INVERSE-FREE SECOND-ORDER MOMENTUM FOR DECENTRALIZED RIEMANNIAN STOCHASTIC OPTIMIZATION ON COMPACT SUBMANIFOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study decentralized stochastic optimization on compact embedded submanifolds and propose decentralized projected Riemannian stochastic Hessian-corrected momentum (DPRSHCM). The method pairs a correction for the full ambient variation of the Riemannian gradient with a projected iterate difference, yielding a quadratic remainder as tangent spaces change. It combines this correction with gradient tracking and metric projection, using one stochastic gradient and one Hessian–vector product per node per iteration. With a diminishing stepsize and momentum parameters independent of the iteration budget, DPRSHCM achieves $\widetilde{\mathcal{O}}(\epsilon^{-3/2})$ per-node oracle complexity for reducing both the expected squared Riemannian gradient norm and mean squared consensus error below ϵ. A constant stepsize chosen for a prescribed budget removes the logarithmic factor. With exact local gradients and stochastic Hessian–vector products, fixed parameters achieve the same joint criterion in $\mathcal{O}(\epsilon^{-1})$ iterations. Experiments on PCA and matrix completion demonstrate the method's effectiveness and the computational benefits of reusing gradient computations in matrix completion.

1 INTRODUCTION

We study decentralized stochastic optimization on a compact embedded submanifold. A network of nodes learns a shared model from local data streams using only neighbor communication:

$$\begin{aligned} \min_{x_1,\dots,x_n\in\mathcal{M}} \quad & \frac{1}{n}\sum_{i=1}^{n} f_i(x_i), \qquad f_i(x) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}\big[F_i(x;\xi_i)\big], \\ \text{s.t.} \quad & x_1 = \cdots = x_n, \end{aligned} \tag{1.1}$$

where $\mathcal{M}$ is the constraint manifold, x_i is the variable at node i, and $F_i(\cdot;\xi_i)$ is its sample loss. Samples are independent across nodes and iterations, and their distributions $\mathcal{D}_i$ may differ. Applications include streaming PCA with orthogonality constraints (Cardot & Degras, 2018; Ye & Zhang, 2021) and subspace estimation for low-rank matrix completion (Boumal & Absil, 2011; 2015). Efficient methods for these problems must control sampling noise and reach consensus with a small number of oracle calls and communication rounds per update. For continuing data streams, the iteration budget may also be unknown in advance.

Existing decentralized Riemannian methods combine local optimization with consensus while respecting the manifold constraint. Retraction-based methods on the Stiefel manifold (Chen et al., 2021) and metric-projection methods on compact embedded submanifolds (Deng & Hu, 2025a) provide two such frameworks. Gradient tracking accounts for differences between local objective gradients. Under stochastic access, methods such as DRSGD and DRSGT also have to control sampling noise (Chen et al., 2021; Zhao et al., 2026). Recursive momentum improves gradient estimation by correcting an accumulated estimate as the iterate moves. DPRSRM applies this approach on compact submanifolds using gradients at two points evaluated with the same sample (Deng & Hu, 2025b; 2026).

The two-point correction creates a computational cost when gradient evaluation requires expensive intermediate calculations that depend on the current point. For the variable-projection formulation

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.

of matrix completion studied here, each point requires a separate least-squares factorization. In addition, DPRSRM's constant-stepsize guarantee selects parameters using a prescribed iteration budget. These observations motivate a method that can reuse computations at one point and support continued optimization without selecting a new schedule for a longer run. Further related work and a comparison of oracle models and rates appear in Appendix B.

Hessian-corrected momentum offers a way to pursue this goal. It uses a gradient and a Hessian–vector product at the current point to correct the estimate (Tran & Cutkosky, 2022). For the matrix completion objective above, the Hessian–vector product reuses the factorization computed for the gradient. We investigate how to use this structure in decentralized manifold optimization, with convergence guarantees for parameters independent of the iteration budget.

However, the extension to manifolds requires accounting for changing tangent spaces. The Riemannian Hessian describes the tangent component of the gradient variation, while an ambient recursion tracks the full gradient difference. The correction must capture this variation along the actual difference between successive iterates. That difference contains movement from both optimization and neighbor mixing, so the random correction error evolves jointly with consensus and gradient-tracking errors.

Our contributions are as follows.

- **Hessian-corrected momentum for decentralized manifold optimization.** We propose decentralized projected Riemannian stochastic Hessian-corrected momentum (DPRSHCM), combining gradient tracking and metric projection with a correction that accounts for changing tangent spaces. The correction pairs the derivative of the Riemannian gradient as an ambient vector field with the iterate difference projected onto the current tangent space. This yields a quadratic remainder in the ambient norm (Lemma 5.1), using one stochastic gradient and one Hessian–vector product at the current point.
- **Convergence without a prescribed iteration budget.** We establish convergence through weighted bounds on successive iterate differences that control the random correction and Taylor errors jointly with consensus and gradient-tracking errors. A schedule independent of the iteration budget achieves $\widetilde{\mathcal{O}}(\epsilon^{-3/2})$ per-node oracle complexity for reducing both the expected squared Riemannian gradient norm and mean squared consensus error below ϵ (Theorem 6.1); a prescribed budget yields $\mathcal{O}(\epsilon^{-3/2})$ complexity with single-sample initialization (Theorem 6.2).
- **Fixed-parameter convergence with exact gradients and stochastic Hessian–vector products.** We show that exact local gradients permit a fixed stepsize and momentum parameter while retaining stochastic Hessian–vector products. The analysis uses the dependence of the correction noise on the iterate difference to control its effect, yielding $\mathcal{O}(\epsilon^{-1})$ iterations to reach the same joint criterion (Theorem 6.3).

Experiments on PCA and matrix completion examine the correction terms, parameter schedules and computational cost.

2 Related work

Metric projection and gradient tracking support decentralized optimization on compact embedded manifolds (Deng & Hu, 2025a). DPRSRM adds recursive momentum using gradients of the same sampled loss at consecutive iterates (Deng & Hu, 2025b). Its $\mathcal{O}(\epsilon^{-3/2})$ guarantee uses parameters selected for a prescribed budget; the revised analysis obtains this rate with single-sample initialization under mean-square oracle regularity (Deng & Hu, 2026).

Euclidean Hessian-corrected momentum approximates the gradient difference using an HVP at the current point (Tran & Cutkosky, 2022). On a manifold, the tangent projector also varies, and neighbor mixing contributes to the iterate difference. DPRSHCM accounts for these effects through the ambient derivative and weighted moment bounds. This supports an iteration-budget-independent schedule, the prescribed-budget rate, and fixed-parameter convergence with exact gradients and stochastic HVPs. Appendix B gives further comparisons.

3 PRELIMINARIES

3.1 NOTATION

We use lowercase letters for vectors or matrices at a single node, bold lowercase letters for network variables formed by stacking all node variables vertically, and uppercase letters for matrices or constants. The sets of real numbers and positive integers are denoted by $\mathbb{R}$ and $\mathbb{N}_+$. For matrices $X, Y \in \mathbb{R}^{d\times r}$, the Euclidean inner product and its induced norm are $\langle X, Y\rangle := \mathrm{tr}(X^\top Y)$ and $\|X\| := \sqrt{\langle X, X\rangle}$, where $\mathrm{tr}(\cdot)$ denotes the matrix trace. For network variables we also use $\|\cdot\|$ for the product Euclidean norm. The identity matrix and the all-ones vector are denoted by I and $\mathbf{1}$, with dimensions determined by context. The symbols $\otimes$ and $\mathrm{col}(\cdot)$ denote the Kronecker product and the vertical stacking operation. For any network variable $\mathbf{z} = \mathrm{col}(z_1, \ldots, z_n)$, write $\hat{z} := n^{-1}\sum_{i=1}^n z_i$ and $\widehat{\mathbf{z}} := \mathbf{1}\otimes\hat{z} = (J\otimes I)\mathbf{z}$, where $J := n^{-1}\mathbf{1}\mathbf{1}^\top$. For a network of n nodes, $W \in \mathbb{R}^{n\times n}$ denotes the symmetric doubly stochastic mixing matrix and $\mathbf{W} := W \otimes I$. Unless stated otherwise, $i \in [n] := \{1, \ldots, n\}$ is the node index and k is the iteration index. $\mathbb{E}$ denotes expectation, $\xi_{i,k}$ is the random sample drawn by node i at iteration k, and $\mathcal{F}_k$ is the σ-algebra generated by all random samples before the start of iteration k. For a smooth map Φ defined near y, $D\Phi(y)[\omega]$ denotes its directional derivative at y along ω.

3.2 COMPACT SUBMANIFOLDS AND PROXIMAL SMOOTHNESS

The feasible set $\mathcal{M}$ is a compact embedded submanifold. For any $x \in \mathcal{M}$, $T_x\mathcal{M}$ and $N_x\mathcal{M}$ denote the tangent and normal spaces of $\mathcal{M}$ at x, and $\mathcal{P}_{T_x\mathcal{M}}$ and $\mathcal{P}_{N_x\mathcal{M}}$ denote the Euclidean orthogonal projections onto them. With the metric induced by the Euclidean inner product, the Riemannian gradient of a smooth function f_i is $\mathrm{grad}\, f_i(x) = \mathcal{P}_{T_x\mathcal{M}}\nabla f_i(x)$, where ∇f_i is the Euclidean gradient. We use $\nabla^2 f_i$ and $\mathrm{Hess}\, f_i$ for the Euclidean and Riemannian Hessians of f_i.

For any $x \in \mathbb{R}^{d\times r}$, the distance from x to $\mathcal{M}$ and the metric projection onto $\mathcal{M}$ are

$$\mathrm{dist}(x, \mathcal{M}) := \inf_{y\in\mathcal{M}} \|y - x\|, \qquad \mathcal{P}_{\mathcal{M}}(x) := \arg\min_{y\in\mathcal{M}} \|y - x\|.$$

For a network variable $\mathbf{x}$ we use the product projection $\mathcal{P}_{\mathcal{M}^n}(\mathbf{x}) := \mathrm{col}(\mathcal{P}_{\mathcal{M}}(x_1), \ldots, \mathcal{P}_{\mathcal{M}}(x_n))$. In general $\mathcal{P}_{\mathcal{M}}$ is set-valued. Given a constant $R > 0$, the open tubular neighborhood of radius R around $\mathcal{M}$ is $U_{\mathcal{M}}(R) := \{x : \mathrm{dist}(x, \mathcal{M}) < R\}$.

Definition 3.1 (R-proximally smooth set, Clarke et al., 1995)**.** If the projection $\mathcal{P}_{\mathcal{M}}(x)$ is single-valued for every $x \in U_{\mathcal{M}}(R)$, then the closed set $\mathcal{M}$ is called an R-proximally smooth set.

Proximal smoothness controls metric projection onto the nonconvex manifold constraint. It makes $\mathcal{P}_{\mathcal{M}}$ Lipschitz continuous on every closed tubular neighborhood of radius less than R and gives the normal inequality. Both properties are recalled in Appendix C. All the iterates analyzed below stay in such a neighborhood. For example, the Stiefel manifold is a 1-proximally smooth set (Deng & Hu, 2025a).

3.3 STATIONARY POINTS

Let $x_1, \ldots, x_n \in \mathcal{M}$ denote the local variables at the nodes and $\mathbf{x} := \mathrm{col}(x_1, \ldots, x_n)$. The Euclidean average $\hat{x} = \frac{1}{n}\sum_{i=1}^n x_i$ need not lie on $\mathcal{M}$. For $\{x_i\}_{i=1}^n \subset \mathcal{M}$, we have

$$\mathcal{P}_{\mathcal{M}}(\hat{x}) = \arg\min_{y\in\mathcal{M}} \sum_{i=1}^n \|y - x_i\|^2,$$

and any element $\bar{x}$ of $\mathcal{P}_{\mathcal{M}}(\hat{x})$ is called an induced arithmetic mean of $\{x_i\}_{i=1}^n$ on the manifold $\mathcal{M}$ (Sarlette & Sepulchre, 2009; Chen et al., 2021). We write $\bar{\mathbf{x}} := \mathbf{1}\otimes\bar{x}$ and define the global objective function $F(z) := \frac{1}{n}\sum_{i=1}^n f_i(z)$.

Definition 3.2 (Stochastic ϵ-stationary point, Chen et al., 2021; Deng & Hu, 2025b)**.** If there exists $\bar{x} \in \mathcal{P}_{\mathcal{M}}(\hat{x})$ such that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|x_i - \bar{x}\|^2] \le \epsilon, \qquad \mathbb{E}[\|\mathrm{grad}\, F(\bar{x})\|^2] \le \epsilon,$$

then the point set $\{x_1, x_2, \ldots, x_n\} \subset \mathcal{M}$ is called a stochastic ϵ-stationary point of the decentralized optimization problem.

The two quantities are called the *consensus error* and the *stationarity error*, respectively.

4 Problem and assumptions

Assumption 4.1 (Manifold and objective smoothness, Deng & Hu, 2025a;b)**.** *The set $\mathcal{M}$ is a compact, smooth (C^∞), embedded, and R-proximally smooth submanifold. Each f_i is twice continuously differentiable on a neighborhood of* $\operatorname{conv}(\mathcal{M})$*, and there exists $L_f > 0$ such that*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L_f \|x - y\|, \qquad x, y \in \operatorname{conv}(\mathcal{M}), \quad i \in [n]. \tag{4.1}$$

The following lemma collects the consequences of Assumption 4.1 needed in our analysis.

Lemma 4.2 (see Deng & Hu (2025a, Section 2))**.** *Under Assumption 4.1, there exist constants $L_g, L_G, L_2 > 0$ such that, for every $i \in [n]$, $x, y \in \mathcal{M}$, and every ambient vector z,*

$$f_i(y) \le f_i(x) + \langle \operatorname{grad} f_i(x), y - x \rangle + \frac{L_g}{2} \|y - x\|^2, \tag{4.2}$$

$$\|\operatorname{grad} f_i(y) - \operatorname{grad} f_i(x)\| \le L_G \|x - y\|, \tag{4.3}$$

$$\left\|\mathcal{P}_{T_x\mathcal{M}} z - \mathcal{P}_{T_y\mathcal{M}} z\right\| \le L_2 \|x - y\| \|z\|. \tag{4.4}$$

By compactness of $\mathcal{M}$ and Assumption 4.1, the constants $G := \max_{i\in[n],\, x\in\mathcal{M}} \|\operatorname{grad} f_i(x)\|$, $G_\nabla := \max_{i\in[n],\, x\in\mathcal{M}} \|\nabla f_i(x)\|$, $M_H := \max_{i\in[n],\, x\in\mathcal{M}} \left\|\nabla^2 f_i(x)\right\|_{\mathrm{op}}$ and $F_* := \min_{x\in\mathcal{M}} F(x)$ are finite, where $\|\cdot\|_{\mathrm{op}}$ is the operator norm. We write $L := \max\{L_g, L_G, G, 1\}$.

Assumption 4.3 (Second-order smoothness, Tran & Cutkosky, 2022)**.** *For each i there exists $\rho_i > 0$ such that, for all $x, y \in \operatorname{conv}(\mathcal{M})$,*

$$\left\|\nabla f_i(x) + \nabla^2 f_i(y)(y - x) - \nabla f_i(y)\right\| \le \frac{\rho_i}{2} \|x - y\|^2. \tag{4.5}$$

We write $\rho := \max_{1\le i\le n} \rho_i$.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the undirected communication graph, where $\mathcal{V} = [n] := \{1, \ldots, n\}$ is the set of agents and $\mathcal{E}$ is the set of communication edges. Let $W \in \mathbb{R}^{n\times n}$ be the mixing matrix associated with $\mathcal{G}$, where $W_{ij} > 0$ if $i = j$ or $(i, j) \in \mathcal{E}$, and $W_{ij} = 0$ otherwise. For the Euclidean consensus penalty

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

write $\nabla\phi_i(\mathbf{x}) := x_i - \sum_{j=1}^{n} W_{ij} x_j$, $\operatorname{grad}\phi_i(\mathbf{x}) := \mathcal{P}_{T_{x_i}\mathcal{M}}(\nabla\phi_i(\mathbf{x}))$.

Assumption 4.4 (Communication network, Chen et al., 2021; Deng & Hu, 2025a)**.** *We assume that the undirected graph $\mathcal{G}$ is connected and W is symmetric and doubly stochastic, i.e.,* (i) $W = W^\top$ *and* $W\mathbf{1} = \mathbf{1}$*;* (ii) $W_{ij} \ge 0$ *and* $1 > W_{ii} > 0$ *for all i, j;* (iii) *the eigenvalues of W lie in* $(-1, 1]$.

The network contraction property implied by Assumption 4.4 is recalled in Appendix E.

Assumption 4.5 (Stochastic oracle, Deng & Hu, 2025b; Tran & Cutkosky, 2022)**.** *Let $\nabla F_i(x; \xi)$ and $\nabla^2 F_i(x; \xi)[u]$ denote the gradient and the Hessian–vector product of the single-sample loss with respect to x. For any $\mathcal{F}_k$-measurable $x \in \mathcal{M}$ and any $\mathcal{F}_k$-measurable tangent vector $u \in T_x\mathcal{M}$, node i can sample $\xi_{i,k}$ and compute $\nabla F_i(x; \xi_{i,k})$ and $\nabla^2 F_i(x; \xi_{i,k})[u]$. The samples are independent across nodes and iterations, and*

$$\mathbb{E}[\nabla F_i(x; \xi_{i,k}) \mid \mathcal{F}_k] = \nabla f_i(x), \tag{4.6}$$

$$\mathbb{E}[\|\nabla F_i(x; \xi_{i,k}) - \nabla f_i(x)\|^2 \mid \mathcal{F}_k] \le \nu_i^2, \tag{4.7}$$

$$\mathbb{E}[\nabla^2 F_i(x; \xi_{i,k})[u] \mid \mathcal{F}_k] = \nabla^2 f_i(x)[u], \tag{4.8}$$

$$\mathbb{E}[\left\|\nabla^2 F_i(x; \xi_{i,k})[u] - \nabla^2 f_i(x)[u]\right\|^2 \mid \mathcal{F}_k] \le \sigma_{\nabla^2,i}^2 \|u\|^2. \tag{4.9}$$

We use the projected stochastic gradient $\operatorname{grad} F_i(x; \xi_{i,k}) := \mathcal{P}_{T_x\mathcal{M}} \nabla F_i(x; \xi_{i,k})$.

5 DECENTRALIZED PROJECTED RIEMANNIAN STOCHASTIC HESSIAN-CORRECTED MOMENTUM METHOD

We construct a momentum correction for the gradient change along the actual decentralized update. This requires accounting for both changing tangent spaces and the iterate difference produced by optimization and neighbor mixing. Differentiating $\operatorname{grad} f_i = \mathcal{P}_{T\cdot\mathcal{M}}\nabla f_i$ along $\omega \in T_y\mathcal{M}$ gives the correction operator

$$\overline{\operatorname{Hess}}\, f_i(y)[\omega] := D(\mathcal{P}_{T\cdot\mathcal{M}})(y)[\omega]\nabla f_i(y) + \mathcal{P}_{T_y\mathcal{M}}\nabla^2 f_i(y)[\omega], \tag{5.1}$$

where $D(\mathcal{P}_{T\cdot\mathcal{M}})(y)[\omega]$ is the derivative of the tangent projector along ω. Each node estimates both terms using the same sample at the current point:

$$\overline{\operatorname{Hess}}\, F_i(y;\xi)[\omega] := D(\mathcal{P}_{T\cdot\mathcal{M}})(y)[\omega]\nabla F_i(y;\xi) + \mathcal{P}_{T_y\mathcal{M}}\nabla^2 F_i(y;\xi)[\omega]. \tag{5.2}$$

The operator has a tangent input and an ambient output. Its tangent part is the Riemannian Hessian, $\operatorname{Hess} f_i(y)[\omega] = \mathcal{P}_{T_y\mathcal{M}}\,\overline{\operatorname{Hess}}\, f_i(y)[\omega]$ (Boumal, 2023, Corollary 5.16). Both output components enter the ambient gradient difference. Lemma D.1 gives conditional unbiasedness and the variance bound $\bar{\sigma}_{H,i}^2 \|\omega\|^2$.

Each update combines neighbor mixing, a tracked gradient step and metric projection. The correction uses the resulting iterate difference, projected onto the current tangent space:

$$u_{i,k} := \mathcal{P}_{T_{x_{i,k}}\mathcal{M}}(x_{i,k} - x_{i,k-1}), \tag{5.3}$$

For $\Delta = x_{i,k} - x_{i,k-1}$, proximal smoothness and orthogonal projection give

$$\|\Delta - u_{i,k}\| \le \frac{\|\Delta\|^2}{2R}, \qquad \|u_{i,k}\| \le \|\Delta\|.$$

The projection removes a second-order normal component of the input, while (5.2) retains the first-order normal variation of the gradient. The norm bound controls the correction variance by $\|\Delta\|^2$. The following lemma establishes the quadratic Taylor remainder (proof in Appendix D).

Lemma 5.1 (Second-order expansion of the Riemannian gradient)**.** *Suppose Assumptions 4.1 and 4.3 hold. Then, for every i and for all $x, y \in \mathcal{M}$,*

$$\left\|\operatorname{grad} f_i(x) + \overline{\operatorname{Hess}}\, f_i(y)[\mathcal{P}_{T_y\mathcal{M}}(y-x)] - \operatorname{grad} f_i(y)\right\| \le L_H \|y-x\|^2, \tag{5.4}$$

where L_H is the constant (D.6) *of Appendix D. It depends only on $\mathcal{M}$, R and the constants of Section 4.*

To display the normal variation explicitly, write $g = \operatorname{grad} f_i$, $P_y = \mathcal{P}_{T_y\mathcal{M}}$ and $u = P_y(y-x)$. Lemma 5.1 gives

$$g(y) - g(x) = \underbrace{\operatorname{Hess} f_i(y)[u]}_{\text{tangent variation}} + \underbrace{(I - P_y)Dg(y)[u]}_{\text{normal variation}} + O(\|y-x\|^2). \tag{5.5}$$

For example, on the unit sphere with $f(x) = a^\top x$ and $u \in T_x\mathcal{M}$,

$$D(\operatorname{grad} f)(x)[u] = -(a^\top u)x - (a^\top x)u, \qquad \operatorname{Hess} f(x)[u] = -(a^\top x)u.$$

At a point with $a^\top x = 0$, the Hessian vanishes while the normal gradient variation $-(a^\top u)x$ can be nonzero. Together, the full derivative and the projected iterate difference give a quadratic gradient-difference remainder and correction variance bounded by the squared iterate difference (Lemma D.1). These bounds support the Hessian-corrected recursion of Tran & Cutkosky (2022).

The estimator at node i is, for $k \ge 1$,

$$q_{i,k} = (1-\tau_{k-1})\left(d_{i,k-1} + \overline{\operatorname{Hess}}\, F_i(x_{i,k};\xi_{i,k})[u_{i,k}]\right) + \tau_{k-1}\operatorname{grad} F_i(x_{i,k};\xi_{i,k}), \tag{5.6}$$

where $\tau_k \in (0,1]$ is a deterministic momentum parameter common to all nodes. Both oracle evaluations use the current point, enabling the factorization reuse in matrix completion discussed in Section 1 and measured in Section 7. We clip the direction to keep it bounded. This keeps the iterates in a neighborhood of $\mathcal{M}$ on which the metric projection is single-valued (Lemma E.6 in Appendix E):

$$d_{i,k} := \operatorname{clip}_{L_d}(q_{i,k}) := \begin{cases} q_{i,k}, & \|q_{i,k}\| \le L_d, \\ L_d q_{i,k}/\|q_{i,k}\|, & \|q_{i,k}\| > L_d. \end{cases} \tag{5.7}$$

Algorithm 1 The DPRSHCM for solving (1.1)

Input: Initial point $\mathbf{x}_0 = \bar{\mathbf{x}}_0 \in \mathcal{M}^n$, $\mathbf{s}_{-1} = \mathbf{d}_{-1} = \mathbf{0}$, stepsizes $\{\alpha_k\}_{k\geq 0}$, momentum parameters $\{\tau_k\}_{k\geq 0} \subset (0, 1]$, clipping radius L_d, horizon K.

1: Sample $\xi_{i,0}$, let $q_{i,0} = \operatorname{grad} F_i(x_{i,0}; \xi_{i,0})$ and $s_{i,0} = d_{i,0} = \operatorname{clip}_{L_d}(q_{i,0})$.
2: $v_{i,0} = \mathcal{P}_{T_{x_{i,0}}\mathcal{M}}(s_{i,0})$.
3: $x_{i,1} = \mathcal{P}_{\mathcal{M}}(\sum_{j=1}^n W_{ij} x_{j,0} - \alpha_0 v_{i,0})$.
4: **for** $k = 1, 2, \ldots, K-1$ **do**
5: Update the projected iterate difference $u_{i,k}$ via (5.3).
6: Update Hessian-corrected gradient estimator $q_{i,k}$ via (5.6).
7: Update the clipped gradient estimator $d_{i,k}$ via (5.7).
8: Update Riemannian gradient tracking $s_{i,k}$ via (5.8).
9: Project onto tangent space: $v_{i,k} = \mathcal{P}_{T_{x_{i,k}}\mathcal{M}}(s_{i,k})$.
10: Update new iterate $x_{i,k+1}$ via (5.10).
11: **end for**

Output: $\{x_{i,\hat{k}}\}_{i=1}^n$, with $\hat{k}$ independent of the oracle samples and uniform on $\mathcal{I}_K$: $\mathcal{I}_K = \{\lceil K/2 \rceil, \ldots, K-1\}$ in Theorems 6.1–6.2 ($K \geq 2$), and $\mathcal{I}_K = \{0, \ldots, K-1\}$ in Theorem 6.3 ($K \geq 1$).

Assumption 5.2 (Clipping radius)**.** *The clipping radius L_d satisfies $L_d \geq G$.*

Since $\mathcal{M}$ is compact, such a bound exists under the smoothness assumptions. For the PCA objective $\|\operatorname{grad} f_i(X)\| \leq \|A_i\|^2$. Section 7 reports the sensitivity of the method to L_d. The direction $d_{i,k}$ enters the decentralized projection framework of Deng & Hu (2025a). Each node exchanges its local variable and its tracking variable with its neighbors through the mixing matrix, takes a descent step along the tangent projection of the tracked direction, and maps the result back to the manifold by the metric projection:

$$s_{i,k} = \sum_{j=1}^n W_{ij} s_{j,k-1} + d_{i,k} - d_{i,k-1}, \tag{5.8}$$

$$v_{i,k} = \mathcal{P}_{T_{x_{i,k}}\mathcal{M}}(s_{i,k}), \tag{5.9}$$

$$x_{i,k+1} = \mathcal{P}_{\mathcal{M}}\left(\sum_{j=1}^n W_{ij} x_{j,k} - \alpha_k v_{i,k}\right). \tag{5.10}$$

The update uses the metric projection directly and requires no explicit vector transport. The complete method is summarized in Algorithm 1.

6 Convergence analysis

The correction bounds of Section 5 depend on iterate differences, which reflect both optimization and neighbor mixing. We therefore analyze gradient estimation, consensus and gradient tracking jointly. Let e_k denote the mean-square error of the stacked gradient estimates and let $\Delta\mathbf{x}_k$ stack the successive iterate differences $\Delta x_{i,k} = x_{i,k+1} - x_{i,k}$. Lemma E.12 gives

$$e_{k+1} \leq (1-\tau_k) e_k + \underbrace{C_H \mathbb{E}\,\|\Delta\mathbf{x}_k\|^2}_{\text{correction noise}} + \underbrace{\frac{C_R}{\tau_k} \mathbb{E} \sum_i \|\Delta x_{i,k}\|^4}_{\text{squared Taylor remainder}} + C_g \tau_k^2, \tag{6.1}$$

where $C_H = 2\sigma_H^2$, $C_R = L_H^2$ and $C_g = 2n\nu^2$, with $\nu^2 := \max_i \nu_i^2$ and $\sigma_H^2 := \max_i \bar{\sigma}_{H,i}^2$ from Lemma D.1. The correction variance gives the second-moment term, while squaring the Taylor remainder in Lemma 5.1 gives the fourth-moment term. Iterate differences measure movement across iterations; consensus error measures disagreement among nodes. Lemma E.8 relates them by bounding the iterate difference through consensus disagreement and the search directions. Lemma E.15

then controls weighted second and fourth moments of these differences by accumulated squared search directions. This lets descent absorb the correction noise and Taylor errors jointly with consensus and tracking errors. Proofs and constants appear in Appendix E and Table 3.

6.1 Diminishing stepsizes

For the continuing-stream setting of Section 1, we fix a schedule that does not require a prescribed iteration budget: $\alpha_k = \alpha_0/(k+1)^{1/3}$ and $\tau_k = 2K_H\alpha_k\alpha_{k+1}$. The parameters δ, K_H and α_0 are chosen, in this order, according to (E.45), independently of K.

Theorem 6.1 (Diminishing-stepsize convergence)**.** *Suppose Assumptions 4.1–5.2 hold, and initialize Algorithm 1 with $x_{1,0} = \cdots = x_{n,0}$. Choose δ, K_H and α_0, in this order, according to* (E.45)*, and use the stepsizes and momentum parameters* (E.47)*. Then, for every integer $K \geq 2$, the output index $\hat{k}$, uniform on $\{\lceil K/2\rceil, \dots, K-1\}$ and independent of the oracle samples, satisfies*

$$\mathbb{E}\left\|\operatorname{grad} F(\bar{x}_{\hat{k}})\right\|^2 \leq \frac{180}{7\alpha_0 K^{2/3}}\left[\mathcal{C}_0 + \mathcal{C}_3\alpha_0^3(1+\log K) + 4\mathcal{G}_4(C_x^{\dim})\alpha_0^4 + \frac{5}{2}\mathcal{C}_5\alpha_0^5\right], \tag{6.2}$$

and

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{x}_{\hat{k}} - \bar{\mathbf{x}}_{\hat{k}}\right\|^2 \leq \frac{2^{2/3}C_x^{\dim}\alpha_0^2}{K^{2/3}}, \tag{6.3}$$

where the constants $\mathcal{C}_0$, $\mathcal{C}_3$, $\mathcal{C}_5$ and $\mathcal{G}_4(C_x^{\dim})$ do not depend on K and are given in (E.61) *in Appendix E. Algorithm 1 uses one stochastic gradient and one stochastic Hessian–vector product per node and iteration, so both oracle complexities per node are $\widetilde{\mathcal{O}}(\epsilon^{-3/2})$ for the criterion in Definition 3.2. Here $\widetilde{\mathcal{O}}$ hides the factor $(1+\log(1/\epsilon))^{3/2}$.*

The same schedule therefore applies for every $K \geq 2$: a run can be extended without changing its update parameters. The theorem gives joint stationarity and consensus guarantees with stochastic gradients and Hessian–vector products evaluated at the current local iterate.

6.2 Constant stepsize

When the budget is known, a constant stepsize removes the logarithmic factor in Theorem 6.1. We set $\alpha_k \equiv \alpha$ and $\tau_k \equiv \tau := 2K_H\alpha^2$, choosing δ, then $K_H \geq K_H^{\mathrm{con}}$, then $\alpha \leq \bar{\alpha}$ according to (E.65). These thresholds are independent of K; the stepsize chosen below depends on the budget.

Theorem 6.2 (Constant-stepsize complexity)**.** *Suppose the assumptions and initialization conditions of Lemma E.19 hold. Fix $K_H \geq K_H^{\mathrm{con}}$ and a scale $a \in (0, \bar{\alpha}]$, with $\bar{\alpha}$ given in* (E.65)*. None of K_H, a and $\bar{\alpha}$ depends on K. For a prescribed horizon $K \geq 2$, set*

$$\alpha_k \equiv \alpha := \frac{a}{K^{1/3}}, \qquad \tau_k \equiv \tau := 2K_H\alpha^2. \tag{6.4}$$

Then

$$\begin{gathered}\mathbb{E}\left[\left\|\operatorname{grad} F(\bar{x}_{\hat{k}})\right\|^2 + \frac{1}{n}\left\|\mathbf{x}_{\hat{k}} - \bar{\mathbf{x}}_{\hat{k}}\right\|^2\right] \leq \frac{\mathcal{C}_{\mathrm{opt}}}{K^{2/3}}, \\ \mathcal{C}_{\mathrm{opt}} := \frac{12(\mathcal{B} + C_\Phi a^3)}{a} + C_x^{\mathrm{con}}a^2,\end{gathered} \tag{6.5}$$

where $\hat{k}$ is the independent uniform output index in Algorithm 1 and $\mathcal{B}$ is given in (E.76)*. Consequently, the stochastic-gradient and Hessian–vector-product oracle complexities per node are $O(\epsilon^{-3/2})$ for the criterion in Definition 3.2.*

The estimator-error decrease in the coupled descent inequality bounds the Lyapunov function at an iterate in the first half of the run. This yields the stated complexity with a single initial sample.

6.3 Constant parameters and exact gradients

The correction noise associated with successive iterate differences in (6.1) also permits fixed parameters when local gradients are exact. We keep $\alpha_k \equiv \alpha$ and $\tau_k \equiv \tau$ independent of the budget, subject to $\alpha \leq \bar{\alpha}_{\tau,\delta}$ in (E.86). For exact local gradients, node i uses (E.83), with $\nu = 0$ and the stochastic Hessian–vector oracle (E.82).

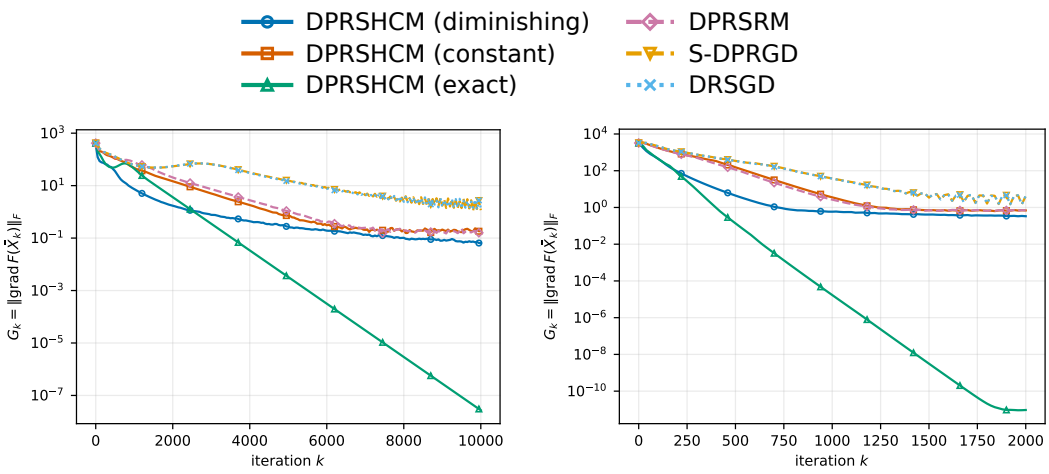

Figure 1: Synthetic PCA: gradient norm, medians over ten streams. Consensus and objective: Figure 3.

Figure 2: Synthetic LRMC: gradient norm, medians over ten instances. Consensus and objective: Figure 4.

Theorem 6.3 (Constant parameters)**.** *Suppose Assumptions 4.1–5.2 hold, and run Algorithm 1 with* $\alpha_k \equiv \alpha$, $\tau_k \equiv \tau$, $x_{1,0} = \cdots = x_{n,0}$ *and* $\mathbf{s}_0 = \mathbf{d}_0$. *Let* δ *satisfy* (E.3), $\tau \in (0, 1]$ *and* $\alpha \in (0, \bar{\alpha}_{\tau,\delta}]$. *Then, for every* $K \geq 1$,

$$\mathbb{E}\left[\left\|\text{grad}\, F(\bar{x}_{\hat{k}})\right\|^2 + \frac{1}{n}\left\|\mathbf{x}_{\hat{k}} - \bar{\mathbf{x}}_{\hat{k}}\right\|^2\right] \leq \frac{\mathcal{C}_{\text{ex}}}{K} + \mathcal{C}_{\text{fl}}\,\tau\nu^2,$$

$$\mathcal{C}_{\text{ex}} := \left(\Delta_0 + \frac{3\alpha\Theta_0}{2n} + \frac{3(1+C_W)\alpha\nu^2}{2\tau}\right)\left(\frac{4}{\alpha} + \frac{16C_c\alpha}{3}\right), \quad \mathcal{C}_{\text{fl}} := (3+4C_W)\left(4 + \frac{16C_c}{3}\right). \tag{6.6}$$

Here $\hat{k}$ *is independent of the oracle samples and uniform on* $\{0, \ldots, K-1\}$. *The bound holds with* $\nu = 0$ *for exact local gradients and the oracle* (E.82). *An* ϵ*-stationary point in the sense of Definition 3.2 is then reached within* $O(\epsilon^{-1})$ *iterations, each using one exact local gradient and one stochastic Hessian–vector product per node.*

The two noise sources have distinct effects. Gradient noise contributes the residual $\mathcal{C}_{\text{fl}}\tau\nu^2$, which vanishes for exact gradients. Hessian noise is multiplied by the iterate difference, as in (6.1), and is controlled through the admissible stepsize. Thus stochastic Hessian–vector products are compatible with the $1/K$ joint rate of deterministic projected gradient tracking (Deng & Hu, 2025a), using fixed parameters and exact local gradients.

7 Numerical Experiments

We evaluate DPRSHCM on decentralized PCA and low-rank matrix completion (LRMC), comparing it with DPRSRM (Deng & Hu, 2025b), S-DPRGD (Deng & Hu, 2025a) and DRSGD (Chen et al., 2021). S-DPRGD uses stochastic gradients in the projected method of Deng & Hu (2025a). The diminishing, constant and exact-gradient settings of DPRSHCM correspond to Theorems 6.1–6.3.

For $\bar{X}_k = \mathcal{P}_{\mathcal{M}}(n^{-1}\sum_i X_{i,k})$, we measure the gradient norm $G_k = \left\|\text{grad}\, F(\bar{X}_k)\right\|_F$ and consensus error $c_k = (n^{-1}\sum_i \left\|X_{i,k} - \bar{X}_k\right\|_F^2)^{1/2}$. Their squared sum corresponds to the joint criterion in Definition 3.2. Reported means average this sum across runs. Appendix F.3 gives data generation, sampling and tuning details.

7.1 DECENTRALIZED PRINCIPAL COMPONENT ANALYSIS

We solve (A.1) with $n = 8$ nodes, 1000 samples per node, $d = 10$ and $r = 5$, using the ER graph with $p = 0.6$ and $K = 10000$. In Figure 1, the diminishing setting reaches the smallest gradient norm among the stochastic methods. Its mean joint residual is 4.2×10^{-3}, compared with 0.037 for DPRSRM (Table 4).

Extending a run without retuning With the same $(\alpha_0, \tau_0) = (3.4 \times 10^{-4}, 0.1)$, the diminishing setting reduces the mean objective gap from 3.4×10^{-3} at $K = 2500$ to 3.0×10^{-7} at $K = 40000$. At $K = 40000$, the constant setting and DPRSRM, using parameters calibrated for $K = 10000$, reach 3.3×10^{-5} and 3.1×10^{-5}, respectively. Table 1 shows continued improvement in gradient norm and consensus as the same diminishing schedule is extended.

Table 1: Terminal measures under increasing horizons without retuning; mean $\pm$ standard deviation over five sampling streams. The diminishing setting keeps its tuned (α_0, τ_0); the last two rows apply the constant parameters calibrated for $K = 10000$ to $K = 40000$.

Method	K	Objective gap	G_k	c_k
DPRSHCM (dim.)	2500	$(3.4\pm2.6)\times10^{-3}$	1.11 ± 0.13	$(9.7\pm0.4)\times10^{-6}$
	5000	$(8.9\pm3.2)\times10^{-5}$	0.269 ± 0.037	$(5.0\pm0.5)\times10^{-6}$
	10000	$(4.5\pm1.9)\times10^{-6}$	0.0643 ± 0.016	$(2.3\pm0.2)\times10^{-6}$
	20000	$(8.0\pm1.7)\times10^{-7}$	0.0279 ± 0.0033	$(1.3\pm0.2)\times10^{-6}$
	40000	$(3.0\pm1.3)\times10^{-7}$	0.0194 ± 0.0049	$(5.6\pm0.9)\times10^{-7}$
DPRSHCM (const.)	40000	$(3.3\pm1.2)\times10^{-5}$	0.195 ± 0.04	$(1.4\pm0.1)\times10^{-5}$
DPRSRM	40000	$(3.1\pm1.1)\times10^{-5}$	0.204 ± 0.04	$(9.1\pm1.6)\times10^{-6}$

7.2 DECENTRALIZED LOW-RANK MATRIX COMPLETION

We solve the unregularized variable-projection model (A.2), which is smooth locally where the observed-row least-squares matrices have full column rank. Ten noisy 50×960 instances with rank-ten signals and observation probability $1/2$ are distributed over eight nodes on a ring, with $K = 2000$. The diminishing setting has the smallest joint residual among stochastic methods on every instance, averaging 0.11 versus 0.50 for DPRSRM (Figure 2 and Table 5). Appendices F.5 and F.6 report real-data and heterogeneous-partition results.

Correction and cost The component ablation in Appendix F.4 shows that removing either correction term increases the PCA residual. On LRMC, the projected HVP alone gives similar residuals, in agreement with the normal direction of the projector-derivative term for this objective. Reusing the least-squares factorization reduces the LRMC correction cost: the constant configuration reaches comparable accuracy to DPRSRM with about 30% less running time (Appendix F.8).

8 CONCLUSIONS AND LIMITATIONS

DPRSHCM combines an ambient gradient correction with decentralized tracking and metric projection. It achieves $\widetilde{\mathcal{O}}(\epsilon^{-3/2})$ oracle complexity with a horizon-independent diminishing schedule, $\mathcal{O}(\epsilon^{-3/2})$ with a horizon-dependent constant stepsize, and $\mathcal{O}(\epsilon^{-1})$ iterations with exact local gradients. The experiments show continued improvement when the diminishing schedule is extended without retuning, lower terminal errors on PCA with random partitions and synthetic LRMC, and lower computational cost for the constant configuration on synthetic LRMC. The analysis uses worst-case local smoothness and oracle bounds, while the admissible stepsize also depends on proximal smoothness. Sharpening these dependencies is a direction for future work.

AI USE STATEMENT

We used generative AI tools to assist with proof writing and checking, implementation, and manuscript revision. For Theorem 6.2, AI assistance helped revise the proof from a large initial batch to a single initial sample. The authors developed the displacement estimate in Lemma E.8; AI tools assisted with checking this argument and the other proofs. The six algorithm implementations and the experiment scripts were written with AI assistance. AI tools also assisted with reference formatting, language editing, conference formatting, and feedback on the analysis and experimental methodology. During revision, they assisted with interpreting the reported results and checking their consistency with the claims and proofs.

Generative AI was not used to develop the theoretical model or conceptual framework, formulate the mathematical claims, propose or refine hypotheses, design the experiments, or generate data. Synthetic instances were sampled from the stated distributions. Translation, dataset cleaning and reformatting, and qualitative or thematic data analysis are not applicable. The authors checked the proofs line by line and checked the implementations against the update rules and original baseline descriptions. The authors take responsibility for the final text, claims, code and results.

REPRODUCIBILITY STATEMENT

The problem and oracle assumptions are stated in Section 4, and the clipping condition in Section 5. Complete proofs of Lemmas 5.1 and D.1 and of Theorems 6.1, 6.2 and 6.3 are given in Appendices D and E, and the constants are listed in Table 3. The experimental setup, the parameter search and the tuned parameters of every method are described in Section 7 and Appendix F.3.

REFERENCES

R. P. Boas, Jr. and C. O. Imoru. Elementary convolution inequalities. *SIAM Journal on Mathematical Analysis*, 6(3):457–471, 1975. doi: 10.1137/0506040.

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

Nicolas Boumal and Pierre-Antoine Absil. RTRMC: A Riemannian trust-region method for low-rank matrix completion. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/37bc2f75bf1bcfe8450a1a41c200364c-Abstract.html.

Nicolas Boumal and Pierre-Antoine Absil. Low-rank matrix completion via preconditioned optimization on the Grassmann manifold. *Linear Algebra and its Applications*, 475:200–239, 2015. doi: 10.1016/j.laa.2015.02.027.

Hervé Cardot and David Degras. Online principal component analysis in high dimension: Which algorithm to choose? *International Statistical Review*, 86(1):29–50, 2018. doi: 10.1111/insr.12220.

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. URL https://proceedings.mlr.press/v139/chen21g.html.

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

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

Kangkang Deng and Jiang Hu. Decentralized projected Riemannian stochastic recursive momentum method for nonconvex optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11):11203–11211, 2025b. doi: 10.1609/aaai.v39i11.33218.

Kangkang Deng and Jiang Hu. Decentralized projected Riemannian stochastic recursive momentum method for smooth optimization on compact submanifolds, 2026. URL https://arxiv.org/abs/2412.02382v3. Revised analysis, arXiv:2412.02382v3.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015. doi: 10.1145/2827872.

Jiang Hu and Kangkang Deng. Improving the communication in decentralized manifold optimization through single-step consensus and compression. *arXiv preprint arXiv:2407.08904*, 2024. doi: 10.48550/arXiv.2407.08904.

Zhenwei Huang, Wen Huang, Pratik Jawanpuria, and Bamdev Mishra. Federated learning on Riemannian manifolds with differential privacy. *Machine Learning*, 115(4):84, 2026a. doi: 10.1007/s10994-026-07018-w. URL https://arxiv.org/abs/2404.10029.

Zhenwei Huang, Wen Huang, Pratik Jawanpuria, and Bamdev Mishra. Riemannian federated learning via averaging gradient streams. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=oEtrDiFOFF.

Danqi Jin, Yitong Chen, Jie Chen, and Wen Zhang. Zeroth-order distributed stochastic optimization over Riemannian manifolds. *IEEE Signal Processing Letters*, 32:4004–4008, 2025. doi: 10.1109/LSP.2025.3620779.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263.

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.

Jiaxiang Li and Shiqian Ma. Federated learning on Riemannian manifolds. *Applied Set-Valued Analysis and Optimization*, 5(2):213–232, 2023. doi: 10.23952/asvao.5.2023.2.07. URL https://arxiv.org/abs/2206.05668.

Bamdev Mishra, Hiroyuki Kasai, Pratik Jawanpuria, and Atul Saroop. A Riemannian gossip approach to subspace learning on Grassmann manifold. *Machine Learning*, 108:1783–1803, 2019. doi: 10.1007/s10994-018-05775-x. URL https://arxiv.org/abs/1705.00467.

Duc Toan Nguyen and César A. Uribe. Intrinsic decentralized stochastic Riemannian optimization on manifolds with bounded sectional curvature. *IEEE Control Systems Letters*, 10:1345–1350, 2026. doi: 10.1109/LCSYS.2026.3706455.

Alain Sarlette and Rodolphe Sepulchre. Consensus optimization on manifolds. *SIAM Journal on Control and Optimization*, 48(1):56–76, 2009. doi: 10.1137/060673400.

Suhail M. Shah. Distributed optimization on Riemannian manifolds for multi-agent networks. *arXiv preprint arXiv:1711.11196*, 2017. doi: 10.48550/arXiv.1711.11196.

Vincent G. Sigillito, Simon P. Wing, Larrie V. Hutton, and K. B. Baker. Ionosphere. UCI Machine Learning Repository, 1989. URL https://archive.ics.uci.edu/dataset/52/ionosphere.

Hoang Tran and Ashok Cutkosky. Better SGD using second-order momentum. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3530–3541, 2022. doi: 10.52202/068431-0255. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/1704fe7aaff33a54802b83a016050ab8-Abstract-Conference.html.

Roberto Tron, Bijan Afsari, and René Vidal. Riemannian consensus for manifolds with bounded curvature. *IEEE Transactions on Automatic Control*, 58(4):921–934, 2013. doi: 10.1109/TAC.2012.2225533.

Hongye Wang, Zhaoye Pan, Chang He, Jiaxiang Li, and Bo Jiang. Federated learning on Riemannian manifolds: A gradient-free projection-based approach. *arXiv preprint arXiv:2507.22855*, 2025a. doi: 10.48550/arXiv.2507.22855. URL https://arxiv.org/abs/2507.22855.

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 and Xin Liu. A variance-reduced stochastic gradient tracking algorithm for decentralized optimization with orthogonality constraints. *Journal of Industrial and Management Optimization*, 19(10):7753–7776, 2023. doi: 10.3934/JIMO.2023018.

Xiuheng Wang, Ricardo Augusto Borsoi, and Cédric Richard. Riemannian diffusion adaptation over graphs with application to online distributed PCA. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 9736–9740, 2024. doi: 10.1109/ICASSP48485.2024.10446975.

Xiuheng Wang, Ricardo Augusto Borsoi, Cédric Richard, and Ali H. Sayed. Riemannian diffusion adaptation for distributed optimization on manifolds. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62401–62422. PMLR, 2025b. URL https://proceedings.mlr.press/v267/wang25h.html.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004. doi: 10.1016/j.sysconle.2004.02.022.

Haishan Ye and Tong Zhang. DeEPCA: Decentralized exact PCA with linear convergence rate. *Journal of Machine Learning Research*, 22(238):1–27, 2021. URL https://www.jmlr.org/papers/v22/21-0298.html.

Jiaojiao Zhang, Jiang Hu, Anthony Man-Cho So, and Mikael Johansson. Nonconvex federated learning on compact smooth submanifolds with heterogeneous data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c6776239514e90cbaa56b575a9f9e05c-Abstract-Conference.html.

Jishu Zhao, Xi Wang, Jinlong Lei, and Shixiang Chen. Distributed stochastic proximal algorithm on Riemannian submanifolds for weakly-convex functions. *arXiv preprint arXiv:2510.22270v2*, 2025. URL https://arxiv.org/abs/2510.22270v2. Revised July 3, 2026.

Jishu Zhao, Xi Wang, and Jinlong Lei. Distributed Riemannian stochastic gradient tracking algorithm on the Stiefel manifold. *Automatica*, 184:112726, 2026. doi: 10.1016/j.automatica.2025.112726.

A MOTIVATING EXAMPLES

The experiments use the following two instances of (1.1).

Decentralized PCA Let $A_i = [a_{i,1}, \dots, a_{i,m_i}]^\top \in \mathbb{R}^{m_i \times d}$ contain the samples at node i, and let A stack the local matrices. The leading r-dimensional eigenspace of $A^\top A$ solves

$$\min_{X_1,\dots,X_n \in \mathrm{St}(d,r)} -\frac{1}{2n} \sum_{i=1}^{n} \mathrm{tr}(X_i^\top A_i^\top A_i X_i) \quad \text{s.t.} \quad X_1 = \cdots = X_n, \tag{A.1}$$

where $\mathrm{St}(d,r) = \{X \in \mathbb{R}^{d\times r} : X^\top X = I_r\}$. Nodes exchange subspace representatives through their neighbors (Cardot & Degras, 2018; Ye & Zhang, 2021; Wang et al., 2025b). For a uniformly sampled $j \in [m_i]$, the loss $F_i(X; j) = -(m_i/2)\left\|X^\top a_{i,j}\right\|^2$ satisfies

$$f_i(X) = \mathbb{E}[F_i(X;\xi_i)] = -\frac{1}{2}\,\mathrm{tr}(X^\top A_i^\top A_i X).$$

The factor m_i makes the sampled gradient unbiased for the unnormalized local objective used in (A.1).

Decentralized low-rank matrix completion Node i holds columns $A_i \in \mathbb{R}^{d\times T_i}$ and observed indices $\Omega_i \subseteq [d]\times[T_i]$. The operator $\mathcal{P}_{\Omega_i}$ retains observed entries and sets the others to zero. Eliminating the coefficient matrix gives the variable-projection model

$$\begin{aligned} \min_{X_1,\dots,X_n\in \mathrm{St}(d,r)} \quad & \frac{1}{2n}\sum_{i=1}^{n} \|\mathcal{P}_{\Omega_i}(X_i V_i(X_i) - A_i)\|_F^2 , \\ \text{s.t.} \quad & X_1 = \cdots = X_n, \qquad V_i(X) \in \arg\min_V \|\mathcal{P}_{\Omega_i}(XV - A_i)\|_F^2 . \end{aligned} \tag{A.2}$$

For each column, full rank of the observed-row submatrix of X yields a unique least-squares solution and a smooth reduced objective locally. This domain condition is distinct from the smoothness on a neighborhood of $\mathrm{conv}(\mathcal{M})$ in Assumption 4.1. The reduced objective is invariant under right orthogonal transformations: Stiefel representatives describe a column subspace on the Grassmann manifold (Boumal & Absil, 2011; 2015; Deng & Hu, 2025a).

B Further related work

Manifold consensus uses ambient distances (Sarlette & Sepulchre, 2009) or intrinsic operations (Tron et al., 2013). Distributed Riemannian gradient methods cover geodesically convex and nonconvex objectives (Shah, 2017); Grassmann gossip combines stochastic updates with a consensus penalty (Mishra et al., 2019). For Stiefel constraints, DESTINY uses an approximate augmented Lagrangian and gradient tracking (Wang & Liu, 2022), while VRSGT uses finite-sum variance reduction (Wang & Liu, 2023). Communication compression reduces exchange costs (Hu & Deng, 2024).

Other stochastic approaches include DRSGT with constant or variable sample sizes (Zhao et al., 2026), diffusion for online PCA (Wang et al., 2024), and zeroth-order methods with weighted Fréchet consensus (Jin et al., 2025). For weakly convex objectives, stochastic proximal methods give $\mathcal{O}(\epsilon^{-2})$ sampled-model solves for a proximal-residual criterion (Zhao et al., 2025). Intrinsic decentralized methods also provide objective-gap bounds in geodesically convex regions (Nguyen & Uribe, 2026).

Federated methods use server–client communication: RFedSVRG applies finite-sum variance reduction (Li & Ma, 2023), drift corrections address heterogeneous data (Zhang et al., 2024), PriRFed introduces tangent-space noise for privacy (Huang et al., 2026a), and RFedAGS aggregates transported gradients (Huang et al., 2026b). Gradient-free federated methods use function evaluations (Wang et al., 2025a). Section 2 discusses the recursive-momentum methods closest to DPRSHCM.

Table 2 compares representative decentralized methods by geometry, oracle cost, stepsize, and communication rounds. The table records differences in oracle models and accuracy criteria.

C Properties of proximally smooth sets

By the basic properties of proximally smooth sets (Clarke et al., 1995, Theorems 4.1 and 4.8), an R-proximally smooth set $\mathcal{M}$ satisfies the following two properties:

1. For any $\delta \in (0, R)$,

$$\|\mathcal{P}_{\mathcal{M}}(x) - \mathcal{P}_{\mathcal{M}}(y)\| \le \frac{R}{R-\delta}\|x-y\|, \qquad \forall x, y \in \overline{U}_{\mathcal{M}}(\delta), \tag{C.1}$$

   where $\overline{U}_{\mathcal{M}}(\delta) := \{x \in \mathbb{R}^{d\times r} : \mathrm{dist}(x,\mathcal{M}) \le \delta\}$

2. For any $x \in \mathcal{M}$, normal vector $v \in N_x\mathcal{M}$, and $y \in \mathcal{M}$,

$$\langle v, y - x\rangle \le \frac{\|v\|}{2R}\|y-x\|^2. \tag{C.2}$$

Therefore, the manifold projection is Lipschitz continuous on any closed tubular neighborhood whose radius is strictly less than R, with Lipschitz constant $\frac{R}{R-\delta}$. In particular, this constant is close to 1 when δ is sufficiently small relative to R. For example, the Stiefel manifold is a 1-proximally smooth set (Deng & Hu, 2025a).

Table 2: Representative decentralized stochastic methods and the results of this paper. Each entry is subject to the assumptions of the cited result.

Method	Manifold	Oracle complexity	Stepsize	Comm. rounds
DRSGD (Chen et al., 2021)	Stiefel	$\widetilde{\mathcal{O}}(\epsilon^{-2})$ SG	Diminishing; $(k+1)^{-1/2}$	$\widetilde{\mathcal{O}}(t\epsilon^{-2})$
DRSGD (Chen et al., 2021)	Stiefel	$\mathcal{O}(\epsilon^{-2})$ SG	Constant; $K^{-1/2}$	$\mathcal{O}(t\epsilon^{-2})$
DRSGT[a] (Zhao et al., 2026)	Stiefel	$\widetilde{\mathcal{O}}(\epsilon^{-2})$ SG	Constant	$\widetilde{\mathcal{O}}(t\epsilon^{-1})$
VRSGT[b] (Wang & Liu, 2023)	Stiefel constraints	$\mathcal{O}(N+\sqrt{N}\epsilon^{-1})$ component gradients	Constant	$\mathcal{O}(\epsilon^{-1})$
DPRSRM (Deng & Hu, 2025b)	Compact embedded submanifolds	$\mathcal{O}(\epsilon^{-3/2})$ SG	Constant; depends on K	$\mathcal{O}(\epsilon^{-3/2})$
Nguyen–Uribe[c] (Nguyen & Uribe, 2026)	Bounded sectional curvature	$\widetilde{\mathcal{O}}(\epsilon^{-2})$ SG	Diminishing; $(k+1)^{-1/2}$	$\widetilde{\mathcal{O}}(\epsilon^{-2})$
Stochastic proximal[d] (Zhao et al., 2025)	Compact embedded submanifolds	$\mathcal{O}(\epsilon^{-2})$ sampled-model solves	Constant; inverse penalty $K^{-1/2}$	$\mathcal{O}(t\epsilon^{-2})$
DPRSHCM (diminishing)	Compact embedded submanifolds	$\widetilde{\mathcal{O}}(\epsilon^{-3/2})$ SG + HVP	Diminishing; $(k+1)^{-1/3}$	$\widetilde{\mathcal{O}}(\epsilon^{-3/2})$
DPRSHCM (constant)	Compact embedded submanifolds	$\mathcal{O}(\epsilon^{-3/2})$ SG + HVP	Constant; $K^{-1/3}$	$\mathcal{O}(\epsilon^{-3/2})$
DPRSHCM (exact gradient)	Compact embedded submanifolds	$\mathcal{O}(\epsilon^{-1})$ FG + HVP	Small constant	$\mathcal{O}(\epsilon^{-1})$

Oracle counts are per node. SG, FG, and HVP denote stochastic gradients, exact local gradients, and stochastic Hessian–vector products. DPRSHCM also evaluates the tangent-projector derivative. Each FG evaluation counts once; each DPRSHCM oracle has the displayed order. Network and regularity constants are suppressed; $\widetilde{\mathcal{O}}$ also hides logarithmic factors.

Communication counts are synchronous neighbor-exchange rounds to reach each row's accuracy criterion, not edge messages or rounds for consensus alone; t counts mixing steps per iteration. Stepsizes refer to optimization updates.

[a]DRSGT uses batch size $N_k = k+1$; its oracle count includes all sampled gradients. [b]VRSGT uses local finite sums of size N, with inner loop length and batch size both of order $\sqrt{N}$ and a full local gradient refresh once per outer iteration. Its iterates are not exactly feasible, and its stationarity gap measures a generalized gradient together with consensus and orthogonality violation. [c]The criterion is expected ergodic objective gap at most ϵ, under geodesic convexity in a common bounded convex region; consensus uses geodesic distance. [d]The criterion is expected squared proximal residual at most ϵ. Each exact sampled-model solve counts once; inner solve costs are excluded.

Unmarked smooth rows use squared-gradient tolerance ϵ. DPRSHCM uses the joint criterion of Definition 3.2. Its two residuals are bounded at the same index. This index lies in the second half of the run for diminishing and constant stepsizes. See Theorems 6.1, 6.2 and 6.3.

D PROOFS FOR THE CORRECTION OPERATOR

Extension of the tangent projector to a tube Since $\mathcal{M}$ is R-proximally smooth, the map

$$\mathcal{Q}(z) := \mathcal{P}_{T_{\mathcal{P}_{\mathcal{M}}(z)}\mathcal{M}}, \qquad z \in U_{\mathcal{M}}(R), \tag{D.1}$$

is well defined and smooth on the tubular neighborhood $U_{\mathcal{M}}(R)$, and restricts to $y \mapsto \mathcal{P}_{T_y\mathcal{M}}$ on $\mathcal{M}$. For $y \in \mathcal{M}$ and an ambient direction ω we write $D(\mathcal{P}_{T\cdot\mathcal{M}})(y)[\omega] := D\mathcal{Q}(y)[\omega]$. For tangent ω this is the derivative of Section 5: take a curve $\gamma \subset \mathcal{M}$ with $\gamma(0) = y$ and $\gamma'(0) = \omega$; then (4.4) gives, for all z,

$$\|D(\mathcal{P}_{T\cdot\mathcal{M}})(y)[\omega]z\| = \lim_{t\to 0} \left\|(\mathcal{P}_{T_{\gamma(t)}\mathcal{M}} - \mathcal{P}_{T_y\mathcal{M}})z\right\| / |t| \le L_2 \|\omega\| \|z\|. \tag{D.2}$$

Since $D\mathcal{P}_{\mathcal{M}}(y) = \mathcal{P}_{T_y\mathcal{M}}$ for $y \in \mathcal{M}$, the chain rule gives

$$D(\mathcal{P}_{T\cdot\mathcal{M}})(y)[\omega] = D(\mathcal{P}_{T\cdot\mathcal{M}})(y)[\mathcal{P}_{T_y\mathcal{M}}\omega], \qquad y \in \mathcal{M}, \tag{D.3}$$

so the derivative annihilates normal directions. Combining (D.3) with (D.2) and $\left\|\mathcal{P}_{T_y\mathcal{M}}\omega\right\| \le \|\omega\|$, for every $y \in \mathcal{M}$ and all ambient ω and z,

$$\|D(\mathcal{P}_{T\cdot\mathcal{M}})(y)[\omega]z\| \le L_2 \|\omega\| \|z\|. \tag{D.4}$$

Finally, set

$$C_Q := \sup\{\left\|D^2\mathcal{Q}(z)\right\| : \operatorname{dist}(z, \mathcal{M}) \le R/2\}. \tag{D.5}$$

This constant is finite because $\mathcal{Q}$ is smooth on $U_{\mathcal{M}}(R)$ and $\overline{U}_{\mathcal{M}}(R/2)$ is compact. The remainder constant of Lemma 5.1 is

$$L_H := \frac{\rho}{2} + G_\nabla \max\left\{\frac{C_Q}{2},\ \frac{2}{R^2} + \frac{L_2}{R}\right\} + L_2 L_f + \frac{M_H}{2R}. \tag{D.6}$$

D.1 PROOF OF LEMMA 5.1

Proof. Fix i and $x, y \in \mathcal{M}$, and write $\mathcal{Q}$ for the extension in (D.1), so that $\mathcal{Q}(y) = \mathcal{P}_{T_y\mathcal{M}}$, $\mathcal{Q}(x) = \mathcal{P}_{T_x\mathcal{M}}$, and $D(\mathcal{P}_{T\cdot\mathcal{M}})(y)[\omega] = D\mathcal{Q}(y)[\omega]$. By (D.3), $D\mathcal{Q}(y)[\mathcal{P}_{T_y\mathcal{M}}(y-x)] = D\mathcal{Q}(y)[y-x]$, and $\mathcal{P}_{T_y\mathcal{M}}(y-x) = (y-x) - \mathcal{P}_{N_y\mathcal{M}}(y-x)$. Hence, by the embedded definition $\operatorname{grad} f_i = \mathcal{P}_{T\cdot\mathcal{M}}\nabla f_i$ and (5.1),

$$\begin{aligned}
&\operatorname{grad} f_i(x) + \overline{\operatorname{Hess}}\, f_i(y)[\mathcal{P}_{T_y\mathcal{M}}(y-x)] - \operatorname{grad} f_i(y) \\
&= \mathcal{P}_{T_y\mathcal{M}}\left(\nabla f_i(x) + \nabla^2 f_i(y)[y-x] - \nabla f_i(y)\right) \\
&\quad + \left(\mathcal{Q}(x) - \mathcal{Q}(y) + D\mathcal{Q}(y)[y-x]\right)\nabla f_i(x) \\
&\quad + D\mathcal{Q}(y)[y-x]\left(\nabla f_i(y) - \nabla f_i(x)\right) \\
&\quad - \mathcal{P}_{T_y\mathcal{M}}\nabla^2 f_i(y)\left[\mathcal{P}_{N_y\mathcal{M}}(y-x)\right],
\end{aligned}$$

where $\mathcal{Q}(x)\nabla f_i(x) = \operatorname{grad} f_i(x)$. The first term is bounded by $(\rho/2)\|y-x\|^2$ by Assumption 4.3 and nonexpansiveness of the projector.

For the second term, if $\|y-x\| \le R$, every point of the chord $y + t(x-y)$ is within $\min(t, 1-t)\|y-x\| \le R/2$ of $\mathcal{M}$. Taylor's theorem on this tubular neighborhood therefore gives $\|\mathcal{Q}(x) - \mathcal{Q}(y) + D\mathcal{Q}(y)[y-x]\| \le C_Q\|y-x\|^2/2$. If $\|y-x\| > R$, use $\|\mathcal{Q}(x) - \mathcal{Q}(y)\| \le 2$ and (D.4) to bound the same expression by $(2/R^2 + L_2/R)\|y-x\|^2$. Multiplication by $\|\nabla f_i(x)\| \le G_\nabla$ gives the second contribution in (D.6).

The third term is at most $L_2 L_f \|y-x\|^2$ by (D.4) and (4.1). For the last term, set $v = \mathcal{P}_{N_y\mathcal{M}}(y-x)$. The proximal normal inequality at y gives

$$\|v\|^2 = \langle -v, x-y\rangle \le \frac{\|v\|}{2R}\|x-y\|^2, \qquad \|v\| \le \frac{\|x-y\|^2}{2R}.$$

Since $\left\|\nabla^2 f_i(y)\right\|_{\mathrm{op}} \le M_H$, the fourth term is at most $M_H\|y-x\|^2/(2R)$. Adding the four contributions proves (5.4). □

D.2 Oracle bounds for the correction operator

Lemma D.1 (Oracle bounds for the projected gradient and correction)**.** *Suppose Assumptions 4.1 and 4.5 hold. Then, for every $\mathcal{F}_k$-measurable $x \in \mathcal{M}$,*

$$\mathbb{E}[\operatorname{grad} F_i(x;\xi_{i,k}) \mid \mathcal{F}_k] = \operatorname{grad} f_i(x), \qquad \mathbb{E}[\|\operatorname{grad} F_i(x;\xi_{i,k}) - \operatorname{grad} f_i(x)\|^2 \mid \mathcal{F}_k] \le \nu_i^2,$$

and, for every $\mathcal{F}_k$-measurable tangent vector $u \in T_x\mathcal{M}$,

$$\mathbb{E}[\overline{\operatorname{Hess}}\, F_i(x;\xi_{i,k})[u] \mid \mathcal{F}_k] = \overline{\operatorname{Hess}}\, f_i(x)[u], \tag{D.7}$$

$$\mathbb{E}[\left\|\overline{\operatorname{Hess}}\, F_i(x;\xi_{i,k})[u] - \overline{\operatorname{Hess}}\, f_i(x)[u]\right\|^2 \mid \mathcal{F}_k] \le \bar{\sigma}_{H,i}^2 \|u\|^2, \tag{D.8}$$

where $\bar{\sigma}_{H,i}^2 := 2L_2^2\nu_i^2 + 2\sigma_{\nabla^2,i}^2$.

We use the node-independent noise levels

$$\nu^2 := \max_{1\le i\le n} \nu_i^2, \qquad \sigma_H^2 := \max_{1\le i\le n} \bar{\sigma}_{H,i}^2. \tag{D.9}$$

The proofs below use the oracle only through Lemma D.1. Every bound derived there remains valid for any oracle that is conditionally unbiased and has second moments bounded by ν^2 and σ_H^2. In particular it remains valid with $\nu^2 = 0$ when the local gradients are exact (Subsection E.2.4).

Proof. Since x and u are $\mathcal{F}_k$-measurable, so are the linear maps $\mathcal{P}_{T_x\mathcal{M}}$ and $D(\mathcal{P}_{T\cdot\mathcal{M}})(x)[u]$. They can therefore be taken out of the conditional expectation. Hence

$$\mathbb{E}[\operatorname{grad} F_i(x;\xi_{i,k}) \mid \mathcal{F}_k] = \mathcal{P}_{T_x\mathcal{M}}\mathbb{E}[\nabla F_i(x;\xi_{i,k}) \mid \mathcal{F}_k] = \mathcal{P}_{T_x\mathcal{M}}\nabla f_i(x) = \operatorname{grad} f_i(x),$$

and, since $\mathcal{P}_{T_x\mathcal{M}}$ is nonexpansive,

$$\mathbb{E}[\|\operatorname{grad} F_i(x;\xi_{i,k}) - \operatorname{grad} f_i(x)\|^2 \mid \mathcal{F}_k] \le \mathbb{E}[\|\nabla F_i(x;\xi_{i,k}) - \nabla f_i(x)\|^2 \mid \mathcal{F}_k] \le \nu_i^2.$$

For the correction operator, the same argument and (5.2) give

$$\mathbb{E}[\overline{\operatorname{Hess}}\, F_i(x;\xi_{i,k})[u] \mid \mathcal{F}_k] = D(\mathcal{P}_{T\cdot\mathcal{M}})(x)[u]\nabla f_i(x) + \mathcal{P}_{T_x\mathcal{M}}\nabla^2 f_i(x)[u] = \overline{\operatorname{Hess}}\, f_i(x)[u].$$

The gradient and Hessian noises use the same sample, so we bound the second moment of their sum by twice the sum of the second moments. By (D.2) and the nonexpansiveness of $\mathcal{P}_{T_x\mathcal{M}}$,

$$\begin{aligned}
&\mathbb{E}[\left\|\overline{\operatorname{Hess}}\, F_i(x;\xi_{i,k})[u] - \overline{\operatorname{Hess}}\, f_i(x)[u]\right\|^2 \mid \mathcal{F}_k] \\
&\le 2L_2^2 \|u\|^2 \, \mathbb{E}[\|\nabla F_i(x;\xi_{i,k}) - \nabla f_i(x)\|^2 \mid \mathcal{F}_k] \\
&\quad + 2\mathbb{E}[\left\|\nabla^2 F_i(x;\xi_{i,k})[u] - \nabla^2 f_i(x)[u]\right\|^2 \mid \mathcal{F}_k] \\
&\le 2L_2^2\nu_i^2 \|u\|^2 + 2\sigma_{\nabla^2,i}^2 \|u\|^2,
\end{aligned}$$

which is (D.8). □

E Convergence analysis: auxiliary lemmas and proofs

This appendix contains the error quantities and constants, the auxiliary lemmas, and the proofs of the results of Section 6. The stacking and averaging conventions are those of Section 3.

Network contraction Under Assumption 4.4, the second-largest singular value σ_2 of W satisfies $\sigma_2 \in [0,1)$, and for every network variable $\mathbf{z}$,

$$\begin{aligned}
\|((W-J)\otimes I)\mathbf{z}\| &\le \sigma_2 \|\mathbf{z} - \widehat{\mathbf{z}}\|, \\
\|((I-W)\otimes I)\mathbf{z}\| &\le (1+\sigma_2)\|\mathbf{z} - \widehat{\mathbf{z}}\| \le 2\|\mathbf{z} - \widehat{\mathbf{z}}\|.
\end{aligned} \tag{E.1}$$

We call this the network contraction property.

Error quantities For convenience in the proofs, write $g_{i,k} := \operatorname{grad} f_i(x_{i,k})$ and $H_{i,k} := \overline{\operatorname{Hess}}\, f_i(x_{i,k})$, and introduce

$$\begin{aligned}
\varepsilon_{i,k} &:= d_{i,k} - g_{i,k}, & \varepsilon^{\mathrm{nc}}_{i,k} &:= q_{i,k} - g_{i,k},\\
\varepsilon^{\mathrm{g}}_{i,k+1} &:= \operatorname{grad} F_i(x_{i,k+1};\xi_{i,k+1}) - g_{i,k+1},\\
\varepsilon^{\mathrm{h}}_{i,k+1} &:= \overline{\operatorname{Hess}}\, F_i(x_{i,k+1};\xi_{i,k+1})[u_{i,k+1}] - H_{i,k+1}[u_{i,k+1}],\\
r_{i,k+1} &:= g_{i,k} + H_{i,k+1}[u_{i,k+1}] - g_{i,k+1}.
\end{aligned}$$

Here nc denotes the error before clipping, the superscripts g and h identify the gradient noise and correction noise, and $r_{i,k+1}$ is the second-order remainder of Lemma 5.1. We also use the scalar quantities

$$\begin{aligned}
e_k &:= \mathbb{E}\left\|\mathbf{d}_k - \mathbf{g}_k\right\|^2, & E_k &:= e_k/n,\\
T_k &:= \frac{1}{n}\mathbb{E}\left\|\mathbf{s}_k - \widehat{\mathbf{d}}_k\right\|^2, & X_k &:= \frac{1}{n}\mathbb{E}\left\|\mathbf{x}_k - \bar{\mathbf{x}}_k\right\|^2,\\
S_k &:= \frac{1}{n}\mathbb{E}\left\|\mathbf{s}_k\right\|^2, & \mathsf{G}_k &:= \mathbb{E}\left\|\operatorname{grad} F(\bar{x}_k)\right\|^2.
\end{aligned}$$

These abbreviations are used in the proofs. The conclusions of lemmas and theorems are stated in terms of the original iterates and gradients. Conditional expectations are taken before the fresh sample is drawn. Thus $x_{i,k+1}$, $d_{i,k}$ and $u_{i,k+1}$ are $\mathcal{F}_{k+1}$-measurable and $\xi_{i,k+1}$ is independent of $\mathcal{F}_{k+1}$.

Consensus neighborhood and constants For $0 < \delta < R$, let

$$\mathcal{N}(\delta) := \{\mathbf{x} \in \mathcal{M}^n : \|\mathbf{x} - \bar{\mathbf{x}}\| \le \delta\}. \tag{E.2}$$

If $\mathbf{x} \in \mathcal{M}^n$ and $\|\mathbf{x} - \mathbf{1}\otimes y\| \le \delta$ for some $y \in \mathcal{P}_{\mathcal{M}}(\hat{x})$, then $\operatorname{dist}(\hat{x},\mathcal{M}) \le \|\hat{x} - y\| \le \delta/\sqrt{n} < R$, so $\bar{x}$ is single-valued on $\mathcal{N}(\delta)$. Throughout the section, δ is restricted by

$$0 < \delta < \min\left\{\frac{R}{3}, \frac{R(1-\sigma_2)}{2}\right\}, \tag{E.3}$$

and we use the constants

$$\begin{aligned}
\ell &:= \frac{R}{R-2\delta}, \qquad \varrho := \ell\sigma_2, \qquad C_c := \frac{\ell^2}{(1-\varrho)^2},\\
\mathcal{K}_c &:= \sum_{m=0}^{\infty} \varrho^m (m+2)^{1/3}, \qquad \mathcal{K} := 2 + 2(1+\sigma_2)\ell\,\mathcal{K}_c.
\end{aligned} \tag{E.4}$$

Here ℓ is the Lipschitz constant of the metric projection on $\overline{U}_{\mathcal{M}}(2\delta)$, $\varrho < 1$ is the network contraction factor after projection, and $\mathcal{K}_c \le \sum_m \varrho^m(m+2) \le 2/(1-\varrho)^2$. We further use

$$D_d := \frac{9L_d^2}{(1-\sigma_2)^2}, \qquad C_W := \frac{6(1+\sigma_2^2)}{(1-\sigma_2^2)^2}, \qquad \Theta_0 := \frac{2nL_d^2}{1-\sigma_2^2}, \qquad \Delta_0 := F(\bar{x}_0) - F_*, \tag{E.5}$$

which bound the tracker (Lemma E.5), the tracking recursion (Lemma E.12), the initial tracking disagreement, and the initial optimality gap, respectively. Finally, the polynomial coefficients of the descent inequality are written through the two functions

$$\begin{aligned}
\mathcal{G}_3(C) &:= \left(\tfrac{3}{2}L^2 + 2L^2L_2^2\right)C + \left(6M_2^2 + 6(2L_2+8Q)^2\right)C^2 + 24Q^2D_d^2,\\
\mathcal{G}_4(C) &:= 2L\left((8Q + 2L_2 + M_2)^2 + M_2^2\right)C^2 + 8Q^2LD_d^2,
\end{aligned} \tag{E.6}$$

where M_2 and Q are the constants of Lemma E.1 and C is a consensus constant that depends on the stepsize regime. Table 3 collects the constants of the analysis.

E.1 Auxiliary lemmas

The estimates in this subsection follow from the geometry of $\mathcal{M}$, clipping and the projected update (5.10). They apply to any estimator clipped at L_d and every positive nonincreasing stepsize sequence.

The following geometric estimates are adapted from Deng & Hu (2025a;b).

Table 3: Constants used in the convergence analysis. The superscripts dim and con refer to the diminishing and the constant stepsize regimes.

Group	Symbol	Meaning	Defined in
Geometry	R	proximal smoothness radius of $\mathcal{M}$	Subsection 3.2
	δ, ℓ, ϱ	consensus radius, projection Lipschitz constant, contraction factor	(E.3), (E.4)
	C_c, $\mathcal{K}_c$, $\mathcal{K}$	accumulated consensus factor, series constants	(E.4)
	L_2, M_2, Q	tangent projector Lipschitz constant, mean-projection and projection curvature constants	(4.4), Lemma E.1
Smoothness	L_g, L_G, L	descent, gradient Lipschitz, common smoothness constant	Lemma 4.2, Section 4
	G, G_∇, M_H	gradient bounds, Hessian bound	Section 4
	ρ, L_H	second-order constant, remainder constant	Assumption 4.3, (D.6)
Algorithm	L_d, D_d	clipping radius, tracker bound	Assumption 5.2, (E.5)
	τ_k, K_H	momentum parameter, momentum scale	(5.6), (E.47)
	C_W, Θ_0, Δ_0	tracking recursion constant, initial tracking disagreement, initial gap	(E.5)
Noise	ν^2, σ_H^2	gradient and Hessian–vector noise levels	(D.9)
Derived	C_x^{dim}, C_x^{con}	consensus constants	(E.18), (E.19)
	C_Δ^{dim}, C_Δ^{con}	displacement constants	(E.42), (E.63)
	$\mathcal{G}_3(\cdot)$, $\mathcal{G}_4(\cdot)$	coefficients of α^3, α^4 in the descent inequality	(E.6)
	K_H^{dim}, K_H^{con}, $\bar{\alpha}$	momentum and stepsize thresholds	(E.46), (E.65)
	Γ_2, Γ_3, P_τ, $\bar{\alpha}_{\tau,\delta}$	exact-gradient descent coefficients, feedback constant, stepsize threshold	(E.84), (E.85), (E.86)

Lemma E.1 (Deng & Hu (2025a, Lemma 5))**.** *Suppose that Assumptions 4.1 and 4.4 hold, and let $0 < \delta < R/3$. Then there exist constants $M_2, Q > 0$, depending only on $\mathcal{M}$ and δ, such that, for every $\mathbf{x} \in \mathcal{N}(\delta)$,*

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

and, for every $x \in \mathcal{M}$ and every ambient vector u satisfying $\|u\| \le 3\delta$,

$$\|\mathcal{P}_{\mathcal{M}}(x+u) - x - \mathcal{P}_{T_x\mathcal{M}}(u)\| \le Q \|u\|^2. \tag{E.8}$$

Lemma E.2 (Sum of the consensus gradients)**.** *Suppose that Assumptions 4.1 and 4.4 hold. Then, for every $\mathbf{x} \in \mathcal{M}^n$ with $\bar{x}$ single-valued,*

$$\left\| \sum_{i=1}^n \operatorname{grad} \phi_i(\mathbf{x}) \right\| \le 2L_2 \|\mathbf{x} - \bar{\mathbf{x}}\|^2. \tag{E.9}$$

Proof. Since W is doubly stochastic, $\sum_{i=1}^n \nabla\phi_i(\mathbf{x}) = \sum_{j=1}^n (1 - \sum_{i=1}^n W_{ij}) x_j = 0$. Hence

$$\sum_{i=1}^n \operatorname{grad} \phi_i(\mathbf{x}) = \sum_{i=1}^n \left(\mathcal{P}_{T_{x_i}\mathcal{M}} - \mathcal{P}_{T_{\bar{x}}\mathcal{M}} \right) \nabla\phi_i(\mathbf{x}).$$

By (4.4) and the Cauchy–Schwarz inequality,

$$\left\| \sum_{i=1}^n \operatorname{grad} \phi_i(\mathbf{x}) \right\| \le L_2 \sum_{i=1}^n \|x_i - \bar{x}\| \, \|\nabla\phi_i(\mathbf{x})\| \le L_2 \|\mathbf{x} - \bar{\mathbf{x}}\| \, \|\nabla\phi(\mathbf{x})\|.$$

Finally, $\nabla\phi(\mathbf{x}) = ((I - W) \otimes I)(\mathbf{x} - \bar{\mathbf{x}})$ because $((I - W) \otimes I)\bar{\mathbf{x}} = 0$, so $\|\nabla\phi(\mathbf{x})\| \le (1 + \sigma_2) \|\mathbf{x} - \bar{\mathbf{x}}\| \le 2 \|\mathbf{x} - \bar{\mathbf{x}}\|$ by (E.1). □

Lemma E.3 (Error contraction under clipping)**.** *For every q and every z with $\|z\| \le L_d$,*

$$\left\|\mathrm{clip}_{L_d}(q) - z\right\| \le \|q - z\|. \tag{E.10}$$

In particular, under Assumption 5.2, $\|d_{i,k} - \mathrm{grad}\, f_i(x_{i,k})\| \le \|q_{i,k} - \mathrm{grad}\, f_i(x_{i,k})\|$ and $\|d_{i,k+1} - d_{i,k}\| \le \|q_{i,k+1} - d_{i,k}\|$ for every $i \in [n]$ and $k \ge 0$.

Proof. The map clip_{L_d} is the Euclidean projection onto the closed ball $\{z : \|z\| \le L_d\}$. It is non-expansive and fixes every point of the ball. Hence $\left\|\mathrm{clip}_{L_d}(q) - z\right\| = \left\|\mathrm{clip}_{L_d}(q) - \mathrm{clip}_{L_d}(z)\right\| \le \|q - z\|$. The two particular cases follow from $\|\mathrm{grad}\, f_i(x_{i,k})\| \le G \le L_d$ and $\|d_{i,k}\| \le L_d$. □

Lemma E.4 (Preservation of network averages)**.** *Suppose that Assumption 4.4 holds and that $\{\mathbf{s}_k, \mathbf{d}_k\}_{k\ge 0}$ satisfies the gradient-tracking update* (5.8)*. If $\widehat{\mathbf{s}}_0 = \widehat{\mathbf{d}}_0$, then*

$$\widehat{\mathbf{s}}_k = \widehat{\mathbf{d}}_k, \qquad k \ge 0. \tag{E.11}$$

Proof. Apply $J \otimes I$ to (5.8). Since $JW = J$, $\widehat{\mathbf{s}}_k - \widehat{\mathbf{d}}_k = \widehat{\mathbf{s}}_{k-1} - \widehat{\mathbf{d}}_{k-1} = 0$ by induction. □

Lemma E.5 (Uniform boundedness of the tracker)**.** *Suppose that Assumption 4.4 holds. Let $\{\mathbf{s}_k, \mathbf{d}_k, \mathbf{v}_k\}_{k\ge 0}$ satisfy* (5.8)–(5.9)*, where $d_{i,k} = \mathrm{clip}_{L_d}(q_{i,k})$ and $\mathbf{s}_0 = \mathbf{d}_0$. Then, for every $k \ge 0$,*

$$\|\mathbf{s}_k\|^2 \le nD_d, \qquad \|\mathbf{v}_k\|^2 \le nD_d, \tag{E.12}$$

with D_d from (E.5)*.*

Proof. Clipping gives $\|\mathbf{d}_k\| \le \sqrt{n}L_d$ and $\|\mathbf{d}_{k+1} - \mathbf{d}_k\| \le 2\sqrt{n}L_d$. By Lemma E.4, the tracking update is

$$\mathbf{s}_{k+1} = ((W - J) \otimes I)\mathbf{s}_k + \widehat{\mathbf{d}}_k + \mathbf{d}_{k+1} - \mathbf{d}_k.$$

Thus $\|\mathbf{s}_{k+1}\| \le \sigma_2 \|\mathbf{s}_k\| + 3\sqrt{n}L_d$. Since $\|\mathbf{s}_0\| \le \sqrt{n}L_d$, summing the geometric series gives $\|\mathbf{s}_k\| \le 3\sqrt{n}L_d/(1 - \sigma_2)$. Squaring proves the tracker bound, and orthogonal tangent projection gives $\|\mathbf{v}_k\| \le \|\mathbf{s}_k\|$. □

The following lemma ensures that the metric projection remains single-valued along the iteration.

Lemma E.6 (Invariance of the consensus neighborhood)**.** *Suppose Assumptions 4.1, 4.4, and 5.2 hold and $\mathbf{s}_0 = \mathbf{d}_0$. Let δ satisfy* (E.3)*, let the stepsizes in Algorithm 1 be positive and nonincreasing, and suppose*

$$\alpha_0 \le \frac{(R(1 - \sigma_2) - 2\delta)\delta}{R\sqrt{nD_d}}. \tag{E.13}$$

If $\mathbf{x}_0 \in \mathcal{N}(\delta)$, then for every $k \ge 0$,

$$\mathbf{x}_k \in \mathcal{N}(\delta), \qquad \sum_{j=1}^{n} W_{ij}x_{j,k} - \alpha_k v_{i,k} \in \overline{U}_{\mathcal{M}}(2\delta), \quad i \in [n].$$

Proof. Since $(R(1 - \sigma_2) - 2\delta)/R < 1$, (E.13) implies $\alpha_k\sqrt{nD_d} \le \delta$ for every k. We argue by induction. Suppose $\mathbf{x}_k \in \mathcal{N}(\delta)$. For each $i \in [n]$, Lemma E.5 gives

$$\begin{aligned}\left\|\sum_{j=1}^{n} W_{ij}x_{j,k} - \alpha_k v_{i,k} - \bar{x}_k\right\| &\le \sum_{j=1}^{n} W_{ij}\|x_{j,k} - \bar{x}_k\| + \alpha_k\|v_{i,k}\| \\ &\le \delta + \alpha_k\sqrt{nD_d} \le 2\delta.\end{aligned}$$

Thus the point sent to $\mathcal{P}_{\mathcal{M}}$ in the update belongs to $\overline{U}_{\mathcal{M}}(2\delta)$. Since $\hat{x}_k \in \overline{U}_{\mathcal{M}}(\delta) \subset \overline{U}_{\mathcal{M}}(2\delta)$ and $2\delta < R$, the projection map is ℓ-Lipschitz on $\overline{U}_{\mathcal{M}}(2\delta)$ by (C.1). Hence, by the optimality of $\bar{\mathbf{x}}_{k+1}$ among consensus points and $\|\mathbf{x}_k - \widehat{\mathbf{x}}_k\| \le \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|$,

$$\begin{aligned}\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \le \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_k\| &= \|\mathcal{P}_{\mathcal{M}^n}(\mathbf{W}\mathbf{x}_k - \alpha_k\mathbf{v}_k) - \mathcal{P}_{\mathcal{M}^n}(\widehat{\mathbf{x}}_k)\| \\ &\le \ell\,\|((W - J) \otimes I)(\mathbf{x}_k - \widehat{\mathbf{x}}_k) - \alpha_k\mathbf{v}_k\| \\ &\le \varrho\,\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \ell\alpha_k\sqrt{nD_d} \le \delta,\end{aligned}$$

where the last step is exactly (E.13). The claim follows by induction. □

The following theorem extends the consensus estimate of Deng & Hu (2025a, Lemma 8) to nonincreasing stepsizes.

Theorem E.7 (Consensus bounds)**.** *Suppose the conditions of Lemma E.6 hold and* $x_{1,0} = \cdots = x_{n,0}$*. Then, for every* $k \geq 0$*,*

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \leq \varrho \|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \ell\alpha_k \|\mathbf{v}_k\|, \tag{E.14}$$

and consequently, for every $k \geq 1$*,*

$$\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| \leq \ell \sum_{t=0}^{k-1} \varrho^{k-1-t}\alpha_t \|\mathbf{v}_t\| \leq \ell\sqrt{nD_d} \sum_{t=0}^{k-1} \varrho^{k-1-t}\alpha_t. \tag{E.15}$$

In particular, suppose

$$\alpha_k = \frac{\alpha_0}{(k+1)^{1/3}}, \qquad k \geq 0, \tag{E.16}$$

where α_0 *satisfies* (E.13)*. Then*

$$\frac{1}{n} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 \leq C_x^{\dim} \alpha_k^2, \qquad k \geq 0, \tag{E.17}$$

where

$$C_x^{\dim} := \ell^2 D_d \mathcal{K}_c^2. \tag{E.18}$$

For $\alpha_k \equiv \alpha$*, the corresponding bound is*

$$\frac{1}{n} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 \leq C_x^{\mathrm{con}} \alpha^2, \qquad C_x^{\mathrm{con}} := D_d C_c = \frac{\ell^2 D_d}{(1-\varrho)^2}. \tag{E.19}$$

Proof. Lemma E.6 ensures that the arguments of $\mathcal{P}_{\mathcal{M}^n}$ lie in the region where the projection is ℓ-Lipschitz. Exactly as in the proof of that lemma, the optimality of $\bar{\mathbf{x}}_{k+1}$ and (E.1) give $\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \leq \ell\sigma_2 \|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \ell\alpha_k \|\mathbf{v}_k\|$. This is (E.14). Iterating it, using $\mathbf{x}_0 = \bar{\mathbf{x}}_0$ and Lemma E.5, proves (E.15). For the stepsizes in (E.16), setting $m = k - 1 - t$ gives

$$\sum_{t=0}^{k-1} \varrho^{k-1-t}\alpha_t = \alpha_k \sum_{m=0}^{k-1} \varrho^m \left(\frac{k+1}{k-m}\right)^{1/3} \leq \alpha_k \sum_{m=0}^{k-1} \varrho^m (m+2)^{1/3} \leq \mathcal{K}_c \alpha_k,$$

where $(k+1)/(k-m) \leq m+2$ for $0 \leq m \leq k-1$. Substitution into (E.15), followed by squaring and division by n, proves the result. For constant α, sum the geometric series in (E.15) and use $\|\mathbf{v}_t\|^2 \leq nD_d$. □

Lemma E.8 (Pathwise displacement bounds)**.** *Suppose the conditions of Theorem E.7 hold. Then, for every* $k \geq 0$*,*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq 2(1+\sigma_2) \|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + 2\alpha_k \|\mathbf{v}_k\|, \tag{E.20}$$

hence $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \leq 8(1+\sigma_2)^2 \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + 8\alpha_k^2 \|\mathbf{v}_k\|^2$*, and*

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq 2\alpha_k \|\mathbf{v}_k\| + 2(1+\sigma_2)\ell \sum_{t=0}^{k-1} \varrho^{k-1-t}\alpha_t \|\mathbf{v}_t\|. \tag{E.21}$$

Moreover, $\sum_{i=1}^n \|x_{i,k+1} - x_{i,k}\|^4 \leq \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^4$ *and* $\|\mathbf{v}_k\|^4 \leq nD_d \|\mathbf{v}_k\|^2$*.*

Proof. The update (5.10) and the optimality of the metric projection imply

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{x}_k\| &\leq \|\mathcal{P}_{\mathcal{M}^n}(\mathbf{W}\mathbf{x}_k - \alpha_k\mathbf{v}_k) - (\mathbf{W}\mathbf{x}_k - \alpha_k\mathbf{v}_k)\| \\
&\quad + \|((I-W)\otimes I)\mathbf{x}_k + \alpha_k\mathbf{v}_k\| \\
&\leq 2\|((I-W)\otimes I)\mathbf{x}_k + \alpha_k\mathbf{v}_k\| \\
&\leq 2(1+\sigma_2)\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + 2\alpha_k\|\mathbf{v}_k\|,
\end{aligned}$$

where the first inequality is the triangle inequality, the second uses that $\mathbf{x}_k \in \mathcal{M}^n$ is a candidate for the projection, and the third uses $((I-W)\otimes I)\bar{\mathbf{x}}_k = 0$ and (E.1). The squared form follows from $(a+b)^2 \leq 2a^2 + 2b^2$, and (E.21) follows by inserting (E.15). The last two claims follow from $\sum_i a_i^2 \leq (\sum_i a_i)^2$ for $a_i \geq 0$ and from Lemma E.5. □

Lemma E.9 (Weighted moment bounds)**.** *Under the hypotheses of Lemma E.8, let $p \in \{2, 4\}$ and let positive deterministic weights w_k satisfy $(w_k/w_t)^{1/p} \le (k-t+1)^{1/3}$ for $0 \le t < k$. Then, for every $K \ge 1$,*

$$\sum_{k=0}^{K-1} w_k \mathbb{E}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^p \le \mathcal{K}^p \sum_{k=0}^{K-1} w_k \alpha_k^p \mathbb{E}\left\|\mathbf{v}_k\right\|^p. \tag{E.22}$$

For constant stepsize α, writing $\mathcal{S}_K = \sum_{k=0}^{K-1} \mathbb{E}\left\|\mathbf{s}_k\right\|^2$, we also have

$$\begin{aligned} \sum_{k<K} \mathbb{E}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2 &\le \mathcal{K}^2 \alpha^2 \mathcal{S}_K, \\ \sum_{k<K} \mathbb{E} \sum_i \left\|x_{i,k+1} - x_{i,k}\right\|^4 &\le \mathcal{K}^4 n D_d \alpha^4 \mathcal{S}_K, \end{aligned} \tag{E.23}$$

and

$$\sum_{k=0}^{K} \mathbb{E}\left\|\mathbf{x}_k - \bar{\mathbf{x}}_k\right\|^2 \le C_c \alpha^2 \mathcal{S}_K. \tag{E.24}$$

Proof. Multiply (E.21) by $w_k^{1/p}$. With $a_t = w_t^{1/p} \alpha_t \left\|\mathbf{v}_t\right\|$, the resulting bound is the convolution of a_t with a nonnegative kernel bounded by

$$h_0 = 2, \qquad h_m = 2(1+\sigma_2)\ell \varrho^{m-1}(m+1)^{1/3} \quad (m \ge 1).$$

Its ℓ_1 norm is $\mathcal{K}$. Extend a_t by zero outside $0 \le t < K$ and apply Young's inequality $\left\|h * a\right\|_{\ell_p} \le \left\|h\right\|_{\ell_1} \left\|a\right\|_{\ell_p}$ (Boas & Imoru, 1975). Taking expectations proves (E.22). For $w_k = 1$ and constant α, use $\left\|\mathbf{v}_k\right\| \le \left\|\mathbf{s}_k\right\|$, $\left\|\mathbf{s}_k\right\|^4 \le n D_d \left\|\mathbf{s}_k\right\|^2$, and $\sum_i \left\|\Delta x_{i,k}\right\|^4 \le \left\|\Delta \mathbf{x}_k\right\|^4$ to obtain (E.23). Applying the same convolution inequality to (E.15), whose kernel has sum $\ell/(1-\varrho) = \sqrt{C_c}$, proves (E.24). □

The following descent inequality adapts the projected-gradient argument of Deng & Hu (2025b) to variable stepsizes. The proof retains the displacement and consensus terms needed for the Hessian correction.

Lemma E.10 (Pathwise descent inequality)**.** *Suppose the conditions of Theorem E.7 hold. Then, for every $k \ge 0$,*

$$\begin{aligned} F(\bar{x}_{k+1}) \le F(\bar{x}_k) &- \frac{\alpha_k}{4}\left\|\operatorname{grad} F(\bar{x}_k)\right\|^2 + \frac{3\alpha_k}{2n}\left(\left\|\mathbf{g}_k - \mathbf{d}_k\right\|^2 + \left\|\widehat{\mathbf{d}}_k - \mathbf{s}_k\right\|^2\right) \\ &+ \left(\frac{3L^2}{2} + 2L^2 L_2^2\right)\frac{\alpha_k}{n}\left\|\mathbf{x}_k - \bar{\mathbf{x}}_k\right\|^2 \\ &+ \frac{3M_2^2 + 6(2L_2 + 8Q)^2}{\alpha_k n^2}\left\|\mathbf{x}_k - \bar{\mathbf{x}}_k\right\|^4 + \frac{3M_2^2}{\alpha_k n^2}\left\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\right\|^4 \\ &+ \frac{2L(8Q + 2L_2 + M_2)^2}{n^2}\left\|\mathbf{x}_k - \bar{\mathbf{x}}_k\right\|^4 + \frac{2LM_2^2}{n^2}\left\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\right\|^4 \\ &+ \frac{\alpha_k}{n}\left(2L\alpha_k - \frac{3}{8}\right)\left\|\mathbf{s}_k\right\|^2 + \frac{8Q^2\alpha_k^3}{n^2}(3 + L\alpha_k)\left\|\mathbf{s}_k\right\|^4. \end{aligned} \tag{E.25}$$

Proof. Fix k and abbreviate $a = \alpha_k$, $g = \operatorname{grad} F(\bar{x}_k)$, $\chi_j = n^{-1}\left\|\mathbf{x}_j - \bar{\mathbf{x}}_j\right\|^2$, $z = n^{-1}\left\|\mathbf{s}_k\right\|^2$, $A = n^{-1}\left\|\mathbf{1} \otimes g - \mathbf{s}_k\right\|^2$, $h = \bar{x}_{k+1} - \bar{x}_k + a\hat{v}_k$, and $q = \bar{x}_{k+1} - \bar{x}_k$. Smoothness (4.2), the polarization identity for $-a\langle g, \hat{s}_k\rangle$, and $\langle g, h\rangle \le a\left\|g\right\|^2/4 + \left\|h\right\|^2/a$ give

$$F(\bar{x}_{k+1}) - F(\bar{x}_k) \le -\frac{a}{4}\left\|g\right\|^2 - \frac{a}{2}z + \frac{a}{2}A + \frac{\left\|h\right\|^2}{a} + \frac{L}{2}\left\|q\right\|^2 + a\langle g, \hat{s}_k - \hat{v}_k\rangle. \tag{E.26}$$

By Jensen's inequality, gradient Lipschitz continuity and nonexpansiveness of $J \otimes I$,

$$A \le 3L^2\chi_k + \frac{3}{n}\left(\left\|\mathbf{g}_k - \mathbf{d}_k\right\|^2 + \left\|\widehat{\mathbf{d}}_k - \mathbf{s}_k\right\|^2\right). \tag{E.27}$$

To control h, write the update as $x_{i,k+1} = \mathcal{P}_{\mathcal{M}}(x_{i,k} - \nabla\phi_i(\mathbf{x}_k) - av_{i,k})$. The neighborhood bound gives $\|\nabla\phi_i(\mathbf{x}_k)\| \le 2\delta$, and (E.13) gives $a\|v_{i,k}\| \le \delta$. Thus (E.8) applies. Averaging the projection remainders, using (E.9), $\|\nabla\phi(\mathbf{x}_k)\| \le 2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|$ and $\|\mathbf{v}_k\| \le \|\mathbf{s}_k\|$, yields

$$\|\hat{x}_{k+1} - \hat{x}_k + a\hat{v}_k\| \le (2L_2 + 8Q)\chi_k + 2Qa^2 z. \tag{E.28}$$

Put $B = 2L_2 + 8Q$. Equation (E.7) at k and $k+1$ gives $\|h\| \le M_2\chi_k + M_2\chi_{k+1} + B\chi_k + 2Qa^2 z$. Group the last two summands and use the three-term squared-norm bound:

$$\|h\|^2 \le (3M_2^2 + 6B^2)\chi_k^2 + 3M_2^2\chi_{k+1}^2 + 24Q^2a^4z^2. \tag{E.29}$$

Since $\|\hat{v}_k\| \le \sqrt{z}$, the four-term bound for $\|q\| \le (B + M_2)\chi_k + M_2\chi_{k+1} + 2Qa^2z + a\sqrt{z}$ gives

$$\|q\|^2 \le 4(B + M_2)^2\chi_k^2 + 4M_2^2\chi_{k+1}^2 + 16Q^2a^4z^2 + 4a^2z. \tag{E.30}$$

Finally, $s_{i,k} - v_{i,k}$ is normal at $x_{i,k}$ and g is tangent at $\bar{x}_k$. Insert $\mathcal{P}_{T_{\bar{x}_k}\mathcal{M}} - \mathcal{P}_{T_{x_{i,k}}\mathcal{M}}$ in their inner product. By (4.4), $\|g\| \le G \le L$ and $\langle u, v\rangle \le 2\|u\|^2 + \|v\|^2/8$,

$$a\langle g, \hat{s}_k - \hat{v}_k\rangle \le 2aL^2L_2^2\chi_k + \frac{a}{8}z. \tag{E.31}$$

Substitution of these four bounds into (E.26) gives (E.25), with the stated coefficients. □

Lemma E.11 (Descent inequality under polynomial stepsizes)**.** *Suppose the polynomial-stepsize hypotheses of Theorem E.7 hold and $\alpha_0 \le 3/(16L)$. Then, for every $k \ge 0$,*

$$\begin{aligned} F(\bar{x}_{k+1}) \le F(\bar{x}_k) - \frac{\alpha_k}{4}\|\operatorname{grad} F(\bar{x}_k)\|^2 + \frac{3\alpha_k}{2n}\left(\|\mathbf{g}_k - \mathbf{d}_k\|^2 + \left\|\widehat{\mathbf{d}}_k - \mathbf{s}_k\right\|^2\right) \\ + \mathcal{G}_3(C_x^{\dim})\alpha_k^3 + \mathcal{G}_4(C_x^{\dim})\alpha_k^4, \end{aligned} \tag{E.32}$$

with $\mathcal{G}_3, \mathcal{G}_4$ from (E.6)*.*

Proof. In (E.25), use $\|\mathbf{x}_j - \bar{\mathbf{x}}_j\|^2/n \le C_x^{\dim}\alpha_k^2$ for $j = k, k+1$, and $\|\mathbf{s}_k\|^2/n \le D_d$. The coefficient $2L\alpha_k - 3/8$ is nonpositive, and the remaining cubic and quartic terms are exactly bounded by $\mathcal{G}_3(C_x^{\dim})\alpha_k^3 + \mathcal{G}_4(C_x^{\dim})\alpha_k^4$. □

E.2 Oracle complexity analysis

Lemma E.10 reduces the oracle-complexity analysis to controlling the estimator error $\|\mathbf{d}_k - \mathbf{g}_k\|^2$ and the tracking disagreement $\left\|\mathbf{s}_k - \widehat{\mathbf{d}}_k\right\|^2$. We first establish one-step bounds that hold for every positive nonincreasing stepsize sequence, and then treat the three parameter regimes. Lemma E.9 supplies the common moment bounds. Diminishing stepsizes require changing potential weights; prescribed-budget constant stepsizes require control of the initial error before selecting the output; exact gradients permit absorption of the accumulated errors with fixed parameters.

E.2.1 Estimator and tracking bounds

Lemma E.12 (One-step estimator and tracking bounds)**.** *Suppose Assumptions 4.1–5.2 hold, $x_{1,0} = \cdots = x_{n,0}$, $\mathbf{s}_0 = \mathbf{d}_0$, the stepsizes are positive, nonincreasing and satisfy* (E.13)*, and $\tau_k \in (0, 1]$ for every k. Then, for every $k \ge 0$,*

$$\begin{aligned} e_{k+1} &\le (1 - \tau_k)e_k + \frac{L_H^2}{\tau_k}\mathbb{E}\sum_{i=1}^n \|x_{i,k+1} - x_{i,k}\|^4 \\ &\quad + 2\sigma_H^2\mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + 2n\tau_k^2\nu^2, \end{aligned} \tag{E.33}$$

$$\begin{aligned} \mathbb{E}\|\mathbf{d}_{k+1} - \mathbf{d}_k\|^2 &\le 3\tau_k^2 e_k + (3L^2 + 2\sigma_H^2)\mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &\quad + 3L_H^2\mathbb{E}\sum_{i=1}^n \|x_{i,k+1} - x_{i,k}\|^4 + 2n\tau_k^2\nu^2, \end{aligned} \tag{E.34}$$

$$nT_{k+1} \le \frac{1 + \sigma_2^2}{2}nT_k + \frac{1 + \sigma_2^2}{1 - \sigma_2^2}\mathbb{E}\|\mathbf{d}_{k+1} - \mathbf{d}_k\|^2. \tag{E.35}$$

Consequently, for every integer $K \geq 1$,

$$\sum_{k=0}^{K-1} \alpha_k n T_k \leq \Theta_0 \alpha_0 + C_W \sum_{k=0}^{K-1} \alpha_k \tau_k^2 e_k + \frac{C_W(3L^2 + 2\sigma_H^2)}{3} \sum_{k=0}^{K-1} \alpha_k \mathbb{E} \left\| \mathbf{x}_{k+1} - \mathbf{x}_k \right\|^2 + C_W L_H^2 \sum_{k=0}^{K-1} \alpha_k \mathbb{E} \sum_{i=1}^{n} \left\| x_{i,k+1} - x_{i,k} \right\|^4 + \frac{2C_W n \nu^2}{3} \sum_{k=0}^{K-1} \alpha_k \tau_k^2. \tag{E.36}$$

The bounds remain valid, with $\nu^2 = 0$, when the local gradients are exact and the Hessian–vector oracle satisfies the two properties of Lemma D.1 with constant σ_H^2.

Proof. Estimator error. Fix $k \geq 0$ and condition on $\mathcal{F}_{k+1}$. The variables $x_{i,k}$, $x_{i,k+1}$, $d_{i,k}$ and $u_{i,k+1}$ are $\mathcal{F}_{k+1}$-measurable. The sample $\xi_{i,k+1}$ is fresh. With the error notation introduced at the beginning of this section, the update (5.6) reads

$$\varepsilon_{i,k+1}^{\mathrm{nc}} = (1 - \tau_k)(\varepsilon_{i,k} + r_{i,k+1}) + (1 - \tau_k)\varepsilon_{i,k+1}^{\mathrm{h}} + \tau_k \varepsilon_{i,k+1}^{\mathrm{g}}, \tag{E.37}$$

which is the estimator decomposition used in Tran & Cutkosky (2022, Lemma 3). Conditionally on $\mathcal{F}_{k+1}$, the first term is measurable and the two oracle noises have zero mean by Lemma D.1, so the cross term vanishes. The gradient and Hessian noises may use the same sample; by $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, Lemma D.1, and $\|u_{i,k+1}\| \leq \|x_{i,k+1} - x_{i,k}\|$,

$$\mathbb{E}\left[\left\| (1 - \tau_k)\varepsilon_{i,k+1}^{\mathrm{h}} + \tau_k \varepsilon_{i,k+1}^{\mathrm{g}} \right\|^2 \,\middle|\, \mathcal{F}_{k+1} \right] \leq 2(1 - \tau_k)^2 \sigma_H^2 \left\| x_{i,k+1} - x_{i,k} \right\|^2 + 2\tau_k^2 \nu^2. \tag{E.38}$$

Lemma E.3 and conditional orthogonality therefore give

$$\mathbb{E}[\|\varepsilon_{i,k+1}\|^2 \mid \mathcal{F}_{k+1}] \leq (1 - \tau_k)^2 \left\| \varepsilon_{i,k} + r_{i,k+1} \right\|^2 + 2(1 - \tau_k)^2 \sigma_H^2 \left\| x_{i,k+1} - x_{i,k} \right\|^2 + 2\tau_k^2 \nu^2. \tag{E.39}$$

For $0 < \tau_k < 1$, Young's inequality with parameter $\tau_k/(1-\tau_k)$ yields

$$(1 - \tau_k)^2 \left\| \varepsilon_{i,k} + r_{i,k+1} \right\|^2 \leq (1 - \tau_k) \left\| \varepsilon_{i,k} \right\|^2 + \tau_k^{-1} \left\| r_{i,k+1} \right\|^2,$$

and the same inequality holds when $\tau_k = 1$. Lemma 5.1 gives $\|r_{i,k+1}\| \leq L_H \|x_{i,k+1} - x_{i,k}\|^2$. Summing (E.39) over the nodes, using $(1-\tau_k)^2 \leq 1$, and taking expectation gives (E.33).

Estimator increment. By Lemma E.3, $\|d_{i,k+1} - d_{i,k}\| \leq \|q_{i,k+1} - d_{i,k}\|$, and by (5.6),

$$q_{i,k+1} - d_{i,k} = g_{i,k+1} - g_{i,k} - \tau_k \varepsilon_{i,k} + (1 - \tau_k) r_{i,k+1} + (1 - \tau_k)\varepsilon_{i,k+1}^{\mathrm{h}} + \tau_k \varepsilon_{i,k+1}^{\mathrm{g}}. \tag{E.40}$$

By (4.3), $L_G \leq L$, Lemma 5.1, and $\|a+b+c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$, the $\mathcal{F}_{k+1}$-measurable part satisfies

$$\left\| g_{i,k+1} - g_{i,k} - \tau_k \varepsilon_{i,k} + (1 - \tau_k) r_{i,k+1} \right\|^2 \leq 3L^2 \left\| x_{i,k+1} - x_{i,k} \right\|^2 + 3\tau_k^2 \left\| \varepsilon_{i,k} \right\|^2 + 3L_H^2 \left\| x_{i,k+1} - x_{i,k} \right\|^4.$$

The stochastic part has conditional mean zero and its conditional second moment is bounded by (E.38) with $(1-\tau_k)^2 \leq 1$. Combining, summing over i, and taking total expectation yields (E.34).

Tracking disagreement. Lemma E.4 gives $\widehat{\mathbf{s}}_k = \widehat{\mathbf{d}}_k$. Hence $\mathbf{s}_k - \widehat{\mathbf{d}}_k$ belongs to the disagreement subspace, and (5.8) gives

$$\mathbf{s}_{k+1} - \widehat{\mathbf{d}}_{k+1} = \big((W - J) \otimes I\big)(\mathbf{s}_k - \widehat{\mathbf{d}}_k) + \big((I - J) \otimes I\big)(\mathbf{d}_{k+1} - \mathbf{d}_k).$$

If $\sigma_2 > 0$, Young's inequality with parameter $(1 - \sigma_2^2)/(2\sigma_2^2)$ and (E.1) give $\left\| \mathbf{s}_{k+1} - \widehat{\mathbf{d}}_{k+1} \right\|^2 \leq \frac{1+\sigma_2^2}{2} \left\| \mathbf{s}_k - \widehat{\mathbf{d}}_k \right\|^2 + \frac{1+\sigma_2^2}{1-\sigma_2^2} \left\| \mathbf{d}_{k+1} - \mathbf{d}_k \right\|^2$; if $\sigma_2 = 0$, the first term vanishes and the bound remains valid. Taking expectation gives (E.35).

Weighted sum. Substitute (E.34) into (E.35), multiply by α_{k+1}, and sum from $k = 0$ to $K-1$. Since $\alpha_{k+1} \leq \alpha_k$, the left-hand side is at least $\sum_{k=0}^{K-1} \alpha_k n T_k - \alpha_0 \mathbb{E} \left\| \mathbf{s}_0 - \widehat{\mathbf{d}}_0 \right\|^2 + \alpha_K \mathbb{E} \left\| \mathbf{s}_K - \widehat{\mathbf{d}}_K \right\|^2$. The

terminal term is nonnegative, and $\mathbf{s}_0 = \mathbf{d}_0$ together with the clipping bound gives $\mathbb{E}\left\|\mathbf{s}_0 - \widehat{\mathbf{d}}_0\right\|^2 = \mathbb{E}\,\|((I-J)\otimes I)\mathbf{d}_0\|^2 \le nL_d^2$. Finally, $1 - \frac{1+\sigma_2^2}{2} = \frac{1-\sigma_2^2}{2}$ and $\frac{3(1+\sigma_2^2)/(1-\sigma_2^2)}{(1-\sigma_2^2)/2} = C_W$. Rearranging and dividing by $(1-\sigma_2^2)/2$ proves (E.36). The leading term is $\frac{2nL_d^2}{1-\sigma_2^2}\alpha_0 = \Theta_0\alpha_0$.

The last claim holds because the argument uses the oracle only through the unbiasedness and the second-moment bounds of Lemma D.1. □

Lemma E.13 (Tracking bound under polynomial stepsizes)**.** *Suppose Assumptions 4.1–5.2 hold, the polynomial-stepsize hypotheses of Theorem E.7 hold, and $\tau_k \in (0,1]$. Then, for every integer $K \ge 1$,*

$$\begin{aligned}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\widehat{\mathbf{d}}_k - \mathbf{s}_k\right\|^2 &\le \Theta_0\alpha_0 + C_W\sum_{k=0}^{K-1}\alpha_k\tau_k^2\mathbb{E}\,\|\mathbf{g}_k - \mathbf{d}_k\|^2 \\ &\quad + \frac{C_W n(3L^2+2\sigma_H^2)C_\Delta^{\dim}}{3}\sum_{k=0}^{K-1}\alpha_k^3 \\ &\quad + C_W L_H^2 n^2 (C_\Delta^{\dim})^2\sum_{k=0}^{K-1}\alpha_k^5 + \frac{2C_W n\nu^2}{3}\sum_{k=0}^{K-1}\alpha_k\tau_k^2,\end{aligned} \tag{E.41}$$

where

$$C_\Delta^{\dim} := 8(1+\sigma_2)^2 C_x^{\dim} + 8D_d. \tag{E.42}$$

Proof. Lemma E.8 and (E.17) give $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \le nC_\Delta^{\dim}\alpha_k^2$. Hence $\sum_i \|x_{i,k+1} - x_{i,k}\|^4 \le n^2(C_\Delta^{\dim})^2\alpha_k^4$. Insert these bounds into (E.36). □

Lemma E.14 (Direction decomposition)**.** *Under the hypotheses of Lemma E.12,*

$$\mathbb{E}\,\|\mathbf{v}_k\|^2 \le 3\mathbb{E}\left\|\mathbf{s}_k - \widehat{\mathbf{d}}_k\right\|^2 + 3e_k + 6n\mathbb{E}\,\|\operatorname{grad} F(\bar{x}_k)\|^2 + 6L^2\mathbb{E}\,\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2. \tag{E.43}$$

For the polynomial stepsizes of Theorem E.7, this gives

$$\mathbb{E}\,\|\mathbf{v}_k\|^2 \le 3\mathbb{E}\left\|\mathbf{s}_k - \widehat{\mathbf{d}}_k\right\|^2 + 3e_k + 6n\mathbb{E}\,\|\operatorname{grad} F(\bar{x}_k)\|^2 + 6nL^2 C_x^{\dim}\alpha_k^2. \tag{E.44}$$

Proof. Orthogonal projection and preservation of the network average give $\|\mathbf{v}_k\| \le \|\mathbf{s}_k\| \le \left\|\mathbf{s}_k - \widehat{\mathbf{d}}_k\right\| + \|\mathbf{d}_k - \mathbf{g}_k\| + \|\widehat{\mathbf{g}}_k\|$. By Jensen's inequality and (4.3),

$$\|\widehat{\mathbf{g}}_k - \mathbf{1}\otimes\operatorname{grad} F(\bar{x}_k)\|^2 \le L^2\,\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2.$$

Squaring the three-term bound and taking expectations proves (E.43); (E.17) gives the specialization. □

E.2.2 Diminishing stepsizes

The parameters of this regime are chosen in the order δ, K_H, α_0 according to

$$\begin{aligned} 0 < \delta &< \min\left\{\frac{R}{3}, \frac{R(1-\sigma_2)}{2}\right\}, \\ K_H &\ge K_H^{\dim}, \\ 0 < \alpha_0 &\le \min\left\{\frac{(R(1-\sigma_2)-2\delta)\delta}{R\sqrt{nD_d}}, \frac{3}{16L}, \frac{2^{-1/3}}{\sqrt{K_H}}\right\},\end{aligned} \tag{E.45}$$

where the momentum threshold, which depends on δ through $\mathcal{K}$ but not on α_0, is

$$K_H^{\dim} := 90\cdot 2^{1/3}\left(\mathcal{K}^2\sigma_H^2 + \sqrt{\mathcal{K}^4\sigma_H^4 + \frac{n\mathcal{K}^4 L_H^2 D_d}{180}}\right). \tag{E.46}$$

The stepsizes and the common momentum parameters are

$$\alpha_k := \frac{\alpha_0}{(k+1)^{1/3}}, \qquad \tau_k := 2K_H\alpha_k\alpha_{k+1}, \qquad k \geq 0, \tag{E.47}$$

so that α_k is the sequence of (E.16) and $K_H = \tau_0/(2\alpha_0\alpha_1)$. Condition (E.45) contains (E.3), (E.13), $\alpha_0 \leq 3/(16L)$, and

$$0 < \tau_k = \frac{2K_H\alpha_0^2}{((k+1)(k+2))^{1/3}} \leq 2^{2/3}K_H\alpha_0^2 \leq 1, \tag{E.48}$$

so all results of Subsections E.1 and E.2.1 apply. The role of $K_H^{\dim}$ is explained in Lemma E.16.

Lemma E.15 (Weighted displacement bounds)**.** *Suppose Assumptions 4.1–5.2 hold, use Algorithm 1 with $x_{1,0} = \cdots = x_{n,0}$, choose the parameters according to* (E.45)*, and use the sequences* (E.47)*. Then, for every integer $K \geq 1$,*

$$\sum_{k=0}^{K-1} \frac{\alpha_k}{\tau_k}\mathbb{E}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2 \leq \frac{2^{1/3}\mathcal{K}^2}{2K_H}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\mathbf{v}_k\right\|^2, \tag{E.49}$$

and

$$\sum_{k=0}^{K-1} \frac{\alpha_k}{\tau_k^2}\mathbb{E}\sum_{i=1}^{n}\left\|x_{i,k+1} - x_{i,k}\right\|^4 \leq \frac{2^{2/3}\mathcal{K}^4 nD_d}{4K_H^2}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\mathbf{v}_k\right\|^2. \tag{E.50}$$

Proof. Apply Lemma E.9 with $(p, w_k) = (2, \alpha_k/\tau_k)$ and $(4, \alpha_k/\tau_k^2)$. For $t < k$, the two weight ratios satisfy

$$\left(\frac{\alpha_k/\tau_k}{\alpha_t/\tau_t}\right)^{1/2} = \left(\frac{k+2}{t+2}\right)^{1/6} \leq (k-t+1)^{1/6},$$

$$\left(\frac{\alpha_k/\tau_k^2}{\alpha_t/\tau_t^2}\right)^{1/4} = \left[\left(\frac{k+1}{t+1}\right)^{1/3}\left(\frac{k+2}{t+2}\right)^{2/3}\right]^{1/4} \leq (k-t+1)^{1/4}.$$

Both satisfy the required exponent bound. Moreover,

$$\frac{\alpha_k^3}{\tau_k} \leq \frac{2^{1/3}}{2K_H}\alpha_k, \qquad \frac{\alpha_k^5}{\tau_k^2} \leq \frac{2^{2/3}}{4K_H^2}\alpha_k.$$

Substitute these into (E.22); for the fourth moment also use $\sum_i \left\|\Delta x_{i,k}\right\|^4 \leq \left\|\Delta\mathbf{x}_k\right\|^4$ and $\left\|\mathbf{v}_k\right\|^4 \leq nD_d\left\|\mathbf{v}_k\right\|^2$. □

Lemma E.16 (Weighted estimator potential)**.** *Suppose the conditions of Lemma E.15 hold. Then, for every integer $K \geq 1$,*

$$\begin{aligned}
&\frac{2}{nK_H\alpha_K}\mathbb{E}\left\|\mathbf{d}_K - \mathbf{g}_K\right\|^2 - \frac{2}{nK_H\alpha_0}\mathbb{E}\left\|\mathbf{d}_0 - \mathbf{g}_0\right\|^2 \\
\leq\ &-\frac{3}{2n}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\mathbf{d}_k - \mathbf{g}_k\right\|^2 + \frac{1}{15n}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\mathbf{s}_k - \widehat{\mathbf{d}}_k\right\|^2 \\
&+\frac{2}{15}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\operatorname{grad}F(\bar{x}_k)\right\|^2 + \mathcal{R}_{\mathrm{e}}\sum_{k=0}^{K-1}\alpha_k^3 + 8\nu^2\sum_{k=0}^{K-1}\alpha_k\tau_k,
\end{aligned} \tag{E.51}$$

where

$$\mathcal{R}_{\mathrm{e}} := \frac{40L_d^2}{219K_H^2\alpha_0^6} + \frac{2L^2C_x^{\dim}}{15}. \tag{E.52}$$

Proof. Step 1: Conditional estimator recursion. By (E.48), $\tau_k \in (0, 1]$, so (E.33) of Lemma E.12 applies:

$$e_{k+1} \leq (1-\tau_k)e_k + \frac{L_H^2}{\tau_k}\mathbb{E}\sum_{i=1}^{n}\left\|x_{i,k+1} - x_{i,k}\right\|^4 + 2\sigma_H^2\mathbb{E}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2 + 2n\tau_k^2\nu^2. \tag{E.53}$$

Step 2: Weighted summation and displacement feedback. Multiply (E.53), after subtracting e_k, by $4\alpha_k/(n\tau_k) = 2/(nK_H\alpha_{k+1})$. The exact change in the weighted estimator potential is

$$\frac{2e_{k+1}}{nK_H\alpha_{k+1}} - \frac{2e_k}{nK_H\alpha_k} = \frac{2(e_{k+1}-e_k)}{nK_H\alpha_{k+1}} + \frac{2}{nK_H}\left(\frac{1}{\alpha_{k+1}} - \frac{1}{\alpha_k}\right)e_k. \tag{E.54}$$

The second term is nonzero because the stepsizes vary. Concavity of $t^{1/3}$ and Young's inequality imply, for every $\mu > 0$,

$$\frac{1}{\alpha_{k+1}} - \frac{1}{\alpha_k} \le \frac{\alpha_k^2}{3\alpha_0^3}, \qquad \alpha_k^2 \le \frac{\mu}{2}\alpha_k + \frac{1}{2\mu}\alpha_k^3.$$

Also, $\|d_{i,k}\| \le L_d$ by the clipping step and $\|g_{i,k}\| \le G \le L_d$, so $e_k \le 4nL_d^2$. Consequently,

$$\frac{2}{nK_H}\left(\frac{1}{\alpha_{k+1}} - \frac{1}{\alpha_k}\right)e_k \le \frac{\mu}{3nK_H\alpha_0^3}\alpha_k e_k + \frac{4L_d^2}{3K_H\mu\alpha_0^3}\alpha_k^3.$$

Substitution into (E.54), followed by summation, gives

$$\begin{aligned}\frac{2e_K}{nK_H\alpha_K} - \frac{2e_0}{nK_H\alpha_0} \le{}& \left(-4 + \frac{\mu}{3K_H\alpha_0^3}\right)\frac{1}{n}\sum_{k=0}^{K-1}\alpha_k e_k + \frac{8\sigma_H^2}{n}\sum_{k=0}^{K-1}\frac{\alpha_k}{\tau_k}\mathbb{E}\left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\|^2 \\ &+ \frac{4L_H^2}{n}\sum_{k=0}^{K-1}\frac{\alpha_k}{\tau_k^2}\mathbb{E}\sum_{i=1}^{n}\left\|x_{i,k+1}-x_{i,k}\right\|^4 \\ &+ \frac{4L_d^2}{3K_H\mu\alpha_0^3}\sum_{k=0}^{K-1}\alpha_k^3 + 8\nu^2\sum_{k=0}^{K-1}\alpha_k\tau_k.\end{aligned} \tag{E.55}$$

For the proof, write

$$\vartheta_H := \frac{4\cdot 2^{1/3}\mathcal{K}^2\sigma_H^2}{nK_H} + \frac{2^{2/3}\mathcal{K}^4L_H^2D_d}{K_H^2}. \tag{E.56}$$

Lemma E.15 bounds the two displacement sums in (E.55) by $\vartheta_H\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\mathbf{v}_k\right\|^2$. In the proof notation, Lemma E.14 reads $\mathbb{E}\left\|\mathbf{v}_k\right\|^2 \le 3e_k + 3nT_k + 6n\mathsf{G}_k + 6nL^2C_x^{\dim}\alpha_k^2$. It follows that

$$\begin{aligned}\frac{2e_K}{nK_H\alpha_K} - \frac{2e_0}{nK_H\alpha_0} \le{}& \left[\frac{-4+\mu/(3K_H\alpha_0^3)}{n} + 3\vartheta_H\right]\sum_{k=0}^{K-1}\alpha_k e_k + 3n\vartheta_H\sum_{k=0}^{K-1}\alpha_k T_k \\ &+ 6n\vartheta_H\sum_{k=0}^{K-1}\alpha_k\mathsf{G}_k + \left(\frac{4L_d^2}{3K_H\mu\alpha_0^3} + 6nL^2C_x^{\dim}\vartheta_H\right)\sum_{k=0}^{K-1}\alpha_k^3 \\ &+ 8\nu^2\sum_{k=0}^{K-1}\alpha_k\tau_k.\end{aligned} \tag{E.57}$$

Step 3: Absorption of the feedback terms. The condition $K_H \ge K_H^{\dim}$ in (E.45) is equivalent, for $K_H > 0$, to

$$K_H^2 - 180\cdot 2^{1/3}\mathcal{K}^2\sigma_H^2K_H - 45n\,2^{2/3}\mathcal{K}^4L_H^2D_d \ge 0,$$

and dividing by $45nK_H^2$ yields $\vartheta_H \le 1/(45n)$. Choose

$$\mu := 3K_H\alpha_0^3\left(\frac{5}{2} - 3n\vartheta_H\right) > 0.$$

Then the coefficient of $\sum_k \alpha_k e_k$ in (E.57) is exactly $-3/(2n)$. Moreover, $3n\vartheta_H \le 1/15$, $6n\vartheta_H \le 2/15$, and

$$\frac{4L_d^2}{3K_H\mu\alpha_0^3} = \frac{4L_d^2}{9K_H^2\alpha_0^6(5/2-3n\vartheta_H)} \le \frac{40L_d^2}{219K_H^2\alpha_0^6},$$

since $5/2 - 3n\vartheta_H \ge 73/30$. Substituting these estimates into (E.57) and expanding e_k, T_k and G_k gives (E.51). □

Lemma E.17 (Descent with estimator cancellation)**.** *Suppose the conditions of Lemma E.15 hold. Then, for every integer $K \geq 1$,*

$$\frac{7}{60}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\operatorname{grad}F(\bar{x}_k)\right\|^2 \leq \Delta_0 + \frac{2}{nK_H\alpha_0}\mathbb{E}\left\|\mathbf{d}_0-\mathbf{g}_0\right\|^2 + \frac{47}{30n}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\mathbf{s}_k-\widehat{\mathbf{d}}_k\right\|^2 + \left(\mathcal{G}_3(C_x^{\dim})+\mathcal{R}_{\mathrm{e}}\right)\sum_{k=0}^{K-1}\alpha_k^3 + \mathcal{G}_4(C_x^{\dim})\sum_{k=0}^{K-1}\alpha_k^4 + 8\nu^2\sum_{k=0}^{K-1}\alpha_k\tau_k. \tag{E.58}$$

Proof. For this proof, let $\mathcal{V}_k := \mathbb{E}F(\bar{x}_k) + 2e_k/(nK_H\alpha_k)$. Condition (E.45) contains $\alpha_0 \leq 3/(16L)$, and the common initialization and the schedule (E.47) give the hypotheses of Lemma E.11. Taking expectation in (E.32), summing, and adding (E.51) cancels the accumulated estimator-error terms and yields

$$\mathcal{V}_K - \mathcal{V}_0 \leq -\frac{7}{60}\sum_{k=0}^{K-1}\alpha_k\mathsf{G}_k + \frac{47}{30}\sum_{k=0}^{K-1}\alpha_k T_k + \left(\mathcal{G}_3(C_x^{\dim})+\mathcal{R}_{\mathrm{e}}\right)\sum_{k=0}^{K-1}\alpha_k^3 + \mathcal{G}_4(C_x^{\dim})\sum_{k=0}^{K-1}\alpha_k^4 + 8\nu^2\sum_{k=0}^{K-1}\alpha_k\tau_k. \tag{E.59}$$

The coefficients are $1/4 - 2/15 = 7/60$ and $3/2 + 1/15 = 47/30$. Since $\mathcal{V}_K \geq F_*$, rearranging and expanding the initial potential proves the result. □

Lemma E.18 (Accumulated gradient bound)**.** *Under the hypotheses of Theorem 6.1, for every integer $K \geq 1$,*

$$\frac{7}{60}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}\left\|\operatorname{grad}F(\bar{x}_k)\right\|^2 \leq \mathcal{C}_0 + \mathcal{C}_3\sum_{k=0}^{K-1}\alpha_k^3 + \mathcal{G}_4(C_x^{\dim})\sum_{k=0}^{K-1}\alpha_k^4 + \mathcal{C}_5\sum_{k=0}^{K-1}\alpha_k^5. \tag{E.60}$$

where, with $A_3 := \frac{1}{3}C_W(3L^2+2\sigma_H^2)C_\Delta^{\dim}$ and $A_5 := C_W L_H^2 n(C_\Delta^{\dim})^2$, the finite constants, independent of K, are

$$\begin{aligned}\mathcal{C}_0 &:= \Delta_0 + \frac{2\nu^2}{K_H\alpha_0} + \frac{47\Theta_0\alpha_0}{30n},\\ \mathcal{C}_3 &:= \mathcal{G}_3(C_x^{\dim}) + \mathcal{R}_{\mathrm{e}} + \frac{47}{30}A_3 + 16K_H\nu^2,\\ \mathcal{C}_5 &:= \frac{47}{30}\left(A_5 + 16C_W K_H^2 L_d^2 + \frac{8}{3}C_W K_H^2\nu^2\right).\end{aligned} \tag{E.61}$$

Proof of Lemma E.18 and Theorem 6.1. Condition (E.45) ensures that Lemmas E.13–E.17 apply simultaneously. Lemma E.13, divided by n, gives

$$\sum_{k=0}^{K-1}\alpha_k T_k \leq \frac{\Theta_0\alpha_0}{n} + \frac{C_W}{n}\sum_{k=0}^{K-1}\alpha_k\tau_k^2 e_k + A_3\sum_{k=0}^{K-1}\alpha_k^3 + A_5\sum_{k=0}^{K-1}\alpha_k^5 + \frac{2C_W\nu^2}{3}\sum_{k=0}^{K-1}\alpha_k\tau_k^2.$$

Since $e_k \leq 4nL_d^2$ and $\alpha_{k+1} \leq \alpha_k$, $\alpha_k\tau_k \leq 2K_H\alpha_k^3$ and $\alpha_k\tau_k^2 \leq 4K_H^2\alpha_k^5$. Moreover, by Lemma E.3 and Lemma D.1, $\mathbb{E}\left\|\mathbf{d}_0-\mathbf{g}_0\right\|^2 \leq \mathbb{E}\left\|\mathbf{q}_0-\mathbf{g}_0\right\|^2 \leq n\nu^2$, so the initial term of (E.58) is at most $2\nu^2/(K_H\alpha_0)$. Substituting these bounds into (E.58) yields (E.60) with the constants (E.61).

Integral comparison gives

$$\sum_{k=0}^{K-1} \alpha_k^3 \le \alpha_0^3(1+\log K), \qquad \sum_{k=0}^{K-1} \alpha_k^4 \le 4\alpha_0^4, \qquad \sum_{k=0}^{K-1} \alpha_k^5 \le \frac{5}{2}\alpha_0^5.$$

Set $m := \lceil K/2 \rceil$. For $K \ge 2$, $\lfloor K/2 \rfloor \ge K/3$, and therefore

$$\lfloor K/2 \rfloor \, \alpha_{K-1} = \frac{\alpha_0 \lfloor K/2 \rfloor}{K^{1/3}} \ge \frac{\alpha_0 K^{2/3}}{3}. \tag{E.62}$$

The accumulated gradient terms are nonnegative, so

$$\mathbb{E}\left\|\operatorname{grad} F(\bar{x}_{\hat{k}})\right\|^2 = \frac{1}{\lfloor K/2 \rfloor} \sum_{k=m}^{K-1} \mathsf{G}_k \le \frac{\sum_{k=m}^{K-1} \alpha_k \mathsf{G}_k}{\lfloor K/2 \rfloor \, \alpha_{K-1}}.$$

Combining this inequality with (E.60), (E.62), and the integral estimates proves (6.2). For every $k \in \{m, \dots, K-1\}$, Theorem E.7 gives

$$X_k \le C_x^{\dim} \alpha_k^2 = \frac{C_x^{\dim} \alpha_0^2}{(k+1)^{2/3}} \le \frac{2^{2/3} C_x^{\dim} \alpha_0^2}{K^{2/3}}.$$

Averaging this bound over the independent uniform output index proves (6.3). Thus both expected residuals in Definition 3.2 are bounded for the same output. Solving the bound of order $(1+\log K)K^{-2/3}$ for K proves the stated oracle complexity, since each node makes K stochastic-gradient queries and $K-1$ Hessian–vector-product queries, including initialization. □

E.2.3 Constant stepsize

Throughout this regime, δ satisfies (E.3), the stepsize is $\alpha_k \equiv \alpha$, and $\tau_k \equiv \tau := 2K_H\alpha^2$. Use C_x^{con} from (E.19) and set

$$C_\Delta^{\mathrm{con}} := 8(1+\sigma_2)^2 C_x^{\mathrm{con}} + 8D_d, \tag{E.63}$$

and the coefficients of the one-step recursions and of the Lyapunov function are

$$\begin{aligned} A_{e,s} &:= 16\sigma_H^2 + 4nC_\Delta^{\mathrm{con}} L_H^2, \qquad A_{t,s} := 8L^2 + \frac{16}{3}\sigma_H^2 + 8nC_\Delta^{\mathrm{con}} L_H^2, \\ \gamma_e &:= 2, \qquad \gamma_t := \frac{4}{1-\sigma_2^2}, \qquad \gamma_x := \frac{2}{1-\varrho^2}\left(1 + 2(1+\sigma_2)^2 A_{e,s}\right), \\ C_{\mathrm{des}} &:= \mathcal{G}_3(C_x^{\mathrm{con}}) + \mathcal{G}_4(C_x^{\mathrm{con}}), \qquad C_\Phi := C_{\mathrm{des}} + 8K_H\nu^2 + \frac{16}{3} C_W K_H^2 \nu^2. \end{aligned} \tag{E.64}$$

The momentum threshold and the admissible stepsize are

$$\begin{aligned} K_H^{\mathrm{con}} &:= \max\left\{1,\, 8\left(2A_{e,s} + \frac{\gamma_x \ell^2 (1+\varrho^2)}{1-\varrho^2}\right)\right\}, \\ \bar{\alpha} &:= \min\left\{1,\, \frac{(R(1-\sigma_2)-2\delta)\delta}{R\sqrt{nD_d}},\, \frac{1}{32L},\, \frac{1}{\sqrt{128Q^2 D_d(3+L)}},\, \frac{1}{\sqrt{2K_H}},\right. \\ &\qquad \left. \frac{1}{(32C_W K_H^2)^{1/4}},\, \frac{1}{\sqrt{8K_H C_W (1+\sigma_2)^2 A_{t,s}}},\, \frac{1}{\sqrt{32 C_W A_{t,s}}}\right\}, \end{aligned} \tag{E.65}$$

so that the parameters are chosen in the order δ, then $K_H \ge K_H^{\mathrm{con}}$, then $\alpha \le \bar{\alpha}$. The thresholds depend on the horizon neither directly nor through α. The bound $\alpha \le \bar{\alpha}$ contains (E.13), $\alpha \le 3/(16L)$, $\alpha \le 1$ and $\tau = 2K_H\alpha^2 \le 1$. All results of Subsections E.1 and E.2.1 therefore apply with $\tau_k \equiv \tau$.

Lemma E.19 (Coupled Lyapunov recursion)**.** *Suppose Assumptions 4.1–5.2 hold and let the iterates be generated by Algorithm 1 with $\alpha_k \equiv \alpha$ and $\tau_k \equiv \tau = 2K_H\alpha^2$, where $x_{1,0} = \cdots = x_{n,0}$, $\mathbf{s}_0 = \mathbf{d}_0$, δ satisfies* (E.3)*, $K_H \ge K_H^{\mathrm{con}}$, and $0 < \alpha \le \bar{\alpha}$. Then the Lyapunov function*

$$\Phi_k := \mathbb{E}F(\bar{x}_k) + \frac{\gamma_e \alpha}{n\tau}\mathbb{E}\left\|\mathbf{d}_k - \mathbf{g}_k\right\|^2 + \frac{\gamma_t \alpha}{n}\mathbb{E}\left\|\mathbf{s}_k - \widehat{\mathbf{d}}_k\right\|^2 + \frac{\gamma_x \alpha}{n\tau}\mathbb{E}\left\|\mathbf{x}_k - \bar{\mathbf{x}}_k\right\|^2 \tag{E.66}$$

satisfies, for every $k \geq 0$,

$$\begin{aligned}\Phi_{k+1} - \Phi_k \leq & -\frac{\alpha}{4}\mathbb{E}\left\|\operatorname{grad} F(\bar{x}_k)\right\|^2 - \frac{\alpha}{8n}\mathbb{E}\left\|\mathbf{s}_k\right\|^2 - \frac{\alpha}{4n}\mathbb{E}\left\|\mathbf{d}_k - \mathbf{g}_k\right\|^2 \\ & - \frac{\alpha}{2n}\mathbb{E}\left\|\mathbf{s}_k - \widehat{\mathbf{d}}_k\right\|^2 - \frac{\alpha}{2n\tau}\mathbb{E}\left\|\mathbf{x}_k - \bar{\mathbf{x}}_k\right\|^2 + C_\Phi \alpha^3.\end{aligned} \tag{E.67}$$

Proof. Step 1: One-step error recursions. By (E.14), $\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\| \leq \varrho\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| + \ell\alpha\|\mathbf{v}_k\|$ holds. Young's inequality with parameter $(1-\varrho^2)/(2\varrho^2)$, the case $\varrho = 0$ being immediate, and $\|\mathbf{v}_k\| \leq \|\mathbf{s}_k\|$ yield

$$X_{k+1} \leq \frac{1+\varrho^2}{2}X_k + \frac{\ell^2(1+\varrho^2)}{1-\varrho^2}\alpha^2 S_k. \tag{E.68}$$

Theorem E.7 gives the pointwise bound (E.19). Lemma E.8 gives

$$\frac{1}{n}\mathbb{E}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2 \leq 8(1+\sigma_2)^2 X_k + 8\alpha^2 S_k, \tag{E.69}$$

With (E.19) and Lemma E.5 it also gives the pathwise bound $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \leq nC_\Delta^{\mathrm{con}}\alpha^2$. Using $\sum_i \|x_{i,k+1} - x_{i,k}\|^4 \leq \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^4$, we therefore obtain

$$\frac{1}{n}\mathbb{E}\sum_{i=1}^n \|x_{i,k+1} - x_{i,k}\|^4 \leq 8nC_\Delta^{\mathrm{con}}\alpha^2\big((1+\sigma_2)^2 X_k + \alpha^2 S_k\big). \tag{E.70}$$

Since $\tau \in (0,1]$, Lemma E.12 applies with $\tau_k \equiv \tau$. Dividing (E.33) by n, substituting (E.69)–(E.70), and using $\tau = 2K_H\alpha^2$ and $K_H \geq 1$, we obtain

$$E_{k+1} \leq (1-\tau)E_k + A_{e,s}\big((1+\sigma_2)^2 X_k + \alpha^2 S_k\big) + 2\tau^2\nu^2. \tag{E.71}$$

Indeed, $\frac{L_H^2}{\tau}\cdot 8nC_\Delta^{\mathrm{con}}\alpha^2 = \frac{4nC_\Delta^{\mathrm{con}}L_H^2}{K_H} \leq 4nC_\Delta^{\mathrm{con}}L_H^2$ and $2\sigma_H^2 \cdot 8 = 16\sigma_H^2$. Likewise, substituting (E.34) into (E.35), dividing by n, and using (E.69)–(E.70) and $\alpha \leq 1$ yield

$$T_{k+1} \leq \frac{1+\sigma_2^2}{2}T_k + \frac{3(1+\sigma_2^2)}{1-\sigma_2^2}\Big[\tau^2 E_k + A_{t,s}\big((1+\sigma_2)^2 X_k + \alpha^2 S_k\big) + \frac{2}{3}\tau^2\nu^2\Big]. \tag{E.72}$$

Step 2: Descent with the direction term retained. Taking expectation in the pathwise descent inequality (E.25), applying (E.19) at iterations k and $k+1$, Lemma E.5 in the form $\mathbb{E}\|\mathbf{s}_k\|^4 \leq nD_d\mathbb{E}\|\mathbf{s}_k\|^2 = n^2 D_d S_k$, and $\alpha \leq 1$, we obtain

$$\begin{aligned}\mathbb{E}[F(\bar{x}_{k+1})] - \mathbb{E}[F(\bar{x}_k)] \leq & -\frac{\alpha}{4}\mathsf{G}_k + \frac{3\alpha}{2}(E_k + T_k) + C_{\mathrm{des}}\alpha^3 \\ & + \alpha\Big[-\frac{3}{8} + 2L\alpha + 8Q^2 D_d(3+L\alpha)\alpha^2\Big]S_k,\end{aligned} \tag{E.73}$$

where the consensus terms of (E.25) are bounded by $\mathcal{G}_3(C_x^{\mathrm{con}})\alpha^3 + \mathcal{G}_4(C_x^{\mathrm{con}})\alpha^4 \leq C_{\mathrm{des}}\alpha^3$. The functions $\mathcal{G}_3$ and $\mathcal{G}_4$ include the nonnegative terms $24Q^2D_d^2$ and $8Q^2LD_d^2$. We use this common upper bound while retaining the direction-dependent quartic term in the bracket. The stepsize bound (E.65) ensures $2L\alpha \leq \frac{1}{16}$ and $8Q^2D_d(3+L\alpha)\alpha^2 \leq \frac{1}{16}$, and therefore reduces (E.73) to

$$\mathbb{E}[F(\bar{x}_{k+1})] - \mathbb{E}[F(\bar{x}_k)] \leq -\frac{\alpha}{4}\mathsf{G}_k - \frac{\alpha}{4}S_k + \frac{3\alpha}{2}(E_k + T_k) + C_{\mathrm{des}}\alpha^3. \tag{E.74}$$

Step 3: Lyapunov coupling. The weights in (E.64) satisfy

$$\gamma_e = 2, \qquad \gamma_t\frac{1-\sigma_2^2}{2} = 2, \qquad \gamma_t\frac{3(1+\sigma_2^2)}{1-\sigma_2^2} = 2C_W,$$
$$\gamma_x\frac{1-\varrho^2}{2} = 2(1+\sigma_2)^2 A_{e,s} + 1.$$

Multiply (E.71), (E.72), and (E.68), written as bounds on $E_{k+1}-E_k$, $T_{k+1}-T_k$, and $X_{k+1}-X_k$, by $\gamma_e\alpha/\tau$, $\gamma_t\alpha$, and $\gamma_x\alpha/\tau$, respectively, and add them to (E.74). After substituting the identities above, we obtain

$$\begin{aligned}\Phi_{k+1}-\Phi_k \le\ & -\frac{\alpha}{4}\mathsf{G}_k+\alpha\left(-\frac{1}{2}+2C_W\tau^2\right)E_k-\frac{\alpha}{2}T_k\\ &+\frac{\alpha}{\tau}\left[-1+2C_W(1+\sigma_2)^2A_{t,s}\tau\right]X_k+C_\Phi\alpha^3\\ &+\alpha\left[-\frac{1}{4}+\frac{A_{e,s}}{K_H}+\frac{\gamma_x\ell^2(1+\varrho^2)}{2K_H(1-\varrho^2)}+2C_WA_{t,s}\alpha^2\right]S_k,\end{aligned}\tag{E.75}$$

where the remainder is $C_{\mathrm{des}}\alpha^3+8K_H\nu^2\alpha^3+\frac{16}{3}C_WK_H^2\nu^2\alpha^5\le C_\Phi\alpha^3$, by $\tau=2K_H\alpha^2$ and $\alpha\le 1$. Together with $\tau=2K_H\alpha^2$, the bounds in (E.65) imply

$$2C_W\tau^2\le\frac{1}{4},\qquad 2C_W(1+\sigma_2)^2A_{t,s}\tau\le\frac{1}{2},\qquad 2C_WA_{t,s}\alpha^2\le\frac{1}{16}.$$

Thus the coefficients of E_k and X_k in (E.75) are at most $-\alpha/4$ and $-\alpha/(2\tau)$, respectively. Finally, $K_H\ge K_H^{\mathrm{con}}$ gives

$$-\frac{1}{4}+\frac{A_{e,s}}{K_H}+\frac{\gamma_x\ell^2(1+\varrho^2)}{2K_H(1-\varrho^2)}+2C_WA_{t,s}\alpha^2\le-\frac{1}{8}.$$

Substitution proves (E.67). □

For a horizon K and a scale $a\in(0,\bar\alpha]$, the theorem below uses the three additional constants

$$\begin{aligned}\Delta_{\mathcal{M}}:=\max_{x\in\mathcal{M}}F(x)-F_*,\qquad C_E:=\frac{12\Delta_0}{a^2}+\frac{6\gamma_e\nu^2}{K_Ha^3}+\frac{12\gamma_tL_d^2}{a}+4C_\Phi a,\\ \mathcal{B}:=\Delta_{\mathcal{M}}+\frac{\gamma_eC_E}{2K_H}+\gamma_taD_d+\frac{\gamma_xC_x^{\mathrm{con}}a}{2K_H},\end{aligned}\tag{E.76}$$

none of which depends on K. The constant $\Delta_{\mathcal{M}}$ replaces the initial gap Δ_0 because the Lyapunov estimate starts at an index in the first half of the run. Compactness bounds the objective gap at that index by $\Delta_{\mathcal{M}}$.

Proof of Theorem 6.2. Step 1: The summed inequality and the required bound on the potential. For any deterministic index $j\in\{0,\dots,K-1\}$, summing (E.67) from $k=j$ to $K-1$ and using $\Phi_K\ge F_*$ gives

$$\begin{aligned}&\frac{\alpha}{4}\sum_{k=j}^{K-1}\mathsf{G}_k+\frac{\alpha}{8}\sum_{k=j}^{K-1}S_k+\frac{\alpha}{4}\sum_{k=j}^{K-1}E_k+\frac{\alpha}{2}\sum_{k=j}^{K-1}T_k+\frac{\alpha}{2\tau}\sum_{k=j}^{K-1}X_k\\ &\qquad\le\Phi_j-F_*+C_\Phi(K-j)\alpha^3.\end{aligned}\tag{E.77}$$

Since $K\alpha^3=a^3$, the desired rate follows if an index $j<K/2$ can be chosen such that Φ_j-F_* is bounded independently of K. With single-sample initialization, $E_0=O(1)$, so its contribution $\gamma_e\alpha E_0/\tau=\gamma_eE_0/(2K_H\alpha)$ scales as $1/\alpha$. The next step uses the first half of the run to obtain $E_j=O(\alpha)$ and bound all three weighted error terms in Φ_j at the same index.

Step 2: A common index with a bounded Lyapunov function. The initialization of Algorithm 1, Lemma E.3, and Lemma D.1 give

$$E_0\le\nu^2,\qquad T_0\le L_d^2,\qquad X_0=0.\tag{E.78}$$

Indeed, clipping fixes the true gradient because $\|g_{i,0}\|\le L_d$; also, $\mathbf{s}_0=\mathbf{d}_0$ and $\widehat{\mathbf{d}}_0=(J\otimes I)\mathbf{d}_0$, so $T_0\le n^{-1}\mathbb{E}\,\|\mathbf{d}_0\|^2\le L_d^2$; common initialization implies $X_0=0$.

Let $N:=\lfloor K/2\rfloor$. Sum (E.67) over $k=0,\dots,N-1$ and use $\Phi_N\ge F_*$. The remaining sums are nonnegative, so we keep only the estimator-error sum on the left-hand side. Then

$$\frac{\alpha}{4}\sum_{k=0}^{N-1}E_k\le\Phi_0-F_*+C_\Phi N\alpha^3\le\Delta_0+\frac{\gamma_e\nu^2}{2K_H\alpha}+\gamma_t\alpha L_d^2+C_\Phi N\alpha^3.\tag{E.79}$$

Because $N \geq K/3$, $K\alpha^3 = a^3$, and $\alpha \leq a$, division by $N\alpha/4$ yields

$$\frac{1}{N}\sum_{k=0}^{N-1} E_k \leq \frac{4\Delta_0}{N\alpha} + \frac{2\gamma_e\nu^2}{K_H N\alpha^2} + \frac{4\gamma_t L_d^2}{N} + 4C_\Phi\alpha^2$$
$$\leq \frac{12\Delta_0\alpha^2}{a^3} + \frac{6\gamma_e\nu^2\alpha}{K_H a^3} + \frac{12\gamma_t L_d^2\alpha^3}{a^3} + 4C_\Phi\alpha^2 \leq C_E\alpha,$$

with C_E as in (E.76). Select a deterministic minimizer of these expected errors, $j \in \operatorname{argmin}_{0\leq k<N} E_k$. Thus

$$E_j \leq C_E\alpha, \qquad j < N \leq K/2. \tag{E.80}$$

The index is used only in the proof; the algorithm continues without a restart or additional oracle calls.

The remaining terms are bounded at this same index. By Lemma E.4, $\widehat{\mathbf{d}}_k = (J\otimes I)\mathbf{s}_k$. Orthogonality and Lemma E.5 then give $T_k \leq S_k \leq D_d$ for every k. The consensus estimate (E.19) gives $X_k \leq C_x^{\mathrm{con}}\alpha^2$. Consequently,

$$\frac{\gamma_e\alpha}{\tau}E_j \leq \frac{\gamma_e C_E}{2K_H}, \qquad \gamma_t\alpha T_j \leq \gamma_t a D_d, \qquad \frac{\gamma_x\alpha}{\tau}X_j \leq \frac{\gamma_x C_x^{\mathrm{con}} a}{2K_H}.$$

Compactness of $\mathcal{M}$ and continuity of F give $\mathbb{E}[F(\bar{x}_j)] - F_* \leq \Delta_{\mathcal{M}} < \infty$. Combining these four bounds in (E.66) and using (E.76) establishes

$$\Phi_j - F_* \leq \mathcal{B}. \tag{E.81}$$

We have $E_j = O(\alpha)$ is sufficient, because its weight in the potential is $\alpha/\tau = 1/(2K_H\alpha)$.

Step 3: Stationarity, consensus, and oracle complexity. Return to (E.77), discard the other nonnegative sums, and use (E.81). We obtain

$$\frac{\alpha}{4}\sum_{k=j}^{K-1}\mathsf{G}_k \leq \mathcal{B} + C_\Phi K\alpha^3 = \mathcal{B} + C_\Phi a^3.$$

Since $j < \lceil K/2\rceil$, the sum over the second half is smaller, and $K - \lceil K/2\rceil = \lfloor K/2\rfloor \geq K/3$ for $K \geq 2$. Hence

$$\frac{1}{\lfloor K/2\rfloor}\sum_{k=\lceil K/2\rceil}^{K-1}\mathsf{G}_k \leq \frac{4(\mathcal{B} + C_\Phi a^3)}{\alpha\lfloor K/2\rfloor} \leq \frac{12\big(\mathcal{B} + C_\Phi a^3\big)}{aK^{2/3}}.$$

At every iteration, (E.19) also gives

$$X_k \leq C_x^{\mathrm{con}}\alpha^2 = \frac{C_x^{\mathrm{con}}a^2}{K^{2/3}}, \qquad 0 \leq k \leq K,$$

Averaging both errors over the independent uniform output index in the second half of the run proves (6.5) with the stated $\mathcal{C}_{\mathrm{opt}}$.

The initialization requires one stochastic-gradient evaluation per node, and each subsequent iteration requires one stochastic-gradient evaluation and one Hessian correction. The index selected in the proof does not affect these counts. Hence, both oracle counts are $O(K)$, and choosing $K = \lceil(\mathcal{C}_{\mathrm{opt}}/\epsilon)^{3/2}\rceil = O(\epsilon^{-3/2})$ yields the claimed complexities. The dependence of $\mathcal{C}_{\mathrm{opt}}$ on the stepsize scale includes $144\Delta_0/(K_H a^3)$ and $144\nu^2/(K_H^2 a^4)$. □

E.2.4 Constant parameters and the exact-gradient regime

This subsection proves Theorem 6.3. The stepsize and the momentum parameter are constant, $\alpha_k \equiv \alpha$ and $\tau_k \equiv \tau \in (0,1]$, and α is not tied to τ. Under Assumption 4.5 the iterates are those of Algorithm 1, and the single-sample initialization gives $e_0 \leq n\nu^2$ by (E.78). The theorem also covers a mixed oracle.

In the exact-gradient regime, each node evaluates its exact local gradient $\operatorname{grad} f_i(x)$. The correction remains stochastic. For every $\mathcal{F}_k$-measurable $x \in \mathcal{M}$ and $u \in T_x\mathcal{M}$, a fresh sample $\zeta_{i,k}$ returns $\mathcal{H}_i(x,\zeta_{i,k})[u]$ with

$$\mathbb{E}[\mathcal{H}_i(x,\zeta_{i,k})[u] \mid \mathcal{F}_k] = \overline{\operatorname{Hess}}\, f_i(x)[u],$$
$$\mathbb{E}\left[\left\|\mathcal{H}_i(x,\zeta_{i,k})[u] - \overline{\operatorname{Hess}}\, f_i(x)[u]\right\|^2 \middle| \mathcal{F}_k\right] \leq \sigma_H^2\|u\|^2. \tag{E.82}$$

The samples are independent across nodes and iterations. These are the oracle properties of the correction (D.7)–(D.8). The gradient is exact, so the two gradient properties of Lemma D.1 hold with $\nu^2 = 0$. When the exact ambient gradient $\nabla f_i(x)$ is available, one realization is

$$\mathcal{H}_i(x, \zeta)[u] = D(\mathcal{P}_{T\cdot\mathcal{M}})(x)[u]\nabla f_i(x) + \mathcal{P}_{T_x\mathcal{M}}\nabla^2 F_i(x; \zeta)[u].$$

The first term is deterministic conditional on $\mathcal{F}_k$; the second is unbiased and has variance at most $\sigma^2_{\nabla^2,i} \|u\|^2$, since orthogonal projection is nonexpansive. Thus one exact ambient gradient supplies both $\operatorname{grad} f_i(x)$ and the projector-derivative term, and one sampled HVP supplies the remaining term. The fully sampled correction $\overline{\operatorname{Hess}} F_i(x; \zeta)[u]$ also satisfies (E.82), with the bound in (D.9); it uses a sampled gradient together with the sampled HVP. The estimator is initialized by $q_{i,0} = \operatorname{grad} f_i(x_{i,0})$ and, for $k \geq 1$, updated by

$$q_{i,k} = (1 - \tau)\left(d_{i,k-1} + \mathcal{H}_i(x_{i,k}, \zeta_{i,k})[u_{i,k}]\right) + \tau \operatorname{grad} f_i(x_{i,k}) \tag{E.83}$$

with a fixed $\tau \in (0, 1]$. The clipping, tracking, tangent projection and manifold projection steps are unchanged. The analysis of Subsections E.1 and E.2.1 uses the oracle only through Lemma D.1, so all its bounds apply with $\nu^2 = 0$.

Throughout this regime, δ satisfies (E.3), $\alpha_k \equiv \alpha$, and $\tau_k \equiv \tau$. We use the constants C_c, $\mathcal{K}$ of (E.4), $C_x^{\mathrm{con}} = D_d C_c$ of (E.19), and

$$\Gamma_2 := \frac{\mathcal{G}_3(C_x^{\mathrm{con}})}{D_d}, \qquad \Gamma_3 := \frac{\mathcal{G}_4(C_x^{\mathrm{con}})}{D_d}, \tag{E.84}$$

together with the feedback coefficients

$$\begin{aligned}
A_e^{\mathrm{ex}} &:= \frac{2\mathcal{K}^2\sigma_H^2}{\tau}\alpha^2 + \frac{\mathcal{K}^4 n D_d L_H^2}{\tau^2}\alpha^4, \\
A_t^{\mathrm{ex}} &:= C_W\tau^2 A_e^{\mathrm{ex}} + \frac{C_W(3L^2 + 2\sigma_H^2)\mathcal{K}^2}{3}\alpha^2 + C_W\mathcal{K}^4 n D_d L_H^2\alpha^4, \\
P_\tau &:= (1 + C_W\tau^2)\left(\frac{2\mathcal{K}^2\sigma_H^2}{\tau} + \frac{\mathcal{K}^4 n D_d L_H^2}{\tau^2}\right) + \frac{C_W(3L^2 + 2\sigma_H^2)\mathcal{K}^2}{3} + C_W\mathcal{K}^4 n D_d L_H^2.
\end{aligned} \tag{E.85}$$

For $\alpha \leq 1$, we have $A_e^{\mathrm{ex}} + A_t^{\mathrm{ex}} \leq P_\tau\alpha^2$. The admissible stepsize is

$$\bar{\alpha}_{\tau,\delta} := \min\left\{1,\ \frac{(R(1 - \sigma_2) - 2\delta)\delta}{R\sqrt{nD_d}},\ \frac{3}{128L},\ \sqrt{\frac{3}{64\Gamma_2}},\ \left(\frac{3}{64\Gamma_3}\right)^{1/3},\ \frac{1}{\sqrt{32P_\tau}}\right\}, \tag{E.86}$$

which depends on τ and δ but not on the horizon.

Proof of Theorem 6.3. The bound $\alpha \leq \bar{\alpha}_{\tau,\delta}$ contains (E.13) and $\alpha \leq 1$, and $\tau \in (0, 1]$. Hence Lemmas E.6, E.8, E.10, E.12 and Theorem E.7 apply with $\alpha_k \equiv \alpha$ and $\tau_k \equiv \tau$. In the exact-gradient regime, Lemma E.12 applies through its last claim, with $\nu^2 = 0$. Write

$$\mathcal{S}_K := \sum_{k=0}^{K-1} \mathbb{E}\|\mathbf{s}_k\|^2 = n\sum_{k=0}^{K-1} S_k.$$

Step 1: Cumulative displacement and consensus bounds. Lemma E.9 gives (E.23)–(E.24), and Theorem E.7 gives (E.19). These bounds retain $\mathcal{S}_K$, allowing the descent inequality to absorb the accumulated correction and consensus errors.

Step 2: Estimator and tracking feedback. Under Assumption 4.5, $e_0 \leq n\nu^2$ by (E.78). In the exact-gradient regime, $L_d \geq G$ and clipping leave the initial gradient unchanged, so $d_{i,0} = \operatorname{grad} f_i(x_{i,0})$ and $e_0 = 0$. In both cases $e_0 \leq n\nu^2$. Summing (E.33) over $k = 0, \ldots, K - 1$, discarding $e_K \geq 0$, dividing by τ, and using (E.23) yield

$$\begin{aligned}
\sum_{k=0}^{K-1} e_k &\leq \frac{e_0}{\tau} + \frac{2\sigma_H^2}{\tau}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 + \frac{L_H^2}{\tau^2}\sum_{k=0}^{K-1}\mathbb{E}\sum_{i=1}^{n}\|x_{i,k+1} - x_{i,k}\|^4 + 2n\tau\nu^2 K \\
&\leq \frac{n\nu^2}{\tau} + A_e^{\mathrm{ex}}\mathcal{S}_K + 2n\tau\nu^2 K.
\end{aligned} \tag{E.87}$$

Dividing (E.36) by α and inserting (E.23) and (E.87) give

$$\begin{aligned}\sum_{k=0}^{K-1} nT_k &\le \Theta_0 + C_W\tau^2 \sum_{k=0}^{K-1} e_k + \frac{C_W(3L^2+2\sigma_H^2)}{3}\sum_{k=0}^{K-1}\mathbb{E}\left\|\mathbf{x}_{k+1}-\mathbf{x}_k\right\|^2 \\ &\quad + C_W L_H^2 \sum_{k=0}^{K-1}\mathbb{E}\sum_{i=1}^{n}\left\|x_{i,k+1}-x_{i,k}\right\|^4 + \frac{2C_W n\tau^2\nu^2}{3}K \\ &\le \Theta_0 + A_t^{\mathrm{ex}}\mathcal{S}_K + C_W n\tau\nu^2 + \left(2C_W\tau^3 + \frac{2C_W\tau^2}{3}\right) n\nu^2 K.\end{aligned} \tag{E.88}$$

The constant Θ_0 accounts for the initial disagreement among the local gradients retained by the initialization $\mathbf{s}_0 = \mathbf{d}_0$. Adding the two bounds and using $\tau \le 1$, we obtain

$$\sum_{k=0}^{K-1}(e_k + nT_k) \le \Theta_0 + \left(A_e^{\mathrm{ex}} + A_t^{\mathrm{ex}}\right)\mathcal{S}_K + \frac{(1+C_W)n\nu^2}{\tau} + \left(2 + \frac{8C_W}{3}\right) n\tau\nu^2 K. \tag{E.89}$$

Step 3: Summed descent inequality. Take expectations in (E.25) and sum over $k = 0, \dots, K-1$. For each fourth-power consensus term, bound one factor $\left\|\mathbf{x}_j - \bar{\mathbf{x}}_j\right\|^2$ by (E.19) and the remaining sum by (E.24). The second-power consensus term is bounded by (E.24) directly, and $\left\|\mathbf{s}_k\right\|^4 \le nD_d\left\|\mathbf{s}_k\right\|^2$. Since $C_x^{\mathrm{con}}C_c = (C_x^{\mathrm{con}})^2/D_d$ and $C_c = C_x^{\mathrm{con}}/D_d$, the resulting coefficients are exactly $\Gamma_2\alpha^3$ and $\Gamma_3\alpha^4$, and we obtain

$$\begin{aligned}\mathbb{E}F(\bar{x}_K) - F(\bar{x}_0) \le\ &-\frac{\alpha}{4}\sum_{k=0}^{K-1}\mathbb{E}\left\|\operatorname{grad} F(\bar{x}_k)\right\|^2 + \frac{3\alpha}{2n}\sum_{k=0}^{K-1}(e_k + nT_k) \\ &-\frac{\alpha}{n}\left[\frac{3}{8} - 2L\alpha - \Gamma_2\alpha^2 - \Gamma_3\alpha^3\right]\mathcal{S}_K.\end{aligned} \tag{E.90}$$

Insert (E.89) and set

$$\gamma^{\mathrm{ex}} := \frac{3}{8} - 2L\alpha - \Gamma_2\alpha^2 - \Gamma_3\alpha^3 - \frac{3}{2}\left(A_e^{\mathrm{ex}} + A_t^{\mathrm{ex}}\right). \tag{E.91}$$

By (E.86), each of $2L\alpha$, $\Gamma_2\alpha^2$, $\Gamma_3\alpha^3$, and $\frac{3}{2}(A_e^{\mathrm{ex}} + A_t^{\mathrm{ex}}) \le \frac{3}{2}P_\tau\alpha^2$ is at most $3/64$, so $\gamma^{\mathrm{ex}} \ge 3/16$. With $F(\bar{x}_K) \ge F_*$, (E.90) becomes

$$\begin{aligned}\frac{\alpha}{4}\sum_{k=0}^{K-1}\mathbb{E}\left\|\operatorname{grad} F(\bar{x}_k)\right\|^2 + \frac{3\alpha}{16n}\mathcal{S}_K &\le \mathcal{B}^{\mathrm{ex}} + (3+4C_W)\alpha\tau\nu^2 K, \\ \mathcal{B}^{\mathrm{ex}} &:= \Delta_0 + \frac{3\alpha\Theta_0}{2n} + \frac{3(1+C_W)\alpha\nu^2}{2\tau}.\end{aligned} \tag{E.92}$$

Step 4: Joint stationarity and complexity. By (E.24) and (E.92),

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\operatorname{grad} F(\bar{x}_k)\right\|^2 + \frac{1}{n}\left\|\mathbf{x}_k - \bar{\mathbf{x}}_k\right\|^2\right] &\le \frac{4}{\alpha}\left[\mathcal{B}^{\mathrm{ex}} + (3+4C_W)\alpha\tau\nu^2 K\right] + \frac{C_c\alpha^2}{n}\mathcal{S}_K \\ &\le \left(\frac{4}{\alpha} + \frac{16C_c\alpha}{3}\right)\left[\mathcal{B}^{\mathrm{ex}} + (3+4C_W)\alpha\tau\nu^2 K\right] \\ &\le \mathcal{C}_{\mathrm{ex}} + \mathcal{C}_{\mathrm{fl}}\,\tau\nu^2 K,\end{aligned}$$

where the last step uses $\alpha \le 1$. Dividing by K and drawing $\hat{k}$ independently and uniformly from $\{0, \dots, K-1\}$ proves (6.6). In the exact-gradient regime $\nu = 0$. Both terms inside the expectation are nonnegative, so this output satisfies the two requirements of Definition 3.2 once $\mathcal{C}_{\mathrm{ex}}/K \le \epsilon$. Each iteration uses one exact local-gradient evaluation and one stochastic Hessian correction per node. This gives the stated complexity. □

F Additional details and results of the numerical experiments

F.1 Terminal measures of the main comparisons

Tables 4 and 5 list the terminal measures of the comparisons of Section 7 on the synthetic PCA instance and on the LRMC instances.

The exact-gradient configuration approaches numerical precision on both problems. The diminishing configuration has the smallest joint residual among the stochastic methods; the constant configuration and DPRSRM reach similar residuals. In noisy LRMC, the distance from the generating subspace stabilizes near 0.017 for every convergent method.

Table 4: Terminal measures for the synthetic PCA instance; mean $\pm$ standard deviation over ten sampling streams.

Method	Objective gap	G_k	c_k	d_s
DPRSHCM (dim.)	$(5.3\pm1.9)\times10^{-6}$	0.0638 ± 0.012	$(2.2\pm0.2)\times10^{-6}$	$(2.4\pm1.0)\times10^{-4}$
DPRSHCM (const.)	$(3.7\pm1.5)\times10^{-5}$	0.206 ± 0.052	$(1.3\pm0.2)\times10^{-5}$	$(4.4\pm0.7)\times10^{-4}$
DPRSHCM (exact)	$(5.0\pm3.5)\times10^{-13}$	$(2.7\pm0.0)\times10^{-8}$	$(6.7\pm0.1)\times10^{-15}$	$(4.3\pm0.0)\times10^{-10}$
DPRSRM	$(4.4\pm5.6)\times10^{-5}$	0.188 ± 0.042	$(9.5\pm0.9)\times10^{-6}$	$(5.6\pm6.3)\times10^{-4}$
S-DPRGD	$(3.5\pm0.9)\times10^{-3}$	1.34 ± 0.47	$(3.0\pm0.2)\times10^{-3}$	$(9.1\pm0.6)\times10^{-3}$
DRSGD	$(3.6\pm0.6)\times10^{-3}$	1.44 ± 0.31	$(2.9\pm0.2)\times10^{-3}$	$(9.0\pm0.4)\times10^{-3}$

Table 5: Terminal measures for the LRMC problems. The synthetic objective gap is relative to the centralized reference value F_{ref}. The synthetic block reports means and standard deviations over ten random instances; the MovieLens block reports the training residual $F(\bar{X}_K)$, the gradient and consensus measures, and the held-out RMSE over three sampling streams.

	Synthetic (ten instances)			
Method	Objective gap	G_k	c_k	d_s
DPRSHCM (dim.)	$(2.0\pm0.3)\times10^{-5}$	0.328 ± 0.037	$(1.9\pm0.1)\times10^{-5}$	0.017 ± 0.0012
DPRSHCM (const.)	$(8.4\pm1.3)\times10^{-5}$	0.694 ± 0.079	$(6.3\pm0.3)\times10^{-5}$	0.017 ± 0.0012
DPRSHCM (exact)	$(2.8\pm1.9)\times10^{-15}$	$(3.4\pm6.2)\times10^{-11}$	$(2.3\pm0.2)\times10^{-15}$	0.017 ± 0.0012
DPRSRM	$(8.6\pm1.8)\times10^{-5}$	0.704 ± 0.1	$(6.3\pm0.2)\times10^{-5}$	0.017 ± 0.0012
S-DPRGD	$(3.6\pm3.0)\times10^{-3}$	4.3 ± 0.59	$(2.8\pm0.1)\times10^{-3}$	0.0171 ± 0.0014
DRSGD	$(4.7\pm6.7)\times10^{-3}$	4.41 ± 0.92	$(2.8\pm0.1)\times10^{-3}$	0.0173 ± 0.0016
	MovieLens-100K			
Method	Residual $F(\bar{X}_K)$	G_k	c_k	Held-out RMSE
DPRSHCM (dim.)	1557 ± 1.5	7.53 ± 0.52	$(1.7\pm0.1)\times10^{-4}$	0.887 ± 0.0056
DPRSHCM (const.)	1557 ± 1.5	9.74 ± 1.2	$(6.9\pm0.2)\times10^{-4}$	0.887 ± 0.0032
DPRSHCM (exact)	$1554\pm9.7\times10^{-5}$	0.555 ± 0.025	0.0161 ± 0.00026	$0.883\pm2.8\times10^{-6}$
DPRSRM	1557 ± 2.4	7.08 ± 2.2	$(3.0\pm0.0)\times10^{-4}$	0.887 ± 0.0049
S-DPRGD	1558 ± 0.12	14.8 ± 2.2	0.0652 ± 0.0021	0.891 ± 0.002
DRSGD	1558 ± 0.047	15.2 ± 0.74	0.067 ± 0.0032	0.891 ± 0.0011

F.2 Consensus and objective convergence

Figures 3 and 4 display the consensus and objective diagnostics for the comparisons in Figures 1 and 2.

F.3 Experimental setup and choice of parameters

Additional measures and synthetic data Besides G_k and c_k in Section 7, we report the objective gap $F(\bar{X}_k) - F_\star$ and the distance $d_s = \min_{Q^\top Q = I_r} \left\| \bar{X}_k Q - X_\star \right\|_F$ to a reference solution.

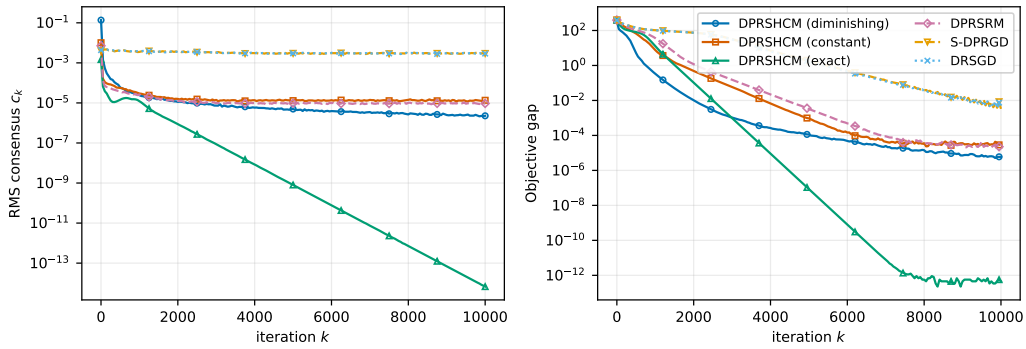

Figure 3: Synthetic PCA on the ER graph with $p = 0.6$ and $K = 10000$: consensus error and objective gap; pointwise medians over ten sampling streams.

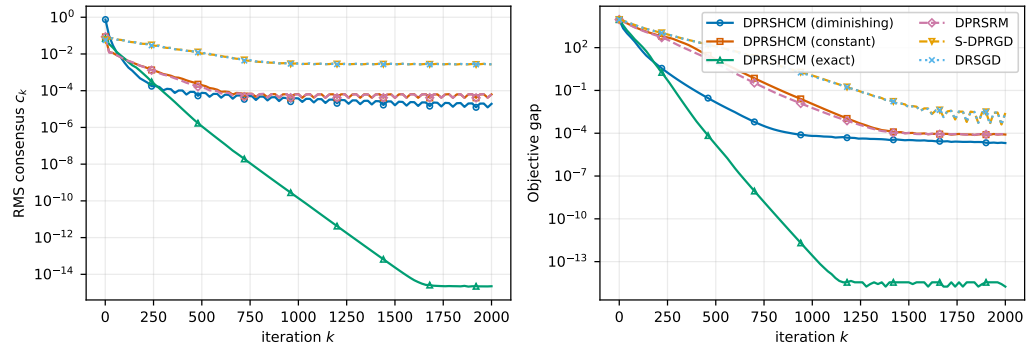

Figure 4: Synthetic LRMC on the ring with $K = 2000$: consensus error, reference objective gap $F(\bar{X}_k) - F_{\mathrm{ref}}$; pointwise medians over ten instances.

Synthetic PCA We solve the decentralized PCA problem (A.1) with local data $A_i \in \mathbb{R}^{m_i \times d}$. For the stacked matrix A the leading r right singular vectors give $X_\star$, and $F_\star = -(2n)^{-1} \sum_{j=1}^{r} \sigma_j(A)^2$.

We set $n = 8$, $m_1 = \cdots = m_n = 1000$, $d = 10$ and $r = 5$. Starting from a standard Gaussian matrix $B \in \mathbb{R}^{8000 \times 10}$ with thin singular value decomposition $B = U\Sigma V^\top$, we form

$$A = U\widetilde{\Sigma}V^\top, \qquad \widetilde{\Sigma} = \sigma_1(B)\,\mathrm{diag}(1, \gamma, \ldots, \gamma^{d-1}), \qquad \gamma = 0.8, \tag{F.1}$$

so γ controls the conditioning of the problem. The rows of A are randomly partitioned into n blocks of equal size. The budget is $K = 10000$, the network is the ER graph with $p = 0.6$, and the parameters are listed in Table 6 in Appendix F.3.

Synthetic LRMC We solve the variable-projection model (A.2), in which $V_i(X_i)$ is the minimum-norm least squares solution. This unregularized model is smooth locally where the observed-row least-squares matrices have full column rank (Appendix A). We generate ten independent instances. Standard Gaussian factors $L \in \mathbb{R}^{50 \times 10}$ and $R \in \mathbb{R}^{10 \times 960}$ give $A = LR + E$ with independent Gaussian noise of standard deviation 0.1, and every entry is observed with probability $1/2$, an oversampling ratio of about 2.4 (Boumal & Absil, 2015). Because of the noise, $X_\star = \mathrm{orth}(L)$ need not minimize the observed-data objective. We compute a reference value F_{ref} by centralized Riemannian gradient descent and report $F(\bar{X}_k) - F_{\mathrm{ref}}$ as the reference gap. The distance d_s is measured from $\mathrm{orth}(L)$. The columns are randomly partitioned among eight nodes. We use the ring network, $K = 2000$ and the parameters of Table 8 in Appendix F.3.

Cost measure All methods perform one mixing step per iteration, so k counts communication rounds and the curves below are plotted against k. Per iteration, DPRSHCM evaluates one stochastic gradient and one Hessian-based correction. DPRSRM draws two stochastic gradients, and S-

DPRGD and DRSGD draw one. Let κ be the measured cost of one Hessian correction relative to one stochastic gradient. The oracle cost of a run in equivalent gradient evaluations is

$$N_{\mathrm{eq}} := N_{\mathrm{grad}} + \kappa N_{\mathrm{hvp}}, \tag{F.2}$$

where N_{grad} and N_{hvp} count sampled rows or columns. Subsection F.8 measures κ and reports N_{eq} for every method. Section 6 counts gradient and Hessian–vector-product oracle calls separately; N_{eq} converts these counts using the measured implementation cost.

Networks and mixing matrix The mixing matrix W is the Metropolis constant edge weight matrix (Xiao & Boyd, 2004). On the ring used in the LRMC experiments every node assigns weight $1/3$ to itself and to each of its two neighbors. We use a ring and connected Erdős–Rényi graphs with edge probability $p = 0.3$ and $p = 0.6$, called ER graphs below. For the eight-node instances of Subsection F.4 the second largest singular values are $\sigma_2 = 0.805$, 0.752 and 0.778. The contraction factor ϱ in (E.4) depends on σ_2. Subsections F.4 and F.7 therefore compare the topologies through σ_2.

Sampling protocol Each node reshuffles its local samples at the beginning of every epoch and traverses them without replacement in batches of ten, except for the MNIST batch size specified below. The initial estimator uses the same batch size. In the exact-gradient setting the gradient uses the entire local dataset and the correction remains sampled. This finite-data protocol uses random reshuffling, whereas the analysis assumes independent sampling with conditional unbiasedness. Within an instance all six methods share the data partition, the initial basis, the network and the batch size. Every method draws its own sampling stream at every node, so no two methods see the same sequence of batches. The initial basis is the polar factor of a standard Gaussian matrix, so $\mathbf{x}_0 = \bar{\mathbf{x}}_0$.

Reported statistics and platform The synthetic PCA instance is solved ten times with independent sampling streams. The Ionosphere comparison uses ten random partitions of the rows, each with its own network, initial basis and sampling streams. We use a subset of 320 rows to form eight equal local blocks. The MNIST and MovieLens instances are solved three times and the heterogeneous MNIST partition twice. The synthetic LRMC comparison uses ten independent random instances. Curves report pointwise medians and tables report means and sample standard deviations over these runs. Each run contributes its last-iteration value to the terminal statistics. Curves are unsmoothed, nonpositive values are omitted on logarithmic axes, and consensus curves start at $k = 1$. All six methods are implemented in Python 3.11 with NumPy 2.4 and run in double precision with one thread per run. All experiments are performed on a cloud virtual machine with two Intel Xeon 2.8 GHz virtual cores and 8 GB of memory, under Linux.

Baselines DPRSRM is implemented as in Deng & Hu (2025b). Its stepsize and momentum parameter depend on the prescribed horizon K. S-DPRGD and DRSGD have a single stepsize. DRSGD uses the polar retraction. All three baselines perform one mixing step per iteration. We therefore compare communication rounds; the number of transmitted scalars also depends on whether a method exchanges a tracking variable.

Grid search Every method is tuned on every problem class by the same two-stage grid search. The pilot instances are separate from the test instances. For the real datasets they use independent partitions, initial bases and networks. The first stage sweeps a coarse grid of (α, τ) at $K = 1000$ on two pilot instances. For the PCA problems the stepsize grid is the half-decade grid $\alpha \in \{10^{-5}, 3 \cdot 10^{-5}, \dots, 10^{-2}\}$ and the momentum grid is $\tau \in \{0.01, 0.05, 0.1, 0.3, 0.5, 0.9, 0.99\}$. The LRMC objectives are larger, so for the LRMC problems the stepsize grid is $\{3 \cdot 10^{-6}, \dots, 3 \cdot 10^{-4}\}$. MNIST has far more samples per node, so its stepsize grid is $\{10^{-6}, \dots, 3 \cdot 10^{-5}\}$ with a coarser momentum grid, and MNIST and MovieLens use one pilot instance, with $K = 500$ for MovieLens. The best three pairs of each method are retained, two for MNIST. The second stage reruns the retained pairs and their stepsize neighbors, with factors between 0.5 and 2, together with the adjacent values of τ. It uses the budget of the reported comparison, $K = 2000$, and more pilot instances: five for the synthetic PCA and Ionosphere instances, three for the synthetic LRMC instances, two for MNIST and one for MovieLens. The synthetic PCA comparison is run at $K = 10000$, so there the retained pairs are rescreened once more at that budget on three pilot instances. The pair with the smallest mean tuning score over the pilots is selected. The tuning score for each pilot is the

median of $G_k^2 + c_k^2$ over the last ten percent of the iterations. Whenever the selected pair lies on the boundary of the explored set, the set is extended in that direction and the stage is repeated. This check is automatic and stops only when the selected pair is interior. The three settings of DPRSHCM are tuned independently. DPRSRM is tuned on the same grids as the constant setting. For the diminishing setting the search is over (α_0, τ_0). The parameters in the tables below are the outcome of this search.

Momentum and clipping The tuned pair determines $K_H = \tau_0/(2\alpha_0\alpha_1)$. Figure 5 varies $K_H \in [1.1 \times 10^4, 5.4 \times 10^6]$ at fixed α_0 on synthetic PCA. The gap is 4.6×10^{-6} at the tuned $K_H = 5.4 \times 10^5$, increasing to 9.6 at the lower endpoint and 1.1×10^{-4} at the upper endpoint. The theoretical conditions couple K_H and the stepsize as in (E.45).

Assumption 5.2 requires $L_d \geq G$. Figure 6 varies L_d from 10 to 10^8. At $L_d = 10$, clipping occurs in 92% of node updates and the gap is 23; for $L_d \geq 1000$, it never activates and all three measures coincide. We use $L_d = 10^8$ for PCA and 5×10^4 for LRMC; clipping is inactive in these comparisons.

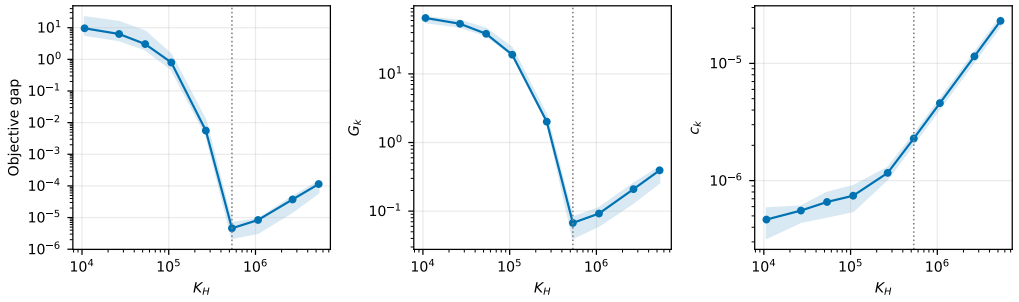

Figure 5: Sensitivity of the terminal objective gap, gradient norm and consensus error to K_H for the diminishing setting on the synthetic PCA instance, with α_0 at its tuned value and five sampling streams; the band is the range over the streams and the dotted line marks the tuned value.

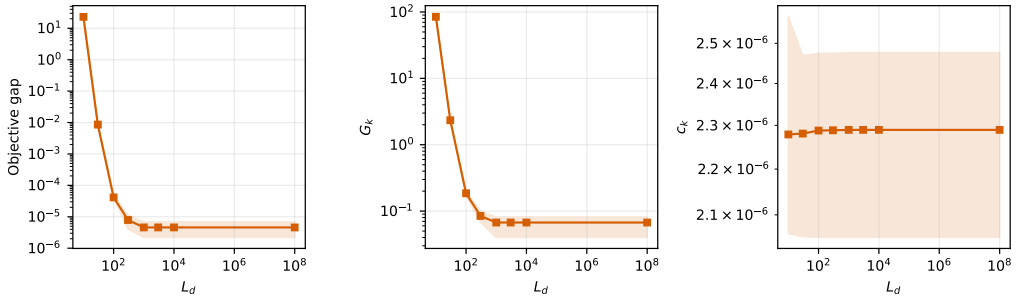

Figure 6: Sensitivity of the terminal measures to the clipping radius L_d on the synthetic PCA instance; diminishing setting, five sampling streams.

Tuned parameters Tables 6, 7 and 8 list the tuned parameters of every method on the synthetic PCA, the real PCA and the LRMC instances.

F.4 The role of the gradient-derivative correction

The correction operator (5.2) consists of two terms,

$$\overline{\mathrm{Hess}}\, F_i(y;\xi)[\omega] = D(\mathcal{P}_{T\cdot\mathcal{M}})(y)[\omega]\nabla F_i(y;\xi) + \mathcal{P}_{T_y\mathcal{M}}\nabla^2 F_i(y;\xi)[\omega].$$

The second term is the projected Hessian–vector product of the Euclidean recursive momentum methods (Tran & Cutkosky, 2022). The first term accounts for the change of the tangent space with

Table 6: Parameters for the synthetic PCA instance. Entries are (α_0, τ_0) for the diminishing setting, (α, τ) for the constant recursive methods, and the stepsize for the stochastic gradient methods.

Method	Synthetic PCA
DPRSHCM (diminishing)	$(3.4 \times 10^{-4},\ 0.1)$
DPRSHCM (constant)	$(2.4 \times 10^{-5},\ 0.001)$
DPRSHCM (exact)	$(3.7 \times 10^{-5},\ 0.9)$
DPRSRM	$(1.7 \times 10^{-5},\ 0.001)$
S-DPRGD	$(1.0 \times 10^{-5},\ –)$
DRSGD	$(1.0 \times 10^{-5},\ –)$

Table 7: Parameters for the real PCA instances.

Method	Ionosphere	MNIST
DPRSHCM (diminishing)	$(1.0 \times 10^{-3},\ 0.99)$	$(2.1 \times 10^{-5},\ 0.3)$
DPRSHCM (constant)	$(1.05 \times 10^{-4},\ 0.01)$	$(2.1 \times 10^{-6},\ 0.01)$
DPRSHCM (exact)	$(4.2 \times 10^{-4},\ 0.99)$	$(7.0 \times 10^{-6},\ 0.5)$
DPRSRM	$(1.05 \times 10^{-4},\ 0.01)$	$(3.5 \times 10^{-6},\ 0.003)$
S-DPRGD	$(1.0 \times 10^{-4},\ –)$	$(2.1 \times 10^{-6},\ –)$
DRSGD	$(1.0 \times 10^{-4},\ –)$	$(2.1 \times 10^{-6},\ –)$

Table 8: Parameters for the LRMC instances.

Method	Synthetic	MovieLens-100K
DPRSHCM (diminishing)	$(1.0 \times 10^{-4},\ 0.5)$	$(6.0 \times 10^{-4},\ 0.3)$
DPRSHCM (constant)	$(1.0 \times 10^{-5},\ 0.01)$	$(5.0 \times 10^{-5},\ 0.01)$
DPRSHCM (exact)	$(3.0 \times 10^{-5},\ 0.99)$	$(2.1 \times 10^{-4},\ 0.99)$
DPRSRM	$(1.0 \times 10^{-5},\ 0.01)$	$(7.0 \times 10^{-5},\ 0.003)$
S-DPRGD	$(7.0 \times 10^{-6},\ –)$	$(5.0 \times 10^{-5},\ –)$
DRSGD	$(7.0 \times 10^{-6},\ –)$	$(5.0 \times 10^{-5},\ –)$

the base point. On the Stiefel manifold, $\mathcal{P}_{T_X\mathcal{M}} Z = Z - X\,\mathrm{sym}(X^\top Z)$ gives

$$D(\mathcal{P}_{T\mathcal{M}})(X)[\omega]Z = -\omega\,\mathrm{sym}(X^\top Z) - X\,\mathrm{sym}(\omega^\top Z), \tag{F.3}$$

so the first term costs a few matrix products.

The order of the remainder By Lemma 5.1, the two terms leave a remainder of second order in the displacement. We compare this corrected remainder with the uncorrected change of the true Riemannian gradient. Setting $y = x_{i,k}$ and $x = x_{i,k-1}$ in (5.4) and recalling (5.3), the remainder is

$$r_{i,k} := \mathrm{grad}\, f_i(x_{i,k-1}) + \overline{\mathrm{Hess}}\, f_i(x_{i,k})[u_{i,k}] - \mathrm{grad}\, f_i(x_{i,k}). \tag{F.4}$$

We evaluate $r_{i,k}$ as a diagnostic with exact local gradients and Hessians every twenty iterations along a run of the diminishing setting. Figure 7 plots $\|r_{i,k}\|$ against $\|x_{i,k} - x_{i,k-1}\|$ for all nodes in log-log scale, together with the uncorrected difference $\|\mathrm{grad}\, f_i(x_{i,k-1}) - \mathrm{grad}\, f_i(x_{i,k})\|$. The fitted slopes are 1.76 against 0.67 on the synthetic PCA instance and 1.95 against 0.83 on the synthetic LRMC instance. The two exponents are empirical slopes along the recorded trajectories. The quadratic remainder bound in Lemma 5.1, after squaring, yields the quartic displacement term controlled by Lemma E.15. DPRSRM instead uses a same-sample gradient difference whose conditional expectation is the true gradient change. The figure compares the Taylor remainder with that true gradient change.

Four estimators To isolate the contribution of each term we compare four estimators. All four update the direction by

$$q_{i,k} = (1-\tau)\big(d_{i,k-1} + C_{i,k}\big) + \tau\,\mathrm{grad}\, F_i(x_{i,k}; \xi_{i,k}), \tag{F.5}$$

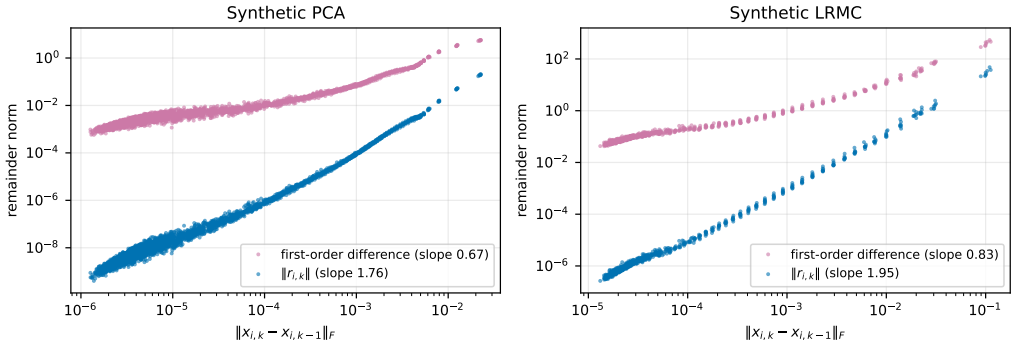

Figure 7: Order of the Taylor remainder: log-log scatter of $\|r_{i,k}\|$ in (F.4) against the displacement $\|x_{i,k} - x_{i,k-1}\|$, compared with the uncorrected true-gradient difference; every point is one node at one diagnostic iteration.

as in (5.6), and differ only in the correction $C_{i,k}$. Write $x = x_{i,k}$, $u = u_{i,k}$ as in (5.3), $g = \nabla F_i(x_{i,k}; \xi_{i,k})$ and $H[u] = \nabla^2 F_i(x_{i,k}; \xi_{i,k})[u]$. Then

$$\begin{aligned} \text{DPRSHCM:} \quad & C_{i,k} = -u\,\mathrm{sym}(x^\top g) - x\,\mathrm{sym}(u^\top g) + \mathcal{P}_{T_x\mathcal{M}} H[u], \\ \text{DPRSHCM-C:} \quad & C_{i,k} = \mathcal{P}_{T_x\mathcal{M}} H[u], \\ \text{DPRSHCM-R:} \quad & C_{i,k} = -u\,\mathrm{sym}(x^\top g) - x\,\mathrm{sym}(u^\top g), \\ \text{DPRSRM:} \quad & C_{i,k} = \mathrm{grad}\, F_i(x_{i,k}; \xi_{i,k}) - \mathrm{grad}\, F_i(x_{i,k-1}; \xi_{i,k}). \end{aligned} \tag{F.6}$$

All four use the momentum parameter τ of the tuned constant setting. We vary the stepsize to examine the effect of the correction at different displacement scales. Figure 8 plots the terminal $G_k^2 + c_k^2$, and Table 9 gives the measures at the tuned stepsize.

On PCA, removing either term increases the residual across stable stepsizes: DPRSHCM-C and DPRSHCM-R give residuals 79–377 and 19–16705 times those of the full correction. DPRSRM stays within a factor of two of DPRSHCM. On LRMC, DPRSRM and DPRSHCM-C are within 10% and 17% of the full correction, respectively, while DPRSHCM-R gives residuals 1.9–3.1 times as large. Table 9 reports the complete terminal measures at the tuned stepsize.

The difference between the objectives follows from the projector derivative. In LRMC, least-squares optimality gives $x^\top g = 0$, so its contribution is the normal vector $-x\,\mathrm{sym}(u^\top g)$. In PCA, $x^\top g = -x^\top A_\xi^\top A_\xi x$ is generally nonzero and $-u\,\mathrm{sym}(x^\top g)$ can have a tangent component. The current tangent projection removes normal components, while momentum and tracking carry them into subsequent iterations with different tangent spaces.

Eight-node networks On the ring and ER graphs with $p = 0.3, 0.6$, the diminishing configuration has terminal gaps in $[5.15, 5.33] \times 10^{-6}$ and gradient norms in $[0.0637, 0.0649]$, within the variation across runs. Their spectral values $\sigma_2 = 0.805, 0.752, 0.778$ are close. Appendix F.7 examines a wider range of network spectra.

F.5 Decentralized PCA on real data

We consider two real datasets of different size. The first is Ionosphere (Sigillito et al., 1989), with 351 observations and 34 features. Each feature is centered and scaled by its standard deviation, and a zero deviation is replaced by one. We split 320 observations among eight nodes in blocks of forty rows. The second is MNIST (LeCun et al., 1998). Its 60000 training images of size 28×28 are scaled by $1/255$, centered and split evenly among eight nodes. Each node then holds a 7500×784 matrix. We compute $r = 5$ principal components on the ER graph with $p = 0.6$. The batch size is ten for Ionosphere and 1500 for MNIST. In the MNIST exact-gradient setting, each update uses all 7500 local samples for the gradient and 1500 for the sampled HVP. The budget is $K = 2000$ in both cases, corresponding to 500 sampling epochs for Ionosphere and 400 for MNIST. Parameters are listed in Table 7 in Appendix F.3.

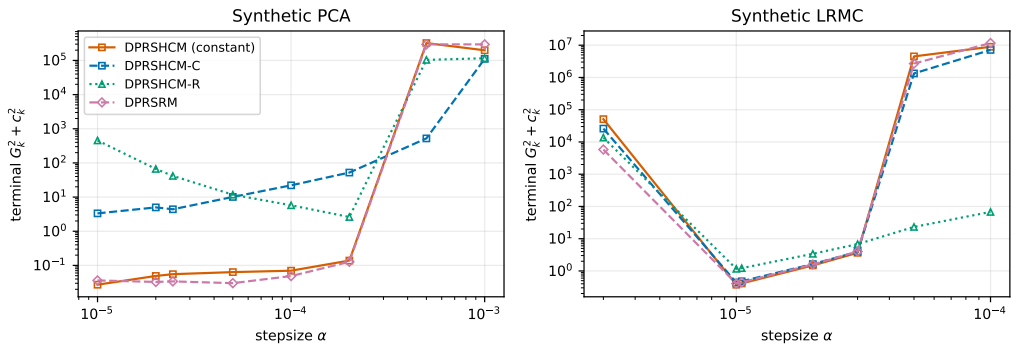

Figure 8: Correction ablation: terminal $G_k^2 + c_k^2$ versus stepsize, with medians over five streams (PCA) and three instances (LRMC). Variants are defined in Appendix F.4. Missing points correspond to divergent runs.

Table 9: Terminal measures of the four estimators in (F.6) at the tuned parameters of the constant setting, $(\alpha, \tau) = (2.4 \times 10^{-5}, 0.001)$ on the synthetic PCA instance and $(1.0 \times 10^{-5}, 0.01)$ on the synthetic LRMC instances; mean $\pm$ standard deviation over five sampling streams and over three instances, respectively. The LRMC gap is relative to F_{ref}.

	Synthetic PCA			
Method	Objective gap	G_k	c_k	d_s
DPRSHCM (const.)	$(4.1\pm1.7)\times10^{-5}$	0.222 ± 0.062	$(1.3\pm0.2)\times10^{-5}$	$(4.7\pm0.4)\times10^{-4}$
DPRSHCM-C	0.0415 ± 0.06	2.51 ± 1.3	$(1.5\pm0.4)\times10^{-4}$	0.0274 ± 0.024
DPRSHCM-R	0.732 ± 1.4	8.31 ± 8.6	$(1.3\pm0.1)\times10^{-5}$	0.0967 ± 0.11
DPRSRM	$(3.2\pm1.3)\times10^{-5}$	0.191 ± 0.045	$(1.3\pm0.1)\times10^{-5}$	$(4.1\pm0.7)\times10^{-4}$
	Synthetic LRMC			
Method	Objective gap	G_k	c_k	d_s
DPRSHCM (const.)	$(7.8\pm0.3)\times10^{-5}$	0.655 ± 0.026	$(6.3\pm0.1)\times10^{-5}$	0.0181 ± 0.00061
DPRSHCM-C	$(8.3\pm0.6)\times10^{-5}$	0.68 ± 0.037	$(6.3\pm0.3)\times10^{-5}$	0.0181 ± 0.0006
DPRSHCM-R	$(2.5\pm0.3)\times10^{-4}$	1.11 ± 0.045	$(6.4\pm0.0)\times10^{-5}$	0.0181 ± 0.00057
DPRSRM	$(7.8\pm1.0)\times10^{-5}$	0.66 ± 0.062	$(6.2\pm0.1)\times10^{-5}$	0.0181 ± 0.00059

Figures 9–10 and Table 10 report the results. On Ionosphere, the diminishing configuration has the smallest mean gap among stochastic methods. For the constant configuration and DPRSRM, nine partitions have median gaps 2.3×10^{-4} and 1.5×10^{-4}, respectively; the remaining partition gives 4.9×10^{-3} and 0.91 and accounts for most of DPRSRM's mean gap. On MNIST, the diminishing configuration reaches a mean joint residual of 0.69, compared with 2.4 for DPRSRM and 11 for the constant configuration. The exact-gradient configuration reaches $G_k = 6.4 \times 10^{-11}$ on MNIST.

The pilot runs also show stepsize sensitivity. Both constant-parameter stochastic methods lose consensus at $\alpha = 10^{-5}$ on MNIST. For exact-gradient Ionosphere, doubling the selected stepsize from 4.2×10^{-4} to 8.4×10^{-4} produces oscillations on two of ten partitions, with terminal consensus errors 0.05 and 0.32.

Heterogeneous local distributions Problem (1.1) allows the local distributions $\mathcal{D}_i$ to differ across nodes. To see how sensitive the methods are, we repeat the MNIST experiment with a partition that separates the local scatter matrices. The images are sorted by their score along the leading principal direction of the full dataset and cut into eight contiguous blocks. Node i receives the images whose score lies in the i-th octile. The Frobenius norms of the local scatter matrices then range over a factor 3.1, against 1.02 under the random partition. All other quantities are unchanged, including the tuned parameters. Figure 11 uses the parameters selected for the random partition. The heterogeneous partition gives objective gaps 49 and 124 for the exact-gradient configuration

and DPRSRM, with consensus errors 0.22 and 0.078. The diminishing configuration reaches gap 2.1×10^{-3} and consensus error 1.6×10^{-5}, compared with 1.5×10^{-5} and 1.5×10^{-5} under random partitioning. The constant configuration gives 2.4×10^{-4} and 8.6×10^{-5}, close to its random-partition results. The stochastic-gradient methods reach gap about 50, $G_k = 1205$ and $c_k = 0.068$. These results measure transfer of fixed parameters across data partitions.

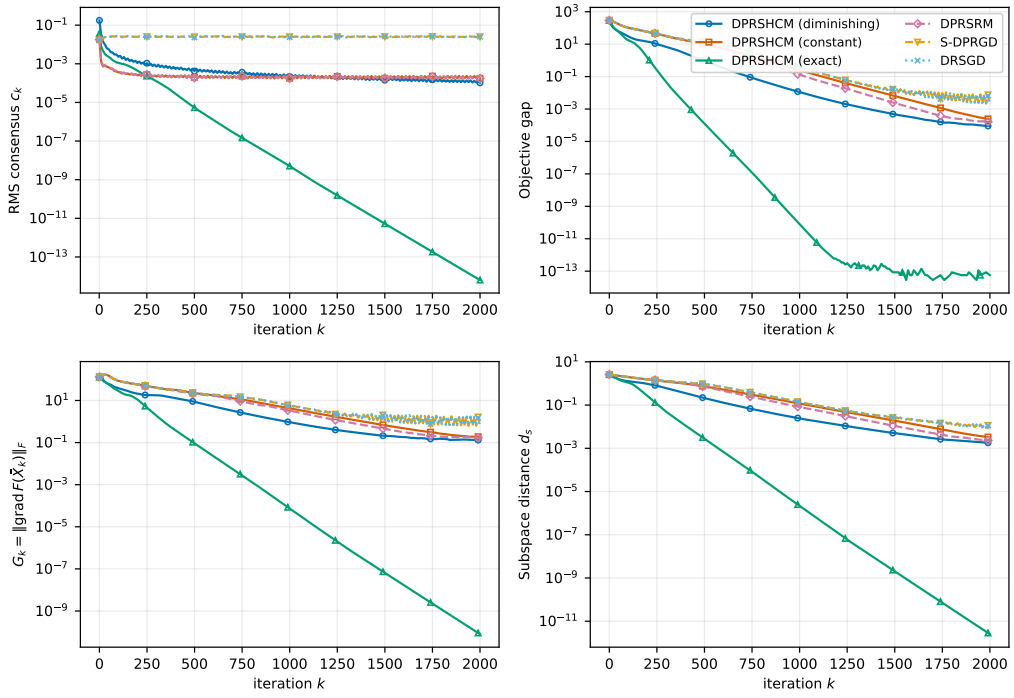

Figure 9: Ionosphere PCA: the six methods with $K = 2000$; pointwise medians over ten random partitions, each with its own network, initial basis and sampling streams.

Table 10: Terminal measures for the real PCA instances; mean ± standard deviation over ten random partitions for Ionosphere and three sampling streams for MNIST.

	Ionosphere			
Method	Objective gap	G_k	c_k	d_s
DPRSHCM (dim.)	$(3.0\pm4.4)\times10^{-4}$	0.153 ± 0.059	$(1.0\pm0.1)\times10^{-4}$	$(3.3\pm2.9)\times10^{-3}$
DPRSHCM (const.)	$(7.7\pm14.8)\times10^{-4}$	0.225 ± 0.11	$(1.8\pm0.1)\times10^{-4}$	$(5.0\pm5.3)\times10^{-3}$
DPRSHCM (exact)	$(5.7\pm388.7)\times10^{-15}$	$(8.4\pm22.5)\times10^{-9}$	$(2.4\pm6.0)\times10^{-13}$	$(3.5\pm9.4)\times10^{-10}$
DPRSRM	0.0913 ± 0.29	0.854 ± 2.1	$(1.7\pm0.1)\times10^{-4}$	0.0294 ± 0.082
S-DPRGD	$(2.6\pm1.7)\times10^{-3}$	0.598 ± 0.12	0.0252 ± 0.0026	0.0101 ± 0.0049
DRSGD	$(2.6\pm1.7)\times10^{-3}$	0.589 ± 0.11	0.0249 ± 0.0018	0.0101 ± 0.0048
	MNIST			
Method	Objective gap	G_k	c_k	d_s
DPRSHCM (dim.)	$(1.5\pm0.2)\times10^{-5}$	0.832 ± 0.047	$(1.5\pm0.0)\times10^{-5}$	$(4.1\pm0.3)\times10^{-5}$
DPRSHCM (const.)	$(2.5\pm0.3)\times10^{-4}$	3.36 ± 0.17	$(7.9\pm0.3)\times10^{-5}$	$(1.6\pm0.2)\times10^{-4}$
DPRSHCM (exact)	$(-3.4\pm2.2)\times10^{-11}$	$(6.4\pm1.8)\times10^{-11}$	$(1.2\pm0.3)\times10^{-15}$	$(1.5\pm0.0)\times10^{-14}$
DPRSRM	$(5.5\pm0.5)\times10^{-5}$	1.53 ± 0.017	$(4.0\pm0.0)\times10^{-5}$	$(8.2\pm0.9)\times10^{-5}$
S-DPRGD	$(7.9\pm0.8)\times10^{-3}$	21 ± 1.4	0.0102 ± 0.00021	$(7.9\pm0.2)\times10^{-4}$
DRSGD	$(8.6\pm1.0)\times10^{-3}$	21.6 ± 1.3	0.0105 ± 0.00018	$(8.5\pm0.5)\times10^{-4}$

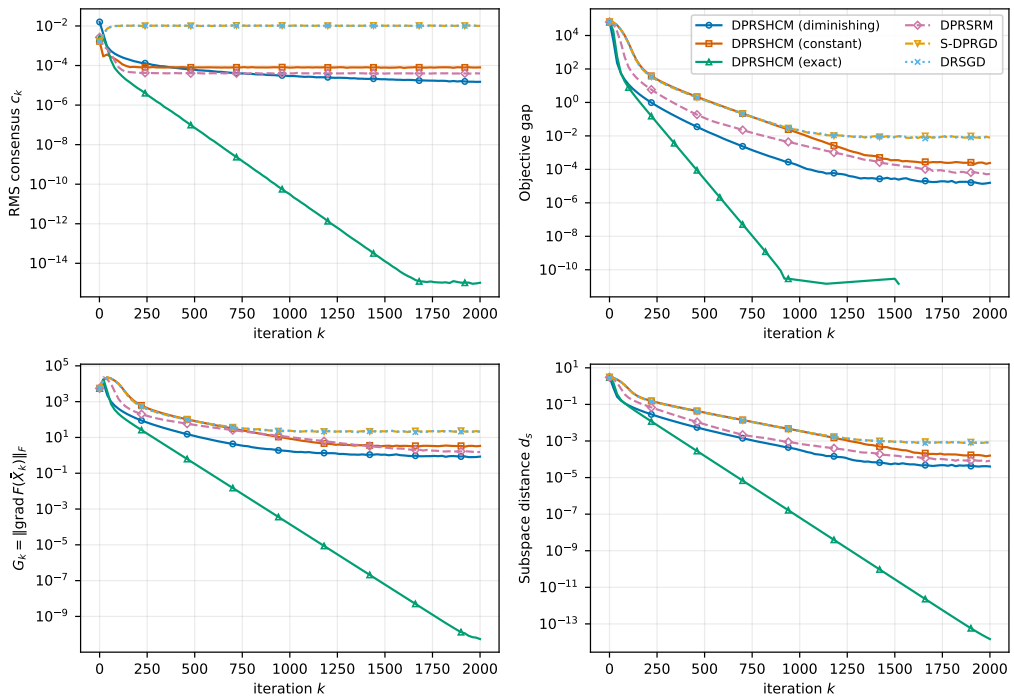

Figure 10: MNIST PCA: the six methods with $K = 2000$ and batch size 1500; pointwise medians over three sampling streams.

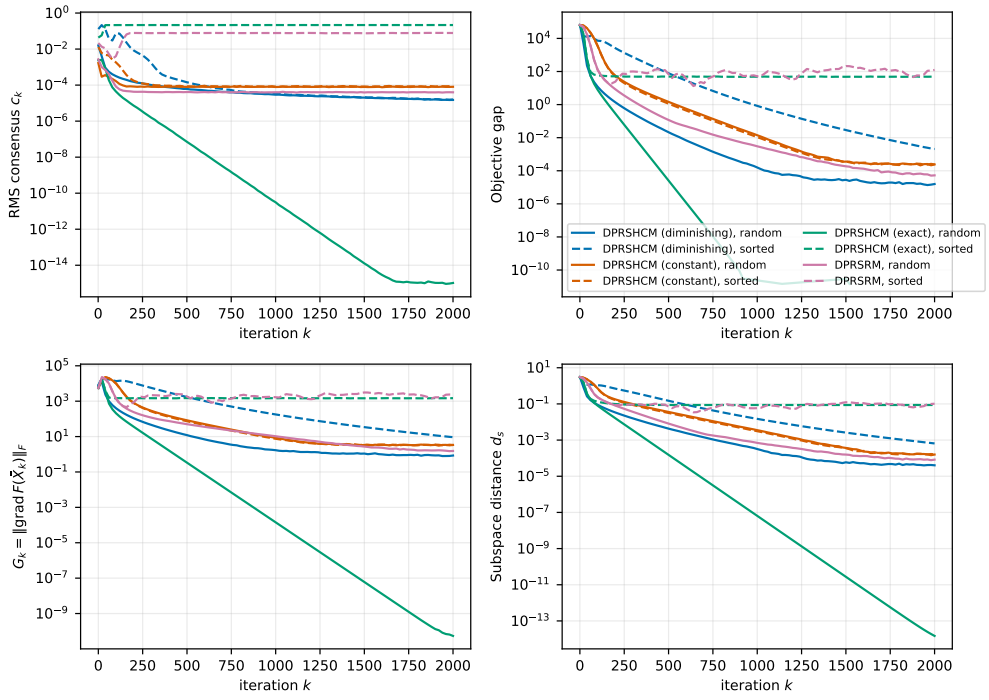

Figure 11: MNIST PCA under the random partition, solid, and under the heterogeneous partition by leading principal score, dashed, which separates the local scatter matrices; the three settings of DPRSHCM and DPRSRM with $K = 2000$.

F.6 DECENTRALIZED LRMC ON MOVIELENS-100K

This subsection continues the LRMC experiment of Section 7.2 on a real dataset. Table 5 reports the terminal measures.

We also solve the problem on MovieLens-100K (Harper & Konstan, 2015). It contains 100000 ratings of 1682 movies by 943 users, a density of about 6.3%. We take the users as rows and the movies as columns. Ten ratings per user are withdrawn before the run to form a held-out set. The unregularized model (A.2) fits the training ratings of every column exactly when the column has few of them, so rows and columns with few training ratings are removed. The retained submatrix contains 325 users and 304 movies, with 38234 training and 1889 held-out ratings. Every training rating is then reduced by a baseline $\mu + b_u + b_j$, the global mean plus a user and a movie bias estimated on the training ratings by alternating shrinkage (Koren et al., 2009), and the same baseline is subtracted from the held-out ratings. The model fits the residual matrix with $r = 3$. Predicting every held-out rating by its baseline alone gives a held-out RMSE of 0.916, which is the reference for the values below. The 304 columns are partitioned randomly among eight nodes. Each node holds 38 columns with between 41 and 291 training ratings per column and a median of 114. No ground-truth subspace is available. We therefore replace d_s by the root-mean-square error on the held-out set and the objective gap by the training residual $F(\bar{X}_k)$. The held-out entries are excluded from every Ω_i, so they enter neither the objective nor the inner least squares solution V_i. Write Ω_i^{test} for the held-out entries at node i. The held-out RMSE at iteration k is

$$\mathrm{RMSE}_k = \left(\frac{\sum_{i=1}^{n} \sum_{(u,j)\in\Omega_i^{\text{test}}} \left(\left[\bar{X}_k V_i(\bar{X}_k) \right]_{uj} - [A_i]_{uj} \right)^2}{\sum_{i=1}^{n} |\Omega_i^{\text{test}}|} \right)^{1/2}. \tag{F.7}$$

Predictions use the training entries of each column and RMSE is measured in rating units. We use the ring network and $K = 2000$. Figure 12 and Table 5 give the results. The six methods attain RMSEs between 0.883 and 0.891, compared with 0.916 for the baseline predictor; training residuals differ by less than 0.3%. The diminishing configuration and DPRSRM both attain RMSE 0.887, with gradient norms 7.5 and 7.1 and consensus errors 1.7×10^{-4} and 3.0×10^{-4}, respectively. The exact-gradient configuration has the smallest gradient norm, 0.56, and RMSE, 0.883, with consensus error 0.016.

F.7 SENSITIVITY STUDIES

All studies in this subsection use the synthetic PCA instance with $K = 10000$, five sampling streams per setting and the parameters of Table 6. The parameters remain fixed to measure sensitivity to each changed problem setting. We compare the three settings of DPRSHCM and DPRSRM.

Number of nodes Figure 13 varies $n \in \{4, 8, 16, 32\}$ on connected ER graphs with $p = 0.6$ and $m_i = 1000$. Thus the total sample count grows with n. The diminishing configuration's gap decreases from 8.1×10^{-6} to 1.1×10^{-6}, and DPRSRM's from 7.5×10^{-5} to 6.3×10^{-6}.

Batch size Figure 14 uses $b \in \{1, 5, 10, 50, m_i\}$. From $b = 1$ to $b = 50$, the diminishing configuration's gap decreases from 7.1×10^{-4} to 4.3×10^{-7}, and DPRSRM's from 3.9×10^{-3} to 2.9×10^{-7}. At $b = m_i$, the diminishing configuration reaches 7.8×10^{-8}; the constant configuration and DPRSRM reach 2.7×10^{-7} with parameters selected for $b = 10$. The exact-gradient configuration uses its separately tuned fixed stepsize and reaches numerical precision.

Spectrum Figure 15 varies $\gamma \in \{0.5, 0.7, 0.8, 0.9, 0.95\}$ in (F.1). This changes the condition number $\gamma^{-(d-1)}$, the eigengap $\sigma_1^2(A)\gamma^{2(r-1)}(1-\gamma^2)$ and the scale of the sampled objectives. Both stochastic momentum methods attain their smallest gap at $\gamma = 0.8$. The diminishing configuration has a smaller gap at every tested value, while both retain large gaps at $\gamma = 0.5$ (5.2 and 8.1). The exact-gradient configuration reaches $G_k = 2.5$ at that value within the fixed budget.

Network topology revisited Subsection F.4 compares three topologies at $n = 8$. To widen the range of σ_2 we repeat the comparison at $n = 16$ with $m_i = 1000$ over the ring, the ER graphs with $p = 0.1, 0.3, 0.6$ and the complete graph. Their second largest singular values are 0.949, 0.962, 0.821, 0.494 and 0.000. Figure 16 plots the terminal consensus error against $1/(1-\sigma_2)$,

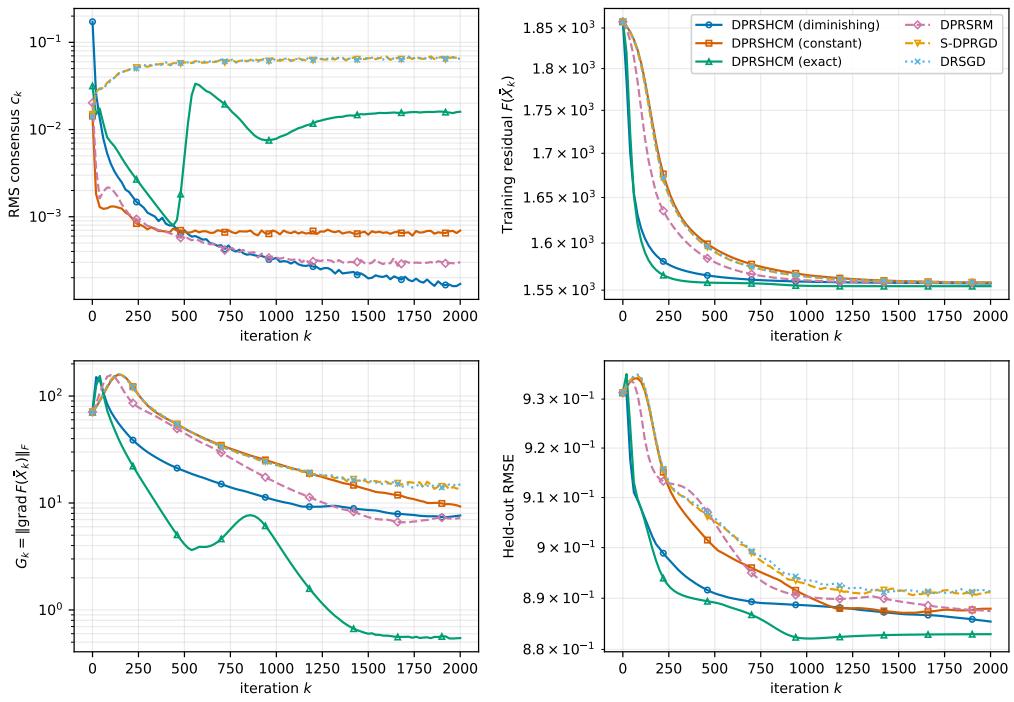

Figure 12: MovieLens-100K on the ring network with $K = 2000$: consensus error, training residual, gradient norm at $\bar{X}_k$ and held-out RMSE; pointwise medians over three sampling streams.

a measure of network connectivity. The theoretical factor $\ell/(1-\varrho)$ in (E.4) also depends on the projection Lipschitz constant ℓ. Over these five networks, log-log fits give slopes 0.73, 0.71 and 0.74 for the diminishing setting, the constant setting and DPRSRM. Their consensus errors increase as connectivity deteriorates. The exact-gradient setting remains near roundoff on every topology. The objective gaps and gradient norms vary across the five topologies by less than the run-to-run variation.

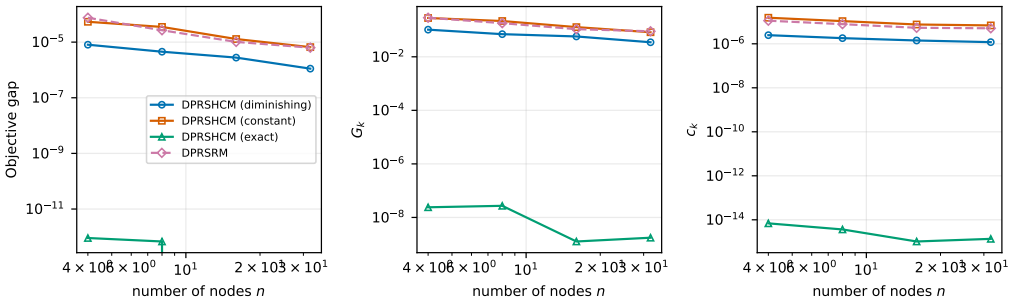

Figure 13: Synthetic PCA: terminal measures, medians, against the number of nodes $n \in \{4, 8, 16, 32\}$ at a fixed local sample size $m_i = 1000$.

F.8 Oracle cost and running time

On synthetic LRMC, each stochastic DPRSHCM update uses ten columns for both the gradient and correction. DPRSRM evaluates gradients at two points on the same ten columns; S-DPRGD and DRSGD use one gradient batch. The exact-gradient configuration uses all 120 local columns for the gradient and ten for the HVP. The gradient requires a thin SVD of each observed-row matrix. The

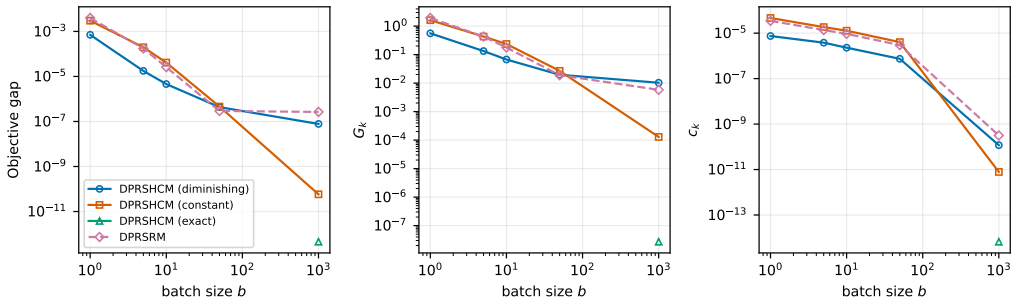

Figure 14: Synthetic PCA: terminal measures, medians, against the local batch size $b \in \{1, 5, 10, 50, m_i\}$; $b = m_i$ is the exact local gradient.

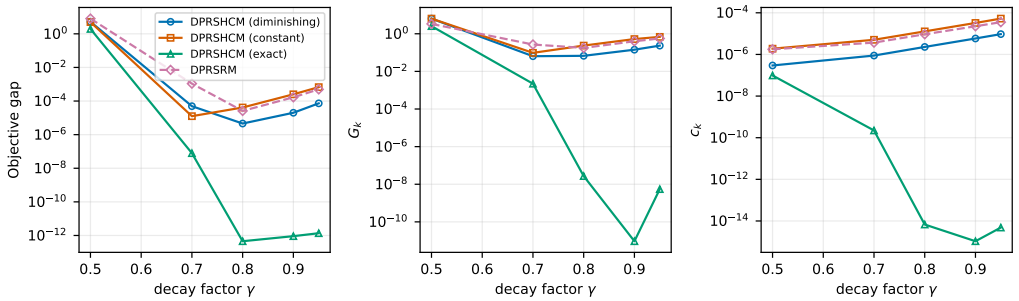

Figure 15: Synthetic PCA: terminal measures, medians, against the singular-value decay factor $\gamma \in \{0.5, 0.7, 0.8, 0.9, 0.95\}$.

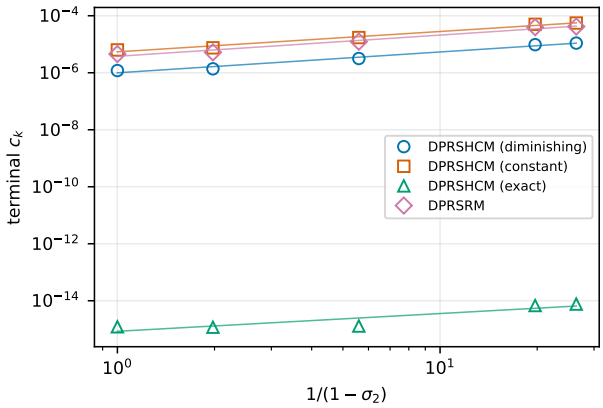

Figure 16: Synthetic PCA with $n = 16$: terminal consensus error, medians, against $1/(1-\sigma_2)$ over five topologies, with least-squares fits in log-log coordinates.

Table 11: Mean terminal residual and running time in the main comparisons (ten PCA streams and ten LRMC instances). Times exclude diagnostics and use single-machine mixing, without inter-machine communication.

Problem	Method	$G_k^2 + c_k^2$	Time (s)
PCA	DPRSHCM (constant)	0.045	9.8
	DPRSRM	0.037	7.8
LRMC	DPRSHCM (constant)	0.49	26.7
	DPRSRM	0.50	38.4

correction differentiates the least-squares solve and reuses this SVD, whereas DPRSRM requires one SVD at each point.

We estimate κ in (F.2) as the additional HVP time, after the gradient on the same batch, divided by the gradient time. We time 10^4 calls for synthetic PCA, 2000 for synthetic LRMC and 200 for MNIST, using each experiment's batch size. The measured ratios are 1.92, 0.36 and 0.72, respectively. Small matrix operations in the projector derivative dominate synthetic PCA; on MNIST, multiplication by the sampled data dominates. Table 12 gives the LRMC counts, including initialization. The stochastic configurations use 32% fewer equivalent gradient evaluations than DPRSRM. The exact-gradient count is 9.1 times the stochastic count; Theorem 6.3 counts each full local gradient as one query.

Mean times for diminishing, constant and exact-gradient DPRSHCM, DPRSRM, S-DPRGD and DRSGD are respectively $26.7, 26.7, 140.4, 38.4, 19.9, 20.1$ seconds on LRMC, and $9.9, 9.8, 10.5, 7.8, 5.5, 6.1$ on synthetic PCA. Table 11 pairs accuracy and time for the constant configuration and DPRSRM. Timings exclude diagnostics and inter-machine communication; all runs use single-machine matrix multiplication for mixing.

Table 12: Oracle queries over a complete run of $K = 2000$ iterations on the synthetic LRMC instances, per node and counted in sampled columns, and the equivalent gradient count (F.2) with $\kappa = 0.36$.

Method	Gradient queries	HVP queries	Equivalent gradient queries
DPRSHCM (stochastic)	20000	19990	27196
DPRSHCM (exact)	240000	19990	247196
DPRSRM	39990	–	39990
S-DPRGD	20000	–	20000
DRSGD	20000	–	20000