Distributed Sparse Regression via Penalization

Yao Ji JIYAO@PURDUE.EDU
School of Industrial Engineering
Purdue University
West Lafayette, IN 47906, USA

Gesualdo Scutari* GSCUTARI@PURDUE.EDU
School of Industrial Engineering
Purdue University
West Lafayette, IN 47906, USA

Ying Sun* YSUN@PSU.EDU
School of Electrical Engineering and Computer Science
The Pennsylvania State University
State College, PA 16802, USA

Harsha Honnappa HONNAPPA@PURDUE.EDU
School of Industrial Engineering
Purdue University
West Lafayette, IN 47906, USA

Abstract

We study sparse linear regression over a network of agents, modeled as an undirected graph (with no centralized node). The estimation problem is formulated as the minimization of the sum of the local LASSO loss functions plus a quadratic penalty of the consensus constraint—the latter being instrumental to obtain distributed solution methods. While penalty-based consensus methods have been extensively studied in the optimization literature, their *statistical* and computational guarantees in the *high dimensional* setting remain unclear. This work provides an answer to this open problem. Our contribution is two-fold. First, we establish statistical consistency of the estimator: under a suitable choice of the penalty parameter, the optimal solution of the penalized problem achieves *near optimal minimax rate* $\mathcal{O}(s\log d/N)$ in ℓ_2-loss, where s is the sparsity value, d is the ambient dimension, and N is the *total* sample size in the network–this matches centralized sample rates. Second, we show that the proximal-gradient algorithm applied to the penalized problem, which naturally leads to distributed implementations, converges linearly up to a tolerance of the order of the centralized statistical error–the rate scales as $\mathcal{O}(d)$, revealing an unavoidable speed-accuracy dilemma. Numerical results demonstrate the tightness of the derived sample rate and convergence rate scalings.

1. Introduction

We study high-dimensional sparse estimation over a network of m agents, modeled as an undirected graph. No centralized agent is assumed in the network; agents can communicate only with their immediate neighbors. Each agent i owns a data set (y_i, X_i), generated

∗. Equal contribution.

according to the linear model

$$y_i = X_i\theta^* + w_i, \tag{1}$$

where $y_i \in \mathbb{R}^n$ is the vector of n observations, $X_i \in \mathbb{R}^{n\times d}$ is the design matrix, $w_i \in \mathbb{R}^n$ is observation noise, and $\theta^* \in \mathbb{R}^d$ is the unknown s-sparse parameter *common* to all local models. In the high-dimensional setting, as postulated here, the ambient dimension d is larger than the total sample size $N = n \cdot m$ and $s << d$.

A standard approach to estimate θ^* from $\{(y_i, X_i)\}_{i=1}^m$ is to solve the LASSO problem, whose Lagrangian form reads

$$\hat{\theta} \in \underset{\theta\in\mathbb{R}^d}{\arg\min}\ \frac{1}{m}\sum_{i=1}^m \frac{1}{2n}\|y_i - X_i\theta\|^2 + \lambda\|\theta\|_1, \tag{2}$$

where $\lambda > 0$ controls the sparsity of the solution $\hat{\theta}$. Since the objective function involves the entire data set $\{(y_i, X_i)\}_{i=1}^m$ across the network, and routing local data to other agents is infeasible (e.g., due to privacy issues) or highly inefficient, Problem (2) cannot be solved by each agent i independently. This calls for the design of distributed algorithms whereby agents alternate computations, based on available local information, with communications with neighboring nodes. To this end, a widely adopted approach is to decompose (2) by introducing local estimates θ_i's of the common variable θ, each one controlled by one agent, and forcing consensus among the agents (e.g., (Nedić et al., 2018)):

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{md}} \frac{1}{m}\sum_{i=1}^m \frac{1}{2n}\|y_i - X_i\theta_i\|^2 + \frac{\lambda}{m}\|\boldsymbol{\theta}\|_1, \quad \text{subject to} \quad V\boldsymbol{\theta} = \mathbf{0}, \tag{3}$$

where $\boldsymbol{\theta} = [\theta_1^\top, \ldots, \theta_m^\top]^\top$ is the 'stack vector' of all the local copies θ_i's, and V is a positive semidefinite consensus-enforcing matrix, i.e., $V\boldsymbol{\theta} = \mathbf{0}$ if and only if all θ_i's are equal.

The objective function in (3) is now (additively) separable in the agents' variables; however, there is still a coupling across the θ_i's, due to the consensus constraint $V\boldsymbol{\theta} = \mathbf{0}$. To resolve this coupling, a widely adopted strategy in the literature of distributed optimization is to employ an inexact penalization of the constraint via a quadratic function. This leads to the following relaxed formulation:

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^{md}}{\arg\min}\ \frac{1}{m}\sum_{i=1}^m \frac{1}{2n}\|y_i - X_i\theta_i\|^2 + \frac{1}{2m\gamma}\|\boldsymbol{\theta}\|_V^2 + \frac{\lambda}{m}\|\boldsymbol{\theta}\|_1, \tag{4}$$

where $\|\boldsymbol{\theta}\|_V^2 \triangleq \boldsymbol{\theta}^\top V\boldsymbol{\theta}$, and $\gamma > 0$ is a free parameter controlling the violation of the consensus constraint $V\boldsymbol{\theta} = \mathbf{0}$. Invoking standard results of penalty methods (see, e.g., Nesterov et al. (2018)), it is not difficult to check that, when $\gamma \downarrow 0$, every limit point of the resulting sequence $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\gamma)$ is a solution of (3). This justifies the use of (4) as an approximation of (3) (for sufficiently small γ).

Problem (4) unlocks distributed solution methods. Here, we consider the proximal gradient algorithm (Nesterov et al., 2018) that, based upon a suitable choice of the matrix V, is readily implementable over the network. This resembles the renowned Distributed Gradient Descent algorithms (DGD) (see Sec. 1.3), which are among the most studied distributed schemes in the literature. Motivated by the popularity of the penalized formulation (4) and associated DGD algorithms, the goal of this paper is to study the statistical properties of the estimator (4) as well as computational guarantees of the aforementioned associated DGD algorithm.

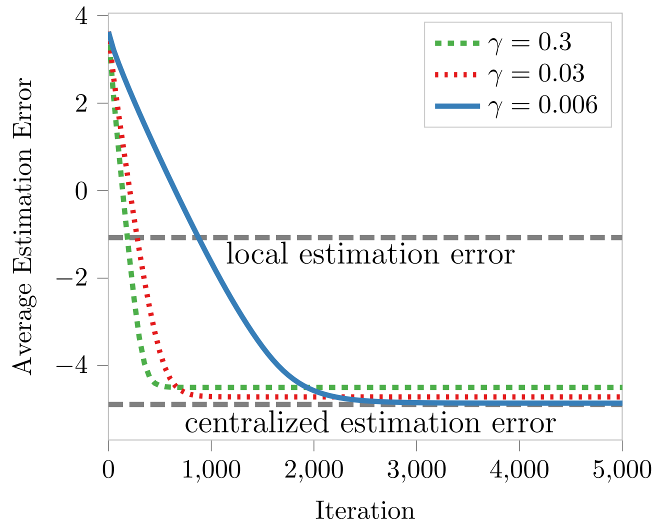

Figure 1: Proximal gradient in the high-dimensional setting (4): linear convergence up to some tolerance; different curves refer to different values of the penalty parameter γ. Notice the speed-accuracy dilemma.

1.1 Challenges and open problems

While penalty-based formulations like (4) and related solution methods have been extensively studied in the optimization literature, the statistical properties of the solution $\hat{\boldsymbol{\theta}}$ in the *high-dimensional* setting ($d >> N$) remain unknown, and so are the convergence guarantees of the proximal gradient algorithm applied to (4). Postponing to Sec. 1.3 a detailed review of the literature, here we point out the following. **Statistics:** classical sample complexity analysis of LASSO error $\|\hat{\theta} - \theta^*\|^2$ for (2) (e.g., (Wainwright, 2019)) is not directly applicable to the penalized problem (4)–for instance, it is unclear whether each agent's error $\|\hat{\theta}_i - \theta^*\|^2$ can match centralized sample complexity. **Distributed optimization:** When it comes to algorithms for solving (4), existing studies are of pure optimization type, lacking of statistical guarantees. If nevertheless invoked to predict convergence of the proximal gradient algorithm applied to (4), they would certify *sublinear* convergence of the optimization error, since the objective function in (4) is not strongly convex on the entire space (recall $d > N$). This results in a pessimistic prediction, as shown by the exploratory experiment in Fig. 1: the average estimation error $(1/m) \cdot \sum_{i=1}^{m} \|\theta_i^t - \theta^*\|^2$ decreases *linearly* up to a tolerance (floor); different curves refer to different values of the penalty parameter γ. The figure also plots the (square) estimation error achieved by solving (2)–termed as centralized estimation error–and the average of the (square) estimation errors achieved by each agent solving the LASSO problem using only its local data–termed as local estimation error. The experiment seems to suggest that statistical error comparable to centralized ones are still achievable where data are distributed over a network. However this requires a sufficiently small γ, and thus results in slow convergence rates.

To the best of our knowledge, a theoretical understanding of these phenomena remains an open problem; questions are abundant, such as: (i) Is centralized statistical consistency (sample complexity $N = o(s \log d)$) provably achievable when data are distributed across the network? What is the role/impact of the network? (ii) Does the (distributed) proximal gradient achieve statistically optimal solutions while converging linearly? (iii) How do sample and convergence rates of the algorithm scale with the model, γ, d, N, and network parameters?

1.2 Major contributions

This work addresses the above questions–our contributions can be summarized as follows.

1) **Statistical analysis of the penalized LASSO problem** (4)**:** We establish non-asymptotic error bounds on the estimation error averaged over the agents, $(1/m)\sum_{i=1}^{m}\|\hat{\theta}_i - \theta^*\|^2$, under proper tuning of λ and γ. Our results are of two types. **(i)** A deterministic bound, under a strong convexity requirement on the objective in (4) restricted to certain directions containing the augmented LASSO error $\hat{\boldsymbol{\theta}} - 1_m \otimes \theta^*$ (cf. Theorem 6)–this bound sheds light on the role of the network and consensus errors (via γ) into the estimation process; and **(ii)** a (sample) convergence rate $(1/m)\sum_{i=1}^{m}\|\hat{\theta}_i - \theta^*\|^2 = \mathcal{O}(s \log d/N)$ (cf. Theorem 7), which holds with high probability (w.h.p.) under standard Gaussian data generation models (cf. Assumption 1).This matches the statistical error of the centralized LASSO estimator (2), showing for the first time that statistical consistency over networks is achievable under the same order of the sample size N used in the centralized setting, *even when the number of local samples n is insufficient.* This is possible because agents communicate over the network.

2) **Algorithmic convergence and statistical guarantees:** To compute such estimators in a distributed fashion, we leverage the proximal-gradient algorithm applied to (4), and study its convergence and statistical properties (cf. Theorems 10 and 12). A major result is proving that, in the setting (ii) above, the algorithm converges *linearly* up to a fixed tolerance which can be driven below the statistical precision of the centralized LASSO problem (2). Specifically, to enter an ε-neighborhood of a statistically optimal solution, it takes
$$\mathcal{O}\left(\frac{1}{1-\rho}\cdot\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\cdot d\, m \log m \cdot \log\frac{1}{\varepsilon}\right)$$
number of communications (iterations), where $\rho \in [0,1)$ is a measure of the connectivity of the network (the smaller ρ is, the more connected the graph); and $\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ is the restricted condition number of the LASSO loss function [see (2)], with $\lambda_{\max}$ (resp. $\lambda_{\min}$) denoting the largest (resp. smallest) eigenvalue of the covariance matrix Σ of the data (cf. Assumption 1). This shows that centralized statistical accuracy is achievable over a given network (without moving data) but at the price of a linear rate (and thus communication cost) that scales as $\mathcal{O}(d)$. This 'speed-accuracy dilemma' is confirmed by our experiments (cf. Sec. 5). A similar phenomenon has been observed previously

in low-dimensional settings (strongly convex losses) (Yuan et al., 2016, Theorem 3). However, our results demonstrating this dilemma in the high-dimensional setting as well, imply that this appears to be a "scarlet letter" of DGD-like algorithms.

1.3 Related works

Statistical analysis: Statistical properties of the LASSO solution $\hat{\theta}$ of (2) along with several other regularized M-estimators have been extensively studied in the literature (see, e.g., (Tibshirani, 1996, Hastie et al., 2015, Wainwright, 2019) and references therein). Introducing suitably restricted notions of strong convexity of the loss–e.g., (Bickel et al., 2009, Candes and Tao, 2006, de Geer and Buhlmann, 2009, Sahand et al., 2012, Wainwright, 2019)–(nonasymptotic) error bounds and sample complexity for such estimators under high-dimensional scaling are established. For instance, for the LASSO estimator (2), statistical errors read $\|\hat{\theta} - \theta^*\|^2 = \mathcal{O}(s \log d/N)$.

These conditions and results for (2) do not transfer directly to the *lifted, penalized* formulation (4)–it is not even clear the relation between $\hat{\theta}$ [cf. (2)] and $\hat{\theta}_1, \ldots, \hat{\theta}_m$ [cf. (4)]. A new solution and statistical analysis is needed for the "augmented" LASSO estimator $\hat{\boldsymbol{\theta}}$ (4), possibly revealing the role of the network on the statistical properties of $\hat{\boldsymbol{\theta}}$.

Centralized optimization algorithms: Referring to solution methods for *centralized* sparse linear regression problems, several studies are available in the literature, including (Becker et al., 2011, Beck and Teboulle, 2009, Bredies and Lorenz, 2008, Hale et al., 2008, Tseng and Yun, 2009, Zhou and So, 2017, Wen et al., 2017, Bolte et al., 2009) and (Agarwal et al., 2012). Since (2) is not strongly convex in a global sense, classical (accelerated) first-order methods like (Becker et al., 2011, Beck and Teboulle, 2009) are known to converge at sublinear rate; others (Bredies and Lorenz, 2008, Hale et al., 2008) are proved to achieve linear convergence if initialized in a neighborhood of the solution of (2); and (Tseng and Yun, 2009, Zhou and So, 2017, Wen et al., 2017, Bolte et al., 2009) showed linear convergence (in particular) of the proximal-gradient algorithm, invoking global regularity conditions of the loss (2), such as the Luo-Tseng's bound (Luo and Tseng, 1993) or the KL property (Bolte et al., 2009, Pan and Liu, 2018). These studies are of pure optimization type–e.g., convergence focuses on iteration complexity of the optimization error, no statistical analysis of the limit points is provided. Furthermore, they are not suitable for the high-dimensional regime (i.e., "d, N growing"). A closer related work is (Agarwal et al., 2012), which establishes global linear convergence of the proximal-gradient algorithm for (2) up to the statistical precision of the model, under a restricted strong convexity (RSC) and restricted smoothness (RSM) assumption. The method is not directly implementable over mesh networks, because of the lack of a centralized node. Furthermore, it is unclear whether RSC/RSM conditions hold for the penalized sum-loss in (4). On the other hand, a naive application of the RSC/RSM to *each* agent's loss $f_i(\theta_i) = (1/2n)\|y_i - X_i\theta_i\|^2$ in (4) (without accounting for the penalty, coupling term $(1/\gamma)\|\boldsymbol{\theta}\|_V^2$), would require a local sample scaling $n = \mathcal{O}(s \log d)$ to hold. This conclusion is unsatisfactory because it would state that the centralized minimax error bound $\|\hat{\theta} - \theta^*\|^2 = \mathcal{O}(s \log d/N)$ is not achievable over networks–a fact that is confuted by our theoretical findings and experiments.

Divide and Conquer (D&C) methods: When it comes to decomposition methods for statistical estimation and inference, the statistical community is best acquainted with D&C

methods. D&C algorithms postulate the existence of a node in the network (a.k.a. *master* node) connected to all the others (termed *worker* nodes), which combines the estimators produced by each worker using its local data set. D&C algorithms for M-estimation in *low-dimension*, covering the asymptotics $d, N \to \infty$ while $d/N \to c \in [0, 1)$, have been extensively studied in the literature; representative examples include Rosenblatt and Nadler (2016), Wang et al. (2018), Chen et al. (2021), Bao and Xiong (2021), Jianqing et al. (2021). More relevant to this work are the D&C methods applicable to sparse linear regression in *high-dimension*, i.e., $d > N$ and $d/N \to \infty$, which include Lee et al. (2015), Battey et al. (2018), Wang et al. (2017), Jordan et al. (2018). Lee et al. (2015), Battey et al. (2018) devised a one-shot approach averaging at the master node "debiased" local LASSO estimators. Wang et al. (2017), Jordan et al. (2018) independently improved the sample complexity of Lee et al. (2015) hinging on ideas from Shamir et al. (2014)–Table 1 provides the sample and communication complexity of these methods, which can be summarized as follows. By performing a single round of communication from the workers to the master node, resulting in a $\mathcal{O}(d)$ communication cost, these algorithms achieve the centralized statistical error $\mathcal{O}(s \log d/N)$ as long as the local sample size is sufficiently large, i.e., $n = \Omega(ms^2 \log d)$ (see Table. 1). Alternatively, this imposes a constraint on the maximum number of workers, i.e., $m = \mathcal{O}(n/(s^2 \log d))$, which limits the range of applicability of these methods to small-size (star) networks. The condition on m can be removed and that on n alleviated at the cost of multiple communication rounds; to our knowledge, the state of the art is Wang et al. (2017) showing that $n = \Omega(s^2 \log d)$ suffices under $\log m$ communication rounds, resulting in a total communication cost $\mathcal{O}(d \log m)$. None of these methods is directly implementable over mesh networks, because of the lack of a centralized node. Naive attempts of decentralizing D&C methods over mesh networks by replacing the exact average at the master node with local consensus updates fail to achieve centralized statistical consistency.

D&C Methods	$n \gtrsim m\, s^2 \log d$	$m\, s^2 \log d \gtrsim n \gtrsim s^2 \log d$
	Communication Cost (one round)	Communication Cost (multiple rounds)
Avg-Debias Lee et al. (2015)	d	✗
Battey et al. (2018)	d	✗
CSL Jordan et al. (2018)	d	✗[1]
EDSL Wang et al. (2017)	d	$d \log m$

Table 1: D&C algorithms for sparse linear regression in the high-dimensional, $d > N$ and $d/N \to \infty$: local sample size and communication cost to achieve the centralized statistical error $\mathcal{O}(s \log d/N)$. For a single communication round, all methods require a condition on the minimum local sample size n; multiple communication rounds can reduce the condition on local sample size n. [1]CSL Jordan et al. (2018) can be extended to multiple rounds of communication to reduce the local sample size using the similar argument as in EDSL Wang et al. (2017).

In contrast to D&C methods, the DGD-like algorithm studied in this paper to solve (4) provably achieves (near) optimal minimax rates with *no conditions on the local sample size*, at a total communication cost however of $\mathcal{O}(d^2)$. This raises the question whether communication costs of $\mathcal{O}(d)$ are achievable in high-dimension over mesh networks by other distributed optimization algorithms, yet with no conditions on the local sample size. Motivated by this work, the study of other methods in high-dimension is the subject of current investigation; see, e.g., the companion work Sun et al. (2022). In fact, as discussed next, there is no study of any other existing distributed algorithm in high-dimension.

Distributed optimization algorithms: Solving the LASSO problem (2) over mesh networks falls under the umbrella of distributed optimization. The literature of distributed optimization methods is vast; given the focus of the paper, we comment next only relevant works on decentralization of the (proximal) gradient method over mesh networks modeled as undirected graphs. Distributed Gradient Descent (DGD) algorithms, including those derived by penalizing consensus constraints as in (4), have been extensively studied in the literature; see, e.g., (Nedić and Ozdaglar, 2009, Nedić et al., 2010, Chen and Sayed, 2012, Sayed, 2014, Chen and Ozdaglar, 2012, Yuan et al., 2016, Zeng and Yin, 2018, Daneshmand et al., 2020, Nedić et al., 2018). Among all, the most relevant distributed scheme to this paper is (Zeng and Yin, 2018), a proximal gradient algorithm. When applied to (4), under the additional assumption of bounded (sub)gradient of the agents' losses (a fact that is not guaranteed), *sublinear* convergence (on the objective value) to the optimal solutions of (4) would be certified (recall that agents' losses are not strongly convex globally). Furthermore, the connection between the solution of the penalized problem (4) and that of the LASSO formulation (2) remains unclear.

While different and not derived directly from (4), the other DGD-like algorithms can be roughly commented as follows: (i) when the agents' loss functions are strongly convex (or the centralized loss satisfies the KL property (Zeng and Yin, 2018, Daneshmand et al., 2020)), differentiable, and there are no constraints, DGD-like schemes equipped with a constant stepsize, converge (only) to a neighborhood of the solution at linear rate (Yuan et al., 2016, Zeng and Yin, 2018, Yuan et al., 2020). Convergence (in objective value) to the exact solution is achieved only using diminishing stepsize rules, thus at the slower sublinear rate (see, e.g., (Zeng and Yin, 2018, Jakovetić et al., 2014)). This speed-accuracy dilemma can be overcomed by correcting explicitly the local gradient direction so that a constant stepsize can be used still preserving convergence to the exact solution; examples include: gradient tracking methods (Qu and Li, 2017, Nedić et al., 2016, Xu et al., 2018, Lorenzo and Scutari, 2016, Sun et al., 2019) and primal-dual schemes (Jakovetić et al., 2011, Shi et al., 2014, Jakovetić, 2019, Jakovetić et al., 2013, Shi et al., 2015a,b), just to name a few.

Distributed Compressed Sensing: Distributed compressed sensing (DCS) is a specific instance of distributed sparse linear regression, where the observation noise is deterministic and bounded by $\|w\|_\infty \leq \sigma$. Various approaches have been proposed for solving DCS, but none of them have been able to achieve a linear convergence rate together with nearly minimax sample complexity. One approach proposed by Fosson et al. (2016) is DJ-IST, a distributed soft thresholding algorithm for recovering jointly sparse signals. They prove that DJ-IST converges to a minimum of a suitable cost functional with concave penalization. However, they do not guarantee a convergence rate nor the minimax optimality of the solution. Other works, such as Patterson et al. (2014) and Azarnia et al. (2020), built upon distributed iterative hard thresholding methods. Patterson et al. (2014) proposes distributed iterative hard thresholding (DHIT) to solve DCS, but it can only recover the sparse ground truth up to $\|w\|$, free of the sample size, indicating a loose bound. Additionally, DHIT requires a global computation step that constructs a spanning tree over the network rooted at a special node. This special node needs information about other nodes that must be trained using a distributed algorithm, leading to high computation and communication costs. Azarnia et al. (2020) proposes an extension of hard thresholding pursuit (HTP) that can be used in a distributed manner. They establish a linear convergence rate for distributed diffusion hard thresholding pursuit (DDHTP) under Restricted Isometry Property, which is more stringent than Assumption 3 required in this paper; yet, the steady-state error of DDHTP does not match the nearly minimax optimal bound $\mathcal{O}(s\sigma^2 \log d/N)$.

The above overview of distributed algorithms shows that there is no study of statistical/computational guarantees in the *high-dimensional* regime. Our comments above on centralized optimization algorithms apply here for all the aforementioned distributed ones: all the convergence results are of pure optimization type and are confuted by our experiments (see Fig. 1). A new analysis is needed to understand the behaviour of distributed algorithms in the high-dimensional regime. This paper represents the first study of a DGD-like algorithm towards this direction.

1.4 Notation and paper organization

The rest of the paper is organized as follows. Sec. 2 introduces the main assumptions on the data model and network setting along with some related consequences. Solution analysis of the penalized LASSO (4) is addressed in Sec. 3–a deterministic error bound, based on a notion of restricted strong convexity, is first established; then near optimal centralized sample complexity is proved under a Gaussian data generation model (Sec. 3.3). Statistical and computational guarantees of the (distributed) proximal gradient algorithms applied to (4) are discussed in Sec. 4. First, linear convergence to the solution of the original LASSO problem up to some tolerance is proved; then, statistical guarantees under the aforementioned Gaussian model are determined (Sec. 4.1). Finally, Sec. 5 provides some experiments validating our theoretical findings while Sec. 6 draws some conclusions. All the proofs of our results are relegated to the appendix.

Notation: Let $[m] \triangleq \{1, \ldots, m\}$, with $m \in \mathbb{N}_{++}$; 1 is the vector of all ones; e_i is the i-th canonical vector; I_d is the $d \times d$ identity matrix (when unnecessary, we omit the subscript); and $\otimes$ denotes the Kronecker product. Given $x_1, \ldots, x_m \in \mathbb{R}^d$, the bold symbol $\mathbf{x} = [x_1^\top, \ldots, x_m^\top]^\top \in \mathbb{R}^{md}$ denotes the stack vector; for any $\mathbf{x} = [x_1^\top, \ldots, x_m^\top]^\top$, we define its block-average as $x_{\text{av}} \triangleq (1/m)\sum_{i=1}^m x_i$, and the disagreement vector $\mathbf{x}_\perp \triangleq [x_{\perp 1}^\top, \ldots, x_{\perp m}^\top]^\top$, with each $x_{\perp i} \triangleq x_i - x_{\text{av}}$. Similarly, for any collection of matrices $X_1, \ldots, X_m \in \mathbb{R}^{n\times d}$, we use bold notation for the stacked matrix $\mathbf{X} = [X_1^\top, \ldots, X_m^\top]^\top$. We order the eigenvalues of any symmetric matrix $A \in \mathbb{R}^{m\times m}$ in nonincreasing fashion, i.e., $\lambda_{\max}(A) = \lambda_1(A) \geq \ldots \geq \lambda_m(A) = \lambda_{\min}(A)$. We use $\|\cdot\|$ to denote the Euclidean norm; when other norms are used, e.g., ℓ_1-norm and ℓ_∞, we will append the associate subscript to $\|\cdot\|$, such as $\|\cdot\|_1$, and $\|\cdot\|_\infty$. Consistently, when applied to matrices, $\|\cdot\|$ denotes the operator norm induced by $\|\cdot\|$. Furthermore, we write $\|x\|_A \triangleq (x^\top A x)^{1/2}$, for any symmetric positive semidefinite matrix. Given $\mathcal{S} \subseteq [d]$ and $y \in \mathbb{R}^d$, we denote by $|\mathcal{S}|$ the cardinality of $\mathcal{S}$ and by $y_\mathcal{S}$ the $|\mathcal{S}|$-dimensional vector containing the entries of y indexed by the elements of $\mathcal{S}$; $\mathcal{S}^c$ is the complement of $\mathcal{S}$. All the log in the paper are intended natural logarithms, unless otherwise stated. Given two univariate random variables X and Y, we say that Y has stochastic dominance over X if $X \preceq^{\text{st}} Y$, meaning $\mathbb{P}(X \leq t) \geq \mathbb{P}(Y \leq t)$, for all $t \in \mathbb{R}$ (Marshall et al., 2011, Page 694).

2. Setup and Background

In this section we introduce the main assumptions on the data model and network setting underlying our analysis along with some related consequences.

2.1 Problem setting

The following quantities associated with (1) will be used throughout the paper:

$$\mathcal{S} \triangleq \operatorname{supp}\{\theta^*\}, \quad s = |\mathcal{S}|, \quad L_{\max} \triangleq \max_{i \in [m]} \lambda_{\max}\left(\frac{X_i^\top X_i}{n}\right). \tag{5}$$

We collect all the local data $\{(y_i, X_i)\}_{i=1}^m$ into the stacked vector measures $\mathbf{y} = [y_1^\top, \ldots, y_m^\top]^\top \in \mathbb{R}^N$ and matrix $\mathbf{X} = [X_1^\top, \ldots, X_m^\top]^\top \in \mathbb{R}^{N \times d}$. The quadratic losses of the centralized LASSO problem (2) and of the penalized one (4) are denoted respectively by

$$F(\theta) \triangleq \frac{1}{2N}\|\mathbf{y} - \mathbf{X}\theta\|^2 \quad \text{and} \quad L(\boldsymbol{\theta}) \triangleq \frac{1}{2N}\sum_{i=1}^m \underbrace{\|y_i - X_i\theta_i\|^2}_{\triangleq f_i(\theta_i)} + \frac{1}{2m\gamma}\|\boldsymbol{\theta}\|_V^2. \tag{6}$$

We recall next the main path/assumptions used to bound the LASSO error $\|\hat{\theta} - \theta^*\|^2$ in the centralized setting (2) (e.g., (Wainwright, 2019)). In the regime $d > N$, F is not strongly convex–the $d \times d$ Hessian matrix $\mathbf{X}^\top\mathbf{X}$ has at most rank N. Nevertheless, $\|\hat{\theta} - \theta^*\|^2$ can be well-controlled requiring strong convexity of F to hold for a subset of directions. The *Restricted Eigenvalue* (RE) condition suffices (Bickel et al., 2009, Candes and Tao, 2006, Wainwright, 2019)

$$\frac{1}{N}\|\mathbf{X}\Delta\|^2 \geq \delta_c\|\Delta\|^2, \quad \forall \Delta \in \mathbb{C}(\mathcal{S}) \triangleq \{\Delta \in \mathbb{R}^d \mid \|\Delta_{\mathcal{S}^c}\|_1 \leq 3\|\Delta_{\mathcal{S}}\|_1\}, \tag{7}$$

where $\delta_c > 0$ is the curvature parameter, and $\mathbb{C}(\mathcal{S})$ captures the set of "sparse" directions of interests. The rationale behind (7) is that, since $\hat{\theta} - \theta^*$ can be proved to belong to $\mathbb{C}(\mathcal{S})$, if F is strongly convex on $\mathbb{C}(\mathcal{S})$–as requested by (7)–then small differences on the loss will translate into bounds on $\|\hat{\theta} - \theta^*\|^2$.

The RE (7) imposes conditions on the design matrix $\mathbf{X}$. The following RSC implies (7).

Lemma 1 *Suppose that F satisfies the following RSC condition with curvature $\mu > 0$ and tolerance $\tau > 0$:*

$$\frac{1}{N}\|\mathbf{X}\Delta\|^2 \geq \frac{\mu}{2}\|\Delta\|^2 - \frac{\tau}{2}\|\Delta\|_1^2, \quad \forall \Delta \in \mathbb{R}^d. \tag{8}$$

Under $\mu/2 - 16s\tau > 0$, the RE (7) holds with $\delta_c = \mu/2 - 16s\tau$.

The practical utility of the RSC condition (8) vs. the RE is that it can be certified with high probability by a variety of random design matrices $\mathbf{X}$. Here we consider the following.

Assumption 1 (Random Gaussian model) *The design matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$ satisfies the following:* (i) *the rows of $\mathbf{X}$ are i.i.d. $\mathcal{N}(0, \Sigma)$; and* (ii) *Σ is positive definite, with minimum eigenvalue $\lambda_{\min}(\Sigma) > 0$.*

Lemma 2 ((Raskutti et al., 2010, Theorem 1)) *Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ be a design matrix satisfying Assumption 1. Then, there exist universal constants $c_0, c_1 > 0$ such that, with probability at least $1 - \exp(-c_0 N)$, the RSC condition (8) holds with parameters*

$$\mu = \lambda_{\min}(\Sigma) \quad \text{and} \quad \tau = 2c_1\zeta_\Sigma \frac{\log d}{N}, \quad \text{with } \zeta_\Sigma \triangleq \max_{i\in[d]} \Sigma_{ii}. \tag{9}$$

2.2 Network setting

We model the network of m agents as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = [m]$ is the set of agents, and $\mathcal{E}$ is the set of the edges; $\{i, j\} \in \mathcal{E}$ if and only if there is a communication link between agent i and agent j. We make the blanket assumption that $\mathcal{G}$ is connected, which is necessary for the convergence of distributed algorithms to a consensual solution.

To solve (4) over $\mathcal{G}$ via gradient descent, each agent should be able to compute the gradient of the objective (w.r.t. its own local variable θ_i) using only information from its immediate neighbours. This imposes some extra conditions on the sparsity pattern of the matrix V. We will use the following widely adopted structure for V.

Assumption 2 *The matrix $V = (I_m - W)\otimes I_d$, where $W \triangleq (w_{ij})_{i,j=1}^m$ satisfies the following:*

(a) It is compliant with $\mathcal{G}$, that is, (i) $w_{ii} > 0, \forall i \in [m]$; (ii) $w_{ij} > 0$, if $\{i, j\} \in \mathcal{E}$; and (iii) $w_{ij} = 0$ otherwise; and

(b) It is symmetric and stochastic, that is, $W1 = 1$ (and thus also $1^\top W = 1^\top$).

It follows from the connectivity of $\mathcal{G}$ and Assumption 2 that

$$V\boldsymbol{\theta} = \mathbf{0} \quad \text{iff} \quad \theta_i = \theta_j, \forall i \neq j \in [m],$$

and

$$\rho \triangleq \max\{|\lambda_2(W)|, |\lambda_{\min}(W)|\} < 1. \tag{10}$$

Roughly speaking, ρ measures how fast the network mixes information (the smaller, the faster). Note that if $\mathcal{G}$ is a complete graph or is a star network, one can choose $W = 11^\top/m$, resulting in $\rho = 0$.

3. Solution Analysis and Statistical Guarantees

This section is devoted to the solution analysis of the penalized LASSO problem (4), establishing nonasymptotic error bounds of $(1/m)\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2$. Our study builds on the following steps.

1) We first determine a suitable restricted set of directions $\mathbb{C}_\gamma(\mathcal{S})$ [cf. (11)] which contains the augmented LASSO error $\hat{\boldsymbol{\theta}} - 1_m \otimes \theta^*$ under certain conditions on the sparsity-enhancing parameter λ [cf. Proposition 3]–the set $\mathbb{C}_\gamma(\mathcal{S})$ plays similar role as $\mathbb{C}(\mathcal{S})$ [cf. (7)] for the centralized LASSO (2), and sheds light on the role of the penalty parameter γ (and thus the consensus errors) on the sparsity pattern of $\hat{\boldsymbol{\theta}}$;

2) We then determine a RSC-like condition [cf. (15)] ensuring that the subset $\mathbb{C}_\gamma(\mathcal{S})$ is well-aligned with the curved directions of the Hessian of the loss L of (4), under a suitable choice of γ controlling the consensus error;

3) Results in the previous steps will translate into bounds on $(1/m)\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2$ [cf. Theorem 6]. Quite interesting, our RSC condition holds w.h.p. under the random model in Assumption 1 (cf. Lemma 5), which will yield *centralized* sample complexity $(1/m)\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 = \mathcal{O}(s\log d/N)$ (cf. Theorem 7).

3.1 The set of (almost) sparse average directions

For each given $\gamma \in (0,1)$, define the set

$$\mathbb{C}_\gamma(\mathcal{S}) \triangleq \{\boldsymbol{\Delta} \in \mathbb{R}^{md} \mid \|(\Delta_{\text{av}})_{\mathcal{S}^c}\|_1 \leq 3\|(\Delta_{\text{av}})_{\mathcal{S}}\|_1 + h(\gamma, \|\boldsymbol{\Delta}_\perp\|)\}, \tag{11}$$

where

$$h(\gamma, \|\boldsymbol{\Delta}_\perp\|) \triangleq -\frac{1-\rho}{m\gamma\lambda}\|\boldsymbol{\Delta}_\perp\|^2 + \Big(2\max_{i\in[m]}\|w_i^\top X_i\|_\infty/(\lambda n) + 2\Big)\sqrt{d/m}\|\boldsymbol{\Delta}_\perp\|. \tag{12}$$

Note that the maximum of $h(\gamma, \bullet)$ over $\mathbb{R}_+$ is a decreasing function of $\gamma > 0$. This suggests that, the sparsity of the average component Δ_{av} of directions $\boldsymbol{\Delta} \in \mathbb{C}_\gamma(\mathcal{S})$ can be controlled by γ; in particular, by decreasing γ one can make Δ_{av} arbitrary "close" to the cone $\mathbb{C}(\mathcal{S})$ of sparse directions of the centralized LASSO (2) [cf. (7)]. The importance of $\mathbb{C}_\gamma(\mathcal{S})$ is captured by the following result.

Proposition 3 *Under Assumption 2 and λ satisfying*

$$\frac{2}{N}\|\mathbf{X}^\top \mathbf{w}\|_\infty \leq \lambda, \tag{13}$$

the augmented LASSO error $\hat{\boldsymbol{\theta}} - 1_m \otimes \theta^*$ lies in $\mathbb{C}_\gamma(\mathcal{S})$.

Proof See Appendix A. ∎

Therefore, the average component of the augmented LASSO error is nearly sparse for sufficiently small γ and large λ. This is a key property that will be used to achieve centralized statistical errors (cf. Sec. 3.3).

3.2 In-network RE condition

We impose a positive curvature on the loss L of (4) [cf. (6)] along suitable chosen directions in $\mathbb{C}_\gamma(\mathcal{S})$. The first-order Taylor expansion of L at $\boldsymbol{\theta}'$ along $\boldsymbol{\theta} - \boldsymbol{\theta}'$, denoted by $\mathcal{T}_L(\boldsymbol{\theta};\boldsymbol{\theta}')$, can be lower bounded as

$$\begin{aligned}\mathcal{T}_L(\boldsymbol{\theta};\boldsymbol{\theta}') &\triangleq L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}') - \langle \nabla L(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}'\rangle \\ &\geq \underbrace{\frac{1}{4}\frac{\|\mathbf{X}(\boldsymbol{\theta} - \boldsymbol{\theta}')_{\text{av}}\|^2}{N}}_{\text{curvature along average}} - \underbrace{\left(\frac{L_{\max}}{2m} - \frac{1-\rho}{2m\gamma}\right)\|(\boldsymbol{\theta} - \boldsymbol{\theta}')_\perp\|^2}_{\text{nonconsensual component}}.\end{aligned} \tag{14}$$

The second term on the RHS of (14) is due to the disagreement of the θ_i's, and can be controlled choosing suitably small γ. In fact, we will prove that a curvature condition as the first term of the RHS of (14) along the directions $\boldsymbol{\theta} - \boldsymbol{\theta}' \in \mathbb{C}_\gamma(\mathcal{S})$ is enough to establish the desired error bounds on the LASSO error $(1/m)\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2$. Specifically, the following RSC-like property suffices.

Assumption 3 (In-network RE) *The loss fuction L satisfies the following RSC condition with curvature $\delta > 0$ and tolerance $\xi > 0$:*

$$\frac{1}{N}\|\mathbf{X}\Delta_{\text{av}}\|^2 \geq \delta\|\Delta_{\text{av}}\|^2 - \xi\, h^2(\gamma, \|\mathbf{\Delta}_\perp\|), \quad \forall \mathbf{\Delta} \in \mathbb{C}_\gamma(\mathcal{S}). \tag{15}$$

reading again the previous paragraph, how about making the LHS of (15) explicit in terms of L , meaning write what the LHS is in terms of L. This would be more consistent with (14)Without the tolerance term, (15) would read as an RE property (strong convexity) of the loss L along the average component of directions in $\mathbb{C}_\gamma(\mathcal{S})$; the tolerance relaxes such requirement, asking for a positive curvature only along the directions in $\mathbb{C}_\gamma(\mathcal{S})$ such that the RHS of (15) is positive. Notice also that along consensual directions $\mathbf{\Delta} \in \mathbb{C}_\gamma(\mathcal{S})$–i.e., such that $\mathbf{\Delta}_\perp = \mathbf{0}$–(15) reduces to RE (7).

The following two results establish sufficient conditions for (15) to hold, for deterministic and random design matrices $\mathbf{X}$–which match those required for the centralized LASSO (see Lemma 1 and Lemma 2).

Lemma 4 *Reinstate Lemma 1, under $\mu/2 - 16s\tau > 0$. Then (15) holds, with $\delta = \mu/2 - 16s\tau$ and $\xi = \tau$, for any fixed $\gamma > 0$.*

Proof See Appendix B.1. ■

Lemma 5 *Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ be a design matrix satisfying Assumption 1. For any N and γ satisfy*

$$N \geq \frac{128 s c_1 \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)}, \quad \textit{and} \quad \gamma > 0, \tag{16}$$

the following holds

$$\frac{1}{N}\|\mathbf{X}\Delta_{\text{av}}\|^2 \geq \frac{\lambda_{\min}(\Sigma)}{4}\|\Delta_{\text{av}}\|^2 - \frac{\lambda_{\min}(\Sigma)}{64s} h^2(\gamma, \|\mathbf{\Delta}_\perp\|), \qquad \forall \mathbf{\Delta} \in \mathbb{C}_\gamma(\mathcal{S}), \tag{17}$$

with probability at least $1 - \exp(-c_0 N)$, where $c_0, c_1 > 0$ are universal constants.

Proof See Appendix B.2. ■

3.3 Error bounds and statistical consistency of the LASSO error of (4)

We are ready to establish consistency and convergence rates for the augmented LASSO estimator $\hat{\boldsymbol{\theta}}$. Our first result is a deterministic upper bound on the average error under the In-network RE condition (15).

Theorem 6 *Consider the augmented LASSO problem (4) under Assumptions 2 and 3. For any fixed λ and γ satisfying respectively*

$$\frac{2}{N}\|\mathbf{X}^\top \mathbf{w}\|_\infty \leq \lambda \quad \textit{and} \quad \gamma \leq \frac{2(1-\rho)}{4L_{\max} + \delta},$$

any solution $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, \ldots, \hat{\theta}_m]^\top$ *satisfies*

$$\frac{1}{m}\sum_{i=1}^{m}\|\hat{\theta}_i - \theta^*\|^2$$
$$\leq \underbrace{\frac{9\lambda^2 s}{\delta^2}}_{\text{centralized error}} + \underbrace{\frac{2\xi d^2\gamma^2(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^4}{\delta\lambda^2 n^4(1-\rho)^2} + \frac{4d\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^2}{\delta n^2[2(1-\rho) - 4L_{\max}\gamma - \delta\gamma]}}_{\text{cost of decentralization}}. \tag{18}$$

Proof See Appendix C. ■

Theorem 6 shows the bound on the LASSO error over the network can be decoupled in two terms–the first one matches that of the centralized LASSO error (see, e.g., (Wainwright, 2019, Th. 7.13))–while the second one quantifies the price to pay due to the decentralization of the optimization and consequent lack of consensus. The explicit dependence on γ shows that the detriment effect of the consensus errors can be controlled by γ: as $\gamma \to 0$, the error bound above approaches that of the centralized LASSO solution. There is however no free lunch; we anticipate that $\gamma \to 0$ affects adversarially the convergence rate of the proximal gradient algorithm applied to problem (4), determining thus a speed-accuracy dilemma.

The next result provides nonasymptotic rates for the LASSO error above, under the random Gaussian model for $\mathbf{X}$ and the noise $\mathbf{w}$ in (1)–optimal centralized convergence rates are achievable by a proper choice of γ.

Theorem 7 *Consider the augmented LASSO problem (4) with* $d \geq 2$ *under Assumptions 2. Suppose that the design matrix* $\mathbf{X}$ *satisfies Assumption 1 and* $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N)$*; the sample size satisfies*

$$N \geq \frac{c_9 s \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)}; \tag{19}$$

and the parameters λ *and* γ *are chosen according to the following*

$$\lambda = c_8\sigma\sqrt{\frac{\zeta_\Sigma t_0 \log d}{N}}, \tag{20}$$

$$\gamma \leq \frac{8n(1-\rho)c_7 s}{32c_4c_7 s\lambda_{\max}(\Sigma)[n + (d + \log m)] + \lambda_{\min}(\Sigma)n\left\{384d\left[c_{26}m\log m\max\left\{1, \frac{2\log md}{nc_{24}}\right\} + 1\right] + c_7 s\right\}}, \tag{21}$$

for some fixed $t_0 > 2$*. Then, any solution* $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, \ldots, \hat{\theta}_m]^\top$ *of problem (4) satisfies*

$$\frac{1}{m}\sum_{i=1}^{m}\|\hat{\theta}_i - \theta^*\|^2 = \left(144 + c_7 + \frac{c_7^2}{512}\right)\frac{c_8^2\sigma^2\zeta_\Sigma t_0}{\lambda_{\min}(\Sigma)^2}\frac{s\log d}{N} = \mathcal{O}\left(\frac{s\log d}{N}\right) \tag{22}$$

with probability at least

$$1 - 8\exp(-c_{25}\log d) - \exp(-c_0 N). \tag{23}$$

$c_0, c_3, c_4, c_6, c_7, c_8, c_9, c_{24}, c_{25}, c_{26}$ *are universal constants (their expression can be found in Table 4 in the appendix).*

Proof See Appendix D. ■

The error bound (22) matches the statistical error of the centralized LASSO estimator in (2)–proving that statistical consistency over networks is achievable under the same order of the sample size N used in the centralized setting, even when the local number n of samples does not suffice. This is possible because agents communicate over the network–the computation of such a solution and associated communication overhead is studied in the next section.

4. Distributed Gradient Descent Algorithm

To compute the statistically optimal estimator $\hat{\boldsymbol{\theta}}$ over networks, we employ the proximal gradient algorithm applied to the penalized formulation (4), which naturally decomposes across the agents. Specifically, at iteration t, $\boldsymbol{\theta}$ is updated by minimizing the first order approximation of the objective function L, which reads

$$\boldsymbol{\theta}^{t+1} = \underset{\|\theta_i\|_1 \leq R \ \forall i \in [m]}{\operatorname{argmin}} \ L(\boldsymbol{\theta}^t) + \langle \nabla L(\boldsymbol{\theta}^t), \boldsymbol{\theta} - \boldsymbol{\theta}^t \rangle + \frac{1}{2\beta m}\|\boldsymbol{\theta} - \boldsymbol{\theta}^t\|^2 + \frac{\lambda}{m}\|\boldsymbol{\theta}\|_1, \tag{24}$$

where we included an extra constraint $\|\theta_i\|_1 \leq R$ to regularize the iterates, and $\beta > 0$ plays the role of the stepsize. The following lemma shows that one can find a sufficiently large R so that the solution of (4) does not change if we add therein the norm ball constraint $\|\theta_i\|_1 \leq R$, $i \in [m]$.

Lemma 8 *Consider Problem (4) under Assumption 2. Further assume that (i) the design matrix* $\mathbf{X}$ *satisfies the RSC condition (8) with* $\delta = \mu/2 - 16\, s\tau > 0$*; (ii)* λ *satisfies (13); and (iii)* γ *satisfies*

$$\gamma \leq \frac{(1-\rho)}{2L_{\max} + \delta + 128(d/s)\delta(\max_{i\in[m]}\|w_i^\top X_i\|_\infty/(\lambda n) + 2\sqrt{m})^2}. \tag{25}$$

Then, $\|\hat{\theta}_i\|_1 \leq R$, $i \in [m]$, *whenever* R *is such that*

$$R \geq \max\left\{\frac{\lambda s}{\delta(1-r)}\left(13 + \frac{1}{32}\sqrt{\frac{2\tau s}{\delta}}\right), \frac{1}{r}\|\theta^*\|_1\right\}, \tag{26}$$

with $r \in (0,1)$.

Proof See Appendix E. ■

Therefore, we can focus on Problem (24) without loss of generality. Notice that the problem is separable in $\{\theta_i\}_{i\in[m]}$; hence, it can be solved distributively from each agent i. Furthermore, the solution can be computed in an explicit form, as determined next.

Lemma 9 *The solution to (24) reads*

$$\theta_i^{t+1} = \begin{cases} \operatorname{prox}_{\beta\lambda\|\cdot\|_1}(\psi_i^t), & \textit{if} \quad \left\|\operatorname{prox}_{\beta\lambda\|\cdot\|_1}(\psi_i^t)\right\|_1 \leq R, \\ \Pi_{\mathcal{B}_{\|\cdot\|_1}(R)}(\psi_i^t), & \textit{otherwise}; \end{cases} \tag{27}$$

where $\mathbb{R} \ni x \mapsto prox_h(x) \in \mathbb{R}$ *is the proximal operator (applied to* ψ_i^t *component-wise), and*

$$\psi_i^t = \left(1 - \frac{\beta}{\gamma}\right)\theta_i^t + \frac{\beta}{\gamma}\left(\sum_{j=1}^m w_{ij}\theta_j^t - \gamma\nabla f_i(\theta_i^t)\right).$$

Proof See Appendix F. ■

Notice that the proximal operation in (27) has a closed-form expression via soft-thresholding (Donoho, 1995) while the projection onto the ℓ_1-ball can be efficiently computed using the procedure in (Duchi et al., 2008). since β now is proportional to m and it will appear with terms divided by m, which will cancel the m, please go over the definition to remove such m from the beginning. For instance, would be enough to just drop the m from the denominator of $\lambda\|\cdot\|_1/m$?Notice I'm not asking to change the definition of the augmented function of the lifted problem [Yao: You mean keep the lifted problem as fixed, but change the definition of beta in (24) as (25), right? In that way, β will not be proportional to m.] exactly, why do you want β to be proportional to m while at the end what matters seems to be the product $\beta\lambda\|\cdot\|_1/m$. Since your β is proportional to m, that m cancels out with the one at the denominator To perform the update (27), each agent i only needs to receive the local estimates θ_j^t from its immediate neighbors.

4.1 Linear convergence to statistical precision

Our convergence rate on the optimization error $\boldsymbol{\theta}^t - \hat{\boldsymbol{\theta}}$ is stated under the RSC condition (8) (with parameters μ and τ), in terms of the contraction coefficient κ and initial optimality gap η_G^0:

$$\mu_{\text{av}} \triangleq \frac{\mu}{8} - 8s\tau, \quad \kappa \triangleq 1 - \frac{\beta\mu_{\text{av}}}{4}, \quad \text{and} \quad \eta_G^0 \triangleq \left(L(\boldsymbol{\theta}^0) + \frac{\lambda}{m}\|\boldsymbol{\theta}^0\|_1\right) - \left(L(\hat{\boldsymbol{\theta}}) + \frac{\lambda}{m}\|\hat{\boldsymbol{\theta}}\|_1\right). \tag{28}$$

Further, denote

$$\varepsilon_{\text{stat}}^2 \triangleq \frac{36}{m}\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 + \frac{\lambda^2 s}{1976\mu^2}. \tag{29}$$

We can now state our convergence result.

Theorem 10 *Consider the augmented LASSO problem (4) under Assumptions 2. Suppose the design matrix* $\mathbf{X}$ *satisfies the RSC condition* (8) *with* $\mu \geq c_{10}s\tau$ *for some sufficiently large constant* $c_{10} > 0$*; and the penalty parameters* λ *and* γ *satisfies*

$$\lambda \geq \max\left\{\frac{2\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N}, 64\tau\|\theta^*\|_1\right\}, \tag{30}$$

$$\gamma \leq \frac{1-\rho}{2L_{\max} + (\mu/2 - 16s\tau)\left(1 + 128(d/s)(\max_{i\in[m]}\|w_i^\top X_i\|_\infty/(\lambda n) + 2\sqrt{m})^2\right)}, \tag{31}$$

respectively. Let $\{\theta_i^t\}_{i\in[m]}$ *be the sequence generated by Algorithm (27) under the following choices of tuning parameters* β *and* R

$$\beta = \frac{\gamma}{\gamma L_{\max} + 1 - \lambda_{\min}(W)} \tag{32}$$

and

$$\max\left\{\frac{56\lambda s}{\mu - 32s\tau}, 2\|\theta^*\|_1\right\} \leq R \leq \frac{\lambda}{32\tau}. \tag{33}$$

Further assume

$$\eta_G^0 \geq 4s\tau \cdot \varepsilon_{\text{stat}}^2. \tag{34}$$

Then, there holds

$$\frac{1}{m}\sum_{i=1}^{m}\|\theta_i^t - \hat{\theta}_i\|^2 \leq \frac{\tau s}{\mu_{\text{av}}} \cdot \varepsilon_{\text{stat}}^2 + \left(\frac{\tau s}{\mu_{\text{av}}}\frac{4\alpha^4}{\lambda^2 s} + \frac{\alpha^2}{\mu_{\text{av}}}\right), \tag{35}$$

for any tolerance parameter α^2 *such that*

$$\min\left\{\frac{R\lambda}{4}, \eta_G^0\right\} \geq \alpha^2 \geq 4s\tau \cdot \varepsilon_{\text{stat}}^2, \tag{36}$$

and for all

$$t \geq \left\lceil \log_2 \log_2\left(\frac{R\lambda}{\alpha^2}\right)\right\rceil \left(1+\frac{L_{\max}\log 2}{\mu_{\text{av}}} + \frac{(1+\rho)\log 2}{\gamma\mu_{\text{av}}}\right) + \left(\frac{L_{\max}}{\mu_{\text{av}}} + \frac{1+\rho}{\gamma\mu_{\text{av}}}\right)\log\left(\frac{\eta_G^0}{\alpha^2}\right). \tag{37}$$

The intervals in (33) *and* (36) *are nonempty.*

Proof See Appendix G. ■

The theorem shows that Algorithm (27) converges at a linear rate to an optimal solution $\hat{\boldsymbol{\theta}}$, up to a tolerance term as specified on the right hand side of (35)–the first term therein depends on the model parameters, while the second one is controlled by α^2. Theorem 12 and (see also Corollary 13) below proves that for the random Gaussian data generation model, the tolerance can be driven below the statistical precision for sufficiently large N.

Remark 11 *Algorithm* (27) *is closely related to the DGD algorithm studied in the literature of distributed optimization (e.g., (Zeng and Yin, 2018, Nedić et al., 2018)). In fact, if in (27) one choose* $\beta = \gamma/2$, *with* γ *satisfying* (31) *[note that this choice of* β *is compatible with* (32)*], the gradient step therein reduces to*

$$\psi_i^t = \frac{1}{2}\left(\theta_i + \sum_{i=1}^{m} w_{ij}\theta_j^t\right) - \frac{\gamma}{2}\nabla f_i(\theta_i^t), \tag{38}$$

which can be viewed as DGD with weight matrix $\frac{1}{2}(I+W)$ *and step size* $\gamma/2$.

Theorem 12 *Consider the augmented LASSO problem (4) with* $d \geq 2$ *under Assumption 2. Suppose the design matrix* $\mathbf{X}$ *satisfies Assumption 1,* $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N)$, *and*

$$N \geq \frac{c_{12} s\zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)}. \tag{39}$$

Choose the penalty parameters λ and γ satisfying respectively

$$\lambda \geq \max\left\{\sigma\sqrt{\frac{6\zeta_\Sigma t_0 \log d}{N}}, \frac{128 s c_1 \zeta_\Sigma \log d}{N}\right\}, \tag{40}$$

and

$$\gamma \leq \frac{1-\rho}{2c_4\lambda_{\max}(\Sigma)\left(1+\frac{d+\log m}{n}\right)+128\lambda_{\min}(\Sigma)dm\cdot\left[\left(2\log m+1\right)/c_{24}+8\right]\max\left\{\frac{2\log md}{nc_{24}},1\right\}}. \tag{41}$$

for some fixed $t_0 > 2$. Let $\{\theta_i^t\}_{i\in[m]}$ be the sequence generated by Algorithm (27) under the following choices of tuning parameters β and R

$$\beta = \frac{\gamma}{\gamma c_4 d/n + 1 - \lambda_m(W)}, \quad \max\left\{\frac{56\lambda s}{\lambda_{\min}(\Sigma) - 64 s c_1 \zeta_\Sigma \log d/N}, 2s\right\} \leq R \leq \frac{\lambda N}{64 c_1 \zeta_\Sigma \log d}, \tag{42}$$

and

$$\eta_G^0 \geq \frac{11339 s c_1 \zeta_\Sigma \log d}{N\lambda_{\min}(\Sigma)^2}\lambda^2 s. \tag{43}$$

Then, with probability at least (23), there holds

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^m \|\theta_i^t - \hat{\theta}_i\|^2\\ \leq&\frac{9}{\lambda_{\min}(\Sigma)}\left(\alpha^2 + \frac{72 s c_1 \zeta_\Sigma \log d}{N}\frac{1}{m}\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 + \frac{s c_1 \zeta_\Sigma \log d}{988N}\frac{\lambda^2 s}{\lambda_{\min}(\Sigma)^2} + \frac{8 s c_1 \zeta_\Sigma \log d}{N}\frac{\alpha^4}{\lambda^2 s}\right),\end{aligned} \tag{44}$$

for any tolerance parameter α^2 such that

$$\min\left\{\frac{R\lambda}{4}, \eta_G^0\right\} \geq \alpha^2 \geq \frac{8 s c_1 \zeta_\Sigma \log d}{N}\left(\frac{36}{m}\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 + \frac{\lambda^2 s}{1976\lambda_{\min}(\Sigma)^2}\right), \tag{45}$$

and for all

$$\begin{aligned}t \geq& \left\lceil \log_2\log_2\left(\frac{R\lambda}{\alpha^2}\right)\right\rceil\left(1 + \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{456 c_4[1+(d+\log m)/n]\log 2}{55} + \frac{456(1+\rho)\log 2}{55\lambda_{\min}(\Sigma)\gamma}\right)\\ &+\left(\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{456 c_4[1+(d+\log m)/n]}{55} + \frac{456(1+\rho)}{55\gamma\lambda_{\min}(\Sigma)}\right)\log\left(\frac{\eta_G^0}{\alpha^2}\right).\end{aligned} \tag{46}$$

c_1, c_4 and c_{12} are universal constants (their expression can be found in Table 4 in the appendix). The range of values of R in (42) is nonempty; and the interval in (45) is nonempty as well, with probability at least (23).

Proof See Appendix H. ■

A suitable choice of the free parameters above leads to the following simplified result, showing linear convergence up to a tolerance that can be made of a higher order than the statistical error.

Corollary 13 *Consider the augmented LASSO problem (4) with $d \geq 2$ under Assumptions 2. Suppose the design matrix $\mathbf{X}$ satisfies Assumption 1, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N)$ and the sample size satisfies*

$$N \geq \max\left\{\frac{c_{12} s \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)}, \frac{c_{13} s^2 \zeta_\Sigma \log d}{\sigma^2}\right\} \quad \textit{and} \quad \frac{d + \log m}{n} \geq 1. \tag{47}$$

Choose the penalty parameters λ and γ satisfying respectively

$$\lambda = c_8 \sigma \sqrt{\frac{\zeta_\Sigma t_0 \log d}{N}} \tag{48}$$

and

$$\gamma \leq \frac{1-\rho}{4c_4 \lambda_{\max}(\Sigma) + 128 \lambda_{\min} dm \cdot \left[(2\log m + 1)/c_{24} + 8\right] \max\left\{\frac{2\log md}{nc_{24}}, 1\right\}}. \tag{49}$$

for some fixed $t_0 \geq 2$. Let $\{\theta_i^t\}_{i\in[m]}$ be the sequence generated by Algorithm (27) under the following choices of β and R

$$\beta = \frac{\gamma}{\gamma c_4 d/n + 1 - \lambda_m(W)}, \quad \max\left\{\frac{c_{15} s \sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{t_0 \zeta_\Sigma \log d}{N}}, 2s\right\} \leq R \leq c_{16}\sigma\sqrt{\frac{t_0 N}{\zeta_\Sigma \log d}}, \tag{50}$$

and

$$\eta_G^0 \geq c_{22}\sigma^2 t_0 \left(\frac{s\zeta_\Sigma \log d}{N\lambda_{\min}(\Sigma)}\right)^2. \tag{51}$$

Then, with probability at least (23), we have

$$\frac{1}{m}\sum_{i=1}^m \|\theta_i^t - \hat{\theta}_i\|^2$$

$$\leq \frac{c_{17}}{\lambda_{\min}(\Sigma)}\alpha^2 + \frac{sc_{18}\zeta_\Sigma \log d}{\lambda_{\min}(\Sigma) N}\frac{1}{m}\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 + \frac{sc_{19}\sigma^2\zeta_\Sigma^2 t_0 \log d}{\lambda_{\min}(\Sigma)N}\frac{s\log d}{\lambda_{\min}(\Sigma)^2 N} + \frac{c_{20}}{\lambda_{\min}(\Sigma)\sigma^2 t_0}\alpha^4, \tag{52}$$

for any tolerance parameter α^2 such that

$$\min\left\{\frac{R\sigma c_8}{4}\sqrt{\frac{\zeta_\Sigma t_0 \log d}{N}}, \eta_G^0\right\} \geq \alpha^2 \geq \frac{8sc_1\zeta_\Sigma \log d}{N}\left(\frac{36}{m}\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 + \frac{c_{21}\sigma^2\zeta_\Sigma t_0}{\lambda_{\min}(\Sigma)^2}\frac{s\log d}{N}\right), \tag{53}$$

and for all

$$t \geq \left\{\left\lceil \log_2 \log_2\left(\frac{R\sigma c_8}{\alpha^2}\sqrt{\frac{\zeta_\Sigma t_0 \log d}{N}}\right)\right\rceil \log 2 + \log\left(\frac{\eta_G^0}{\alpha^2}\right)\right\} \cdot \frac{dm}{1-\rho}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \cdot c_{23}\left(\frac{2\log m + 1}{c_{24}} + 8\right)\max\left\{\frac{2\log md}{nc_{24}}, 1\right\}. \tag{54}$$

The range of value of R in (50) is nonempty; and the interval in (53) is nonempty as well, with probability at least (23). The expression of the universal constants appearing above can be found in Table 4 in the appendix.

Proof See Appendix I. ■

It is not difficult to check that, in the above setting, Theorem 7 holds (in particular, (48) implies (21), whenever the free parameters $c_7 \geq 1/8$); hence, by (22), we have $\frac{1}{m}\sum_{i=1}^{m}\|\hat{\theta}_i - \theta^*\|^2 = \mathcal{O}(\frac{s\log d}{N})$. Therefore, whenever the sample size $N = o(s\log d)$–a condition that is required for statistical consistency of any centralized method by minimax results (see, e.g., (Raskutti et al., 2011)), the (lower bound of the) tolerance α^2 in (53) and thus the overall residual error in (52) is of smaller order than the statistical error $\mathcal{O}(\frac{s\log d}{N})$. Therefore, in this setting, a total number of communications (iterations) of here

$$\mathcal{O}\left(\frac{1}{1-\rho}\cdot\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\cdot d\, m\,\log m\cdot\log\frac{1}{\alpha^2}\right) \tag{55}$$

is sufficient to drive the iterates generated by Algorithm (27) within $\mathcal{O}(\alpha^2)$ of an optimal solution $\widehat{\boldsymbol{\theta}}$ (in the sense of (52)), and thus to an estimate of θ^* within the statistical error. This matches centralized statistical accuracy achievable by the LASSO estimator $\hat{\theta}$ in (2).

The expression (55) sheds light on the impact of the problem and network parameters on the convergence. Specifically, the following comments are in order.

(i) **Network dependence/scaling:** The term $1/(1-\rho) > 1$ captures the effect of the network; as expected, weakly connected networks (i.e., as $\rho \to 1$) call for more rounds of communication to achieve the prescribed accuracy. Recall that $\rho = \rho(m)$ is a function of the number of agents m (and the specific topology under consideration). Hence, the term

$$\frac{m\log m}{1-\rho(m)} \tag{56}$$

shows how the number of rounds of communications on the network scales with m, for a given graph topology (determining $\rho(m)$). Table 2 provides some estimates of the scaling of $1/(1-\rho(m))$ for different graphs, when the Metropolis-Hastings rule is used for the gossip matrix W (Nedić et al., 2018). Some graphs, for instance the Erdős-Rényi, exhibit a more favorable scaling than others, such as line graphs. Note that (56) does not capture the overall *cost* of communications, which also depends on how dense the graph is. For instance, counting as one channel use per communication, with each edge shared between two nodes, the Erdős-Rényi graph (with $p = \log m/m$) calls for a total channel use (i.e., across all nodes) per communication round of $\mathcal{O}(m\log m)$. In contrast, the complete graph requires $\mathcal{O}(m^2)$ channel uses per communication round, even though both graphs display a scaling of (56) with m, and thus the total number of communication rounds, of the same order. See Fig. 5 in Sec. 5 for an experimental exploration on (56).

(ii) **Population condition number:** The ratio $\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ is the condition number of the covariance matrix of the data; it can be interpreted as the restricted condition number of the LASSO loss function $F(\theta)$ [see (6)]. Therefore, as expected, ill-conditioned problems call for more iterations (communication) to achieve the prescribed accuracy.

(iii) **Speed-accuracy dilemma:** We proved that centralized statistical accuracy, as for the LASSO estimator $\hat{\theta}$ in (2), is achievable over networks via the distributed algorithm (27), even if agents locally do not have enough data for statistically consistency. This is possible thanks to the "help" of the network via information mixing. However, (55) shows that, no matter how fast the network propagates information (how small ρ is), the total communication rounds needed to achieve a prescribed accuracy scales as $\mathcal{O}(d)$. Our numerical results will support the tightness of such a scaling. We thus discover that the DGD-like scheme suffers in the high-dimensional regime of similar speed-accuracy dilemma as known when applied to strongly convex, smooth losses (low dimensional case) (Nedić et al., 2018). This seems to be unavoidable and a consequence of the structural updates employed by the algorithm.

The speed-accuracy dilemma arises from the penalty approach (4), where the parameter γ balances the average consensus and local optimization dynamics. Achieving exact consensus requires a diminishing step-size $\gamma = \gamma^k$, resulting in sublinear convergence rates to the solution of the original problem (3). In contrast, using a constant stepsize leads to faster (linear) rates but produces inexact solutions of (3) (with decreasing error as γ decreases).I do not see the point of discussing here the diminishing stepsize. ALso, your last comment above about inexact solutions does not justify the speed-accuracy dilemma. Actually the point is exactly that we may not need consensus to acheive statistical optimality. What you wrote abote does not explain anything To explain the speed-accuracy dilemma in the high-dimensional case, we show in Theorem 7 that choosing the consensus regularization parameter γ to scale as $\mathcal{O}(1/d)$ (as in equation (21)) ensures that the solution of the penalized problem lies within the centralized statistical error $\mathcal{O}(s \log d/N)$, which is minimax optimal in high-dimensional cases. Moreover, in Corollary 13, we demonstrate that with such a choice of γ, (27) yields a linear convergence rate up to the centralized statistical error ball, with the rate depending linearly on the dimension d.

	path	2-d grid	complete	p-Erdős-Rényi	p-Erdős-Rényi
$(1-\rho)^{-1}$	$\mathcal{O}(m^2)$	$\mathcal{O}(m \log m)$	$\mathcal{O}(1)$	$\mathcal{O}(1)$ $[p = \log m/m]$	$\mathcal{O}(1)$ $[p = \mathcal{O}(1)]$

Table 2: Scaling of $(1-\rho(m))^{-1}$ with agents' number m, for different graph topologies.

5. Numerical Results

In this section, we provide some experiments on synthetic and real data. The former are instrumental to validate our theoretical findings. Specifically, in this regard, we provide three sets of validating simulations. 1) On the statistical error front, we show that with a proper choice of γ, the solution of the distributed formulation (4) achieves the statistical accuracy of the centralized LASSO estimator (Theorem 7). We also validate the dependency of γ with the dimension d [cf. (21)]. 2) On the computational front, we demonstrate that the DGD-like algorithm (27) displays linear convergence up to the level of statistical precision.

3) Finally, combining these results, we illustrate the speed-accuracy dilemma, as predicted by Theorem 12. We then proceed to experiment on real data, showing that statistical accuracy of the centralized LASSO estimator (Theorem 7) is achievable by the distributed method (4), still at the cost of a convergence rate scaling with $\mathcal{O}(d)$. All the experiments were run on a server equipped with Intel(R) Xeon(R) CPU E5-2699A v4 @ 2.40GHz.

Experimental setup (synthetic data): The ground truth θ^* is set by randomly sampling a multivariate Gaussian $\mathcal{N}(0, I_d)$ and thresholding the smallest $d - s$ elements to zero. The noise vector $\mathbf{w}$ is assumed to be multivariate Gaussian $\mathcal{N}(\mathbf{0}, 0.25 I_N)$. We construct $\mathbf{X} \in \mathbb{R}^{N \times d}$ by independently generating each row $x_i \in \mathbb{R}^d$, adopting the following procedure (Agarwal et al., 2012): let $z_1, \ldots, z_{d-1}$ be i.i.d. standard normal random variables, set $x_{i,1} = z_1/\sqrt{1 - 0.25^2}$ and $x_{i,t+1} = 0.25 x_{i,t} + z_t$, for $t = 1, 2, \ldots, d-1$. It can be verified that all the eigenvalues of $\Sigma = \mathrm{cov}(x_i)$ lie within the interval $[0.64, 2.85]$. We partition $(\mathbf{X}, \mathbf{y})$ as $\mathbf{X} = [X_1^\top, X_2^\top, \ldots, X_m^\top]^\top$ and $\mathbf{y} = [y_1^\top, \ldots, y_m^\top]^\top$, and agent i owns the dataset portion (X_i, y_i) and we have m agents in total. We simulate an undirected graph $\mathcal{G}$ following the Erdös-Rényi model $G(m, p)$, where m is the number of agents and p is the probability that an edge is independently included in the graph. The coefficient of the matrix W are chosen according to the Lazy Metropolis rule (Olshevsky, 2017). All results are using Monte Carlo with 30 repetitions.

1) Statistical accuracy verification (Theorem 7). We set $N = 220, m = 20, d = 400$, and consider two types of graphs, namely a fully connected ($\rho = 0.4897$) and a weakly connected graph, the latter generated as Erdös-Rényi graph with edge probability $p = 0.1$, resulting in $\rho \approx 0.95$.

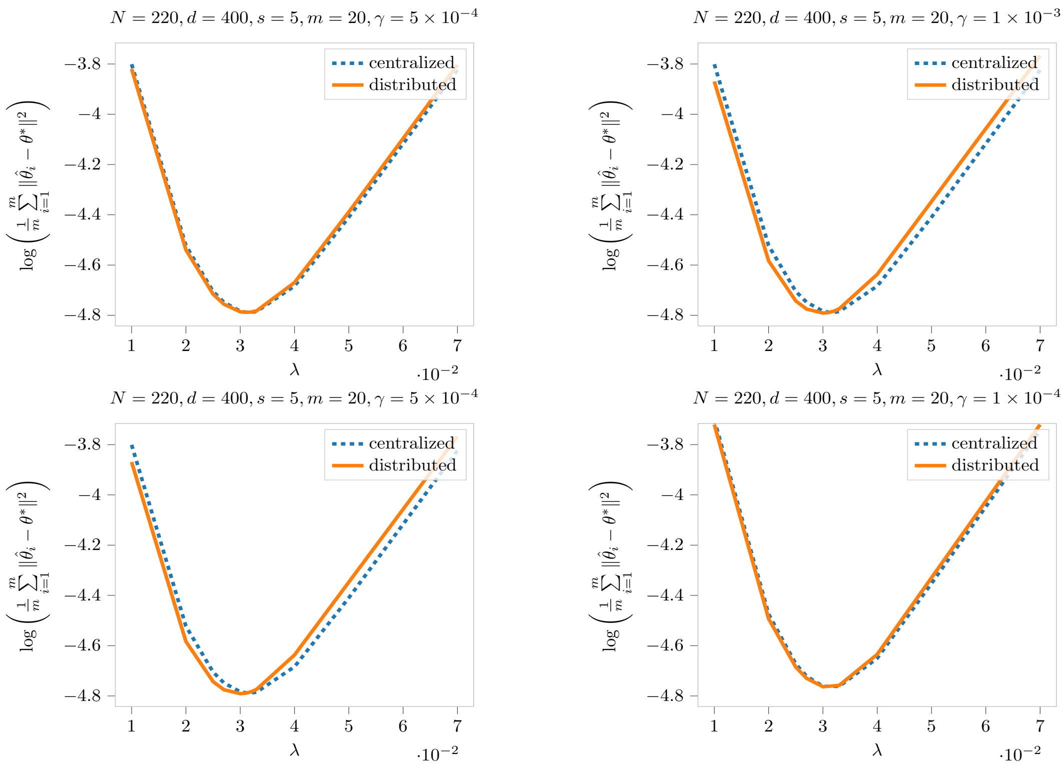

Figure 2: Statistical error of the estimator $\widehat{\boldsymbol{\theta}}$ [see (4)] and the centralized LASSO estimator $\hat{\theta}$ [see (2)] versus λ, using synthetic data; **First row:** fully connected graph ($\rho = 0.4897$); **Second row:**. Erdös-Rényi graph with $p = 0.1$, ($\rho \approx 0.95$). Notice that our theory explains the behaviour of the curves only for values of $\lambda \geq 0.033$ (as required by (21)).

Fig. 2 plots the log-average statistical error $\log\left(\sum_{i=1}^{m}\|\hat{\theta}_i - \theta^*\|^2/m\right)$ versus λ for the fully- and weakly-connected graphs (resp.), and contrasts it with the centralized LASSO log-error $\log\left(\|\hat{\theta} - \theta^*\|^2\right)$. The following comments are in order. **(i)** A careful choice of γ $(= 5 \times 10^{-4})$ is required to ensure that the distributed penalty LASSO recovers the centralized ℓ_2-error; however, with the same choice of γ, the solution achieved by the distributed method (4) over weakly connected graph can not recover the the statistical accuracy of the centralized LASSO estimator. **(ii)** The weakly connected graph requires a smaller γ $(= 1 \times 10^{-4})$ to recover the centralized statistical error; which is consistent with the dependence of γ on ρ as in (21). **(iii)** The range of λ guaranteeing the minimal ℓ_2- error in both the centralized and distributed penalty LASSO is comparable, as predicted by condition (19) on λ.

2) Validating $\gamma = \mathcal{O}(1/d)$ in (21) **(Theorem 7).** Fig. 3 plots the log-average statistical error $\log\left(\sum_{i=1}^{m}\|\hat{\theta}_i - \theta^*\|^2/m\right)$ versus γ, for three choices of (N, d, s, m), corresponding to increasing values of d and roughly constant $s \log d/N$ (and so the centralized statistical error). The left panel shows results on a fully connected graph while the right panel on an Erdös-Rényi graph with $p = 0.1$, ($\rho \approx 0.87$). As reference, we also plot the statistical error of the centralized LASSO estimator $\hat{\theta}$ (dashed-line curves). The figure shows that, as d increases, a smaller γ is required to preserve the centralized statistical estimates (resulting in the vertical dashed line). The scaling of such a γ is roughly $\mathcal{O}(1/d)$–slower scaling will result in increasing statistical errors, validating the dimension-dependence of recovery predicted in (21). Notice also that a weaker connected graph requires smaller γ to recover the centralized statistical error, which is consistent with the dependence of γ on ρ as proved in (21).

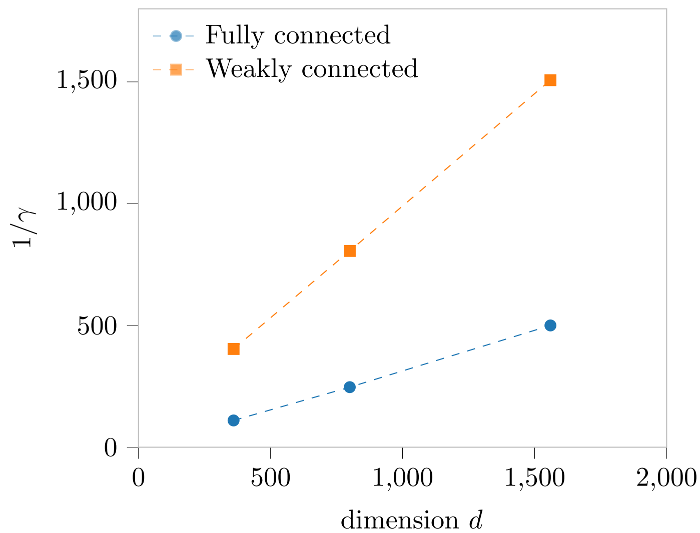

Figure 3: Validating $\gamma = \mathcal{O}((1-\rho)/d)$ as in (21): Inverse of critical γ to retain centralized statistical consistency versus dimension d, using synthetic data; As d varying, the critical value of $1/\gamma$ depends linearly on d. The slope of the weakly connected network is larger than that of the fully connected one, providing further validation for the scaling of $\gamma = \mathcal{O}((1-\rho)/d)$.

3) Linear convergence and the speed-accuracy dilemma (Theorem 12). Fig. 4 plots the log average optimization error versus the number of iterations generated by the distributed proximal-gradient Algorithm (27), in the same setting of Fig. 3. As predicted by Theorem 12, linear convergence within the centralized statistical error is achievable when $d, N \to \infty$ and $s \log d/N = o(1)$, but at a rate scaling with $\mathcal{O}(d)$, revealing the the speed-accuracy dilemma.

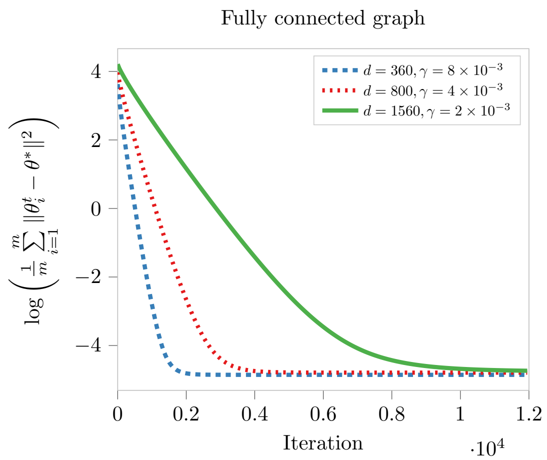

Figure 4: Linear convergence of Algorithm (27) up to the centralized statistical error: estimation error generated by Algorithm (24) versus iterations (communications), using synthetic data. **Left panel:** fully connected graph ($\rho = 0.4897$); **Right panel:**. Erdös-Rényi graph with $p = 0.1$, ($\rho \approx 0.87$). As predicted by our theory, the scaling of γ to recover centralized statistical consistency is $\gamma = \Theta(1/d)$: As d roughly doubles, going from 360 to 800, γ decreases by half. The same scaling is observed when d goes from 800 to 1560, revealing the the speed-accuracy dilemma.

4) Experimental exploration on the dependence of communications on network size m. We conducted experiments on five different network typologies, including three deterministic graphs (i.e., complete graph, path graph, and star graph) and two random graphs (i.e., Erdös-Rényi graph with $p = \mathcal{O}(1)$ and $p = \mathcal{O}(\log m/m)$). For each topology, we gradually increased the number of nodes from $m = \{10, 25, 40, 50, 100\}$ while keeping the total sample size N fixed at 200 and the dimension d fixed at 400. We performed a grid search on the parameter γ, and plotted the critical communications for each combination of m and graph type, which allows us to reach the centralized statistical error. Note for Erdös-Rényi graph with $p = \mathcal{O}(\log m/m)$ and $p = \mathcal{O}(1)$, we choose $p = \log m/m$ and $p = 0.1$, under such choices of p, the connectivity of the Erdös-Rényi graph are of $\mathcal{O}(1)$ with high probability (Nedić et al., 2018, Proposition 5).

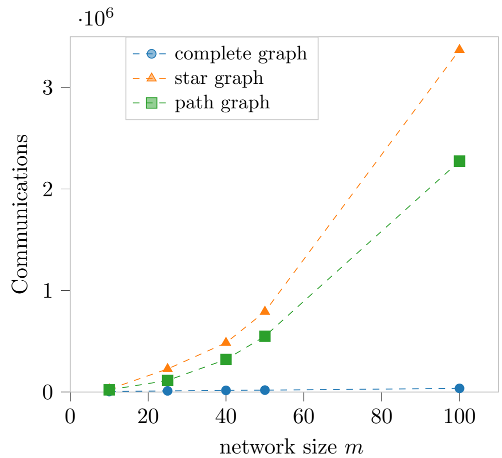

Figure 5: Network size m versus the number of communications to reach the centralized accuracy, using synthetic data. It shows that the communication cost for certain typologies, such as path and star graphs, grows polynomially (more than linearly) with the number of nodes.

Figure 5 illustrates the relationship between the number of nodes m and the number of communications required for different graph topologies, with fixed values of s, d, and N. While our theory cannot provide a tight linear bound on the best graph topology (i.e., complete graph), it does show that the communication cost for general topologies, such as path and star graphs, grows polynomially (more than linearly) with the number of nodes. This observation partially justifies our theoretical formula for the universal dependence on m of $m \log m/(1-\rho(m))$ in (56), which is reasonably good, but we cannot prove that it is tight because we cannot derive a lower bound. However, for some unfavorable topologies, we can conclude that the communication cost grows polynomially fast.

Experiment on real data. We test our findings on the dataset `eyedata` in the `NormalBetaPrime` package (Bai and Ghosh, 2019). This dataset contains gene expression data of $d = 200$ genes, and $N = 120$ samples. Data originate from microarray experiments of mammalian eye tissue samples. We randomly divide the dataset into training sample set with size $N_{\text{train}} = 80$ and test dataset with size $N_{\text{test}} = 40$. We partition the training data into $m = 10$ subsets. Each agent i owns the data set portion with size 8. We run Monte Carlo simulations, with 30 repetitions. Since we do not have access of the ground truth θ^*, we replace the ℓ_2 statistical error and the ℓ_2 optimization error with the MSE errors

$$\texttt{MSE}^{\infty} \triangleq \frac{1}{mN_{\text{test}}} \sum_{i=1}^{m} \|y^*_{\text{test}} - \hat{y}_i\|^2 \quad \text{and} \quad \texttt{MSE}^{t} \triangleq \frac{1}{mN_{\text{test}}} \sum_{i=1}^{m} \|y^*_{\text{test}} - y^t_i\|^2, \tag{57}$$

respectively, where y^*_{test} is the output of the test set, and $\hat{y}_i = X_i \hat{\theta}_i$, $i \in [m]$, are the model forecasts; $\hat{y}^t_i = X_i \theta^t_i$, $i \in [m]$, are output at iteration t. $m = 1$ corresponds to the centralized case, with $\hat{\mathbf{y}} = \mathbf{X}\hat{\theta}$.

Our first experiment is meant to check whether the solution of the penalized problem (4) matches the solution of the centralized LASSO via a proper choice of γ. Fig. 6 plots the MSE (log scale) vs. γ achieved by Algorithm (27) over a fully connected graph (**left panel**) and a weakly Erdös-Rényi graph with $p = 0.1$, resulting in $\rho \approx 0.95$ (**right panel**). The results confirm what we have already observed on synthetic data.

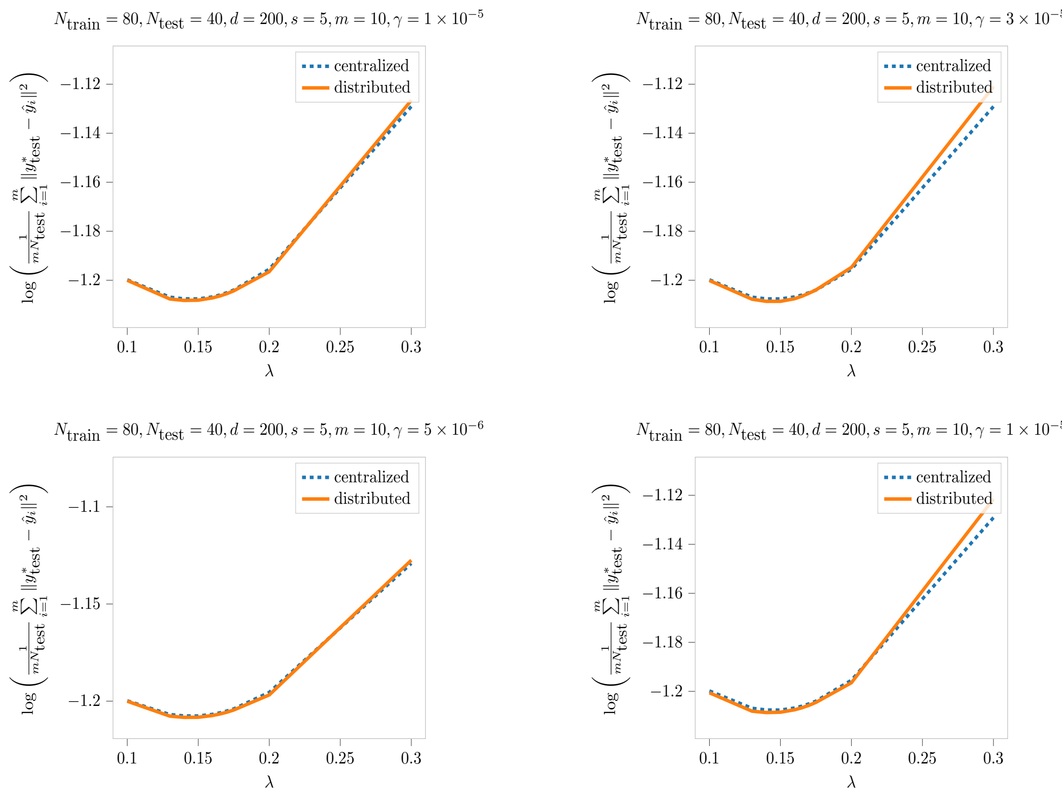

Figure 6: $\texttt{MSE}^{\infty}$ defined in (57) associated with the estimator $\widehat{\boldsymbol{\theta}}$ [see (4)] and the centralized LASSO estimator $\hat{\theta}$ [see (2)] versus λ using the dataset `eyedata` in the `NormalBeta-Prime` package. **First row:** fully connected graph ($\rho = 0.4897$); **Second row:** Erdös-Rényi graph with $p = 0.1$, ($\rho \approx 0.96$). Notice that our theory explains the behaviour of the curves only for values of $\lambda \geq 0.15$ (as required by (21)).

Our second experiment on real data is to validate the speed-accuracy dilemma, postulated by our theory and already validated on synthetic data (cf. Fig.4). Fig. 7 plots the log average optimization error versus the number of iterations generated by Algorithm (27), in the same network setting as for Fig. 6; different curves refer to different values of the penalty parameter γ. Since θ^* is no longer available when using real data, we heuristically set R in the projection (27) as $R = \max_{1 \leq i \leq m} \|\hat{\theta}_i\|_1$. This max-quantity can be obtained locally by each agent by running a min-consensus algorithms, requiring a number of communications of the order of the diameter of the network. The figure still shows linear convergence up to some tolerance, which is of the order of the MSE error in (57). Even on real data the speed-accuracy dilemma is evident.

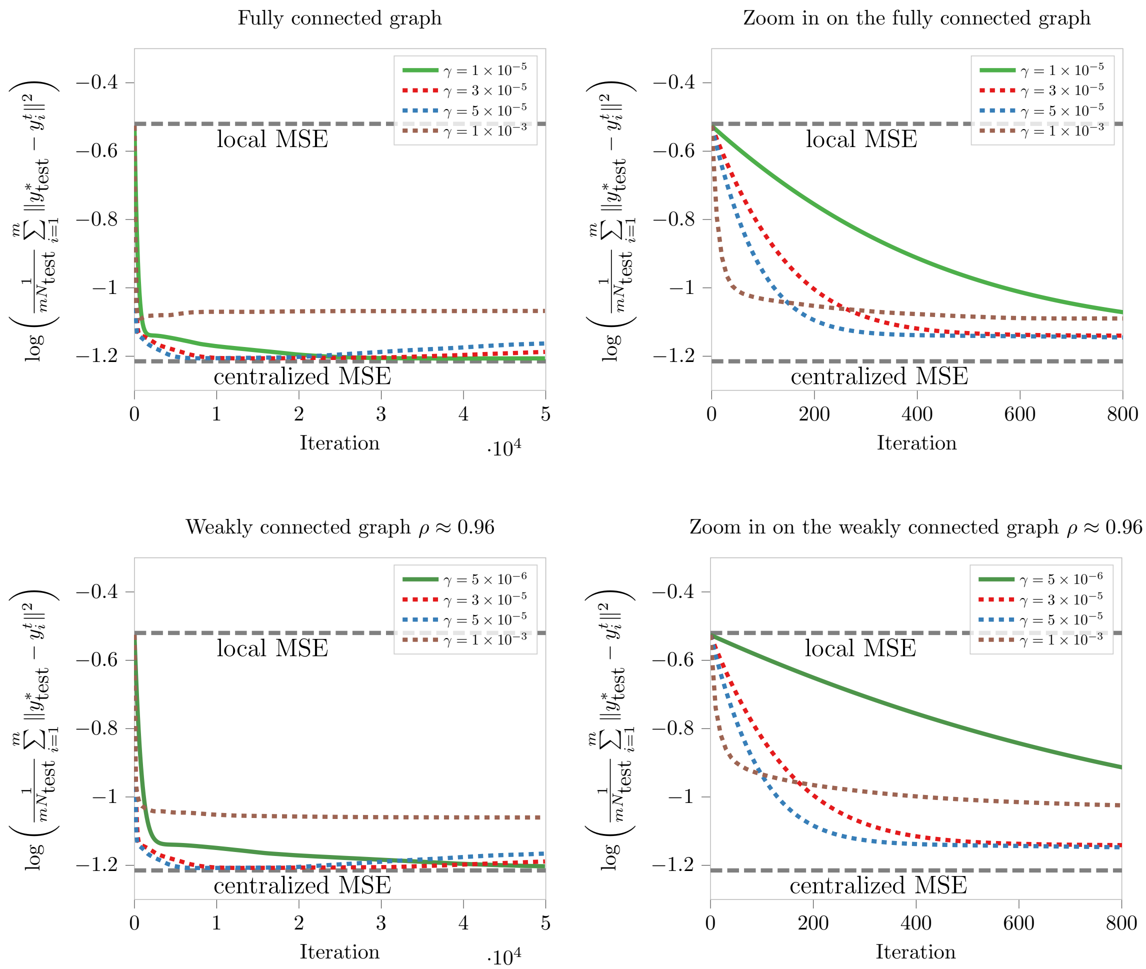

Figure 7: Linear convergence of Algorithm (27) up to the centralized statistical error, for different values of γ, using the dataset `eyedata` in the `NormalBeta-Prime` package: $\texttt{MSE}^t$ defined in (57) versus the number of iterations (communications). **First row:** fully connected graph ($\rho = 0.4897$); **Second row:** Erdös-Rényi graph with $p = 0.1$, ($\rho \approx 0.96$). **Left panel:** iterations up to 5×10^4. **Right panel:** zoom in on the iterations up to 8×10^2.

6. Concluding Remarks

We studied sparse linear regression over mesh networks. We established statistical and computational guarantees in the high-dimensional regime of a penalty-based consensus formulation and associated distributed proximal gradient method. This is the first attempt of studying the behaviour of a distributed method in the high-dimensional regime; our interest in penalty-based formulations to decentralize the optimization/estimation was motivated by their popularity and early adoption in the literature of distributed optimization (low-dimensional regime). We proved that optimal sample complexity $\mathcal{O}(s \log d/N)$ for the distributed estimator is achievable over networks, even when *local sample size is not sufficient for statistical consistency*. This contrasts with D&C methods which impose a condition on the local samples size (let alone they are readily implementable over mesh networks). On the computational side, such statistically optimal estimates can be achieved by the distributed proximal-gradient algorithm applied to the penalized problem, which converges at linear rate–such a rate however scales as $\mathcal{O}(1/d)$, no matter how "good" the network connectivity is, resulting in a total communication cost of $\mathcal{O}(d^2)$.

We claim that this unfavorable communication cost is unavoidable for such penalty-based methods, because they lack of any mechanism mixing directly local gradients (they only average iterates). This raises the question whether communication costs of $\mathcal{O}(d)$ are achievable in high-dimension over mesh networks by other distributed, iterative algorithms, yet with no conditions on the local sample size. A first study towards this direction is the companion work (Sun et al., 2022), where the projected gradient algorithm (Sun et al., 2019) based on gradient tracking is studied in the high-dimensional setting. The analysis of other distributed methods employing other forms of gradient correction, such as primal-dual method as in (Jakovetić et al., 2011, Shi et al., 2014, Jakovetić, 2019, Jakovetić et al., 2013, Shi et al., 2015a,b) remains an interesting topic for future investigation.

7. Acknowledge

The work of Ji, Scutari, and Sun has been partially supported by the Office of Naval Research, under the Grant # N00014-21-1-2673.

Appendix

In this appendix we present the proofs of the results in the paper. We will use the same notation as in the paper along with the following additional definitions.

Recall the statistical error $\hat{\boldsymbol{\nu}} \triangleq \hat{\boldsymbol{\theta}} - 1_m \otimes \theta^*$. For any $\boldsymbol{\theta} \in \mathbb{R}^{md}$ partitioned as $\boldsymbol{\theta} = [\theta_1, \dots, \theta_m]$, with each $\theta_i \in \mathbb{R}^d$, we define

$$\boldsymbol{\nu} \triangleq \boldsymbol{\theta} - 1_m \otimes \theta^*. \tag{58}$$

When needed, we decompose $\boldsymbol{\theta}$, and accordingly $\boldsymbol{\nu}$, in its average and orthogonal component

$$\boldsymbol{\nu} = 1_m \otimes \nu_{\text{av}} + \boldsymbol{\nu}_\perp, \quad \text{with} \quad \nu_{\text{av}} = \frac{1}{m}\sum_{i=1}^m \nu_i. \tag{59}$$

In particular, when $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, we will write for the augmented LASSO error

$$\hat{\boldsymbol{\nu}} \triangleq \hat{\boldsymbol{\theta}} - 1_m \otimes \theta^* \quad \text{and} \quad \hat{\boldsymbol{\nu}} = 1_m \otimes \hat{\nu}_{\text{av}} + \hat{\boldsymbol{\nu}}_\perp, \tag{60}$$

whereas when $\boldsymbol{\theta} = \boldsymbol{\theta}^t$, with $\boldsymbol{\theta}^t$ being the iterates generated by Algorithm (24) , we will write

$$\boldsymbol{\nu}^t \triangleq \boldsymbol{\theta}^t - 1_m \otimes \theta^* \quad \text{and} \quad \boldsymbol{\nu}^t = 1_m \otimes \nu_{\text{av}}^t + \boldsymbol{\nu}_\perp^t. \tag{61}$$

Finally, the optimization error (along with its decomposition in average and orthogonal component) is denoted by

$$\boldsymbol{\Delta}^t \triangleq \boldsymbol{\theta}^t - \hat{\boldsymbol{\theta}} \quad \text{and} \quad \boldsymbol{\Delta}^t = 1_m \otimes \Delta_{\text{av}}^t + \boldsymbol{\Delta}_\perp^t. \tag{62}$$

Table 4 below summarizes all the universal constants used in the paper along with their range of values and associated constraints.

universal constant	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7
value	>0	>0	>0	>0	$\max\{2, 4c_2^2, 4c_2^2 c_3^{-1}\}$	>32	$c_5/32-1$	free

universal constant	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}
value	$\geq \sqrt{6}$	$\max\{128c_1, c_5\}$	1824	$3648c_1$	$\max\{3648c_1, c_5\}$	$2731c_1^2/t_0$	1152	$57\sqrt{6}c_8$

universal constant	c_{16}	c_{17}	c_{18}	c_{19}	c_{20}	c_{21}	c_{22}	c_{23}
value	$\sqrt{6}c_8/64c_1$	9	$72c_1c_{17}$	$c_1c_8^2c_{17}/988$	$8c_1c_{17}/c_8^2$	$c_8^2/1976$	$11339c_1c_8^2$	$21130c_4$

Table 3: Universal constants used in the paper.

universal constant	c_{24}	c_{25}	c_{26}
value	>0	$c_{25} = \min\{1, c_3, c_6, (t_0-2)/2\}$	$c_8^2 c_{24}$

Table 4: Universal constants used in the paper.

Appendix A. Proof of Proposition 3

For the sake of convenience, let us rewrite the objective function in (4) as

$$G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \frac{\lambda}{m}\|\boldsymbol{\theta}\|_1, \quad \text{with} \quad L(\boldsymbol{\theta}) = \frac{1}{2N}\sum_{i=1}^{m}\|y_i - X_i\theta_i\|^2 + \frac{1}{2m\gamma}\|\boldsymbol{\theta}\|_V^2. \tag{63}$$

By the optimality of $\hat{\boldsymbol{\theta}}$, it follows

$$\begin{aligned}
&G(\hat{\boldsymbol{\theta}}) \leq G(1_m \otimes \theta^*) \\
\Leftrightarrow &\frac{1}{2N}\sum_{i=1}^{m}\|y_i - X_i\hat{\theta}_i\|^2 + \frac{1}{2m\gamma}\|\hat{\boldsymbol{\theta}}\|_V^2 + \frac{\lambda}{m}\|\hat{\boldsymbol{\theta}}\|_1 \\
&\leq \frac{1}{2N}\sum_{i=1}^{m}\|y_i - X_i\theta^*\|^2 + \underbrace{\frac{1}{2m\gamma}\|1_m \otimes \theta^*\|_V^2}_{=0 \ \text{(Assumption 2)}} + \frac{\lambda}{m}\|1_m \otimes \theta^*\|_1.
\end{aligned}$$

Using $y_i = X_i\theta^* + w_i$ and the fact that θ^* is $\mathcal{S}$-sparse, we can write

$$\frac{1}{N}\sum_{i=1}^{m}\|X_i\hat{\theta}_i - X_i\theta^*\|^2$$

$$
\begin{aligned}
&\leq \frac{2}{N}\sum_{i=1}^{m} w_i^\top X_i(\hat{\theta}_i-\theta^*)+\frac{2\lambda}{m}\sum_{i=1}^{m}(\|\theta^*\|_1-\|\hat{\theta}_i\|_1)-\frac{1}{m\gamma}\|\hat{\boldsymbol{\theta}}\|_V^2\\
&= \frac{2}{N}\sum_{i=1}^{m} w_i^\top X_i(\hat{\theta}_i-\theta^*)+\frac{2\lambda}{m}\sum_{i=1}^{m}(\|\theta^*_{\mathcal{S}}\|_1-\|(\hat{\theta}_i)_{\mathcal{S}}\|_1-\|(\hat{\theta}_i)_{\mathcal{S}^c}\|_1)-\frac{1}{m\gamma}\|\hat{\boldsymbol{\theta}}\|_V^2\\
&\leq \frac{2}{N}\sum_{i=1}^{m} w_i^\top X_i\hat{\nu}_i+\frac{2\lambda}{m}\sum_{i=1}^{m}(\|(\hat{\nu}_i)_{\mathcal{S}}\|_1-\|(\hat{\nu}_i)_{\mathcal{S}^c}\|_1)-\frac{1}{m\gamma}\|\hat{\boldsymbol{\theta}}\|_V^2.
\end{aligned}
$$

Using the factorization $\hat{\boldsymbol{\nu}}=1_m\otimes\hat{\nu}_{\mathrm{av}}+\hat{\boldsymbol{\nu}}_\perp$, the above bounds reads

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{m}\|X_i\hat{\theta}_i-X_i\theta^*\|^2\\
&\leq \frac{2}{N}\sum_{i=1}^{m} w_i^\top X_i(\hat{\nu}_{\mathrm{av}}+\hat{\nu}_{\perp i})+\frac{2\lambda}{m}\sum_{i=1}^{m}(\|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}}+(\hat{\nu}_{\perp i})_{\mathcal{S}}\|_1-\|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}^c}+(\hat{\nu}_{\perp i})_{\mathcal{S}^c}\|_1)-\frac{1}{m\gamma}\|\hat{\boldsymbol{\theta}}\|_V^2\\
&\overset{(a)}{\leq} \frac{2}{N}w^\top\mathbf{X}\hat{\nu}_{\mathrm{av}}+\frac{2}{N}\sum_{i=1}^{m} w_i^\top X_i\hat{\nu}_{\perp i}+\frac{2\lambda}{m}\sum_{i=1}^{m}(\|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}}\|_1-\|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}^c}\|_1)+\frac{2\lambda}{m}\sum_{i=1}^{m}\|\hat{\nu}_{\perp i}\|_1-\frac{1}{m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|_V^2,
\end{aligned}
$$

where in (a) we used $\sum_{i=1}^{m} w_i^\top X_i\hat{\nu}_{\mathrm{av}}=\mathbf{w}^\top\mathbf{X}\,\hat{\nu}_{\mathrm{av}}$ and $\|\hat{\boldsymbol{\theta}}\|_V^2=\|\hat{\boldsymbol{\theta}}-1_m\otimes\theta^*\|_V^2=\|\hat{\boldsymbol{\nu}}_\perp\|_V^2$.

We bound now the two terms $\mathbf{w}^\top\mathbf{X}\hat{\nu}_{\mathrm{av}}$ and $\sum_{i=1}^{m} w_i^\top X_i\hat{\nu}_{\perp i}$. We have

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{m}\|X_i\hat{\theta}_i-X_i\theta^*\|^2\\
&\overset{\text{Hölder's}}{\leq} \frac{2}{N}\|\mathbf{w}^\top\mathbf{X}\|_\infty\|\hat{\nu}_{\mathrm{av}}\|_1+\max_{i\in[m]}\|w_i^\top X_i\|_\infty\frac{2}{N}\sum_{i=1}^{m}\|\hat{\nu}_{\perp i}\|_1\\
&\qquad+\frac{2\lambda}{m}\sum_{i=1}^{m}(\|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}}\|_1-\|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}^c}\|_1)+\frac{2\lambda}{m}\sum_{i=1}^{m}\|\hat{\nu}_{\perp i}\|_1-\frac{1}{m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|_V^2\\
&\overset{(10),(13)}{\leq} 3\lambda\|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}}\|_1-\lambda\|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}^c}\|_1\underbrace{-\frac{1-\rho}{m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|^2+\left(\max_{i\in[m]}\|w_i^\top X_i\|_\infty\frac{2}{N}+\frac{2\lambda}{m}\right)\sum_{i=1}^{m}\|\hat{\nu}_{\perp i}\|_1}_{\text{Term II}}.
\end{aligned}
\tag{64}
$$

Finally, we bound Term II. Since we have no sparsity information on $\hat{\boldsymbol{\nu}}_\perp$, we can only assert that

$$
\|\hat{\nu}_{\perp i}\|_1\leq\sqrt{d}\|\hat{\nu}_{\perp i}\|,\quad\forall i\in[m].
$$

Hence,

$$
\text{Term II}\leq-\frac{1-\rho}{m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|^2+\left(\max_{i\in[m]}\|w_i^\top X_i\|_\infty\frac{2}{N}+\frac{2\lambda}{m}\right)\sum_{i=1}^{m}\sqrt{d}\|\hat{\nu}_{\perp i}\|
$$

$$= -\frac{1-\rho}{m\gamma}\|\hat{\boldsymbol{\nu}}_{\perp}\|^2 + \lambda\left(\max_{i\in[m]}\|w_i^\top X_i\|_\infty \frac{2}{\lambda N} + \frac{2}{m}\right)\sum_{i=1}^{m}\sqrt{d}\|\hat{\nu}_{\perp i}\|$$

$$\overset{(12)}{=} \lambda\, h(\gamma, \|\hat{\boldsymbol{\nu}}_{\perp}\|). \tag{65}$$

Using (65) in (64), we finally obtain

$$\frac{1}{N\lambda}\sum_{i=1}^{m}\|X_i\hat{\theta}_i - X_i\theta^*\|^2 \le 3\|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}}\|_1 - \|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}^c}\|_1 + h(\gamma, \|\hat{\boldsymbol{\nu}}_{\perp}\|),$$

implying $3\|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}}\|_1 - \|(\hat{\nu}_{\mathrm{av}})_{\mathcal{S}^c}\|_1 + h(\gamma, \|\hat{\boldsymbol{\nu}}_{\perp}\|) \ge 0$, which concludes the proof. □

Appendix B. Proof of Lemma 4 and Lemma 5

B.1 Proof of Lemma 4

The RSC condition (8) at $\Delta_{\mathrm{av}} \in \mathbb{R}^d$,

$$\frac{1}{N}\|\mathbf{X}\Delta_{\mathrm{av}}\|^2 \ge \frac{\mu}{2}\|\Delta_{\mathrm{av}}\|^2 - \frac{\tau}{2}\|\Delta_{\mathrm{av}}\|_1^2.$$

Let $\boldsymbol{\Delta} \in \mathbb{C}_\gamma(\mathcal{S})$, so that

$$\|(\Delta_{\mathrm{av}})_{\mathcal{S}^c}\|_1 \le 3\|(\Delta_{\mathrm{av}})_{\mathcal{S}}\|_1 + h(\gamma, \|\boldsymbol{\Delta}_{\perp}\|),$$

where $h(\gamma, \|\boldsymbol{\Delta}_{\perp}\|)$ is defined in (12). Substituting this inequality into the RSC above, it yields

$$\frac{1}{N}\|\mathbf{X}\Delta_{\mathrm{av}}\|^2 \ge \left(\frac{\mu}{2} - 16s\tau\right)\|\Delta_{\mathrm{av}}\|^2 - \tau h^2(\gamma, \|\boldsymbol{\Delta}_{\perp}\|). \tag{66}$$

Therefore, (17) holds with $\delta = \mu/2 - 16s\tau$ and $\xi = \tau$. □

B.2 Proof of Lemma 5

Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ be a design matrix satisfying Assumption 1. RSC condition (Raskutti et al., 2010, Theorem 1) implies that there exist $c_0, c_1 > 0$, such that for all $\Delta_{\mathrm{av}} \in \mathbb{R}^d$,

$$\frac{1}{N}\|\mathbf{X}\Delta_{\mathrm{av}}\|^2 \ge \frac{1}{2}\|\Sigma^{\frac{1}{2}}\Delta_{\mathrm{av}}\|^2 - \frac{c_1\zeta_\Sigma \log d}{N}\|\Delta_{\mathrm{av}}\|_1^2 \tag{67}$$

holds with probability at least $1 - \exp(-c_0 N)$. Furthermore, by condition (c) of $\mathbf{X}$, we have

$$\|\Sigma^{\frac{1}{2}}\Delta_{\mathrm{av}}\|^2 \ge \lambda_{\min}(\Sigma)\|\Delta_{\mathrm{av}}\|^2. \tag{68}$$

Let $\boldsymbol{\Delta} \in \mathbb{C}_\gamma(\mathcal{S})$, that is,

$$\|(\Delta_{\mathrm{av}})_{\mathcal{S}^c}\|_1 \le 3\|(\Delta_{\mathrm{av}})_{\mathcal{S}}\|_1 + h(\gamma, \|\Delta_{\perp}\|). \tag{69}$$

Substituting (68) and (69) into (67), yields

$$\frac{\|\mathbf{X}\Delta_{\mathrm{av}}\|^2}{N} \ge \frac{\lambda_{\min}(\Sigma)}{2}\|\Delta_{\mathrm{av}}\|^2 - \frac{c_1\zeta_\Sigma \log d}{N}(32s\|\Delta_{\mathrm{av}}\|^2 + 2h^2(\gamma, \|\boldsymbol{\Delta}_{\perp}\|))$$

$$
\begin{aligned}
&= \left(\frac{\lambda_{\min}(\Sigma)}{2} - \frac{32 s c_1 \zeta_\Sigma \log d}{N}\right)\|\Delta_{\text{av}}\|^2 - \frac{2c_1\zeta_\Sigma \log d}{N} h^2(\gamma, \|\boldsymbol{\Delta}_\perp\|) \\
&\geq \frac{\lambda_{\min}(\Sigma)}{4}\|\Delta_{\text{av}}\|^2 - \frac{\lambda_{\min}(\Sigma)}{64s} h^2(\gamma, \|\boldsymbol{\Delta}_\perp\|), \quad \text{for} \quad N \geq \frac{128 s c_1 \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)}.
\end{aligned}
$$

This completes the proof. □

Appendix C. Proof of Theorem 6

Our starting point toward the upper bound on the average LASSO error $(1/m)\sum_{i=1}^m \|\hat{\nu}_i\|^2$ is lower- and upper-bounding the average of local errors $(1/N)\sum_{i=1}^m \|X_i\hat{\nu}_i\|^2$ while decomposing $\hat{\boldsymbol{\theta}}$ in its average component and orthogonal one. This decomposition is instrumental to separate in the desired final bound a term of the same order of the centralized LASSO error from the (additive) perturbation due to the lack of exact consensus.

- **Step 1: Establishing the upper bound of** $(1/N)\sum_{i=1}^m \|X_i\hat{\nu}_i\|^2$.

We start with the optimality condition of Problem (4). By optimality of $\hat{\boldsymbol{\theta}}$, it follows that

$$
\frac{1}{N}\sum_{i=1}^m \left(X_i\hat{\theta}_i - y_i\right)^\top X_i\hat{\nu}_i \leq \frac{\lambda}{m}(\|1_m \otimes \theta^*\|_1 - \|\hat{\boldsymbol{\theta}}\|_1) + \frac{1}{2m\gamma}(\underbrace{\|1_m \otimes \theta^*\|_V^2}_{=0 \ \text{(Assumption 2)}} - \|\hat{\boldsymbol{\theta}}\|_V^2). \quad (70)
$$

We can then write

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^m \|X_i\hat{\nu}_i\|^2 \\
&= \frac{2}{N}\sum_{i=1}^m \Big((X_i\hat{\theta}_i - y_i) + (y_i - X_i\theta^*)\Big)^\top X_i\hat{\nu}_i - \frac{1}{N}\sum_{i=1}^m \|X_i\hat{\nu}_i\|^2 \\
&\overset{(70)}{\leq} \frac{2}{N}\sum_{i=1}^m (y_i - X_i\theta^*)^\top X_i\hat{\nu}_i + \frac{2\lambda}{m}(\|1_m \otimes \theta^*\|_1 - \|\hat{\boldsymbol{\theta}}\|_1) - \frac{1}{m\gamma}\|\hat{\boldsymbol{\theta}}\|_V^2 - \frac{1}{N}\sum_{i=1}^m \|X_i\hat{\nu}_i\|^2.
\end{aligned}
$$

Using $y_i = X_i\theta^* + w_i$ and the fact that θ^* is $\mathcal{S}$-sparse, we can write

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^m \|X_i\hat{\nu}_i\|^2 \\
\leq &\frac{2}{N}\sum_{i=1}^m w_i^\top X_i\hat{\nu}_i + \frac{2\lambda}{m}\sum_{i=1}^m (\|\theta^*_{\mathcal{S}}\|_1 - \|(\hat{\theta}_i)_{\mathcal{S}}\|_1 - \|(\hat{\theta}_i)_{\mathcal{S}^c}\|_1) - \frac{1}{m\gamma}\|\hat{\boldsymbol{\theta}}\|_V^2 - \frac{1}{N}\sum_{i=1}^m \|X_i\hat{\nu}_i\|^2 \\
\leq &\frac{2}{N}\sum_{i=1}^m w_i^\top X_i\hat{\nu}_i + \frac{2\lambda}{m}\sum_{i=1}^m (\|(\hat{\nu}_i)_{\mathcal{S}}\|_1 - \|(\hat{\nu}_i)_{\mathcal{S}^c}\|_1) - \frac{1}{m\gamma}\|\hat{\boldsymbol{\theta}}\|_V^2 - \frac{1}{N}\sum_{i=1}^m \|X_i\hat{\nu}_i\|^2.
\end{aligned}
$$

Introducing the decomposition $\hat{\boldsymbol{\nu}} = 1_m \otimes \hat{\nu}_{\text{av}} + \hat{\boldsymbol{\nu}}_\perp$, the above bound reads

$$
\frac{2}{N}\sum_{i=1}^m \|X_i\hat{\nu}_i\|^2
$$

$$\leq \frac{2}{N}\sum_{i=1}^{m} w_i^\top X_i(\hat{\nu}_{\text{av}} + \hat{\nu}_{\perp i}) + \frac{2\lambda}{m}\sum_{i=1}^{m}(\|(\hat{\nu}_{\text{av}})_{\mathcal{S}} + (\hat{\nu}_{\perp i})_{\mathcal{S}}\|_1 - \|(\hat{\nu}_{\text{av}})_{\mathcal{S}^c} + (\hat{\nu}_{\perp i})_{\mathcal{S}^c}\|_1) - \frac{1}{m\gamma}\|\hat{\boldsymbol{\theta}}\|_V^2$$

$$\overset{(a)}{\leq} \frac{2}{N}\mathbf{w}^\top \mathbf{X}\hat{\nu}_{\text{av}} + \frac{2}{N}\sum_{i=1}^{m} w_i^\top X_i \hat{\nu}_{\perp i} + \frac{2\lambda}{m}\sum_{i=1}^{m}(\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1 - \|(\hat{\nu}_{\text{av}})_{\mathcal{S}^c}\|_1) + \frac{2\lambda}{m}\sum_{i=1}^{m}\|\hat{\nu}_{\perp i}\|_1 - \frac{1}{m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|_V^2,$$

where in (a) we used $\sum_{i=1}^{m} w_i^\top X_i \hat{\nu}_{\text{av}} = \mathbf{w}^\top \mathbf{X}\,\hat{\nu}_{\text{av}}$ and $\|\hat{\boldsymbol{\theta}}\|_V^2 = \|\hat{\boldsymbol{\theta}} - 1_m \otimes \theta^*\|_V^2 = \|\hat{\boldsymbol{\nu}}_\perp\|_V^2$.

We bound now the two terms $\mathbf{w}^\top \mathbf{X}\hat{\nu}_{\text{av}}$ and $\sum_{i=1}^{m} w_i^\top X_i \hat{\nu}_{\perp i}$. We have

$$\begin{aligned}
&\frac{2}{N}\sum_{i=1}^{m}\|X_i\hat{\theta}_i - X_i\theta^*\|^2 \\
&\overset{\text{Hölder's}}{\leq} \frac{2}{N}\|\mathbf{w}^\top \mathbf{X}\|_\infty \|\hat{\nu}_{\text{av}}\|_1 + \max_{i\in[m]}\|w_i^\top X_i\|_\infty \frac{2}{N}\sum_{i=1}^{m}\|\hat{\nu}_{\perp i}\|_1 \\
&\quad + \frac{2\lambda}{m}\sum_{i=1}^{m}(\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1 - \|(\hat{\nu}_{\text{av}})_{\mathcal{S}^c}\|_1) + \frac{2\lambda}{m}\sum_{i=1}^{m}\|\hat{\nu}_{\perp i}\|_1 - \frac{1}{m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|_V^2 \\
&\overset{(10),(13)}{\leq} 3\lambda\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1 - \lambda\|(\hat{\nu}_{\text{av}})_{\mathcal{S}^c}\|_1 - \frac{1-\rho}{m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|^2 + \left(\max_{i\in[m]}\|w_i^\top X_i\|_\infty \frac{2}{N} + \frac{2\lambda}{m}\right)\sum_{i=1}^{m}\|\hat{\nu}_{\perp i}\|_1
\end{aligned}$$

We further relax the bound by dropping $-\lambda\|(\hat{\nu}_{\text{av}})_{\mathcal{S}^c}\|_1$ and enlarging $\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1 \leq \|\hat{\nu}_{\text{av}}\|_1$ while revealing the term $\frac{9\lambda^2 s}{2\delta}$ which is of the order of the centralized LASSO error:

$$\begin{aligned}
&\frac{2}{N}\sum_{i=1}^{m}\|X_i\hat{\nu}_i\|^2 \\
&\leq 3\lambda\sqrt{s}\|\hat{\nu}_{\text{av}}\| - \frac{1-\rho}{m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|^2 + \left(\max_{i\in[m]}\|w_i^\top X_i\|_\infty \frac{2}{N} + \frac{2\lambda}{m}\right)\|\hat{\boldsymbol{\nu}}_\perp\|_1 \\
&= 2\cdot\frac{3\lambda\sqrt{s}}{\sqrt{2\delta}}\cdot\sqrt{\frac{\delta}{2}}\|\hat{\nu}_{\text{av}}\| - \frac{1-\rho}{m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|^2 + \left(\max_{i\in[m]}\|w_i^\top X_i\|_\infty \frac{2}{N} + \frac{2\lambda}{m}\right)\|\hat{\boldsymbol{\nu}}_\perp\|_1 \\
&\leq \frac{9\lambda^2 s}{2\delta} + \frac{\delta}{2}\|\hat{\nu}_{\text{av}}\|^2 - \frac{1-\rho}{m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|^2 + \left(\max_{i\in[m]}\|w_i^\top X_i\|_\infty \frac{2}{N} + \frac{2\lambda}{m}\right)\|\hat{\boldsymbol{\nu}}_\perp\|_1 \\
&\overset{(15)}{\leq} \frac{9\lambda^2 s}{2\delta} + \frac{1}{2}\left(\frac{\|\mathbf{X}\hat{\nu}_{\text{av}}\|^2}{N} + \xi h^2(\gamma, \|\hat{\boldsymbol{\nu}}_\perp\|)\right) - \frac{1-\rho}{m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|^2 + \left(\max_{i\in[m]}\|w_i^\top X_i\|_\infty \frac{2}{N} + \frac{2\lambda}{m}\right)\|\hat{\boldsymbol{\nu}}_\perp\|_1.
\end{aligned} \tag{71}$$

• **Step 2: Establishing the lower bound of** $(1/N)\sum_{i=1}^{m}\|X_i\hat{\nu}_i\|^2$.

Invoking the decomposition $\hat{\nu}_i = \hat{\nu}_{\text{av}} + \hat{\nu}_{\perp\, i}$, $i \in [m]$, along with the Young's inequality, we can write

$$\frac{2}{N}\sum_{i=1}^{m}\|X_i(\hat{\nu}_{\text{av}} + \hat{\nu}_{\perp\, i})\|^2 \;\geq \frac{1}{N}\sum_{i=1}^{m}\|X_i\hat{\nu}_{\text{av}}\|^2 - \frac{2}{N}\sum_{i=1}^{m}\|X_i\hat{\nu}_{\perp i}\|^2$$

$$
\begin{aligned}
&= \frac{1}{N}\|\mathbf{X}\hat{\nu}_{\text{av}}\|^2 - \frac{2}{N}\sum_{i=1}^{m}\|X_i\hat{\nu}_{\perp i}\|^2 \\
&\overset{(15)}{\geq} \frac{1}{2}[\delta\|\hat{\nu}_{\text{av}}\|^2 - \xi h^2(\gamma, \|\hat{\boldsymbol{\nu}}_{\perp}\|)] + \frac{1}{2N}\|\mathbf{X}\hat{\nu}_{\text{av}}\|^2 - \frac{2}{N}\sum_{i=1}^{m}\|X_i\hat{\nu}_{\perp i}\|^2.
\end{aligned}
\tag{72}
$$

- **Step 3: Lower bound $\leq$ Upper bound.**

  Chaining (71) and (72) while adding $\frac{\delta}{2m}\|\hat{\boldsymbol{\nu}}_{\perp}\|^2$ on both sides, yield

$$
\begin{aligned}
&\frac{1}{2}\delta\|\hat{\nu}_{\text{av}}\|^2 + \frac{\delta}{2m}\|\hat{\boldsymbol{\nu}}_{\perp}\|^2 \\
\leq& \frac{9\lambda^2 s}{2\delta} + \xi h^2(\gamma, \|\hat{\boldsymbol{\nu}}_{\perp}\|) + \left(\frac{2L_{\max}}{m} + \frac{\delta}{2m} - \frac{1-\rho}{m\gamma}\right)\|\hat{\boldsymbol{\nu}}_{\perp}\|^2 + \left(\max_{i\in[m]}\|w_i^{\top}X_i\|_{\infty}\frac{2}{N} + \frac{2\lambda}{m}\right)\|\hat{\boldsymbol{\nu}}_{\perp}\|_1 \\
\leq& \frac{9\lambda^2 s}{2\delta} + \xi h_{\max}^2 + \underbrace{\left(\frac{2L_{\max}}{m} + \frac{\delta}{2m} - \frac{1-\rho}{m\gamma}\right)\|\hat{\boldsymbol{\nu}}_{\perp}\|^2 + \left(\max_{i\in[m]}\|w_i^{\top}X_i\|_{\infty}\frac{2}{N} + \frac{2\lambda}{m}\right)\sqrt{md}\|\hat{\boldsymbol{\nu}}_{\perp}\|}_{\triangleq h_1(\gamma, \|\hat{\boldsymbol{\nu}}_{\perp}\|)},
\end{aligned}
\tag{73}
$$

where in the last inequality we used $L_{\max} = \max_{i\in[m]}\lambda_{\max}(X_i^{\top}X_i/n)$ [cf. (5)], and the following upper bound for $h(\gamma, \|\hat{\boldsymbol{\nu}}_{\perp}\|)$

$$
h(\gamma, \|\hat{\boldsymbol{\nu}}_{\perp}\|) \leq h_{\max} \triangleq \frac{d\gamma}{\lambda(1-\rho)}\left(\frac{\max_{i\in[m]}\|w_i^{\top}X_i\|_{\infty}}{n} + \lambda\right)^2. \tag{74}
$$

Under

$$
\gamma \leq \frac{2(1-\rho)}{4L_{\max} + \delta}, \tag{75}
$$

we have

$$
\frac{2L_{\max}}{m} + \frac{\delta}{2m} - \frac{1-\rho}{m\gamma} \leq 0.
$$

Hence, $h_1(\gamma, \|\hat{\boldsymbol{\nu}}_{\perp}\|)$ is a quadratic function of $\|\hat{\boldsymbol{\nu}}_{\perp}\|$ opening downward, and it can be upper bounded over $\mathbb{R}^+$ as

$$
h_{1\max} \triangleq \frac{2d\gamma(\max_{i\in[m]}\|w_i^{\top}X_i\|_{\infty}/n + \lambda)^2}{2(1-\rho) - 4L_{\max}\gamma - \delta\gamma}. \tag{76}
$$

Using (76) in (73), we finally obtain

$$
\begin{aligned}
&\frac{\|\hat{\boldsymbol{\nu}}\|^2}{m} \\
\leq& \frac{9\lambda^2 s}{\delta^2} + \frac{2\xi h_{\max}^2 + 2h_{1\max}}{\delta} \\
\leq& \frac{9\lambda^2 s}{\delta^2} + \frac{2\xi d^2\gamma^2}{\delta\lambda^2(1-\rho)^2}\left(\frac{\max_{i\in[m]}\|w_i^{\top}X_i\|_{\infty}}{n} + \lambda\right)^4 + \frac{4d\gamma(\max_{i\in[m]}\|w_i^{\top}X_i\|_{\infty}/n + \lambda)^2}{\delta[2(1-\rho) - 4L_{\max}\gamma - \delta\gamma]}.
\end{aligned}
$$

This concludes the proof. $\square$

Appendix D. Proof of Theorem 7

The proof builds on the following four steps: **1)** We first consider as source of randomness only the design matrix $\mathbf{X}$ (cf. Assumption 1) while keeping $\mathbf{w}$ fixed, deriving a high-probability bound for $L_{\max}$ in (5); **2)** We then fix $\mathbf{X}$ and consider the randomness coming from the noise $\mathbf{w}$, providing high-probability bounds for the noise-dependent terms $\|\mathbf{X}^\top \mathbf{w}\|_\infty/N$ and $\max_{1\le i\le m}\|X_i^\top w_i\|_\infty/n$; **3)** We then combine the previous two results via the union bound and establish a lower bound on λ for (13) to hold with high probability; **4)** Finally, we use the bound in **3)** to obtain the final error bound on the ℓ_2-LASSO error.

Let $\mathbb{P}$ be a probability measure on the product sample space $\mathbb{R}^{N\times d}\otimes\mathbb{R}^N$. For brevity, we use the same notation for the marginal distributions on $\mathbb{R}^{N\times d}$ and $\mathbb{R}^N$.

• **Step 1: Randomness from $\mathbf{X}$.**

We define three "good" events so that the largest eigenvalue of $(1/n)X_i^\top X_i$, smallest eigenvalue of $(1/N)\mathbf{X}^\top\mathbf{X}$ and the norm of the columns of $\mathbf{X}$ are well-controlled. We prove next that these events jointly occur with high probability. Specifically, let

$$A_1 \triangleq \left\{\mathbf{X}\in\mathbb{R}^{N\times d}\;\middle|\; L_{\max}\le c_4\lambda_{\max}(\Sigma)\left(1+\frac{d+\log m}{n}\right)\right\}, \tag{77}$$

$$A_2 \triangleq \left\{\mathbf{X}\in\mathbb{R}^{N\times d}\;\middle|\;\mathbf{X}\text{ satisfies (17)}\right\},\ \text{ and }\ A_3 \triangleq \left\{\mathbf{X}\in\mathbb{R}^{N\times d}\;\middle|\;\max_{j=1,\ldots,d}\frac{1}{\sqrt{N}}\|\mathbf{X}e_j\|\le\sqrt{\frac{3\zeta_\Sigma}{2}}\right\}, \tag{78}$$

where $c_4>0$ is a universal constant (see (83)), and we recall from (5) and (9) that $L_{\max}\triangleq\max_{i\in[m]}\lambda_{\max}(X_i^\top X_i/n)$, and $\zeta_\Sigma\triangleq\max_{i\in[d]}\Sigma_{ii}$, respectively. We proceed to bounding $\mathbb{P}(A_1)$, $\mathbb{P}(A_2)$, and $\mathbb{P}(A_3)$.

(i) Bounding $\mathbb{P}(A_1)$**:** Recall that $\mathbf{X}=[X_1^\top,\ldots,X_m^\top]^\top$, and $\mathbf{X}$ satisfies Assumption 1. Thus, $\{X_i\}_{i\in[m]}$ are i.i.d random matrices, with i.i.d. rows drawn from $\mathcal{N}(0,\Sigma)$. By (Vershynin, 2012, Remark 5.40) it follows that the following holds with probability at least

$$1-2\exp\{-c_3t^2\}, \tag{79}$$

for all $t\ge 0$

$$\left\|\frac{1}{n}X_i^\top X_i-\Sigma\right\|\le\max\{a,a^2\}\|\Sigma\|,\ \text{ where } a\triangleq c_2\left(\sqrt{\frac{d}{n}}+\frac{t}{\sqrt{n}}\right), \tag{80}$$

with constants c_3 and $c_2>0$. Given (80) and using the triangle inequality, we have

$$\begin{aligned}\left\|\frac{1}{n}X_i^\top X_i\right\| &= \left\|\frac{1}{n}X_i^\top X_i-\Sigma+\Sigma\right\|\\ &\le\left\|\frac{1}{n}X_i^\top X_i-\Sigma\right\|+\left\|\Sigma\right\|\\ &\le\lambda_{\max}(\Sigma)\max\{a,a^2\}+\|\Sigma\|\\ &=\lambda_{\max}(\Sigma)\max\{a,a^2\}+\lambda_{\max}(\Sigma).\end{aligned} \tag{81}$$

Applying the union bound we obtain the following bound for $L_{\max}$

$$\mathbb{P}\left(L_{\max}\le\lambda_{\max}(\Sigma)(1+\max\{a,a^2\})\right)\ge 1-m\cdot 2\exp\{-c_3t^2\}. \tag{82}$$

Setting $t = \sqrt{d + c_3^{-1}\log m}$, yields

$$a = c_2\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{d + c_3^{-1}\log m}{n}}\right).$$

Therefore, we conclude

$$\begin{aligned} L_{\max} &\leq \lambda_{\max}(\Sigma)\left(1 + a + a^2\right) \leq \lambda_{\max}(\Sigma)\left(1 + a\right)^2 \leq 2\lambda_{\max}(\Sigma)\left(1 + a^2\right) \\ &\leq 2\lambda_{\max}(\Sigma)\left(1 + 2c_2^2\left(\frac{d}{n} + c_3^{-1}\frac{\log m}{n}\right)\right) \\ &\leq c_4\lambda_{\max}(\Sigma)\left(1 + \frac{d + \log m}{n}\right), \end{aligned} \tag{83}$$

with probability at least $1 - 2\exp(-c_3 d)$ and

$$c_4 = \max\{2, 4c_2^2, 4c_2^2 c_3^{-1}\} \geq 2. \tag{84}$$

(ii) Bounding $\mathbb{P}(A_2)$: This follows immediately from Lemma 5: if $N \geq \frac{128 s c_1 \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)}$ and $\gamma > 0$, then

$$\mathbb{P}(A_2^c) \leq \exp(-c_0 N). \tag{85}$$

(iii) Bounding $\mathbb{P}(A_3)$: Recall Assumption 1. It follows that $\mathbf{X}e_j$ is an isotropic Gaussian random vector in $\mathbb{R}^N$ with $\mathcal{N}(0, \Sigma_{jj})$ entries. Hence, $\|\mathbf{X}e_j\|^2/\Sigma_{jj}$ is a chi-squared random variable with degree N. Then, applying the standard bound for chi-squared random variables (Wainwright, 2019, Example 2.11) we have

$$\mathbb{P}\left(\left|\frac{1}{N}\left\|\frac{\mathbf{X}e_j}{\sqrt{\Sigma_{jj}}}\right\|^2 - 1\right| \geq t\right) \leq 2\exp(-Nt^2/8), \qquad \text{for all } t \in (0,1). \tag{86}$$

Taking $t = \frac{1}{2}$ in (86), we have

$$\mathbb{P}\left(\frac{1}{N}\left\|\frac{\mathbf{X}e_j}{\sqrt{\Sigma_{jj}}}\right\|^2 \geq \frac{3}{2}\right) \leq \mathbb{P}\left(\left|\frac{1}{N}\left\|\frac{\mathbf{X}e_j}{\sqrt{\Sigma_{jj}}}\right\|^2 - 1\right| \geq \frac{1}{2}\right) \leq 2\exp(-N/32). \tag{87}$$

Applying the union bound, we obtain

$$\begin{aligned} \mathbb{P}\left(\max_{j\in[d]}\frac{1}{N}\left\|\frac{\mathbf{X}e_j}{\sqrt{\Sigma_{jj}}}\right\|^2 \geq \frac{3}{2}\right) &\leq d\,\mathbb{P}\left(\left|\frac{1}{N}\left\|\frac{\mathbf{X}e_j}{\sqrt{\Sigma_{jj}}}\right\|^2 - 1\right| \geq \frac{1}{2}\right) \\ &\leq 2d\exp(-N/32) \\ &= 2\exp(-N/32 + \log d). \end{aligned} \tag{88}$$

Therefore, for all $N \geq c_5 \log d$, with $c_5 > 32$, we have

$$\begin{aligned} \mathbb{P}\left(\max_{j\in[d]}\frac{\|\mathbf{X}e_j\|^2}{N} \leq \frac{3}{2}\zeta_\Sigma\right) &\geq 1 - 2\exp[-(c_5/32)\log d + \log d] \\ &= 1 - 2\exp(-c_6\log d), \text{ where } c_6 = c_5/32 - 1 > 0. \end{aligned} \tag{89}$$

Combining the conditions on N, we have

$$N \geq \frac{c_9 s \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)} \overset{(a)}{\geq} \max\left\{\frac{128 s c_1 \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)}, c_5 \log d\right\}, \tag{90}$$

where $c_9 = \max\{128c_1, c_5\}$, and in (a) we used $s \geq 1, \zeta_\Sigma \geq \lambda_{\min}(\Sigma)$.

Finally, we combine (83), (85), and (89); using the union bound again we obtain

$$\mathbb{P}(A_1^c \cup A_2^c \cup A_3^c) \leq \mathbb{P}(A_1^c) + \mathbb{P}(A_2^c) + \mathbb{P}(A_3^c) \leq 2\exp(-c_3 d) + \exp(-c_0 N) + 2\exp(-c_6 \log d).$$

Define $A \triangleq A_1 \cap A_2 \cap A_3$, we have

$$\mathbb{P}(A) \geq 1 - 2\exp(-c_3 d) - \exp(-c_0 N) - 2\exp(-c_6 \log d). \tag{91}$$

• **Step 2: Randomness from $\mathbf{w}$.** We start with bounding $\|\mathbf{X}^\top \mathbf{w}\|_\infty$. For fixed $\mathbf{X} \in A$, and $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 I_N)$, recall $\mathbf{X} = [X_1^\top, \ldots, X_m^\top]^\top$, and $\mathbf{w} = [w_1^\top, \ldots, w_m^\top]^\top$, where for each agent $i \in [m]$, $X_i \in \mathbb{R}^{n \times d}$ is the design matrix, $w_i \in \mathbb{R}^n$ is observation noise. Then, for any $i \in [m]$, and $j \in [d]$, it follows that

$$\frac{\mathbf{w}^\top \mathbf{X} e_j}{N}\Bigg|_{\mathbf{X} \in A} \sim \mathcal{N}\left(0, \frac{\sigma^2}{N} \cdot \frac{\|\mathbf{X} e_j\|^2}{N}\right) \qquad \text{and} \qquad \frac{w_i^\top X_i e_j}{N}\Bigg|_{\mathbf{X} \in A} \sim \mathcal{N}\left(0, \frac{\sigma^2}{N} \cdot \frac{\|X_i e_j\|^2}{N}\right). \tag{92}$$

Note that

$$\max_{i \in [m]} \max_{j \in [d]} \frac{\|X_i e_j\|^2}{N} \leq \max_{j \in [d]} \frac{1}{m} \sum_{i=1}^m \frac{\|X_i e_j\|^2}{n} = \max_{j \in [d]} \frac{\|\mathbf{X} e_j\|^2}{N}, \tag{93}$$

due to $\frac{1}{m}\sum_{i=1}^m \frac{\|X_i e_j\|^2}{n} = \frac{\|\mathbf{X} e_j\|^2}{N}$.

By definition, for all $\mathbf{X} \in A \subseteq A_3$, $2\|\mathbf{X} e_j\|^2/(3\zeta_\Sigma N) \leq 1$ and, by (93), $2\|X_i e_j\|^2/(3\zeta_\Sigma N) \leq 1$. Therefore, combining it with (92), we obtain

$$\begin{aligned} &\sqrt{\frac{2}{3\zeta_\Sigma}}\frac{\mathbf{w}^\top \mathbf{X} e_j}{N}\Bigg|_{\mathbf{X} \in A} \sim \mathcal{N}\left(0, \frac{\sigma^2}{N} \cdot \frac{2\|\mathbf{X} e_j\|^2}{3\zeta_\Sigma N}\right), \qquad \text{where} \qquad \frac{2\|\mathbf{X} e_j\|^2}{3\zeta_\Sigma N} \leq 1; \quad \text{and} \\ &\sqrt{\frac{2}{3\zeta_\Sigma}}\frac{w_i^\top X_i e_j}{N}\Bigg|_{\mathbf{X} \in A} \sim \mathcal{N}\left(0, \frac{\sigma^2}{N} \cdot \frac{2\|X_i e_j\|^2}{3\zeta_\Sigma N}\right), \qquad \text{where} \qquad \frac{2\|X_i e_j\|^2}{3\zeta_\Sigma N} \leq 1. \end{aligned} \tag{94}$$

Denote $p_{\mathbf{X}}(x)$ and $p_{X_i}(x_i)$ as the density of $\sqrt{2/(3\zeta_\Sigma)}\mathbf{w}^\top \mathbf{X} e_j/N$ and $\sqrt{2/3\zeta_\Sigma} w_i^\top X_i e_j/N$, respectively. Let $Z \sim \mathcal{N}(0, \sigma^2/N)$, and $p_Z(z)$ denotes its density. Then, since $2\|\mathbf{X} e_j\|^2/(3\zeta_\Sigma N) \leq 1$, and $2\|X_i e_j\|^2/(3\zeta_\Sigma N) \leq 1$, then, we conclude $p_{\mathbf{X}}(0) \geq p_Z(0)$, as well as $p_{X_i}(0) \geq p_Z(0)$. (Horn, 1988, Theorem 1) implies

$$\left|\sqrt{\frac{2}{3\zeta_\Sigma}}\frac{\mathbf{w}^\top \mathbf{X} e_j}{N}\Bigg|_{\mathbf{X} \in A}\right| \preceq^{\text{st}} |Z|, \quad \text{as well as} \quad \left|\sqrt{\frac{2}{3\zeta_\Sigma}}\frac{w_i^\top X_i e_j}{N}\Bigg|_{\mathbf{X} \in A}\right| \preceq^{\text{st}} |Z|. \tag{95}$$

Therefore, (95) implies

$$\mathbb{P}\left(\frac{|\mathbf{w}^\top \mathbf{X} e_j|}{N} \geq x\sqrt{\frac{3\zeta_\Sigma}{2}} \;\Bigg|\; \mathbf{X} \in A\right)$$

$$
\begin{aligned}
=&\mathbb{P}\bigg(\sqrt{\frac{2}{3\zeta_\Sigma}}\frac{|\mathbf{w}^\top\mathbf{X}e_j|}{N}\geq x \;\bigg|\; \mathbf{X}\in A\bigg)\\
\leq&\mathbb{P}(|Z|\geq x)\\
\leq&2\exp\bigg(-\frac{Nx^2}{2\sigma^2}\bigg).
\end{aligned}
\tag{96}
$$

Notice that $\|\mathbf{X}^\top\mathbf{w}\|_\infty/N=\max_{j\in[d]}|\mathbf{w}^\top\mathbf{X}e_j|/N$. Hence, if we take $x=\sigma\sqrt{t_0\log d/N}$, with $t_0>2$, the union bound implies

$$
\begin{aligned}
&\mathbb{P}\bigg(\frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N}\geq\sigma\sqrt{\frac{t_0\log d}{N}}\sqrt{\frac{3\zeta_\Sigma}{2}} \;\bigg|\; \mathbf{X}\in A\bigg)\\
=&\mathbb{P}\bigg(\sqrt{\frac{2}{3\zeta_\Sigma}}\frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N}\geq\sigma\sqrt{\frac{t_0\log d}{N}} \;\bigg|\; \mathbf{X}\in A\bigg)\\
\leq&2\exp\bigg(-\frac{N}{2\sigma^2}\frac{\sigma^2t_0\log d}{N}+\log d\bigg)\\
=&2\exp\bigg(-\frac{1}{2}(t_0-2)\log d\bigg).
\end{aligned}
\tag{97}
$$

Define

$$
D_1\triangleq\bigg\{\mathbf{w}\in\mathbb{R}^N \;\bigg|\; \frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N}\leq\sigma\sqrt{\frac{t_0\log d}{N}}\sqrt{\frac{3\zeta_\Sigma}{2}}\bigg\}
\tag{98}
$$

Therefore, $\mathbb{P}(D_1\mid\mathbf{X}\in A)\geq 1-2\exp\left(-\frac{1}{2}(t_0-2)\log d\right)$. Combining it with (91), we have

$$
\begin{aligned}
\mathbb{P}(A\cap D_1)=&\mathbb{P}(D_1\mid A)\mathbb{P}(A)\\
\geq&[1-2\exp\{-[(t_0-2)\log d]/2\}][1-2\exp(-c_3d)-\exp(-c_0N)-2\exp(-c_6\log d)]\\
=&[1-2\exp(-c_3d)-\exp(-c_0N)-2\exp(-c_6\log d)]-2\exp\{-[(t_0-2)\log d]/2\\
&+2\exp\{-[(t_0-2)\log d]/2\}\left(2\exp(-c_3d)+\exp(-c_0N)+2\exp(-c_6\log d)]\right)\\
\geq&1-2\exp(-c_3d)-\exp(-c_0N)-2\exp(-c_6\log d)-2\exp\{-[(t_0-2)\log d]/2.
\end{aligned}
$$

It remains to bound $\max_{i\in[m]}\|X_i^\top w_i\|_\infty$. Since X_i, $i\in[m]$, are independent, the columns of X_i are n dimensional i.i.d Gaussian random vectors, each element has variance at most ζ_Σ, and the elements of $w_i\sim\mathcal{N}(0,\sigma^2I_n)$. Then each element of $X_i^\top w_i$ is the sum of n independent sub-exponential random variables with sub-exponential norm at most $\sigma\sqrt{\zeta_\Sigma}$. Applying (Vershynin, 2012, Proposition 5.16) and the union bound, we obtain

$$
\mathbb{P}\left(\frac{1}{n}\max_{i\in[m]}\|X_i^\top w_i\|_\infty\leq t\right)\geq 1-2\exp\left(-c_{24}\min\left\{\frac{t^2}{\sigma^2\zeta_\Sigma},\frac{t}{\sigma\sqrt{\zeta_\Sigma}}\right\}n+\log md\right),\quad t\geq 0,
$$

for some $c_{24}>0$. Thus, under $2\log md\leq c_{24}n$ and $t=\sigma\sqrt{\frac{2\zeta_\Sigma\log md}{nc_{24}}}$,

$$
\mathbb{P}\left(\frac{1}{n}\max_{i\in[m]}\|X_i^\top w_i\|_\infty\leq\sigma\sqrt{\frac{2\zeta_\Sigma\log md}{nc_{24}}}\right)
$$

$$\geq 1 - 2\exp\left(-c_{24}\min\left\{\frac{2\sigma^2\zeta_\Sigma \log md}{c_{24}n\sigma^2\zeta_\Sigma}, \frac{\sigma\sqrt{2\zeta_\Sigma \log md}}{\sqrt{c_{24}n\zeta_\Sigma}\sigma}\right\}n + \log md\right)$$
$$\geq 1 - 2\exp\left(-\log d\right), \tag{99}$$

while, under $2\log md > c_{24}n$ and $t = \frac{2\sigma\sqrt{\zeta_\Sigma}\log md}{nc_{24}}$, it holds

$$\mathbb{P}\left(\frac{1}{n}\max_{i\in[m]}\|X_i^\top w_i\|_\infty \leq \frac{2\sigma\sqrt{\zeta_\Sigma}\log md}{nc_{24}}\right)$$
$$\geq 1 - 2\exp\left(-c_{24}\min\left\{\frac{4\sigma^2\zeta_\Sigma \log^2 md}{c_{24}^2n^2\sigma^2\zeta_\Sigma}, \frac{2\sigma\sqrt{\zeta_\Sigma}\log md}{c_{24}n\sigma\sqrt{\zeta_\Sigma}}\right\}n + \log md\right)$$
$$\geq 1 - 2\exp\left(-\log d\right). \tag{100}$$

Combining (99) and (100), we have

$$\mathbb{P}\left(\frac{1}{n}\max_{i\in[m]}\|X_i^\top w_i\|_\infty \leq \sigma\sqrt{\zeta_\Sigma}\max\left\{\frac{2\log md}{nc_{24}}, \sqrt{\frac{2\log md}{nc_{24}}}\right\}\right) \geq 1 - 2\exp\left(-\log d\right). \tag{101}$$

Define

$$D_2 \triangleq \left\{\mathbf{w}\in\mathbb{R}^N \;\middle|\; \frac{1}{n}\max_{i\in[m]}\|X_i^\top w_i\|_\infty \leq \sigma\sqrt{\zeta_\Sigma}\max\left\{\frac{2\log md}{nc_{24}}, \sqrt{\frac{2\log md}{nc_{24}}}\right\}\right\},$$

and $D \triangleq D_1 \cap D_2$. Then, chaining (91), (101), and (97), we finally get

$$\begin{aligned}
&\mathbb{P}(A\cap D)\\
\geq& 1 - 2\exp(-c_3 d) - \exp(-c_0 N) - 2\exp(-c_6\log d) - 2\exp\{-[(t_0-2)\log d]/2 - 2\exp(-\log d)\\
=& 1 - 8\exp(-c_{25}\log d) - \exp(-c_0N).
\end{aligned} \tag{102}$$

where $c_{25} = \min\{1, c_3, c_6, (t_0-2)/2\}$.

- **Step 3: Sufficient condition on λ for** (13) **to hold with high probability.**

  We first recall (13) for convenience.

$$\frac{2}{N}\|\mathbf{X}^\top\mathbf{w}\|_\infty \leq \lambda.$$

Combining it with the high probability upper bound for $\|\mathbf{X}^\top\mathbf{w}\|_\infty/N$ in (97) (**Step 2**), we conclude that, if λ satisfies

$$\lambda = c_8\sigma\sqrt{\frac{\zeta_\Sigma t_0\log d}{N}}, \tag{103}$$

with $c_8 \geq \sqrt{6}$, then (13) holds with probability at least (102).

- **Step 4: Bounding the statistical error under** (21)**.**

  Recall the deterministic error bounds in Theorem 6: for any fixed λ satisfying (13) and $\gamma \leq \frac{2(1-\rho)}{4L_{\max}+\delta}$,

$$\frac{1}{m}\sum_{i=1}^m\|\hat{\nu}_i\|^2$$

$$\stackrel{(18)}{\leq} \frac{9\lambda^2 s}{\delta^2} + \frac{2\xi d^2\gamma^2}{\delta\lambda^2(1-\rho)^2}\left(\frac{\max_{i\in[m]}\|w_i^\top X_i\|_\infty}{n} + \lambda\right)^4 + \frac{4d\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty/n + \lambda)^2}{\delta[2(1-\rho) - 4L_{\max}\gamma - \delta\gamma]}.$$

In **Step 3**, we provided a sufficient condition on λ to guarantee (13) holds with probability at least as (102). Now we proceed to provide a sufficient condition on γ, not only to guarantee $\gamma \leq 2(1-\rho)/(4L_{\max}+\delta)$, but also contribute to restricting the error term above within the centralized statistical error, which is of the order $\mathcal{O}(\lambda^2 s)$.

By Lemma 5, if $N \geq 128 s c_1 \zeta_\Sigma \log d/\lambda_{\min}(\Sigma)$, with probability at least $1-\exp(-c_0 N)$, the in-network RE condition holds with $\delta = \lambda_{\min}(\Sigma)/4$ and $\xi = \lambda_{\min}(\Sigma)/(64s)$. Combining this with the high probability upper bound derived on $L_{\max}$ in (83) and the high probability upper bound derived for $\max_{i\in[m]}\|w_i^\top X_i\|_\infty/n$ in (101), we have

$$\frac{1}{m}\sum_{i=1}^m \|\hat{\nu}_i\|^2 \leq \frac{144\lambda^2 s}{\lambda_{\min}(\Sigma)^2} + \underbrace{\frac{d^2\gamma^2\left(\sigma\sqrt{\zeta_\Sigma}\max\left\{\frac{2\log md}{nc_{24}}, \sqrt{\frac{2\log md}{nc_{24}}}\right\} + \lambda\right)^4}{8\lambda^2 s(1-\rho)^2}}_{\texttt{Term I}}$$
$$+ \underbrace{\frac{8d\gamma\left(\sigma\sqrt{\zeta_\Sigma}\max\left\{\frac{2\log md}{nc_{24}}, \sqrt{\frac{2\log md}{nc_{24}}}\right\} + \lambda\right)^2}{\lambda_{\min}(\Sigma)\left(1-\rho-4\gamma c_4\lambda_{\max}(\Sigma)\left(1+\frac{d+\log m}{n}\right) - \frac{\gamma\lambda_{\min}(\Sigma)}{8}\right)}}_{\texttt{Term II}}$$

with probability larger than (102).

It remains to prove that condition (21) on γ is sufficient for `Term I` and `Term II` to be within $\mathcal{O}(\lambda^2 s)$. Notice that

$$\texttt{Term I} \leq \texttt{Term II}^2 \cdot \frac{\lambda_{\min}(\Sigma)^2}{512\lambda^2 s}. \tag{104}$$

Thus it is sufficient to bound only `Term II`. Enforcing `Term II` $\leq c_7\lambda^2 s/\lambda_{\min}(\Sigma)^2$, where c_7 is a numerical constant, we derive the following sufficient condition on γ to ensure `Term II` $\leq c_7\lambda^2 s/\lambda_{\min}(\Sigma)^2$:

$$\gamma \leq \frac{1-\rho}{8\lambda_{\min}(\Sigma)\frac{d}{c_7 s}\left[\frac{\sigma\sqrt{\zeta_\Sigma}}{\lambda}\max\left\{\frac{2\log md}{nc_{24}}, \sqrt{\frac{2\log md}{nc_{24}}}\right\} + 1\right]^2 + 4c_4\lambda_{\max}(\Sigma)[1+\frac{d+\log m}{n}] + \frac{\lambda_{\min}\Sigma}{8}}. \tag{105}$$

Thus, under (105), we have

$$\texttt{Term II} \leq c_7\frac{\lambda^2 s}{\lambda_{\min}(\Sigma)^2} \Rightarrow \texttt{Term I} \leq \frac{c_7^2\lambda^4 s^2}{\lambda_{\min}(\Sigma)^4}\frac{\lambda_{\min}(\Sigma)^2}{512\lambda^2 s} = \frac{c_7^2\lambda^2 s}{512\lambda_{\min}(\Sigma)^2}.$$

Therefore, the final statistical error satisfies

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m \|\hat{\nu}_i\|^2 &\leq \left(144 + c_7 + \frac{c_7^2}{512}\right)\frac{\lambda^2 s}{\lambda_{\min}(\Sigma)^2} \\ &= \left(144 + c_7 + \frac{c_7^2}{512}\right)\frac{c_8^2\sigma^2\zeta_\Sigma t_0}{\lambda_{\min}(\Sigma)^2}\frac{s\log d}{N}\end{aligned}$$

$$= \mathcal{O}\left(\frac{s\log d}{N}\right).$$

Since λ satisfies (103), we have

$$\begin{aligned}
&\frac{\sigma\sqrt{\zeta_\Sigma}}{\lambda}\max\left\{\frac{2\log md}{nc_{24}}, \sqrt{\frac{2\log md}{nc_{24}}}\right\} \\
\leq &\frac{1}{c_8}\sqrt{\frac{2m\log md}{c_{24}t_0\log d}}\max\left\{1, \sqrt{\frac{2\log md}{nc_{24}}}\right\}
\end{aligned} \tag{106}$$

Substituting (106) into (105), we have the following sufficient condition for (105)

$$\gamma \leq \frac{8n(1-\rho)c_7 s}{32c_4c_7s\lambda_{\max}(\Sigma)[n+(d+\log m)] + \lambda_{\min}(\Sigma)n\left\{64d\left[\frac{1}{c_8}\sqrt{\frac{2m\log md}{c_{24}t_0\log d}}\max\left\{1, \sqrt{\frac{2\log md}{nc_{24}}}\right\}+1\right]^2 + c_7 s\right\}}.$$

Notice that

$$\begin{aligned}
&64d\left[\frac{1}{c_8}\sqrt{\frac{2m\log md}{c_{24}t_0\log d}}\max\left\{1, \sqrt{\frac{2\log md}{nc_{24}}}\right\}+1\right]^2 \\
&\leq 64d\left[\frac{4m\log md}{c_8^2c_{24}t_0\log d}\max\left\{1, \frac{2\log md}{nc_{24}}\right\}+2\right] \\
&\leq 128d\left[\frac{2m\log md}{c_8^2c_{24}t_0\log d}\max\left\{1, \frac{2\log md}{nc_{24}}\right\}+1\right] \\
&\leq 128d\left[\frac{m\log md}{c_8^2c_{24}\log d}\max\left\{1, \frac{2\log md}{nc_{24}}\right\}+1\right], \\
&= 128d\left[\left(\frac{2m\log m}{2c_8^2c_{24}\log d}+\frac{m}{c_8^2c_{24}}\right)\max\left\{1, \frac{2\log md}{nc_{24}}\right\}+1\right], \\
&\leq 128d\left[\frac{3m\log m}{c_8^2c_{24}}\max\left\{1, \frac{2\log md}{nc_{24}}\right\}+1\right], \\
&\leq 128d\left[3c_{26}m\log m\max\left\{1, \frac{2\log md}{nc_{24}}\right\}+1\right],
\end{aligned} \tag{107}$$

thus, we have (21) is sufficient for (105).

This completes the proof. □

Appendix E. Proof of Lemma 8

Using (59), each $\|\hat{\theta}_i\|_1$ can be bounded as

$$\|\hat{\theta}_i\|_1 \leq \|\hat{\theta}_i - \theta^*\|_1 + \|\theta^*\|_1 = \|\hat{\nu}_{\text{av}} + \hat{\nu}_{\perp i}\|_1 + \|\theta^*\|_1 \leq \|\hat{\nu}_{\text{av}}\|_1 + \sqrt{d}\,\|\hat{\nu}_{\perp i}\| + \|\theta^*\|_1. \tag{108}$$

We bound next $\|\hat{\nu}_{\text{av}}\|_1$. By Proposition 3, any solution $\hat{\boldsymbol{\theta}}$ of (4) satisfies

$$\|(\hat{\nu}_{\text{av}})_{\mathcal{S}^c}\|_1 \leq 3\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1 + h(\gamma, \|\hat{\boldsymbol{\nu}}_\perp\|).$$

Therefore,

$$\begin{aligned}
\|\hat{\nu}_{\text{av}}\|_1 \leq & 4\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1 + h(\gamma, \|\hat{\boldsymbol{\nu}}_\perp\|) \\
\leq & 4\sqrt{s}\|\hat{\nu}_{\text{av}}\| + h(\gamma, \|\hat{\boldsymbol{\nu}}_\perp\|) \\
\leq & 4\sqrt{s}\frac{\|\hat{\boldsymbol{\nu}}\|}{\sqrt{m}} + h(\gamma, \|\hat{\boldsymbol{\nu}}_\perp\|) \\
\overset{(a)}{\leq} & 4\sqrt{s}\sqrt{\frac{9\lambda^2 s}{\delta^2} + \frac{2\tau d^2\gamma^2(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^4}{\delta\lambda^2 n^4(1-\rho)^2} + \frac{4d\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^2}{\delta n^2[2(1-\rho) - 4L_{\max}\gamma - \delta\gamma]}} \\
& + h(\gamma, \|\hat{\boldsymbol{\nu}}_\perp\|) \\
\leq & \frac{12\lambda s}{\delta} + \sqrt{\frac{32s\tau}{\delta}}\frac{d\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^2}{\lambda n^2(1-\rho)} + \sqrt{\frac{64sd\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^2}{\delta n^2[2(1-\rho) - 4L_{\max}\gamma - \delta\gamma]}} \\
& + h(\gamma, \|\hat{\boldsymbol{\nu}}_\perp\|),
\end{aligned} \tag{109}$$

where in (a) we used Theorem 6 and the fact that the RSC (8) implies the in-network RE (15) [cf. Lemma 4], with $\xi = \tau$ and $\delta = \mu/2 - 16s\tau > 0$.

Substituting (109) in (108) yields

$$\begin{aligned}
\|\hat{\theta}_i\|_1 \leq & \frac{12\lambda s}{\delta} + \sqrt{\frac{32s\tau}{\delta}}\frac{d\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^2}{\lambda n^2(1-\rho)} + \sqrt{\frac{64sd\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^2}{\delta n^2[2(1-\rho) - 4L_{\max}\gamma - \delta\gamma]}} \\
& + \underbrace{h(\gamma, \|\hat{\boldsymbol{\nu}}_\perp\|) + \sqrt{d}\|\hat{\boldsymbol{\nu}}_\perp\|}_{\triangleq h_2(\gamma, \|\hat{\boldsymbol{\nu}}_\perp\|)} + \|\theta^*\|_1 \\
\overset{\text{(a)}}{\leq} & \frac{12\lambda s}{\delta} + \sqrt{\frac{32s\tau}{\delta}}\frac{d\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^2}{\lambda n^2(1-\rho)} + \sqrt{\frac{64sd\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^2}{\delta n^2[2(1-\rho) - 4L_{\max}\gamma - \delta\gamma]}} \\
& + \frac{d\gamma(2\max_{i\in[m]}\|w_i^\top X_i\|_\infty + (2+\sqrt{m})\lambda n)^2}{4\lambda n^2(1-\rho)} + \|\theta^*\|_1 \\
\overset{\text{(b)}}{\leq} & \frac{12\lambda s}{\delta} + \sqrt{\frac{32s\tau}{\delta}}\underbrace{\frac{d\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^2}{\lambda n^2(1-\rho)}}_{\texttt{Term I}=h_{\max}} + \sqrt{\frac{64s}{\delta}}\underbrace{\sqrt{\frac{d\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^2}{n^2[2(1-\rho) - 4L_{\max}\gamma - \delta\gamma]}}}_{\texttt{Term II}} \\
& + \underbrace{\frac{d\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + 2\sqrt{m}\lambda n)^2}{\lambda n^2(1-\rho)}}_{\texttt{Term III}} + \|\theta^*\|_1,
\end{aligned} \tag{110}$$

where in (a) we bounded $h_2(\gamma, \bullet)$ on $\mathbb{R}$ as

$$\begin{aligned}
h_2(\gamma, \|\hat{\boldsymbol{\nu}}_\perp\|) & \overset{(12)}{=} -\frac{1-\rho}{\lambda m\gamma}\|\hat{\boldsymbol{\nu}}_\perp\|^2 + \left(2\max_{i\in[m]}\|w_i^\top X_i\|_\infty/(\lambda n) + 2\right)\sqrt{d/m}\|\hat{\boldsymbol{\nu}}_\perp\| + \sqrt{d}\|\hat{\boldsymbol{\nu}}_\perp\| \\
& \leq \frac{d\gamma\left(2\max_{i\in[m]}\|w_i^\top X_i\|_\infty + (2+\sqrt{m})\lambda n\right)^2}{4\lambda n^2(1-\rho)};
\end{aligned}$$

and in (b) we enlarged $2 + \sqrt{m} \leq 4\sqrt{m}$.

We bound now `Term I`–`Term III` using condition (25) on γ. We have the following:

$$h_{\max} = \texttt{Term I} \le \frac{\lambda s}{128\,\delta}, \tag{111}$$

$$\texttt{Term II} \le \sqrt{\frac{d\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + 2\sqrt{m}\lambda n)^2}{n^2[2(1-\rho) - 4L_{\max}\gamma - \delta\gamma]}} \le \sqrt{\frac{\lambda^2 s}{256\,\delta}},$$

$$\texttt{Term III} \le \frac{\lambda\, s}{128\,\delta}.$$

Substituting the above bounds in (110) we obtain

$$\begin{aligned}
\|\hat{\theta}_i\|_1 \le & \frac{12\lambda s}{\delta} + \sqrt{\frac{32\tau s}{\delta}}\frac{\lambda s}{128\delta} + \frac{\lambda s}{2\delta} + \frac{\lambda s}{128\delta} + \|\theta^*\|_1 \\
\le & \frac{\lambda s}{\delta}\left(13 + \frac{1}{32}\sqrt{\frac{2\tau s}{\delta}}\right) + \|\theta^*\|_1 \\
\overset{(26)}{\le} & (1-r)\cdot R + r\cdot R = R.
\end{aligned}$$

This completes the proof. $\square$.

Appendix F. Proof of Lemma 9

Since $L(\boldsymbol{\theta}) \triangleq \frac{1}{m}\sum_{i=1}^m f_i(\theta_i) + \frac{1}{2m\gamma}\|\boldsymbol{\theta}\|_V^2$, we have

$$\nabla L(\boldsymbol{\theta}^t) = \frac{1}{m}\begin{bmatrix}\nabla f_1(\theta_1^t)\\ \vdots \\ \nabla f_m(\theta_m^t)\end{bmatrix} + \frac{1}{m\gamma}\left((I-W)\otimes I_d\right)\boldsymbol{\theta}^t.$$

Substituting the expression of $\nabla L(\boldsymbol{\theta}^t)$ into Problem (24), it is not hard to see it is separable in the θ_i's, and the update of θ_i given as

$$\begin{aligned}
\theta_i^{t+1} &= \operatorname*{arg\,min}_{\|\theta_i\|_1\le R}\ \left\langle \frac{1}{m}\nabla f_i(\theta_i^t) + \frac{1}{m\gamma}\left(\theta_i^t - \sum_{j=1}^m w_{ij}\theta_j^t\right), \theta_i - \theta_i^t\right\rangle + \frac{1}{2\beta m}\|\theta_i - \theta_i^t\|^2 + \frac{\lambda}{m}\|\theta_i\|_1 \\
&= \operatorname*{arg\,min}_{\|\theta_i\|_1\le R}\ \frac{1}{2}\left\|\theta_i - \theta_i^t + \beta\nabla f_i(\theta_i^t) + \frac{\beta}{\gamma}\left(\theta_i^t - \sum_{j=1}^m w_{ij}\theta_j^t\right)\right\|^2 + \beta\lambda\|\theta_i\|_1.
\end{aligned}$$

The problem boils down to solving

$$\begin{aligned}
\min_{\theta_i}\quad & \|\theta_i - \psi_i^t\|^2 + \lambda'\|\theta_i\|_1 \\
\text{s.t.}\quad & \|\theta_i\|_1 \le R,
\end{aligned} \tag{112}$$

with $\lambda' \triangleq \frac{2\beta\lambda}{m}$ and ψ_i^t defined in (27).

To solve (112) we first drop the constraint $\|\theta_i\|_1 \le R$. The minimizer of the objective function is given by

$$\tilde{\theta}_i = \text{prox}_{\frac{\lambda'}{2}\|\cdot\|_1}(\psi_i^t). \tag{113}$$

Note that $\tilde{\theta}_i$ can be computed in closed form by soft-thresholding ψ_i^t.

Case 1: $\tilde{\theta}_i$ satisfies the constraint in (112), i.e., $\|\tilde{\theta}_i\|_1 \leq R$. We conclude that $\tilde{\theta}_i$ is a solution of (112).

Case 2: $\tilde{\theta}_i$ violates the constraint in (112), i.e., $\|\tilde{\theta}_i\|_1 > R$. Then, the constraint must be active at the optimal point of (112). Hence, Problem (112) is equivalent to

$$\begin{aligned} \min_{\theta_i} \quad & \|\theta_i - \psi_i^t\|^2 \\ \text{s.t.} \quad & \|\theta_i\|_1 = R, \end{aligned} \tag{114}$$

where we dropped the term $\lambda'\|\theta_i\|_1$ in the objective function, since it is constant on the constraint set. Since (113) can be computed in closed form by soft-thresholding ψ_i^t, we conclude $\|\psi_i^t\|_1 \geq \|\tilde{\theta}_i\|_1 > R$, and thus the convex problem with constraint (114) is equivalent to

$$\begin{aligned} \min_{\theta_i} \quad & \|\theta_i - \psi_i^t\|^2 \\ \text{s.t.} \quad & \|\theta_i\|_1 \leq R. \end{aligned} \tag{115}$$

Combining the two cases completes the proof. $\square$

Appendix G. Proof of Theorem 10

Recall the factorization of the objective function by G and L as introduced in (63)

$$G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \frac{\lambda}{m}\|\boldsymbol{\theta}\|_1, \quad \text{with} \quad L(\boldsymbol{\theta}) = \frac{1}{2N}\sum_{i=1}^{m}\|y_i - X_i\theta_i\|^2 + \frac{1}{2m\gamma}\|\boldsymbol{\theta}\|_V^2.$$

We begin (**Step 1**) proving a weaker result than Theorem 10, that is, linear convergence of the error $G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}})$, up to the tolerance as on the RHS of (35)–this is Theorem 14 below. Then (**Step 2**), leveraging the curvature property of G along the trajectory of the algorithm (see Lemma 16 in Appendix J.1), we transfer the rate decay of $G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}})$ on that of the iterates error $\|\boldsymbol{\theta}^t - \hat{\boldsymbol{\theta}}\|$, which completes the proof of Theorem 10.

- **Step 1: On linear convergence of the optimality gap $G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}})$.**

Recall the definition of $\varepsilon_{\text{stat}}^2$ and μ_{av} as given in (29) and (28), respectively.

Theorem 14 *Instate the setting of Theorem 10. There holds:*

$$G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}}) \leq \alpha^2, \tag{116}$$

for any tolerance parameter α^2 such that

$$\min\left\{\frac{R\lambda}{4}, \eta_G^0\right\} \geq \alpha^2 \geq 4s\tau \cdot \varepsilon_{\text{stat}}^2, \tag{117}$$

and for all

$$t \geq \left\lceil \log_2\log_2\left(\frac{R\lambda}{\alpha^2}\right)\right\rceil\left(1 + \frac{L_{\max}\log 2}{\mu_{\text{av}}} + \frac{(1+\rho)\log 2}{\gamma\mu_{\text{av}}}\right) + \left(\frac{L_{\max}}{\mu_{\text{av}}} + \frac{1+\rho}{\gamma\mu_{\text{av}}}\right)\log\left(\frac{\eta_G^0}{\alpha^2}\right). \tag{118}$$

Furthermore, the interval in (117) is nonempty.

Proof See Appendix J. ∎

• **Step 2: On linear convergence of the optimality gap $\|\boldsymbol{\theta}^t - \hat{\boldsymbol{\theta}}\|$.**

We can now proceed to prove Theorem 10. Given Theorem 14, it is sufficient to show that (35) holds.

Recall the shorthand for the optimization error, $\boldsymbol{\Delta}^t = \boldsymbol{\theta}^t - \hat{\boldsymbol{\theta}}$. At high-level the idea is to construct a lower bound of $G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}})$ as a function of $\|\boldsymbol{\Delta}^t\|^2$ by exploiting, under the RSC condition (8), the curvature property of G along a restricted set of directions. Specifically, we use the following curvature property proved in Lemma 16 (cf. Sec. J.1), which holds under the more stringent setting of Theorem 14[1]: for all $t \geq T$,

$$\mu_{\text{av}}\|\Delta_{\text{av}}^t\|^2 - f(\|\boldsymbol{\Delta}_{\perp}^t\|) \leq G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}}) + \frac{\tau}{4}(v^2 + 8h_{\max}^2), \tag{119}$$

where $f(\|\boldsymbol{\Delta}_{\perp}^t\|)$ is defined as [cf. (14)],

$$f(\|\boldsymbol{\Delta}_{\perp}^t\|) = \left(\frac{L_{\max}}{2m} - \frac{1-\rho}{2m\gamma}\right)\|\boldsymbol{\Delta}_{\perp}^t\|^2,$$

v^2 is given by [cf. (145)]

$$v^2 = 144s\|\hat{\nu}_{\text{av}}\|_2^2 + 4\min\left\{\frac{2\eta}{\lambda}, 2R\right\}^2, \quad \text{with} \quad \eta = \alpha^2, \tag{120}$$

and $h_{\max}$ is defined as [cf. (74)]

$$h_{\max} = \frac{d\gamma}{\lambda(1-\rho)}\left(\frac{\max_{i\in[m]}\|w_i^\top X_i\|_\infty}{n} + \lambda\right)^2.$$

We proceed now to bound the LHS and RHS of (119). The goal is to lower bound the LHS by a quantity proportional to $\|\boldsymbol{\Delta}^t\|^2$, so that (119) will provide the desired bound of $\|\boldsymbol{\Delta}^t\|^2$ in terms of the optimization gap $G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}})$ (up to a tolerance). The following bound of $f(\|\boldsymbol{\Delta}_{\perp}^t\|)$, which is a consequence of (174) serves the scope:

$$f(\|\boldsymbol{\Delta}_{\perp}^t\|) \leq -\frac{\mu_{\text{av}}}{m}\|\boldsymbol{\Delta}_{\perp}^t\|^2.$$

We also upper bound the RHS of (119) to further simplify the final expression; specifically, we use

$$h_{\max} \overset{(111)}{\leq} \frac{\lambda s}{64(\mu - 32s\tau)} \tag{121}$$

and

$$144s\|\hat{\nu}_{\text{av}}\|^2 + 4\min\left\{\frac{2\alpha^2}{\lambda}, 2R\right\}^2 \leq \frac{144s\|\hat{\boldsymbol{\nu}}\|^2}{m} + \frac{16\alpha^4}{\lambda^2},$$

where the inequality follows from $\|\hat{\nu}_{\text{av}}\|^2 \leq \|\hat{\boldsymbol{\nu}}\|^2/m$ and the fact that $\alpha^2 \leq R\lambda/4$ [cf. (36)].

1. Specifically, in the proof of Theorem 14, we showed that condition on R as in (33) is more stringent than (26) in Lemma 16–see Fact 1 in Appendix J.2.

Using the above bounds along with (116) in (119) and rearranging terms yield: for all $t \geq T$,

$$\mu_{\text{av}} \frac{\|\boldsymbol{\Delta}^t\|^2}{m} \leq \alpha^2 + \frac{\tau}{4}\left(\frac{144 s\|\hat{\boldsymbol{\nu}}\|^2}{m} + 8\left(\frac{\lambda s}{64(\mu - 32 s\tau)}\right)^2 + \frac{16\alpha^4}{\lambda^2}\right)$$
$$\leq \alpha^2 + \frac{36 s\tau\|\hat{\boldsymbol{\nu}}\|^2}{m} + \frac{\tau s \lambda^2 s}{1976\mu^2} + \frac{4\tau\alpha^4}{\lambda^2}, \tag{122}$$

where the last inequality follows from $\mu \geq c_{10} s\tau$ with $c_{10} = 1824$. This proves (35), and completes the proof. □

Appendix H. Proof of Theorem 12

Given the postulated random data model and the conclusions of Theorem 10, it is sufficient to prove the following: **Step 1:** Under (40) on λ, condition (30) holds with high-probability; **Step 2:** Condition (41) on γ is sufficient for (31) to hold with high probability; **Step 3:** Under condition (39) on N, any R in (42) satisfies also (33); furthermore, the interval of values of R in (42) is nonempty; **Step 4:** Any α^2 in (45) satisfies (36) with high-probability; and the range for α^2 in (45) is nonempty with high-probability; **Step 5:** Given the bound on the statistical error as in (35) and for all t satisfying (37), we conclude that (44) holds, for all t satisfying (46), with high-probability.

• **Step 1: Sufficient condition on λ for** (30) **to hold with high probability.** To prove this result, we follow a similar path as introduced in the proof of Theorem 7.

(i) Randomness from X. This step is the same as **Step 1** in the proof of Theorem 7 (cf. Appendix D), except for the definition of the event A_2 replaced here with A_2', defined as

$$A_2' \triangleq \left\{\mathbf{X} \in \mathbb{R}^{N\times d} \,\middle|\, \mathbf{X} \text{ satisfies RSC (8) with parameters } (\mu,\tau) = \left(\lambda_{\min}(\Sigma), 2c_1\zeta_\Sigma \frac{\log d}{N}\right)\right\}, \tag{123}$$

where $\zeta_\Sigma = \max_{i\in[d]} \Sigma_{ii}$. Lemma 2 implies

$$\mathbb{P}(A_2') \geq 1 - \exp(-c_0 N). \tag{124}$$

Define $A' = A_1 \cap A_2' \cap A_3$, where A_1 and A_3 are defined in (77) and (78) (cf. Appendix D), respectively, and recall here for convenience

$$A_1 \triangleq \left\{\mathbf{X} \in \mathbb{R}^{N\times d} \,\middle|\, L_{\max} \leq c_4 \lambda_{\max}(\Sigma)\left(1 + \frac{d + \log m}{n}\right)\right\} \text{ and}$$
$$A_3 \triangleq \left\{\mathbf{X} \in \mathbb{R}^{N\times d} \,\middle|\, \max_{j=1,\ldots,d} \frac{1}{\sqrt{N}}\|\mathbf{X}e_j\| \leq \sqrt{\frac{3\zeta_\Sigma}{2}}\right\}.$$

Then, similar to under condition (90), there holds (91), since N satisfies (39) with $c_{12} = \max\{c_9, c_{11}\}$ $c_9 = \max\{128c_1, c_5\}$, and $c_{11} = 3648 c_1$, we have

$$\mathbb{P}(A') \geq 1 - 2\exp(-c_3 d) - \exp(-c_0 N) - 2\exp(-c_6 \log d). \tag{125}$$

(ii) Randomness from w. This step follows **Step 2** as in the proof of Theorem 7 (cf. Appendix D) and thus is not duplicated. In particular, recalling the definitions of D_1, D_2, and D therein for convenience

$$D_1 \triangleq \left\{ \mathbf{w} \in \mathbb{R}^N \;\middle|\; \frac{\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N} \leq \sigma \sqrt{\frac{t_0 \log d}{N}} \sqrt{\frac{3\zeta_\Sigma}{2}} \right\},$$

$$D_2 \triangleq \left\{ \mathbf{w} \in \mathbb{R}^N \;\middle|\; \frac{\max_{i\in[m]} \|X_i^\top w_i\|_\infty}{n} \leq \sigma\sqrt{\zeta_\Sigma} \max\left\{ \frac{2\log md}{nc_{24}}, \sqrt{\frac{2\log md}{nc_{24}}} \right\} \right\}, \quad \text{and} \quad D \triangleq D_1 \cap D_2;$$

and following the same reasoning as in **Step 2** of the proof of Theorem 7, we have, for all $t_0 \geq 2$,

$$\mathbb{P}(A' \cap D) \geq 1 - 8\exp(-c_{25} \log d) - \exp(-c_0 N). \tag{126}$$

(iii) Sufficient condition on λ for (30) to hold with high probability. Recall (30) for convenience,

$$\lambda \geq \max\left\{ \frac{2\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N}, 64\tau\|\theta^*\|_1 \right\}.$$

Combining it with the high probability upper bound for $\|\mathbf{X}^\top \mathbf{w}\|_\infty/N$ derived in (126), we conclude the following: suppose $\lambda \geq \sigma\sqrt{\frac{6\zeta_\Sigma t_0 \log d}{N}}$, then for any tuple $(\mathbf{X}, \mathbf{w}) \in A' \cap D$, and any $t_0 > 2$,, since $2\|\mathbf{X}^\top \mathbf{w}\|/N \leq \sigma\sqrt{\frac{6\zeta_\Sigma t_0 \log d}{N}}$, it follows that $\lambda \geq 2\|\mathbf{X}^\top \mathbf{w}\|_\infty/N$.

That is,

$$\begin{aligned} \mathbb{P}\left(\lambda \geq \frac{2\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N} \right) &\geq \mathbb{P}(A' \cap D) \\ &\overset{(126)}{\geq} 1 - 8\exp(-c_{25} \log d) - \exp(-c_0 N). \end{aligned}$$

Furthermore, for any tuple $(\mathbf{X}, \mathbf{w}) \in A' \cap D$, (123) implies $\tau = 2c_1\zeta_\Sigma \log d/N$. Therefore it follows that if $\lambda \geq \frac{128sc_1\zeta_\Sigma \log d}{N}$, then $64\tau\|\theta^*\|_1$. We conclude that for any $t_0 > 2$, display (40), i.e.

$$\lambda \geq \max\left\{ \sigma\sqrt{\frac{6\zeta_\Sigma t_0 \log d}{N}}, \frac{128sc_1\zeta_\Sigma \log d}{N} \right\},$$

is sufficient for (30) to hold with probability at least (126).

- **Step 2:** (41) **is sufficient for** (31) **to hold with high probability.**

Recall (31) for convenience,

$$\gamma \leq \frac{1-\rho}{2L_{\max} + (\mu/2 - 16s\tau)\left(1 + 128(d/s)(\max_{i\in[m]}\|w_i^\top X_i\|_\infty/(\lambda n) + 2\sqrt{m})^2\right)}.$$

In order to derive a sufficient condition on γ to ensure (31) holds with high probability, we leverage **Step 1 (i)** above, where we derived high probability bounds for $L_{\max}$, $\max_{i\in[m]}\|w_i^\top X_i\|_\infty/n$, and λ. Specifically, substituting into (31) the bounds on $L_{\max}$ [as in

(83)], $\max_{i\in[m]}\|X_i^\top w_i\|_\infty/n$ [as in (126)], and the explicit expression of the RSC parameters (μ,τ) [as in (123)], we conclude that if

$$\gamma \le \frac{1-\rho}{2c_4\lambda_{\max}(\Sigma)\left(1+\frac{d+\log m}{n}\right)+\left(\frac{\lambda_{\min}(\Sigma)}{2}-\frac{32sc_1\zeta_\Sigma\log d}{N}\right)\left[1+\frac{128d}{s}\left(\frac{\sigma}{\lambda}\sqrt{\zeta_\Sigma}\max\left\{\frac{2\log md}{nc_{24}},\sqrt{\frac{2\log md}{nc_{24}}}\right\}+2\sqrt{m}\right)^2\right]}, \tag{127}$$

then (31) holds with probability at least (23).

We conclude this step by showing that (41) is sufficient for (127). Specifically, since

$$\begin{aligned}
&\frac{\sigma}{\lambda}\sqrt{\zeta_\Sigma}\max\left\{\frac{2\log md}{nc_{24}},\sqrt{\frac{2\log md}{nc_{24}}}\right\}\\
&\overset{(40)}{\le}\min\left\{\sqrt{\frac{N}{6t_0\log d}}\max\left\{\frac{2\log md}{nc_{24}},\sqrt{\frac{2\log md}{nc_{24}}}\right\},\frac{N\sigma}{128sc_1\zeta_\Sigma\log d}\sqrt{\zeta_\Sigma}\max\left\{\frac{2\log md}{nc_{24}},\sqrt{\frac{2\log md}{nc_{24}}}\right\}\right\}\\
&\le\sqrt{\frac{m\log md}{3t_0c_{24}\log d}}\max\left\{\sqrt{\frac{2\log md}{nc_{24}}},1\right\},
\end{aligned}$$

and

$$\frac{\lambda_{\min}(\Sigma)}{2}\ge\frac{\lambda_{\min}(\Sigma)}{2}-\frac{32sc_1\zeta_\Sigma\log d}{N}\overset{(39)}{\ge}0,$$

we obtain the more conservative condition

$$\gamma \le \frac{1-\rho}{2c_4\lambda_{\max}(\Sigma)\left(1+\frac{d+\log m}{n}\right)+\frac{\lambda_{\min}(\Sigma)}{2}\left[1+\frac{128d}{s}\left(\sqrt{\frac{m\log md}{3t_0c_{24}\log d}}\max\left\{\sqrt{\frac{2\log md}{nc_{24}}},1\right\}+2\sqrt{m}\right)^2\right]}. \tag{128}$$

We proceed to further simplify (128) by strengthening the condition. Observe that

$$\begin{aligned}
&1+\frac{128d}{s}\left(\sqrt{\frac{m\log md}{3t_0c_{24}\log d}}\max\left\{\sqrt{\frac{2\log md}{nc_{24}}},1\right\}+2\sqrt{m}\right)^2\\
&\overset{(a)}{\le}1+\frac{128d}{s}\cdot\left(\sqrt{\frac{\log md}{3t_0c_{24}\log d}}+2\right)^2 m\max\left\{\sqrt{\frac{2\log md}{nc_{24}}},1\right\}^2\\
&\overset{(b)}{\le}1+\frac{128d}{s}\cdot\left[\frac{2}{3c_{24}}\left(2\log m+1\right)+8\right]m\max\left\{\frac{2\log md}{nc_{24}},1\right\}\\
&\overset{(c)}{\le}256dm\cdot\left[\left(2\log m+1\right)/c_{24}+8\right]\max\left\{\frac{2\log md}{nc_{24}},1\right\},
\end{aligned}\tag{129}$$

where (a) follows from

$$\left(\sqrt{\frac{m\log md}{3t_0c_{24}\log d}}\max\left\{\sqrt{\frac{2\log md}{nc_{24}}},1\right\}+2\sqrt{m}\right)\le\left(\sqrt{\frac{\log md}{3t_0c_{24}\log d}}+2\right)\sqrt{m}\max\left\{\sqrt{\frac{2\log md}{nc_{24}}},1\right\};$$

(b) is due to

$$
\begin{aligned}
&\left(\sqrt{\frac{\log md}{3t_0c_{24}\log d}}+2\right)^2 m\max\left\{\sqrt{\frac{2\log md}{nc_{24}}},1\right\}^2\\
&\le\left(\sqrt{\frac{\log m}{6c_{24}\log d}}+\sqrt{\frac{1}{6c_{24}}}+2\right)^2 m\max\left\{\sqrt{\frac{2\log md}{nc_{24}}},1\right\}^2\\
&=\left[\sqrt{\frac{1}{6c_{24}}}\left(\sqrt{\frac{2\log m}{2\log d}}+1\right)+2\right]^2 m\max\left\{\sqrt{\frac{2\log md}{nc_{24}}},1\right\}^2\\
&\overset{\text{(d)}}{\le}\left[\sqrt{\frac{1}{6c_{24}}}\left(\sqrt{2\log m}+1\right)+2\right]^2 m\max\left\{\sqrt{\frac{2\log md}{nc_{24}}},1\right\}^2\\
&\le\left[\frac{2}{3c_{24}}\left(2\log m+1\right)+8\right] m\max\left\{\frac{2\log md}{nc_{24}},1\right\},
\end{aligned}
\tag{130}
$$

and (d) is due to $1 \le 2\log d$, for $d \ge 2$; and in (c), we upper bound both 1 and $128d/s \cdot \left[\frac{2}{3c_{24}}\left(2\log m+1\right)+8\right] m\max\left\{\frac{2\log md}{nc_{24}},1\right\}$ by $128d \cdot \left[\left(2\log m+1\right)/c_{24}+8\right]\max\left\{\frac{2\log md}{nc_{24}},1\right\}$. Using (129) and further simplification, we obtain the final condition (41) for (31) to hold with probability at least (23).

• **Step 3: Ensuring there exists an R fulfilling** (33).

Substituting in (33) the explicit expression of the RSC parameters (μ,τ) [under the event in (123)] as well as $\|\theta^*\|_1 = s$, we conclude that (33) holds with probability at least $1-\exp(-c_0N)$, whenever R satisfies display (42),

$$
\max\left\{\frac{56\lambda s}{\lambda_{\min}(\Sigma)-64sc_1\zeta_\Sigma\log d/N},2s\right\}\le R\le\frac{\lambda N}{64c_1\zeta_\Sigma\log d}.
$$

We now show that the interval (42) is non-empty. Since N satisfies (39) with $c_{12} = \max\{128c_1, c_5, c_{11}\}$, and $c_{11} = 3648c_1$, there holds

$$
\frac{56\lambda s}{\lambda_{\min}(\Sigma)-64sc_1\zeta_\Sigma\log d/N}\le\frac{\lambda N}{64c_1\zeta_\Sigma\log d}. \tag{131}
$$

Furthermore, (40) [**Step 1 (iii)**] implies

$$
\lambda\ge\frac{128sc_1\zeta_\Sigma\log d}{N},
$$

and thus

$$
2s\le\frac{\lambda N}{64c_1\zeta_\Sigma\log d}. \tag{132}
$$

By (131) and (132), we infer that (42) is non-empty.

• **Step 4:** (45) **is sufficient for** (36) **to hold with high-probability.** Substituting in (36) the explicit expression of the RSC parameters (μ,τ) [under the event (123)], we conclude that (45) is in fact sufficient for (36) to hold with probability at least (126).

It remains to prove that the (random) interval (45) is non-empty with high probability, which we do next. To this end, we upper bound the statistical error $\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2/m$, under (36). Recall that, with probability at least (126), (31) holds. Therefore, we can invoke Theorem 6 to bound the statistical error, and write: with probability at least (126),

$$
\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 \\
&\overset{\text{Thm. 6}}{\leq} \frac{9\lambda^2 s}{\delta^2} + \frac{2\xi}{\delta} \underbrace{\frac{d^2\gamma^2(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^4}{\lambda^2 n^4 (1-\rho)^2}}_{(\texttt{Term I})^2 \texttt{ in } (110)} + \frac{4}{\delta} \underbrace{\frac{d\gamma(\max_{i\in[m]}\|w_i^\top X_i\|_\infty + \lambda n)^2}{n^2[2(1-\rho) - 4L_{\max}\gamma - \delta\gamma]}}_{(\texttt{Term II})^2 \texttt{ in } (110)} \\
&\overset{(111)}{\leq} \frac{9\lambda^2 s}{\delta^2} + \frac{2\xi}{\delta}\left(\frac{\lambda s}{128\,\delta}\right)^2 + \frac{4}{\delta}\frac{\lambda^2 s}{256\,\delta} \\
&\overset{(a)}{=} \frac{9\lambda^2 s}{(\mu/2 - 16s\tau)^2} + \frac{2\tau}{\mu/2 - 16s\tau}\left(\frac{\lambda s}{128\,(\mu/2 - 16s\tau)}\right)^2 + \frac{4}{\mu/2 - 16s\tau}\frac{\lambda^2 s}{256\,(\mu/2 - 16s\tau)} \\
&\overset{(123)}{\leq} \frac{1}{(\lambda_{\min}(\Sigma) - 64sc_1\zeta_\Sigma \log d/N)^2}\left(36\lambda^2 s + \frac{16}{128^2}\frac{2sc_1\zeta_\Sigma \log d}{N}\frac{\lambda^2 s}{(\lambda_{\min}(\Sigma) - 64sc_1\zeta_\Sigma \log d/N)} + \frac{16}{256}\lambda^2 s\right),
\end{aligned}
$$

where in (a) we used $\xi = \tau$ and $\delta = \mu/2 - 16s\tau > 0$ (due to Lemma 4).

Thus, with probability at least (126), we can upper bound the lower interval bound in (45) by

$$
\begin{aligned}
&\frac{8sc_1\zeta_\Sigma \log d}{N}\left(\frac{36}{m}\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 + \frac{\lambda^2 s}{1976\lambda_{\min}(\Sigma)^2}\right) \\
&\leq \frac{8\cdot 36 sc_1\zeta_\Sigma \log d}{N(\lambda_{\min}(\Sigma) - 64sc_1\zeta_\Sigma \log d/N)^2}\left(36\lambda^2 s + \frac{16}{128^2}\frac{2sc_1\zeta_\Sigma \log d}{N}\frac{\lambda^2 s}{(\lambda_{\min}(\Sigma) - 64sc_1\zeta_\Sigma \log d/N)} + \frac{16}{256}\lambda^2 s\right) \\
&\quad + \frac{8sc_1\zeta_\Sigma \log d}{N}\frac{\lambda^2 s}{1976(\lambda_{\min}(\Sigma) - 64sc_1\zeta_\Sigma \log d/N)^2} \\
&\leq \frac{288 sc_1\zeta_\Sigma \log d}{N(\lambda_{\min}(\Sigma) - 64sc_1\zeta_\Sigma \log d/N)^2}\left(\frac{sc_1\zeta_\Sigma \log d}{512N}\frac{\lambda^2 s}{(\lambda_{\min}(\Sigma) - 64sc_1\zeta_\Sigma \log d/N)} + 37\lambda^2 s\right).
\end{aligned}
\tag{133}
$$

Using the bound on N given by (39)

$$
N \geq \frac{c_{12} s\zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)}, \quad \text{with} \quad c_{12} = \max\{3648c_1, c_5\}
$$

we obtain

$$
64sc_1\zeta_\Sigma \log d/N \leq \lambda_{\min}(\Sigma)/57.
$$

Substituting into the inequality above we have

$$
\frac{8sc_1\zeta_\Sigma \log d}{N}\left(\frac{36}{m}\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 + \frac{\lambda^2 s}{1976\lambda_{\min}(\Sigma)^2}\right)
$$

$$\leq \frac{288 s c_1 \zeta_\Sigma \log d}{N(\lambda_{\min}(\Sigma) - 64 s c_1 \zeta_\Sigma \log d/N)\left(\frac{56}{57}\lambda_{\min}(\Sigma)\right)}\left(\lambda^2 s + 37\lambda^2 s\right).$$

Applying again the lowerbound on N we can further get

$$\begin{aligned} & N(\lambda_{\min}(\Sigma) - 64 s c_1 \zeta_\Sigma \log d/N) \\ & \qquad \geq \max\left\{\frac{c_{12} s \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)}(\lambda_{\min}(\Sigma) - 64 s c_1 \zeta_\Sigma \log d/N), N \cdot \frac{56}{57}\lambda_{\min}(\Sigma)\right\}, \end{aligned} \tag{134}$$

and thus

$$\begin{aligned} & \frac{8 s c_1 \zeta_\Sigma \log d}{N}\left(\frac{36}{m}\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 + \frac{\lambda^2 s}{1976 \lambda_{\min}(\Sigma)^2}\right) \\ \leq & \frac{288 s c_1 \zeta_\Sigma \log d}{\frac{56}{57}\lambda_{\min}(\Sigma)}\left(38\lambda^2 s\right) \cdot \min\left\{\frac{\lambda_{\min}(\Sigma)}{c_{12} s \zeta_\Sigma \log d}(\lambda_{\min}(\Sigma) - 64 s c_1 \zeta_\Sigma \log d/N)^{-1}, \frac{57}{56}\lambda_{\min}(\Sigma)^{-1}\right\} \\ \leq & \min\left\{4(\lambda_{\min}(\Sigma) - 64 s c_1 \zeta_\Sigma \log d/N)^{-1}\lambda^2 s, \quad 11339 \cdot \frac{s c_1 \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)^2}\left(\lambda^2 s\right)\right\} \leq \min\left\{\frac{\lambda R}{4}, \eta_G^0\right\}. \end{aligned}$$

The last inequality follows from the conditions on R and η_G^0 given by (42) and (43), respectively.

• **Step 5:** (44) **holds, for all** t **satisfying** (46), with high probability.

Building on the conclusions of the previous steps and Theorem 10, to prove the statement of this step, it is sufficient to show that the RHS of (44) [resp. of (46)] is an upper bound of the RHS of (35) [resp. (37)] that holds with high probability.

We begin with the RHS of (35): with probability at least (126), there holds,

$$\begin{aligned} & \frac{1}{\mu/8 - 8 s\tau}\alpha^2 + \frac{36 s\tau \|\hat{\boldsymbol{\nu}}\|^2}{m(\mu/8 - 8 s\tau)} + \frac{\tau s \lambda^2 s}{1976 \mu^2(\mu/8 - 8 s\tau)} + \frac{4\tau\alpha^4}{\lambda^2(\mu/8 - 8 s\tau)} \\ \overset{(a)}{\leq} & \frac{456}{55\lambda_{\min}(\Sigma)}\left(\alpha^2 + \frac{72 s c_1 \zeta_\Sigma \log d}{N}\frac{1}{m}\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 + \frac{s c_1 \zeta_\Sigma \log d}{988 N}\frac{\lambda^2 s}{\lambda_{\min}(\Sigma)^2} + \frac{8 s c_1 \zeta_\Sigma \log d}{N}\frac{\alpha^4}{\lambda^2 s}\right) \\ \leq & \frac{9}{\lambda_{\min}(\Sigma)}\left(\alpha^2 + \frac{72 s c_1 \zeta_\Sigma \log d}{N}\frac{1}{m}\sum_{i=1}^m \|\hat{\theta}_i - \theta^*\|^2 + \frac{s c_1 \zeta_\Sigma \log d}{988 N}\frac{\lambda^2 s}{\lambda_{\min}(\Sigma)^2} + \frac{8 s c_1 \zeta_\Sigma \log d}{N}\frac{\alpha^4}{\lambda^2 s}\right), \end{aligned}$$

where in (a) we use the following fact (which holds probability at least (126))

$$\frac{\mu}{8} - 8 s\tau = \frac{\lambda_{\min}(\Sigma)}{8} - 16 s c_1 \zeta_\Sigma \frac{\log d}{N} \overset{(39)}{\geq} \frac{\lambda_{\min}(\Sigma)}{8} - \frac{16 s c_1 \zeta_\Sigma \log d \lambda_{\min}(\Sigma)}{3648 c_1 s \zeta_\Sigma \log d} = \frac{55 \lambda_{\min}(\Sigma)}{456}.$$

Next, we bound the RHS of (37), invoking the high probability bound for $L_{\max}$ [as in (83)] and the explicit expression of the RSC parameters (μ, τ) [under (123)]. We have the following

$$
\begin{aligned}
&\left\lceil \log_2 \log_2 \left(\frac{R\lambda}{\alpha^2}\right)\right\rceil \left(1 + \frac{L_{\max} \log 2}{\mu_{\text{av}}} + \frac{(1+\rho)\log 2}{\gamma \mu_{\text{av}}}\right) + \left(\frac{L_{\max}}{\mu_{\text{av}}} + \frac{1+\rho}{\gamma\mu_{\text{av}}}\right) \log\left(\frac{\eta_G^0}{\alpha^2}\right) \\
\leq & \left\lceil \log_2 \log_2 \left(\frac{R\lambda}{\alpha^2}\right)\right\rceil \left(1 + \frac{c_4 \lambda_{\max}(\Sigma)[1 + (d+\log m)/n]\log 2}{\lambda_{\min}(\Sigma)/8 - \lambda_{\min}(\Sigma)/228} + \frac{(1+\rho)\log 2}{\gamma[\lambda_{\min}(\Sigma)/8 - \lambda_{\min}(\Sigma)/228]}\right) \\
& + \left(\frac{c_4 \lambda_{\max}(\Sigma)[1 + (d+\log m)/n]}{\lambda_{\min}(\Sigma)/8 - \lambda_{\min}(\Sigma)/228} + \frac{1+\rho}{\gamma[\lambda_{\min}(\Sigma)/8 - \lambda_{\min}(\Sigma)/228]}\right) \log\left(\frac{\eta_G^0}{\alpha^2}\right) \\
= & \left\lceil \log_2 \log_2 \left(\frac{R\lambda}{\alpha^2}\right)\right\rceil \left(1 + \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \frac{456 c_4 [1 + (d+\log m)/n]\log 2}{55} + \frac{456(1+\rho)\log 2}{55\lambda_{\min}(\Sigma)\gamma}\right) \\
& + \left(\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \frac{456 c_4[1 + (d+\log m)/n]}{55} + \frac{456(1+\rho)}{55\gamma\lambda_{\min}(\Sigma)}\right) \log\left(\frac{\eta_G^0}{\alpha^2}\right).
\end{aligned}
$$

This completes the proof. □

Appendix I. Proof of Corollary 13

The corollary is a customization of Theorem 12, under the (feasible) choices of N, λ and γ as in the statement of the corollary.

• **Step 1: On the choices of λ and γ.** We show that, under (47), (48) and (49) are special instances of (40) and (41), respectively.

Since N satisfies (47), with $c_{12} = \max\{3648\, c_1, c_5\}$ and $c_{13} = 2731 c_1^2/t_0$, there holds

$$\sigma\sqrt{\frac{6\zeta_\Sigma t_0 \log d}{N}} \geq \frac{128 s c_1 \zeta_\Sigma \log d}{N}. \tag{135}$$

Therefore, (40) reduces to

$$\lambda \geq \sigma\sqrt{\frac{6\zeta_\Sigma t_0 \log d}{N}},$$

which is satisfied by the choice of λ as in (48), with c_8 being any constant such that $c_8 \geq \sqrt{6}$.

Consider now the condition on γ as in (41). Since we are interested in the high-dimensional regime where $N \ll d$, we assume that $d + \log m \geq n$. Using this, we can lower bound the RHS of (41) and readily obtain the more stringent condition on γ as in (49), with $c_{14} = 1152$.

• **Step 2: Condition on R in (50) implies (42).** Using again (47), we can upper bound the lower interval of R in (42) as

$$\frac{56\lambda s}{\lambda_{\min}(\Sigma) - 64 s c_1 \zeta_\Sigma \log d / N} \overset{(47)}{\leq} \frac{56\lambda s}{\lambda_{\min}(\Sigma) - \lambda_{\min}(\Sigma)/57} \overset{(48)}{=} \frac{57 c_8 s}{\lambda_{\min}(\Sigma)} \sigma \sqrt{\frac{6 t_0 \zeta_\Sigma \log d}{N}}.$$

Using (48), the upper interval in the same condition reads

$$\frac{\lambda N}{64 c_1 \zeta_\Sigma \log d} = \frac{c_8}{64 c_1}\sigma\sqrt{\frac{6 t_0 N}{\zeta_\Sigma \log d}}.$$

Therefore, (50) is sufficient for (42) to hold, with

$$c_{15} = 57\sqrt{6}c_8, \quad \text{and} \quad c_{16} = \frac{\sqrt{6}c_8}{64c_1}.$$

It remains to show that the range of value of R in (50) is nonempty. Using again (47) yields

$$N \geq \frac{c_{12} s \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)} \quad \Rightarrow \quad \frac{c_8}{64c_1}\sigma\sqrt{\frac{6t_0 N}{\zeta_\Sigma \log d}} \geq \frac{57\sqrt{6}c_8}{\lambda_{\min}(\Sigma)} s\sigma\sqrt{\frac{t_0\zeta_\Sigma \log d}{N}},$$

and

$$N \geq \frac{c_{13} s^2 \zeta_\Sigma \log d}{\sigma^2} \quad \Rightarrow N \geq \frac{c_{13} s^2 \zeta_\Sigma \log d}{c_8^2\sigma^2} \quad \Rightarrow \frac{\sqrt{6}c_8}{64c_1}\sigma\sqrt{\frac{t_0 N}{\zeta_\Sigma \log d}} \geq 2s,$$

which leads to the desired conclusion.

• **Step 3:** (45) **reduces to** (53)**, under** (48)**.** The statement follows from a direct substitution in (45) of the expression of λ as in (48). In fact, we have

$$\frac{\lambda^2 s}{1976\lambda_{\min}(\Sigma)^2} = \frac{c_8^2\sigma^2\zeta_\Sigma t_0 s \log d}{1976 N\lambda_{\min}(\Sigma)^2} = \frac{c_{21}\sigma^2\zeta_\Sigma t_0}{\lambda_{\min}(\Sigma)^2}\frac{s\log d}{N},$$

where $c_{21} = c_8^2/1976$, and

$$\frac{R\lambda}{4} = \frac{R\sigma c_8}{4}\sqrt{\frac{\zeta_\Sigma t_0 \log d}{N}}, \quad \eta_G^0 \geq \frac{11339 s c_1 \zeta_\Sigma \log d}{N\lambda_{\min}(\Sigma)^2}\lambda^2 s = c_{22}\sigma^2 t_0\left(\frac{s\zeta_\Sigma \log d}{N\lambda_{\min}(\Sigma)}\right)^2,$$

where $c_{22} = 11339 c_1 c_8^2$. Notice that the (random) interval (53) is non-empty, with probability at least (23). This follows readily from the fact that (45) is nonempty with the same probability, for all R satisfies (42) (Theorem 12), and that (50) is sufficient for (42) to hold (**Step 2** above).

• **Step 4:** (52) **holds with high probability, for all** t **satisfying** (54)**.** Eq. (52) follows readily from (44) by substitution of the values of λ and γ as in (48) and (49), respectively; and defining the following constants

$$c_{17} = 9, \quad c_{18} = 72c_1c_{17}, \quad c_{19} = c_1c_8^2c_{17}/988, \quad \text{and} \quad c_{20} = 8c_1c_{17}/c_8^2.$$

We conclude the proof showing that (54) is a stronger condition than (46). Using (48) and (49), the RHS of (46) reads

$$\left\lceil \log_2\log_2\left(\frac{c_8\sigma R}{\alpha^2}\sqrt{\frac{\zeta_\Sigma t_0\log d}{N}}\right)\right\rceil\left(1 + \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{456c_4[1+(d+\log m)/n]\log 2}{55}\right)$$
$$+\left\{\left\lceil \log_2\log_2\left(\frac{c_8\sigma R}{\alpha^2}\sqrt{\frac{\zeta_\Sigma t_0\log d}{N}}\right)\right\rceil\log 2 + \log\left(\frac{\eta_G^0}{\alpha^2}\right)\right\}$$
$$\times\frac{456(1+\rho)}{55\lambda_{\min}(\Sigma)}\cdot\frac{2c_4\lambda_{\max}(\Sigma)\left(1+\frac{d+\log m}{n}\right) + 128\lambda_{\min} dm\cdot\left[\left(2\log m + 1\right)/c_{24} + 8\right]\max\left\{\frac{2\log md}{nc_{24}}, 1\right\}}{1-\rho}$$

$$
\begin{aligned}
&+ \frac{456c_4[1+(d+\log m)/n]}{55}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\log\left(\frac{\eta_G^0}{\alpha^2}\right)\\
\overset{\rho\le 1}{\le} & \left\lceil \log_2\log_2\left(\frac{c_8\sigma R}{\alpha^2}\sqrt{\frac{\zeta_\Sigma t_0\log d}{N}}\right)\right\rceil\left(1+\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{456c_4[1+(d+\log m)/n]\log 2}{55}\right)\\
&+\left\{\left\lceil \log_2\log_2\left(\frac{c_8\sigma R}{\alpha^2}\sqrt{\frac{\zeta_\Sigma t_0\log d}{N}}\right)\right\rceil\log 2+\log\left(\frac{\eta_G^0}{\alpha^2}\right)\right\}\\
&\times\frac{912}{55}\cdot\frac{1}{1-\rho}\cdot\left(2c_4\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\left(1+\frac{d+\log m}{n}\right)+128dm\cdot\left[\left(2\log m+1\right)/c_{24}+8\right]\max\left\{\frac{2\log md}{nc_{24}},1\right\}\right)\\
&+ \frac{456c_4[1+(d+\log m)/n]}{55}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\log\left(\frac{\eta_G^0}{\alpha^2}\right)\\
\overset{(84)}{\le} & \left\lceil \log_2\log_2\left(\frac{c_8\sigma R}{\alpha^2}\sqrt{\frac{\zeta_\Sigma t_0\log d}{N}}\right)\right\rceil\underbrace{\left(c_4\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{456[1+(d+\log m)/n]\log 2+55}{55}\right)}_{\triangleq\texttt{term I}}\\
&+\left\{\left\lceil \log_2\log_2\left(\frac{c_8\sigma R}{\alpha^2}\sqrt{\frac{\zeta_\Sigma t_0\log d}{N}}\right)\right\rceil\log 2+\log\left(\frac{\eta_G^0}{\alpha^2}\right)\right\}\\
&\times\frac{912}{55}\cdot\frac{1}{1-\rho}\cdot\left[4c_4\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{d+\log m}{n}+128dm\cdot\left(\frac{2\log m+1}{c_{24}}+8\right)\max\left\{\frac{2\log md}{nc_{24}},1\right\}\right]\\
&+\underbrace{\frac{456c_4[1+(d+\log m)/n]}{55}\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\log\left(\frac{\eta_G^0}{\alpha^2}\right)}_{\texttt{term II}}. \qquad (136)
\end{aligned}
$$

Using $(d+\log m)/n\ge 1$ and $\rho\in(0,1)$, we can bound `term I` and `term II` as

$$
\begin{aligned}
\texttt{term I} &\le c_4\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{1}{1-\rho}\frac{(921\log 2+55)}{55}\frac{d+\log m}{n}\\
&\le(\log 2)\,c_4\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{1}{1-\rho}\frac{1001}{55}\frac{d+\log m}{n},
\end{aligned}
$$

and

$$
\texttt{term II}\le c_4\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{1}{1-\rho}\frac{912}{55}\frac{d+\log m}{n}\log\left(\frac{\eta_G^0}{\alpha^2}\right).
$$

Using the above bounds along with $(d+\log m)/n\le dm$, we can further bound (136) as

$$
\begin{aligned}
&\left\lceil \log_2\log_2\left(\frac{c_8\sigma R}{\alpha^2}\sqrt{\frac{\zeta_\Sigma t_0\log d}{N}}\right)\right\rceil\log 2\left(c_4\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{1}{1-\rho}\frac{1001}{55}\frac{d+\log m}{n}\right)\\
&+\left\{\left\lceil \log_2\log_2\left(\frac{c_8\sigma R}{\alpha^2}\sqrt{\frac{\zeta_\Sigma t_0\log d}{N}}\right)\right\rceil\log 2+\log\left(\frac{\eta_G^0}{\alpha^2}\right)\right\}\\
&\times\frac{912}{55}\cdot\frac{1}{1-\rho}\cdot\left[4c_4\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}\frac{d+\log m}{n}+128dm\cdot\left(\frac{2\log m+1}{c_{24}}+8\right)\max\left\{\frac{2\log md}{nc_{24}},1\right\}\right]
\end{aligned}
$$

$$+ c_4 \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \frac{1}{1-\rho} \frac{912}{55} \frac{d + \log m}{n} \log\left(\frac{\eta_G^0}{\alpha^2}\right)$$
$$\leq \left\{ \left\lceil \log_2 \log_2 \left(\frac{R\sigma c_8}{\alpha^2} \sqrt{\frac{\zeta_\Sigma t_0 \log d}{N}} \right) \right\rceil \log 2 + \log\left(\frac{\eta_G^0}{\alpha^2}\right) \right\}$$
$$\cdot \frac{c_{23}}{1-\rho} \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} dm \cdot \left(\frac{2\log m + 1}{c_{24}} + 8 \right) \max\left\{ \frac{2\log md}{nc_{24}}, 1 \right\} \tag{137}$$

where $c_{23} = 22222\, c_4$. This proves (54), and completes the proof of the corollary. □

Appendix J. Proof of Theorem 14

At high level the proof is organized in the following two steps. **(Step 1)** Under the following event, a tolerance $\eta > 0$ and an iteration number T are given such that

$$G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}}) \leq \eta, \quad \forall t \geq T, \tag{138}$$

we establish a sufficient decrease of the optimization error $G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}})$ in the form

$$G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}}) \leq \kappa^{t-T}(G(\boldsymbol{\theta}^T) - G(\hat{\boldsymbol{\theta}})) + \text{tolerance}, \quad \forall t \geq T, \tag{139}$$

for suitable $\kappa \in (0, 1)$ and tolerance > 0–this is proved in Lemma 17 (cf. Appendix J.1). Then, (**Step 2**) we divide the iterations $t = 0, 1, 2, \ldots,$ into a series of disjoint epochs $[T_k, T_{k+1})$, with $0 = T_0 \leq T_1 \leq \cdots$, each one with associated η_k, with $\eta_0 \geq \eta_1 \geq \cdots$. The tuples $\{(\eta_k, T_k)\}$ are constructed so that $G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}}) \leq \eta_k$, for all $t \geq T_k$. This permits to apply recursively (139) with smaller and smaller values of η_k, till the error $G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}})$ is driven below a desired threshold. This second step, formalized in Proposition 18 (cf. Appendix J.2), leverages (Agarwal et al., 2012, Th. 2).

J.1 Step 1: Sufficient decrease of the optimization error under (138)

The error decrease in the form (139) is formally stated in Lemma 17 below. It requires two intermediate technical results, namely: (i) Lemma 15, which restricts the (average of the) optimization error $\boldsymbol{\Delta}^t = \boldsymbol{\theta}^t - \hat{\boldsymbol{\theta}}$ to a set of "almost" sparse directions; and (ii) Lemma 16, which establishes a curvature property of G along such trajectories.

Lemma 15 (On the sparsity of Δ_{av}^t) *Consider Problem (4) under Assumption 2. Further assume that (i) the design matrix $\mathbf{X}$ satisfies the RSC condition (8) with $\delta = \mu/2 - 16\, s\tau > 0$; (ii) λ satisfies (13); and (iii) γ satisfies (31). Let $\{\boldsymbol{\theta}^t\}$ be the sequence generated by Algorithm* (24) *with R chosen such that*

$$R \geq \max\left\{ \frac{\lambda s}{\delta(1-r)} \left(13 + \frac{1}{32}\sqrt{\frac{2\tau s}{\delta}} \right), \frac{1}{r}\|\theta^*\|_1 \right\}, \tag{140}$$

for some $r \in (0, 1)$. Under condition (138) with parameters (T, η), the following holds: for any $t \geq T$,

$$\|(\Delta_{\text{av}}^t)_{\mathcal{S}^c}\|_1 \leq 3\|(\Delta_{\text{av}}^t)_{\mathcal{S}}\|_1 + 6\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1 + 2h_{\max} + \min\left\{ \frac{2\eta}{\lambda}, 2R \right\}, \tag{141}$$

where [cf. (74)]

$$h_{\max} = \frac{d\gamma}{\lambda(1-\rho)} \left(\frac{\max_{i\in[m]} \|w_i^\top X_i\|_\infty}{n} + \lambda \right)^2. \tag{142}$$

Proof See Appendix K.1. ■

Invoking the RSC condition (8), the next lemma links the objective- and the iterate-errors along (141).

Lemma 16 (Curvature along (141)**)** *Instate the assumptions of Lemma 15. Under condition (138) with parameters (T, η), the following holds: for any $t \geq T$,*

$$\left\{ \begin{array}{l} \left(\frac{\mu}{8} - 8\tau s\right) \|\Delta_{\text{av}}^t\|^2 \leq G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}}) + f(\|\boldsymbol{\Delta}_\perp^t\|) + \frac{\tau}{4}(v^2 + 8h_{\max}^2), \\ \left(\frac{\mu}{8} - 8\tau s\right) \|\Delta_{\text{av}}^t\|^2 \leq \mathcal{T}_L(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}^t) + f(\|\boldsymbol{\Delta}_\perp^t\|) + \frac{\tau}{4}(v^2 + 8h_{\max}^2), \end{array} \right. \tag{143}$$

where $\mathcal{T}_L(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}^t)$ is the first order Taylor error of L at $\boldsymbol{\theta}^t$ along the direction $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^t$ [cf. (14)],

$$f(\|\boldsymbol{\Delta}_\perp^t\|) \triangleq \left(\frac{L_{\max}}{2m} - \frac{1-\rho}{2m\gamma} \right) \|\boldsymbol{\Delta}_\perp^t\|^2, \tag{144}$$

and

$$v^2 \triangleq 144 s \|\hat{\nu}_{\text{av}}\|^2 + 4 \min\left\{ \frac{2\eta}{\lambda}, 2R \right\}^2. \tag{145}$$

Proof See Appendix K.2. ■

Using Lemma 16, we are now ready to formally prove (139).

Lemma 17 (Descent of the objective function) *Instate the assumptions of Lemma 15, under the stronger condition $\frac{\mu}{8} - 8\tau s > 0$ and the additional assumption that β is chosen so that*

$$\beta \leq \frac{m\gamma}{\gamma L_{\max} + 1 - \lambda_{\min}(W)}. \tag{146}$$

Under (138) with parameters (T, η), the following holds:

$$G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}}) \leq \kappa^{t-T}(G(\boldsymbol{\theta}^T) - G(\hat{\boldsymbol{\theta}})) + \tau(36 s \|\hat{\nu}_{\text{av}}\|^2 + 2h_{\max}^2 + \epsilon^2), \quad \forall t \geq T, \tag{147}$$

where

$$\kappa \triangleq \left(1 - \frac{\beta}{m} \left(\frac{\mu}{8} - 8\tau s \right) \right) \in \left(0, \frac{1}{2} \right) \quad \textit{and} \quad \epsilon = \min\left\{ \frac{2\eta}{\lambda}, 2R \right\}. \tag{148}$$

Proof See Appendix K.3. ■

Note the structure of the tolerance term in (147): $s\|\hat{\nu}_{\text{av}}\|^2$ is of the order of the statistical error; $h_{\max}^2$ is due to the lack of consensus on the agents trajectories θ_i^t's, it can be controlled by carefully choosing γ; and ϵ^2 is a function of the threshold η. In Step 2 below we show that, since $\kappa < 1$, one can eventually driven the error ϵ^2 below the threshold $\mathcal{O}(s\|\hat{\nu}_{\text{av}}\|^2 + h_{\max}^2)$.

J.2 Step 2: Recursive application of Lemma 17

As anticipated, the key idea is to divide the iterations $t = 0, 1, 2, \ldots,$ into a series of disjoint epochs $[T_k, T_{k+1})$, with $T_k \leq T_{k+1}$, each one with associated η_k, such that (i) $G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}}) \leq \eta_k$, for all $t \geq T_k$; and (ii) $\eta_0 \geq \eta_1 \geq \cdots$. This permits to apply recursively Lemma 17 with smaller and smaller values of η_k, till the error $G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}})$ is driven below the threshold $4\tau(36s\|\hat{\nu}_{\text{av}}\|^2 + 2h_{\max}^2)$. This construction follows the same argument as in the proof of (Agarwal et al., 2012, Th. 2) with minor adjustments (Lemma 4 therein is replaced with our Lemma 17) and thus is omitted.

Proposition 18 ((Agarwal et al., 2012, Th. 2)) *Instate the setting of Lemma 17. Further assume,*

$$R \leq \frac{\lambda}{32\tau}. \tag{149}$$

Then, there holds

$$G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}}) \leq \alpha^2,$$

for any tolerance α^2 such that

$$\min\left\{\frac{R\lambda}{4}, \eta_G^0\right\} \geq \alpha^2 \geq 4\tau(36s\|\hat{\nu}_{\text{av}}\|^2 + 2h_{\max}^2), \tag{150}$$

and for all

$$t \geq \left\lceil \log_2 \log_2\left(\frac{R\lambda}{\alpha^2}\right)\right\rceil \left(1 + \frac{\log 2}{\log 1/\kappa}\right) + \frac{\log(\eta_G^0/\alpha^2)}{\log 1/\kappa}, \tag{151}$$

where $\eta_G^0 = G(\boldsymbol{\theta}^0) - G(\hat{\boldsymbol{\theta}})$.

Equipped with Proposition 18, we can now complete the proof of Theorem 14. It remains to show the following facts:

- **Fact 1:** The lower bound condition on R as in (33) is more stringent than that in (140), under a proper choice of $r \in (0, 1)$; and the interval in (33) is nonempty;
- **Fact 2:** The range of α in (117) is contained in that of (150); and the interval (117) is nonempty;
- **Fact 3:** (118) is sufficient for (151).

We prove these facts next.

- **Fact 1:** Choosing $r = 1/2$, the lower bound condition on R in (140) reads

$$R \geq \max\left\{\frac{2\lambda s}{\mu/2 - 16s\tau}\left(13 + \frac{1}{32}\sqrt{\frac{2\tau s}{\mu/2 - 16s\tau}}\right), 2\|\theta^*\|_1\right\}, \tag{152}$$

Recalling $\mu \geq c_{10}s\tau = 1824s\tau$, the following holds for the lower bound in (152):

$$\frac{2\lambda s}{\mu/2 - 16s\tau}\left(13 + \frac{1}{32}\sqrt{\frac{2\tau s}{\mu/2 - 16s\tau}}\right) \leq \frac{56\lambda s}{\mu - 32s\tau}.$$

Therefore,

$$\max\left\{\frac{2\lambda s}{\mu/2 - 16s\tau}\left(13 + \frac{1}{32}\sqrt{\frac{2\tau s}{\mu/2 - 16s\tau}}\right), 2\|\theta^*\|_1\right\} \leq \max\left\{\frac{56\lambda s}{\mu - 32s\tau}, 2\|\theta^*\|_1\right\},$$

which proves the desired implication.

Finally, notice that the interval in (33) is non-empty. This is a consequence of (i) the fact

$$\frac{56\lambda s}{\mu - 32s\tau} \le \frac{\lambda}{32\tau},$$

due to $\mu \ge c_{10}s\tau = 1824s\tau$; and (ii) the condition $\lambda \ge 64\tau\|\theta^*\|_1$, due to (30).

- **Fact 2:** Using the condition on γ as in (31), we have

$$\begin{aligned}
4\tau(36s\|\hat{\nu}_{\text{av}}\|^2 + 2h_{\max}^2) \le\; & 4\tau\left(36s\frac{\sum_{i=1}^m\|\hat{\nu}_i\|^2}{m} + 2h_{\max}^2\right)\\
\overset{(111)}{\le}\; & 4\tau\left(36s\frac{\sum_{i=1}^m\|\hat{\nu}_i\|^2}{m} + \frac{2\lambda^2s^2}{128^2\,(\mu/2-16\,s\tau)^2}\right)\\
\overset{\mu\ge c_{10}s\tau}{\le}\; & 4\tau\left(36s\frac{\sum_{i=1}^m\|\hat{\nu}_i\|^2}{m} + \frac{2\lambda^2s^2(114)^2}{128^2\,(56)^2\mu^2}\right)\\
\le\; & 4s\tau\left(\frac{36}{m}\sum_{i=1}^m\|\hat{\nu}_i\|^2 + \frac{\lambda^2 s}{1976\mu^2}\right).
\end{aligned}$$

Therefore, the range of α in (117) is included in that of (150).

It remains to show that the range of α^2 in (117) is nonempty, which is a consequence of the following chain of inequalities.

$$\begin{aligned}
& 4\tau\left(36s\frac{\|\hat{\boldsymbol{\nu}}\|^2}{m} + \frac{\lambda^2s^2}{1976\mu^2}\right)\\
\overset{\substack{\text{Th.6}\\ \xi=\tau\ (\text{Lm. 4})}}{\le}\; & 144\tau s\left(\frac{9\lambda^2 s}{(\mu/2-16s\tau)^2} + \frac{2\tau}{\mu/2-16s\tau}\underbrace{\frac{d^2\gamma^2(\max_{1\le i\le m}\|w_i^\top X_i\|_\infty + \lambda n)^4}{\lambda^2n^4(1-\rho)^2}}_{=\texttt{Term I}^2\text{ [see (110)]}}\right.\\
& \left.+\frac{4}{\mu/2-16s\tau}\underbrace{\frac{d\gamma(\max_{1\le i\le m}\|w_i^\top X_i\|_\infty + \lambda n)^2}{n^2[2(1-\rho) - 4L_{\max}\gamma - (\mu/2-16s\tau)\gamma]}}_{=\texttt{Term II}^2\text{ [see (110)]}} + \frac{\lambda^2 s}{36\cdot 1976(\mu-32s\tau)^2}\right)\\
\overset{(111)}{\le}\; & 144\tau s\left(\frac{9\lambda^2 s}{(\mu/2-16s\tau)^2} + \frac{\tau s}{\mu/2-16s\tau}\,\frac{\lambda^2 s}{8192(\mu/2-16s\tau)^2} + \frac{\lambda^2 s}{64(\mu/2-16s\tau)^2}\right.\\
& \left.+\frac{\lambda^2 s}{36\cdot 7904(\mu/2-16s\tau)^2}\right)\\
\overset{(a)}{\le}\; & 144\tau s\left(\frac{9\lambda^2 s}{(\mu/2-16s\tau)^2} + \frac{1}{896}\,\frac{\lambda^2 s}{8192(\mu/2-16s\tau)^2} + \frac{\lambda^2 s}{64(\mu/2-16s\tau)^2}\right.\\
& \left.+\frac{\lambda^2 s}{36\cdot 7904(\mu/2-16s\tau)^2}\right)\\
<\; & \frac{144\tau s}{(\mu/2-16s\tau)^2}\cdot 10\lambda\cdot\lambda\,s
\end{aligned}$$

$$
\begin{aligned}
&\overset{(33)}{\leq} \frac{1440\tau s\lambda}{(\mu/2-16s\tau)^2}\cdot\frac{\mu-32s\tau}{56}R\\
&= \frac{1440\tau s}{28(\mu/2-16s\tau)}\lambda R\\
&\overset{(b)}{<} \frac{\lambda R}{17} < \frac{\lambda R}{4},
\end{aligned}
\tag{153}
$$

where (a) and (b) follow from $\mu/2-16s\tau \geq 896s\tau$, due to $\mu \geq c_{10}s\tau$, with $c_{10}=1824$. This together with (34) shows that the range of α^2 in (117) is non-empty.

• **Fact 3:** We obtain (118) from (151) by upper bounding the right hand side of (151). To this end, we first lower bound $\log(1/\kappa)$ as:

$$
\log\left(\frac{1}{\kappa}\right) \overset{(32),(148)}{=} \log\left(\frac{1}{1-\frac{\gamma(\mu/8-8\tau s)}{\gamma L_{\max}+1-\lambda_{\min}(W)}}\right) \overset{(10)}{\geq} \log\left(\frac{1}{1-\frac{\gamma(\mu/8-8\tau s)}{\gamma L_{\max}+1+\rho}}\right) \geq \frac{\gamma(\mu/8-8\tau s)}{\gamma L_{\max}+1+\rho}.
\tag{154}
$$

Using (154) in (151) and using $\eta_G^0 \geq \alpha^2$ [due to (150)], we can upper bound the right hand side as

$$
\left\lceil \log_2\log_2\left(\frac{R\lambda}{\alpha^2}\right)\right\rceil\left(1+\frac{(\gamma L_{\max}+1+\rho)\log 2}{\gamma(\mu/8-8\tau s)}\right)+\frac{(\gamma L_{\max}+1+\rho)}{\gamma(\mu/8-8\tau s)}\log\left(\frac{\eta_G^0}{\alpha^2}\right),
$$

which proves (118).

□

Appendix K. Proofs of auxiliary Lemmata in Sec. J

K.1 Proof of Lemma 15

Recalling the definitions of $\boldsymbol{\Delta}^t$, $\boldsymbol{\nu}^t$, and $\widehat{\boldsymbol{\nu}}$ as given in (62), (61) and (60), respectively, we have $\boldsymbol{\Delta}^t = \boldsymbol{\theta}^t - \hat{\boldsymbol{\theta}} = \boldsymbol{\nu}^t - \hat{\boldsymbol{\nu}}$. Therefore, $\Delta_{\text{av}}^t = \nu_{\text{av}}^t - \hat{\nu}_{\text{av}}$. We can then bound the desired quantity $\|(\Delta_{\text{av}}^t)_{\mathcal{S}^c}\|_1$ as

$$
\|(\Delta_{\text{av}}^t)_{\mathcal{S}^c}\|_1 \leq \|(\nu_{\text{av}}^t)_{\mathcal{S}^c}\|_1 + \|(\hat{\nu}_{\text{av}})_{\mathcal{S}^c}\|_1.
\tag{155}
$$

We prove below the following upper bounds for $\|(\nu_{\text{av}}^t)_{\mathcal{S}^c}\|_1$ and $\|(\hat{\nu}_{\text{av}})_{\mathcal{S}^c}\|_1$

$$
\begin{cases}
\|(\nu_{\text{av}}^t)_{\mathcal{S}^c}\|_1 \leq 3\|(\nu_{\text{av}}^t)_{\mathcal{S}}\|_1 + h(\gamma,\|\boldsymbol{\nu}_\perp^t\|) + \min\left\{\frac{2\eta}{\lambda}, 2R\right\},\\
\|(\hat{\nu}_{\text{av}})_{\mathcal{S}^c}\|_1 \leq 3\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1 + h(\gamma,\|\hat{\boldsymbol{\nu}}_\perp\|).
\end{cases}
\tag{156}
$$

Using (156) in (155) and the triangle inequality yields the desired result

$$
\begin{aligned}
\|(\Delta_{\text{av}}^t)_{\mathcal{S}^c}\|_1 &\leq 3\|(\nu_{\text{av}}^t)_{\mathcal{S}}\|_1 + h(\gamma,\|\boldsymbol{\nu}_\perp^t\|) + \min\left\{\frac{2\eta}{\lambda}, 2R\right\} + 3\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1 + h(\gamma,\|\hat{\boldsymbol{\nu}}_\perp\|)\\
&\leq 3(\|(\Delta_{\text{av}}^t)_{\mathcal{S}}\|_1 + 2\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1) + h(\gamma,\|\boldsymbol{\nu}_\perp^t\|) + h(\gamma,\|\hat{\boldsymbol{\nu}}_\perp\|) + \min\left\{\frac{2\eta}{\lambda}, 2R\right\}.
\end{aligned}
$$

$$\overset{(74)}{\leq} 3\|(\Delta_{\rm av}^t)_{\mathcal{S}}\|_1 + 6\|(\hat{\nu}_{\rm av})_{\mathcal{S}}\|_1 + 2h_{\max} + \min\left\{\frac{2\eta}{\lambda}, 2R\right\}.$$

We prove next (156). From the optimality of $\hat{\boldsymbol{\theta}}$, we deduce

$$G(\hat{\boldsymbol{\theta}}) - G(1_m \otimes \theta^*) \leq 0, \tag{157}$$

which together with (138) implies

$$G(\boldsymbol{\theta}^t) - G(1_m \otimes \theta^*) \leq \eta, \quad \forall t \geq T. \tag{158}$$

Hence, for any $t \geq T$, there holds

$$\begin{aligned}
&\frac{1}{2N}\sum_{i=1}^m \|X_i\theta_i^t - y_i\|^2 + \frac{1}{2m\gamma}\|\boldsymbol{\theta}^t\|_V^2 + \frac{\lambda}{m}\|\boldsymbol{\theta}^t\|_1 \\
\leq &\frac{1}{2N}\|\mathbf{X}\theta^* - y\|^2 + \frac{1}{2m\gamma}\|1_m \otimes \theta^*\|_V^2 + \frac{\lambda}{m}\|1_m \otimes \theta^*\|_1 + \eta, \\
\Leftrightarrow & \\
&\frac{1}{2N}\sum_{i=1}^m \|X_i\theta_i^t - y_i\|^2 + \frac{1}{2m\gamma}\|1_m \otimes \theta^* + \boldsymbol{\nu}^t\|_V^2 + \frac{\lambda}{m}\|1_m \otimes \theta^* + \boldsymbol{\nu}^t\|_1 \\
\leq &\frac{1}{2N}\|\mathbf{X}\theta^* - y\|^2 + \frac{1}{2m\gamma}\|1_m \otimes \theta^*\|_V^2 + \frac{\lambda}{m}\|1_m \otimes \theta^*\|_1 + \eta.
\end{aligned} \tag{159}$$

Subtracting $\sum_{i=1}^m \langle \frac{1}{N} X_i^\top (X_i\theta^* - y_i), \nu_i^t\rangle$ from both sides and rearranging terms, we obtain

$$\begin{aligned}
&\underbrace{-\sum_{i=1}^m \left\langle \frac{1}{N}X_i^\top(X_i\theta^* - y_i), \nu_i^t\right\rangle + \frac{1}{2m\gamma}\|1_m \otimes \theta^*\|_V^2 - \frac{1}{2m\gamma}\|1_m \otimes \theta^* + \boldsymbol{\nu}^t\|_V^2 + \eta}_{\texttt{Term I}} \\
\geq &\frac{1}{2N}\sum_{i=1}^m \|X_i(\theta^* + \nu_i^t) - y_i\|^2 - \frac{1}{2N}\|\mathbf{X}\theta^* - y\|^2 - \sum_{i=1}^m \left\langle \frac{1}{N}X_i^\top(X_i\theta^* - y_i), \nu_i^t\right\rangle \\
&+ \frac{\lambda}{m}(\|1_m \otimes \theta^* + \boldsymbol{\nu}^t\|_1 - \|1_m \otimes \theta^*\|_1) \\
\geq &\frac{\lambda}{m}\underbrace{(\|1_m \otimes \theta^* + \boldsymbol{\nu}^t\|_1 - \|1_m \otimes \theta^*\|_1)}_{\texttt{Term II}},
\end{aligned} \tag{160}$$

where the last inequality follows from convexity of $\sum_{i=1}^m \|X_i\theta_i - y_i\|^2/(2N)$.

We proceed upper (resp. lower) bounding `Term I` (resp. `Term II`). We have

$$\begin{aligned}
\texttt{Term I} &= -\sum_{i=1}^m \left\langle \frac{1}{N}X_i^\top(X_i\theta^* - y_i), \nu_{\rm av}^t + \nu_{\perp i}^t\right\rangle - \frac{1}{2m\gamma}\|\boldsymbol{\nu}_\perp^t\|_V^2 \\
&= \frac{1}{N}\mathbf{w}^\top \mathbf{X}\nu_{\rm av}^t + \frac{1}{N}\sum_{i=1}^m w_i^\top X_i \nu_{\perp i}^t - \frac{1}{2m\gamma}\|\boldsymbol{\nu}_\perp^t\|_V^2 \\
&\overset{(13)}{\leq} \frac{\lambda}{2}\|\nu_{\rm av}^t\|_1 + \frac{1}{N}\max_{i\in[m]}\|X_i^\top w_i\|_\infty \|\boldsymbol{\nu}_\perp^t\|_1 - \frac{1}{2m\gamma}\|\boldsymbol{\nu}_\perp^t\|_V^2.
\end{aligned} \tag{161}$$

To lower bound `Term II` we decompose $\theta^* + \nu_i^t$ as

$$\theta^* + \nu_i^t = \theta^*_{\mathcal{S}} + \theta^*_{\mathcal{S}^c} + (\nu_i^t)_{\mathcal{S}} + (\nu_i^t)_{\mathcal{S}^c}.$$

Then, for each i, we have

$$\begin{aligned}
\|\theta^* + \nu_i^t\|_1 - \|\theta^*\|_1 &\overset{\text{triangle inequality}}{\geq} \|\theta^*_{\mathcal{S}} + (\nu_i^t)_{\mathcal{S}^c}\|_1 - \|\theta^*_{\mathcal{S}^c} + (\nu_i^t)_{\mathcal{S}}\|_1 - \|\theta^*\|_1 \\
&\overset{\text{decomposibility}}{=} \|\theta^*_{\mathcal{S}}\|_1 + \|(\nu_i^t)_{\mathcal{S}^c}\|_1 - \|(\nu_i^t)_{\mathcal{S}}\|_1 - \|\theta^*\|_1 \\
&\overset{\text{triangle inequality}}{\geq} \|(\nu_{\text{av}}^t)_{\mathcal{S}^c}\|_1 - \|(\nu_{\perp i}^t)_{\mathcal{S}^c}\|_1 - \|(\nu_{\text{av}}^t)_{\mathcal{S}}\|_1 - \|(\nu_{\perp i}^t)_{\mathcal{S}}\|_1 \\
&\overset{\text{decomposibility}}{=} (\|(\nu_{\text{av}}^t)_{\mathcal{S}^c}\|_1 - \|(\nu_{\text{av}}^t)_{\mathcal{S}}\|_1) - \|\nu_{\perp i}^t\|_1.
\end{aligned} \tag{162}$$

Using (162) and (161) in (160), yields

$$\|(\nu_{\text{av}}^t)_{\mathcal{S}^c}\|_1 \leq 3\|(\nu_{\text{av}}^t)_{\mathcal{S}}\|_1 + h(\gamma, \|\boldsymbol{\nu}_\perp^t\|) + \frac{2\eta}{\lambda}. \tag{163}$$

On the other hand, since $\|\theta_i^t\|_1 \leq R$ [by construction, see (27)] and $\|\theta^*\|_1 < R$ [by condition (26)], we have

$$\|(\nu_{\text{av}}^t)_{\mathcal{S}^c}\|_1 \leq \|\nu_{\text{av}}^t\|_1 \leq \|\theta^*\|_1 + \|\theta_{\text{av}}^t\|_1 < R + R = 2R. \tag{164}$$

Using (164), we can then strengthen (163) as

$$\begin{aligned}
\|(\nu_{\text{av}}^t)_{\mathcal{S}^c}\|_1 &\leq \min\left\{3\|(\nu_{\text{av}}^t)_{\mathcal{S}}\|_1 + h(\gamma, \|\boldsymbol{\nu}_\perp^t\|) + \frac{2\eta}{\lambda}, 2R\right\} \\
&\leq 3\|(\nu_{\text{av}}^t)_{\mathcal{S}}\|_1 + h(\gamma, \|\boldsymbol{\nu}_\perp^t\|) + \min\left\{\frac{2\eta}{\lambda}, 2R\right\},
\end{aligned}$$

which proves the first inequality in (156).

The proof of the second inequality in (156) follows the same steps and use the fact that $\|\hat{\theta}_i\|_1 \leq R$, for all $i \in [m]$ (Lemma 8). $\square$

K.2 Proof of Lemma 16

To bound the (average component of the) optimization error in terms of the function optimality gap we leverage the curvature property of L [under the RSC condition (8)] along the trajectory of the algorithm. We explicitly use the fact that the trajectory lies in the set described by (141) [cf. Lemma 15].

By Taylor's formula of L at $\hat{\boldsymbol{\theta}}$ along $\boldsymbol{\Delta}^t$, we have

$$\mathcal{T}_L(\boldsymbol{\theta}^t; \hat{\boldsymbol{\theta}}) \triangleq L(\boldsymbol{\theta}^t) - L(\hat{\boldsymbol{\theta}}) - \langle \nabla L(\hat{\boldsymbol{\theta}}), \boldsymbol{\Delta}^t \rangle \geq \underbrace{\frac{1}{4}\frac{\|\mathbf{X}\Delta_{\text{av}}^t\|^2}{N}}_{\text{curvature along average}} - \underbrace{\left(\frac{L_{\max}}{2m} - \frac{1-\rho}{2m\gamma}\right)\|\boldsymbol{\Delta}_\perp^t\|^2}_{\text{nonconsensual component}}. \tag{165}$$

Recalling that $G(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \frac{\lambda}{m}\|\boldsymbol{\theta}\|_1$, by the optimality of $\hat{\boldsymbol{\theta}}$, it follows

$$\langle \boldsymbol{\Delta}^t, \nabla L(\hat{\boldsymbol{\theta}}) \rangle + \frac{\lambda}{m}\|\boldsymbol{\theta}^t\|_1 - \frac{\lambda}{m}\|\hat{\boldsymbol{\theta}}\|_1 \geq 0. \tag{166}$$

We can then write

$$
\begin{aligned}
G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}}) &\overset{(166)}{\geq} \mathcal{T}_L(\boldsymbol{\theta}^t; \hat{\boldsymbol{\theta}}) \\
&\overset{(165)}{\geq} \frac{1}{4}\frac{\|\mathbf{X}\Delta_{\text{av}}^t\|^2}{N} - \left(\frac{L_{\max}}{2m} - \frac{1-\rho}{2m\gamma}\right)\|\boldsymbol{\Delta}_{\perp}^t\|^2 \\
&\overset{\text{RSC (8)}}{\geq} \frac{1}{4}\left(\frac{\mu}{2}\|\Delta_{\text{av}}^t\|^2 - \frac{\tau}{2}\|\Delta_{\text{av}}^t\|_1^2\right) - \left(\frac{L_{\max}}{2m} - \frac{1-\rho}{2m\gamma}\right)\|\boldsymbol{\Delta}_{\perp}^t\|^2 \\
&\overset{\text{(a)}}{\geq} \frac{1}{4}\left(\frac{\mu}{2}\|\Delta_{\text{av}}^t\|^2 - \frac{\tau}{2}(64s\|\Delta_{\text{av}}^t\|^2 + 2v^2 + 16h_{\max}^2)\right) - \left(\frac{L_{\max}}{2m} - \frac{1-\rho}{2m\gamma}\right)\|\boldsymbol{\Delta}_{\perp}^t\|^2,
\end{aligned}
\tag{167}
$$

where in (a) we used (141) (Lemma 15) as

$$
\begin{aligned}
\|\Delta_{\text{av}}^t\|_1^2 &\overset{(141)}{\leq} \left(4\|(\Delta_{\text{av}}^t)_{\mathcal{S}}\|_1 + 6\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1 + 2h_{\max} + \min\left\{\frac{2\eta}{\lambda}, 2R\right\}\right)^2 \\
&\leq 4(4\|(\Delta_{\text{av}}^t)_{\mathcal{S}}\|_1)^2 + 4(6\|(\hat{\nu}_{\text{av}})_{\mathcal{S}}\|_1)^2 + 4(2h_{\max})^2 + 4\left(\min\left\{\frac{2\eta}{\lambda}, 2R\right\}\right)^2 \\
&\leq 64s\|\Delta_{\text{av}}^t\|^2 + 2v^2 + 16h_{\max}^2,
\end{aligned}
$$

where $v^2 = 144s\|\hat{\nu}_{\text{av}}\|^2 + 4\min\left\{\frac{2\eta}{\lambda}, 2R\right\}^2$.

Reorganizing the terms in (167) yields the first inequality in (143).

Similar arguments apply to derive the second inequality in (143) by noticing that, for quadratic L, we have

$$\mathcal{T}_L(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}^t) = \mathcal{T}_L(\boldsymbol{\theta}^t; \hat{\boldsymbol{\theta}}).$$

This concludes the proof. □

K.3 Proof of Lemma 17

The proof follows descent arguments (see, e.g., (Nesterov, 2007)), suitably coupled with the curvature property established in Lemma 16 to achieve contraction up to a controllable tolerance.

Recall the algorithmic update (24)

$$\boldsymbol{\theta}^{t+1} = \underset{\|\theta_i\|_1 \leq R,\ \forall i\in[m]}{\operatorname{argmin}} \left\{G_t(\boldsymbol{\theta}) \triangleq L(\boldsymbol{\theta}^t) + \langle \nabla L(\boldsymbol{\theta}^t), \boldsymbol{\theta} - \boldsymbol{\theta}^t\rangle + \frac{1}{2\beta}\|\boldsymbol{\theta} - \boldsymbol{\theta}^t\|^2 + \frac{\lambda}{m}\|\boldsymbol{\theta}\|_1\right\}.$$

By definition of $\boldsymbol{\theta}^{t+1}$, we have $G_t(\boldsymbol{\theta}^{t+1}) \leq G_t(\boldsymbol{\theta})$, for all feasible $\boldsymbol{\theta}$. Recall that, by Lemma 8, $\hat{\boldsymbol{\theta}}$ is feasible. Hence, $\boldsymbol{\theta}_\omega \triangleq \omega\hat{\boldsymbol{\theta}} + (1-\omega)\boldsymbol{\theta}^t$ is feasible as well, for any $\omega \in (0,1)$. Therefore,

$$
\begin{aligned}
&G_t(\boldsymbol{\theta}^{t+1}) \\
&\leq G_t(\boldsymbol{\theta}_\omega) \\
&= L(\boldsymbol{\theta}^t) + \omega\,\langle \nabla L(\boldsymbol{\theta}^t), \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^t\rangle + \frac{\omega^2}{2\beta}\|\boldsymbol{\Delta}^t\|^2 + \frac{\lambda}{m}\|\boldsymbol{\theta}_\omega\|_1
\end{aligned}
$$

$$= (1-\omega)L(\boldsymbol{\theta}^t) + \omega L(\hat{\boldsymbol{\theta}}) - \omega \underbrace{\left[L(\hat{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^t) - \langle \nabla L(\boldsymbol{\theta}^t), \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^t\rangle\right]}_{\mathcal{T}_L(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}^t)} + \frac{\omega^2}{2\beta}\|\boldsymbol{\Delta}^t\|^2 + \frac{\lambda}{m}\|\boldsymbol{\theta}_\omega\|_1$$

$$\overset{(143)}{\leq} (1-\omega)\left(L(\boldsymbol{\theta}^t) + \frac{\lambda}{m}\|\boldsymbol{\theta}^t\|_1\right) + \omega\left(L(\hat{\boldsymbol{\theta}}) + \frac{\lambda}{m}\|\hat{\boldsymbol{\theta}}\|_1\right)$$
$$+ \omega\left(f(\|\boldsymbol{\Delta}_\perp^t\|) + \frac{\tau}{4}(v^2 + 8h_{\max}^2) - \left(\frac{\mu}{8} - 8\tau s\right)\|\Delta_{\text{av}}^t\|^2\right) + \frac{\omega^2}{2\beta}\|\boldsymbol{\Delta}^t\|^2$$
$$= (1-\omega)G(\boldsymbol{\theta}^t) + \omega G(\hat{\boldsymbol{\theta}}) + \omega f(\|\boldsymbol{\Delta}_\perp^t\|) + \omega\frac{\tau}{4}(v^2 + 8h_{\max}^2) + \frac{\omega^2}{2\beta}\|\boldsymbol{\Delta}^t\|^2 - \omega\left(\frac{\mu}{8} - 8\tau s\right)\|\Delta_{\text{av}}^t\|^2. \tag{168}$$

We proceed to relate $G(\boldsymbol{\theta}^{t+1})$ with $G_t(\boldsymbol{\theta}^{t+1})$.

$$G(\boldsymbol{\theta}^{t+1})$$
$$= G_t(\boldsymbol{\theta}^{t+1}) - \frac{1}{2\beta}\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|^2 + \underbrace{L(\boldsymbol{\theta}^{t+1}) - L(\boldsymbol{\theta}^t) - \langle \nabla L(\boldsymbol{\theta}^t), \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\rangle}_{=\frac{1}{2N}\sum_{i=1}^m \|X_i(\theta_i^{t+1} - \theta_i^t)\|^2 + \frac{1}{2m\gamma}\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|_V^2}$$
$$\overset{(168)}{\leq} G(\boldsymbol{\theta}^t) - \omega(G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}})) + \omega f(\|\boldsymbol{\Delta}_\perp^t\|) + \omega\frac{\tau}{4}(v^2 + 8h_{\max}^2) + \frac{\omega^2}{2\beta}\|\boldsymbol{\Delta}^t\|^2 - \omega\left(\frac{\mu}{8} - 8\tau s\right)\|\Delta_{\text{av}}^t\|^2$$
$$+ \frac{1}{2N}\sum_{i=1}^m \|X_i(\theta_i^{t+1} - \theta_i^t)\|^2 + \frac{1}{2m\gamma}\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|_V^2 - \frac{1}{2\beta}\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|^2. \tag{169}$$

Subtracting $G(\hat{\boldsymbol{\theta}})$ from both sides of the above inequality and denoting the function gap as $\eta_G^t = G(\boldsymbol{\theta}^t) - G(\hat{\boldsymbol{\theta}})$, we have

$$\eta_G^{t+1} \leq (1-\omega)\,\eta_G^t + \omega f(\|\boldsymbol{\Delta}_\perp^t\|) + \omega\frac{\tau}{4}(v^2 + 8h_{\max}^2) + \frac{\omega^2}{2\beta}\|\boldsymbol{\Delta}^t\|^2 - \omega\left(\frac{\mu}{8} - 8\tau s\right)\|\Delta_{\text{av}}^t\|^2$$
$$+ \frac{1}{2N}\sum_{i=1}^m \|X_i(\theta_i^{t+1} - \theta_i^t)\|^2 + \frac{1}{2m\gamma}\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|_V^2 - \frac{1}{2\beta}\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|^2$$
$$\leq (1-\omega)\eta_G^t + \omega f(\|\boldsymbol{\Delta}_\perp^t\|) + \omega\frac{\tau}{4}(v^2 + 8h_{\max}^2) + \frac{\omega^2}{\beta}\|\boldsymbol{\Delta}_\perp^t\|^2$$
$$+ \underbrace{\omega\left(\frac{\omega\, m}{\beta} - \left(\frac{\mu}{8} - 8\tau s\right)\right)\|\Delta_{\text{av}}^t\|^2}_{\leq 0,\ \text{for } 0 \leq \omega \leq \frac{\beta}{m}\left(\frac{\mu}{8} - 8\tau s\right)} + \underbrace{\frac{1}{2}\left(\frac{L_{\max}}{m} + \frac{1 - \lambda_{\min}(W)}{m\gamma} - \frac{1}{\beta}\right)}_{\leq 0 \text{ [condition on } \beta \text{ in (146)]}}\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|^2$$
$$\overset{(a)}{\leq} \underbrace{(1-\omega)\eta_G^t}_{\texttt{Term I}} + \underbrace{\omega f(\|\boldsymbol{\Delta}_\perp^t\|) + \frac{\omega^2}{\beta}\|\boldsymbol{\Delta}_\perp^t\|^2}_{\texttt{Term II}} + \underbrace{\omega\frac{\tau}{4}(v^2 + 8h_{\max}^2)}_{\texttt{Term III}}, \tag{170}$$

where (a) holds under the condition $0 \leq \omega \leq \frac{\beta}{m}\left(\frac{\mu}{8} - 8\tau s\right)$. Note that $\mu/8 - 8\tau s > 0$ by assumption; hence the interval for ω is non-empty. Furthermore, $\omega \in (0, 1/2)$, due to the following

$$\frac{\beta}{m}\left(\frac{\mu}{8} - 8\tau s\right) \overset{(146)}{\leq} \frac{\gamma}{\gamma L_{\max} + 1 - \lambda_{\min}(W)} \cdot \left(\frac{\mu}{8} - 8\tau s\right) \overset{(b)}{<} \frac{1}{2}, \tag{171}$$

where in (b) we used the following lower bound for $(1 - \lambda_{\min})/\gamma$:

$$
\begin{aligned}
&\frac{1 - \lambda_{\min}(W)}{\gamma} \\
&\overset{(31)}{\geq} \frac{1 - \lambda_{\min}(W)}{1 - \rho} \left(2L_{\max} + \frac{\mu}{2} - 16s\tau + \frac{128d}{s} \left(\frac{\mu}{2} - 16s\tau \right) \left(\frac{\max_{i \in [m]} \| w_i^\top X_i \|_\infty}{\lambda n} + 2\sqrt{m} \right)^2 \right) \\
&\overset{(10)}{\geq} \left(2L_{\max} + \frac{\mu}{2} - 16s\tau + \frac{128d}{s} \left(\frac{\mu}{2} - 16s\tau \right) \left(\frac{\max_{i \in [m]} \| w_i^\top X_i \|_\infty}{\lambda n} + 2\sqrt{m} \right)^2 \right) \\
&\geq 2 \left(\frac{\mu}{8} - 8\tau s \right).
\end{aligned}
$$

In (170), `Term I` captures the geometric decrease of the objective error, for any $\omega < 1$; `Term II` is due to consensus errors and it is controllable by choosing a sufficiently small network regularizer γ; finally, `Term III` is due to the lack of strong convexity, determining a nonzero tolerance on the achievable objective error.

We choose ω to minimize the contraction factor in `Term I`, resulting in

$$
\omega = \frac{\beta}{m} \left(\frac{\mu}{8} - 8\tau s \right). \tag{172}
$$

Under this choice we can bound `Term I`–`Term III` as follows.

• `Term I`:

$$
\texttt{Term I} = \underbrace{\left(1 - \frac{\beta}{m} \left(\frac{\mu}{8} - 8\tau s \right) \right)}_{\kappa} \eta_G^t. \tag{173}
$$

Note that $\kappa \in (0, 1/2)$, due to (171).

• `Term II`: Using the upper bound of γ in (25), we can bound $f(\|\mathbf{\Delta}_\perp^t\|)$ [cf. (144)] as

$$
f(\|\mathbf{\Delta}_\perp^t\|) \leq - \left(\frac{L_{\max}}{2m} + \frac{\mu - 32\,s\tau}{4m} + \frac{32d(\mu - 32\,s\tau)}{sm} (\max_{i \in [m]} \| w_i^\top X_i \|_\infty / (\lambda n) + 2\sqrt{m})^2 \right) \|\mathbf{\Delta}_\perp^t\|^2. \tag{174}
$$

Therefore,

$$
\begin{aligned}
\texttt{Term II} &\overset{(172),(174)}{\leq} -\frac{\beta}{m} \left(\frac{\mu}{8} - 8\tau s \right) \frac{1}{m} \left(\frac{\mu}{4} - 8s\tau \right) \|\mathbf{\Delta}_\perp^t\|^2 + \frac{\beta}{m^2} \left(\frac{\mu}{8} - 8\tau s \right)^2 \|\mathbf{\Delta}_\perp^t\|^2 \\
&\leq -\frac{\beta}{m^2} \left(\frac{\mu}{8} - 8\tau s \right)^2 \|\mathbf{\Delta}_\perp^t\|^2 \leq 0.
\end{aligned} \tag{175}
$$

• `Term III`:

$$
\texttt{Term III} = \frac{\beta}{m} \left(\frac{\mu}{8} - 8\tau s \right) \tau \left(36s \|\hat{\nu}_{\text{av}}\|_2^2 + 2h_{\max}^2 + \min \left\{ \frac{2\eta}{\lambda}, 2R \right\}^2 \right). \tag{176}
$$

Using (173), (175), and (176) in (170), we finally obtain: for all $t \geq T$,

$$\begin{aligned}\eta_G^{t+1} &\leq \kappa\, \eta_G^t + \frac{\beta}{m}\left(\frac{\mu}{8} - 8\tau s\right)\tau\left(36s\|\hat{\nu}_{\mathrm{av}}\|^2 + 2h_{\max}^2 + \min\left\{\frac{2\eta}{\lambda}, 2R\right\}^2\right) \\ &\leq \kappa^{t-T}\eta_G^T + \frac{1}{1-\kappa}\cdot\frac{\beta}{m}\left(\frac{\mu}{8} - 8\tau s\right)\tau\left(36s\|\hat{\nu}_{\mathrm{av}}\|^2 + 2h_{\max}^2 + \min\left\{\frac{2\eta}{\lambda}, 2R\right\}^2\right) \qquad (177) \\ &= \kappa^{t-T}\eta_G^T + \tau\left(36s\|\hat{\nu}_{\mathrm{av}}\|^2 + 2h_{\max}^2 + \min\left\{\frac{2\eta}{\lambda}, 2R\right\}^2\right).\end{aligned}$$

This completes the proof. □

References

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, pages 2452–2482, April 2012.

Ghanbar Azarnia, Mohammad Ali Tinati, Abbas Ali Sharifi, and Hamid Shiri. Incremental and diffusion compressive sensing strategies over distributed networks. *Digital Signal Processing*, 101:102732, June 2020.

R. Bai and M. Ghosh. Normalbetaprime: Normal beta prime prior. 2019. URL `https://CRAN.R-project.org/package=NormalBetaPrime`. R package version 2.2.

Y. Bao and W. Xiong. One-round communication efficient distributed m-estimation. In *International Conference on Artificial Intelligence and Statistics*, 2021.

H. Battey, F. Jianqing, L. Han, L. Junwei, and Z. Ziwei. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3):1352 – 1382, June 2018.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, January 2009.

S. Becker, J. Bobin, and E. J. Candes. Nesta: a fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4:1–39, April 2011.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, February 2009.

J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizeations of lojasiewicz inequalities: subgradient flows, talweg, convexity. *The Transactions of the American Mathematical Society*, 322:3319–3363, June 2009.

K. Bredies and D. A. Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14:813–837, September 2008.

E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies. *IEEE Transactions on Information Theory*, 52(12):5406–5425, January 2006.

A. I. Chen and A. Ozdaglar. A fast distributed proximal-gradient method. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 601–608, April 2012.

J. Chen and A. H. Sayed. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305, August 2012.

X. Chen, W. Liu, and Y. Zhang. First-order newton-type estimator for distributed estimation and inference. *Journal of the American Statistical Association*, pages 1–17, 2021.

A. Daneshmand, G. Scutari, and V. Kungurtsev. Second-order guarantees of distributed gradient algorithms. *SIAM Journal on Optimization*, 30(4):3029–3068, January 2020.

S. Van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3(4):1360–1392, October 2009.

D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, November 1995.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279, July 2008.

Sophie M Fosson, Javier Matamoros, Carles Antón-Haro, and Enrico Magli. Distributed recovery of jointly sparse signals under communication constraints. *IEEE Transactions on Signal Processing*, 64(13):3470–3482, 2016.

E. T. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: methodology and convergence. *SIAM Journal on Optimization*, 19:1107–1130, October 2008.

T. Hastie, R. Tibshirani, and M. J. Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

P. S. Horn. On the stochastic ordering of absolute univariate Gaussian random variables. *The Annals of Statistics*, 16(3):1327–1329, September 1988.

D. Jakovetić. A unification and generalization of exact distributed first-order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5:31–46, September 2019.

D. Jakovetić, J. Xavier, and JMF. Moura. Cooperative convex optimization in networked systems: augmented lagrangian algorithms with directed gossip communication. *IEEE Transactions on Signal Processing*, 59(8):3889–3902, July 2011.

D. Jakovetić, JMF. Moura, and J. Xavier. Linear convergence rate of a class of distributed augmented lagrangian algorithms. *IEEE Transactions on Automatic Control*, 60(4):922–936, July 2013.

D. Jakovetić, J. Xavier, and JMF. Moura. Fast distributed gradient methods. *IEEE Transactions on Automatic Control*, 59:1131–1146, December 2014.

F. Jianqing, G. Yongyi, and W. Kaizheng. Communication-efficient accurate statistical estimation. *Journal of the American Statistical Association*, 0(0):1–11, August 2021.

M. I. Jordan, J. D. Lee, and Y. Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2018.

J. D. Lee, Y. Sun, Q. Liu, and J.E. Taylor. Communication-efficient sparse regression: a one-shot approach. *Journal of Machine Learning Research*, January 2015.

P. Di Lorenzo and G. Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2:1–1, February 2016.

Z. Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46:157–178, March 1993.

A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and Its Applications*, volume 143. Springer, second edition, 2011.

A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, January 2009.

A. Nedić, A. Ozdaglar, and P. A. Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55(4):922–938, April 2010.

A. Nedić, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27:2597–2633, July 2016.

A. Nedić, A. Olshevsky, and M. G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106:953–976, September 2018.

Y. Nesterov. Gradient methods for minimizing composite objective function. *Research Papers in Economics*, January 2007.

Y. Nesterov et al. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

A. Olshevsky. Linear time average consensus and distributed optimization on fixed graphs. *SIAM Journal on Control and Optimization*, 55(6):3990–4014, December 2017.

S. Pan and Y. Liu. Metric subregularity of subdifferential and KL property of exponent 1/2. *arxiv preprint, arXiv:1812.00558v3*, 322, 2018.

Stacy Patterson, Yonina C Eldar, and Idit Keidar. Distributed compressed sensing for static and time-varying networks. *IEEE Transactions on Signal Processing*, 62(19):4931–4946, 2014.

G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5:1245–1260, April 2017.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, August 2010.

G. Raskutti, M. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory*, 57:6976–6994, November 2011.

J. Rosenblatt and B. Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

N. Sahand, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, November 2012.

A. H. Sayed. Adaptation, learning, and optimization over networks. *Foundations and Trends in Machine Learning*, 7:311–801, January 2014.

O. Shamir, N. Srebro, and T. Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International Conference on Machine Learning*, pages 1000–1008. PMLR, 2014.

W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin. On the linear convergence of the admm in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62: 1750–1761, July 2014.

W. Shi, Q. Ling, G. Wu, and W. Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, November 2015a.

W. Shi, Q. Ling, G. Wu, and W. Yin. A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing*, 63(22):6013–6023, November 2015b.

Y Sun, A Daneshmand, and G Scutari. Distributed optimization based on gradient-tracking revisited: enhancing convergence rate via surrogation. *arXiv preprint, arXiv:1905.02637*, 2019.

Y Sun, Maros M., G Scutari, and Guang C. High-dimensional inference over networks: linearly convergence algorithms and statistical guarantees. *arXiv:2201.08507*, 2022.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, January 1996.

P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117:387–423, March 2009.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, 2012.

M. J. Wainwright. *High-Dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press, 2019.

J. Wang, M. Kolar, N. Srebro, and T. Zhang. Efficient distributed learning with sparsity. In *International Conference on Machine Learning*, pages 3636–3645. PMLR, May 2017.

S. Wang, F. Roosta, P. Xu, and W. W. Mahoney. Giant: Globally improved approximate newton method for distributed optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

B. Wen, X. Chen, and T. Pong. Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM Journal on Optimization*, 27:124–145, December 2017.

J. Xu, S. Zhu, Y. C. Soh, and L. Xie. Convergence of asynchronous distributed gradient methods over stochastic networks. *IEEE Transactions on Automatic Control*, 63(2): 434–448, July 2018.

K. Yuan, Q. Ling, and W. Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, January 2016.

K. Yuan, S. Alghunaim, B. Ying, and A. H. Sayed. On the influence of bias-correction on distributed stochastic optimization. *IEEE Transactions on Signal Processing*, 68:4352–4367, July 2020.

J. Zeng and W. Yin. On nonconvex decentralized gradient descent. *IEEE Transactions on Signal Processing*, 66(11):2834–2848, June 2018.

Z. Zhou and A. Man-Cho So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165:689–728, December 2017.