Delta_S^{t,i}) + \Delta_S^{t,i} \odot \Delta_S^{t,i} \right. \\
& \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\Delta_{S^c}^{t,i} \odot \Delta_{S^c}^{t,i}) \right) \\
& - 4\eta \frac{1}{m} \sum_{i=1}^m \Delta^{t,i} \odot \left(\left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{u}}_S^t - \mathbf{w}^*) \right. \\
& \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\bar{\mathbf{u}}_{S^c}^t \odot \bar{\mathbf{u}}_{S^c}^t) - \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right) \\
& - 4\eta \frac{1}{m} \sum_{i=1}^m \Delta^{t,i} \odot \left(2 \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i}) + 2\bar{\mathbf{u}}_S^t \odot \Delta_S^{t,i} \right. \\
& \left. + 2\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\bar{\mathbf{u}}_{S^c}^t \odot \Delta_{S^c}^{t,i}) + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\Delta_S^{t,i} \odot \Delta_S^{t,i}) + \Delta_S^{t,i} \odot \Delta_S^{t,i} \right. \\
& \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\Delta_{S^c}^{t,i} \odot \Delta_{S^c}^{t,i}) \right).
\end{aligned} \tag{19}$$

Thus, this would obtain the recursion (10) for support averaged signal.

The recursion of optimization error on non-support $\mathcal{S}^c$ becomes

$$\begin{aligned}
\bar{\mathbf{u}}_{\mathcal{S}^c}^{t+1} = & \bar{\mathbf{u}}_{\mathcal{S}^c}^t - 4\eta \bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \left(\frac{\mathbf{X}^T \mathbf{X}}{N} (\bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \bar{\mathbf{u}}_{\mathcal{S}^c}^t) - \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} + \left(\frac{\mathbf{X} \mathbf{X}^T}{N} - \mathbf{I} \right) (\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*) \right) \\
& - 4\eta \bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \frac{1}{m} \sum_{i=1}^m \left(2 \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \boldsymbol{\Delta}_{\mathcal{S}}^{t,i}) \right. \\
& + 2 \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \boldsymbol{\Delta}_{\mathcal{S}^c}^{t,i}) + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\boldsymbol{\Delta}_{\mathcal{S}}^{t,i} \odot \boldsymbol{\Delta}_{\mathcal{S}}^{t,i}) \\
& \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\boldsymbol{\Delta}_{\mathcal{S}^c}^{t,i} \odot \boldsymbol{\Delta}_{\mathcal{S}^c}^{t,i}) \right) \\
& - 4\eta \frac{1}{m} \sum_{i=1}^m \boldsymbol{\Delta}_{\mathcal{S}^c}^{t,i} \odot \left(\left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*) \right. \\
& \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \bar{\mathbf{u}}_{\mathcal{S}^c}^t) - \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right) \\
& - 4\eta \frac{1}{m} \sum_{i=1}^m \boldsymbol{\Delta}_{\mathcal{S}^c}^{t,i} \odot \left(2 \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \boldsymbol{\Delta}_{\mathcal{S}}^{t,i}) \right. \\
& + 2 \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \boldsymbol{\Delta}_{\mathcal{S}^c}^{t,i}) + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) (\boldsymbol{\Delta}_{\mathcal{S}}^{t,i} \odot \boldsymbol{\Delta}_{\mathcal{S}}^{t,i}) \\
& \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} (\boldsymbol{\Delta}_{\mathcal{S}^c}^{t,i} \odot \boldsymbol{\Delta}_{\mathcal{S}^c}^{t,i}) \right). \tag{20}
\end{aligned}$$

Rearranging the above equality would obtain the (11). $\square$

The following lemma shows the recursion of consensus error on support $\mathcal{S}$ and non-support $\mathcal{S}^c$.

Lemma 6. Consider the sequence $\{\mathbf{u}^{t,i}\}$ generated according to (3) and (4) by DGD for solving loss function in (2), the consensus error $\boldsymbol{\Delta}^t$ on support $\mathcal{S}$ and non-support $\mathcal{S}^c$ have following recursion

$$\begin{aligned}
\|\boldsymbol{\Delta}_{\mathcal{S}}^{t+1}\|_{\infty} \leq & \rho \|\boldsymbol{\Delta}_{\mathcal{S}}^t\|_{\infty} \left(1 + 4\eta \left((\sqrt{s}\delta_{\max} + 1) \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} + 2d (\bar{\mathbf{u}}_{\mathcal{S}^c}^t + \boldsymbol{\Delta}_{\mathcal{S}^c}^t)^2 \right. \right. \\
& \left. \left. + 2 (\sqrt{s}\delta_{\max} + 1) (\|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} + \|\boldsymbol{\Delta}_{\mathcal{S}}^t\|_{\infty})^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \right) \right) \\
& + 4\rho\eta \|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} \cdot \left(\sqrt{s} (\delta_{\max} + \delta) \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \right. \\
& \left. + \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_{\infty} + 2d (\bar{\mathbf{u}}_{\mathcal{S}^c}^t + \boldsymbol{\Delta}_{\mathcal{S}^c}^t)^2 \right); \tag{21}
\end{aligned}$$

$$\begin{aligned}
\|\boldsymbol{\Delta}_{\mathcal{S}^c}^{t+1}\|_{\infty} \leq & \rho \|\boldsymbol{\Delta}_{\mathcal{S}^c}^t\|_{\infty} \left(1 + 4\eta \left(\sqrt{s}\delta_{\max} \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} + \sqrt{s}\delta_{\max} (\|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} + \|\boldsymbol{\Delta}_{\mathcal{S}}^t\|_{\infty})^2 \right. \right. \\
& \left. \left. + 6d (\|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty} + \|\boldsymbol{\Delta}_{\mathcal{S}^c}^t\|_{\infty})^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \right) \right) \\
& + 4\rho\eta \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty} \cdot \left(\sqrt{s} (\delta_{\max} + \delta) \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} + \sqrt{s}\delta_{\max} \|\boldsymbol{\Delta}_{\mathcal{S}}^t\|_{\infty} \right. \\
& \left. \cdot (\|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} + \|\boldsymbol{\Delta}_{\mathcal{S}}^t\|_{\infty}) + 4d \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty}^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} + \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_{\infty} \right). \tag{22}
\end{aligned}$$

Proof. Based on the iteration of DGD, there is

$$\begin{aligned}
\mathbf{U}^{t+1} \left(\mathbf{I}_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T \right) &= (\mathbf{U}^t - \eta \nabla \mathbf{f}(\mathbf{U}^t)) \mathbf{W} \left(\mathbf{I}_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T \right) \\
&= (\mathbf{U}^t - \eta \nabla \mathbf{f}(\mathbf{U}^t)) \left(\mathbf{W} - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T \right) \\
&= \left(\mathbf{U}^t - \bar{\mathbf{U}}^t - \eta \nabla \mathbf{f}(\mathbf{U}^t) + \eta \nabla \mathbf{F}(\bar{\mathbf{U}}^t) \right) \left(\mathbf{W} - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T \right). \quad (23)
\end{aligned}$$

Thus, the consensus error on support $\mathcal{S}$ has recursion as follows

$$\begin{aligned}
\|\Delta_{\mathcal{S}}^{t+1}\|_{\infty} &\stackrel{(i)}{\leq} \rho \|\Delta_{\mathcal{S}}^t\|_{\infty} + \rho \eta \max_i \|\mathbf{1}_{\mathcal{S}} \odot (\nabla f_i(\mathbf{u}^{t,i}) - \nabla F(\bar{\mathbf{u}}^t))\|_{\infty} \\
&\leq \rho \|\Delta_{\mathcal{S}}^t\|_{\infty} + \rho \eta \max_i \|\mathbf{1}_{\mathcal{S}} \odot (\nabla f_i(\mathbf{u}^{t,i}) - \nabla f_i(\bar{\mathbf{u}}^t))\|_{\infty} \\
&\quad + \rho \eta \max_i \|\mathbf{1}_{\mathcal{S}} \odot (\nabla f_i(\bar{\mathbf{u}}^t) - \nabla F(\bar{\mathbf{u}}^t))\|_{\infty} \\
&\stackrel{(ii)}{\leq} \rho \|\Delta_{\mathcal{S}}^t\|_{\infty} + 4\rho\eta \|\Delta_{\mathcal{S}}^t\|_{\infty} \cdot ((\sqrt{s}\delta_{\max} + 1) \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} + \\
&\quad + 2d \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty}^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} + 2(\sqrt{s}\delta_{\max} + 1) \|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} \|\Delta_{\mathcal{S}}^t\|_{\infty} \\
&\quad + 4d \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty} \|\Delta_{\mathcal{S}^c}^t\|_{\infty} + (\sqrt{s}\delta_{\max} + 1) \|\Delta_{\mathcal{S}}^t\|_{\infty}^2 + 2d \|\Delta_{\mathcal{S}^c}^t\|_{\infty}^2) \\
&\quad + 4\rho\eta \|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} \cdot \left(2(\sqrt{s}\delta_{\max} + 1) \|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} \|\Delta_{\mathcal{S}}^t\|_{\infty} + 4d \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty} \|\Delta_{\mathcal{S}^c}^t\|_{\infty} \right. \\
&\quad + (\sqrt{s}\delta_{\max} + 1) \|\Delta_{\mathcal{S}}^t\|_{\infty}^2 + 2d \|\Delta_{\mathcal{S}}^t\|_{\infty}^2 + (\sqrt{s}(\delta_{\max} + \delta) + 2) \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} \\
&\quad \left. + 2d \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty}^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} + \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_{\infty} \right) \\
&\stackrel{(iii)}{\leq} \rho \|\Delta_{\mathcal{S}}^t\|_{\infty} \left(1 + 4\eta \left((\sqrt{s}\delta_{\max} + 1) \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} + 2d (\bar{\mathbf{u}}_{\mathcal{S}^c}^t + \Delta_{\mathcal{S}^c}^t)^2 \right. \right. \\
&\quad \left. \left. + 2(\sqrt{s}\delta_{\max} + 1) (\|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} + \|\Delta_{\mathcal{S}}^t\|_{\infty})^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \right) \right) \\
&\quad + 4\rho\eta \|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} \cdot \left(\sqrt{s}(\delta_{\max} + \delta) \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \right. \\
&\quad \left. + \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_{\infty} + 2d (\bar{\mathbf{u}}_{\mathcal{S}^c}^t + \Delta_{\mathcal{S}^c}^t)^2 \right), \quad (24)
\end{aligned}$$

where (i) is due to the defined spectral gap of network in Assumption 1 and (ii) uses the gradient difference formula in (17) and

$$\begin{aligned}
\nabla f_i(\bar{\mathbf{u}}^t) - \nabla F(\bar{\mathbf{u}}^t) &= \bar{\mathbf{u}}^t \odot \left(\frac{4\mathbf{X}_i^T \mathbf{X}_i}{n} (\bar{\mathbf{u}}^t \odot \bar{\mathbf{u}}^t - \mathbf{w}^*) - \frac{4\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right) \\
&\quad - \bar{\mathbf{u}}^t \odot \left(\frac{4\mathbf{X}^T \mathbf{X}}{N} (\bar{\mathbf{u}}^t \odot \bar{\mathbf{u}}^t - \mathbf{w}^*) - \frac{4\mathbf{X}^T \boldsymbol{\xi}}{N} \right) \quad (25) \\
&= \bar{\mathbf{u}}^t \odot \left(4 \left(\left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) - \left(\frac{\mathbf{X}^T \mathbf{X}}{N} - \mathbf{I} \right) \right) (\bar{\mathbf{u}}^t \odot \bar{\mathbf{u}}^t - \mathbf{w}^*) \right. \\
&\quad \left. + \left(\frac{4\mathbf{X}^T \boldsymbol{\xi}}{N} - \frac{4\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right) \right)
\end{aligned}$$

(26)

and local and global RIP conditions. The (iii) is summing up terms involved $\|\Delta_{\mathcal{S}}^{t+1}\|_{\infty}$ and $\|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty}$ separately.

The recursion of consensus error on non-support $\mathcal{S}^c$ part is as follows

$$\begin{aligned}
\|\Delta_{\mathcal{S}^c}^{t+1}\|_{\infty} &\leq \rho \|\Delta_{\mathcal{S}^c}^t\|_{\infty} + \rho\eta \max_i \|\mathbf{1}_{\mathcal{S}^c} \odot (\nabla f_i(\mathbf{u}_i^t) - \nabla F(\bar{\mathbf{u}}^t))\|_{\infty} \\
&\leq \rho \|\Delta_{\mathcal{S}^c}^t\|_{\infty} + \rho\eta \max_i \|\mathbf{1}_{\mathcal{S}^c} \odot (\nabla f_i(\mathbf{u}_i^t) - \nabla f_i(\bar{\mathbf{u}}^t))\|_{\infty} \\
&\quad + \rho\eta \max_i \|\mathbf{1}_{\mathcal{S}^c} \odot (\nabla f_i(\bar{\mathbf{u}}^t) - \nabla F(\bar{\mathbf{u}}^t))\|_{\infty} \\
&\leq \rho \|\Delta_{\mathcal{S}^c}^t\|_{\infty} + 4\rho\eta \|\Delta_{\mathcal{S}^c}^t\|_{\infty} \cdot (\sqrt{s}\delta_{\max} \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} + \\
&\quad + 2d \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty}^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} + 2\sqrt{s}\delta_{\max} \|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} \|\Delta_{\mathcal{S}}^t\|_{\infty} \\
&\quad + 4d \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty} \|\Delta_{\mathcal{S}^c}^t\|_{\infty} + \sqrt{s}\delta_{\max} \|\Delta_{\mathcal{S}}^t\|_{\infty}^2 + 2d \|\Delta_{\mathcal{S}^c}^t\|_{\infty}^2) \\
&\quad + 4\rho\eta \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty} \cdot \left(2\sqrt{s}\delta_{\max} \|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} \|\Delta_{\mathcal{S}}^t\|_{\infty} + 4d \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty} \|\Delta_{\mathcal{S}^c}^t\|_{\infty} \right. \\
&\quad + \sqrt{s}\delta_{\max} \|\Delta_{\mathcal{S}}^t\|_{\infty}^2 + 2d \|\Delta_{\mathcal{S}^c}^t\|_{\infty}^2 + \sqrt{s}(\delta_{\max} + \delta) \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} \\
&\quad + 4d \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty}^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} + \left. \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_{\infty} \right) \\
&\leq \rho \|\Delta_{\mathcal{S}^c}^t\|_{\infty} \left(1 + 4\eta \left(\sqrt{s}\delta_{\max} \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} + \sqrt{s}\delta_{\max} (\|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} + \|\Delta_{\mathcal{S}}^t\|_{\infty})^2 \right. \right. \\
&\quad + 6d (\|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty} + \|\Delta_{\mathcal{S}^c}^t\|_{\infty})^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \left. \right) \\
&\quad + 4\rho\eta \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty} \cdot \left(\sqrt{s}(\delta_{\max} + \delta) \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} + \sqrt{s}\delta_{\max} \|\Delta_{\mathcal{S}}^t\|_{\infty} \right. \\
&\quad + (\|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} + \|\Delta_{\mathcal{S}}^t\|_{\infty}) + 4d \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty}^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} + \left. \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_{\infty} \right). \quad (27)
\end{aligned}$$

□

The following proposition shows the dynamics of average variable $\bar{\mathbf{u}}^t$ and consensus error $\Delta_{\mathcal{S}}^{t,i}$ in the form of an inductive hypothesis. Before showing the proposition, we define the following quantities. We define $T := \frac{1}{\eta w_{\max}^*} \log \frac{1}{\alpha^4}$ and for any integer $k \geq -1$, $T_k := 2^k T$ and $\bar{T}_k := \sum_{i=0}^k T_i$ with $\bar{T}_{-1} = 0$ where T_k denotes the number of iterations between $(k-1)$ -th and k -th induction step. Defining $K := \left\lceil \log_2 \frac{w_{\max}^*}{\zeta} \right\rceil$ as the number of induction steps, $B_k := \frac{w_{\max}^*}{40 \times 2^k}$ denotes the upper bound of perturbed error in $(k-1)$ -th induction step and constant scale parameter $\beta := \frac{32}{1-\sqrt{\rho}}$.

Proposition 2. *With the same setting as Theorem 1, the following claims hold $k = 0, 1, \dots, K-1$ steps.*

- (a) For $\bar{T}_{k-1} \leq t < \bar{T}_k$ that $\forall k \in [K]$, there is $\|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_{\infty} \leq \frac{w_{\max}^*}{2^k}$.
- (b) For $\forall k \in [K]$, there is $\|\bar{\mathbf{u}}_{\mathcal{S}}^{\bar{T}_{k-1}} \odot \bar{\mathbf{u}}_{\mathcal{S}}^{\bar{T}_{k-1}} - \mathbf{w}^*\|_{\infty} \leq \frac{w_{\max}^*}{2^k}$.
- (c) For $\bar{T}_{k-1} \leq t < \bar{T}_k$ that $\forall k \in [K]$, it has $\|\Delta_{\mathcal{S}}^t\|_{\infty} \leq 4\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_{\infty} B_k$. In addition, the refined element-wise bound is $|\Delta_j^t| \leq 4\beta\rho^{\frac{3}{4}}\eta |\bar{u}_j^t| B_k, \forall j \in \mathcal{S}$.

- (d) For $\bar{T}_{k-1} \leq t < \bar{T}_k$ that $\forall k \in [K]$, it has $\|\Delta_{\mathcal{S}^c}^t\|_\infty \leq 4\beta\rho^{\frac{3}{4}}\eta\|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_\infty B_k$. In addition, the refined element-wise bound is $|\Delta_j^t| \leq 4\beta\rho^{\frac{3}{4}}\eta|\bar{u}_j^t|B_k, \forall j \in \mathcal{S}^c$.
- (e) For $\forall k \in [K]$ and $\forall j \in \mathcal{S}$, $\alpha^3 \leq \bar{u}_j^{T_{k-1}} \leq w_j^* + 4B_k$.

Proof. Proof idea: Inductions (a), (b), (e) indicate that if the connectivity of the network is sufficiently well (ρ is small enough), the trajectory of the averaged signal $\bar{\mathbf{u}}^t$ would mimic that of the centralized case [36]. Different from the centralized setting, these three claims are based on inductions (c) and (d), which guarantee that the consensus error along both support $\mathcal{S}$ and non-support $\mathcal{S}^c$ can be controlled based on the magnitude of the respective signals. We utilize this property to reparameterize consensus error by Hadamard product based on the averaged signal. Thus, the perturbed error terms induced by the decentralized network in the recursion of the averaged signal can be quantitatively through the reparameterized consensus error. Then conditions on network connectivity ρ and step size η can guarantee that the averaged signal in decentralized would have properties in inductions (a), (b), (e) based on inductions (c), (d).

Base case: As the initialization $\mathbf{u}^{0,i} = \alpha \mathbf{1}_d, \forall i \in [m]$. Due to the condition on α , the base case is true.

Induction Step: If the above (a)-(e) induction hypotheses hold all until some $0 \leq k \leq K-1$, we should prove they still hold at $k+1$ -th induction step.

(a) The magnitude of $\mathbf{p}^t$ in (12) under this induction step can be bounded based on inductions (c), (d). $\forall \bar{T}_{k-1} \leq t < t+1 < \bar{T}_k$, if $\|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_\infty$ keep same order as initialization, then there is

$$\begin{aligned} \|\mathbf{p}_{\mathcal{S}}^t\|_\infty &\leq 8\sqrt{s}\delta_{\max}\beta\rho^{\frac{3}{4}}\eta B_k \|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_\infty^2 + (\sqrt{s}\delta_{\max} + 3) \left(4\beta\rho^{\frac{3}{4}}\eta B_k \|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_\infty\right)^2 \\ &\quad + 2d \left(8\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_\infty^2 B_k + \left(4\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_\infty B_k\right)^2\right) \\ &\leq \frac{3\sqrt{\rho}B_k}{16}, \end{aligned} \quad (28)$$

where the first inequality is based on Lemma 2 and Lemma 3, the last inequality is due to step size, value of β , $\|\bar{\mathbf{u}}_{\mathcal{S}}^t\|_\infty^2 \leq 2w_{\max}^*$, network connectivity condition and global RIP condition that $\rho^{\frac{1}{4}}\sqrt{s}\delta_{\max} \leq \sqrt{s}\delta \leq 1$.

For the perturbation $\mathbf{q}^t$ in (13), which is an error term outside the multiplicative updates in (10), based on induction (a), fine-grained upper in (c), there is $\forall j \in \mathcal{S}$

$$\begin{aligned} |q_j^t| &\leq 4\beta\rho^{\frac{3}{4}}\eta B_k |\bar{u}_j| \left(\sqrt{s}\delta_{\max} \|\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*\|_\infty + \frac{3\sqrt{\rho}B_k}{16} + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) \\ &\leq \frac{\sqrt{\rho}B_k |\bar{u}_j|}{32}, \end{aligned} \quad (29)$$

where the first inequality is due to (28) by comparing formula of $\mathbf{q}^t$ with $\mathbf{p}^t$ and last inequality is due to step size condition and such that $\rho^{\frac{1}{4}} \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \leq \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_\infty < w_{\max}^*$. Then $\mathbf{q}^t$ could be reparameterized as $\mathbf{q}^t = \mathbf{r}_q^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t$ where $\|\mathbf{r}_q^t\|_\infty \leq \frac{\sqrt{\rho}B_k}{16}$ for $\forall t$ that $\bar{T}_{k-1} \leq t < \bar{T}_k$, the perturbed optimization recursion on support over decentralized network in (10) becomes

$$(\bar{\mathbf{u}}_{\mathcal{S}}^{t+1})^2 = (\bar{\mathbf{u}}_{\mathcal{S}}^t)^2 \odot (\mathbf{1}_d - 4\eta (\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^* + \mathbf{E}_2^t + \mathbf{E}_3^t + \mathbf{p}^t + \mathbf{r}_q^t))^2. \quad (30)$$

where the perturbation errors $\mathbf{E}_2^t$ and $\mathbf{E}_3^t$ are defined as

$$\mathbf{E}_2^t = \left(\frac{\mathbf{X}^T \mathbf{X}}{N} - \mathbf{I} \right) (\bar{\mathbf{u}}_{\mathcal{S}}^t \odot \bar{\mathbf{u}}_{\mathcal{S}}^t - \mathbf{w}^*)$$

$$\mathbf{E}_3^t = \frac{\mathbf{X}^T \mathbf{X}}{N} (\bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \bar{\mathbf{u}}_{\mathcal{S}^c}^t) - \frac{\mathbf{X}^T \boldsymbol{\xi}}{N}. \quad (31)$$

Because there is $\|\mathbf{E}_2^t\|_\infty + \|\mathbf{E}_3^t\|_\infty + \|\mathbf{p}^t\|_\infty + \|\mathbf{r}_q^t\|_\infty \leq B_k$, which is based on the upper bound in (70). Then the proof is divided into the following two cases based on the magnitude of the element in $\mathbf{w}_{\mathcal{S}^+}^*$.

(1) $\forall j$ that $w_j^* \geq 20B_k$, based on induction hypothesis (e) that $\left(\bar{u}_j^{\bar{T}^{k-1}}\right)^2 \leq w_j^* + 4B_k$, there is $\left(\bar{u}_j^{\bar{T}^{k-1}}\right)^2 \leq \frac{6}{5}w_j^*$, which illustrates that it satisfies the conditions in Lemma B.10 in [36]. Then because induction hypothesis (a) and (b) are true until t -th iteration, then if $B_k < \left\| \left(\bar{u}_j^t\right)^2 - w_j^* \right\|_\infty \leq \frac{w_{\max}^*}{2^k}$, then $\left\| \left(\bar{u}_j^{t+1}\right)^2 - w_j^* \right\|_\infty \leq \left\| \left(\bar{u}_j^t\right)^2 - w_j^* \right\|_\infty \leq \frac{w_{\max}^*}{2^k}$, else if $\left\| \left(\bar{u}_j^t\right)^2 - w_j^* \right\|_\infty \leq B_k$, then $\left\| \left(\bar{u}_j^{t+1}\right)^2 - w_j^* \right\|_\infty \leq B_k$. Combined with two cases, we can conclude that (a) also holds for $(t+1)$ -th iteration for j that $w_j^* \geq 20B_k$.

(2) For arbitrary j -th elements whose magnitude is not sufficiently larger than the perturbation that $w_j^* \leq 20B_k$, based on the upper bound in induction (e), perturbation bound and monotonic property in Lemma B.6 in [36], we can guarantee that $\left(\bar{u}_j^t\right)^2$ would keep staying in $(0, w_j^* + 4B_k]$. With condition $w_j^* \leq 20B_k$, we can conclude that $\left\| \left(\bar{u}_j^{t+1}\right)^2 - w_j^* \right\|_\infty \leq \max\{w_j^*, 4B_k\} \leq 20B_k \leq \frac{w_{\max}^*}{2^k}$.

Combined these two cases would finish proof of (a).

(b) To prove this statement, we should guarantee that there are sufficient iterative steps in the $(k-1)$ -th induction that can make $\left\| \left(\bar{\mathbf{u}}_{\mathcal{S}^+}^{\bar{T}^{k-1}}\right)^2 - \mathbf{w}^* \right\|_\infty$ decrease at least by half from the beginning iteration of current induction stage to that of next induction stage. The proof is also divided into two cases.

(1) The one case is that $\forall j$ that it already has $\left| \left(\bar{u}_j^{\bar{T}^{k-1}}\right)^2 - w_j^* \right| < \frac{w_{\max}^*}{2^{k+1}}$, then with similar proof in (a), we can guarantee that $\forall t \geq \bar{T}_{k-1}$, $\left| \left(\bar{u}_j^t\right)^2 - w_j^* \right| < \frac{w_{\max}^*}{2^{k+1}}$. Thus, for these supports, we prove the $(k+1)$ -th induction also holds.

(2) The second case is $\forall j$ that there is $\left| \left(\bar{u}_j^{\bar{T}^{k-1}}\right)^2 - w_j^* \right| \geq \frac{w_{\max}^*}{2^{k+1}} = 20B_k$. Based on the upper bound in induction (e) that $\left(\bar{u}_j^{\bar{T}^{k-1}}\right)^2 \leq w_j^* + 4B_k$, then $\bar{u}_j^{\bar{T}^{k-1}}, w_j^*$ must satisfy $0 \leq \left(\bar{u}_j^{\bar{T}^{k-1}}\right)^2 \leq w_j^* - 20B_k, w_j^* \geq 20B_k$, respectively. This means $\bar{u}_j^{\bar{T}^{k-1}}$ is far away from w_j^* at least $20B_k$ distance. According to Lemma B.12 in [36], to achieve $\left| \left(\bar{u}_j^{\bar{T}^{k-1}}\right)^2 - w_j^* \right| \leq 20B_k$, the sufficient condition for the number of iterations t in current induction stage is $t \geq \frac{15}{32\eta w_j^*} \log \frac{(w_j^*)^2}{19 \left(\bar{u}_j^{\bar{T}^{k-1}}\right)^2 B_k}$. Now we verify the setting of T_k as follows

$$\begin{aligned} T_k &= \frac{2^k}{\eta w_{\max}^*} \log \frac{1}{\alpha^4} \\ &\geq \frac{1}{40\eta B_k} \log \left(\frac{3(w_{\max}^*)^2}{\zeta} \cdot \frac{1}{\alpha^3} \right) \\ &\geq \frac{1}{2\eta w_j^*} \log \left(\frac{3(w_j^*)^2}{\zeta} \cdot \frac{1}{\left(\bar{u}_j^{\bar{T}^{k-1}}\right)^2} \right) \end{aligned}$$

$$\geq \frac{1}{2\eta w_j^*} \log \frac{(w_j^*)^2}{16 \left(\bar{u}_j^{\bar{T}_{k-1}}\right)^2 B_k}, \quad (32)$$

where the first inequality is due to the definition of B_k and small initialization condition that $\alpha \leq \frac{\zeta}{3(w_{\max}^*)^2}$ and the second inequality is due to $w_j^* \geq 20B_k$ and lower bound in induction (e) that $\left(\bar{u}_j^{\bar{T}_{k-1}}\right)^2 \geq \alpha^3$. The last inequality is because $\frac{\zeta}{3} \leq \frac{16}{40} \frac{w_{\max}^*}{2^{K-1}} \leq \frac{16}{40} \frac{w_{\max}^*}{2^k} = 16B_k$. Thus, combining the above two cases and similar proof, we can conclude that $\forall t \geq \bar{T}_k$, there is $\left\|(\bar{\mathbf{u}}_S^t)^2 - \mathbf{w}^*\right\|_\infty \leq \frac{w_{\max}^*}{2^{k+1}}$. This completes the proof of induction (b).

(c) To make the consensus error satisfy the above induction, $\forall t$ that $\bar{T}_{k-1} \leq t < t+1 < \bar{T}_k$, based on one step iteration in (24) of Lemma 6, induction (a), (c) and step size condition $\eta \leq \frac{1-\sqrt{\rho}}{64\rho^{\frac{1}{2}} w_{\max}^*}$, it has

$$\begin{aligned} & \sqrt{\rho} \left(1 + 4\eta \left((\sqrt{s}\delta_{\max} + 1) \left\| \bar{\mathbf{u}}_S^t \odot \bar{\mathbf{u}}_S^t - \mathbf{w}^* \right\|_\infty + 2d \left(\bar{\mathbf{u}}_{S^c}^t + \boldsymbol{\Delta}_{S^c}^t \right)^2 + 2 \left(\sqrt{s}\delta_{\max} + 1 \right) \right. \right. \\ & \quad \left. \left. \cdot \left(\left\| \bar{\mathbf{u}}_S^t \right\|_\infty + \left\| \boldsymbol{\Delta}_S^t \right\|_\infty \right)^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) \right) \leq 1 + \frac{1-\sqrt{\rho}}{2}. \end{aligned} \quad (33)$$

Then the recursion (24) in Lemma 6 becomes

$$\begin{aligned} \left\| \boldsymbol{\Delta}_S^{t+1} \right\|_\infty & \leq 4 \left(\frac{1+\sqrt{\rho}}{2} \right) \beta \rho^{\frac{3}{4}} \eta \left\| \bar{\mathbf{u}}_S^t \right\|_\infty B_k + 4\rho^{\frac{3}{4}} \eta \left\| \bar{\mathbf{u}}_S^t \right\|_\infty B_k \\ & \leq \frac{4}{(1-c_{10}(1-\sqrt{\rho}))^2} \left(\left(\frac{1+\sqrt{\rho}}{2} \right) \beta + 1 \right) \rho^{\frac{3}{4}} \eta \left\| \bar{\mathbf{u}}_S^{t+1} \right\|_\infty B_k, \end{aligned} \quad (34)$$

where the last inequality is based on induction such as if $\left\| \boldsymbol{\Delta}_S^t \right\|_\infty \leq 4\beta\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_S^t \right\|_\infty B_k$, then based (30) and step size condition $\eta \leq \frac{c_{10}(1-\sqrt{\rho})}{w_{\max}^*}$, there is

$$\left\| \bar{\mathbf{u}}_S^{t+1} \right\|_\infty \geq \left\| \bar{\mathbf{u}}_S^t \right\|_\infty (1 - c_{10}(1 - \sqrt{\rho}))^2. \quad (35)$$

To guarantee that (34) holds induction (c), the sufficient condition is

$$\frac{\left(\frac{1+\sqrt{\rho}}{2} \right) \beta + 1}{(1 - c_{10}(1 - \sqrt{\rho}))^2} \leq \beta. \quad (36)$$

The c_{10} should be chosen that $\frac{1+\sqrt{\rho}}{2(1-c_{10}(1-\sqrt{\rho}))^2} < 1$, which means that c_{10} should satisfy $c_{10} \leq \frac{1-\sqrt{\frac{1+\sqrt{\rho}}{2}}}{1-\sqrt{\rho}}$. Then we can set $c_{10} = \frac{1-\left(\frac{1+\sqrt{\rho}}{2}\right)^{\frac{1}{4}}}{1-\sqrt{\rho}}$, which results in $\frac{1}{(1-c_{10}(1-\sqrt{\rho}))^2} = \sqrt{\frac{2}{1+\sqrt{\rho}}}$. Based on (36), the lower bound for β is $\beta \geq \frac{2\left(\sqrt{2(1+\sqrt{\rho})}+1+\sqrt{\rho}\right)}{1-\rho}$.

Unfolding the recursion in (24) from the beginning of $(k-1)$ -th induction to the beginning of k -th induction based on (34) and combining induction (a) would have

$$\begin{aligned} \left\| \boldsymbol{\Delta}_S^{\bar{T}_k} \right\|_\infty & \leq \left(\frac{1+\sqrt{\rho}}{2} \right)^{2^k T} \left\| \boldsymbol{\Delta}_S^{\bar{T}_{k-1}} \right\|_\infty + 4\rho^{\frac{3}{4}} \eta B_k \sum_{i=0}^{2^k T-1} \left(\frac{1+\sqrt{\rho}}{2} \right)^{2^k T-1-i} \left\| \bar{\mathbf{u}}_S^{\bar{T}_{k-1}+i} \right\|_\infty \\ & \leq \frac{4\sqrt{2}\rho^{\frac{3}{4}} \eta B_k}{\sqrt{1+\sqrt{\rho}}} \sum_{i=0}^{2^k T-1} \left(\frac{1+\sqrt{\rho}}{2} \cdot \sqrt{\frac{2}{1+\sqrt{\rho}}} \right)^{2^k T-1-i} \left\| \bar{\mathbf{u}}_S^{\bar{T}_k} \right\|_\infty \\ & \quad + 4\beta\rho^{\frac{3}{4}} \eta \left(\frac{1+\sqrt{\rho}}{2} \right)^{2^k T} \left\| \bar{\mathbf{u}}_S^{\bar{T}_{k-1}} \right\|_\infty B_k \end{aligned}$$

$$\begin{aligned}
&\leq \frac{4\sqrt{2}\rho^{\frac{3}{4}}\eta B_k}{\sqrt{1+\sqrt{\rho}}} \sum_{i=0}^{2^k T-1} \left(\sqrt{\frac{1+\sqrt{\rho}}{2}} \right)^{2^k T-1-i} \left\| \bar{\mathbf{u}}_S^T \right\|_{\infty} \\
&\quad + 4\beta\rho^{\frac{3}{4}}\eta \left(\frac{1+\sqrt{\rho}}{2} \cdot \sqrt{\frac{2}{1+\sqrt{\rho}}} \right)^{2^k T} \left\| \bar{\mathbf{u}}_S^T \right\|_{\infty} B_k \\
&\leq 4\beta\rho^{\frac{3}{4}}\eta \left(\sqrt{\frac{1+\sqrt{\rho}}{2}} \right)^{2^k T} \left\| \bar{\mathbf{u}}_S^T \right\|_{\infty} B_k + \frac{2\sqrt{2} \left(1 + \sqrt{\frac{1+\sqrt{\rho}}{2}} \right)}{(1-\sqrt{\rho})\sqrt{1+\sqrt{\rho}}} \cdot 4\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_S^T \right\|_{\infty} B_k \\
&\leq 4\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_S^T \right\|_{\infty} \left(\beta \left(\sqrt{\frac{1+\sqrt{\rho}}{2}} \right)^{2^k T} + \frac{8}{1-\sqrt{\rho}} \right) B_k, \tag{37}
\end{aligned}$$

where both the second and third inequalities use lower bound in (35). To guarantee that the last inequality satisfies the induction (c), the β and ρ should satisfy the following condition

$$B_k \left(\beta \left(\sqrt{\frac{1+\sqrt{\rho}}{2}} \right)^{2^k T} + \frac{8}{1-\sqrt{\rho}} \right) \leq B_{k+1}\beta. \tag{38}$$

One sufficient condition for achieving above inequality is $\frac{8}{1-\sqrt{\rho}} \leq \frac{\beta}{4}$. Combine above lower bound of β , we can verify that $\beta = \frac{32}{1-\sqrt{\rho}}$ satisfies this condition and following inequality is attained

$$\left(\sqrt{\frac{1+\sqrt{\rho}}{2}} \right)^{\frac{1}{\eta w_{\max}^*} \log \frac{1}{\alpha^4}} \leq \frac{1}{4}, \tag{39}$$

which implies that the step size η should satisfy

$$\eta \leq \frac{\log \frac{1}{\alpha^4} \ln \left(1 + \sqrt{\frac{1+\sqrt{\rho}}{2}} - 1 \right)}{-2 \ln 2w_{\max}^*}. \tag{40}$$

One sufficient condition for achieving inequality is $\eta \leq \frac{\log \frac{1}{\alpha^4} \left(1 - \sqrt{\frac{1+\sqrt{\rho}}{2}} \right)}{4w_{\max}^*}$ based on inequality that $\ln \left(1 + \sqrt{\frac{1+\sqrt{\rho}}{2}} - 1 \right) \leq \sqrt{\frac{1+\sqrt{\rho}}{2}} - 1$. Combining all the above conditions on η , we obtain the upper bound as shown in condition (5).

(d) $\forall t$ that $\bar{T}_{k-1} \leq t < t+1 < \bar{T}_k$, based on one step iteration in (27) of Lemma 6 and inductions (a), (c), there is

$$\begin{aligned}
\left\| \Delta_{S^c}^{t+1} \right\|_{\infty} &\leq 4 \left(\frac{1+\sqrt{\rho}}{2} \right) \gamma \rho^{\frac{3}{4}} \eta \left\| \bar{\mathbf{u}}_{S^c}^t \right\|_{\infty} B_k + 4\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_{S^c}^t \right\|_{\infty} \\
&\quad \cdot \left(B_k + 4\sqrt{s}\delta\beta\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_S^t \right\|_{\infty} B_k \left(4\beta\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_S^t \right\|_{\infty} B_k + \left\| \bar{\mathbf{u}}_S^t \right\|_{\infty} \right) \right) \\
&\leq 4 \left(\frac{1+\sqrt{\rho}}{2} \right) \gamma \rho^{\frac{3}{4}} \eta \left\| \bar{\mathbf{u}}_{S^c}^t \right\|_{\infty} B_k + 4\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_{S^c}^t \right\|_{\infty} \left(B_k + 4\sqrt{s}\delta\beta\rho^{\frac{3}{4}}\eta B_k \left\| \bar{\mathbf{u}}_S^t \right\|_{\infty}^2 \right) \\
&\leq 4 \left(\frac{1+\sqrt{\rho}}{2} \right) \gamma \rho^{\frac{3}{4}} \eta \left\| \bar{\mathbf{u}}_{S^c}^t \right\|_{\infty} B_j + 4\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_{S^c}^t \right\|_{\infty} B_k \\
&\leq \frac{4}{(1-c_{10}(1-\sqrt{\rho}))^2} \left(\left(\frac{1+\sqrt{\rho}}{2} \right) \gamma + 1 \right) \rho^{\frac{3}{4}} \eta \left\| \bar{\mathbf{u}}_{S^c}^{t+1} \right\|_{\infty} B_k, \tag{41}
\end{aligned}$$

where the second inequality is due to $4\beta\rho^{\frac{3}{4}}\eta B_j \leq 4 \times \frac{32}{1-\sqrt{\rho}} \times \frac{1-\sqrt{\rho}}{256w_{\max}^*} B_j \leq 1$ and third inequality is due to global RIP condition on δ that is order of $\frac{1}{\sqrt{s}}$ and $\beta\eta \left\| \bar{\mathbf{u}}_S^t \right\|_{\infty}^2 \leq 1$, the last inequality is

based on induction such that if $\|\Delta_{S^c}^t\|_\infty \leq 4\gamma\rho^{\frac{3}{4}}\eta\|\bar{\mathbf{u}}_{S^c}^t\|_\infty B_k$, then based (45), if $\eta \leq \frac{c_{10}(1-\sqrt{\rho})}{w_{\max}^*}$, we have

$$\|\bar{\mathbf{u}}_{S^c}^{t+1}\|_\infty \geq \|\bar{\mathbf{u}}_{S^c}^t\|_\infty (1 - c_{10}(1 - \sqrt{\rho}))^2. \quad (42)$$

With the same derivation in proving induction (c), the lower bound for γ is the same as that of β .

Then also unrolling the recursion in (27) from beginning of $(k-1)$ -th induction to beginning of k -th induction and combining induction (a) would have

$$\begin{aligned} \|\Delta_{S^c}^{\bar{T}_k}\|_\infty &\leq \left(\frac{1+\sqrt{\rho}}{2}\right)^{2^k T} \|\Delta_{S^c}^{\bar{T}_{k-1}}\|_\infty + 4\rho^{\frac{3}{4}}\eta B_k \sum_{i=0}^{2^k T-1} \left(\frac{1+\sqrt{\rho}}{2}\right)^{2^k T-1-i} \|\bar{\mathbf{u}}_{S^c}^{\bar{T}_{k-1}+i}\|_\infty \\ &\leq 4\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^c}^{\bar{T}_k}\|_\infty B_k \left(\gamma \left(\sqrt{\frac{1+\sqrt{\rho}}{2}} \right)^{2^k T} + \frac{8}{1-\sqrt{\rho}} \right). \end{aligned} \quad (43)$$

The derivation of above inequality is similar with (37). To guarantee the above two inequality satisfy induction (d), the γ and ρ can have the same value as in (38) that $\gamma = \beta = \frac{32}{1-\sqrt{\rho}}$.

(e) For the upper bound, the proof is divided into three cases.

(1) One of case is $\left(\bar{u}_j^{\bar{T}_{k-1}}\right)^2 \leq w_j^* + B_k$, based on the Lemma B.6 in [36], we can conclude that $\forall t \geq \bar{T}_{k-1}$, $\left(\bar{u}_j^t\right)^2$ would keep below $w_j^* + B_k$.

(2) Another case is $w_j^* + B_k \leq \left(\bar{u}_j^{\bar{T}_{k-1}}\right)^2 \leq w_j^* + 2B_k$, in this case $\forall t \geq \bar{T}_{k-1}$, Either $0 \leq \left(\bar{u}_j^t\right)^2 \leq w_j^* + B_k$, which can use result in case (1) that $\forall t' \geq t$, $0 \leq \left(\bar{u}_j^{t'}\right)^2 \leq w_j^* + B_k$ or that $w_j^* + B_k \leq \left(\bar{u}_j^t\right)^2 \leq w_j^* + 2B_k$, which would guarantee that $\left(\bar{u}_j^{t+1}\right)^2 \leq \left(\bar{u}_j^t\right)^2$.

(3) The last case is $w_j^* + 2B_k < \left(\bar{u}_j^{\bar{T}_{k-1}}\right)^2 \leq w_j^* + 4B_k$, then based on Lemma B.14 [36], after the sufficient number of iterations that $\forall t \geq \frac{1}{10\eta B_k} = \frac{4 \times 2^k}{\eta w_{\max}^*}$ and we can check that T_k is large enough that satisfies this condition, it can keep that $\left(\bar{u}_j^{\bar{T}_{k-1}+t}\right)^2 \leq w_j^* + 2B_k$.

Following the above three cases, we can guarantee that there exists $\left(\bar{u}_j^{\bar{T}_k}\right)^2 \leq w_j^* + \frac{1}{2} \times 4B_k = w_j^* + 4B_{k+1}$.

For the lower bound, we can also have bound $\|\mathbf{g}^t\|_\infty \leq \frac{3\sqrt{\rho}B_k}{8}$ by comparing formulas between $\mathbf{p}^t$ and $\mathbf{g}^t$, $\forall t$ that $\bar{T}_{k-1} \leq t < \bar{T}_k$. For perturbation $\mathbf{f}^t$, based on induction (a), fine-grained bound in (d), $\forall j \in S^c$ it has

$$\begin{aligned} |f_j^t| &\leq 4\beta\rho^{\frac{3}{4}}\eta|\bar{u}_j^t|B_k \left(\sqrt{s}\delta_{\max} \|\bar{\mathbf{u}}_{S^c}^t \odot \bar{\mathbf{u}}_{S^c}^t - \mathbf{w}^*\|_\infty + \frac{3\sqrt{\rho}B_k}{16} + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) \\ &\leq \frac{\sqrt{\rho}B_k|\bar{u}_j^t|}{16}, \end{aligned} \quad (44)$$

where the first inequality is due to bound for $\|\mathbf{g}^t\|_\infty$ and the last inequality is the same reason as (29). $\mathbf{f}^t$ can also be reparameterized as $\mathbf{f}^t = \mathbf{r}_f^t \odot \bar{\mathbf{u}}_{S^c}^t$ where $\|\mathbf{r}_f^t\|_\infty \leq \frac{\sqrt{\rho}B_k}{16}$ for $\forall t$ that $\bar{T}_{k-1} \leq t < \bar{T}_k$, the perturbed optimization recursion on non-support S^c over decentralized network in (11) becomes

$$\left(\bar{\mathbf{u}}_{S^c}^{t+1}\right)^2 = \left(\bar{\mathbf{u}}_{S^c}^t\right)^2 \odot \left(\mathbf{1}_d - 4\eta \left(\mathbf{E}_2^t + \mathbf{E}_3^t + \mathbf{g}^t + \mathbf{r}_f^t\right)\right)^2. \quad (45)$$

Based on the similar lower bound (79) and upper bound in (80), we can conclude that $\forall t \leq \mathcal{O}\left(\frac{1}{\eta\zeta} \log \frac{1}{\alpha^4}\right)$, there is

$$\prod_{i=0}^t \left(1 + 8\eta \|\mathbf{E}_2^t\|_\infty + \|\mathbf{E}_3^t\|_\infty + \|\mathbf{g}^t\|_\infty + \|\mathbf{r}_f^t\|_\infty\right)^2 \leq \frac{1}{\alpha}, \quad (46)$$

where this inequality guarantees that the averaged signal on non-support remains $\|\bar{\mathbf{u}}_{S^c}^t\|_\infty \leq \sqrt{\alpha}$ until early stopping. $\forall j \in \mathcal{S}$, we use j_k to denote the largest index that $w_j^* \leq B_{j_k} + \alpha^3$. As $B_0 = w_{\max}^*$, the existence of j_k is guaranteed. Then for $t = 0, \dots, \bar{T}_{j_k} - 1$ and based on (30), the $(\bar{u}_j^t)^2$ would shrinkage from initialization α^2 , to obtain the lower bound, we should consider the maximum shrinkage as follows

$$\begin{aligned} \left(\bar{u}_j^{\bar{T}_{j_k}}\right)^2 &\geq \alpha^2 \prod_{i=0}^{\bar{T}_{j_k}-1} \left(1 - 4\eta \left((\bar{u}_j^i)^2 + \|\mathbf{E}_2^i\|_\infty + \|\mathbf{E}_3^i\|_\infty + \|\mathbf{g}^i\|_\infty + \|\mathbf{r}_f^i\|_\infty\right)\right)^2 \\ &\geq \alpha^3, \end{aligned} \quad (47)$$

where the last inequality is due to (46), $(\bar{u}_j^i)^2 \leq \alpha^2, \forall i = 1, \dots, \bar{T}_{j_k} - 1$ and step size condition that $\eta \left((\bar{u}_j^i)^2 + \|\mathbf{E}_2^i\|_\infty + \|\mathbf{E}_3^i\|_\infty + \|\mathbf{g}^i\|_\infty + \|\mathbf{r}_f^i\|_\infty\right) \leq \frac{3}{40}$ and $(1 - 5x)(1 + 8x) \geq 1, \forall x \in [0, \frac{3}{40}]$. Thus, we have $(\bar{u}_j^t)^2 \geq \alpha^3$ for $\forall t = 0, \dots, \bar{T}_{j_k}$.

For $t = \bar{T}_{j_k} + 1, \dots, \bar{T}_{j_k+1} - 1$ and $\forall j \in \mathcal{S}$, let consider the auxiliary iterations that $(\hat{u}_j^t)^2 = (\hat{u}_j^{t-1})^2 \odot \left(1 - 4\eta \left((\hat{u}_j^{t-1})^2 - (w_j^* - B_{j_k+1})\right)\right)^2$ where $\hat{u}_j^{\bar{T}_{j_k}} = \bar{u}_j^{\bar{T}_{j_k}} \geq \alpha^3$ as above proved. Based on definition of j_k , there is $w_j^* - B_{j_k+1} \geq \alpha^3$. According to monotonic property in Lemma B.6 in [36], we can guarantee that $\left(\hat{u}_j^{\bar{T}_{j_k+1}}\right)^2 \in \left[\min \left\{\hat{u}_j^{\bar{T}_{j_k}}, w_j^* - B_{j_k+1}\right\}, \max \left\{\hat{u}_j^{\bar{T}_{j_k}}, w_j^* - B_{j_k+1}\right\}\right]$ that ensures that $\left(\hat{u}_j^{\bar{T}_{j_k+1}}\right)^2 \geq \left(\hat{u}_j^{\bar{T}_{j_k}}\right)^2 \geq \alpha^3$ where the first inequality is due to the squeezing property in Lemma B.9 in [36]. Then we can follow the same analysis to prove that left i -th induction step that $i = j_k + 2, \dots, m - 1$ based on the monotonic property of B_k . □

A.4 General Case

This section considers the general setting where the ground truth $\mathbf{w}^*$ includes both positive and negative elements in its support. The analysis here is more complex than in the previous section due to the presence of $\mathbf{v}^{t,i}$ and its consensus error terms. The following formulates the recursion of perturbed average variable $\bar{\mathbf{u}}^t$ and similar derivation could be applied to $\bar{\mathbf{v}}^t$.

Lemma 7. Consider the sequence $\{\mathbf{u}_i^t, \mathbf{v}_i^t\}, i \in [m]$ generated according to (3) and (4) by DGD for solving problem (2), the average signal $\{\bar{\mathbf{u}}^t\}$ on positive support $\mathcal{S}^+$, negative support $\mathcal{S}^-$ and non-support $\mathcal{S}^c$ are updated according to the following formulas

$$\begin{aligned} \bar{\mathbf{u}}_{S^+}^{t+1} &= \bar{\mathbf{u}}_{S^+}^t \odot \left(\mathbf{1}_d - 4\eta \left((\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* \right) - 4\eta \frac{\mathbf{X}^T \mathbf{X}}{N} \left((\bar{\mathbf{u}}_{S^c}^t)^2 - (\bar{\mathbf{v}}_{S^c}^t)^2 + (\bar{\mathbf{u}}_{S^-}^t)^2 - (\bar{\mathbf{v}}_{S^+}^t)^2 \right) \right. \\ &\quad \left. + 4\eta \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} - 4\eta \left(\frac{\mathbf{X}^T \mathbf{X}}{N} - \mathbf{I} \right) \left((\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* - (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right) - 4\eta \mathbf{p}_u^t \right) - 4\eta \mathbf{q}_u^t; \end{aligned} \quad (48)$$

$$\bar{\mathbf{u}}_{S^-}^{t+1} = \bar{\mathbf{u}}_{S^-}^t \odot \left(\mathbf{1}_d - 4\eta \left(-(\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* + \frac{\mathbf{X}^T \mathbf{X}}{N} \left((\bar{\mathbf{u}}_{S^c}^t)^2 - (\bar{\mathbf{v}}_{S^c}^t)^2 + (\bar{\mathbf{u}}_{S^-}^t)^2 - (\bar{\mathbf{v}}_{S^+}^t)^2 \right) \right) \right)$$

$$-\frac{\mathbf{X}^T \boldsymbol{\xi}}{N} + \left(\frac{\mathbf{X}^T \mathbf{X}}{N} - \mathbf{I} \right) \left((\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* - (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right) - 4\eta \mathbf{y}_u^t \Big) - 4\eta \mathbf{z}_u^t; \quad (49)$$

$$\begin{aligned} \bar{\mathbf{u}}_{S^c}^{t+1} = & \bar{\mathbf{u}}_{S^c}^t \odot \left(\mathbf{1}_d - 4\eta \left(\frac{\mathbf{X}^T \mathbf{X}}{N} \left((\bar{\mathbf{u}}_{S^c}^t)^2 - (\bar{\mathbf{v}}_{S^c}^t)^2 + (\bar{\mathbf{u}}_{S^-}^t)^2 - (\bar{\mathbf{v}}_{S^+}^t)^2 \right) - \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right. \right. \\ & \left. \left. + \left(\frac{\mathbf{X}^T \mathbf{X}}{N} - \mathbf{I} \right) \left((\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* - (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right) \right) - 4\eta \mathbf{g}_u^t \right) - 4\eta \mathbf{f}_u^t, \end{aligned} \quad (50)$$

where the perturbed error terms $\mathbf{p}_u^t, \mathbf{q}_u^t, \mathbf{y}_u^t, \mathbf{z}_u^t, \mathbf{g}_u^t, \mathbf{f}_u^t$ induced from decentralized network are defined as

$$\begin{aligned} \mathbf{p}_u^t := & \frac{1}{m} \sum_{i=1}^m \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left(2\bar{\mathbf{u}}_{S^+}^t \odot \boldsymbol{\Delta}_{u,S^+}^{t,i} + \left(\boldsymbol{\Delta}_{u,S^+}^{t,i} \right)^2 - 2\bar{\mathbf{v}}_{S^-}^t \odot \boldsymbol{\Delta}_{v,S^-}^{t,i} - \left(\boldsymbol{\Delta}_{v,S^-}^{t,i} \right)^2 \right) \\ & + 3 \left(\boldsymbol{\Delta}_{u,S^+}^{t,i} \right)^2 + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left(2\bar{\mathbf{u}}_{S^c}^t \odot \boldsymbol{\Delta}_{u,S^c}^{t,i} + \left(\boldsymbol{\Delta}_{u,S^c}^{t,i} \right)^2 + 2\bar{\mathbf{u}}_{S^-}^t \odot \boldsymbol{\Delta}_{u,S^-}^{t,i} + \left(\boldsymbol{\Delta}_{u,S^-}^{t,i} \right)^2 \right. \\ & \left. - 2\bar{\mathbf{v}}_{S^c}^t \odot \boldsymbol{\Delta}_{v,S^c}^{t,i} - \left(\boldsymbol{\Delta}_{v,S^c}^{t,i} \right)^2 - 2\bar{\mathbf{v}}_{S^+}^t \odot \boldsymbol{\Delta}_{v,S^+}^{t,i} - \left(\boldsymbol{\Delta}_{v,S^+}^{t,i} \right)^2 \right); \end{aligned} \quad (51)$$

$$\begin{aligned} \mathbf{q}_u^t := & \frac{1}{m} \sum_{i=1}^m \boldsymbol{\Delta}_{u,S^+}^{t,i} \odot \left(\left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left((\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* - (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right. \right. \\ & \left. \left. + 2\bar{\mathbf{u}}_{S^+}^t \odot \boldsymbol{\Delta}_{u,S^+}^{t,i} + \left(\boldsymbol{\Delta}_{u,S^+}^{t,i} \right)^2 - 2\bar{\mathbf{v}}_{S^-}^t \odot \boldsymbol{\Delta}_{v,S^-}^{t,i} - \left(\boldsymbol{\Delta}_{v,S^-}^{t,i} \right)^2 \right) \right. \\ & \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left((\bar{\mathbf{u}}_{S^c}^t + \boldsymbol{\Delta}_{u,S^c}^{t,i})^2 - (\bar{\mathbf{v}}_{S^c}^t + \boldsymbol{\Delta}_{v,S^c}^{t,i})^2 + (\bar{\mathbf{u}}_{S^-}^t + \boldsymbol{\Delta}_{u,S^-}^{t,i})^2 - (\bar{\mathbf{v}}_{S^+}^t + \boldsymbol{\Delta}_{v,S^+}^{t,i})^2 \right) \right. \\ & \left. + \left(\boldsymbol{\Delta}_{u,S^+}^{t,i} \right)^2 - \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right); \end{aligned} \quad (52)$$

$$\mathbf{g}_u^t := \mathbf{p}_u^t - \frac{3}{m} \sum_{i=1}^m \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left(\boldsymbol{\Delta}_{u,S^+}^{t,i} \right)^2; \quad (53)$$

$$\begin{aligned} \mathbf{f}_u^t := & \frac{1}{m} \sum_{i=1}^m \boldsymbol{\Delta}_{u,S^c}^{t,i} \odot \left(\left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left((\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* - (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right. \right. \\ & \left. \left. + 2\bar{\mathbf{u}}_{S^+}^t \odot \boldsymbol{\Delta}_{u,S^+}^{t,i} + \left(\boldsymbol{\Delta}_{u,S^+}^{t,i} \right)^2 - 2\bar{\mathbf{v}}_{S^-}^t \odot \boldsymbol{\Delta}_{v,S^-}^{t,i} - \left(\boldsymbol{\Delta}_{v,S^-}^{t,i} \right)^2 \right) - \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right. \\ & \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left((\bar{\mathbf{u}}_{S^c}^t + \boldsymbol{\Delta}_{u,S^c}^{t,i})^2 - (\bar{\mathbf{v}}_{S^c}^t + \boldsymbol{\Delta}_{v,S^c}^{t,i})^2 + (\bar{\mathbf{u}}_{S^-}^t + \boldsymbol{\Delta}_{u,S^-}^{t,i})^2 - (\bar{\mathbf{v}}_{S^+}^t + \boldsymbol{\Delta}_{v,S^+}^{t,i})^2 \right) \right); \end{aligned} \quad (54)$$

$$\mathbf{y}_u^t := \mathbf{g}_u^t - \frac{1}{m} \sum_{i=1}^m \left(\boldsymbol{\Delta}_{v,S^-}^{t,i} \right)^2; \quad (55)$$

$$\begin{aligned} \mathbf{z}_u^t := & \frac{1}{m} \sum_{i=1}^m \boldsymbol{\Delta}_{u,S^-}^{t,i} \odot \left(-2\bar{\mathbf{v}}_{S^-}^t \odot \boldsymbol{\Delta}_{v,S^-}^{t,i} - \left(\boldsymbol{\Delta}_{v,S^-}^{t,i} \right)^2 + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left((\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* - (\bar{\mathbf{v}}_{S^-}^t)^2 \right. \right. \\ & \left. \left. - \mathbf{w}_{S^-}^* + 2\bar{\mathbf{u}}_{S^+}^t \odot \boldsymbol{\Delta}_{u,S^+}^{t,i} + \left(\boldsymbol{\Delta}_{u,S^+}^{t,i} \right)^2 - 2\bar{\mathbf{v}}_{S^-}^t \odot \boldsymbol{\Delta}_{v,S^-}^{t,i} - \left(\boldsymbol{\Delta}_{v,S^-}^{t,i} \right)^2 \right) \right. \\ & \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left((\bar{\mathbf{u}}_{S^c}^t + \boldsymbol{\Delta}_{u,S^c}^{t,i})^2 - (\bar{\mathbf{v}}_{S^c}^t + \boldsymbol{\Delta}_{v,S^c}^{t,i})^2 + (\bar{\mathbf{u}}_{S^-}^t + \boldsymbol{\Delta}_{u,S^-}^{t,i})^2 - (\bar{\mathbf{v}}_{S^+}^t + \boldsymbol{\Delta}_{v,S^+}^{t,i})^2 \right) \right. \\ & \left. + \left(\boldsymbol{\Delta}_{u,S^+}^{t,i} \right)^2 - \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right); \end{aligned} \quad (56)$$

Proof. The average of optimization for $\mathbf{u}^{t,i}, i \in [m]$ under DGD are as follows

$$\begin{aligned}\bar{\mathbf{u}}^{t+1} &= \bar{\mathbf{u}}^t - \frac{\eta}{m} \sum_{i=1}^m \nabla_{\mathbf{u}} f_i(\mathbf{u}^{t,i}, \mathbf{v}^{t,i}) \\ &= \bar{\mathbf{u}}^t - \frac{\eta}{m} \sum_{i=1}^m \nabla_{\mathbf{u}} f_i(\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t) + \frac{\eta}{m} \sum_{i=1}^m \underbrace{\left(\nabla_{\mathbf{u}} f_i(\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t) - \nabla_{\mathbf{u}} f_i(\bar{\mathbf{u}}^t, \mathbf{v}^{t,i}) \right)}_{\Pi_{1,i}} \\ &\quad + \frac{\eta}{m} \sum_{i=1}^m \underbrace{\left(\nabla_{\mathbf{u}} f_i(\bar{\mathbf{u}}^t, \mathbf{v}^{t,i}) - \nabla_{\mathbf{u}} f_i(\mathbf{u}^{t,i}, \mathbf{v}^{t,i}) \right)}_{\Pi_{2,i}}.\end{aligned}\quad (57)$$

Based on error decomposition on positive support $\mathcal{S}^+$, negative support $\mathcal{S}^-$ and non-support $\mathcal{S}^c$, there has

$$\begin{aligned}\nabla_{\mathbf{u}} F(\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t) &= 4\bar{\mathbf{u}}^t \odot \left((\bar{\mathbf{u}}_{\mathcal{S}^+}^t)^2 - (\bar{\mathbf{v}}_{\mathcal{S}^-}^t)^2 - \mathbf{w}^* + \left(\frac{\mathbf{X}^T \mathbf{X}}{N} - \mathbf{I} \right) \left((\bar{\mathbf{u}}_{\mathcal{S}^+}^t)^2 - (\bar{\mathbf{v}}_{\mathcal{S}^-}^t)^2 - \mathbf{w}^* \right) \right. \\ &\quad \left. + \frac{\mathbf{X}^T \mathbf{X}}{N} \left((\bar{\mathbf{u}}_{\mathcal{S}^c}^t)^2 - (\bar{\mathbf{v}}_{\mathcal{S}^c}^t)^2 + (\bar{\mathbf{u}}_{\mathcal{S}^-}^t)^2 - (\bar{\mathbf{v}}_{\mathcal{S}^+}^t)^2 \right) - \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right).\end{aligned}\quad (58)$$

With same decomposition on $\mathcal{S}^+, \mathcal{S}^-, \mathcal{S}^c$ for $\Pi_{1,i}$ and $\Pi_{2,i}$, there are

$$\begin{aligned}\Pi_{1,i} &= 4\bar{\mathbf{u}}^t \odot \left(2\bar{\mathbf{v}}_{\mathcal{S}^-}^t \odot \Delta_{\mathbf{v}, \mathcal{S}^-}^{t,i} + \left(\Delta_{\mathbf{v}, \mathcal{S}^-}^{t,i} \right)^2 + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left(2\bar{\mathbf{v}}_{\mathcal{S}^-}^t \odot \Delta_{\mathbf{v}, \mathcal{S}^-}^{t,i} + \left(\Delta_{\mathbf{v}, \mathcal{S}^-}^{t,i} \right)^2 \right) \right. \\ &\quad \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left(2\bar{\mathbf{v}}_{\mathcal{S}^c}^t \odot \Delta_{\mathbf{v}, \mathcal{S}^c}^{t,i} + \left(\Delta_{\mathbf{v}, \mathcal{S}^c}^{t,i} \right)^2 + 2\bar{\mathbf{v}}_{\mathcal{S}^+}^t \odot \Delta_{\mathbf{v}, \mathcal{S}^+}^{t,i} + \left(\Delta_{\mathbf{v}, \mathcal{S}^+}^{t,i} \right)^2 \right) \right); \\ \Pi_{2,i} &= 4\bar{\mathbf{u}}^t \odot \left(-2\bar{\mathbf{u}}_{\mathcal{S}^+}^t \odot \Delta_{\mathbf{u}, \mathcal{S}^+}^{t,i} - \left(\Delta_{\mathbf{u}, \mathcal{S}^+}^{t,i} \right)^2 + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left(-2\bar{\mathbf{u}}_{\mathcal{S}^+}^t \odot \Delta_{\mathbf{u}, \mathcal{S}^+}^{t,i} - \left(\Delta_{\mathbf{u}, \mathcal{S}^+}^{t,i} \right)^2 \right) \right. \\ &\quad \left. + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left(-2\bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \Delta_{\mathbf{u}, \mathcal{S}^c}^{t,i} - \left(\Delta_{\mathbf{u}, \mathcal{S}^c}^{t,i} \right)^2 - 2\bar{\mathbf{u}}_{\mathcal{S}^-}^t \odot \Delta_{\mathbf{u}, \mathcal{S}^-}^{t,i} - \left(\Delta_{\mathbf{u}, \mathcal{S}^-}^{t,i} \right)^2 \right) \right) \\ &\quad - 4\Delta_{\mathbf{u}}^{t,i} \odot \left((\bar{\mathbf{u}}_{\mathcal{S}^+}^t)^2 - (\bar{\mathbf{v}}_{\mathcal{S}^-}^t)^2 - \mathbf{w}^* + 2\bar{\mathbf{u}}_{\mathcal{S}^+}^t \odot \Delta_{\mathbf{u}, \mathcal{S}^+}^{t,i} + \left(\Delta_{\mathbf{u}, \mathcal{S}^+}^{t,i} \right)^2 - \Delta_{\mathbf{v}, \mathcal{S}^-}^{t,i} \left(2\bar{\mathbf{v}}_{\mathcal{S}^-}^t + \Delta_{\mathbf{v}, \mathcal{S}^-}^{t,i} \right) \right. \\ &\quad \left. + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left((\bar{\mathbf{u}}_{\mathcal{S}^+}^t)^2 - (\bar{\mathbf{v}}_{\mathcal{S}^-}^t)^2 - \mathbf{w}^* + 2\bar{\mathbf{u}}_{\mathcal{S}^+}^t \odot \Delta_{\mathbf{u}, \mathcal{S}^+}^{t,i} + \left(\Delta_{\mathbf{u}, \mathcal{S}^+}^{t,i} \right)^2 - 2\bar{\mathbf{v}}_{\mathcal{S}^-}^t \odot \Delta_{\mathbf{v}, \mathcal{S}^-}^{t,i} \right. \right. \\ &\quad \left. \left. - \left(\Delta_{\mathbf{v}, \mathcal{S}^-}^{t,i} \right)^2 \right) + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left((\bar{\mathbf{u}}_{\mathcal{S}^c}^t)^2 - (\bar{\mathbf{v}}_{\mathcal{S}^c}^t)^2 + 2\bar{\mathbf{u}}_{\mathcal{S}^c}^t \odot \Delta_{\mathbf{u}, \mathcal{S}^c}^{t,i} + \left(\Delta_{\mathbf{u}, \mathcal{S}^c}^{t,i} \right)^2 - 2\bar{\mathbf{v}}_{\mathcal{S}^c}^t \odot \Delta_{\mathbf{v}, \mathcal{S}^c}^{t,i} \right. \right. \\ &\quad \left. \left. - \left(\Delta_{\mathbf{v}, \mathcal{S}^c}^{t,i} \right)^2 + (\bar{\mathbf{u}}_{\mathcal{S}^-}^t)^2 - (\bar{\mathbf{v}}_{\mathcal{S}^+}^t)^2 + 2\bar{\mathbf{u}}_{\mathcal{S}^-}^t \odot \Delta_{\mathbf{u}, \mathcal{S}^-}^{t,i} + \left(\Delta_{\mathbf{u}, \mathcal{S}^-}^{t,i} \right)^2 - 2\bar{\mathbf{v}}_{\mathcal{S}^+}^t \odot \Delta_{\mathbf{v}, \mathcal{S}^+}^{t,i} \right. \right. \\ &\quad \left. \left. - \left(\Delta_{\mathbf{v}, \mathcal{S}^+}^{t,i} \right)^2 - \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right) \right).\end{aligned}\quad (59)$$

Substituting (59) into (57) would obtain the perturbed recursion of averaged $\bar{\mathbf{u}}_{\mathcal{S}^+}^t, \bar{\mathbf{u}}_{\mathcal{S}^-}^t, \bar{\mathbf{u}}_{\mathcal{S}^c}^t$ in Lemma 7. $\square$

The following lemma separately bounds the consensus errors on $\mathcal{S}^+, \mathcal{S}^-$, and $\mathcal{S}^c$ by the corresponding magnitudes of parameters, which is different from the current analysis in the decentralized optimization literature.

Lemma 8. Consider the sequence $\{\mathbf{u}^{t,i}, \mathbf{v}^{t,i}\}, i \in [m]$ generated according to (3) and (4) by DGD for solving (2), the consensus error of $\Delta_{\mathbf{u}}^t$ on positive support S^+ , negative support S^- and non-support S^c have the following recursions

$$\begin{aligned} \|\Delta_{\mathbf{u}, S^+}^{t+1}\|_{\infty} &\leq \rho \|\Delta_{\mathbf{u}, S^+}^t\|_{\infty} \cdot \left(1 + 4\eta \left((\sqrt{s}\delta_{\max} + 1) \left(\left\| (\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* \right\|_{\infty} + \left\| (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right\|_{\infty} \right) \right. \right. \\ &\quad + 2d \left(\left(\|\bar{\mathbf{u}}_{S^c}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^c}^t\|_{\infty} \right)^2 + \left(\|\bar{\mathbf{v}}_{S^c}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^c}^t\|_{\infty} \right)^2 \right. \\ &\quad + \left. \left(\|\bar{\mathbf{u}}_{S^-}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^-}^t\|_{\infty} \right)^2 + \left(\|\bar{\mathbf{v}}_{S^+}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^+}^t\|_{\infty} \right)^2 \right) \\ &\quad \left. + 2(\sqrt{s}\delta_{\max} + 1) \left(\|\bar{\mathbf{u}}_{S^+}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^+}^t\|_{\infty} \right)^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \right) \\ &\quad + 4\rho\eta \|\bar{\mathbf{u}}_{S^+}^t\|_{\infty} \cdot \left(\sqrt{s}\delta_{\max} \left(2\|\bar{\mathbf{v}}_{S^-}^t\|_{\infty} \|\Delta_{\mathbf{v}, S^-}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^-}^t\|_{\infty}^2 \right) + C_{err}^t \right); \end{aligned} \quad (60)$$

$$\begin{aligned} \|\Delta_{\mathbf{u}, S^-}^{t+1}\|_{\infty} &\leq \rho \|\Delta_{\mathbf{u}, S^-}^t\|_{\infty} \cdot \left(1 + 4\rho\eta \left((\sqrt{s}\delta_{\max} + 1) \left(\left\| (\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* \right\|_{\infty} + \left\| (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right\|_{\infty} \right) \right. \right. \\ &\quad + 2\|\bar{\mathbf{u}}_{S^+}^t\|_{\infty} \|\Delta_{\mathbf{u}, S^+}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^+}^t\|_{\infty}^2 + 2\|\bar{\mathbf{v}}_{S^-}^t\|_{\infty} \|\Delta_{\mathbf{v}, S^-}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^-}^t\|_{\infty}^2 \left. \right) \\ &\quad + 4d \left(\left(\|\bar{\mathbf{u}}_{S^c}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^c}^t\|_{\infty} \right)^2 + \left(\|\bar{\mathbf{v}}_{S^c}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^c}^t\|_{\infty} \right)^2 \right. \\ &\quad + \left. \left(\|\bar{\mathbf{u}}_{S^-}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^-}^t\|_{\infty} \right)^2 + \left(\|\bar{\mathbf{v}}_{S^+}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^+}^t\|_{\infty} \right)^2 \right) + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \left. \right) \\ &\quad + 4\rho\eta \|\bar{\mathbf{u}}_{S^-}^t\|_{\infty} \cdot \left((\sqrt{s}\delta_{\max} + 1) \cdot \left(2\|\bar{\mathbf{u}}_{S^+}^t\|_{\infty} \|\Delta_{\mathbf{u}, S^+}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^+}^t\|_{\infty}^2 \right. \right. \\ &\quad \left. \left. + 2\|\bar{\mathbf{v}}_{S^-}^t\|_{\infty} \|\Delta_{\mathbf{v}, S^-}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^-}^t\|_{\infty}^2 \right) + C_{err} \right); \end{aligned} \quad (61)$$

$$\begin{aligned} \|\Delta_{\mathbf{u}, S^c}^{t+1}\|_{\infty} &\leq \rho \|\Delta_{\mathbf{u}, S^c}^t\|_{\infty} \cdot \left(1 + 4\eta \left(\sqrt{s}\delta_{\max} \left(\left\| (\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* \right\|_{\infty} + \left\| (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right\|_{\infty} \right) \right. \right. \\ &\quad + 2\|\bar{\mathbf{u}}_{S^+}^t\|_{\infty} \|\Delta_{\mathbf{u}, S^+}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^+}^t\|_{\infty}^2 + 2\|\bar{\mathbf{v}}_{S^-}^t\|_{\infty} \|\Delta_{\mathbf{v}, S^-}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^-}^t\|_{\infty}^2 \left. \right) \\ &\quad + 4d \left(\left(\|\bar{\mathbf{u}}_{S^c}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^c}^t\|_{\infty} \right)^2 + \left(\|\bar{\mathbf{v}}_{S^c}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^c}^t\|_{\infty} \right)^2 \right. \\ &\quad + \left. \left(\|\bar{\mathbf{u}}_{S^-}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^-}^t\|_{\infty} \right)^2 + \left(\|\bar{\mathbf{v}}_{S^+}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^+}^t\|_{\infty} \right)^2 \right) + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \left. \right) \\ &\quad + 4\rho\eta \|\bar{\mathbf{u}}_{S^c}^t\|_{\infty} \cdot \left(\sqrt{s}\delta_{\max} \cdot \left(2\|\bar{\mathbf{u}}_{S^+}^t\|_{\infty} \|\Delta_{\mathbf{u}, S^+}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^+}^t\|_{\infty}^2 \right. \right. \\ &\quad \left. \left. + 2\|\bar{\mathbf{v}}_{S^-}^t\|_{\infty} \|\Delta_{\mathbf{v}, S^-}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^-}^t\|_{\infty}^2 \right) + C_{err}^t \right), \end{aligned} \quad (62)$$

where the C_{err}^t is defined as

$$\begin{aligned} C_{err}^t &:= \sqrt{s}(\delta_{\max} + \delta) \cdot \left(\left\| (\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* \right\|_{\infty} + \left\| (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right\|_{\infty} \right) \\ &\quad + 2d \left(\left(\|\bar{\mathbf{u}}_{S^c}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^c}^t\|_{\infty} \right)^2 + \left(\|\bar{\mathbf{v}}_{S^c}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^c}^t\|_{\infty} \right)^2 \right. \\ &\quad + \left. \left(\|\bar{\mathbf{u}}_{S^-}^t\|_{\infty} + \|\Delta_{\mathbf{u}, S^-}^t\|_{\infty} \right)^2 + \left(\|\bar{\mathbf{v}}_{S^+}^t\|_{\infty} + \|\Delta_{\mathbf{v}, S^+}^t\|_{\infty} \right)^2 \right) \\ &\quad + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} + \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_{\infty}. \end{aligned} \quad (63)$$

Proof. Based on the updating of DGD, there is

$$\begin{aligned}
\mathbf{U}^{t+1} \left(\mathbf{I}_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T \right) &= (\mathbf{U}^t - \eta \nabla_{\mathbf{u}} \mathbf{f}(\mathbf{U}^t, \mathbf{V}^t)) \mathbf{W} \left(\mathbf{I}_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T \right) \\
&= \left(\mathbf{U}^t - \nabla_{\mathbf{u}} \mathbf{f}(\mathbf{U}^t, \mathbf{V}^t) + \nabla_{\mathbf{u}} \mathbf{F}(\bar{\mathbf{U}}^t, \bar{\mathbf{V}}^t) \right) \left(\mathbf{W} - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T \right) \\
&= \left(\mathbf{U}^t - \nabla_{\mathbf{u}} \mathbf{f}(\mathbf{U}^t, \mathbf{V}^t) + \nabla_{\mathbf{u}} \mathbf{f}(\bar{\mathbf{U}}^t, \mathbf{V}^t) \right. \\
&\quad \left. + \nabla_{\mathbf{u}} \mathbf{f}(\bar{\mathbf{U}}^t, \bar{\mathbf{V}}^t) - \nabla_{\mathbf{u}} \mathbf{f}(\bar{\mathbf{U}}^t, \mathbf{V}^t) \right. \\
&\quad \left. + \nabla_{\mathbf{u}} \mathbf{F}(\bar{\mathbf{U}}^t, \bar{\mathbf{V}}^t) - \nabla_{\mathbf{u}} \mathbf{f}(\bar{\mathbf{U}}^t, \bar{\mathbf{V}}^t) \right) \left(\mathbf{W} - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T \right). \quad (64)
\end{aligned}$$

Thus, there is

$$\begin{aligned}
\|\Delta_{\mathbf{u}, \mathcal{S}^+}^{t+1}\|_{\infty} &\leq \rho \|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty} + \rho \eta \max_i \|\mathbf{1}_{\mathcal{S}^+} \odot (\nabla_{\mathbf{u}} f_i(\mathbf{u}_i^t, \mathbf{v}_i^t) - \nabla_{\mathbf{u}} F(\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t))\|_{\infty} \\
&\leq \rho \|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty} + \rho \eta \max_i \|\mathbf{1}_{\mathcal{S}^+} \odot \Pi_{1,i}\|_{\infty} + \rho \eta \max_i \|\mathbf{1}_{\mathcal{S}^+} \odot \Pi_{2,i}\|_{\infty} \\
&\quad + \rho \eta \max_i \|\mathbf{1}_{\mathcal{S}^+} \odot (\nabla_{\mathbf{u}} f_i(\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t) - \nabla_{\mathbf{u}} F(\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t))\|_{\infty} \\
&\leq \rho \|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty} + 4\rho \eta \|\bar{\mathbf{u}}_{\mathcal{S}^+}^t\|_{\infty} \left(\sqrt{s} \delta_{\max} \left(2 \|\bar{\mathbf{v}}_{\mathcal{S}^-}^t\|_{\infty} \|\Delta_{\mathbf{v}, \mathcal{S}^-}^t\|_{\infty} + \|\Delta_{\mathbf{v}, \mathcal{S}^-}^t\|_{\infty}^2 \right) \right. \\
&\quad \left. + 2d \left(2 \|\bar{\mathbf{v}}_{\mathcal{S}^c}^t\|_{\infty} \|\Delta_{\mathbf{v}, \mathcal{S}^c}^t\|_{\infty} + \|\Delta_{\mathbf{v}, \mathcal{S}^c}^t\|_{\infty}^2 + 2 \|\bar{\mathbf{v}}_{\mathcal{S}^+}^t\|_{\infty} \|\Delta_{\mathbf{v}, \mathcal{S}^+}^t\|_{\infty} + \|\Delta_{\mathbf{v}, \mathcal{S}^+}^t\|_{\infty}^2 \right) \right. \\
&\quad \left. + \|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty} (\|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty} + 2 \|\bar{\mathbf{u}}_{\mathcal{S}^+}^t\|_{\infty}) + \sqrt{s} \delta_{\max} \|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty} \left(2 \|\bar{\mathbf{u}}_{\mathcal{S}^+}^t\|_{\infty} + \|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty} \right) \right. \\
&\quad \left. + 2d \left(2 \|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty} \|\Delta_{\mathbf{u}, \mathcal{S}^c}^t\|_{\infty} + \|\Delta_{\mathbf{u}, \mathcal{S}^c}^t\|_{\infty}^2 + 2 \|\bar{\mathbf{u}}_{\mathcal{S}^-}^t\|_{\infty} \|\Delta_{\mathbf{u}, \mathcal{S}^-}^t\|_{\infty} + \|\Delta_{\mathbf{u}, \mathcal{S}^-}^t\|_{\infty}^2 \right) \right. \\
&\quad \left. + 4\rho \eta \|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty} \cdot \left(\left(\|\bar{\mathbf{u}}_{\mathcal{S}^+}^t\|_{\infty}^2 - \mathbf{w}_{\mathcal{S}^+}^* \right)_{\infty} + 2 \|\bar{\mathbf{u}}_{\mathcal{S}^+}^t\|_{\infty} \|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty} + \|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty}^2 \right) \right. \\
&\quad \left. + \sqrt{s} \delta_{\max} \left(\left(\|\bar{\mathbf{u}}_{\mathcal{S}^+}^t\|_{\infty}^2 - \mathbf{w}_{\mathcal{S}^+}^* \right)_{\infty} + \left(\|\bar{\mathbf{v}}_{\mathcal{S}^-}^t\|_{\infty}^2 - \mathbf{w}_{\mathcal{S}^-}^* \right)_{\infty} + 2 \|\bar{\mathbf{u}}_{\mathcal{S}^+}^t\|_{\infty} \|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty} \right. \right. \\
&\quad \left. \left. + \|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty}^2 \right) + 2d \left(\left(\|\bar{\mathbf{u}}_{\mathcal{S}^c}^t\|_{\infty} + \|\Delta_{\mathbf{u}, \mathcal{S}^c}^t\|_{\infty} \right)^2 + \left(\|\bar{\mathbf{v}}_{\mathcal{S}^c}^t\|_{\infty} + \|\Delta_{\mathbf{v}, \mathcal{S}^c}^t\|_{\infty} \right)^2 \right. \right. \\
&\quad \left. \left. + \left(\|\bar{\mathbf{u}}_{\mathcal{S}^-}^t\|_{\infty} + \|\Delta_{\mathbf{u}, \mathcal{S}^-}^t\|_{\infty} \right)^2 + \left(\|\bar{\mathbf{v}}_{\mathcal{S}^+}^t\|_{\infty} + \|\Delta_{\mathbf{v}, \mathcal{S}^+}^t\|_{\infty} \right)^2 \right) + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \right) \\
&\quad + 4\rho \eta \|\bar{\mathbf{u}}_{\mathcal{S}^+}^t\|_{\infty} \cdot \left(\sqrt{s} \delta_{\max} \left(\left(\|\bar{\mathbf{u}}_{\mathcal{S}^+}^t\|_{\infty}^2 - \mathbf{w}_{\mathcal{S}^+}^* \right)_{\infty} + \left(\|\bar{\mathbf{v}}_{\mathcal{S}^-}^t\|_{\infty}^2 - \mathbf{w}_{\mathcal{S}^-}^* \right)_{\infty} \right) \right. \\
&\quad \left. + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} + \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_{\infty} \right), \quad (65)
\end{aligned}$$

where the second inequality is due to (64) and the last inequality is substituting into the formula $\Pi_{1,i}, \Pi_{2,i}$ in (59) and gradient difference at averaged pair. Merging terms involved $\|\bar{\mathbf{u}}_{\mathcal{S}^+}^t\|_{\infty}$ and $\|\Delta_{\mathbf{u}, \mathcal{S}^+}^t\|_{\infty}$ separately would obtain the result in (60). Performing analogous proof would yield results on negative support $\mathcal{S}^-$ in (49) and non-support $\mathcal{S}^c$ in (50). $\square$

The following proposition describes the dynamics of $\mathbf{u}^{t,i}, \mathbf{v}^{t,i}$ by conducting inductive hypothesis for both $\mathbf{u}^{t,i}$ and $\mathbf{v}^{t,i}$, which is different with centralized setting [36]. Because the complicated consensus errors prevent transferring the proof of non-negative case to general case trivially.

Proposition 3. *This proposition inherits Proposition 2 for the general case with considering dynamic of $\mathbf{v}_i^t$ for learning negative part signal $\mathbf{w}_{\mathcal{S}^-}^*$. Recall the definitions of $T, T_k, \bar{T}_k, B_k, K, \beta$ in Proposition 2, the following statements hold in each induction step.*

- (a) $\forall t$ that $\bar{T}_{k-1} \leq t < \bar{T}_k$ with $\forall k \in [K]$, there is $\max \left\{ \left\| \left(\bar{\mathbf{u}}_{\mathcal{S}^+}^t \right)^2 - \mathbf{w}_{\mathcal{S}^+}^* \right\|_{\infty}, \left\| \left(\bar{\mathbf{v}}_{\mathcal{S}^-}^t \right)^2 - \mathbf{w}_{\mathcal{S}^-}^* \right\|_{\infty} \right\} \leq \frac{w_{\max}^*}{2^k}$.

- (b) For $\forall k \in [K]$, there is $\max \left\{ \left\| \left(\bar{\mathbf{u}}_{\mathcal{S}^+}^{\bar{T}_{k-1}} \right)^2 - \mathbf{w}_{\mathcal{S}^+}^* \right\|_{\infty}, \left\| \left(\bar{\mathbf{v}}_{\mathcal{S}^-}^{\bar{T}_{k-1}} \right)^2 - \mathbf{w}_{\mathcal{S}^-}^* \right\|_{\infty} \right\} \leq \frac{w_{\max}^*}{2^k}$.
- (c) $\forall t$ that $\bar{T}_{k-1} \leq t < \bar{T}_k$ with $\forall k \in [K]$, there are $\max_i \left\| \Delta_{\mathbf{u}, \mathcal{S}^+}^{t,i} \right\|_{\infty} \leq 4\beta\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_{\mathcal{S}^+}^t \right\|_{\infty} B_k$ and $\max_i \left\| \Delta_{\mathbf{v}, \mathcal{S}^-}^{t,i} \right\|_{\infty} \leq 4\beta\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{v}}_{\mathcal{S}^-}^t \right\|_{\infty} B_k$.
- (d) For any $\bar{T}_{k-1} \leq t < \bar{T}_k$ that $\forall k \in [K]$, there are $\max_i \left\| \Delta_{\mathbf{u}, \mathcal{S}^c}^{t,i} \right\|_{\infty} \leq 4\beta\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_{\mathcal{S}^c}^t \right\|_{\infty} B_k$ and $\max_i \left\| \Delta_{\mathbf{v}, \mathcal{S}^c}^{t,i} \right\|_{\infty} \leq 4\beta\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{v}}_{\mathcal{S}^c}^t \right\|_{\infty} B_k$.
- (e) For any $\bar{T}_{k-1} \leq t < \bar{T}_k$ that $\forall k \in [K]$, there are $\max_i \left\| \Delta_{\mathbf{u}, \mathcal{S}^-}^{t,i} \right\|_{\infty} \leq 4\beta\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_{\mathcal{S}^-}^t \right\|_{\infty} B_k$ and $\max_i \left\| \Delta_{\mathbf{v}, \mathcal{S}^+}^{t,i} \right\|_{\infty} \leq 4\beta\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{v}}_{\mathcal{S}^+}^t \right\|_{\infty} B_k$.
- (f) $\forall t$ that $\bar{T}_{k-1} \leq t < \bar{T}_k$ with $k \in [K]$, there are refined element-wise bounds for consensus errors as $|\Delta_{\mathbf{u}, j}^t| \leq 4\beta\rho^{\frac{3}{4}}\eta |\bar{u}_j| B_k$ and $|\Delta_{\mathbf{v}, j}^t| \leq 4\beta\rho^{\frac{3}{4}}\eta |\bar{v}_j| B_k, \forall j \in \mathcal{S}$.
- (g) For $\forall j \in \mathcal{S}^+$ and $k \in [K]$, there is $\alpha^3 \leq \left(\bar{\mathbf{u}}_j^{\bar{T}_{k-1}} \right)^2 \leq w_j^* + 4B_k$. For $\forall j \in \mathcal{S}^-$ and $k \in [K]$, there is $\alpha^3 \leq \left(\bar{\mathbf{v}}_j^{\bar{T}_{k-1}} \right)^2 \leq w_j^* + 4B_k$.
- (h) $\forall t \leq \mathcal{O} \left(\frac{1}{\eta\zeta} \log \frac{1}{\alpha} \right)$, $\max \left\{ \left\| \bar{\mathbf{u}}_{\mathcal{S}^c}^t \right\|_{\infty}, \left\| \bar{\mathbf{v}}_{\mathcal{S}^c}^t \right\|_{\infty} \right\} \leq \sqrt{\alpha}$.
- (i) $\forall t \leq \mathcal{O} \left(\frac{1}{\eta\zeta} \log \frac{1}{\alpha} \right)$ and $\forall j \in \mathcal{S}$, there is $\bar{\mathbf{u}}_j^t \bar{\mathbf{v}}_j^t \leq \alpha^{\frac{3}{2}}$.
- (j) $\forall t \leq \mathcal{O} \left(\frac{1}{\eta\zeta} \log \frac{1}{\alpha} \right)$, it has $\max \left\{ \left\| \bar{\mathbf{u}}_{\mathcal{S}^-}^t \right\|_{\infty}, \left\| \bar{\mathbf{v}}_{\mathcal{S}^+}^t \right\|_{\infty} \right\} \leq \sqrt{\alpha}$.

Proof. Proof idea: The key difference in the proof between centralized and decentralized settings lies in the fact that we cannot directly transfer the proof from the simpler non-negative $\mathbf{w}^*$ to the general $\mathbf{w}^*$. This is because the error terms induced in (48), (49), (50) by the decentralized network in Lemma 7 prevent obtaining the results in inductions (i), (j) without induction steps as Lemma B.16 in [36], which allows to apply proof of non-negative case to the general case directly in the centralized case. Therefore, we have to conduct the comprehensive induction process for the general $\mathbf{w}^*$, which can ensure that the magnitudes of $\bar{\mathbf{u}}^t$ on the negative support $\mathcal{S}^-$ and $\bar{\mathbf{v}}^t$ on the positive support $\mathcal{S}^+$ remain small up to the early stopping iteration steps. The inductions (a)-(h) show that the averaged signal $\bar{\mathbf{u}}^t$ on the positive support $\mathcal{S}^+$ and the averaged signal $\bar{\mathbf{v}}^t$ on the negative support exhibit dynamics similar to those in non-negative case, as outlined in Proposition 2.

Base case: As the initialization $\mathbf{u}^{0,i} = \mathbf{v}^{0,i} = \alpha \mathbf{1}_d, \forall i \in [m]$. Due to the condition on α , the base case is true.

Induction Step: Under the assumption that all above induction hypotheses hold until some $0 \leq k \leq K-1$, we should prove they still should hold at $k+1$ -th induction.

(a) We should prove that $\forall t \in \{\bar{T}_{k-1}, \dots, \bar{T}_k - 1\}$, the condition $\left\| \left(\bar{\mathbf{u}}_{\mathcal{S}^+}^t \right)^2 - \mathbf{w}_{\mathcal{S}^+}^* \right\|_{\infty} \leq \frac{w_{\max}^*}{2^k}$ and $\left\| \left(\bar{\mathbf{v}}_{\mathcal{S}^-}^t \right)^2 - \mathbf{w}_{\mathcal{S}^-}^* \right\|_{\infty} \leq \frac{w_{\max}^*}{2^k}$ still holds. If the condition is true for t -th iteration, then based on claims (c), (d), (e), (h) and (j) the $\left\| \mathbf{p}_u^t \right\|_{\infty}$ in (51) under current induction step could be bounded as

$$\begin{aligned} \left\| \mathbf{p}_u^t \right\|_{\infty} &\leq 8\sqrt{s}\delta_{\max}\beta\rho^{\frac{3}{4}}B_k\eta \left(\left\| \bar{\mathbf{u}}_{\mathcal{S}^+}^t \right\|_{\infty}^2 + \left\| \bar{\mathbf{v}}_{\mathcal{S}^-}^t \right\|_{\infty}^2 \right) + \sqrt{s}\delta_{\max} \left(4\beta\rho^{\frac{3}{4}}B_k\eta \left\| \bar{\mathbf{v}}_{\mathcal{S}^-}^t \right\|_{\infty} \right)^2 \\ &\quad + \left(\sqrt{s}\delta_{\max} + 3 \right) \left(4\beta\rho^{\frac{3}{4}}\eta \left\| \bar{\mathbf{u}}_{\mathcal{S}^+}^t \right\|_{\infty} B_k \right)^2 + 2d \left(8\beta\rho^{\frac{3}{4}}\eta B_k + \left(4\beta\rho^{\frac{3}{4}}\eta B_k \right)^2 \right) \\ &\quad \cdot \left(\left\| \bar{\mathbf{u}}_{\mathcal{S}^c}^t \right\|_{\infty}^2 + \left\| \bar{\mathbf{v}}_{\mathcal{S}^c}^t \right\|_{\infty}^2 + \left\| \bar{\mathbf{u}}_{\mathcal{S}^-}^t \right\|_{\infty}^2 + \left\| \bar{\mathbf{v}}_{\mathcal{S}^+}^t \right\|_{\infty}^2 \right) \\ &\leq 32\sqrt{s}\delta_{\max}\beta\rho^{\frac{3}{4}}B_k\eta w_{\max}^* + 32\sqrt{s}\delta_{\max}\beta^2\rho^{\frac{3}{2}}B_k^2\eta^2 w_{\max}^* + 32 \left(\sqrt{s}\delta_{\max} + 3 \right) \beta^2\rho^{\frac{3}{2}}\eta^2 B_k^2 w_{\max}^* \end{aligned}$$

$$\begin{aligned}
& + 2d \left(8\beta\rho^{\frac{3}{4}}\eta B_k + 16\beta^2\rho^{\frac{3}{2}}\eta^2 B_k^2 \right) \left(\|\bar{\mathbf{u}}_{S^c}^t\|_\infty^2 + \|\bar{\mathbf{v}}_{S^c}^t\|_\infty^2 + \|\bar{\mathbf{u}}_{S^-}^t\|_\infty^2 + \|\bar{\mathbf{v}}_{S^+}^t\|_\infty^2 \right) \\
& \leq \frac{3(\sqrt{s}\delta_{\max} + 3)\rho^{\frac{3}{4}}B_k}{16} + \frac{d\alpha\rho^{\frac{3}{4}}}{32}, \tag{66}
\end{aligned}$$

where the first inequality is due to definition of local RIP condition and Lemma 2 and Lemma 3, second inequality is due to hypothesis (a) holds at t -th iteration that $\|\bar{\mathbf{u}}_{S^+}^t\|_\infty^2 \leq 2w_{\max}^*$, $\|\bar{\mathbf{v}}_{S^-}^t\|_\infty^2 \leq 2w_{\max}^*$, the last inequality is due to definitions of β, B_k , step size condition and hypothesis that $\|\bar{\mathbf{u}}_{S^c}^t\|_\infty, \|\bar{\mathbf{v}}_{S^c}^t\|_\infty, \|\bar{\mathbf{u}}_{S^-}^t\|_\infty, \|\bar{\mathbf{v}}_{S^+}^t\|_\infty$ would keep below $\sqrt{\alpha}$ up to early stopping under hypothesis (h) and (j).

The element-wise bound for $\mathbf{q}_u^t$ in (52) under current induction step can be bounded as follows that $\forall j \in S^+$, there is

$$\begin{aligned}
|q_{u,j}^t| & \leq 4\beta\rho^{\frac{3}{4}}\eta|\bar{u}_j^t|B_k \left(\sqrt{s}\delta_{\max} \left(\left\| (\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* \right\|_\infty + \left\| (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right\|_\infty \right) \right. \\
& \quad \left. + \frac{3(\sqrt{s}\delta_{\max} + 3)\rho^{\frac{3}{4}}B_k}{16} + 8d \left(\|\bar{\mathbf{u}}_{S^c}^t\|_\infty^2 + \|\bar{\mathbf{v}}_{S^c}^t\|_\infty^2 + \|\bar{\mathbf{u}}_{S^-}^t\|_\infty^2 + \|\bar{\mathbf{v}}_{S^+}^t\|_\infty^2 \right) + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) \\
& \leq 4\beta\rho^{\frac{1}{2}}\eta|\bar{u}_j^t|B_k \left(\rho^{\frac{1}{4}}\sqrt{s}\delta_{\max} \left(\left\| (\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* \right\|_\infty + \left\| (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right\|_\infty \right) \right. \\
& \quad \left. + \frac{3(\sqrt{s}\delta_{\max} + 3)\rho B_k}{16} + 32d\rho^{\frac{1}{4}}\alpha + \rho^{\frac{1}{4}}\max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) \\
& \leq \frac{\rho^{\frac{1}{2}}B_k|\bar{u}_j^t|}{32}, \tag{67}
\end{aligned}$$

where the first inequality is based on refined bound in hypothesis (c), result in (66) and inequality $\|\mathbf{x} + \mathbf{y}\|_\infty^2 \leq 2\|\mathbf{x}\|_\infty^2 + 2\|\mathbf{y}\|_\infty^2$ and step size condition that $4\beta\rho^{\frac{3}{4}}\eta B_k < 1$. The second inequality uses hypothesis (h) and (j). The last inequality is because of hypothesis (a), the step size condition, network connectivity condition, small initialization α and definition of B_k .

Then $\mathbf{q}_u^t$ could be reparameterized as $\mathbf{q}_u^t = \mathbf{r}_{q_u}^t \odot \bar{\mathbf{u}}_{S^+}^t$ where $\|\mathbf{r}_{q_u}^t\|_\infty \leq \frac{\sqrt{\rho}B_k}{32}, \forall t, \bar{T}_{k-1} \leq t < \bar{T}_k$. Then the perturbed recursion for $\bar{\mathbf{u}}^t$ on positive support S^+ over decentralized network in (48) becomes

$$(\bar{\mathbf{u}}_{S^+}^{t+1})^2 = (\bar{\mathbf{u}}_{S^+}^t)^2 \odot \left(\mathbf{1}_d - 4\eta \left((\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* + \mathbf{E}_{2s}^t + \mathbf{E}_{3s}^t + \mathbf{p}_u^t + \mathbf{r}_{q_u}^t \right) \right)^2, \tag{68}$$

where the new perturbed error terms $\mathbf{E}_{2s}^t, \mathbf{E}_{3s}^t$ are defined as

$$\begin{aligned}
\mathbf{E}_{2s}^t & := \left(\frac{\mathbf{X}^T \mathbf{X}}{N} - \mathbf{I} \right) \left((\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* - (\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right) \\
\mathbf{E}_{3s}^t & := \frac{\mathbf{X}^T \mathbf{X}}{N} \left((\bar{\mathbf{u}}_{S^c}^t)^2 - (\bar{\mathbf{v}}_{S^+}^t)^2 + (\bar{\mathbf{u}}_{S^-}^t)^2 - (\bar{\mathbf{v}}_{S^c}^t)^2 \right) - \frac{\mathbf{X}^T \boldsymbol{\xi}}{N}. \tag{69}
\end{aligned}$$

Then the total error terms in (68) can be bounded as follows

$$\begin{aligned}
\|\mathbf{E}_{2s}^t\|_\infty + \|\mathbf{E}_{3s}^t\|_\infty + \|\mathbf{p}_u^t\|_\infty + \|\mathbf{r}_{q_u}^t\|_\infty & \stackrel{(i)}{\leq} 2\sqrt{s}\delta \frac{w_{\max}^*}{2^k} + 8d\alpha + \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_\infty \\
& \quad + \frac{3(\sqrt{s}\delta_{\max} + 3)\rho^{\frac{3}{4}}B_k}{16} + \frac{d\alpha\rho^{\frac{3}{4}}}{32} + \frac{\rho^{\frac{1}{2}}B_k}{32} \\
& \stackrel{(ii)}{\leq} C_\gamma \frac{w_{\max}^*}{2^k} + C_b \cdot \frac{2}{C_b} \max \left\{ \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_\infty, \epsilon \right\} + \frac{B_k}{2} \\
& \stackrel{(iii)}{\leq} C_\gamma \frac{w_{\max}^*}{2^k} + C_b \zeta + \frac{B_k}{2} \\
& \stackrel{(iv)}{\leq} (C_\gamma + 2C_b) \frac{w_{\max}^*}{2^k} + \frac{B_k}{2}
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(v)}{\leq} \frac{1}{2} \cdot \frac{1}{40} \frac{w_{\max}^*}{2^k} + \frac{B_k}{2} \\
&\stackrel{(vi)}{\leq} B_k,
\end{aligned} \tag{70}$$

where (i) is because of Lemma 2 and Lemma 3 under the global RIP condition and induction hypothesis (a) and substituting the results in (66) and (67). (ii) is due to condition on global RIP parameter that $\delta \leq \frac{C_\gamma}{2\sqrt{s}(\log \lceil \frac{w_{\max}^*}{\zeta} \rceil + 1)}$, condition $\alpha \leq \frac{\zeta}{(3d+1)^2}$ and network connectivity condition $\rho \leq \frac{1}{36(\sqrt{s}\delta_{\max}+3)^{\frac{4}{3}}}$. (iii) is due to definition of ζ and (iv) is based on definition of K that $\zeta \leq \frac{w_{\max}^*}{2^{K-1}} \leq \frac{w_{\max}^*}{2^{k-1}}$. (v) is because definition of B_k and C_γ, C_b are determined later that are small enough that $C_\gamma + 2C_b \leq \frac{1}{80}$.

Based on iteration (68) and perturbation bound in (70), we would prove the $\|(\bar{\mathbf{u}}_{S^+}^{t+1})^2 - \mathbf{w}_{S^+}^*\|_\infty \leq \frac{w_{\max}^*}{2^k}$ should be also true based on the proof of hypothesis (a) in Proposition 2 by replacing $\mathcal{S}$ with $\mathcal{S}^+$. We can obtain similar results for $\bar{\mathbf{v}}_{S^-}^t$. Combined two cases and induction on t , we have conclusion that $\forall t \geq \bar{T}_{k-1}$, there is

$$\max \left\{ \|(\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^*\|_\infty, \|(\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^*\|_\infty \right\} \leq \frac{w_{\max}^*}{2^k}. \tag{71}$$

Thus, we finish the proof of induction (a).

(b) Comparing the (68), (70) with (30), the proof of this hypothesis can follow the proof of hypothesis (b) in Proposition 2.

(c) $\forall t$ that $\bar{T}_{k-1} \leq t < t+1 < \bar{T}_k$, based on one step iteration in (60) and induction (a), there is

$$\begin{aligned}
\|\Delta_{\mathbf{u}, S^+}^{t+1}\|_\infty &\stackrel{(i)}{\leq} \rho \|\Delta_{\mathbf{u}, S^+}^t\|_\infty \cdot \left(1 + 4\eta \left(2(\sqrt{s}\delta_{\max} + 1) w_{\max}^* + 8d \left(\|\bar{\mathbf{u}}_{S^c}^t\|_\infty^2 + \|\bar{\mathbf{v}}_{S^c}^t\|_\infty^2 \right. \right. \right. \\
&\quad \left. \left. + \|\bar{\mathbf{u}}_{S^-}^t\|_\infty^2 + \|\bar{\mathbf{v}}_{S^+}^t\|_\infty^2 \right) + 8(\sqrt{s}\delta_{\max} + 1) \|\bar{\mathbf{u}}_{S^+}^t\|_\infty^2 + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) \\
&\quad + 4\rho\eta \|\bar{\mathbf{u}}_{S^+}^t\|_\infty \cdot \left(4\sqrt{s}\delta_{\max}\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{v}}_{S^-}^t\|_\infty B_k \cdot \left(2\|\bar{\mathbf{v}}_{S^-}^t\|_\infty + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{v}}_{S^-}^t\|_\infty B_k \right) \right. \\
&\quad \left. + 4\sqrt{s}\delta_{\max} \frac{w_{\max}^*}{2^k} + 8d \left(\|\bar{\mathbf{u}}_{S^c}^t\|_\infty^2 + \|\bar{\mathbf{v}}_{S^c}^t\|_\infty^2 + \|\bar{\mathbf{u}}_{S^-}^t\|_\infty^2 + \|\bar{\mathbf{v}}_{S^+}^t\|_\infty^2 \right) \right. \\
&\quad \left. + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty + \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_\infty \right) \\
&\stackrel{(ii)}{\leq} \rho^{\frac{3}{4}} \|\Delta_{\mathbf{u}, S^+}^t\|_\infty \cdot \left(1 + 4\eta \left(18\rho^{\frac{1}{4}} w_{\max}^* (\sqrt{s}\delta_{\max} + 1) + 32\rho^{\frac{1}{4}} d\alpha \right. \right. \\
&\quad \left. \left. + \rho^{\frac{1}{4}} \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) + 4\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^+}^t\|_\infty \cdot \left(16\sqrt{s}\delta_{\max}\beta\rho\eta \|\bar{\mathbf{v}}_{S^-}^t\|_\infty^2 B_k \right. \right. \\
&\quad \left. \left. + 4\rho^{\frac{1}{4}}\sqrt{s}\delta_{\max} \frac{w_{\max}^*}{2^k} + 32\rho^{\frac{1}{4}} d\alpha + 2\rho^{\frac{1}{4}} \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) \right) \\
&\stackrel{(iii)}{\leq} \rho^{\frac{3}{4}} \|\Delta_{\mathbf{u}, S^+}^t\|_\infty \cdot \left(1 + 80\eta w_{\max}^* \right) + 4\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^+}^t\|_\infty \cdot \left(16\sqrt{s}\delta_{\max}\beta\rho\eta \|\bar{\mathbf{v}}_{S^-}^t\|_\infty^2 B_k + \frac{3}{4} B_k \right) \\
&\stackrel{(iv)}{\leq} 4 \left(\frac{1 + \sqrt{\rho}}{2} \right) \beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^+}^t\|_\infty B_k + 4\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^+}^t\|_\infty B_k \\
&\stackrel{(v)}{\leq} \frac{4}{(1 - c_{10}(1 - \sqrt{\rho}))^2} \left(\left(\frac{1 + \sqrt{\rho}}{2} \right) \beta + 1 \right) \rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^+}^{t+1}\|_\infty B_k,
\end{aligned} \tag{72}$$

where (i) is due to the claim (a) and step size condition that $4\beta\rho^{\frac{3}{4}}\eta B_k \leq 1$ and $\delta \leq \delta_{\max}$. (ii) is due to $4\beta\rho^{\frac{3}{4}}\eta B_k \leq 1$, condition that all $\|\bar{\mathbf{u}}_{S^c}^t\|_\infty, \|\bar{\mathbf{v}}_{S^c}^t\|_\infty, \|\bar{\mathbf{u}}_{S^+}^t\|_\infty, \|\bar{\mathbf{v}}_{S^+}^t\|_\infty$ are not larger than $\sqrt{\alpha}$ before early stopping in hypothesis (h), (j) and $\left\|\frac{\mathbf{X}^T \boldsymbol{\xi}}{N}\right\|_\infty \leq \max_i \left\|\frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n}\right\|_\infty$.

The (iii) is because $\rho^{\frac{1}{4}} \leq \min \left\{ \frac{1}{\sqrt{s}\delta_{\max}+1}, \frac{\delta}{8\delta_{\max}}, \frac{\left\|\frac{\mathbf{X}^T \boldsymbol{\xi}}{N}\right\|_\infty}{8 \max_i \left\|\frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n}\right\|_\infty} \right\}$, $\alpha \leq \frac{\epsilon}{(12d+1)^2}$ and $\max \left\{ \left\|\frac{\mathbf{X}^T \boldsymbol{\xi}}{N}\right\|_\infty, \epsilon \right\} \leq w_{\max}^*$. The (iv) is because in the current induction step hypothesis (c) holds in t -th iteration and step size condition that satisfies $\eta \leq \frac{1-\sqrt{\rho}}{160\rho^{\frac{1}{4}}w_{\max}^*}$ such that $\rho^{\frac{1}{2}} \left(\rho^{\frac{1}{4}} + 80\rho^{\frac{1}{4}}\eta w_{\max}^* \right) \leq (1 - (1 - \rho^{\frac{1}{2}}))(1 + \frac{1-\sqrt{\rho}}{2}) \leq \frac{1+\sqrt{\rho}}{2}$ and $16\sqrt{s}\delta_{\max}\beta\rho\eta \|\bar{\mathbf{v}}_{S^+}^t\|_\infty^2 \leq \frac{B_k}{4}$ where $\|\bar{\mathbf{v}}_{S^+}^t\|_\infty^2 \leq 2w_{\max}^*$ is according to claim (a). The last inequality (v) is based on the (68) and step size condition $\eta \leq \frac{c_{10}(1-\sqrt{\rho})}{w_{\max}^*}$ such that $4\eta \left(\left(\bar{\mathbf{u}}_{S^+}^t \right)^2 - w_{S^+}^* \right)_\infty + B_k \leq 1$ based on hypothesis (a). Comparing the formula of (72) with (34), we can follow the poof of hypothesis (c) in Proposition 2 to finish the proof.

(d) The proof is similar to that of hypothesis (c). $\forall t$ that $\bar{T}_{k-1} \leq t < t+1 \leq \bar{T}_k$, according to one step iteration in (49) and hypothesis (a), we have

$$\begin{aligned}
\left\| \boldsymbol{\Delta}_{\mathbf{u}, S^c}^{t+1} \right\|_\infty &\leq \rho \left\| \boldsymbol{\Delta}_{\mathbf{u}, S^c}^t \right\|_\infty \cdot \left(1 + 4\eta \left(\sqrt{s}\delta_{\max} \left(2w_{\max}^* + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^+}^t\|_\infty B_k \left(2\|\bar{\mathbf{u}}_{S^+}^t\|_\infty \right. \right. \right. \right. \\
&\quad \left. \left. \left. + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^+}^t\|_\infty B_k \right) + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{v}}_{S^+}^t\|_\infty B_k \left(2\|\bar{\mathbf{v}}_{S^+}^t\|_\infty + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{v}}_{S^+}^t\|_\infty B_k \right) \right) \right. \\
&\quad \left. + 64d\alpha + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) + 4\rho\eta \|\bar{\mathbf{u}}_{S^c}^t\|_\infty \cdot \left(\sqrt{s}\delta_{\max} \left(4\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^+}^t\|_\infty B_k \right. \right. \\
&\quad \cdot \left(2\|\bar{\mathbf{u}}_{S^+}^t\|_\infty + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^+}^t\|_\infty B_k \right) + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{v}}_{S^+}^t\|_\infty B_k \cdot \left(2\|\bar{\mathbf{v}}_{S^+}^t\|_\infty \right. \\
&\quad \cdot \left. \left. + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{v}}_{S^+}^t\|_\infty B_k \right) \right) + 4\sqrt{s}\delta_{\max} \frac{w_{\max}^*}{2^k} + 32d\alpha + 2 \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \Bigg) \\
&\leq \rho^{\frac{3}{4}} \left\| \boldsymbol{\Delta}_{\mathbf{u}, S^c}^t \right\|_\infty \cdot \left(1 + 4\eta \left(\rho^{\frac{1}{4}} \sqrt{s}\delta_{\max} \left(2w_{\max}^* + 16\beta\rho^{\frac{3}{4}}\eta B_k \left(\|\bar{\mathbf{u}}_{S^+}^t\|_\infty^2 + \|\bar{\mathbf{v}}_{S^+}^t\|_\infty^2 \right) \right) \right. \right. \\
&\quad \left. \left. + \frac{4}{9}\zeta + \rho^{\frac{1}{4}} \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) \right) + 4\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^c}^t\|_\infty \cdot \left(16\sqrt{s}\delta_{\max}\beta\rho\eta B_k \left(\|\bar{\mathbf{u}}_{S^+}^t\|_\infty^2 + \|\bar{\mathbf{v}}_{S^+}^t\|_\infty^2 \right) \right. \\
&\quad \left. + 4\rho^{\frac{1}{4}} \sqrt{s}\delta_{\max} \frac{w_{\max}^*}{2^k} + \frac{\zeta}{4} + 2\rho^{\frac{1}{4}} \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{N} \right\|_\infty \right) \\
&\leq 4 \left(\frac{1+\sqrt{\rho}}{2} \right) \beta\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^c}^t\|_\infty B_k + 4\rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^c}^t\|_\infty B_k \\
&\leq \frac{4}{(1-c_{10}(1-\sqrt{\rho}))^2} \left(\left(\frac{1+\sqrt{\rho}}{2} \right) \beta + 1 \right) \rho^{\frac{3}{4}}\eta \|\bar{\mathbf{u}}_{S^c}^{t+1}\|_\infty B_k, \tag{73}
\end{aligned}$$

where the first two inequalities have the same reasons in (72). The third inequality is due to $16\beta\rho^{\frac{3}{4}}B_k \left(\|\bar{\mathbf{u}}_{S^+}^t\|_\infty^2 + \|\bar{\mathbf{v}}_{S^+}^t\|_\infty^2 \right) \leq \frac{8}{5}w_{\max}^*$ and $\rho^{\frac{1}{4}}\sqrt{s}\delta_{\max} \leq \frac{\sqrt{s}\delta}{8} \leq \frac{1}{640}$ due to upper bound of global RIP condition in Proposition 1 with $C_\gamma \leq \frac{1}{80}$. The last inequality is due to (79). Then we can use analogized proof for hypothesis (c) to finish the proof.

(e) The proof is similar to hypothesis (d). $\forall t$ that $\bar{T}_{k-1} \leq t < t+1 \leq \bar{T}_k$, based on one step iteration in (50) and hypothesis (a), there is

$$\begin{aligned} \|\Delta_{u,S^-}^{t+1}\|_\infty &\leq \rho \|\Delta_{u,S^-}^t\|_\infty \cdot \left(1 + 4\rho\eta \left((\sqrt{s}\delta_{\max} + 1) \left(2w_{\max}^* + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{u}_{S^+}^t\|_\infty B_k (2\|\bar{u}_{S^+}^t\|_\infty \right. \right. \right. \\ &\quad \left. \left. + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{u}_{S^+}^t\|_\infty B_k \right) + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{v}_{S^-}^t\|_\infty B_k \left(2\|\bar{v}_{S^-}^t\|_\infty + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{v}_{S^-}^t\|_\infty B_k \right) \right) \\ &\quad \left. + 64d\alpha + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) + 4\rho\eta \|\bar{u}_{S^-}^t\|_\infty \cdot \left((\sqrt{s}\delta_{\max} + 1) \left(4\beta\rho^{\frac{3}{4}}\eta \|\bar{u}_{S^+}^t\|_\infty B_k \right. \right. \\ &\quad \cdot \left(2\|\bar{u}_{S^+}^t\|_\infty + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{u}_{S^+}^t\|_\infty B_k \right) + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{v}_{S^-}^t\|_\infty B_k (2\|\bar{v}_{S^-}^t\|_\infty \\ &\quad \left. \left. + 4\beta\rho^{\frac{3}{4}}\eta \|\bar{v}_{S^-}^t\|_\infty B_k \right) \right) + 2\sqrt{s}(\delta + \delta_{\max}) \frac{w_{\max}^*}{2^k} + 32d\alpha + 2 \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) \\ &\leq \frac{4}{(1 - c_{10}(1 - \sqrt{\rho}))^2} \left(\left(\frac{1 + \sqrt{\rho}}{2} \right) \beta + 1 \right) \rho^{\frac{3}{4}}\eta \|\bar{u}_{S^-}^{t+1}\|_\infty B_k, \end{aligned} \quad (74)$$

where the last inequality and the left proof have analogized derivations in proof of hypothesis (d).

(f) To absorb the perturbed error $\mathbf{q}_u^t$ in (48), we need to prove the fine-grained bound for consensus error in hypotheses (c), (d), (d). The idea is that we focus on the $\forall j \in \mathcal{S}^+$, then the upper bound on the consensus error on j -th entry is

$$\begin{aligned} |\Delta_{u,j}^{t+1}| &\leq \rho |\Delta_{u,j}^t| \cdot \left(1 + 4\eta \left(\sqrt{s}\delta_{\max} \left(\left\| (\bar{u}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* \right\|_\infty + \left\| -(\bar{v}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right\|_\infty \right. \right. \right. \\ &\quad \left. \left. + 2\|\bar{u}_{S^+}^t\|_\infty \|\Delta_{u,S^+}^t\|_\infty + \|\Delta_{u,S^+}^t\|_\infty^2 + 2\|\bar{v}_{S^-}^t\|_\infty \|\Delta_{v,S^-}^t\|_\infty + \|\Delta_{v,S^-}^t\|_\infty^2 \right) \right. \\ &\quad \left. + |(\bar{u}_j^t)^2 - w_j^*| + 2(\bar{u}_j^t)^2 + 3|\bar{u}_j^t| \|\Delta_{u,j}^t\|_\infty + \|\Delta_{u,j}^t\|_\infty^2 + 24d\alpha + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) \\ &\quad + 4\rho\eta |\bar{u}_j^t| \cdot \left(\sqrt{s}\delta_{\max} \left(2\|\bar{u}_{S^+}^t\|_\infty \|\Delta_{u,S^+}^t\|_\infty + \|\Delta_{u,S^+}^t\|_\infty^2 + 2\|\bar{v}_{S^-}^t\|_\infty \|\Delta_{v,S^-}^t\|_\infty \right. \right. \\ &\quad \left. \left. + \|\Delta_{v,S^-}^t\|_\infty^2 \right) + \sqrt{s}(\delta_{\max} + \delta) \left(\left\| (\bar{u}_{S^+}^t)^2 - \mathbf{w}_{S^+}^* \right\|_\infty + \left\| -(\bar{v}_{S^-}^t)^2 - \mathbf{w}_{S^-}^* \right\|_\infty \right) \right. \\ &\quad \left. + 40d\alpha + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty + \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_\infty \right) \\ &\leq \frac{4}{(1 - c_{10}(1 - \sqrt{\rho}))^2} \left(\left(\frac{1 + \sqrt{\rho}}{2} \right) \beta + 1 \right) \rho^{\frac{3}{4}}\eta B_k |\bar{u}_j^{t+1}|, \end{aligned} \quad (75)$$

where the last inequality is due to substituting the crude bound and analogous derivation in (72). Following the proof of crude bound would finish the proof. The proof for $\forall j \in \mathcal{S}^-$ and $\forall j \in \mathcal{S}^c$ can combine this proof with proofs of hypotheses (e) and (d), respectively. The fine-grained bound for v^t can be proved by using analogous proofs for u^t .

(g) The proof can follow the proof of hypothesis (e) in Proposition 2 by replacing the $\mathcal{S}$ with $\mathcal{S}^+$. The proof of $\bar{v}_{S^-}^t$ is analogous to $\bar{u}_{S^+}^t$.

(h) $\forall t$ that $\bar{T}_{k-1} \leq t < t+1 < \bar{T}_k$. First, we bound the perturbation error $\|\mathbf{g}_u^t\|_\infty, \|\mathbf{f}_u^t\|_\infty$ in (50) induced from decentralized network

$$\begin{aligned} \|\mathbf{g}_u^t\|_\infty &\leq \|\mathbf{p}_u^t\|_\infty + 3\sqrt{s}\delta_{\max} \left(4\beta\rho^{\frac{3}{4}}\eta \|\bar{u}_{S^+}^t\|_\infty B_k \right)^2 \\ &\leq \frac{3(\sqrt{s}\delta_{\max} + 3)\rho^{\frac{3}{4}}B_k}{16} + \frac{d\alpha\rho^{\frac{3}{4}}}{32} + 96\sqrt{s}\delta_{\max}\rho^{\frac{3}{2}}\beta^2\eta^2w_{\max}^*B_k^2 \end{aligned}$$

$$\leq \frac{\rho^{\frac{3}{4}} B_k (150 (\sqrt{s} \delta_{\max} + 3) + 3\sqrt{\rho})}{800} + \frac{d\alpha \rho^{\frac{3}{4}}}{32}, \quad (76)$$

where the second inequality is substituting the upper bound in (66) and $\|\bar{\mathbf{u}}_{S^+}^t\|_{\infty} \leq 2w_{\max}^*$ due to hypothesis (a). $\forall j \in S^c$, there is

$$\begin{aligned} |f_j^t| &\leq 4\beta \rho^{\frac{3}{4}} \eta |\bar{u}_j^t| B_k \cdot \left(\sqrt{s} \delta_{\max} \cdot \left(80B_k + 16\beta \rho^{\frac{3}{4}} \eta B_k \left(\|\bar{\mathbf{u}}_{S^+}^t\|_{\infty}^2 + \|\bar{\mathbf{v}}_{S^-}^t\|_{\infty}^2 \right) \right) \right. \\ &\quad \left. + 32d\alpha + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \right) \\ &\leq \frac{\sqrt{\rho} B_k |\bar{u}_j^t|}{32}, \end{aligned} \quad (77)$$

where the first inequality is due to hypotheses (a), (c)-(f), (h), (g) and step size condition. The second inequality is due to the definition of B_k , the condition on ρ , and small initialization for α .

Then $\mathbf{f}_u^t$ could be reparameterized as $\mathbf{f}_u^t = \mathbf{r}_{f_u}^t \odot \bar{\mathbf{u}}_{S^c}^t$, where $\|\mathbf{r}_{f_u}^t\|_{\infty} \leq \frac{\sqrt{\rho} B_k}{32}, \forall t$ that $\bar{T}_{k-1} \leq t < \bar{T}_k$. Then the perturbed recursion for $\bar{\mathbf{u}}^t$ on non-support S^c over decentralized network in (50) becomes

$$(\bar{\mathbf{u}}_{S^c}^{t+1})^2 = (\bar{\mathbf{u}}_{S^c}^t)^2 \odot (\mathbf{1}_d - 4\eta (\mathbf{E}_{2s}^t + \mathbf{E}_{3s}^t + \mathbf{g}_u^t + \mathbf{r}_{f_u}^t))^2. \quad (78)$$

Thus, we can conduct similar calculation in (70) and obtain the upper bound $\|\mathbf{E}_{2s}^t\|_{\infty} + \|\mathbf{E}_{3s}^t\|_{\infty} + \|\mathbf{g}_u^t\|_{\infty} + \|\mathbf{r}_{f_u}^t\|_{\infty} \leq B_k$. Based on the step size condition that $\eta \leq \frac{c_{10}(1-\sqrt{\rho})}{w_{\max}^*}$, therefore there is a similar result in (35) as

$$\|\bar{\mathbf{u}}_{S^c}^{t+1}\|_{\infty} \geq \|\bar{\mathbf{u}}_{S^c}^t\|_{\infty} (1 - c_{10} (1 - \sqrt{\rho}))^2. \quad (79)$$

Based on the recursion in (78), $\forall k$ -th induction stage, we have the following upper bound

$$\begin{aligned} &\prod_{t=0}^{\bar{T}_k-1} \left(1 + 8\eta \left(\|\mathbf{E}_{2s}^t\|_{\infty} + \|\mathbf{E}_{3s}^t\|_{\infty} + \|\mathbf{g}_u^t\|_{\infty} + \|\mathbf{r}_{f_u}^t\|_{\infty} \right) \right)^2 \\ &\stackrel{(i)}{\leq} \prod_{t=0}^{\bar{T}_k-1} (1 + 8\eta \cdot 4C_b \zeta)^2 \left(1 + 8\eta \cdot \frac{C_{\gamma}}{K} \left(\|(\bar{\mathbf{u}}_{S^+}^t)^2 - \mathbf{w}_{S^+}^*\|_{\infty} + \|(\bar{\mathbf{v}}_{S^-}^t)^2 - \mathbf{w}_{S^-}^*\|_{\infty} \right) \right)^2 \\ &\quad \cdot \left(1 + 8\eta \left(\frac{\rho^{\frac{3}{4}} B_k (150 (\sqrt{s} \delta_{\max} + 3) + 3\sqrt{\rho})}{800} + \frac{d\alpha \rho^{\frac{3}{4}}}{32} \right) \right)^2 + (1 + 8\eta \|\mathbf{r}_{f_u}^t\|_{\infty})^2 \\ &\stackrel{(ii)}{\leq} (1 + 32\eta C_b \zeta)^{4T_k} \left(1 + 4\eta \cdot \frac{2C_{\gamma} w_{\max}^*}{K \times 2^k} \right)^{4(k+1)T_k} \cdot \prod_{t=0}^{\bar{T}_k-1} \left(1 + 8\eta \cdot \frac{\rho^{\frac{3}{4}} (\sqrt{s} \delta_{\max} + 4)}{200} \cdot \frac{w_{\max}^*}{2^k} \right)^2 \\ &\quad \cdot \left(1 + 8\eta \rho^{\frac{3}{4}} \sqrt{s} \delta_{\max} B_k \right)^2 \left(1 + 8\eta \cdot \rho^{\frac{3}{4}} \left(32d\alpha + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \right) \right)^2 \\ &\stackrel{(iii)}{\leq} (1 + 32\eta C_b \zeta)^{8T_k} \left(1 + 4\eta \cdot \frac{2C_{\gamma} w_{\max}^*}{K \times 2^k} \right)^{12(k+1)T_k} \\ &\stackrel{(iv)}{\leq} (1 + 32\eta C_b \zeta)^{8T_{K-1}} \left(1 + 4\eta \cdot \frac{2C_{\gamma} w_{\max}^*}{K \times 2^{K-1}} \right)^{12KT_{K-1}} \\ &\stackrel{(v)}{\leq} \left(1 + 4\eta \cdot \frac{8C_b \zeta}{K} \right)^{8KT_{K-1}} \left(1 + 4\eta \cdot \frac{2C_{\gamma} w_{\max}^*}{K \times 2^{K-1}} \right)^{12KT_{K-1}} \\ &\stackrel{(vi)}{\leq} \left(1 + 4\eta \cdot \frac{2C_{\gamma} w_{\max}^*}{K \times 2^{K-1}} \right)^{20KT_{K-1}} \\ &\leq \frac{1}{\alpha} \end{aligned} \quad (80)$$

where (i) is due to $\forall x_1, x_2, \dots, x_n \geq 0$, there is $1 + \sum_{i=1}^n x_i \leq \prod_{i=1}^n (1 + x_i)$, substituting the upper bound in (76) and definition of global RIP parameter. For (ii), we use $\bar{T}_k \leq 2T_k$ based on definition of $\bar{T}_k, T_k$, induction (a) with Lemma B.13 in [36], step size condition, small initialization for α , definition of B_k and refined upper bound for $\|\mathbf{r}_{f_u}^t\|_\infty$ as $\|\mathbf{f}_u^t\|_\infty \leq \rho^{\frac{3}{4}} \left(\sqrt{s} \delta_{\max} B_k + 32d\alpha + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \right) \|\bar{\mathbf{u}}_{S^c}^t\|_\infty$. (iii) is because network connectivity condition that $\rho \leq \min \left\{ \left(\frac{\sqrt{s} \delta}{\sqrt{s} \delta_{\max} + 4} \right)^{\frac{4}{3}}, \frac{\|\mathbf{X}^T \boldsymbol{\xi}\|_\infty}{8m \max_i \|\mathbf{X}_i^T \boldsymbol{\xi}_i\|_\infty} \right\}$. (iv), (v) are due to condition $i \leq K - 1$ and inequality $1 + \sum_{i=1}^K x_i \leq \prod_{i=1}^K (1 + x_i)$ that $x_1 = \dots, x_K = \frac{32\eta C_b \zeta}{K}$. (vi) is based on the definition of K that $\zeta \leq \frac{w_{\max}^*}{2^{K-1}}$ and condition that $4C_b \leq C_\gamma$. The last inequality results from the Lemma A.2 in [36] with upper bound is set as $B = \frac{2C_\gamma w_{\max}^*}{K \times 2^{K-1}}$ and initialization $x_0 = \alpha^2$, then for $t \leq \frac{K \times 2^{K-1}}{64\eta C_\gamma w_{\max}^*} \log \frac{1}{\alpha^4}$, there is

$$\left(1 + 4\eta \cdot \frac{2C_\gamma w_{\max}^*}{K \times 2^{K-1}} \right)^t \leq \frac{1}{\alpha} \quad (81)$$

Thus, we select universal constant $C_\gamma \leq \frac{1}{1280}$ such that $\frac{K \times 2^{K-1}}{64\eta C_\gamma w_{\max}^*} \log \frac{1}{\alpha^4} \geq 20K \times T_{K-1}$, which would lead to the last inequality.

(i) Concentrate on the iteration by DGD algorithm on support $\mathcal{S}$, based on (57), we have the recursive formula as

$$\begin{aligned} \bar{\mathbf{u}}_S^{t+1} &= \bar{\mathbf{u}}_S^t \odot \left(\mathbf{1}_d - 4\eta \left((\bar{\mathbf{u}}_S^t)^2 - (\bar{\mathbf{v}}_S^t)^2 - \mathbf{w}^* \right) + \mathbf{E}_{d1}^t + \mathbf{E}_{d2}^t + \mathbf{E}_{d3}^t \right) \\ &\quad - \frac{4\eta}{m} \sum_{i=1}^m \Delta_{\mathbf{u}, \mathcal{S}}^{t,i} \odot \mathbf{E}_{d4}^{t,i}, \end{aligned} \quad (82)$$

where the $\mathbf{E}_{d1}^t, \mathbf{E}_{d2}^t, \mathbf{E}_{d3}^t$ and $\mathbf{E}_{d4}^{t,i}$ are defined as

$$\begin{aligned} \mathbf{E}_{d1}^t &:= \left(\frac{\mathbf{X}^T \mathbf{X}}{N} - \mathbf{I} \right) \left((\bar{\mathbf{u}}_S^t)^2 - (\bar{\mathbf{v}}_S^t)^2 - \mathbf{w}^* \right); \\ \mathbf{E}_{d2}^t &:= \frac{\mathbf{X}^T \mathbf{X}}{N} \left((\bar{\mathbf{u}}_{S^c}^t)^2 - (\bar{\mathbf{v}}_{S^c}^t)^2 \right) - \frac{\mathbf{X}^T \boldsymbol{\xi}}{N}; \\ \mathbf{E}_{d3}^t &:= - \sum_{i=1}^m \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left(\left((\bar{\mathbf{u}}_S^t)^2 - (\mathbf{u}_S^{t,i})^2 + (\mathbf{v}_S^{t,i})^2 - (\bar{\mathbf{v}}_S^t)^2 \right) \right. \\ &\quad \left. + \left((\bar{\mathbf{u}}_{S^c}^t)^2 - (\mathbf{u}_{S^c}^{t,i})^2 + (\mathbf{v}_{S^c}^{t,i})^2 - (\bar{\mathbf{v}}_{S^c}^t)^2 \right) \right); \\ \mathbf{E}_{d4}^{t,i} &:= (\mathbf{u}_S^{t,i})^2 - (\mathbf{v}_S^{t,i})^2 + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \left((\mathbf{u}_S^{t,i})^2 - (\mathbf{v}_S^{t,i})^2 - \mathbf{w}^* \right) \\ &\quad + \frac{\mathbf{X}_i^T \mathbf{X}_i}{n} \left((\mathbf{u}_{S^c}^{t,i})^2 - (\mathbf{v}_{S^c}^{t,i})^2 \right) - \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n}. \end{aligned} \quad (83)$$

For the negative part $\bar{\mathbf{v}}^t$, with similar decomposition in (57), there is

$$\begin{aligned} \bar{\mathbf{v}}^{t+1} &= \bar{\mathbf{v}}^t - \frac{\eta}{m} \sum_{i=1}^m \nabla_{\mathbf{v}} f_i(\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t) + \frac{\eta}{m} \sum_{i=1}^m (\nabla_{\mathbf{v}} f_i(\bar{\mathbf{u}}^t, \bar{\mathbf{v}}^t) - \nabla_{\mathbf{v}} f_i(\mathbf{u}^{t,i}, \bar{\mathbf{v}}^t)) \\ &\quad + \frac{\eta}{m} \sum_{i=1}^m (\nabla_{\mathbf{v}} f_i(\mathbf{u}^{t,i}, \bar{\mathbf{v}}^t) - \nabla_{\mathbf{v}} f_i(\mathbf{u}^{t,i}, \mathbf{v}^{t,i})). \end{aligned} \quad (84)$$

After tedious calculation, its formula is as follows

$$\bar{\mathbf{v}}_S^{t+1} = \bar{\mathbf{v}}_S^t \odot \left(\mathbf{1}_d + 4\eta \left((\bar{\mathbf{u}}_S^t)^2 - (\bar{\mathbf{v}}_S^t)^2 - \mathbf{w}^* \right) + \mathbf{E}_{d1}^t + \mathbf{E}_{d2}^t + \mathbf{E}_{d3}^t \right)$$

$$+ \frac{4\eta}{m} \sum_{i=1}^m \Delta_{v,S}^{t,i} \odot E_{d4}^{t,i}. \quad (85)$$

Based on induction (c) and (d), we can reparameterize $\Delta_{u,S}^{t,i}$ and $\Delta_{v,S}^{t,i}$ as $\Delta_{u,S}^{t,i} = \mathbf{r}_{u,S}^{t,i} \odot \bar{\mathbf{u}}_S^t$ and $\Delta_{v,S}^{t,i} = \mathbf{r}_{v,S}^{t,i} \odot \bar{\mathbf{v}}_S^t$, where $\max \left\{ \left\| \mathbf{r}_{u,S}^{t,i} \right\|_{\infty}, \left\| \mathbf{r}_{v,S}^{t,i} \right\|_{\infty} \right\} \leq 4\beta\rho^{\frac{3}{4}}\eta B_k$ for $\bar{T}_{k-1} \leq t < \bar{T}_k$. Thus, the (82) and (85) can be reformulated as

$$(\bar{\mathbf{u}}_S^{t+1})^2 = (\bar{\mathbf{u}}_S^t)^2 \odot \left(\mathbf{1}_d - 4\eta \left((\bar{\mathbf{u}}_S^t)^2 - (\bar{\mathbf{v}}_S^t)^2 - \mathbf{w}^* \right) + \mathbf{E}_{d1}^t + \mathbf{E}_{d2}^t + \mathbf{E}_{d3}^t \right) - 4\eta \frac{\sum_{i=1}^m \mathbf{r}_{u,S}^{t,i} \odot \mathbf{E}_{d4}^{t,i}}{m} \quad (86)$$

$$(\bar{\mathbf{v}}_S^{t+1})^2 = (\bar{\mathbf{v}}_S^t)^2 \odot \left(\mathbf{1}_d + 4\eta \left((\bar{\mathbf{u}}_S^t)^2 - (\bar{\mathbf{v}}_S^t)^2 - \mathbf{w}^* \right) + \mathbf{E}_{d1}^t + \mathbf{E}_{d2}^t + \mathbf{E}_{d3}^t \right) + 4\eta \frac{\sum_{i=1}^m \mathbf{r}_{v,S}^{t,i} \odot \mathbf{E}_{d4}^{t,i}}{m}. \quad (87)$$

Based on hypotheses (a), (c)-(f), (j) and step size condition, we have the following bound for additional perturbed error terms

$$\begin{aligned} & \max \left\{ \left\| \mathbf{r}_{u,S}^{t,i} \odot \mathbf{E}_{d4}^{t,i} \right\|_{\infty}, \left\| \mathbf{r}_{v,S}^{t,i} \odot \mathbf{E}_{d4}^{t,i} \right\|_{\infty} \right\} \leq 4\beta\rho^{\frac{3}{4}}\eta B_k \left(4 \left(\left\| \bar{\mathbf{u}}_S^t \right\|_{\infty}^2 + \left\| \bar{\mathbf{v}}_S^t \right\|_{\infty}^2 \right) \right. \\ & \quad \left. + 2\sqrt{s}\delta_{\max} \left(w_{\max}^* + \left\| \bar{\mathbf{u}}_S^t \right\|_{\infty}^2 + \left\| \bar{\mathbf{v}}_S^t \right\|_{\infty}^2 \right) + 16d\alpha + \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \right) \\ & \leq 4\beta\sqrt{\rho}\eta B_k \left(\rho^{\frac{1}{4}} (16 + 10\sqrt{s}\delta_{\max}) w_{\max}^* + 16d\alpha + \rho^{\frac{1}{4}} \max_i \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_{\infty} \right) \\ & \leq 40\beta\sqrt{\rho}\eta \cdot \frac{w_{\max}^*}{40 \times 2^k} \cdot \rho^{\frac{1}{4}} (2 + \sqrt{s}\delta_{\max}) w_{\max}^* \\ & \leq \frac{C_{\gamma} w_{\max}^*}{K \times 2^k}, \end{aligned} \quad (88)$$

where the last inequality is due to network connectivity condition that $\rho \leq \min \left\{ \left(\frac{\sqrt{s}\delta}{\sqrt{s}\delta_{\max} + 4} \right)^{\frac{4}{3}}, \frac{\left\| \mathbf{X}^T \boldsymbol{\xi} \right\|_{\infty}}{8m \max_i \left\| \mathbf{X}_i^T \boldsymbol{\xi}_i \right\|_{\infty}} \right\}$.

Then multiplying (86) and (87), we can obtain that

$$\begin{aligned} (\bar{\mathbf{u}}_S^{t+1} \odot \bar{\mathbf{v}}_S^{t+1})^2 & \leq (\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{v}}_S^t)^2 \odot \left(1 - \left(4\eta \left((\bar{\mathbf{u}}_S^t)^2 - (\bar{\mathbf{v}}_S^t)^2 - \mathbf{w}^* + \mathbf{E}_{d1}^t + \mathbf{E}_{d2}^t + \mathbf{E}_{d3}^t \right) \right)^2 \right. \\ & \quad \left. + 8\eta \max \left\{ \left\| \mathbf{r}_{u,S}^{t,i} \odot \mathbf{E}_{d4}^{t,i} \right\|_{\infty}, \left\| \mathbf{r}_{v,S}^{t,i} \odot \mathbf{E}_{d4}^{t,i} \right\|_{\infty} \right\} \right. \\ & \quad \left. + \left(4\eta \max \left\{ \left\| \mathbf{r}_{u,S}^{t,i} \odot \mathbf{E}_{d4}^{t,i} \right\|_{\infty}, \left\| \mathbf{r}_{v,S}^{t,i} \odot \mathbf{E}_{d4}^{t,i} \right\|_{\infty} \right\} \right)^2 \right)^2 \\ & \stackrel{(i)}{\leq} (\bar{\mathbf{u}}_S^t \odot \bar{\mathbf{v}}_S^t)^2 \odot \left(1 + 4\eta \cdot \frac{C_{\gamma} w_{\max}^*}{K \times 2^k} \right)^4 \\ & \stackrel{(ii)}{\leq} \alpha^4 \odot \left(1 + 4\eta \cdot \frac{C_{\gamma}}{K} 2^{-K+1} w_{\max}^* \right)^{8KT_{K-1}} \\ & \stackrel{(iii)}{\leq} \alpha^3, \end{aligned} \quad (89)$$

where the (i) is due to (88) and $4\eta \left\| (\bar{\mathbf{u}}_S^t)^2 - (\bar{\mathbf{v}}_S^t)^2 - \mathbf{w}^* + \mathbf{E}_{d1}^t + \mathbf{E}_{d2}^t + \mathbf{E}_{d3}^t \right\|_{\infty} \leq 1$. The (ii) is due to induction (a) and Lemma B.13 in [36]. The reason for (iii) is the same as the that of last inequality in (80) because of the setting of step size.

(j) Recall the recursive formula of $\bar{\mathbf{u}}_{S^-}^t$ in (49) and compare the definition of the perturbation $\mathbf{y}_u^t$ with perturbations $\mathbf{p}_u^t, \mathbf{g}_u^t$, we can obtain $\|\mathbf{y}_u^t\|_\infty \leq \frac{150\rho^{\frac{3}{4}}B_k(\sqrt{s}\delta_{\max}+4)}{600} + \frac{\rho^{\frac{3}{4}}d\alpha}{32}$. Compare the outer perturbation $\mathbf{z}_u^t$ with outer perturbations $\mathbf{f}_u^t, \mathbf{q}_u^t$, we can obtain $\|\mathbf{z}_u^t\|_\infty \leq \frac{\sqrt{\rho}B_k\|\bar{\mathbf{u}}_{S^-}^t\|_\infty}{16}$. With similar reparameterization as $\mathbf{z}_u^t = \mathbf{r}_{z_u}^t \odot \bar{\mathbf{u}}_{S^-}^t$, the (49) would become

$$(\bar{\mathbf{u}}_{S^-}^{t+1})^2 = (\bar{\mathbf{u}}_{S^-}^t)^2 \odot \left(\mathbf{1}_d - 4\eta \left(-\mathbf{w}_{S^-}^* - (\bar{\mathbf{v}}_{S^-}^t)^2 + \mathbf{E}_{2s}^t + \mathbf{E}_{3s}^t + \mathbf{y}_u^t + \mathbf{r}_{z_u}^t \right) \right)^2 \quad (90)$$

$\forall j \in S^-$ and $\forall t \leq \mathcal{O}\left(\frac{1}{\eta\zeta} \log \frac{1}{\alpha}\right)$. Let $0 \leq \tau \leq t$ be the largest τ such that $(\bar{v}_j^\tau)^2 > -w_j^*$. If there is no such τ exists or $\tau = t$, then

$$\begin{aligned} (\bar{u}_j^t)^2 &\leq (\bar{u}_j^{\tau-1})^2 (1 + 4\eta (\|\mathbf{E}_{2s}^{t-1}\|_\infty + \|\mathbf{E}_{3s}^{t-1}\|_\infty + \|\mathbf{y}_s^{t-1}\|_\infty + \|\mathbf{r}_{z_s}^{t-1}\|_\infty))^2 \\ &\leq \alpha, \end{aligned} \quad (91)$$

where the last inequality is due to the similar bound in (80).

If $\tau < t$, then unrolling (90) to τ -th iteration would have

$$\begin{aligned} (\bar{u}_j^t)^2 &= (\bar{u}_j^\tau)^2 \prod_{i=\tau}^{t-1} \left(1 - 4\eta \left(-w_j^* - (\bar{v}_j^i)^2 + E_{2s,j}^i + E_{3s,j}^i + y_{s,j}^i + r_{z_s,j}^i \right) \right)^2 \\ &\stackrel{(i)}{\leq} \frac{\alpha^2}{4} \left(1 - 4\eta \left(-w_j^* - (\bar{v}_j^\tau)^2 + E_{2s,j}^\tau + E_{3s,j}^\tau + y_{s,j}^\tau + r_{z_s,j}^\tau \right) \right)^2 \\ &\quad \cdot \prod_{i=\tau+1}^{t-1} \left(1 - 4\eta \left(-w_j^* - (\bar{v}_j^i)^2 + E_{2s,j}^i + E_{3s,j}^i + y_{s,j}^i + r_{z_s,j}^i \right) \right)^2 \\ &\stackrel{(ii)}{\leq} \alpha^2 \prod_{i=\tau+1}^{t-1} (1 + 4\eta (\|\mathbf{E}_{2s}^i\|_\infty + \|\mathbf{E}_{3s}^i\|_\infty + \|\mathbf{y}_s^i\|_\infty + \|\mathbf{r}_{z_s}^i\|_\infty))^2 \\ &\stackrel{(iii)}{\leq} \alpha, \end{aligned} \quad (92)$$

where (i) is based on condition for initialization that $\alpha \leq \frac{w_{\min}^*}{4}$ and induction (h) that $\bar{u}_j^\tau \leq \frac{\alpha^{\frac{3}{2}}}{\bar{v}_j^\tau} \leq \frac{\alpha^{\frac{3}{2}}}{\sqrt{-w_j^*}} \leq \frac{\alpha}{2}$. (ii) is based on the induction (a), step size condition that $\left(1 - 4\eta \left(-w_j^* - (\bar{v}_j^\tau)^2 + E_{2s,j}^\tau + E_{3s,j}^\tau + y_{s,j}^\tau + r_{z_s,j}^\tau \right) \right)^2 \leq 4$ and $\forall i > \tau$ that $(\bar{v}_j^i)^2 < -w_j^*$ which is based on definition of τ . The last inequality has the same reason as (91).

A.5 Proof of Main Results

A.5.1 Proof of Theorem 1

Based on the proof of hypothesis (a) in Proposition 3, we can conclude that $\forall t \geq \bar{T}_K$, there would be $\max \left\{ \left\| \left(\bar{\mathbf{u}}_{S^+}^{\bar{T}_{K-1}} \right)^2 - \mathbf{w}_{S^+}^* \right\|_\infty, \left\| \left(\bar{\mathbf{v}}_{S^-}^{\bar{T}_{K-1}} \right)^2 - \mathbf{w}_{S^-}^* \right\|_\infty \right\} \leq \frac{w_{\max}^*}{2^K} \leq \zeta$, where the last inequality is due to the definition of K . The total computational complexity is the value of $\bar{T}_{K-1} = \frac{2^K-1}{\eta w_{\max}^*} \log \frac{1}{\alpha^4} = \mathcal{O}\left(\frac{1}{\eta\zeta} \log \frac{1}{\alpha}\right)$.

Due to the definition of ζ in the Theorem 1, if $\zeta \geq w_{\max}^*$, then our result holds at $t = 0$ based on the small initialization condition. When $\zeta \leq w_{\max}^*$, we consider the two cases where the first case is the magnitudes of parameters is strong enough that $\zeta = \frac{1}{5}w_{\min}^* \geq 960\zeta$, where the value 960 is due to under the condition $2C_b + C_\gamma \leq \frac{1}{80}$, $4C_b \leq C_\gamma$, we set $C_b = \frac{1}{480}$, $C_\gamma = \frac{1}{120}$, we can use run DGD T_1' iterations to obtain the estimator $\max \left\{ \left\| \left(\bar{\mathbf{u}}_{S^+}^{\bar{T}_{K-1}} \right)^2 - \mathbf{w}_{S^+}^* \right\|_\infty, \left\| \left(\bar{\mathbf{v}}_{S^-}^{\bar{T}_{K-1}} \right)^2 - \mathbf{w}_{S^-}^* \right\|_\infty \right\} \leq \frac{w_{\max}^*}{2^K} \leq \frac{1}{5}w_{\min}^*$, then based on Proposition 2 in [36], we can also run DGD with another $\frac{45}{32\eta w_{\min}^*} \log \frac{w_{\min}^*}{\epsilon}$ iterations to obtain the dimension-independent bound $|\bar{w}_j^t - w_j^*| \leq \max \{ \sqrt{s}\delta \max_{i \in S} B_i, B_j, \epsilon \}, \forall j \in$

S where the $B_j := \left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \odot \mathbf{1}_j \right\|_\infty$. The total iterations are $\bar{T}_{K-1} + \frac{45}{32\eta w_{\min}^*} \log \frac{w_{\min}^*}{\epsilon} \leq \mathcal{O}\left(\frac{1}{\eta w_{\min}^*} \log \frac{1}{\alpha}\right) = \mathcal{O}\left(\frac{1}{\eta \zeta} \log \frac{1}{\alpha}\right)$ where the first inequality is due to the condition of α .

The second case is $\zeta = 960\varsigma \geq \frac{1}{5}w_{\min}^*$, then running of DGD with $\mathcal{O}\left(\frac{1}{\eta \zeta}\right)$ iterations would obtain the result in Theorem 1.

A.5.2 Proof of Corollary 1

Because the entries of global design matrix generated from i.i.d. 1-Sub-Gaussian distribution, we have the upper bound of global RIP parameter as $\delta \lesssim \frac{1}{\sqrt{s}}$ probability at least $1 - \frac{1}{8d^3}$, which is based on the Lemma 1 and sample size lower bound in Corollary 1 matches the condition of sample complexity in Lemma 1. For the local design matrix $\{\mathbf{X}_i/\sqrt{n}\}_{i=1}^m$, local sample size satisfies $n = \frac{N}{m} \gtrsim \left(\sqrt{\frac{m}{\delta}}\right)^{-2} (s \ln \frac{ed}{s} + \ln(dn))$, we can bound the local RIP parameter as $\delta_{\max} \leq \sqrt{\frac{m}{\delta}}$ with probability at least $1 - \frac{1}{8d^3}$ based on the union bound and Lemma 1. Based on condition in (6), the $\rho \lesssim \frac{1}{m^4}$ holds with probability at least $1 - \frac{1}{4d^3}$. Based on the Lemma 4, we have the upper bound for the noise level as $\left\| \frac{\mathbf{X}^T \boldsymbol{\xi}}{N} \right\|_\infty \lesssim \sigma \sqrt{\frac{\log d}{N}}$ with probability at least $1 - \frac{1}{8d^3}$. Thus, with probability at least $1 - \frac{3}{8d^3}$, we have $\|\bar{\mathbf{w}}^t - \mathbf{w}^*\|_2^2 \lesssim s\epsilon^2 + \frac{d\epsilon^2}{(2d+1)^2} \lesssim \frac{s\sigma^2 \log d}{N}$ where in the last inequality we select $\epsilon = 4\sigma \sqrt{\frac{2\log(2d)}{N}}$ in Theorem 1.

A.5.3 Proof of Proposition 1

Because each agent has same initialization $\mathbf{u}^{0,i} = \mathbf{v}^{0,i} = \alpha \mathbf{1}_d, \forall i \in [m]$, after one step of local gradient descent from (3) and (4), there are

$$\begin{aligned} \mathbf{u}^{1,i} &= \alpha \odot \left(1 + 4\eta \left(\mathbf{w}^* + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \mathbf{w}^* + \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right) \right) \\ \mathbf{v}^{1,i} &= \alpha \odot \left(1 - 4\eta \left(\mathbf{w}^* + \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \mathbf{w}^* + \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right) \right). \end{aligned} \quad (93)$$

Because the perturbed error bound

$$\left\| \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \mathbf{w}^* \right\|_\infty + \left\| \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n} \right\|_\infty \lesssim \phi := \sqrt{s} \delta_{\max} w_{\max} + \sigma \sqrt{\frac{\log d}{n}} \quad (94)$$

holds with probability at least $1 - \frac{3}{8d^3}$ based on Lemma 2 and Lemma 4, we denote $\boldsymbol{\nu}^i = \left(\frac{\mathbf{X}_i^T \mathbf{X}_i}{n} - \mathbf{I} \right) \mathbf{w}^* + \frac{\mathbf{X}_i^T \boldsymbol{\xi}_i}{n}$.

Consider $\mathbf{u}^{1,i}$, for $\forall p \in \mathcal{S}^+, \forall q \in \mathcal{S}^c$, there is $w_p^* - |\nu_p^i| > \phi > |\nu_q^i| > 0$ based on the condition in Proposition 1. In addition, $\forall j \in \mathcal{S}^-$, there is $-|w_j^*| + |\nu_j^i| < 0$ based on the condition on Proposition 1. Thus, the growth of elements on positive support $\mathcal{S}^+$ would be larger than these of $\mathcal{S}^-, \mathcal{S}^c$, and the Trun_k operator would identify the $\mathcal{S}^+$. The analogous analysis could also applied to $\mathbf{v}^{1,i}$ that the Trun_k operator would identify the $\mathcal{S}^-$. Because each agent can identify the $\mathcal{S}^+$ and $\mathcal{S}^-$ and based on results in Proposition 3, the Trun_k would also obtain the optimal statistical error with probability at least $1 - \frac{3}{8d^3}$. $\square$

B Appendix.B

B.1 Kernel to rich regime transition

Previous works show a transition from the kernel regime to the rich regime by varying the initialization scale in the gradient descent method [38]. Experimental results in Fig. 7 show that the transition phenomenon also appears in DGD. We can observe that when we increase the initialization scale α gradually, DGD would converge to the minimal ℓ_2 norm solution $\mathbf{w}_{\ell_2}^*$. On the contrary, when we

decrease the α , the DGD would converge to the minimal ℓ_1 norm solution (sparse solution) $w_{\ell_1}^*$. Thus, Fig. 7 demonstrates the existence of phase transition from kernel to rich regime for DGD when decreasing initialization α . Since we focus on sparse recovery, the small initialization would achieve this aim with better generalization performance.

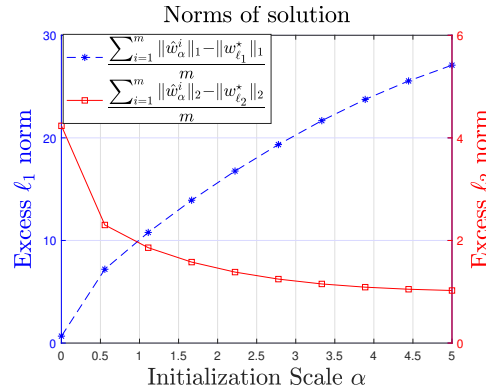


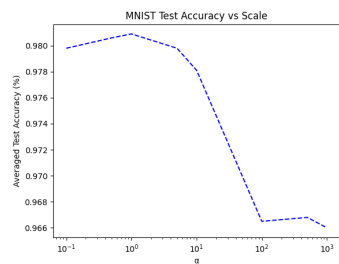
Figure 7: We plot $\frac{\sum_{i=1}^m \|\hat{w}_\alpha^i\|_1 - \|w_{\ell_1}^*\|_1}{m}$ in the blue and $\frac{\sum_{i=1}^m \|\hat{w}_\alpha^i\|_2 - \|w_{\ell_2}^*\|_2}{m}$ in red vs. α , where $\hat{w}_\alpha^i$ denote the convergent solution by DGD in i -th agent. The setting is $d = 2000$, $k = 2$, $N = 20$, $m = 20$, $\rho = 0.2135$ and magnitude of sparse signal is 10.

B.2 More implicit regularizations of DSGD

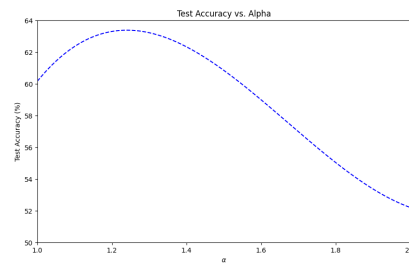
To verify the widespread existence of implicit regularization in decentralized optimization, we add two extra experiments on general overparameterized neural network architecture trained by DSGD, which are motivated from [38].

(1) The first one is that we use vanilla decentralized SGD(DSGD) to train a depth-2, 5000 hidden ReLU network with the cross-entropy loss on the MNIST dataset until each agent model reaches almost 100% training accuracy. The number of agents is 10 and network connectivity is $\rho = 0.178$. 60000 total training samples and 10000 test samples are uniformly allocated to agents. To evaluate the implicit regularization of SGD under varying initialization scales, the network weights are initialized as αw_0 , $w_0 \sim \mathcal{N}(0, \sqrt{\frac{2}{n_{\text{in}}}})$, which is suggested by [14] and n_{in} denotes the number of units in the last layer. Each agent uses the same batch size 256 to train in DSGD. The step sizes were optimally tuned for each α individually to achieve the best validation error. We plot the average test error (which is defined as the summation of the test error of each agent's model, then divided by the number of agents) vs. α in Fig. 8(a). The figure shows a visible phase transition for generalization ($\approx 98\%$ for $\alpha \leq 6$, and $\approx 96.6\%$ error for $\alpha \geq 100$). Fig. 8(a) shows that the transition from the kernel regime to the rich regime by varying the initialization scale may also exist in complex fully connected neural networks.

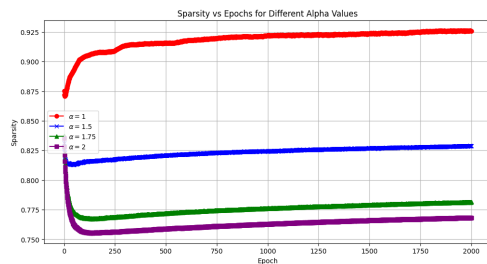
(2) The second one is that we use vanilla DSGD with batch size 128 to train the VGG11-like deep convolutional neural network on CIFAR10 with small step size 10^{-4} for 2000 epochs such that each local model achieves almost 100% training accuracy. The network setting is the same as the first experiment. The VGG11-like architecture is the same as [38]. Weights were initialized using Uniform He initialization multiplied by α . The Fig. 8(b) plots the average test accuracy vs. α . In addition, we adopt the sparse feature learning measure [4] to monitor the sparsity of learned features in all agents along the epoch in Fig. 8(c). From Fig. 8(c) and Fig. 8(d), we can observe that the implicit regularization of DGD under small initialization may promote neural networks to learn sparse features that can have better generalization performance, which can be regarded as the complementary observation in [4] under decentralized learning setting. How to prove this kind of implicit regularization of DSGD theoretically will be an interesting future work.



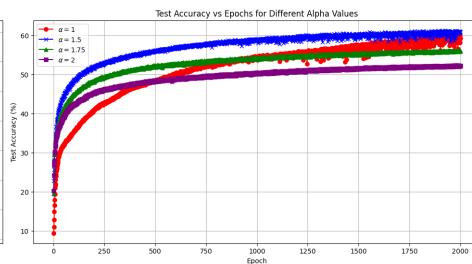
(a) MNIST test accuracy vs α



(b) CIFAR10 test accuracy vs α



(c) Feature sparsity vs α



(d) Training curves for CIFAR10

Figure 8: (a) DSGD trained ReLU network on **MNIST**. (b), (c), (d) DSGD trained VGG11-like network on **CIFAR10**.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer:[NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.