Acceleration in Distributed Sparse Regression

Marie Maros
School of Industrial Engineering
Purdue University
mmaros@purdue.edu

Gesualdo Scutari
School of Industrial Engineering
Purdue University
gscutari@purdue.edu

Abstract

We study acceleration for distributed sparse regression in *high-dimensions*, which allows the parameter size to exceed and grow faster than the sample size. When applicable, existing distributed algorithms employing acceleration perform poorly in this setting, theoretically and numerically. We propose an accelerated distributed algorithm suitable for high-dimensions. The method couples a suitable instance of accelerated Nesterov's proximal gradient with consensus and gradient-tracking mechanisms, aiming at estimating locally the gradient of the empirical loss while enforcing agreement on the local estimates. Under standard assumptions on the statistical model and tuning parameters, the proposed method is proved to globally converge at *linear* rate to an estimate that is within the *statistical precision* of the model. The iteration (i.e., gradient oracle) complexity scales as $\mathcal{O}(\sqrt{\kappa})$, while the communications per iteration are at most $\widetilde{\mathcal{O}}(\log m/(1-\rho))$, where κ is the restricted condition number of the empirical loss, m is the number of agents, and $\rho \in [0,1)$ measures the network connectivity. As by-product of our design, we also report an accelerated method for high-dimensional estimation over master-worker architectures, which is of independent interest and compares favorably with existing works.

1 Introduction

Consider a multiagent system composed of m machines, connected throughout a fixed, undirected graph–we will refer to such networks as *mesh* networks, in contrast to centralized topologies such as master-worker architectures, where there exists a node connected to all the others. The ultimate goal is to solve the stochastic (quadratic) optimization (learning) problem

$$\theta^\star \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left\{ \bar{f}(\theta) \triangleq \frac{1}{2}\mathbb{E}_{(x,y)\sim\mathbb{P}}[(x^\top\theta - y)^2] \right\}. \tag{1}$$

Each of the m agents has access to n i.i.d observations, drawn from the unknown, common distribution $\mathbb{P}$ on $\mathcal{Z} \subseteq \mathbb{R}^p$, and collected in the set $\mathcal{S}_i$. The associated local empirical estimate of $\bar{f}$ reads

$$f_i(\theta) \triangleq \frac{1}{2n}\sum_{j\in\mathcal{S}_i}(x_j^\top\theta - y_j)^2, \tag{2}$$

where $x_j \in \mathbb{R}^d$ is the vector of predictors and $y_j \in \mathbb{R}$ the associated response. The overall Empirical Risk Minimization (ERM) over the network is then

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\Omega:\mathcal{R}(\theta)\le R} \left\{ f(\theta) \triangleq \frac{1}{m}\sum_{i=1}^m f_i(\theta) \right\}, \tag{3}$$

where $\mathcal{R}$ is a (convex) regularizer (with $R > 0$) and Ω is some (convex) subset of $\mathbb{R}^d$.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

We study estimators under *high-dimensional scaling*, meaning that the ambient dimension d exceeds (and grows faster than) the total sample size $N = m \cdot n$. The desired parameter $\theta^\star$ is assumed to lie in a smaller subset of Ω or is well approximated by a member of it–such structural constraints are enforced by the regularizer $\mathcal{R}$ in (3). Examples include sparse linear models with ℓ_1 regularizers, low-rank matrix recovery via nuclear norm, and matrix regression with soft-rank constraints [1].

The goal is to compute an estimate of $\theta^\star$ by solving the ERM (3) that is within the statistical error of the model $\|\widehat{\theta} - \theta^\star\|$. We target first-order methods. Despite the vast literature of distributed optimization (see Sec. 1.1), only very recently such guarantees have been established on mesh networks for the LASSO problem in the projected form [26] [as (3)] and in the Lagrangian form [11]. The benchmark is the scheme in [26], termed NetLASSO: under restricted notions of strong convexity and smoothness of f (see Sec. 2)–which hold with high probability for a variety of data generation models–as well as $s \log d/N = o(1)$–a condition necessary for statistical consistency in s-spare linear regression–the iterates generated by NetLASSO reach an ε-neighborhood of a statistically optimal solution in

$$\widetilde{\mathcal{O}}\left(\kappa \, \frac{\log m}{1-\rho} \, \log 1/\varepsilon\right) \quad \text{number of communications,} \tag{4}$$

where $\kappa \geq 1$ is the condition number of f in (3) restricted to sparse directions, and $\rho \in [0, 1)$ measures the connectivity of the network–the smaller the more connected the graph ($\widetilde{\mathcal{O}}$ hides log-factors on optimization parameters). This shows linear rate up to statistical precision, whose dependence on κ matches that of the centralized Projected Gradient Algorithm (PGA) [1].

Since the bottleneck in distributed systems is often represented by the cost of communications (in comparison with local computations) [6, 15], a key question is whether the communication complexity (4) is improvable while still preserving the same *statistical* guarantees and local computational complexity (first-order methods). This is particularly sensible when dealing with ill-conditioned population losses, resulting in very large values of κ. To improve the dependency on κ, a natural approach is employing some form of acceleration. Despite the vast literature of accelerated methods in distributed optimization, this remains an open problem in *high-dimensions*, as elaborated next.

1.1 Distributed acceleration in high-dimensions: Challenges and open problems

Accelerating first-order methods over mesh networks (undirected graphs) has received attention in the last few years; distributed schemes in the primal domain include [22, 32, 33, 14, 12, 24] while [25, 28, 13] are accelerated dual or penalty-based methods. They are suitable to solve distributed, *unconstrained, smooth* optimization problems, with the exception of [33], which decentralizes the accelerated proximal-gradient method, and thus is applicable also to the constrained ERM (3).

Albeit different, when applied to the minimization of a L_f-smooth and μ_f-strongly convex loss $f = 1/m \sum_{i=1}^m f_i$, with each f_i being L_i-smooth and μ_i-strongly convex, they all provably achieve linear convergence to the minimizer of f, some [25, 13, 14, 12, 28] with communication complexity scaling with $\sqrt{\kappa_\ell}$ ($\kappa_\ell = \max_i L_i / \min_i \mu_i$ is the local condition number) while others [32, 33, 24] with $\sqrt{\kappa}_f$ ($\kappa_f = L/\mu$ is the condition number of f). Note that in general $\kappa_f < \kappa_\ell$; hence, the latter group is preferable to the former. For weakly convex losses f_i's, convergence is certified at *sublinear* rate, the optimality gap on the objective value vanishes as $\mathcal{O}(1/t^2)$, where t is the iteration index.

When applied to the constrained ERM (3) in high-dimensions, these convergence results are unsatisfactory. **First**, they are of pure optimization type, and do not provide any statistical guarantee–computing below the statistical noise would result in a waste of resources. **Second**, for fixed d and N, with $d > N$ (high-dimensions), they would certify sublinear convergence rates, as the empirical loss f is convex but not strongly convex *globally*–the $d \times d$ Hessian matrix $\nabla^2 f$ has rank at most N. This provides pessimistic, non-informative predictions, which contrast with (4), proving instead that linear rates up to the statistical precision are achievable by nonaccelerated distributed methods. **Third**, under the scaling $d/N \to \infty$ (d growing faster than N)–the typical regime in high-dimensions [29, 1]–convergence analyses in the aforementioned papers break down; they all require *global smoothness* of the local f_i's and global f losses, a property that does not hold when $d/N \to \infty$. If fact, for commonly used designs of predictors x_i's, the Lipschitz constant of ∇f grows indefinitely with d/N [29]. **Fourth**, if nevertheless simulated, the only distributed accelerated scheme applicable to the constrained ERM (3), that is DPAG [33], performs poorly in high-dimensions, as Fig. 1 shows. In the figure we plot the average

estimation error, defined as $(1/m)\sum_{i=1}^{m}\|\theta_i^t-\theta^\star\|^2$, versus the iterations indexed by t, where θ_i^t is the estimate of $\theta^\star$ from agent i at iteration t generated by DPAG solving a projected LASSO over a mesh network (a special instance of (3) with $\mathcal{R}(\theta)=\|\theta\|_1$), and $\theta^\star$ is the population s-sparse minimizer in (1). We contrast DPAG with the new distributed accelerated method proposed in this paper. As benchmark, we also plot the estimation error generated by NetLASSO (non-accelerated method) [26]. Despite the good performance of DPAG in low-dimensions (i.e., $N > d$) [33], the figure shows that acceleration as in DPAG is no longer effective in high-dimensions ($d > N$); DPAG achieves estimates of $\theta^\star$ than are worse than those achievable by non accelerated schemes, yielding estimation errors that are much larger than centralized statistical ones. Further, in our experiment we observed that this gap grows with d/N. In contrast, the proposed distributed acceleration exhibits linear convergence up to *centralized* statistical precision, at a rate faster than the nonaccelerated NetLASSO.

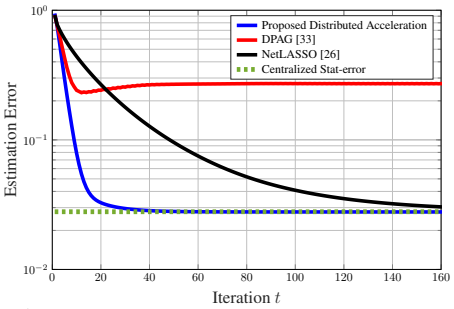

Figure 1: Projected LASSO: Average estimation error vs. iterations. Existing distributed accelerations (DPAG) may not work in high-dimension! The proposed accelerated method outperforms nonaccelerated distributed schemes (NetLASSO) while achieving centralized statistical accuracy.

We conjecture that the unsatisfactory performance of DPAG [33] is due to the following facts: **(i)** the empirical loss of LASSO is strongly convex only over sparse vectors (see Sec. 2). Hence, fast convergence is expected only when traveling such a curved region of the landscape; **(ii)** the scheme [33] can be interpreted as a decentralization of the accelerated proximal gradient based on the Nesterov's first method [21]; as such, it does not implement any mechanism enforcing the trajectory at which gradients are evaluated to be (approximately) sparse. This suggests that acceleration in the distributed high-dimensional setting needs to be rethought–new designs and convergence analyses are needed. This is what this paper is about.

1.2 Main contributions

Algorithm design: We propose a distributed algorithm for high-dimensional estimation achieving acceleration provably. The method decentralizes Nesterov's accelerated proximal gradient (form III) [20] via consensus dynamics and a gradient tracking mechanism [7, 30, 19], the former enforcing an agreement on the agents' estimates and the latter aiming at tracking locally the gradient of f.

The rationale to hinge on Nesterov's third scheme rather than form I, as used instead so far in all accelerated distributed algorithms, is to enforce feasibility (with respect to the norm ball constraint $\mathcal{R}(\theta)\le R$) of the momentum-sequence generated by the algorithm, the one over which gradients are evaluated. For instance, when solving a constrained LASSO (i.e., $\mathcal{R}(\theta)=\|\theta\|_1$), this yields sequences that are approximately sparse, thus traveling the landscape of an "almost" strongly convex and smooth loss. In fact, our analysis shows that this is the key enabler to retain fast, linear convergence to statistically optimal solutions in high-dimension, where f is no longer strongly convex globally.

Statistical-computational guarantees: Under standard notion of restricted strong convexity and smoothness of the empirical loss, which hold with high probability for a variety of statistical models, and proper algorithm tuning, the iterates generated by the proposed scheme converge at *linear* rate to a limit point that is within a fixed tolerance from the unknown $\theta^\star$. When customized to specific statistical models–including sparse linear models with ℓ_1 regularization, low-rank matrix recovery with nuclear norm, and matrix completion–the tolerance becomes of the order of the statistical error $\|\hat\theta-\theta^\star\|$. Specifically, for such models and arbitrary network connectivity ρ, the iterates generated by the algorithm enter an ε-neighborhood of a *statistically optimal* estimate of $\theta^\star$ after

$$\mathcal{O}\left(\sqrt{\kappa}\,\log 1/\varepsilon\right) \quad \text{and} \quad \widetilde{\mathcal{O}}\left(\sqrt{\kappa}\,\frac{\log m}{1-\rho}\,\log 1/\varepsilon\right) \tag{5}$$

numbers of iterations (gradient evaluations) and communications, respectively, where $\tilde{\mathcal{O}}$ hides log-factors on optimization parameters (recall κ is the restricted condition number). This improves on the communication complexity (4) of NetLASSO [26] applied to ℓ_1-constrained LASSO, showing the more favorable square-root scaling on κ, typical of accelerated methods in low-dimension. Furthermore, the rates in (5) are showed to be invariant under high-dimensional scaling $d/N\to\infty$, as long as the statistical error $\|\hat\theta-\theta^\star\|$ remains constant.

A special case: A by-product of the proposed method is its customization to master-worker architectures, which is of independent interest and compares favorable with existing accelerated centralized designs [16, 31], proposed for ℓ_1-least squares (in Lagrangian form); they are *double-loop* methods, combining the accelerated proximal gradient method [27, 5, 20] with homotopy continuation.

2 Setup and Background

This section states the main assumptions on the ERM (3) and the network underlying our derivations.

• **RSC & RSM:** We anticipated that, in high-dimension, f in (3) is not strongly convex uniformly (as $d > N$). However, strong convexity and smoothness hold along a restricted set of directions, for a variety of statistical estimations [29], which is enough to employ fast convergence and favorable statistical guarantees of the PGA [1]. Here we postulate the same properties for the landscape of f.

Assumption 1 (Global RSC & RSM [1])**.** *The loss function f in (3) satisfies*

(i) *Restricted Strong Convexity (RSC) with respect to $\mathcal{R}$ with (positive) parameters $(\mu/2, \tau_\mu)$ over Ω:*

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{\mu}{4}\|x - y\|_2^2 - \frac{\tau_\mu}{2}\mathcal{R}^2(x - y), \quad \forall x, y \in \Omega; \tag{6}$$

(ii) *Restricted Smoothness (RSM) with respect to $\mathcal{R}$ with (positive) parameters $(2L, \tau_L)$ over Ω:*

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq L\|x - y\|_2^2 + \frac{\tau_L}{2}\mathcal{R}^2(x - y), \quad \forall x, y \in \Omega. \tag{7}$$

The restricted condition number is defined as $\kappa \triangleq L/\mu$ and it is assumed $\kappa \geq 1$.

In the distributed setting, the RSM property is required to hold also locally.

Assumption 2 (local RSM)**.** *Each loss function f_i satisfies local RSM with respect to $\mathcal{R}$ with (positive) parameters $(2\gamma_\ell, \tau_\ell)$ over Ω:*

$$f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle \leq \gamma_\ell\|x - y\|_2^2 + \frac{\tau_\ell}{2}\mathcal{R}^2(x - y), \quad \forall x, y \in \Omega; \tag{8}$$

• **Decomposable regularizer $\mathcal{R}$:** This notion plays a key role for statistical consistency of $\hat{\theta}$ in high-dimension [1]. Decomposability is defined with respect to two subspaces of Ω, namely: $\mathcal{M}$, known as the *parameter* subspace, capturing constraints from the model (e.g., vectors with a particular support or a subspace of low-rank matrices); and $\bar{\mathcal{M}}^\perp$, referred as *perturbation* subspace, representing the deviations from the model subspace $\mathcal{M}$. We also require Lipschitz continuity of $\mathcal{R}$ over $\bar{\mathcal{M}}$.

Assumption 3. ***(i)*** *The regularizer $\mathcal{R} : \Omega \to \mathbb{R}$ is a norm, with $\Omega \subseteq \mathbb{R}^d$ convex and $0 \in \Omega$ without loss of generality. Furthermore, given a subspace pair $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$, such that $\mathcal{M} \subseteq \bar{\mathcal{M}}$:*

(ii) *$\mathcal{R}$ is $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$-decomposable, that is, $\mathcal{R}(x + y) = \mathcal{R}(x) + \mathcal{R}(y)$, for all $x \in \mathcal{M}$ and $y \in \bar{\mathcal{M}}^\perp$;*

(iii) *When $\bar{\mathcal{M}} \neq \{0\}$, $\mathcal{R}$ is $\Psi(\bar{\mathcal{M}})$-Lipschitz over $\bar{\mathcal{M}}$ with respect to some norm $\| \bullet \|$: $\Psi(\bar{\mathcal{M}}) \triangleq \sup_{\theta \in \bar{\mathcal{M}} \setminus \{0\}} \mathcal{R}(\theta)/\|\theta\|$. If $\bar{\mathcal{M}} = \{0\}$, we set $\Psi(\{0\}) = 0$.*

Regularizers sharing the above properties over certain pairs $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$ are ubiquitous in M-estimation problems [29]; examples include the ℓ_1 norm, the (overlap) group norm, and the nuclear norm.

• **Statistical models:** RSC and RSM properties can be certified with high probability by a variety of random data generations. Here, we consider the following widely used statistical models, which cover a variety of estimation tasks, such as sparse linear models with ℓ_1 regularization, low-rank matrix recovery with nuclear norm, and matrix regression with soft-rank constraints [17, 1, 29].

Assumption 4. *The random predictors $x_i \in \mathbb{R}^d$ are i.i.d. and fulfill one of the following conditions:*

(i) $x_i \sim \mathcal{N}(0, \Sigma)$ *and are i.i.d., for some $\Sigma \succ 0$;*

(ii) x_i *are $(\tau^2, \Sigma)-$sub-Gaussian and i.i.d., with $\Sigma \succ 0$;*

(iii) $x_i = e_j$ *where $j \sim$ uniform$[1, d]$ and i.i.d.*

Lemma 1. *Under Assumption 4(i), there exist universal constants $c_0, c_1, c_2 > 0$ such that, with probability at least $1 - c_0 \exp(-c_1 N + \log m)$, $N \geq 10$, Assumptions 1 hold with parameters*

$$\mu = \lambda_{\min}(\Sigma), \quad \tau_\mu = c_2 \mathbb{E}[\mathcal{R}^*(x_i)]^2/N \quad \textit{and} \quad L = \lambda_{\max}(\Sigma), \quad \tau_L = c_2 \mathbb{E}[\mathcal{R}^*(x_i)]^2/N,$$

and so does Assumption 2 with parameters $\gamma_\ell = (2m+1)\lambda_{\max}(\Sigma)$ and $\tau_\ell = c_2(\mathbb{E}[\mathcal{R}^*(x_i)]^2)/n$, where $\mathcal{R}^*$ is the dual norm of $\mathcal{R}$. When $\mathcal{R}(\cdot) = \|\cdot\|_1$, $(\mathbb{E}[\mathcal{R}^*(x_i)])^2 \leq 9(\max_j(\Sigma)_{jj}) \log(d)$.

The proof of Lemma 1 follows from that of [1, Prop. 1]. Similar results under Assumptions 4(ii)-(iii) can be found in the supplementary material, see Appendix A.7.

• **On the network:** Agents are connected throughout a communication network, modelled as an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where the vertices $\mathcal{V} = [m] \triangleq \{1, \ldots, m\}$ correspond to the agents and $\mathcal{E}$ is the set of edges of the graph; $(i,j) \in \mathcal{E}$ if and only if there is a communication link between agents i and j. The set of neighbors of each agent i is denoted by $\mathcal{N}_i \triangleq \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\} \cup \{i\}$. In the proposed algorithms, agents will average neighbouring information using weights as below.

Assumption 5 (Gossip matrices)**.** *Define $W = [w_{ij}]_{ij=1}^m$; the following hold:* ***(i)*** *W is $\mathcal{G}$-congruent, that is, $w_{ij} > 0$, if $(i,j) \in \mathcal{E}$; otherwise $w_{ij} = 0$; furthermore, $w_{ii} > 0$, for all $i \in [m]$;* ***(ii)*** *$W = W^\top$ and $W1 = 1$ (stochastic); and* ***(iii)*** *there holds $\rho \triangleq \|W - J\|_2 < 1$, where $J \triangleq 11^\top/m$.*

Assumption 5 is quite standard in the literature of distributed algorithms and is satisfied by several weight matrices; see, e.g., [18]. Note that $\rho < 1$ by construction, as long as $\mathcal{G}$ is connected. Roughly speaking, ρ measures how fast the network mixes information; the smaller ρ, the faster the mixing.

3 Algorithm Design

At the core of our distributed accelerated algorithm there is an instance of Nesterov's acceleration proximal gradient method, which we identified to be suitable for high-dimensional estimation. We begin presenting this method in the centralized setting, along with its statistical guarantees (Sec. 3.1); we then proceed to design and analyze a suitable decentralization to mesh networks (Sec. 3.2).

3.1 Warm-up: A Nesterov's acceleration (form III) for high-dimensional estimation

Consider (3) in the centralized setting (e.g., master-worker systems). Assumption 1 shows that f is strongly convex and smooth only along a restricted set of directions. To enable fast convergence, the algorithm design should produce trajectories (approximately) traveling such portions of the landscape of f. To handle this situation while achieving acceleration we propose to use Nesterov's method written in form III [9, Algorithm 11] equipped with the projected operator to enforce feasibility of all the sequences generated by the algorithm. This ensures that gradients are always evaluated at feasible points of (3), which we prove unlock linear convergence (up to a tolerance) at accelerated rate.

Given feasible θ^0 and v^0, the main iterate $\{\theta^t, v^t, z^t\}$ of the algorithm reads: for any $t = 0, 1, \ldots,$

$$\begin{aligned} z^t &= \frac{1}{\beta+1}\,\theta^t + \frac{\beta}{\beta+1}\,v^t, \\ v^{t+1} &= \Pi_{\Omega \cap \mathcal{R}(v) \leq R}\left((1-\beta)v^t + \beta z^t - \frac{\alpha}{\beta}\nabla f(z^t)\right); \\ \theta^{t+1} &= \beta\, v^{t+1} + (1-\beta)\theta^t. \end{aligned} \tag{9}$$

where $\alpha \in (0,1)$ and $\beta \in (0,1)$ are parameters to tune, and $\Pi_{\Omega \cap \mathcal{R}(v) \leq R}(\bullet)$ denotes the Euclidean projection (of its argument) onto the convex set $\{\theta \in \Omega \,:\, \mathcal{R}(v) \leq R\}$; for specific estimators such as ℓ_1-constrained or Lagrangian LASSO, this projection has a closed form expression [8, 11]. Notice that, under feasibility of θ^0 and v^0, the sequences $\{\theta^t\}$, $\{v^t\}$ and $\{\theta^t\}$ are feasible, for all t.

A wide variety of accelerated methods exists in the centralized setting (see, e.g., [9] for a recent overview), and most variants settle the infeasibility of the z^t sequence using two projection (proximal) operations per iteration (see, e.g., [3]). The algorithm in (9) performs instead one projection per iteration. Variations of the theme can be found in a number of references, including [2, 9, 27, 10]. Existing analyses however are not adequate to capture the behaviour of these methods in high-dimensions, including (9); as discussed in Sec. 1.1 for distributed implementations, they would

provide pessimist predictions and break down under the scaling $d/N \to \infty$. As by-product of our novel convergence analysis in the distributed setting, Theorem 1 below establishes, under RSC and RSM properties (Assumption 1), global convergence of the algorithm in (9) at linear (accelerated) rate to a fixed tolerance. This result is of independent interest, and complements existing accelerated schemes for ℓ_1-least squares estimations (LASSO in Lagrangian form) [16, 31]. These are double-loop methods, combining accelerated proximal gradient with a homotopy continuation mechanism.

Leveraging Assumptions 1, 3, let us introduce the tolerance Δ^2 and the initial optimality gap V^0:

$$\Delta^2 \triangleq \left(2\mathcal{R}\left(\Pi_{\mathcal{M}^\perp}(\theta^\star)\right) + 2\mathcal{R}(\Delta^\star) + \Psi(\bar{\mathcal{M}})\|\Delta^\star\|\right)^2, \; V^0 \triangleq f(\theta^0) - f(\widehat{\theta}) + \frac{\mu}{8}\|v^0 - \widehat{\theta}\|^2, \quad (10)$$

where $\Delta^\star \triangleq \theta^\star - \widehat{\theta}$ is the satistical error, $\Pi_{\mathcal{M}^\perp}$ denotes the Euclidean projection onto the orthogonal complement of the subspace $\mathcal{M}$, and $\Psi(\bar{\mathcal{M}})$ is defined in Assumption 3. Convergence is stated next.

Theorem 1. *Given the ERM* (3)*, suppose (i) f satisfies the RSC/RSM conditions (Assumption 1); (ii) $\mathcal{R}$ is a decomposable regularizer with respect to chosen pair of subspaces $(\mathcal{M}, \bar{\mathcal{M}})$ (Assumption 3); and (iii) $(\mathcal{M}, \bar{\mathcal{M}})$ and RSC/RSM parameters (μ, τ_μ) and (L, τ_L) are such that $\Psi^2(\bar{\mathcal{M}}) \max\{\tau_L, \tau_\mu\} \leq C_0$, for some constant $C_0 > 0$. Let $\{\theta^t\}_{t\geq 1}$ be the sequence generated using Algorithm* (9)*, with tuning $\alpha = 2L$ and $\beta = \sqrt{1/(8\kappa)}$. Then, for any solution $\widehat{\theta}$ of* (3) *for which $\mathcal{R}(\widehat{\theta}) = R$, we have*

$$\|\theta^t - \widehat{\theta}\|^2 \leq \frac{8}{\mu}\left(1 - \sqrt{\frac{1}{16\kappa}}\right)^t V^0 + \mathcal{O}\left(\frac{\tau_\mu + \tau_L}{\mu}\Delta^2\right). \quad (11)$$

This proves linear convergence at the accelerated rate $1 - \sqrt{1/(16\kappa)}$, up to a fixed tolerance. As it will be showed directly in the distributed setting (see Sec. 4), when considering specific statistical models under $\Psi^2(\bar{\mathcal{M}}) \max\{\tau_L, \tau_\mu\} = o(1)$–a condition necessary for statistical consistency [1, 29]–the residual error in (11) becomes *smaller than the statistical precision*, proving thus convergence to statistically optimal solutions. This improves the convergence rate of the PGA [1], with Algorithm (9) gaining acceleration while retaining the same statistical precision and computational complexity.

Finally, we remark that, while Theorem 1 has been stated for quadratic ERM, the same conclusions hold for nonquadratic losses, satisfying RSC/RSM conditions as in Assumption 1.

3.2 Distributed accelerated proximal gradient method for high-dimensional estimation

Equipped with Algorithm (9), we are ready to introduce our accelerated method for distributed high-dimensional estimations in the form (3), as given in Algorithm 1 below, and discussed next.

What prevents (9) from being implemented on mesh networks is the lack of knowledge of ∇f at the agents' sides. Agents' losses f_i's are decoupled by introducing local estimates θ_i, controlled (updated) by the intended agent, along with the associated sequences z_i and v_i, which play the same role of z and v as in (9): first, they are locally updated as the related ones in (9), followed by the consensus step (16), enforcing agreement among the momentum sequences and local estimates. The update of the v-variables, as per (9) would require the knowledge of ∇f from each agent, an information that is not available. Hence, ∇f is replaced by a local estimate y_i [see (13)]; the tracking of ∇f via each y_i is employed by the dynamic consensus mechanism (a.k.a. gradient-tracking) [7, 30, 19] in (14). In fact, it is not difficult to to check that the average of the y_i^t's tracks the average of the local gradients $\nabla f_i(z_i^t)$, i.e.,

$$\frac{1}{m}\sum_{i=1}^m y_i^t = \frac{1}{m}\sum_{i=1}^m \nabla f_i(z_i^t). \quad (12)$$

Assuming that consensus on the z_i's and the y_i's variables is asymptotically achieved, i.e., $\|z_i^t - z_j^t\| \to 0$ and $\|y_i^t - y_j^t\| \to 0$ as $t \to \infty$, for all $i, j = 1, \ldots m$ (a fact that will be proved), (12) would yield the desired gradient tracking property $\|\nabla f(z_i^t) - y_i^t\| \to 0$ as $t \to \infty$, for all $i = 1, \ldots, m$.

4 Statistical-Computational Guarantees

This section studies convergence of Algorithm 1. Our first result (Theorem 2) is of deterministic type: We establish conditions for the estimation error $(1/m)\sum_{i=1}^m \|\theta_i^t - \widehat{\theta}\|^2$ to shrink linearly

Algorithm 1

Data: Set $\theta_i^0 = v_i^0 = z^{-1} = 0$ and $y^{-1} = \nabla f_i(0)$ for all $i \in [m]$; $\alpha \geq 2L$ and $\beta \in (0,1)$;
Iterate $t = 0, 1, \ldots$

[S.1]: Consensus and gradient-tracking: Each agent i performs

$$z_i^t = \sum_{j \in \mathcal{N}_i} w_{ij} \left(\frac{1}{\beta+1}\, \theta_j^t + \frac{\beta}{1+\beta}\, v_j^t \right), \tag{13}$$

$$y_i^t = \sum_{j \in \mathcal{N}_i} w_{ij} \left(y_j^{t-1} + \nabla f_j(z_j^t) - \nabla f_j(z_j^{t-1}) \right); \tag{14}$$

[S.2]: Local Optimization and updates: Each agent i performs

$$v_i^{t+1} = \Pi_{\Omega \cap \mathcal{R}(v) \leq R} \left((1-\beta) v_i^t + \beta z_i^t - \frac{\alpha}{\beta} y_i^t \right); \tag{15}$$

$$\theta_i^{t+1} = \beta\, v_i^{t+1} + (1-\beta)\theta_i^t. \tag{16}$$

up to a tolerance. These conditions hold with high-probability for a variety of statistical models (Assumption 4), proving linear convergence (at accelerated rate) within minimax statistical rates.

Theorem 2. *Consider the ERM (3) over mesh networks, under the same assumptions (i)-(iii) as in Theorem 1; furthermore, (iv) suppose f_i satisfies the local RSM condition (Assumption 2), for all $i \in [m]$. Let $\{(\theta_i^t)_{i\in[m]}\}_{t\geq 1}$ be the sequence generated by Algorithm 1, with tuning $\alpha = 2L$ and $\beta = \sqrt{1/(8\kappa)}$, and weight matrix W satisfying Assumption 5, with ρ such that*

$$\rho \leq C_1 \min \left\{ \sqrt{\kappa^{-1}}, \frac{L\,\kappa^{-1}}{m^{3/2}\left(\gamma_\ell + \Psi^2(\bar{\mathcal{M}})(\tau_\ell + \tau_L)\right)} \right\}, \tag{17}$$

where $C_1 > 0$ is a universal constant (see Theorem 3 in Appendix A.4).

Then, for any solution $\widehat{\theta}$ of (3) for which $\mathcal{R}(\widehat{\theta}) = R$, we have: for some $C_2 > 0$,

$$\frac{1}{m}\sum_{i=1}^m \|\theta_i^t - \widehat{\theta}\|^2 \leq \frac{8}{\mu}\left(1 - \sqrt{\frac{1}{32\kappa}}\right)^t L^0 + \mathcal{O}\Big(\underbrace{\frac{\tau_\mu + \tau_L}{\mu}\,\Delta^2}_{\textit{centralized precision}} + \underbrace{\frac{\tau_\ell\sqrt{m}\rho}{\mu}\,\Delta^2}_{\textit{network error}} \Big). \tag{18}$$

where $L^0 \in (0, \infty)$ captures the initial optimality gap and dependence on optimization and RSC/RSM parameters (its expression can be found in Theorem 3, Appendix A.4).

This shows linear convergence at rate $1 - \sqrt{1/(32\kappa)}$ up to a tolerance. The first term of this tolerance matches that achieved in the centralized setting (see (11)) while the second term is an extra error due to the decentralization of the estimation process–it vanishes when $\rho = 0$ (fully connected network). In the next section, we will show that for the statistical models captured by Assumption (i), the overall tolerance can be made of a higher-order of the centralized statistical error. The condition on ρ in (17) calls for a sufficiently connected network. When it is not satisfied by the given network and gossip matrix W, it can be enforced running multiple rounds of communications, affecting the total communication complexity by log-factor terms. More specifically, given W with associated $\rho = \|W - J\|_2$ larger than what required in (17), one can build another matrix $\bar{W} = W^K$ and choose the positive integer number K so that $\bar{\rho} \triangleq \|\bar{W} - J\|_2 = \rho^K$ satisfies (17); it is sufficient to set

$$K = \left\lceil \frac{\log\left(C_1 \kappa m^{2.5}(1 + C_2/L)\right)}{1-\rho} \right\rceil.$$

The use of $\bar{W}$ instead of W in Algorithm 1 corresponds to employing K rounds of communications (consensus and tracking updates) per iteration t, each one using the weights W.

4.1 Statistical Guarantees

We apply now Theorem 2 to specific statistical models, establishing convergence and statistical guarantees with overwhelming probability. Corollary 1 studies the case of sparse linear regression,

under both strong and weak sparsity. Corollary 2 addresses the low-rank matrix recovery problem. Sub-Gaussian linear regression (Corollary 3) and matrix completion (Corollary 4) are discussed in the supporting material (Appendix A.3).

• **Sparse regression:** Suppose $\theta^\star$ in (1) is sparse, i.e., $\|\theta^\star\|_q \leq R_q$, with $q \in [0,1]$ ($q = 0$ corresponds to hard-sparsity, while $q > 0$ models weak sparsity). The regression model at agent i is: $y_j = x_j^\top \theta^\star + w_j$, $j \in \mathcal{S}_i$, where $|\mathcal{S}_i| = n$, $w_j \sim \mathcal{N}(0, \sigma^2)$, and x_j satisfies Assumption 4(i). Define

$$\chi_N \triangleq \frac{\max_i([\Sigma]_{ii})}{\lambda_{\min}(\Sigma)} R_q \left(\frac{\log d}{N}\right)^{1-\frac{q}{2}}, \quad q \in [0,1] \text{ and } R_0 = s.$$

Corollary 1. *Consider the ERM 3 solving the linear regression problem above, with $\mathcal{R}(\cdot) = \|\cdot\|_1$, $\|\theta^\star\|_1 \leq R = \|\widehat{\theta}\|_1$, and $\Omega \equiv \mathbb{R}^d$; let $N = \Omega((\eta_\Sigma R_q)^{1/(1-q/2)} \log d)$. Let $\{(\theta_i^t)_{i=1}^m\}_{t\geq 0}$ be the sequence generated by Algorithm 1 with tuning as in Theorem 2, where* (17) *becomes*

$$\rho \leq C_1 \min\left\{\sqrt{\kappa^{-1}}, \frac{\kappa^{-1}}{m^{2.5}\big(1 + C_2/L\big)}\right\}, \quad \textit{for some } C_1,\, C_2 > 0. \tag{19}$$

Then, with probability at least $1 - c_0 \exp\left(-c_1 N + \log(m)\right)$, for some $c_0,\, c_1 > 0$, it holds

$$\frac{1}{m}\sum_{i=1}^m \|\theta_i^t - \widehat{\theta}\|^2 \leq \frac{8}{\lambda_{\min}(\Sigma)}\left(1 - \sqrt{\frac{1}{32\kappa}}\right)^t L^0 + \mathcal{O}\left(\chi_N\left(R_q\left(\frac{\log d}{N}\right)^{1-\frac{q}{2}} + \|\Delta^\star\|^2\right)\right), \tag{20}$$

where $L^0 > 0$ depends on the the initial optimality gap and statistical parameters (see Appendix A.6)

A few comments are in order. **(i)** Linear convergence at rate $1 - \mathcal{O}(\kappa^{-1/2})$ is certified, up to a tolerance that, under the sample-size $N = \Omega((\eta_\Sigma R_q)^{1/(1-q/2)} \log d)$, is of a higher order than the centralized statistical error $\|\Delta^\star\|^2$. In fact, the additional term $R_q(\frac{\log d}{N})^{1-q/2}$, due to the statistical nonidentifiability of linear regression over the ℓ_q-ball and thus not improvable, is no larger than $\|\Delta^\star\|^2$ with high-probability. **(ii)** The aforementioned scaling on N is unavoidable, as it matches that necessary for any (centralized) method to be statistically consistent over ℓ_q-ball [23]. This proves that statistically optimal estimates are achievable over mesh networks even when the local sample size n violates information theoretical lower bounds (as long as N satisfies so). This is possible thanks to the network mixing local information at sufficiently fast rate, as required by (43). **(iii)** As anticipated, condition (43), when not implicitly satisfied, can be met running multiple rounds of communications/iteration; it can be shown that $\widetilde{\mathcal{O}}(\log m/(1-\rho))$ are enough, resulting in an overall communication complexity as in (5) to reach an ε-neighborhood of a statistically optimal estimate.

• **Low-rank matrix recovery:** We consider the compressed sensing version of the matrix regression over meshed networks, as formulated next. Let $\Theta^\star$ in (1) satisfy $\|\Theta^\star\|_q \leq R_q$, with $q \in [0,1]$ and $R_q > 0$. In that the case $q = 0$ corresponds to matrices with rank at most r; we will have $R_0 = r$. The case $q > 0$ results in approximately low-rank of $\Theta^\star$. Consider observations $y_j = \langle X_j, \Theta^\star\rangle + w_j$, $j \in \mathcal{S}_i$, $|\mathcal{S}_i| = n$, where $X_j \in \mathbb{R}^{p\times p}$ is such that $\mathrm{vec}(X_j)$ satisfies Assumption 4.(i), $w_j \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. and independent of X_j. Define

$$\chi_N \triangleq \frac{\max_i([\Sigma]_{ii})}{\lambda_{\min}} R_q \left(\frac{p}{N}\right)^{1-\frac{q}{2}}, \quad q \in [0,1] \text{ and } R_0 = r.$$

Corollary 2. *Consider the ERM* (3) *solving the low-rank matrix regression problem above with $\mathcal{R}(\Theta) = \|\Theta\|_1$, $\|\Theta^\star\|_1 \leq R = \|\hat{\theta}\|_1$, and $\Omega \equiv \mathbb{R}^{p\times p}$; let $N = \Omega\left(p(\eta_\Sigma R_q)^{1/(1-q/2)}\right)$. Let $\{(\theta_i^t)_{i=1}^m\}_{t\geq 0}$ be the sequence generated by Algorithm 1 with tuning as in Theorem 2, including* (17) *for some $C_1, C_2 > 0$. Then, with probability at least $1 - c_0 \exp\left(-c_1 N + \log(m)\right)$, for some $c_0,\, c_1 > 0$, it holds:*

$$\frac{1}{m}\sum_{i=1}^m \|\theta_i^t - \widehat{\theta}\|^2 \leq \frac{8}{\lambda_{\min}(\Sigma)}\left(1 - \sqrt{\frac{1}{32\kappa}}\right)^t L^0 + \mathcal{O}\left(\chi_N\left(R_q\left(\frac{p}{N}\right)^{1-\frac{q}{2}} + \|\Delta^\star\|^2\right)\right),$$

where $L^0 > 0$ depends on the the initial optimality gap and statistical parameters (see Appendix A.6).

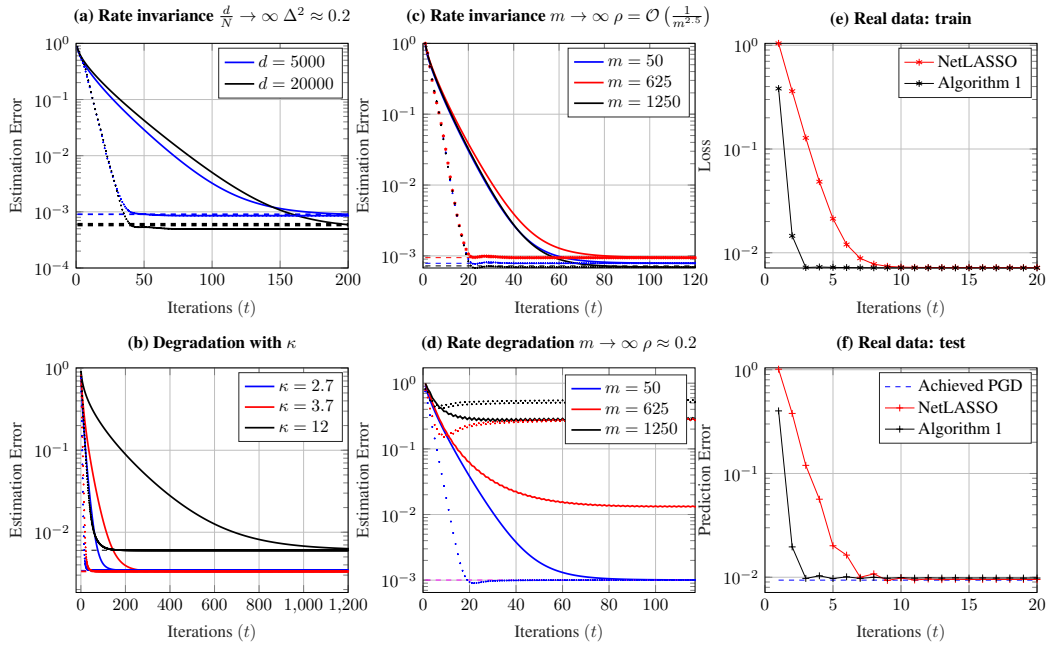

Figure 2: Algorithm 1 (dotted-lines) vs. NetLASSO (solid-lines). **(a)** Linear convergence vs iterations up to statistical precision (dashed) Δ^2, for s, d, N growing, such that $\Delta^2 \approx 0.2$. **(b)** Estimation error vs iterations for different values of κ. **(c)** Estimation error vs iterations for growing m such that $\rho = \mathcal{O}(m^{-2.5})$. **(d)** Same as (c) but with $\rho = \approx 0.2$. **(e)-(f):** `eyedata` data set, (e) training error vs. iterations, (f) testing error vs. iterations.

Although quantitative aspects of the sample rates are different, Corollary 2 is analogous to Corollary 1, which we refer to for specific comments. As therein, also here the final tolerance is of a higher-order of the statistical error of the model.

Overall the above statistical guarantees (along with those discussed in the supporting material for other estimation problems, see Appendix A.3) complement [1], matching those therein but over meshed networks while improving convergences rate of the PGA gaining acceleration.

5 Numerical Results

Numerical results are presented in this section validating our theoretical findings on synthetic and real data. Simulations are performed on a computer with an Intel i7-8650U CPU @ 1.9 GHz using 16 GB RAM.

Synthetic data: We simulate the sparse linear regression model as in Corollary 1, with $q = 0$. We set $R_0 = s$, $\sigma^2 = 0.25$ and $R = \|\theta^\star\|^2$. The covariates $x_i \in \mathbb{R}^d$ are generated as in [1] $x_{i,j} = z_{i,j} + \omega x_{i,j-1}$, with $x_{i,1} = z_{i,1}$, where each $z_{i,j} \sim \mathcal{N}(0,1)$ and i.i.d. This implies that $\mu \geq \mu' \triangleq (1+\omega)^{-2}$ and $L \leq L' \triangleq 2((1+\omega)^2(1+\omega))^{-1}$. We will denote $\Delta^2 \triangleq (s \log d)/N$, which is of the order of the statistical precision of the model. The network is generated using an Erdős-Rényi graph, with m agents and link activation probability p. The specific values of each of the parameters is given in each specific experiment. All curves are averaged over 10 independent Montecarlo simulations. We now proceed to verify the different aspects of Corollary 1.

1) Linear convergence to optimal estimates: Fig 2.(a) plots the normalized estimation error $\sum_i \|\theta_i^t - \theta^\star\|^2 / \|\theta^\star\|^2$ versus iterations, generated by Algorithm 1 (dashed-lines) and NetLASSO [26] (solid lines), solving the regression problem, with $\omega = 0.5$ (yielding $\mu = 0.44$ and $L = 5.33$); different curves refer to different values of s, d, N, with d growing faster than N, such that $\Delta^2 \approx 0.2$, and $m = 5$. In both algorithms we set the stepsize $\alpha = 2L'$; the momentum parameter β is set to $\beta = (8\mu'/L')^{-1/2}$. Condition (43) is enforced via $T = 41$ rounds of communications/iteration for both algorithms. In **Fig. 2.(b)** we plot the same quantities now for different condition numbers κ, obtained choosing $\omega \in \{0.1, 0.2, 0.5\}$. The following comments are in order: **(i)** As predicted by (20), Algorithm 1 convergences at linear rate to statistically optimal estimates; although it has the same computational cost per iteration of the nonaccelerated NetLASSO, it is much faster, especially when Σ is ill-conditioned (large κ), which proves the advantages of acceleration also in high-dimensions.

(ii) As d and N increase, as long as Δ^2 is constant, the rate and reached statistical accuracy remain roughly invariant. (Fig. 2.(a)), which is a resirable property.

2) On the scaling of ρ with m as in (43): Fig. 2.(c)-(d) plots the same quantities as in Fig. 2(a), now with fixed $d = 5000$, $s = \lceil \sqrt{d} \rceil$, $N = n \cdot m = 2500$, and varying $m \in \{50, 625, 1250\}$. The random graph is generated so that in plot (c) $\rho \approx \mathcal{O}(m^{-2.5})$ (this is achieved setting $p = 0.8$) while in plot (d) $\rho \approx 0.2$ for all m (achieved setting $p \in \{0.7, 0.4, 0.22\}$ for the associated values of m). The number of communications/iteration is set to 7 for both algorithms. The two plots show that ρ should decrease with m if one wants to achieve centralized statistical consistency; $\rho \approx \mathcal{O}(m^{-2.5})$, as predicted by (43), seems to be sufficient. On the other hand, a ρ constant with m breaks down the performance of both distributed algorithms.

(B) Real data: We test here NetLASSO and Algorithm 1 on the data set `eyedata` in the `NormalBetaPrime` package [4]; $d = 200$ and $N = 120$. We generated $\bar{W}$ for $m = 10$ and $p = 0.9$. To achieve the required connectivity we set $W = \bar{W}^7$. The data set is split between test- and training-data, corresponding to $N_{\text{test}} = 40$ and $N_{\text{train}} = 80$, respectively. Both data sets are equally distributed across $m = 10$ agents. In Fig. 2.(e) we display quantitites relevant to the training phase. Specifically, we plot the objective function value along the iterates generated by NetLASSO and Algorithm 1. We observe that both schemes achieve a similar final objective function value and converge linearly, with Algorithm 1 at much faster rate. Fig. 2.(f) shows the performance on the test data. We implemented the following procedure. Denote by y the output of the test-set and by X the test covariates. We build predictors $y_i^t = X\theta_i^t$, $i \in [m]$, where θ_i^t is the estimate of $\theta^\star$ from agent i at time t, generated by the two algorithms when running on the training set. The figure plots $(N_{\text{test}} m)^{-1} \sum_{i=1}^m \|y - y_i^t\|_2^2$ versus iterations, for NetLASSO and Algorithm 1. The dashed line corresponds to $(N_{\text{test}} m)^{-1} \|y - \hat{y}\|^2$, where $\hat{y} = X\widehat{\theta}$ and $\widehat{\theta}$ denotes the estimator obtained via the PGA. The performance on test-data are consistent with what observed already on training-data.

6 Conclusions

We proposed a decentralized accelerated algorithm to solve quadratic high dimensional estimation problems over mesh networks whose empirical losses are neither strongly convex nor Lipschitz smooth globally. To employ acceleration in this setting, the design hinges on careful considerations regarding the directions traversed by the schemes, enforcing sparsity of the iterate and momentum sequences. Global convergence to statistically optimal solutions is proved at liner rate, proportional to $\sqrt{\kappa}$, with a communication cost per iteration at most $\tilde{\mathcal{O}}\left(\log m/(1-\rho)\right)$. It is unclear whether this communication cost is improvable, for instance, whether the log-dependence on the number of agents and the transmission of all d gradient/iterate components can be relaxed. This is left to future investigations.

Acknowledgments and Disclosure of Funding

Funding in direct support of this work: ONR Grant N. N00014-21-1-2673.

References

[1] Alekh Agarwal, Sahand Negahban, and Martin Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012.

[2] Alfred Auslender and Mark Teboulle. Interior gradient and proximal methods in convex and conic optimization. *SIAM J. Optimization*, 16:697–725, 2006.

[3] Michel Baes. Estimate sequence methods: extensions and approximations. *Institute for Operations Research, ETH, Zürich, Switzerland*, 2009.

[4] Ray Bai and Malay Ghosh. On the beta prime prior for scale parameters in high-dimensional bayesian regression models. *arXiv preprint arXiv:1807.06539*, 2018.

[5] Amir Beck and Mark Teboulle. A fast iterative shrinkage-threshold algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[6] Ron Bekkerman, Mikhail Bilenko, and John Langford. *Scaling up Machine Learning: Parallel and Distributed Approaches*. Cambridge University Press, 2011.

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

[8] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279, July 2008.

[9] Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021.

[10] Alexander Gasnikov and Yuri Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58:48–64, 2018.

[11] Yao Ji, Gesualdo Scutari, Ying Sun, and Harsha Honnappa. Distributed sparse regression via penalization. *arXiv preprint, arXiv:2111.06530*, Nov. 2021.

[12] Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[13] Huan Li, Cong Fang, Wotao Yin, and Zhouchen Lin. A sharp convergence rate analysis for distributed accelerated gradient methods. *arXiv preprint, arXiv:1810.01053*, 2018.

[14] Huan Li and Zhouchen Lin. Revisiting extra for smooth distributed optimization. *SIAM J. on Optimization*, 30(3):1795–1821, 2020.

[15] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 5330–5340, 2017.

[16] Qihang Lin and Lin Xiao. An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. *Comput. Optim. Appl.*, 60:633–674, 2015.

[17] Po-Ling Loh and Martin Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

[18] Angelia Nedić, Alex Olshevsky, and Mike Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106:953–976, 2018.

[19] Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27:2597–2633, July 2016.

[20] Yuri Nesterov. Gradient methods for minimizing composite objective function. *Math. Program. (Ser.B)*, 140:125–161, 2013.

[21] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[22] Guannan Qu and Na Li. Accelerated distributed nesterov gradient descent. *IEEE Transactions on Automatic Control*, 65(6):2566–2581, 2019.

[23] Garvesh Raskutti, Martin Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

[24] Alexander Rogozin, Vladislav Lukoshkin, Alexander Gasnikov, Dmitry Kovalev, and Egor Shulgin. Towards accelerated rates for distributed optimization over time-varying networks. In *International Conference on Optimization and Applications (OPTIMA 2021)*, pages 258–272, 2021.

[25] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3027–3036, 2017.

[26] Ying Sun, Marie Maros, Gesualdo Scutari, and Guang Cheng. High-dimensional inference over networks: Linear convergence and statistical guarantees. *ArXiv preprint, arXiv:2201.08507*, Jan 2022.

[27] Paul Tseng. On accelerated proximal gradient methods for convex–concave optimization. *manuscript*, 2008.

[28] César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. A dual approach for optimal algorithms in distributed optimization over networks. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–37, 2020.

[29] Martin Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[30] Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Convergence of asynchronous distributed gradient methods over stochastic networks. *IEEE Trans. Automat. Contr.*, 63(2):434–448, 2017.

[31] Tianbao Yang, Lijun Zhang, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. A simple homotopy proximal mapping algorithm for compressive sensing. *Machine Learning*, 108:1019–1056, 2019.

[32] Haishan Ye, Luo Luo, Ziang Zhou, and Tong Zhang. Multi-consensus decentralized accelerated gradient descent. *arXiv preprint, arXiv:2005.00797*, 2020.

[33] Haishan Ye, Ziang Zhou, Luo Luo, and Tong Zhang. Decentralized accelerated proximal gradient descent. In *Proc. of the Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The limitations of our work follow from the provided assumptions.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Supplementary Material
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]
(b) Did you mention the license of the assets? [N/A]
(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Appendix

This appendix contains supplementary theoretical results as well as the proofs of the main results presented in the main paper. We organize this content as follows.

Sec. A.1 contains the list of the notation used in the paper and the rest of this appendix.

Sec. A.2 contains additional numerical simulations.

Sec. A.3 complements results in Sec. 4.1, establishing guarantees of Algorithm 1 applied to sub-Gaussian linear regression (Corollary 4) and convex matrix completion (Corollary 3).

Sec. A.4 Contains a more general version of Theorem 2 along with its proof..

Sec. A.6 Contains proofs of Corollaries 1-4

Sec. A.7 Contains some intermediate technical results instrumental for the proofs presented in Section A.4.

A.1 Notation

Problem size:

Symbols	Location	Description
d	(1)	Problem dimension
n	(2)	Number of local samples
m	(3)	Number of agents
$N = n \cdot m$	Section 1	Total number of samples

Relevant functions:

Symbols	Location	Description
f_i	(2)	Local empirical loss
f	(3)	Global empirical loss
$\boldsymbol{f}$	(21)	Stacked empirical risk
$\boldsymbol{F}$	(22)	Stacked global empirical risk

Population curvature parameters:

Symbols	Location	Description
κ	Section 1	$\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ (See Section A.4)

Network quantities:

Symbols	Location	Description
ρ	Assumption 5	Network connectivity
W	Assumption 5	Gossip matrix
$\mathbf{W}$	Section A.4	$W \otimes I_d$
J	Section 5	$J = \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^\top$
$\mathbf{J}$	Section A.4	$\mathbf{J} = J \otimes I_d$

Structure promotion:

Symbols	Location	Description
$\mathcal{R}$	(3)	Norm constraint
R	(3)	Constraint radius
$\mathcal{M}$	Assumption 3	Model subspace
$\mathcal{M}^\perp$	Assumption 3	$\mathcal{R} - \ell_2$ Lipschitz constant $\mathcal{M}$
$\mathcal{R}^*$	Section A.4	Dual norm to $\mathcal{R}$

Tolerances:

Symbols	Location	Description
γ_ℓ, τ_ℓ	Assumption 2	Local RSM constant and tolerance
μ, τ_μ	Assumption 1	Global RSC constant and tolerance
L, τ_L	Assumption 1	Global RSM constant and tolerance

Other symbols:

Symbols	Location	Description
$\mathcal{S}_i$	(2)	Indexes of samples corresponding to agent i
α	Algorithm 1	Step-size
β	Algorithm 1	Momentum term

We boldface stacked quantities of any type, vector and matrices. For vectors, the stacked is obtained grouping agents' variables into a single vector; for instance, we write $\boldsymbol{\theta} \triangleq [\theta_1^\top, \ldots, \theta_m^\top]^\top$, where θ_i is the local variable owned by agent i. For gossip matrices $A \in \mathbb{R}^{m\times m}$ we define their "augmented" instance as

$$\mathbf{A} \triangleq A \otimes I_d,$$

where $\otimes$ denotes the Kronecker product.

Following the same rationale, we define the stacked local empirical risk as

$$\boldsymbol{f}(\boldsymbol{\theta}) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_i(\theta_i), \tag{21}$$

while the stacked average loss is defined as

$$\boldsymbol{F}(\boldsymbol{\theta}) \triangleq \frac{1}{m}\sum_{i=1}^{m} f(\theta_i). \tag{22}$$

Other useful notation is the following: $1_m \in \mathbb{R}^m$ is the vector of all 1's. We define $J \triangleq \frac{1}{m}1_m 1_m^\top$; it is the projection onto the consensus space. Given a matrix Σ, we denote by $\lambda_i(\Sigma)$ and $\sigma_i(\Sigma)$ its ith largest eigenvalue and singular value, respectively–with $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ (resp. $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$) being the largest and smallest eigenvalues (resp. singular values), respectively. We denote by $\mathcal{R}^*$ the dual norm of the norm $\mathcal{R}$. For any $\Theta \in \mathbb{R}^{p\times p}$, $\|\Theta\|_1 = \sum_{j=1}^{p}\sigma_j(\Theta)$ is the nuclear norm; $\|\Theta\|_\infty = \max_{i,j\in[p]}|\Theta_{i,j}|$; $|||\Theta|||_\infty = \max_{i\in[p]}\sum_{j\in[p]}|\Theta_{i,j}|$; $|||\Theta|||_2$ is the spectral norm; and $\|\Theta\|_F$ is the Frobenius norm. Finally, denote by $\|\mathbf{x}\|_{\mathbf{A}}^2$ the semi-norm induced by the p.s.d. matrix $\mathbf{A}$, i.e. $\|\mathbf{x}\|_{\mathbf{A}}^2 \triangleq \langle \mathbf{x}, \mathbf{A}\mathbf{x}\rangle$.

A.2 Additional Experiments

This section complements the results presented in Sec. 5. In particular we show the invariance of the performance of Algorithm 1 under the asymptotic scaling "d/N growing", as long as $\frac{s\log d}{N}$ remains constant.

Consider the sparse linear regression model as described in Sec. 5, with $q = 0$, $\sigma^2 = 0.25$ and $R = \|\theta^\star\|^2$. The covariates $x_{i,j} \sim \mathcal{N}(0,1)$ are i.i.d. leading to $\mu = L = 1$. All the curves are averaged over 10 independent Montecarlo simulations. The network is generated using an Erdős-Rényi graph, with $m = 5$ and $p = 0.5$. To achieve a connectivity $\rho \approx 0.02$ we run 14 rounds of gradient tracking and consensus per iteration. Figure 3 plots the normalized estimation error $\frac{\sum_{i=1}^{m}\|\theta_i^t - \theta^\star\|^2}{m\|\theta^\star\|^2}$ versus iterations, generated by Algorithm 1 and NetLASSO (solid lines) solving the regression problem with $d = \{10^4, 10^5, 10^6\}$, $n = \{80, 100, 120\}$, $\alpha = 1/2$ and $\beta = \sqrt{1/8}$. We observe that as predicted by the theory, the achieved statistical precision and convergence rate remain of the same order, under the scaling $\frac{d}{N}\uparrow$. Furthermore, Figure 3 shows a gain using acceleration in high dimension even when $\kappa = 1$.

Real data: We test the NetLASSO and Algorithm 1 on the data set `Housing`[1], with $d = 13$, and total sample size $N_{\text{test+train}} = 500$. The network is generated using an Erdős-Rényi graph with $m = 5$

[1]The data set can be found `https://www.cs.ubc.ca/ schmidtm/Software/lasso.html`

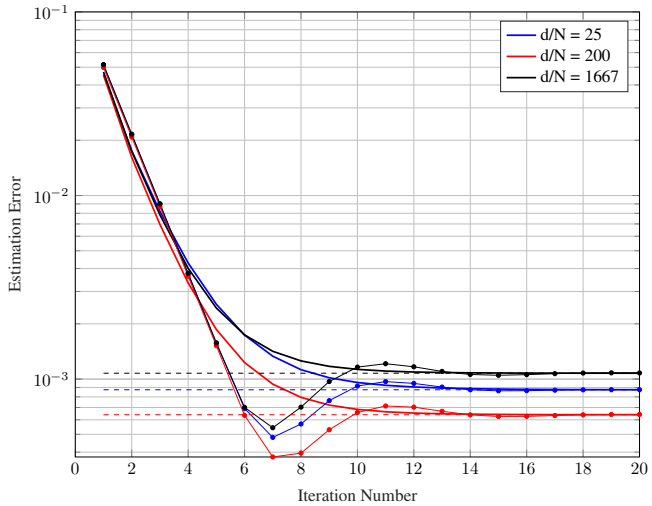

Figure 3: Estimation error versus interations for different values of $\frac{d}{N}$; $d = 10^4$, d^5, and d^6; $n = 80$, 100 and 120, with $m = 5$. The network has connectivity $\rho = 0.023$; $\mu = L = 1$, the stepsize is then set to $1/2$, and $\beta = \sqrt{1/8}$. Solid line curves refer to NetLASSO while dotted-line curves represent the accelerated variant. Observe that even when the population has $\kappa = 1$ acceleration yields a gain (as predicted by the theory). Dotted-line curves indicate statistical precision achieved using centralized PGD.

and $p = 0.5$; 14 communication rounds are performed to achieve a connectivity of $\rho \approx 0.02$. The data set is split between training and test samples, $N_{\text{train}} = 10$ and $N_{\text{test}} = 490$, respectively. We denote by X_i^{train}, y_i^{train} (resp. X_i^{test}, y_i^{test}) the training (resp. testing) sample pairs assigned to agent i . Observe that we choose N_{train} small to illustrate the performance when the objective function is not strongly convex. The left plot in Figure 4 reports the performance in terms of

$$\frac{\sum_{i=1}^m \left\| X_i^{\text{train}} \theta_i^t - y_i^{\text{train}} \right\|^2}{\sum_{i=1}^m \left\| X_i^{\text{train}} \bar{\boldsymbol{\theta}} - y_i^{\text{train}} \right\|^2}$$

where $\bar{\boldsymbol{\theta}}$ is the solution obtained via PGD (centralized), and θ_i^t denote the iterates generated by either NetLASSO (black-line curves) or Algorithm 1 (red-line curves). The right plot in Figure 4 shows the performance of the two algorithms in terms of test/prediction error, defined as

$$\frac{\sum_{i=1}^m \| X_i^{\text{test}} \theta_i^t - y_i^{\text{test}} \|^2}{\sum_{i=1}^m \| X_i^{\text{train}} \bar{\boldsymbol{\theta}} - y_i^{\text{train}} \|^2},$$

where θ_i^t are obtained using the NetLASSO and Algorithm 1 (black-line and red-line curves, respectively) using the train data set, where $\bar{\theta}$ is obtained using the centralized PGD on the train data set. Observe that both test and train errors are well behaved while the test error ends up being slightly larger than the train error.

A.3 Additional Statistical Models and guarantees

This section contains statistical-computation guarantees of Algorithm 1 applied to (i) the sparse linear regression, with covariantes x_i drawn from a sub-Gaussian distribution; and (ii) the convex matrix completion problem.

• **Sparse sub-Gaussian regression:** consider the sparse regression model described in Section 4 where x_i are i.i.d. and now $(\tau^2, \Sigma)-$sub-Gaussian, i.e., fulfill Assumption 4.(ii). Denote by $X \triangleq [x_1^\top; \ldots; x_N^\top]$ the matrix containing the samples x_i row-wise, and assume that each column of X has ℓ_2-norm equal to $C > 0$. This assumption is standard when dealing with sub-Gaussian measurement

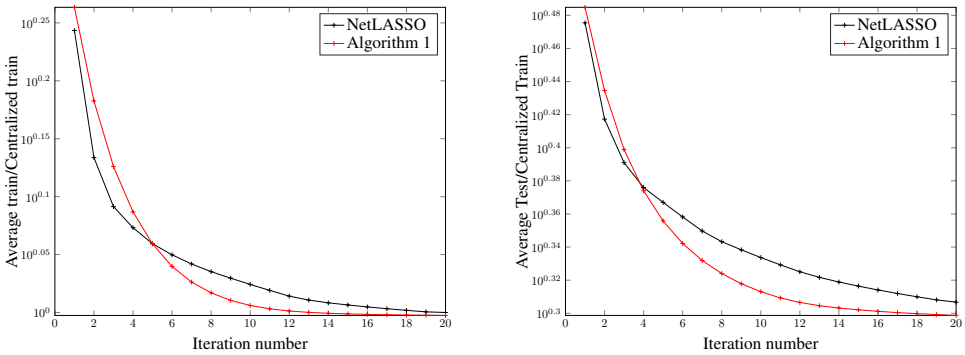

Figure 4: `Housing` data set. Left panel: normalized training error v.s. iterations; Right panel: normalized testing error v.s. iterations.

matrices [29]. Finally, define

$$\chi_N \triangleq \frac{\max\left\{\frac{\tau^4}{\lambda_{\min}(\Sigma)}, \lambda_{\min}(\Sigma)\right\}}{\lambda_{\min}(\Sigma)} R_q \left(\frac{\log d}{N}\right)^{1-\frac{q}{2}}, \quad q \in [0,1] \text{ and } R_0 = s.$$

Corollary 3. *Consider the ERM* (3) *solving the linear regression problem described above, with* $\mathcal{R}(\cdot) = \|\cdot\|_1$, $\|\theta^\star\|_1 \leq R = \|\widehat{\theta}\|_1$ *and* $\Omega \equiv \mathbb{R}^d$; *and let*

$$N = \Omega\left((R_q \max\left\{\frac{\tau^4}{\lambda_{\min}(\Sigma)}, \lambda_{\min}(\Sigma)\right\})^{1/(1-q/2)} \log d\right).$$

Let $\{(\theta_i^t)_{i=1}^m\}_{t\geq 0}$ *be the sequence generated by Algorithm 1 with tuning as in Theorem 2, where* (17) *becomes*

$$\rho \leq C_1 \min\left\{\sqrt{\kappa^{-1}}, \frac{\kappa^{-1}}{m^{2.5}(1+C_2/L)}\right\}, \quad \textit{for some } C_1,\, C_2 > 0.$$

Then, with probability at least $1 - c_0 \exp\left(-c_1 N \min\{\mu^2/\tau^4, 1\} + \log(m)\right)$, *for some* $c_0, c_1, > 0$, *it holds*

$$\frac{1}{m}\sum_{i=1}^m \|\theta_i^t - \widehat{\theta}\|^2 \leq \frac{8}{\lambda_{\min}(\Sigma)}\left(1 - \sqrt{\frac{1}{32\kappa}}\right)^t L^0 + \mathcal{O}\left(\chi_N\left(R_q\left(\frac{\log d}{N}\right)^{1-\frac{q}{2}} + \|\Delta^\star\|^2\right)\right).$$

Proof. See Appendix A.6.3. □

• **Matrix Completion:** Let y_j be a noisy measurement of $[\Theta^\star]_{a(j),b(j)}$, which corresponds to the $a(j), b(j)$ element of the matrix $\Theta^\star \in \mathbb{R}^{p\times p}$, which is selected uniformly at random from its $d = p^2$ elements. $\Theta^\star$ is assumed to be near low-rank, in the sense that $\|\Theta^\star\|_q \leq R_q$ with $q \in (0,1]$ where $\|\Theta^\star\|_q \triangleq \sum_{i=1}^p \sigma_i(\Theta^\star)^q$. Let $X_j \in \mathbb{R}^{p\times p}$, with $[X_j]_{a(j),b(j)} = 1$ and remaining entries equal to zero. Then, $y_j = \langle X_j, \Theta^\star\rangle + (\sigma/d)w_j$, $j \in \mathcal{S}_i$, $i \in [m]$, and w_j are sub-exponential with parameter 1 and independent from X_i. Our goal is to compute an estimate of $\Theta^\star$. To achieve this, we cast the problem into an ERM of the form (3) where we choose $\Omega = \{\Theta \in \mathbb{R}^{p\times p} : \|\Theta\|_\infty \leq \omega/p\}$ with $\omega \geq 1$. Observe that by vectorizing the quantities involved in the inner product, Assumption 4(iii) is fulfilled.

Corollary 4. *Consider the ERM 3 solving the matrix completion described above, with* $\mathcal{R}(\cdot) = \|\cdot\|_1$, $\|\Theta^\star\|_1 \leq R = \|\widehat{\Theta}\|_1$ *and* $\Omega = \{\Theta : \|\Theta\|_\infty \leq \frac{\omega}{p}\}$, *with* $\omega \geq 1$; *and set* $N = \Omega\left(p \log p\right)$. *Let* $\{(\Theta_i^t)_{i=1}^m\}_{t\geq 0}$ *be the sequence generated by Algorithm 1 with tuning as in Theorem 2, where* (17) *becomes*

$$\rho \leq \frac{C_1}{m^5}, \quad \textit{for some } C_1 > 0.$$

Then, with probability at least $1 - c_0(m+1)\exp(-c_1 p \log p)$ *for some* $c_0,\ c_1 > 0$ *it holds*

$$\frac{1}{m}\sum_{i=1}^{m}\|\theta_i^t - \widehat{\theta}\|^2 \leq 8\left(1 - \frac{1}{4\sqrt{2}}\right)^t L^0 + \mathcal{O}\left(R_q \omega^{2-q}\left(\frac{p\log p}{N}\right)^{1-\frac{q}{2}} + \|\Delta^\star\|^2\right).$$

Proof. See Appendix A.6.4. □

A.4 Proof of Theorem 2

Using the notation introduced in Appendix A.1, we start rewriting Algorithm 1 in the "augmented" vector-matrix form:

$$\mathbf{w}^t = \frac{1}{\beta+1}\boldsymbol{\theta}^t + \frac{\beta}{1+\beta}\mathbf{v}^t \tag{23a}$$

$$\mathbf{z}^t = \mathbf{W}\mathbf{w}^t \tag{23b}$$

$$\mathbf{y}^t = \mathbf{W}\left(\mathbf{y}^{t-1} + \left(\nabla \boldsymbol{f}(\mathbf{z}^t) - \nabla \boldsymbol{f}(\mathbf{z}^{t-1})\right)\right) \tag{23c}$$

$$\mathbf{v}^{t+1} = \operatorname*{argmin}_{v_i \in \Omega : \mathcal{R}(v_i) \leq R, \forall i}\left\{m\langle \mathbf{y}^t, \mathbf{v}\rangle + \frac{\beta\alpha}{2}\|\mathbf{v} - (1-\beta)\mathbf{v}^t - \beta\mathbf{z}^t\|^2\right\} \tag{23d}$$

$$\boldsymbol{\theta}^{t+1} = \beta\mathbf{v}^{t+1} + (1-\beta)\boldsymbol{\theta}^t, \tag{23e}$$

where $\mathbf{z}^{-1} = \mathbf{W}\mathbf{w}^{-1}$, and $\mathbf{w}^{-1}$ is any arbitrary vector, with each of the d-component being feasible for (3). Notice that, with a slight abuse of notation, we still used the (stacked) vector $\mathbf{y}$ to denote the tracking variables as in (23c), although in (23d) they are m times those defined in (14), due to the $1/m$ factor in the definition of $\boldsymbol{f}$ (see (21)). This explains the coefficient m in front of $\langle \mathbf{y}^t, \mathbf{v}\rangle$ in (23d).

We will study convergence leveraging the following Lyapunov-like function along the iterates of the algorithm:

$$V^t \triangleq \boldsymbol{F}(\mathbf{W}\boldsymbol{\theta}^t) - \boldsymbol{F}(\mathbf{W}\widehat{\boldsymbol{\theta}}) + \frac{\beta^2\alpha}{2m}\|\boldsymbol{\theta}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \frac{\beta^2\alpha}{2m}\|\mathbf{v}^t - \widehat{\boldsymbol{\theta}}\|^2. \tag{24}$$

For convenience, define

$$\boldsymbol{g}(\mathbf{w}) \triangleq \boldsymbol{F}(\mathbf{W}\mathbf{w}) + \frac{\beta^2\alpha}{2m}\|\mathbf{w}\|^2_{\mathbf{I}-\mathbf{W}^2}.$$

It is not difficult to check that, under Assumption 1, $\mathbf{g}$ inherits the RSC and RSM properties of f, in the following sense:

$$\frac{\mu}{4m}\|\mathbf{x}-\mathbf{y}\|^2_{\mathbf{W}^2} + \frac{\beta^2\alpha}{2m}\|\mathbf{x}-\mathbf{y}\|^2_{\mathbf{I}-\mathbf{W}^2} - \varepsilon_\ell(\mathbf{x},\mathbf{y}) \leq \boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}(\mathbf{y}) - \langle \nabla \boldsymbol{g}(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle, \tag{25a}$$

$$\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}(\mathbf{y}) - \langle \nabla \boldsymbol{g}(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \leq \frac{L}{m}\|\mathbf{x}-\mathbf{y}\|^2_{\mathbf{W}^2} + \frac{\beta^2\alpha}{2m}\|\mathbf{x}-\mathbf{y}\|^2_{\mathbf{I}-\mathbf{W}^2} + \varepsilon_u(\mathbf{x},\mathbf{y}), \tag{25b}$$

for all $\mathbf{x}, \mathbf{y} \in \Omega^m$, where

$$\varepsilon_\ell(\mathbf{x},\mathbf{y}) \triangleq \frac{\tau_\mu}{2m}\sum_{i=1}^{m}\mathcal{R}\left(\sum_{j=1}^{m} w_{i,j}(x_j - y_j)\right) \quad \text{and} \quad \varepsilon_u(\mathbf{x},\mathbf{y}) \triangleq \frac{\tau_L}{2m}\sum_{i=1}^{m}\mathcal{R}^2\left(\sum_{j=1}^{m} w_{i,j}(x_j - y_j)\right). \tag{26}$$

Using the above properties, our first step is to establish descent on the Lyapunov function up to some error; formally, we have the following.

Proposition 1. *For any* $\alpha \geq 2L$*, the following holds:*

$$V^{t+1} \leq (1-\beta)\,V^t - \frac{\beta}{2m}\left(\frac{\mu}{2} - \beta^2\alpha\right)\|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2 + \beta\varepsilon_\ell(\widehat{\boldsymbol{\theta}}, \mathbf{w}^t) + \varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t) + \langle \boldsymbol{\varepsilon}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle, \tag{27}$$

for all $t = 0, 1, \ldots$*, where*

$$\boldsymbol{\varepsilon}^t \triangleq \mathbf{y}^t + \frac{\beta^2\alpha}{m}(\mathbf{W} - \mathbf{I})\mathbf{z}^t - \mathbf{W}\nabla \boldsymbol{F}(\mathbf{W}\mathbf{w}^t). \tag{28}$$

Proof. See Appendix A.7.1. □

The rest of the proof consists of controlling the error terms

$$\beta\varepsilon_\ell(\widehat{\boldsymbol{\theta}}, \mathbf{w}^t) + \varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t) + \beta\langle \boldsymbol{\varepsilon}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle.$$

The errors $\varepsilon_\ell(\widehat{\boldsymbol{\theta}}, \mathbf{w}^t)$ and $\varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t)$ can be controlled using the decomposability of the regularizer–Lemma 2 below provides the desired bound–whereas $\langle \boldsymbol{\varepsilon}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle$ requires a suitable analysis of the tracking variables $\mathbf{y}^t$–Proposition 2 below summarizes the result.

Lemma 2. *Let*

$$\nu^2 \triangleq 2\left(2\mathcal{R}\left(\Pi_{\mathcal{M}^\perp}(\theta^\star)\right) + 2\mathcal{R}(\theta^\star - \widehat{\theta}) + \Psi(\bar{\mathcal{M}})\|\widehat{\theta} - \theta^\star\|\right)^2. \tag{29}$$

The following bounds hold for $\varepsilon_\ell(\widehat{\boldsymbol{\theta}}, \mathbf{w}^t)$ and $\varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t)$:

$$\varepsilon_\ell(\widehat{\boldsymbol{\theta}}, \mathbf{w}^t) \leq \frac{1}{2m}\tau_\mu 8\Psi^2(\bar{\mathcal{M}})\|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2 + \frac{\tau_\mu}{2}\nu^2, \tag{30}$$

$$\begin{aligned}\varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t) \leq{}& \frac{1}{2m}24\tau_L\Psi^2(\bar{\mathcal{M}})\beta^2\left(\|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 + \beta^2\|\mathbf{w}^t - \widehat{\boldsymbol{\theta}}\|^2_{\mathbf{W}^2} + (1-\beta)^2\|\mathbf{v}^t - \widehat{\boldsymbol{\theta}}\|^2\right)\\ &+ \frac{\tau_L}{2}9\beta^2\nu^2.\end{aligned} \tag{31}$$

Proof. See Appendix A.7.2. □

Proposition 2. *The following bound holds for $\langle \boldsymbol{\varepsilon}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}\rangle$:*

$$\begin{aligned}\langle \varepsilon^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}\rangle \leq{}& \frac{A(\rho)}{2m}\|\mathbf{v} - \widehat{\boldsymbol{\theta}}\|^2 + \frac{A(\rho)}{2m}\|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2 + \frac{B(\rho)}{2m}\|\boldsymbol{\theta}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \frac{B(\rho)}{2m}\beta\|\mathbf{v}^t - \widehat{\boldsymbol{\theta}}\|^2\\ &+ \frac{R(\rho)}{2m}\sum_{i=0}^{t-1}\rho^{t-1-i}\|\mathbf{z}^i - \widehat{\boldsymbol{\theta}}\|^2 + \frac{C}{2m}\rho^{t+1} + C(\rho)\nu^2,\end{aligned}$$

for any $\mathbf{v} = [v_1^\top, \ldots, v_m^\top]^\top$, with feasible $v_1, \ldots, v_m$, where ν^2 is defined in (29), and

$$\begin{aligned}A(\rho) &\triangleq \beta^2\alpha\rho + 3m\sqrt{m}\rho\left(\left(3 + \frac{2\rho}{1-\rho}\right)(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell) + 2L + 8\Psi^2(\bar{\mathcal{M}})\tau_L\right),\\ B(\rho) &\triangleq \frac{\beta^2\alpha\rho}{1+\beta} + \frac{3m\sqrt{m}\rho}{1+\beta}\left(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell\right),\\ R(\rho) &\triangleq 6m\sqrt{m}\rho\left(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell\right),\\ C &\triangleq 4R\sqrt{m}m\sum_{i=1}^m \mathcal{R}^*(\nabla f_i(\theta^\star)) + 3m\sqrt{m}(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}}))\tau_\ell\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|^2,\\ C(\rho) &\triangleq 9\sqrt{m}m\rho\left(\left(3 + \frac{2\rho}{1-\rho}\right)\tau_\ell + \tau_L\right).\end{aligned} \tag{32}$$

Proof. See Appendix A.7.3. □

We can now combine the above intermediate results and prove convergence of Algorithm 1, stated in Theorem 3 below–Theorem 2 (in the main paper) follows as special case (see Remark 1).

To state the theorem, it is convenient to introduce the following quantities:

$$B \triangleq \boldsymbol{F}(\boldsymbol{\theta}^0) - \boldsymbol{F}(\widehat{\boldsymbol{\theta}}) + \frac{\mu}{16m}\|\mathbf{v}^0 - \widehat{\boldsymbol{\theta}}\|^2 + \frac{\beta\mu}{2m}\|\mathbf{z}^{-1} - \widehat{\boldsymbol{\theta}}\|^2 + \frac{\mu}{16m}\|\boldsymbol{\theta}^0\|^2_{\mathbf{I}-\mathbf{W}^2}, \tag{33}$$

$$\lambda \triangleq \frac{1 - \sqrt{\frac{1}{8\kappa}} + \left(1 - \sqrt{\frac{1}{8\kappa}}\right)\left(\frac{24\Psi^2(\bar{\mathcal{M}})\tau_L}{2L} - \frac{\sqrt{8\kappa}\tau(\rho)}{2L}\right) + \frac{\sqrt{8\kappa}\tau(\rho)}{L}}{1 - \frac{24\tau_L\Psi^2(\bar{\mathcal{M}})}{2L} - \frac{\sqrt{8\kappa}\tau(\rho)}{2L}} \tag{34}$$

where

$$\tau(\rho) \triangleq 3m\sqrt{m}\rho\left(\left(4+\frac{2\rho}{1-\rho}\right)\left(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell\right) + 3L + 12\Psi^2(\bar{\mathcal{M}})\tau_L\right). \tag{35}$$

Notice that the rate λ in (34) matches that of the centralized PGA applied to the ERM (3) [1], with an improved dependence on κ, which gains the square root (typical of accelerated methods).

Theorem 3 (Convergence of Algorithm 1)**.** *Given the ERM* (3)*, suppose* (i) *f satisfies the RSC/RSM conditions (Assumption 1); (ii) $\mathcal{R}$ is a decomposable regularizer with respect to a chosen pair of subspaces $(\mathcal{M}, \bar{\mathcal{M}})$ (Assumption 3); (iii) f_i satisfies the local RSM condition (Assumption 2) for all $i \in [m]$. Let $\{(\theta_i^t)_{i\in[m]}\}_{t\geq 1}$ be the sequence generated by Algorithm 1, with tuning $\alpha = 2L$ and $\beta = \sqrt{1/(8\kappa)}$ and gossip matrix satisfying $W \succeq 0$ and Assumption 5, with ρ such that*

$$\rho \leq \min\left\{\frac{56}{128}\beta, \frac{\kappa^{-1}}{384m\sqrt{m}\left(L^{-1}(16\gamma_\ell + 64\Psi^2(\bar{\mathcal{M}})\tau_\ell + 12\Psi^2(\bar{\mathcal{M}})\tau_L) + 3\right)}\right\}. \tag{36}$$

If $(\mathcal{M}, \bar{\mathcal{M}})$ and RSC/RSM parameters (μ, τ_μ) and (L, τ_L) are such that

$$\frac{\mu}{8} \geq 8\tau_\mu\Psi^2(\bar{\mathcal{M}}) + 24\tau_L\Psi^2(\bar{\mathcal{M}}), \tag{37}$$

then, for any solution $\widehat{\theta}$ of (3) *for which $\mathcal{R}(\widehat{\theta}) = R$, there holds*

$$\begin{aligned}
\frac{1}{m}\left\|\boldsymbol{\theta}^{t+1} - \widehat{\boldsymbol{\theta}}\right\|^2 \leq{}& \frac{8}{\mu}(\lambda+\rho)^{t+1}B \\
&+ (\lambda+\rho)^{t+1}\frac{2\sqrt{2}\rho}{\sqrt{\mu L}}\left(4R\sqrt{m}\sum_{i=1}^m \mathcal{R}^*(\nabla f_i(\theta^\star)) + 3\sqrt{m}(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell)\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|^2\right) \\
&+ \left(\frac{2\sqrt{2}}{(1-\lambda-\rho)\sqrt{\mu L}}\left(9m\sqrt{m}\rho\left(\left(3+\frac{2}{1-\rho}\right)\tau_\ell + \tau_L\right)\right) + \frac{4}{\mu}\tau_\mu\right) \\
&\times 2\left(2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^\star)) + 2\mathcal{R}(\theta^\star - \widehat{\theta}) + \Psi(\bar{\mathcal{M}})\|\theta^\star - \widehat{\theta}\|\right)^2
\end{aligned} \tag{38}$$

where B and λ are defined in (33) and (34), respectively.

Proof. Combining Proposition 1 with Lemma 2 and Proposition 2 yields

$$\begin{aligned}
&\boldsymbol{F}(\mathbf{W}\boldsymbol{\theta}^{t+1}) - \boldsymbol{F}(\mathbf{W}\widehat{\boldsymbol{\theta}}) + \underbrace{\frac{\beta^2\alpha}{2m}}_{\triangleq q'_\theta}\|\boldsymbol{\theta}^{t+1}\|^2_{\mathbf{I}-\mathbf{W}^2} + \underbrace{\left(\frac{\beta^2\alpha}{2m} - \frac{\beta A(\rho)}{2m} - \frac{\beta^2 24\tau_L\Psi^2(\bar{\mathcal{M}})}{2m}\right)}_{\triangleq q'_v}\|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 \\
&+ \underbrace{\left(\frac{\beta\mu}{2m} - \frac{\beta^3\alpha}{2m} - \frac{\beta}{2m}8\tau_\mu\Psi^2(\bar{\mathcal{M}}) - \frac{24\tau_L\beta^4}{2m}\Psi^2(\bar{\mathcal{M}}) - \frac{\beta A(\rho)}{2m}\right)}_{\triangleq q'_z}\|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2 \leq \\
&(1-\beta)\left(\boldsymbol{F}(\mathbf{W}\boldsymbol{\theta}^t) - \boldsymbol{F}(\mathbf{W}\widehat{\boldsymbol{\theta}})\right) + \underbrace{\left((1-\beta)\frac{\beta^2\alpha}{2m} + \frac{B(\rho)\beta}{2m}\right)}_{\triangleq q''_\theta}\|\boldsymbol{\theta}^t\|^2_{\mathbf{I}-\mathbf{W}^2} \\
&+ \underbrace{\left(\frac{(1-\beta)\beta^2\alpha}{2m} + \frac{24\tau_L\Psi^2(\bar{\mathcal{M}})}{2m}\beta^2(1-\beta)^2 + \frac{B(\rho)\beta^2}{2m}\right)}_{\triangleq q''_v}\|\mathbf{v}^t - \widehat{\boldsymbol{\theta}}\|^2 \\
&+ \underbrace{\frac{\beta R(\rho)}{2m}}_{\triangleq q''_z}\|\mathbf{z}^{t-1} - \widehat{\boldsymbol{\theta}}\|^2 + \underbrace{\frac{\beta R(\rho)}{2m}\sum_{i=0}^{t-2}\rho^{t-i-1}\|\mathbf{z}^i - \widehat{\boldsymbol{\theta}}\|^2 + \frac{\beta C}{2m}\rho^{t+1} + (\beta C(\rho) + \frac{\beta}{2}(\tau_\mu + 9\tau_L\beta))\nu^2}_{\triangleq \delta^t}.
\end{aligned} \tag{39}$$

Introducing

$$\begin{aligned} B^{t+1} &\triangleq \boldsymbol{F}(\mathbf{W}\boldsymbol{\theta}^{t+1}) - \boldsymbol{F}(\widehat{\boldsymbol{\theta}}) + q'_\theta \|\boldsymbol{\theta}^{t+1}\|^2_{\mathbf{I}-\mathbf{W}^2} + q'_v \|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 + q'_z \|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2 \\ &= \boldsymbol{g}(\boldsymbol{\theta}^{t+1}) + q'_v \|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 + q'_z \|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2, \end{aligned} \tag{40}$$

and

$$\bar{\lambda} = \max\left\{1-\beta, \frac{q''_\theta}{q'_\theta}, \frac{q''_v}{q'_v}, \frac{q''_z}{q'_z}\right\}, \quad \text{for} \quad q'_z,\, q'_\theta,\, q'_v > 0, \tag{41}$$

(39) yields

$$B^{t+1} \leq \bar{\lambda} B^t + \delta^t.$$

We will enforce shortly conditions on the system parameters via q'_θ, q'_v, q'_z and q''_θ, q''_v, q''_z so that $\bar{\lambda} \in (0,1)$. Hence, hereafter, we will assume $\bar{\lambda} \in (0,1)$.

Using $q''_z \leq \bar{\lambda}\, q'_z$, due to (41), and the expression of δ^t (see (39)), we can write

$$B^{t+1} \leq \bar{\lambda} B^t + q'_z\, \bar{\lambda} \sum_{i=0}^{t-2} \rho^{t-i-1} \|\mathbf{z}^i - \widehat{\boldsymbol{\theta}}\|^2 + \frac{\beta C}{2m}\rho^{t+1} + \left(\beta C(\rho) + \frac{\beta}{2}\left(\tau_\mu + 9\tau_L \beta\right)\right)\nu^2. \tag{42}$$

To study the dynamics of (42), define the sequence $\{S^t\}$ as

$$S^t = \rho\, S^{t-1} + \sqrt{\rho\bar{\lambda}} B^{t-1}, \quad t = 1, 2 \ldots; \quad \text{and} \quad S^0 = 0.$$

Notice that

$$S^t = \sqrt{\rho\bar{\lambda}} \sum_{i=0}^{t-1} \rho^{t-1-i} B^i, \quad t = 1, 2, \ldots.$$

Consequently, we can write: for any $t = 0, 1, \ldots,$

$$\begin{aligned} B^{t+1} &\leq \bar{\lambda} B^t + \sqrt{\rho\bar{\lambda}} S^t + \frac{\beta C}{2m}\rho^{t+1} + \left(\beta C(\rho) + \frac{\beta}{2}(\tau_\mu + \tau_L\beta)\right)\nu^2, \\ S^{t+1} &\leq \rho S^t + \sqrt{\bar{\lambda}\rho} B^t; \end{aligned}$$

and

$$\left\| \begin{matrix} B^{t+1} \\ S^{t+1} \end{matrix} \right\| \leq \left\| \begin{pmatrix} \bar{\lambda} & \sqrt{\bar{\lambda}\rho} \\ \sqrt{\bar{\lambda}\rho} & \rho \end{pmatrix} \right\|_2 \left\| \begin{matrix} B^t \\ S^t \end{matrix} \right\| + \frac{\beta C}{2m}\rho^t + \left(\beta C(\rho) + \frac{\beta}{2}(\tau_\mu + 9\tau_L\beta)\right)\nu^2.$$

The matrix above has spectral norm equal to $\bar{\lambda} + \rho$. Consequently, under

$$\bar{\lambda} + \rho < 1, \tag{43}$$

by telescoping back to $t = 0$ while using $S^0 = 0$, we obtain

$$\begin{aligned} B^{t+1} &\leq \left(\bar{\lambda} + \rho\right)^{t+1} B^0 + \frac{\beta C}{2m} \sum_{i=0}^{t} \rho^{i+1} (\bar{\lambda} + \rho)^{t-i} + \frac{1}{1 - \bar{\lambda} - \rho}\left(\beta C(\rho) + \frac{\beta}{2}(\tau_\mu + 9\tau_L\beta)\right)\nu^2 \\ &\leq (\bar{\lambda} + \rho)^{t+1} B^0 + \frac{\beta C}{2m}(\bar{\lambda} + \rho)^{t+1} \frac{\rho}{\bar{\lambda}} + \frac{1}{1 - \bar{\lambda} - \rho}\left(\beta C(\rho) + \frac{\beta}{2}(\tau_\mu + 9\tau_L\beta)\right)\nu^2. \end{aligned} \tag{44}$$

Using (40), we can lower bound B^{t+1} as

$$\begin{aligned} B^{t+1} &\geq \boldsymbol{g}(\boldsymbol{\theta}^{t+1}) - \boldsymbol{g}(\widehat{\boldsymbol{\theta}}) \\ &\overset{(a)}{\geq} \frac{\mu}{4m}\|\boldsymbol{\theta}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2_{\mathbf{W}^2} + \frac{\beta^2\alpha}{2m}\|\boldsymbol{\theta}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2_{\mathbf{I}-\mathbf{W}^2} - \varepsilon_\ell(\boldsymbol{\theta}^{t+1}, \widehat{\boldsymbol{\theta}}) \\ &\overset{(30)}{\geq} \frac{\beta^2\alpha}{2m}\|\boldsymbol{\theta}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 + \left(\frac{\mu}{4m} - \frac{\tau_\mu}{2m} 8\Psi^2(\bar{\mathcal{M}}) - \frac{\beta^2\alpha}{2m}\right)\|\boldsymbol{\theta}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2_{\mathbf{W}^2} - \frac{\tau_\mu}{2}\nu^2, \end{aligned} \tag{45}$$

where (a) follows from (25a) and $\langle \nabla \boldsymbol{g}(\widehat{\boldsymbol{\theta}}), \boldsymbol{\theta}^{t+1} - \widehat{\boldsymbol{\theta}} \rangle \geq 0$.

Chaining (44) and (45), we finally obtain

$$\begin{aligned}
&\|\boldsymbol{\theta}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 + \left(\frac{\mu}{2\beta^2\alpha} - \frac{\tau_\mu}{\beta^2\alpha^2} 8\Psi^2(\bar{\mathcal{M}}) - 1 \right) \|\boldsymbol{\theta}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2_{\mathbf{W}^2} \leq \frac{2m}{\beta^2\alpha} (\bar{\lambda} + \rho)^{t+1} V^0 \\
&+ (\bar{\lambda} + \rho)^{t+1} \frac{\rho}{\bar{\lambda}\beta\alpha} \left(4R\sqrt{m}m \sum_{i=1}^{m} \mathcal{R}^*(\nabla f_i(\theta^\star)) + 3m\sqrt{m}(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell) \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|^2 \right) \\
&+ \frac{2m}{\beta\alpha(1 - \bar{\lambda} - \rho)} \left(9m\sqrt{m}\rho \left(\left(3 + \frac{2}{1-\rho} \right) \tau_\ell + \tau_L \right) + \frac{\tau_\mu + 9\tau_L}{2} \right) \nu^2 + \frac{m}{\beta^2\alpha} \frac{\tau_\mu}{2} \nu^2.
\end{aligned} \tag{46}$$

We proceed now to establish sufficient conditions for $\bar{\lambda}$ in (41) to be strictly less than one as well as for (43) to hold, which have been subsumed in the above derivations.

Substituting the expressions of q'_θ, q'_v, q'_z and q''_θ, q''_v, q''_z (as in (39)), in (41), we have

$$\begin{aligned}
&\frac{\beta^2\alpha}{2m} > 0, \\
&\frac{\beta^2\alpha}{2m} - \frac{\beta A(\rho)}{2m} - \frac{\beta^2 24\tau_L \Psi^2(\bar{\mathcal{M}})}{2m} > 0, \\
&\frac{\beta\mu}{2m} - \frac{\beta^3\alpha}{2m} - \frac{\beta}{2m} 8\tau_\mu \Psi^2(\bar{\mathcal{M}}) - \frac{24\tau_L\beta^4\Psi^2(\bar{\mathcal{M}})}{2m} - \frac{\beta A(\rho)}{2m} > 0, \\
&\bar{\lambda} = \max\Bigg\{ 1 - \beta, \frac{(1-\beta)\beta^2\alpha + B(\rho)\beta}{\beta^2\alpha}, \frac{(1-\beta)\beta^2\alpha + 24\tau_L\Psi^2(\bar{\mathcal{M}})\beta^2(1-\beta)^2 + B(\rho)\beta^2}{\beta^2\alpha - \beta A(\rho) - \beta^2 24\tau_L\Psi^2(\bar{\mathcal{M}})}, \\
&\qquad \frac{\beta R(\rho)}{\beta\mu - \beta^3\alpha - \beta 8\tau_\mu\Psi^2(\bar{\mathcal{M}}) - 24\tau_L\beta^4\Psi^2(\bar{\mathcal{M}}) - \beta A(\rho)} \Bigg\}.
\end{aligned}$$

Sufficient conditions are

$$\beta\alpha - A(\rho) - \beta 24\tau_L\Psi^2(\bar{\mathcal{M}}) > 0, \tag{47a}$$

$$\mu - \beta^2\alpha - 8(\tau_\mu + 24\beta^3\tau_L)\Psi^2(\bar{\mathcal{M}}) - A(\rho) > 0, \tag{47b}$$

$$\begin{aligned}
\bar{\lambda} = \max\Bigg\{ &1 - \beta + \frac{(1-\beta)(\beta A(\rho) + \beta^2 24\tau_L\Psi^2(\bar{\mathcal{M}})) + 24\Psi^2(\bar{\mathcal{M}})\beta^2(1-\beta)^2 + B(\rho)\beta}{\beta^2\alpha - \beta A(\rho) - \beta^2 24\tau_L\Psi^2(\bar{\mathcal{M}})}, \\
&\frac{R(\rho)}{\mu - \beta^2\alpha - 8\psi^2(\bar{\mathcal{M}}) - 24\tau_L\beta^3\Psi^2(\bar{\mathcal{M}}) - \beta A(\rho)} \Bigg\}.
\end{aligned} \tag{47c}$$

By enforcing

$$\frac{R(\rho)}{\mu - \beta^2\alpha - 8\tau_\mu\Psi^2(\bar{\mathcal{M}}) - 24\tau_L\beta^3\Psi^2(\bar{\mathcal{M}}) - A(\rho)} \leq 1 - \beta, \tag{48}$$

we have

$$\bar{\lambda} \leq 1 - \beta + \frac{(1-\beta)A(\rho) + B(\rho) + 48\beta(1-\beta)\tau_L\Psi^2(\bar{\mathcal{M}})}{\beta\alpha - A(\rho) - 24\beta\tau_L\Psi^2(\bar{\mathcal{M}})}. \tag{49}$$

For (48) to hold it is sufficient that

$$\mu - \beta^2\alpha - 8\tau_\mu\Psi^2(\bar{\mathcal{M}}) - 24\tau_L\beta^3\Psi^2(\bar{\mathcal{M}}) - A(\rho) \geq \frac{\mu}{2}, \tag{50a}$$

$$R(\rho) \leq \frac{\mu}{2}(1-\beta). \tag{50b}$$

Notice that (50a) also implies (47b).

Choose

$$\beta^2 = \frac{\mu}{8L}. \tag{51}$$

Sufficient conditions for (50) to hold are

$$\frac{\mu}{4} \geq \max\left\{ 8\tau_\mu\Psi^2(\bar{\mathcal{M}}) + 24\tau_L\Psi^2(\bar{\mathcal{M}}), A(\rho), R(\rho) \right\}. \tag{52}$$

Using now the expression of $A(\rho)$ and $R(\rho)$ (see (32)) and enforcing

$$\rho \leq \frac{1}{2}, \tag{53}$$

(52) is implied by

$$\beta^2\alpha\rho + 3m\sqrt{m}\rho\left(7(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell) + 2L + 8\Psi^2(\bar{\mathcal{M}})\tau_L\right) \leq \frac{\mu}{4}, \tag{54a}$$

$$8\tau_\mu\Psi^2(\bar{\mathcal{M}}) + 24\tau_L\Psi^2(\bar{\mathcal{M}}) \leq \frac{\mu}{4}. \tag{54b}$$

The following additional condition on ρ is sufficient for (54a):

$$\rho \leq \frac{\mu}{12m\sqrt{m}\left(14\gamma_\ell + 56\Psi^2(\bar{\mathcal{M}})\tau_\ell + 2L + 8\Psi^2(\bar{\mathcal{M}})\tau_L\right) + \mu}. \tag{55}$$

Observe that under (52), β given by (51) and $\alpha \geq 2L$, the LHS of (47a) is lower bounded by $\sqrt{\frac{\mu L}{2}} - \frac{\mu}{4} - \sqrt{\frac{\mu}{2L}}\frac{\mu}{8}$; therefore (47a) is fulfilled, for any values of μ and L.

It remains to show that (43) holds. We claim that, using therein the upper bound of $\bar{\lambda}$ in (49), with β given by (51), (43) holds under

$$24\tau_L\Psi^2(\bar{\mathcal{M}}) \leq \frac{\mu}{8}, \tag{56}$$

as long as ρ satisfies (36). In fact, first, notice that using definition of the parameters $A(\rho)$ and $B(\rho)$ it holds

$$\bar{\lambda} \leq \lambda,$$

where λ is given by (34). Then, substituting (36) in (35) and using $\tau(\rho) \leq \frac{\mu}{128}$, we deduce

$$\bar{\lambda} \leq \lambda \leq 1 - \frac{7}{8}\beta + \beta^2(1-\beta) \leq 1 - \frac{7-\sqrt{8}}{8}\beta,$$

which proves (43).

Finally, notice that (36) is sufficient for (53) and (55).

To summarize, we proved (46), under (36) and (54b), (56), with (54b), (56) implied by (37). The final convergence as in (38) follows from (46), lower bounding the LHS by $\frac{1}{m}\|\boldsymbol{\theta}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2$ and using (52) and $\bar{\lambda}^{-1} \leq 2$. □

Remark 1 (Theorem 2 as special case of Theorem 3)**.** *Using $\tau(\rho) \leq \frac{\mu}{128}$, we can bound λ in* (34) *as*

$$\lambda \leq 1 - \sqrt{\frac{1}{8\kappa}} + \left(1 - \sqrt{\frac{1}{8\kappa}}\right)\frac{48\tau_L\Psi^2(\bar{\mathcal{M}})}{\frac{127}{64}L - 24\tau_L\Psi^2(\bar{\mathcal{M}})} + \frac{\sqrt{8\mu L}/64}{\frac{127}{64}L - 24\tau_L\Psi^2(\bar{\mathcal{M}})}.$$

Under $24\tau_L\Psi^2(\bar{\mathcal{M}}) \leq \frac{\sqrt{2}\mu}{14}$ and $\rho \leq \frac{\sqrt{\kappa^{-1}}}{18}$, one gets

$$\lambda + \rho \leq 1 - \sqrt{\frac{1}{32\kappa}}.$$

This yields Theorem 2.

A.5 Proof of Theorem 1

Theorem 1 is proved as special case of the theorem below–Remark 2 to follow elaborate on the connection.

Theorem 4. *Given the ERM* (3)*, suppose (i) f satisfies the RSC/RSM conditions (Assumption 1); (ii) $\mathcal{R}$ is a decomposable regularizer with respect to a chosen pair of subspaces $(\mathcal{M}, \bar{\mathcal{M}})$ (Assumption 3); (iii) $(\bar{\mathcal{M}}, \bar{\mathcal{M}})$ and the RSC/RSM parameters (μ, τ_μ) and (L, τ_L) are such that*

$$\frac{\mu}{8} \geq 8\tau_\mu\Psi^2(\bar{\mathcal{M}}) + 24\tau_L\Psi^2(\bar{\mathcal{M}}). \tag{57}$$

Let $\{\theta^t\}_{t\geq 1}$ be the sequence generated by the algorithm in (9) *with tuning $\alpha = 2L$ and $\beta = \sqrt{1/(8\kappa)}$. Then, it holds*

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \frac{8}{\mu}\lambda^{t+1}\left(f(\theta^0) - f(\widehat{\theta}) + \frac{\mu}{8}\|v^0 - \widehat{\theta}\|^2\right) \tag{58}$$
$$+ \frac{4}{\mu}\left(\frac{1}{1-\lambda}(\beta\tau_\mu + 9\tau_L\beta^2) + \tau_\mu\right) \times 2\left(2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^\star)) + 2\mathcal{R}(\theta^\star - \widehat{\theta}) + \Psi(\bar{\mathcal{M}})\|\widehat{\theta} - \theta^\star\|\right)^2,$$

with

$$\lambda \triangleq \frac{1 - \sqrt{\frac{1}{8\kappa}} + \left(1 - \sqrt{\frac{1}{8\kappa}}\right)\left(\frac{24\Psi^2(\bar{\mathcal{M}})\tau_L}{2L}\right)}{1 - \frac{24\tau_L\Psi^2(\bar{\mathcal{M}})}{2L}}.$$

Proof. We apply Lemma 1, with $\boldsymbol{g} = f$; we can write

$$f(\theta^{t+1}) - f(\widehat{\theta}) + \frac{\beta^2\alpha}{2}\|v^{t+1} - \widehat{\theta}\|^2 \leq (1-\beta)(f(\theta^t) - f(\widehat{\theta})) + \frac{(1-\beta)\beta^2\alpha}{2}\|v^t - \widehat{\theta}\|^2$$
$$- \left(\frac{\beta\mu}{4} - \frac{\beta^3\alpha}{2}\right)\|\widehat{\theta} - z^t\|^2 + \beta\frac{\tau_\mu}{2}\mathcal{R}^(\widehat{\theta} - z^t) + \varepsilon_u\frac{\tau_L}{2}\mathcal{R}^2(\theta^{t+1} - z^t).$$

Using Lemma 2 we have

$$\frac{\tau_\mu}{2}\mathcal{R}^2(\widehat{\theta} - z^t) \leq \frac{\tau_\mu}{2}\left(8\Psi^2(\bar{\mathcal{M}})\|\widehat{\theta} - z^t\|^2 + \nu^2\right),$$
$$\frac{\tau_L}{2}\mathcal{R}^2(\theta^{t+1} - z^t) \leq \frac{\tau_L}{2}\left(24\beta^2\Psi^2(\bar{\mathcal{M}})\left(\|v^{t+1} - \widehat{\theta}\|^2 + (1-\beta)^2\|v^t - \widehat{\theta}\|^2 + \beta^2\|z^t - \widehat{\theta}\|^2\right) + 9\beta^2\nu^2\right).$$

Therefore,

$$f(x^{t+1}) - f(\widehat{\theta}) + \left(\frac{\beta^2\alpha}{2} - \frac{24\tau_L}{2}\beta^2\Psi^2(\bar{\mathcal{M}})\right)\|v^{t+1} - \widehat{\theta}\|^2$$
$$\leq (1-\beta)(f(x^t) - f(\widehat{\theta})) + \left(\frac{(1-\beta)\beta^2\alpha}{2} + \frac{24\tau_L\beta^2}{2}(1-\beta)^2\Psi^2(\bar{\mathcal{M}})\right)\|v^t - \widehat{\theta}\|^2 + \left(\frac{\beta\tau_\mu}{2} + \frac{9\tau_L}{2}\beta^2\right)\nu^2$$
$$- \left(\frac{\beta\mu}{4} - \frac{\beta^3\alpha}{2} - \frac{24\tau_L\beta^4\Psi^2(\bar{\mathcal{M}})}{2} - \frac{\beta\tau_\mu}{2}8\Psi^2(\bar{\mathcal{M}})\right)\|z^t - \widehat{\theta}\|^2.$$

Enforcing

$$\beta^2\alpha + 24\tau_L\beta^3\Psi^2(\bar{\mathcal{M}}) + 8\Psi^2(\bar{\mathcal{M}})\tau_\mu \leq \frac{\mu}{2}, \tag{59}$$
$$(1-\beta)\beta^2\alpha + 24\tau_L\beta^2(1-\beta)^2\Psi^2(\bar{\mathcal{M}}) \leq \bar{\lambda}\underbrace{\left(\beta^2\alpha - 24\tau_L\beta^2\Psi^2(\bar{\mathcal{M}})\right)}_{\triangleq 2q_v}, \tag{60}$$

for some $\bar{\lambda} \in [1-\beta, 1)$ (to be determined), yields

$$f(\theta^{t+1}) - f(\widehat{\theta}) + q_v\|v^{t+1} - \widehat{\theta}\|^2 \leq \bar{\lambda}\left(f(x^t) - f(\widehat{\theta}) + q_v\|v^t - \widehat{\theta}\|^2\right) + \left(\frac{\beta\tau_\mu}{2} + \frac{9\tau_L}{2}\beta^2\right)\nu^2,$$

whereby telescoping we have: for $t = 0, 1, \ldots,$

$$f(\theta^{t+1}) - f(\widehat{\theta}) \leq \bar{\lambda}^{t+1}\left(f(\theta^0) - f(\widehat{\theta}) + q_v\|v^0 - \widehat{\theta}\|^2\right) + \frac{1}{1-\bar{\lambda}}\left(\frac{\beta\tau_\mu}{2} + \frac{9\tau_L}{2}\beta^2\right)\nu^2. \tag{61}$$

Invoking the RSC of f and using optimimality of $\widehat{\theta}$, we can write

$$f(\theta^{t+1}) - f(\widehat{\theta}) \geq \frac{\mu}{4}\|\theta^{t+1} - \widehat{\theta}\|^2 - \frac{\tau_\mu}{2}\mathcal{R}^2(\theta^{t+1} - \widehat{\theta})$$
$$\overset{(30)}{\geq} \frac{\mu}{4}\|\theta^{t+1} - \widehat{\theta}\|^2 - \frac{\tau_\mu}{2}\left(8\Psi^2(\bar{\mathcal{M}})\|\theta^{t+1} - \widehat{\theta}\|^2 + \nu^2\right). \tag{62}$$

Chaining (61) and (62) under

$$8\tau_\mu \Psi^2(\bar{\mathcal{M}}) \leq \frac{\mu}{4}, \tag{63}$$

and using $q_v \leq \frac{\beta^2\alpha}{2}$, we obtain

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \frac{8}{\mu}\lambda^{t+1}\left(f(\theta^0) - f(\widehat{\theta}) + \frac{\beta^2\alpha}{2}\|v^0 - \widehat{\theta}\|^2\right) + \frac{8}{\mu}\left(\frac{1}{1-\lambda}\left(\frac{\beta\tau_\mu}{2} + \frac{9\tau_L}{2}\beta^2\right) + \frac{\tau_\mu}{2}\right)\nu^2.$$

Observe that the smallest $\bar{\lambda}$ one can choose while satisfying (60) is

$$\bar{\lambda} = \max\left\{1-\beta, \frac{(1-\beta)\alpha + 24\tau_L(1-\beta)^2\Psi^2(\bar{\mathcal{M}})}{\alpha - 24\tau_L\Psi^2(\bar{\mathcal{M}})}\right\},$$

and thus

$$\bar{\lambda} = 1-\beta + \frac{24\tau_L((1-\beta)^2 + (1-\beta))\Psi^2(\bar{\mathcal{M}})}{\alpha - 24\tau_L\Psi^2(\bar{\mathcal{M}})} \leq \frac{1-\beta + (1-\beta)\left(\frac{24\Psi^2(\bar{\mathcal{M}})\tau_L}{2L}\right)}{1 - \frac{24\tau_L\Psi^2(\bar{\mathcal{M}})}{2L}}.$$

To summarize, the above convergence has been achieved under (63) and (59), implied by (57). The upper bound on the rate as above is compliant with (60). Notice that (57) also guarantees that the RHS of (60) is strictly positive. □

Remark 2. *By further assuming that* $24\Psi^2(\bar{\mathcal{M}})\tau_L \leq \frac{\sqrt{2}\mu}{14}$ *it holds that*

$$\lambda \leq 1 - \sqrt{\frac{1}{16\kappa}}.$$

This yields the result in Theorem 1.

A.6 Proofs of Corollaries 1-4

A.6.1 Proof of Corollary 1

In the setting of Corollary 1 the following holds

$$\Psi^2(\bar{\mathcal{M}}) \leq \left(\frac{\log d}{N}\right)^{-\frac{q}{2}} R_q, \qquad \|\Pi_{\mathcal{M}^\perp}(\theta^\star)\|_1 \leq \left(\frac{\log d}{N}\right)^{\frac{1-q}{2}} R_q, \tag{64a}$$

$$\gamma_\ell = (2m+1)L, \qquad \tau_\ell = \eta_\Sigma c_0 \frac{\log d}{n}, \tag{64b}$$

$$\mu = \lambda_{\min}(\Sigma), \qquad L = \lambda_{\max}(\Sigma), \tag{64c}$$

$$\tau_\mu = \tau_L = \eta_\Sigma c_0 \frac{\log d}{N}, \qquad \max_{i\in[m]} \|\nabla f_i(\theta^\star)\|_1 \leq c_1\sqrt{m\eta_\sigma\sigma^2\frac{log(md)}{N}}, \tag{64d}$$

with probability at least

$$1 - c_2 \exp\left(-c_3 N + \log(m)\right),$$

for some universal constants c_0, c_1, c_2, $c_3 > 0$ and all $q \in [0,1]$. The proof of the above result follows from [26], [11] with minor modifications, hence it is omitted.

Invoking Theorem 3 given that

$$R_q c_4 \eta_\Sigma \left(\frac{\log d}{N}\right)^{1-\frac{q}{2}} \leq \mu,$$

and the network

$$\rho \leq \frac{\kappa^{-1}}{384m\sqrt{m}(L^{-1}(16(2m+1)) + c_5(m+1)\mu + 3)},$$

for some universal constants $C_4,\ c_5 > 0$, with the parameters (64) yields

$$\frac{1}{m}\|\boldsymbol{\theta}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 \leq \frac{8}{\mu}\left(1 - \sqrt{\frac{1}{32\kappa}}\right)^{t+1}\left(f(0) - f(\widehat{\theta}) + \frac{\mu}{16m}\|\widehat{\boldsymbol{\theta}}\|^2 + \frac{\beta\mu}{2m}\|\widehat{\boldsymbol{\theta}}\|^2\right)$$
$$+ \frac{2\sqrt{2}}{\mu}\left(1 - \sqrt{\frac{1}{32\kappa}}\right)^{t+1}\frac{\kappa^{-3/2}}{96m\sqrt{m}\,(16(2m+1))}\times$$
$$\left(4R\sqrt{m}m\sqrt{c_2 m\eta_\Sigma\sigma^2\frac{\log(md)}{N}} + 3\sqrt{m}(2(2m+1)L + 8\mu)\|\boldsymbol{\theta}^\star - \widehat{\boldsymbol{\theta}}\|^2\right)$$
$$+ 24\left(\frac{2\sqrt{2}}{\mu}\frac{\kappa^{-3/2}}{96m\sqrt{m}((16(2m+1)))}9m\sqrt{m}\left(c_0\eta_\Sigma\frac{\log d}{N}(7m+1)\right) + \frac{4}{\mu}c_0\eta_\Sigma\frac{\log d}{N}\right)\times$$
$$\times\left(R_q^2\left(\frac{\log d}{N}\right)^{1-q} + \mathcal{R}^2(\theta^\star - \widehat{\theta}) + \left(\frac{\log d}{N}\right)^{-\frac{q}{2}}R_q\|\theta^\star - \widehat{\theta}\|^2\right),$$

whereby noticing that $\kappa^{-1} \leq 1$ and invoking Lemma 5 in the supplementary material of [1] which states

$$\mathcal{R}(\widehat{\theta} - \theta^\star) \leq 2\Psi(\bar{\mathcal{M}})\|\widehat{\theta} - \theta^\star\| + \mathcal{R}\left(\Pi_{\mathcal{M}^\perp}(\theta^\star)\right)$$

yields the desired result. □

A.6.2 Proof of Corollary 2

In the setting of Corollary 1 we have that for $q \in [0, 1]$

$$\Psi^2(\bar{\mathcal{M}}) \leq R_q\left(\frac{p}{N}\right)^{-\frac{q}{2}}, \qquad \|\Pi_{\mathcal{M}^\perp}(\Theta^\star)\|_1 \leq R_q\left(\frac{p}{N}\right)^{\frac{1-q}{2}},$$
$$\gamma_\ell = (2m+1)L \qquad \tau_\ell = c_0\eta_\Sigma\frac{p}{n},$$
$$\mu = \lambda_{\min}(\Sigma) \qquad L = \lambda_{\max}(\Sigma),$$
$$\tau_\mu = \tau_L = c_0\eta_\Sigma\frac{\log d}{N}, \qquad \max_{i\in[m]}\|\nabla f_i(\Theta^\star)\| \leq c_1\sigma\eta_\Sigma^{1/2}m\sqrt{\frac{p}{N}}.$$

The results above follow from small modifications of [1], [29], [26] and applying the union bound. Then, the desired result follows by taking the same steps as in the proof of Corollary 1. □

A.6.3 Proof of Corollary 3

In the setting of Corollary 3 we have that for $q \in [0, 1]$

$$\Psi^2(\bar{\mathcal{M}}) \leq \left(\frac{\log d}{N}\right)^{-\frac{q}{2}}R_q, \qquad \|\Pi_{\mathcal{M}^\perp}(\theta^\star)\|_1 \leq \left(\frac{\log d}{N}\right)^{\frac{1-q}{2}}R_q,$$
$$\gamma_\ell = \frac{\lambda_{\max}(\Sigma)}{2} + \frac{m\lambda_{\min}(\Sigma)}{2}, \qquad \tau_\ell = c_0\max\left\{\frac{\tau^4}{\lambda_{\min}(\Sigma)}, \lambda_{\min}(\Sigma)\right\}\frac{\log d}{n},$$
$$\mu = \lambda_{\min}(\Sigma), \qquad L = \frac{\mu}{2} + \frac{\lambda_{\max}(\Sigma)}{2},$$
$$\tau_\mu = \tau_L = c_0\max\left\{\frac{\tau^4}{\lambda_{\min}(\Sigma)}, \lambda_{\min}(\Sigma)\right\}\frac{\log d}{N}, \quad \max_{i\in[m]}\|\nabla f_i(\theta^\star)\|_\infty \leq Cc_1\sqrt{\frac{\sigma m\log(md)}{N}},$$

with the probability given in the Corollary's statement. The above follow from small modifications of the statistical results in [17], [11], and using the union bound. Then, following the same steps as in the proof of Corollary 1 yields the desired result. □

A.6.4 Proof of Corollary 4

The matrix completion case deviates slightly from the other cases as the empirical risk does not fulfill exactly Assumptions 1 and 2, but a slightly modified version [1]. Observe that this is consistent with the results obtained for plain projected gradient descent in [1]. For the described model it holds that for all $i \in [m]$ with probability at least

$$1 - \exp\left(-p \log p\right) - c_1 m \exp(-p \log p),$$

and for all $V \in \mathbb{R}^{p \times p}$ and $V_i \in \mathbb{R}^{p \times p}$, it holds

$$\left\langle \mathrm{vec}(V), \frac{XX^\top}{N} \mathrm{vec}(V) \right\rangle \leq \|V\|_F^2 + c_2 p \|V\|_\infty \|V\|_1 \sqrt{\frac{p \log p}{N}} + c_3 \left(p \|V\|_\infty \sqrt{\frac{p \log p}{N}} \right)^2, \tag{65a}$$

$$\left\langle \mathrm{vec}(V), \frac{XX^\top}{N} \mathrm{vec}(V) \right\rangle \geq \|V\|_F^2 - c_2 p \|V\|_\infty \|V\|_1 \sqrt{\frac{p \log p}{N}} + c_3 \left(p \|V\|_\infty \sqrt{\frac{p \log p}{N}} \right)^2, \tag{65b}$$

$$\left\langle \mathrm{vec}(V_i), \frac{X_{\mathcal{S}_i} X_{\mathcal{S}_i}^\top}{n} \mathrm{vec}(V_i) \right\rangle \leq \|V_i\|_F^2 + c_2 p m \|V_i\|_\infty \|V_i\|_1 \sqrt{\frac{p \log p}{N}} + c_3 \left(p m \|V\|_\infty \sqrt{\frac{p \log p}{N}} \right)^2, \tag{65c}$$

$$\left\langle \mathrm{vec}(V_i), \frac{X_{\mathcal{S}_i} X_{\mathcal{S}_i}^\top}{n} \mathrm{vec}(V_i) \right\rangle \geq \|V_i\|_F^2 - c_2 p m \|V_i\|_\infty \|V_i\|_1 \sqrt{\frac{p \log p}{N}} - c_3 \left(p m \|V\|_\infty \sqrt{\frac{p \log p}{N}} \right)^2, \tag{65d}$$

where the statements involving X are a restatement of Proposition 2 in the supplementary material of [1] and the remaining can be established by combining the proof of Proposition 2 in [1], the strategy employed to establish the RSM in [26] and the union bound. By setting $V = \Theta_1 - \Theta_2$ such that both $\Theta_1,\ \Theta_2 \in \Omega$ one can establish

$$\begin{aligned} &f(\Theta_1) - f(\Theta_2) - \langle \nabla f(\Theta_2), \Theta_1 - \Theta_2 \rangle \leq \|\Theta_1 - \Theta_2\|_F^2 \\ &+ c_2 p \|\Theta_1 - \Theta_2\|_\infty \|\Theta_1 - \Theta_2\|_1 \sqrt{\frac{p \log p}{N}} + c_3 \left(p \|\Theta_1 - \Theta_2\|_\infty \sqrt{\frac{p \log p}{N}}, \right)^2 \end{aligned}$$

and an analogous lower bound. Observe that due to belonging to Ω it holds that

$$\|\Theta_1 - \Theta_2\|_\infty \leq 2 \frac{\omega}{p},$$

implying that the RSC and RSS hold upto a tolerance $c_3 \omega^2 \frac{p \log p}{N}$. For such a model, the procedure to establish convergence is identical to that in the proof of Theorem 3 with the difference that the additional tolerance needs to be absorbed with the misspecification error terms (ν^2).

The remaining terms are given by

$$\begin{aligned} &\Psi^2(\bar{\mathcal{M}}) \leq R_q \omega^{-q} \left(\frac{p \log p}{N} \right)^{-\frac{q}{2}}, \\ &\|\Pi_{\mathcal{M}^\perp}(\Theta^\star)\|_1 \leq R_q \omega^{1-q} \left(\frac{p \log p}{N} \right)^{\frac{1-q}{2}}, \\ &\max_{i \in [m]} \mathcal{R}^* \left(\nabla f_i(\Theta^\star) \right) \leq c_1 m \sigma \sqrt{\frac{p \log p}{N}}, \end{aligned}$$

which can be found in [1],[29].

A.7 Supporting results in the proof of Theorem 3

A.7.1 Proof of Proposition 1

From (25b) it follows

$$
\begin{aligned}
\boldsymbol{g}(\boldsymbol{\theta}^{t+1}) &\leq \boldsymbol{g}(\mathbf{w}^t) + \langle \nabla \boldsymbol{g}(\mathbf{w}^t), \boldsymbol{\theta}^{t+1} - \mathbf{w}^t \rangle + \frac{L}{m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{W}^2} + \varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t) \\
&\quad + \frac{\beta^2\alpha}{2m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{I}-\mathbf{W}^2} \\
&\overset{(23e)}{=} (1-\beta)\left(\boldsymbol{g}(\mathbf{w}^t) + \langle \nabla \boldsymbol{g}(\mathbf{w}^t), \boldsymbol{\theta}^t - \mathbf{w}^t \rangle\right) + \beta\left(\boldsymbol{g}(\mathbf{w}^t) + \langle \nabla \boldsymbol{g}(\mathbf{w}^t), \mathbf{v}^{t+1} - \mathbf{w}^t \rangle\right) \\
&\quad + \frac{L}{m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{W}^2} + \frac{\beta^2\alpha}{2m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t) \\
&\leq (1-\beta)\boldsymbol{g}(\boldsymbol{\theta}^t) + \beta\left(\boldsymbol{g}(\mathbf{w}^t) + \langle \nabla \boldsymbol{g}(\mathbf{w}^t), \mathbf{v}^{t+1} - \mathbf{w}^t \rangle\right) \\
&\quad + \frac{L}{m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{W}^2} + \frac{\beta^2\alpha}{2m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t).
\end{aligned} \tag{66}
$$

Using (25a) while subtracting $-\boldsymbol{g}(\widehat{\boldsymbol{\theta}})$ on both sides yields

$$
\begin{aligned}
\boldsymbol{g}(\boldsymbol{\theta}^{t+1}) - \boldsymbol{g}(\widehat{\boldsymbol{\theta}}) &\leq (1-\beta)(\boldsymbol{g}(\boldsymbol{\theta}^t) - \boldsymbol{g}(\widehat{\boldsymbol{\theta}})) + \beta\langle \nabla \boldsymbol{g}(\mathbf{w}^t), \mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}} \rangle \\
&\quad - \frac{\beta\mu}{4m}\|\mathbf{w}^t - \widehat{\boldsymbol{\theta}}\|^2_{\mathbf{W}^2} - \frac{\beta^3\alpha}{2m}\|\mathbf{w}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \beta\varepsilon_\ell(\widehat{\boldsymbol{\theta}}, \mathbf{w}^t) \\
&\quad + \frac{L}{m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{W}^2} + \frac{\beta^2\alpha}{2m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t).
\end{aligned} \tag{67}
$$

Invoking the optimality of $\mathbf{v}^{t+1}$ in (23d), we have

$$
\begin{aligned}
&\langle \nabla \boldsymbol{g}(\mathbf{w}^t), \mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}} \rangle \\
&\leq \left\langle \mathbf{y}^t - \nabla \boldsymbol{g}(\mathbf{w}^t) + \frac{\beta\alpha}{m}\left(\mathbf{v}^{t+1} - (1-\beta)\mathbf{v}^t - \beta\mathbf{z}^t\right), \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1} \right\rangle \\
&= \left\langle \mathbf{y}^t + \frac{\beta^2\alpha}{m}(\mathbf{I}-\mathbf{W})\mathbf{w}^t - \nabla \boldsymbol{g}(\mathbf{w}^t) + \frac{\alpha\beta}{m}\left(\mathbf{v}^{t+1} - (1-\beta)\mathbf{v}^t - \beta\mathbf{w}^t\right), \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1} \right\rangle \\
&\overset{(28)}{=} \left\langle \boldsymbol{\varepsilon}^t + \frac{\alpha\beta}{m}\left(\mathbf{v}^{t+1} - (1-\beta)\mathbf{v}^t - \beta\mathbf{w}^t\right), \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1} \right\rangle.
\end{aligned} \tag{68}
$$

Using (68) in (67) we obtain

$$
\begin{aligned}
\boldsymbol{g}(\boldsymbol{\theta}^{t+1}) - \boldsymbol{g}(\widehat{\boldsymbol{\theta}}) &\leq (1-\beta)\left(\boldsymbol{g}(\boldsymbol{\theta}^t) - \boldsymbol{g}(\widehat{\boldsymbol{\theta}})\right) + \frac{\beta^2\alpha}{m}\langle \mathbf{v}^{t+1} - (1-\beta)\mathbf{v}^t - \beta\mathbf{w}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1} \rangle \\
&\quad + \beta\left\langle \boldsymbol{\varepsilon}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1} \right\rangle - \frac{\beta\mu}{4m}\|\mathbf{w}^t - \widehat{\boldsymbol{\theta}}\|^2_{\mathbf{W}^2} - \frac{\beta^3\alpha}{2m}\|\mathbf{w}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \beta\varepsilon_\ell(\widehat{\boldsymbol{\theta}}, \mathbf{w}^t) \\
&\quad + \frac{L}{m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{W}^2} + \frac{\beta^2\alpha}{2m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t).
\end{aligned} \tag{69}
$$

We use now

$$
\beta\left(\mathbf{v}^{t+1} - (1-\beta)\mathbf{v}^t - \beta\mathbf{w}^t\right) \overset{(23e)}{=} \boldsymbol{\theta}^{t+1} - \mathbf{w}^t \tag{70}
$$

to bound the term

$$
\begin{aligned}
&\frac{\beta^2\alpha}{m}\langle \mathbf{v}^{t+1} - (1-\beta)\mathbf{v}^t - \beta\mathbf{w}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1} \rangle \\
&= \frac{\beta^2\alpha}{2m}\|\widehat{\boldsymbol{\theta}} - (1-\beta)\mathbf{v}^t - \beta\mathbf{w}^t\|^2 - \frac{\beta^2\alpha}{2m}\|\mathbf{v}^{t+1} - (1-\beta)\mathbf{v}^t - \beta\mathbf{w}^t\|^2 - \frac{\beta^2\alpha}{2m}\|\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\|^2 \\
&\overset{(70)}{=} \frac{\beta^2\alpha}{2m}\|\widehat{\boldsymbol{\theta}} - (1-\beta)\mathbf{v}^t - \beta\mathbf{w}^t\|^2 - \frac{\alpha}{2m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2 - \frac{\beta^2\alpha}{2m}\|\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\|^2 \\
&\leq \frac{\beta^3\alpha}{2m}\|\mathbf{w}^t - \widehat{\boldsymbol{\theta}}\|^2 + \frac{(1-\beta)\beta^2\alpha}{2m}\|\widehat{\boldsymbol{\theta}} - \mathbf{v}^t\|^2 - \frac{\alpha}{2m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2 - \frac{\beta^2\alpha}{2m}\|\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\|^2.
\end{aligned} \tag{71}
$$

Combining (69) and (71) yields

$$\begin{aligned}\boldsymbol{g}(\boldsymbol{\theta}^{t+1}) - \boldsymbol{g}(\widehat{\boldsymbol{\theta}}) \leq (1-\beta)\left(\boldsymbol{g}(\boldsymbol{\theta}^t) - \boldsymbol{g}(\widehat{\boldsymbol{\theta}})\right) - \frac{\beta}{2m}\left(\frac{\mu}{2} - \beta^2\alpha\right)\|\mathbf{w}^t - \widehat{\boldsymbol{\theta}}\|^2_{\mathbf{W}^2} \\ + \frac{(1-\beta)\beta^2\alpha}{2m}\|\widehat{\boldsymbol{\theta}} - \mathbf{v}^t\|^2 - \frac{\beta^2\alpha}{2m}\|\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\|^2 - \frac{\alpha}{2m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2 \\ + \frac{L}{m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{W}^2} + \frac{\beta^2\alpha}{2m}\|\boldsymbol{\theta}^{t+1} - \mathbf{w}^t\|^2_{\mathbf{I}-\mathbf{W}^2} \\ + \beta\langle \boldsymbol{\varepsilon}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle + \beta\varepsilon_\ell(\widehat{\boldsymbol{\theta}}, \mathbf{w}^t) + \varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t).\end{aligned}$$

Using $\alpha\mathbf{I} \succeq 2L\mathbf{W}^2 + \beta^2\alpha(\mathbf{I} - \mathbf{W}^2)$, due to $\alpha \geq 2L$, rearranging terms, and using the definitions of V^t, we obtain the desired result as in (27). □

A.7.2 Proof of Lemma 2

Given $\varepsilon_\ell(\widehat{\boldsymbol{\theta}}, \mathbf{w}^t)$ as in (26), we notice that the argument points z_i^t and $\sum_{j=1}^m w_{i,j}\theta_j^{t+1}$ are feasible for all t and $i \in [m]$. We can then apply [1, Lemma 1] and conclude

$$\varepsilon_\ell(\widehat{\boldsymbol{\theta}}, \mathbf{w}^t) \leq \frac{1}{2m}\tau_\mu 8\Psi^2(\bar{\mathcal{M}})\|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2 + \frac{\tau_\mu}{2}\nu^2.$$

Consider now $\varepsilon_u(\boldsymbol{\theta}^{t+1}, \mathbf{w}^t)$ (see (26)). Since

$$\mathbf{W}\boldsymbol{\theta}^{t+1} - \mathbf{z}^t \overset{(70)}{=} \beta\mathbf{W}\left(\mathbf{v}^{t+1} - (1-\beta)\mathbf{v}^t - \beta\mathbf{w}^t\right)$$

and $\mathcal{R}$ is a norm, it holds

$$\begin{aligned}\mathcal{R}^2\left(\sum_{j=1}^m w_{i,j}(\theta_j^{t+1} - z_i^t)\right) &= \beta^2\mathcal{R}^2\left(\sum_{j=1}^m w_{i,j}\left(\beta v_j^{t+1} - (1-\beta)v_j^t - \beta w_j^t\right)\right) \\ &\leq 3\beta^2\left(\mathcal{R}^2\left(\sum_{j=1}^m w_{i,j}v_j^{t+1} - \widehat{\theta}\right) + (1-\beta)^2\mathcal{R}^2\left(\sum_{j=1}^m w_{i,j}v_j^t - \widehat{\theta}\right) + \beta^2\mathcal{R}^2\left(z_j^t - \widehat{\theta}\right)\right).\end{aligned}$$

Again, invoking feasibility of the z, v-iterates and [1, Lemma 1] we deduce (31). □

A.7.3 Proof of Proposition 2

By definition

$$\langle \boldsymbol{\varepsilon}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle = \underbrace{\langle \mathbf{y}^t - \mathbf{W}\nabla\mathbf{F}(\mathbf{W}\mathbf{w}^t), \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle}_{\texttt{term I}} + \underbrace{\frac{\beta^2\alpha}{m}\langle(\mathbf{W}^2 - \mathbf{W})\mathbf{w}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle}_{\texttt{term II}}. \tag{72}$$

We bound the two terms separately.

For `term II`, we have

$$\begin{aligned}\texttt{term II} &= \frac{\beta^2\alpha}{m}\langle(\mathbf{W} - \mathbf{I})(\mathbf{w}^t), (\mathbf{W} - \mathbf{J})(\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1})\rangle \\ &\overset{(a)}{\leq} \frac{\beta^2\alpha\rho}{2m}\|\mathbf{w}^t\|^2_{\mathbf{I}-\mathbf{W}} + \frac{\beta^2\alpha\rho}{2m}\|\mathbf{v}^{t+1}\|^2_{\mathbf{I}-\mathbf{W}} \\ &\overset{(23a)}{\leq} \frac{\beta^2\alpha\rho}{2m(1+\beta)}\|\boldsymbol{\theta}^t\|^2_{\mathbf{I}-\mathbf{W}} + \frac{\beta^3\alpha\rho}{2m(1+\beta)}\|\mathbf{v}^t\|^2_{\mathbf{I}-\mathbf{W}} + \frac{\beta^2\alpha\rho}{2m}\|\mathbf{v}^{t+1}\|^2_{\mathbf{I}-\mathbf{W}}\end{aligned}$$

where in (a) we have used the fact that $(\mathbf{I} - \mathbf{W})^{1/2}$ and $(\mathbf{W} - \mathbf{J})^{1/2}$ commute.

We work now with `term I`. Due to the favorable behavior along the consensus direction, we bound `term I` considering separately the inner product along the consensus space and its orthogonal complement, that is,

$$\texttt{term I} = \underbrace{\langle \mathbf{y}^t - \mathbf{W}\nabla\boldsymbol{F}(\mathbf{W}\mathbf{w}^t), \mathbf{J}(\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1})\rangle}_{\texttt{term I(a)}} + \underbrace{\langle \mathbf{y}^t - \mathbf{W}\nabla\boldsymbol{F}(\mathbf{W}\mathbf{w}^t), (\mathbf{I} - \mathbf{J})(\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1})\rangle}_{\texttt{term I(b)}}.$$

We will leverage the following properties:

$$\mathbf{J}\mathbf{y}^t = \mathbf{J}\nabla \boldsymbol{f}(\mathbf{W}\mathbf{w}^t), \tag{73a}$$

$$\mathbf{J}\nabla \boldsymbol{f}(\mathbf{J}\mathbf{w}^t) = \mathbf{J}\nabla \boldsymbol{F}(\mathbf{J}\mathbf{w}^t). \tag{73b}$$

where in (73a) we explicitly used the initialization $\mathbf{y}^{-1} = \nabla \boldsymbol{f}(\mathbf{W}\mathbf{w}^{-1})$, and (73b) follows by inspection.

We proceed working with `term I(a)`, as follows:

$$\begin{aligned}
\texttt{term I(a)} &\overset{(73a)}{=} \langle \nabla \boldsymbol{f}(\mathbf{W}\mathbf{w}^t) - \nabla \boldsymbol{F}(\mathbf{W}\mathbf{w}^t), \mathbf{J}(\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1})\rangle \\
&\overset{(73b)}{=} \langle \nabla \boldsymbol{f}(\mathbf{W}\mathbf{w}^t) - \nabla \boldsymbol{f}(\mathbf{J}\mathbf{w}^t) + \nabla \boldsymbol{F}(\mathbf{J}\mathbf{w}^t) - \nabla \boldsymbol{F}(\mathbf{W}\mathbf{w}^t), \mathbf{J}(\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1})\rangle \\
&\overset{(a)}{=} \langle (\nabla^2 \boldsymbol{f})(\mathbf{W} - \mathbf{J})(\mathbf{w}^t - \widehat{\boldsymbol{\theta}}), \mathbf{J}(\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1})\rangle,
\end{aligned}$$

where in (a) we used the following two facts: the functions are quadratic, with $\nabla^2 \boldsymbol{f}$ denoting the Hessian of the quadratic function $\boldsymbol{f}$; and $(\mathbf{I} - \mathbf{J})(\nabla^2 \boldsymbol{F})\mathbf{J} = \mathbf{0}$.

We move to `term I(b)`. Using

$$\mathbf{y}^t = \mathbf{W}\nabla \boldsymbol{f}(\mathbf{W}\mathbf{w}^t) + \sum_{i=0}^{t-1} (\mathbf{W} - \mathbf{I})\mathbf{W}^{t-i}\nabla \boldsymbol{f}(\mathbf{W}\mathbf{w}^i)$$

we have:

$$\begin{aligned}
\texttt{term I(b)} &= -\langle (\mathbf{W} - \mathbf{J})^{t+1}\nabla \boldsymbol{f}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle \\
&\quad + \langle (\mathbf{W} - \mathbf{J})(\nabla \boldsymbol{f}(\mathbf{W}\mathbf{w}^t) - \nabla \boldsymbol{f}(\widehat{\boldsymbol{\theta}}) - \nabla \boldsymbol{F}(\mathbf{W}\mathbf{w}^t)), \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle \\
&\quad + \sum_{i=0}^{t-1} \langle (\mathbf{W} - \mathbf{J})^{t-i} \left(\mathbf{W}(\nabla \boldsymbol{f}(\mathbf{W}\mathbf{w}^i) - \nabla \boldsymbol{f}(\widehat{\boldsymbol{\theta}})) + \nabla \boldsymbol{f}(\widehat{\boldsymbol{\theta}}) - \nabla \boldsymbol{f}(\mathbf{W}\mathbf{w}^i)\right), \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle \\
&\overset{(a)}{=} -\langle (\mathbf{W} - \mathbf{J})^{t+1}\nabla \boldsymbol{f}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle \\
&\quad + \langle (\nabla^2 \boldsymbol{f})(\mathbf{z}^t - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J})(\widehat{\boldsymbol{\theta}} - \mathbf{v})\rangle + \langle (\nabla^2 \boldsymbol{F})(\mathbf{z}^t - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J})(\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}})\rangle \\
&\quad + \sum_{i=0}^{t-1} \langle (\mathbf{W} - \mathbf{J})^{t-i}(\mathbf{W} - \mathbf{I})(\nabla^2 \boldsymbol{f})(\mathbf{z}^i - \widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle.
\end{aligned}$$

where in (a) we used the fact that the functions are quadratic and that $\mathbf{J}\nabla \boldsymbol{F}(\widehat{\boldsymbol{\theta}}) = \nabla \boldsymbol{F}(\widehat{\boldsymbol{\theta}})$).

Using the above expressions of `term I(a)` and `term I(b)`, `term I` reads

$$\begin{aligned}
\texttt{term I} &= \langle (\nabla^2 \boldsymbol{f})(\mathbf{W} - \mathbf{J})(\mathbf{w}^t - \widehat{\boldsymbol{\theta}}), \mathbf{J}(\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1})\rangle - \langle (\mathbf{W} - \mathbf{J})^{t+1}\nabla \boldsymbol{f}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle \\
&\quad + \langle (\nabla^2 \boldsymbol{f})(\mathbf{z}^t - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J})(\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1})\rangle + \langle (\nabla^2 \boldsymbol{F})(\mathbf{z}^t - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J})(\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}})\rangle \\
&\quad + \sum_{i=0}^{t-1} \langle (\mathbf{W} - \mathbf{J})^{t-i}(\mathbf{W} - \mathbf{I})(\nabla^2 \boldsymbol{f})(\mathbf{z}^i - \widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\rangle.
\end{aligned}$$

To properly control `term I`, we need to bounds terms therein having the following structure:

$$|\langle (\nabla^2 \boldsymbol{f})(\mathbf{w} - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J})^k(\widehat{\boldsymbol{\theta}} - \mathbf{v})\rangle| \quad \text{and} \quad |\langle (\nabla^2 \boldsymbol{f})(\mathbf{J})(\mathbf{w} - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J})^k(\widehat{\boldsymbol{\theta}} - \mathbf{v})\rangle\rangle|,$$

$$|\langle (\nabla^2 \boldsymbol{F})(\mathbf{w} - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J})(\mathbf{v} - \widehat{\boldsymbol{\theta}})\rangle|,$$

$$\langle (\mathbf{W} - \mathbf{J})^k \nabla \boldsymbol{f}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}} - \mathbf{v}\rangle,$$

for $\mathbf{w}$ and $\mathbf{v}$ with feasible d-blocks, and integer k.

The following lemma provides suitable bounds for these terms.

Lemma 3. *For any* $\mathbf{w} = [w_1^\top, \dots, w_m]^\top$ *and* $\mathbf{v} = [v_1^\top, \dots, v_m]^\top$*, with* w_i, v_i*'s feasible for the ERM* (3)*, the following hold:*

$$
\begin{aligned}
&|\langle (\nabla^2 \boldsymbol{f})(\mathbf{w} - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J})^k (\widehat{\boldsymbol{\theta}} - \mathbf{v}) \rangle| \\
&\leq \frac{3m\sqrt{m}\rho^k (2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell)}{2m} \left(\|\mathbf{w} - \widehat{\boldsymbol{\theta}}\|^2 + \|\mathbf{v} - \widehat{\boldsymbol{\theta}}\|^2 \right) + \frac{m^2\sqrt{m}\rho^k \tau_\ell}{2m} 18\nu^2, \qquad (74) \\
&|\langle (\nabla^2 \boldsymbol{F})(\mathbf{w} - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J}))\widehat{\boldsymbol{\theta}} - \mathbf{v}) \rangle| \\
&\leq \frac{3m^2\sqrt{m}\rho \left(2L + 8\tau_L \Psi^2(\bar{\mathcal{M}})\right)}{2m} \left(\|\mathbf{w} - \widehat{\boldsymbol{\theta}}\|^2 + \|\mathbf{v} - \widehat{\boldsymbol{\theta}}\|^2 \right) + \frac{m^2\sqrt{m}\rho \tau_L}{2m} 18\nu^2, \qquad (75)
\end{aligned}
$$

and

$$
\begin{aligned}
&\langle (\mathbf{W} - \mathbf{J})^k \nabla \boldsymbol{f}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}} - \mathbf{v} \rangle \qquad (76) \\
&\leq 2R\sqrt{m} m \rho^k \frac{\sum_{i=1}^m \mathcal{R}^*(\nabla f_i(\theta^\star))}{m} + \frac{3m\sqrt{m}\rho^k (2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell)}{2m} \left(\|\mathbf{v} - \widehat{\boldsymbol{\theta}}\|^2 + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|^2 \right) \\
&\quad + \frac{m^2\sqrt{m}\rho^k \tau_\ell}{2m} 18\nu^2. \qquad (77)
\end{aligned}
$$

Proof. See Appendix A.7.4. □

Using (74), we have

$$
\begin{aligned}
&\langle (\nabla^2 \boldsymbol{f})(\mathbf{W} - \mathbf{J})(\mathbf{w}^t - \widehat{\boldsymbol{\theta}}), \mathbf{J}(\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}) \rangle \\
&\leq \frac{3m\sqrt{m}\rho (2\gamma_\ell + 3\Psi^2(\bar{\mathcal{M}})\tau_\ell)}{2m} \left(\|\mathbf{w}^t - \widehat{\boldsymbol{\theta}}\|^2 + \|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 \right) + \frac{m\sqrt{m}\rho \tau_\ell}{2m} 18\nu^2 \\
&\leq \frac{3m^2\sqrt{m}\rho (2\gamma_\ell + 3\Psi^2(\bar{\mathcal{M}})\tau_\ell)}{2m} \left(\frac{1}{1+\beta} \|\boldsymbol{\theta}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \frac{\beta}{1+\beta} \|\mathbf{v}^t - \widehat{\boldsymbol{\theta}}\|^2 + \|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 \right) + \frac{m^2\sqrt{m}\rho \tau_\ell}{2m} 18\nu^2,
\end{aligned}
$$

where in the last inequality we used

$$
\|\mathbf{w}^t - \widehat{\boldsymbol{\theta}}\|^2 = \|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2 + \|\mathbf{w}^t\|^2_{\mathbf{I}-\mathbf{W}^2} \leq \|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2 + \frac{1}{1+\beta} \|\boldsymbol{\theta}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \frac{\beta}{1+\beta} \|\mathbf{v}^t\|^2_{\mathbf{I}-\mathbf{W}^2}.
$$

Using similar path to bound the other terms in `term I` and combining together the final bounds of `term I` and `term II`, we obtain

$$
\begin{aligned}
&\langle \boldsymbol{\varepsilon}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1} \rangle \leq \frac{3m\sqrt{m}\rho (2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell)}{2m} \left(\frac{1}{1+\beta} \|\boldsymbol{\theta}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \frac{\beta}{1+\beta} \|\mathbf{v}^t - \widehat{\boldsymbol{\theta}}\|^2 + \|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 \right) \\
&+ \frac{m^2\sqrt{m}\rho \tau_\ell}{2m} 18\nu^2 + 2R\sqrt{m} m \rho^{t+1} \sum_{i=1}^m \frac{\mathcal{R}^*(\nabla f_i(\theta^\star))}{m} \\
&+ \frac{3m\sqrt{m}\rho^{t+1} (2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell)}{2m} \left(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|^2 + \|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 \right) + \frac{m^2\sqrt{m}\rho^{t+1} \tau_\ell}{2m} 18\nu^2 \\
&+ \frac{3m\sqrt{m}\rho (2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell)}{2m} \left(\|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2 + \|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 \right) + \frac{m^2\sqrt{m}\rho \tau_\ell}{2m} 18\nu^2 \\
&+ \frac{3m\sqrt{m}\rho (2L + 8\tau_L \Psi^2(\bar{\mathcal{M}}))}{2m} \left(\|\widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1}\|^2 + \|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2 \right) + \frac{m^2\sqrt{m}\rho \tau_L}{2m} 18\nu^2 \\
&+ \sum_{i=0}^{t-1} (1+\rho)\rho^{t-i} \left(\frac{3m\sqrt{m}(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell)}{2m} \left(\|\mathbf{z}^i - \widehat{\boldsymbol{\theta}}\|^2 + \|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2 \right) + \frac{m^2\sqrt{m}\tau_\ell}{2m} 18\nu^2 \right) \\
&+ \frac{\beta^2 \alpha\rho}{2m(1+\beta)} \|\boldsymbol{\theta}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \frac{\beta^3 \alpha\rho}{2m(1+\beta)} \|\mathbf{v}^t - \widehat{\boldsymbol{\theta}}\|^2 + \frac{\beta^2 \alpha\rho}{2m} \|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2.
\end{aligned}
$$

By properly grouping terms, we obtain the desired result

$$\langle \boldsymbol{\varepsilon}^t, \widehat{\boldsymbol{\theta}} - \mathbf{v}^{t+1} \rangle \leq 2R\sqrt{m}m\rho^{t+1} \sum_{i=1}^{m} \frac{\mathcal{R}^*(\nabla f_i(\theta^\star))}{m} + \frac{3m\sqrt{m}\rho^{t+1}(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell)}{2m} \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|^2$$
$$+ \left(\frac{\beta^2\alpha\rho}{2m} + \frac{3m\sqrt{m}\rho}{2m} \left(\left(3 + \frac{2\rho}{1-\rho} \right) (2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell) + (2L + 8\tau_L\Psi^2(\bar{\mathcal{M}})) \right) \right) \|\mathbf{v}^{t+1} - \widehat{\boldsymbol{\theta}}\|^2$$
$$+ \frac{3m\sqrt{m}\rho}{2m} \left(2(\gamma_\ell + L) + 8\Psi^2(\bar{\mathcal{M}})(\tau_\ell + \tau_L) \right) \|\mathbf{z}^t - \widehat{\boldsymbol{\theta}}\|^2$$
$$+ \frac{3m\sqrt{m}(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell)}{2m} \sum_{i=0}^{t-1} (1+\rho)\rho^{t-i} \|\mathbf{z}^i - \widehat{\boldsymbol{\theta}}\|^2$$
$$\left(\frac{\beta^2\alpha\rho}{2m} + \frac{3m\sqrt{m}\rho}{2m} \left(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell \right) \right) \left(\frac{1}{1+\beta} \|\boldsymbol{\theta}^t\|^2_{\mathbf{I}-\mathbf{W}^2} + \frac{\beta}{1+\beta} \|\mathbf{v}^t - \widehat{\boldsymbol{\theta}}\|^2 \right)$$
$$+ \frac{m^2\sqrt{m}\rho}{2m} 18\Psi^2(\bar{\mathcal{M}})\nu^2 \left(\left(3 + \frac{2\rho}{1-\rho} \right) \tau_\ell + \tau_L \right).$$

□

A.7.4 Proof of Lemma 3

For convenience, denote by $b^k_{i,j}$ the i, j element of the matrix $(W - J)^k$. Then, it holds

$$|\langle (\nabla^2 \boldsymbol{f})(\mathbf{w} - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J})^k (\widehat{\boldsymbol{\theta}} - \mathbf{v}) \rangle| = \frac{1}{m} \left| \sum_{i=1}^{m} \left\langle \left(\nabla^2 f_i \right) (w_i - \widehat{\theta}), \sum_{j=1}^{m} b^k_{i,j} (\widehat{\theta} - v_j) \right\rangle \right|$$
$$\leq \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{m} |b^k_{i,j}| \left| \left\langle \left(\nabla^2 f_i \right) (w_i - \widehat{\theta}), \widehat{\theta} - v_j \right\rangle \right|.$$

Then, under Assumption 2 it follows that

$$\left| \left\langle \left(\nabla^2 f_i \right) (w_i - \widehat{\theta}), \widehat{\theta} - v_i \right\rangle \right| \leq \frac{1}{2} \left| \langle (\nabla^2 f_i)(w_i - \widehat{\theta}), w_i - \widehat{\theta} \rangle \right| + \frac{1}{2} \left| \left\langle \left(\nabla^2 f_i \right) (v_j - \widehat{\theta}), v_k - \widehat{\theta} \right\rangle \right|$$
$$+ \frac{1}{2} \left| \left\langle \left(\nabla^2 f_i \right) (w_j - v_i), w_j - v_i \right\rangle \right|$$
$$\leq \gamma_\ell \left(\|w_i - \widehat{\theta}\|^2 + \|v_j - \widehat{\theta}\|^2 + \|w_i - v_j\|^2 \right)$$
$$+ \frac{\tau_\ell}{2} \left(\mathcal{R}^2(w_i - \widehat{\theta}) + \mathcal{R}^2(v_j - \widehat{\theta}) + \mathcal{R}^2(v_j - w_i) \right).$$

Under Assumption 5, it holds $|b_{i,j}| \leq \sqrt{m}\rho^k$ [26], yielding

$$|\langle (\nabla^2 \boldsymbol{f})(\mathbf{w} - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J})^k (\widehat{\boldsymbol{\theta}} - \mathbf{v}) \rangle|$$
$$\leq \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{\sqrt{m}\rho^k 6\gamma_\ell}{2m} \left(\|w_i - \widehat{\theta}\|^2 + \|v_j - \widehat{\theta}\|^2 \right) + \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{\sqrt{m}\rho^k \tau_\ell}{2m} \left(3\mathcal{R}^2(w_i - \widehat{\theta}) + 3\mathcal{R}^2(v_j - \widehat{\theta}) \right).$$

Because w_i and v_j are feasible for all $j, i \in [m]$, we can use in [1, Lemma 1] and obtain

$$|\langle (\nabla^2 \boldsymbol{f})(\mathbf{w} - \widehat{\boldsymbol{\theta}}), (\mathbf{W} - \mathbf{J})^k (\widehat{\boldsymbol{\theta}} - \mathbf{v}) \rangle|$$
$$\leq \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{6\sqrt{m}\rho^k \gamma_\ell}{2m} \left(\|w_i - \widehat{\theta}\|^2 + \|v_j - \widehat{\theta}\|^2 \right)$$
$$+ \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{3\sqrt{m}\rho^k \tau_\ell}{2m} \left(8\Psi^2(\bar{\mathcal{M}})(\|w_i - \widehat{\theta}\|^2 + \|v_j - \widehat{\theta}\|^2) + 6\nu^2 \right)$$
$$= \frac{3m\sqrt{m}\rho^k(2\gamma_\ell + 8\Psi^2(\bar{\mathcal{M}})\tau_\ell)}{2m} \left(\|\mathbf{w} - \widehat{\boldsymbol{\theta}}\|^2 + \|\mathbf{v} - \widehat{\boldsymbol{\theta}}\|^2 \right) + \frac{\sqrt{m}m^2\rho^k\tau_\ell}{2m} 18\nu^2,$$

where ν^2 is defined in (29).

The proof of (75) follows identical steps as above, with the only difference of using the RSM property in Assumption 1.

We prove now (77). Denote by $b_{i,j}^k$ the i, j element of the matrix $(W - J)^k$. We have

$$\begin{aligned}
&\langle (\mathbf{W} - \mathbf{J})^k \nabla \boldsymbol{f}(\widehat{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}} - \mathbf{v} \rangle = \langle (\mathbf{W} - \mathbf{J})^k \nabla \boldsymbol{f}(\boldsymbol{\theta}^\star), \widehat{\boldsymbol{\theta}} - \mathbf{v} \rangle + \langle (\mathbf{W} - \mathbf{J})^k (\nabla^2 \boldsymbol{f})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star), \widehat{\boldsymbol{\theta}} - \mathbf{v} \rangle \\
&= \frac{1}{m} \sum_{i=1}^m \sum_{j=1}^m b_{i,j}^k \langle \nabla f_j(\theta^\star), \widehat{\theta} - v_i \rangle + \langle (\mathbf{W} - \mathbf{J})^k (\nabla^2 \boldsymbol{f})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star), \widehat{\boldsymbol{\theta}} - \mathbf{v} \rangle. \qquad (78)
\end{aligned}$$

We now proceed to upper bound the terms on the RHS. Under Assumption 5 it can be shown that $|b_{i,j}^k| \leq \sqrt{m}\rho^k$ [26], therefore using Hölder's inequality along with feasibility of each v_i and $\widehat{\theta}$, we obtain

$$\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^m b_{i,j}^k \langle \nabla f_j(\theta^\star), \widehat{\theta} - v_i \rangle \leq 2R\sqrt{m} m \rho^k \frac{\sum_{i=1}^m \mathcal{R}^*(\nabla f_i(\theta^\star))}{m}.$$

For the second term in (78) we can use the upper bound in (74). □