# PLADOS: A PARAMETER-FREE LANDING ALGORITHM FOR DECENTRALIZED OPTIMIZATION ON THE STIEFEL MANIFOLD

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Decentralized optimization on the Stiefel manifold has broad applications in machine learning and signal processing. We propose PLADOS, a parameter-free landing algorithm for decentralized optimization on the Stiefel manifold. Each iteration combines local stochastic gradient updates with a single retraction-free communication step over the communication graph. A key convex-like property of the intersection of the nonconvex Stiefel manifold and consensus constraints is characterized through a restricted secant inequality for the penalty-only scheme, ensuring local contraction of the manifold consensus error without introducing additional parameters that require tuning. For $n$ nodes using independent local batches of size $b$, we prove that PLADOS achieves linear speedup with respect to the number of nodes with an asymptotic convergence rate of $O(\sigma/\sqrt{nbK} + (n\xi^2)^{1/3}/K^{2/3})$ after $K$ iterations, where $\sigma^2 > 0$ is the sampling variance and $\xi^2$ quantifies gradient heterogeneity. Experiments demonstrate stability across tested stepsizes and networks, time efficiency, and linear speedup with respect to the number of nodes.

## 1 INTRODUCTION

Decentralized optimization enables a network of $n$ nodes to learn from local data using only neighbor communication (Tsitsiklis et al., 1986; Nedic & Ozdaglar, 2009; Duchi et al., 2011). Distributing both data and computation makes this framework useful for large-scale learning and signal processing. Many such applications also require orthonormal decision matrices, including principal component analysis (Chen et al., 2021), low-rank estimation (Deng & Hu, 2025a), orthogonal representation learning (Huang et al., 2018), and low-rank adaptation (Zhang et al., 2026). These problems motivate decentralized stochastic optimization on the Stiefel manifold (Boumal, 2023, Section 7.3):

$$\min_{X\in \mathrm{St}(p,r)} f(X) := \frac{1}{n}\sum_{i=1}^{n} f_i(X), \qquad \mathrm{St}(p,r) := \{X \in \mathbb{R}^{p\times r} : X^\top X = I_r\}, \tag{1}$$

where node $i$ accesses its local objective through a stochastic oracle and communicates with its neighbors over a connected graph.

Solving equation (1) requires the nodes to reach consensus while satisfying a nonconvex orthogonality constraint. Even when all local matrices have orthonormal columns, their weighted average can leave the Stiefel manifold. Retractions restore feasibility after each update, but the required QR or polar factorizations add computation at every node (Chen et al., 2021; Deng & Hu, 2025a). Retraction-free landing methods offer a cheaper way to reduce constraint violation (Ablin & Peyré, 2022; Zhang et al., 2026). In a decentralized algorithm, however, landing must work jointly with communication and optimization.

Stochastic gradients and heterogeneous data make this interaction more delicate. Independent local samples reduce the variance of the averaged gradient, while differences between local objectives drive the nodes apart. On the Stiefel manifold, disagreement also affects feasibility and the descent of the projected network average. A convergence analysis must therefore control these errors together while preserving the variance reduction from independent local batches. This is essential for obtaining linear speedup with respect to the number of nodes, a central benefit of decentralized stochastic optimization (Lian et al., 2017; Tang et al., 2018).



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.

Table 1: Comparison with representative stochastic decentralized and retraction-free methods.

| Method | Decentralized | Stiefel manifold constraints | Retraction-free | Parameter-free | Heterogeneity assumption | Linear node speedup |
|---|---|---|---|---|---|---|
| EF-Landing (Song et al., 2025) | × | ✓ | ✓ | × | × | ✓ |
| Man-LoRA (Zhang et al., 2026) | × | ✓ | ✓ | ✓ | × | × |
| DRSGD (Chen et al., 2021) | ✓ | ✓ | × | × | × | ✓ |
| **PLADOS** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Note.* "Parameter-free" means that constraint-related stepsizes or parameters have explicit values and require no tuning; PLADOS uses $\gamma = 1/3$. "Heterogeneity assumption" records an explicit bound on cross-worker gradient heterogeneity, not whether heterogeneous data are allowed.

This raises the question: can a simple retraction-free decentralized algorithm control both consensus and orthogonality errors and achieve stochastic linear speedup with respect to the number of nodes? We answer this question with PLADOS, a parameter-free landing algorithm that combines a penalty-only scheme with one communication round and one local stochastic gradient step per iteration. To the best of our knowledge, PLADOS is the first decentralized stochastic landing method with proven convergence on the Stiefel manifold.

### 1.1 Related Work

**Decentralized optimization in Euclidean space.** Classical consensus and gradient methods laid the foundations (Tsitsiklis et al., 1986; Nedic & Ozdaglar, 2009; Duchi et al., 2011; Yuan et al., 2016). D-PSGD and subsequent analyses establish nonconvex stochastic linear speedup and characterize topology, compression, and heterogeneity effects (Lian et al., 2017; Koloskova et al., 2019; Tang et al., 2018; Alghunaim & Yuan, 2022; Takezawa et al., 2026). EXTRA, exact diffusion, NIDS, and gradient tracking use auxiliary information for exact convergence or reduced local drift (Shi et al., 2015; Yuan et al., 2018; Li et al., 2019; Nedić et al., 2017; Qu & Li, 2017; Li et al., 2022; Pu & Nedić, 2021; Koloskova et al., 2021; Xin et al., 2021; Liu et al., 2025). In D-PSGD, heterogeneity contributes a $K^{-2/3}$ term, while independent batches reduce stochastic error (Alghunaim & Yuan, 2022). By contrast, $D^2$ uses a history correction to remove dependence on global gradient heterogeneity (Tang et al., 2018). Extending these Euclidean techniques to nonconvex manifold constraints is nontrivial.

**Retraction-based decentralized optimization on manifolds.** Early work studied consensus intrinsically on manifolds (Sarlette & Sepulchre, 2009; Tron et al., 2012). On the Stiefel manifold, DRSGD combines stochastic Riemannian steps with retractions and multi-step consensus, while DRGTA adds gradient tracking (Chen et al., 2021). DPRGD, DPRGT, and DPRSRM extend projected Riemannian techniques to compact embedded submanifolds and stochastic recursive momentum (Deng & Hu, 2025a;b). Proximal tracking, Riemannian EXTRA, local-consensus, conjugate-gradient, natural-gradient, compressed, and quantized variants further broaden this class (Wang et al., 2025; Wu et al., 2025; Hu et al., 2025b; Chen et al., 2023; Hu et al., 2025a; Hu & Deng, 2024; Chen et al., 2025). Repeated factorizations and vector transport can limit scalability, especially on GPUs (Song et al., 2025; Zhang et al., 2026).

**Retraction-free optimization on the Stiefel manifold.** Centralized infeasible, landing, and constraint-dissolving methods avoid orthonormalization (Gao et al., 2019; Ablin & Peyré, 2022; Xiao et al., 2024; Zhang et al., 2026). Centrally coordinated EF-Landing combines stochastic landing with compressed communication (Song et al., 2025). Decentralized examples include the approximate augmented-Lagrangian method DESTINY, landing with gradient tracking in DRFGT, and the deterministic primal–dual RF-EXTRA correction (Wang & Liu, 2022; Sun et al., 2024; Li & Hu, 2026). Stochastic linear speedup with respect to the number of nodes without gradient tracking remains a key challenge for decentralized methods with independent mini-batches and heterogeneous objectives. Table 1 summarizes the main properties of representative methods and PLADOS. Penalty selection is another central issue in constrained optimization: Byrd & Nocedal (1998) asked at the International Congress of Mathematicians 1998 whether sensitivity to penalty parameters could limit large-scale exact penalty methods. Our PLADOS answers this question for the considered constrained problem by a landing correction that requires no problem-dependent penalty tuning.

### 1.2 CONTRIBUTIONS

Our contributions are threefold.

- **A parameter-free landing algorithm.** We propose PLADOS for decentralized optimization on the Stiefel manifold. Each iteration combines local stochastic gradient updates with a single retraction-free communication step. The landing coefficient is universally fixed at $\gamma = 1/3$, independently of unknown problem constants, so no additional feasibility parameter requires tuning.
- **Linear convergence of the penalty-only scheme.** The feasible set intersects the convex consensus subspace with the nonconvex product Stiefel manifold. We establish a restricted secant inequality (RSI) for the penalty-only scheme near this intersection, yielding local Q-linear convergence of the distance to the feasible set.
- **Linear speedup with respect to the number of nodes.** With independent local mini-batches, we prove linear speedup with respect to the number of nodes with heterogeneous local data, once the stochastic term dominates the heterogeneity and network transients. The asymptotic rate is $O(\sigma/\sqrt{nbK} + (n\xi^2)^{1/3}/K^{2/3})$ after $K$ iterations, where $\sigma^2 > 0$ is the sampling variance and $\xi^2$ quantifies gradient heterogeneity. This holds for a fixed network or a uniformly gapped network family with uniformly bounded problem and oracle constants.

## 2 PROBLEM FORMULATION AND PLADOS

We study decentralized optimization on the Stiefel manifold in equation (1). For a stack $\mathbf{X} = (X_1, \dots, X_n)$, define

$$\widehat{X} = \frac{1}{n}\sum_i X_i, \qquad \bar{X} = P_{\mathrm{St}}(\widehat{X}), \qquad \bar{\mathbf{X}} = \mathbf{1} \otimes \bar{X}, \qquad d(\mathbf{X}) = \left\|\mathbf{X} - \bar{\mathbf{X}}\right\|_F.$$

Thus $d(\mathbf{X})$ is the distance to the consensus Stiefel manifold $\mathcal{C} = \{\mathbf{1} \otimes Q : Q \in \mathrm{St}(p, r)\}$. The projection $P_{\mathrm{St}}(Y) = Y(Y^\top Y)^{-1/2}$, defined for full-column-rank $Y$, is the nearest-point polar projection (Higham, 1986). For any ambient matrix $U$, write $P_X U = U - X\,\mathrm{sym}(X^\top U)$, where $\mathrm{sym}(A) = (A + A^\top)/2$ for a square matrix $A$. At a feasible $X$, this is the tangent projection and $\mathrm{grad}\, f(X) = P_X \nabla f(X)$ (Boumal, 2023, Section 7.3).

### 2.1 ASSUMPTIONS

We state the assumptions used in the analysis. Let $\mathcal{F}_k$ denote the history immediately before the time-$k$ mini-batches are sampled. The initialization $\mathbf{X}_0$ is deterministic; its local tube condition is stated in Theorem 3.7.

**Assumption 2.1** (Objective regularity and stochastic oracle)**.** *Each $f_i$ has a $C^2$ extension to a neighborhood of a fixed compact tube around $\mathrm{St}(p, r)$. Before iteration $k$, node $i$ draws a fresh mini-batch $\mathcal{B}_{i,k} = \{\zeta_{i,k}^{(1)}, \dots, \zeta_{i,k}^{(b)}\}$ and forms*

$$S_{i,k} = P_{X_{i,k}}\left(\frac{1}{b}\sum_{\ell=1}^{b} \nabla_X F_i\left(X_{i,k}; \zeta_{i,k}^{(\ell)}\right)\right), \tag{2}$$

*where $F_i(\cdot; \zeta)$ is a local sample loss. Conditional on $\mathcal{F}_k$, the nodewise estimators are independent across nodes and satisfy*

$$\mathbb{E}\left[\frac{1}{b}\sum_{\ell=1}^{b} \nabla_X F_i(X_{i,k}; \zeta_{i,k}^{(\ell)}) \,\middle|\, \mathcal{F}_k\right] = \nabla f_i(X_{i,k}). \tag{3}$$

*The batch gradient satisfies $\|b^{-1}\sum_{\ell=1}^{b} \nabla_X F_i(X_{i,k}; \zeta_{i,k}^{(\ell)})\|_F \le G$ almost surely in the tube. Define $G_{i,k} := P_{X_{i,k}} \nabla f_i(X_{i,k})$. Linearity of $P_{X_{i,k}}$ gives the first equality below, and we assume the projected variance bound in the second:*

$$\mathbb{E}[S_{i,k} \mid \mathcal{F}_k] = G_{i,k}, \qquad \mathbb{E}[\|S_{i,k} - G_{i,k}\|_F^2 \mid \mathcal{F}_k] \le \frac{\sigma^2}{b}. \tag{4}$$

The oracle conditions separate the mean gradient from sampling noise. Independent local batches reduce the variance of the averaged noise to $\sigma^2/(nb)$. The almost-sure bound controls tube invariance and higher-order remainders; smooth finite-sum losses satisfy it on a compact neighborhood.

**Assumption 2.2** (Data heterogeneity)**.** *For every feasible $X$,*

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(X)-\nabla f(X)\|_F^2 \le \xi^2. \tag{5}$$

The quantity $\xi^2$ measures variation among local gradients evaluated at the same feasible matrix. Smoothness separately controls the effect of evaluating those gradients at different local iterates.

**Assumption 2.3** (Communication network)**.** *The nodes communicate through a symmetric doubly stochastic matrix $W$ whose graph is connected. With $J = \mathbf{1}\mathbf{1}^\top/n$, we have $\sigma_2(W) := \|W - J\|_2 < 1$; write $\mathbf{W} = W \otimes I_p$ and $\mathbf{J} = J \otimes I_p$.*

The spectral gap $1 - \sigma_2(W)$ determines how effectively communication reduces disagreement between nodes. These are standard regularity, oracle, heterogeneity, and connectivity assumptions in decentralized stochastic optimization (Chen et al., 2021; Tang et al., 2018).

### 2.2 The PLADOS algorithm

For a coefficient $\gamma \in (0, 1)$, define the landing map

$$\phi_\gamma(X) = X - \gamma X(X^\top X - I_r). \tag{6}$$

Throughout the paper we fix $\gamma = 1/3$, write $\phi = \phi_{1/3}$, and set $\Phi(\mathbf{X}) = (\phi(X_1), \dots, \phi(X_n))$. This universally prescribed landing coefficient gives the parameter-free feasibility correction. Section 3 proves that it contracts the normal feasibility error while leaving every feasible point unchanged.

**Algorithm 1** PLADOS

**Input:** communication weights $w_{ij}$, batch size $b$, initial point $X_{i,0}$, horizon $K$, and either $\alpha_0$ or $\alpha_K$

**Stepsize:** either $\alpha_k = \alpha_0/\sqrt{k+1}$ with $\alpha_0 \le \bar{\alpha}$, or, for a known horizon, a constant stepsize $\alpha_k \equiv \alpha_K \le \bar{\alpha}$

1: **for** $k = 0, \dots, K-1$ **do**
2: draw $\mathcal{B}_{i,k} = \{\zeta_{i,k}^{(\ell)}\}_{\ell=1}^{b}$
3: compute $S_{i,k}$ by equation (2)
4: $Y_{i,k} = \phi(X_{i,k}) = X_{i,k} - \frac{1}{3}X_{i,k}(X_{i,k}^\top X_{i,k} - I_r)$
5: communicate $Y_{i,k}$ with neighbors to form $\sum_j w_{ij} Y_{j,k}$
6: $X_{i,k+1} = \sum_j w_{ij} Y_{j,k} - \alpha_k S_{i,k}$
7: **end for**

In stacked form, Algorithm 1 is

$$\mathbf{X}_{k+1} = \mathbf{W}\Phi(\mathbf{X}_k) - \alpha_k \mathbf{S}_k. \tag{7}$$

One iteration uses one communication round, one mini-batch gradient, and the products $X_i^\top X_i$ and $X_i(X_i^\top X_i - I_r)$. The landing map costs $O(pr^2)$ local arithmetic and consists entirely of matrix products and additions. Figure 1 illustrates the mechanism. Within the invariant tube, mixing promotes consensus, while landing with fixed $\gamma = 1/3$ reduces feasibility error. With a diminishing stepsize, the iterates approach the Stiefel manifold and achieve vanishing consensus and average stationarity errors under our assumptions.

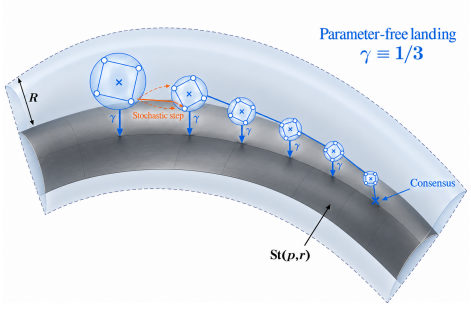



Figure 1: Stochastic landing with $\gamma = 1/3$. Crosses mark network averages.

# 3 CONVERGENCE ANALYSIS

We establish local linear convergence of the penalty-only scheme, then control stochastic disagreement and objective descent. Together, these results yield convergence with diminishing stepsizes and linear speedup with respect to the number of nodes under a horizon-dependent stepsize. Full proofs and a sufficient stepsize cap appear in Appendices A–C.

## 3.1 PENALTY-ONLY SCHEME AND ITS LINEAR CONVERGENCE

The penalty-only scheme $F(\mathbf{X}) = \mathbf{W}\Phi(\mathbf{X})$ combines the polynomial landing map with neighbor communication. We first study how its iteration reduces the distance to the consensus Stiefel manifold.

Let $\Psi(\mathbf{X}) = \mathbf{X} - F(\mathbf{X})$, and let $\bar{\mathbf{X}}$ be the nearest point to $\mathbf{X}$ in $\mathcal{C}$. Set $\rho_\star := \max\{\sigma_2(W), 1/3\}$. The following lemma gives the local geometric property used in the stochastic analysis.

**Lemma 3.1** (Restricted secant inequality). *For every $\rho \in (\rho_\star, 1)$, there is a radius $R > 0$ such that, whenever $d(\mathbf{X}) \le R$, the nearest point $\bar{\mathbf{X}} \in \mathcal{C}$ is unique and*

$$\left\langle \Psi(\mathbf{X}), \mathbf{X} - \bar{\mathbf{X}} \right\rangle \ge \frac{1}{2} \left\| \Psi(\mathbf{X}) \right\|_F^2 + \frac{1-\rho^2}{2} d(\mathbf{X})^2. \tag{8}$$

Although $\mathcal{C}$ intersects a convex consensus subspace with a nonconvex product Stiefel manifold, the RSI gives a convex-like property near the feasible set: the residual is aligned with the direction from the nearest feasible point. Expanding $\|F(\mathbf{X}) - \bar{\mathbf{X}}\|_F^2$ then yields the following linear convergence result with the prescribed $\gamma = 1/3$.

**Theorem 3.2** (Linear convergence of the penalty-only scheme). *For every $\rho \in (\rho_\star, 1)$, after possibly decreasing the radius $R$ in Lemma 3.1, whenever $d(\mathbf{X}) \le R$,*

$$\left\| F(\mathbf{X}) - \bar{\mathbf{X}} \right\|_F \le \rho \left\| \mathbf{X} - \bar{\mathbf{X}} \right\|_F, \qquad d(F(\mathbf{X})) \le \rho d(\mathbf{X}). \tag{9}$$

*Consequently, the fixed-point iteration $\mathbf{X}^{t+1} = F(\mathbf{X}^t)$, initialized with $d(\mathbf{X}^0) \le R$, obeys*

$$d(\mathbf{X}^t) \le \rho^t d(\mathbf{X}^0), \qquad t = 0, 1, \ldots. \tag{10}$$

The local factor $\rho$ can be any number above $\rho_\star = \max\{\sigma_2(W), 1/3\}$ on a sufficiently small tube. Contraction along the normal bundle gives Q-linear decay of the distance to $\mathcal{C}$, controlling consensus and orthogonality errors together.

## 3.2 ONE-STEP DESCENT

We relate feasibility control to objective descent by separating tangent motion from normal perturbations, then bound the effects of landing, sampling noise, and disagreement.

**Lemma 3.3** (Polar projection expansion). *Let $\bar{X} = P_{\mathrm{St}}(\widehat{X})$, and suppose the singular values of $\widehat{X}$ lie in $[s, S] \subset (0, \infty)$. For all sufficiently small $\Delta$,*

$$P_{\mathrm{St}}(\widehat{X} + \Delta) - \bar{X} = \mathcal{L}_{\widehat{X}}[P_{\bar{X}}\Delta] + \mathcal{R}_{\widehat{X}}(\Delta), \tag{11}$$

*where $\mathcal{L}_{\widehat{X}} = DP_{\mathrm{St}}(\widehat{X})$ on $T_{\bar{X}}\,\mathrm{St}$ is uniformly invertible and*

$$\left\| \mathcal{R}_{\widehat{X}}(\Delta) \right\|_F \le C_p \left\| P_{\bar{X}}\Delta \right\|_F \left\| \Delta \right\|_F. \tag{12}$$

The linear map $L_{\widehat{X}}$ is characterized by a Sylvester equation (Higham, 1986; Gawlik & Leok, 2017); Appendix A.1 proves equation (12) and the uniform operator bounds. Combining Theorem 3.2 with the tangent–normal splitting gives the stochastic disagreement recursion.

**Lemma 3.4** (Stochastic disagreement recursion). *There are $C_1 > 0$ and a cap $\bar{\alpha} > 0$ such that, whenever $\mathbf{X}_k$ is in the radius-$R$ tube and $\alpha_k \le \bar{\alpha}$,*

$$\mathbb{E}[d(\mathbf{X}_{k+1})^2 \mid \mathcal{F}_k] \le \frac{3+\rho^2}{4} d(\mathbf{X}_k)^2 + C_1 n \alpha_k^2 \big(\xi^2 + \tfrac{\sigma^2}{b}\big) + C_1 n \alpha_k^4 \big(\left\|\mathrm{grad}\, f(\bar{X}_k)\right\|_F^2 + \tfrac{\sigma^2}{nb}\big). \tag{13}$$

*Moreover, for either nonincreasing schedule stated below, the iterates stay in the tube when $d(\mathbf{X}_0) \le R/2$ and $\alpha_0 \le (1-\rho)R/(2\sqrt{n}G)$.*

The $\alpha_k^2$ term in equation (13) is the usual effect of heterogeneity and independent local noise. A gradient shared by all nodes is instead a common tangent direction; its distance to $\mathcal{C}$ starts at order $\alpha_k^2$, producing the $\alpha_k^4$ term after squaring.

Double stochasticity of the communication matrix gives the exact average update

$$\widehat{X}_{k+1} = \widehat{X}_k + \widehat{F}_k - \alpha_k \widehat{S}_k, \qquad \widehat{F}_k = \frac{1}{n}\sum_i \phi(X_{i,k}) - \widehat{X}_k. \tag{14}$$

Conditional independence yields $\mathbb{E}[\|n^{-1}\sum_i (S_{i,k} - G_{i,k})\|_F^2 \mid \mathcal{F}_k] \le \sigma^2/(nb)$, while smoothness, with a tube-uniform constant $L_g$, gives

$$\left\|\widehat{G}_k - \operatorname{grad} f(\bar{X}_k)\right\|_F^2 \le \frac{L_g^2}{n} d(\mathbf{X}_k)^2. \tag{15}$$

At $\bar{X}_k$, the local gradients average to the global gradient; hence the error in equation (15) is controlled by the disagreement between local iterates.

**Lemma 3.5** (Landing-average bounds)**.** *There are tube-uniform constants $L_a, C_a > 0$ such that*

$$\left\|\widehat{F}_k\right\|_F^2 \le \frac{L_a^2}{n} d(\mathbf{X}_k)^2, \qquad \left\|P_{\bar{X}_k}\widehat{F}_k\right\|_F^2 \le \frac{C_a}{n^2} d(\mathbf{X}_k)^4. \tag{16}$$

The first bound controls the ambient Taylor remainder; the sharper second bound states that landing perturbs the average only to second order in tangent directions. Appendix B records their direct derivation from the explicit polynomial landing map. Applying Lemma 3.3 to equation (14) gives the following descent estimate.

**Lemma 3.6** (One-step descent)**.** *After possibly reducing $\bar{\alpha}$, there are constants $C_2, C_3 > 0$ such that every iterate in the invariant tube satisfies*

$$\begin{aligned}\mathbb{E}[f(\bar{X}_{k+1}) - f(\bar{X}_k)] \le -C_3\alpha_k \mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 + \frac{C_2\alpha_k}{n}\mathbb{E}d(\mathbf{X}_k)^2 + \frac{C_2\alpha_k^2\sigma^2}{nb} \\ + \frac{C_2}{n}\mathbb{E}d(\mathbf{X}_k)^4 + \frac{C_2}{\sqrt{n}}\sqrt{\mathbb{E}d(\mathbf{X}_k)^4\,\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2}.\end{aligned} \tag{17}$$

The factor $\alpha_k$ before $\mathbb{E}d(\mathbf{X}_k)^2$ comes from the tangent projection in equation (11). The fourth-order and mixed terms are reduced in Appendix C using tube invariance and equation (13).

### 3.3 Error accumulation and rates

Summing the disagreement recursion controls its contribution to descent. The higher-order terms can then be absorbed under a common stepsize cap; Appendix C gives the accumulation argument. Define $\mathcal{E}_0 = f(\bar{X}_0) - \inf_{X\in\mathrm{St}} f(X) + d(\mathbf{X}_0)^2 + d(\mathbf{X}_0)/\sqrt{n}$.

**Theorem 3.7** (Unified stochastic bound)**.** *Under Assumptions 2.1–2.3, suppose the deterministic initialization satisfies $d(\mathbf{X}_0) \le R/2$, and let $0 < \alpha_k \le \bar{\alpha}$ be either constant or $\alpha_0/\sqrt{k+1}$. Constants $C_0, C_1, C_2$, independent of $K, b, \sigma, \xi$, satisfy*

$$\frac{\sum_{k<K}\alpha_k \mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2}{\sum_{k<K}\alpha_k} \le \frac{C_0\mathcal{E}_0}{\sum_{k<K}\alpha_k} + \frac{C_1\sigma^2\sum_{k<K}\alpha_k^2}{nb\sum_{k<K}\alpha_k} + \frac{C_2 n(\xi^2+\sigma^2/b)\sum_{k<K}\alpha_k^3}{\sum_{k<K}\alpha_k}. \tag{18}$$

The terms in equation (18) represent optimization progress, averaged sampling noise, and accumulated network disagreement, respectively.

**Corollary 3.8** (Square-root diminishing stepsize)**.** *For $\alpha_k = \alpha_0/\sqrt{k+1}$,*

$$\min_{k<K}\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 \le \frac{C_0\mathcal{E}_0}{\alpha_0\sqrt{K}} + \frac{C_1\alpha_0\sigma^2(1+\log K)}{nb\sqrt{K}} + \frac{C_2 n\alpha_0^2(\xi^2+\sigma^2/b)}{\sqrt{K}}. \tag{19}$$

For a horizon-dependent constant step, fix a problem-scale constant $c > 0$, independent of $K, n, b, \sigma, \xi$, and set

$$\alpha_K = \min\left\{\bar{\alpha}, \frac{c}{1+\sigma\sqrt{K/(nb)} + (n\xi^2K)^{1/3}}\right\}. \tag{20}$$

**Corollary 3.9** (Linear speedup). *Use $\alpha_k \equiv \alpha_K$ from equation (20) for $K$ iterations. For a fixed network, or a uniformly gapped network family with uniformly bounded problem and oracle constants, every $\sigma, \xi \geq 0$, including $\sigma = \xi = 0$, satisfies*

$$\frac{1}{K}\sum_{k<K}\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 = O\left(\frac{\sigma}{\sqrt{nbK}} + \frac{(n\xi^2)^{1/3}}{K^{2/3}} + \frac{n^2}{K}\right). \tag{21}$$

The stochastic term $\sigma/\sqrt{nbK}$ gives linear speedup with respect to the number of nodes when it dominates the other two terms. The term $n^2/K$ is a network-dependent transient. If $\sigma > 0$, it is dominated by the variance term once $K \gtrsim n^5 b/\sigma^2$; if $\xi > 0$, it is dominated by the heterogeneity term once $K \gtrsim n^5/\xi^2$. After either applicable threshold, the remaining two terms recover the classical $K^{-1/2}$ and $K^{-2/3}$ iteration exponents of tracker-free decentralized SGD (Tang et al., 2018; Lian et al., 2017).

## 4 Experiments

We evaluate PLADOS on principal component analysis (PCA) and low-rank matrix completion, examining stepsize and network sensitivity, scalability, data heterogeneity, and computational efficiency against DRSGD (Chen et al., 2021). All runs use $\gamma = 1/3$ and 1 communication round with Metropolis weights (Xiao et al., 2006). Curves are medians over 5 seeds for PCA robustness and scalability, and 3 seeds for matrix completion and heterogeneity. Appendix D provides implementation details.

### 4.1 Decentralized PCA

Node $i$ stores $A_i \in \mathbb{R}^{m_i \times p}$ and minimizes $f_i(X) = -(2m_i)^{-1}\|A_i X\|_F^2$ over $X \in \operatorname{St}(p, r)$, with $r = 5$. We evaluate the projected average $\bar{X}_k = P_{\operatorname{St}}(n^{-1}\sum_i X_{i,k})$ using $\|\operatorname{grad} f(\bar{X}_k)\|_F$ and the subspace distance

$$d_s(\bar{X}_k, X^\star) = \min_{QQ^\top = Q^\top Q = I_r} \|\bar{X}_k - X^\star Q\|_F. \tag{22}$$

Here $X^\star$ spans the leading $r$-dimensional principal subspace. Figure 2 uses the population covariance as the reference; the remaining PCA experiments use the global empirical covariance.

**Synthetic PCA.** For stepsize and network sensitivity, we set $n = 8$, $m_i = 1000$ for all $i$, $p = 10$, and $r = 5$. The rows of each $A_i$ are independently sampled from a zero-mean Gaussian distribution whose covariance has eigenvalues $1, 0.8, \ldots, 0.8^{p-1}$. With batch size 100, each run performs 60000 updates, or 6000 local epochs. Figure 2(a,b) compares PLADOS and DRSGD using $\alpha_k = \alpha_0/\sqrt{k+1}$, with $\alpha_0 \in \{0.01, 0.03, 0.05, 0.08\}$, on an Erdős–Rényi graph (Erdős & Rényi, 1959; Gilbert, 1959) with independent edge probability $0.6$, denoted ER(0.6). Panels (c,d) fix $\alpha_0 = 0.03$ and compare complete, star, ER(0.4), ER(0.6), and ER(0.8) graphs. In panels (a,b), larger values of $\alpha_0$ give faster initial progress, and PLADOS reaches comparable errors in less wall-clock time than DRSGD. In panels (c,d), the trajectories of both metrics nearly overlap across the 5 graphs at $\alpha_0 = 0.03$.

For scalability, we set $p = 100$ and $m_i = 60000/n$, and compare $n = 8, 16, 32$ nodes under a common epoch budget (Chen et al., 2021). With a ring graph and batch size $b = 1$, each run completes 300 epochs, with the same total sample budget across node counts. We use the constant stepsize

$$\alpha_{\text{syn}}(n) = \widehat{\alpha}_n\sqrt{n/300}, \tag{23}$$

with $(\widehat{\alpha}_8, \widehat{\alpha}_{16}, \widehat{\alpha}_{32}) = (0.003, 0.004, 0.006)$. Figure 3(a,b) shows nearly linear speedup in wall-clock time over the tested range of node counts.

**MNIST PCA.** We use the MNIST handwritten-digit dataset (LeCun et al., 1998). The scalability experiment uses all 60000 training images. Each image is vectorized in $\mathbb{R}^p$ with $p = 784$, and pixel values are divided by 255. The node counts, local sample sizes, ring graph, batch size, and epoch budget match the synthetic scalability experiment. We use the constant stepsize

$$\alpha_{\text{MNIST}}(n) = \frac{\widehat{\alpha}_n\sqrt{n}}{10000\sqrt{300}}, \tag{24}$$

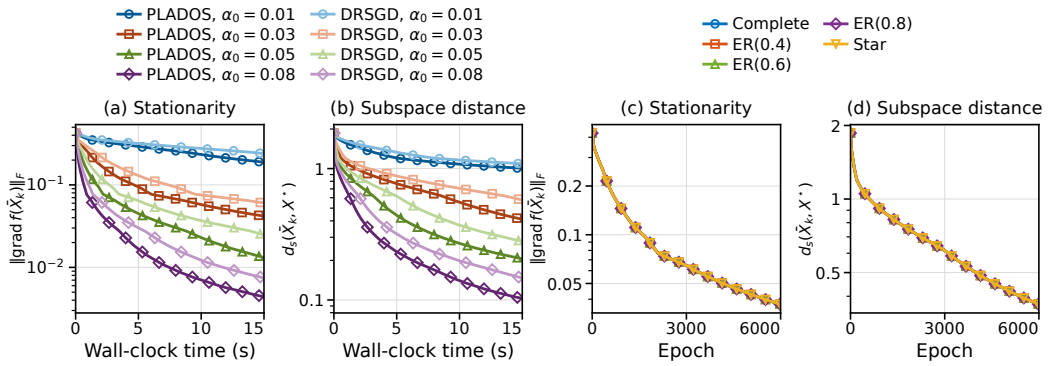



Figure 2: Synthetic PCA: PLADOS and DRSGD at matched diminishing stepsizes (a,b), and PLADOS on different networks (c,d).

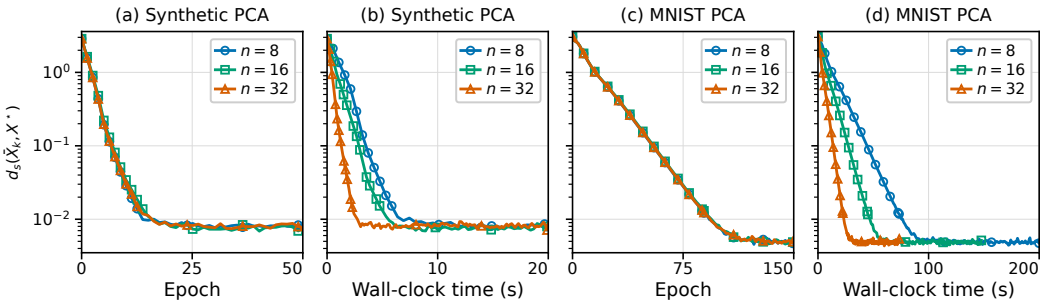



Figure 3: PCA scalability with 8, 16, and 32 MPI nodes. All runs complete 300 epochs.

with $(\widehat{\alpha}_8, \widehat{\alpha}_{16}, \widehat{\alpha}_{32}) = (1.0, 1.4, 2.0)$. Figure 3(c,d) shows aligned epoch curves and nearly linear wall-clock scaling over the tested range. Appendix D.3 details the speedup analysis.

To study data heterogeneity, we adapt the label-based partitioning of Hsieh et al. (2020). Images are sorted by label into 8 equal blocks. Each node retains a randomly selected fraction $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ of its block; the rest is pooled and randomly redistributed in equal shares. A larger $\rho$ induces greater data heterogeneity across nodes through increased label skew. The endpoints $\rho = 0$ and $\rho = 1$ correspond to random assignment and label-based blocks, respectively, while the global objective remains unchanged.

The 8 ring nodes use exact local gradients, $\sigma = 0$, and a shared constant stepsize $\alpha = 7 \times 10^{-4}$. Figure 4 shows the Riemannian gradient norm and subspace distance over epochs 4000–10000. Larger data heterogeneity generally yields larger Riemannian gradient norms and subspace distances. Over the full budget of 15000 updates, runs with larger data heterogeneity exhibit nonzero residual plateaus, illustrating convergence to a neighborhood at the fixed stepsize, as discussed in Appendix D.4. This agrees with the residual term in appendix equation (107).

### 4.2 Decentralized Low-Rank Matrix Completion

We recover a low-rank matrix from partial, noisy observations. Let $Y = [y_1, \ldots, y_m] \in \mathbb{R}^{p \times m}$ collect the noisy columns. The columns are partitioned among $n$ nodes, with $\mathcal{J}_i$ denoting the indices assigned to node $i$ and $m_i = |\mathcal{J}_i|$. Node $i$ minimizes, over $X \in \mathrm{St}(p, r)$,

$$f_i(X) = \frac{1}{m_i} \sum_{j \in \mathcal{J}_i} \min_{v \in \mathbb{R}^r} \frac{1}{2} \|X_{\Omega_j} v - y_{j,\Omega_j}\|_2^2, \tag{25}$$

where $\Omega_j \subseteq \{1, \ldots, p\}$ contains the observed row indices of column $y_j$, and $X_{\Omega_j}$ consists of the corresponding rows of $X$.

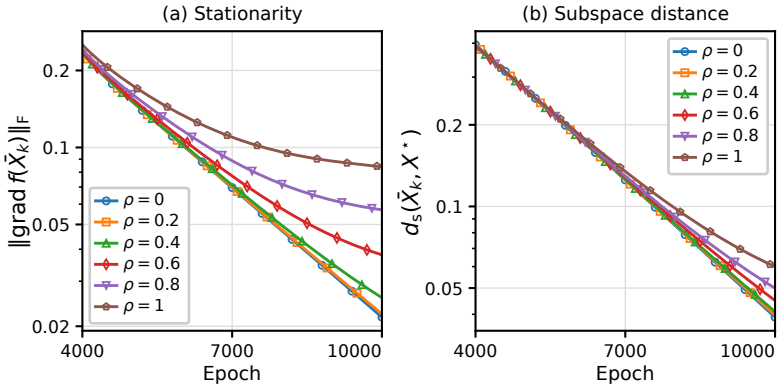



Figure 4: MNIST PCA with heterogeneous partitions.

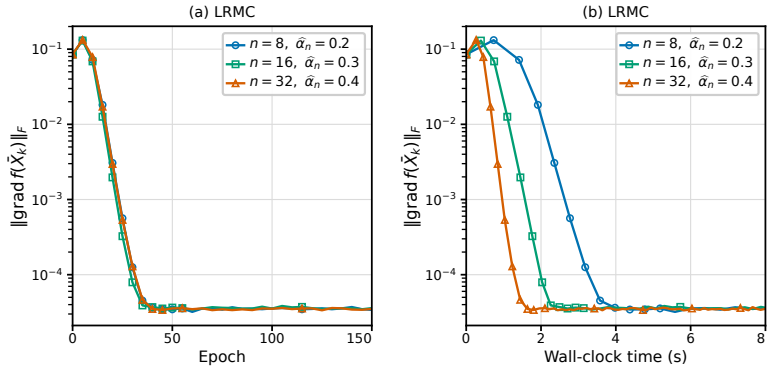



Figure 5: LRMC scalability with 8, 16, and 32 MPI nodes.

We set $p = 100$, $m = 12000$, $r = 5$, and compare $n = 8, 16, 32$ nodes. The data are generated as $Y = UV + E$, where $U \in \mathrm{St}(p, r)$ is the polar factor of a Gaussian $p \times r$ matrix, $V \in \mathbb{R}^{r\times m}$ has independent standard Gaussian entries, and $E \in \mathbb{R}^{p\times m}$ has independent zero-mean Gaussian entries with standard deviation $10^{-3}$. Each entry is observed independently with probability $0.2$. Columns are randomly divided equally among the nodes, so $m_i = m/n$. With a ring graph and batch size $b = 12$, each run completes 800 epochs, with the same total sample budget across node counts. We use the constant stepsize

$$\alpha_{\mathrm{LRMC}}(n) = \widehat{\alpha}_n \sqrt{n/800}, \tag{26}$$

with $(\widehat{\alpha}_8, \widehat{\alpha}_{16}, \widehat{\alpha}_{32}) = (0.2, 0.3, 0.4)$. Figure 5 shows aligned epoch curves and nearly linear speedup in wall-clock time over the tested range of node counts. The displayed ranges are 150 epochs and 8 seconds; all runs complete 800 epochs. Additional experiments in Appendix D show stable consensus behavior (Figure 6) and further demonstrate linear speedup (Figure 7).

## 5 Conclusion

PLADOS makes the plain decentralized stochastic-gradient recursion retraction free on the Stiefel manifold while preserving its leading stochastic and heterogeneity scales. Its parameter-free landing choice $\gamma = 1/3$ yields an RSI and a local contraction of distance to the consensus Stiefel manifold along normal directions; a tangent-preserving polar expansion then transfers this feasibility control to stochastic descent. The resulting guarantee accommodates independent local mini-batches and exhibits linear speedup with respect to the number of nodes after the stated network- and heterogeneity-dependent transient, without a gradient tracker or repeated orthonormalization. Experiments demonstrate nearly linear wall-clock speedup and show that larger data heterogeneity generally yields larger stationarity residuals at a fixed stepsize.

## AI USE STATEMENT

Generative AI tools were used to aid and polish the writing, including language editing and checks of mathematical exposition. The authors take full responsibility for the accuracy, integrity, and originality of the paper.

## REPRODUCIBILITY STATEMENT

Appendices A–C provide the proofs, and Appendix D describes the experimental implementation, stepsize scaling, evaluation procedures, and heterogeneous data construction. The code package is available from the authors upon reasonable request.

## REFERENCES

Pierre Ablin and Gabriel Peyré. Fast and accurate optimization on the orthogonal manifold without retraction. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 5636–5657. PMLR, 2022.

Sulaiman A. Alghunaim and Kun Yuan. A unified and refined convergence analysis for non-convex decentralized learning. *IEEE Transactions on Signal Processing*, 70:3264–3279, 2022. doi: 10.1109/TSP.2022.3184770.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Richard H. Byrd and Jorge Nocedal. Active set and interior methods for nonlinear optimization. In *Proceedings of the International Congress of Mathematicians, Berlin 1998*, volume III of *Documenta Mathematica, Extra Volume ICM 1998*, pp. 667–676. Deutsche Mathematiker-Vereinigung, 1998. doi: 10.4171/DMS/1-3/64.

Jun Chen, Haishan Ye, Mengmeng Wang, Tianxin Huang, Guang Dai, Ivor W. Tsang, and Yong Liu. Decentralized Riemannian conjugate gradient method on the Stiefel manifold. *arXiv preprint arXiv:2308.10547*, 2023.

Jun Chen, Lina Liu, Tianyi Zhu, Yong Liu, Guang Dai, Yunliang Jiang, and Ivor W Tsang. Decentralized optimization on compact submanifolds by quantized Riemannian gradient tracking. *IEEE Transactions on Signal Processing*, 2025.

Shixiang Chen, Alfredo Garcia, Mingyi Hong, and Shahin Shahrampour. Decentralized Riemannian gradient descent on the Stiefel manifold. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1594–1605. PMLR, 2021.

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.

John Duchi, Alekh Agarwal, and Martin Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 2011.

P. Erdős and A. Rényi. On random graphs. I. *Publicationes Mathematicae Debrecen*, 6(3–4):290–297, 1959. doi: 10.5486/PMD.1959.6.3-4.12.

Bin Gao, Xin Liu, and Ya-xiang Yuan. Parallelizable algorithms for optimization problems with orthogonality constraints. *SIAM Journal on Scientific Computing*, 41(3):A1949–A1983, 2019.

Evan S. Gawlik and Melvin Leok. Iterative computation of the Fréchet derivative of the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1354–1379, 2017. doi: 10.1137/16M108971X.

Edgar N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959. doi: 10.1214/aoms/1177706098.

Gene H. Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–432, 1973. doi: 10.1137/0710036.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. doi: 10.1137/0907079.

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-IID data quagmire of decentralized machine learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4387–4398. PMLR, 2020.

Jiang Hu and Kangkang Deng. Improving the communication in decentralized manifold optimization through single-step consensus and compression. *arXiv preprint arXiv:2407.08904*, 2024.

Jiang Hu, Kangkang Deng, and Quanzheng Li. Decentralized Riemannian natural gradient methods with Kronecker product approximations. *Journal of the Operations Research Society of China*, 2025a.

Jiang Hu, Jiaojiao Zhang, and Kangkang Deng. Achieving local consensus over compact submanifolds. *IEEE Transactions on Automatic Control*, 70(9):5750–5763, 2025b. doi: 10.1109/TAC.2025.3545711.

Lei Huang, Xianglong Liu, Bo Lang, Adams Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent Stiefel manifolds in deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *Proceedings of the International Conference on Machine Learning*, 2019.

Anastasia Koloskova, Tao Lin, and Sebastian U. Stich. An improved analysis of gradient tracking for decentralized machine learning. In *Advances in Neural Information Processing Systems*, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Huaqing Li, Lifeng Zheng, Zheng Wang, Yu Yan, Liping Feng, and Jing Guo. S-diging: A stochastic gradient tracking algorithm for distributed optimization. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(1):53–65, 2022.

Shu Li and Jiang Hu. A retraction-free EXTRA method for decentralized optimization on the Stiefel manifold. *arXiv preprint arXiv:2604.23754*, 2026.

Zhi Li, Wei Shi, and Ming Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67(17): 4494–4506, 2019.

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*, volume 30, 2017.

Yue Liu, Tao Lin, Anastasia Koloskova, and Sebastian U. Stich. Decentralized gradient tracking with local steps. *Optimization Methods and Software*, 2025.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on automatic control*, 54(1):48–61, 2009.

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

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 2021.

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*, 2009.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Yilong Song, Peijin Li, Bin Gao, and Kun Yuan. Distributed retraction-free and communication-efficient optimization on the Stiefel manifold. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56281–56318. PMLR, 2025.

Youbang Sun, Shixiang Chen, Alfredo Garcia, and Shahin Shahrampour. Retraction-free decentralized non-convex optimization with orthogonal constraints. *arXiv preprint arXiv:2405.11590*, 2024. doi: 10.48550/arXiv.2405.11590.

Yuki Takezawa, Anastasia Koloskova, and Sebastian U. Stich. Improved convergence analysis of topology dependence in decentralized SGD. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. $D^2$: Decentralized training over decentralized data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4848–4856. PMLR, 2018.

Roberto Tron, Bijan Afsari, and René Vidal. Riemannian consensus for manifolds with bounded curvature. *IEEE Transactions on Automatic Control*, 2012.

John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE transactions on automatic control*, 31(9): 803–812, 1986.

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. doi: 10.1109/TSP.2022.3182883.

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.

Jiayuan Wu, Zhanwang Deng, Jiang Hu, Weijie Su, and Zaiwen Wen. Riemannian EXTRA: Communication-efficient decentralized optimization over compact submanifolds with data heterogeneity. *arXiv preprint arXiv:2505.15537*, 2025.

Lin Xiao, Stephen Boyd, and Sanjay Lall. Distributed average consensus with time-varying Metropolis weights. Unpublished manuscript, 2006.

Nachuan Xiao, Xin Liu, and Kim-Chuan Toh. Dissolving constraints for Riemannian optimization. *Mathematics of Operations Research*, 2024.

Ran Xin, Usman A. Khan, and Soummya Kar. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 2021.

Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 2055–2060. IEEE, 2015.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

Kun Yuan, Bicheng Ying, Xiaochuan Zhao, and Ali H. Sayed. Exact diffusion for distributed optimization and learning — Part I: Algorithm development. *IEEE Transactions on Signal Processing*, 2018.

Yuan Zhang, Jiang Hu, Zhijian Lai, Lin Lin, and Zaiwen Wen. Retraction-free optimization over the Stiefel manifold for the LoRA fine-tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–12, 2026. doi: 10.1109/TPAMI.2026.3711595.

## APPENDIX CONTENTS

**A Geometric estimates** **14**

A.1 Polar projection with a product remainder . . . . . 14

A.2 Parameter-free landing contraction and the RSI . . . . . 16

**B Stochastic recursion and one-step descent** **18**

B.1 Noise and projected-gradient estimates . . . . . 18

B.2 The disagreement recursion . . . . . 20

B.3 Landing average and polar radii . . . . . 22

B.4 One-step descent . . . . . 23

**C Error accumulation and rate proofs** **25**

C.1 Stable-energy weighted disagreement summation . . . . . 25

C.2 Higher-order accumulation . . . . . 26

C.3 One unified stepsize cap and the master bound . . . . . 27

C.4 Corollaries and the explicit network transient . . . . . 28

**D Additional Experimental Details and Results** **28**

D.1 Supplementary experimental settings . . . . . 28

D.2 Additional results on stochastic matrix completion . . . . . 29

D.3 Additional scalability results . . . . . 30

D.4 MNIST PCA with heterogeneous partitions . . . . . 30

## A GEOMETRIC ESTIMATES

All norms in this appendix are Frobenius norms unless an operator norm is displayed explicitly. We first prove the polar expansion used for the projected network average and then prove the parameter-free landing contraction. The coefficient is kept as a general $\gamma \in (0, 1)$ in the latter argument and is set to $1/3$ only after the dependence on $\gamma$ is visible.

### A.1 POLAR PROJECTION WITH A PRODUCT REMAINDER

**Lemma A.1.** *For matrices $A, B$ of the same size and every $\eta > 0$,*

$$2\langle A, B\rangle \le \eta \|A\|_F^2 + \eta^{-1}\|B\|_F^2, \tag{27}$$

$$\|A + B\|_F^2 \le (1+\eta)\|A\|_F^2 + (1+\eta^{-1})\|B\|_F^2. \tag{28}$$

*For nonnegative $a, b$,*

$$(a+b)^4 \le (1+\eta)^3 a^4 + (1+\eta^{-1})^3 b^4. \tag{29}$$

*Proof.* For equation (29), apply Jensen's inequality (Boyd & Vandenberghe, 2004, Section 3.1.8) to $t \mapsto t^4$ with weights $\lambda = (1+\eta)^{-1}$ and $1-\lambda = \eta/(1+\eta)$:

$$(a+b)^4 \le \lambda^{-3} a^4 + (1-\lambda)^{-3} b^4,$$

which is equation (29). □

The polar derivative is standard (Gawlik & Leok, 2017). We use its normal-fiber structure to obtain the product remainder below.

**Lemma A.2.** *Let $X = QH$ be the right polar decomposition, where $Q = P_{\mathrm{St}}(X)$, $H = (X^\top X)^{1/2} = H^\top \succ 0$, and $\sigma(H) \subset [s_-, s_+]$ with $s_- > 0$. If $\|\Delta\|_F \le s_-/2$, then*

$$P_{\mathrm{St}}(X+\Delta) - Q = \mathcal{L}_X[P_Q\Delta] + \mathcal{R}_X(\Delta), \tag{30}$$

$$\|\mathcal{L}_X\|_{\mathrm{op}} \le s_-^{-1}, \tag{31}$$

$$\langle Z, \mathcal{L}_X[Z]\rangle \ge s_+^{-1}\|Z\|_F^2, \qquad Z \in T_Q\,\mathrm{St}, \tag{32}$$

$$\|\mathcal{R}_X(\Delta)\|_F \le C_p\,\|P_Q\Delta\|_F\,\|\Delta\|_F. \tag{33}$$

*Here $P_Q\Delta = \Delta - Q\,\mathrm{sym}(Q^\top\Delta)$, $\mathcal{L}_X = DP_{\mathrm{St}}(X)|_{T_Q\,\mathrm{St}}$, and $C_p$ depends only on $s_-$ and $s_+$.*

*Proof.* The identities $Q^\top Q = I_r$ and $X = QH$ give

$$X^\top X = HQ^\top QH = H^2, \qquad \sigma_j(X) = \sqrt{\lambda_j(X^\top X)} = \lambda_j(H) = \sigma_j(H), \quad j = 1,\dots,r. \tag{34}$$

The last equality uses $H \succ 0$ (Higham, 1986, Lemma 2.1). Thus the spectral bounds imposed on $H$ are exactly the bounds on the singular values of $X$.

Only in this proof, put

$$\Delta_T := P_Q\Delta, \qquad \Delta_N := Q\,\mathrm{sym}(Q^\top\Delta).$$

Then $\Delta_T \in T_Q\,\mathrm{St}$, $\Delta_N \in N_Q\,\mathrm{St}$, and $\Delta = \Delta_T + \Delta_N$. These two components are orthogonal. Indeed, if $A = \mathrm{sym}(Q^\top\Delta)$, then

$$\langle QA, \Delta_T\rangle = \mathrm{tr}(AQ^\top\Delta_T) = 0,$$

because $A$ is symmetric whereas $Q^\top\Delta_T$ is skew symmetric.

The normal segment has constant polar factor. For $0 \le t \le 1$,

$$X + t\Delta_N = Q(H + tA), \qquad \lambda_{\min}(H+tA) \ge s_- - \|\Delta\|_F \ge s_-/2.$$

Thus $H + tA \succ 0$, and uniqueness of the right polar decomposition gives

$$P_{\mathrm{St}}(X + t\Delta_N) = Q, \qquad 0 \le t \le 1. \tag{35}$$

Every point on the segments below has smallest singular value at least $s_-/2$, by Weyl's inequality, and spectral norm at most $s_+ + s_-/2$. The polar factor is smooth on the full-column-rank set, so its second derivative is uniformly bounded on the resulting compact set. Choose $C_p$ to be twice such a bound. By equation (35) and the fundamental theorem of calculus,

$$\begin{aligned} P_{\mathrm{St}}(X+\Delta) - P_{\mathrm{St}}(X) &= \int_0^1 DP_{\mathrm{St}}(X + \Delta_N + t\Delta_T)[\Delta_T]dt \\ &= DP_{\mathrm{St}}(X)[\Delta_T] + \mathcal{R}_X(\Delta), \end{aligned}$$

where

$$\mathcal{R}_X(\Delta) := \int_0^1 \{DP_{\mathrm{St}}(X + \Delta_N + t\Delta_T) - DP_{\mathrm{St}}(X)\}[\Delta_T]dt.$$

Applying the fundamental theorem once more to the derivative yields

$$\begin{aligned} \|\mathcal{R}_X(\Delta)\|_F &\le \frac{C_p}{2}\int_0^1 (\|\Delta_N\|_F + t\,\|\Delta_T\|_F)\,\|\Delta_T\|_F\; dt \\ &\le C_p\,\|\Delta\|_F\,\|\Delta_T\|_F. \end{aligned}$$

This proves equations (30) and (33). In particular, a purely normal perturbation has zero remainder and does not change the polar factor.

It remains to derive the formula for $\mathcal{L}_X$ and prove its two bounds. For $Z \in T_Q\,\mathrm{St}$, write the right polar decomposition of $X + tZ$ as $Q(t)H(t)$ for $t$ near zero, and put $\dot Q = Q'(0)$, $\dot H = H'(0)$, and $\Theta = Q^\top\dot Q$. Differentiating $Q(t)^\top Q(t) = I_r$ shows that $\Theta^\top = -\Theta$, while differentiating $X + tZ = Q(t)H(t)$ gives $Z = \dot QH + Q\dot H$. Left multiplication by $Q^\top$, followed by subtraction of the transpose, eliminates the symmetric matrix $\dot H$ and gives

$$H\Theta + \Theta H = Q^\top Z - Z^\top Q = 2Q^\top Z. \tag{36}$$

Because every eigenvalue sum $h_i + h_j$ is positive, this equation has a unique skew-symmetric solution. Projecting the differentiated factorization onto the orthogonal complement of $Q$ gives $(I - QQ^\top)\dot{Q} = (I - QQ^\top)ZH^{-1}$. Consequently,

$$\mathcal{L}_X[Z] = Q\Theta + (I - QQ^\top)ZH^{-1}, \tag{37}$$

in agreement with the standard Fréchet derivative formula (Higham, 1986; Gawlik & Leok, 2017). Diagonalize $H = V\,\mathrm{diag}(h_1, \ldots, h_r)V^\top$ and set $B = V^\top Q^\top ZV$. Because $Q^\top Z$ is skew symmetric,

$$(V^\top \Theta V)_{ij} = \frac{2B_{ij}}{h_i + h_j}.$$

Consequently,

$$\|\Theta\|_F \le s_-^{-1} \left\|Q^\top Z\right\|_F, \qquad \left\|(I - QQ^\top)ZH^{-1}\right\|_F \le s_-^{-1} \left\|(I - QQ^\top)Z\right\|_F.$$

The two terms in equation (37) are orthogonal, as are the corresponding components of $Z$; squaring and adding proves equation (31). The same eigenbasis calculation gives

$$\left\langle Q^\top Z, \Theta \right\rangle = \sum_{i,j} \frac{2B_{ij}^2}{h_i + h_j} \ge s_+^{-1} \left\|Q^\top Z\right\|_F^2.$$

Together with $H^{-1} \succeq s_+^{-1} I$ in the orthogonal component, this proves equation (32). Since the derivative of the Stiefel-valued curve $P_{\mathrm{St}}(X + tZ)$ lies in $T_Q\,\mathrm{St}$, the map $\mathcal{L}_X$ is an endomorphism of this finite-dimensional tangent space. Cauchy–Schwarz and equation (32) further give

$$\|\mathcal{L}_X[Z]\|_F \ge s_+^{-1} \|Z\|_F.$$

Thus $\mathcal{L}_X$ is injective and hence invertible on $T_Q\,\mathrm{St}$, with $\left\|\mathcal{L}_X^{-1}\right\|_{\mathrm{op}} \le s_+$, which justifies the uniform invertibility asserted in Lemma 3.3. □

### A.2 Parameter-free landing contraction and the RSI

For this subsection, let $\gamma \in (0, 1)$ and define

$$\phi_\gamma(X) = X - \gamma X(X^\top X - I_r), \qquad \Phi_\gamma(\mathbf{X}) = (\phi_\gamma(X_1), \ldots, \phi_\gamma(X_n)), \qquad F_\gamma = \mathbf{W}\Phi_\gamma.$$

**Lemma A.3.** *At $\bar{\mathbf{X}} = \mathbf{1} \otimes Q \in \mathcal{C}$,*

$$N_{\bar{\mathbf{X}}}\mathcal{C} = (\mathbf{1} \otimes N_Q\,\mathrm{St}) \oplus (\mathbf{1}^\perp \otimes \mathbb{R}^{p\times r}), \tag{38}$$

*and the sum is orthogonal, where $N_Q\,\mathrm{St} = \{QA : A = A^\top\}$.*

*Proof.* The embedded Stiefel tangent space is $T_Q\,\mathrm{St} = \{Z : Q^\top Z + Z^\top Q = 0\}$, whose Frobenius-orthogonal complement is $N_Q\,\mathrm{St} = \{QA : A = A^\top\}$ (Boumal, 2023, Section 7.3). Hence $T_{\bar{\mathbf{X}}}\mathcal{C} = \{\mathbf{1} \otimes Z : Z \in T_Q\,\mathrm{St}\}$. For a stack $H = (H_1, \ldots, H_n)$, let $\widehat{H} = n^{-1}\sum_i H_i$. Then

$$\sum_i \langle H_i, Z\rangle = n\left\langle \widehat{H}, Z\right\rangle \quad \text{for every } Z \in T_Q\,\mathrm{St}.$$

Thus $H \in N_{\bar{\mathbf{X}}}\mathcal{C}$ if and only if $\widehat{H} \in N_Q\,\mathrm{St}$. In that case,

$$H = \mathbf{1} \otimes \widehat{H} + (H - \mathbf{1} \otimes \widehat{H}),$$

where the second term has zero network average. This proves equation (38); orthogonality follows from the zero average. □

**Lemma A.4.** *Let $\rho_\star(\gamma) = \max\{\sigma_2(W), |1 - 2\gamma|\}$. At every $\bar{\mathbf{X}} \in \mathcal{C}$,*

$$DF_\gamma(\bar{\mathbf{X}})[u] = u, \qquad u \in T_{\bar{\mathbf{X}}}\mathcal{C}, \tag{39}$$

$$\left\|DF_\gamma(\bar{\mathbf{X}})[v]\right\|_F \le \rho_\star(\gamma)\|v\|_F, \qquad v \in N_{\bar{\mathbf{X}}}\mathcal{C}. \tag{40}$$

*Moreover, choose $\rho \in (\rho_\star(\gamma), 1)$, and suppose*

$$\gamma(6R + 3R^2) \le \rho - \rho_\star(\gamma). \tag{41}$$

*If $\bar{\mathbf{X}} \in \mathcal{C}$, $v \in N_{\bar{\mathbf{X}}}\mathcal{C}$, $\|v\|_F = 1$, and $\left\|\mathbf{Y} - \bar{\mathbf{X}}\right\|_F \le R$, then*

$$\|DF_\gamma(\mathbf{Y})[v]\|_F \le \rho. \tag{42}$$

*Proof.* Write $\bar{\mathbf{X}} = \mathbf{1} \otimes Q$. At the feasible block $Q$,

$$D\phi_\gamma(Q)[H] = H - \gamma Q(H^\top Q + Q^\top H).$$

Every ambient matrix has the orthogonal decomposition $H = P_Q H + Q\,\mathrm{sym}(Q^\top H)$. The derivative is the identity on the first component and multiplication by $1 - 2\gamma$ on the second. In particular, its operator norm is one.

If $u \in T_{\bar{\mathbf{X}}}\mathcal{C}$, then $u = \mathbf{1} \otimes H$ for some $H \in T_Q\,\mathrm{St}$. The blockwise landing derivative fixes $u$, and $W\mathbf{1} = \mathbf{1}$, so communication also fixes it. This proves equation (39).

By Lemma A.3, write the orthogonal decomposition

$$v = \mathbf{1} \otimes QA + z, \qquad A = A^\top, \qquad \widehat{z} = 0.$$

Communication fixes the consensus-normal component, on which the landing derivative acts by $1 - 2\gamma$. The same blockwise derivative maps the zero-average subspace into itself: its network average is $D\phi_\gamma(Q)[\widehat{z}] = 0$. Afterward $W - J$ contracts it by $\sigma_2(W) = \|W - J\|_2$. The two images remain orthogonal, and hence

$$\begin{aligned}\left\|DF_\gamma(\bar{\mathbf{X}})[v]\right\|_F^2 &\le |1 - 2\gamma|^2 \left\|\mathbf{1} \otimes QA\right\|_F^2 + \sigma_2(W)^2 \left\|z\right\|_F^2 \\ &\le \rho_\star(\gamma)^2 \left\|v\right\|_F^2.\end{aligned}$$

By homogeneity this proves equation (40) for every normal vector.

It remains to control the derivative away from $\mathcal{C}$. For one block, write $Y = Q + E$. Direct differentiation gives

$$D\phi_\gamma(Y)[H] = H - \gamma\{H(Y^\top Y - I_r) + Y(H^\top Y + Y^\top H)\}.$$

The first difference from the corresponding expression at $Q$ is bounded by $(2\left\|E\right\|_F + \left\|E\right\|_F^2)\left\|H\right\|_F$. Expanding the second difference into

$$E(H^\top Q + Q^\top H) + Q(H^\top E + E^\top H) + E(H^\top E + E^\top H)$$

bounds it by $(4\left\|E\right\|_F + 2\left\|E\right\|_F^2)\left\|H\right\|_F$. Therefore

$$\left\|(D\phi_\gamma(Y) - D\phi_\gamma(Q))[H]\right\|_F \le \gamma(6R + 3R^2)\left\|H\right\|_F \tag{43}$$

whenever $\left\|E\right\|_F \le R$. Applying this blockwise, using $\left\|W\right\|_2 = 1$, and then combining it with the contraction at $\bar{\mathbf{X}}$ gives

$$\left\|DF_\gamma(\mathbf{Y})[v]\right\|_F \le \{\rho_\star(\gamma) + \gamma(6R + 3R^2)\}\left\|v\right\|_F \le \rho,$$

which is equation (42). □

*Proof of Lemma 3.1 and Theorem 3.2.* By the smooth nearest-point projection theorem (Boumal, 2023, Theorem 5.53), shrink $R$, if necessary, so that the radius-$R$ neighborhood is a tubular neighborhood of $\mathcal{C}$. Let $\bar{\mathbf{X}} = \mathbf{1} \otimes \bar{X}$ be the unique nearest point to $\mathbf{X}$. The tubular-neighborhood property gives $\mathbf{X} - \bar{\mathbf{X}} \in N_{\bar{\mathbf{X}}}\mathcal{C}$, and every point of the normal segment

$$\bar{\mathbf{X}} + t(\mathbf{X} - \bar{\mathbf{X}}), \qquad 0 \le t \le 1,$$

is within distance $R$ of $\bar{\mathbf{X}}$. Moreover, $F_\gamma(\bar{\mathbf{X}}) = \bar{\mathbf{X}}$. The fundamental theorem of calculus and Lemma A.4 therefore give the stronger, fixed-representative estimate

$$\begin{aligned}\left\|F_\gamma(\mathbf{X}) - \bar{\mathbf{X}}\right\|_F &= \left\|\int_0^1 DF_\gamma(\bar{\mathbf{X}} + t(\mathbf{X} - \bar{\mathbf{X}}))[\mathbf{X} - \bar{\mathbf{X}}]dt\right\|_F \\ &\le \rho\left\|\mathbf{X} - \bar{\mathbf{X}}\right\|_F.\end{aligned} \tag{44}$$

Taking the distance from $F_\gamma(\mathbf{X})$ to the nearest point in $\mathcal{C}$ can only decrease the left-hand side, so $d(F_\gamma(\mathbf{X})) \le \rho d(\mathbf{X})$.

Now set $\gamma = 1/3$, for which $\rho_\star(1/3) = \max\{\sigma_2(W), 1/3\}$. For example, after also imposing the tubular-radius restriction, any

$$0 < R \le \min\left\{\frac{1}{8}, -1 + \sqrt{1 + \rho - \max\{\sigma_2(W), 1/3\}}\right\}$$

satisfies equation (41). This proves the contraction theorem.

Finally, let $F = F_{1/3}$ and $\Psi(\mathbf{X}) = \mathbf{X} - F(\mathbf{X})$. With $e = \mathbf{X} - \bar{\mathbf{X}}$ and $r = \Psi(\mathbf{X})$, the fixed-representative estimate is $\|e - r\|_F^2 \le \rho^2 \|e\|_F^2$. Expanding its left-hand side and rearranging gives the restricted secant inequality

$$\left\langle \Psi(\mathbf{X}), \mathbf{X} - \bar{\mathbf{X}} \right\rangle \ge \frac{1}{2} \|\Psi(\mathbf{X})\|_F^2 + \frac{1-\rho^2}{2} d(\mathbf{X})^2. \tag{45}$$

Thus both the contraction and the RSI use the same fixed nearest representative. Conversely, substituting the RSI back into the exact identity

$$\left\|F(\mathbf{X}) - \bar{\mathbf{X}}\right\|_F^2 = d(\mathbf{X})^2 + \|\Psi(\mathbf{X})\|_F^2 - 2\left\langle \Psi(\mathbf{X}), \mathbf{X} - \bar{\mathbf{X}} \right\rangle$$

recovers the fixed-representative contraction, so the two estimates are equivalent on this tube.

Finally, if $\mathbf{X}^{t+1} = F(\mathbf{X}^t)$ and $d(\mathbf{X}^0) \le R$, then $d(\mathbf{X}^{t+1}) \le \rho d(\mathbf{X}^t) \le R$. Induction gives $d(\mathbf{X}^t) \le \rho^t d(\mathbf{X}^0)$. At $\gamma = 1/3$, equations (39) and (40) also give, at every $\bar{\mathbf{X}} \in \mathcal{C}$,

$$DF(\bar{\mathbf{X}})[u] = u \quad (u \in T_{\bar{\mathbf{X}}}\mathcal{C}), \qquad \left\|DF(\bar{\mathbf{X}})|_{N_{\bar{\mathbf{X}}}\mathcal{C}}\right\|_{\mathrm{op}} \le \rho_\star < 1. \tag{46}$$

This identifies the tangent and normal behavior of the linearization. □

## B Stochastic recursion and one-step descent

Let $\mathcal{F}_k$ be the sigma-field generated by the initialization and all mini-batches (and any oracle randomness) strictly before time $k$. Equivalently, it is the history immediately before the time-$k$ mini-batches are drawn. Thus $\mathbf{X}_k$, $\bar{X}_k$, and all population gradients evaluated at $\mathbf{X}_k$ are $\mathcal{F}_k$-measurable. Compactness and Assumptions 2.1–2.3 give finite constants $L_g, L_f, G_f$ such that, throughout the tube,

$$\|P_X \nabla f_i(X) - P_Y \nabla f_i(Y)\|_F \le L_g \|X - Y\|_F, \tag{47}$$

$$\|\mathrm{grad}\, f(Q)\|_F \le G_f, \qquad Q \in \mathrm{St}, \tag{48}$$

$$f(Q') - f(Q) \le \langle \mathrm{grad}\, f(Q), Q' - Q \rangle + \frac{L_f}{2} \|Q' - Q\|_F^2, \qquad Q, Q' \in \mathrm{St}. \tag{49}$$

These bounds follow from the stated local regularity on the compact tube. Since $P_X$ is uniformly bounded there, we enlarge $G$ so that $G \ge 1$ bounds both the batch gradient and $\|S_{i,k}\|_F$. This common bound will be used in the pathwise tube-invariance argument.

### B.1 Noise and projected-gradient estimates

Define

$$G_{i,k} := P_{X_{i,k}} \nabla f_i(X_{i,k}), \qquad \Xi_{i,k} := S_{i,k} - G_{i,k},$$

and let $\mathbf{G}_k$ and $\mathbf{\Xi}_k$ denote the corresponding stacks. The raw-oracle assumption, the $\mathcal{F}_k$-measurability of $X_{i,k}$, and linearity of $U \mapsto P_{X_{i,k}} U$ give

$$\mathbb{E}[S_{i,k} \mid \mathcal{F}_k] = P_{X_{i,k}} \mathbb{E}\left[ \frac{1}{b} \sum_{\ell=1}^{b} \nabla_X F_i(X_{i,k}; \zeta_{i,k}^{(\ell)}) \,\middle|\, \mathcal{F}_k \right] = P_{X_{i,k}} \nabla f_i(X_{i,k}) = G_{i,k}.$$

Thus $\mathbf{S}_k = \mathbf{G}_k + \mathbf{\Xi}_k$ and $\mathbb{E}[\Xi_{i,k} \mid \mathcal{F}_k] = 0$. Because conditionally independent raw mini-batches remain conditionally independent after the deterministic maps $U \mapsto P_{X_{i,k}} U$, we have

$$\mathbb{E}[\Xi_{i,k} \mid \mathcal{F}_k] = 0, \tag{50}$$

$$\mathbb{E}[\left\|\widehat{\mathbf{\Xi}}_k\right\|_F^2 \mid \mathcal{F}_k] \le \frac{\sigma^2}{nb}, \tag{51}$$

$$\mathbb{E}[\|(I - \mathbf{J})\mathbf{\Xi}_k\|_F^2 \mid \mathcal{F}_k] \le \frac{n\sigma^2}{b}. \tag{52}$$

Indeed, expanding the first squared norm conditionally gives

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_i \Xi_{i,k}\right\|_F^2 \,\middle|\, \mathcal{F}_k\right] = \frac{1}{n^2}\sum_i \mathbb{E}[\|\Xi_{i,k}\|_F^2 \mid \mathcal{F}_k] + \frac{1}{n^2}\sum_{i\neq j} \mathbb{E}[\langle \Xi_{i,k}, \Xi_{j,k}\rangle \mid \mathcal{F}_k].$$

For $i \neq j$, conditional independence and equation (50) give

$$\mathbb{E}[\langle \Xi_{i,k}, \Xi_{j,k}\rangle \mid \mathcal{F}_k] = \langle \mathbb{E}[\Xi_{i,k} \mid \mathcal{F}_k], \mathbb{E}[\Xi_{j,k} \mid \mathcal{F}_k]\rangle = 0.$$

This proves equation (51). Since $I - \mathbf{J}$ is an orthogonal projector,

$$\mathbb{E}[\|(I-\mathbf{J})\boldsymbol{\Xi}_k\|_F^2 \mid \mathcal{F}_k] \le \sum_i \mathbb{E}[\|\Xi_{i,k}\|_F^2 \mid \mathcal{F}_k],$$

which proves equation (52).

At the common feasible point $\bar{X}_k$, linearity of $P_{\bar{X}_k}$ and the definition of $f$ give $n^{-1}\sum_i \operatorname{grad} f_i(\bar{X}_k) = \operatorname{grad} f(\bar{X}_k)$. Moreover, nonexpansiveness of the feasible tangent projection and the heterogeneity assumption imply

$$\sum_i \left\|\operatorname{grad} f_i(\bar{X}_k) - \operatorname{grad} f(\bar{X}_k)\right\|_F^2 \le \sum_i \left\|\nabla f_i(\bar{X}_k) - \nabla f(\bar{X}_k)\right\|_F^2 \le n\xi^2. \tag{53}$$

By equation (47) and Jensen's inequality,

$$\begin{aligned}\left\|\widehat{G}_k - \operatorname{grad} f(\bar{X}_k)\right\|_F^2 &= \left\|\frac{1}{n}\sum_i (G_{i,k} - \operatorname{grad} f_i(\bar{X}_k))\right\|_F^2 \\ &\le \frac{1}{n}\sum_i \left\|G_{i,k} - \operatorname{grad} f_i(\bar{X}_k)\right\|_F^2 \le \frac{L_g^2}{n} d(\mathbf{X}_k)^2.\end{aligned} \tag{54}$$

For the centered stack, insert the local gradients evaluated at $\bar{X}_k$. Using $\|I-\mathbf{J}\|_2 \le 1$, $\|A+B\|_F^2 \le 2\|A\|_F^2 + 2\|B\|_F^2$, and equation (53) gives

$$\begin{aligned}\|(I-\mathbf{J})\mathbf{G}_k\|_F^2 &\le 2\sum_i \left\|G_{i,k} - \operatorname{grad} f_i(\bar{X}_k)\right\|_F^2 + 2\sum_i \left\|\operatorname{grad} f_i(\bar{X}_k) - \operatorname{grad} f(\bar{X}_k)\right\|_F^2 \\ &\le 2L_g^2 d(\mathbf{X}_k)^2 + 2n\xi^2.\end{aligned}$$

The population stack is measurable and the noise has zero conditional mean, so the conditional cross term vanishes. Together with equation (52), this yields

$$\mathbb{E}[\|(I-\mathbf{J})\mathbf{S}_k\|_F^2 \mid \mathcal{F}_k] \le 2L_g^2 d(\mathbf{X}_k)^2 + 2n\xi^2 + \frac{n\sigma^2}{b}. \tag{55}$$

The same orthogonality for the averaged noise gives

$$\begin{aligned}\mathbb{E}[\left\|\widehat{S}_k\right\|_F^2 \mid \mathcal{F}_k] &= \left\|\widehat{G}_k\right\|_F^2 + \mathbb{E}[\left\|\widehat{\boldsymbol{\Xi}}_k\right\|_F^2 \mid \mathcal{F}_k] \\ &\le 2\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 + \frac{2L_g^2}{n} d(\mathbf{X}_k)^2 + \frac{\sigma^2}{nb}.\end{aligned} \tag{56}$$

Finally, the almost-sure oracle bound, enlarged at the start of this appendix, implies the pathwise estimates

$$\|\mathbf{S}_k\|_F \le \sqrt{n}G, \qquad \left\|\widehat{S}_k\right\|_F \le G. \tag{57}$$

### B.2 THE DISAGREEMENT RECURSION

The next estimate controls the normal component of the stochastic direction at $\bar{X}_k$. Using $(I - P_{\bar{X}_k})P_{\bar{X}_k} = 0$, we obtain

$$(I - P_{\bar{X}_k})S_{i,k} = (I - P_{\bar{X}_k})(P_{X_{i,k}} - P_{\bar{X}_k}) \left( \frac{1}{b} \sum_{\ell=1}^{b} \nabla_X F_i(X_{i,k}; \zeta_{i,k}^{(\ell)}) \right).$$

From

$$(P_X - P_{\bar{X}_k})U = -(X - \bar{X}_k)\,\mathrm{sym}(X^\top U) - \bar{X}_k\,\mathrm{sym}((X - \bar{X}_k)^\top U)$$

and $\|X\|_2 \le 1 + R$ in the tube, we obtain

$$\left\|(I - P_{\bar{X}_k})S_{i,k}\right\|_F \le L_P G \left\|X_{i,k} - \bar{X}_k\right\|_F, \qquad L_P := 2 + R. \tag{58}$$

**Lemma B.1.** *Fix an iteration $k$ and let $\mathbf{Y}_k = F(\mathbf{X}_k)$. If $\alpha_k \left\|\widehat{S}_k\right\|_F \le 1/2$, then, for every $\eta > 0$,*

$$\begin{aligned} d(\mathbf{Y}_k - \alpha_k \mathbf{S}_k)^2 \le{}& (1+\eta) \left\|\mathbf{Y}_k - \mathbf{1} \otimes \bar{X}_k\right\|_F^2 \\ &+ C_t(\eta)\alpha_k^2 \left\{ \|(I - \mathbf{J})\mathbf{S}_k\|_F^2 + \sum_i \left\|(I - P_{\bar{X}_k})S_{i,k}\right\|_F^2 \right\} \\ &+ C_t(\eta) n \alpha_k^4 \left\|\widehat{S}_k\right\|_F^4, \end{aligned} \tag{59}$$

*where $C_t(\eta) = 2(1 + \eta^{-1}) \max\{1, C_p^2\}$.*

*Proof.* Use the feasible comparison point $P_{\mathrm{St}}(\bar{X}_k - \alpha_k P_{\bar{X}_k}\widehat{S}_k)$. Lemma A.2, applied at $\bar{X}_k$, has $\mathcal{L}_{\bar{X}_k}[P_{\bar{X}_k}\widehat{S}_k] = P_{\bar{X}_k}\widehat{S}_k$. Hence

$$\begin{aligned} &P_{\mathrm{St}}(\bar{X}_k - \alpha_k P_{\bar{X}_k}\widehat{S}_k) \\ &\quad = \bar{X}_k - \alpha_k P_{\bar{X}_k}\widehat{S}_k + \mathcal{R}_{\bar{X}_k}(-\alpha_k P_{\bar{X}_k}\widehat{S}_k), \\ &\left\|\mathcal{R}_{\bar{X}_k}(-\alpha_k P_{\bar{X}_k}\widehat{S}_k)\right\|_F \le C_p \alpha_k^2 \left\|P_{\bar{X}_k}\widehat{S}_k\right\|_F^2. \end{aligned} \tag{60}$$

The condition $\alpha_k \left\|\widehat{S}_k\right\|_F \le 1/2$ guarantees the radius required in that lemma. Direct subtraction gives

$$\begin{aligned} &\mathbf{Y}_k - \alpha_k \mathbf{S}_k - \mathbf{1} \otimes P_{\mathrm{St}}(\bar{X}_k - \alpha_k P_{\bar{X}_k}\widehat{S}_k) \\ &\quad = \mathbf{Y}_k - \mathbf{1} \otimes \bar{X}_k - \alpha_k(\mathbf{S}_k - \mathbf{1} \otimes P_{\bar{X}_k}\widehat{S}_k) \\ &\qquad - \mathbf{1} \otimes \mathcal{R}_{\bar{X}_k}(-\alpha_k P_{\bar{X}_k}\widehat{S}_k). \end{aligned}$$

The middle bracket has the exact orthogonal decomposition

$$\mathbf{S}_k - \mathbf{1} \otimes P_{\bar{X}_k}\widehat{S}_k = (I - \mathbf{J})\mathbf{S}_k + \mathbf{1} \otimes (I - P_{\bar{X}_k})\widehat{S}_k. \tag{61}$$

Indeed, the first term has zero network average and the second is a consensus stack. Jensen's inequality further gives

$$n \left\|(I - P_{\bar{X}_k})\widehat{S}_k\right\|_F^2 \le \sum_i \left\|(I - P_{\bar{X}_k})S_{i,k}\right\|_F^2. \tag{62}$$

Apply equation (28) first to the initial error and the sum of the last two terms, and then use $\|B + C\|_F^2 \le 2\|B\|_F^2 + 2\|C\|_F^2$. Equations (60)– (62), together with $\left\|P_{\bar{X}_k}\widehat{S}_k\right\|_F \le \left\|\widehat{S}_k\right\|_F$, give equation (59). Finally, distance to $\mathcal{C}$ is no larger than the norm to the particular feasible stack $\mathbf{1} \otimes P_{\mathrm{St}}(\bar{X}_k - \alpha_k P_{\bar{X}_k}\widehat{S}_k)$, which completes the proof. □

*Proof of Lemma 3.4.* Put $\mathbf{Y}_k = F(\mathbf{X}_k)$ and use the current nearest representative $\bar{\mathbf{X}}_k = \mathbf{1} \otimes \bar{X}_k$. The fixed-representative contraction equation (44) gives

$$\left\|\mathbf{Y}_k - \bar{\mathbf{X}}_k\right\|_F \le \rho d(\mathbf{X}_k). \tag{63}$$

Moreover, equation (58) gives

$$\sum_i \left\|(I - P_{\bar{X}_k})S_{i,k}\right\|_F^2 \le L_P^2 G^2 d(\mathbf{X}_k)^2. \tag{64}$$

Temporarily impose $\alpha_k G \le 1/2$; the fiber restriction equation (80) below is strictly stronger. Apply Lemma B.1 with $\eta = (1-\rho^2)/(2\rho^2)$. Conditional expectation, followed by equations (55) and (56), and the pathwise inequality $\left\|\widehat{S}_k\right\|_F^4 \le G^2 \left\|\widehat{S}_k\right\|_F^2$, yields

$$\begin{aligned}\mathbb{E}[d(\mathbf{X}_{k+1})^2 \mid \mathcal{F}_k] \le \frac{1+\rho^2}{2} d(\mathbf{X}_k)^2 + C_t \alpha_k^2 \left\{(2L_g^2 + L_P^2 G^2) d(\mathbf{X}_k)^2 + 2n\xi^2 + \frac{n\sigma^2}{b}\right\} \\ + C_t n G^2 \alpha_k^4 \left\{2\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 + \frac{2L_g^2}{n} d(\mathbf{X}_k)^2 + \frac{\sigma^2}{nb}\right\},\end{aligned} \tag{65}$$

where $C_t = 2(1+\rho^2)(1-\rho^2)^{-1}\max\{1, C_p^2\}$. Let

$$C_d := C_t\{2L_g^2 + L_P^2 G^2 + 2G^2 L_g^2\}. \tag{66}$$

Because $\alpha_k \le 1$, all multiples of $d(\mathbf{X}_k)^2$ after $(1+\rho^2)/2$ in equation (65) are at most $C_d \alpha_k^2 d(\mathbf{X}_k)^2$. Therefore the explicit condition

$$\alpha_k \le \alpha_d := \sqrt{\frac{1-\rho^2}{4C_d}} \tag{67}$$

gives

$$\frac{1+\rho^2}{2} + C_d \alpha_k^2 \le \frac{3+\rho^2}{4} < 1. \tag{68}$$

The explicit choice

$$C_1 := 2C_t \max\{1, G^2\} \tag{69}$$

dominates every remaining coefficient, independently of $k, n, b, \sigma, \xi$, and gives

$$\begin{aligned}\mathbb{E}[d(\mathbf{X}_{k+1})^2 \mid \mathcal{F}_k] \le \frac{3+\rho^2}{4} d(\mathbf{X}_k)^2 + C_1 n \alpha_k^2 \left(\xi^2 + \frac{\sigma^2}{b}\right) \\ + C_1 n \alpha_k^4 \left(\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 + \frac{\sigma^2}{nb}\right),\end{aligned} \tag{70}$$

which is the claimed stochastic recursion.

It remains to justify that every estimate above is applied inside the tube. Nearest-point minimality and equations (44) and (57) give the pathwise inequality

$$\begin{aligned}d(\mathbf{X}_{k+1}) &\le d(F(\mathbf{X}_k)) + \alpha_k \left\|\mathbf{S}_k\right\|_F \\ &\le \rho d(\mathbf{X}_k) + \alpha_k \sqrt{n} G.\end{aligned} \tag{71}$$

If $d(\mathbf{X}_0) \le R/2$ and

$$\alpha_0 \le \alpha_t := \frac{(1-\rho)R}{2\sqrt{n}G}, \tag{72}$$

then the induction hypothesis $d(\mathbf{X}_k) \le R$ implies

$$d(\mathbf{X}_{k+1}) \le \rho R + \frac{1-\rho}{2} R = \frac{1+\rho}{2} R < R.$$

The initialization starts the induction, proving tube invariance. □

### B.3 LANDING AVERAGE AND POLAR RADII

For the algorithmic value $\gamma = 1/3$, write

$$\widehat{F}_k = -\frac{\gamma}{n}\sum_i X_{i,k}(X_{i,k}^\top X_{i,k} - I_r).$$

Since $\bar{X}_k^\top \bar{X}_k = I_r$,

$$X_{i,k}^\top X_{i,k} - I_r = (X_{i,k} - \bar{X}_k)^\top X_{i,k} + \bar{X}_k^\top (X_{i,k} - \bar{X}_k).$$

Thus, on the radius-$R$ tube,

$$\left\|X_{i,k}^\top X_{i,k} - I_r\right\|_F \le (2+R)\left\|X_{i,k} - \bar{X}_k\right\|_F, \qquad \|X_{i,k}\|_2 \le 1 + R. \tag{73}$$

Set

$$L_a := \gamma(1+R)(2+R), \qquad C_a := \gamma^2(2+R)^2. \tag{74}$$

Using $\|\sum_i A_i\|_F^2 \le n\sum_i \|A_i\|_F^2$ gives

$$\left\|\widehat{F}_k\right\|_F^2 \le \frac{L_a^2}{n} d(\mathbf{X}_k)^2. \tag{75}$$

For the tangent component, symmetry of $X_{i,k}^\top X_{i,k} - I_r$ implies $P_{\bar{X}_k}[\bar{X}_k(X_{i,k}^\top X_{i,k} - I_r)] = 0$, and hence

$$P_{\bar{X}_k}[X_{i,k}(X_{i,k}^\top X_{i,k} - I_r)] = P_{\bar{X}_k}[(X_{i,k} - \bar{X}_k)(X_{i,k}^\top X_{i,k} - I_r)].$$

The feasible tangent projection is nonexpansive, so

$$\left\|P_{\bar{X}_k}\widehat{F}_k\right\|_F \le \frac{\gamma(2+R)}{n}\sum_i \left\|X_{i,k} - \bar{X}_k\right\|_F^2.$$

After squaring,

$$\left\|P_{\bar{X}_k}\widehat{F}_k\right\|_F^2 \le \frac{C_a}{n^2} d(\mathbf{X}_k)^4 \le \frac{C_a}{n} d(\mathbf{X}_k)^4. \tag{76}$$

Jensen's inequality gives $\left\|\widehat{X}_k - \bar{X}_k\right\|_F \le d(\mathbf{X}_k)/\sqrt{n}$. Write the right polar decomposition as $\widehat{X}_k = \bar{X}_k H_k$, where $H_k = (\widehat{X}_k^\top \widehat{X}_k)^{1/2} \succ 0$. By equation (34),

$$\sigma_j(\widehat{X}_k) = \lambda_j(H_k) = \sigma_j(H_k), \qquad j = 1, \dots, r. \tag{77}$$

Moreover, $\|H_k - I_r\|_F = \left\|\widehat{X}_k - \bar{X}_k\right\|_F$, because $\bar{X}_k^\top \bar{X}_k = I_r$. Hence every eigenvalue of $H_k$ lies in $[r_-, r_+]$, where

$$r_- := 1 - \frac{R}{\sqrt{n}}, \qquad r_+ := 1 + \frac{R}{\sqrt{n}}. \tag{78}$$

In particular, $R \le 1/8$ gives $r_- \ge 7/8$. The averaged update is

$$\widehat{X}_{k+1} - \widehat{X}_k = \Delta_k := \widehat{F}_k - \alpha_k \widehat{S}_k. \tag{79}$$

Equations (57) and (75) imply

$$\|\Delta_k\|_F \le \frac{L_a R}{\sqrt{n}} + \alpha_k G.$$

The numerator in the following quantity is positive:

$$\alpha_f := \frac{r_-/2 - L_a R/\sqrt{n}}{G} > 0. \tag{80}$$

Indeed, $r_-/2 \ge 7/16$, whereas $L_a R/\sqrt{n} \le (51/64)(1/8) = 51/512 < 7/16$. Consequently, $\alpha_k \le \alpha_f$ ensures

$$\|\Delta_k\|_F \le r_-/2, \tag{81}$$

so Lemma A.2 applies at every iteration. Notice also that $\alpha_f G < 1/2$; hence the same cap guarantees the polar radius used in Lemma B.1.

### B.4 ONE-STEP DESCENT

*Proof of Lemma 3.6.* Lemma A.2 and equation (79) give the exact identity

$$\bar{X}_{k+1} - \bar{X}_k = \mathcal{L}_{\widehat{X}_k}[P_{\bar{X}_k}\Delta_k] + \mathcal{R}_{\widehat{X}_k}(\Delta_k). \tag{82}$$

Substitution into equation (49) produces four terms:

$$\begin{aligned}\mathbb{E}[f(\bar{X}_{k+1}) - f(\bar{X}_k)] \le{}& \mathbb{E}\left\langle \operatorname{grad} f(\bar{X}_k), \mathcal{L}_{\widehat{X}_k}[P_{\bar{X}_k}\widehat{F}_k]\right\rangle - \alpha_k \mathbb{E}\left\langle \operatorname{grad} f(\bar{X}_k), \mathcal{L}_{\widehat{X}_k}[P_{\bar{X}_k}\widehat{S}_k]\right\rangle \\ &+ \mathbb{E}\left\langle \operatorname{grad} f(\bar{X}_k), \mathcal{R}_{\widehat{X}_k}(\Delta_k)\right\rangle + \frac{L_f}{2}\mathbb{E}\left\|\bar{X}_{k+1} - \bar{X}_k\right\|_F^2. \end{aligned} \tag{83}$$

We bound these terms in their displayed order, keeping every absorption coefficient explicit.

First decompose $\widehat{S}_k = \widehat{G}_k + \widehat{\mathbf{\Xi}}_k$. All factors in the inner product containing the current noise, except the noise itself, are $\mathcal{F}_k$-measurable. Linearity and equation (50) therefore give

$$\mathbb{E}[\left\langle \operatorname{grad} f(\bar{X}_k), \mathcal{L}_{\widehat{X}_k}[P_{\bar{X}_k}\widehat{\mathbf{\Xi}}_k]\right\rangle \mid \mathcal{F}_k] = 0.$$

Since $P_{\bar{X}_k} \operatorname{grad} f(\bar{X}_k) = \operatorname{grad} f(\bar{X}_k)$,

$$P_{\bar{X}_k}\widehat{G}_k = \operatorname{grad} f(\bar{X}_k) + P_{\bar{X}_k}(\widehat{G}_k - \operatorname{grad} f(\bar{X}_k)).$$

Using equations (32), (31), and (54), then applying Cauchy–Schwarz, gives

$$\mathbb{E}\left\langle \operatorname{grad} f(\bar{X}_k), \mathcal{L}_{\widehat{X}_k}[P_{\bar{X}_k}\widehat{S}_k]\right\rangle \ge \frac{1}{r_+}\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 - \frac{L_g}{r_-}\sqrt{\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 \frac{\mathbb{E}d(\mathbf{X}_k)^2}{n}}.$$

The scalar inequality $ab \le a^2/(4r_+) + r_+ b^2$ now yields

$$-\alpha_k \mathbb{E}\left\langle \operatorname{grad} f(\bar{X}_k), \mathcal{L}_{\widehat{X}_k}[P_{\bar{X}_k}\widehat{S}_k]\right\rangle \le -\frac{3\alpha_k}{4r_+}\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 + \frac{r_+ L_g^2 \alpha_k}{r_-^2 n}\mathbb{E}d(\mathbf{X}_k)^2. \tag{84}$$

For the landing term, equation (31), Cauchy–Schwarz, and the weaker form of equation (76) give

$$\mathbb{E}\left\langle \operatorname{grad} f(\bar{X}_k), \mathcal{L}_{\widehat{X}_k}[P_{\bar{X}_k}\widehat{F}_k]\right\rangle \le \frac{\sqrt{C_a}}{r_-\sqrt{n}}\sqrt{\mathbb{E}d(\mathbf{X}_k)^4\, \mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2}. \tag{85}$$

For the polar remainder, Lemma A.2 gives

$$\left\|\mathcal{R}_{\widehat{X}_k}(\Delta_k)\right\|_F \le C_p \left\|P_{\bar{X}_k}\Delta_k\right\|_F \left\|\Delta_k\right\|_F.$$

Applying $\|A+B\|_F^2 \le 2\|A\|_F^2 + 2\|B\|_F^2$ to both squared factors and multiplying the results gives

$$\begin{aligned}\left\|P_{\bar{X}_k}\Delta_k\right\|_F^2 \left\|\Delta_k\right\|_F^2 \le 4\{ &\left\|P_{\bar{X}_k}\widehat{F}_k\right\|_F^2 \left\|\widehat{F}_k\right\|_F^2 + \alpha_k^2 \left\|P_{\bar{X}_k}\widehat{F}_k\right\|_F^2 \left\|\widehat{S}_k\right\|_F^2 \\ &+ \alpha_k^2 \left\|P_{\bar{X}_k}\widehat{S}_k\right\|_F^2 \left\|\widehat{F}_k\right\|_F^2 + \alpha_k^4 \left\|P_{\bar{X}_k}\widehat{S}_k\right\|_F^2 \left\|\widehat{S}_k\right\|_F^2\}. \end{aligned} \tag{86}$$

The four products in equation (86) are bounded one by one by equations (75), (76), (57), and (56). After taking expectation, the result is

$$\begin{aligned}\mathbb{E}[\left\|P_{\bar{X}_k}\Delta_k\right\|_F^2 \left\|\Delta_k\right\|_F^2] \le{}& \frac{4C_a(G^2 + L_a^2 R^2)}{n}\mathbb{E}d(\mathbf{X}_k)^4 + \frac{4G^2 L_a^2 \alpha_k^2}{n}\mathbb{E}d(\mathbf{X}_k)^2 \\ &+ 8G^2\alpha_k^4 \mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 + \frac{8G^2 L_g^2 \alpha_k^4}{n}\mathbb{E}d(\mathbf{X}_k)^2 + 4G^2\alpha_k^4\frac{\sigma^2}{nb}. \end{aligned} \tag{87}$$

For example, the last product in equation (86) is at most $4G^2\alpha_k^4 \left\|\widehat{S}_k\right\|_F^2$; applying equation (56) produces exactly the last three terms in equation (87). The other three products give the first two terms, after using $d(\mathbf{X}_k) \le R$ and $n \ge 1$.

By Cauchy–Schwarz and $\sqrt{x_1+\cdots+x_m} \le \sum_j \sqrt{x_j}$, the first and third terms of equation (87) contribute, respectively,

$$\frac{2C_p\sqrt{C_a(G^2+L_a^2R^2)}}{\sqrt{n}}\sqrt{\mathbb{E}d(\mathbf{X}_k)^4\,\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2}, \tag{88}$$

$$2\sqrt{2}C_pG\alpha_k^2\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2. \tag{89}$$

For the other three square-root products, Young's inequality gives

$$\begin{aligned}
&2C_pGL_a\alpha_k\sqrt{\frac{\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2\,\mathbb{E}d(\mathbf{X}_k)^2}{n}}\\
&\qquad\le \frac{\alpha_k}{16r_+}\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2+\frac{16r_+C_p^2G^2L_a^2\alpha_k}{n}\mathbb{E}d(\mathbf{X}_k)^2, \qquad (90)\\
&2\sqrt{2}C_pGL_g\alpha_k^2\sqrt{\frac{\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2\,\mathbb{E}d(\mathbf{X}_k)^2}{n}}\\
&\qquad\le \frac{\alpha_k}{16r_+}\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2+\frac{32r_+C_p^2G^2L_g^2\alpha_k^3}{n}\mathbb{E}d(\mathbf{X}_k)^2,\\
&2C_pG\alpha_k^2\sqrt{\frac{\sigma^2}{nb}\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2}\\
&\qquad\le \frac{\alpha_k}{16r_+}\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2+16r_+C_p^2G^2\alpha_k^3\frac{\sigma^2}{nb}.
\end{aligned}$$

Since $\alpha_k \le 1$, the two $\alpha_k^3$ remainders are bounded by the required $\alpha_k\mathbb{E}d(\mathbf{X}_k)^2/n$ and $\alpha_k^2\sigma^2/(nb)$ terms. These three applications spend exactly $3\alpha_k/(16r_+)$ of the descent coefficient.

It remains to bound the chord term. Equations (82), (31), (33), and (81) give

$$\left\|\bar{X}_{k+1}-\bar{X}_k\right\|_F \le \left(r_-^{-1}+\frac{C_pr_-}{2}\right)\left\|P_{\bar{X}_k}\Delta_k\right\|_F.$$

Also, $\left\|P_{\bar{X}_k}\Delta_k\right\|_F^2 \le 2\left\|P_{\bar{X}_k}\widehat{F}_k\right\|_F^2+2\alpha_k^2\left\|P_{\bar{X}_k}\widehat{S}_k\right\|_F^2$. Consequently, by equations (76) and (56),

$$\begin{aligned}
\frac{L_f}{2}\mathbb{E}\left\|\bar{X}_{k+1}-\bar{X}_k\right\|_F^2 \le L_f\left(r_-^{-1}+\frac{C_pr_-}{2}\right)^2\Bigg\{&\frac{C_a}{n}\mathbb{E}d(\mathbf{X}_k)^4+2\alpha_k^2\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2\\
&+\frac{2L_g^2\alpha_k^2}{n}\mathbb{E}d(\mathbf{X}_k)^2+\alpha_k^2\frac{\sigma^2}{nb}\Bigg\}. \qquad (91)
\end{aligned}$$

We now collect the coefficients. The main descent in equation (84) is $-12\alpha_k/(16r_+)$. The three Young inequalities above add $3\alpha_k/(16r_+)$. The only remaining positive multiples of $\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2$ are equation (89) and the second term of equation (91). Put

$$C_g := 2\sqrt{2}C_pG+2L_f\left(r_-^{-1}+\frac{C_pr_-}{2}\right)^2, \qquad \alpha_g := \frac{1}{16r_+C_g}. \tag{92}$$

Then $\alpha_k \le \alpha_g$ implies $C_g\alpha_k^2 \le \alpha_k/(16r_+)$, leaving $-8\alpha_k/(16r_+) = -\alpha_k/(2r_+)$. Thus we may take

$$C_3 := \frac{1}{2r_+}. \tag{93}$$

All other coefficients were displayed above. We may take

$$C_2 := \max\left\{\begin{array}{l}\dfrac{r_+L_g^2}{r_-^2}+16r_+C_p^2G^2(L_a^2+2L_g^2)+2L_f\left(r_-^{-1}+\dfrac{C_pr_-}{2}\right)^2L_g^2,\\[2ex] 16r_+C_p^2G^2+L_f\left(r_-^{-1}+\dfrac{C_pr_-}{2}\right)^2, \quad L_f\left(r_-^{-1}+\dfrac{C_pr_-}{2}\right)^2C_a,\\[2ex] \dfrac{\sqrt{C_a}}{r_-}+2C_p\sqrt{C_a(G^2+L_a^2R^2)}\end{array}\right\}.$$

Equations (83)– (91) now give

$$\mathbb{E}[f(\bar{X}_{k+1}) - f(\bar{X}_k)] \le -C_3\alpha_k \mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 + \frac{C_2\alpha_k}{n}\mathbb{E}d(\mathbf{X}_k)^2 + \frac{C_2\alpha_k^2\sigma^2}{nb} + \frac{C_2}{n}\mathbb{E}d(\mathbf{X}_k)^4 + \frac{C_2}{\sqrt{n}}\sqrt{\mathbb{E}d(\mathbf{X}_k)^4\,\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2}, \tag{94}$$

which proves Lemma 3.6. □

The constants $C_2$ and $C_3$ are fixed entirely by this one-step argument: $C_3$ is the descent coefficient left after the four explicit absorptions, while $C_2$ dominates the variance, disagreement, fourth-order, and mixed remainders. The stable scalar energy estimate used in Appendix C only accumulates the already established disagreement recursion.

# C Error accumulation and rate proofs

All constants in this appendix depend only on the compact tube, smoothness and oracle bounds, and the fixed network spectral gap. They are independent of $K, b, \sigma, \xi$. The numbered constants $C_0, C_1, C_2, C_3$ in the main statements denote common enlargements of the fixed constants below.

## C.1 Stable-energy weighted disagreement summation

We use the following scalar energy estimate from Xu et al. (2015, Lemma 2) to accumulate the stable disagreement recursion, omitting the unused forcing term at the final index.

**Lemma C.1** (Stable scalar energy estimate)**.** *Suppose nonnegative sequences satisfy* $v_{k+1} \le \eta v_k + w_k$, *where* $0 < \eta < 1$. *Then, for every* $K \ge 1$,

$$\left(\sum_{k=0}^{K} v_k^2\right)^{1/2} \le \frac{\sqrt{2}}{1-\eta}\left(\sum_{k=0}^{K-1} w_k^2\right)^{1/2} + \sqrt{\frac{2}{1-\eta^2}}\, v_0. \tag{95}$$

Let $\alpha_k$ be either constant or $\alpha_0/\sqrt{k+1}$, so that $0 < \alpha_{k+1} \le \alpha_k \le \alpha_0 \le 1$. Set

$$D_k := \mathbb{E}d(\mathbf{X}_k)^2, \qquad \mathcal{G}_k := \mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2, \qquad A := \xi^2 + \frac{\sigma^2}{b}.$$

Taking total expectation in equation (70), multiplying by $\alpha_{k+1}$, taking square roots, and using $\sqrt{x_1 + \cdots + x_m} \le \sum_j \sqrt{x_j}$ gives, for $v_k := \sqrt{\alpha_k D_k}$,

$$v_{k+1} \le \frac{\sqrt{3+\rho^2}}{2}\, v_k + \sqrt{C_1 nA}\,\alpha_k^{3/2} + \sqrt{C_1 n}\,\alpha_k^{5/2}\sqrt{\mathcal{G}_k} + \sqrt{\frac{C_1\sigma^2}{b}}\,\alpha_k^{5/2}. \tag{96}$$

For example, $\sqrt{\alpha_{k+1}\alpha_k^2} \le \alpha_k^{3/2}$ and $\sqrt{\alpha_{k+1}\alpha_k^4} \le \alpha_k^{5/2}$. Only these two index-shift inequalities are needed.

Apply Lemma C.1 with $\eta = \sqrt{3+\rho^2}/2$, and use the triangle inequality for the three forcing sequences in equation (96). Squaring the resulting four-term bound with $(a_1 + \cdots + a_4)^2 \le 4\sum_j a_j^2$ yields

$$\sum_{k=0}^{K} \alpha_k D_k \le C_E\left\{\alpha_0 D_0 + C_1 nA \sum_{k<K} \alpha_k^3 + C_1 n \sum_{k<K} \alpha_k^5 \mathcal{G}_k + \frac{C_1\sigma^2}{b}\sum_{k<K}\alpha_k^5\right\}, \tag{97}$$

where the explicit fixed choice

$$C_E := 4\max\left\{\frac{2}{(1-\sqrt{3+\rho^2}/2)^2}, \frac{8}{1-\rho^2}\right\}$$

is independent of $K, b, \sigma, \xi$ (and is uniform over a network family whenever $\rho$ is bounded uniformly below one). Finally,

$$\alpha_k^5\mathcal{G}_k \le \alpha_0^4\alpha_k\mathcal{G}_k,$$
$$\frac{\sigma^2}{b}\alpha_k^5 \le n\left(\xi^2 + \frac{\sigma^2}{b}\right)\alpha_k^3,$$

because $\alpha_k \le 1$ and $n \ge 1$. The prescribed initialization is deterministic, so $D_0 = d(\mathbf{X}_0)^2$; for a random initialization the same proof simply replaces this quantity by its expectation. Dropping the nonnegative $k = K$ term on the left and taking $C_0 \ge C_E \max\{1, C_1\}$ proves

$$\sum_{k<K} \alpha_k \mathbb{E} d(\mathbf{X}_k)^2 \le C_0 \left\{ \alpha_0 d(\mathbf{X}_0)^2 + n \left( \xi^2 + \frac{\sigma^2}{b} \right) \sum_{k<K} \alpha_k^3 + n\alpha_0^4 \sum_{k<K} \alpha_k \mathbb{E} \left\| \operatorname{grad} f(\bar{X}_k) \right\|_F^2 \right\}. \tag{98}$$

Conditional independence gives the one-step $1/(nb)$ variance bound, and the energy argument retains the powers of $n$, $b$, and $\alpha_k$ required by the master bound. The final $\alpha_k^5$ noise term is deliberately upper-bounded by the displayed $nA\alpha_k^3$ network term.

### C.2 Higher-order accumulation

**Lemma C.2** (Higher-order accumulation)**.** *Suppose tube invariance holds and the stepsize is either constant or $\alpha_k = \alpha_0/\sqrt{k+1}$, with $0 < \alpha_0 \le 1$. There is a fixed $C_h > 0$ such that, for every $\varepsilon > 0$ and $K \ge 1$,*

$$\begin{aligned} &\frac{C_2}{n} \sum_{k<K} \mathbb{E} d(\mathbf{X}_k)^4 + \frac{C_2}{\sqrt{n}} \sum_{k<K} \sqrt{\mathbb{E} d(\mathbf{X}_k)^4 \, \mathbb{E} \left\| \operatorname{grad} f(\bar{X}_k) \right\|_F^2} \\ &\quad \le C_h \left\{ \frac{d(\mathbf{X}_0)^2}{n} + \frac{d(\mathbf{X}_0)}{\sqrt{n}} \right\} + \varepsilon \sum_{k<K} \alpha_k \mathbb{E} \left\| \operatorname{grad} f(\bar{X}_k) \right\|_F^2 \\ &\qquad + C_h \left( \alpha_0 + \frac{1}{\varepsilon} \right) \sum_{k<K} \alpha_k \mathbb{E} d(\mathbf{X}_k)^2. \end{aligned} \tag{99}$$

*Proof.* Iterating the pathwise recursion equation (71) gives

$$d(\mathbf{X}_k) \le \rho^k d(\mathbf{X}_0) + \sqrt{n} G \sum_{\ell=0}^{k-1} \rho^{k-1-\ell} \alpha_\ell. \tag{100}$$

For a constant schedule, the convolution in equation (100) is at most $\alpha_k/(1-\rho)$. For $\alpha_k = \alpha_0/\sqrt{k+1}$, set $j = k-1-\ell$. For $0 \le j \le k-1$,

$$\frac{\alpha_{k-1-j}}{\alpha_k} = \sqrt{\frac{k+1}{k-j}} \le \sqrt{j+2},$$

where the squared inequality follows from $(j+2)(k-j) - (k+1) = (j+1)(k-j-1) \ge 0$. Therefore both schedules obey

$$d(\mathbf{X}_k) \le \rho^k d(\mathbf{X}_0) + \sqrt{n} G \alpha_k \sum_{j=0}^{\infty} \rho^j \sqrt{j+2}. \tag{101}$$

The series in equation (101) is finite by the ratio test. Its lower bound $1/(1-\rho)$ also covers constant stepsizes.

Squaring equation (101), multiplying by $d(\mathbf{X}_k)^2$, and then using the pathwise tube bound $d(\mathbf{X}_k) \le R$ gives

$$\begin{aligned} \mathbb{E} d(\mathbf{X}_k)^4 &\le 2R^2 \rho^{2k} d(\mathbf{X}_0)^2 \\ &\quad + 2nG^2 \alpha_k^2 \left( \sum_{j=0}^{\infty} \rho^j \sqrt{j+2} \right)^2 \mathbb{E} d(\mathbf{X}_k)^2. \end{aligned} \tag{102}$$

Consequently, geometric summation and $\alpha_k^2 \le \alpha_0 \alpha_k$ imply

$$\begin{aligned} \frac{C_2}{n} \sum_{k<K} \mathbb{E} d(\mathbf{X}_k)^4 &\le \frac{2C_2 R^2}{1-\rho^2} \frac{d(\mathbf{X}_0)^2}{n} \\ &\quad + 2C_2 G^2 \left( \sum_{j=0}^{\infty} \rho^j \sqrt{j+2} \right)^2 \alpha_0 \sum_{k<K} \alpha_k \mathbb{E} d(\mathbf{X}_k)^2. \end{aligned} \tag{103}$$

Next, take the square root of equation (102), use $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ and $\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 \le G_f^2$, and multiply by $C_2/\sqrt{n}$. This yields, term by term,

$$\begin{aligned}
&\frac{C_2}{\sqrt{n}}\sqrt{\mathbb{E}d(\mathbf{X}_k)^4\,\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2} \\
&\quad\le \frac{\sqrt{2}C_2RG_f}{\sqrt{n}}\rho^k d(\mathbf{X}_0) + \sqrt{2}C_2G\left(\sum_{j=0}^{\infty}\rho^j\sqrt{j+2}\right)\alpha_k\sqrt{\mathbb{E}d(\mathbf{X}_k)^2\,\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2} \\
&\quad\le \frac{\sqrt{2}C_2RG_f}{\sqrt{n}}\rho^k d(\mathbf{X}_0) + \varepsilon\alpha_k\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 \\
&\qquad + \frac{C_2^2G^2}{2\varepsilon}\left(\sum_{j=0}^{\infty}\rho^j\sqrt{j+2}\right)^2\alpha_k\mathbb{E}d(\mathbf{X}_k)^2,
\end{aligned}$$

where the last line is Young's inequality. Summing this display and using $\sum_{k<K}\rho^k \le (1-\rho)^{-1}$, then combining it with equation (103), proves equation (99) after choosing $C_h$ no smaller than the finitely many displayed fixed coefficients. □

### C.3 One unified stepsize cap and the master bound

Sum equation (94) from $k=0$ to $K-1$, use $f(\bar{X}_K) \ge \inf_{X\in\mathrm{St}} f(X)$, and apply Lemma C.2 with $\varepsilon = C_3/4$. Moving this part of the gradient sum to the left gives directly

$$\begin{aligned}
\frac{3C_3}{4}\sum_{k<K}\alpha_k\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 \le f(\bar{X}_0) - \inf_{X\in\mathrm{St}} f(X) + C_h\left\{\frac{d(\mathbf{X}_0)^2}{n} + \frac{d(\mathbf{X}_0)}{\sqrt{n}}\right\} \\
+ \frac{C_2\sigma^2}{nb}\sum_{k<K}\alpha_k^2 + \left\{C_2 + C_h\left(\alpha_0 + \frac{4}{C_3}\right)\right\}\sum_{k<K}\alpha_k\mathbb{E}d(\mathbf{X}_k)^2.
\end{aligned} \tag{104}$$

Here $n^{-1} \le 1$ was used for the disagreement term in equation (94). Inserting equation (98) into equation (104), and using $\alpha_0 \le 1$, shows that the only term returned to the gradient sum is at most

$$C_0 n\left\{C_2 + C_h\left(1 + \frac{4}{C_3}\right)\right\}\alpha_0^4\sum_{k<K}\alpha_k\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2.$$

Thus every local restriction and the final feedback absorption are enforced by the single sufficient cap

$$\bar{\alpha} := \min\left\{1,\ \frac{(1-\rho)R}{2\sqrt{n}G},\ \frac{r_-/2 - L_aR/\sqrt{n}}{G},\ \sqrt{\frac{1-\rho^2}{4C_d}},\ \frac{1}{16r_+C_g},\ \left[\frac{C_3}{4C_0n\{C_2 + C_h(1+4/C_3)\}}\right]^{1/4}\right\}. \tag{105}$$

All entries are positive. In order, they enforce normalization, tube invariance, the polar-fiber radii, the disagreement margin, the linear descent absorption, and the final error-feedback absorption.

If $\alpha_0 \le \bar{\alpha}$, the last entry of equation (105) makes the returned coefficient at most $C_3/4$. Moving it to the left of equation (104) leaves at least $C_3/2$. After enlarging the numbered constants, we obtain

$$\begin{aligned}
\frac{C_3}{2}\sum_{k<K}\alpha_k\mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 \le C_0\left\{f(\bar{X}_0) - \inf_{X\in\mathrm{St}} f(X) + d(\mathbf{X}_0)^2 + \frac{d(\mathbf{X}_0)}{\sqrt{n}}\right\} \\
+ \frac{C_1\sigma^2}{nb}\sum_{k<K}\alpha_k^2 + C_2n\left(\xi^2 + \frac{\sigma^2}{b}\right)\sum_{k<K}\alpha_k^3.
\end{aligned} \tag{106}$$

Dividing by $(C_3/2)\sum_{k<K}\alpha_k$, and using the definitions in the main text, proves Theorem 3.7.

### C.4 COROLLARIES AND THE EXPLICIT NETWORK TRANSIENT

For $\alpha_k = \alpha_0/\sqrt{k+1}$, elementary integral comparisons give

$$\sum_{k<K} \alpha_k \geq \alpha_0\sqrt{K}, \qquad \sum_{k<K} \alpha_k^2 \leq \alpha_0^2(1+\log K), \qquad \sum_{k<K} \alpha_k^3 \leq 3\alpha_0^3.$$

Substitution into equation (18) proves Corollary 3.8; the minimum of nonnegative values is no larger than their positive weighted average.

For a constant step $\alpha \leq \bar{\alpha}$, equation (18) gives

$$\frac{1}{K}\sum_{k<K} \mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 \leq \frac{C_0\mathcal{E}_0}{\alpha K} + \frac{C_1\alpha\sigma^2}{nb} + C_2 n\alpha^2\left(\xi^2 + \frac{\sigma^2}{b}\right). \tag{107}$$

For the horizon-dependent choice equation (20), the definition of the minimum and $\max\{x,y\} \leq x+y$ give, for every $K \geq 1$,

$$\frac{1}{\alpha_K K} \leq \left(\frac{1}{\bar{\alpha}} + \frac{1}{c}\right)\frac{1}{K} + \frac{\sigma}{c\sqrt{nbK}} + \frac{(n\xi^2)^{1/3}}{cK^{2/3}}, \tag{108}$$

$$\frac{\alpha_K\sigma^2}{nb} \leq \frac{c\sigma^2}{nb\{1+\sigma\sqrt{K/(nb)}\}} \leq \frac{c\sigma}{\sqrt{nbK}}, \tag{109}$$

$$n\alpha_K^2\xi^2 \leq \frac{c^2 n\xi^2}{\{1+(n\xi^2K)^{1/3}\}^2} \leq \frac{c^2(n\xi^2)^{1/3}}{K^{2/3}}, \tag{110}$$

$$\frac{n\alpha_K^2\sigma^2}{b} \leq \frac{c^2 n\sigma^2}{b\{1+\sigma\sqrt{K/(nb)}\}^2} \leq \frac{c^2 n^2}{K}. \tag{111}$$

The second inequality uses $z^2/(1+z) \leq z$, the third uses $z^3/(1+z)^2 \leq z$, and the fourth uses $z^2/(1+z)^2 \leq 1$, all for $z \geq 0$. Under the fixed-network or uniformly gapped-family conditions of Corollary 3.9, the explicit cap equation (105) gives $\bar{\alpha}^{-1} = O(\sqrt{n})$. Substitution into equation (107), together with $\sqrt{n}/K \leq n^2/K$, therefore proves the always-valid bound

$$\frac{1}{K}\sum_{k<K} \mathbb{E}\left\|\operatorname{grad} f(\bar{X}_k)\right\|_F^2 = O\left(\frac{\sigma}{\sqrt{nbK}} + \frac{(n\xi^2)^{1/3}}{K^{2/3}} + \frac{n^2}{K}\right). \tag{112}$$

For fixed problem and network constants, the last term is no larger than the sampling term when $\sigma > 0$ and $K \geq n^5 b/\sigma^2$, or no larger than the heterogeneity term when $\xi > 0$ and $K \geq n^5/\xi^2$. Whenever either condition holds, the last term in equation (112) is absorbed by one of the first two, which gives the post-transient two-term form discussed after Corollary 3.9.

## D ADDITIONAL EXPERIMENTAL DETAILS AND RESULTS

### D.1 SUPPLEMENTARY EXPERIMENTAL SETTINGS

All nodes start from the same polar projection of a Gaussian matrix. PCA gradients use the empirical covariance, and subspace distances are computed by orthogonal Procrustes alignment (Chen et al., 2021). The reference in Figure 2 is the population principal subspace; Figures 3 and 4 use the full empirical principal subspace. Synthetic covariance eigenvalues are $0.8^j$, $j = 0, \ldots, p-1$, in a random orthogonal basis. For the network comparison, ER graphs threshold shared edge uniforms, giving nested edge sets; graphs are generated independently of the data.

Figure 2 and LRMC use fresh uniform minibatches, independent across iterations and nodes, without replacement within each batch. The heterogeneity experiment uses exact local gradients.

The LRMC scalability data distributions and dimensions are specified in Section 4.2. Each stochastic gradient solves the sampled columnwise least-squares problems. A column contributes $(X_{\Omega_j}v_j(X) - y_{j,\Omega_j})v_j(X)^\top$ on observed rows and zero elsewhere, averaged over the sampled columns. The reduced loss is smooth where the observed submatrices $X_{\Omega_j}$ have full column rank (Golub & Pereyra, 1973).

Stepsize legends show coefficients before scaling. Figure 2 uses $\alpha_k = \alpha_0/\sqrt{k+1}$, with no additional scaling. Synthetic PCA, MNIST PCA, and LRMC scalability multiply $\widehat{\alpha}_n$ by $\sqrt{n/300}$, $\sqrt{n/300}/10000$, and $\sqrt{n/800}$, respectively. The LRMC coefficients $(0.2, 0.3, 0.4)$ give actual stepsizes $(0.02, 0.03\sqrt{2}, 0.08)$. Heterogeneity uses the unscaled constant $7 \times 10^{-4}$. In Figure 2, DRSGD uses 1 communication round, consensus stepsize 1, and the polar retraction (Chen et al., 2021). Both methods share data, initialization, minibatch streams, and stepsizes; their execution order alternates.

Figure 2 simulates all nodes in 1 Python process on the server. Figures 3 and 5 use MPI neighbor communication on the same server, with 1 physical core per node and single-threaded BLAS. Algorithm time includes sampling, gradient evaluation, communication, and updates, and excludes initialization and diagnostics. Figure 2 and LRMC scalability use 20 untimed warm-up updates on a separate state.

Figure 2 records diagnostics every 50 updates. LRMC scalability records the initial state, the first update, and approximately every 5 epochs thereafter. Both figures pair median cumulative time with median error at the same recorded iteration, without interpolation. PCA scalability uses pointwise medians on common epoch or time grids.

### D.2 Additional results on stochastic matrix completion

We compare PLADOS and DRSGD (Chen et al., 2021) on the LRMC problem in equation (25), with $p = 100$, $m = 1000$, $r = 5$, and $n = 8$. Data follow the construction in Section 4.2. Each entry is observed independently with probability $10r(p+m-r)/(pm) = 0.5475$, so the expected number of observations is 10 times the degrees of freedom of a rank-$r$ matrix. Each node holds 125 columns and draws fresh batches of 12 columns on a ring. Both methods perform 3000 updates using $\alpha_k = n\alpha_0/\sqrt{k+1}$, with $\alpha_0 \in \{0.04, 0.06, 0.08\}$; legends show $\alpha_0$ before scaling. DRSGD uses 1 communication round, consensus stepsize 1, and the polar retraction. The methods share data, initialization, and minibatch streams. All nodes are simulated in 1 process on the same local machine, with single-threaded BLAS and the timing exclusions in Appendix D.1.

Figure 6 reports the Riemannian gradient norm and the consensus error $\|\mathbf{X}_k - \bar{\mathbf{X}}_k\|_F$, where $\bar{\mathbf{X}}_k = \mathbf{1} \otimes \bar{X}_k$. Each point pairs median cumulative wall-clock time with median error at the same iteration over 3 runs. Across the 3 stepsizes, PLADOS rapidly reduces consensus error and maintains low disagreement as the steps diminish. Compared with DRSGD, PLADOS shows a slight advantage, while both methods reach comparable final Riemannian gradient norms. These results show stable consensus behavior for the tested stepsizes without retractions.

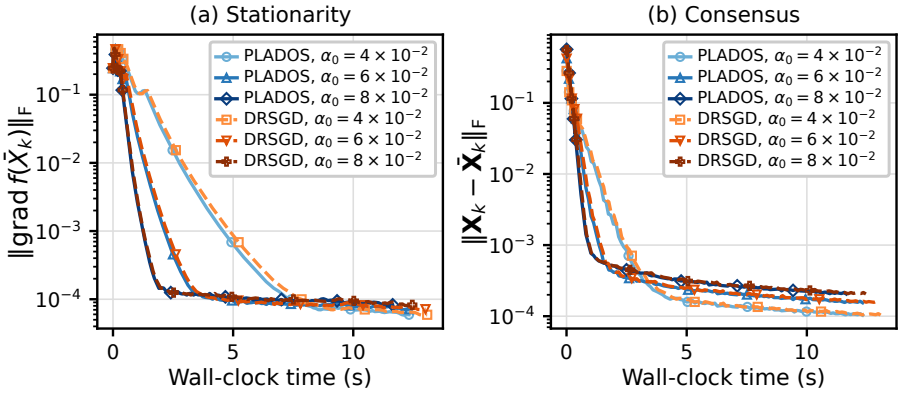



Figure 6: Stochastic matrix completion: PLADOS and DRSGD at 3 matched diminishing stepsizes. Curves show medians over 3 runs.

### D.3 Additional scalability results

Each task uses the same global dataset and initialization across node counts within a trial. MNIST is divided by 255 without centering or whitening, and equal shards follow a random permutation. PCA runs complete 300 epochs with batch size 1, giving the same total budget of 18 million sample evaluations across node counts. LRMC runs complete 800 epochs with batch size 12, giving 9.6 million sample evaluations and $K_n = 800 \cdot 12000/(12n)$ updates per node. The observation mask is also fixed across node counts.

PCA stepsizes were fixed before testing. For LRMC, the coefficients $(0.2, 0.3, 0.4)$ were chosen after an initial comparison with coefficient $0.2$ at all node counts, so the actual step approximately doubles when the node count doubles. The 8-node runs are reused, while the 16- and 32-node runs were performed in a later session on the same server.

For repetition $s$, let $t_{n,s}(k)$ denote the recorded cumulative algorithm time and let $\mathcal{K}_{n,s}$ be the saved iteration indices. For PCA, we define the first recorded time to the common accuracy threshold by

$$T_{n,s} := \min\{t_{n,s}(k) : k \in \mathcal{K}_{n,s},\ d_s(\bar{X}_k, X^\star) \le 0.01\}.$$

For LRMC, the criterion is $\|\operatorname{grad} f(\bar{X}_k)\|_F \le 10^{-3}$. All runs reach the corresponding threshold. The ratio $T_{8,s}/T_{n,s}$ is computed within each trial before taking the median and interquartile range (IQR). There is no interpolation of hitting times; diagnostics are recorded every epoch for PCA and approximately every 5 epochs for LRMC. Figure 7 shows nearly linear practical scaling for MNIST PCA and LRMC, and substantial acceleration with greater variability for synthetic PCA. The comparisons use the stated stepsize configurations.

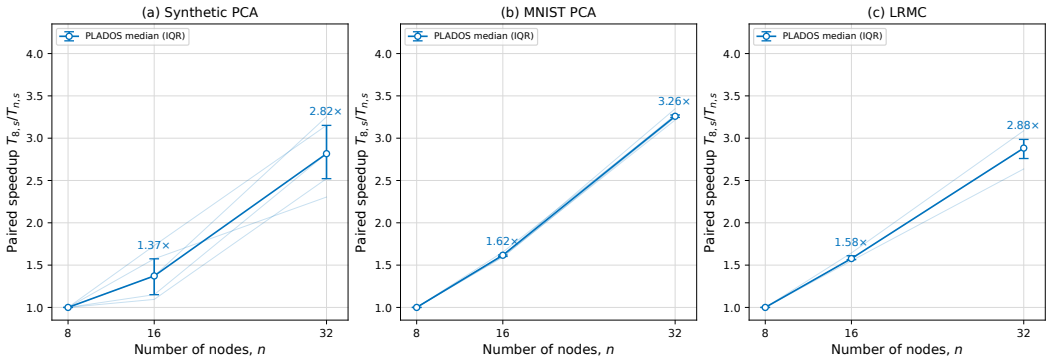



Figure 7: Paired time-to-accuracy speedup relative to 8 nodes, at $d_s \le 0.01$ for PCA and $\|\operatorname{grad} f(\bar{X}_k)\|_F \le 10^{-3}$ for LRMC. Faint lines show individual runs; thick lines and numeric labels show medians, and error bars show IQR.

### D.4 MNIST PCA with heterogeneous partitions

We adapt the label-partitioned fraction of Hsieh et al. (2020) to MNIST PCA. A fixed balanced pool contains 50000 training images, 5000 per digit. Images are vectorized in $\mathbb{R}^p$ with $p = 784$ and divided by 255 without centering; we set $r = 5$. This pool and its global empirical reference subspace are fixed across partitions, keeping the global optimization problem unchanged.

Images are randomly ordered within each digit, concatenated in label order, and split into 8 blocks of 6250. Each block is shuffled once. At level $\rho$, node $i$ retains its first $\lfloor 6250\rho \rfloor$ images; the remaining images are pooled, randomly permuted, and redistributed equally. Thus every node always has 6250 images and the global pool is preserved. We use $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$: $\rho = 0$ gives random assignment and $\rho = 1$ gives label-sorted blocks, with labels potentially shared across nodes. The retained fraction $\rho$ is a partition parameter, not the theoretical bound $\xi$; even a random finite partition can have nonzero gradient heterogeneity.

Each node caches $C_i = A_i^\top A_i/6250$ in double precision and uses the exact empirical gradient $-C_i X_{i,k}$, so $\sigma = 0$. The graph is an 8-node ring with Metropolis weights, $\gamma = 1/3$, and the same

actual stepsize $\alpha = 7 \times 10^{-4}$ for every $\rho$. This value was chosen in exploratory runs before the 15000-update comparison. Within each trial, all nodes and partition levels share the same Gaussian polar initialization. All 8 nodes are simulated in 1 process.

Each epoch consists of 1 full-gradient update. All 18 runs complete 15000 epochs, with diagnostics recorded at every update. Figure 4 uses pointwise medians without interpolation and the same 4000–10000 window for all curves and both metrics. Larger data heterogeneity generally yields larger late-run gradient norms and subspace distances. The full runs exhibit nonzero plateaus, consistent with the fixed-step residual in equation (107): for $\sigma = 0$,

$$\frac{1}{K} \sum_{k<K} \mathbb{E} \left\| \operatorname{grad} f(\bar{X}_k) \right\|_F^2 \leq \frac{C_0 \mathcal{E}_0}{\alpha K} + C_2 n \alpha^2 \xi^2.$$

This bound concerns average squared stationarity. The observed plateaus illustrate convergence to a neighborhood at the fixed stepsize.