DECENTRALIZED DYNAMIC SMOOTHING WITH PARTIAL GRADIENT TRACKING ON COMPACT SUBMANIFOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study decentralized composite optimization on compact embedded submanifolds, where each local objective contains a smooth agent-specific loss and a common convex nonsmooth regularizer. Tracking the full smoothed gradient introduces a Lipschitz factor that diverges as the smoothing parameter decreases. We propose a decentralized smoothing method with partial gradient tracking (DSPGT), which tracks only the heterogeneous smooth gradients and evaluates the common smoothing gradient locally. A Lyapunov function containing the local regularizer values controls the latter term through the Jensen decrease induced by mixing. For a uniformly sampled tail iterate, the expected smoothed stationarity residual and consensus error are $\mathcal{O}(K^{-1/3})$. For Moreau smoothing, these bounds give a generalized ε-stationary point of the original problem at every agent in $\mathcal{O}(\varepsilon^{-3})$ projection-oracle calls and communication rounds, matching the centralized order. Across the 18 reported settings on MNIST and CIFAR-10, DSPGT has the lowest median residual.

1 INTRODUCTION

In decentralized optimization, agents minimize an aggregate objective using local computation and communication with neighboring agents (Lian et al., 2017; Sun et al., 2020). We consider the case in which the decision variable also lies on a compact embedded submanifold and the objective contains a common nonsmooth regularizer:

$$\min_{x_i \in \mathcal{M},\ x_1=\cdots=x_n} \frac{1}{n}\sum_{i=1}^{n}\big(f_i(x_i)+h(x_i)\big), \tag{1.1}$$

where $\mathcal{M} \subset \mathbb{R}^{d\times r}$ is a compact C^3 embedded submanifold, f_i is a smooth local loss, and h is a common ambient-convex, possibly nonsmooth regularizer shared by all agents. This formulation covers $\ell_{2,1}$-regularized decentralized PCA on the Stiefel manifold: f_i is determined by the covariance available at node i, and $h(X) = \lambda\|X\|_{2,1}$ is invariant under a change of orthogonal basis. Related manifold constraints also arise in decentralized PCA, orthogonal embeddings, and subspace learning (Ye & Zhang, 2021; Chen et al., 2021). Thus, sites can estimate a shared feature subspace from local covariances without pooling their sample matrices.

For smooth objectives, decentralized manifold optimization has been studied through Riemannian descent, gradient tracking, augmented Lagrangian methods, and projected consensus (Chen et al., 2021; Wang & Liu, 2022; Deng & Hu, 2025a). The nonsmooth case has received less attention, and the available projection-based smoothing and subgradient analyses have $\mathcal{O}(\varepsilon^{-4})$ oracle complexity (Wang et al., 2024; Wang & Liu, 2023).[1] However, in the centralized setting, several approaches, including smoothing, augmented-Lagrangian, and ADMM-type methods, attain $\mathcal{O}(\varepsilon^{-3})$ complexity (Beck & Rosset, 2023; Peng et al., 2023; Deng et al., 2024). This leaves a gap for decentralized methods for solving (1.1): can they attain the centralized $\mathcal{O}(\varepsilon^{-3})$ order?

[1] DR-ProxGT attains an $\mathcal{O}(\varepsilon^{-2})$ outer-iteration bound, but each iteration uses a stronger composite tangent-space proximal oracle and its proximal certificate differs from the generalized stationarity considered here (Wang et al., 2025). Section 2 gives a fuller comparison.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this gap, we consider a smoothing framework. THANOS also smooths the nonsmooth regularizer and handles the Stiefel constraint through an infeasible exact-penalty construction inherited from DESTINY's approximate augmented Lagrangian (Wang & Liu, 2022; 2023). Their analysis gives $\mathcal{O}(\varepsilon^{-4})$ complexity. We instead use a projection-based framework that keeps every local iterate on the manifold. We show that directly combining smoothing with projected gradient tracking does not recover the centralized $\mathcal{O}(\varepsilon^{-3})$ rate. The difficulty arises in the network recursion. For standard smoothings, $\mathrm{Lip}(\nabla h_\mu) = \mathcal{O}(1/\mu)$. If the full smoothed gradient is tracked, this diverging factor enters the tracking and consensus errors through increments of the form

$$\mathrm{grad} h_{\mu_{k+1}}(x_{i,k+1}) - \mathrm{grad} h_{\mu_k}(x_{i,k}),$$

whose spatial variation scales as $1/\mu_k$. To recover a stationary point of the original nonsmooth problem, the smoothing parameter must satisfy $\mu_k \to 0$. The factor $1/\mu_k$ therefore grows with k, which requires smaller stepsizes in the full-tracking analysis and increases the complexity.

We therefore propose DSPGT, which separates gradient tracking for the heterogeneous local losses from local evaluation of the common smoothed regularizer. This separation removes the factor $1/\mu_k$ from the network recursions and recovers the centralized $\mathcal{O}(\varepsilon^{-3})$ order.

Our contributions are:

- We develop DSPGT, a projection-based dynamic-smoothing method that tracks only the heterogeneous smooth gradients and evaluates the common smoothing gradient locally. A Lyapunov function containing the local regularizer values, together with ambient convexity and double-stochastic mixing, keeps the network recursions independent of $1/\mu_k$. Each iteration uses one neighbor exchange and one nearest-point projection.
- We establish horizon-free step sizes giving $\mathcal{O}(K^{-1/3})$ smoothed stationarity and consensus errors. For Moreau smoothing, the network output and every local output satisfy generalized ε-stationarity of the original problem in $\mathcal{O}(\varepsilon^{-3})$ projection-oracle calls and communication rounds. This improves the existing $\mathcal{O}(\varepsilon^{-4})$ smoothing and subgradient bounds and matches the centralized order.
- We evaluate DSPGT and three baselines in 264 reported runs on MNIST and frozen CIFAR-10 ResNet-18 features over ring, grid, and expander networks with 8–32 agents. DSPGT obtains the lowest median residual in all fourteen MNIST and all four CIFAR-10 settings, and reaches the joint residual–consensus target in 38 of 42 MNIST runs and 11 of 12 CIFAR-10 runs. The topology and tracking studies connect the observed behavior with the convergence analysis.

2 Related work and oracle context

Decentralized optimization and manifold constraints. Euclidean decentralized optimization uses consensus correction and gradient tracking to remove the steady-state bias created by heterogeneous local objectives (Nedic et al., 2017; Qu & Li, 2017; Sun et al., 2020). For representation learning with orthogonality constraints, the iterates cannot simply be averaged in the ambient space. Early work studied intrinsic consensus on manifolds (Sarlette & Sepulchre, 2009; Tron et al., 2012). Optimization methods later combined multi-step consensus with Riemannian descent or gradient tracking on the Stiefel manifold (Chen et al., 2021), and projected mixing extended the analysis to general compact embedded submanifolds (Deng & Hu, 2025a). Recent stochastic variants address finite-sum sampling and recursive momentum while retaining smooth local objectives (Deng & Hu, 2025b; Zhao et al., 2026). Specialized algorithms such as DeEPCA can recover centralized-like linear convergence for the unregularized eigenproblem (Ye & Zhang, 2021), but do not cover a general nonsmooth composite regularizer.

Centralized nonsmooth manifold optimization. Three first-order approaches are most relevant. Riemannian subgradient methods apply to ambient weakly convex functions and attain $\mathcal{O}(\varepsilon^{-4})$ complexity for their stationarity measure (Li et al., 2021). Retraction-based proximal-gradient methods, represented by ManPG, exploit $f + h$ more directly by solving a composite subproblem in a tangent space (Chen et al., 2020). For stochastic online objectives, R-ProxSPB attains an $\mathcal{O}(\varepsilon^{-3})$ incremental first-order oracle complexity (Wang et al., 2022). Dynamic-smoothing methods instead follow

gradients of $f + h_{\mu_k}$ with $\mu_k \downarrow 0$, requiring only a smoothing oracle and a manifold step. Beck & Rosset (2023) obtain a $k^{-1/3}$ stationarity rate for a broad dynamic-smoothing scheme, while Peng et al. (2023) develop smoothing-gradient methods on compact embedded submanifolds. These centralized analyses give an $\mathcal{O}(\varepsilon^{-3})$ projection-oracle order, but do not include network disagreement.

Decentralized nonsmooth manifold optimization. Existing decentralized methods follow the same three approaches. DRSM applies weakly convex Riemannian subgradients on the Stiefel manifold and has complexity $\mathcal{O}(\varepsilon^{-4}\log^2(\varepsilon^{-2}))$ (Wang et al., 2024). THANOS smooths the regularizer and tracks the entire smoothed gradient; it is explicit and projection based, but its current analysis gives $\mathcal{O}(\varepsilon^{-4})$ and is specific to the Stiefel manifold (Wang & Liu, 2023). DR-ProxGT treats general compact submanifolds and improves the outer iteration order to $\mathcal{O}(\varepsilon^{-2})$ by using a composite tangent-space proximal oracle (Wang et al., 2025).

3 DSPGT: SMOOTHING–NETWORK COUPLING AND PARTIAL TRACKING

3.1 GEOMETRY, NETWORK, AND GENERALIZED STATIONARITY

We endow $\mathcal{M}$ with the Riemannian metric induced by the ambient Euclidean inner product. For $x \in \mathcal{M}$, let $T_x\mathcal{M}$ denote the tangent space and let $\mathcal{P}_{T_x\mathcal{M}}$ be the orthogonal projector onto $T_x\mathcal{M}$. If g is smooth in a neighborhood of $\mathcal{M}$, then $\mathrm{grad} g(x) = \mathcal{P}_{T_x\mathcal{M}}(\nabla g(x))$. We write $\mathcal{P}_{\mathcal{M}}(y)$ for a nearest-point projection of y onto $\mathcal{M}$. Since $\mathcal{M}$ is a compact C^3 embedded submanifold, this projection is single valued and smooth in a tubular neighborhood of $\mathcal{M}$ (Clarke et al., 1995; Absil & Malick, 2012). The conditions below ensure that every projection used by the algorithm is taken in this neighborhood; the corresponding radius and projection estimates are given in Appendix A.

Let W be a symmetric doubly stochastic mixing matrix on a connected network, with positive diagonal, $W_{ij} = 0$ for non-neighbors, and $\sigma_2 := \|W - n^{-1}\mathbf{1}\mathbf{1}^\top\|_2 < 1$. For $\mathbf{x} = \mathrm{col}(x_1, \dots, x_n)$, write $\hat{x} = n^{-1}\sum_i x_i$ for the ambient network average, $\bar{x} = \mathcal{P}_{\mathcal{M}}(\hat{x})$ for the projected network mean, and $\bar{\mathbf{x}} = \mathbf{1}_n \otimes \bar{x}$. We assume that each f_i has a Lipschitz gradient near $\mathcal{M}$, and that the smoothing family h_μ satisfies, uniformly in $\mu > 0$,

$$|h_\mu - h| \le c_h\mu, \qquad \|\nabla h_\mu\| \le G_h, \qquad \mathrm{Lip}(\nabla h_\mu) \le \kappa_h/\mu, \tag{3.1}$$

together with gradient consistency as $\mu \downarrow 0$. The full assumptions, tubular-neighborhood construction, and smoothing consistency statement are in Appendix A.

Let $\bar{f} = n^{-1}\sum_i f_i$ and $F_\mu = \bar{f} + h_\mu$. The splitting $\min_{x\in\mathcal{M}, y} \bar{f}(x) + h(y)$ subject to $x = y$ has KKT conditions $\mathcal{P}_{T_x\mathcal{M}}(\nabla\bar{f}(x) + z) = 0$, $z \in \partial h(y)$, and $x = y$. This motivates the following finite-accuracy notion, related to generalized stationarity in Peng et al. (2023).

Definition 3.1 (Distributed generalized stationarity)**.** *A random network output $\mathbf{x} \in \mathcal{M}^n$, whose mean has a unique projection $\bar{x}$, is a distributed generalized ε-stationary point if there exist auxiliary variables y, z, jointly measurable with the network output, such that $z \in \partial h(y)$ almost surely and*

$$\begin{aligned} &\mathbb{E}\|\mathcal{P}_{T_{\bar{x}}\mathcal{M}}(\nabla\bar{f}(\bar{x}) + z)\| \le \varepsilon, \\ &\|\bar{x} - y\| \le \varepsilon, \qquad \max_i \|x_i - \bar{x}\| \le \varepsilon \quad \textit{almost surely}. \end{aligned} \tag{3.2}$$

For deterministic outputs, omit the expectation. A single output $x \in \mathcal{M}$ is generalized ε-stationary if the same conditions hold with $\bar{x} = x$ and without the consensus condition.

Here y is an ambient auxiliary point. The definition controls tangent stationarity, splitting feasibility, and consensus; at zero tolerance it reduces to the KKT conditions of the original problem.

3.2 SMOOTHING–NETWORK COUPLING UNDER FULL TRACKING

Let $\mathbf{G}_{\mu_k}$ stack $\mathrm{grad} f_i(x_{i,k}) + \mathrm{grad} h_{\mu_k}(x_{i,k})$ over all agents. A direct full-gradient tracker obeys a recursion whose perturbation includes

$$\|\mathbf{G}_{\mu_{k+1}} - \mathbf{G}_{\mu_k}\| \lesssim \left(C + \frac{\kappa_h}{\mu_{k+1}}\right)\|\mathbf{x}_{k+1} - \mathbf{x}_k\| + \text{change of smoothing parameter}. \tag{3.3}$$

The same factor appears when one directly compares local smoothing gradients with $\mathrm{grad} h_{\mu_k}(\bar{x}_k)$. The method below removes the first occurrence from the tracker and postpones the second until cumulative consensus control is available.

3.3 DSPGT WITH PARTIAL GRADIENT TRACKING

DSPGT initializes the local iterates at a common point and tracks only the heterogeneous local gradients $\mathrm{grad} f_i$. Each agent evaluates the common smoothed-regularizer gradient $\mathrm{grad} h_{\mu_k}$ locally. Algorithm 1 gives the complete update and the shared tail-iterate output rule.

Algorithm 1 DSPGT: decentralized smoothing with partial gradient tracking

Require: Budget K, step sizes $\{\alpha_k\}$, smoothing parameters $\{\mu_k\}$, common $x_{i,0} = x_0 \in \mathcal{M}$, and a shared random seed

1: $s_{i,0} \leftarrow \mathrm{grad} f_i(x_{i,0})$ at every agent
2: **for** $k = 0, \dots, K-1$ **do**
3: $\quad v_{i,k} \leftarrow \mathcal{P}_{T_{x_{i,k}}\mathcal{M}}(s_{i,k}) + \mathrm{grad} h_{\mu_k}(x_{i,k})$
4: $\quad x_{i,k+1} \leftarrow \mathcal{P}_{\mathcal{M}}(\sum_j W_{ij} x_{j,k} - \alpha_k v_{i,k})$
5: $\quad s_{i,k+1} \leftarrow \sum_j W_{ij} s_{j,k} + \mathrm{grad} f_i(x_{i,k+1}) - \mathrm{grad} f_i(x_{i,k})$
6: **end for**
7: Sample shared $R_K \sim \mathrm{Unif}(\{\lceil K/2 \rceil, \dots, K-1\})$ and return x_{i,R_K}

Each iteration uses local evaluations of $\mathrm{grad} f_i$ and $\mathrm{grad} h_{\mu_k}$, one synchronized neighbor exchange of the pair $(x_{i,k}, s_{i,k})$, and one nearest-point projection. No residual has to be evaluated globally to choose the output; the shared index can be implemented by synchronized reservoir sampling.

For the analysis, let $\mathbf{s}_k = \mathrm{col}(s_{1,k}, \dots, s_{n,k})$, $\mathbf{v}_k = \mathrm{col}(v_{1,k}, \dots, v_{n,k})$, and $\widehat{\mathbf{G}}_k = \mathbf{1}_n \otimes n^{-1} \sum_i \mathrm{grad} f_i(x_{i,k})$.

The distinction from full tracking can be seen directly by decomposing its regularizer increment as

$$\begin{aligned} &\mathrm{grad} h_{\mu_{k+1}}(x_{i,k+1}) - \mathrm{grad} h_{\mu_k}(x_{i,k}) \\ &\quad = \big[\mathrm{grad} h_{\mu_{k+1}}(x_{i,k+1}) - \mathrm{grad} h_{\mu_{k+1}}(x_{i,k})\big] + \big[\mathrm{grad} h_{\mu_{k+1}}(x_{i,k}) - \mathrm{grad} h_{\mu_k}(x_{i,k})\big]. \end{aligned}$$

The first bracket carries the diverging spatial factor κ_h/μ_{k+1}, while the second is caused by changing the smoothing level. Neither appears in the partial tracker in Algorithm 1; its perturbation contains only $\mathrm{grad} f_i(x_{i,k+1}) - \mathrm{grad} f_i(x_{i,k})$. The tracker and consensus recursions below therefore have constants independent of μ_k.

3.4 LOCAL REGULARIZER POTENTIAL

The tracker conservation identity $n^{-1} \sum_i s_{i,k} = n^{-1} \sum_i \mathrm{grad} f_i(x_{i,k})$ and Lipschitz continuity of the local gradients give a recursion with constants independent of μ_k. To handle the local smoothing corrections, define

$$\Psi_\mu(\mathbf{x}) = \bar{f}(\bar{x}) + H_\mu(\mathbf{x}), \qquad H_\mu(\mathbf{x}) = \frac{1}{n} \sum_i h_\mu(x_i). \tag{3.4}$$

For $y_i = \sum_j W_{ij} x_j$, convexity and double stochasticity imply

$$H_\mu(\mathbf{x}) - \frac{1}{n} \sum_i h_\mu(y_i) \geq \frac{\mu}{2\kappa_h n} \sum_{i,j} W_{ij} \|\nabla h_\mu(x_j) - \nabla h_\mu(y_i)\|^2 \geq 0. \tag{3.5}$$

This Jensen gap bounds the smoothing-gradient difference produced by mixing. It is the step that replaces the standard mean-point comparison and prevents the network part of the descent inequality from inheriting $1/\mu_k$.

4 CONVERGENCE GUARANTEE

Let R_K be uniform on $\mathcal{I}_K = \{\lceil K/2 \rceil, \dots, K-1\}$ and independent of the algorithmic history. The appendix derives the geometric constants and the intermediate estimates used below; the following statement gives the admissible parameter range and the rate tradeoff that determines the schedule.

To state these thresholds, let $R > 0$, $r_{\mathcal{P}} \in (0, R)$, and $L_{\mathcal{P}} > 0$ be the uniform geometric constants for the normal inequality, projection expansion, and tangent-projector Lipschitz bound in Appendix A. Choose

$$0 < \delta < \min\left\{\frac{R}{4}, \frac{r_{\mathcal{P}}}{3}, \frac{R(1-\sigma_2)}{2}\right\}, \qquad \ell_\delta = \frac{R}{R-2\delta}, \qquad \rho = \ell_\delta \sigma_2 < 1. \tag{4.1}$$

Using the constants in the assumptions, define

$$\begin{aligned} L_g &= L_f + \frac{L_G}{2\delta}, & L &= \max\{L_G, L_g, L_f + L_G L_{\mathcal{P}}\}, \\ L_\mu &= L + G_h L_{\mathcal{P}} + \frac{\kappa_h}{\mu} + \frac{L_G + G_h}{2\delta}, & B_s &= \frac{3L}{1-\sigma_2} + G_h. \end{aligned} \tag{4.2}$$

Theorem 4.1 (Horizon-free polynomial schedules)**.** *Suppose the assumptions in Appendix A hold. Fix $p \in (0,1)$, $K \geq 2$, and $\mu_0 > 0$, and let the agents start from a common point $x_{1,0} = \cdots = x_{n,0} \in \mathcal{M}$. Set*

$$\mu_k = \frac{\mu_0}{(k+1)^p}, \qquad \alpha_k = \frac{\alpha_0}{(k+1)^p}.$$

There exist $c_0, c_1 > 0$, depending on the fixed problem, manifold, network, and p, but independent of μ_0, α_0, and K, such that the conclusions below hold whenever

$$0 < \alpha_0 < \min\left\{\frac{c_0}{1+L_{\mu_0}}, \frac{\delta}{\sqrt{n}B_s}, \frac{(1-\rho)\delta}{\ell_\delta \sqrt{n} B_s}, c_1\right\}. \tag{4.3}$$

Then

$$\begin{aligned} &\mathbb{E}\|\mathrm{grad} F_{\mu_{R_K}}(\bar{x}_{R_K})\|^2 = \mathcal{O}((K+1)^{p-1}), && \mu_{R_K} = \mathcal{O}((K+1)^{-p}), \\ &\mathbb{E}\left[\frac{1}{n}\|\mathbf{x}_{R_K} - \bar{\mathbf{x}}_{R_K}\|^2\right] = \mathcal{O}((K+1)^{-2p}). \end{aligned}$$

The same stationarity bound holds at some deterministic $k_K \in \mathcal{I}_K$. Moreover, with $m_K = \lceil K/2 \rceil$, Lemma E.5 gives the pointwise tail bound

$$\max_i \|x_{i,k} - \bar{x}_k\| \leq d_K := \sqrt{n D_{c,p}} B_s \alpha_{m_K - 1}, \quad k \in \mathcal{I}_K,$$

where $D_{c,p} = \ell_\delta^2 (\sum_{j \geq 0} \rho^j (j+1)^p)^2$.

Balancing the stationarity and consensus rates gives $p = 1/3$. Hence

$$\mathbb{E}\left[\|\mathrm{grad} F_{\mu_{R_K}}(\bar{x}_{R_K})\|^2 + \frac{1}{n}\|\mathbf{x}_{R_K} - \bar{\mathbf{x}}_{R_K}\|^2\right] = \mathcal{O}((K+1)^{-2/3}), \qquad \mu_{R_K} = \mathcal{O}((K+1)^{-1/3}). \tag{4.4}$$

The next result converts these estimates into generalized stationarity of the original problem and its actual local outputs.

Proposition 4.2 (Generalized stationarity of the original problem)**.** *Under Theorem 4.1, suppose h is convex and globally Lipschitz and h_μ is its Moreau envelope. Let S_K be the squared smoothed-gradient upper bound in that theorem and set*

$$\begin{aligned} \eta_K &= \max\{\sqrt{S_K}, G_h \mu_{m_K}, d_K\}, \\ C_\star &= \max\{1 + C_{\mathrm{loc}}, 2\}, \qquad C_{\mathrm{loc}} = L_f + L_{\mathcal{P}}(L_G + G_h). \end{aligned}$$

The network output $\mathbf{x}_{R_K}$ is distributed generalized η_K-stationary, and every local output x_{i,R_K} is generalized $C_\star \eta_K$-stationary. Both reach tolerance ε in

$$K = \mathcal{O}\Big(\max\{\varepsilon^{-2/(1-p)}, \varepsilon^{-1/p}\}\Big).$$

In particular, $p = 1/3$ gives $\mathcal{O}(\varepsilon^{-3})$ projection-oracle calls and communication rounds per agent, with constants independent of K and ε.

Proof. Take $y_K = \mathrm{prox}_{\mu_{R_K} h}(\bar{x}_{R_K})$ and $z_K = \nabla h_{\mu_{R_K}}(\bar{x}_{R_K})$. Then $z_K \in \partial h(y_K)$ and $\|\bar{x}_{R_K} - y_K\| \leq G_h \mu_{m_K}$. Jensen's inequality and the pointwise consensus bound give Equation (3.2). The same y_K, z_K certify each local output: tangent-projector and gradient Lipschitz continuity bound its tangent residual by the mean-point residual plus $C_{\mathrm{loc}} d_K$, while its splitting error is at most $d_K + G_h \mu_{m_K}$. The full conversion is proved in Proposition D.4. □

Proof outline. The complete proof is in the appendix; we summarize the four dependencies that determine the rate. First, smooth-gradient tracking and projected consensus give stable cumulative recursions with constants independent of μ_k:

$$\frac{1}{n}\sum_{k=1}^{K}\frac{\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^2}{\alpha_{k-1}}\le\frac{C}{n}\sum_{k<K}\alpha_k\|\mathbf{v}_k\|^2,\qquad\frac{1}{n}\sum_{k<K}\|\mathbf{s}_k-\widehat{\mathbf{G}}_k\|^2\le C+\frac{C\alpha_0}{n}\sum_{k<K}\alpha_k\|\mathbf{v}_k\|^2. \tag{4.5}$$

Second, the Jensen gap in Equation (3.5), the smooth-loss descent, and the projection expansion yield, after absorbing the network terms,

$$\sum_{k<K}\alpha_k\|\widetilde{q}_k\|^2+\frac{1}{n}\sum_{k<K}\alpha_k\|\mathbf{v}_k\|^2\le C, \tag{4.6}$$

where $\widetilde{q}_k$ is the average local smoothed tangent gradient. Third, only at the final conversion do we compare $\widetilde{q}_k$ with the mean-point residual. The resulting $1/\mu_k$ factor multiplies a weighted consensus sum, and the coupling $\alpha_k/\mu_k=\alpha_0/\mu_0$ allows this term to be absorbed. Finally, uniform averaging over the last half gives $\mathcal{O}(K^{p-1})$ mean-point stationarity and pointwise tail consensus $\mathcal{O}(K^{-2p})$, which proves Theorem 4.1.

Role of partial tracking in the complexity bound. Partial tracking confines the intrinsic $\mathrm{Lip}(\nabla h_{\mu_k})=\mathcal{O}(1/\mu_k)$ factor to the final conversion from the average local direction to the mean-point residual. It is therefore absent from the recursively amplified tracker and consensus estimates in Equation (4.5). At the final conversion, the factor multiplies the controlled weighted consensus error and is absorbed because $\alpha_k/\mu_k=\alpha_0/\mu_0$. The common regularizer and the convex Jensen gap are therefore required for this argument.

For Moreau smoothing, Proposition 4.2 counts smoothing-gradient evaluations, nearest-point projections, and communication rounds for generalized stationarity of the original problem. Partial tracking is the mechanism that improves this theoretical upper bound: it keeps the factor $\mathcal{O}(1/\mu_k)$ out of the recursively amplified network errors and thereby permits the $\mathcal{O}(\varepsilon^{-3})$ projection-oracle rate.

5 Experiments

We test decentralized $\ell_{2,1}$-regularized subspace learning on MNIST (LeCun et al., 1998) and on frozen ImageNet ResNet-18 features from CIFAR-10 (Krizhevsky, 2009; He et al., 2016). Each network solves

$$\min_{X_i\in\mathrm{St}(d,10),\ X_1=\cdots=X_n}\frac{1}{n}\sum_{i=1}^{n}\left(-\frac{1}{2m_i}\|A_iX_i\|_F^2+\lambda\|X_i\|_{2,1}\right),$$

where $A_i\in\mathbb{R}^{m_i\times d}$ contains the centered data at agent i and $C_i=A_i^\top A_i/m_i$. We use $d=784$ for MNIST and $d=512$ for CIFAR-10. For MNIST, the two regularization levels use $\lambda=0.003$ and $\lambda=0.03$, respectively. CIFAR uses $\lambda=0.3$ in both split conditions. The topology study uses 8-agent ring/expander, 16-agent ring/grid/expander, and 32-agent ring/expander networks under IID/medium and label-Dirichlet-0.1/strong conditions. Every method receives the same partition, initialization, graph, $K=2500$ rounds, and one synchronized neighbor exchange per round.

We compare DSPGT with THANOS (Wang & Liu, 2023), DRSM (Wang et al., 2024), and DR-ProxGT (Wang et al., 2025). DSPGT uses the same fixed schedule in every reported setting:

$$\mu_k=0.1(k+1)^{-1/3},\qquad\alpha_k=\begin{cases}0.6/\widehat{L}(k+1)^{-1/3}, & 0\le k<1250,\\ 12.5/\widehat{L}(k+1)^{-1/3}, & 1250\le k<2500.\end{cases}$$

where $\widehat{L}=\max_i\|C_i\|_2$. The reported experiments comprise 264 runs: 168 MNIST topology, 48 partial/full-tracking, and 48 CIFAR runs. We report medians over completed runs, completion counts, and target-attainment counts. DR-ProxGT uses a tangent-proximal KKT tolerance of 10^{-6}. All scheduled runs remain in the target-attainment denominators.

Table 1: Aggregate results over the MNIST topology study and the CIFAR-10 feature study. Wins count settings with the smallest median over completed runs; Target counts all scheduled runs that satisfy the fixed residual–consensus criterion. Completed reports terminal-metric records over scheduled runs.

Data	Method	Completed	Res. wins	Cons. wins	Target
MNIST	**DSPGT**	42/42	14	12	38/42
	THANOS	39/42	0	1	4/42
	DRSM	42/42	0	1	0/42
	DR-ProxGT	42/42	0	0	18/42
CIFAR-10	**DSPGT**	12/12	4	2	11/12
	THANOS	11/12	0	0	0/12
	DRSM	12/12	0	1	0/12
	DR-ProxGT	6/12	0	1	3/12

At the projected network mean $\bar{X} = \mathcal{P}_{\mathrm{St}}(n^{-1}\sum_i X_i)$, every method is evaluated with the same nonsmooth objective, RMS consensus error, orthogonality error, and smoothed stationarity residual

$$\mathcal{R}_{\mathrm{eval}}(\bar{X}) = \left\| \mathcal{P}_{T_{\bar{X}}\mathcal{M}} \left(-\frac{1}{n}\sum_i C_i \bar{X} + \nabla h_{10^{-3}}(\bar{X}) \right) \right\|_F .$$

The joint target requires a residual no larger than one quarter of its initial value and consensus error at most 10^{-4}. Runs that do not attain both thresholds within $K = 2500$ are counted as misses. Since every method uses one neighbor exchange per iteration, we compare progress in communication rounds. CIFAR features are extracted once in CPU `eval` mode, and the resulting subspaces are evaluated with a ridge probe whose coefficient is 10^{-4}.

MNIST: scale and topology. DSPGT obtains the smallest setting-level median residual over completed runs in all fourteen topology–condition settings and the smallest median consensus error in twelve. It reaches the joint target in $38/42$ runs, compared with $18/42$ for DR-ProxGT, $4/42$ for THANOS, and none for DRSM. Its aggregate median residual is 0.0108, versus 0.101 for DR-ProxGT, 0.501 for THANOS, and 0.805 for DRSM. The aggregate trajectories in Figure 1 show that these terminal differences arise within the common communication budget rather than from different stopping rules.

The effect of connectivity is most visible at 32 agents. Under strong heterogeneity, the expander gives a median residual of 0.0121 and consensus error 5.57×10^{-9}; the ring gives 0.0388 and 4.47×10^{-5}. Across 8, 16, and 32-agent expanders in the same condition, the median residual stays between 0.0119 and 0.0122. Across the tested range, the expander maintains this residual range as the network grows from 8 to 32 agents.

CIFAR-10 frozen features. DSPGT has the smallest setting-level median residual in all four 32-agent settings and reaches the joint target in $11/12$ runs, compared with $3/12$ for DR-ProxGT and none for THANOS or DRSM. Its aggregate median residual is 0.0877, versus 0.441, 0.710, and 4.70, respectively. Probe accuracy ranges from 0.629 to 0.631 for DSPGT and is comparable to THANOS and the completed DR-ProxGT runs. DSPGT completes all twelve runs; baseline completion counts are reported in Tables 1 and 2.

Tracking checks. We compare partial and full tracking on 32-agent ring and expander networks with label-Dirichlet concentrations 0.1 and 0.03 and $\lambda = 0.03$. Table 3 reports the terminal residual, RMS consensus error, and tracking error separately for each version in four groups of six matched instances, together with an overall summary. Each group combines the three seeds under both fixed polynomial schedules used in the stress test. The two versions have nearly identical terminal metrics within each group. Both attain smaller consensus and tracking errors on the expanders than on the rings. Partial tracking preserves the observed finite-budget endpoints across these networks and heterogeneity levels. Its principal role is theoretical: it keeps the smoothing-dependent Lipschitz factor out of the recursive tracker and consensus bounds, which improves the resulting complexity upper bound.

Table 2: CIFAR-10 frozen-feature medians over completed runs. EV is test explained variance, probe accuracy is reported as a fraction, and Completed reports terminal-metric records over the three scheduled runs.

Graph	Split	Method	Completed	Residual	Consensus	EV	Probe acc.
expander	non-IID	DSPGT	3/3	0.0828	2.87e-08	0.374	0.629
expander	non-IID	THANOS	3/3	0.302	0.00286	0.375	0.63
expander	non-IID	DRSM	3/3	9.66	2.67e-06	0.037	0.502
expander	non-IID	DR-ProxGT	0/3	–	–	–	–
expander	IID	DSPGT	3/3	0.0599	4.93e-09	0.375	0.631
expander	IID	THANOS	2/3	0.585	0.417	0.374	0.631
expander	IID	DRSM	3/3	0.289	0.0243	0.375	0.63
expander	IID	DR-ProxGT	3/3	0.725	0.229	0.375	0.631
ring	non-IID	DSPGT	3/3	0.285	5.69e-05	0.374	0.63
ring	non-IID	THANOS	3/3	0.441	0.0344	0.375	0.631
ring	non-IID	DRSM	3/3	9.71	3.49e-05	0.0373	0.503
ring	non-IID	DR-ProxGT	0/3	–	–	–	–
ring	IID	DSPGT	3/3	0.0902	2.16e-05	0.374	0.631
ring	IID	THANOS	3/3	0.66	0.0134	0.375	0.631
ring	IID	DRSM	3/3	0.279	0.0856	0.375	0.63
ring	IID	DR-ProxGT	3/3	0.697	1.79e-05	0.375	0.631

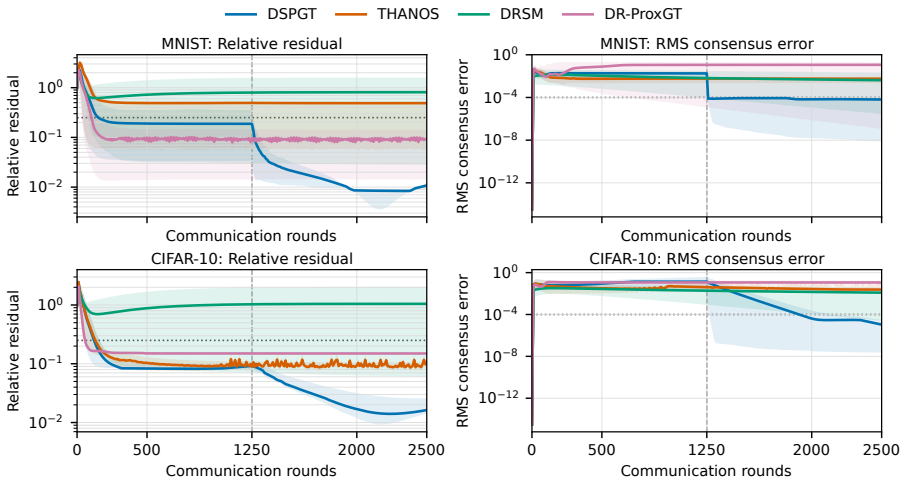

Figure 1: Convergence across the topology–condition settings. At each checkpoint, we first take the median over seeds within a setting and then aggregate the setting medians with equal weight; shading gives the interquartile range across settings. Residuals are normalized by their initial values. Horizontal dotted lines mark the fixed target thresholds, and the vertical dashed line marks DSPGT's stepsize switch at iteration 1250. Curves aggregate completed trajectories; Table 1 reports completion counts.

Table 3: Partial/full tracking comparison on 32-agent MNIST networks with $\lambda = 0.03$ and $K = 2500$. Values are terminal metrics summarized as median [minimum, maximum]. Each group pools seeds 110–112 under the two fixed polynomial schedules ($p = 1/3$ and $p = 1/2$).

Network	Dirichlet α	Tracking	Pairs	Residual	RMS consensus	Tracking error
Ring	0.1	Partial	6	2.863 $[2.463, 3.291]$	4.36×10^{-4} $[7.98 \times 10^{-5}, 1.21 \times 10^{-3}]$	0.0358 $[0.0172, 0.0639]$
Ring	0.1	Full	6	2.863 $[2.463, 3.291]$	4.37×10^{-4} $[7.98 \times 10^{-5}, 1.21 \times 10^{-3}]$	0.0358 $[0.0172, 0.0639]$
Ring	0.03	Partial	6	2.862 $[2.461, 3.305]$	3.33×10^{-4} $[6.70 \times 10^{-5}, 8.12 \times 10^{-4}]$	0.0278 $[0.0153, 0.0427]$
Ring	0.03	Full	6	2.862 $[2.461, 3.305]$	3.33×10^{-4} $[6.70 \times 10^{-5}, 8.13 \times 10^{-4}]$	0.0278 $[0.0153, 0.0427]$
Expander	0.1	Partial	6	2.866 $[2.470, 3.278]$	1.72×10^{-6} $[3.57 \times 10^{-7}, 4.09 \times 10^{-6}]$	0.00310 $[0.00177, 0.00468]$
Expander	0.1	Full	6	2.866 $[2.470, 3.278]$	1.72×10^{-6} $[3.57 \times 10^{-7}, 4.09 \times 10^{-6}]$	0.00310 $[0.00177, 0.00468]$
Expander	0.03	Partial	6	2.863 $[2.463, 3.297]$	1.63×10^{-6} $[3.23 \times 10^{-7}, 3.28 \times 10^{-6}]$	0.00295 $[0.00171, 0.00400]$
Expander	0.03	Full	6	2.863 $[2.463, 3.297]$	1.63×10^{-6} $[3.23 \times 10^{-7}, 3.28 \times 10^{-6}]$	0.00295 $[0.00171, 0.00400]$
Overall	–	Partial	24	2.866 $[2.461, 3.305]$	3.56×10^{-5} $[3.23 \times 10^{-7}, 1.21 \times 10^{-3}]$	0.0100 $[0.00171, 0.0639]$
Overall	–	Full	24	2.866 $[2.461, 3.305]$	3.56×10^{-5} $[3.23 \times 10^{-7}, 1.21 \times 10^{-3}]$	0.0100 $[0.00171, 0.0639]$

Full seed-level endpoints, work counts, and solver-status records are given in the appendix and supplement.

6 Conclusion and scope

We introduced DSPGT for decentralized composite optimization on compact embedded submanifolds. Partial tracking removes the smoothing-dependent Lipschitz factor from the recursive network bounds, while the local-regularizer Lyapunov function gives an $\mathcal{O}(\varepsilon^{-3})$ rate. It is the best-known decentralized dynamic-smoothing projection-oracle order and matches the centralized order. In the tested MNIST and CIFAR-10 settings, DSPGT has the lowest median residual and the most joint residual–consensus target hits. The analysis covers a common ambient-convex regularizer, a static connected network, consensual initialization in a local tubular neighborhood, sufficient spectral connectivity for one-step mixing, and an exact nearest-point projection oracle. For Moreau smoothing, it establishes finite-accuracy generalized stationarity of the original problem. The experiments study $\ell_{2,1}$-regularized subspace learning on two image datasets, and every reported result uses seeds 110–112.

AI use statement

We used OpenAI Codex to assist with manuscript drafting and language editing, literature searches, proof checking. The authors take full responsibility for the final manuscript, including the correctness of the theoretical claims and proofs, the validity of the experimental methodology and reported results, the accuracy of citations, and the accompanying code and supplementary materials.

References

Pierre-Antoine Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012. doi: {10.1137/100802529}.

Amir Beck and Israel Rosset. A dynamic smoothing technique for a class of nonsmooth optimization problems on manifolds. *SIAM Journal on Optimization*, 33(3):1473–1493, 2023. doi: 10.1137/22M1489447.

Shixiang Chen, Shiqian Ma, Anthony Man-Cho So, and Tong Zhang. Proximal gradient method for nonsmooth optimization over the Stiefel manifold. *SIAM Journal on Optimization*, 30(1): 210–239, 2020.

Shixiang Chen, Alfredo Garcia, Mingyi Hong, and Shahin Shahrampour. Decentralized Riemannian gradient descent on the Stiefel manifold. In *International Conference on Machine Learning*, pp. 1594–1605. PMLR, 2021.

Francis H Clarke, Ronald J Stern, and Peter R Wolenski. Proximal smoothness and the lower-C2 property. *Journal of Convex Analysis*, 2(1-2):117–144, 1995.

Kangkang Deng and Jiang Hu. Decentralized projected Riemannian gradient method for smooth optimization on compact submanifolds embedded in the Euclidean space. *Numerische Mathematik*, 157(6):2173–2210, 2025a. doi: {10.1007/s00211-025-01497-0}.

Kangkang Deng and Jiang Hu. Decentralized Projected Riemannian Stochastic Recursive Momentum Method for Nonconvex Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11203–11211, 2025b. doi: {10.1609/aaai.v39i11.33218}.

Kangkang Deng, Jiang Hu, Jiayuan Wu, and Zaiwen Wen. Oracle Complexities of Augmented Lagrangian Methods for Nonsmooth Manifold Optimization. *arXiv preprint arXiv:2404.05121*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The MNIST database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`, 1998. Accessed 2026-08-12.

Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man-Cho So. Weakly convex optimization over Stiefel manifold using Riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021.

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*, pp. 5336–5346, 2017.

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

Zheng Peng, Weihe Wu, Jiang Hu, and Kangkang Deng. Riemannian smoothing gradient type algorithms for nonsmooth optimization problem on compact Riemannian submanifold embedded in Euclidean space. *Applied Mathematics & Optimization*, 88(3):85, 2023.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

Alain Sarlette and Rodolphe Sepulchre. Consensus optimization on manifolds. *SIAM Journal on Control and Optimization*, 48(1):56–76, 2009.

Haoran Sun, Songtao Lu, and Mingyi Hong. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *International Conference on Machine Learning*, pp. 9217–9228. PMLR, 2020.

Roberto Tron, Bijan Afsari, and René Vidal. Riemannian consensus for manifolds with bounded curvature. *IEEE Transactions on Automatic Control*, 58(4):921–934, 2012.

Bokun Wang, Shiqian Ma, and Lingzhou Xue. Riemannian stochastic proximal gradient methods for nonsmooth optimization over the Stiefel manifold. *Journal of Machine Learning Research*, 23(106):1–33, 2022. URL `https://www.jmlr.org/papers/v23/21-0314.html`.

Jinxin Wang, Jiang Hu, Shixiang Chen, Zengde Deng, and Anthony Man-Cho So. Decentralized non-smooth optimization over the Stiefel manifold. In *2024 IEEE 13th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 1–5. IEEE, 2024. doi: 10.1109/SAM60225.2024.10636572.

Lei Wang and Xin Liu. Decentralized optimization over the Stiefel manifold by an approximate augmented Lagrangian function. *IEEE Transactions on Signal Processing*, 70:3029–3041, 2022.

Lei Wang and Xin Liu. Smoothing Gradient Tracking for Decentralized Optimization over the Stiefel Manifold with Non-Smooth Regularizers. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 126–132. IEEE, 2023.

Lei Wang, Le Bao, and Xin Liu. A Decentralized Proximal Gradient Tracking Algorithm for Composite Optimization on Riemannian Manifolds. *Journal of Machine Learning Research*, 26(106):1–37, 2025. URL https://www.jmlr.org/papers/volume26/24-1989/24-1989.html.

Haishan Ye and Tong Zhang. DeEPCA: Decentralized exact PCA with linear convergence rate. *The Journal of Machine Learning Research*, 22(1):10777–10803, 2021.

Jishu Zhao, Xi Wang, and Jinlong Lei. Distributed Riemannian Stochastic Gradient Tracking Algorithm on the Stiefel Manifold. *Automatica*, 184:112726, 2026. doi: {10.1016/j.automatica.2025.112726}.

A Full technical details and proofs

A.1 Notation

Let $\mathbf{1}_n \in \mathbb{R}^n$ denote the all-ones vector and set $J := n^{-1}\mathbf{1}_n\mathbf{1}_n^\top$. We write $\mathbf{x} := \mathrm{col}(x_1, \ldots, x_n)$ for the collection of local variables and define $f(\mathbf{x}) := n^{-1}\sum_{i=1}^n f_i(x_i)$. For $\mu > 0$, define

$$\mathcal{F}_\mu(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n (f_i(x_i) + h_\mu(x_i)), \qquad F_\mu(z) := \frac{1}{n}\sum_{i=1}^n f_i(z) + h_\mu(z).$$

Let $\mathbf{W} := W \otimes I_d \in \mathbb{R}^{nd\times nd}$ and $\mathbf{J} := J \otimes I_d$, where $\otimes$ denotes the Kronecker product. The algorithm always uses the original neighbor-sparse matrix W, never a power W^t. Its primal and tracker averages can be computed in one synchronized communication round by sending the pair $(x_{i,k}, s_{i,k})$. The manifold $\mathcal{M}$ carries the metric induced by the ambient Euclidean inner product $\langle\cdot,\cdot\rangle$, and $\|\cdot\|$ denotes the Euclidean norm. We use $\mathcal{M}^n$ for the n-fold Cartesian product and $[n] = \{1, \ldots, n\}$. For $x \in \mathcal{M}$, the tangent and normal spaces are $T_x\mathcal{M}$ and $N_x\mathcal{M}$. We write ∇f and $\mathrm{grad} f$ for the Euclidean and Riemannian gradients of a smooth function.

B Preliminaries

For $y \in \mathbb{R}^{d\times r}$, let

$$\mathrm{dist}(y, \mathcal{M}) := \inf_{x\in\mathcal{M}} \|y - x\|, \qquad \mathcal{P}_\mathcal{M}(y) := \operatorname*{argmin}_{x\in\mathcal{M}} \|y - x\|,$$

and set $U_\mathcal{M}(R) := \{y : \mathrm{dist}(y, \mathcal{M}) < R\}$. A closed set is R-proximally smooth if $\mathcal{P}_\mathcal{M}$ is single valued on $U_\mathcal{M}(R)$. Every compact C^3 embedded submanifold has positive reach and hence is proximally smooth (Clarke et al., 1995). Let $R > 0$ be a proximal smoothness radius of $\mathcal{M}$. Compactness and C^3 regularity also give a radius $r_\mathcal{P} \in (0, R)$ and $Q > 0$ such that the uniform projection expansion in Lemma D.6 holds whenever $\|u\| \le r_\mathcal{P}$. We choose $\delta > 0$ below so that $2\delta < R$. Then

$$\|\mathcal{P}_\mathcal{M}(y) - \mathcal{P}_\mathcal{M}(z)\| \le \frac{R}{R-2\delta}\|y - z\|, \qquad y, z \in \bar{U}_\mathcal{M}(2\delta), \tag{B.1}$$

$$|\langle v, y - x\rangle| \le \frac{\|v\|}{4\delta}\|y - x\|^2, \qquad x, y \in \mathcal{M},\ v \in N_x\mathcal{M}. \tag{B.2}$$

The absolute value in (B.2) follows by applying the proximal normal inequality to both v and $-v$ and using $2\delta < R$.

For $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{M}^n$, define the Euclidean average $\hat{x} := n^{-1} \sum_i x_i$. Whenever $\hat{x} \in U_{\mathcal{M}}(R)$, write $\bar{x} := \mathcal{P}_{\mathcal{M}}(\hat{x})$ and $\bar{\mathbf{x}} := \mathbf{1}_n \otimes \bar{x}$. Let

$$F(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + h(x), \qquad F_\mu(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + h_\mu(x).$$

Definition 3.1 describes generalized stationarity of the original problem. For the general smoothing analysis, we also use the following certificate with a squared tolerance δ_{sm}, distinct from the unsquared tolerance ε.

Definition B.1. *For $\delta_{\mathrm{sm}}, \mu > 0$, $\{x_i\}_{i=1}^n \subset \mathcal{M}$ is a $(\delta_{\mathrm{sm}}, \mu)$-smoothed stationary certificate if*

$$\frac{1}{n} \sum_{i=1}^{n} \|x_i - \bar{x}\|^2 \le \delta_{\mathrm{sm}}, \qquad \|\mathrm{grad} F_\mu(\bar{x})\|^2 \le \delta_{\mathrm{sm}}.$$

A randomized collection is an expected $(\delta_{\mathrm{sm}}, \bar{\mu})$-smoothed stationary certificate if the expectations of the two left-hand sides are at most δ_{sm} and its smoothing parameter is at most $\bar{\mu}$ almost surely. A point $x \in \mathcal{M}$ is stationary for the original problem if

$$0 \in \mathcal{P}_{T_x \mathcal{M}} \left(\frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x) + \partial h(x) \right),$$

where ∂h is the ambient convex subdifferential.

C Auxiliary single-round projected-consensus estimates

This section adapts the single-step projection argument of (Deng & Hu, 2025b, Theorem 3.6 and Section 4.1) to the bounded directions used later. Unlike the earlier multi-step module, every update below uses W itself rather than a power W^t.

Let $G = (\mathcal{V}, \mathcal{E})$ be a connected undirected network with $n \ge 2$, and let W be its mixing matrix.

Assumption C.1. *$W = W^\top \ge 0$, $W\mathbf{1}_n = \mathbf{1}_n$, and $0 < W_{ii} < 1$. Moreover,*

$$\sigma_2 := \|W - J\|_2 \in [0, 1), \qquad J := \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top.$$

Choose the local radius so that

$$0 < \delta < \min \left\{ \frac{R}{4}, \frac{r_{\mathcal{P}}}{3}, \frac{R(1-\sigma_2)}{2} \right\}, \qquad \ell_\delta := \frac{R}{R - 2\delta}, \qquad \rho := \ell_\delta \sigma_2 < 1. \tag{C.1}$$

Define

$$\mathcal{N}_\delta := \left\{ \mathbf{x} \in \mathcal{M}^n : \begin{array}{c} \hat{x} \in U_{\mathcal{M}}(R), \ \bar{x} = \mathcal{P}_{\mathcal{M}}(\hat{x}), \\ \|\mathbf{x} - \bar{\mathbf{x}}\| \le \delta \end{array} \right\},$$

$$\phi(\mathbf{x}) := \frac{1}{4} \sum_{i,j=1}^{n} W_{ij} \|x_i - x_j\|^2, \qquad \nabla \phi_i(\mathbf{x}) = x_i - \sum_j W_{ij} x_j.$$

The following estimate isolates both norms of the perturbation that are needed in the single-step proof.

Lemma C.2 (Invariant neighborhood)**.** *Let $B_\infty, B_2 > 0$, $\mathbf{x}_k \in \mathcal{N}_\delta$, $\max_i \|u_{i,k}\| \le B_\infty$, and $\|\mathbf{u}_k\| \le \sqrt{n} B_2$. If*

$$0 \le \alpha_k < \min \left\{ \frac{\delta}{B_\infty}, \frac{(1-\rho)\delta}{\ell_\delta \sqrt{n} B_2} \right\}, \tag{C.2}$$

then

$$x_{i,k+1} = \mathcal{P}_{\mathcal{M}} \left(\sum_j W_{ij} x_{j,k} - \alpha_k u_{i,k} \right)$$

is well defined, $\mathbf{x}_{k+1} \in \mathcal{N}_\delta$, every projection argument belongs to $U_{\mathcal{M}}(2\delta)$, and

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \le \rho \|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \ell_\delta \alpha_k \|\mathbf{u}_k\|. \tag{C.3}$$

Proof. Set $z_{i,k} := \sum_j W_{ij} x_{j,k} - \alpha_k u_{i,k}$. Since $\max_j \|x_{j,k} - \bar{x}_k\| \le \|\mathbf{x}_k - \bar{\mathbf{x}}_k\| \le \delta$,

$$\|z_{i,k} - \bar{x}_k\| \le \sum_j W_{ij} \|x_{j,k} - \bar{x}_k\| + \alpha_k B_\infty < 2\delta.$$

Thus $z_{i,k} \in U_{\mathcal{M}}(2\delta)$ and its projection is unique. Moreover, $\hat{x}_k = n^{-1} \sum_i x_{i,k}$ lies in $\bar{U}_{\mathcal{M}}(\delta)$ and $\mathcal{P}_{\mathcal{M}}(\hat{x}_k) = \bar{x}_k$. Applying (B.1) componentwise gives

$$\begin{aligned} \|\mathbf{x}_{k+1} - \mathbf{1}_n \otimes \bar{x}_k\| &\le \ell_\delta \|\mathbf{W}\mathbf{x}_k - \alpha_k \mathbf{u}_k - \mathbf{1}_n \otimes \hat{x}_k\| \\ &\le \ell_\delta \sigma_2 \|\mathbf{x}_k - \mathbf{1}_n \otimes \hat{x}_k\| + \ell_\delta \alpha_k \|\mathbf{u}_k\| \\ &\le \rho \|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \ell_\delta \alpha_k \|\mathbf{u}_k\|. \end{aligned}$$

Here the last inequality uses the fact that the Euclidean average is the best common Euclidean point, so $\|\mathbf{x}_k - \mathbf{1}_n \otimes \hat{x}_k\| \le \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|$. Condition (C.2) makes the last display strictly smaller than δ. Jensen's inequality therefore gives $\operatorname{dist}(\hat{x}_{k+1}, \mathcal{M}) \le n^{-1/2} \|\mathbf{x}_{k+1} - \mathbf{1}_n \otimes \bar{x}_k\| < \delta < R$, so $\bar{x}_{k+1} = \mathcal{P}_{\mathcal{M}}(\hat{x}_{k+1})$ is well defined. Since this projected mean minimizes the sum of squared distances from the new local copies over all common points of $\mathcal{M}$,

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \le \|\mathbf{x}_{k+1} - \mathbf{1}_n \otimes \bar{x}_k\| < \delta.$$

This proves both (C.3) and neighborhood invariance. No division by σ_2 is used, so the exact-averaging case is included. □

For later reference, the unperturbed update satisfies

$$x_{i,k+1} = \mathcal{P}_{\mathcal{M}} \left(\sum_j W_{ij} x_{j,k} \right). \tag{C.4}$$

Theorem C.3 (Single-communication projected-consensus contraction)**.** *If $\mathbf{x}_0 \in \mathcal{N}_\delta$, then the iterates of* (C.4) *remain in $\mathcal{N}_\delta$ and*

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \le \rho \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|, \qquad \rho = \frac{R\sigma_2}{R - 2\delta} < 1. \tag{C.5}$$

Proof. Apply Lemma C.2 with $\alpha_k = 0$. □

D Auxiliary geometric and smoothing estimates

Assumption D.1. *For constants $L_f, L_G > 0$, each f_i has an L_f-Lipschitz Euclidean gradient on $\operatorname{conv}(\mathcal{M})$, and $\max_{x \in \mathcal{M}} \|\nabla f_i(x)\| \le L_G$.*

Assumption D.2. *For each $\mu > 0$, h_μ is convex and continuously differentiable on the ambient space. There are constants $\kappa_h, c_h, G_h > 0$, independent of μ, such that*

$$|h_\mu(u) - h(u)| \le c_h \mu, \quad |h_\mu(u) - h_\nu(u)| \le c_h |\mu - \nu|,$$

$$\|\nabla h_\mu(u) - \nabla h_\mu(v)\| \le \frac{\kappa_h}{\mu} \|u - v\|, \qquad \|\nabla h_\mu(u)\| \le G_h.$$

If $\mu_k \downarrow 0$, $x_k \to x$, and $\nabla h_{\mu_k}(x_k) \to \xi$, then $\xi \in \partial h(x)$.

For the finite-accuracy conversion, we use the Moreau envelope

$$h_\mu(u) = \min_y \left\{ h(y) + \frac{\|u - y\|^2}{2\mu} \right\}, \qquad \mu > 0,$$

of a finite-valued convex Lipschitz function h. This is a special case of Assumption D.2. For the regularizer used in the experiments, the Moreau envelope of $\lambda \|\cdot\|_{2,1}$ satisfies this assumption with $c_h = \lambda^2 d/2$, $\kappa_h = 1$, and $G_h = \lambda\sqrt{d}$.

Proposition D.3 (Consistency as smoothing vanishes)**.** *Let $(\mathbf{x}_\ell, \mu_\ell)$ be $(\delta_{\mathrm{sm},\ell}, \mu_\ell)$-smoothed stationary certificates with $\delta_{\mathrm{sm},\ell} \to 0$ and $\mu_\ell \to 0$. Every accumulation point of the mean points $\bar{x}_\ell$ is stationary for F, and all local copies converge to that point along the same subsequence.*

Proof. Along a subsequence, compactness gives $\bar{x}_\ell \to x_\star$. Boundedness of ∇h_{μ_ℓ} gives a further subsequence converging to some ξ, and Assumption D.2 yields $\xi \in \partial h(x_\star)$. Continuity of the tangent projector and of ∇f_i then gives

$$\mathcal{P}_{T_{x_\star}\mathcal{M}}\left(\frac{1}{n}\sum_i \nabla f_i(x_\star) + \xi\right) = 0.$$

The consensus part of Definition B.1 gives convergence of every local copy. □

Proposition D.4 (Finite-accuracy conversion and local outputs)**.** *Let $\mathbf{x} \in \mathcal{M}^n$ be a random network output with a well-defined projected mean $\bar{x}$. Suppose Assumption D.1 and the tangent-projector bound* (D.3) *hold. Suppose h is convex and globally Lipschitz, h_μ is its Moreau envelope, and $\|\nabla h_\mu\| \le G_h$. Assume $0 < \mu \le \bar{\mu}$ almost surely and*

$$\mathbb{E}\|\mathrm{grad} F_\mu(\bar{x})\|^2 \le S, \qquad \max_i \|x_i - \bar{x}\| \le d \quad \textit{almost surely},$$

where $S, d, \bar{\mu}$ are deterministic bounds. Set

$$y = \mathrm{prox}_{\mu h}(\bar{x}), \qquad z = \frac{\bar{x} - y}{\mu} = \nabla h_\mu(\bar{x}), \qquad \eta = \max\{\sqrt{S}, G_h\bar{\mu}, d\}.$$

Then $\mathbf{x}$ is distributed generalized η-stationary in the sense of Definition 3.1. For every agent, the same y, z satisfy $z \in \partial h(y)$ and

$$\mathbb{E}\|\mathcal{P}_{T_{x_i}\mathcal{M}}(\nabla \bar{f}(x_i) + z)\| \le \sqrt{S} + C_{\mathrm{loc}} d, \tag{D.1}$$

$$\|x_i - y\| \le d + G_h\bar{\mu} \quad \textit{almost surely}, \tag{D.2}$$

where $C_{\mathrm{loc}} = L_f + L_{\mathcal{P}}(L_G + G_h)$. Thus every local output is generalized $C_\star\eta$-stationary, with $C_\star = \max\{1 + C_{\mathrm{loc}}, 2\}$.

Proof. The proximal subproblem is strongly convex and has a unique minimizer. Its optimality condition and the Moreau gradient identity give

$$0 \in \partial h(y) + \frac{y - \bar{x}}{\mu}, \qquad z \in \partial h(y), \qquad z = \nabla h_\mu(\bar{x}).$$

The proximal mapping is continuous jointly in its input and positive parameter, so y, z are measurable on the output probability space. Moreover, $\|z\| \le G_h$ and $\|\bar{x} - y\| = \mu\|z\| \le G_h\bar{\mu}$ almost surely. Since

$$\mathcal{P}_{T_{\bar{x}}\mathcal{M}}(\nabla \bar{f}(\bar{x}) + z) = \mathrm{grad} F_\mu(\bar{x}),$$

Cauchy–Schwarz gives an expected tangent residual at most $\sqrt{S}$. Together with the consensus bound, this establishes the network claim.

For the local claim, write $P_i = \mathcal{P}_{T_{x_i}\mathcal{M}}$ and $P_{\bar{x}} = \mathcal{P}_{T_{\bar{x}}\mathcal{M}}$. The exact decomposition

$$\begin{aligned} P_i(\nabla \bar{f}(x_i) + z) = {} & P_{\bar{x}}(\nabla \bar{f}(\bar{x}) + z) \\ & + P_i(\nabla \bar{f}(x_i) - \nabla \bar{f}(\bar{x})) \\ & + (P_i - P_{\bar{x}})(\nabla \bar{f}(\bar{x}) + z) \end{aligned}$$

and the bounds $\|P_i\|_{\mathrm{op}} \le 1$, $\|P_i - P_{\bar{x}}\|_{\mathrm{op}} \le L_{\mathcal{P}}\|x_i - \bar{x}\|$, and $\|\nabla \bar{f}(\bar{x})\| \le L_G$ imply

$$\|P_i(\nabla \bar{f}(x_i) + z)\| \le \|\mathrm{grad} F_\mu(\bar{x})\| + C_{\mathrm{loc}}\|x_i - \bar{x}\|.$$

Taking expectations proves (D.1). The triangle inequality gives $\|x_i - y\| \le \|x_i - \bar{x}\| + \|\bar{x} - y\|$, proving (D.2). Both inequalities hold for every agent with the same witnesses and constants, independently of μ. □

Because $\mathcal{M}$ is compact and C^2, the tangent-projector map is Lipschitz: there is $L_{\mathcal{P}} > 0$ such that

$$\|\mathcal{P}_{T_x\mathcal{M}} - \mathcal{P}_{T_y\mathcal{M}}\|_{\mathrm{op}} \le L_{\mathcal{P}}\|x - y\|, \qquad x, y \in \mathcal{M}. \tag{D.3}$$

The following standard consequences are stated with the radii needed later.

Lemma D.5 (Smoothness bounds). *Under Assumption D.1, for $x, y \in \mathcal{M}$,*

$$f_i(y) - f_i(x) \le \langle \mathrm{grad} f_i(x), y - x \rangle + \frac{L_g}{2}\|y - x\|^2, \tag{D.4}$$

$$\|\mathrm{grad} f_i(x) - \mathrm{grad} f_i(y)\| \le (L_f + L_G L_{\mathcal{P}})\|x - y\|, \tag{D.5}$$

where $L_g := L_f + L_G/(2\delta)$.

Lemma D.6 (Uniform projection expansion). *There is $Q > 0$ such that, for every $x \in \mathcal{M}$ and $\|u\| \le 3\delta$,*

$$\|\mathcal{P}_{\mathcal{M}}(x + u) - x - \mathcal{P}_{T_x\mathcal{M}} u\| \le Q\|u\|^2. \tag{D.6}$$

Lemma D.7 (Euclidean and manifold means). *If $\mathbf{x} \in \mathcal{N}_\delta$, then*

$$\|\bar{x} - \hat{x}\| \le \frac{M_2}{n}\|\mathbf{x} - \bar{\mathbf{x}}\|^2 \tag{D.7}$$

for a constant $M_2 > 0$ depending only on $\mathcal{M}$ and δ.

Lemmas D.5–D.7 follow from compactness, (B.2), and a uniform Taylor expansion of the nearest-point projection; see (Deng & Hu, 2025a). The strict radius $3\delta < r_{\mathcal{P}}$ keeps the Taylor constant uniform.

E Decentralized dynamic smoothing with partial gradient tracking

We present the dynamic smoothing method and its horizon-free stepsize guarantee. For $\mu > 0$, let

$$\begin{aligned} F_{i,\mu}(x) &:= f_i(x) + h_\mu(x), \\ \nabla F_{i,\mu}(x) &= \nabla f_i(x) + \nabla h_\mu(x), \\ \mathrm{grad} F_{i,\mu}(x) &:= \mathrm{grad} f_i(x) + \mathrm{grad} h_\mu(x), \\ \mathrm{grad} h_\mu(x) &:= \mathcal{P}_{T_x\mathcal{M}}(\nabla h_\mu(x)). \end{aligned}$$

E.1 The algorithm

The method uses projected consensus and partial gradient tracking. Each agent communicates only $x_{i,k}$ and $s_{i,k}$ with its neighbors and evaluates the common dynamic-smoothing correction $\mathrm{grad} h_{\mu_k}(x_{i,k})$ locally. The smoothing term is not tracked because the Lipschitz constant of ∇h_{μ_k} grows as $\mathcal{O}(1/\mu_k)$.

Partial versus full tracking. A natural full-tracking counterpart would update a tracker by

$$s^{\mathrm{full}}_{i,k+1} = \sum_{j=1}^{n} W_{ij} s^{\mathrm{full}}_{j,k} + \mathrm{grad} F_{i,\mu_{k+1}}(x_{i,k+1}) - \mathrm{grad} F_{i,\mu_k}(x_{i,k}).$$

Its regularizer increment contains both

$$\mathrm{grad} h_{\mu_k}(x_{i,k+1}) - \mathrm{grad} h_{\mu_k}(x_{i,k}) \quad \text{and} \quad \mathrm{grad} h_{\mu_{k+1}}(x_{i,k+1}) - \mathrm{grad} h_{\mu_k}(x_{i,k+1}).$$

The first difference is bounded by a Lipschitz factor of order $1/\mu_k$, and the second reflects the changing smoothing parameter. Thus, a full tracker couples network disagreement directly with a smoothing gradient whose Lipschitz constant increases as μ_k decreases. The partial tracker instead uses

$$s_{i,k+1} = \sum_{j=1}^{n} W_{ij} s_{j,k} + \mathrm{grad} f_i(x_{i,k+1}) - \mathrm{grad} f_i(x_{i,k}), \qquad r_{i,k} = \mathrm{grad} h_{\mu_k}(x_{i,k}),$$

and adds $r_{i,k}$ only to the local search direction. This is valid because h_{μ_k} is common across agents: gradient tracking is needed to recover the average of the heterogeneous local gradients, not to represent an identical regularizer available locally at every node. The tracking recursion is therefore governed by the fixed smoothness of f_i, while the dependence on μ_k is addressed in the descent

Algorithm 2 DSPGT: decentralized smoothing with partial gradient tracking

Require: A consensual initial point $x_{1,0} = \cdots = x_{n,0} \in \mathcal{M}$; iteration budget $K \geq 2$; step sizes $\{\alpha_k\}_{k=0}^{K-1}$; smoothing parameters $\{\mu_k\}_{k=0}^{K}$; and a shared random seed.

1: Set $s_{i,0} = \mathrm{grad} f_i(x_{i,0})$ on each node $i \in [n]$.
2: **for** $k = 0, \ldots, K-1$ **do**
3: Set $r_{i,k} = \mathrm{grad} h_{\mu_k}(x_{i,k})$ and $v_{i,k} = \mathcal{P}_{T_{x_{i,k}}\mathcal{M}}(s_{i,k}) + r_{i,k}$.
4: Update $x_{i,k+1} = \mathcal{P}_{\mathcal{M}}(\sum_j W_{ij} x_{j,k} - \alpha_k v_{i,k})$.
5: Track only the local smooth gradient:
$s_{i,k+1} = \sum_j W_{ij} s_{j,k} + \mathrm{grad} f_i(x_{i,k+1}) - \mathrm{grad} f_i(x_{i,k})$.
6: **end for**
7: Using the shared seed, sample R_K uniformly from $\mathcal{I}_K = \{\lceil K/2 \rceil, \ldots, K-1\}$ and return x_{i,R_K} at node i.

analysis. Here $\alpha_k > 0$ is the iteration-dependent step size. The tracker estimates the network average of the local smooth gradients, while every agent computes $\mathrm{grad} h_{\mu_k}(x_{i,k})$ locally.

A shared tail index needs no global residual evaluation and can be implemented by storing the tail iterates or by synchronized reservoir sampling. One projection-oracle iteration consists of local evaluations of $\mathrm{grad} f_i$ and $\mathrm{grad} h_{\mu_k}$, one neighbor-mixing round, and one exact nearest-point projection onto $\mathcal{M}$.

E.2 Convergence analysis

This subsection gives the complexity result for Algorithm 2. The descent estimate for the smoothed objective contains the local curvature scale $L_{\mu_k} = \mathcal{O}(1/\mu_k)$. Partial tracking prevents an additional $1/\mu_k$ factor from entering the tracking recursion and its induced consensus control. Below, Lemma E.3 involves only differences of $\mathrm{grad} f_i$, Lemma E.5 yields a consensus constant independent of μ_k, and the common smoothing correction is subsequently controlled through the Jensen gap of its ambient convex smoothing. For later use, define

$$L := \max\{L_G, L_g, L_f + L_G L_{\mathcal{P}}\}. \tag{E.1}$$

Here L_f and L_G are from Assumption D.1, while L_g and $L_{\mathcal{P}}$ are defined in Lemma D.5 and (D.3), respectively. For the smoothed objective, we use the shorthand

$$L_\mu := L + G_h L_{\mathcal{P}} + \frac{\kappa_h}{\mu} + \frac{L_G + G_h}{2\delta},$$

so $L_\mu = \mathcal{O}(1/\mu)$. The following theorem gives an anytime stepsize schedule and the optimized complexity obtained by balancing the tracking, consensus, and smoothing errors.

Theorem E.1 (Horizon-free stepsize complexity)**.** *Fix $p \in (0,1)$. Let $K \geq 2$, and let $\{\mathbf{x}_k\}_k$ be generated by Algorithm 2, where $s_{i,k}$ tracks only $\mathrm{grad} f_i$ and $v_{i,k} = \mathcal{P}_{T_{x_{i,k}}\mathcal{M}}(s_{i,k}) + \mathrm{grad} h_{\mu_k}(x_{i,k})$. Suppose that Assumptions C.1, D.1, and D.2 hold. For fixed $\mu_0, \alpha_0 > 0$, set*

$$\mu_k := \frac{\mu_0}{(k+1)^p}, \qquad \alpha_k := \frac{\alpha_0}{(k+1)^p},$$

$$\hat{q}_k := \frac{1}{n}\sum_{i=1}^{n} \mathrm{grad} f_i(\bar{x}_k) + \mathrm{grad} h_{\mu_k}(\bar{x}_k),$$

$$\tilde{q}_k := \frac{1}{n}\sum_{i=1}^{n} \left(\mathrm{grad} f_i(x_{i,k}) + \mathrm{grad} h_{\mu_k}(x_{i,k})\right).$$

$$D_s := \frac{(3L)^2}{(1-\sigma_2)^2}, \qquad B_s := \sqrt{D_s} + G_h,$$

$$\Delta_{\mu,K} := \sum_{k=0}^{K-1} |\mu_{k+1} - \mu_k|.$$

Let R_K be uniform on $\mathcal{I}_K := \{\lceil K/2\rceil, \dots, K-1\}$ and independent of the algorithmic history. Choose δ as in (C.1). *If $\mathbf{x}_0 \in \mathcal{N}_\delta$, $x_{1,0} = \cdots = x_{n,0}$ and*

$$0 < \alpha_0 < \min\left\{\frac{c_0}{1+L_{\mu_0}}, \frac{\delta}{\sqrt{n}B_s}, \frac{(1-\rho)\delta}{\ell_\delta\sqrt{n}B_s}, c_1\right\}, \tag{E.2}$$

where $c_0, c_1 > 0$ may depend on the fixed problem, manifold, network, and on p, but are independent of μ_0, α_0, and K. Notice that

$$\frac{(1-\rho)\delta}{\ell_\delta\sqrt{n}B_s} = \frac{\big(R(1-\sigma_2)-2\delta\big)\delta}{R\sqrt{n}B_s}.$$

Thus the two network-dependent bounds in (E.2) *are exactly the single-mixing counterparts of the $t = 1$ bounds in (Deng & Hu, 2025b, Theorem 3.6), with B_s additionally accounting for the bounded common smoothing correction. Then*

$$\begin{aligned}
\mathbb{E}\|\tilde{q}_{R_K}\|^2 &\le C\frac{1+c_h\Delta_{\mu,K}}{|\mathcal{I}_K|\alpha_{K-1}},\\
\mathbb{E}\|\mathrm{grad}F_{\mu_{R_K}}(\bar{x}_{R_K})\|^2 &\le C\frac{1+c_h\Delta_{\mu,K}}{|\mathcal{I}_K|\alpha_{K-1}},\\
\mathbb{E}\left[\frac{1}{n}\|\mathbf{x}_{R_K}-\bar{\mathbf{x}}_{R_K}\|^2\right] &\le C\alpha^2_{\lceil K/2\rceil-1}, \qquad \mu_{R_K} \le \mu_{\lceil K/2\rceil}.
\end{aligned}$$

Here and below, C may depend on the fixed manifold, network, number of agents, problem constants, p, and μ_0, but it is independent of K. The same estimates imply the existence of a deterministic $k_K \in \mathcal{I}_K$ satisfying the mean-point stationarity bound; the consensus bound holds pointwise for every $k \in \mathcal{I}_K$.

In particular, choosing $p = 1/3$ gives the horizon-free schedules $\mu_k = \mu_0/(k+1)^{1/3}$ and $\alpha_k = \alpha_0/(k+1)^{1/3}$. The randomized output then satisfies

$$\begin{aligned}
\mathbb{E}\left[\|\mathrm{grad}F_{\mu_{R_K}}(\bar{x}_{R_K})\|^2 + \frac{1}{n}\|\mathbf{x}_{R_K}-\bar{\mathbf{x}}_{R_K}\|^2\right] &= \mathcal{O}((K+1)^{-2/3}),\\
\mu_{R_K} &= \mathcal{O}((K+1)^{-1/3}).
\end{aligned}$$

Consequently, the randomized rule obtains an expected $(\delta_{\mathrm{sm}}, \mathcal{O}(\delta_{\mathrm{sm}}^{1/2}))$-smoothed certificate in $\mathcal{O}(\delta_{\mathrm{sm}}^{-3/2})$ per-agent projection-oracle iterations. Equivalently, for an unsquared residual tolerance ε, setting $\delta_{\mathrm{sm}} = \varepsilon^2$ gives $K = \mathcal{O}(\varepsilon^{-3})$ and $\mu_{R_K} = \mathcal{O}(\varepsilon)$. Because each iteration uses one multiplication by W, the communication complexities are $\mathcal{O}(\delta_{\mathrm{sm}}^{-3/2})$ and $\mathcal{O}(\varepsilon^{-3})$, respectively. Proposition D.3 applies to any sequence of the deterministic certificates furnished by the existence statement as their tolerances and smoothing parameters vanish. For Moreau smoothing, Proposition 4.2 also yields finite-accuracy generalized stationarity of the original problem at the network output and every local output. The schedules $\{\alpha_k\}$ and $\{\mu_k\}$ depend only on the current iteration and therefore do not require the stopping horizon K in advance.

Complexity for the original problem. For the output of Theorem E.1, set $m_K = \lceil K/2\rceil$ and use

$$S_K = C\frac{1+c_h\Delta_{\mu,K}}{|\mathcal{I}_K|\alpha_{K-1}}, \qquad d_K = \sqrt{nD_{c,p}}B_s\alpha_{m_K-1}, \qquad \bar{\mu}_K = \mu_{m_K}.$$

Theorem E.1 bounds the expected squared tangent residual by S_K. Equation (E.10) bounds the maximum local deviation by d_K, and monotonicity gives $\mu_{R_K} \le \bar{\mu}_K$ almost surely. Proposition D.4 therefore applies with these three deterministic bounds. Since $\Delta_{\mu,K} \le \mu_0$, $|\mathcal{I}_K| \ge K/3$, and $\alpha_{K-1} = \alpha_0 K^{-p}$, there is $A > 0$, independent of K, such that

$$\eta_K := \max\{\sqrt{S_K}, G_h\bar{\mu}_K, d_K\} \le A\big(K^{-(1-p)/2} + K^{-p}\big).$$

For $0 < \varepsilon \le 1$, taking

$$K \ge \max\left\{2, \left(\frac{2C_\star A}{\varepsilon}\right)^{2/(1-p)}, \left(\frac{2C_\star A}{\varepsilon}\right)^{1/p}\right\}$$

ensures $C_\star \eta_K \le \varepsilon$. Thus the network output and every local output meet their respective generalized ε-stationarity conditions. The exponent is minimized at $p = 1/3$, yielding $\mathcal{O}(\varepsilon^{-3})$ iterations. Each iteration uses one nearest-point projection and one synchronized neighbor exchange, so the projection and communication complexities have the same order. The constants may depend on the fixed problem, manifold, network, n, p, α_0, and μ_0, but not on K or ε. The deterministic index k_K in Theorem E.1 admits the same conversion with its non-expected residual bound.

Comparison with proximal tracking. For Moreau smoothing, Theorem E.1 and Proposition 4.2 count smoothing-gradient evaluations, projection-oracle calls, and communication rounds to obtain generalized stationarity of the original problem. The auxiliary subgradient condition holds exactly, while tangent stationarity, splitting feasibility, and consensus are controlled. DR-ProxGT Wang et al. (2025) uses a composite tangent-space proximal oracle and a proximal certificate. The two bounds therefore count different local oracles.

The proof is given in Subsection E.2.2. The recalled projected-consensus estimate is used as a geometric tool, while the new part of the analysis controls dynamic smoothing and partial tracking and couples a small residual with a vanishing smoothing parameter at one output.

E.2.1 Two network lemmas

The first lemma replaces an ill-defined second derivative of $x \mapsto \mathcal{P}_{T_x\mathcal{M}}$ outside $\mathcal{M}$ by the intrinsic Lipschitz constant in (D.3).

Lemma E.2. *Let $L_2 := 2L_{\mathcal{P}}$. For every $\mathbf{x} \in \mathcal{M}^n$,*

$$\left\| \sum_{i=1}^{n} \mathrm{grad}\phi_i(\mathbf{x}) \right\| \le L_2 \|\mathbf{x} - \bar{\mathbf{x}}\|^2. \tag{E.3}$$

Proof. Set $a_i := \nabla\phi_i(\mathbf{x}) = x_i - \sum_j W_{ij} x_j$. Double stochasticity gives $\sum_i a_i = 0$. Hence

$$\sum_i \mathrm{grad}\phi_i(\mathbf{x}) = \sum_i (\mathcal{P}_{T_{x_i}\mathcal{M}} - \mathcal{P}_{T_{\bar{x}}\mathcal{M}}) a_i.$$

By (D.3), Cauchy–Schwarz, and $\|I - W\|_2 \le 2$,

$$\left\| \sum_i \mathrm{grad}\phi_i(\mathbf{x}) \right\| \le L_{\mathcal{P}} \|\mathbf{x} - \bar{\mathbf{x}}\| \|\mathbf{a}\| \le 2L_{\mathcal{P}} \|\mathbf{x} - \bar{\mathbf{x}}\|^2.$$

□

E.2.2 Proof of Theorem E.1

For ease of notation, define

$$\hat{g}_k := \frac{1}{n} \sum_{i=1}^{n} \mathrm{grad} f_i\left(x_{i,k}\right), \quad \hat{\mathbf{G}}_k := \left(\mathbf{1}_n \otimes I_d\right) \hat{g}_k,$$
$$\mathbf{G}_k := [\mathrm{grad} f_1(x_{1,k})^\top, \ldots, \mathrm{grad} f_n(x_{n,k})^\top]^\top,$$
$$\mathbf{v}_k := [v_{1,k}^\top, \ldots, v_{n,k}^\top]^\top, \quad \mathbf{s}_k := [s_{1,k}^\top, \ldots, s_{n,k}^\top]^\top.$$

In this subsection the tracking variable is reserved for the local smooth gradients. We further denote

$$r_{i,k} := \mathrm{grad} h_{\mu_k}(x_{i,k}), \qquad \mathbf{r}_k := [r_{1,k}^\top, \ldots, r_{n,k}^\top]^\top,$$
$$\tilde{q}_k := \frac{1}{n} \sum_{i=1}^{n} \big(\mathrm{grad} f_i(x_{i,k}) + r_{i,k}\big),$$
$$\hat{q}_k := \frac{1}{n} \sum_{i=1}^{n} \mathrm{grad} f_i(\bar{x}_k) + \mathrm{grad} h_{\mu_k}(\bar{x}_k).$$

Thus $v_{i,k} = \mathcal{P}_{T_{x_{i,k}}\mathcal{M}}(s_{i,k}) + r_{i,k}$. The quantity $\tilde{q}_k$ is the local averaged smoothed residual defined by the decentralized iterates, while the stationarity residual at the mean point is $\|\mathrm{grad} F_{\mu_k}(\bar{x}_k)\| = \|\hat{q}_k\|$. We first relate the tracking error $\mathbf{s}_k - \hat{\mathbf{G}}_k$ to the motion of the primal iterates.

Lemma E.3. *For the iterates generated by Algorithm 2 and every $k \geq 0$,*

$$\|\mathbf{G}_{k+1} - \mathbf{G}_k\| \leq 4L\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + 2L\alpha_k\|\mathbf{v}_k\|, \tag{E.4}$$

$$\|\mathbf{s}_{k+1} - \hat{\mathbf{G}}_{k+1}\| \leq \sigma_2\|\mathbf{s}_k - \hat{\mathbf{G}}_k\| + \|\mathbf{G}_{k+1} - \mathbf{G}_k\|. \tag{E.5}$$

Proof. The recursion for $s_{i,k}$ contains only $\mathrm{grad} f_i(x_{i,k+1}) - \mathrm{grad} f_i(x_{i,k})$. Hence the regularizer enters this lemma only through the displacement of the primal iterates, represented by $\mathbf{v}_k$. The Lipschitz continuity of $\mathbf{G}$ yields

$$\|\mathbf{G}_{k+1} - \mathbf{G}_k\| \leq L\|\mathbf{x}_{k+1} - \mathbf{x}_k\|.$$

Moreover,

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}_k\| &\leq \|\mathbf{x}_{k+1} - \mathbf{x}_k + (I_{nd} - \mathbf{W})\mathbf{x}_k + \alpha_k\mathbf{v}_k\| + \|(I_{nd} - \mathbf{W})\mathbf{x}_k + \alpha_k\mathbf{v}_k\| \\ &\leq 2\|(I_{nd} - \mathbf{W})\mathbf{x}_k + \alpha_k\mathbf{v}_k\| \leq 4\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + 2\alpha_k\|\mathbf{v}_k\|,\end{aligned}$$

The first inequality is the triangle inequality. For the second, the projection residual is no larger than the distance to the feasible point $\mathbf{x}_k$; the last estimate uses $\|I - \mathbf{W}\|_2 \leq 2$. It follows from the definition of $\hat{\mathbf{G}}_{k+1}$ that

$$\begin{aligned}\mathbf{s}_{k+1} - \hat{\mathbf{G}}_{k+1} &= ((I_n - J) \otimes I_d)\mathbf{s}_{k+1} \\ &= ((I_n - J) \otimes I_d)((W \otimes I_d)\mathbf{s}_k + \mathbf{G}_{k+1} - \mathbf{G}_k) \\ &= ((W - J) \otimes I_d)(\mathbf{s}_k - \hat{\mathbf{G}}_k) + ((I_n - J) \otimes I_d)(\mathbf{G}_{k+1} - \mathbf{G}_k).\end{aligned}$$

Using the spectral property of W, we conclude that (E.5) holds. □

We next establish the uniform tracker bound needed for neighborhood invariance.

Lemma E.4. *Let $\{\mathbf{x}_k\}_k$ be generated by Algorithm 2 under Assumptions C.1, D.1, and D.2. Define $D_s = (3L)^2/(1-\sigma_2)^2$ and $B_s = \sqrt{D_s} + G_h$. If $x_{1,0} = \cdots = x_{n,0}$, $0 \leq \alpha_k \leq \alpha_0$ for every k, and*

$$\alpha_0 < \min\left\{\frac{\delta}{\sqrt{n}B_s}, \frac{(1-\rho)\delta}{\ell_\delta\sqrt{n}B_s}\right\}, \tag{E.6}$$

then $\mathbf{x}_k \in \mathcal{N}_\delta$ and

$$\|\mathbf{s}_k\|^2 \leq nD_s, \qquad k \geq 0. \tag{E.7}$$

Proof. Double stochasticity and the initialization of the tracker imply, by telescoping its recursion, that

$$\frac{1}{n}\sum_{i=1}^n s_{i,k} = \frac{1}{n}\sum_{i=1}^n \mathrm{grad} f_i(x_{i,k}) = \hat{g}_k, \qquad k \geq 0. \tag{E.8}$$

The tracker bound does not use neighborhood invariance. Since $\|\mathbf{s}_0\| \leq \sqrt{n}L$, tracker conservation gives $\hat{\mathbf{s}}_k = \hat{\mathbf{G}}_k$, and hence

$$\begin{aligned}\|\mathbf{s}_{k+1} - \hat{\mathbf{G}}_k\| &\leq \sigma_2\|\mathbf{s}_k\| + \|\mathbf{G}_{k+1} - \mathbf{G}_k\| \\ &\leq \sigma_2\|\mathbf{s}_k\| + 2\sqrt{n}L, \\ \|\mathbf{s}_{k+1}\| &\leq \sigma_2\|\mathbf{s}_k\| + 3\sqrt{n}L.\end{aligned}$$

Induction yields $\|\mathbf{s}_k\| \leq 3\sqrt{n}L/(1-\sigma_2) = \sqrt{nD_s}$. Consequently $\|\mathbf{v}_k\| \leq \|\mathbf{s}_k\| + \sqrt{n}G_h \leq \sqrt{n}B_s$ and also $\max_i \|v_{i,k}\| \leq \sqrt{n}B_s$. Starting from consensus, apply Lemma C.2 inductively with $B_\infty = \sqrt{n}B_s$ and $B_2 = B_s$. Since $\alpha_k \leq \alpha_0$, condition (E.6) is exactly (C.2). Thus all projections are unique and $\mathbf{x}_k \in \mathcal{N}_\delta$ for every k. □

With the above lemma, we have the following result on the consensus error.

Lemma E.5. *Let $\{\mathbf{x}_k\}$ be generated by Algorithm 2. Under the conditions of Lemma E.4, suppose additionally that $\alpha_k = \alpha_0/(k+1)^p$ for some $p \in (0,1)$.*

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \le \rho\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \ell_\delta \alpha_k \|\mathbf{v}_k\|. \tag{E.9}$$

Moreover, define

$$C_{p,\rho} := \sum_{m=0}^{\infty} \rho^m (m+1)^p, \qquad D_{c,p} := \ell_\delta^2 C_{p,\rho}^2.$$

Then, for every $k \ge 1$,

$$\frac{1}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 \le D_{c,p} B_s^2 \alpha_{k-1}^2. \tag{E.10}$$

Proof. The recursion is (C.3) with $\mathbf{u}_k = \mathbf{v}_k$. By (E.7) and Assumption D.2, $\|\mathbf{v}_k\| \le \sqrt{n} B_s$. Since the initial copies agree, unrolling (E.9) gives, for $k \ge 1$,

$$\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| \le \ell_\delta \sqrt{n} B_s \sum_{j=0}^{k-1} \rho^{k-1-j} \alpha_j.$$

Writing $m = k - 1 - j$ and using

$$\frac{\alpha_j}{\alpha_{k-1}} = \left(\frac{k}{k-m}\right)^p \le (m+1)^p$$

proves (E.10). The series defining $C_{p,\rho}$ is finite because $\rho < 1$; hence $D_{c,p}$ is independent of k and of the stopping horizon. □

Lemma E.6 (Uniform projected-mixing expansion)**.** *Let $\mathbf{x} \in \mathcal{N}_\delta$, and let $v_i \in T_{x_i}\mathcal{M}$ satisfy $\max_i \|v_i\| \le B_\infty$. Define*

$$y_i := \sum_j W_{ij} x_j, \quad a_i := y_i - x_i, \quad z_i := y_i - \alpha v_i, \quad x_i^+ := \mathcal{P}_{\mathcal{M}}(z_i),$$

where $0 \le \alpha B_\infty < \delta$, and assume $\mathbf{x}^+ := (x_1^+, \dots, x_n^+) \in \mathcal{N}_\delta$. Then there is a constant $C_G > 0$, depending only on $\mathcal{M}$, δ, and W, such that

$$\|\mathcal{P}_{N_{x_i}\mathcal{M}} a_i\| \le \frac{1}{4\delta} \sum_j W_{ij} \|x_j - x_i\|^2, \tag{E.11}$$

$$x_i^+ - x_i = \mathcal{P}_{T_{x_i}\mathcal{M}} a_i - \alpha v_i + e_i, \quad \|e_i\| \le Q\|a_i - \alpha v_i\|^2, \tag{E.12}$$

$$\frac{1}{n}\sum_i \operatorname{dist}(z_i, \mathcal{M}) \le \frac{C_G}{n}\|\mathbf{x} - \bar{\mathbf{x}}\|^2 + \frac{C_G \alpha^2}{n}\|\mathbf{v}\|^2, \tag{E.13}$$

$$\|\bar{x}^+ - \bar{x} + \alpha\hat{v}\| \le \frac{C_G}{n}\left(\|\mathbf{x} - \bar{\mathbf{x}}\|^2 + \|\mathbf{x}^+ - \bar{\mathbf{x}}^+\|^2 + \alpha^2\|\mathbf{v}\|^2\right), \tag{E.14}$$

$$\|\bar{x}^+ - \bar{x}\|^2 \le \frac{C_G}{n}\left(\|\mathbf{x} - \bar{\mathbf{x}}\|^2 + \|\mathbf{x}^+ - \bar{\mathbf{x}}^+\|^2 + \alpha^2\|\mathbf{v}\|^2\right), \tag{E.15}$$

where $\hat{v} = n^{-1}\sum_i v_i$. All constants are independent of the smoothing parameter and of the horizon.

Proof. For a unit normal $w \in N_{x_i}\mathcal{M}$, (B.2) gives

$$|\langle w, a_i\rangle| \le \frac{1}{4\delta}\sum_j W_{ij}\|x_j - x_i\|^2.$$

Taking the supremum over w proves (E.11). Since $\mathbf{x} \in \mathcal{N}_\delta$, every $\|x_j - \bar{x}\| \le \delta$ and hence $\|a_i\| \le 2\delta$. Therefore $\|a_i - \alpha v_i\| < 3\delta < r_{\mathcal{P}}$. Lemma D.6 therefore gives (E.12) with the same Q for all i and all iterations.

The spectral bound $\|I - W\|_2 \le 2$ yields $\sum_i \|a_i\|^2 \le 4\|\mathbf{x} - \bar{\mathbf{x}}\|^2$. Moreover,

$$z_i - x_i^+ = \mathcal{P}_{N_{x_i}\mathcal{M}} a_i - e_i.$$

Summing this identity, using (E.11), double stochasticity, and

$$\sum_i \|a_i - \alpha v_i\|^2 \le 2\sum_i \|a_i\|^2 + 2\alpha^2\|\mathbf{v}\|^2,$$

proves (E.13).

Because $\sum_i a_i = 0$, Lemma E.2 and (E.12) give

$$\|\hat{x}^+ - \hat{x} + \alpha\hat{v}\| \le \frac{C_G}{n}\left(\|\mathbf{x} - \bar{\mathbf{x}}\|^2 + \alpha^2\|\mathbf{v}\|^2\right).$$

Adding the two mean corrections from Lemma D.7 proves (E.14). Finally, $\|\hat{v}\|^2 \le \|\mathbf{v}\|^2/n$; squaring the last display and using $\|\mathbf{x}-\bar{\mathbf{x}}\|^2/n \le \delta^2$, $\|\mathbf{x}^+-\bar{\mathbf{x}}^+\|^2/n \le \delta^2$, and $\alpha^2\|\mathbf{v}\|^2/n \le \alpha^2 B_\infty^2 < \delta^2$ proves (E.15) after enlarging C_G. □

Role of partial tracking in the analysis. The absence of κ_h/μ_k in the recursion (E.5) and in the consensus constants $D_s, D_{c,p}$ is a consequence of the algorithmic split, rather than a notational convenience. If the tracker increment included $\mathrm{grad}h_{\mu_k}$, bounding its variation across different local iterates would introduce the smoothing Lipschitz modulus κ_h/μ_k into these network-error estimates. Partial tracking confines this dependence to the descent argument below, where convexity provides the Jensen-gap control required for the local smoothing correction.

We combine the geometric estimates in Section D with the preceding network bounds to obtain sufficient decrease of a dynamically smoothed potential.

For the convex-smoothing analysis below, set

$$\bar{f}(z) := \frac{1}{n}\sum_{i=1}^n f_i(z), \qquad H_\mu(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n h_\mu(x_i), \qquad \Psi_\mu(\mathbf{x}) := \bar{f}(\bar{x}) + H_\mu(\mathbf{x}).$$

Unlike $F_\mu(\bar{x})$, the potential Ψ_μ evaluates the common regularizer at the local iterates, where its directions are computed in Algorithm 2. At a consensus point the two potentials coincide.

Lemma E.7 (Descent with ambient convex smoothing)**.** *Let $\{\mathbf{x}_k\}_k$ be generated by Algorithm 2 and set*

$$y_{i,k} := \sum_{j=1}^n W_{ij}x_{j,k}.$$

Under Assumptions C.1, D.1, and D.2,

$$H_{\mu_k}(\mathbf{x}_k) - \frac{1}{n}\sum_{i=1}^n h_{\mu_k}(y_{i,k}) \ge 0.$$

Suppose the neighborhood and mixing conditions of Lemma E.4 hold. If

$$\alpha_k\left(1 + \frac{\kappa_h}{\mu_k}\right) \le c_0$$

for a sufficiently small constant $c_0 > 0$, then there exist constants $c_q, c_v, C > 0$, independent of μ_k, such that

$$\begin{aligned}\Psi_{\mu_{k+1}}(\mathbf{x}_{k+1}) \le \Psi_{\mu_k}(\mathbf{x}_k) &- c_q\alpha_k\|\tilde{q}_k\|^2 - \frac{c_v\alpha_k}{n}\|\mathbf{v}_k\|^2 \\ &- \frac{1}{4}\left(H_{\mu_k}(\mathbf{x}_k) - \frac{1}{n}\sum_{i=1}^n h_{\mu_k}(y_{i,k})\right) \\ &+ \frac{C}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{C}{n}\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2 \\ &+ \frac{C\alpha_k}{n}\|\mathbf{s}_k - \hat{\mathbf{G}}_k\|^2 + c_h|\mu_{k+1} - \mu_k|.\end{aligned} \tag{E.16}$$

Thus, in the descent estimate for the local-regularizer potential, no term of the form $\left(L + G_hL_{\mathcal{P}} + \kappa_h/\mu_k\right)^2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2/(nL_{\mu_k})$ occurs. For conversion to the mean-point stationarity measure only, one still has

$$\|\hat{q}_k - \tilde{q}_k\| \le \left(L + G_hL_{\mathcal{P}} + \frac{\kappa_h}{\mu_k}\right)\frac{1}{\sqrt{n}}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|. \tag{E.17}$$

Proof. Write $h_k := h_{\mu_k}$ and $H_k := H_{\mu_k}$. Ambient convexity and the double stochasticity of W imply

$$\frac{1}{n}\sum_{i=1}^{n} h_k(y_{i,k}) \le \frac{1}{n}\sum_{i,j=1}^{n} W_{ij} h_k(x_{j,k}) = H_k(\mathbf{x}_k), \tag{E.18}$$

so the Jensen gap in (E.16) is nonnegative.

We first obtain a quantitative use of this Jensen decrease. Since h_k is convex and has gradient Lipschitz modulus κ_h/μ_k, it satisfies

$$h_k(u) - h_k(v) - \langle \nabla h_k(v), u - v\rangle \ge \frac{\mu_k}{2\kappa_h}\|\nabla h_k(u) - \nabla h_k(v)\|^2,$$

applying this inequality with $u = x_{j,k}$ and $v = y_{i,k}$, weighting by W_{ij}, and using $\sum_j W_{ij}(x_{j,k} - y_{i,k}) = 0$ gives

$$H_k(\mathbf{x}_k) - \frac{1}{n}\sum_{i=1}^{n} h_k(y_{i,k}) \ge \frac{\mu_k}{2\kappa_h n}\sum_{i,j=1}^{n} W_{ij}\|\nabla h_k(x_{j,k}) - \nabla h_k(y_{i,k})\|^2.$$

Let $\omega_W := \min_i W_{ii} > 0$, which follows from Assumption C.1. Retaining the $j = i$ terms yields

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla h_k(y_{i,k}) - \nabla h_k(x_{i,k})\|^2 \le \frac{2\kappa_h}{\omega_W \mu_k}\left(H_k(\mathbf{x}_k) - \frac{1}{n}\sum_{i=1}^{n} h_k(y_{i,k})\right). \tag{E.19}$$

Set $z_{i,k} := y_{i,k} - \alpha_k v_{i,k}$. The ambient descent inequality for h_k and (E.18) give

$$\begin{aligned}\frac{1}{n}\sum_i h_k(z_{i,k}) - H_k(\mathbf{x}_k) \le\ & -\left(H_k(\mathbf{x}_k) - \frac{1}{n}\sum_i h_k(y_{i,k})\right)\\ & - \frac{\alpha_k}{n}\sum_i \langle \nabla h_k(y_{i,k}), v_{i,k}\rangle\\ & + \frac{\kappa_h \alpha_k^2}{2\mu_k n}\|\mathbf{v}_k\|^2.\end{aligned} \tag{E.20}$$

Since $v_{i,k} \in T_{x_{i,k}}\mathcal{M}$,

$$\langle \nabla h_k(x_{i,k}), v_{i,k}\rangle = \langle r_{i,k}, v_{i,k}\rangle.$$

By (E.19), Cauchy–Schwarz, and Young's inequality,

$$\begin{aligned}&-\frac{\alpha_k}{n}\sum_i \langle \nabla h_k(y_{i,k}) - \nabla h_k(x_{i,k}), v_{i,k}\rangle\\ &\quad\le \frac{1}{4}\left(H_k(\mathbf{x}_k) - \frac{1}{n}\sum_i h_k(y_{i,k})\right) + \frac{C\kappa_h\alpha_k^2}{\mu_k n}\|\mathbf{v}_k\|^2.\end{aligned}$$

Consequently,

$$\begin{aligned}\frac{1}{n}\sum_i h_k(z_{i,k}) - H_k(\mathbf{x}_k) \le\ & -\frac{3}{4}\left(H_k(\mathbf{x}_k) - \frac{1}{n}\sum_i h_k(y_{i,k})\right)\\ & - \frac{\alpha_k}{n}\sum_i \langle r_{i,k}, v_{i,k}\rangle\\ & + \frac{C\kappa_h\alpha_k^2}{\mu_k n}\|\mathbf{v}_k\|^2.\end{aligned} \tag{E.21}$$

We next pass from $z_{i,k}$ to $x_{i,k+1} = \mathcal{P}_{\mathcal{M}}(z_{i,k})$. Lemma E.4 supplies $\mathbf{x}_k, \mathbf{x}_{k+1} \in \mathcal{N}_\delta$, $\max_i \|v_{i,k}\| \le \sqrt{n}B_s$, and $\alpha_k\sqrt{n}B_s < \delta$. Thus every hypothesis of Lemma E.6 is satisfied. In particular, (E.13) holds with a constant uniform in k, μ_k, and K. Since $\|\nabla h_k\| \le G_h$, h_k is G_h-Lipschitz in the ambient space. Hence

$$H_k(\mathbf{x}_{k+1}) - \frac{1}{n}\sum_i h_k(z_{i,k}) \le \frac{C}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{C\alpha_k^2}{n}\|\mathbf{v}_k\|^2.$$

Combining this estimate with (E.21) gives

$$H_k(\mathbf{x}_{k+1}) - H_k(\mathbf{x}_k) \le -\frac{3}{4}\left(H_k(\mathbf{x}_k) - \frac{1}{n}\sum_i h_k(y_{i,k})\right) - \frac{\alpha_k}{n}\sum_i \langle r_{i,k}, v_{i,k}\rangle + \frac{C\kappa_h \alpha_k^2}{\mu_k n}\|\mathbf{v}_k\|^2 + \frac{C}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{C\alpha_k^2}{n}\|\mathbf{v}_k\|^2. \tag{E.22}$$

It remains to combine the local smooth losses with (E.22). Let

$$g_k^c := \mathrm{grad}\bar{f}(\bar{x}_k) = \frac{1}{n}\sum_{i=1}^n \mathrm{grad} f_i(\bar{x}_k).$$

By the Riemannian quadratic upper bound for $\bar{f}$,

$$\bar{f}(\bar{x}_{k+1}) - \bar{f}(\bar{x}_k) \le \langle g_k^c, \bar{x}_{k+1} - \bar{x}_k\rangle + \frac{L_g}{2}\|\bar{x}_{k+1} - \bar{x}_k\|^2.$$

The uniform mean expansion (E.14), together with the boundedness of g_k^c on $\mathcal{M}$, implies

$$\langle g_k^c, \bar{x}_{k+1} - \bar{x}_k\rangle \le -\frac{\alpha_k}{n}\sum_i \langle g_k^c, v_{i,k}\rangle + \frac{C}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{C}{n}\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2 + \frac{C\alpha_k^2}{n}\|\mathbf{v}_k\|^2.$$

The squared-step estimate (E.15) gives directly

$$\|\bar{x}_{k+1} - \bar{x}_k\|^2 \le \frac{C}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{C}{n}\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2 + \frac{C\alpha_k^2}{n}\|\mathbf{v}_k\|^2.$$

Consequently,

$$\bar{f}(\bar{x}_{k+1}) - \bar{f}(\bar{x}_k) \le -\frac{\alpha_k}{n}\sum_i \langle g_k^c, v_{i,k}\rangle + \frac{C}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{C}{n}\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2 + \frac{C\alpha_k^2}{n}\|\mathbf{v}_k\|^2. \tag{E.23}$$

Every constant in (E.23) is independent of μ_k.

For $p_{i,k} := \mathcal{P}_{T_{x_{i,k}}\mathcal{M}}(g_k^c) + r_{i,k}$, the tangency of $v_{i,k}$ implies

$$\langle g_k^c + r_{i,k}, v_{i,k}\rangle = \langle p_{i,k}, v_{i,k}\rangle.$$

Moreover, by the Lipschitz continuity of the local smooth gradients,

$$\begin{aligned}\frac{1}{n}\sum_i \|p_{i,k} - v_{i,k}\|^2 &= \frac{1}{n}\sum_i \|\mathcal{P}_{T_{x_{i,k}}\mathcal{M}}(g_k^c - s_{i,k})\|^2 \\ &\le 2\|g_k^c - \hat{g}_k\|^2 + \frac{2}{n}\|\mathbf{s}_k - \hat{\mathbf{G}}_k\|^2 \\ &\le \frac{C}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{C}{n}\|\mathbf{s}_k - \hat{\mathbf{G}}_k\|^2.\end{aligned}$$

Therefore,

$$-\frac{\alpha_k}{n}\sum_i \langle p_{i,k}, v_{i,k}\rangle \le -\frac{3\alpha_k}{4n}\|\mathbf{v}_k\|^2 + \frac{C\alpha_k}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{C\alpha_k}{n}\|\mathbf{s}_k - \hat{\mathbf{G}}_k\|^2. \tag{E.24}$$

Adding (E.22) and (E.23), then applying (E.24), shows that

$$\begin{aligned}\Psi_{\mu_k}(\mathbf{x}_{k+1}) - \Psi_{\mu_k}(\mathbf{x}_k) \le & -\frac{3}{4}\left(H_k(\mathbf{x}_k) - \frac{1}{n}\sum_i h_k(y_{i,k})\right) \\ & -\frac{\alpha_k}{n}\left(\frac{3}{4} - C\alpha_k\left(1 + \frac{\kappa_h}{\mu_k}\right)\right)\|\mathbf{v}_k\|^2 \\ & +\frac{C}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 \\ & +\frac{C}{n}\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2 \\ & +\frac{C\alpha_k}{n}\|\mathbf{s}_k - \hat{\mathbf{G}}_k\|^2.\end{aligned}$$

Reducing c_0 if necessary gives a negative coefficient at most $-\alpha_k\|\mathbf{v}_k\|^2/(2n)$.

Finally, conservation of the tracker average gives

$$\hat{v}_k - \tilde{q}_k = \frac{1}{n}\sum_i \left(\mathcal{P}_{T_{x_{i,k}}\mathcal{M}} - \mathcal{P}_{T_{\bar{x}_k}\mathcal{M}}\right)\left(s_{i,k} - \mathrm{grad} f_i(x_{i,k})\right),$$

and consequently

$$\|\hat{v}_k - \tilde{q}_k\|^2 \le \frac{C}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2.$$

Since $\|\hat{v}_k\|^2 \le \|\mathbf{v}_k\|^2/n$, it follows that

$$\|\tilde{q}_k\|^2 \le \frac{2}{n}\|\mathbf{v}_k\|^2 + \frac{C}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2.$$

Splitting the negative direction-energy term into two equal parts and using this estimate gives constants $c_q, c_v > 0$ such that the current-μ_k version of (E.16) holds. The change from μ_k to μ_{k+1} costs at most $c_h|\mu_{k+1} - \mu_k|$ in H_μ by Assumption D.2. This proves (E.16).

The conversion estimate (E.17) follows from the Lipschitz continuity of $\mathrm{grad} f_i$ and

$$\|\mathrm{grad} h_{\mu_k}(x) - \mathrm{grad} h_{\mu_k}(y)\| \le \left(G_h L_{\mathcal{P}} + \frac{\kappa_h}{\mu_k}\right)\|x - y\|.$$

It is not used in the descent of Ψ_μ and is needed only when the local residual is translated to the mean-point residual. □

With these preparations, we give the proof of Theorem E.1.

Proof of Theorem E.1. We first record the cancellation relating the average algorithmic direction to the local smoothed residual. The double stochasticity of W and the initialization of $s_{i,0}$ imply

$$\frac{1}{n}\sum_{i=1}^n s_{i,k} = \frac{1}{n}\sum_{i=1}^n \mathrm{grad} f_i(x_{i,k}) = \hat{g}_k, \qquad k \ge 0. \tag{E.25}$$

Since $\sum_i \left(s_{i,k} - \mathrm{grad} f_i(x_{i,k})\right) = 0$, with $\hat{v}_k := n^{-1}\sum_i v_{i,k}$ we have

$$\begin{aligned}\hat{v}_k - \tilde{q}_k &= \frac{1}{n}\sum_{i=1}^n \mathcal{P}_{T_{x_{i,k}}\mathcal{M}}\left(s_{i,k} - \mathrm{grad} f_i(x_{i,k})\right) \\ &= \frac{1}{n}\sum_{i=1}^n \left(\mathcal{P}_{T_{x_{i,k}}\mathcal{M}} - \mathcal{P}_{T_{\bar{x}_k}\mathcal{M}}\right)\left(s_{i,k} - \mathrm{grad} f_i(x_{i,k})\right), \qquad \text{(E.26)} \\ \|\hat{v}_k - \tilde{q}_k\|^2 &\le \frac{C}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2.\end{aligned}$$

Indeed, Lemma E.4 and Assumption D.1 give the uniform bound $\|s_{i,k} - \mathrm{grad} f_i(x_{i,k})\| \le \sqrt{nD_s} + L$; the final inequality follows from the Lipschitz continuity of the tangent projection. The common smoothed regularizer cancels exactly in (E.26).

We next establish weighted cumulative estimates for the variable-step recursions. The following elementary convolution fact will be used twice. Let $a_k = a_0/(k+1)^p$, let $0 \le \gamma < 1$, and suppose

$$z_{k+1} \le \gamma z_k + a_k u_k, \qquad z_0 = 0.$$

Then there exists $C_{p,\gamma} < \infty$ such that, for every K,

$$\sum_{k=1}^{K} \frac{z_k^2}{a_{k-1}} \le C_{p,\gamma} \sum_{k=0}^{K-1} a_k u_k^2. \tag{E.27}$$

Indeed, after setting $\zeta_k = z_k/\sqrt{a_{k-1}}$ and $v_j = \sqrt{a_j}u_j$, unrolling the recursion gives

$$\zeta_k \le \sum_{j=0}^{k-1} \gamma^{k-1-j} \sqrt{\frac{a_j}{a_{k-1}}}\, v_j.$$

For $m = k-1-j$, $a_j/a_{k-1} = (k/(k-m))^p \le (m+1)^p$. Thus the convolution kernel is bounded by $\{\gamma^m (m+1)^{p/2}\}_{m\ge 0}$, which is in ℓ_1. Young's convolution inequality proves (E.27).

Apply this fact to (E.9) with $a_k = \alpha_k$, $z_k = \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|$, and $u_k = \ell_\delta \|\mathbf{v}_k\|$. We obtain

$$\frac{1}{n}\sum_{k=1}^{K} \frac{\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2}{\alpha_{k-1}} \le \frac{C}{n} \sum_{k=0}^{K-1} \alpha_k \|\mathbf{v}_k\|^2. \tag{E.28}$$

In particular, since $\alpha_{k-1} \le \alpha_0$,

$$\frac{1}{n}\sum_{k=0}^{K} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 \le \frac{C\alpha_0}{n} \sum_{k=0}^{K-1} \alpha_k \|\mathbf{v}_k\|^2. \tag{E.29}$$

Set $T_k := \|\mathbf{s}_k - \hat{\mathbf{G}}_k\|$. Equations (E.5) and (E.4) give

$$T_{k+1} \le \sigma_2 T_k + 4L\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + 2L\alpha_k \|\mathbf{v}_k\|.$$

The usual unweighted convolution inequality, followed by (E.29), yields

$$\frac{1}{n}\sum_{k=0}^{K-1} T_k^2 \le C + \frac{C\alpha_0}{n} \sum_{k=0}^{K-1} \alpha_k \|\mathbf{v}_k\|^2. \tag{E.30}$$

Here the constant includes T_0^2/n and is independent of K.

The coupled schedules satisfy

$$\alpha_k \left(1 + \frac{\kappa_h}{\mu_k}\right) = \frac{\alpha_0}{(k+1)^p} + \frac{\kappa_h \alpha_0}{\mu_0} \le \alpha_0 \left(1 + \frac{\kappa_h}{\mu_0}\right). \tag{E.31}$$

After reducing c_0 if necessary, condition (E.2) therefore implies the stepsize requirement in Lemma E.7 at every iteration. Furthermore, compactness of $\mathcal{M}$ and Assumption D.2 give the explicit uniform lower bound

$$\Psi_{\mu_k}(\mathbf{x}_k) \ge \min_{x\in\mathcal{M}} \bar{f}(x) + \min_{x\in\mathcal{M}} h(x) - c_h \mu_0, \qquad 0 \le k \le K.$$

Sum the descent inequality over $k = 0, \dots, K-1$. Dropping its nonnegative Jensen-gap sum and applying (E.28)–(E.30) gives

$$\begin{aligned} &c_q \sum_{k=0}^{K-1} \alpha_k \|\tilde{q}_k\|^2 + \frac{c_v}{n} \sum_{k=0}^{K-1} \alpha_k \|\mathbf{v}_k\|^2 \\ &\le C + c_h \Delta_{\mu,K} + \frac{C\alpha_0}{n} \sum_{k=0}^{K-1} \alpha_k \|\mathbf{v}_k\|^2 + C\alpha_0 + \frac{C\alpha_0^2}{n} \sum_{k=0}^{K-1} \alpha_k \|\mathbf{v}_k\|^2. \end{aligned} \tag{E.32}$$

After reducing the constant in the bound $\alpha_0 < c_1$ in (E.2), the terms involving $\|\mathbf{v}_k\|^2$ on the right-hand side are absorbed into the left-hand side. Hence

$$\sum_{k=0}^{K-1} \alpha_k \|\tilde{q}_k\|^2 + \frac{1}{n} \sum_{k=0}^{K-1} \alpha_k \|\mathbf{v}_k\|^2 \le C\left(1 + c_h \Delta_{\mu,K}\right). \tag{E.33}$$

Since $|\mathcal{I}_K| \geq K/3$ for $K \geq 2$, the same estimate over the last half of the run and $\alpha_k \geq \alpha_{K-1}$ for $k \in \mathcal{I}_K$ give

$$\mathbb{E}\|\tilde{q}_{R_K}\|^2 = \frac{1}{|\mathcal{I}_K|}\sum_{k\in\mathcal{I}_K}\|\tilde{q}_k\|^2 \leq \mathcal{O}\left(\frac{1+c_h\Delta_{\mu,K}}{|\mathcal{I}_K|\alpha_{K-1}}\right).$$

Finally, set $M_k := L + G_h L_{\mathcal{P}} + \kappa_h/\mu_k$. Because

$$\alpha_k M_k \leq \alpha_0\left(L + G_h L_{\mathcal{P}} + \frac{\kappa_h}{\mu_0}\right) =: \Gamma, \tag{E.34}$$

(E.17) implies

$$\begin{aligned}\sum_{k=0}^{K-1}\alpha_k\|\hat{q}_k\|^2 &\leq 2\sum_{k=0}^{K-1}\alpha_k\|\tilde{q}_k\|^2 + \frac{2\Gamma^2}{n}\sum_{k=1}^{K-1}\frac{\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^2}{\alpha_k}\\ &\leq 2\sum_{k=0}^{K-1}\alpha_k\|\tilde{q}_k\|^2 + \frac{2^{p+1}\Gamma^2}{n}\sum_{k=1}^{K-1}\frac{\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^2}{\alpha_{k-1}}\\ &\leq C(1+c_h\Delta_{\mu,K}).\end{aligned}$$

Here the $k=0$ consensus error is zero, the second inequality uses $\alpha_{k-1}/\alpha_k \leq 2^p$, and the last uses (E.28) and (E.33). Since $\|\hat{q}_k\| = \|\mathrm{grad}F_{\mu_k}(\bar{x}_k)\|$ and R_K is uniform on $\mathcal{I}_K$, averaging only over this set proves

$$\mathbb{E}\|\mathrm{grad}F_{\mu_{R_K}}(\bar{x}_{R_K})\|^2 \leq \mathcal{O}\left(\frac{1+c_h\Delta_{\mu,K}}{|\mathcal{I}_K|\alpha_{K-1}}\right).$$

The same average guarantees the existence of an index $k_K \in \mathcal{I}_K$ satisfying this stationarity bound. Since $\{\mu_k\}$ is nonincreasing, both μ_{R_K} and μ_{k_K} are at most $\mu_{\lceil K/2\rceil}$. Finally, the pointwise consensus bound in Lemma E.5 yields, for every $k \in \mathcal{I}_K$,

$$\frac{1}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 \leq D_{c,p}B_s^2\alpha^2_{\lceil K/2\rceil-1}.$$

For general $p \in (0,1)$, $\Delta_{\mu,K} \leq \mu_0$ and

$$\frac{1+c_h\Delta_{\mu,K}}{|\mathcal{I}_K|\alpha_{K-1}} = \mathcal{O}(K^{p-1}), \qquad \alpha^2_{\lceil K/2\rceil-1} = \mathcal{O}(K^{-2p}).$$

Balancing K^{p-1} and K^{-2p} gives the unique exponent $p = 1/3$. Thus the randomized output satisfies

$$\mathbb{E}\left[\|\mathrm{grad}F_{\mu_{R_K}}(\bar{x}_{R_K})\|^2 + \frac{1}{n}\|\mathbf{x}_{R_K} - \bar{\mathbf{x}}_{R_K}\|^2\right] = \mathcal{O}((K+1)^{-2/3}),$$
$$\mu_{R_K} = \mathcal{O}((K+1)^{-1/3}).$$

The deterministic existence index satisfies the corresponding non-expected bounds. Applying Proposition D.3 to any sequence of those deterministic certificates establishes the stated vanishing-smoothing conclusion.

□

F Numerical experiments

F.1 Benchmark experiments

The reported comparisons contain 264 runs: 168 MNIST topology, 48 partial/full tracking, and 48 CIFAR-10 runs. Method-specific selectors and the DSPGT empirical schedule were fixed before evaluation. All reported results, tables, and trajectories use only seeds 110–112. Each selector remains fixed within its split and regularization condition, with the physical step rescaled only by the corresponding $\widehat{L}$. Ten baseline runs do not produce terminal metrics: THANOS has SVD nonconvergence on all three seeds of the 32-agent MNIST expander IID/medium setting and on one

corresponding CIFAR run; DR-ProxGT fails the common tangent-proximal KKT tolerance of 10^{-6} on all six non-IID CIFAR runs.

The common target is the first recorded iterate at which the residual is at most one quarter of its initial value and RMS consensus is at most 10^{-4}. Runs that do not reach the target use the full $K = 2500$ budget in target-work summaries. The raw records include graph and partition metadata, physical steps after $\widehat{L}$ normalization, trajectories, payloads, work counters, and test-utility fields. Tables 4 and 2 report setting-level medians over runs that return terminal metrics under the stated solver criteria. The "Completed" column gives this count among the three scheduled evaluation seeds. Table 3 reports terminal residual, consensus, and tracking-error medians and ranges for the partial and full variants over matched instances. Every scheduled run remains in the target-attainment denominator. The supplement provides seed-level trajectories, solver status, work counters, and software information.

Table 4: MNIST setting-level medians over completed runs. Target work is the payload at first joint-target attainment, or the full-budget payload otherwise.

n	Graph	Split	Method	Completed	Residual	Consensus	Payload	Target work
8	expander	non-IID/S	DSPGT	3/3	0.0119	4.36e-10	1.57e+08	7.97e+07
8	expander	non-IID/S	THANOS	3/3	0.56	0.0018	1.57e+08	1.57e+08
8	expander	non-IID/S	DRSM	3/3	1.83	1.34e-06	7.84e+07	7.84e+07
8	expander	non-IID/S	DR-ProxGT	3/3	0.145	2.71e-09	1.57e+08	1.88e+07
8	expander	IID/M	DSPGT	3/3	0.00944	8.38e-05	1.57e+08	7.9e+07
8	expander	IID/M	THANOS	3/3	0.0611	4.44e-05	1.57e+08	6.27e+06
8	expander	IID/M	DRSM	3/3	0.0406	0.00825	7.84e+07	7.84e+07
8	expander	IID/M	DR-ProxGT	3/3	0.0141	0.127	1.57e+08	1.57e+08
8	ring	non-IID/S	DSPGT	3/3	0.0119	6.44e-09	7.84e+07	3.98e+07
8	ring	non-IID/S	THANOS	3/3	0.536	0.00631	7.84e+07	7.84e+07
8	ring	non-IID/S	DRSM	3/3	1.83	2.93e-06	3.92e+07	3.92e+07
8	ring	non-IID/S	DR-ProxGT	3/3	0.145	1.37e-08	7.84e+07	1.88e+07
8	ring	IID/M	DSPGT	3/3	0.00858	9.06e-05	7.84e+07	3.95e+07
8	ring	IID/M	THANOS	3/3	0.0536	0.00222	7.84e+07	7.84e+07
8	ring	IID/M	DRSM	3/3	0.0281	0.0155	3.92e+07	3.92e+07
8	ring	IID/M	DR-ProxGT	3/3	0.0132	0.191	7.84e+07	7.84e+07
16	expander	non-IID/S	DSPGT	3/3	0.0122	2.93e-09	1.57e+08	7.97e+07
16	expander	non-IID/S	THANOS	3/3	0.547	0.00305	1.57e+08	1.57e+08
16	expander	non-IID/S	DRSM	3/3	1.64	1.46e-06	7.84e+07	7.84e+07
16	expander	non-IID/S	DR-ProxGT	3/3	0.145	5.01e-08	1.57e+08	3.14e+07
16	expander	IID/M	DSPGT	3/3	0.00827	9.22e-05	1.57e+08	7.9e+07
16	expander	IID/M	THANOS	3/3	0.0607	0.000201	1.57e+08	1.57e+08
16	expander	IID/M	DRSM	3/3	0.0332	0.0159	7.84e+07	7.84e+07
16	expander	IID/M	DR-ProxGT	3/3	0.014	0.213	1.57e+08	1.57e+08
16	grid	non-IID/S	DSPGT	3/3	0.0121	1.07e-08	7.84e+07	3.98e+07
16	grid	non-IID/S	THANOS	3/3	0.518	0.00596	7.84e+07	7.84e+07
16	grid	non-IID/S	DRSM	3/3	1.64	2.82e-06	3.92e+07	3.92e+07
16	grid	non-IID/S	DR-ProxGT	3/3	0.145	1.62e-07	7.84e+07	2.6e+07
16	grid	IID/M	DSPGT	3/3	0.00953	9.91e-05	7.84e+07	3.95e+07
16	grid	IID/M	THANOS	3/3	0.0581	0.0105	7.84e+07	7.84e+07
16	grid	IID/M	DRSM	3/3	0.0283	0.0231	3.92e+07	3.92e+07
16	grid	IID/M	DR-ProxGT	3/3	0.0264	0.276	7.84e+07	7.84e+07
16	ring	non-IID/S	DSPGT	3/3	0.012	5.64e-08	7.84e+07	4.01e+07
16	ring	non-IID/S	THANOS	3/3	0.522	0.00572	7.84e+07	7.84e+07
16	ring	non-IID/S	DRSM	3/3	1.64	7.13e-06	3.92e+07	3.92e+07
16	ring	non-IID/S	DR-ProxGT	3/3	0.145	5.26e-07	7.84e+07	3.7e+07
16	ring	IID/M	DSPGT	3/3	0.00839	8.93e-05	7.84e+07	3.95e+07
16	ring	IID/M	THANOS	3/3	0.0512	0.00269	7.84e+07	7.84e+07
16	ring	IID/M	DRSM	3/3	0.0281	0.0393	3.92e+07	3.92e+07
16	ring	IID/M	DR-ProxGT	3/3	0.0131	0.173	7.84e+07	7.84e+07
32	expander	non-IID/S	DSPGT	3/3	0.0121	5.57e-09	1.57e+08	7.97e+07
32	expander	non-IID/S	THANOS	3/3	0.502	0.00709	1.57e+08	1.57e+08
32	expander	non-IID/S	DRSM	3/3	1.65	1.62e-06	7.84e+07	7.84e+07
32	expander	non-IID/S	DR-ProxGT	3/3	0.145	9.9e-08	1.57e+08	3.95e+07
32	expander	IID/M	DSPGT	3/3	0.00905	0.000102	1.57e+08	7.9e+07
32	expander	IID/M	THANOS	0/3	–	–	–	–
32	expander	IID/M	DRSM	3/3	0.0294	0.024	7.84e+07	7.84e+07
32	expander	IID/M	DR-ProxGT	3/3	0.0427	0.306	1.57e+08	1.57e+08
32	ring	non-IID/S	DSPGT	3/3	0.0388	4.47e-05	7.84e+07	5.71e+07
32	ring	non-IID/S	THANOS	3/3	0.512	0.0195	7.84e+07	7.84e+07
32	ring	non-IID/S	DRSM	3/3	1.66	2.56e-05	3.92e+07	3.92e+07
32	ring	non-IID/S	DR-ProxGT	3/3	1.15	1.27	7.84e+07	7.84e+07
32	ring	IID/M	DSPGT	3/3	0.00824	8.58e-05	7.84e+07	6.4e+07
32	ring	IID/M	THANOS	3/3	0.0471	0.0127	7.84e+07	7.84e+07
32	ring	IID/M	DRSM	3/3	0.0279	0.0786	3.92e+07	3.92e+07
32	ring	IID/M	DR-ProxGT	3/3	0.0131	0.101	7.84e+07	7.84e+07

The tracking comparison uses 24 seed-matched partial/full pairs across 32-agent ring and expander networks and label-Dirichlet concentrations 0.1 and 0.03. Each of the four groups contains six pairs, combining three seeds and two fixed polynomial schedules. Table 3 reports terminal metrics separately for the partial and full variants as medians and minimum–maximum ranges in each group

and over all 24 pairs. Both versions have an overall median residual of 2.866, RMS consensus error of 3.56×10^{-5}, and tracking error of 0.0100 at the reported precision.

F.2 Implementation and trajectory details

MNIST images are scaled to $[0, 1]$ and centered by their training mean. The IID allocation gives every agent equal capacity; the non-IID allocation draws label proportions from a Dirichlet distribution with concentration 0.1 and then enforces the same capacity. Each agent precomputes $C_i = A_i^\top A_i / m_i$, so one optimization round uses $-C_i X_i$ without rescanning the images. MNIST uses $\lambda = 0.003$ for medium regularization and $\lambda = 0.03$ for strong regularization. The row penalty is invariant under $X \mapsto XQ$ for orthogonal Q, and its proximal mapping acts rowwise:

$$[\mathrm{prox}_{\mu\lambda\|\cdot\|_{2,1}}(X)]_{j,:} = \left(1 - \frac{\mu\lambda}{\|X_{j,:}\|_2}\right)_+ X_{j,:}.$$

The corresponding Moreau gradient is $\mu^{-1}(X - \mathrm{prox}_{\mu\lambda\|\cdot\|_{2,1}}(X))$.

For CIFAR-10, a pretrained ImageNet ResNet-18 in evaluation mode maps each image to its 512-dimensional global-average-pooled feature. We use the standard ImageNet resize, crop, and normalization associated with the pretrained weights. Feature extraction is performed once before optimization. The resulting subspaces are evaluated by explained variance, reconstruction error, and a ridge linear probe fitted on the training features. We use $\lambda = 0.3$ for both CIFAR split conditions.

Figure 2 complements the aggregate curves in the main text with four settings that cross the two datasets with nominal and strongly heterogeneous conditions on ring and expander graphs. Lines are seed medians and bands span the three seeds. The CIFAR ring/non-IID panels omit DR-ProxGT because its three runs do not meet the common tangent-proximal KKT tolerance.

Reproducibility. All methods use the same data partition, initialization, graph, diagnostic checkpoints, and 2500 communication rounds within an instance. DSPGT and the three baselines are evaluated with the common residual and consensus metrics defined in the main text. The supplementary records contain every seed-level trajectory, terminal metric, solver status, communication payload, and local work counter.

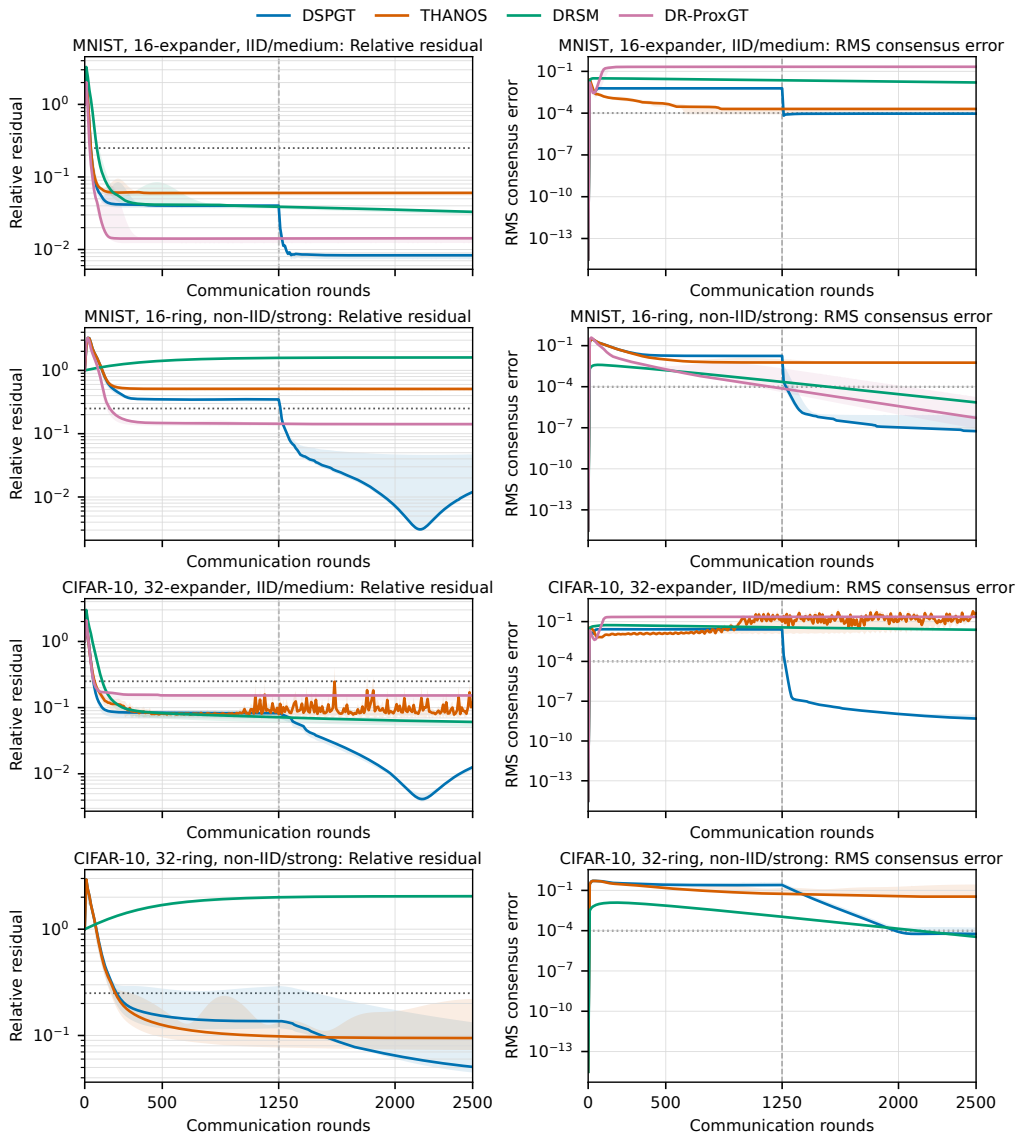

Figure 2: Seed-level convergence ranges for four representative settings. Residuals are normalized by their initial values; horizontal dotted lines mark the joint-target thresholds and the vertical dashed line marks DSPGT's fixed stepsize switch at iteration 1250. The panels cover both datasets, ring and expander connectivity, and nominal and strongly heterogeneous conditions.