STABILITY AND GENERALIZATION OF DECENTRALIZED RIEMANNIAN STOCHASTIC GRADIENT DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalization bounds for Euclidean decentralized SGD rely on the nonexpansiveness of averaging. Manifold projection can amplify perturbations caused by replacing a training sample, complicating the extension of these guarantees to manifold constraints. We analyze stochastic decentralized projected Riemannian gradient descent (DPRGD) on compact smooth embedded manifolds without boundary. By bounding projection expansion through consensus error, we derive a stability-based generalization bound whose additional geometric factor is controlled by the sum of squared step sizes. Motivated by this analysis, we propose DPRGD-ATC, an adapt-then-combine variant that averages locally updated models before projection. Under the same assumptions, step sizes, and communication budget, we prove a tighter generalization bound by reducing the terms associated with consensus error. A local curvature construction further shows that DPRGD-ATC can have strictly lower sensitivity to a single-sample replacement. We also establish empirical stationarity guarantees for both methods and compare their network-dependent error terms. Experiments on synthetic PCA across three geometries and on Fashion-MNIST with up to 32 agents illustrate these geometric effects.

1 INTRODUCTION

Decentralized learning trains a shared model by combining local gradient updates with communication between neighboring agents (Lian et al., 2017; Koloskova et al., 2020). Algorithmic stability connects the effect of replacing one training sample to generalization error (Bousquet & Elisseeff, 2002; Hardt et al., 2016). This approach has led to generalization guarantees for decentralized SGD (Richards & Rebeschini, 2020; Sun et al., 2021; Zhu et al., 2022; Le Bars et al., 2024; Zeng & Lei, 2025). In particular, Euclidean decentralized updates admit worst-case stability bounds comparable to centralized SGD, even though the agents may hold different models during training (Le Bars et al., 2024).

Manifold constraints arise naturally in decentralized principal component analysis and subspace learning. Agents holding different observations cooperate to estimate orthonormal directions or a low-dimensional subspace, whose reconstruction error on new observations determines its statistical performance (Chen et al., 2021; Wang et al., 2024; 2025). Averaging local orthonormal models generally violates the constraint, so projected algorithms restore feasibility after communication and local updates (Deng & Hu, 2025a). These methods have been analyzed primarily through consensus and optimization convergence. Their guarantees do not quantify how manifold projection and consensus error jointly affect sensitivity to a training-sample replacement and the resulting generalization error.

To see where the Euclidean argument applies, consider m agents and two runs whose datasets differ in one sample. Let x_j, x'_j be the corresponding models at agent j, and let $Q \in \mathbb{R}^{m\times m}$ be a nonnegative doubly stochastic mixing matrix. Then

$$\frac{1}{m}\sum_i \left\| \sum_j Q_{ij}(x_j - x'_j) \right\| \le \frac{1}{m}\sum_j \left\| x_j - x'_j \right\|. \tag{1}$$

Thus averaging does not increase the average perturbation between the two runs. This property does not depend on the consensus error, the difference between agents' models within one run.

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.

Nearest-point projection onto a manifold generally loses this nonexpansiveness. Write $\mathcal{P}_{\mathcal{M}}$ for nearest-point projection onto $\mathcal{M}$, and let $\tau > 0$ be a lower bound on its reach, the distance within which nearest points are unique. For ambient points u, v at distance at most $r < \tau$ from $\mathcal{M}$ (Leobacher & Steinicke, 2021; Deng & Hu, 2025a),

$$\|\mathcal{P}_{\mathcal{M}}(u) - \mathcal{P}_{\mathcal{M}}(v)\| \leq \frac{1}{1 - r/\tau} \|u - v\| . \tag{2}$$

The amplification can occur even on the unit circle $\mathcal{M} = \mathbb{S}^1$. For distinct $p, q \in \mathbb{S}^1$ and $u = (1-r)p$, $v = (1-r)q$, where $0 < r < 1$, normalization gives

$$\frac{\|\mathcal{P}_{\mathcal{M}}(u) - \mathcal{P}_{\mathcal{M}}(v)\|}{\|u - v\|} = \frac{1}{1 - r} > 1.$$

Thus the expansion in (2) is an actual geometric effect. Using a fixed radius would multiply the same factor at every step. We instead control the distance of each pre-projection point from $\mathcal{M}$ through consensus error and second-order estimates for tangent updates. This makes the accumulated geometric cost depend on squared step sizes. It also motivates DPRGD-ATC: following the Euclidean adapt-then-combine principle (Richards & Rebeschini, 2020; Le Bars et al., 2024), we average locally updated models before projection to reduce consensus error.

Informal theorem. Let $\varepsilon_{\mathrm{E}}(T)$ be the basic Euclidean stability bound after T steps, and let $\varepsilon_{\mathrm{B}}(T)$ and $\varepsilon_{\mathrm{A}}(T)$ denote the bounds for baseline DPRGD and its ATC variant, respectively. Under common regularity and mixing conditions, with the same sufficiently small steps η_t fixed in advance, a constant C_{geom} depending only on geometry, gradient bounds, and the network satisfies

$$\varepsilon_{\mathrm{A}}(T) \leq \varepsilon_{\mathrm{B}}(T) \leq \exp\left(C_{\mathrm{geom}} \sum_{t<T} \eta_t^2 \right) \varepsilon_{\mathrm{E}}(T). \tag{3}$$

Thus square-summable steps keep the additional multiplier bounded, and DPRGD-ATC improves the bound at the same communication cost. Theorems 2–3 give the precise guarantees; Appendix A.4 derives (3). Our contributions are:

- **Manifold generalization bounds.** We extend the Euclidean stability recursion to stochastic DPRGD on compact smooth embedded manifolds. A joint induction controls consensus error and keeps projection well defined along both runs on neighboring datasets. Second-order geometric bounds on the distance before projection make the accumulated projection expansion depend on squared step sizes. The resulting generalization bound retains the Euclidean $1/N$ prefactor, where N is the total sample size, and isolates the additional geometric factor.
- **DPRGD-ATC and its guarantees.** Guided by this factor, we propose an adapt-then-combine variant that mixes local gradient updates before projection. We control the normal components introduced by mixing gradients from different tangent spaces and prove that the smaller consensus error reduces the squared consensus and mixed terms by factors σ^2 and 2σ, where $\sigma < 1/2$ is the mixing contraction factor. We also establish empirical stationarity and complexity guarantees, with improved network-related constants at the same communication cost.
- **Lower sample sensitivity under a local construction.** For every compact smooth embedded manifold without boundary of dimension at least one, we construct a local learning problem in which DPRGD-ATC has strictly lower sensitivity to a single-sample replacement. The leading sensitivity ratio is determined by normal curvature and admits an exact circle formula. Experiments on synthetic sphere, Stiefel, and Grassmann PCA and on Fashion-MNIST with up to 32 agents evaluate geometry, sample sensitivity, and generalization separately.

1.1 Related work

Stability and decentralized generalization. Hardt et al. (2016) analyze SGD by coupling runs on neighboring datasets. Richards & Rebeschini (2020) establish graph-independent generalization bounds for an adapt-then-combine distributed method and combine them with optimization

estimates. Sun et al. (2021) study decentralized SGD under convex and nonconvex losses; Zhu et al. (2022) derive topology-dependent bounds. Le Bars et al. (2024) sharpen this analysis for both update orders and recover centralized-style worst-case bounds. Zeng & Lei (2025) obtain training-error-dependent stability bounds under weaker regularity assumptions. Zhu et al. (2025) study data-influence propagation through the network. We bound the distance of each pre-projection point from the manifold using consensus error and second-order geometric estimates for averaging and tangent updates. Coupling these estimates with the stability recursion yields a geometric factor controlled by squared step sizes and quantifies the improvement from changing the update order.

Decentralized constrained optimization. Consensus and optimization errors are central to decentralized gradient analysis (Yuan et al., 2016; Shi et al., 2015). Even convex constraints change the evolution of the average iterate under projection (Choi & Kim, 2025). For manifolds, Riemannian consensus (Tron et al., 2013), decentralized Stiefel optimization (Chen et al., 2021), and projected methods on compact submanifolds (Deng & Hu, 2025a) provide convergence guarantees. We analyze the stochastic version of DPRGD in Deng & Hu (2025a). The method in Chen et al. (2021) instead projects the consensus direction onto a tangent space and uses a retraction. Retraction-free methods (Sun et al., 2024; Goyens & Feppon, 2026) organize the tangent and normal updates differently. REXTRA uses a correction recursion for constant-step optimization while communicating local iterates (Wu et al., 2025); DPRSRM combines recursive stochastic gradients with tracking (Deng & Hu, 2025b), and VRSGT uses variance reduction for orthogonality-constrained optimization (Wang & Liu, 2023). These methods control optimization errors through history or additional states. Our comparison isolates how update order changes replace-one stability within DPRGD, using the same gradient, projection, and communication budgets.

Update order and diffusion methods. Euclidean adapt-then-combine updates appear as Variant A of D-SGD in Le Bars et al. (2024). Hosseini et al. (2026) also give stability-based generalization bounds for adversarial ATC diffusion training. On manifolds, Wang et al. (2024) combine retracted local estimates through an intrinsic mean, while Wang et al. (2025) use exponential-map updates and tangent-space consensus. Our DPRGD-ATC variant averages locally updated models before a single nearest-point projection. We establish its improvement over DPRGD in the stability upper bound and compare their local sample sensitivities and empirical stationarity guarantees.

2 Problem, DPRGD, and stability

Let $\mathcal{M}$ be a nonempty compact C^3 embedded manifold without boundary in a Euclidean space, with the induced metric. Matrix spaces use the Frobenius norm. Each of m agents holds $S_i = (z_{i1}, \dots, z_{in}) \in \mathcal{Z}^n$, where $\mathcal{Z}$ is the sample space; write $S = (S_1, \dots, S_m)$ and $N = mn$. For a sample loss $\ell : \mathcal{M} \times \mathcal{Z} \to \mathbb{R}$, define

$$F_i(x) = \frac{1}{n}\sum_{r=1}^{n} \ell(x, z_{ir}), \qquad F_S(x) = \frac{1}{m}\sum_{i} F_i(x), \qquad \min_{x\in\mathcal{M}} F_S(x).$$

When all samples are i.i.d. from a distribution $\mathcal{D}$, the population risk is $F(x) = \mathbb{E}_{z\sim\mathcal{D}}\ell(x, z)$. We write $\operatorname{grad}\ell(x, z) = \operatorname{grad}_x \ell(x, z) \in T_x\mathcal{M}$ for the Riemannian gradient with respect to the model x. Tangent vectors are represented in the ambient space. The symbol ∇ will be reserved for gradients of functions defined on an open ambient set.

Assumption 1 (Uniform regularity)**.** *The reach of $\mathcal{M}$ is at least $\tau > 0$. Choose $0 < h < \tau/4$ and let $U_h = \{u : \operatorname{dist}(u, \mathcal{M}) < h\}$. Nearest-point projection is C^2 on a neighborhood of $\overline{U}_h$, with constants*

$$C_0 = \frac{\tau}{\tau - h}, \qquad C_1 > 0, \qquad \sup_{u\in\overline{U}_h} \left\|D^2\mathcal{P}_{\mathcal{M}}(u)\right\| \le C_1.$$

For every $z \in \mathcal{Z}$, $\ell(\cdot, z)$ is C^2 and takes values in $[0, M_0]$. For all $x, y \in \mathcal{M}$,

$$\|\operatorname{grad}\ell(x, z)\| \le G, \qquad |\ell(x, z) - \ell(y, z)| \le L_0 \|x - y\|,$$
$$\|\operatorname{grad}\ell(x, z) - \operatorname{grad}\ell(y, z)\| \le L_g \|x - y\|,$$

where $G > 0$ and all constants are independent of the dataset.

Here $D\mathcal{P}_{\mathcal{M}}$ and $D^2\mathcal{P}_{\mathcal{M}}$ are the first and second ambient derivatives, with their induced operator norms. Compactness and C^3 regularity supply a positive reach and finite projection constants (Leobacher & Steinicke, 2021); the uniform loss bounds are assumptions on the sample family.

Communication and updates. Let $\mathbf{1}_m \in \mathbb{R}^m$ be the all-ones vector, $J = \mathbf{1}_m\mathbf{1}_m^\top/m$, and let Q be a fixed, data-independent, nonnegative symmetric doubly stochastic matrix with $\sigma = \|Q - J\|_2 < 1/2$. For a nonnegative symmetric doubly stochastic one-round matrix W, we may use $Q = W^k$, where k is a fixed number of communication rounds. Writing $\lambda = \|W - J\|_2 < 1$ gives $\sigma = \lambda^k$; the training communication cost is kT rounds. For a lazy ring, maintaining a fixed $\sigma < 1/2$ requires $k = \Theta(m^2)$ (Appendix A.9). All agents start at the same data-independent $x_0 \in \mathcal{M}$. The step sequence (η_t) is deterministic and independent of the data and sampled indices. At iteration t, agent i draws I_i^t uniformly from $\{1, \dots, n\}$, independently across agents and iterations, and computes $g_i^t = \operatorname{grad} \ell(x_i^t, z_{i,I_i^t})$.

The stochastic DPRGD update is

$$x_i^{t+1} = \mathcal{P}_{\mathcal{M}}\left(\sum_j Q_{ij} x_j^t - \eta_t g_i^t\right). \tag{4}$$

Replacing g_i^t by $\operatorname{grad} F_i(x_i^t)$ recovers the original full-gradient DPRGD (Deng & Hu, 2025a). Each agent evaluates one local gradient, communicates its model through k rounds, and projects once. We use B to index the DPRGD bounds below.

Average and feasible output. Define the ambient average and its projected output by

$$\bar{x}^t = \frac{1}{m}\sum_i x_i^t, \qquad \widehat{x}^t = \mathcal{P}_{\mathcal{M}}(\bar{x}^t). \tag{5}$$

The average need not belong to $\mathcal{M}$, whereas $\widehat{x}^t$ is feasible whenever its projection is unique. Our trajectory estimates guarantee this property. Write $\mathcal{A}_T(S;\mathcal{I}) = \widehat{x}^T$ for the algorithm's output after T steps, with sampled index array $\mathcal{I} = (I_i^t)_{i,t}$. In statistical statements the dataset dependence is explicit; in optimization statements we condition on a fixed dataset.

Stability and generalization. Datasets $S \sim S'$ are neighbors if one example at one agent is replaced. The two runs use the same initialization, steps, and sampled indices.

Definition (Uniform expected stability)**.** *An algorithm $\mathcal{A}$ has stability at most ϵ if*

$$\sup_{S\sim S',\, z\in\mathcal{Z}} \mathbb{E}_{\mathcal{I}}|\ell(\mathcal{A}(S;\mathcal{I}), z) - \ell(\mathcal{A}(S';\mathcal{I}), z)| \le \epsilon. \tag{6}$$

The generalization gap is $\operatorname{gen}(\mathcal{A}, S;\mathcal{I}) := F(\mathcal{A}(S;\mathcal{I})) - F_S(\mathcal{A}(S;\mathcal{I}))$. For i.i.d. data, the replace-one argument (Bousquet & Elisseeff, 2002; Hardt et al., 2016) turns the stability definition into the expected-generalization guarantee

$$\left|\mathbb{E}_{S,\mathcal{I}} \operatorname{gen}(\mathcal{A}, S;\mathcal{I})\right| \le \epsilon.$$

This bounds the absolute expected gap; experiments also report the distinct quantity $\mathbb{E}|\operatorname{gen}|$. Stability is uniform over admissible fixed datasets, including heterogeneous ones; the generalization conclusion uses common-distribution sampling. Appendix A.3 gives the exchangeability proof.

3 Stability of stochastic DPRGD

We first analyze (4). The task is to control the coefficient in (2) using the algorithm's consensus error. Stack the vectorized node variables as rows of X. Define

$$E(X) = m^{-1/2}\|X - JX\|_F, \qquad D(X) = \max_{i,j}\|x_i - x_j\| \le \sqrt{2m}E(X).$$

Here $E(X)$ is the root-mean-square consensus error and $D(X)$ is the maximum node distance. Neither compares different datasets. With $a = 1/(2\tau) + C_1/2$, define

$$r_{\mathrm{B}}(D, \eta) = aD^2 + C_1 D\eta G + \frac{C_1}{2}\eta^2 G^2. \tag{7}$$

For $D(X) \le D$ and $D + \eta G \le h$, each point before projection in (4) lies within distance $r_{\mathrm{B}}(D, \eta)$ of $\mathcal{M}$. The first term comes from averaging feasible points, the second from the interaction between the mixing offset and tangent step, and the last from the tangent step alone. Lemma 10 proves this estimate.

Let $\eta_* > 0$ be the common admissible step threshold defined explicitly in (25) in Appendix A.2. For $0 \le \eta_t \le \eta_*$, define deterministic bounds

$$e_0 = 0, \qquad e_{t+1} = 2\sigma e_t + 2\eta_t G, \qquad \delta_t = \sqrt{2m}e_t. \tag{8}$$

These bounds depend on the graph and step sequence, and hold for every admissible dataset and index array. For fixed geometric constants and σ, the sufficient threshold η_* scales as $m^{-1/2}$.

Theorem 2 (DPRGD stability and generalization)**.** *Under Assumption 1 and the setup above, DPRGD is well defined, $E(X^t) \le e_t$, and $D(X^t) \le \delta_t$. Its points before projection and ambient means $\bar{x}^t$ lie within distance $h/4$ of $\mathcal{M}$. Define the distance and projection factors*

$$r_{\mathrm{B},t} = r_{\mathrm{B}}(\delta_t, \eta_t), \qquad c_{\mathrm{B},t} = (1 - r_{\mathrm{B},t}/\tau)^{-1}, \qquad o_{\mathrm{B},T} = (1 - a\delta_T^2/\tau)^{-1}.$$

The feasible output $\widehat{x}^T$ satisfies (6) *with bound* $\min\{M_0, \varepsilon_{\mathrm{B}}(T)\}$*, where*

$$\varepsilon_{\mathrm{B}}(T) = \frac{2L_0 G o_{\mathrm{B},T}}{N} \sum_{s<T} \eta_s c_{\mathrm{B},s} \prod_{t=s+1}^{T-1} c_{\mathrm{B},t}(1 + \eta_t L_g). \tag{9}$$

For i.i.d. training samples, the same bound controls the absolute expected generalization gap.

How the proof handles projection. The first step jointly bounds consensus error and ensures unique projection: a coarse 2-Lipschitz estimate gives $E(X^{t+1}) \le 2\sigma E(X^t) + 2\eta_t G$, while the common step condition keeps the iterates in the neighborhood where this estimate applies. The second step uses the quadratic distance estimate (7) to replace that coarse constant by the finer $c_{\mathrm{B},t}$ in the statistical analysis. This distinction retains the dependence on the step sizes.

For neighboring datasets, let $x_i^t, x_i'^t$ be the coupled iterates and define $d_t = \mathbb{E}_{\mathcal{I}}[m^{-1} \sum_i \|x_i^t - x_i'^t\|]$. The replaced example is sampled with probability $1/n$ at its node. Double stochasticity and the two-point projection estimate give

$$d_{t+1} \le c_{\mathrm{B},t} \left[(1 + \eta_t L_g) d_t + \frac{2\eta_t G}{N} \right]. \tag{10}$$

The distance bound holds for both runs and every index array, so $c_{\mathrm{B},t}$ is deterministic. Iterating this inequality and projecting the final mean gives (9). Thus consensus error enters stability through a controlled projection factor, rather than an uncontrolled expansion at each iteration.

Comparison with Euclidean stability bounds. Removing the projection factor from (10) gives the standard smooth-loss perturbation recursion used by Hardt et al. (2016) and by Le Bars et al. (2024, Appendix C). Starting from $d_0 = 0$ and unrolling it yields the reference bound

$$\varepsilon_{\mathrm{E}}(T) = \frac{2L_0 G}{N} \sum_{s<T} \eta_s \prod_{t=s+1}^{T-1} (1 + \eta_t L_g).$$

This expression retains the full step sequence; the refined bounded-loss result in Le Bars et al. (2024, Theorem 4.1) additionally exploits the time before a replaced sample is first selected. Here we compare the same basic recursion before and after manifold projection:

$$\varepsilon_{\mathrm{B}}(T) \le o_{\mathrm{B},T} \exp\left(\frac{2}{\tau} \sum_{t<T} r_{\mathrm{B},t} \right) \varepsilon_{\mathrm{E}}(T). \tag{11}$$

The consensus bound satisfies

$$\sum_{t<T} \delta_t^2 \le \frac{8mG^2}{(1 - 2\sigma)^2} \sum_{t<T} \eta_t^2.$$

Cauchy–Schwarz controls the mixed term in (7) by the same squared-step sum. The additional geometric multiplier is therefore uniformly bounded for square-summable schedules. The total bound still contains the Euclidean nonconvex gradient propagation factor.

Table 1: Consensus and distance bounds under the same assumptions. Both updates use one gradient and one projection per agent and the same communication block.

Quantity	DPRGD	DPRGD-ATC
Consensus error bound	e_t	σe_t
Squared consensus term	$a\delta_t^2$	$a\sigma^2\delta_t^2$
Mixed term	$C_1\delta_t\eta_t G$	$2C_1\sigma\delta_t\eta_t G$
Local step term	$C_1\eta_t^2 G^2/2$	$C_1\eta_t^2 G^2/2$

4 Stability improvement through adapt-then-combine

The squared consensus and mixed terms in (7) increase the projection factors in the DPRGD generalization bound. In (4), local gradient steps introduce differences after communication. To reduce these differences before projection, we propose DPRGD-ATC, which communicates the locally updated models:

$$x_i^{t+1} = \mathcal{P}_{\mathcal{M}}\left(\sum_j Q_{ij}(x_j^t - \eta_t g_j^t)\right). \tag{12}$$

The initialization, sampling rule, steps, and output (5) are unchanged. Each agent still evaluates one gradient and projects once, communicating $x_i^t - \eta_t g_i^t$ through the same k rounds. We use A to index this variant. For $Q = J$, all its states coincide, giving centralized projected minibatch SGD; DPRGD can leave consensus even with this exact average.

The gain requires accounting for the different tangent spaces. A gradient tangent at one agent can have a normal component at another. Controlling tangent-space variation gives the distance bound

$$r_{\mathrm{A}}(D,\eta) = aD^2 + 2C_1 D\eta G + \frac{C_1}{2}\eta^2 G^2. \tag{13}$$

At a fixed node distance D, this has an additional mixed term. The following result combines it with ATC's smaller consensus error.

Theorem 3 (Stability and generalization bound of DPRGD-ATC)**.** *Under the assumptions and step restriction of Theorem 2, DPRGD-ATC is well defined, with $E(X_{\mathrm{A}}^t) \le \sigma e_t$ and $D(X_{\mathrm{A}}^t) \le \sigma\delta_t$. Define*

$$r_{\mathrm{A},t} = r_{\mathrm{A}}(\sigma\delta_t,\eta_t), \quad c_{\mathrm{A},t} = (1 - r_{\mathrm{A},t}/\tau)^{-1}, \quad o_{\mathrm{A},T} = (1 - a\sigma^2\delta_T^2/\tau)^{-1}.$$

The feasible output $\widehat{x}_{\mathrm{A}}^T$ satisfies (6) *with bound* $\min\{M_0, \varepsilon_{\mathrm{A}}(T)\}$*, where*

$$\varepsilon_{\mathrm{A}}(T) = \frac{2L_0 G o_{\mathrm{A},T}}{N}\sum_{s<T}\eta_s c_{\mathrm{A},s}\prod_{t=s+1}^{T-1} c_{\mathrm{A},t}(1+\eta_t L_g). \tag{14}$$

For i.i.d. training samples, the same bound controls the absolute expected generalization gap. Compared with DPRGD under the same assumptions and step sequence,

$$\begin{aligned} r_{\mathrm{B},t} - r_{\mathrm{A},t} &= a(1-\sigma^2)\delta_t^2 + C_1(1-2\sigma)\delta_t\eta_t G \ge 0, \\ \varepsilon_{\mathrm{A}}(T) &\le \varepsilon_{\mathrm{B}}(T). \end{aligned} \tag{15}$$

The condition $\sigma < 1/2$ both closes the consensus recursion and orders the mixed terms in (15). Appendix A.9 proves separate stability guarantees under the weaker condition $\sigma < (1 - h/(4\tau))$ by sharpening the projection estimate and adjusting the steps.

The DPRGD-ATC consensus recursion is $E(X_{\mathrm{A}}^{t+1}) \le 2\sigma(E(X_{\mathrm{A}}^t) + \eta_t G)$. It gives a factor σ relative to the common DPRGD bound. Consequently the squared consensus term is reduced by σ^2 and the mixed term by $2\sigma < 1$, as summarized in Table 1. The local squared-step term is shared. These reductions give a smaller stability upper bound and hence a tighter expected generalization guarantee under identical step and communication conditions.

4.1 A LOCAL CURVATURE EFFECT ON SAMPLE SENSITIVITY

Theorem 3 compares stability upper bounds. We now compare actual sample sensitivity in a one-step problem. Three agents start together and average exactly. The outer agents take opposite tangent steps; one middle-agent sample is perturbed. DPRGD projects the outer steps separately, shifting their mean and amplifying the perturbation at the output projection. DPRGD-ATC cancels these steps before projection.

The relevant curvature is described by the second fundamental form. Let Π_x project orthogonally onto $T_x\mathcal{M}$. For a smooth curve γ in $\mathcal{M}$ with $\gamma(0) = x$ and $\gamma'(0) = v$, define

$$\mathrm{II}_x(v, v) = (I - \Pi_x)\gamma''(0).$$

This normal component of acceleration depends only on x, v, not on the chosen curve. For unit v, its norm measures the extrinsic curvature in that direction.

Proposition 4 (Local comparison of sample sensitivity)**.** *Let $\mathcal{M}$ be a compact C^3 embedded manifold without boundary, with $\dim \mathcal{M} \geq 1$. Choose $x \in \mathcal{M}$ and a unit $v \in T_x\mathcal{M}$ with $\mathrm{II}_x(v, v) \neq 0$. Set $m = 3$, $Q = J$, and $x_i^0 = x$. Use $\ell(y, z) = C - \langle z, y\rangle$, with $z \in \{sv : |s| \leq 1\}$ and $C = \max_{y\in\mathcal{M}} |\langle v, y\rangle|$. Node 1 contains n copies of v, node 2 of 0, and node 3 of $-v$. Replace one sample at node 2 by ϵv to obtain S^ϵ. For $\nu \in \{\mathrm{B}, \mathrm{A}\}$ define $\Phi_\nu(\epsilon) = \mathbb{E}_{\mathcal{I}}\ell(\widehat{x}_\nu^1(S^\epsilon), v)$. For full local gradients or uniform stochastic gradients and every sufficiently small admissible $\eta > 0$, $|\Phi'_{\mathrm{A}}(0)| = \eta/(3n)$. Moreover, as $\eta \downarrow 0$,*

$$\frac{|\Phi'_{\mathrm{B}}(0)|}{|\Phi'_{\mathrm{A}}(0)|} = 1 + \frac{\eta^2}{3}\left\|\mathrm{II}_x(v, v)\right\|^2 + o(\eta^2). \tag{16}$$

Such a point and direction exist on every manifold satisfying these geometric assumptions. The loss family satisfies Assumption 1; for sufficiently small $\eta > 0$, the derivative ratio is strictly greater than one.

The derivative $|\Phi'_\nu(0)|$ measures sensitivity of the expected test loss. Thus DPRGD-ATC has a smaller absolute loss change under sufficiently small nonzero replacements, relative to each method's unperturbed output. This complements Theorem 3 with an actual local sensitivity comparison. Differentiating the output projection gives the squared-curvature term.

An exact circle instance. For $\mathcal{M} = \mathbb{S}^1$, $x = e_1$, and $v = e_2$ (the standard coordinate vectors), $C = 1$ and the construction gives

$$|\Phi'_{\mathrm{B}}(0)| = \frac{\eta}{n(1 + 2/\sqrt{1+\eta^2})}, \qquad \frac{|\Phi'_{\mathrm{B}}(0)|}{|\Phi'_{\mathrm{A}}(0)|} = \frac{3}{1 + 2/\sqrt{1+\eta^2}} = 1 + \frac{\eta^2}{3} + O(\eta^4). \tag{17}$$

On the circle, the separately projected outer steps have a shorter mean, whose normalization amplifies the sample perturbation. Appendix A.5 proves both formulas; Figure 1(b) evaluates the circle instance.

5 CONVERGENCE TO EMPIRICAL STATIONARY POINTS

Both DPRGD and DPRGD-ATC admit empirical stationarity guarantees under the trajectory conditions used for stability. We compare their optimization errors on a fixed training dataset S, so this section requires no distributional assumption on S. Since the ambient average $\bar{x}^t$ need not be feasible, we analyze $H = F_S \circ \mathcal{P}_{\mathcal{M}}$ and transfer its gradient bound to the feasible output $\widehat{x}^t$. Lemma 13 gives $\left\|D^2 H\right\| \leq L_H := C_1 G + C_0^2 L_g$ on $\overline{U}_h$ and $\|\mathrm{grad}\, F_S(\widehat{x}^t)\|^2 \leq \chi \|\nabla H(\bar{x}^t)\|^2$, where $\chi = (1 + C_1 h)^2$.

For a constant $0 < \eta \leq \min\{\eta_*, 1/(6L_H)\}$, define the consensus, mean-distance, and projection-distance bounds

$$E_{\mathrm{B}}^* = \frac{2\eta G}{1 - 2\sigma}, \qquad E_{\mathrm{A}}^* = \sigma E_{\mathrm{B}}^*, \qquad D_\nu^* = \sqrt{2m}E_\nu^*,$$
$$q_\nu = a(D_\nu^*)^2, \qquad r_\nu^* = r_\nu(D_\nu^*, \eta), \qquad V_\nu = L_g E_\nu^* + (L_g + C_1 G)q_\nu + r_\nu^*/\eta,$$

Here $\nu = \mathrm{B}$ denotes DPRGD and $\nu = \mathrm{A}$ denotes DPRGD-ATC. The quantity V_ν bounds the update error due to consensus and projection, separately from the centered gradient noise.

Theorem 5 (Empirical stationarity of DPRGD and DPRGD-ATC)**.** *Under Assumption 1, the communication and initialization conditions of Section 2, and the constant-step restriction above, each method $\nu \in \{\mathrm{B}, \mathrm{A}\}$ satisfies, for every fixed dataset S and $T \geq 1$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{I}} \left\|\operatorname{grad} F_S(\widehat{x}_\nu^t)\right\|^2 \leq \chi \left[\frac{2(F_S(x_0) - \min_{\mathcal{M}} F_S)}{\eta T} + \frac{5}{2} V_\nu^2 + \frac{3 L_H \eta G^2}{m}\right]. \tag{18}$$

Stochastic gradients use the independent uniform sampling of Section 2. For full local gradients, the last term is zero.

The three terms represent the initial objective gap, consensus and projection error, and sampling noise, respectively. The factor $1/m$ comes from averaging independent centered gradient errors across agents. Equivalently, the left side is $\mathbb{E}_{\mathcal{I},R} \left\|\operatorname{grad} F_S(\widehat{x}_\nu^R)\right\|^2$ for R uniform on $\{0, \ldots, T-1\}$, independently of the data and $\mathcal{I}$. Thus the theorem concerns the average over iterates, or a uniformly selected iterate.

Proof mechanism. Double stochasticity gives a common mean-update identity: a gradient step plus a projection correction. The trajectory estimates bound the resulting error in the update direction by V_ν. Smooth descent for H and summation then give (18). Only the gradient noise is centered; the projection correction is controlled by its norm, accounting for its dependence on sampled gradients (Appendix A.7).

Corollary 6 (Comparison of network-related errors)**.** *Under the same assumptions, let $V_{\mathrm{loc}} = C_1 \eta G^2/2$. Then*

$$V_{\mathrm{A}} \leq V_{\mathrm{loc}} + 2\sigma(V_{\mathrm{B}} - V_{\mathrm{loc}}) \leq V_{\mathrm{B}}. \tag{19}$$

Hence the right side of (18) *for ATC is no larger than that for DPRGD, with a reduction of $5\chi(V_{\mathrm{B}}^2 - V_{\mathrm{A}}^2)/2$.*

The initial-gap and noise terms are identical for the two methods; ATC improves the consensus and projection error bound V_ν. Exact averaging ($\sigma = 0$) leaves $V_{\mathrm{A}} = V_{\mathrm{loc}}$. For fixed problem and network parameters, $V_\nu = O(\eta)$, so the bound has order $1/(\eta T) + \eta/m + \eta^2$. A fixed positive step therefore gives a residual error; choosing the step as a function of the run length yields the following rates.

Corollary 7 (Stationarity complexity)**.** *Fix the problem and network parameters and choose $0 < c \leq \min\{\eta_*, 1/(6L_H)\}$. For both DPRGD and DPRGD-ATC, $\eta = cT^{-1/2}$ gives an $O(T^{-1/2})$ expected squared gradient bound with stochastic gradients; $\eta = cT^{-1/3}$ gives $O(T^{-2/3})$ with full gradients. Thus $\mathbb{E}\left\|\operatorname{grad} F_S(\widehat{x}_\nu^R)\right\|^2 \leq \epsilon^2$ requires at most $O(\epsilon^{-4})$ and $O(\epsilon^{-3})$ iterations, respectively. The chosen step is constant throughout each run.*

ATC improves network-related constants at the same complexity orders; Appendix A.8 gives operation counts. For i.i.d. data, averaging stability over R also bounds the random output's expected loss generalization gap (Appendix A.3); its stationarity guarantee concerns F_S.

6 Numerical experiments

We evaluate the geometric mechanism, stability bounds, and observed statistical behavior separately. Appendix B specifies the synthetic PCA problems and paired protocols.

Geometric mechanism. Numerical differentiation verifies the circle sensitivity ratio in Proposition 4 to 4.8×10^{-14}. Averaging probes on sphere, Stiefel, and Grassmann give distance-versus-consensus slopes 2.0001, 2.0001, and 2.0000, respectively, supporting the predicted quadratic scaling of the distance from the averaged point to the manifold (Figure 1).

Consensus error and distance before projection. We compare 20 paired datasets on each of $\mathbb{S}^7$, $\mathrm{St}(8, 2)$, and rank-two Grassmann projectors. Four agents hold 40 samples each and use $T = 512$ updates with two ring communication rounds ($\sigma = 1/4$). Both methods share data, initialization, sampled indices, and a step satisfying the common theoretical threshold.

ATC's maximum observed consensus error is about 22% of baseline, and its maximum projection distance is 44–49% of baseline. All trajectories satisfy the geometric bounds (Figure 2).

Table 2: Uniform stability upper bounds at $T = 512$ and $N = 160$, with common steps and communication. Reduction is relative to the DPRGD bound.

Manifold	DPRGD ε_{B}	DPRGD-ATC ε_{A}	Reduction
$\mathbb{S}^7$	0.02789	0.02250	19%
$\mathrm{St}(8,2)$	0.03104	0.02281	27%
$\mathrm{Gr}(8,2)$	0.02952	0.01891	36%

Table 3: Fashion-MNIST on $\mathrm{St}(784, 8)$, with $N = 4096$, $T = 2048$, and a common $\eta = 0.1$. Means over 20 paired seeds; B and A denote DPRGD and DPRGD-ATC. $G_{\mathrm{abs}} = \mathbb{E}|F_{\mathrm{ref}} - F_S|$ is estimated across runs. The last column is the ratio of mean terminal consensus errors.

		Test loss		$10^3 G_{\mathrm{abs}}$		
m	k	B	A	B	A	$E_{\mathrm{A}}/E_{\mathrm{B}}$
4	2	0.32912	0.32913	2.1214	2.1202	0.208
8	2	0.32677	0.32677	2.0250	2.0247	0.499
16	2	0.32527	0.32527	2.0701	2.0697	0.686
32	2	0.32459	0.32458	2.0513	2.0518	0.803
32	144	0.32458	0.32457	2.0515	2.0518	0.065

Stability bound comparison. Under this common protocol, DPRGD-ATC lowers the uniform bounds by 19–36% (Table 2). The sample-size and Euclidean gradient-propagation factors are identical; the improvement comes from the geometric terms associated with consensus error.

The bounds are uniform over neighboring datasets and test examples. Distribution-averaged gap differences at these small steps have paired intervals containing zero (Appendix B.3). Appendix B.4 reports full-gradient optimization results.

Real image data and larger networks. Fashion-MNIST (Xiao et al., 2017) uses all 784 pixels on $\mathrm{St}(784, 8)$, with 4096 training images distributed across 4–32 ring nodes. A separate development split selects the common $\eta = 0.1$; 20 fresh paired seeds evaluate $T = 2048$, $k = 2$. Table 3 reports official-test loss and absolute gaps relative to the finite sampling distribution. At 32 nodes, $k = 144$ additionally restores $\sigma \leq 1/4$.

Both methods reduce test loss from 0.9883 to 0.325–0.329. At $k = 2$, ATC lowers terminal consensus error by 20–79% and maximum projection distance by 62–64%, with comparable learning progress. Absolute-gap differences are of order 10^{-6} and all paired intervals include zero. Appendix E gives intervals, trajectories, and communication costs; these larger-step runs are outside the theoretical sufficient conditions.

Fixed-baseline real-data comparison. Appendices C–D report larger-step synthetic comparisons and fixed-baseline tuning. On Fashion-MNIST, a configuration selected independently of the test set preserves learning progress and uses half the communication of fixed DPRGD. Across 50 fresh pairs per node count, the pooled absolute reference gap changes by -0.679×10^{-6}, with paired interval $[-1.324, -0.079] \times 10^{-6}$, and mean replace-one sensitivity decreases by 0.592%. Appendix F gives the selection rules, per-node results, and costs.

7 Conclusion

Manifold projection couples consensus error to sample sensitivity. Controlling this interaction yields a stability-based generalization bound for stochastic DPRGD with an explicit geometric factor. DPRGD-ATC reduces this factor under common steps and communication, while retaining empirical stationarity guarantees with improved network-related constants. A local curvature construction demonstrates strictly lower sample sensitivity, connecting the geometric analysis to an actual statistical effect.

AI USE STATEMENT

We used OpenAI ChatGPT and Codex to assist with literature search and analysis, research framing, algorithm development, formulation of mathematical claims, proof construction and checking, experimental design, implementation, analysis and interpretation of numerical results, and manuscript drafting and revision. This assistance also covered code for generating synthetic data, processing the public image dataset, preparing figures, and checking references. Numerical results were computed by running the supplied Python programs. Verification included checks of proof steps, replay of saved algorithm updates, recomputation of statistics from raw arrays, and comparison of bibliographic records with primary sources. The authors take responsibility for the final text, mathematical claims, code, and results, including all AI-assisted contributions.

REFERENCES

Thomas Bendokat, Ralf Zimmermann, and P.-A. Absil. A Grassmann manifold handbook: Basic geometry and computational aspects. *Advances in Computational Mathematics*, 50 (1), 2024. doi: 10.1007/s10444-023-10090-8. URL https://doi.org/10.1007/s10444-023-10090-8. Article 6.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002. URL https://www.jmlr.org/papers/v2/bousquet02a.html.

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.

Woocheol Choi and Jimyeong Kim. On the convergence analysis of the decentralized projected gradient descent method. *SIAM Journal on Optimization*, 35(3):1673–1702, 2025. doi: 10.1137/23m1562032. URL https://doi.org/10.1137/23m1562032.

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

Kangkang Deng and Jiang Hu. Decentralized projected riemannian stochastic recursive momentum method for nonconvex optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11203–11211, 2025b. doi: 10.1609/aaai.v39i11.33218. URL https://ojs.aaai.org/index.php/AAAI/article/view/33218.

Evan S. Gawlik and Melvin Leok. Iterative computation of the Fréchet derivative of the polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1354–1379, 2017. doi: 10.1137/16m108971x. URL https://doi.org/10.1137/16m108971x.

Florentin Goyens and Florian Feppon. The Riemannian landing method: From projected gradient flows to SQP, 2026. URL https://arxiv.org/abs/2603.24309. arXiv:2603.24309.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234. PMLR, 2016. URL https://proceedings.mlr.press/v48/hardt16.html.

Hesam Hosseini, Ying Cao, and Ali H. Sayed. Stability and generalization of adversarial diffusion training. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 926–930. IEEE, 2026. doi: 10.1109/icassp55912.2026.11465093. URL https://doi.org/10.1109/icassp55912.2026.11465093.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In *Proceedings of the 37th*

International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pp. 5381–5393. PMLR, 2020. URL https://proceedings.mlr.press/v119/koloskova20a.html.

Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Kevin Scaman, and Giovanni Neglia. Improved stability and generalization guarantees of the decentralized SGD algorithm. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26215–26240. PMLR, 2024. URL https://proceedings.mlr.press/v235/le-bars24a.html.

Gunther Leobacher and Alexander Steinicke. Existence, uniqueness and regularity of the projection onto differentiable manifolds. *Annals of Global Analysis and Geometry*, 60(3):559–587, 2021. doi: 10.1007/s10455-021-09788-z. URL https://doi.org/10.1007/s10455-021-09788-z.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f75526659f31040afeb61cb7133e4e6d-Paper.pdf.

Dominic Richards and Patrick Rebeschini. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 21(34):1–44, 2020. URL https://www.jmlr.org/papers/v21/18-638.html.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015. doi: 10.1137/14096668x. URL https://doi.org/10.1137/14096668x.

Tao Sun, Dongsheng Li, and Bao Wang. Stability and generalization of decentralized stochastic gradient descent. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9756–9764, 2021. doi: 10.1609/aaai.v35i11.17173. URL https://doi.org/10.1609/aaai.v35i11.17173.

Youbang Sun, Shixiang Chen, Alfredo Garcia, and Shahin Shahrampour. Retraction-free decentralized non-convex optimization with orthogonal constraints, 2024. URL https://arxiv.org/abs/2405.11590. Accepted for publication in IEEE Transactions on Automatic Control; arXiv:2405.11590.

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. URL https://doi.org/10.1109/tac.2012.2225533.

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. URL https://www.aimsciences.org/article/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 *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9736–9740. IEEE, 2024. doi: 10.1109/icassp48485.2024.10446975. URL https://doi.org/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, 2025. URL https://proceedings.mlr.press/v267/wang25h.html.

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

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. doi: 10.48550/arXiv.1708.07747. URL https://arxiv.org/abs/1708.07747.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016. doi: 10.1137/130943170. URL https://doi.org/10.1137/130943170.

Shuang Zeng and Yunwen Lei. Stability and generalization analysis of decentralized SGD: Sharper bounds beyond Lipschitzness and smoothness. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74098–74132. PMLR, 2025. URL https://proceedings.mlr.press/v267/zeng25d.html.

Tongtian Zhu, Fengxiang He, Lan Zhang, Zhengyang Niu, Mingli Song, and Dacheng Tao. Topology-aware generalization of decentralized SGD. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27479–27503. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhu22d.html.

Tongtian Zhu, Wenhao Li, Can Wang, and Fengxiang He. DICE: Data influence cascade in decentralized learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vRYcHLqKgx.

A COMPLETE PROOFS

All norms below are induced Euclidean norms or Frobenius norms. A stacked state has one node per row; tensor-valued states are vectorized only for notation. Projection acts separately on each node. In each method's proof, y_i^t denotes the argument of $\mathcal{P}_{\mathcal{M}}$ in its update: $y_i^t = \sum_j Q_{ij} x_j^t - \eta_t g_i^t$ for DPRGD and $y_i^t = \sum_j Q_{ij}(x_j^t - \eta_t g_j^t)$ for DPRGD-ATC. We write Π_x for the orthogonal projector onto $T_x\mathcal{M}$ and $\Pi_x^\perp = I - \Pi_x$. The constants in Assumption 1 apply to the closure of the indicated neighborhood; derivatives there are understood from the larger projection domain U_τ.

A.1 GEOMETRIC ESTIMATES

Lemma 8 (Normal chords and projection). *For $p, q \in \mathcal{M}$ and $n \in N_p\mathcal{M}$,*

$$|\langle n, q-p\rangle| \le \frac{\|n\|}{2\tau}\|q-p\|^2, \qquad \left\|\Pi_p^\perp(q-p)\right\| \le \frac{\|q-p\|^2}{2\tau}. \tag{20}$$

If $u, v \in U_\tau$ have distances r_u, r_v from $\mathcal{M}$, then

$$\|\mathcal{P}_{\mathcal{M}}(u) - \mathcal{P}_{\mathcal{M}}(v)\| \le \left(1 - \frac{r_u + r_v}{2\tau}\right)^{-1}\|u - v\|. \tag{21}$$

In particular $\|D\mathcal{P}_{\mathcal{M}}(u)\| \le C_0$ on $\overline{U}_h$ and $D\mathcal{P}_{\mathcal{M}}(x) = \Pi_x$ for $x \in \mathcal{M}$.

Proof. For a unit normal $\widehat{n}$ at p, the normal-ray property gives $\mathcal{P}_{\mathcal{M}}(p + s\widehat{n}) = p$ for $0 < s < \tau$. Specifically, the second assertion of Proposition 2 in Leobacher & Steinicke (2021) identifies the projection along each normal ray up to its frontier distance, and Proposition 6 bounds that distance below by the reach. These results apply since $\mathcal{M}$ is a compact C^3 embedded manifold without boundary. Comparing the squared distances from $p + s\widehat{n}$ to p and q gives

$$s^2 \le \|q - p - s\widehat{n}\|^2 = \|q-p\|^2 + s^2 - 2s\langle \widehat{n}, q-p\rangle.$$

Let $s \uparrow \tau$ and use both signs of $\widehat{n}$ to obtain the first inequality, including $n = 0$ directly. Taking the supremum over unit normals gives the second inequality.

Put $p = \mathcal{P}_{\mathcal{M}}(u)$, $q = \mathcal{P}_{\mathcal{M}}(v)$, $n = u - p$, and $n' = v - q$. These residuals are normal at their projections. Therefore

$$\begin{aligned}\langle u - v, p - q\rangle &= \|p-q\|^2 + \langle n, p-q\rangle - \langle n', p-q\rangle \\ &\ge \left(1 - \frac{r_u + r_v}{2\tau}\right)\|p-q\|^2.\end{aligned}$$

Cauchy–Schwarz proves (21); the factor is positive. Taking $v \to u$ gives the derivative bound. At a manifold point, projection differentiates to identity on tangent directions and zero on normal directions, proving $D\mathcal{P}_{\mathcal{M}}(x) = \Pi_x$. □

Lemma 9 (Tangent steps and tangent-space variation). *If $x \in \mathcal{M}$, $v \in T_x\mathcal{M}$, and $\|v\| \le h$, then*

$$\operatorname{dist}(x + v, \mathcal{M}) \le \frac{C_1}{2}\|v\|^2. \tag{22}$$

If $x, y \in \mathcal{M}$ with $\|x - y\| \le h$, then

$$\|\Pi_x - \Pi_y\| \le C_1\|x - y\|. \tag{23}$$

For any ambient increment u with $\|u\| \le h$,

$$\operatorname{dist}(x + u, \mathcal{M}) \le \left\|\Pi_x^\perp u\right\| + \frac{C_1}{2}\|\Pi_x u\|^2. \tag{24}$$

Proof. The segment $x + sv$, $0 \le s \le 1$, lies in $\overline{U}_h$. Taylor's formula and $D\mathcal{P}_{\mathcal{M}}(x)v = v$ give

$$\mathcal{P}_{\mathcal{M}}(x + v) = x + v + \int_0^1 (1-s)D^2\mathcal{P}_{\mathcal{M}}(x + sv)[v, v]\, ds.$$

Taking norms proves (22). The chord from x to y lies in $\overline{U}_h$ as well. Integrating $D^2\mathcal{P}_{\mathcal{M}}$ along it, and using $D\mathcal{P}_{\mathcal{M}}(x) = \Pi_x$ at both endpoints, proves (23). Finally, distance to a closed set is 1-Lipschitz. Apply this fact to $x + u$ and $x + \Pi_x u$, and then apply (22). □

Lemma 10 (Distances before projection and at the output)**.** *Suppose $x_i \in \mathcal{M}$, $D(X) \le D$, $g_i \in T_{x_i}\mathcal{M}$, $\|g_i\| \le G$, and $D + \eta G \le h$. Then*

$$\operatorname{dist}\left(\sum_j Q_{ij}x_j - \eta g_i, \mathcal{M}\right) \le r_{\mathrm{B}}(D,\eta), \quad \operatorname{dist}\left(\sum_j Q_{ij}(x_j - \eta g_j), \mathcal{M}\right) \le r_{\mathrm{A}}(D,\eta).$$

Also $\operatorname{dist}(m^{-1}\sum_i x_i, \mathcal{M}) \le aD^2$.

Proof. Rewrite the mixed point relative to x_i. For $s_i = \sum_j Q_{ij}(x_j - x_i)$, Lemma 8 and nonnegative row weights imply

$$\|s_i\| \le D, \qquad \left\|\Pi_{x_i}^{\perp} s_i\right\| \le \frac{D^2}{2\tau}.$$

For the baseline, $u_i = s_i - \eta g_i$ has normal part at most $D^2/(2\tau)$ and tangent part at most $D + \eta G$. Lemma 9 bounds its distance by

$$\frac{D^2}{2\tau} + \frac{C_1}{2}(D + \eta G)^2 = aD^2 + C_1 D\eta G + \frac{C_1}{2}\eta^2 G^2.$$

For DPRGD-ATC, tangency of g_j at x_j gives

$$\left\|\Pi_{x_i}^{\perp} g_j\right\| = \left\|(\Pi_{x_j} - \Pi_{x_i})g_j\right\| \le C_1 DG.$$

Thus $u_i = s_i - \eta\sum_j Q_{ij}g_j$ has normal part at most $D^2/(2\tau) + \eta C_1 DG$ and tangent part at most $D + \eta G$. Adding these contributions proves the DPRGD-ATC formula. For the output mean, take uniform row weights and $\eta = 0$ in the baseline calculation. □

A.2 Projection domain and consensus error

Lemma 11 (Averaging identity)**.** *For arbitrary ambient node values u_i and $\overline{u} = m^{-1}\sum_i u_i$,*

$$E(U)^2 = \frac{1}{m}\sum_i \|u_i - \overline{u}\|^2 = \frac{1}{2m^2}\sum_{i,j}\|u_i - u_j\|^2.$$

Consequently, if a map is L-Lipschitz on every pair of node values, then $E(\operatorname{map}(U)) \le LE(U)$. Also $D(U) \le \sqrt{2m}E(U)$.

Proof. Expand the pairwise squared distances and cancel the cross terms. Applying the pairwise Lipschitz inequality proves the second assertion. For $i \ne j$,

$$\|u_i - u_j\|^2 \le 2\|u_i - \overline{u}\|^2 + 2\|u_j - \overline{u}\|^2 \le 2\sum_r \|u_r - \overline{u}\|^2 = 2mE(U)^2.$$

The case $i = j$ is immediate; maximizing proves the diameter estimate. □

Explicit common step threshold. For the proofs below, abbreviate $\rho = 2\sigma < 1$ and define

$$S_* = \frac{2\sqrt{2m}}{1-\rho}, \qquad K_* = aS_*^2 + C_1 S_* + \frac{C_1}{2}, \qquad \eta_* = \min\left\{\frac{h}{4G(S_*+1)}, \sqrt{\frac{h}{4K_*G^2}}\right\}. \quad (25)$$

These constants ensure both the step-length and distance-to-manifold conditions in the next induction.

Lemma 12 (Uniform feasible trajectories)**.** *For $0 \le \eta_t \le \eta_*$ and common feasible initialization, every run of both methods satisfies the domain and consensus error conclusions of Theorems 2 and 3. The output means also lie within distance $h/4$ of $\mathcal{M}$.*

Proof. Put $\overline{\eta} = \eta_*$. Equation (8) gives

$$e_t = 2G\sum_{s<t}\rho^{t-1-s}\eta_s \le \frac{2\overline{\eta}G}{1-\rho}, \qquad \delta_t \le \delta_* := S_*\overline{\eta}G.$$

The common restriction implies

$$\delta_* + \overline{\eta} G \le h/4, \qquad r_{\mathrm{B}}(\delta_*, \overline{\eta}) = K_* \overline{\eta}^2 G^2 \le h/4. \tag{26}$$

Furthermore, because $0 \le \sigma < 1/2$,

$$r_{\mathrm{A}}(\sigma D, \eta) = a\sigma^2 D^2 + 2C_1 \sigma D \eta G + \frac{C_1}{2} \eta^2 G^2 \le r_{\mathrm{B}}(D, \eta).$$

At time zero, both consensus errors vanish. Suppose inductively $E(X_{\mathrm{B}}^t) \le e_t$ and $E(X_{\mathrm{A}}^t) \le \sigma e_t$. Lemma 11 gives the corresponding diameter bounds. Lemma 10 and (26) show that the points before projection in both methods are at distance at most $h/4 < \tau/2$ from $\mathcal{M}$. Their projections are unique and, by Lemma 8, pairwise at most 2-Lipschitz.

Let Y_ν^t stack the points y_i^t before projection of method ν, and let G_t stack its local gradients. Then $m^{-1/2} \|(I - J)G_t\|_F \le G$, since an orthogonal projector contracts the Frobenius norm. Using $QJ = JQ = J$ gives

$$E(Y_{\mathrm{B}}^t) \le \sigma E(X_{\mathrm{B}}^t) + \eta_t G,$$
$$E(Y_{\mathrm{A}}^t) \le \sigma \big(E(X_{\mathrm{A}}^t) + \eta_t G \big).$$

After projection, these yield

$$E(X_{\mathrm{B}}^{t+1}) \le \rho e_t + 2\eta_t G = e_{t+1},$$
$$E(X_{\mathrm{A}}^{t+1}) \le \rho \sigma e_t + 2\sigma \eta_t G = \sigma e_{t+1}.$$

This closes the joint induction for the trajectory domain and consensus error. Finally, Lemma 10 bounds the mean distances by $a\delta_t^2$ and $a\sigma^2\delta_t^2$, respectively; both are at most $h/4$ by (26). □

A.3 Replace-one stability and expected generalization

We use the stability definition (6), sharing the index array between neighboring runs. All suprema range over the sample family in Assumption 1.

Proof of Theorem 2. The pathwise assertions follow from Lemma 12. Fix neighboring datasets differing at entry j_* of node i_*. Couple their initial state and indices, using the same fixed step sequence. Write $\Delta_{i,t} = \|x_i^t - x_i'^t\|$ and let B_t be the indicator of $I_{i_*}^t = j_*$. The gradient regularity and a triangle inequality give

$$\big\| g_i^t - g_i'^t \big\| \le L_g \Delta_{i,t} + 2G \mathbf{1}_{\{i = i_*\}} B_t.$$

Indeed, compare the two gradients first at the same sample and then, only for the replaced index, at different samples at the same point. For DPRGD, the difference between y_i^t and $y_i'^t$ satisfies

$$\big\| y_i^t - y_i'^t \big\| \le \sum_j Q_{ij} \Delta_{j,t} + \eta_t L_g \Delta_{i,t} + 2\eta_t G \mathbf{1}_{\{i = i_*\}} B_t.$$

Each run's points y_i^t are within distance $r_{\mathrm{B},t}$ of the manifold. Lemma 8 compares arbitrary pairs in this tube and therefore controls their projections throughout training. The inequality holds pathwise, including the dependence between the current perturbation and gradient. Sum over i, divide by m, and condition on the joint history. Column stochasticity and $\mathbb{E}[B_t \mid \text{history}] = 1/n$ give (10).

Since $d_0 = 0$, recursion gives

$$d_T \le \frac{2G}{N} \sum_{s<T} \eta_s c_{\mathrm{B},s} \prod_{t=s+1}^{T-1} c_{\mathrm{B},t} (1 + \eta_t L_g).$$

Both output means are within distance $a\delta_T^2$ of $\mathcal{M}$, so Lemma 8 and the triangle inequality imply

$$\mathbb{E} \big\| \widehat{x}^T - \widehat{x}'^T \big\| \le o_{\mathrm{B},T} \mathbb{E} \big\| \bar{x}^T - \bar{x}'^T \big\| \le o_{\mathrm{B},T} d_T.$$

The L_0-Lipschitz property proves (9); bounded loss supplies the cap M_0.

For completeness, enumerate the full dataset as $z_1, \ldots, z_N$ and let z'_k be an independent copy. Let $S^{(k)}$ replace z_k by z'_k. Exchangeability of (z_k, z'_k) gives

$$\begin{aligned}\mathbb{E}[F(\mathcal{A}(S)) - F_S(\mathcal{A}(S))] &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}[\ell(\mathcal{A}(S), z'_k) - \ell(\mathcal{A}(S), z_k)] \\ &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}[\ell(\mathcal{A}(S^{(k)}), z_k) - \ell(\mathcal{A}(S), z_k)].\end{aligned}$$

Apply (6) and the triangle inequality. This establishes the stated bound on the absolute expected signed generalization gap. □

Proof of Theorem 3. Lemma 12 proves the uniform domain and smaller consensus error bound. With the same neighboring-run coupling, DPRGD-ATC satisfies

$$\left\|y_i^t - y_i'^t\right\| \le \sum_j Q_{ij}\left[(1+\eta_t L_g)\Delta_{j,t} + 2\eta_t G \mathbf{1}_{\{j=i_*\}} B_t\right].$$

Column stochasticity makes the averaged injection again $2\eta_t G/N$. Applying Lemma 8 at radius $r_{\mathrm{A},t}$ yields the same recursion with $c_{\mathrm{A},t}$. The output radius is $a\sigma^2\delta_T^2$. The preceding proof therefore establishes stability and expected generalization for DPRGD-ATC.

Subtract the two radius polynomials to obtain (15). Every factor in the sum-product stability expression is nonnegative. The inequalities $r_{\mathrm{A},t} \le r_{\mathrm{B},t}$ and $o_{\mathrm{A},T} \le o_{\mathrm{B},T}$ thus prove $\varepsilon_{\mathrm{A}}(T) \le \varepsilon_{\mathrm{B}}(T)$. Applying the same cap M_0 preserves this ordering. The ordered quantities are the explicit stability upper bounds in the theorem. □

Full-gradient and random-output variants. For full local gradients, replacing one sample changes $\operatorname{grad} F_{i_*}$ at a fixed point by at most $2G/n$. Thus the averaged one-step injection is again $2\eta_t G/(mn)$, and the same proof gives the same stability expression with no sampling expectation. For a random output, let R have fixed probabilities p_t on a finite set of iterations, independently of the data and sampled indices, and share R between the coupled runs. Conditioning on R gives, for every neighboring pair and test sample z,

$$\mathbb{E}_{R,\mathcal{I}}|\ell(\widehat{x}_\nu^R(S), z) - \ell(\widehat{x}_\nu^R(S'), z)| \le \sum_t p_t \min\{M_0, \varepsilon_\nu(t)\}.$$

The preceding exchangeability argument gives the same bound on the absolute expected generalization gap. Uniform $p_t = 1/T$ on $0 \le t < T$ gives the random-output rule in the optimization theorem.

A.4 The accumulated geometric factor

All radius ratios are below $1/2$. The elementary inequality $-\log(1-u) \le 2u$ for $0 \le u \le 1/2$ yields

$$\prod_{t=s}^{T-1} c_{\nu,t} \le \exp\left(\frac{2}{\tau}\sum_{t<T} r_{\nu,t}\right).$$

Extract this common upper bound from the stability sum to prove (11) and its DPRGD-ATC counterpart.

To control the radius sums, extend the finite step sequence by zero and use

$$\delta_t = 2\sqrt{2m}G\sum_{s<t}\rho^{t-1-s}\eta_s.$$

By weighted Cauchy–Schwarz,

$$\left(\sum_{s<t}\rho^{t-1-s}\eta_s\right)^2 \le \frac{1}{1-\rho}\sum_{s<t}\rho^{t-1-s}\eta_s^2.$$

Summing over $t < T$ gives the displayed squared consensus error bound in the main text. Put $U = \sum_{t<T} \eta_t^2$. Another application of Cauchy–Schwarz gives

$$\sum_{t<T} r_{\mathrm{B},t} \le G^2 \left[\frac{8am}{(1-\rho)^2} + \frac{2C_1\sqrt{2m}}{1-\rho} + \frac{C_1}{2} \right] U,$$
$$\sum_{t<T} r_{\mathrm{A},t} \le G^2 \left[\frac{8a\sigma^2 m}{(1-\rho)^2} + \frac{4C_1\sigma\sqrt{2m}}{1-\rho} + \frac{C_1}{2} \right] U.$$

The output factors are uniformly bounded by the common step restriction. To include the final output projection in (3), weighted Cauchy–Schwarz also gives

$$\delta_T^2 \le \frac{8mG^2}{1-\rho} \sum_{s<T} \rho^{T-1-s} \eta_s^2 \le \frac{8mG^2}{1-\rho} U.$$

Since $a\delta_T^2/\tau < 1/2$, we have $o_{\mathrm{B},T} \le \exp(2a\delta_T^2/\tau)$. Combining this estimate with (11) and the preceding radius sum proves (3) with

$$C_{\mathrm{geom}} = \frac{2G^2}{\tau} \left[\frac{8am}{(1-\rho)^2} + \frac{8am + 2C_1\sqrt{2m}}{1-\rho} + \frac{C_1}{2} \right].$$

The first inequality in (3) is Theorem 3. Thus this constant includes the output projection and is independent of T and the dataset. The radius sums above display the individual terms reduced by ATC. For an affine constraint, nearest projection is nonexpansive, averages and tangent steps remain feasible, and the geometric factors can instead be taken as one. The affine case therefore recovers the Euclidean propagation bound.

A.5 Local curvature and sample sensitivity

Proof of Proposition 4. Let Π_x be the tangent projector and set $B = \mathrm{II}_x(v, v)$. Nearest projection is C^2 near x, and $D\mathcal{P}_{\mathcal{M}}(x)v = v$. Differentiating $\mathcal{P}_{\mathcal{M}}(\gamma(t)) = \gamma(t)$ twice along a manifold curve with $\gamma(0) = x$, $\gamma'(0) = v$ gives $D^2\mathcal{P}_{\mathcal{M}}(x)[v, v] = B$.

For the loss $\ell(y, z) = C - \langle z, y \rangle$, $\operatorname{grad} \ell(x, sv) = -sv$. Before replacing the sample, put

$$p_\pm = \mathcal{P}_{\mathcal{M}}(x \pm \eta v), \qquad q_\eta = \frac{p_+ + x + p_-}{3}.$$

The DPRGD output is $\mathcal{P}_{\mathcal{M}}(q_\eta)$ and the ATC output is x. For full gradients, the middle node's local gradient is $-(\epsilon/n)v$. Differentiating its projected step at $\epsilon = 0$ yields $(\eta/n)v$, so the two loss derivatives are

$$\Phi'_{\mathrm{B}}(0) = -\frac{\eta}{3n} \left\langle v, D\mathcal{P}_{\mathcal{M}}(q_\eta)v \right\rangle, \qquad \Phi'_{\mathrm{A}}(0) = -\frac{\eta}{3n}. \tag{27}$$

For stochastic gradients, the changed sample is selected at the middle node with probability $1/n$. Conditional on selection, the same calculation replaces ϵ/n by ϵ; otherwise the output is independent of ϵ. Differentiation of this finite weighted sum gives exactly (27). Each algorithm's loss change is measured relative to its own output at $\epsilon = 0$.

Second-order Taylor expansion at x gives

$$p_\pm = x \pm \eta v + \frac{\eta^2}{2} B + o(\eta^2), \qquad q_\eta = x + \frac{\eta^2}{3} B + o(\eta^2).$$

For a normal vector $n \in N_x\mathcal{M}$, define the shape operator S_n by $\langle S_n v, w \rangle = \langle n, \mathrm{II}_x(v, w) \rangle$. To identify the mixed projection derivative, choose a tangent field w and differentiate the normality identity $\langle u - \mathcal{P}_{\mathcal{M}}(u), w(\mathcal{P}_{\mathcal{M}}(u)) \rangle = 0$. It gives

$$(I - S_{u - \mathcal{P}_{\mathcal{M}}(u)}) D\mathcal{P}_{\mathcal{M}}(u) = \Pi_{\mathcal{P}_{\mathcal{M}}(u)}.$$

Along the normal ray $u = x + sn$, Lemma 8 gives $\mathcal{P}_{\mathcal{M}}(u) = x$ for sufficiently small $|s|$. Differentiating $(I - sS_n)D\mathcal{P}_{\mathcal{M}}(x + sn) = \Pi_x$ at $s = 0$, then applying it to v, yields $D^2\mathcal{P}_{\mathcal{M}}(x)[n, v] = S_n v$.

Hence

$$D\mathcal{P}_{\mathcal{M}}(q_\eta)v = v + \frac{\eta^2}{3} S_B v + o(\eta^2),$$
$$\langle v, D\mathcal{P}_{\mathcal{M}}(q_\eta)v\rangle = 1 + \frac{\eta^2}{3}\langle B, \mathrm{II}_x(v,v)\rangle + o(\eta^2) = 1 + \frac{\eta^2}{3}\|B\|^2 + o(\eta^2).$$

This inner product is positive for sufficiently small η, so (27) proves (16). When $B \neq 0$, its ratio is strictly larger than one. For each such fixed step, differentiability then gives a strict inequality between the absolute loss changes for sufficiently small nonzero replacements. Under stochastic sampling, the coupled loss change is zero on a miss and deterministic on a hit. Its expected absolute value therefore equals the absolute change in expected loss, connecting this construction directly to (6).

To verify existence of a curved direction, maximize $y \mapsto \|y - c\|^2$ on $\mathcal{M}$, for any fixed ambient point c, and call a maximizer x. Its first-order condition makes $x - c$ normal. For any unit tangent v and a smooth manifold curve with this velocity, the second-order condition gives

$$0 \geq \frac{1}{2}\frac{d^2}{dt^2}\|\gamma(t) - c\|^2\Big|_{t=0} = 1 + \langle x - c, \mathrm{II}_x(v,v)\rangle .$$

Thus $\mathrm{II}_x(v,v) \neq 0$. The assumption $\dim \mathcal{M} \geq 1$ ensures that a unit tangent exists.

Finally, compactness gives finite $C = \max_{y\in\mathcal{M}} |\langle v, y\rangle|$. It is positive because $\frac{d}{dt}\langle v, \gamma(t)\rangle|_{t=0} = 1$ for a curve with initial velocity v. Thus $0 \leq \ell(y, sv) \leq 2C$. The family has $L_0 = G = 1$. Its gradient is $-s\Pi_y v$. Lemma 9 bounds $\|\Pi_y - \Pi_{y'}\|$ by $C_1\|y - y'\|$ for chords of length at most h; for longer chords, the bound $2 \leq (2/h)\|y - y'\|$ applies. Therefore $L_g = \max\{C_1, 2/h\}$ is valid. The loss is smooth and all constants are uniform for $|s| \leq 1$. Choosing η below the common step threshold supplies the full setting of Assumption 1. □

Proof of the circle formula (17). For $\ell(x,v) = 1 - v^\top x$ on $\mathbb{S}^1$, $\operatorname{grad}\ell(x,v) = -(I - xx^\top)v$. At the common initial point, the full local gradients are $-e_2$, $-(\epsilon/n)e_2$, and e_2. Write $r = \sqrt{1+\eta^2}$. After one baseline step the three states are

$$x^1_{1,\mathrm{B}} = (e_1 + \eta e_2)/r, \quad x^1_{2,\mathrm{B}} = \frac{e_1 + \eta\epsilon e_2/n}{\sqrt{1 + (\eta\epsilon/n)^2}}, \quad x^1_{3,\mathrm{B}} = (e_1 - \eta e_2)/r.$$

The sum of these states at $\epsilon = 0$ is $(1 + 2/r)e_1$, and its derivative is $(\eta/n)e_2$. Normalization of a vector ce_1 differentiates to $(I - e_1e_1^\top)/c$. The derivative of the baseline output is consequently $\eta e_2/[n(1 + 2/r)]$.

DPRGD-ATC first averages the locally updated models, so all its states and its output equal

$$\widehat{x}^1_{\mathrm{A}}(S^\epsilon) = \frac{e_1 + \eta\epsilon e_2/(3n)}{\sqrt{1 + (\eta\epsilon/(3n))^2}}.$$

Its derivative at zero is $\eta e_2/(3n)$. Applying the test loss $1 - e_2^\top x$ proves both expressions. Taylor expansion of $(1+\eta^2)^{-1/2}$ proves the ratio expansion. The ratio exceeds one for every $\eta > 0$. Both test losses equal one at zero and are differentiable there, so the strict derivative inequality persists for the magnitudes of their changes at sufficiently small nonzero ϵ.

For stochastic gradients, all samples at the outer nodes are identical. At the middle node, the replaced sample is selected with probability $1/n$. Conditional on a hit, the same calculation uses ϵ in place of ϵ/n in the middle gradient. Its output derivatives are $\eta e_2/(1 + 2/r)$ and $\eta e_2/3$. Conditional on a miss, both outputs are e_1 for every ϵ. The expected loss is a finite weighted sum of smooth functions, so differentiation commutes with this expectation. Multiplication by $1/n$ recovers exactly the two derivatives in the circle formula. The expected losses agree at zero, and the smaller ATC derivative magnitude gives strictly lower sample sensitivity for sufficiently small nonzero replacements under stochastic sampling.

The sample family satisfies $M_0 = 2$, $G = 1$, $L_0 = 1$, and, for example, $L_g = 2$, since $\left\|xx^\top - yy^\top\right\| \leq 2\|x - y\|$. Thus the construction lies in the main assumptions for every sufficiently small positive η; its explicit formulas also hold beyond that sufficient step range. It uses actual single-entry replacements with $|\epsilon| \leq 1$ and a common data-independent initialization. □

A.6 SMOOTHNESS OF THE PROJECTED OBJECTIVE

Lemma 13 (Exact extension). *Let $H = F_S \circ \mathcal{P}_{\mathcal{M}}$. On $\overline{U}_h$,*

$$\nabla H(u) = D\mathcal{P}_{\mathcal{M}}(u) \operatorname{grad} F_S(\mathcal{P}_{\mathcal{M}}(u)), \qquad \left\|D^2 H(u)\right\| \le L_H = C_1 G + C_0^2 L_g,$$

and

$$\|\operatorname{grad} F_S(\mathcal{P}_{\mathcal{M}}(u))\|^2 \le (1 + C_1 h)^2 \|\nabla H(u)\|^2 . \tag{28}$$

The smoothness descent inequality with constant L_H holds on any segment contained in $\overline{U}_h$.

Proof. Write $p = \mathcal{P}_{\mathcal{M}}(u)$ and $\mathfrak{g}(p) = \operatorname{grad} F_S(p)$, viewed as an ambient-vector-valued C^1 field on $\mathcal{M}$. The image of $D\mathcal{P}_{\mathcal{M}}(u)$ lies in $T_p\mathcal{M}$, since projection takes values in $\mathcal{M}$. As $u - p$ is normal at p, differentiation gives

$$D\big(\tfrac{1}{2}\operatorname{dist}(u, \mathcal{M})^2\big)[a] = \langle u - p, a - D\mathcal{P}_{\mathcal{M}}(u)a\rangle = \langle u - p, a\rangle .$$

Thus projection is the gradient of $u \mapsto \|u\|^2/2 - \operatorname{dist}(u, \mathcal{M})^2/2$, and $D\mathcal{P}_{\mathcal{M}}(u)$ is self-adjoint. For any ambient direction a, the chain rule gives

$$DH(u)[a] = \langle \mathfrak{g}(p), D\mathcal{P}_{\mathcal{M}}(u)a\rangle = \langle D\mathcal{P}_{\mathcal{M}}(u)\mathfrak{g}(p), a\rangle .$$

This proves the gradient formula. To bound the derivative of $\mathfrak{g}$, choose a smooth manifold curve γ with $\gamma(0) = p$ and $\gamma'(0) = v \in T_p\mathcal{M}$. The chord-Lipschitz assumption, inherited by the empirical average, gives

$$\|D\mathfrak{g}(p)[v]\| = \lim_{s\to 0} \frac{\|\mathfrak{g}(\gamma(s)) - \mathfrak{g}(p)\|}{|s|} \le L_g \|v\| .$$

Differentiating the gradient formula now yields

$$D(\nabla H)(u)[a] = D^2\mathcal{P}_{\mathcal{M}}(u)[a, \mathfrak{g}(p)] + D\mathcal{P}_{\mathcal{M}}(u)D\mathfrak{g}(p)[D\mathcal{P}_{\mathcal{M}}(u)a].$$

The two terms have norms at most $C_1 G\|a\|$ and $C_0^2 L_g \|a\|$, respectively. Taking the supremum over unit a proves the Hessian bound. For a segment $u + sd$ contained in $\overline{U}_h$, Taylor's formula consequently gives

$$H(u + d) - H(u) - \langle \nabla H(u), d\rangle = \int_0^1 (1 - s) D^2 H(u + sd)[d, d]\, ds \le \frac{L_H}{2}\|d\|^2 .$$

For the gradient transfer, set $n = u - p$. Define the self-adjoint shape operator $S_n : T_p\mathcal{M} \to T_p\mathcal{M}$ by $\langle S_n v, w\rangle = \langle n, \mathrm{II}_p(v, w)\rangle$, where $\mathrm{II}_p(v, w)$ is the normal component of the ambient derivative of a tangent field with value w in direction v. Differentiating $\mathcal{P}_{\mathcal{M}}(\gamma(s)) = \gamma(s)$ twice at zero gives

$$D^2\mathcal{P}_{\mathcal{M}}(p)[v, v] + \mathrm{II}_p\gamma''(0) = \gamma''(0), \qquad D^2\mathcal{P}_{\mathcal{M}}(p)[v, v] = \mathrm{II}_p(v, v).$$

It follows that $|\langle S_n v, v\rangle| \le C_1 \|n\| \|v\|^2$. Self-adjointness therefore implies $\|S_n\| \le C_1\|n\|$.

For any local tangent field w, differentiate $\langle u - \mathcal{P}_{\mathcal{M}}(u), w(\mathcal{P}_{\mathcal{M}}(u))\rangle = 0$ in an arbitrary ambient direction a:

$$\begin{aligned} 0 &= \langle a - D\mathcal{P}_{\mathcal{M}}(u)a, w(p)\rangle + \langle n, Dw(p)[D\mathcal{P}_{\mathcal{M}}(u)a]\rangle \\ &= \langle \Pi_p a - D\mathcal{P}_{\mathcal{M}}(u)a + S_n D\mathcal{P}_{\mathcal{M}}(u)a, w(p)\rangle . \end{aligned}$$

Here the second equality uses the definition of II and the tangency of $D\mathcal{P}_{\mathcal{M}}(u)a$. Since $w(p)$ can be any tangent vector, this proves

$$(I - S_n)D\mathcal{P}_{\mathcal{M}}(u) = \Pi_p.$$

Apply this identity to $\mathfrak{g}(p)$ and use $\Pi_p\mathfrak{g}(p) = \mathfrak{g}(p)$ to obtain

$$\mathfrak{g}(p) = (I - S_n)\nabla H(u), \qquad \|\mathfrak{g}(p)\| \le (1 + C_1\|n\|)\|\nabla H(u)\| .$$

Squaring and using $\|n\| \le h$ proves (28). □

A.7 PROOF OF EMPIRICAL STATIONARITY FOR BOTH METHODS

Proof of Theorem 5. Fix either method and suppress its subscript. Work conditionally on the fixed training dataset; the argument is uniform in that dataset. Let $\mathcal{F}_t$ contain the history before the fresh sampled indices. Put $\overline{g}_t = m^{-1}\sum_i g_i^t$ and

$$\overline{g}_t^0 = \frac{1}{m}\sum_i \operatorname{grad} F_i(x_i^t), \qquad \xi_t = \overline{g}_t - \overline{g}_t^0.$$

Set $\zeta_{i,t} = g_i^t - \operatorname{grad} F_i(x_i^t)$. Conditional on $\mathcal{F}_t$, these vectors are independent and centered, since each fresh index is sampled uniformly. Therefore the cross terms vanish and

$$\begin{aligned}\mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_t] &= \frac{1}{m^2}\sum_i \mathbb{E}[\|\zeta_{i,t}\|^2 \mid \mathcal{F}_t] \\ &= \frac{1}{m^2}\sum_i \Big(\mathbb{E}[\|g_i^t\|^2 \mid \mathcal{F}_t] - \|\operatorname{grad} F_i(x_i^t)\|^2\Big) \le \frac{G^2}{m}.\end{aligned}$$

Together with conditional centering, this gives

$$\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_t] \le G^2/m. \tag{29}$$

For full gradients $\xi_t = 0$. Define $b_i^t = \mathcal{P}_{\mathcal{M}}(y_i^t) - y_i^t$ and $\overline{b}_t = m^{-1}\sum_i b_i^t$. Both update orders have the same mean identity, by column stochasticity:

$$\bar{x}^{t+1} = \bar{x}^t - \eta\overline{g}_t + \overline{b}_t, \qquad \|\overline{b}_t\| \le r_\nu^*. \tag{30}$$

The preceding domain proof gives $E(X_t) \le E_\nu^*$ and $\operatorname{dist}(\bar{x}^t, \mathcal{M}) \le q_\nu$. Let $p_t = \mathcal{P}_{\mathcal{M}}(\bar{x}^t)$. Then

$$\frac{1}{m}\sum_i \|x_i^t - p_t\| \le E_\nu^* + q_\nu,$$

and hence

$$\|\overline{g}_t^0 - \operatorname{grad} F_S(p_t)\| \le L_g(E_\nu^* + q_\nu).$$

The normal segment from p_t to $\bar{x}^t$ lies in the projection neighborhood. Integrating $D^2\mathcal{P}_{\mathcal{M}}$ along it gives $\|D\mathcal{P}_{\mathcal{M}}(\bar{x}^t) - \Pi_{p_t}\| \le C_1 q_\nu$. Consequently

$$\|\overline{g}_t^0 - \nabla H(\bar{x}^t)\| \le L_g E_\nu^* + (L_g + C_1 G)q_\nu.$$

Writing

$$u_t = \overline{g}_t^0 - \nabla H(\bar{x}^t) - \overline{b}_t/\eta$$

gives $\|u_t\| \le V_\nu$ and the exact recursion

$$\bar{x}^{t+1} = \bar{x}^t - \eta\big(\nabla H(\bar{x}^t) + \xi_t + u_t\big). \tag{31}$$

The error u_t can depend on the fresh noise through projection; we use only its deterministic norm bound.

To justify descent on the nonconvex tube, every point on the segment from $\bar{x}^t$ to $\bar{x}^{t+1}$ has distance to $\mathcal{M}$ at most

$$q_\nu + \|\bar{x}^{t+1} - \bar{x}^t\| \le q_\nu + \eta G + r_\nu^* \le 3h/4 < h.$$

Thus Lemma 13 applies on the full segment. Set $v_t = \nabla H(\bar{x}^t)$. Smoothness, (31), and the inequality $\|v+\xi+u\|^2 \le 3(\|v\|^2 + \|\xi\|^2 + \|u\|^2)$ imply

$$\begin{aligned}H(\bar{x}^{t+1}) - H(\bar{x}^t) \le\ & -\eta\|v_t\|^2 - \eta\langle v_t, \xi_t\rangle - \eta\langle v_t, u_t\rangle \\ & + \frac{3L_H\eta^2}{2}(\|v_t\|^2 + \|\xi_t\|^2 + \|u_t\|^2).\end{aligned}$$

Young's inequality bounds $-\langle v_t, u_t\rangle \le \|v_t\|^2/4 + \|u_t\|^2$. Conditional expectation removes only the ξ_t linear term. Since $L_H\eta \le 1/6$, we obtain

$$\mathbb{E}[H(\bar{x}^{t+1}) - H(\bar{x}^t)] \le -\frac{\eta}{2}\mathbb{E}\|v_t\|^2 + \frac{5}{4}\eta V_\nu^2 + \frac{3}{2}L_H\eta^2 G^2/m.$$

Summing for $0 \le t < T$ telescopes the left-hand side. Since $H(\bar{x}^0) = F_S(x_0)$ and $H(\bar{x}^T) \ge \min_{\mathcal{M}} F_S$, we obtain

$$\frac{\eta}{2} \sum_{t<T} \mathbb{E} \left\| v_t \right\|^2 \le F_S(x_0) - \min_{\mathcal{M}} F_S + \frac{5}{4} \eta T V_\nu^2 + \frac{3 L_H \eta^2 T G^2}{2m}.$$

Dividing by $\eta T/2$ gives

$$\frac{1}{T} \sum_{t<T} \mathbb{E} \left\| v_t \right\|^2 \le \frac{2(F_S(x_0) - \min_{\mathcal{M}} F_S)}{\eta T} + \frac{5}{2} V_\nu^2 + \frac{3 L_H \eta G^2}{m}.$$

Since $\widehat{x}_\nu^t = p_t$, the gradient-transfer estimate (28) gives the average-iterate bound. Equivalently, for R uniform on $\{0, \dots, T-1\}$ independently of the sampled indices,

$$\mathbb{E} \left\| \operatorname{grad} F_S(\widehat{x}_\nu^R) \right\|^2 = \frac{1}{T} \sum_{t<T} \mathbb{E} \left\| \operatorname{grad} F_S(p_t) \right\|^2 \le \frac{\chi}{T} \sum_{t<T} \mathbb{E} \left\| v_t \right\|^2 .$$

Substitution proves (18), conditionally on every admissible training dataset. Full gradients set $\xi_t = 0$ throughout and remove the variance contribution. Finally, $E_{\mathrm{A}}^* \le E_{\mathrm{B}}^*$, $q_{\mathrm{A}} \le q_{\mathrm{B}}$, and the radius comparison gives $r_{\mathrm{A}}^* \le r_{\mathrm{B}}^*$; every coefficient in V_ν is nonnegative, so $V_{\mathrm{A}} \le V_{\mathrm{B}}$. □

Proof of Corollary 6. Write $E = E_{\mathrm{B}}^*$ and $D = D_{\mathrm{B}}^*$. Expanding the definitions of $V_{\mathrm{B}}, V_{\mathrm{A}}$ gives

$$V_{\mathrm{B}} - V_{\mathrm{A}} = (1-\sigma) L_g E + (1-\sigma^2) a D^2 \left(L_g + C_1 G + \frac{1}{\eta} \right) + (1 - 2\sigma) C_1 D G \ge 0.$$

Subtract the common $V_{\mathrm{loc}} = C_1 \eta G^2/2$ and compute

$$2\sigma(V_{\mathrm{B}} - V_{\mathrm{loc}}) - (V_{\mathrm{A}} - V_{\mathrm{loc}}) = \sigma L_g E + (2\sigma - \sigma^2) a D^2 \left(L_g + C_1 G + \frac{1}{\eta} \right) \ge 0.$$

This proves (19), since $V_{\mathrm{B}} \ge V_{\mathrm{loc}}$ and $2\sigma < 1$. All remaining terms in the two stationarity bounds coincide, so their difference is $5\chi(V_{\mathrm{B}}^2 - V_{\mathrm{A}}^2)/2$. □

A.8 Stationarity and operation complexity

Proof of Corollary 7. Put $\overline{\eta} = \min\{\eta_*, 1/(6L_H)\}$ and $\Delta = F_S(x_0) - \min_{\mathcal{M}} F_S \ge 0$. To display the dependence of V_ν on the step, define

$$e_{\mathrm{B}} = \frac{2G}{1-2\sigma}, \qquad e_{\mathrm{A}} = \sigma e_{\mathrm{B}}, \qquad d_\nu = \sqrt{2m} e_\nu, \qquad \kappa_{\mathrm{B}} = 1, \quad \kappa_{\mathrm{A}} = 2.$$

Then $E_\nu^* = e_\nu \eta$, $D_\nu^* = d_\nu \eta$, and direct substitution gives

$$V_\nu(\eta) = \left(L_g e_\nu + a d_\nu^2 + \kappa_\nu C_1 d_\nu G + \frac{C_1 G^2}{2} \right) \eta + (L_g + C_1 G) a d_\nu^2 \eta^2.$$

All coefficients are nonnegative. For $0 < \eta \le \overline{\eta}$, it follows that

$$V_\nu(\eta) \le K_\nu \eta, \qquad K_\nu = L_g e_\nu + a d_\nu^2 + \kappa_\nu C_1 d_\nu G + \frac{C_1 G^2}{2} + (L_g + C_1 G) a d_\nu^2 \overline{\eta}.$$

These constants do not depend on T or the target accuracy. They explicitly retain the dependence on m, σ, and the manifold and loss constants. Corollary 6, applied at $\eta = \overline{\eta}$, gives $K_{\mathrm{A}} \le K_{\mathrm{B}}$.

Fix $0 < c \le \overline{\eta}$ and an integer $T \ge 1$. For stochastic gradients, substitute $\eta = cT^{-1/2}$ into Theorem 5 to obtain

$$\begin{aligned} \mathbb{E} \left\| \operatorname{grad} F_S(\widehat{x}_\nu^R) \right\|^2 &\le \chi \left[\left(\frac{2\Delta}{c} + \frac{3 L_H c G^2}{m} \right) T^{-1/2} + \frac{5}{2} K_\nu^2 c^2 T^{-1} \right] \\ &\le C_{\nu,\mathrm{sg}} T^{-1/2}, \qquad C_{\nu,\mathrm{sg}} = \chi \left[\frac{2\Delta}{c} + \frac{3 L_H c G^2}{m} + \frac{5}{2} K_\nu^2 c^2 \right]. \end{aligned}$$

Here the second inequality uses $T^{-1} \le T^{-1/2}$. With full gradients the variance term vanishes. Substituting $\eta = cT^{-1/3}$ gives

$$\mathbb{E}_R \left\| \operatorname{grad} F_S(\widehat{x}_\nu^R) \right\|^2 \le C_{\nu,\mathrm{fg}} T^{-2/3}, \qquad C_{\nu,\mathrm{fg}} = \chi \left[\frac{2\Delta}{c} + \frac{5}{2} K_\nu^2 c^2 \right].$$

For any $\epsilon > 0$, sufficient horizons are therefore

$$T_{\nu,\mathrm{sg}} = \max\left\{1, \left\lceil (C_{\nu,\mathrm{sg}}/\epsilon^2)^2 \right\rceil\right\}, \qquad T_{\nu,\mathrm{fg}} = \max\left\{1, \left\lceil (C_{\nu,\mathrm{fg}}/\epsilon^2)^{3/2} \right\rceil\right\}.$$

The steps are fixed during each run and satisfy the common restriction. If a dataset-independent choice of horizon is desired, replace Δ by M_0 in the constants, which remains valid by boundedness of the loss. This also makes the horizon compatible with the stability setup. At a common choice of c, $K_{\mathrm{A}} \le K_{\mathrm{B}}$ implies the corresponding ordering of both complexity constants. □

Operation counts. Each stochastic iteration evaluates one sample gradient and one nearest-point projection at each node, and performs k synchronous neighbor communication rounds. We count a full local gradient as n sample-gradient evaluations. Thus the total sample-gradient counts over all nodes are mT and mnT, respectively; both methods use mT node projections and kT training communication rounds. Each round exchanges one ambient model-sized vector across each active directed communication link. The counts describe gradient and projection calls; the arithmetic cost of an individual projection depends on the manifold and its representation.

To implement the random output, draw and share R independently before training, retain x_i^R at each node, and perform one final sum reduction to compute $\bar{x}^R$, followed by a projection at the output node. On a fixed spanning tree this takes $m-1$ model-vector transmissions, with a further $m-1$ if the output is broadcast to all nodes. This final aggregation is in addition to the kT training rounds and does not depend on T. These counts give the operation complexity for the feasible output used in the theorem.

A.9 Stability under weaker mixing

The separate stability guarantees require less mixing than the ordering of their bounds. Let

$$L_h = \frac{1}{1 - h/(4\tau)}, \qquad b_m = \sqrt{2m}, \qquad L_h \sigma < 1,$$

and retain all other assumptions of Section 2, replacing only its condition $\sigma < 1/2$ by the displayed condition. Define

$$e_0^{(L)} = 0, \qquad e_{t+1}^{(L)} = L_h \sigma e_t^{(L)} + L_h \eta_t G, \qquad \delta_t^{(L)} = b_m e_t^{(L)}, \tag{32}$$
$$S_L = \frac{b_m L_h}{1 - L_h \sigma}, \quad K_{\mathrm{B},L} = a S_L^2 + C_1 S_L + C_1/2, \quad K_{\mathrm{A},L} = a\sigma^2 S_L^2 + 2C_1 \sigma S_L + C_1/2.$$

Put $K_L = \max\{K_{\mathrm{B},L}, K_{\mathrm{A},L}\}$ and

$$\eta_L = \min\left\{ \frac{h}{4G(S_L+1)}, \sqrt{\frac{h}{4K_L G^2}} \right\}. \tag{33}$$

Theorem 14 (Individual bounds under weaker mixing)**.** *For $0 \le \eta_t \le \eta_L$, both algorithms have unique projections throughout their trajectories and at their outputs. Their consensus errors satisfy*

$$E(X_{\mathrm{B}}^t) \le e_t^{(L)}, \qquad E(X_{\mathrm{A}}^t) \le \sigma e_t^{(L)}.$$

Their points before projection and output means have distance at most $h/4$ from $\mathcal{M}$. Substituting $\delta_t^{(L)}$ for δ_t in the definitions of $r_{\nu,t}, c_{\nu,t}, o_{\nu,T}$ and in (9)*,* (14) *gives valid uniform expected stability bounds for each method. For common-distribution i.i.d. samples, the same bounds control the absolute expected generalization gaps.*

Proof. The deterministic recursion gives

$$e_t^{(L)} = L_h G \sum_{s<t} (L_h\sigma)^{t-1-s} \eta_s \le \frac{L_h \eta_L G}{1 - L_h \sigma}, \qquad \delta_t^{(L)} \le S_L \eta_L G.$$

Thus (33) ensures

$$\delta_t^{(L)} + \eta_t G \le h/4, \qquad r_{\mathrm{B}}(\delta_t^{(L)}, \eta_t) \le K_{\mathrm{B},L}\eta_L^2 G^2 \le h/4,$$

and $r_{\mathrm{A}}(\sigma\delta_t^{(L)}, \eta_t) \le K_{\mathrm{A},L}\eta_L^2 G^2 \le h/4$.

We prove feasibility and consensus control together. Both consensus errors vanish at initialization. At step t, the inductive consensus bounds and Lemma 11 give diameters at most $\delta_t^{(L)}$ and $\sigma\delta_t^{(L)}$. The preceding step-length condition permits Lemma 10; its radius bounds place every point before projection within distance $h/4 < \tau$ of $\mathcal{M}$. Hence projection is unique, and Lemma 8 gives the pairwise Lipschitz constant L_h on these points. With Y_ν^t denoting their stack, double stochasticity and the gradient norm bound yield

$$E(Y_{\mathrm{B}}^t) \le \sigma E(X_{\mathrm{B}}^t) + \eta_t G,$$
$$E(Y_{\mathrm{A}}^t) \le \sigma(E(X_{\mathrm{A}}^t) + \eta_t G).$$

The averaging identity in Lemma 11 now gives

$$E(X_{\mathrm{B}}^{t+1}) \le L_h\sigma e_t^{(L)} + L_h\eta_t G = e_{t+1}^{(L)},$$
$$E(X_{\mathrm{A}}^{t+1}) \le L_h\sigma(\sigma e_t^{(L)} + \eta_t G) = \sigma e_{t+1}^{(L)}.$$

This closes the induction without using an ordering of the two radius bounds. Lemma 10 also bounds the output-mean distances by $a(\delta_t^{(L)})^2$ and $a\sigma^2(\delta_t^{(L)})^2$, both at most $h/4$ by the respective $K_{\nu,L}$ bounds.

The same deterministic estimates apply to every index array on either neighboring dataset. The coupling argument in Appendix A.3 therefore gives

$$d_{t+1} \le c_{\nu,t}\big[(1 + \eta_t L_g)d_t + 2\eta_t G/N\big], \qquad d_0 = 0.$$

Unrolling and applying the output factor $o_{\nu,T}$ proves the stated stability expressions. Bounded loss supplies the cap M_0, and the exchangeability argument in Appendix A.3 proves the generalization statement. □

Since $h < \tau/4$, $L_h < 16/15$, and $\sigma < 15/16$ is a sufficient mixing condition for this theorem. Any fixed $\sigma < 1$ can instead be accommodated by choosing $0 < h < \min\{\tau/4, 4\tau(1-\sigma)\}$ and using the associated projection constants and step restriction. The comparison in Theorem 3 additionally uses

$$r_{\mathrm{B},t} - r_{\mathrm{A},t} = a(1-\sigma^2)(\delta_t^{(L)})^2 + C_1(1-2\sigma)\delta_t^{(L)}\eta_t G.$$

For $\sigma > 1/2$, this expression need not be nonnegative for arbitrary step sequences. The weaker mixing theorem consequently supplies individual guarantees, while $\sigma < 1/2$ ensures their ordering.

For a symmetric doubly stochastic one-round matrix W with $\lambda = \|W - J\|_2 \in (0,1)$, $Q = W^k$ satisfies $\sigma = \lambda^k$. Thus the smallest integer giving $\sigma < 1/2$ is $\lfloor \log(1/2)/\log\lambda \rfloor + 1$. For the lazy ring with weights $1/2$ on each node and $1/4$ on each of its two neighbors, the eigenvalues are $\cos^2(\pi j/m)$, $j = 0, \dots, m-1$. Hence $\lambda = \cos^2(\pi/m)$ and maintaining a fixed contraction factor requires $k = \Theta(m^2)$. Both update orders use the same k rounds per gradient step; relaxing the target contraction reduces this communication constant but also changes the sufficient step threshold.

B Manifold instances and experimental details

The experiments study a single-sample sensitivity construction, trajectory-level geometry, and statistical behavior under stochastic training. Sphere, Stiefel, and Grassmann instantiate a shared PCA loss family in three embeddings, each with exact nearest-point projection.

B.1 Feasible models and exact operators

For $z \in \mathbb{R}^d$ with $\|z\| \le 1$, consider the following instances. All embeddings use their induced Euclidean metric.

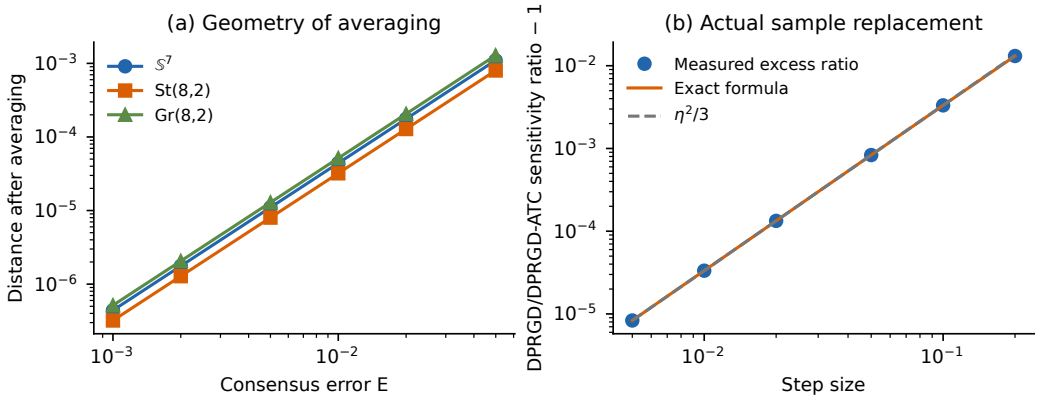


Figure 1: Geometry and sample sensitivity. (a) Averaging feasible points produces a distance from the manifold that is second order in the consensus error. (b) A single-sample replacement under stochastic sampling gives lower sensitivity for DPRGD-ATC in the circle instance (17). Points use exact hit/miss expectations and numerical differentiation; lines show the formula and its leading term.

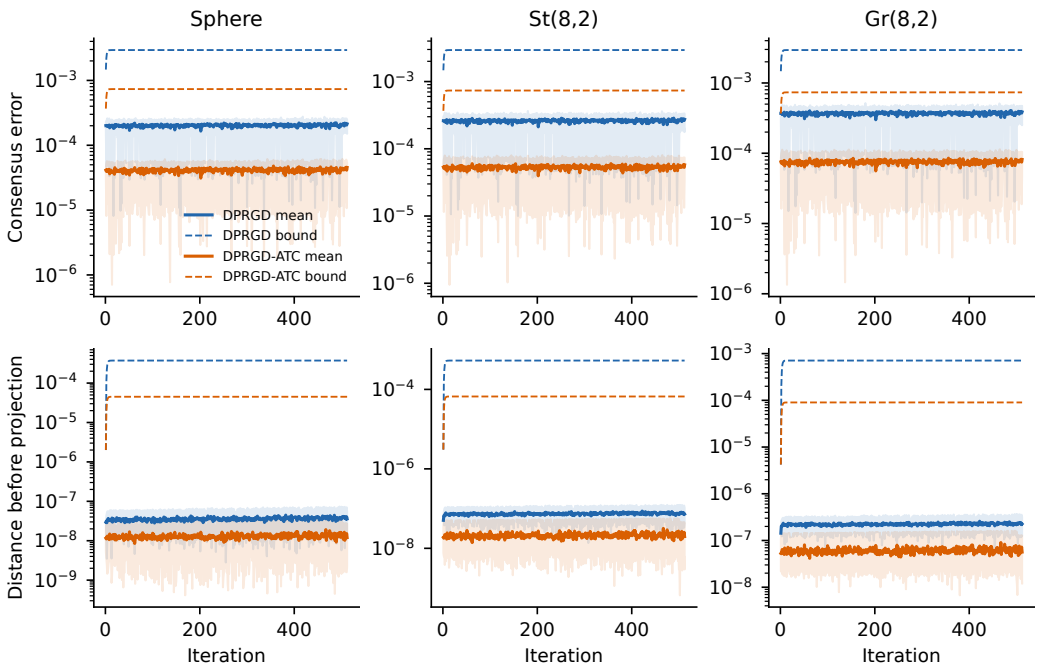


Figure 2: DPRGD-ATC reduces consensus error (top) and distance before projection (bottom) during stochastic training on all three geometries. Solid curves average 20 original-data runs, shading spans their minimum–maximum range, and dashed curves are the deterministic upper bounds.

Sphere. For $x \in \mathbb{S}^{d-1}$, set $\ell(x, z) = 1 - (x^\top z)^2$. The nearest projection and gradient are

$$\mathcal{P}_{\mathcal{M}}(y) = y/\|y\|, \qquad \operatorname{grad} \ell(x, z) = -2(I - xx^\top)zz^\top x.$$

The loss is in $[0, 1]$ and satisfies $L_0 = 2$, $G = 2$, $L_g = 8$. The sphere has reach one. We use the common valid lower bound $\tau = 1/2$ across the three geometries.

Stiefel manifold. For $X \in \mathrm{St}(d,r) = \{X : X^\top X = I_r\}$, use $\ell(X,z) = 1 - \left\|X^\top z\right\|^2$. If $Y = U\Sigma V^\top$ is a thin singular-value decomposition with positive singular values, then

$$\mathcal{P}_{\mathcal{M}}(Y) = UV^\top, \qquad \operatorname{grad}\ell(X,z) = -2zz^\top X + 2XX^\top zz^\top X.$$

The same constants $M_0 = 1$, $L_0 = 2$, $G = 2$, $L_g = 8$ apply. To verify the Lipschitz claims, expand the difference of the loss into two bilinear terms. For the gradient, the linear term has Lipschitz constant 2, and the product $XX^\top zz^\top X$ has constant at most 3 by a three-term telescoping expansion. This gives $2 + 2 \cdot 3 = 8$. The gradient norm bound follows from $\operatorname{grad}\ell(X,z) = -2(I - XX^\top)zz^\top X$.

Every matrix at Frobenius distance less than $1/2$ from the Stiefel manifold has smallest singular value greater than $1/2$, by comparison with a nearest feasible matrix and the singular-value perturbation inequality. Its nearest orthonormal factor is uniquely the polar factor. Thus $\tau = 1/2$ supplies a valid reach lower bound for the common step condition. The polar derivative can be evaluated through a Sylvester equation (Gawlik & Leok, 2017).

Grassmann projectors. Represent a rank-r subspace by its orthogonal projector $X = X^\top = X^2$, $\operatorname{tr} X = r$, in the Euclidean space of symmetric matrices. Take $\ell(X,z) = 1 - z^\top Xz$ and $A = zz^\top$. Then

$$\operatorname{grad}\ell(X,z) = -(XA + AX - 2XAX), \qquad \mathcal{P}_{\mathcal{M}}(Y) = U_r U_r^\top,$$

where U_r contains the leading r eigenvectors of symmetric Y. The loss is in $[0,1]$ and admits $L_0 = 1$, $G = 1$, $L_g = 6$. In a basis splitting the range and null space of X, the gradient has off-diagonal blocks $-(Xz)((I-X)z)^\top$ and its transpose. Its squared norm is $2\left\|Xz\right\|^2\left\|(I-X)z\right\|^2 \le 1/2$. Telescoping the expression for $\operatorname{grad}\ell$ gives the stated Lipschitz constant. At distance less than $1/2$ from a rank-r projector, the leading and trailing eigenvalues lie respectively above and below $1/2$, so the nearest rank-r projector is unique. Again $\tau = 1/2$ is valid. The projector embedding represents the subspace independently of its basis (Bendokat et al., 2024).

For the training runs, set $h = 0.1$, $\tau = 0.5$, and use the following explicit upper bounds for $\sup_{\overline{U}_h}\left\|D^2\mathcal{P}_{\mathcal{M}}\right\|$:

$$C_{1,\mathrm{sphere}} = \frac{6}{(1-h)^2}, \quad C_{1,\mathrm{St}} = \frac{3(1+h)}{(1-h)^3} + \frac{3(1+h)^3}{(1-h)^5}, \quad C_{1,\mathrm{Gr}} = \frac{1}{(1/2-h)^3}. \tag{34}$$

These explicit bounds are dimension independent in the Frobenius norm and determine the common admissible step. The updates evaluate exact projections; the projection derivatives enter the analysis through the constants derived below.

Derivation of the projection constants. For the sphere, differentiating normalization twice yields three bilinear terms with coefficient at most $1/\left\|y\right\|^2$ and a fourth term with coefficient at most $3/\left\|y\right\|^2$. Since $\|y\| \ge 1 - h$, their sum proves the first bound.

For Stiefel, write $\mathcal{P}_{\mathcal{M}}(Y) = Yf(B)$, $B = Y^\top Y$, $f(B) = B^{-1/2}$, and set $s = 1 - h$, $u = 1 + h$. In the closed tube, $\sigma_{\min}(Y) \ge s$ and $\|Y\|_2 \le u$. The resolvent identity

$$f(B) = \frac{1}{\pi}\int_0^\infty t^{-1/2}(B + tI)^{-1}\,dt$$

and differentiation under the integral give

$$\|Df(B)[H]\|_F \le \frac{1}{2s^3}\|H\|_F, \quad \left\|D^2 f(B)[H,K]\right\|_F \le \frac{3}{4s^5}\|H\|_F\|K\|_F.$$

The scalar integrals are $\pi/(2s^3)$ and $3\pi/(8s^5)$; the second derivative has two resolvent products. Also $\|DB(Y)[H]\|_F \le 2u\|H\|_F$ and $\left\|D^2B(Y)[H,K]\right\|_F \le 2\|H\|_F\|K\|_F$. Applying the product and chain rules to $Yf(B)$ gives two terms of size u/s^3, a term of size $3u^3/s^5$, and a term of size u/s^3. This proves $3u/s^3 + 3u^3/s^5$.

For Grassmann projectors, every symmetric Y in the closed tube has spectrum within h of $\{0,1\}$ and exactly r eigenvalues near one. Let Γ be the counterclockwise circle of radius $1/2$ centered at one. The spectral projection is

$$\mathcal{P}_{\mathcal{M}}(Y) = \frac{1}{2\pi\mathrm{i}}\int_\Gamma (zI - Y)^{-1}\,dz.$$

On this circle the resolvent norm is at most $(1/2 - h)^{-1}$. Its second derivative is the sum of two products containing three resolvents and the two perturbation matrices. The contour has length π, so the Frobenius bilinear bound is $(1/2 - h)^{-3}$. This also proves uniform smoothness of the projection throughout the chosen tube.

B.2 Data, communication, and execution

We use $d = 8$, $m = 4$, and rank two for the matrix instances. Samples are the standard coordinate vectors $e_1, \ldots, e_8$ with probabilities $p = (8, 5, 2, 1, 1, 1, 1, 1)/20$. The first column of the normalized Sylvester Hadamard matrix initializes the sphere; its first two columns initialize Stiefel, and their projector initializes Grassmann. These points are data independent.

The original training study uses $Q = W^2$, with W the four-cycle matrix having diagonal entries $1/2$ and neighbor entries $1/4$. Each iteration includes two neighbor exchanges, with no projection between communication rounds. We set $\tau = 0.5$, $h = 0.1$, use the derivative bounds in (34), and retain the executed steps $\eta = 3.6764705882 \times 10^{-4}$ for sphere and Stiefel and $7.3529411765 \times 10^{-4}$ for Grassmann. These steps satisfy the revised common threshold. The gradients are evaluated at the stored feasible states. DPRGD uses (4), whereas DPRGD-ATC uses (12). The Stiefel geometry and normalized quadratic model use the implementation accompanying Chen et al. (2021); the two updates are implemented explicitly. Computation uses double precision and synchronous-node simulation on one CPU server, with three independent jobs processed concurrently.

For each geometry, the stochastic study uses 20 datasets at $n = 40$, each paired with a 512-step index array. Its seed for unit $j \in \{0, \ldots, 19\}$ is $24000 + j$; training labels are sampled from p, then a fresh replacement is independently sampled from p. The replacement is at position zero of node zero. Indices are generated by seed $724000 + j$, uniformly with replacement at each node and time. Every method and neighboring run uses the identical index array. Two of the 20 replacements are neutral on each geometry and are retained. There are $3 \times 20 \times 2 \times 2 = 240$ stochastic trajectories.

For the six full-gradient optimization runs, each node has $n = 40$ samples with exact local proportions

$$p_1 = p + (e_1 - e_2)/20, \quad p_2 = p - (e_1 - e_2)/20, \quad p_3 = p + (e_3 - e_4)/20, \quad p_4 = p - (e_3 - e_4)/20.$$

Both methods use 2048 iterations and the same common step as in the stochastic study. All full gradients are calculated from the exact empirical covariance sufficient statistic. Together the training experiments contain 246 trajectories and 135,168 iteration steps.

Figure 1(a) measures the geometry of averaging feasible points. At the common reference point, generate two unit tangent directions using seed 810. For each $s \in \{0.001, 0.002, 0.005, 0.01, 0.02, 0.05\}$, form four feasible points $\mathcal{P}_{\mathcal{M}}(x_0 \pm sH_1)$ and $\mathcal{P}_{\mathcal{M}}(x_0 \pm sH_2)$, mix with W, and measure the maximum normal distance against $E(X)$. These 18 observations use the same projection operators as the current paper. These geometric probes use one round of W; the training trajectories use $Q = W^2$, satisfying $\sigma < 1/2$.

For Figure 1(b), use the circle formula (17) with $n = 20$ and $\eta \in \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$. The stochastic expectation is computed exactly from a hit of probability $1/n$ and a miss of probability $1 - 1/n$. Centered derivatives at replacement widths 0.002 and 0.001 are combined by Richardson extrapolation. Full-gradient derivatives are also evaluated, with widths 0.02 and 0.01. Their maximum ratio errors are 4.8×10^{-14} and 2.4×10^{-13}, respectively. All displayed steps lie in the validity range of the explicit single-step construction. The multi-step training runs use the separate common sufficient condition stated above.

B.3 Statistics and complete numerical values

For each geometry, the training comparison uses 20 i.i.d. datasets with $n = 40$ examples per agent and $T = 512$ stochastic iterations. Each paired unit includes its independently replaced dataset and both methods. We record $\widehat{s}_\nu$, the mean of $\mathbb{E}_{z \sim p} |\ell(\widehat{x}_\nu^T(S), z) - \ell(\widehat{x}_\nu^T(S'), z)|$ over the 20 units, computing the test-distribution expectation exactly. The steps are 3.6765×10^{-4} for sphere and Stiefel and 7.3529×10^{-4} for Grassmann, all within the common sufficient limits.

Table 4 quantifies the reduction in uniform stability bounds alongside the measured sample sensitivities. The sensitivity ratios are close to one, with paired differences of order 10^{-9}. The larger

Table 4: Stability bounds and measured stochastic sensitivity at $T = 512$, $N = 160$. $\widehat{s}$ is the distribution-averaged absolute loss change, averaged over paired runs.

Manifold	ε_{B}	ε_{A}	$10^4\widehat{s}_{\mathrm{B}}$	$\widehat{s}_{\mathrm{A}}/\widehat{s}_{\mathrm{B}}$
$\mathbb{S}^7$	0.02789	0.02250	2.6990	1.000030
$\mathrm{St}(8,2)$	0.03104	0.02281	5.2580	1.000008
$\mathrm{Gr}(8,2)$	0.02952	0.01891	5.2580	1.000008

Table 5: Signed generalization gaps, in units of 10^{-4}, and the DPRGD-ATC-minus-DPRGD paired difference, in units of 10^{-8}. Brackets are 95% paired-unit bootstrap intervals.

Manifold	Method	10^4 mean gap	10^4 interval
$\mathbb{S}^7$	DPRGD	1.5289	[-3.868, 7.554]
$\mathbb{S}^7$	DPRGD-ATC	1.5289	[-3.868, 7.554]
$\mathrm{St}(8,2)$	DPRGD	0.6928	[-8.713, 10.669]
$\mathrm{St}(8,2)$	DPRGD-ATC	0.6928	[-8.713, 10.669]
$\mathrm{Gr}(8,2)$	DPRGD	0.6928	[-8.713, 10.669]
$\mathrm{Gr}(8,2)$	DPRGD-ATC	0.6928	[-8.713, 10.669]
Manifold	Paired difference	10^8 mean	10^8 interval
$\mathbb{S}^7$	DPRGD-ATC−DPRGD	0.2367	[-0.829, 1.236]
$\mathrm{St}(8,2)$	DPRGD-ATC−DPRGD	0.3841	[-0.658, 1.413]
$\mathrm{Gr}(8,2)$	DPRGD-ATC−DPRGD	0.3849	[-0.654, 1.413]

uniform bounds cover all admissible neighboring datasets and test samples, whereas the measured sensitivities average over the test distribution and paired runs. The trajectory measurements below identify reduced consensus error and projection distance as the geometric improvements associated with the smaller bounds.

The mean signed generalization gaps for DPRGD are 1.529×10^{-4}, 6.928×10^{-5}, and 6.928×10^{-5}. DPRGD-ATC differs by 2.37×10^{-9}, 3.84×10^{-9}, and 3.85×10^{-9}. All gap intervals and paired difference intervals for both gaps and sensitivities include zero. These estimates describe similar observed statistical behavior in this training protocol, alongside the reductions in geometry-related bounds.

Population risk is computed exactly from the eight-point distribution. The signed gap is $F(\widehat{x}^T) - F_S(\widehat{x}^T)$. Paired sensitivity averages the absolute loss change over this same distribution. Bootstrap intervals use 4096 resamples of the 20 complete dataset/replacement/index units, with random seed 81271. The resampling is paired across algorithms. These are pointwise descriptive 95% percentile intervals. The independent resampling unit is the complete dataset/replacement/index tuple, including its full trajectory and test-distribution evaluation.

The paired sensitivity differences (DPRGD-ATC minus DPRGD) have means 8.06×10^{-9}, 4.18×10^{-9}, and 4.22×10^{-9} on sphere, Stiefel, and Grassmann. Their respective intervals are $[-7.44, 23.79] \times 10^{-9}$, $[-12.29, 19.91] \times 10^{-9}$, and $[-12.26, 19.95] \times 10^{-9}$. These paired estimates quantify the small observed differences in Table 4.

Figure 2 shows reduced consensus error and distance to the manifold with DPRGD-ATC. The maximum consensus errors over the original runs are $(2.762, 3.652, 5.165) \times 10^{-4}$ for DPRGD and $(0.6151, 0.8031, 1.1357) \times 10^{-4}$ for DPRGD-ATC. The maximum distances before projection are $(0.7347, 1.2409, 3.6792) \times 10^{-7}$ and $(0.3562, 0.5700, 1.6165) \times 10^{-7}$. All stored-state consensus errors, distances before projection, and mean distances in original and replaced trajectories are within the specified bounds up to rounding residuals below 10^{-14}. The stochastic path bounds use the realized replacement-hit indicators, whose conditional expectations give the $1/n$ selection probability in the stability theorem.

Table 6: Full-gradient optimization after 2048 iterations. The mean squared gradient is averaged over iterations $t = 0, \ldots, 2047$, as in the random-iterate theorem. B_{opt} is the right side of Theorem 5.

Manifold	Method	Final loss	Final $\|\operatorname{grad} F_S\|^2$	Mean $\|\operatorname{grad} F_S\|^2$	B_{opt}
$\mathbb{S}^7$	DPRGD	0.818518	0.088316	0.075011	10.487
$\mathbb{S}^7$	DPRGD-ATC	0.818518	0.088316	0.075011	2.339
$\mathrm{St}(8,2)$	DPRGD	0.648186	0.148566	0.135218	29.783
$\mathrm{St}(8,2)$	DPRGD-ATC	0.648186	0.148566	0.135218	5.212
$\mathrm{Gr}(8,2)$	DPRGD	0.648186	0.074283	0.067609	19.773
$\mathrm{Gr}(8,2)$	DPRGD-ATC	0.648186	0.074283	0.067609	3.757

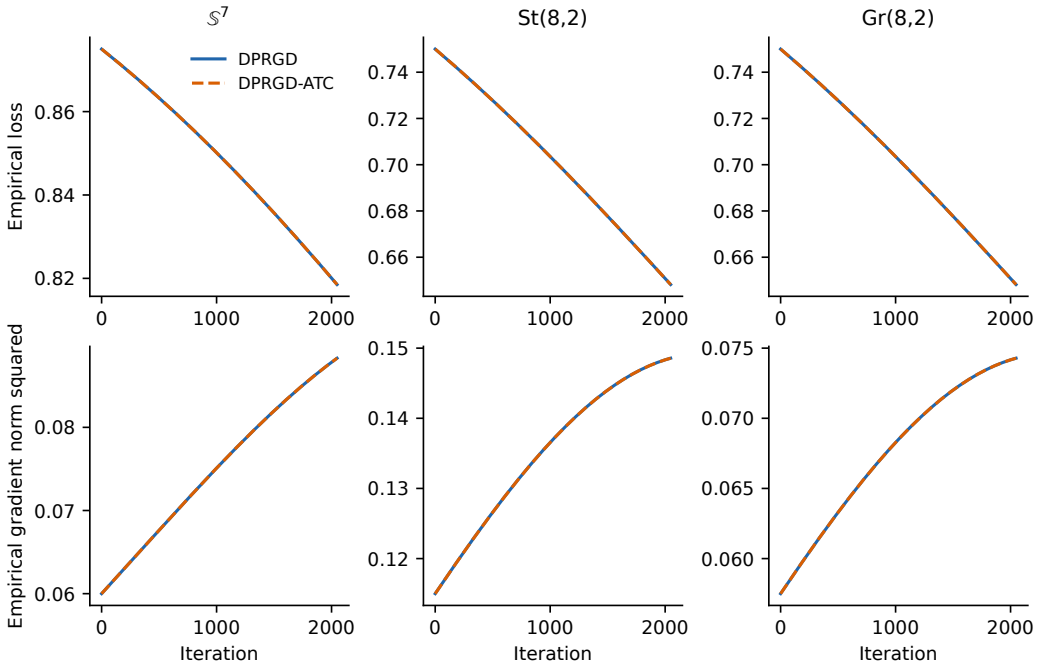


Figure 3: Full-gradient objective and squared gradient norm under common step sizes and communication. The two methods show nearly overlapping trajectories; the objective decreases on all three geometries.

B.4 Empirical optimization

Six full-gradient trajectories use the same steps and communication for 2048 iterations on fixed heterogeneous data. Both methods decrease the empirical loss from 0.875 to 0.8185 on the sphere and from 0.750 to 0.6482 on the matrix instances. DPRGD-ATC achieves this objective progress with smaller evaluated stationarity bounds: 2.339, 5.212, and 3.757, compared with 10.487, 29.783, and 19.773 for DPRGD. The table reports these bounds together with the final and time-averaged squared gradients, connecting the optimization comparison to Theorem 5.

Figure 3 complements the objective comparison with gradient trajectories. Squared gradient norms start at 0.060, 0.115, and 0.0575 and increase over this horizon as the losses decrease. The time-averaged squared gradients remain below the evaluated stationarity bounds; the table makes their different scales explicit. For exact averaging, the $Q = J$ centralized projected-minibatch identity is verified on one common-initialization update per geometry, with maximum entrywise discrepancy below 2.3×10^{-16}.

Table 7: Mean sample sensitivity $\widehat{s}$ with shared parameters. Parentheses give 95% paired-bootstrap intervals for the reduction. Values below 10^{-14} are at numerical precision.

Setting	DPRGD	DPRGD-ATC	Reduction
$\mathrm{St}(8,2)$, i.i.d.	6.30×10^{-3}	$< 10^{-14}$	$\approx 100\%$
$\mathrm{St}(8,2)$, heterogeneous	6.37×10^{-2}	8.59×10^{-3}	86.5% $(56.7, 100)$
$\mathrm{St}(64,4)$, i.i.d.	1.10×10^{-4}	3.73×10^{-5}	66.2% $(36.1, 86.8)$

B.5 Geometric scope and communication

The analysis applies to smooth compact embedded manifolds without boundary, with positive reach and the stated projection regularity. Proposition 4 constructs a local problem with strictly lower sample sensitivity for DPRGD-ATC on every such manifold of dimension at least one. The circle supplies the exact formula evaluated in the sensitivity experiment; Stiefel and Grassmann training uses polar and spectral projection. Each embedding provides explicit constants for the same two-point projection estimate, connecting the three experimental instances to the common analysis.

The network enters the common admissible step and the consensus error constants. For a base matrix W, choose the exchange count k so that $Q = W^k$ satisfies $\|Q - J\| < 1/2$. This sufficient spectral condition supports joint control of the projection domain and consensus error and the coefficient comparison. Theorem 3 then compares the two updates at the same communication cost, step size, and horizon, isolating the effect of averaging the locally updated models.

C Extended experiments

The extended study examines the parameter dependence of sample sensitivity while keeping the DPRGD and DPRGD-ATC update equations fixed. It supplements the theorem-condition experiments in Appendix B with larger steps, different communication counts, and a higher-dimensional instance.

C.1 Data and parameter selection

Samples remain coordinate vectors, so the population risk and test-distribution mean sensitivity are evaluated exactly. For dimension d, let $w = (8, 5, 2, 1, \dots, 1) \in \mathbb{R}^d$ and define $p_j = w_j^\alpha / \sum_r w_r^\alpha$. The development grid uses $\alpha \in \{1/2, 1, 2\}$. Common-distribution runs draw every node's samples from p. Heterogeneous runs draw at node $i \in \{0, 1, 2, 3\}$ from $p_i = (p + \mathrm{shift}_i(p))/2$, where shift_i circularly shifts the coordinates by i. Their population risk and test sensitivity use $\overline{p} = m^{-1} \sum_i p_i$. The heterogeneous statistics concern this mixture distribution, whereas our stated i.i.d. generalization theorem concerns a common sampling distribution.

The initialization is either the first r columns of the normalized Sylvester Hadamard matrix or the orthonormal factor of a Gaussian $d \times r$ matrix generated with seed $1701 + d + r$. Sphere uses $r = 1$ and Grassmann uses the corresponding rank-r projector. Initializations are independent of the training data. Each unit samples its dataset, then independently replaces the first sample at node zero from that node's distribution. Both methods and both neighboring datasets use the same uniformly sampled index array; neutral replacements are retained.

The first development stage uses $\mathrm{St}(8, 2)$, $m = 4$, $n = 40$, the original distribution, and Hadamard initialization. It tests $\eta \in \{3.6764705882 \times 10^{-4}, 0.01, 0.03, 0.1, 0.3, 1\}$ with $k \in \{1, 2, 4\}$ ring communication rounds. Runs extend to 2048 iterations, with candidate outputs at 128, 512, and 2048. The three best distinct step/communication pairs are then crossed with the three spectra, two initializations, and two distribution types. Four seeds, 81000–81003, are used throughout development; repeated configurations are counted once. No algorithm state or correction term is added.

Selection first requires both methods to decrease their mean population and empirical risks by at least 10^{-4} from initialization. ATC's mean risk may exceed the baseline by at most $\max\{0.05F_{\mathrm{B}}, 10^{-3}\}$, applied separately to population and empirical risk. Among eligible configurations with baseline mean sensitivity above 10^{-10}, selection maximizes the relative reduction in mean sensitivity. Con-

Table 8: Held-out means for shared larger-step configurations. Population risk uses p for i.i.d. data and the equal-node mixture for heterogeneous data. Each setting uses 40 paired units.

Distribution	Method	Sensitivity	Population	Empirical	Absolute gap
i.i.d.	DPRGD	6.30×10^{-3}	0.35338	0.34666	0.02659
i.i.d.	DPRGD-ATC	$< 10^{-14}$	0.35000	0.34328	0.02672
Heterogeneous	DPRGD	6.37×10^{-2}	0.54839	0.54239	0.02434
Heterogeneous	DPRGD-ATC	8.59×10^{-3}	0.51781	0.52375	0.02875

figurations and output iterations are frozen before validation. An independently tuned DPRGD configuration minimizes its development sensitivity subject to the learning condition on the same distribution and initialization; numerical-zero baselines remain in that comparison.

The shared i.i.d. configuration selects $\eta = 1, k = 1, T = 128$; the heterogeneous configuration selects $\eta = 1, k = 1, T = 2048$. Both retain $\alpha = 1$ and Hadamard initialization. Their one-round mixing has $\sigma = 1/2$, outside the strict mixing condition of the main comparison theorem; the selected step also exceeds its sufficient threshold. Validation uses 40 independent units with seeds 82000–82039 and 83000–83039, respectively; their index generators use the data seed plus 10^6. The independently tuned DPRGD settings are $(\eta, k, T) = (0.3, 1, 512)$ for i.i.d. data and $(1, 2, 128)$ for heterogeneous data. Those comparisons report their different iteration and communication budgets explicitly. No validation outcome is used to choose a new configuration.

All runs use double precision. The parameterized implementation reproduces all four trajectories in one saved pair on each of the three original geometries to below 10^{-13}. Larger-step experiments record consensus error, projection distances, projection margins, risk, and gradient norms in addition to the paired statistical measurements. There are 434 distinct paired configurations across development, validation, and transfer, each containing four trajectories, with no failed runs. Their larger-step statistics are empirical measurements rather than evaluations of a theorem outside its hypotheses.

The primary numerical environment is Linux with NumPy 1.26.4 and SciPy 1.11.4. Replaying all 434 records in that environment reproduces the saved results. Long, large-step heterogeneous DPRGD trajectories can differ across BLAS/LAPACK implementations. An independent Windows replay of all 40 heterogeneous pairs gives a mean sensitivity reduction of 87.1%, consistent with the primary aggregate, although individual trajectories differ. The reported tables use the fixed primary environment throughout.

C.2 Held-out sensitivity and generalization

For each independent unit, let $s_\nu = \sum_j \overline{p}_j |\ell(\widehat{x}_\nu(S), e_j) - \ell(\widehat{x}_\nu(S'), e_j)|$. Reported sensitivities average s_ν over units. We distinguish the mean signed gap $\mathbb{E}[F - F_S]$ from the mean absolute gap $\mathbb{E}|F - F_S|$. Intervals use 5000 paired bootstrap resamples of complete units, with seed 91271. Reduction is $1 - \overline{s}_{\mathrm{A}}/\overline{s}_{\mathrm{B}}$, rather than an average of per-unit ratios. Percentage intervals are omitted when a bootstrap denominator is at most 10^{-10}; absolute differences remain reported.

For i.i.d. data, the paired sensitivity difference (ATC minus DPRGD) is -0.006297, with interval $[-0.015845, -0.001013]$. All 40 ATC sensitivities are below 10^{-14}; eight replacements are neutral. For heterogeneous data, the mean difference is -0.055095, with interval $[-0.083391, -0.029516]$; fourteen replacements are neutral. The corresponding mean sensitivity reduction is 86.5%, with interval $[56.7\%, 100\%]$. ATC's sensitivities are below 10^{-14} in 38 of these 40 units, while two replacements change the selected subspace. Thus the heterogeneous average includes occasional finite changes, rather than a uniform per-unit reduction.

These categorical PCA instances admit coordinate subspaces on which every individual sample gradient vanishes. At a common such subspace, both updates remain fixed. Reaching the same subspace for neighboring datasets makes the measured loss change numerically zero; it does not imply zero uniform stability over all admissible datasets. The independently selected i.i.d. DPRGD setting reaches the same regime, with mean sensitivity 5.77×10^{-17} and population risk 0.35000 after 512 iterations. The shared ATC configuration reaches it after 128 iterations, using the same one-

Table 9: Transfer without retuning: ten paired units per geometry, with $\eta = 1, k = 1, T = 128$.

Geometry	Method	Sensitivity	Population risk	Absolute gap
$\mathbb{S}^7$	DPRGD	2.36×10^{-2}	0.62622	0.01842
$\mathbb{S}^7$	DPRGD-ATC	4.28×10^{-10}	0.61500	0.02375
$\mathrm{Gr}(8, 2)$	DPRGD	$< 10^{-14}$	0.35000	0.03000
$\mathrm{Gr}(8, 2)$	DPRGD-ATC	$< 10^{-14}$	0.35000	0.03000
$\mathrm{St}(64, 4)$	DPRGD	1.10×10^{-4}	0.80629	0.01970
$\mathrm{St}(64, 4)$	DPRGD-ATC	3.73×10^{-5}	0.81056	0.02332

round communication pattern. For heterogeneous data, independently tuned DPRGD has sensitivity 0.017433 with 128 iterations and two rounds per step, compared with ATC's 0.008594 at 2048 one-round steps; the paired sensitivity-difference interval $[-0.03277, 0.01615]$ includes zero. This comparison has different optimization budgets.

The mean absolute generalization gap is 0.02659 versus 0.02672 on i.i.d. data and 0.02434 versus 0.02875 on heterogeneous data (DPRGD versus ATC). Paired differences are 0.000133 with interval $[0.000013, 0.000252]$ and 0.004409 with interval $[-0.001491, 0.010566]$. Mean signed gaps are 0.006723 versus 0.006719 and 0.005994 versus -0.005938, respectively. These observations separate reduced sample sensitivity from realized generalization-gap behavior.

C.3 Dimension and geometry transfer

The selected i.i.d. setting, $\eta = 1, k = 1, T = 128$, is transferred without retuning to sphere, Grassmann, and $\mathrm{St}(64, 4)$ using seeds 84000–84009. The higher-dimensional probabilities use the same weights $(8, 5, 2, 1, \ldots, 1)$ normalized in dimension 64, and the initialization uses four Hadamard columns. The projection-derivative constants in (34) are independent of dimension and rank in the stated Frobenius norm; the larger step remains outside the theorem's sufficient threshold. The sphere and projector embeddings retain their own induced-metric gradients, so equal step sizes need not give equal subspace dynamics. On $\mathrm{St}(64, 4)$, sensitivity decreases by 66.2% with interval $[36.1\%, 86.8\%]$, while population risk changes from 0.80629 to 0.81056. Both methods decrease population and empirical risk from initialization and satisfy the predefined risk tolerance. On the sphere, the absolute paired sensitivity difference is reported in the supplementary statistics; its percentage interval is omitted because a bootstrap denominator reaches the numerical threshold. Both Grassmann sensitivities are at numerical precision. These outcomes show how the same update order interacts with dimension, metric, and the finite sample distribution.

C.4 Communication and an optimization baseline

The network study reuses the original 20 Stiefel paired datasets and sampling arrays, retaining their step and 512 iterations. Increasing the lazy-ring communication count from $k = 2$ to 3 and 4 changes σ from $1/4$ to $1/8$ and $1/16$. Figure 4 plots the mean across runs of each trajectory's maximum consensus error and maximum distance before projection. The deterministic ATC envelope retains its factor σ; the measured ratios need not equal that bound coefficient. The added runs satisfy the same main-theorem conditions. Their paired sensitivity differences remain of order 10^{-9} with intervals containing zero.

We also compare full-gradient DPRGD and DPRGD-ATC with DRGTA (Chen et al., 2021) on $\mathrm{St}(8, 2)$. The DRGTA implementation retains its tangent consensus step, retraction, and recursive gradient tracker; its consensus step is one. A four-step candidate grid $\{0.01, 0.03, 0.1, 0.3\}$ uses four development seeds and selects each method's step by mean final empirical risk. All select 0.3. Ten fresh datasets, seeds 85000–85009, are evaluated for 2048 updates with $Q = W^2$. An independent four-process execution of the upstream DRGTA class agrees with the batched recurrence over 16 updates to 6.7×10^{-16}, after normalizing the local summed objective to the paper's mean loss.

All three methods reach comparable final objective and gradient values on these data. DRGTA evaluates an initial full gradient and communicates both its model and tracker; its count is $n(T + 1)$ sample gradients per node and twice the model-only payload at common k, T. The comparison

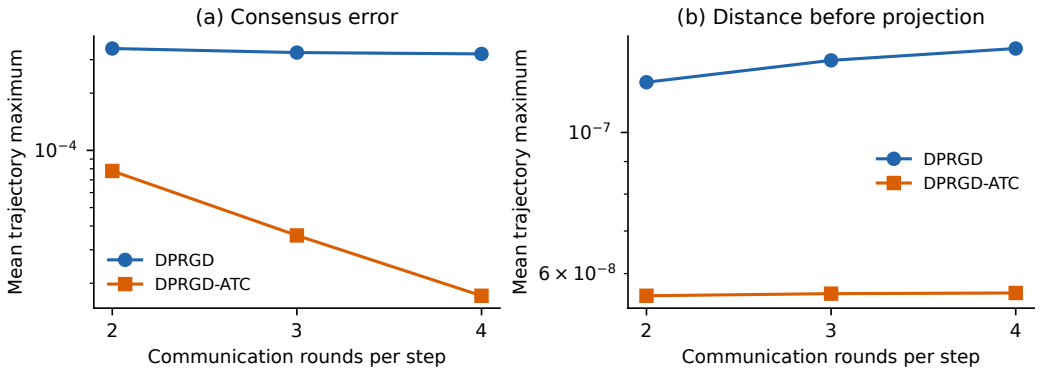


Figure 4: Changing communication at fixed steps and data. Each point averages the maximum over a trajectory, across 20 original-data runs. The horizontal axis counts neighbor exchanges per gradient step; increasing it also increases communication cost.

Table 10: Full-gradient optimization, averaged over ten held-out datasets. Communication counts algorithmic scalar payload over directed ring edges, excluding diagnostic aggregation. Sample-gradient counts are algorithmic counts per node, excluding diagnostic evaluations.

Method	Empirical risk	Squared gradient	Sample gradients	Scalars sent
DPRGD	0.34750	$< 10^{-28}$	81,920	524,288
DPRGD-ATC	0.34750	$< 10^{-28}$	81,920	524,288
DRGTA	0.34750	$< 10^{-28}$	81,960	1,048,576

describes this implementation and network, rather than assigning the same communication structure to other correction methods such as REXTRA.

D GENERALIZATION WITH A FIXED DPRGD CONFIGURATION

This experiment studies the mean absolute generalization gap while holding DPRGD at the original Stiefel configuration: $d = 8$, $r = 2$, $m = 4$, $n = 40$, $\eta_{\mathrm{B}} = 3.6764705882 \times 10^{-4}$, $k_{\mathrm{B}} = 2$, and $T_{\mathrm{B}} = 512$. The coordinate distribution and Hadamard initialization are unchanged. DPRGD-ATC varies its predetermined steps, communication count, and terminal iteration, retaining the same loss, gradient, projection, and projected-mean output.

D.1 STATISTICAL TARGET AND LEARNING PROGRESS

Let p and $\widehat{p}$ be the population and global empirical coordinate frequencies. For a Stiefel output X, put $v(X) = \mathrm{diag}(XX^{\top})$. The risks and signed gap satisfy the exact identities

$$F(X) = 1 - p^{\top} v(X), \qquad F_S(X) = 1 - \widehat{p}^{\top} v(X), \qquad F(X) - F_S(X) = (\widehat{p} - p)^{\top} v(X). \tag{35}$$

At the common Hadamard initialization, $v(X^0) = (r/d)\mathbf{1}$, so every dataset has zero generalization gap. Training changes these coordinate weights and creates the measured gap. To compare learning progress, an ATC candidate is eligible only when its mean population-risk decrease and mean empirical-risk decrease from initialization are each at least 95% of the corresponding fixed-DPRGD decrease. The same criterion is applied on fresh validation datasets.

The primary endpoint is the mean of $|F(\widehat{x}) - F_S(\widehat{x})|$ over independent dataset/replacement/index units. We also report the signed gap, replace-one sensitivity, both risks, squared empirical gradient, consensus error, and resource counts. The absolute gap differs from the magnitude of the expected signed gap controlled by the displayed stability-to-generalization argument. Parameter changes can affect optimization progress, sampling variability, and consensus simultaneously; the original common-parameter experiments isolate the update-order comparison.

D.2 Search and independent validation

The constant-step grid is

$$\eta \in \{10^{-6}, 10^{-5}, 10^{-4}, \eta_{\mathrm{B}}, 10^{-3}, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10\},$$

with $k \in \{1, 2, 4, 8, 16\}$ and candidate horizons $T \in \{32, 128, 512, 2048, 8192, 32768\}$. Each step/communication configuration runs a single trajectory to the largest horizon, saving its prescribed prefixes. Eight development seeds, 86000–86007, are shared across all candidates.

The six best distinct eligible initial-step/communication pairs are also tested with

$$\eta_t = \eta_0(1 + t/\kappa)^{-a}, \qquad a \in \{1/2, 1\}, \quad \kappa \in \{32, 512, 4096\},$$

and with a one-time multiplication of the step by 0.1 or 0.01 after 512 updates. All schedules are fixed before their runs. A drop schedule evaluated at $T \leq 512$ is identified with its equivalent constant-step candidate. The cumulative step and squared-step sum are recorded. For fixed cumulative step and horizon, a constant sequence minimizes the squared-step sum, so decay is assessed through its actual dynamics rather than presumed to improve this quantity.

Eligible candidates are ordered by mean absolute gap, then population risk, communication rounds, and iteration count. The top eight overall and up to eight satisfying $T \leq 512$ and $kT \leq 1024$ form the screening set, with duplicate configurations removed. Sixteen new seeds, 87000–87015, assess only these nominated terminal configurations. The main and resource-constrained selections are then frozen before evaluation on 100 new paired units, seeds 88000–88099. The latter selection uses no more stochastic gradient calls or model communication than the baseline. Different communication and iteration budgets of the main selection are reported explicitly.

Every method uses the same training data, independent replacement at node zero's first sample, and the common prefix of the sampled-index array. The index generator uses the data seed plus 10^6, with the saved integer arrays specifying the realized draws. Neutral replacements remain in the sample. Projection validity is checked, and a candidate with invalid paired units is not summarized conditionally on its successful units. Paired bootstrap intervals use 10000 resamples, seed 91272, and ascending data-seed order. The principal comparison is ATC minus DPRGD; a successful primary comparison requires a negative upper endpoint of its 95% interval and both learning conditions. A 20% relative decrease is a target effect size, not a selection or reporting assumption.

The main frozen configuration is transferred without retuning to sphere and Grassmann using 20 new seeds, 89000–89019, per geometry. Their DPRGD baselines retain the original parameters, including the doubled Grassmann step $7.3529411765 \times 10^{-4}$. These transfer comparisons assess the same fixed-baseline question under each embedding.

D.3 Results and interpretation

The search completes all 65 constant-step and 48 scheduled-step configurations, yielding 642 distinct terminal candidates after removing equivalent prefixes. Of these, 410 meet the development learning criterion, and 14 distinct configurations enter independent screening. All 136 batch tasks, including fixed baselines and final comparisons, complete without invalid paired units. Figure 5 shows the development measurements.

The frozen main configuration uses

$$\eta_t = \frac{3.6764705882 \times 10^{-4}}{\sqrt{1 + t/4096}}, \qquad k = 16, \qquad T = 512.$$

The resource-constrained selection uses the same schedule and horizon with $k = 2$. Table 11 reports the independent validation results. The main candidate decreases the mean absolute gap by 3.0738%, with 95% interval $[3.0496\%, 3.0980\%]$. Its paired absolute difference is -5.4684×10^{-5}, with interval $[-6.3793, -4.6216] \times 10^{-5}$. The resource-constrained candidate gives 3.0737% reduction, with interval $[3.0495\%, 3.0979\%]$, and paired difference -5.4683×10^{-5}, with interval $[-6.3791, -4.6214] \times 10^{-5}$. Both satisfy the predefined learning criterion; neither reaches the 20% target.

The baseline population and empirical risks are 0.726992 and 0.726120. The corresponding main-ATC risks are 0.727700 and 0.726854. Its cumulative step decreases from 0.188235 to 0.182705,

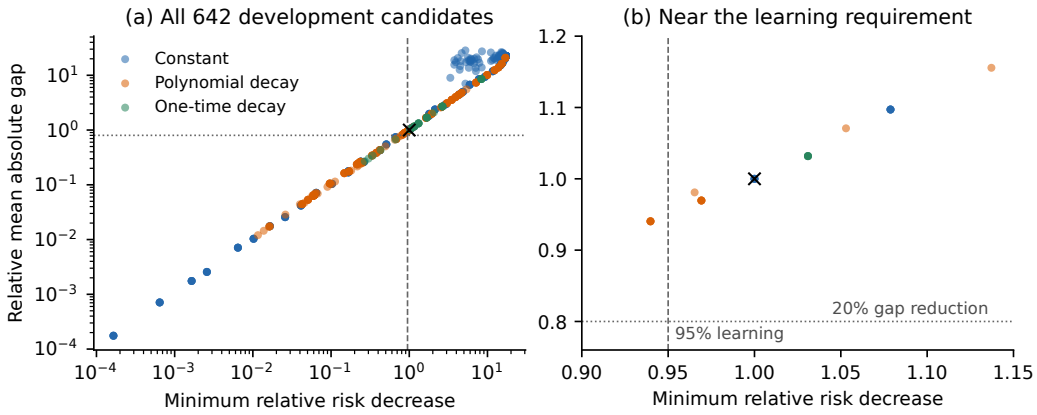


Figure 5: All 642 development candidates and a closer view near the learning requirement. The horizontal coordinate is the smaller of the population- and empirical-risk decrease ratios relative to fixed DPRGD; the vertical coordinate is the mean absolute-gap ratio. The dashed line gives the 95% learning threshold, the dotted line gives the target 20% gap reduction, and the cross marks the baseline. Both axes in the left panel are logarithmic.

Table 11: Independent validation on 100 paired Stiefel datasets. All runs have $T = 512$. The two progress columns are population- and empirical-risk decreases divided by the corresponding fixed-DPRGD decrease.

Method	k	10^3 absolute gap	Population	Empirical
DPRGD	2	1.77908	1.00000	1.00000
ATC, main	16	1.72439	0.96922	0.96923
ATC, resource	2	1.72440	0.96922	0.96923

and its squared-step sum decreases from 6.9204×10^{-5} to 6.5216×10^{-5}. Population and empirical learning progress retain 96.9220% and 96.9232% of baseline. Thus the approximately 3.07% gap reduction accompanies approximately 3.08% less learning progress. This is consistent with the gap–progress dependence in (35); the experiment does not isolate a comparable improvement at identical learning progress.

Each method uses 512 stochastic gradient calls per node. DPRGD and resource-constrained ATC use 1024 neighbor exchanges and 131072 directed-edge communication scalars; main ATC uses 8192 exchanges and 1048576 scalars. Increasing communication eightfold changes the mean absolute gap by only 1.42×10^{-9} between the two selected ATC configurations. The resource-constrained configuration therefore attains essentially the same observed gap at the original resource budget.

On sphere, direct transfer decreases the mean absolute gap from 0.00130045 to 0.00125860, a 3.218% reduction with interval $[3.169\%, 3.261\%]$, while retaining 96.866% and 96.860% of the population and empirical learning progress. On Grassmann, it decreases the gap from 0.00229286 to 0.00107763, but retains only 47.231% and 47.219% of the corresponding learning progress. The Grassmann transfer fails the learning criterion. Its original baseline uses twice the Stiefel step under the projector metric, whereas transfer preserves the selected ATC schedule unchanged.

E Real image data and larger networks

E.1 Data, loss, and evaluation distribution

We use Fashion-MNIST (Xiao et al., 2017), retaining all 784 pixels and learning an eight-dimensional subspace on $\mathrm{St}(784, 8)$. A fixed permutation of its 60000 training images, generated with seed 93000, assigns 10000 images to development and 50000 to the final sampling pool. The official 10000 test images are separate. Pixels are divided by 255, centered using the development

mean, and normalized to unit Euclidean norm; a zero centered image remains zero. Labels are not used. Preprocessing is frozen before the final experiments.

For a transformed image z, the loss is $\ell(X, z) = 1 - \|X^\top z\|^2$. Its ambient gradient is $V = -2z(z^\top X)$ and its Riemannian gradient is $V - X\,\mathrm{sym}(X^\top V)$, where $\mathrm{sym}(A) = (A + A^\top)/2$. Both algorithms retain their stated updates, polar nearest-point projection, and projected-mean output. Their common initialization is the thin orthonormal factor of a 784×8 standard Gaussian matrix generated with seed 93001.

Let C_{ref} be the average of $zz^\top$ over the final sampling pool. Conditional on the frozen preprocessing and this pool, define the uniform finite reference distribution and its exact risk

$$F_{\mathrm{ref}}(X) = 1 - \mathrm{tr}(X^\top C_{\mathrm{ref}} X).$$

Each training dataset consists of $N = 4096$ independent draws with replacement from this distribution. The primary statistical endpoint is the mean of $|F_{\mathrm{ref}}(\widehat{X}) - F_S(\widehat{X})|$ over paired units. We separately report the empirical risk F_{test} on all official test images and its difference from F_S. Thus F_{ref} is an exactly evaluated finite-distribution risk, whereas F_{test} is a held-out test average. The displayed expected-generalization theorem and this mean absolute-gap endpoint remain distinct statistical statements.

For neighboring datasets, position zero at node zero is redrawn independently from the same sampling pool. Sample sensitivity is the mean over all official test images of $|\ell(\widehat{X}(S), z) - \ell(\widehat{X}(S'), z)|$. Neutral replacements remain included. The test set is not used for parameter selection.

E.2 Network, selection, and paired protocol

The main configurations use $m \in \{4, 8, 16, 32\}$ nodes with $n = N/m$ samples each, $T = 2048$ updates, and $k = 2$ rounds on a lazy ring. Each node places weight $1/2$ on itself and $1/4$ on each neighbor. For a given seed, the same global sample list is reshaped across node counts. A fifth configuration uses $m = 32$ and the smallest k with $\|W^k - J\|_2 \le 1/4$, namely $k = 144$. Its contraction factor is approximately 0.2490, compared with 0.9809 for two rounds on the same graph.

Four development seeds, 94000–94003, evaluate constant steps $\{0.001, 0.01, 0.03, 0.1, 0.3\}$ on $m = 4$ and $m = 32$ with $k = 2$. Selection minimizes terminal empirical risk averaged over the two algorithms, both node counts, and all four seeds. Eligible steps require valid projections and mean empirical descent for each algorithm. Ties favor the smaller step. This selects the common step $\eta = 0.1$, which is frozen for all five final settings. Twenty new seeds, 95000–95019, provide paired validation; no final result changes the selection.

Within each unit, both algorithms and neighboring datasets share the same sampled local indices. They are generated from a 2048×32 uniform array using seed equal to the data seed plus 10^6, taking the first m columns and mapping each entry to one of n local indices. Outputs and node states are saved at iterations $0, 128, 512, 1024, 2048$, with additional diagnostics every 128 updates. Paired percentile intervals use 10000 resamples, seed 96000, and ascending data-seed order. Each interval concerns its stated setting and endpoint.

The primary environment uses NumPy 1.26.4 and SciPy 1.11.4 on Linux with double precision. Each of at most four workers uses one numerical-library thread. Nodes are simulated, and multiplying by W^k evaluates the prescribed repeated mixing. Communication accounting includes all k rounds: each algorithm uses mT stochastic gradients in total, T per node, and $2mkTdr$ directed-edge scalars, excluding diagnostic evaluations and final output aggregation. At fixed T, larger networks evaluate more gradients globally; these are network-size comparisons at a common local iteration budget.

The selected step exceeds the sufficient trajectory/stability thresholds of the analysis. The real-data results therefore describe empirical behavior under the original update rules. We evaluate geometric diagnostics and statistical quantities directly, without reporting an inapplicable stability bound.

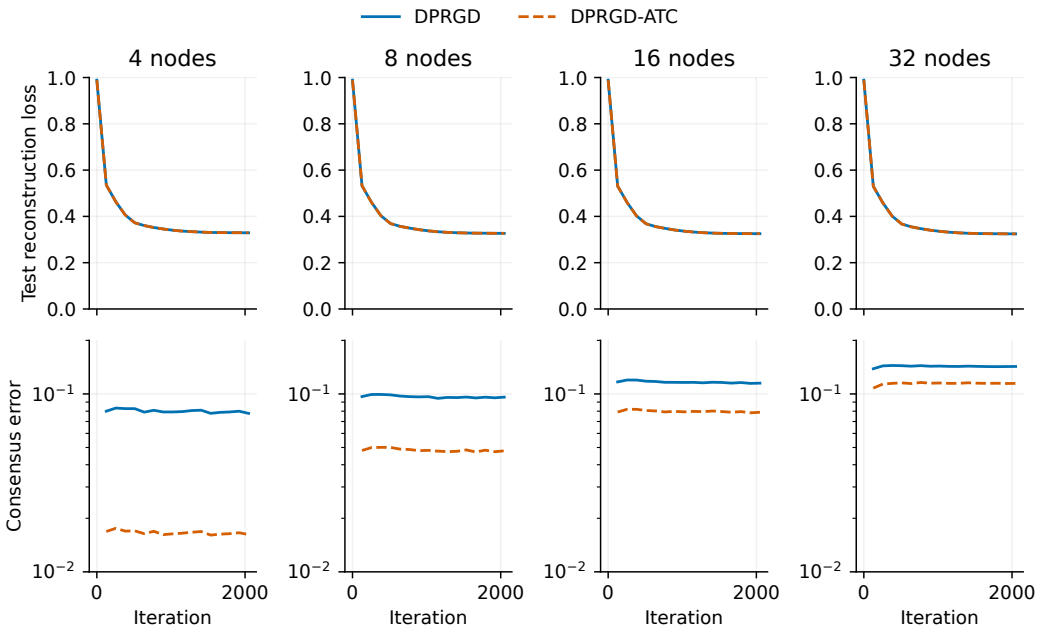


Figure 6: Fashion-MNIST training with $k = 2$. Curves average 20 paired datasets. The top row gives official-test reconstruction loss; the bottom row gives consensus error, using the same logarithmic scale across network sizes. Both algorithms use $\eta = 0.1$.

E.3 Results and resource comparison

All 40 development and 100 final paired configurations complete, giving 560 original and neighboring trajectories. Every saved projection is valid; the implementation checks projection margins at each update. The four network-size experiments reuse the same 20 global validation datasets, so results across network sizes are paired as well as the two algorithms within each setting.

Figure 6 shows that the two learning curves nearly coincide while their consensus errors separate. For $m = 4, 8, 16, 32$ and $k = 2$, mean terminal ATC consensus errors are respectively 0.208, 0.499, 0.686, and 0.803 times the DPRGD values. Across the five settings, ATC retains 99.9983%–100.0015% of the baseline reference-risk decrease. The measured geometric reductions therefore coexist with comparable learning progress.

Table 12 and Figure 7 show the paired statistical comparisons. Every absolute reference-gap interval contains zero; all official-test absolute-gap difference intervals also contain zero. Mean test sensitivities are approximately 0.9–1.2×10^{-4}. Their paired differences are below 10^{-6} in magnitude. The eight-node sensitivity interval excludes zero with a reduction of about 0.80%; the remaining sensitivity intervals include zero. These per-setting intervals describe small distribution-averaged differences alongside the larger consensus and projection effects.

At 32 nodes, increasing k from 2 to 144 changes ATC's terminal consensus error from 0.11460 to 0.00577, while its mean absolute reference gap changes from 0.00205181 to 0.00205181 at the displayed precision. Compared with DPRGD at the same $k = 144$, ATC reduces terminal consensus error by 93.5% and maximum projection distance by 92.3%. Tables 13 and 14 quantify the communication and geometric effects separately.

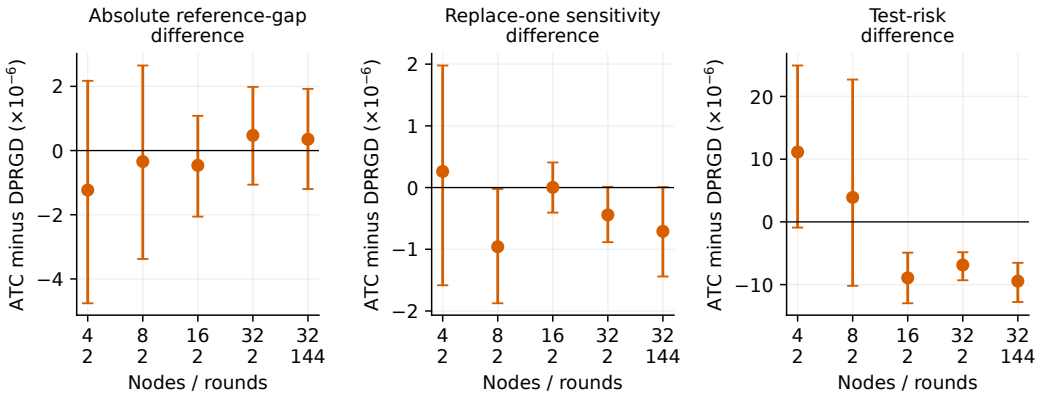


Figure 7: Paired differences on Fashion-MNIST with 95% bootstrap intervals. Negative values favor DPRGD-ATC for the indicated quantity. Tick labels give nodes (top) and rounds per update (bottom). The final setting uses stronger mixing on the same 32-node ring.

Table 12: Paired differences, DPRGD-ATC minus DPRGD, on Fashion-MNIST. All entries are in units of 10^{-6}; brackets give 95% paired bootstrap intervals. Sensitivity averages absolute replace-one loss changes over the official test images.

m	k	Absolute reference gap	Sensitivity	Test loss
4	2	-1.232 $[-4.758, 2.173]$	0.260 $[-1.583, 1.978]$	11.144 $[-0.911, 24.934]$
8	2	-0.343 $[-3.378, 2.650]$	-0.959 $[-1.875, -0.022]$	3.905 $[-10.195, 22.692]$
16	2	-0.461 $[-2.059, 1.082]$	0.003 $[-0.406, 0.407]$	-8.924 $[-12.976, -4.909]$
32	2	0.473 $[-1.063, 1.984]$	-0.443 $[-0.885, 0.011]$	-6.872 $[-9.308, -4.840]$
32	144	0.351 $[-1.199, 1.921]$	-0.709 $[-1.441, 0.009]$	-9.446 $[-12.776, -6.526]$

Table 13: Fashion-MNIST network and resource counts for either algorithm. Gradients are total stochastic calls across nodes; each node evaluates 2048. Communication scalars sum over directed edges and all rounds, in billions. Diagnostic and final output aggregation costs are excluded.

m	k	σ	Gradients	Rounds	Scalars (10^9)
4	2	0.250000	8,192	4,096	0.206
8	2	0.728553	16,384	4,096	0.411
16	2	0.925328	32,768	4,096	0.822
32	2	0.980878	65,536	4,096	1.644
32	144	0.249038	65,536	294,912	118.380

Table 14: Geometric diagnostics on Fashion-MNIST. Consensus error is measured at $T = 2048$; distance is each trajectory's maximum distance before projection, averaged over the 20 original datasets.

		Consensus error		Projection distance	
m	k	DPRGD	DPRGD-ATC	DPRGD	DPRGD-ATC
4	2	0.07778	0.01619	0.00684	0.00244
8	2	0.09595	0.04790	0.00684	0.00254
16	2	0.11500	0.07889	0.00693	0.00264
32	2	0.14269	0.11460	0.00709	0.00258
32	144	0.08844	0.00577	0.00907	0.00070

F FIXED-BASELINE TUNING ON REAL IMAGE DATA

F.1 PROTOCOL AND SELECTION

We retain the Fashion-MNIST preprocessing, $\mathrm{St}(784, 8)$ loss, $N = 4096$ samples, initialization, and projected-mean output from Appendix E. DPRGD is fixed at $\eta = 0.1$, $k = 2$, and $T = 2048$. Only DPRGD-ATC's step sizes, communication rounds and horizon vary. This comparison evaluates tuning against the fixed baseline, complementing the common-parameter comparison in Table 3.

On the development images, we test constant steps

$$\eta \in \{10^{-4}, 10^{-3}, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10\}.$$

Communication uses $k \in \{1, 2, 4, 8, 16\}$ rounds, with horizons $T \in \{512, 1024, 2048, 4096, 8192\}$. Each step/communication pair shares one trajectory across horizons. Four development seeds, 97000–97003, are used at $m = 4$ and 32. The six best distinct initial-step/communication pairs additionally test $\eta_t = \eta_0(1 + t/\kappa)^{-p}$ with $p \in \{1/2, 1\}$ and $\kappa \in \{128, 1024, 4096\}$, and a drop by 0.1 or 0.01 after 1024 updates.

At each development node count, a candidate must attain at least 99% of DPRGD's average reference-risk decrease and at least 99% of its average empirical-risk decrease. Eligible candidates are ranked by mean absolute reference gap, with equal weight on the two node counts. Ties favor lower reference risk, fewer communication rounds, and fewer gradient calls. The top eight candidates and the top four satisfying $T \le 2048$ and $kT \le 4096$ are screened on development seeds 98000–98015. This freezes one main configuration and one resource-restricted configuration. The official test set is not used for selection.

The final comparison uses seeds 99000–99049, each evaluated at $m \in \{4, 8, 16, 32\}$. A seed determines a shared global dataset across node counts and common sampling prefixes across methods. Each complete seed, including all four node counts, is one bootstrap unit. The primary statistic is

$$G_\nu = \frac{1}{4R} \sum_{m \in \{4,8,16,32\}} \sum_{s=1}^{R} \left| F_{\mathrm{ref}}(\widehat{X}_{\nu,m,s}) - F_{S_s}(\widehat{X}_{\nu,m,s}) \right|, \qquad R = 50.$$

We compute 10000 paired bootstrap resamples with seed 100000. The primary success criterion requires a strictly negative 95% interval for $G_{\mathrm{ATC}} - G_{\mathrm{DPRGD}}$ and both 99% descent conditions at every node count. Node-specific intervals describe the individual settings; they are not simultaneous intervals. Test risk, signed gap, sensitivity, squared gradient and geometric diagnostics are evaluated separately.

F.2 SELECTED CONFIGURATIONS AND RESULTS

The main configuration uses constant step $\eta = 0.1$, $k = 1$, and $T = 2048$. The main and resource-restricted selections coincide; one set of validation runs evaluates both roles.

The constant-step development leader used $\eta = 0.03$ and $T = 4096$. Its approximately 7.04% improvement on the four initial seeds did not retain its ranking on the sixteen screening seeds. The frozen choice follows the screening results.

The recorded search contains 792 completed development trajectories and 416 completed screening trajectories, each paired with its neighboring dataset. The final phase contains 400 completed original/neighboring pairs. The counts of numerical failures and timed-out jobs are 0 and 0, respectively.

Main tuned configuration: the pooled absolute gap changes from 0.0023402 to 0.0023395, a relative reduction of 0.029%. The paired difference is -0.679×10^{-6} with interval $[-1.324, -0.079]\times10^{-6}$. The smallest descent ratio across both risks and all node counts is approximately 100.001%. The pooled interval excludes zero, and both learning-progress conditions hold at each node count.

Mean replace-one sensitivity decreases by 0.592%, with difference interval $[-1.328, -0.167] \times 10^{-6}$. The signed reference-gap interval is $[-0.672, 0.624]\times10^{-6}$, and the absolute test-gap interval is $[-0.765, 0.642] \times 10^{-6}$; both contain zero. The positive primary result concerns the absolute finite-reference gap.

Table 15: Main tuned configuration on Fashion-MNIST, using 50 paired datasets at each node count. Gaps and their paired differences are scaled by 10^3. Brackets give 95% intervals. Progress is the smaller of the reference-risk and empirical-risk descent ratios relative to DPRGD. The pooled row weights node counts equally.

m	G_{B}	G_{A}	ΔG	Interval	Progress (%)
4	2.3283	2.3258	-0.0025	$[-0.0045, -0.0004]$	100.001
8	2.3327	2.3331	0.0003	$[-0.0010, 0.0016]$	100.001
16	2.3393	2.3386	-0.0007	$[-0.0016, 0.0002]$	100.001
32	2.3604	2.3605	0.0001	$[-0.0008, 0.0010]$	100.001
All	2.3402	2.3395	-0.0007	$[-0.0013, -0.0001]$	—

Table 16: Mean risks and replace-one sensitivity for the fixed-baseline Fashion-MNIST comparison. Sensitivity is scaled by 10^4 and averaged over the official test images. B is DPRGD and M is the selected ATC configuration, which also satisfies the resource restriction.

Method / m	Reference risk	Training risk	Test risk	10^4 sensitivity	$\|\operatorname{grad} F_S\|^2$
B / 4	0.32745	0.32550	0.32915	1.511	0.00327
M / 4	0.32744	0.32549	0.32915	1.495	0.00327
B / 8	0.32489	0.32295	0.32663	1.401	0.00165
M / 8	0.32488	0.32295	0.32663	1.393	0.00165
B / 16	0.32356	0.32162	0.32531	1.074	0.00083
M / 16	0.32356	0.32162	0.32530	1.072	0.00083
B / 32	0.32292	0.32097	0.32467	0.889	0.00040
M / 32	0.32292	0.32096	0.32466	0.886	0.00040

F.3 Costs, diagnostics, and interpretation

For a configuration with horizon T and k ring rounds, each node uses T sample gradients, the network uses mT sample gradients, and communication uses kT rounds and $2mkT \cdot 784 \cdot 8$ directed scalars. Thus the main configuration's gradient and communication multipliers relative to DPRGD are 1 and 0.5, respectively. Final global output aggregation and diagnostic evaluations are excluded from these training counts. Nodes are simulated on one server.

The search varies step-size schedules and horizons, while the final selection keeps both at their baseline values and halves the communication rounds. This comparison measures the joint practical effect of ATC's update order and its communication choice. The matched-parameter experiments retain their role in isolating projection order. Because the selected step 0.1 lies outside the sufficient conditions of the stability theorem, this subsection reports observed statistical, optimization, and geometric quantities without assigning that guarantee to the real-data runs.

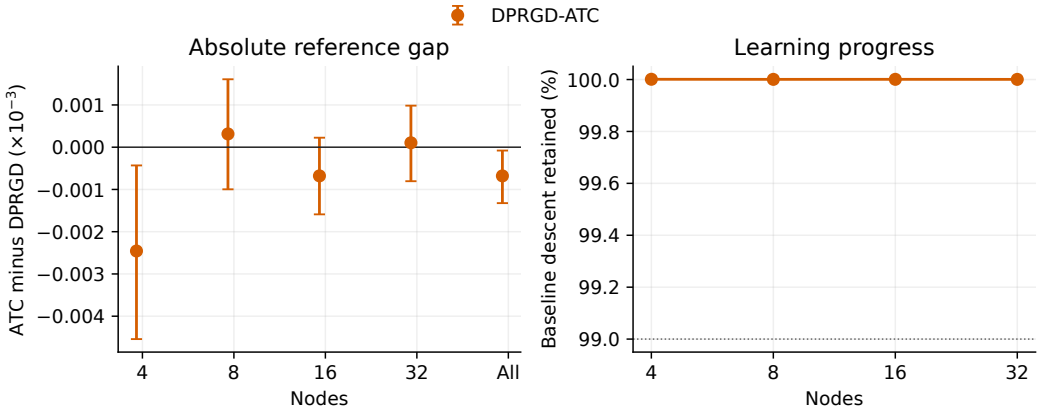


Figure 8: Fixed-baseline Fashion-MNIST tuning. Left: absolute reference-gap differences, with 95% paired intervals; negative values favor ATC. Right: the smaller of the two mean learning-progress ratios, with the 99% selection threshold. Parameters are fixed across node counts. The pooled statistical interval resamples complete data seeds.

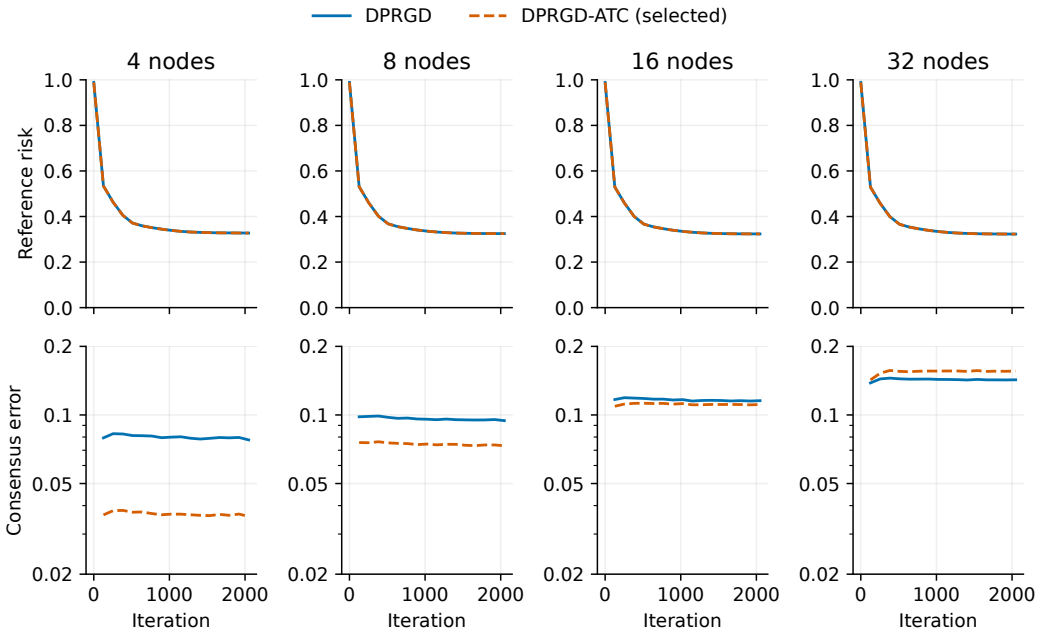


Figure 9: Mean reference-risk and consensus trajectories over the 50 final datasets. Each curve stops at its configuration's selected horizon. The horizontal axis counts local updates; communication costs additionally depend on the selected number of rounds per update.