Distributed (ATC) Gradient Descent for High Dimension Sparse Regression

Yao Ji, *Student Member, IEEE,* Gesualdo Scutari, *Fellow, IEEE,* Ying Sun, *Member, IEEE,* and Harsha Honnappa

***Abstract*—We study linear regression from data distributed over a network of agents (with no master node) by means of LASSO estimation, in *high-dimension*, which allows the ambient dimension to grow faster than the sample size. While there is a vast literature of distributed algorithms applicable to the problem, statistical and computational guarantees of most of them remain unclear in high dimension. This paper provides a first statistical study of the Distributed Gradient Descent (DGD) in the Adapt-Then-Combine (ATC) form. Our theory shows that, under standard notions of restricted strong convexity and smoothness of the loss functions–which hold with high probability for standard data generation models–suitable conditions on the network connectivity and algorithm tuning, DGD-ATC converges globally at a *linear* rate to an estimate that is within the *centralized* statistical precision of the model. In the worst-case scenario, the total number of communications to statistical optimality grows logarithmically with the ambient dimension, which improves on the communication complexity of DGD in the Combine-Then-Adapt (CTA) form, scaling linearly with the dimension. This reveals that mixing gradient information among agents, as DGD-ATC does, is critical in high-dimensions to obtain favorable rate scalings.**

***Index Terms*—Distributed optimization, high-dimension statistics, linear convergence, sparse linear regression.**

I. Introduction

We study sparse linear regression over a network of m agents, modeled as an undirected graph. In particular, no centralized node is assumed in the network; agents can communicate only with their immediate neighbors–we refer to these networks as *mesh* networks. Each agent i locally owns n linear measurements of an s-sparse signal $\theta^* \in \mathbb{R}^d$ *common* to all local models:

$$y_i = X_i\theta^* + w_i, \quad i = 1, \dots, m, \tag{1}$$

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. The total sample size over the network is $N = m \cdot n$. We are interested in the *high-dimensional* setting: the ambient dimension d is larger (and grows faster) than the total sample size N and $s \ll d$ [44].

The LASSO estimator of θ^* based on all N samples reads

$$\widehat{\theta} \in \underset{\theta\in\mathbb{R}^d\,:\,\|\theta\|_1\le R}{\arg\min} F(\theta) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_i(\theta), \tag{2}$$

with

$$f_i(\theta) \triangleq \frac{1}{2n}\|y_i - X_i\theta\|^2,$$

where the sparsity information is encoded via the projection onto the ℓ_1 ball. Solutions methods for (2) have been extensively studied in the centralized setting (e.g., on master-worker architecture); see, e.g., [4], [6], [21], [22], [45]. Our focus here is on first-order methods; more specifically, the benchmark is the Projected Gradient Descent (PGD) [2], whose performance when applied to (2) can be summarized as follows: under (sub-)Gaussian random matrix designs (see Assumption 2 for other statistical models) and conditions for statistical consistency–$s\log d/N = o(1)$–the iterates $\{\theta^t\}$ generated by the PGD (starting from θ^0) satisfy with high-probability:

$$\|\theta^t - \widehat{\theta}\|^2 \le r^t\,\|\theta^0 - \widehat{\theta}\|^2 + o(\|\widehat{\theta} - \theta^\star\|^2), \tag{3}$$

with

$$r = \frac{1 - {\kappa_\Sigma}^{-1} + \mathcal{O}(s\log d/N)}{1 - \mathcal{O}(s\log d/N)},$$

where in the expression of the rate r we neglected universal constants and $\kappa_\Sigma \ge 1$ is the condition number of the covariance matrix of the covariates (see Assumption 2). Notice that the rate r is invariant to the ambient dimension d under high-dimension scaling $s, d/N \to \infty$ as long as $s\log d/N$ remains constant. In words: the optimization error $\|\theta^t - \widehat{\theta}\|^2$ decays *linearly* with rate r, up to a tolerance of a smaller order than $\|\widehat{\theta} - \theta^\star\|^2$. Therefore every limit point of $\{\theta^t\}$ is within the statistical error from $\theta^\star$. This is the best one can hope for, statistically (ignoring lower order terms) and computationally (within first-order, non accelerated methods).

The PGD is not implementable on mesh networks: agents cannot compute locally the full gradient ∇F, as they do not have access to the entire data set, and sharing data across the network is either infeasible (e.g., due to privacy issues) or highly inefficient (e.g., due to excessive communication overhead). A natural question is whether statistical/computational guarantees similar to those of PGD can be mimicked by some *distributed* algorithms. Of particular interest is the regime wherein the local sample size n is below information theoretical bounds while the total one N is sufficient for statistical consistency.

Decentralized versions of PGD have been extensively studied in the literature of distributed optimization (see Sec. I-B for a review of the relevant works); with no doubts, Distributed Gradient Descent (DGD) algorithms are among the most popular ones [11], [12], [29], [30], [33]. Roughly speaking they are of two types, based upon the information mixed

Ji, Scutari, and Honnappa are with the School of Industrial Engineering, Purdue University, West Lafayette, IN 47906, USA (e-mails: <jiyao, gscutari,honnappa>@purdue.edu).

Sun is with the School of Electrical Engineering and Computer Science, The Pennsylvania State University, State College, PA 16802, USA, USA (e-mail: ybs5190@psu.edu).

Manuscript received July 3, 2022; revised December 28, 2022.

locally by the agents, namely: the DGD in the *Combine-then-Adapt* (DGD-CTA) form [29], [30] and the DGD in the *Adapt-Then-Combine* update (DGD-ATC) [11], [12], [33]. DGD-CTA averages local parameters vectors whereas DGD-ATC averages *both* local parameter vectors and gradients. More specifically, when applied to the LASSO problem (2), DGD-CTA and DGD-ATC updates read for all $t = 1, 2, \dots,$

$$\texttt{DGD-CTA}: \theta_i^t = \prod_{\|\theta_i\|_1 \leq R} \left(\sum_{j=1}^m w_{ij} \theta_j^{t-1} - \gamma \nabla f_i(\theta_i^{t-1}) \right), \tag{4}$$

and

$$\texttt{DGD-ATC}: \theta_i^t = \prod_{\|\theta_i\|_1 \leq R} \left(\sum_{j=1}^m w_{ij} \left(\theta_j^{t-1} - \gamma \nabla f_j(\theta_j^{t-1}) \right) \right), \tag{5}$$

respectively, where θ_i^t is the estimate from agent i of the common variable θ at iteration t; $\prod_{\|\cdot\|_1 \leq R}(\bullet)$ denotes the Euclidean projection of its argument onto the ℓ_1-ball $\{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq R\}$, with $R > 0$; $\gamma \in (0, 1]$ is the stepsize; and w_{ij}'s are suitably chosen nonnegative weights, such that $w_{ij} = 0$ if there is no link between i and j. In words, each agent i in DGD-CTA builds its local update first performing one step of mixing with the neighbors' estimates θ_j^t (termed consensus step)–aiming at enforcing asymptotic agreement among all the variables–followed by a "correction" based on its own gradient ∇f_i, and finally projected onto the ℓ_1-ball to enforce sparsity. On the other hand, the updates in DGD-ATC swap the order of consensus and optimization steps, mixing thus local parameters *and* gradients.

Despite their popularity, statistical-computational guarantees of DGD-CTA and DGD-ATC remain elusive in *high-dimension*. Existing studies are of pure optimization type–lacking statistical properties of the limit points of the iterates (4) and (5); furthermore, they are suitable only for *low-dimensions* (see Sec. I-B for details). In the companion paper [20] we provide the first statistical analysis of DGD-CTA in *high-dimension*; this work complements [20] studying DGD-ATC, offering thus a comparative analysis of the two schemes in high-dimension. More specifically, in [20], we studied the statistical properties of DGD-CTA applied to the LASSO problem in *Lagrangian* form. Same conclusions can be proved for the LASSO problem in the *projected* form (2): For standard statistical models of predictors and stepsize $\gamma = \mathcal{O}(d^{-1})$, the iterates generated by DGD-CTA enter, with high-probability, an ε-neighborhood of a statistically optimal estimate of $\theta^\star$ in

$$\mathcal{O}\left(\kappa_\Sigma \frac{d\, m^2 \, \log m}{1-\rho} \cdot \log \frac{1}{\varepsilon} \right) \quad \text{communications (iterations)}, \tag{6}$$

where $\rho \in [0, 1)$ is a measure of the connectivity of the network, the smaller ρ, the more connected the graph. This result is of the same type of (3), showing thus that *centralized* statistical accuracy is achievable over mesh networks at *linear* rate. However, such a rate scales undesirably as $\mathcal{O}(d)$, which contrasts with the rate-invariant property of PGD in the centralized setting, as shown by (3). This is a consequence of the stepsize choice $\gamma = \mathcal{O}(d^{-1})$. Numerical results in [20] confirm that $\gamma = \mathcal{O}(d^{-1})$ cannot be eased, if one aims for centralized statistical accuracy.

The role of γ is to control consensus errors, induced by the use of local gradients ∇f_j in the updates (4) rather than the full gradient ∇F. A natural question is then whether mixing the gradient along with the iterations, as in DGD-ATC, will improve the rate dependence on the ambient dimension. Understanding statistical-computational guarantees of DGD-ATC in high-dimension as well as whether it improves over DGD-CTA are open problems. This work provides an answer to these questions, complementing the study in the companion paper [20] of DGD-CTA in high-dimension.

A. Main contributions

Our contributions can be summarized as follows:

(i) Linear convergence up to a tolerance: Under suitably restricted notions of strong convexity and smoothness of F [2] (see Sec. II)–which hold with high probability for a variety of statistical models underlying (1)–we identify tuning recommendation ensuring the iterates generated by DGD-ATC to converge at *linear* rate to a limit point that is within a fixed tolerance from the centralized LASSO solution $\widehat{\theta}$–see Theorem 2.
Notice that, given the different nature of the CTA- and ATC-form updates, the analysis in [20] is not usable to study DTC-ATC: CTA-updates (4) can be interpreted as those of the (proximal) gradient algorithm applied to a centralized, lifted, penalized optimization problem associated to the original LASSO formulation. This is not the case for the ATC-iterates (5), which calls for an ad-hoc analysis directly in the distributed domain.

(ii) Statistical-computational guarantees: When customized to standard statistical models underlying (1) (see Assumption 2), convergence results in (i) hold with high probability (see Theorem 3); for instance, for (sub-)Gaussian predictors and under $s \log d/N = o(1)$ (needed for centralized statistical consistency) and

$$\rho \leq \texttt{poly}^{-1}(d, m, \kappa_\Sigma), \tag{7}$$

DGD-ATC with stepsize $\gamma = \mathcal{O}(1)$ enters with high-probability an ε-neighborhood of a statistically optimal estimate of $\theta^\star$ in

$$\mathcal{O}\left(\kappa_\Sigma \, \log \frac{1}{\varepsilon} \right) \quad \text{communications (gradient evaluations)}, \tag{8}$$

where we recall κ_Σ is the condition number of the covariance matrix Σ of the data (see Assumption 2). This rate matches (the order of) the one of the centralized PGD [2].
When not met by the given graph and gossip matrix, condition (7) on ρ, asking for a sufficiently connected network, can be enforced via multiple rounds of communications per iteration t, resulting in

$$\widetilde{\mathcal{O}}\left(\kappa_\Sigma \frac{\log(d\, m)}{1-\rho} \log \frac{1}{\varepsilon} \right) \quad \text{communications}, \tag{9}$$

where $\widetilde{\mathcal{O}}$ hides log-factors on optimization parameters but m and d. This improves over the complexity of DGD-CTA in (6), exhibiting a log-scaling with the ambient dimension d versus a much less favorable linear one in DGD-CTA. Numerical results show that these scalings are fairly tight (see Sec. V).

Our analysis reveals an interesting, yet discovered, feature of DGD-ATC versus DGD-CTA in high-dimensions: mixing the local gradients along with the estimates, as DGD-ATC employs, unlocks the use of stepsize values independent on d (as in the PGD), yielding the more favorable rate scaling with d as in (9) ($\log d$ versus d) while still achieving centralized statistical errors. The burden on controlling consensus errors is left to the network, which explains condition (7) on ρ, ensuring sufficiently fast mixing. On the other hand, lacking the gradient mixing, DGD-CTA does not enjoy this feature: no matter how small ρ is, the stepsize γ cannot be freed by the dependence on d. This fact cannot be inferred by existing comparative studies of DGD-CTA and DGD-ATC (e.g., [41]), all performed in the *low-dimensional* setting.

B. Related works

DGD-like methods: As anticipated, closely related to this paper are the works that studied DGD algorithms in the CTA [13], [27], [29], [30], [50], [51] and ATC [11], [12], [33], [49] forms. When applied to the minimization of an average loss $f(\theta) = 1/m \sum_{i=1}^m f_i(\theta)$, convergence guarantees of these distributed methods can be roughly summarized as follows: (i) for strongly convex and smooth losses f_i (or f satisfying the KL property [13], [51]), both type of schemes using a constant stepsize converge at linear rate, but only to a neighborhood of the minimizer of the average-loss f [50], [51], and the size of the neighborhood scales as $\mathcal{O}(\gamma)$ (and is monotonic on ρ for DGD-ATC). Convergence to the exact minimizer is achieved employing diminishing stepsize rules, at the price of slower sublinear rate [19], [51]. (ii) When the loss functions are weakly convex, sublinear convergence is certified for both methods, using a diminishing stepsize. A comparison between DGD schemes in the ATC and CTA form can be found in [41].

These results are unsatisfactory when applied to the LASSO problem (2), and do not provide any statistical guarantee. Specifically, **(i)** for fix d and N, they would predict sublinear convergence rate, as the loss F is convex but not strongly convex (recall $d > N$); this would lead to the misleading conclusion that, differently from the PGD in the centralized setting, fast convergence to LASSO estimators is not achievable over mesh networks by DGD-like algorithms, a fact that contrasts with empirical evidences (see Sec. V) showing linear convergence of both DGD-CTA and DGD-ATC, up to some tolerance. **(ii)** When d grows faster than N–the typical situation in high-dimension–the aforementioned studies break down. In fact, they all require *global smoothness* of the loss functions f_i's and F, a property that no longer holds under the scaling $d/N \to \infty$: for commonly used designs of predictors x_i's, the Lipschitz constant of ∇F grows indefinitely with d/N [44].

A statistical study of a DGD scheme that resembles DGD-CTA, applied to the LASSO problem in the Lagrangian form over mesh networks, was recently developed in the companion work [20]: linear convergence to statistically optimal solutions at rate as in (6) was certified. Statistical-computational guarantees of DGD-ATC remains an open problem in high-dimension, which are the contribution of this paper. We remark that the convergence analysis we put forth here is different from that in [20] for DGD-CTA, since the latter can be reinterpreted as the centralized gradient method applied to a lifted, penalized formulation, and thus builds on the convergence analysis of the PGD in high-dimensions. There exists no such interpretation for DGD-ATC in (5), which calls for a different line of analysis.

Beyond DGD methods: The literature of distributed optimization is plenty of schemes but DGD; they differ from plain DGD for implementing some form of correction of the local gradient direction, via distributed tracking mechanisms of the full gradient [25], [28], [31], [39], [40], [48] or using dual variables [16]–[18], [35]–[37]. A detailed discussion of these methods goes beyond the scope of this work–we refer the readers to the excellent tutorial [27] for more details. Here, we only remark that, as for DGD-like methods, convergence analyses of these other methods lack of statistical arguments and break down in high-dimension. The only exception we are aware of is the recent work [39], which studied convergence of a distributed gradient-tracking (DGT) method, applied to the LASSO problem over networks. In contrast to the DGD methods (4) and (5), in DGT, agents employ a correction of their local directions ∇f_i forming a local estimate of the average gradient ∇F. This is achieved via a suitably designed dynamic consensus mechanism on the local gradients (a.k.a. gradient tracking) [25], [48]. Under some technical assumptions, the scheme is proved to reach a neighborhood of a statistically optimal estimate of the unknown, sparse parameter at a linear rate matching that of the centralized proximal gradient up to $\mathcal{O}(s \log d/N)$.

The above overview shows that, despite the popularity of DGD(-ATC) algorithms in the literature, the understanding of its statistical and computational guarantees (along with its comparison with DGD-CTA) in high-dimension remain elusive. This paper addresses this open problem, shedding light on the role of the network in the statistical computational tradeoffs of DGD algorithms.

C. Notation and paper organization

The rest of the paper is organized as follows. Sec. II introduces the main assumptions on the data model and network setting. Convergence of DGD-ATC is studied in Sec. III– under RSC and RSM, linear convergence of the optimization error is proved up to a tolerance. Sec. IV particularizes the convergence results in Sec. II to the statistical model under Assumption 2: linear convergence up to centralized statistical precision is certified with high probability. Numerical results supporting the theoretical findings are provided in Sec. V.

Notation: Let $[m] \triangleq \{1, \dots, m\}$, $m \in \mathbb{N}_{++}$; 1 is the vector of all ones; $e_i \in \mathbb{R}^d$ is the i-th canonical vector; I_d is the $d \times d$ identity matrix (when unnecessary, we omit the subscript); $\otimes$

denotes the Kronecker product; and $A \succ 0$ (resp. $A \succeq 0$) stands for A being positive definite (resp. semidefinite). 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 $\|\cdot\|_\infty$; with a slight abuse of notation, we still use $\|\bullet\|_0$ to denote the cardinality function. Consistently, when applied to matrices, $\|\cdot\|$ denotes the operator norm induced by $\|\cdot\|$. 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 $\mathcal{S}$. Let $\mathbb{B}_p(R)$ denote the ℓ_p-norm ball with radius R, for any $p = 0, 1, 2, \ldots$; consistently with the adopted notation, $\mathbb{B}_0(R)$ is the set of vectors with sparsity at most R. Finally, we recall that, for a random variable X, the ψ_1-Orlicz norm is defined as $\|X\|_{\psi_1} = \inf\left\{t > 0 : \mathbb{E}[\exp(|X|/t)] \leq 2\right\}$ [43, Definition 2.7.5]. Furthermore, $\|X\|_{\psi_1} < \infty$ is equivalent to X belonging to the class of sub-exponential random variables [43, Proposition 2.7.1].

The following quantities associated with (2) will be used throughout the paper:

$$\mathcal{S} \triangleq \operatorname{supp}\{\theta^*\}, \quad s = |\mathcal{S}|, \quad L_{\max} \triangleq \max_{i\in[m]} \lambda_{\max}\left(\frac{X_i^\top X_i}{n}\right). \tag{10}$$

Finally, we collect all the local data $(y_i, X_i)_{i=1}^m$ into the stacked vector $\mathbf{y} = [y_1^\top, \ldots, y_m^\top]^\top \in \mathbb{R}^N$ and matrix $\mathbf{X} = [X_1^\top, \ldots, X_m^\top]^\top \in \mathbb{R}^{N\times d}$.

II. Setup and background

In this section we introduce the main assumptions on the data model and network setting.

A. Problem setting

In the high-dimension setting, $d > N$, the empirical loss F in (2) is not strongly convex uniformly–the $d \times d$ Hessian matrix $\mathbf{X}^\top\mathbf{X}/N$ has at most rank N. However, strong convexity and smoothness hold along a restricted set of directions, which is enough to employ fast convergence and favorable statistical guarantees of the PGD in the centralized setting [2]. Here we postulate the same properties for the landscape of F, as stated next.

Assumption 1 (RSC/RSM condition [2])**.** *F in (2) satisfies the Restricted Strong Convexity (RSC) property with curvature $\mu > 0$ and tolerance $\tau_\mu > 0$:*

$$\frac{1}{N}\|\mathbf{X}\Delta\|^2 \geq \frac{\mu}{2}\|\Delta\|^2 - \frac{\tau_\mu}{2}\|\Delta\|_1^2, \quad \forall \Delta \in \mathbb{R}^d, \tag{11}$$

and the Restricted Smoothness property (RSM) with parameter $L > 0$ and tolerance $\tau_L > 0$:

$$\frac{1}{N}\|\mathbf{X}\Delta\|^2 \leq \frac{L}{2}\|\Delta\|^2 + \frac{\tau_L}{2}\|\Delta\|_1^2, \quad \forall \Delta \in \mathbb{R}^d. \tag{12}$$

It is assumed $L \geq \mu$.

The RSC/RSM conditions above are certified with high probability by a variety of random design matrices $\mathbf{X}$. Here we consider the following.

Assumption 2. *Suppose the design matrix $\mathbf{X}$ satisfies one of the following random designs:*

- **(a) Gaussian model:** *The rows of $\mathbf{X} \in \mathbb{R}^{N\times d}$ are i.i.d. $\mathcal{N}(0, \Sigma)$, with $\Sigma \succ 0$. Let $\kappa_\Sigma \triangleq \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ denote the condition number of Σ;*
- **(b) Sub-Gaussian model:** *The rows of $\mathbf{X} \in \mathbb{R}^{N\times d}$ are centered i.i.d. sub-Gaussian with parameters (Σ_x, σ_x^2), where $\Sigma_x \succ 0$;*
- **(c) Sub-exponential model:** *The entries of the matrix $\mathbf{X}$ are centered independent sub-exponential random variables centered with variance one and $\|\mathbf{X}_{ij}\|_{\psi_1} \leq \psi$, for all $i \in [N]$ and $j \in [d]$, and finite $\psi > 0$.*

Lemma 1. *Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ be a random design matrix, the following hold:*

- **(a) Gaussian model:** *[32, Theorem 1] Under Assumption 2(a), there exist universal constants $c_0, c_1 > 0$ such that, with probability at least $1 - \exp(-c_0 N)$, the RSC condition* (11) *and RSM condition* (12) *hold with parameters*

$$(\mu, \tau_\mu) = \left(\lambda_{\min}(\Sigma), 2c_1\zeta_\Sigma\frac{\log d}{N}\right)$$
$$\text{and} \quad (L, \tau_L) = \left(4\lambda_{\max}(\Sigma), 2c_1\zeta_\Sigma\frac{\log d}{N}\right), \tag{13}$$

  respectively, with $\zeta_\Sigma \triangleq \max_{i\in[d]} \Sigma_{ii}$;
- **(b) Sub-Gaussian model:** *[24, Lemma 1] Under Assumption 2(b) and*

$$N \geq \frac{4}{c_2} s \log d \max\left\{\frac{\sigma_x^2}{\lambda_{\min}^2(\Sigma_x)}, 1\right\}, \tag{14}$$

  with probability at least

$$1 - 2\exp\left(-\frac{c_2}{2}N\min\left\{\frac{\lambda_{\min}^2(\Sigma_x)}{\sigma_x^4}, 1\right\}\right), \tag{15}$$

  the RSC condition (11) *and RSM condition* (12) *hold with parameters*

$$(\mu, \tau_\mu) = \left(\frac{\lambda_{\min}(\Sigma_x)}{2}, \frac{c_3\sigma_x^4\log d}{\lambda_{\min}(\Sigma_x)N}\right)$$
$$\text{and} \quad (L, \tau_L) = \left(\frac{3\lambda_{\max}(\Sigma_x)}{2}, \frac{c_3\sigma_x^4\log d}{\lambda_{\min}(\Sigma_x)N}\right), \tag{16}$$

  respectively;
- **(c) Sub-exponential model:** *Under Assumption 2(c) and*

$$N \geq \frac{\psi^4}{c_4^2}\log^2 d, \tag{17}$$

with probability at least

$$1 - c_5 \exp\left(-c_4\sqrt{s}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right)\right) - 2\exp\left(-\frac{10422c_4\sqrt{N}}{\psi^2}\right), \tag{18}$$

the RSC condition (11) *and RSM condition* (12) *hold with parameters*

$$(\mu, \tau_\mu) = \left(\frac{10395}{10422} - 27c_5\psi^2\sqrt{\frac{s}{N}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right), 54c_5\psi^2\sqrt{\frac{1}{sN}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right) + \frac{54}{10422s}\right),$$

$$(L, \tau_L) = \left(\frac{10449}{10422} + 27c_5\psi^2\sqrt{\frac{s}{N}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right), 54c_5\psi^2\sqrt{\frac{1}{sN}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right) + \frac{54}{10422s}\right),$$

respectively, where $c_4, c_5 > 0$ are universal constants.

Proof. The statement (c) is proved in Appendix B. □

Sub-Gaussian ensemble covers several types of random design matrices, including general bounded random [44, Theorem 2.2.6], Bernoulli [5], [26], and Gaussian random designs [32]. Sub-exponential designs capture random designs with with heavier tails than sub-Gaussians [14], [38]; examples include element-wise square of sub-Gaussian [43, Lemma 2.7.6], element-wise product of sub-Gaussians [43, Lemma 2.7.7], and Johnson-Lindenstrauss random projection for dimension reduction [9, Lemma 1].

B. Network setting

The network of agents is modeled 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 standard assumption on that $\mathcal{G}$ is connected, which is necessary for the convergence of distributed algorithms. Given the DGD-ATC scheme (5), we make the following standard assumption on the weight matrix $W \triangleq (w_{ij})_{i,j=1}^m$, where $\mathcal{P}_K$ denotes the set of polynomials with degree no larger than $K = 1, 2, \ldots$.

Assumption 3. *[On the weight matrix W] The matrix $W = (w_{ij})_{i,j=1}^m$ belongs to the following class $W = P_K(\overline{W})$, where $P_K \in \mathcal{P}_K$, with $P_K(1) = 1$, and $\overline{W} \triangleq (\bar{w}_{ij})_{i,j=1}^m$ has a sparsity pattern compliant with $\mathcal{G}$, that is* **(i)** *$\bar{w}_{ii} > 0$, for all $i \in [m]$;* **(ii)** *$\bar{w}_{ij} > 0$, if $(i, j) \in \mathcal{E}$; and $\bar{w}_{ij} = 0$ otherwise. Furthermore, $\overline{W}$ is symmetric and stochastic, that is, $\overline{W}1 = 1$ (and thus also $1^\top\overline{W} = 1^\top$). Define $\rho \triangleq \|W - 11^\top/m\|$.*

It follows from Assumption 3 that

$$\rho = \max\{\lambda_2(W), |\lambda_{\min}(W)|\} < 1. \tag{19}$$

Roughly speaking, ρ measures how fast the network mixes information (the smaller, the faster).

Several rules for choosing $\overline{W}$ have been proposed in the literature satisfying Assumption 3, such as the Laplacian, the Metropolis-Hasting, and the maximum-degree weights rules; see, e.g., [27] and references therein. When $K > 1$, K rounds of communications per iteration t are employed in the DGD updates (4) and (5) (one iteration is counted as one computation of the gradient). This can be performed, for instance, using in each of the K communication exchanges the same given reference matrix $\overline{W}$ (satisfying Assumption 3), with associated $\bar{\rho} = \|\overline{W} - 11^\top/m\| < 1$, resulting in $W = \overline{W}^K$. Such a W satisfies Assumption 3, with $\rho = \|\overline{W}^K - 11^\top/m\| = \|(\overline{W} - 11^\top/m)^K\| = \bar{\rho}^K$. Faster information mixing can be obtained using suitably designed polynomials $P_K(\overline{W})$, such as Chebyshev [3], [34] or Jacobi polynomials [7].

III. Convergence Analysis

This section provides our first convergence result of DGD-ATC in high-dimension: under RSC and RSM, linear convergence of the optimization error $(1/m)\sum_{i=1}^m\|\theta_i^t - \widehat{\theta}\|^2$ is proved up to a tolerance of $\mathcal{O}(\|\widehat{\theta} - \theta^*\|^2)$.

We begin introducing the key quantities instrumental to state convergence of DGD-ATC. Recalling the parameters in the RSC/RSM condition (Assumption 1), let us define

$$r_\rho \triangleq \max\left\{\sqrt{\frac{1 - \frac{1}{\kappa} + \frac{8s(2\tau_L + \tau_\mu)}{L}}{1 - \frac{16s\tau_L}{L}}} + \rho^{1/2}\left(\frac{2L_{\max}}{L} + \frac{1}{2} + \frac{\tau_L d}{2L}\right), 2\rho^{1/2}\left(1 + \frac{L_{\max}}{L}\right)\right\}, \tag{20}$$

which will determine the convergence rate of DGD-ATC. The initial optimality gap is defined as

$$\eta_0 \triangleq 2\|\theta_{\text{av}}^0 - \widehat{\theta}\|, \tag{21}$$

for given $\theta_i^0, i \in [m]$. Finally we introduce the tolerance on the final optimization error:

$$\Delta_{\text{stat}} \triangleq \Delta_{\text{cent}} + \Delta_{\text{dist}}, \tag{22}$$

where

$$\Delta_{\text{cent}} \triangleq 8\sqrt{\frac{2(4\tau_L + \tau_\mu)}{L - 16s\tau_L}}\left(\|\widehat{\theta} - \theta^*\|_1 + \sqrt{s}\|\widehat{\theta} - \theta^*\|\right)$$

and

$$\Delta_{\text{dist}} \triangleq 8\rho^{1/2}\left(\frac{L_{\max}}{L} + \frac{1}{2} + \frac{\tau_L d}{2L}\right)\|\widehat{\theta} - \theta^*\| + \frac{8\rho^{1/2}d^{1/2}}{L}\left(\frac{\max_{i\in[m]}\|X_i^\top w_i\|_\infty}{n} + \frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N}\right).$$

Notice that Δ_{stat} is composed of a network independent term, Δ_{cent} (matching centralized statistical precision), and a network dependent one Δ_{dist}.

Theorem 2. *Consider the LASSO problem* (2) *under Assumption 1. Let $\{\boldsymbol{\theta}^t\}$ be the iterates generated by DGD-ATC* (5) *from arbitrary, consensual initialization $\boldsymbol{\theta}^0$ (e.g., $\theta_i^0 = 0$, for all $i = 1, \ldots, m$) using a gossip matrix W satisfying Assumption 3, and stepsize $\gamma = 1/L$. Suppose that*

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}\|\theta_i^{t+1}-\widehat{\theta}\|^2} \le \eta^0 r^{t+1} + \underbrace{\frac{1}{1-r}\cdot 24\sqrt{\frac{2s(4\tau_L+\tau_\mu)}{L-16s\tau_L}}\,\|\widehat{\theta}-\theta^*\|}_{\text{centralized error}} + \underbrace{\frac{\rho^{1/2}g(d,m)}{1-r}\left(\|\widehat{\theta}-\theta^*\|+\sqrt{\frac{\log d}{\log md}}\cdot\frac{\max_{i\in[m]}s^{1/2}\|X_i^\top w_i\|_\infty}{\mu N}+\frac{s^{1/2}\|\mathbf{X}^\top\mathbf{w}\|_\infty}{\mu N}\right)}_{\text{cost of decentralization}}. \quad (26)$$

the RSC/RSM parameters (μ,τ_μ), (L,τ_L) *and the network connectivity* ρ *are such that* $r_\rho<1$. *Then, for any optimum* $\widehat{\theta}$ *of the problem* (2) *for which* $\|\widehat{\theta}\|_1=R$, *we have*

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}\|\theta_i^{t+1}-\widehat{\theta}\|^2} \le \eta^0 r_\rho^{t+1}+\frac{\Delta_{\text{stat}}}{1-r_\rho}, \quad \forall\, t=0,1,\ldots. \quad (23)$$

Proof. See Sec. III-A. □

Theorem 2 certifies linear convergence of DGD-ATC at rate r_ρ, up to some tolerance. Both r_ρ and Δ_{stat} depend on the RSC/RSM parameters τ_μ,τ_L, the problem-related parameters $\kappa,d,s,L_{\max}$, the network connectivity ρ and network size m.

The next corollary establishes explicit conditions on the these parameters, in particular on ρ, for the rate r_ρ to be of the same order of that of the centralized PGD [2, Theorem 1] and the tolerance Δ_{stat} to match centralized statistical precision $\mathcal{O}(\|\widehat{\theta}-\theta^*\|)$. Specifically, introducing

$$r \triangleq \sqrt{\left(1-\frac{1}{2\kappa}+\frac{8s(2\tau_L+\tau_\mu)}{L}\right)\left(1-\frac{24s\tau_L}{L}\right)^{-1}},$$
$$g(d,m) \triangleq \frac{32L_{\max}}{L}+20+\frac{4\tau_L d}{L}+8m\sqrt{\frac{d\log md}{s\log d}}, \quad (24)$$

we have the following.

Corollary 2.1. *Instate the setting of Theorem 2. In addition, suppose* $R\le\|\theta^*\|_1$,

$$\mu > 80s\tau_L+16s\tau_\mu, \quad \text{and} \quad \rho\le\frac{c_6}{\kappa^2 g^2(d,m)}, \quad (25)$$

where $c_6\in(0,1]$ *is some universal constant. Then, for any optimum* $\widehat{\theta}$ *of the LASSO problem* (2) *for which* $\|\widehat{\theta}\|_1=R$, *Eq.* (26) *at the top of the page holds, for all* $t=0,1,\ldots$.

Proof. See Appendix D. □

The following comments are in order.

(i) On the linear rate r**:** The contraction coefficient r determining the linear decay of the optimization error depends, as expected, on the restricted condition number κ and the RSC/RSM tolerance parameters τ_μ,τ_L, the latter due to the lack of strong convexity and smoothness in a global sense. Notice that this rate is of the same order of that of the centralized PGD applied to the LASSO problem (2) [2, Theorem 1] and improves on existing analyses of DGD-ATC [11], [12], [33], [49] whose convergence to a solution of (2) is certified only at *sublinear* rate, due to the lack of strong convexity in the global sense (see Sec. I-B). When F is μ-strongly convexity and L-smooth *globally*, i.e., $\tau_\mu=\tau_L=0$, the expression of r reduces to $\sqrt{1-1/(2\kappa)}$, with $\kappa=L/\mu$ becoming the condition number of F. This recovers the well-known convergence rate of DGD-ATC in low-dimension ($N>d$) [49].

(ii) On the tolerance error: The tolerance in (26) consists of a network independent and a network dependent term. The smaller ρ, the smaller the overall tolerance. When customized to the centralized setting [2]–$\rho=0$ and $s(\tau_\mu+\tau_L)=o(1)$, with the latter condition necessary for statistical consistency– the overall tolerance reduces to that achievable by the PGD, that is, $o(\|\widehat{\theta}-\theta^*\|)$ [2]. When $\rho\neq0$, we will show in Sec. IV that the overall tolerance in (26) can be made of the order of the centralized statistical error $\|\widehat{\theta}-\theta^*\|$.

(iii) On the condition (25) on ρ**:** To ensure convergence to statistical precision at rate of the order of the centralized PGD, condition (25) on ρ is required. Roughly speaking, (25) calls for the network to be sufficiently connected–the more ill conditioned the problem (the larger κ) or the larger m (network size), the smaller ρ is required. When the network topology is given and ρ does not satisfy (25), one can still enforce it by employing multiple rounds of communications. The communication complexity will be studied explicitly in the next section, where convergence is specialized to the statistical model.

A. Proof of Theorem 2

We decompose the iterates $\boldsymbol{\theta}^{t+1}$ generated by DGD-ATC into the average process θ_{av}^{t+1} and consensus error dynamic $\boldsymbol{\theta}_\perp^{t+1}$, for all $t\ge0,1,\ldots,$

$$\theta_{\text{av}}^{t+1}=\frac{1}{m}\sum_{i=1}^{m}\prod_{\|\theta_i\|_1\le R}\left(\sum_{j=1}^{m}w_{ij}\left(\theta_j^t-\gamma\nabla f_j(\theta_j^t)\right)\right), \quad (27)$$

and

$$\boldsymbol{\theta}_\perp^{t+1}=\boldsymbol{\theta}^{t+1}-1_m\otimes\theta_{\text{av}}^{t+1}. \quad (28)$$

The average estimation error is controlled by these two terms, according to

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}\|\theta_i^{t+1}-\widehat{\theta}\|^2}\le\|\theta_{\text{av}}^{t+1}-\widehat{\theta}\|+m^{-1/2}\|\boldsymbol{\theta}_\perp^{t+1}\|. \quad (29)$$

In Proposition 8 (see Appendix C1) and Proposition 10 (see Appendix C2), we prove the following bounds for $\|\theta_{\text{av}}^{t+1}-\widehat{\theta}\|$ and $\|\boldsymbol{\theta}_{\perp}^{t+1}\|$, respectively:

$$\|\theta_{\text{av}}^{t+1}-\widehat{\theta}\| \leq r_{\text{av}}\|\theta_{\text{av}}^{t}-\widehat{\theta}\| + m^{-1/2}\left(\rho+\frac{L_{\max}}{L}\right)\|\boldsymbol{\theta}_{\perp}^{t}\| + \frac{\Delta_{\text{cent}}}{4}+\varepsilon_{\rho}, \tag{30}$$

$$\|\boldsymbol{\theta}_{\perp}^{t+1}\| \leq \rho\left(1+\frac{L_{\max}}{L}\right)\|\boldsymbol{\theta}_{\perp}^{t}\| + \frac{\rho m^{1/2}L_{\max}}{L}\|\theta_{\text{av}}^{t}-\widehat{\theta}\| + m^{1/2}\cdot\varepsilon_{\rho}, \tag{31}$$

where the rate r_{av}, tolerance Δ_{cent}, and error ε_{ρ} are defined as

$$r_{\text{av}} \triangleq \sqrt{\left(1-\kappa^{-1}+\frac{8s(2\tau_L+\tau_{\mu})}{L}\right)\left(1-\frac{16s\tau_L}{L}\right)^{-1}} + \rho\left(\frac{L_{\max}}{L}+\frac{1}{2}+\frac{\tau_L d}{2L}\right), \tag{32}$$

$$\Delta_{\text{cent}} = 8\sqrt{\frac{2(4\tau_L+\tau_{\mu})}{L-16s\tau_L}}\left(\|\widehat{\theta}-\theta^*\|_1+\sqrt{s}\|\widehat{\theta}-\theta^*\|\right), \tag{33}$$

and

$$\varepsilon_{\rho} \triangleq \rho\left(\frac{L_{\max}}{L}+\frac{1}{2}+\frac{\tau_L d}{2L}\right)\|\widehat{\theta}-\theta^*\| + \frac{\rho d^{1/2}}{L}\left(\frac{\max_{i\in[m]}\|X_i^{\top}w_i\|_{\infty}}{n}+\frac{\|\mathbf{X}^{\top}\mathbf{w}\|_{\infty}}{N}\right), \tag{34}$$

respectively. Notice that $r_{\text{av}} \leq r_{\rho}$. Since, under the assumptions of the theorem, the RSC/RSM parameters (μ,τ_{μ}), (L,τ_L), the problem parameters $d, L_{\max}$, and the network parameters (m,ρ) are such that $r_{\rho}\in(0,1)$, it follows that $r_{\text{av}}\in(0,1)$.

Combining (30) and (31) yields, for any $a>0$,

$$\begin{aligned}
&\|\theta_{\text{av}}^{t+1}-\widehat{\theta}\| + a\|\boldsymbol{\theta}_{\perp}^{t+1}\| \\
&\leq \left(r_{\text{av}}+\frac{a\rho m^{1/2}L_{\max}}{L}\right)\|\theta_{\text{av}}^{t}-\widehat{\theta}\| \\
&\quad + \frac{\Delta_{\text{cent}}}{4} + \left(1+am^{1/2}\right)\cdot\varepsilon_{\rho} \\
&\quad + \left[\rho m^{-1/2}+\frac{m^{-1/2}L_{\max}}{L}+a\left(\rho+\frac{\rho L_{\max}}{L}\right)\right]\|\boldsymbol{\theta}_{\perp}^{t}\| \\
&\overset{\text{(a)}}{\leq} r_{\max}\left(\|\theta_{\text{av}}^{t}-\widehat{\theta}\| + a\|\boldsymbol{\theta}_{\perp}^{t}\|\right) + \frac{\Delta_{\text{cent}}}{4} + \left(1+am^{1/2}\right)\cdot\varepsilon_{\rho},
\end{aligned}$$

where in (a) we defined

$$r_{\max} \triangleq \max\left\{r_{\text{av}}+\frac{a\rho m^{1/2}L_{\max}}{L}, a^{-1}\left(\rho m^{-1/2}+\frac{m^{-1/2}L_{\max}}{L}\right)+\rho+\frac{\rho L_{\max}}{L}\right\}. \tag{35}$$

The first element in (35) is a non-increasing function of a while the second element is a non-decreasing one. We can thus minimize $r_{\max}$ by choosing a such that

$$r_{\text{av}}+\frac{a\rho m^{1/2}L_{\max}}{L} = a^{-1}\left(\rho m^{-1/2}+\frac{m^{-1/2}L_{\max}}{L}\right)+\rho + \frac{\rho L_{\max}}{L},$$

which reads

$$a^2\cdot\frac{\rho m^{1/2}L_{\max}}{L} + a\cdot\left[\sqrt{\frac{1-\frac{\mu}{L}+\frac{8s(2\tau_L+\tau_{\mu})}{L}}{1-\frac{16s\tau_L}{L}}} + \rho\left(\frac{\tau_L d}{2L}-\frac{1}{2}\right)\right] - \left(\rho m^{-1/2}+\frac{m^{-1/2}L_{\max}}{L}\right) = 0.$$

To keep the expression of a simple, instead of solving the second-order equation above, we choose an a that preserves the same scaling on ρ and m of the solution, yielding $a=\rho^{-1/2}m^{-1/2}$. With this choice, $r_{\max}$ reads

$$\begin{aligned}
&r_{\max} \\
&= \max\left\{r_{\text{av}}+\frac{\rho^{1/2}L_{\max}}{L}, \rho^{1/2}\left(\rho+\frac{L_{\max}}{L}\right)+\rho+\frac{\rho L_{\max}}{L}\right\} \\
&\overset{(32),\rho\leq 1}{\leq} \max\left\{\sqrt{\frac{1-\frac{1}{\kappa}+\frac{8s(2\tau_L+\tau_{\mu})}{L}}{1-\frac{16s\tau_L}{L}}} + \rho^{1/2}\left(\frac{2L_{\max}}{L}+\frac{1}{2}+\frac{\tau_L d}{2L}\right), 2\rho^{1/2}\left(1+\frac{L_{\max}}{L}\right)\right\} \\
&\overset{(20)}{=} r_{\rho}.
\end{aligned}$$

Therefore, we can bound $\|\theta_{\text{av}}^{t+1}-\widehat{\theta}\|$ and $\|\boldsymbol{\theta}_{\perp}^{t+1}\|$ in (30) and (31) as

$$\begin{aligned}
\|\theta_{\text{av}}^{t+1}-\widehat{\theta}\| &\leq r_{\rho}^{t}\left(\|\theta_{\text{av}}^{1}-\widehat{\theta}\|+\rho^{-1/2}m^{-1/2}\|\boldsymbol{\theta}_{\perp}^{1}\|\right) + \frac{\Delta_{\text{cent}}}{4(1-r_{\rho})} + \frac{\left(1+\rho^{-1/2}\right)\varepsilon_{\rho}}{1-r_{\rho}}, \\
\|\boldsymbol{\theta}_{\perp}^{t+1}\| &\leq r_{\rho}^{t}\left(\rho^{1/2}m^{1/2}\|\theta_{\text{av}}^{1}-\widehat{\theta}\|+\|\boldsymbol{\theta}_{\perp}^{1}\|\right) + \frac{\rho^{1/2}m^{1/2}\Delta_{\text{cent}}}{4(1-r_{\rho})} + \frac{\rho^{1/2}m^{1/2}\left(1+\rho^{-1/2}\right)\varepsilon_{\rho}}{1-r_{\rho}}.
\end{aligned}$$

Using the above bounds in (29), we obtain

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}\|\theta_i^{t+1}-\widehat{\theta}\|^2} \leq \eta_1\left(1+\rho^{1/2}\right)r_{\rho}^{t} + \frac{\left(1+\rho^{1/2}\right)}{1-r_{\rho}}\left[\frac{\Delta_{\text{cent}}}{4}+\left(1+\rho^{-1/2}\right)\varepsilon_{\rho}\right], \tag{36}$$

where η_1 is a term related to the optimality gap at iteration 1, defined as

$$\eta_1 \triangleq \|\theta_{\text{av}}^{1}-\widehat{\theta}\| + \rho^{-1/2}m^{-1/2}\|\boldsymbol{\theta}_{\perp}^{1}\|.$$

We further bound η_1 as follows:

$$\begin{aligned}
&\eta_1 \\
&\overset{(30),(31)}{\leq} \left(r_{\text{av}} + \frac{\rho^{1/2} L_{\max}}{L}\right) \|\theta_{\text{av}}^0 - \widehat{\theta}\| + \Big(\rho m^{-1/2} \\
&\quad + \frac{m^{-1/2} L_{\max}}{L} + \rho^{1/2} m^{-1/2} + \frac{\rho^{1/2} m^{-1/2} L_{\max}}{L}\Big) \underbrace{\|\boldsymbol{\theta}_{\perp}^0\|}_{=0} \\
&\quad + \frac{\Delta_{\text{cent}}}{4} + \left(1 + \rho^{-1/2}\right) \varepsilon_\rho \\
&\overset{(32)}{=} \sqrt{\frac{\left(1 - \kappa^{-1} + \frac{8s(2\tau_L + \tau_\mu)}{L}\right)}{\left(1 - \frac{16 s \tau_L}{L}\right)}} \|\theta_{\text{av}}^0 - \widehat{\theta}\| \\
&\quad + \left[\rho\left(\frac{L_{\max}}{L} + \frac{1}{2} + \frac{\tau_L d}{2L}\right) + \frac{\rho^{1/2} L_{\max}}{L}\right] \|\theta_{\text{av}}^0 - \widehat{\theta}\| \\
&\quad + \frac{\Delta_{\text{cent}}}{4} + \left(1 + \rho^{-1/2}\right) \varepsilon_\rho \\
&\overset{\rho \leq 1}{\leq} \sqrt{\frac{\left(1 - \kappa^{-1} + \frac{8s(2\tau_L + \tau_\mu)}{L}\right)}{\left(1 - \frac{16 s \tau_L}{L}\right)}} \|\theta_{\text{av}}^0 - \widehat{\theta}\| \\
&\quad + \rho^{1/2}\left(\frac{2L_{\max}}{L} + \frac{1}{2} + \frac{\tau_L d}{2L}\right) \|\theta_{\text{av}}^0 - \widehat{\theta}\| + \frac{\Delta_{\text{cent}}}{4} \\
&\quad + \left(1 + \rho^{-1/2}\right) \varepsilon_\rho \\
&\overset{(20)}{\leq} \frac{r_\rho \eta^0}{2} + \frac{\Delta_{\text{cent}}}{4} + \left(1 + \rho^{-1/2}\right) \varepsilon_\rho, \qquad (37)
\end{aligned}$$

where η^0 is defined in (21).

Chaining (37) with (36), we finally obtain

$$\begin{aligned}
&\sqrt{\frac{1}{m}\sum_{i=1}^m \|\theta_i^{t+1} - \widehat{\theta}\|^2} \\
&\leq \frac{\eta^0}{2}\left(1 + \rho^{1/2}\right) r_\rho^{t+1} \\
&\quad + \frac{2\left(1 + \rho^{1/2}\right)}{1 - r_\rho}\left[\frac{\Delta_{\text{cent}}}{4} + \left(1 + \rho^{-1/2}\right)\varepsilon_\rho\right] \\
&\overset{(34)}{=} \frac{\eta^0}{2}\left(1 + \rho^{1/2}\right) r_\rho^{t+1} + \frac{\left(1 + \rho^{1/2}\right)}{1 - r_\rho}\frac{\Delta_{\text{cent}}}{2} + \frac{2}{1 - r_\rho}(\rho \\
&\quad + \rho^{1/2} + \rho^{\frac{3}{2}} + \rho) \cdot \left[\left(\frac{L_{\max}}{L} + \frac{1}{2} + \frac{\tau_L d}{2L}\right)\|\widehat{\theta} - \theta^*\| \right. \\
&\quad \left. + \frac{d^{1/2}}{L}\left(\frac{\max_{i\in[m]} \|X_i^\top w_i\|_\infty}{n} + \frac{\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N}\right)\right] \\
&\overset{\rho \leq 1}{\leq} \eta^0 r_\rho^{t+1} + \frac{\Delta_{\text{cent}}}{1 - r_\rho} \\
&\quad + \frac{8\rho^{1/2}}{1 - r_\rho}\left(\frac{L_{\max}}{L} + \frac{1}{2} + \frac{\tau_L d}{2L}\right)\|\widehat{\theta} - \theta^*\| \\
&\quad + \frac{8\rho^{1/2}}{1 - r_\rho}\frac{d^{1/2}}{L}\left(\frac{\max_{i\in[m]} \|X_i^\top w_i\|_\infty}{n} + \frac{\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N}\right).
\end{aligned}$$

This concludes the proof. □

IV. Guarantees for sparse linear regression

W present now some consequences of Corollary 2.1, customized to the sparse linear regression model. We provide nonasymptotic convergence rate for the DGD-ATC under the random Gaussian model for the data $\mathbf{X}$ [Assumption 2(a)] and random noise vector $\mathbf{w}$. Of particular interest is the scaling of the communication complexity and final tolerance under $s, d/N \to \infty$ and $s \log d/N = \mathcal{O}(1)$. Results are of probabilistic type, as a consequence of Lemma 1(a), which certifies RSC and RSM to hold with high probability. Statistical and computational guarantees of the same flavor are established also for the other random matrix designs in Assumption 2, and discussed in Appendix A.

We preliminary define the following quantities:

$$r = \sqrt{\left(1 - \frac{1}{8\kappa_\Sigma} + \chi(\Sigma)\right)\left(1 - \chi(\Sigma)\right)^{-1}}, \qquad (38)$$

where

$$\chi(\Sigma) \triangleq \frac{12 c_1 \zeta_\Sigma}{\lambda_{\max}(\Sigma)} \cdot \frac{s \log d}{N}, \quad \text{with} \quad \zeta_\Sigma = \max_{i\in[d]} \Sigma_{ii}; \qquad (39)$$

$$\begin{aligned}
g(d, m) \triangleq &16 c_8 \cdot \frac{d + \log m}{n} + 20 + 2c_1 \zeta_\Sigma \cdot \frac{d \log d}{N \lambda_{\max}(\Sigma)} \\
&+ 8m\sqrt{\frac{d \log md}{s \log d}}, \qquad (40)
\end{aligned}$$

where $c_8 \geq 2$ is an universal constant. Furthermore, the centralized tolerance reduces to

$$\Delta \triangleq 24\sqrt{\frac{\chi(\Sigma)}{1 - \chi(\Sigma)}}. \qquad (41)$$

Using the above notations, convergence of DGD-ATC is stated next.

Theorem 3. *Consider the LASSO problem* (2)*, with design matrix* $\mathbf{X}$ *satisfying Assumption 2(a), noise vector* $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N)$*, and regularization parameter* $R \leq \|\theta^*\|_1$*. Furthermore, let*

$$N \geq \frac{c_{13} s \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)} \quad \text{and} \quad \frac{d + \log m}{n} > 1. \qquad (42)$$

Let $\{\boldsymbol{\theta}^t\}$ *be the iterates generated by DGD-ATC* (5)*, using arbitrary, consensual initialization* $\boldsymbol{\theta}^0$*, stepsize* $\gamma = 1/(4\lambda_{\max}(\Sigma))$*, and gossip matrix* W *satisfying Assumption 3 with* ρ *such that*

$$\rho \leq \frac{c_6}{\kappa_\Sigma^2 \, g^2(d, m)}. \qquad (43)$$

Then, for any optimum $\widehat{\theta}$ *of the problem* (2) *for which* $\|\widehat{\theta}\|_1 = R$*, and all* $t = 0, 1, \ldots,$ *there holds*

$$\begin{aligned}
&\sqrt{\frac{1}{m}\sum_{i=1}^m \|\theta_i^{t+1} - \widehat{\theta}\|^2} \\
&\leq \eta^0 r^{t+1} + \underbrace{\frac{\Delta}{1 - r} \cdot \frac{5\sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{6 c_{11} s \zeta_\Sigma \log d}{N}}}_{o\left(\sqrt{s \log d/N}\right)} \\
&\quad + \underbrace{\frac{\rho^{1/2} g(d, m)}{1 - r} \frac{6\sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{6 c_{11} s \zeta_\Sigma \log d}{N}}}_{\mathcal{O}\left(\sqrt{s \log d/N}\right)} \qquad (44)
\end{aligned}$$

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

TABLE I: Scaling of $(1-\rho(m))^{-1}$ with network size m, for different graph topologies.

with probability at least

$$[1-4\exp(-c_{12}\log d)][1-\exp(-c_{12}N)-4\exp(-c_{12}\log d)]. \tag{45}$$

The universal constants above are: $c_1, c_7 > 0, c_6 \in (0,1]$, $c_8 \geq 2$, $c_9 > 32$, $c_{10} = c_9/32 - 1, c_{11} > 2$, $c_{12} = \min\{c_7, c_{10}, (c_{11}-2)/2\}$, *and* $c_{13} = \max\{192c_1, c_9\}$.

Proof. See Sec. IV-B. □

A. Discussion

The following comments are in order.

(i) Linear convergence to statistically optimal estimates: (44) certifies linear convergence at rate r of the average optimization error to an estimate within the statistical precision of the model, that is, $\|\widehat{\theta} - \theta^*\| = \mathcal{O}(\sqrt{s\log d/N})$. Notice that when $\rho = 0$ (fully connected networks or star-topologies), the statistical ball improves to $o(\sqrt{s\log d/N})$, matching that of the centralized PGD [2], see (3). For a fixed network (satisfying (43)), the dependency of r on the ambient dimension d, the total sample size N, and sparsity level s is only through the ratio $s\log d/N$ (see (38)). This implies that such a rate is invariant under the high-dimensional scaling $s, d/N \to \infty$ and $s\log d/N = \mathcal{O}(1)$. Notice also that, under $s\log d/N = o(1)$ and ρ satisfying (43), an ε-neighborhood of a statistically optimal solution is achieved in $\mathcal{O}\left(\kappa_\Sigma \log(1/\varepsilon)\right)$ number of iterations (communications). This is of the same order of the rate of the centralized PGD [2].

(ii) Near optimal sample complexity: The above statistical guarantees are achieved under condition (42) on the total sample size N. This is nearly minimax optimal as $N = \Omega(s\log(d/s))$. This proves that centralized statistical consistency is achieved also when local sample sizes n do not satisfy such a condition. This is possible thanks to the information mixing employed throughout the network, and thus at some communication cost, which will be quantified next.

(iii) On the communication complexity and scaling: As anticipated, condition (43) on ρ, when not met by the given graph and gossip matrix W, can be enforced via multiple rounds of communications per iteration. In fact, given W satisfying Assumption 3, with $\rho = \|W - 11^\top/m\| < 1$, one can build the new matrix W^K with $\|W^K - 11^\top/m\| = \rho^K$, and any $K = 1, 2, \ldots$. As discussed in Sec. II-B, this matrix still satisfies Assumption 3 and, when used in the update (5) instead of W, corresponds to employing K rounds of communications per gradient evaluation, each time using the gossip matrix W. Now one can choose K so that ρ^K satisfies (43), resulting in

$$K = \left\lceil \frac{\log\left(\kappa_\Sigma^2\, g^2(m,d)/c_2\right)}{\log(1/\rho)} \right\rceil = \mathcal{O}\left(\frac{\log\left(d\; m\,\kappa_\Sigma\,(1+\lambda_{\min}(\Sigma))\right)}{1-\rho}\right) \tag{46}$$

communications per iteration. We remark that the dependence on $1-\rho$ in (46) can be improved to $\sqrt{1-\rho}$ if K-order Chebyshev polynomials are used as gossip matrix; see, e.g., [34], [47].

Using (46), we then conclude that an ε-neighborhood of a stationary optimal solution is reached in at most

$$\mathcal{O}\left(\kappa_\Sigma \frac{\log\left(d\; m\,\kappa_\Sigma\,(1+\lambda_{\min}(\Sigma))\right)}{1-\rho}\log(1/\varepsilon)\right) \tag{47}$$

communications. This improves on the communication complexity of DGD-CTA [20] (see (6)), showing a more favorable log-scaling with the ambient dimension d and the network size m. These bounds are fairly tight, as confirmed by our experiments in Sec. V.

(iv) Network dependence/scaling. According to (47) (see also (44)), the network topology affects the convergence rate of DGD-ATC as well as the statistical accuracy through the terms $\log(d\,m)/\,(1-\rho)^{-1}$ and $\rho^{1/2}g(d,m)$, respectively. As expected, larger m or $\rho \in [0,1)$ yields more communications and larger estimation error. Notice that $\rho = \rho(m)$ itself is a function of m (and the network topology).

Referring to (44), recall that, under (43), $\rho(m)^{1/2}g(d,m) = \mathcal{O}(1/\kappa_\Sigma)$. Thus, (44) remains within the centralized statistical error $\mathcal{O}\left(\sqrt{s\log d/N}\right)$ even for increasing ρ and m, as long as (43) is enforced via K communication rounds per iteration, with K given by (46). Therefore statistically consistency is preserved at the cost of more communications.

To quantify the scaling of the communication complexity (46) with m, Table. I provides the dependence of $(1-\rho(m))^{-1}$ therein with m for different graphs, when the weights in W are chosen according to the lazy Metropolis rule [27]. For instance, complete graphs and Erdös-Rényi graphs have the favorable scaling $(1-\rho(m))^{-1} = \mathcal{O}(1)$, in contrast with path graphs ($\mathcal{O}(m^2)$) or 2-d grid graphs ($\mathcal{O}(m\log m)$). While this is informative of the impact of the specific topology and network size on the total number of communications as in (47), it does not capture the entire cost of communications from the agents. For instance, denser networks are expected to generate more traffic. In this light, counting each edge as one channel use in each communication, a measure of communication cost might be the total channel uses to ε-solutions. It is not difficult to check that for complete graphs or Erdös-Rényi graphs with

edge probability $p = \mathcal{O}(1)$, such a communication cost reads $\mathcal{O}(m^2)$ total channel uses while for Erdös-Rényi graph with $p = \log m/m$, it reduces to a more favorable $\widetilde{\mathcal{O}}(m)$ ($\widetilde{\mathcal{O}}$ hides log-factors in the communication complexity).

(v) Comparison with CTA-DGD. Theorem 3 shows that, when the network connectivity ρ is sufficiently small (see (43)), DGD-ATC can adopt constant stepsize $\gamma = \mathcal{O}(1)$, converging to a neighborhood of $\widehat{\theta}$ of size $\mathcal{O}(\sqrt{s \log d/N})$. This is in sharp contrast to the convergence result of DGD-CTA, which requires the stepsize $\gamma = \mathcal{O}(1/d)$ regardless of the network connectivity [20].

To explain the phenomenon intuitively, adding and subtracting the centralized gradient ∇F, we can rewrite (4) and (5), respectively as, for all $t = 1, 2, \ldots,$

$$\texttt{DGD-CTA:} \quad \theta_i^t = \prod_{\|\theta_i\|_1 \le R} \Big(\sum_{j=1}^m w_{ij} \theta_j^{t-1} - \gamma \nabla F(\theta_i^{t-1}) + \gamma \underbrace{(\nabla F(\theta_i^{t-1}) - \nabla f_i(\theta_i^{t-1}))}_{\text{gradient discrepancy}} \Big), \quad (48)$$

and

$$\texttt{DGD-ATC:} \quad \theta_i^t = \prod_{\|\theta_i\|_1 \le R} \Big(\sum_{j=1}^m w_{ij} \left(\theta_j^{t-1} - \gamma \nabla F(\theta_j^{t-1})\right) + \gamma \sum_{j=1}^m w_{ij} \underbrace{\left(\nabla F(\theta_j^{t-1}) - \nabla f_j(\theta_j^{t-1})\right)}_{\text{gradient discrepancy}} \Big). \quad (49)$$

Note that, without the gradient discrepancy term–which is generally $\mathcal{O}(d)$–both algorithms coincide with centralized PGD with consensual initialization, i.e., $\theta_i^0 = \theta_j^0$, for all $i, j \in [m]$. For CTA, the impact of the gradient discrepancy term is controlled by requiring the stepsize $\gamma = \mathcal{O}(1/d)$ to attain a solution of the same statistical precision as the centralized PGD. As for ATC, in addition to being multiplied by γ, the gradient discrepancy term is further averaged by network consensus. This provides the opportunity to control the gradient discrepancy leveraging both γ and the network connectivity ρ: one can choose $\gamma = \mathcal{O}(1)$ while requiring ρ inversely proportional to d.

The extra degree of freedom to control the error in ATC using the network leads to significant improvements over CTA in the high-dimensional setting when $d \to \infty$. Limited by the stepsize choice $\gamma = \mathcal{O}(1/d)$, the computation and communication complexity of CTA grow linearly with d. On the other hand, the the computation complexity of ATC is independent of d, thanks to the larger stepsize choice $\gamma = \mathcal{O}(1)$ and the fact that the network connectivity can be improved exponentially by running multiple communication steps, yielding more favorably communication complexity scaling as $\mathcal{O}(\log d)$.

B. Proof of Theorem 3

The proof is based on the following four steps: **1)** We fix $\mathbf{w}$ and consider as source of randomness the design matrix $\mathbf{X}$ (cf. Lemma 1 Gaussian model) only, deriving a high-probability upper bound for $L_{\max}$ defined in (10) and proving that $F(\theta)$ in (2) satisfies RSC and RSM conditions (Assumption 1) with high-probability; **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 show that (43) is sufficient for the condition on ρ in (25) to hold with high-probability; **4)** Under (42), we show that $\mu > 80 s\tau_L + 16 s\tau_\mu$ holds with high-probability; and finally **5)** given the bound on the optimality gap as in (26), we conclude that (44) holds with high-probability, for all $\widehat{\theta}$ satisfying $\|\widehat{\theta}\|_1 = R$.

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

$$L_{\max} = \max_{i \in [m]} \lambda_{\max}(X_i^\top X_i/n) \quad \text{and} \quad \zeta_\Sigma = \max_{i \in [d]} \Sigma_{ii}.$$

$$\begin{aligned}
&\frac{c_6 L^2}{\kappa^2 \left(32 L_{\max} + 20L + 4\tau_L d + 8mL\sqrt{\frac{d \log md}{s \log d}}\right)^2} \\
&\ge \frac{16 c_6 \lambda_{\max}^2(\Sigma)}{16\kappa_\Sigma^2 \left(32 c_8 \lambda_{\max}(\Sigma)\left(1 + \frac{d + \log m}{n}\right) + 80\lambda_{\max}(\Sigma) + 8c_1 \zeta_\Sigma \frac{d \log d}{N} + 32 m \lambda_{\max}(\Sigma)\sqrt{\frac{d \log md}{s \log d}}\right)^2} \\
&\overset{d+\log m > n}{\ge} \frac{c_6 \lambda_{\max}^2(\Sigma)}{\kappa_\Sigma^2 \left(64 c_8 \lambda_{\max}(\Sigma)\frac{(d + \log m)}{n} + 80\lambda_{\max}(\Sigma) + \frac{8 c_1 \zeta_\Sigma d \log d}{N} + 32 m \lambda_{\max}(\Sigma)\sqrt{\frac{d \log md}{s \log d}}\right)^2} \\
&= \frac{c_6}{16\kappa_\Sigma^2 \left[16 c_8 \cdot \frac{d + \log m}{n} + 20 + 2c_1 \zeta_\Sigma \cdot \frac{d \log d}{N \lambda_{\max}(\Sigma)} + 8m\sqrt{\frac{d \log md}{s \log d}}\right]^2} \\
&\overset{(40)}{\ge} \frac{c_6}{\kappa_\Sigma^2 g^2(d, m)}.
\end{aligned} \quad (53)$$

Define the following events:

$$A_1 \triangleq \left\{ \mathbf{X} \in \mathbb{R}^{N\times d} \;\middle|\; L_{\max} \le c_8 \lambda_{\max}(\Sigma)\left(1+\frac{d+\log m}{n}\right)\right\},$$

$$A_2 \triangleq \left\{ \mathbf{X} \in \mathbb{R}^{N\times d} \;\middle|\; \mathbf{X} \text{ satisfies (11) and (12)}\right\},$$

$$A_3 \triangleq \left\{ \mathbf{X} \in \mathbb{R}^{N\times d} \;\middle|\; \max_{j\in[d]} \frac{1}{\sqrt{N}}\|\mathbf{X}e_j\| \le \sqrt{\frac{3\zeta_\Sigma}{2}}\right\},$$

where $c_8 \ge 2$ is some universal constant, chosen as in (50) below. We prove next that these events occur jointly with high-probability.

(i) Bounding $\mathbb{P}(A_1)$ and $\mathbb{P}(A_3)$: Using [20, Theorem 7, (83)] we infer that there exist universal constants $c_7 > 0$ and $c_8 \ge 2$ such that, with probability at least $1-2\exp(-c_7 d)$, it holds

$$L_{\max} \le c_8 \lambda_{\max}(\Sigma)\left(1+\frac{d+\log m}{n}\right). \tag{50}$$

In addition, [20, Theorem 7, (89)] shows that there exists a universal constant $c_9 > 32$, such that, for all $N \ge c_9 \log d$, we have

$$\mathbb{P}\left(\max_{j\in[d]} \frac{\|\mathbf{X}e_j\|^2}{N} \le \frac{3}{2}\zeta_\Sigma\right) \ge 1-2\exp(-c_{10}\log d), \tag{51}$$

where $c_{10} = c_9/32 - 1 > 0$.

(ii) Bounding $\mathbb{P}(A_2)$: This follows readily from (13) in Lemma 1.

Define $A \triangleq A_1 \cap A_2 \cap A_3$. Combining (50), (51), (13) and using the union bound, we obtain

$$\mathbb{P}(A) \ge 1-2\exp(-c_7 d) - \exp(-c_0 N) - 2\exp(-c_{10}\log d).$$

• **Step 2: Randomness from $\mathbf{w}$.** We fix now $\mathbf{X} \in A$ and consider $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 I_N)$. Define

$$D_1 \triangleq \left\{ \mathbf{w} \in \mathbb{R}^N \;\middle|\; \frac{\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N} \le \sigma\sqrt{\frac{3\zeta_\Sigma}{2}}\sqrt{\frac{c_{11}\log d}{N}}\right\},$$

$$D_2 \triangleq \left\{ \mathbf{w} \in \mathbb{R}^N \;\middle|\; \frac{\max_{i\in[m]}\|X_i^\top w_i\|_\infty}{n} \le \sigma\sqrt{\frac{3\zeta_\Sigma}{2}}\sqrt{\frac{c_{11} m\log md}{n}}\right\},$$

and $D \triangleq D_1 \cap D_2$. Following similar steps as to get [20, Theorem 7, (99)], we deduce that, for all $c_{11} > 2$,

$$\begin{aligned}
&\mathbb{P}(A\cap D)\\
&\ge [1-4\exp\{-[(c_{11}-2)\log d]/2\}][1-\exp(-c_0 N)\\
&\quad -2\exp(-c_7 d) - 2\exp(-c_{10}\log d)]\\
&\ge [1-4\exp(-c_{12}\log d)]\\
&\quad \cdot [1-\exp(-c_{12}N) - 4\exp(-c_{12}\log d)].
\end{aligned} \tag{52}$$

where $c_{12} = \min\{c_7, c_{10}, (c_{11}-2)/2\}$.

• **Step 3: $\gamma = 1/4\lambda_{\max}(\Sigma)$ is sufficient for $\gamma = 1/L$ to hold with high-probability.** It follows from (13) that $\gamma = 1/(4\lambda_{\max}(\Sigma))$ is sufficient for γ to equal $1/L$ with probability at least $1-\exp(-c_0 N)$.

• **Step 4: Condition (25) on ρ holds with high probability under (43).** Substituting into (25) the expressions of (μ, τ_μ) and (L, τ_L) (see (13)), $\kappa = L/\mu$, and the high probability upper bound for $L_{\max}$ (see (50)), yields with probability at least (52), (53) at the bottom of the previous page holds.

Therefore, (43) is sufficient for (25) to hold with probability at least (52).

• **Step 5: (42) is sufficient to guarantee $\mu > 80 s\tau_L + 16 s\tau_\mu$.** Substituting into (24) the expression of (μ, τ_μ) and (L, τ_L) (see (13)) yields, with probability at least $1-\exp(-c_0 N)$,

$$r = \sqrt{\left(1 - \frac{1}{8\kappa_\Sigma} + \chi(\Sigma)\right)(1-\chi(\Sigma))^{-1}},$$

where $\chi(\Sigma)$ is defined in (39). In addition, if

$$N \ge 192 c_1 s\zeta_\Sigma \frac{\log d}{\lambda_{\min}(\Sigma)}, \tag{54}$$

then $\mu > 80 s\tau_L + 16 s\tau_\mu$ holds with probability at least $1-\exp(-c_0 N)$. Chaining (54) with $N \ge c_9 \log d$, we conclude that (42) is sufficient for both of them to hold. This can been seen from

$$N \ge \frac{c_{13} s\zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)} \overset{(a)}{\ge} \max\left\{\frac{192 s c_1 \zeta_\Sigma \log d}{\lambda_{\min}(\Sigma)}, c_9 \log d\right\}, \tag{55}$$

where $c_{13} = \max\{192 c_1, c_9\}$, and in (a) we used $s \ge 1$ and $\zeta_\Sigma \ge \lambda_{\min}(\Sigma)$.

• **Step 6: (44) holds with high-probability for all $\widehat{\theta}$ satisfying $\|\widehat{\theta}\|_1 = R \le \|\theta^*\|_1$. Step 1-5** and $R \le \|\theta^*\|_1$ imply that (26) holds with high-probability. Substituting into (26) the expressions of r from (38), (μ, τ_μ), (L, τ_L) from (13), the upper bound of $L_{\max}$ as in (50), and the upper bound for $\max_{i\in[m]}\|X_i^\top w_i\|_\infty/N$, $\|\mathbf{X}^\top \mathbf{w}\|_\infty/N$, the following holds with probability at least (52),

$$\begin{aligned}
&\sqrt{\frac{1}{m}\sum_{i=1}^m \|\theta_i^{t+1} - \widehat{\theta}\|^2}\\
&\overset{(41)}{\le} \eta^0 r^{t+1} + \frac{\Delta}{1-r}\cdot\|\widehat{\theta}-\theta^*\|\\
&\quad + \frac{\rho^{1/2} g(d,m)}{1-r}\cdot\left(\|\widehat{\theta}-\theta^*\| + \frac{\sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{6 c_{11} s\zeta_\Sigma \log d}{N}}\right),
\end{aligned} \tag{56}$$

where Δ is defined in (41). Invoking [15, Theorem11.1], we have

$$\|\widehat{\theta}-\theta^*\| \le \frac{8\sqrt{s}}{\mu - 4 s\tau_\mu}\frac{\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N}. \tag{57}$$

Substituting (μ, τ_μ), and the upper bound for $\|\mathbf{X}^\top \mathbf{w}\|_\infty/N$ into (57), the following holds with probability at least (52),

$$\begin{aligned}
\|\widehat{\theta}-\theta^*\| &\le \frac{8\sqrt{s}}{\lambda_{\min}(\Sigma) - 8 s c_1 \zeta_\Sigma \log d/N}\sigma\sqrt{\frac{3\zeta_\Sigma}{2}}\sqrt{\frac{c_{11}\log d}{N}}\\
&\overset{(42),(55)}{\le} \frac{8\sqrt{s}}{\lambda_{\min}(\Sigma) - 8\lambda_{\min}(\Sigma)/192}\sigma\sqrt{\frac{3\zeta_\Sigma}{2}}\sqrt{\frac{c_{11}\log d}{N}}\\
&\le \frac{5}{\lambda_{\min}(\Sigma)}\sigma\sqrt{\frac{6 c_{11} s\zeta_\Sigma \log d}{N}}.
\end{aligned} \tag{58}$$

Chaining (56) with (58) completes the proof. □

V. Numerical Results

In this section, we provide some experiments on synthetic and real data; results on synthetic data are meant to validate our theoretical findings. We run simulations on a server equipped with Intel(R) Xeon(R) CPU E5-2699A v4 @ 2.40GHz. We organize the experiments as follows:

1) Our first simulation shows that, with a proper choice of ρ, DGD-ATC exhibits linear convergence up to centralized statistical precision; also both the rate and tolerance are invariant to $s\log d/N$. This validates (44);
2) Our second experiment aims at checking the dependence of ρ on the ambient dimension d, problem condition number κ_Σ and network size m, supporting (43);
3) We contrast DGD-ATC and DGD-CTA; experiments confirm a communication complexity of the two schemes scaling as predicted by (9) and (6), respectively;
4) We conclude the section by testing DGD-ATC and DGD-CTA on high-dimension real data, showing that DGD-ATC achieves centralized MSE error at a fast linear rate, while DGD-CTA exhibits a speed accuracy dilemma.

Experimental setup (synthetic data): Given (1), the ground truth θ^* is generated by randomly sampling a multivariate Gaussian $\mathcal{N}(0, I_d)$ and thresholding the smallest $d-s$ elements to zero. The noise vector w follows $\mathcal{N}(\mathbf{0}, 0.25I_N)$. Each row of $\mathbf{X} \in \mathbb{R}^{N\times d}$ is independently generated, according to the following procedure [2]. Let $z_1, \ldots, z_{d-1}$ be i.i.d. $\mathcal{N}(0,1)$, for a fixed correlation $\omega \in [0,1)$, set $x_{i,1} = z_1/\sqrt{1-\omega^2}$ and $x_{i,t+1} = \omega x_{i,t} + z_t$, for $t \in [d-1]$ and $i \in [N]$. It can be verified that all the eigenvalues of $\Sigma = \mathrm{cov}(x_i)$ lie within the interval $[1/(1+\omega)^2, 2/[(1-\omega)^2(1+\omega)]]$. Therefore, the closer ω to one, the larger the condition number κ_Σ. 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 coefficients of the weight matrix W used in all distributed algorithms are chosen according to the Metropolis–Hastings rule [23]. The stepsize γ of DGD-ATC is set to $\gamma = (1-\omega)^2(1+\omega)/8 \leq 1/(4\lambda_{\max}(\Sigma))$. Results are averaged over 30 Monte Carlo repetitions.

1) Linear convergence up to centralized statistical precision (Fig. 1). Fig. 1 plots the estimation error vs. the iteration, for growing $(N,d) = \{(240,400), (560,6400), (860,51200)\}$, fixed $m = 20$, and $s = \lfloor \log d \rfloor$, so that the statistical precision $s\log d/N \approx 0.125$; $\rho = 0.2$ satisfies (43). We observe the following: **(i)** DGD-ATC shrinks linearly up to the centralized LASSO error, as predicted by (44); and **(ii)** both convergence rate and tolerance remain invariant under s, d, N growing and fixed $s\log d/N$; this is consistent with the dependencies on the rate and tolerance as shown in (38) and (44), respectively. Note that this implies that N can significantly exceed the total communications to statistical optimality. For instance, in Fig. 1, the number of communications to reach centralized statistical consistency is of the order of the hundreds, and remains so even when the sample size N is about 3.5 times larger ($N = 860$).

2) Scaling of ρ with d, κ_Σ and m (Fig. 2): We validate here the aforementioned scaling of ρ as given in (43). **(i)**

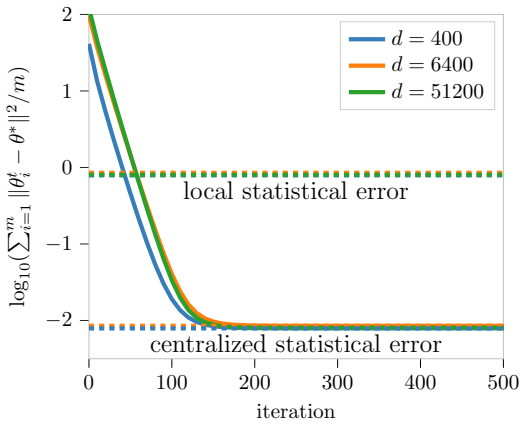

Fig. 1: Estimation error vs. iteration, for different values of s, d, N such that $s\log d/N \approx 0.125$; $m = 20$, $\rho = 0.2$.

Scaling of ρ with d: Fig. 2(a) plots the ratio between the centralized statistical error and estimation error achieved by DGD-ATC versus ρ, for different values of (N, d, s) (as in Fig. 1) and fixed m. The figure shows that, as expected from (43), as d grows, centralized statistical errors are achieved at the price of smaller values of ρ. In other words, ρ cannot be constant with d but vanishing. Fig. 2(b) investigates more in details the scaling of ρ with d. Specifically, we plot the values of ρ (in log scale) to achieve the centralized statistical error within 3% precision versus the ambient dimension d (log-scale). The plot shows a linear dependence of $\log\rho$ with $\log d$, thus validating (43). **(ii) Dependence of ρ on κ_Σ.** Fig. 2(c) plots the estimation error versus iterations for different values of ω and fixed N, d, s, m, ρ, resulting in different κ_Σ. We choose $\omega \to 1$ to make the impact of κ_Σ in (43) dominant. We fix $(N, m, d, s) = (240, 10, 400, 5)$, and simulate a network with $\rho \approx 0.6$, for all $\omega \in \{0.92, 0.93, 0.95\}$. The value of ρ is sufficiently small to achieve centralized statistical error for $\omega = 0.92$. However the figure shows that, to keep centralized statistical consistency, larger values of ω (thus larger $\lambda_{\max}(\Sigma)$) call for smaller values of ρ–a constant ρ instead breaks statistical optimality of the algorithm. **(iii) Scaling of ρ with m:** Fig. 2(d)&(e) plots the estimation error versus iterations for different values of m, and fixed s, d, N. The edge probabilities of the Erdös-Rényi model are set so that, in the subplot (d), ρ remains approximately constant (equal to 0.2) for all values $m \in \{20, 500, 1000\}$ while in the subplot (e) $\rho \in \{0.32, 0.063, 0.045\}$. The two figures show that, to achieve centralized statistical errors, ρ cannot stay constant with m but need to scale roughly as $\rho = \mathcal{O}(1/\sqrt{m})$. While this is consistent with (43), which asks for a vanishing ρ with m, the rate suggested by (43), $\rho = \mathcal{O}(1/m^2)$, seems to be conservative.

3) Communication complexity: CTA-DGD vs. ATC-DGD (Fig. 3): Fig. 3 compares communication complexity of CTA-DGD [20] and ATC-DGD. Panel (a) plots the average estimation error versus the total number of communications. Multiple rounds of communications per gradient evaluation are used in ATC-DGD to enforce condition (43) on ρ, when it is not met by the given graph and gossip matrix W. Both

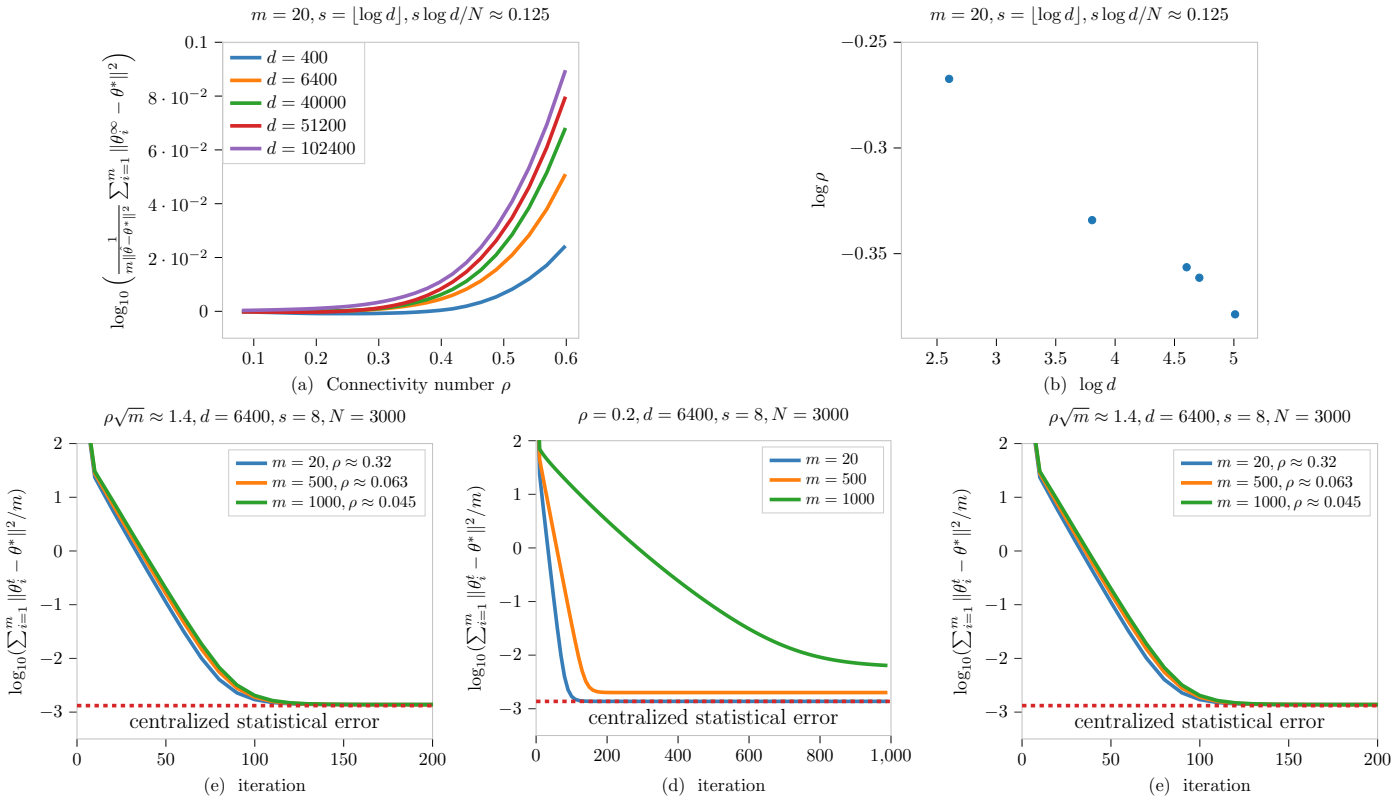

Fig. 2: Scaling of ρ with d, κ_Σ, and m. **(a)**: Ratio between solution error of DGD-ATC and centralized statistical error vs. ρ, for varying s, d, N and fixed m. **(b)**: ρ versus d (log-log scale) to achieve the centralized statistical error within 3% precision. **(c)**: Estimation error versus iterations, for different values of κ_Σ by varying ω, and fixed N, d, s, m, ρ. **(d)**: Estimation error versus iterations, for growing m and ρ fixed. **(e)** Estimation error versus iterations for growing m and $\rho \approx 1/\sqrt{m}$.

schemes achieve centralized statistical errors at linear rate, with ATC-DGD being much faster than CTA-DGD. Panel (b) aims at validating the scaling of the communcation complexity of CTA-DGD and ATC-DGD with d, as predicted by (6) and (9), respectively. We plot the total number of communications needed to reach centralized statistical precision. The figure is obtained generating $(\mathbf{X}, \mathbf{y})$, using different values of $d = \{400, 800, \cdots, 51200\}$, $s = \lfloor \log d \rfloor$, and N chosen accordingly to keep roughly the same statistical precision. We started with a weakly connected graph, $\rho = 0.9$, and, for any chosen d, we run the least number of communications/iteration for ATC-DGD to achieve centralized statistical errors.

The figure shows that the total number of communications scales *logarithmically* with d for ATC-DGD, as predicted by (9), and *linearly* with d for CTA-DGD, as proved in (6). This validates our theoretical findings and supports the conclusion that mixing gradient information among agents, as ATC-DGD does, is critical to save communications.

Experiment on real data. We test the performance of CTA-DGD and ATC-DGD on the dataset `E2006-tfidf` in the `LIBSVM` library [10], which consists of financial risk data from thousands of U.S. companies. There are in total $d = 150360$ features, and $N = 19395$ samples, with $N_{\text{train}} = 16087$ and $N_{\text{test}} = 3308$. We normalize the training data such that each dimension has mean zero and variance $1/N_{\text{train}}$. The testing data is normalized using the statistics computed on the training data. We partition the training data into $m = 10$ subsets. Each agent i owns the training data set portion with size 1608 (we drop 7 samples randomly to divide the sample evenly). 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

$$\begin{aligned} \texttt{MSE}^\infty &\triangleq \frac{1}{mN_{\text{test}}}\sum_{i=1}^m \|y_{\text{test}}^* - \hat{y}_i\|^2 \quad \text{and} \\ \texttt{MSE}^t &\triangleq \frac{1}{mN_{\text{test}}}\sum_{i=1}^m \|y_{\text{test}}^* - y_i^t\|^2, \end{aligned} \tag{59}$$

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; $y_i^t = X_i\theta_i^t$, $i \in [m]$, are the outputs at iteration t; and $\hat{\mathbf{y}} = \mathbf{X}\hat{\theta}$ is the output generated by the PDG ($m = 1$) in the centralized setting. The tuning of the other parameters is the following. We set the projection radius R by grid search to the value yielding the smallest $\texttt{MSE}^\infty$. The stepsize of ATC-DGD is chosen by grid search to achieve the fastest empirical convergence rate while reaching the centralized MSE. The number of communications/iteration of ACT-DGD is set to $K = 18$, resulting being the the least number to achieve centralized MSE over a weakly connected graph with $\rho = 0.9$. For CTA-DGD [20], we tested a few stepsize values; however, because of the size of the problem, even fairly small values are not enough to drive CTA-DGD to achieve centralized MSE within the a reasonable number of

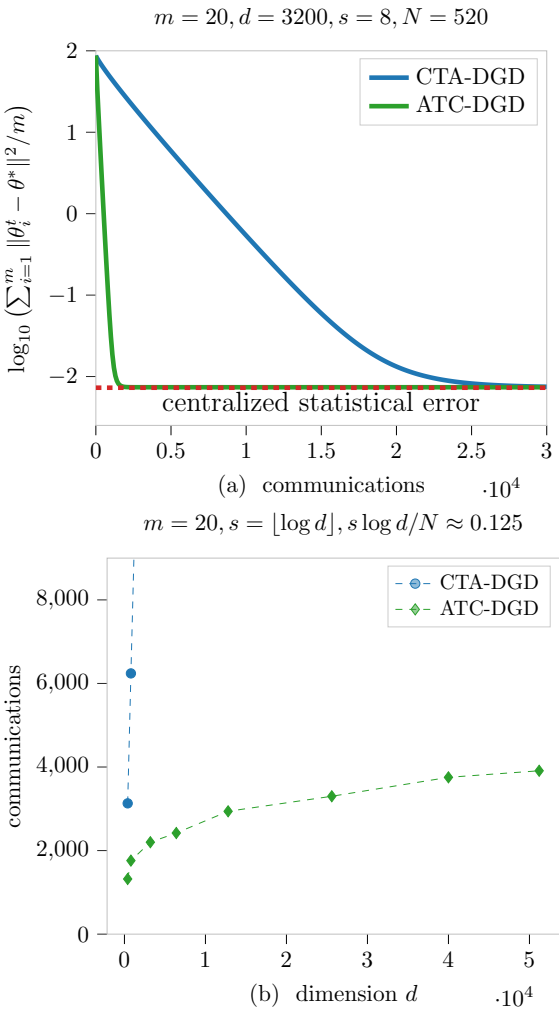

Fig. 3: ATC-DGD vs. CTA-DGD. **(a):** Estimation error vs. total communications. **(b):** Communications to centralized statistical precision vs. dimension $d(> N)$.

communications–this is due to unfavorable linear scaling of the communications with the ambient dimension d.

Fig. 4 plots the $\texttt{MSE}^t$ versus the number of communications. ATC-DGD achieves centralized MSE at linear rate within 2000 communications, while CTA-DGD lacks behind, exhibiting a speed accuracy dilemma: smaller MSE errors are achieved (by using smaller and smaller stepsize values) at the cost of slow convergence.

VI. Conclusions

We established statistical and computational guarantees of the DGD algorithm in the ATC-form, applied to a distributed instance of the projected LASSO problem over mesh networks wherein each agent owns only a subset of data. Under near optimal (total) sample complexity–e.g., $N = \Omega(s \log d)$ for (sub)-Gaussian predictors–DGD-ATC provably achieves statistically optimal estimates at linear rate–the rate is of the same order of that of PGD solving the LASSO problem in a centralized fashion using all data samples N. For worst-case networks i.e., sparse topologies, the communication complexity–the number of communications for statistical consistency–

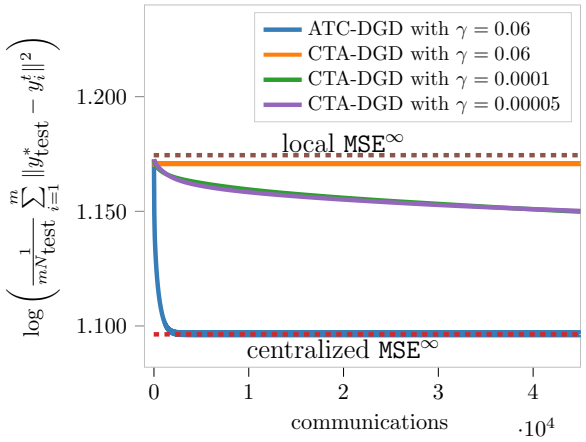

Fig. 4: $\texttt{MSE}^t$ vs.communications for ATC-DGD and CTA-DGD, using the dataset `E2006-tfidf` in the `LIBSVM` library.

scales logarithmically with the ambient dimension d. This showed a significant improvement over DGD in the CTA-form, whose communication complexity scales linearly with d [20]. This difference is sensible in high-dimensions, where typically $s, d, N \to \infty$, with $d > N$. We showed that this is due to the fact that, in the ATC updates, the stepsize can be chosen as $\gamma = \mathcal{O}(1)$, as long as $\rho \leq \texttt{poly}^{-1}(d)$, resulting in a logarithmic number of communications per iteration with respect to d. On the other hand, the CTA updates lack mixing local gradients; because of that, centralized statistical errors can be achieved only under stepsize $\gamma = \mathcal{O}(1/d)$, resulting in a number of iteration- and communication-scaling proportional to d.

At the high-level, this work along with the companion papers [20], [39] showed that when it comes to distributed algorithms applied to high-dimensional statistical problems, classical analyses in the literature of distributed optimization–which are based on sole optimization arguments–are no longer adequate; new studies are needed bringing statistical thinking in distributed optimization. Hopefully this paper will inspire new studies of distributed algorithms beyond DGD under this lens (e.g., distributed primal-dual methods), whose statistical guarantees remain unknown in high-dimension.

Appendix

A. Statistical and computational guarantees for other statistical models

In this section, we present the statistical and computational guarantees of DGD-ATC (the the counterpart of Theorem 3) for random design matrices $\mathbf{X}$ following a sub-Gaussian [Assumption 2(b)] or sub-exponential [Assumption 2(c)] distribution–the two cases are discussed in Sec. A1 and Sec A2, respectively.

A1. Sub-Gaussian ensemble (Assumption 2(b)): Suppose that $\mathbf{X}$ satisfies Assumption 2(b). Define the following quantities:

$$r = \sqrt{\left(1 - \frac{1}{6\kappa_{\Sigma_x}} + \chi(\Sigma_x)\right)(1 - \chi(\Sigma_x))^{-1}}, \qquad (60)$$

where

$$\chi(\Sigma_x) \triangleq \frac{16c_3\sigma_x^4}{\lambda_{\max}(\Sigma_x)\lambda_{\min}(\Sigma_x)} \cdot \frac{s\log d}{N}, \tag{61}$$

$$g(d,m) \triangleq 128c_{15} \cdot \frac{d+\log m}{n} + 40 + \frac{8c_3\sigma_x^4 d\log d}{\lambda_{\max}(\Sigma_x)\lambda_{\min}(\Sigma_x)N} + 24m\sqrt{\frac{d\log md}{s\log d}}, \tag{62}$$

and $c_{15} \geq 2$ is an universal constant. Finally, the centralized tolerance reads

$$\Delta \triangleq 24\sqrt{\frac{\chi(\Sigma_x)}{1-\chi(\Sigma_x)}}. \tag{63}$$

Using the above notations, convergence of DGD-ATC is stated as follows.

Theorem 4. *Consider the LASSO problem* (2)*, where the design matrix* $\mathbf{X}$ *satisfies Assumption 2(b), the noise vector* $\mathbf{w}$ *is sub-Gaussian with parameters* $(\sigma^2 I_N, \sigma^2)$*, and the regularization parameter satisfies* $R \leq \|\theta^*\|_1$*. Furthermore, let*

$$N \geq c_{19} s\log d \max\left\{\frac{\sigma_x^2}{\lambda_{\min}^2(\Sigma_x)}, 1\right\} \quad \text{and} \quad \frac{d+\log m}{n} > 1. \tag{64}$$

Let $\{\boldsymbol{\theta}^t\}$ *be the iterates generated by DGD-ATC* (5)*, using arbitrary, consensual initialization* $\boldsymbol{\theta}^0$*, stepsize* $\gamma = 2/(3\lambda_{\max}(\Sigma_x))$*, and gossip matrix* W *satisfying Assumption 3 with* ρ *such that*

$$\rho \leq \frac{c_6}{\kappa_{\Sigma_x}^2 g^2(d,m)}. \tag{65}$$

Then, for any optimum $\widehat{\theta}$ *of* (2) *for which* $\|\widehat{\theta}\|_1 = R$*, and all* $t = 0, 1, \ldots,$ *there holds*

$$\begin{aligned}
&\sqrt{\frac{1}{m}\sum_{i=1}^m \|\theta_i^{t+1} - \widehat{\theta}\|^2} \\
&\leq \eta^0 r^{t+1} + \underbrace{\frac{\Delta}{1-r} \cdot \frac{17c_{16}\sigma\sigma_x}{\lambda_{\min}(\Sigma_x)}\sqrt{\frac{s\log d}{N}}}_{o\left(\sqrt{s\log d/N}\right)} \\
&\quad + \underbrace{\frac{\rho^{1/2} g(d,m)}{1-r} \cdot \frac{19c_{16}\sigma\sigma_x}{\lambda_{\min}(\Sigma_x)}\sqrt{\frac{s\log d}{N}}}_{\mathcal{O}\left(\sqrt{s\log d/N}\right)}
\end{aligned} \tag{66}$$

with probability at least

$$[1 - 4\exp(-c_{18}\log d)] \cdot \left[1 - 2\exp\left(-c_{18}N\min\left\{\frac{\lambda_{\min}^2(\Sigma_x)}{\sigma_x^4}, 1\right\}\right) - 4\exp(-c_{18}d)\right]. \tag{67}$$

The universal constants above are: $c_2, c_3 > 0, c_6 \in (0,1]$, $c_{14} > 0$, $c_{15} \geq 2$, $c_{16} > 3, c_{17} = c_{16}^2/(2 + c_{16}\sqrt{2}) - 1, c_{18} = \min\{c_2/2, c_{14}, c_{17}\}$ *and* $c_{19} = \max\{192c_3, 4/c_2\}$.

Remark 4.1. *When customized to the Gaussian case, the sub-Gaussian parameter reduces to the variance; hence, a natural candidate for* σ_x^2 *is the largest variance, i.e.,* $\sigma_x^2 = \zeta_\Sigma$*. In this case, Theorem 4 recovers the guarantees as established in Theorem 3 for Gaussian random ensemble by noticing* $\zeta_\Sigma \geq \lambda_{\min}(\Sigma)$.

Proof. The proof follow the same logic of that of Theorem 3; the difference is in Steps 1 and 2 wherein we use now the RSC/RSM conditions for sub-Gaussian random variables and the Bernstein inequality [46] to bound $\|\mathbf{X}^\top\mathbf{w}\|_\infty$ and $\max_{i\in m}\|X_i^\top w_i\|$. Next, we then present only the proof of Steps 1 and 2.

• **Step 1: Randomness from X**. Recall $L_{\max} = \max_{i\in[m]}\lambda_{\max}(X_i^\top X_i/n)$. Define the following events:

$$A_1 \triangleq \left\{\mathbf{X} \in \mathbb{R}^{N\times d} \;\middle|\; L_{\max} \leq c_{15}\lambda_{\max}(\Sigma_x)\left(1 + \frac{d+\log m}{n}\right)\right\},$$
$$A_2 \triangleq \left\{\mathbf{X} \in \mathbb{R}^{N\times d} \;\middle|\; \mathbf{X} \text{ satisfies (11) and (12)}\right\},$$

where $c_{15} \geq 2$ is a universal constant (see (50)). We prove next that these two events occur jointly with high-probability.

(i) Bounding $\mathbb{P}(A_1)$: By [42, Remark 5.40], the following holds with probability at least $1 - 2\exp(-c_{14}d)$:

$$L_{\max} \leq c_{15}\lambda_{\max}(\Sigma)\left(1 + \frac{d+\log m}{n}\right), \tag{68}$$

for some universal constants $c_{14} > 0$ and $c_{15} \geq 2$.

(ii) Bounding $\mathbb{P}(A_2)$: This follows readily from (16) (see Lemma 1).

Define $A \triangleq A_1 \cap A_2$. Combining (i) and (ii) and using the union bound, we obtain

$$\mathbb{P}(A) \geq 1 - 2\exp(-c_{14}d) - 2\exp\left(-\frac{c_2}{2}N\min\left\{\frac{\lambda_{\min}^2(\Sigma_x)}{\sigma_x^4}, 1\right\}\right).$$

• **Step 2: Randomness from w**. We fix now $\mathbf{X} \in A$ and treat $\mathbf{w}$ as sub-Gaussian vector with parameters $(\sigma^2 I_N, \sigma^2)$. Define

$$D_1 \triangleq \left\{\mathbf{w} \in \mathbb{R}^N \;\middle|\; \frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N} \leq c_{16}\sigma\sigma_x\sqrt{\frac{\log d}{N}}\right\},$$
$$D_2 \triangleq \left\{\mathbf{w} \in \mathbb{R}^N \;\middle|\; \frac{\max_{i\in[m]}\|X_i^\top w_i\|_\infty}{n} \leq c_{16}\sigma\sigma_x\sqrt{\frac{m\log md}{n}}\right\},$$

and $D \triangleq D_1 \cap D_2$. Since each pair of X_i and w_i are independent, and the columns of X_i are n dimensional i.i.d sub-Gaussian random vectors, we deduce that each element of $X_i^\top w_i$ is the sum of n independent sub-exponential random variables with sub-exponential parameters $(\sqrt{2}\sigma\sigma_x, \sqrt{2}\sigma\sigma_x)$ [44, Exercise 2.13]. Applying Bernstein's inequality [46, Lemma 2.2.11] and the union bound, we have

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

$$\sqrt{\frac{1}{m}\sum_{i=1}^{m}\|\theta_i^{t+1}-\widehat{\theta}\|^2} \le \eta^0 r_\psi^{t+1} + \underbrace{\frac{\Delta}{1-r_\psi}\cdot 10\sqrt{s}\sigma\psi\max\left\{\frac{2\log d}{Nc_{21}},\sqrt{\frac{2\log d}{Nc_{21}}}\right\} + 25\sqrt{s}\sigma\psi\max\left\{\frac{2\log d}{Nc_{21}},\sqrt{\frac{2\log d}{Nc_{21}}}\right\}}_{\mathcal{O}\left(\sqrt{s}\max\left\{\frac{\log d}{N},\sqrt{\frac{\log d}{N}}\right\}\right)}, \quad (79)$$

Take $t = c_{16}\sigma\sigma_x\sqrt{N\log md}$ and any $c_{16} > 3$, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\in[m]}\|X_i^\top w_i\|_\infty \le c_{16}\sigma\sigma_x\sqrt{N\log md}\right)\\
&\ge 1-2\exp\left[-\left(\frac{c_{16}^2 N}{2n+c_{21}\sqrt{2N\log md}}-1\right)\log md\right]\\
&\overset{N\ge\log md}{\ge} 1-2\exp\left[-\left(\frac{c_{16}^2 N}{2N+c_{16}N\sqrt{2}}-1\right)\log md\right]\\
&\ge 1-2\exp\left[-c_{17}\log d\right],
\end{aligned}$$

where $c_{17} = \frac{c_{16}^2}{2+c_{16}\sqrt{2}} - 1 > 0$.

Similarly, for $\mathbf{X}^\top\mathbf{w}$, it holds

$$\mathbb{P}\left(\|\mathbf{X}^\top\mathbf{w}\|_\infty \le c_{16}\sigma\sigma_x\sqrt{N\log d}\right) \ge 1-2\exp\left[-c_{17}\log d\right].$$

Therefore,

$$\begin{aligned}
&\mathbb{P}(A\cap D)\\
&= \mathbb{P}(D|A)\mathbb{P}(A) \quad (69)\\
&\ge [1-4\exp(-c_{17}\log d)]\times\\
&\left[1-2\exp(-c_{14}d)-2\exp\left(-\frac{c_2}{2}N\min\left\{\frac{\lambda_{\min}^2(\Sigma_x)}{\sigma_x^4},1\right\}\right)\right]\\
&\ge [1-4\exp(-c_{18}\log d)]\times\\
&\left[1-2\exp\left(-c_{18}N\min\left\{\frac{\lambda_{\min}^2(\Sigma_x)}{\sigma_x^4},1\right\}\right)-4\exp(-c_{18}d)\right]. \quad (70)
\end{aligned}$$

where $c_{18} = \min\{c_2/2, c_{14}, c_{17}\}$. □

A2. Sub-exponential ensemble: Consider now the sparse regression model with the random matrix $\mathbf{X}$ satisfying Assumption 2(c). Define the following quantities:

$$r_\psi = \sqrt{\left(1-\frac{1}{2\kappa_\psi}+\chi_\psi\right)(1-\chi_\psi)^{-1}}, \quad (71)$$

where

$$\chi_\psi \triangleq 24\cdot\frac{54c_5\psi^2\sqrt{\frac{s}{N}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right)+\frac{54}{10422}}{\frac{10449}{10422}+27c_5\psi^2\sqrt{\frac{s}{N}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right)}, \quad (72)$$

and

$$\kappa_\psi \triangleq \frac{\frac{10449}{10422}+27c_5\psi^2\sqrt{\frac{s}{N}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right)}{\frac{10395}{10422}-27c_5\psi^2\sqrt{\frac{s}{N}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right)}. \quad (73)$$

For any $\epsilon > 0$, define

$$\begin{aligned}
&g(d,m,\epsilon)\\
&\triangleq \frac{64c_{20}d^{1+2\epsilon}}{n\left[\frac{10449}{10422}+27c_5\psi^2\sqrt{\frac{s}{N}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right)\right]}\\
&\quad+\left(20+8m\sqrt{\frac{d\log md}{s\log d}}\right)\\
&\quad+\frac{216c_5d\psi^2\sqrt{\frac{1}{sN}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right)+\frac{4d}{3s}}{\left[\frac{10449}{10422}+27c_5\psi^2\sqrt{\frac{s}{N}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right)\right]}, \quad (74)
\end{aligned}$$

where $c_{20} \ge 2$ is an universal constant. Furthermore, let

$$\Delta \triangleq 24\sqrt{\frac{\chi_\psi}{1-\chi_\psi}}. \quad (75)$$

Using the above notations, convergence of DGD-ATC is proved next.

Theorem 5. *Consider the LASSO problem* (2)*, where the design matrix* $\mathbf{X}$ *satisfies Assumption 2(c), the noise vector* $\mathbf{w}$ *is deterministic with bounded entries* $\|\mathbf{w}\|_\infty \le \sigma$*, and the regularization parameter satisfies* $R \le \|\theta^*\|_1$*. Furthermore, let*

$$N \ge \max\left\{\frac{\psi^4}{c_4^2}\log^2 d, 10422^2c_5^2\psi^4 s\log^2\left(\frac{ed}{s}\sqrt{\frac{N}{s}}\right)\right\}. \quad (76)$$

Let $\{\boldsymbol{\theta}^t\}$ *be the iterates generated by DGD-ATC* (5)*, using arbitrary, consensual initialization* $\boldsymbol{\theta}^0$*, stepsize*

$$\gamma = \frac{6}{7+162c_5\psi^2\sqrt{\frac{s}{N}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right)}, \quad (77)$$

and gossip matrix W *satisfying Assumption 3 with* ρ *such that*

$$\rho \le \frac{c_6}{\kappa_\psi^2 g^2(d,m,\epsilon)}, \quad (78)$$

for the given $\epsilon > 0$*. Then, for any optimum* $\widehat{\theta}$ *of* (2) *for which* $\|\widehat{\theta}\|_1 = R$*, and* $t = 0, 1, \ldots$*, Eq.* (79) *at the top of the page holds, with probability at least*

$$\begin{aligned}
&1-5\exp(-\log d)-\exp\left(-c_4\sqrt{s}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right)\right)\\
&-3\exp\left(-\frac{c_4\sqrt{N}}{\psi^2}\right)-2\exp\left(-c_{21}\min\left\{\frac{d^{2\epsilon}}{\psi^2},\frac{d^{\frac{1}{2}+\epsilon}}{\psi}\right\}\right). \quad (80)
\end{aligned}$$

The universal constants above are: $c_4, c_5 > 0, c_6 \in (0,1]$, $c_{20} \ge 2$*, and* $c_{21} > 0$.

Proof. Similarly to the proof of Theorem 4, in Step 1 and 2 we now use the RSC/RSM conditions for sub-exponential random

$$\mathbb{P}\left(L_{\max} \le c_{20}\left(1+\frac{d^{1+2\epsilon}\log md}{n}\right)\right) \ge (1-\exp(-\log md))\left(1-2\exp\left(-c_{21}\min\left\{\frac{d^{2\epsilon}}{\psi^2},\frac{d^{\frac{1}{2}+\epsilon}}{\psi}\right\}\right)\right)$$
$$\ge 1-\exp(-\log md)-2\exp\left(-c_{21}\min\left\{\frac{d^{2\epsilon}}{\psi^2},\frac{d^{\frac{1}{2}+\epsilon}}{\psi}\right\}\right). \quad (82)$$

variables and the concentration inequality [42, Theorem 5.44] to bound the heavy tail random variable $L_{\max}$ while leveraging Bernstein-type inequality [46] to bound the sub-exponential random variables $\|\mathbf{X}^\top \mathbf{w}\|_\infty$ and $\max_{i\in m}\|X_i^\top w_i\|$.

• **Step 1: Bounding** $L_{\max}$. Define the following events:

$$A_1 \triangleq \left\{\mathbf{X}\in\mathbb{R}^{N\times d}\;\middle|\; L_{\max}\le c_{20}\left(1+\frac{d^{1+2\epsilon}\log md}{n}\right)\right\},$$
$$A_2 \triangleq \left\{\mathbf{X}\in\mathbb{R}^{N\times d}\;\middle|\; \mathbf{X}\text{ satisfies (11) and (12)}\right\},$$

where $c_{20}\ge 2$ is a universal constant (see (50)), and $\epsilon>0$ is arbitrary. We prove next that these events occur jointly with high-probability. **(i) Bounding** $\mathbb{P}(A_1)$**:** Under Assumption 2(c), the d entries of each row of X_i, that is, $e_j^\top X_i e_k$, $k\in[d]$, are independent centered sub-exponential variables. Bernstein's inequality implies

$$\mathbb{P}\left\{\|e_j^\top X_i\|\ge d\cdot t\right\}$$
$$=\mathbb{P}\left\{\sqrt{\sum_{k=1}^{d}|e_j^\top X_i e_k|^2}\ge dt\right\}$$
$$\le \mathbb{P}\left\{\left|\sum_{k=1}^{d}|e_j^\top X_i e_k|\right|\ge d\cdot t\right\}$$
$$\overset{\text{[42, Prop. 5.16]}}{\le} 2\exp\left(-c_{21}\min\left\{\frac{t^2}{\psi^2},\frac{t}{\psi}\right\}d\right), \quad (81)$$

for all $j\in[n]$ and $i\in[m]$, and some universal constant $c_{21}>0$. For any given $\epsilon>0$, set $t=d^{-\frac{1}{2}+\epsilon}$; then,

$$\mathbb{P}\left\{\|e_j^\top X_i\|\ge d^{\frac{1}{2}+\epsilon}\right\}\le 2\exp\left(-c_{21}\min\left\{\frac{d^{2\epsilon}}{\psi^2},\frac{d^{\frac{1}{2}+\epsilon}}{\psi}\right\}\right).$$

By [42, Theorem 5.44] it follows that for any $j\in[n]$, and $i\in[m]$, if $\|e_j^\top X_i\|\le d^{\frac{1}{2}+\epsilon}$, then, the following holds with probability at least $1-\exp\{-c_{22}t^2+\log d\}$,

$$\left\|\frac{1}{n}X_i^\top X_i - I\right\|\le\max\{a,a^2\},\quad \text{with}\quad a\triangleq t\,\frac{d^{\frac{1}{2}+\epsilon}}{\sqrt{n}},$$

for any given $t\ge 0$ and some constant $c_{22}>0$; which implies (with the same probability)

$$\left\|\frac{1}{n}X_i^\top X_i\right\|\le\left\|\frac{1}{n}X_i^\top X_i - I\right\|+\left\|I\right\|\le\max\{a,a^2\}+1.$$

Applying the union bound, the following bound holds for $L_{\max}$:

$$\mathbb{P}\left(L_{\max}\le(1+\max\{a,a^2\})\;\middle|\;\|e_j^\top X_i\|\le d^{\frac{1}{2}+\epsilon},\forall j\in[n]\right)$$
$$\ge 1-\exp\{-c_{22}t^2+\log md\}.$$

Setting $t=\sqrt{2c_{22}^{-1}\log md}$, yields

$$a=\sqrt{2c_{22}^{-1}\log md}\cdot\frac{d^{\frac{1}{2}+\epsilon}}{\sqrt{n}}.$$

Therefore, we conclude, under $\|e_j^\top X_i\|\le d^{\frac{1}{2}+\epsilon},\forall j\in[n]$,

$$L_{\max}\le\left(1+a+a^2\right)\le(1+a)^2\le 2\left(1+a^2\right)$$
$$\le 2\left(1+2c_{22}^{-1}\log md\cdot\frac{d^{1+2\epsilon}}{n}\right)$$
$$\le c_{20}\left(1+\frac{d^{1+2\epsilon}\log md}{n}\right),$$

with probability at least $1-\exp(-\log md)$ and $c_{20}=\max\{2,4c_{22}^{-1}\}\ge 2$.

Chaining it with (81), we conclude that, for any given $\epsilon>0$, Eq. (82) at the top of the page holds.

(ii) Bounding $\mathbb{P}(A_2)$**:** This follows immediately from (19) in Lemma 1.

Define $A\triangleq A_1\cap A_2$. Combining (i), (ii) and using the union bound, we obtain, under (17),

$$\mathbb{P}(A)$$
$$\ge 1-\exp\left(-c_4\sqrt{s}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}}\right)\right)-3\exp\left(-\frac{c_4\sqrt{N}}{\psi^2}\right)$$
$$-\exp(-\log md)-2\exp\left(-c_{21}\min\left\{\frac{d^{2\epsilon}}{\psi^2},\frac{d^{\frac{1}{2}+\epsilon}}{\psi}\right\}\right). \quad (83)$$

• **Step 2: Bounding** $\max_{i\in[m]}\|X_i^\top w_i\|_\infty$ **and** $\|\mathbf{X}^\top\mathbf{w}\|_\infty$. Since X_i, $i\in[m]$, are independent and the columns of X_i are n dimensional i.i.d sub-exponential random vectors, each element of $X_i^\top w_i$ is the sum of n independent sub-exponential random variables with ψ_1-norm at most $\sigma\psi$. Applying [42, Proposition 5.16] and the union bound, we obtain

$$\mathbb{P}\left(\frac{1}{n}\max_{i\in[m]}\|X_i^\top w_i\|_\infty\le t\right)$$
$$\ge 1-2\exp\left(-c_{21}\min\left\{\frac{t^2}{\sigma^2\psi^2},\frac{t}{\sigma\psi}\right\}n+\log md\right),\quad t\ge 0.$$

Thus, under $2\log md\le c_{21}n$ and $t=\sigma\psi\sqrt{\frac{2\log md}{nc_{21}}}$,

$$\mathbb{P}\left(\frac{1}{n}\max_{i\in[m]}\|X_i^\top w_i\|_\infty\le\sigma\psi\sqrt{\frac{2\log md}{nc_{21}}}\right)$$
$$\ge 1-2\exp\left(-c_{21}\min\left\{\frac{2\sigma^2\psi^2\log md}{c_{21}n\sigma^2\psi^2},\frac{\sigma\psi\sqrt{2\log md}}{\sqrt{c_{21}n}\sigma\psi}\right\}n+\log md\right)$$
$$\ge 1-2\exp(-\log d), \quad (84)$$

$$D_1 \triangleq \left\{ \mathbf{X} \in \mathbb{R}^{N\times d} \;\middle|\; \frac{1}{n}\max_{i\in[m]} \|X_i^\top w_i\|_\infty \le \sigma\psi \max\left\{ \frac{2\log md}{nc_{21}}, \sqrt{\frac{2\log md}{nc_{21}}} \right\} \right\},$$
$$D_2 \triangleq \left\{ \mathbf{X} \in \mathbb{R}^{N\times} \;\middle|\; \frac{1}{N}\|\mathbf{X}^\top \mathbf{w}\|_\infty \le \sigma\psi \max\left\{ \frac{2\log d}{Nc_{21}}, \sqrt{\frac{2\log d}{Nc_{21}}} \right\} \right\}, \quad \text{and} \quad D \triangleq D_1 \cap D_2. \tag{88}$$

while, under $2\log md > c_{21}n$ and $t = \frac{2\sigma\psi\log md}{nc_{21}}$, it holds

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\max_{i\in[m]} \|X_i^\top w_i\|_\infty \le \frac{2\sigma\psi\log md}{nc_{21}} \right) \\
&\ge 1 - 2\exp\left(-c_{21}\min\left\{ \frac{4\sigma^2\psi^2\log^2 md}{c_{21}^2 n^2\sigma^2\psi^2}, \frac{2\sigma\psi\log md}{c_{21}n\sigma\psi} \right\} n \right. \\
&\qquad \left. + \log md \right) \\
&\ge 1 - 2\exp\left(-\log d\right). \qquad (85)
\end{aligned}$$

Combining (84) and (85), we have

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\max_{i\in[m]} \|X_i^\top w_i\|_\infty \le \sigma\psi \max\left\{ \frac{2\log md}{nc_{21}}, \sqrt{\frac{2\log md}{nc_{21}}} \right\} \right) \\
&\ge 1 - 2\exp\left(-\log d\right). \qquad (86)
\end{aligned}$$

Similarly, we can prove

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{N}\|\mathbf{X}^\top \mathbf{w}\|_\infty \le \sigma\psi \max\left\{ \frac{2\log d}{Nc_{21}}, \sqrt{\frac{2\log d}{Nc_{21}}} \right\} \right) \\
&\ge 1 - 2\exp\left(-\log d\right). \qquad (87)
\end{aligned}$$

Define D_1 and D_2 as in (88) at the top of the page. Then, chaining (83), (86), and (87), we finally get

$$\begin{aligned}
&\mathbb{P}(A \cap D) \ge \\
&1 - 5\exp\left(-\log d\right) - \exp\left(-c_4\sqrt{s}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}} \right) \right) \\
&- 3\exp\left(-\frac{c_4\sqrt{N}}{\psi^2} \right) - 2\exp\left(-c_{21}\min\left\{ \frac{d^{2\epsilon}}{\psi^2}, \frac{d^{\frac{1}{2}+\epsilon}}{\psi} \right\} \right).
\end{aligned}$$

□

B. Proof of Lemma 1(c)

We begin recalling that, for any random matrix $\mathbf{X}$ satisfying Assumption 1(c), the Restricted Isometry Property (RIP) holds with high-probability [1, Theorem 3.3]. Then, we present a lemma translating the RIP to RSC/RSM conditions.

Definition 1 (RIP [8])**.** *The matrix $\mathbf{X}$ is said to satisfy the Restricted Isometry Property (RIP) with constant $r_s > 0$ if*

$$(1 - r_s)\|\Delta\|^2 \le \frac{1}{N}\|\mathbf{X}\Delta\|^2 \le (1 + r_s)\|\Delta\|^2 \tag{89}$$

holds for all s-sparse vectors $\Delta \in \mathbb{R}^d$.

Chaining [1, Theorem 3.3] (setting therein the free parameter parameter $\theta' = 1/10422$) with [1, Lemma 3.5], we infer the following high-probability result for sub-exponential design matrices $\mathbf{X}$.

Lemma 6. *Let $\mathbf{X}$ be a random matrix satisfying Assumption 2(c), and N such that* (17) *holds. Then, with probability at least*

$$\begin{aligned}
&1 - c_5\exp\left(-c_4\sqrt{s}\log\left(\frac{e\,d\sqrt{N}}{s\sqrt{s}} \right) \right) \\
&- \mathbb{P}\left(\max_{j\le d} \left| \frac{\|\mathbf{X}e_j\|^2}{N} - 1 \right| > \frac{1}{10422} \right),
\end{aligned}$$

$\mathbf{X}$ satisfies the RIP condition, with constant

$$r_s \le c_5\psi^2\sqrt{\frac{s}{N}}\log\left(\frac{ed\sqrt{N}}{s\sqrt{s}} \right) + \frac{1}{10422},$$

where $c_4, c_5 > 0$ are universal constants.

We proceed with bounding

$$\mathbb{P}\left(\max_{j\le d} \left| \frac{\|\mathbf{X}e_j\|^2}{N} - 1 \right| > \frac{1}{10422} \right).$$

Notice that each element of $\mathbf{X}e_j$ is sub-exponential with variance 1, thus $\frac{\|\mathbf{X}e_j\|^2}{N} - 1$ is a symmetric Weibull variable. Using [1, Lemma 3.7], we have

$$\mathbb{P}\left(\max_{j\le d} \left| \frac{\|\mathbf{X}e_j\|^2}{N} - 1 \right| > \frac{1}{10422} \right) \le 2\exp\left(-c_4\frac{10422\sqrt{N}}{\psi^2} \right).$$

We are ready to translate the RIP property to the RSC/RSM conditions.

Lemma 7. *Suppose $\mathbf{X}$ satisfies the RIP with parameter $r_s > 0$. Then, $\mathbf{X}$ satisfies the RSC and RSM properties with parameters*

$$\begin{aligned}
(\mu, \tau_\mu) &= (1 - 27r_s, 54r_s/s) > 0, \\
(L, \tau_L) &= (1 + 27r_s, 54r_s/s).
\end{aligned}$$

Proof. From the RIP of $\mathbf{X}$ it follows

$$\left| \frac{1}{N}\|\mathbf{X}\theta\|^2 - \|\theta\|^2 \right| \le r_s\|\theta\|^2, \quad \forall\theta \in \mathbb{B}_0(s).$$

Therefore, since $\mathbb{B}_0(s) \cap \mathbb{B}_2(1) \subset \mathbb{B}_0(s)$, it holds

$$\left| \theta^\top \left(\frac{\mathbf{X}^\top\mathbf{X}}{N} - I \right)\theta \right| \le r_s, \quad \forall\theta \in \mathbb{B}_0(s) \cap \mathbb{B}_2(1).$$

Applying [24, Lemma 12], we have

$$\left| \theta^\top \left(\frac{\mathbf{X}^\top\mathbf{X}}{N} - I \right)\theta \right| \le 27r_s\left(\|\theta\|^2 + \frac{2}{s}\|\theta\|_1^2 \right) \qquad \forall\theta \in \mathbb{R}^d,$$

which proves the RSC and RSM properties. □

Chaining Lemma 6 and Lemma 7 concludes the proof of Lemma 1(c).

C. Auxiliary Results in the Proof of Theorem 2

This section contains some intermediate results used in the proof of Theorem 2, namely: a bound of $\|\theta_{\text{av}}^{t+1} - \widehat{\theta}\|$ (Proposition 8, Appendix C1) and of $\|\boldsymbol{\theta}_{\perp}^{t+1}\|$ (Proposition 10, Appendix C2).

It is convenient to introduce the following extra notation, which will be used in the proofs of the results in this section. Given the stacked quantities (see Sec. I-C) $\mathbf{y} = [y_1^\top, \dots, y_m^\top]^\top \in \mathbb{R}^N$, $\mathbf{X} = [X_1^\top, \dots, X_m^\top]^\top \in \mathbb{R}^{N\times d}$, and $\boldsymbol{\theta} = [\theta_1^\top, \dots, \theta_m^\top]^\top$, let us define the stacked loss

$$f(\boldsymbol{\theta}) \triangleq \sum_{i=1}^{m} f_i(\theta_i), \tag{90}$$

where f_i is defined in (2). Thus, the stacked gradient reads

$$\nabla f(\boldsymbol{\theta}) = \begin{bmatrix} \nabla f_1(\theta_1) \\ \vdots \\ \nabla f_m(\theta_m) \end{bmatrix} = \frac{1}{n} \begin{bmatrix} X_1^\top (X_1\theta_1 - y_1) \\ \vdots \\ X_m^\top (X_m\theta_m - y_m) \end{bmatrix}. \tag{91}$$

We will also use the following bound

$$\lambda_{\max}\left(\frac{\mathbf{X}^\top\mathbf{X}}{N}\right) \le \frac{L}{2} + \frac{\tau_L d}{2}, \tag{92}$$

which is a consequence of the RSM condition (12), that is, for all $\Delta \in \mathbb{R}^d$, there holds

$$\frac{1}{N}\|\mathbf{X}\Delta\|^2 \le \frac{L}{2}\|\Delta\|^2 + \frac{\tau_L}{2}\|\Delta\|_1^2 \le \left(\frac{L}{2} + \frac{\tau_L d}{2}\right)\|\Delta\|^2.$$

C1. Proposition 8: The proposition provides an upper bound of $\|\theta_{\text{av}}^{t+1} - \widehat{\theta}\|$, used in (30).

Proposition 8. *In the setting of Theorem 2, for all $t = 0, 1, \dots,$ the following bound holds for $\|\theta_{\text{av}}^{t+1} - \widehat{\theta}\|$:*

$$\|\theta_{\text{av}}^{t+1} - \widehat{\theta}\| \le r_{\text{av}}\|\theta_{\text{av}}^{t} - \widehat{\theta}\| + m^{-1/2}\left(\rho + \frac{L_{\max}}{L}\right)\|\boldsymbol{\theta}_{\perp}^{t}\| + \frac{\Delta_{\text{cent}}}{4} + \varepsilon_\rho. \tag{93}$$

Proof. Recall the definition of θ_{av}^{t+1} from (27). Adding and subtracting terms we can rewrite it as, for all $t = 0, 1, \dots,$

$$\theta_{\text{av}}^{t+1} = \prod_{\|\theta\|_1\le R}\left(\theta_{\text{av}}^{t} - \gamma\nabla F(\theta_{\text{av}}^{t})\right) + \frac{1}{m}\sum_{i=1}^{m}\gamma\varepsilon_i^t, \tag{94}$$

where

$$\varepsilon_i^t \triangleq \frac{1}{\gamma}\left[\prod_{\|\theta\|_1\le R}\left(\sum_{j=1}^{m} w_{ij}\left(\theta_j^t - \gamma\nabla f_j(\theta_j^t)\right)\right) - \prod_{\|\theta\|_1\le R}\left(\theta_{\text{av}}^{t} - \gamma\nabla F(\theta_{\text{av}}^{t})\right)\right]. \tag{95}$$

This allows one to interpret $\theta_{\text{av}}^{t+1} - (1/m)\sum_{i=1}^m \gamma\varepsilon_i^t$ as the outcome of one iteration of the (centralized) PGD applied to (2) at θ_{av}^t. Choosing $\gamma = 1/L$ and using the one-step descent inequality [2, Eq. (54)] we obtain

$$\begin{aligned}
&\left\|\theta_{\text{av}}^{t+1} - \frac{1}{Lm}\sum_{i=1}^{m}\varepsilon_i^t - \widehat{\theta}\right\|^2 \\
&\le \frac{1 - \kappa^{-1} + 8s(2\tau_L + \tau_\mu)/L}{1 - 16s\tau_L/L}\|\theta_{\text{av}}^t - \widehat{\theta}\|^2 \\
&\quad + \frac{2(4\tau_L + \tau_\mu)/L}{1 - 16s\tau_L/L}\left(2\|\widehat{\theta} - \theta^*\|_1 + 2\sqrt{s}\|\widehat{\theta} - \theta^*\|\right)^2. \quad (96)
\end{aligned}$$

Our next result is a bound on the error $\|(1/Lm)\sum_{i=1}^m \varepsilon_i^t\|$.

Lemma 9. *For $\left\|\frac{1}{m}\sum_{i=1}^{m}\gamma\varepsilon_i^t\right\|$, Eq. (97) at the bottom of the page holds.*

Proof. See Appendix C3. □

Using the triangle inequality

$$\left\|\theta_{\text{av}}^{t+1} - \widehat{\theta} - \frac{1}{Lm}\sum_{i=1}^{m}\varepsilon_i^t\right\| \ge \|\theta_{\text{av}}^{t+1} - \widehat{\theta}\| - \left\|\frac{1}{Lm}\sum_{i=1}^{m}\varepsilon_i^t\right\|,$$

and applying Lemma 9 with $\gamma = 1/L$ yield

$$\begin{aligned}
&\|\theta_{\text{av}}^{t+1} - \widehat{\theta}\| \\
&\le \sqrt{\frac{1 - \kappa^{-1} + 8s(2\tau_L + \tau_\mu)/L}{1 - 16s\tau_L/L}}\|\theta_{\text{av}}^t - \widehat{\theta}\| + \left\|\frac{1}{Lm}\sum_{i=1}^{m}\varepsilon_i^t\right\| \\
&\quad + \sqrt{\frac{2(4\tau_L + \tau_\mu)/L}{1 - 16s\tau_L/L}}\left(2\|\widehat{\theta} - \theta^*\|_1 + 2\sqrt{s}\|\widehat{\theta} - \theta^*\|\right) \\
&\overset{\text{Lem. 9}}{\le} \left[\sqrt{\frac{1 - \kappa^{-1} + 8s(2\tau_L + \tau_\mu)/L}{1 - 16s\tau_L/L}} + \rho\left(\frac{L_{\max}}{L} + \frac{1}{2} + \frac{\tau_L d}{2L}\right)\right]\|\theta_{\text{av}}^t - \widehat{\theta}\| + m^{-1/2}\left(\rho + \frac{L_{\max}}{L}\right)\|\boldsymbol{\theta}_{\perp}^t\| \\
&\quad + \sqrt{\frac{2(4\tau_L + \tau_\mu)/L}{1 - 16s\tau_L/L}}\left(2\|\widehat{\theta} - \theta^*\|_1 + 2\sqrt{s}\|\widehat{\theta} - \theta^*\|\right) \\
&\quad + \rho\left(\frac{L_{\max}}{L} + \frac{1}{2} + \frac{\tau_L d}{2L}\right)\|\widehat{\theta} - \theta^*\| \\
&\quad + \frac{\rho\, d^{1/2}}{L}\left(\frac{\max_{i\in[m]}\|X_i^\top w_i\|_\infty}{n} + \frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N}\right) \\
&\overset{\text{(a)}}{=} r_{\text{av}}\|\theta_{\text{av}}^t - \widehat{\theta}\| + m^{-1/2}\left(\rho + \frac{L_{\max}}{L}\right)\|\boldsymbol{\theta}_{\perp}^t\| + \frac{\Delta_{\text{cent}}}{4} + \varepsilon_\rho,
\end{aligned}$$

where in (a), we use the definition of $r_{\text{av}}, \Delta_{\text{cent}}$, and ε_ρ in (32), (33), and (34), respectively.

This completes the proof. □

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\gamma\varepsilon_i^t\right\| \le m^{-1/2}\left(\rho + \gamma L_{\max}\right)\|\boldsymbol{\theta}_{\perp}^t\| + \rho\gamma\left(L_{\max} + \frac{L}{2} + \frac{\tau_L d}{2}\right)\left(\|\widehat{\theta} - \theta_{\text{av}}^t\| + \|\widehat{\theta} - \theta^*\|\right) + \rho\gamma d^{1/2}\left(\frac{m\max_{i\in[m]}\|X_i^\top w_i\|_\infty}{\|\mathbf{X}^\top\mathbf{w}\|_\infty} + 1\right)\frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N}, \qquad \forall\ t = 0, 1, \dots. \tag{97}$$

C2. Proposition 10: The proposition provides an upper bound on the consensus error, used in (31).

Proposition 10. *In the setting of Theorem 2, the following bound holds for* $\|\boldsymbol{\theta}_\perp^{t+1}\|$*, for all* $t = 0, 1, \ldots,$

$$\begin{aligned}&\|\boldsymbol{\theta}_\perp^{t+1}\| \\ &\leq \rho\left(1+\frac{L_{\max}}{L}\right)\|\boldsymbol{\theta}_\perp^t\| + \frac{\rho m^{1/2} L_{\max}}{L}\|\theta_{\text{av}}^t - \widehat{\theta}\| + m^{1/2}\cdot\varepsilon_\rho.\end{aligned} \tag{98}$$

Proof. We start rewriting the DGD-ATC as follows: for all $t = 0, 1, \ldots,$

$$\begin{cases} \boldsymbol{\theta}^{t+1/2} = (W\otimes I_d)\left(\boldsymbol{\theta}^t - \gamma\nabla f(\boldsymbol{\theta}^t)\right) \\ \theta_i^{t+1} \;= \underset{\|\theta\|_1\leq R}{\prod}\theta_i^{t+1/2}, \qquad \text{for all} \;\; i\in[m]. \end{cases} \tag{99}$$

We observe that $(1/m)\|\boldsymbol{\theta}_\perp\|^2$ can be interpreted as the variance of a discrete random variable taking values $\theta_1, \ldots, \theta_m$ with uniform probability. Using (99), we can then write

$$\begin{aligned}&\frac{1}{m}\|\boldsymbol{\theta}_\perp^{t+1}\|^2 \\ &= \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m\frac{1}{2}\left\|\theta_i^{t+1}-\theta_j^{t+1}\right\|^2 \\ &\overset{\text{(a)}}{\leq} \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m\frac{1}{2}\left\|\theta_i^{t+1/2}-\theta_j^{t+1/2}\right\|^2 \\ &= \frac{1}{m}\left\|\boldsymbol{\theta}_\perp^{t+1/2}\right\|^2,\end{aligned}$$

where (a) follows from the non-expansiveness of the projection operator.

We proceed to bound $\|\boldsymbol{\theta}_\perp^{t+1/2}\|$ as follows:

$$\begin{aligned}&\left\|\boldsymbol{\theta}_\perp^{t+1/2}\right\| \\ &\overset{(99)}{=} \left\|\left(\left(W-\frac{1}{m}1_m1_m^\top\right)\otimes I_d\right)\left(\boldsymbol{\theta}^t-\gamma\nabla f(\boldsymbol{\theta}^t)\right)\right\| \\ &\overset{\text{(b)}}{=} \left\|\left(\left((W-\frac{1}{m}1_m1_m^\top)\otimes I_d\right)\right.\right. \\ &\qquad \left.\left.\cdot\left[\boldsymbol{\theta}_\perp^t - \gamma\left(\nabla f(\boldsymbol{\theta}^t) - 1_m\otimes\nabla F(\widehat{\theta})\right)\right]\right\| \\ &\overset{(19)}{\leq} \rho\|\boldsymbol{\theta}_\perp^t\| + \rho\gamma\left\|\nabla f(\boldsymbol{\theta}^t) - 1_m\otimes\nabla F(\widehat{\theta})\right\| \\ &\leq \rho\|\boldsymbol{\theta}_\perp^t\| + \rho\gamma\left\|\nabla f(\boldsymbol{\theta}^t) - \nabla f(1_m\otimes\theta_{\text{av}}^t)\right\| \\ &\quad + \rho\gamma\left\|\nabla f(1_m\otimes\theta_{\text{av}}^t) - \nabla f(1_m\otimes\widehat{\theta})\right\| \\ &\quad + \rho\gamma\|\nabla f(1_m\otimes\widehat{\theta}) - 1_m\otimes\nabla F(\widehat{\theta})\|,\end{aligned} \tag{100}$$

where (b) follows from Assumption 3 that $W1_m = 1_m$.

To bound $\left\|\nabla f(1_m\otimes\widehat{\theta}) - 1_m\otimes\nabla F(\widehat{\theta})\right\|$, we insert the points $\nabla f(1_m\otimes\theta^*)$, $1_m\otimes\nabla F(\theta^*)$, and write

$$\begin{aligned}&\left\|\nabla f(1_m\otimes\widehat{\theta}) - 1_m\otimes\nabla F(\widehat{\theta})\right\| \\ &\leq \left\|\nabla f(1_m\otimes\widehat{\theta}) - \nabla f(1_m\otimes\theta^*)\right\| \\ &\quad + \|\nabla f(1_m\otimes\theta^*) - 1_m\otimes\nabla F(\theta^*)\| \\ &\quad + \left\|1_m\otimes\nabla F(\theta^*) - 1_m\otimes\nabla F(\widehat{\theta})\right\| \\ &\overset{(2),(91)}{=} \sqrt{\sum_{i=1}^m\left\|\frac{X_i^\top X_i}{n}(\widehat{\theta}-\theta^*)\right\|^2} + m^{1/2}\left\|\frac{1}{N}\mathbf{X}^\top\mathbf{X}(\widehat{\theta}-\theta^*)\right\| \\ &\quad + \left\|\frac{1}{n}\begin{bmatrix}X_1^\top w_1\\ \vdots \\ X_m^\top w_m\end{bmatrix} - \frac{1}{N}\begin{bmatrix}\mathbf{X}^\top\mathbf{w}\\ \vdots\\ \mathbf{X}^\top\mathbf{w}\end{bmatrix}\right\| \\ &\leq \sqrt{\sum_{i=1}^m\left\|\frac{X_i^\top X_i}{n}(\widehat{\theta}-\theta^*)\right\|^2} + m^{1/2}\,\frac{\max_{i\in[m]}\|X_i^\top w_i\|}{n} \\ &\quad + m^{1/2}\left\|\frac{1}{N}\mathbf{X}^\top\mathbf{X}(\widehat{\theta}-\theta^*)\right\|.\end{aligned} \tag{101}$$

Plugging in the expression of f, F and (101) into (100) gives

$$\begin{aligned}&\left\|\boldsymbol{\theta}_\perp^{t+1/2}\right\| \\ &\leq \rho\|\boldsymbol{\theta}_\perp^t\| + \rho\gamma\sqrt{\sum_{i=1}^m\left\|\frac{X_i^\top X_i}{n}(\theta_i^t-\theta_{\text{av}}^t)\right\|^2} \\ &\quad + \rho\gamma\sqrt{\sum_{i=1}^m\left\|\frac{X_i^\top X_i}{n}(\widehat{\theta}-\theta_{\text{av}}^t)\right\|^2} \\ &\quad + \rho\gamma\sqrt{\sum_{i=1}^m\left\|\frac{X_i^\top X_i}{n}(\widehat{\theta}-\theta^*)\right\|^2} \\ &\quad + \rho\gamma m^{1/2}\,\frac{\max_{i\in[m]}\|X_i^\top w_i\|}{n} + \rho\gamma m^{1/2}\,\frac{\|\mathbf{X}^\top\mathbf{w}\|}{N} \\ &\quad + \rho\gamma m^{1/2}\left\|\frac{1}{N}\mathbf{X}^\top\mathbf{X}(\widehat{\theta}-\theta^*)\right\| \\ &\overset{(10)}{\leq} \rho\|\boldsymbol{\theta}_\perp^t\| + \rho\gamma\sqrt{\sum_{i=1}^m L_{\max}^2\|\theta_i^t-\theta_{\text{av}}^t\|^2} \\ &\quad + \rho\gamma\sqrt{\sum_{i=1}^m L_{\max}^2\left\|\widehat{\theta}-\theta_{\text{av}}^t\right\|^2} \\ &\quad + \rho\gamma\sqrt{\sum_{i=1}^m L_{\max}^2\left\|\widehat{\theta}-\theta^*\right\|^2} + \rho\gamma m^{1/2}\,\frac{\max_{i\in[m]}\|X_i^\top w_i\|}{n} \\ &\quad + \rho\gamma m^{1/2}\,\frac{\|\mathbf{X}^\top\mathbf{w}\|}{N} + \rho\gamma m^{1/2}\left\|\frac{1}{N}\mathbf{X}^\top\mathbf{X}(\widehat{\theta}-\theta^*)\right\|.\end{aligned} \tag{102}$$

It remains to relate $\max_{i\in[m]}\|X_i^\top w_i\|$ and $\|\mathbf{X}^\top\mathbf{w}\|$ to the statistical error bound. Using norm bound $\|x\| \leq d^{1/2}\|x\|_\infty$,

for any $x \in \mathbb{R}^d$, thus, we have

$$\frac{\max_{i\in[m]} \|X_i^\top w_i\|}{n} + \frac{\|\mathbf{X}^\top \mathbf{w}\|}{N} \leq d^{1/2} \left(\frac{\max_{i\in[m]} \|X_i^\top w_i\|_\infty}{n} + \frac{\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N} \right). \quad (103)$$

Using (92), (103) in (102), we obtain

$$\begin{aligned}
&\left\|\boldsymbol{\theta}_\perp^{t+1/2}\right\| \\
&\leq (\rho + \rho\gamma L_{\max}) \|\boldsymbol{\theta}_\perp^t\| + \rho\gamma m^{1/2} L_{\max} \|\widehat{\theta} - \theta_{\text{av}}^t\| \\
&\quad + \rho\gamma m^{1/2} \left(L_{\max} + \frac{L}{2} + \frac{\tau_L d}{2} \right) \|\widehat{\theta} - \theta^*\| \\
&\quad + \rho\gamma m^{1/2} d^{1/2} \left(\frac{\max_{i\in[m]} \|X_i^\top w_i\|_\infty}{n} + \frac{\|\mathbf{X}^\top \mathbf{w}\|_\infty}{N} \right).
\end{aligned}$$

Letting $\gamma = 1/L$ completes the proof. □

C3. Proof of Lemma 9: We decompose ε_i^t defined in (95) as

$$\varepsilon_i^t = \frac{1}{\gamma} \left[\prod_{\|\theta\|_1 \leq R} \left(\theta_{\text{av}}^t - \gamma \nabla F(\theta_{\text{av}}^t) - \gamma \epsilon_i^t \right) - \prod_{\|\theta\|_1 \leq R} \left(\theta_{\text{av}}^t - \gamma \nabla F(\theta_{\text{av}}^t) \right) \right],$$

where

$$\epsilon_i^t \triangleq -\frac{1}{\gamma} \sum_{j=1}^m w_{ij} \left(\theta_j^t - \gamma \nabla f_j(\theta_j^t) \right) + \frac{1}{\gamma} \theta_{\text{av}}^t - \nabla F(\theta_{\text{av}}^t). \quad (104)$$

Thus

$$\begin{aligned}
&\left\| \frac{1}{m} \sum_{i=1}^m \gamma \varepsilon_i^t \right\| \\
&\overset{(a)}{\leq} \frac{1}{m} \sum_{i=1}^m \|\gamma \epsilon_i^t\| \\
&\overset{(104)}{=} \frac{1}{m} \sum_{i=1}^m \left\| \sum_{j=1}^m w_{ij} \left(\theta_j^t - \gamma \nabla f_j(\theta_j^t) \right) - \theta_{\text{av}}^t + \gamma \nabla F(\theta_{\text{av}}^t) \right\| \\
&\leq \underbrace{\frac{1}{m} \sum_{i=1}^m \left\| \sum_{j=1}^m w_{ij} (\theta_j^t - \theta_{\text{av}}^t) \right\|}_{\texttt{Term I}} \\
&\quad + \underbrace{\frac{\gamma}{m} \sum_{i=1}^m \left\| \sum_{j=1}^m w_{ij} \nabla f_j(\theta_j^t) - \nabla F(\theta_{\text{av}}^t) \right\|}_{\texttt{Term II}},
\end{aligned}$$

where in (a) we used the non-expansiveness of the projection operator.

We proceed bounding `Term I` and `Term II`:

$$\begin{aligned}
\texttt{Term I} &\leq \sqrt{\frac{1}{m} \sum_{i=1}^m \left\| \sum_{j=1}^m w_{ij} \left(\theta_j^t - \theta_{\text{av}}^t \right) \right\|^2} \\
&= m^{-1/2} \left\| (W \otimes I_d) \left(\boldsymbol{\theta}^t - 1_m \otimes \theta_{\text{av}}^t \right) \right\| \\
&= m^{-1/2} \left\| \left(\left(W - \frac{1}{m} 1_m 1_m^\top \right) \otimes I_d \right) \left(\boldsymbol{\theta}^t - 1_m \otimes \theta_{\text{av}}^t \right) \right\| \\
&\overset{(19)}{\leq} \rho m^{-1/2} \|\boldsymbol{\theta}^t - 1_m \otimes \theta_{\text{av}}^t\| \quad (105)
\end{aligned}$$

and

$$\begin{aligned}
&\texttt{Term II} \\
&\leq \frac{\gamma}{m} \sum_{i=1}^m \left\| \sum_{j=1}^m w_{ij} \left(\nabla f_j(\theta_j^t) - \nabla f_j(\theta_{\text{av}}^t) \right) \right\| \\
&\quad + \frac{\gamma}{m} \sum_{i=1}^m \left\| \sum_{j=1}^m w_{ij} \left(\nabla f_j(\theta_{\text{av}}^t) - \nabla F(\theta_{\text{av}}^t) \right) \right\| \\
&\leq \gamma m^{-1/2} \left\| (W \otimes I_d)[\nabla f(\boldsymbol{\theta}^t) - \nabla f(1_m \otimes \theta_{\text{av}}^t)] \right\| \\
&\quad + \gamma m^{-1/2} \left\| (W \otimes I_d)[\nabla f(1_m \otimes \theta_{\text{av}}^t) - 1_m \otimes \nabla F(\theta_{\text{av}}^t)] \right\| \\
&\overset{(b)}{=} \gamma m^{-1/2} \left\| (W \otimes I_d)[\nabla f(\boldsymbol{\theta}^t) - \nabla f(1_m \otimes \theta_{\text{av}}^t)] \right\| \\
&\quad + \gamma m^{-1/2} \left\| \left(\left(W - \frac{1}{m} 1_m 1_m^\top \right) \otimes I_d \right) \left(\nabla f(1_m \otimes \theta_{\text{av}}^t) \right. \right. \\
&\qquad \left. \left. - 1_m \otimes \nabla F(\theta_{\text{av}}^t) \right) \right\| \\
&\overset{(19)}{\leq} \gamma m^{-1/2} \left\| (W \otimes I_d)[\nabla f(\boldsymbol{\theta}^t) - \nabla f(1_m \otimes \theta_{\text{av}}^t)] \right\| \\
&\quad + \rho\gamma m^{-1/2} \left\| \nabla f(1_m \otimes \theta_{\text{av}}^t) - 1_m \otimes \nabla F(\theta_{\text{av}}^t) \right\| \\
&\leq \gamma m^{-1/2} \left\| \nabla f(\boldsymbol{\theta}^t) - \nabla f(1_m \otimes \theta_{\text{av}}^t) \right\| \\
&\quad + \rho\gamma m^{-1/2} \left\| \nabla f(1_m \otimes \theta_{\text{av}}^t) - \nabla f(1_m \otimes \widehat{\theta}) \right\| \\
&\quad + \rho\gamma m^{-1/2} \left\| \nabla f(1_m \otimes \widehat{\theta}) - 1_m \otimes \nabla F(\widehat{\theta}) \right\| \\
&\quad + \rho\gamma m^{-1/2} \left\| 1_m \otimes \nabla F(\widehat{\theta}) - 1_m \otimes \nabla F(\theta_{\text{av}}^t) \right\| \\
&\overset{(91)}{=} \gamma m^{-1/2} \sqrt{\sum_{i=1}^m \left\| \frac{X_i^\top X_i}{n} (\theta_i^t - \theta_{\text{av}}^t) \right\|^2} \\
&\quad + \rho\gamma m^{-1/2} \sqrt{\sum_{i=1}^m \left\| \frac{X_i^\top X_i}{n} (\widehat{\theta} - \theta_{\text{av}}^t) \right\|^2} \\
&\quad + \rho\gamma m^{-1/2} \left\| \nabla f(1_m \otimes \widehat{\theta}) - 1_m \otimes \nabla F(\widehat{\theta}) \right\| \\
&\quad + \rho\gamma \left\| \frac{1}{N} \mathbf{X}^\top \mathbf{X} (\widehat{\theta} - \theta_{\text{av}}^t) \right\|, \quad (106)
\end{aligned}$$

where (b) follows from

$$\left(\frac{1}{m} 1_m 1_m^\top \otimes I_d \right) \nabla f(1_m \otimes \theta_{\text{av}}^t) = 1_m \otimes \nabla F(\theta_{\text{av}}^t).$$

Substituting (10), (101), (103) into (106), we have

$$\begin{aligned}&\texttt{Term II}\\ &\overset{(92)}{\leq} \gamma m^{-1/2} L_{\max} \|\boldsymbol{\theta}_{\perp}^{t}\| \\ &\quad + \rho\gamma \left(L_{\max} + \frac{L}{2} + \frac{\tau_L d}{2}\right) \left(\left\|\widehat{\theta} - \theta_{\text{av}}^{t}\right\| + \left\|\widehat{\theta} - \theta^{*}\right\|\right) \\ &\quad + \rho\gamma d^{1/2} \left(\frac{\max_{i\in[m]} \|X_i^{\top} w_i\|_{\infty}}{n} + \frac{\|\mathbf{X}^{\top}\mathbf{w}\|_{\infty}}{N}\right).\end{aligned} \tag{107}$$

The proof is completed combining the upper bounds of `Term I` and `Term II` as in (105) and (107), respectively. □

D. Proof of Corollary 2.1.

We begin showing that, under (25),

$$r_{\rho} \leq r, \tag{108}$$

where r_{ρ} and r are defined in (20) and (24), respectively. Define

$$\begin{aligned}\delta \triangleq &\sqrt{\frac{1 - (2\kappa)^{-1} + 8s(2\tau_L + \tau_{\mu})/L}{1 - 24s\tau_L/L}} \\ &- \sqrt{\frac{1 - \kappa^{-1} + 8s(2\tau_L + \tau_{\mu})/L}{1 - 16s\tau_L/L}}.\end{aligned}$$

Since

$$r_{\rho} \leq \sqrt{\frac{1 - \frac{1}{\kappa} + \frac{8s(2\tau_L + \tau_{\mu})}{L}}{1 - \frac{16s\tau_L}{L}}} + \rho^{1/2}\left(\frac{4L_{\max}}{L} + \frac{5}{2} + \frac{\tau_L d}{2L}\right),$$

it is sufficient to prove

$$\begin{aligned}&\sqrt{\frac{1 - \frac{1}{\kappa} + \frac{8s(2\tau_L + \tau_{\mu})}{L}}{1 - \frac{16s\tau_L}{L}}} + \rho^{1/2}\left(\frac{4L_{\max}}{L} + \frac{5}{2} + \frac{\tau_L d}{2L}\right) \leq r \\ \Leftrightarrow \quad &\rho^{1/2}\left(\frac{4L_{\max}}{L} + \frac{5}{2} + \frac{\tau_L d}{2L}\right) \leq \delta \\ \Leftrightarrow \quad &\rho \leq \left(\frac{2L\delta}{8L_{\max} + 5L + \tau_L d}\right)^2.\end{aligned} \tag{109}$$

To this end, we proceed lower bounding δ as

$$\begin{aligned}\delta \;&\geq \sqrt{\frac{1 - (2\kappa)^{-1} + 8s(2\tau_L + \tau_{\mu})/L}{1 - 16s\tau_L/L}} \\ &\quad - \sqrt{\frac{1 - \kappa^{-1} + 8s(2\tau_L + \tau_{\mu})/L}{1 - 16s\tau_L/L}} \\ &\overset{(a)}{\geq} \frac{1}{\sqrt{1 + 8s(2\tau_L + \tau_{\mu})/L}} \cdot \frac{1}{4\kappa} \\ &\overset{(b)}{\geq} \frac{1}{\sqrt{1 + \frac{1}{2\kappa}}} \cdot \frac{1}{4\kappa} \\ &= \frac{1}{\sqrt{16\kappa^2 + 8\kappa}},\end{aligned} \tag{110}$$

where in (a) we dropped the negative terms $-(2\kappa)^{-1}, -\kappa^{-1}$, and $-16s\tau_L/L$; and in (b) we used $\mu > 80s\tau_L + 16s\tau_{\mu}$.

Combining (110) with (109), we conclude that (25) is sufficient for (109):

$$\begin{aligned}&\left(\frac{2L\delta}{8L_{\max} + 5L + \tau_L d}\right)^2 \\ &\overset{(110)}{\geq} \frac{L^2}{2(2\kappa^2 + \kappa)\left(8L_{\max} + 5L + \tau_L d\right)^2} \\ &\overset{\kappa>1}{\geq} \frac{L^2}{6\kappa^2\left(8L_{\max} + 5L + \tau_L d\right)^2} \\ &\overset{(24)}{\geq} \frac{8}{3\kappa^2 g^2(d,m)} \\ &\overset{c_6\in(0,1]}{\geq} \frac{c_6}{\kappa^2 g^2(d,m)},\end{aligned}$$

where $c_6 \in (0,1]$ is a free parameter. In addition, using $\mu > 80s\tau_L + 16s\tau_{\mu}$, yields $r < 1$.

It remains to derive the expression of the tolerance error as in the RHS of (26), given that in (23). Using the expression

of Δ_{stat} we have:

$$
\begin{aligned}
&\frac{\Delta_{\text{stat}}}{1-r_\rho} \\
&\overset{(108)}{\leq} \frac{1}{1-r}\cdot 8\sqrt{\frac{2(4\tau_L+\tau_\mu)}{L-16s\tau_L}}\left(\|\widehat{\theta}-\theta^*\|_1+\sqrt{s}\|\widehat{\theta}-\theta^*\|\right) \\
&\quad+\frac{8\rho^{1/2}}{1-r}\cdot\left[\left(\frac{L_{\max}}{L}+\frac{1}{2}+\frac{\tau_L d}{2L}\right)\|\widehat{\theta}-\theta^*\|\right. \\
&\quad\left.+\frac{d^{1/2}}{L}\left(\frac{\max_{i\in[m]}\|X_i^\top w_i\|_\infty}{n}+\frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N}\right)\right] \\
&\overset{R\leq\|\theta^*\|_1}{\leq} \frac{1}{1-r}\cdot 24\sqrt{\frac{2s(4\tau_L+\tau_\mu)}{L-16s\tau_L}}\|\widehat{\theta}-\theta^*\| \\
&\quad+\frac{8\rho^{1/2}}{1-r}\cdot\left[\left(\frac{L_{\max}}{L}+\frac{1}{2}+\frac{\tau_L d}{2L}\right)\|\widehat{\theta}-\theta^*\|\right. \\
&\quad\left.+\frac{d^{1/2}}{L}\left(\frac{\max_{i\in[m]}\|X_i^\top w_i\|_\infty}{n}+\frac{\|\mathbf{X}^\top\mathbf{w}\|_\infty}{N}\right)\right] \\
&= \frac{1}{1-r}\cdot 24\sqrt{\frac{2s(4\tau_L+\tau_\mu)}{L-16s\tau_L}}\|\widehat{\theta}-\theta^*\| \\
&\quad+\frac{8\rho^{1/2}}{1-r}\left(\frac{L_{\max}}{L}+\frac{1}{2}+\frac{\tau_L d}{2L}\right)\|\widehat{\theta}-\theta^*\| \\
&\quad+\frac{8\rho^{1/2}d^{1/2}m}{(1-r)\kappa s^{1/2}}\left(\frac{\max_{i\in[m]}s^{1/2}\|X_i^\top w_i\|_\infty}{\mu N}\right. \\
&\quad\left.+\frac{1}{m}\frac{s^{1/2}\|\mathbf{X}^\top\mathbf{w}\|_\infty}{\mu N}\right) \\
&\leq \frac{1}{1-r}\cdot 24\sqrt{\frac{2s(4\tau_L+\tau_\mu)}{L-16s\tau_L}}\|\widehat{\theta}-\theta^*\| \\
&\quad+\frac{8\rho^{1/2}}{1-r}\left(\frac{L_{\max}}{L}+\frac{1}{2}+\frac{\tau_L d}{2L}\right)\|\widehat{\theta}-\theta^*\| \\
&\quad+\frac{8\rho^{1/2}d^{1/2}m}{(1-r)\kappa s^{1/2}}\left(\frac{\max_{i\in[m]}s^{1/2}\|X_i^\top w_i\|_\infty}{\mu N}\right. \\
&\quad\left.+\frac{s^{1/2}\|\mathbf{X}^\top\mathbf{w}\|_\infty}{\mu N}\right) \\
&\overset{(24)}{\leq} \frac{1}{1-r}\cdot 24\sqrt{\frac{2s(4\tau_L+\tau_\mu)}{L-16s\tau_L}}\|\widehat{\theta}-\theta^*\| \\
&\quad+\frac{\rho^{1/2}g(d,m)}{1-r}\left(\|\widehat{\theta}-\theta^*\|+\frac{s^{1/2}\|\mathbf{X}^\top\mathbf{w}\|_\infty}{\mu N}\right. \\
&\quad\left.+\sqrt{\frac{\log d}{\log md}}\cdot\frac{\max_{i\in[m]}s^{1/2}\|X_i^\top w_i\|_\infty}{\mu N}\right).
\end{aligned}
$$

This completes the proof.

ACKNOWLEDGMENT

The work of Ji, Scutari, and Sun has been supported bythe Office of Naval Research, under the grant N. N00014-21-1-2673.

REFERENCES

[1] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34(1):61–88, August 2011.

[2] A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, April 2012.

[3] W. Auzinger and J. Melenk. Iterative solution of large linear systems. *Lecture notes, TU Wien*, 2011.

[4] Y. Bao and W. Xiong. One-round communication efficient distributed m-estimation. *International Conference on Artificial Intelligence and Statistics*, 130:46–54, April 2021.

[5] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, June 2008.

[6] H. Battey, J. Fan, H. Liu, J. Lu, and Z. Zhu. Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3):1352, May 2018.

[7] Raphael Berthier, Francis Bach, and Pierre Gaillard. Accelerated gossip in networks of given dimension using jacobi polynomial iterations. *SIAM J. on Mathematics of Data Science*, 1:24–47, 2020.

[8] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, May 2008.

[9] D. Chafaï, O. Guédon, G. Lecué, and A/ Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*, volume 37. Société Mathématique de France France, July 2012.

[10] C. Chang and C. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology*, 2(3):1–27, April 2011.

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

[12] 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.

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

[14] M. Genzel and C. Kipp. Generic error bounds for the generalized lasso with sub-exponential data. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2):1–55, 2022.

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

[16] 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.

[17] 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.

[18] 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.

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

[20] Y. Ji, G. Scutari, Y. Sun, and H. Honnappa. Distributed sparse regression via penalization. *arXiv preprint: arXiv:2111.06530*, November 2021.

[21] M. I. Jordan, J. D. Lee, and Y. Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114:668–681, November 2018.

[22] J. D. Lee, Q. Liu, Y. Sun, and J. E. Taylor. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144, January 2017.

[23] X. Lin, S. Boyd, and S. J. Kim. Distributed average consensus with least-mean-square deviation. *Journal of Parallel and Distributed Computing*, 67:33–46, January 2007.

[24] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in neural information processing systems*, 24, 2011.

[25] 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.

[26] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, February 2008.

[27] 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.

[28] 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.

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

[30] 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.

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

[32] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(19):2241–2259, August 2010.

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

[34] K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *International Conference on Machine Learning*, pages 3027–3036. PMLR, May 2017.

[35] 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, April 2015.

[36] 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 2015.

[37] 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.

[38] V. Sivakumar, A. Banerjee, and P. K. Ravikumar. Beyond sub-gaussian measurements: High-dimensional structured estimation with sub-exponential designs. *Advances in neural information processing systems*, 28, 2015.

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

[40] Y. Sun, G. Scutari, and A. Daneshmand. Distributed optimization based on gradient-tracking revisited: enhancing convergence rate via surrogation. *SIAM Journal on Optimization*, 32(2):354–385, June 2022.

[41] S. Y. Tu and A. H. Sayed. Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks. *IEEE Transactions on Signal Processing*, 60(12):6217–6234, May 2012.

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

[43] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications In Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[44] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[45] J. Wang, M. Kolar, N. Srebro, and T. Zhang. Efficient distributed learning with sparsity. *International Conference on Machine Learning*, 70:3636–3645, August 2017.

[46] J. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

[47] J. Xu, Y. Tian, Y. Sun, and G. Scutari. Distributed algorithms for composite optimization: Unified framework and convergence analysis. *IEEE Transactions on Signal Processing*, 69:3555–3570, May 2021.

[48] 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.

[49] 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.

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

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

Yao Ji Yao Ji (Student Member, IEEE) received the B.Sc., M.Sc., in School of Mathematic Science from Beijing Normal University in 2016 and 2019 respectively. She is currently pursuing the Ph.D. degree in Industrial Engineering, Purdue, advised by Gesualdo Scutari and Harsha Honnappa. Her research interests include statistical learning over networks and distributed optimization theory.

Gesualdo Scutari Gesualdo Scutari (Fellow, IEEE) received the electrical engineering and Ph.D. degrees (Hons.) from the University of Rome "La Sapienz" Rome, Italy, in 2001 and 2005, respectively. He is a Professor with the School of Industrial Engineering, Purdue University, West Lafayette, IN, USA, and he is a Purdue Faculty Scholar. His research interests include optimization, equilibrium programming, and their applications to signal processing and machine learning. He was a recipient of the 2013 NSF CAREER Award, the 2015 IEEE Signal Processing Society Young Author Best Paper Award, and the 2020 IEEE Signal Processing Society Best Paper Award. He serves as an IEEE Signal Processing Distinguish Lecturer (2023-2024). He served on the editorial broad of several IEEE journals and he is currently an Associate Editor of SIAM Journal on Optimization.

Ying Sun Ying Sun (Member, IEEE) received the B.E. degree in electronic information from the Huazhong University of Science and Technology, Wuhan, China, in 2011, and the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology in 2016. She was a postdoc researcher with the School of Industrial Engineering, Purdue University from 2016 to 2020. Currently, she is an assistant professor in the Department of Electrical Engineering at The Pennsylvania State University. Her research interests include statistical signal processing, optimization algorithms and machine learning. She is a co-recipient of a student best paper at IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) 2017, and a recipient of the 2020 IEEE Signal Processing Society Young Author Best Paper Award.

Harsha Honnappa Harsha Honnappa received Ph.D. degree in electrical engineering from University of Souithern California. He is an associate Professor in the School of Industrial Engineering at Purdue University. His research interests lie broadly in applied probability, stochastic optimization, simulation methodology and machine learning.