WEAKLY CONVEX OPTIMIZATION ON MANIFOLDS WITH HEAVY-TAILED ORACLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly convex optimization on manifolds underlies nonsmooth learning problems with orthogonality constraints. Existing Riemannian stochastic subgradient guarantees rely on bounded second moments; we study unbiased oracles with only a finite p-th central moment, $1 < p \leq 2$. Our main result is high-probability convergence of Riemannian stochastic subgradient descent (RSSG) on compact embedded manifolds with globally defined smooth retractions. Heavy-tailed steps can leave the region where local retraction bounds apply. We overcome this mismatch by controlling the iterates' actual movement on the compact manifold, without clipping or assuming finite oracle variance. Its oracle complexity is $\widetilde{O}(\varepsilon^{-2p/(p-1)})$, recovering the classical ε^{-4} order at $p = 2$. Two complementary variants bound every tangent step. GeoClip-RSSG couples clipping and stepsize selection to attain the same oracle exponent within a prescribed retraction neighborhood. GeoNorm-RSSG normalizes mini-batch subgradients, attaining the same iteration-rate order with parameter choices independent of p, σ, at an additional sample cost. Experiments support the predicted convergence trends and demonstrate the methods' value for subspace recovery from heavy-tailed data.

1 INTRODUCTION

Weakly convex optimization on manifolds models robust subspace recovery and orthogonal dictionary learning (Li et al., 2021). It combines geometric constraints with convex nonsmooth losses or nonconvex losses with Lipschitz gradients (Davis & Drusvyatskiy, 2019; Boumal, 2023). We consider

$$\min_{x \in \mathcal{M}} f(x), \qquad f(x) = \mathbb{E}_\zeta[f(x, \zeta)], \tag{1}$$

where f is weakly convex in the ambient Euclidean space and $\mathcal{M}$ is a compact smooth embedded manifold. Stiefel constraints and dual principal component pursuit (DPCP) provide our main geometric and application examples.

For $\mathcal{M} = \mathbb{R}^d$, stochastic subgradient and model-based methods achieve ε-accuracy in Moreau stationarity with $O(\varepsilon^{-4})$ oracle calls under bounded second moments (Davis & Drusvyatskiy, 2019). For manifold constraints, Li et al. (2021) establish the same order in expectation using constrained proximal stationarity and square-integrable sample Lipschitz constants, with extensions from Stiefel to compact embedded manifolds. Extensions cover proximally smooth sets and randomized submanifold updates (Davis et al., 2025; Cheung et al., 2025). We ask how this theory changes when an unbiased tangent oracle satisfies only

$$\mathbb{E}_t[\widetilde{v}_t] = v_t \in \partial_{\mathcal{M}} f(x_t), \qquad \mathbb{E}_t \left\|\widetilde{v}_t - v_t\right\|^p \leq \sigma^p, \quad 1 < p \leq 2, \tag{2}$$

where $\mathbb{E}_t$ conditions on the history. This allows infinite variance. Euclidean weakly convex optimization already has finite-moment results: unclipped stochastic subgradients converge in expectation on bounded convex domains, and clipping gives high-probability guarantees (Zhu et al., 2026, Theorems 4.2 and 5.1). Can the Riemannian update attain a high-probability guarantee under the same finite-moment condition?

To see the difficulty, let $\widehat{x}$ minimize $f(y) + (\bar{\rho}/2)\left\|y - x\right\|^2$ over $\mathcal{M}$. A retraction $\mathcal{R}_x$ returns a tangent step ξ to the manifold. Unlike convex projection, it can increase the distance used in the

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

Table 1: High-probability bounds for averaged squared proximal residual on compact embedded manifolds. Iterations T and total oracle calls Q are sufficient for residual at most ε^2. $\widetilde{O}$ suppresses logarithmic confidence and problem constants. The second column indicates whether the parameter choices use p, σ. GeoNorm uses $B = T$. Global and local refer to the domain on which the given retraction is used.

Method	Uses p, σ	Retraction	T	Q
RSSG	Yes	Global	$\widetilde{O}(\varepsilon^{-2p/(p-1)})$	$\widetilde{O}(\varepsilon^{-2p/(p-1)})$
GeoClip	Yes	Local	$\widetilde{O}(\varepsilon^{-2p/(p-1)})$	$\widetilde{O}(\varepsilon^{-2p/(p-1)})$
RSSG	No	Global	$\widetilde{O}(\varepsilon^{-4/(p-1)})$	$\widetilde{O}(\varepsilon^{-4/(p-1)})$
GeoClip	No	Local	$\widetilde{O}(\varepsilon^{-4/(p-1)})$	$\widetilde{O}(\varepsilon^{-4/(p-1)})$
GeoNorm	No	Local	$\widetilde{O}(\varepsilon^{-2p/(p-1)})$	$\widetilde{O}(\varepsilon^{-4p/(p-1)})$

proximal comparison:

$$\underbrace{\|\mathcal{R}_x(\xi) - \widehat{x}\|^2 - \|x + \xi - \widehat{x}\|^2}_{\text{retraction distance error}} \le \kappa_{\mathcal{R}} \|\xi\|^2, \qquad \|\xi\| \le r_{\mathcal{R}}. \tag{3}$$

QR retraction can have positive error at arbitrarily small steps (Proposition K.2). A finite p-th moment also permits steps outside any prescribed local radius. Thus the Euclidean finite-moment analysis requires a distance comparison valid for the retracted update (Li et al., 2021; Zhu et al., 2026).

Compactness controls the large-step part of this comparison. For a globally defined smooth retraction, local quadratic accuracy and bounded diameter imply

$$\|\mathcal{R}_x(\xi) - y\|^2 - \|x - y\|^2 - 2\langle \xi, x - y\rangle \le C_{\mathcal{R},p} \|\xi\|^p, \qquad y \in \mathcal{M},\ \xi \in \mathrm{T}_x\mathcal{M}. \tag{4}$$

Small steps use (3); for large steps the left-hand side grows at most linearly. This bound retains the unbiased linear term and makes the remainder integrable. For the proximal problem above, let $\Phi_{\bar\rho}(x)$ denote its optimal value and $\Theta_{\bar\rho}(x) = \bar\rho \|x - \widehat{x}\|$ its stationarity residual (Section 3). RSSG with stepsize η then satisfies the one-step descent bound

$$\mathbb{E}_t \Phi_{\bar\rho}(x_{t+1}) \le \Phi_{\bar\rho}(x_t) - c\eta \Theta_{\bar\rho}(x_t)^2 + C\eta^p. \tag{5}$$

Here $c > 0$ and $C \ge 0$ are independent of t, η. Summing (5) gives the expectation bound in Theorem F.2. Compactness and (4) also give bounded proximal-value increments with conditional variance $O(\eta^p)$. Concentrating these increments yields the high-probability guarantee in Theorem 5.2.

Our main result retains the RSSG update with a global retraction. Two complementary variants bound the tangent steps: GeoClip couples its threshold and stepsize through $\eta\Lambda \le r_{\mathcal{R}}$, while GeoNorm normalizes a mini-batch subgradient to separate step length from direction estimation. The known Stiefel polar comparison (Li et al., 2021; Chen et al., 2021) gives explicit constants; QR is also covered by the general analysis. Table 1 compares the methods on the same high-probability criterion, including schedules that do not use p or σ. For $p < 2$, these parameter-independent choices yield a better iteration bound for batched GeoNorm and a better oracle bound for single-sample methods.

Our contributions are as follows.

- We establish high-probability convergence of unclipped RSSG under finite p-th moments, with $\widetilde{O}(\varepsilon^{-2p/(p-1)})$ oracle complexity on compact manifolds. A global retraction estimate controls both proximal descent and the variance needed for concentration.
- We derive high-probability bounds for GeoClip and GeoNorm that couple finite-moment noise control with a prescribed retraction radius. Accounting for clipping bias and mini-batch directional error yields convergent bounded-step variants and quantifies the resulting iteration and oracle costs.
- Without knowledge of p, σ, our bounds distinguish GeoNorm's iteration advantage from the single-sample methods' oracle advantage for $p < 2$. Exact proximal-stationarity measurements and heavy-tailed DPCP experiments examine convergence, geometric step control, and fixed-budget recovery.

2 RELATED WORK

Weakly convex optimization on manifolds. Proximal stationarity supports stochastic subgradient analysis in Euclidean space (Davis & Drusvyatskiy, 2019) and on compact embedded manifolds (Li et al., 2021). Extensions cover proximally smooth constraints and randomized submanifold updates (Davis et al., 2025; Cheung et al., 2025); smoothing and variance reduction exploit composite structure (Peng et al., 2023; Deng et al., 2025). Li et al. (2021, Section 6) already cover compact embedded manifolds. Retaining their RSSG update and proximal criterion, we establish finite-moment high-probability convergence by controlling retraction errors and proximal-value increments.

Weakly convex optimization under heavy-tailed noise. Zhu et al. (2026) give finite-moment expectation guarantees for unclipped SGD on bounded convex domains (Theorem 4.2). Their high-probability guarantees use sub-Weibull noise for unclipped SGD (Theorem 4.4), or clipping under finite moments (Theorem 5.1). We establish finite-moment high-probability convergence of unclipped RSSG with retractions on compact manifolds. Related unclipped analyses give expectation guarantees on bounded domains (Liu, 2026b), for Hölder-smooth objectives (Fatkhullin et al., 2025), and for differentiable upper-smooth objectives (Liu, 2026a). Nonsmooth finite-moment analysis also uses online-to-nonconvex reductions (Liu et al., 2024).

Clipping and normalization under heavy-tailed noise. Clipping controls fluctuations in convex and smooth nonconvex optimization (Zhang et al., 2020; Cutkosky & Mehta, 2021; Nguyen et al., 2023; Gorbunov et al., 2024). Normalized mini-batch methods admit choices independent of the noise parameters in Euclidean smooth optimization (Hübler et al., 2025; Liu & Zhou, 2025); Riemannian normalization also appears for geodesically convex objectives (Dodd et al., 2024). We couple these operations to weakly convex proximal descent, accounting for clipping bias or directional error within the prescribed retraction radius. Appendix N also distinguishes this oracle model from heavy-tailed manifold data and smooth zeroth-order settings (Shen et al., 2023; Brunnstrom et al., 2024; Ma et al., 2026).

3 PROBLEM SETTING AND GEOMETRIC PRELIMINARIES

We study (1) on a compact C^2 embedded submanifold $\mathcal{M} \subset \mathbb{R}^d$, with the induced Euclidean metric, tangent space $\mathrm{T}_x\mathcal{M}$, and normal space $\mathrm{N}_x\mathcal{M}$. Weak convexity is imposed on the ambient objective. The Stiefel instance $\mathrm{St}(n,r) = \{X \in \mathbb{R}^{n\times r} : X^\top X = I_r\}$ uses the Frobenius inner product. The reach $\tau = \mathrm{reach}(\mathcal{M}) > 0$ is the largest radius for which every point in the tubular neighborhood has a unique nearest point on $\mathcal{M}$ (Federer, 1959).

Assumption 3.1 (Ambient weak convexity and bounded subgradients)**.** *The function $f : \mathbb{R}^d \to \mathbb{R}$ is lower semicontinuous and ρ-weakly convex: $f + (\rho/2)\|\cdot\|^2$ is convex. For every $x \in \mathcal{M}$ and every ambient limiting subgradient $g \in \partial f(x)$, $\|g\| \le G$. The value $f_\star := \min_{x\in\mathcal{M}} f(x)$ is finite.*

Write $\partial_{\mathcal{M}} f(x) = P_{\mathrm{T}_x\mathcal{M}} \partial f(x)$ for the Riemannian subdifferential.

Assumption 3.2 (Conditional finite-moment oracle)**.** *Given the history $\mathcal{F}_{t-1}$, the oracle returns $\widetilde{v}_t \in \mathrm{T}_{x_t}\mathcal{M}$ with*

$$\mathbb{E}_t \widetilde{v}_t = v_t \in \partial_{\mathcal{M}} f(x_t), \qquad \mathbb{E}_t \|\widetilde{v}_t - v_t\|^p \le \sigma^p, \quad 1 < p \le 2, \tag{6}$$

almost surely, where $\mathbb{E}_t = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ and $\sigma < \infty$.

Here G bounds the population subgradients; the sample oracle may be unbounded and have infinite variance. A batch consists of B conditionally independent oracle calls at the same point, with the same conditional mean.

3.1 PROXIMAL STATIONARITY

The reach-explicit inequality below extends the subgradient comparison of Li et al. (2021, Theorem 1); see Cheung et al. (2025, Lemma 4.1). Its proof is in Appendix B.

Lemma 3.3 (Reach-explicit restricted weak convexity)**.** *Under Assumption 3.1, set $q_{\mathcal{M}} := \rho + G/\tau$. For all $x, y \in \mathcal{M}$ and all $v \in \partial_{\mathcal{M}} f(x)$ induced by an ambient $g \in \partial f(x)$,*

$$f(y) \ge f(x) + \langle v, y - x\rangle - \frac{q_{\mathcal{M}}}{2}\|y - x\|^2. \tag{7}$$

For $\bar{\rho} > q_{\mathcal{M}}$ define the constrained proximal value

$$\Phi_{\bar{\rho}}(x) := \min_{y \in \mathcal{M}} \left\{ f(y) + \frac{\bar{\rho}}{2} \|y - x\|^2 \right\}. \tag{8}$$

Compactness gives nonempty compact sets of minimizers. The argmin map has a measurable graph; fix a measurable selection $\widehat{x}(x)$ and define

$$\Theta_{\bar{\rho}}(x) := \bar{\rho} \|x - \widehat{x}(x)\|. \tag{9}$$

The residual certifies a nearby approximate stationary point: $\|x - \widehat{x}(x)\| = \Theta_{\bar{\rho}}(x)/\bar{\rho}$ and $\operatorname{dist}(0, \partial_{\mathcal{M}} f(\widehat{x}(x))) \le \Theta_{\bar{\rho}}(x)$ (Appendix C). Complexity counts the oracle calls needed to make the averaged squared residual at most ε^2, in expectation or with high probability.

3.2 Retraction distance control

A retraction maps a tangent step back to $\mathcal{M}$ and satisfies $\mathcal{R}_x(0) = x$, $D\mathcal{R}_x(0) = \mathrm{Id}$. The following condition records its effect on the distance comparisons used in proximal descent.

Definition 3.4 (Retraction distance control)**.** *We say that a retraction satisfies the distance bound with constants $(r_{\mathcal{R}}, \kappa_{\mathcal{R}})$ if $r_{\mathcal{R}} \in (0, \infty]$, $\kappa_{\mathcal{R}} \ge 0$, and*

$$\mathfrak{D}_{\mathcal{R}}(x, y; \xi) := \|\mathcal{R}_x(\xi) - y\|^2 - \|x + \xi - y\|^2 \le \kappa_{\mathcal{R}} \|\xi\|^2 \tag{10}$$

for $x, y \in \mathcal{M}$ and $\xi \in \mathrm{T}_x\mathcal{M}$ with $\|\xi\| \le r_{\mathcal{R}}$.

Smooth retractions on compact manifolds admit uniform local constants; Appendix A gives an explicit construction. For the Stiefel polar retraction (23), the known comparison $\left\|\mathcal{R}_X^{\mathrm{pol}}(\Xi) - Y\right\|_F^2 \le \|X + \Xi - Y\|_F^2$ (Li et al., 2021, Lemma 1) gives $(r_{\mathcal{R}}, \kappa_{\mathcal{R}}) = (\infty, 0)$; see Chen et al. (2021, Lemma 2.3) and Appendix B. QR is also globally defined, but can have positive distance error (Appendix K).

4 Stochastic subgradient methods under heavy tails

We study the RSSG update and two bounded-step variants, GeoClip and GeoNorm. Our main contribution is RSSG's finite-moment high-probability analysis; the variants provide deterministic step control. Define $\mathrm{clip}_\lambda(z) = \min\{1, \lambda/\|z\|\} z$ and $\mathrm{nrm}(z) = z/\|z\|$ for $z \ne 0$, setting both to zero at the origin. The three methods share the update

$$x_{t+1} = \mathcal{R}_{x_t}(-d_t), \qquad d_t = \begin{cases} \eta \widetilde{v}_t, & \text{RSSG}, \\ \eta \, \mathrm{clip}_\Lambda(\widetilde{v}_t), & \text{GeoClip}, \\ \alpha \, \mathrm{nrm}\left(B^{-1} \sum_{j=1}^{B} \widetilde{v}_{t,j} \right), & \text{GeoNorm}. \end{cases} \tag{11}$$

RSSG uses the sample's direction and magnitude, preserving conditional unbiasedness before retraction. With a globally defined retraction, every sample yields a valid iterate. The question is whether unbounded tangent steps permit proximal descent and concentration under finite moments. Section 5.2 answers this for the update of Li et al. (2021), using compactness and local retraction accuracy.

GeoClip provides direct control of every tangent step using one oracle call. Clipping at Λ gives $\|d_t\| \le \eta\Lambda$, so $\eta\Lambda \le r_{\mathcal{R}}$ keeps the update in a prescribed neighborhood. For a statistical threshold Λ_{stat}, set $\Lambda = \min\{\Lambda_{\mathrm{stat}}, r_{\mathcal{R}}/\eta\}$ and choose the parameters jointly so that $\Lambda \ge 2G$. This requires $\Lambda_{\mathrm{stat}} \ge 2G$ and $\eta \le r_{\mathcal{R}}/(2G)$ when $G > 0$. A smaller threshold reduces the step bound but increases the clipping-bias bound. Theorem 5.3 balances this bias and the stochastic fluctuations with a feasible stepsize. The method therefore uses the given local retraction directly, with a deterministic bound on each tangent step.

GeoNorm separates step length from direction estimation. Normalizing the mean of B conditionally independent tangent samples gives $\|d_t\| \le \alpha$, hence $\alpha \le r_{\mathcal{R}}$ enforces the same geometric constraint without a clipping threshold. Averaging reduces the mean-subgradient error at order $B^{-(p-1)/p}$, which our analysis converts into a bound on the directional error in descent. The choices

$\alpha = \alpha_0/\sqrt{T}$, $0 < \alpha_0 \le r_{\mathcal{R}}$, and $B = T$ use neither p nor σ. Single-sample RSSG and GeoClip also admit parameter choices independent of p, σ, with different iteration and oracle bounds (Section 5.5). At $B = 1$, GeoNorm uses the sample direction and takes more updates at a fixed sample budget. Algorithm 1 in Appendix L specifies all three procedures and their parameter conditions.

5 Convergence under finite-moment noise

Throughout this section, set $\Delta = \Phi_{\bar\rho}(x_1) - f_\star$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. We first derive the distance comparison needed for proximal descent. RSSG uses compactness to control unbounded tangent steps and concentrate proximal-value increments. GeoClip and GeoNorm bound the steps directly, then control the bias or directional error introduced by their updates. All results concern averaged squared proximal stationarity. For a uniform output index, this is the conditional expectation over the index given the trajectory.

5.1 From distance control to proximal descent

The common starting point is to use the proximal point at x as a comparator after the update. Its squared-distance change bounds the change in $\Phi_{\bar\rho}$. Restricted weak convexity supplies the descent term, while the retraction distance bound controls the remainder. For $g = v + e$, $v \in \partial_{\mathcal{M}} f(x)$, and $x^+ = \mathcal{R}_x(-\eta g)$ with $\eta \|g\| \le r_{\mathcal{R}}$, Lemma D.1 gives

$$\Phi_{\bar\rho}(x^+) - \Phi_{\bar\rho}(x) \le -\frac{\eta(\bar\rho - q_{\mathcal{M}})}{2\bar\rho}\Theta_{\bar\rho}(x)^2 + \bar\rho\eta\,\langle e, \widehat{x}(x) - x\rangle + \frac{\bar\rho(1+\kappa_{\mathcal{R}})}{2}\eta^2\|g\|^2. \tag{12}$$

Weak convexity and curvature enter through $q_{\mathcal{M}} = \rho + G/\tau$; the retraction contributes $1 + \kappa_{\mathcal{R}}$ and the admissible radius. This quadratic bound applies when the tangent step stays within $r_{\mathcal{R}}$, as enforced by GeoClip and GeoNorm. For RSSG, we first replace it with a global p-order bound that is integrable under the oracle assumption.

5.2 Expectation and high-probability convergence of RSSG

Compactness controls the large steps to which the local quadratic bound does not apply. Combined with local distance control, it gives the following estimate for a globally defined retraction.

Lemma 5.1 (p-order distance remainder)**.** *Let $\mathcal{M}$ be compact with diameter $D_{\mathcal{M}}$, and let $\mathcal{R}$ be defined on the entire tangent bundle. Suppose Definition 3.4 holds for some finite radius $r_{\mathcal{R}} > 0$ and constant $\kappa_{\mathcal{R}} \ge 0$. For every $x, y \in \mathcal{M}$, $\xi \in \mathrm{T}_x\mathcal{M}$, and $1 < p \le 2$,*

$$\|\mathcal{R}_x(\xi) - y\|^2 - \|x - y\|^2 \le 2\langle \xi, x - y\rangle + C_{\mathcal{R},p}\|\xi\|^p, \tag{13}$$

where $C_{\mathcal{R},p}$ depends on $D_{\mathcal{M}}, r_{\mathcal{R}}, \kappa_{\mathcal{R}}, p$. Equation (34) *in Appendix F gives an explicit choice.*

The local distance bound controls the remainder by $(1 + \kappa_{\mathcal{R}})\|\xi\|^2$ for $\|\xi\| \le r_{\mathcal{R}}$. For larger steps, compactness bounds it by $D_{\mathcal{M}}^2 + 2D_{\mathcal{M}}\|\xi\|$. Dividing the two bounds by $\|\xi\|^p$ gives the global p-order bound. Every globally defined C^2 retraction on a compact manifold admits suitable local constants (Proposition A.1).

Substituting $\xi = -\eta\widetilde{v}_t$ into Lemma 5.1 preserves the unbiased descent term. Since $\mathbb{E}_t\|\widetilde{v}_t\|^p \le (G + \sigma)^p$, telescoping gives the expectation bound in Theorem F.2 (Appendix F). We next state the high-probability result.

The same distance bound also provides the variance estimate needed for high-probability convergence. Taking $y = x$ in (13) bounds the squared displacement by $C_{\mathcal{R},p}\|\xi\|^p$. Moreover, on $\mathcal{M}$, $\Phi_{\bar\rho}$ is $\bar\rho D_{\mathcal{M}}$-Lipschitz and takes values in $[f_\star, f_\star + \bar\rho D_{\mathcal{M}}^2/2]$. Its centered increment therefore has bounded range and conditional variance of order $\eta^p(G + \sigma)^p$, even when the oracle has infinite variance. Concentrating these increments gives the following result.

Theorem 5.2 (High-probability convergence of RSSG)**.** *Suppose Assumptions 3.1–3.2 hold, $\bar\rho > q_{\mathcal{M}}$, and $\mathcal{R}$ is a globally defined smooth retraction satisfying* (13)*. Run RSSG from deterministic x_1. For $G + \sigma > 0$, choose $\eta = \eta_0/((G + \sigma)T^{1/p})$ with $\eta_0 > 0$. For $\delta \in (0, 1)$ and $\ell = \log(1/\delta)$, with*

probability at least $1-\delta$*,*

$$\frac{1}{T}\sum_{t=1}^{T}\Theta_{\bar{\rho}}(x_t)^2 \le C_{\mathrm{raw}}(G+\sigma)\frac{1+\sqrt{\ell}+\ell}{T^{(p-1)/p}}, \tag{14}$$

where C_{raw} *is given in Appendix F and depends on* $\bar{\rho}, q_{\mathcal{M}}, D_{\mathcal{M}}, C_{\mathcal{R},p}, \Delta, \eta_0$*, but not on* T *or* δ*. Thus* $\widetilde{O}(\varepsilon^{-2p/(p-1)})$ *oracle calls suffice for averaged squared proximal stationarity at most* ε^2 *with probability* $1-\delta$*.*

Both polar and QR are globally defined smooth retractions: for Stiefel tangent ξ, $x+\xi$ has full column rank and its QR factor with positive diagonal depends smoothly on the trial matrix. Positive QR distance error changes the constant, while compactness preserves the rate. Polar admits the explicit choice $C_{\mathcal{R},p} = (4D_{\mathcal{M}})^{2-p}$ (Corollary F.1). Its distance comparison sharpens the remainder bound within the general analysis.

5.3 High-probability convergence of GeoClip

GeoClip keeps each tangent step within $r_{\mathcal{R}}$, so the quadratic descent bound (12) applies directly. Increasing Λ reduces the clipping-bias bound $O(\sigma^p\Lambda^{1-p})$ and increases the centered-second-moment bound $O(\sigma^p\Lambda^{2-p})$. The stepsize must also satisfy $\eta\Lambda \le r_{\mathcal{R}}$. The following joint choice balances these statistical terms while respecting the geometric constraint. The complete finite-horizon bound and clipping estimates are in Appendix D.

Theorem 5.3 (Rate and oracle complexity)**.** *Suppose Assumptions 3.1–3.2 hold,* $\mathcal{R}$ *satisfies Definition 3.4,* $G>0$*, and* $\bar{\rho} > q_{\mathcal{M}}$*. Run GeoClip with*

$$\Lambda_T = \max\{2G, \sigma T^{1/p}\}, \qquad \eta_T = \eta_0 \min\left\{\Lambda_T^{-1}, (G\sqrt{T})^{-1}\right\}, \tag{15}$$

where $0 < \eta_0 \le r_{\mathcal{R}}$ *and* $\delta \in (0,1)$*. Since* $\eta_T\Lambda_T \le \eta_0$*, Algorithm 1 uses* Λ_T *without further geometric truncation. Then, with probability at least* $1-\delta$*,*

$$\frac{1}{T}\sum_{t=1}^{T}\Theta_{\bar{\rho}}(x_t)^2 \le C_{\mathrm{geo}}\left(1+\sqrt{\log\frac{4}{\delta}}+\log\frac{4}{\delta}\right)\max\left\{\frac{\sigma}{T^{(p-1)/p}}, \frac{G}{\sqrt{T}}\right\}, \tag{16}$$

where C_{geo} *is displayed in Appendix J and depends on* Δ/η_0*,* $\bar{\rho}$*,* G*,* $q_{\mathcal{M}}$*,* η_0*, and* $\kappa_{\mathcal{R}}$*, but not on* T *or* δ*. Consequently,* $T = \widetilde{O}(\max\{\sigma^{p/(p-1)}\varepsilon^{-2p/(p-1)}, G^2\varepsilon^{-4}\})$ *oracle calls suffice.*

5.4 Normalization with unknown moment parameters

GeoNorm controls step length by normalization and directional accuracy by averaging. The batch mean has expected error at most $2\sigma B^{-(p-1)/p}$, by a vector von Bahr–Esseen bound. Normalization introduces a factor $1/\|v_t\|$ in the direction error. Restricted weak convexity and proximal minimality give $\langle v_t, x_t - \widehat{x}_t\rangle \ge (\bar{\rho}-q_{\mathcal{M}})\|x_t-\widehat{x}_t\|^2/2$. Hence $\|x_t-\widehat{x}_t\|/\|v_t\| \le 2/(\bar{\rho}-q_{\mathcal{M}})$ when $v_t \ne 0$. This controls the directional error in the proximal comparison even near stationarity. The normalized steps also give bounded fluctuations; Appendix I proves the resulting bound.

Theorem 5.4 (Parameter choices independent of p, σ)**.** *Suppose Assumptions 3.1–3.2 hold with* $G>0$*,* $\bar{\rho} > q_{\mathcal{M}}$*,* $\mathcal{R}$ *satisfying Definition 3.4,* $\delta \in (0,1)$*, and conditionally independent batches. For* $0 < \alpha \le r_{\mathcal{R}}$*, with probability at least* $1-\delta$*,*

$$\frac{1}{T}\sum_{t=1}^{T}\Theta_{\bar{\rho}}(x_t)^2 \le C_{\mathrm{n}}\left[\frac{\Delta}{\alpha T}+\alpha+\sigma B^{-(p-1)/p}+\sqrt{\frac{\log(1/\delta)}{T}}\right], \tag{17}$$

where C_{n} *depends only on* $\bar{\rho}, G, q_{\mathcal{M}}, \kappa_{\mathcal{R}}$*; Theorem D.3 gives the exact coefficients. With* $B=T$ *and* $\alpha = \alpha_0/\sqrt{T}$*,* $0 < \alpha_0 \le r_{\mathcal{R}}$*, the bound has order* $T^{-1/2}(1+\sqrt{\log(1/\delta)})+\sigma T^{-(p-1)/p}$*. These choices use neither* p *nor* σ *and take* T^2 *oracle calls.*

For the Stiefel polar retraction, Lemma B.1 sets $\kappa_{\mathcal{R}} = 0$ and removes the restriction on α_0. With $Q = T^2$ samples, this gives $\widetilde{O}(\varepsilon^{-4p/(p-1)})$ oracle complexity. The general bound (17) also quantifies the directional error for a fixed batch size.

5.5 Iteration and oracle costs with unknown moment parameters

Single-sample methods also admit choices independent of p, σ: $\eta_T = \eta_0 T^{-1/2}$ for RSSG, and $\Lambda_T = \max\{2G, \sqrt{T}\}$, $\eta_T = \eta_0/\Lambda_T$ for GeoClip, with $\eta_0 \le r_{\mathcal{R}}$ in the latter case. Under their respective assumptions, both give high-probability order $\widetilde{O}(T^{-(p-1)/2})$ and oracle complexity $\widetilde{O}(\varepsilon^{-4/(p-1)})$ (Corollary J.1, Appendix J). For $1 < p < 2$, GeoNorm with $B = T$ has a better iteration bound, $\widetilde{O}(\varepsilon^{-2p/(p-1)})$, at the higher oracle bound $\widetilde{O}(\varepsilon^{-4p/(p-1)})$. At $p = 2$, all three bounds give order ε^{-4} iterations, while $B = T$ uses order ε^{-8} samples.

6 Experiments

We evaluate the iteration rates of GeoClip and batched GeoNorm, compare all three updates at matched sample budgets, and examine how retraction distance control affects their trajectories. RSSG, GeoClip, and GeoNorm are evaluated as the updates in (11). Comparators include statistical clipping, a tangent geometric median of eight block means (eight samples per block), and full-batch Riemannian subgradient descent. Synthetic comparisons share the data, initialization, and oracle-index stream across methods. Primary DPCP stepsizes are selected on independent validation seeds. The linear and DPCP objectives are ambient-convex instances of the weakly convex problem class. Online-noise experiments evaluate the finite-moment oracle model, and finite-dataset experiments evaluate recovery from heavy-tailed observations. Appendix M gives the full protocol and additional results.

Stationarity under heavy-tailed oracles. For $f(X) = \langle A, X\rangle_F$ on $\mathrm{St}(32, r)$, $r \in \{1, 4\}$, the proximal point is $\mathrm{polar}(\bar{\rho}X - A)$, allowing exact evaluation of $\Theta_{\bar{\rho}}$. We use horizons 64–2048, four noise families (Student-t, Pareto, symmetric stable, and Gaussian), and finite moment orders $p \in \{1.2, 1.5, 1.8, 2\}$, with distribution parameters chosen to ensure the specified moment exists. Figure 1 shows the Pareto curves over 50 seeds. All 48 GeoClip and GeoNorm ($B = T$) Theil–Sen fits across the full matrix meet the exponent tolerance of 0.1. The largest fitted-minus-predicted exponent is 0.0082. These empirical slopes agree with the GeoClip and batched GeoNorm iteration orders in Theorems 5.3 and 5.4.

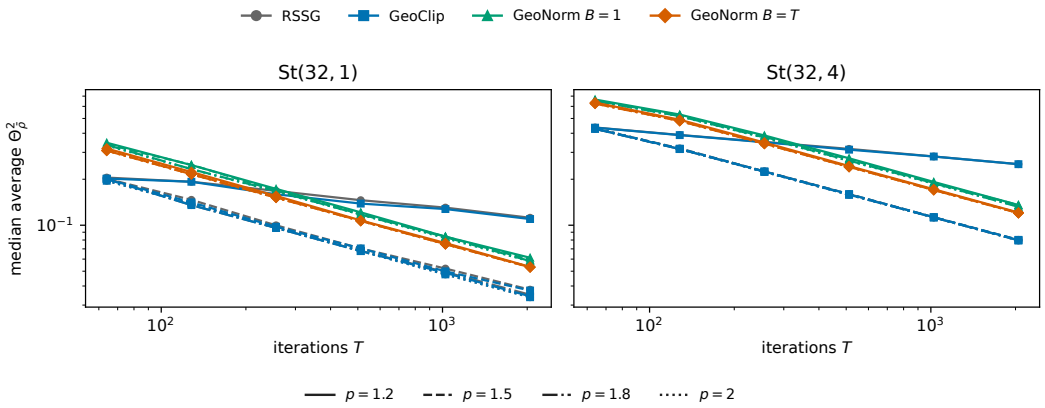

Figure 1: Heavy-tail convergence on the linear Stiefel problem. Curves show the median running average of exact squared proximal stationarity under Pareto noise (50 seeds). Both axes are logarithmic. RSSG shares GeoClip's stepsize schedule. The 48-fit GeoClip and batched GeoNorm analysis also covers Student-t, stable, and Gaussian noise.

Subspace recovery at a fixed oracle budget. DPCP learns a subspace normal basis through a nonsmooth objective on Stiefel. It connects the weakly convex formulation to robust recovery. We minimize its empirical objective $F(X) = N^{-1}\sum_{i=1}^{N} \left\|y_i^\top X\right\|_2$ with $(n, r, N) = (100, 5, 5000)$. Table 2 reports all six primary settings at $Q = 2^{18}$ oracle calls and 20 paired seeds. Single-sample

methods take Q updates, and mini-batch methods satisfy $TB = Q$. In particular, GeoNorm with $B = T$ takes 512 updates of 512 samples, and full-batch RSGM takes 52 updates (260,000 calls). With polar retraction, GeoClip equals statistical clipping, so their identical results share a row.

Table 2: Synthetic DPCP at 2^{18} oracle calls: median projector error and IQR over 20 paired seeds. All six primary settings are shown. Bold marks the lowest median in each column. A dagger marks a Holm-adjusted $p < 0.05$ in the comparisons of GeoClip and GeoNorm ($B = T$) against RSSG. Lower is better.

	Pareto			Student-t		
p	1.2	1.5	1.8	1.2	1.8	2.0
Outlier fraction	0.8	0.5	0.2	0.5	0.5	0.5
RSSG	0.9259 [0.9185,0.9426]	0.2094 [0.1499,0.3022]	0.0346 [0.0278,0.0435]	0.6543 [0.6049,0.7200]	0.3014 [0.2079,0.3282]	0.2307 [0.1967,0.2618]
GeoClip / StatClip	0.9204 [0.9097,0.9389]	0.1599† [0.1370,0.1943]	0.0346 [0.0278,0.0435]	0.4560† [0.4340,0.5461]	0.2146† [0.1932,0.2500]	0.1974† [0.1611,0.2322]
GeoNorm ($B = 1$)	**0.4666** [0.4254,0.5637]	**0.0119** [0.0116,0.0125]	**0.0072** [0.0069,0.0076]	**0.0124** [0.0118,0.0126]	**0.0119** [0.0116,0.0124]	**0.0125** [0.0118,0.0128]
GeoNorm ($B = 64$)	0.9139 [0.8985,0.9299]	0.0853 [0.0737,0.0931]	0.0369 [0.0342,0.0389]	0.1859 [0.1621,0.2356]	0.1078 [0.1042,0.1166]	0.1009 [0.0923,0.1105]
GeoNorm ($B = T$)	0.9373 [0.9211,0.9416]	0.0910† [0.0838,0.0956]	0.0421 [0.0403,0.0428]	0.5636† [0.5013,0.7133]	0.1081† [0.1014,0.1249]	0.0976† [0.0894,0.1223]
Tangent MoM	0.8156 [0.7882,0.8311]	0.0334 [0.0314,0.0373]	0.0178 [0.0172,0.0187]	0.0513 [0.0461,0.0582]	0.0399 [0.0376,0.0418]	0.0403 [0.0387,0.0418]
Full-batch RSGM	0.9465 [0.9416,0.9552]	0.6484 [0.5990,0.6800]	0.0477 [0.0417,0.0558]	0.8261 [0.8091,0.8464]	0.7625 [0.7094,0.7942]	0.6986 [0.6547,0.7276]

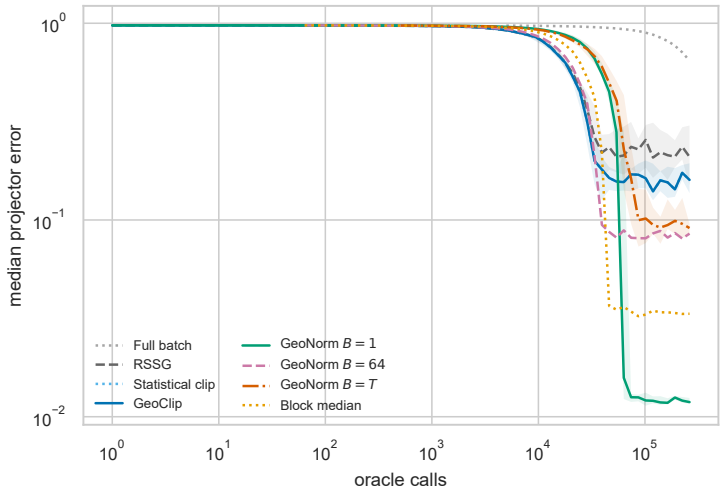

Figure 2: Sample efficiency for Pareto $p = 1.5$ DPCP with 50% outliers. Curves show median projector error over the same 20 paired seeds. Shading gives IQRs for RSSG, GeoClip, and GeoNorm ($B = 1, T$). All methods share a budget of 2^{18} oracle calls; batch sizes change the number of updates. GeoClip and statistical clipping coincide under polar.

GeoNorm ($B = 1$) attains the lowest median in every setting (Table 2). For Student-t noise with $p = 1.2$, its error is 0.0124, compared with 0.6543 for RSSG and 0.4560 for GeoClip; under the hardest Pareto setting it lowers the median from 0.9259 to 0.4666. GeoClip and GeoNorm ($B = T$) improve on RSSG with positive paired intervals in four and three of the five $p < 2$ settings, respectively, with three Holm-significant heavy-tail comparisons for each. At Pareto $p = 1.5$, the paired gains are 0.047 for GeoClip and 0.103 for GeoNorm ($B = T$). Appendix M reports all intervals.

Figure 2 shows the matched-budget trajectories. At the common budget, the final medians are 0.0119, 0.0853, and 0.0910 for $B = 1, 64, T$. Smaller batches permit more updates. Single-sample

normalization also has an exact invariance for radial contamination. For the DPCP tangent subgradient $v_X(y)$ and any $c > 0$,

$$v_X(cy) = c\, v_X(y), \qquad \mathrm{nrm}(v_X(cy)) = \mathrm{nrm}(v_X(y)). \tag{18}$$

Thus $B = 1$ removes positive radial multipliers sample by sample. The recovery comparison reflects this invariance and update frequency; Figure 1 evaluates the general-oracle $B = T$ guarantee.

Retraction and step-size diagnostics. An orientation-sensitive linear-Stiefel objective distinguishes QR and polar updates and evaluates step-amplitude control (Figure 3). At $\eta\Lambda = 1$, QR-StatClip exceeds the controlled radius in 73.8% and 86.7% of steps for $p = 1.2$ and $p = 1.8$, whereas QR-GeoClip never does. The median maximum defect falls from 0.039 to 0.012 and from 0.035 to 0.013, with stationarity gains of 0.205 and 0.166. The cap is inactive at $\eta\Lambda = 0.2$. Appendix M gives full trajectory statistics. The QR and circle examples in Appendix K characterize positive distance errors.

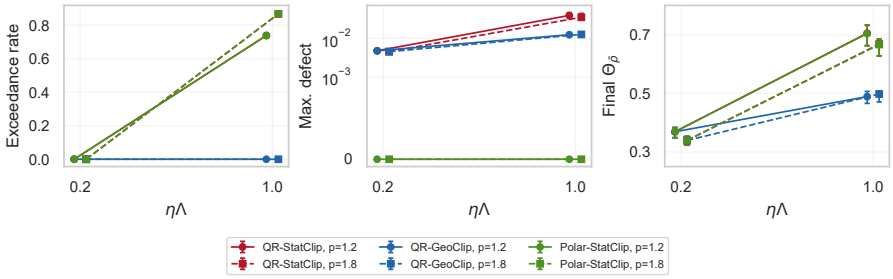

Figure 3: Local-cap ablation over 50 seeds. At $\eta\Lambda = 1$ GeoClip prevents QR radius exceedance and lowers the defect and stationarity measure; at 0.2 the cap is inactive. The radius is the prescribed neighborhood for QR distance control; polar has global distance control.

Subspace recovery from motion trajectories. Across all 155 Hopkins155 sequences (Tron & Vidal, 2007), RSSG and GeoClip attain the lowest category-level medians among the three updates (Table 6). For checker sequences at $p = 1.2$, GeoClip reduces the median error from 0.1274 to 0.1249. These sequence-weighted results show the effectiveness of single-sample updates on motion trajectories. Repeated runs on 30 sequences quantify variability across corruption, initialization, and sampling seeds.

7 Conclusion

We establish high-probability convergence of unclipped RSSG for weakly convex optimization on compact embedded manifolds under finite p-th moments. Local retraction accuracy and compactness control both proximal descent and proximal-value fluctuations, without changing the update. GeoClip jointly selects clipping thresholds and stepsizes to keep every tangent step within a prescribed neighborhood. GeoNorm separates step length from direction estimation through normalized mini-batches. Bounds with and without knowledge of p, σ quantify their iteration and oracle costs, recovering the single-sample ε^{-4} order at $p = 2$. Exact-stationarity and subspace-recovery experiments relate these guarantees to convergence trends, radial invariance, and fixed-budget performance.

AI USE STATEMENT

Generative AI tools assisted with literature discovery, developing and checking mathematical arguments, experimental design, result analysis, and drafting and editing. The authors take responsibility for the final text, claims, proofs, code, and empirical results.

ETHICS STATEMENT

We use synthetic data and public trajectories without human-subject attributes.

REFERENCES

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

Jesper Brunnstrom, Marc Moonen, and Filip Elvander. Robust signal and noise covariance matrix estimation using riemannian optimization. In *European Signal Processing Conference*, pp. 291–295, 2024.

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.

Andy Yat-Ming Cheung, Jinxin Wang, Man-Chung Yue, and Anthony Man-Cho So. Randomized submanifold subgradient method for optimization over Stiefel manifolds, 2025. URL https://arxiv.org/abs/2409.01770.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4883–4895, 2021.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019. doi: 10.1137/18M1178244.

Damek Davis, Dmitriy Drusvyatskiy, and Zhan Shi. Stochastic optimization over proximally smooth sets. *SIAM Journal on Optimization*, 35(1):157–179, 2025. doi: 10.1137/20M1320225.

Kangkang Deng, Zheng Peng, and Weihe Wu. Single-loop $\mathcal{O}(\epsilon^{-3})$ stochastic smoothing algorithms for nonsmooth Riemannian optimization, 2025. URL https://arxiv.org/abs/2505.09485.

Daniel Dodd, Louis Sharrock, and Christopher Nemeth. Learning-rate-free stochastic optimization over Riemannian manifolds. In *International Conference on Machine Learning*, volume 235, pp. 11105–11148. PMLR, 2024.

Ilyas Fatkhullin, Florian Hübler, and Guanghui Lan. Can SGD handle heavy-tailed noise?, 2025. URL https://arxiv.org/abs/2508.04860.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

Adrien Fradin, Abdurakhmon Sadiev, Laurent Condat, and Peter Richtárik. Tight lower bounds and optimal algorithms for stochastic nonconvex optimization with heavy-tailed noise. In *International Conference on Artificial Intelligence and Statistics*, volume 300, pp. 2548–2556. PMLR, 2026.

Eduard Gorbunov, Abdurakhmon Sadiev, Marina Danilova, Samuel Horvath, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. In *International Conference on Machine Learning*, 2024.

Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed SGD. In *International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 2413–2421. PMLR, 2025.

Hideaki Iiduka. Muon converges under heavy-tailed noise: Nonconvex hölder-smooth empirical risk minimization. *arXiv preprint arXiv:2603.15059*, 2026.

Xiao Li, Shixiang Chen, Zengde Deng, Qing Qu, Zhihui Zhu, and Anthony Man-Cho So. Weakly convex optimization over stiefel manifold using riemannian subgradient-type methods. *SIAM Journal on Optimization*, 31(3):1605–1634, 2021. doi: 10.1137/20M1321000.

Langqi Liu, Yibo Wang, and Lijun Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In *International Conference on Machine Learning*, pp. 32122–32138, 2024.

Zijian Liu. In-expectation convergence of stochastic gradient methods under heavy-tailed noise, 2026a. URL https://arxiv.org/abs/2606.00520.

Zijian Liu. Online convex optimization with heavy tails: Old algorithms, new regrets, and applications. In *International Conference on Algorithmic Learning Theory*, volume 313, pp. 1–47, 2026b.

Zijian Liu. Clipped gradient methods for nonsmooth convex optimization under heavy-tailed noise: A refined analysis. In *International Conference on Learning Representations*, 2026c.

Zijian Liu and Zhengyuan Zhou. Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, 2025.

Yinpu Ma, Cunlin Li, and Shiyu Zhang. Projection-robust zeroth-order optimization on manifolds under heavy-tailed oracle noise. *Journal of King Saud University Computer and Information Sciences*, 38(7):805, 2026. doi: 10.1007/s44443-026-01272-7.

Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tails. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zheng Peng, Weihe Wu, Jiang Hu, and Kangkang Deng. Riemannian smoothing gradient type algorithms for nonsmooth optimization problem on compact riemannian submanifold embedded in euclidean space. *Applied Mathematics & Optimization*, 88(3):85, 2023. doi: 10.1007/s00245-023-10061-x.

Yinan Shen, Jingyang Li, Jian-Feng Cai, and Dong Xia. Computationally efficient and statistically optimal robust high-dimensional linear regression, 2023. URL https://arxiv.org/abs/2305.06199.

Roberto Tron and René Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. doi: 10.1109/CVPR.2007.382974.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011. doi: 10.1214/ECP.v16-1624.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303, 1965.

Ryusei Yamada, Naoki Sato, and Hideaki Iiduka. Vanilla SGD with momentum survives heavy-tailed noise: Convergence analysis without gradient clipping or normalization. In *Uncertainty in Artificial Intelligence*, volume 337, pp. 7670–7700. PMLR, 2026.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems*, volume 33, pp. 15383–15393, 2020.

Tianxi Zhu, Yi Xu, Qi Wang, and Xiangyang Ji. Stochastic gradient methods under heavy-tailed noises in weakly convex optimization. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64202.

A LOCAL RETRACTION DISTANCE BOUNDS

A retraction satisfies $\mathcal{R}_x(0) = x$ and $D\mathcal{R}_x(0) = \mathrm{Id}$ and maps tangent steps near zero back to the manifold. For a closed convex set $\mathcal{X}$, projection provides the global comparison

$$\left\|x^+ - y\right\|^2 \le \|x + \xi - y\|^2 - \left\|x + \xi - x^+\right\|^2, \qquad x^+ = \Pi_{\mathcal{X}}(x + \xi), \quad y \in \mathcal{X}. \tag{19}$$

Retraction-based analysis uses a distance-error bound in its place. Global second-order bounds for smooth, globally defined retractions on compact manifolds already appear in Li et al. (2021, Section 6). The condition below also accommodates locally defined retractions and records the constants used in our finite-moment bounds. Definition 3.4 states this distance control in the main text. The Stiefel polar retraction satisfies it with $(r_{\mathcal{R}}, \kappa_{\mathcal{R}}) = (\infty, 0)$. The local distance bound supplies the small-step part of Lemma 5.1 and the per-step bound for GeoClip and GeoNorm.

Proposition A.1 (Uniform local distance bound)**.** *Every C^2 retraction defined on a neighborhood of the zero section of the tangent bundle of a compact embedded manifold admits a finite radius $r_{\mathcal{R}} > 0$ and $c_{\mathcal{R}} < \infty$ such that, uniformly in x,*

$$\|\mathcal{R}_x(\xi) - (x + \xi)\| \le c_{\mathcal{R}} \|\xi\|^2 \quad \textit{for } \|\xi\| \le r_{\mathcal{R}}. \tag{20}$$

Writing $D_{\mathcal{M}} = \operatorname{diam}(\mathcal{M})$, Definition 3.4 then holds with

$$\kappa_{\mathcal{R}} = 2c_{\mathcal{R}}(D_{\mathcal{M}} + r_{\mathcal{R}}) + c_{\mathcal{R}}^2 r_{\mathcal{R}}^2. \tag{21}$$

For a globally defined retraction, these constants and compactness give the RSSG bound in Theorem F.2. For a local retraction, GeoClip's condition $\eta\Lambda \le r_{\mathcal{R}}$ and GeoNorm's condition $\alpha \le r_{\mathcal{R}}$ suffice for Theorems D.2 and D.3. The reach enters through $q_{\mathcal{M}} = \rho + G/\tau$, and the retraction error through $\kappa_{\mathcal{R}}$. For polar retraction neither a finite radius nor a positive defect coefficient is needed.

Appendix K constructs a finite-moment oracle whose steps leave any prescribed local radius with positive probability. It also gives a quadratic positive distance defect for standard QR retraction on $\mathrm{St}(3, 2)$ and an affine descent reversal for a linear objective. A second-order circle retraction also has positive distance error. These constructions identify the distance errors and radius restrictions addressed by bounded-step updates. Global RSSG convergence instead uses the p-order remainder, including for retractions with positive distance error.

A supplied local retraction can also be extended globally by rescaling:

$$\widetilde{\mathcal{R}}_x(\xi) = \mathcal{R}_x\left(\frac{\xi}{\sqrt{1 + \|\xi\|^2 / r_{\mathcal{R}}^2}}\right).$$

The inner argument has norm strictly below $r_{\mathcal{R}}$, is smooth, equals zero at $\xi = 0$, and has derivative the identity there. Thus $\widetilde{\mathcal{R}}$ is a globally defined retraction with the same regularity. Applying RSSG to this extension changes the given update through radial rescaling; GeoClip and GeoNorm use the supplied local map directly with a prescribed bound on every tangent step.

B GEOMETRIC PROOFS

B.1 PROOF OF LEMMA 3.3

Positive reach gives the normal chord inequality

$$|\langle n, y - x\rangle| \le \frac{\|n\|}{2\tau} \|y - x\|^2, \qquad x, y \in \mathcal{M},\ n \in \mathrm{N}_x\mathcal{M}. \tag{22}$$

It bounds the normal component of the ambient subgradient.

Proof. Let $g \in \partial f(x)$ induce $v = P_{\mathrm{T}_x\mathcal{M}} g \in \partial_{\mathcal{M}} f(x)$ and write $n = P_{\mathrm{N}_x\mathcal{M}} g$. Ambient ρ-weak convexity gives

$$f(y) \ge f(x) + \langle g, y - x\rangle - \frac{\rho}{2} \|y - x\|^2.$$

By (22) and $\|n\| \le \|g\| \le G$,

$$\langle n, y - x\rangle \ge -\frac{G}{2\tau} \|y - x\|^2.$$

Substituting $g = v + n$ proves (7). □

B.2 Proof of Proposition A.1

Proof. The zero section $\{(x,0) : x \in \mathcal{M}\}$ is compact and lies in the open domain of the retraction. There is therefore $r_{\mathcal{R}} > 0$ such that the closed tangent disk bundle $\{(x,\xi) : x \in \mathcal{M},\ \xi \in \mathrm{T}_x\mathcal{M},\ \|\xi\| \le r_{\mathcal{R}}\}$ lies in that domain. Indeed, otherwise there would be points outside the domain with step norms tending to zero; compactness of their base points would give a limit on the zero section, contradicting openness. The disk bundle is a closed bounded subset of the ambient product space, hence compact. Continuity of the second derivative in the fiber variable gives a uniform bound $2c_{\mathcal{R}}$ on its operator norm there. Taylor's integral formula and the retraction identities yield

$$\mathcal{R}_x(\xi) - x - \xi = \int_0^1 (1-s) D^2_\xi \mathcal{R}_x(s\xi)[\xi,\xi]\, ds,$$

which proves (20). Let $e = \mathcal{R}_x(\xi) - (x+\xi)$. Expanding the squared norms,

$$\mathfrak{D}_{\mathcal{R}}(x,y;\xi) = 2\langle x+\xi-y, e\rangle + \|e\|^2.$$

For $\|\xi\| \le r_{\mathcal{R}}$, $\|x+\xi-y\| \le D_{\mathcal{M}} + r_{\mathcal{R}}$ and $\|e\| \le c_{\mathcal{R}}\|\xi\|^2$. Therefore,

$$\begin{aligned}\mathfrak{D}_{\mathcal{R}}(x,y;\xi) &\le 2c_{\mathcal{R}}(D_{\mathcal{M}} + r_{\mathcal{R}})\|\xi\|^2 + c^2_{\mathcal{R}}\|\xi\|^4 \\ &\le \left[2c_{\mathcal{R}}(D_{\mathcal{M}} + r_{\mathcal{R}}) + c^2_{\mathcal{R}} r^2_{\mathcal{R}}\right]\|\xi\|^2.\end{aligned}$$

□

B.3 Global polar distance comparison

The polar retraction on Stiefel is

$$\mathcal{R}^{\mathrm{pol}}_X(\Xi) = (X+\Xi)(I_r + \Xi^\top\Xi)^{-1/2}, \qquad \Xi \in \mathrm{T}_X\mathrm{St}(n,r). \tag{23}$$

It is globally defined because $(X+\Xi)^\top(X+\Xi) = I_r + \Xi^\top\Xi \succ 0$. Its distance comparison gives the constants $(r_{\mathcal{R}}, \kappa_{\mathcal{R}}) = (\infty, 0)$.

Lemma B.1 (Global polar distance comparison)**.** *For every $X, Y \in \mathrm{St}(n,r)$ and every $\Xi \in \mathrm{T}_X\mathrm{St}(n,r)$,*

$$\left\|\mathcal{R}^{\mathrm{pol}}_X(\Xi) - Y\right\|^2_F \le \|X+\Xi-Y\|^2_F. \tag{24}$$

Proof. Write $Z = X + \Xi$. Since $X^\top\Xi + \Xi^\top X = 0$, $Z^\top Z = I + \Xi^\top\Xi$. Let $H = (Z^\top Z)^{1/2} \succeq I$ and $R = ZH^{-1} = \mathcal{R}^{\mathrm{pol}}_X(\Xi)$, so $Z = RH$. For $Y \in \mathrm{St}(n,r)$, set $Q = Y^\top R$. Since $\|Q\|_2 \le 1$, the symmetric part satisfies $\mathrm{sym}(Q) \preceq I$. With $A = H - I \succeq 0$,

$$\begin{aligned}&\|Y-Z\|^2_F - \|Y-R\|^2_F \\ &= \mathrm{tr}(H^2 - I) - 2\,\mathrm{tr}(Q(H-I)) \\ &= 2\,\mathrm{tr}((I - \mathrm{sym}\,Q)A) + \mathrm{tr}(A^2) \ge 0.\end{aligned}$$

This proves the claimed inequality. □

C Proximal stationarity

Lemma C.1 (Existence, distance, and stationarity)**.** *Under Assumption 3.1, let $\bar\rho > q_{\mathcal{M}}$ and let $\widehat{x}$ be any minimizer in* (8)*. Then*

$$\|\widehat{x} - x\| \le \frac{2G}{\bar\rho - q_{\mathcal{M}}}, \tag{25}$$

$$\mathrm{dist}(0, \partial_{\mathcal{M}} f(\widehat{x})) \le \bar\rho\,\|\widehat{x} - x\|. \tag{26}$$

Proof. Existence follows from lower semicontinuity and compactness. Put $d = \|\widehat{x} - x\|$. Proximal minimality with comparator x gives

$$f(\widehat{x}) - f(x) + \frac{\bar\rho}{2} d^2 \le 0. \tag{27}$$

Choose $v \in \partial_{\mathcal{M}} f(x)$ induced by an ambient subgradient of norm at most G. Lemma 3.3 and Cauchy–Schwarz give

$$f(\widehat{x}) - f(x) \geq -Gd - \frac{q_{\mathcal{M}}}{2} d^2.$$

Combining with (27) gives $((\bar{\rho} - q_{\mathcal{M}})/2)d^2 \leq Gd$, proving (25).

At a constrained proximal minimizer, the smooth-manifold Fermat rule yields an ambient $\widehat{g} \in \partial f(\widehat{x})$ such that

$$P_{\mathrm{T}_{\widehat{x}}\mathcal{M}} \left(\widehat{g} + \bar{\rho}(\widehat{x} - x)\right) = 0.$$

It follows that

$$\|P_{\mathrm{T}_{\widehat{x}}\mathcal{M}} \widehat{g}\| = \bar{\rho} \, \|P_{\mathrm{T}_{\widehat{x}}\mathcal{M}}(x - \widehat{x})\| \leq \bar{\rho} d,$$

which proves (26). □

D Complete descent and finite-horizon bounds

The following statements retain all coefficients used in the main-text rates. Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ and $\Delta = \Phi_{\bar{\rho}}(x_1) - f_\star$.

Lemma D.1 (Proximal descent with retraction distance error)**.** *Suppose Assumption 3.1 holds, $\bar{\rho} > q_{\mathcal{M}}$, and $\mathcal{R}$ satisfies Definition 3.4. Let $v \in \partial_{\mathcal{M}} f(x)$, $g = v + e \in \mathrm{T}_x\mathcal{M}$, $\|g\| \leq \Lambda$, and $x^+ = \mathcal{R}_x(-\eta g)$ with $\eta\Lambda \leq r_{\mathcal{R}}$. Then*

$$\begin{aligned}\Phi_{\bar{\rho}}(x^+) - \Phi_{\bar{\rho}}(x) \leq & -\frac{\eta(\bar{\rho} - q_{\mathcal{M}})}{2\bar{\rho}} \Theta_{\bar{\rho}}(x)^2 + \bar{\rho}\eta \left\langle e, \widehat{x}(x) - x \right\rangle \\ & + \frac{\bar{\rho}(1 + \kappa_{\mathcal{R}})}{2} \eta^2 \|g\|^2 . \end{aligned} \tag{28}$$

Moreover,

$$\|\widehat{x}(x) - x\| \leq D_{\mathrm{p}} := \frac{2G}{\bar{\rho} - q_{\mathcal{M}}}. \tag{29}$$

For a clipped oracle, write

$$g_t - v_t = u_t + b_t, \qquad u_t = g_t - \mathbb{E}_t[g_t], \quad b_t = \mathbb{E}_t[g_t] - v_t, \tag{30}$$

where $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$. The clipping bounds below hold with universal constants $c_b = 4$ and $c_2 = 32$:

$$\|b_t\| \leq c_b \sigma^p \Lambda^{1-p}, \quad \mathbb{E}_t \|u_t\|^2 \leq c_2 \sigma^p \Lambda^{2-p}, \quad \|u_t\| \leq 2\Lambda. \tag{31}$$

Theorem D.2 (Finite-horizon high-probability bound)**.** *Suppose Assumptions 3.1–3.2 hold. Let $\bar{\rho} > q_{\mathcal{M}}$, let $\delta \in (0, 1)$, and run Algorithm 1 with effective threshold $\Lambda \geq 2G$ (hence $\eta\Lambda \leq r_{\mathcal{R}}$). Define $\Delta = \Phi_{\bar{\rho}}(x_1) - f_\star$, $M_2 = 2G^2 + c_2\sigma^p\Lambda^{2-p}$, and $\ell = \log(4/\delta)$. Then, with probability at least $1 - \delta$,*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \Theta_{\bar{\rho}}(x_t)^2 \leq \frac{2\bar{\rho}}{\bar{\rho} - q_{\mathcal{M}}} \Bigg[& \frac{\Delta}{\eta T} + \bar{\rho} D_{\mathrm{p}} \left(c_b \sigma^p \Lambda^{1-p} + \frac{2\Lambda\ell}{3T} + \sqrt{\frac{2c_2\sigma^p\Lambda^{2-p}\ell}{T}} \right) \\ & + \bar{\rho}(1 + \kappa_{\mathcal{R}})\eta \left(M_2 + \frac{\Lambda^2 \ell}{T} \right) \Bigg]. \end{aligned} \tag{32}$$

The three stochastic terms in (32) respectively control clipping bias, the martingale inner product, and the clipped second moment. Geometry appears only through $q_{\mathcal{M}}$, $r_{\mathcal{R}}$, and $\kappa_{\mathcal{R}}$.

Theorem D.3 (High-probability bound for GeoNorm-RSSG)**.** *Suppose Assumptions 3.1–3.2 hold with $G > 0$, $\bar{\rho} > q_{\mathcal{M}}$, $\delta \in (0, 1)$, and conditionally independent mini-batches. Run Algorithm 1 with batch size B and $0 < \alpha \leq r_{\mathcal{R}}$. Let $\Delta = \Phi_{\bar{\rho}}(x_1) - f_\star$. Then, with probability at least $1 - \delta$,*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \Theta_{\bar{\rho}}(x_t)^2 \leq \frac{2\bar{\rho}G}{\bar{\rho} - q_{\mathcal{M}}} \Bigg[& \frac{\Delta}{\alpha T} + \frac{\bar{\rho}(1 + \kappa_{\mathcal{R}})\alpha}{2} + \frac{8\bar{\rho}\sigma}{\bar{\rho} - q_{\mathcal{M}}} B^{-(p-1)/p} \\ & + \frac{8\bar{\rho}G}{\bar{\rho} - q_{\mathcal{M}}} \sqrt{\frac{2\log(1/\delta)}{T}} \Bigg]. \end{aligned} \tag{33}$$

E PROOF OF THE PROXIMAL DESCENT LEMMA

Proof of Lemma D.1. Write $\widehat{x} = \widehat{x}(x)$ and $d = \|\widehat{x} - x\|$. Using $\widehat{x}$ as a competitor in the proximal value at x^+,

$$\Phi_{\bar{\rho}}(x^+) - \Phi_{\bar{\rho}}(x) \le \frac{\bar{\rho}}{2}\left(\left\|\widehat{x} - x^+\right\|^2 - d^2\right).$$

By the retraction distance bound with $\xi = -\eta g$,

$$\begin{aligned}\left\|\widehat{x} - x^+\right\|^2 &\le \|\widehat{x} - (x - \eta g)\|^2 + \kappa_{\mathcal{R}}\eta^2 \|g\|^2 \\ &= d^2 + 2\eta\langle g, \widehat{x} - x\rangle + (1 + \kappa_{\mathcal{R}})\eta^2 \|g\|^2.\end{aligned}$$

For the true-subgradient component, Lemma 3.3 gives

$$\langle v, \widehat{x} - x\rangle \le f(\widehat{x}) - f(x) + \frac{q_{\mathcal{M}}}{2}d^2.$$

Proximal minimality gives $f(\widehat{x}) - f(x) \le -(\bar{\rho}/2)d^2$, so

$$\langle v, \widehat{x} - x\rangle \le -\frac{\bar{\rho} - q_{\mathcal{M}}}{2}d^2.$$

Substitute $g = v + e$ and $d = \Theta_{\bar{\rho}}(x)/\bar{\rho}$ to obtain (28). The distance bound is Lemma C.1. □

F FINITE-MOMENT DISTANCE CONTROL AND RSSG CONVERGENCE

An admissible constant in Lemma 5.1 is

$$C_{\mathcal{R},p} = \max\left\{(1 + \kappa_{\mathcal{R}})r_{\mathcal{R}}^{2-p},\ \frac{D_{\mathcal{M}}^2 + 2D_{\mathcal{M}}r_{\mathcal{R}}}{r_{\mathcal{R}}^p}\right\}. \tag{34}$$

Polar retraction gives a sharper expression without a local radius.

Corollary F.1 (Polar remainder constant)**.** *On* $\mathrm{St}(n, r)$ *with polar retraction,* (13) *holds with* $C_{\mathcal{R},p} = (4D_{\mathcal{M}})^{2-p}$*, where* $D_{\mathcal{M}} = \mathrm{diam}(\mathrm{St}(n, r)) \le 2\sqrt{r}$.

The remainder is bounded by $\min\{\|\xi\|^2, 4D_{\mathcal{M}}\|\xi\|\}$: polar comparison gives the quadratic bound, while compactness and $\left\|\mathcal{R}_x^{\mathrm{pol}}(\xi) - x\right\| \le \|\xi\|$ give the linear bound.

Theorem F.2 (Finite-moment convergence of RSSG)**.** *Suppose Assumptions 3.1 and 3.2 hold on the compact embedded manifold* $\mathcal{M}$*, and let* $\bar{\rho} > q_{\mathcal{M}}$*. Let* $\mathcal{R}$ *be a globally defined smooth retraction satisfying* (13) *with constant* $C_{\mathcal{R},p}$*. Run the RSSG row of* (11) *with a deterministic initial point* x_1*, and constant stepsize* $\eta > 0$*. Then*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Theta_{\bar{\rho}}(x_t)^2 \le \frac{2\bar{\rho}}{\bar{\rho} - q_{\mathcal{M}}}\left[\frac{\Delta}{\eta T} + \frac{\bar{\rho}}{2}C_{\mathcal{R},p}(G + \sigma)^p\eta^{p-1}\right]. \tag{35}$$

For $G + \sigma > 0$ *and any* $\eta_0 > 0$*, the choice* $\eta = \eta_0/((G + \sigma)T^{1/p})$ *gives*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Theta_{\bar{\rho}}(x_t)^2 = O\Big((G + \sigma)T^{-(p-1)/p}\Big), \tag{36}$$

with the constant obtained directly from (35)*. For Stiefel polar retraction, take* $C_{\mathcal{R},p} = (4D_{\mathcal{M}})^{2-p}$.

Proof of Lemma 5.1. Fix $x, y \in \mathcal{M}$ and $\xi \in \mathrm{T}_x\mathcal{M}$, and let $s = \|\xi\|$. For $s = 0$, the retraction identity makes both sides of (13) zero. If $0 < s \le r_{\mathcal{R}}$, Definition 3.4 and expansion of the Euclidean square give

$$\begin{aligned}\|\mathcal{R}_x(\xi) - y\|^2 - \|x - y\|^2 - 2\langle\xi, x - y\rangle &\le (1 + \kappa_{\mathcal{R}})s^2 \\ &\le (1 + \kappa_{\mathcal{R}})r_{\mathcal{R}}^{2-p}s^p.\end{aligned}$$

If $s > r_{\mathcal{R}}$, global definition ensures that $\mathcal{R}_x(\xi) \in \mathcal{M}$. Since $\|\mathcal{R}_x(\xi) - y\| \le D_{\mathcal{M}}$ and $\|x - y\| \le D_{\mathcal{M}}$, Cauchy–Schwarz yields

$$\begin{aligned}\|\mathcal{R}_x(\xi) - y\|^2 - \|x - y\|^2 - 2\langle \xi, x - y\rangle &\le D_{\mathcal{M}}^2 + 2D_{\mathcal{M}} s \\ &\le \left(D_{\mathcal{M}}^2 r_{\mathcal{R}}^{-p} + 2D_{\mathcal{M}} r_{\mathcal{R}}^{1-p}\right) s^p.\end{aligned}$$

The second inequality uses $s^{-p} \le r_{\mathcal{R}}^{-p}$ and $s^{1-p} \le r_{\mathcal{R}}^{1-p}$, since $p > 1$. Taking the larger of the constants from the two cases proves (13)–(34). □

Proof of Corollary F.1. All norms and inner products in this proof are Frobenius. Fix $x, y \in \mathrm{St}(n, r)$ and $\xi \in \mathrm{T}_x\mathrm{St}(n, r)$, and put $x^+ = \mathcal{R}_x^{\mathrm{pol}}(\xi)$. Lemma B.1 gives

$$\left\|x^+ - y\right\|^2 - \|x - y\|^2 - 2\langle \xi, x - y\rangle \le \|\xi\|^2 .$$

Taking $y = x$ in the same lemma gives $\|x^+ - x\| \le \|\xi\|$. Since x, x^+, y lie on a manifold of diameter $D_{\mathcal{M}}$, the reverse triangle inequality yields

$$\begin{aligned}\left|\left\|x^+ - y\right\|^2 - \|x - y\|^2\right| &\le \left(\left\|x^+ - y\right\| + \|x - y\|\right)\left\|x^+ - x\right\| \\ &\le 2D_{\mathcal{M}} \|\xi\| .\end{aligned}$$

Cauchy–Schwarz therefore gives a second upper bound,

$$\left\|x^+ - y\right\|^2 - \|x - y\|^2 - 2\langle \xi, x - y\rangle \le 4D_{\mathcal{M}} \|\xi\| .$$

Combining the two bounds, and using $1 < p \le 2$, gives

$$\left\|x^+ - y\right\|^2 - \|x - y\|^2 \le 2\langle \xi, x - y\rangle + \min\{\|\xi\|^2, 4D_{\mathcal{M}} \|\xi\|\} \le 2\langle \xi, x - y\rangle + (4D_{\mathcal{M}})^{2-p} \|\xi\|^p . \tag{37}$$

For completeness, the last inequality follows by splitting at $\|\xi\| = 4D_{\mathcal{M}}$: below that value, $\|\xi\|^2 \le (4D_{\mathcal{M}})^{2-p} \|\xi\|^p$; above it, $4D_{\mathcal{M}} \|\xi\| \le (4D_{\mathcal{M}})^{2-p} \|\xi\|^p$. The zero-step case also satisfies the inequality. □

Proof of Theorem F.2. Now condition on $\mathcal{F}_{t-1}$ and take $y = \widehat{x}(x_t)$ and $\xi = -\eta\widetilde{v}_t$. Proximal minimality at x_{t+1} and (13) imply

$$\Phi_{\bar\rho}(x_{t+1}) - \Phi_{\bar\rho}(x_t) \le \bar\rho\eta\langle \widetilde{v}_t, \widehat{x}(x_t) - x_t\rangle + \frac{\bar\rho}{2} C_{\mathcal{R},p}\eta^p \|\widetilde{v}_t\|^p .$$

The proximal selection is measurable with respect to the history. Assumption 3.2, conditional Minkowski's inequality, and $\|v_t\| \le G$ yield

$$\left(\mathbb{E}_t \|\widetilde{v}_t\|^p\right)^{1/p} \le \|v_t\| + \left(\mathbb{E}_t \|\widetilde{v}_t - v_t\|^p\right)^{1/p} \le G + \sigma.$$

In particular, the first moment exists, so conditional unbiasedness can be used in the inner product. Restricted weak convexity and proximal minimality give

$$\begin{aligned}\langle v_t, \widehat{x}(x_t) - x_t\rangle &\le f(\widehat{x}(x_t)) - f(x_t) + \frac{q_{\mathcal{M}}}{2} \|\widehat{x}(x_t) - x_t\|^2 \\ &\le -\frac{\bar\rho - q_{\mathcal{M}}}{2} \|\widehat{x}(x_t) - x_t\|^2 .\end{aligned}$$

Consequently,

$$\mathbb{E}_t\Phi_{\bar\rho}(x_{t+1}) - \Phi_{\bar\rho}(x_t) \le -\frac{\eta(\bar\rho - q_{\mathcal{M}})}{2\bar\rho}\Theta_{\bar\rho}(x_t)^2 + \frac{\bar\rho}{2} C_{\mathcal{R},p}\eta^p(G + \sigma)^p.$$

All terms are integrable: f is continuous and finite on the compact manifold, the proximal value is bounded there, and the oracle has a finite p-th moment. Taking total expectations, summing over $t = 1, \ldots, T$, and using $\Phi_{\bar\rho}(x_{T+1}) \ge f_\star$ gives (35). Substitution of the stated stepsize gives (36). If $G + \sigma = 0$, the oracle is zero almost surely and $0 \in \partial_{\mathcal{M}} f(x)$ everywhere; restricted weak convexity and $\bar\rho > q_{\mathcal{M}}$ then force $\widehat{x}(x) = x$, so the stationarity measure is identically zero. □

Lemma F.3 (Proximal-value increments)**.** *Under the assumptions of Theorem F.2, write* $D_{\mathcal{M}} = \operatorname{diam}(\mathcal{M})$ *and* $Z_t = \Phi_{\bar\rho}(x_{t+1}) - \mathbb{E}_t \Phi_{\bar\rho}(x_{t+1})$*. For every* $x, z \in \mathcal{M}$*,*

$$f_\star \le \Phi_{\bar\rho}(x) \le f_\star + \frac{\bar\rho D_{\mathcal{M}}^2}{2}, \tag{38}$$

$$|\Phi_{\bar\rho}(x) - \Phi_{\bar\rho}(z)| \le \bar\rho D_{\mathcal{M}} \|x - z\|. \tag{39}$$

Consequently, $\mathbb{E}_t Z_t = 0$ *and, almost surely,*

$$|Z_t| \le \frac{\bar\rho D_{\mathcal{M}}^2}{2}, \qquad \mathbb{E}_t Z_t^2 \le \bar\rho^2 D_{\mathcal{M}}^2 C_{\mathcal{R},p} \eta^p (G+\sigma)^p. \tag{40}$$

Proof. Every value minimized in (8) is at least $f_\star$. Choose a global minimizer $y_\star \in \mathcal{M}$ of f. Using it as a competitor gives

$$\Phi_{\bar\rho}(x) \le f_\star + \frac{\bar\rho}{2} \|y_\star - x\|^2 \le f_\star + \frac{\bar\rho D_{\mathcal{M}}^2}{2},$$

which proves (38). Next use a minimizer $\widehat{x}(z)$ in the proximal problem at z as a competitor at x. The reverse triangle inequality gives

$$\begin{aligned}\Phi_{\bar\rho}(x) - \Phi_{\bar\rho}(z) &\le \frac{\bar\rho}{2}\left(\|\widehat{x}(z) - x\|^2 - \|\widehat{x}(z) - z\|^2\right) \\ &\le \frac{\bar\rho}{2}\left(\|\widehat{x}(z) - x\| + \|\widehat{x}(z) - z\|\right)\|x - z\| \le \bar\rho D_{\mathcal{M}} \|x - z\|.\end{aligned}$$

Interchanging x and z proves (39).

Both $\Phi_{\bar\rho}(x_{t+1})$ and its conditional expectation lie in the interval in (38). Subtracting them gives the bound on $|Z_t|$, and centering gives $\mathbb{E}_t Z_t = 0$. Since $\Phi_{\bar\rho}(x_t)$ is history measurable, conditional variance minimization and (39) imply

$$\begin{aligned}\mathbb{E}_t Z_t^2 &\le \mathbb{E}_t \left(\Phi_{\bar\rho}(x_{t+1}) - \Phi_{\bar\rho}(x_t)\right)^2 \\ &\le \bar\rho^2 D_{\mathcal{M}}^2 \mathbb{E}_t \|x_{t+1} - x_t\|^2.\end{aligned}$$

Set $y = x_t$ and $\xi = -\eta\widetilde{v}_t$ in (13). The affine term vanishes, giving

$$\|x_{t+1} - x_t\|^2 \le C_{\mathcal{R},p} \eta^p \|\widetilde{v}_t\|^p.$$

Conditional Minkowski gives $\mathbb{E}_t \|\widetilde{v}_t\|^p \le (G+\sigma)^p$ as in the expectation proof. Substitution proves the conditional variance bound. □

For any constant stepsize $\eta > 0$, the high-probability finite-horizon bound used in Theorem 5.2 is

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \Theta_{\bar\rho}(x_t)^2 \le \frac{2\bar\rho}{\bar\rho - q_{\mathcal{M}}}\Bigg[&\frac{\Delta}{\eta T} + \frac{\bar\rho}{2} C_{\mathcal{R},p}(G+\sigma)^p \eta^{p-1} \\ &+ \bar\rho D_{\mathcal{M}} \sqrt{\frac{2C_{\mathcal{R},p}(G+\sigma)^p \eta^{p-2}\ell}{T}} + \frac{\bar\rho D_{\mathcal{M}}^2 \ell}{3\eta T}\Bigg].\end{aligned} \tag{41}$$

where $\ell = \log(1/\delta)$.

Proof of Theorem 5.2. The case $D_{\mathcal{M}} = 0$ is immediate because all iterates and proximal points coincide. Suppose $D_{\mathcal{M}} > 0$. Lemma F.3 makes $\sum_{t=1}^{k} Z_t$ a martingale with increments bounded by $b = \bar\rho D_{\mathcal{M}}^2/2$ and predictable quadratic variation bounded by

$$V = T\bar\rho^2 D_{\mathcal{M}}^2 C_{\mathcal{R},p} \eta^p (G+\sigma)^p.$$

The scalar Freedman inequality (Tropp, 2011, Theorem 1.1) implies, for $\ell = \log(1/\delta)$,

$$\sum_{t=1}^{T} Z_t \le \sqrt{2V\ell} + \frac{2b\ell}{3} = \bar\rho D_{\mathcal{M}} \sqrt{2TC_{\mathcal{R},p}\eta^p (G+\sigma)^p \ell} + \frac{\bar\rho D_{\mathcal{M}}^2 \ell}{3} \tag{42}$$

with probability at least $1-\delta$. Specifically, the inequality gives an upper-tail bound $\exp\{-u^2/[2(V+bu/3)]\}$; setting $u = \sqrt{2V\ell} + 2b\ell/3$ makes the exponent at most $-\ell$. If $V = 0$, the martingale increments vanish almost surely and (42) holds directly.

The conditional descent estimate in the proof of Theorem F.2 and the definition of Z_t give the pathwise inequality

$$\Phi_{\bar\rho}(x_{t+1}) - \Phi_{\bar\rho}(x_t) \le -\frac{\eta(\bar\rho - q_{\mathcal{M}})}{2\bar\rho}\Theta_{\bar\rho}(x_t)^2 + \frac{\bar\rho}{2}C_{\mathcal{R},p}\eta^p(G+\sigma)^p + Z_t.$$

Summing and using $\Phi_{\bar\rho}(x_{T+1}) \ge f_\star$ yields

$$\frac{\eta(\bar\rho - q_{\mathcal{M}})}{2\bar\rho}\sum_{t=1}^{T}\Theta_{\bar\rho}(x_t)^2 \le \Delta + \frac{\bar\rho}{2}TC_{\mathcal{R},p}\eta^p(G+\sigma)^p + \sum_{t=1}^{T}Z_t.$$

Insert (42) and divide by $\eta T(\bar\rho - q_{\mathcal{M}})/(2\bar\rho)$ to obtain (41).

For $G + \sigma > 0$, substitution of the stated stepsize gives the more explicit rate

$$\frac{1}{T}\sum_{t=1}^{T}\Theta_{\bar\rho}(x_t)^2 \le \frac{2\bar\rho(G+\sigma)}{\bar\rho - q_{\mathcal{M}}}T^{-(p-1)/p}\Bigg[\frac{\Delta}{\eta_0} + \frac{\bar\rho}{2}C_{\mathcal{R},p}\eta_0^{p-1} + \bar\rho D_{\mathcal{M}}\sqrt{2C_{\mathcal{R},p}\eta_0^{p-2}\ell} + \frac{\bar\rho D_{\mathcal{M}}^2\ell}{3\eta_0}\Bigg].$$

Thus one admissible constant in (14) is

$$C_{\mathrm{raw}} = \frac{2\bar\rho}{\bar\rho - q_{\mathcal{M}}}\max\left\{\frac{\Delta}{\eta_0} + \frac{\bar\rho}{2}C_{\mathcal{R},p}\eta_0^{p-1},\ \bar\rho D_{\mathcal{M}}\sqrt{2C_{\mathcal{R},p}\eta_0^{p-2}},\ \frac{\bar\rho D_{\mathcal{M}}^2}{3\eta_0}\right\}. \tag{43}$$

Requiring the right-hand side of (14) to be at most ε^2 proves the oracle complexity claim. When $G + \sigma = 0$, the zero-oracle argument from Theorem F.2 gives zero residual at every iteration. □

G Clipping and concentration lemmas

Lemma G.1 (Clipping decomposition)**.** *Let $\widetilde{v} = v + \xi$ with $\|v\| \le G$, $\mathbb{E}[\xi \mid \mathcal{F}] = 0$, and $\mathbb{E}[\|\xi\|^p \mid \mathcal{F}] \le \sigma^p$. If $g = \mathrm{clip}_\Lambda(\widetilde{v})$ and $\Lambda \ge 2G$, then the decomposition $g - v = u + b$, with $u = g - \mathbb{E}[g \mid \mathcal{F}]$ and $b = \mathbb{E}[g \mid \mathcal{F}] - v$, satisfies* (31)*. Moreover,*

$$\mathbb{E}[\|g\|^2 \mid \mathcal{F}] \le M_2 := 2G^2 + c_2\sigma^p\Lambda^{2-p}. \tag{44}$$

Proof. The clipping map is the Euclidean projection onto the ball of radius Λ. Because $\|v\| \le G \le \Lambda/2$, clipping occurs only if $\|\xi\| \ge \Lambda/2$. On this event,

$$\|g - \widetilde{v}\| \le \|\widetilde{v}\| \le G + \|\xi\| \le 2\|\xi\|.$$

The pointwise bound $u\mathbf{1}_{\{u\ge a\}} \le a^{1-p}u^p$ for $u \ge 0$, $a > 0$, gives

$$\begin{aligned}\|b\| = \|\mathbb{E}[g - \widetilde{v} \mid \mathcal{F}]\| &\le 2\mathbb{E}[\|\xi\|\ \mathbf{1}_{\{\|\xi\|\ge\Lambda/2\}} \mid \mathcal{F}] \\ &\le 2(\Lambda/2)^{1-p}\mathbb{E}[\|\xi\|^p \mid \mathcal{F}] \le 4\sigma^p\Lambda^{1-p}.\end{aligned}$$

Also, projection nonexpansiveness gives $\|g - v\| \le \|\xi\|$ when $\|\xi\| < \Lambda/2$, while always $\|g - v\| \le \Lambda + G \le 3\Lambda/2$. We obtain

$$\begin{aligned}\mathbb{E}[\|g - v\|^2 \mid \mathcal{F}] &\le \mathbb{E}[\|\xi\|^2\mathbf{1}_{\{\|\xi\|<\Lambda/2\}} \mid \mathcal{F}] + \frac{9\Lambda^2}{4}\mathbb{P}(\|\xi\| \ge \Lambda/2 \mid \mathcal{F}) \\ &\le (\Lambda/2)^{2-p}\sigma^p + \frac{9\Lambda^2}{4}(2/\Lambda)^p\sigma^p \le 10\sigma^p\Lambda^{2-p}.\end{aligned}$$

Conditional variance is minimized by the conditional mean, so $\mathbb{E}\|u\|^2 \le \mathbb{E}\|g - v\|^2 \le c_2\sigma^p\Lambda^{2-p}$ for $c_2 = 32$. Since $\|g\| \le \Lambda$, $\|u\| \le \|g\| + \|\mathbb{E}g\| \le 2\Lambda$. The bound $\|g\|^2 \le 2\|v\|^2 + 2\|g - v\|^2$ then proves (44) after enlarging c_2 if needed. □

Lemma G.2 (Bounded-increment concentration)**.** *Let $(X_t)_{t=1}^T$ be real martingale differences with $|X_t| \le b$ almost surely and $\sum_{t=1}^T \mathbb{E}_t X_t^2 \le V$ almost surely, where b, V are deterministic nonnegative constants. Then, for every $\ell > 0$,*

$$\Pr\left\{\sum_{t=1}^T X_t > \sqrt{2V\ell} + \frac{b\ell}{3}\right\} \le e^{-\ell}.$$

Proof. If $V = 0$ or $b = 0$, all increments vanish almost surely. Otherwise, for $0 < \lambda < 3/b$, the exponential series, conditional centering, and $k! \ge 2 \cdot 3^{k-2}$ for $k \ge 2$ give

$$\begin{aligned}\mathbb{E}_t e^{\lambda X_t} &\le 1 + \mathbb{E}_t X_t^2 \sum_{k=2}^{\infty} \frac{\lambda^k b^{k-2}}{k!} \\ &\le \exp\left\{\frac{\lambda^2 \mathbb{E}_t X_t^2}{2(1-\lambda b/3)}\right\}.\end{aligned}$$

Consequently,

$$\exp\left\{\lambda \sum_{t=1}^k X_t - \frac{\lambda^2}{2(1-\lambda b/3)} \sum_{t=1}^k \mathbb{E}_t X_t^2\right\}, \quad 0 \le k \le T,$$

is a nonnegative supermartingale with initial value one. The deterministic variance bound therefore yields

$$\mathbb{E}\exp\left(\lambda \sum_{t=1}^T X_t\right) \le \exp\left\{\frac{\lambda^2 V}{2(1-\lambda b/3)}\right\}.$$

Markov's inequality bounds the upper tail at u by $\exp\{-\lambda u + \lambda^2 V/[2(1-\lambda b/3)]\}$. Set

$$s = \sqrt{2\ell/V}, \qquad \lambda = \frac{s}{1 + (b/3)s}, \qquad u = \sqrt{2V\ell} + \frac{b\ell}{3}.$$

Then $0 < \lambda < 3/b$ and direct substitution makes the exponent $-\ell$, proving the bound. □

Lemma G.3 (Two scalar concentration bounds)**.** *Let d_t be $\mathcal{F}_{t-1}$-measurable with $\|d_t\| \le D_{\mathrm{p}}$. Under Lemma G.1, with probability at least $1 - \delta/2$,*

$$\frac{1}{T}\sum_{t=1}^T \langle d_t, u_t\rangle \le \frac{2D_{\mathrm{p}}\Lambda\ell}{3T} + D_{\mathrm{p}}\sqrt{\frac{2c_2\sigma^p\Lambda^{2-p}\ell}{T}}. \tag{45}$$

With probability at least $1 - \delta/2$,

$$\frac{1}{T}\sum_{t=1}^T \|g_t\|^2 \le 2M_2 + \frac{2\Lambda^2\ell}{T}. \tag{46}$$

Here $\ell = \log(4/\delta)$.

Proof. For $X_t = \langle d_t, u_t\rangle$, we have $\mathbb{E}_t X_t = 0$, $|X_t| \le 2D_{\mathrm{p}}\Lambda$, and

$$\sum_{t=1}^T \mathbb{E}_t X_t^2 \le c_2 D_{\mathrm{p}}^2 \sigma^p \Lambda^{2-p} T.$$

Lemma G.2 with $b = 2D_{\mathrm{p}}\Lambda$ and $V = c_2 D_{\mathrm{p}}^2 \sigma^p \Lambda^{2-p} T$ gives (45), with failure probability at most $\delta/4$.

For $Z_t = \|g_t\|^2$, $0 \le Z_t \le \Lambda^2$ and $\mathbb{E}_t Z_t \le M_2$. Scalar Bernstein/Freedman applied to $Z_t - \mathbb{E}_t Z_t$ gives

$$\frac{1}{T}\sum Z_t \le M_2 + \sqrt{\frac{2\Lambda^2 M_2 \ell}{T}} + \frac{2\Lambda^2\ell}{3T}.$$

Using $\sqrt{2ab} \le a + b/2$ and enlarging constants yields (46). □

H Proof of Theorem D.2

Proof. Apply Lemma D.1 with $e_t = u_t + b_t$ and sum from $t = 1$ to T. Since $\Phi_{\bar\rho}(x_{T+1}) \geq f_\star$,

$$\frac{\eta(\bar\rho - q_{\mathcal{M}})}{2\bar\rho} \sum_{t=1}^{T} \Theta_{\bar\rho}(x_t)^2 \leq \Delta + \bar\rho\eta \sum_{t=1}^{T} \langle u_t, \widehat{x}_t - x_t\rangle + \bar\rho\eta \sum_{t=1}^{T} \langle b_t, \widehat{x}_t - x_t\rangle + \frac{\bar\rho(1+\kappa_{\mathcal{R}})}{2}\eta^2 \sum_{t=1}^{T} \|g_t\|^2 .$$

By (29), $\|\widehat{x}_t - x_t\| \leq D_{\mathrm{p}}$. The clipping-bias term is at most $TD_{\mathrm{p}}c_b\sigma^p\Lambda^{1-p}$. Apply Lemma G.3 to the martingale and squared-gradient sums, use a union bound, divide by $\eta T(\bar\rho - q_{\mathcal{M}})/(2\bar\rho)$. The factor $1/2$ in the quadratic remainder cancels the factor two in (46), which gives (32). □

I Proof of the normalized-method bound

Lemma I.1 (Mini-batch deviation)**.** *Let $\xi_1, \dots, \xi_B$ be conditionally independent tangent vectors with $\mathbb{E}_t\xi_j = 0$ and $\mathbb{E}_t \|\xi_j\|^p \leq \sigma^p$ for $p \in (1, 2]$. Then*

$$\mathbb{E}_t \left\| B^{-1} \textstyle\sum_{j=1}^{B} \xi_j \right\| \leq 2\sigma B^{-(p-1)/p}. \tag{47}$$

Proof. The scalar von Bahr–Esseen inequality (von Bahr & Esseen, 1965) implies that, for conditionally independent centered vectors in a Euclidean space and $p \in [1, 2]$,

$$\mathbb{E}_t \left\| \textstyle\sum_{j=1}^{B} \xi_j \right\|^p \leq 2 \sum_{j=1}^{B} \mathbb{E}_t \|\xi_j\|^p .$$

To obtain the vector form, let z be an independent standard Gaussian vector and $a_p = \mathbb{E}|z_1|^p$, then $\mathbb{E}_z |\langle z, w\rangle|^p = a_p \|w\|^p$. Condition on z, apply the scalar von Bahr–Esseen inequality to $\langle z, \xi_j\rangle$, and average over z; the factor a_p cancels. The argument is conditional on the history; the batch vectors retain independence and zero means under this conditioning. Jensen's inequality gives

$$\mathbb{E}_t \left\| B^{-1} \textstyle\sum_j \xi_j \right\| \leq B^{-1}(2B\sigma^p)^{1/p} \leq 2\sigma B^{-(p-1)/p}.$$

□

Lemma I.2 (Normalization stability)**.** *For $b \neq 0$ and $\mathrm{nrm}(0) = 0$,*

$$\|\mathrm{nrm}(a) - \mathrm{nrm}(b)\| \leq 2 \min\left\{1, \frac{\|a-b\|}{\|b\|}\right\}. \tag{48}$$

Proof. The left-hand side is at most two. If $a \neq 0$, the triangle and reverse triangle inequalities give

$$\left\| \frac{a}{\|a\|} - \frac{b}{\|b\|} \right\| \leq \left\| \frac{a}{\|a\|} - \frac{a}{\|b\|} \right\| + \frac{\|a-b\|}{\|b\|} = \frac{|\,\|b\| - \|a\|\,|}{\|b\|} + \frac{\|a-b\|}{\|b\|} \leq \frac{2\,\|a-b\|}{\|b\|}.$$

If $a = 0$, the left-hand side is one and $\|a - b\| = \|b\|$, so the same bound holds. □

Proof of Theorem D.3. Write $\widehat{x}_t = \widehat{x}(x_t)$, $d_t = x_t - \widehat{x}_t$, and $c = (\bar\rho - q_{\mathcal{M}})/2$. Restricted weak convexity and proximal minimality imply

$$\langle v_t, d_t\rangle \geq c\,\|d_t\|^2 . \tag{49}$$

If $d_t \neq 0$, this also gives $\|v_t\| \geq c\,\|d_t\|$. Since $\|v_t\| \leq G$,

$$\langle \mathrm{nrm}(v_t), d_t\rangle \geq \frac{c}{G}\,\|d_t\|^2 . \tag{50}$$

The same inequality holds when $d_t = 0$. The alignment bound (49) also forces $d_t = 0$ when $v_t = 0$. Let $h_t = \mathrm{nrm}(\bar{v}_t)$ and

$$q_t = \langle \mathrm{nrm}(v_t) - h_t, d_t \rangle .$$

By Lemma I.2, (29), and (49),

$$|q_t| \le 2D_{\mathrm{p}}, \qquad q_t \le \frac{2}{c} \|\bar{v}_t - v_t\| . \tag{51}$$

Lemma I.1 therefore yields

$$\mathbb{E}_t q_t \le \frac{4\sigma}{c} B^{-(p-1)/p}. \tag{52}$$

Use $\widehat{x}_t$ as a competitor in the proximal value at x_{t+1} and apply the retraction distance bound to $-\alpha h_t$. Because $\|h_t\| \le 1$ and $\alpha \le r_{\mathcal{R}}$,

$$\Phi_{\bar{\rho}}(x_{t+1}) - \Phi_{\bar{\rho}}(x_t) \le -\bar{\rho}\alpha \langle h_t, d_t \rangle + \frac{\bar{\rho}(1+\kappa_{\mathcal{R}})}{2}\alpha^2 \tag{53}$$

$$\le -\frac{\alpha c}{\bar{\rho} G}\Theta_{\bar{\rho}}(x_t)^2 + \bar{\rho}\alpha q_t + \frac{\bar{\rho}(1+\kappa_{\mathcal{R}})}{2}\alpha^2. \tag{54}$$

The differences $q_t - \mathbb{E}_t q_t$ form a martingale difference sequence and have absolute value at most $4D_{\mathrm{p}}$. Azuma's inequality and (52) show that, with probability at least $1-\delta$,

$$\frac{1}{T}\sum_{t=1}^{T} q_t \le \frac{4\sigma}{c} B^{-(p-1)/p} + 4D_{\mathrm{p}}\sqrt{\frac{2\log(1/\delta)}{T}}. \tag{55}$$

Sum (54), use $\Phi_{\bar{\rho}}(x_{T+1}) \ge f_\star$, divide by $\alpha c T/(\bar{\rho} G)$, and substitute $c = (\bar{\rho} - q_{\mathcal{M}})/2$ and $D_{\mathrm{p}} = 2G/(\bar{\rho} - q_{\mathcal{M}})$ to obtain (33). □

Proof of Theorem 5.4. The bound in (33) implies (17) with

$$C_{\mathrm{n}} = \frac{2\bar{\rho} G}{\bar{\rho} - q_{\mathcal{M}}} \max\left\{1, \frac{\bar{\rho}(1+\kappa_{\mathcal{R}})}{2}, \frac{8\bar{\rho}}{\bar{\rho} - q_{\mathcal{M}}}, \frac{8\sqrt{2}\bar{\rho} G}{\bar{\rho} - q_{\mathcal{M}}}\right\}.$$

Each coefficient in (33) is at most C_{n}, and all four terms are nonnegative. In particular, $B = T$ gives

$$\frac{1}{T}\sum_{t=1}^{T}\Theta_{\bar{\rho}}(x_t)^2 = O\left(\frac{1+\sqrt{\log(1/\delta)}}{\sqrt{T}} + \frac{\sigma}{T^{(p-1)/p}}\right) \tag{56}$$

for the stated stepsize. Substitute $B = T$ and $\alpha = \alpha_0 T^{-1/2}$ into (33). The initialization and quadratic-retraction terms are proportional to $T^{-1/2}$, the mini-batch term is $\sigma T^{-(p-1)/p}$, and the martingale term is $T^{-1/2}\sqrt{\log(1/\delta)}$. Each of the T iterations uses $B = T$ oracle calls, so the total is T^2. □

J RATE SIMPLIFICATION

Corollary J.1 (Single-sample schedules without tail parameters)**.** *Under the respective problem and retraction assumptions of Theorems 5.2 and D.2, use $\eta_T = \eta_0 T^{-1/2}$ for RSSG, or $\Lambda_T = \max\{2G, \sqrt{T}\}$, $\eta_T = \eta_0/\Lambda_T$ for GeoClip. Here $\eta_0 > 0$, with $\eta_0 \le r_{\mathcal{R}}$ for GeoClip. Neither schedule uses p or σ. With probability $1-\delta$,*

$$\frac{1}{T}\sum_{t=1}^{T}\Theta_{\bar{\rho}}(x_t)^2 \le C\big(1 + \sqrt{\log(4/\delta)} + \log(4/\delta)\big)T^{-(p-1)/2},$$

where C may depend on p, σ and problem parameters, but not T, δ. Thus $Q = T = \widetilde{O}(\varepsilon^{-4/(p-1)})$ suffices.

Proof of Theorem 5.3. Let

$$\psi_T := \max\left\{\sigma T^{-(p-1)/p}, GT^{-1/2}\right\}.$$

The schedule (15) satisfies

$$\frac{1}{\eta_T T} \le \frac{1}{\eta_0}\max\left\{\frac{\Lambda_T}{T}, \frac{G}{\sqrt{T}}\right\} \le \frac{2}{\eta_0}\psi_T, \qquad \sigma^p \Lambda_T^{1-p} \le \psi_T,$$

$$\sqrt{\frac{\sigma^p \Lambda_T^{2-p}}{T}} \le 2\psi_T, \qquad \frac{\Lambda_T}{T} \le 2\psi_T,$$

$$\eta_T G^2 \le \eta_0 \psi_T, \qquad \eta_T \sigma^p \Lambda_T^{2-p} \le \eta_0 \psi_T,$$

$$\eta_T \frac{\Lambda_T^2}{T} \le 2\eta_0 \psi_T.$$

Substitution in (32) proves (16); one admissible choice is

$$C_{\mathrm{geo}} = \frac{8\bar{\rho}}{\bar{\rho} - q_{\mathcal{M}}}\left[\frac{\Delta}{\eta_0} + \bar{\rho} D_{\mathrm{p}}(1 + c_b + \sqrt{2c_2}) + \bar{\rho}(1 + \kappa_{\mathcal{R}})\eta_0(2 + c_2)\right]. \tag{57}$$

Requiring both components of ψ_T to be at most a constant multiple of ε^2 yields the complexity in Theorem 5.3. □

J.1 Proof of Corollary J.1

Proof. For RSSG, substituting $\eta_T = \eta_0 T^{-1/2}$ into (41), and writing $\ell = \log(4/\delta)$, gives

$$\frac{1}{T}\sum_{t=1}^{T} \Theta_{\bar{\rho}}(x_t)^2 \le \frac{2\bar{\rho}}{\bar{\rho} - q_{\mathcal{M}}}\left[\frac{\Delta}{\eta_0} T^{-1/2} + \frac{\bar{\rho}}{2} C_{\mathcal{R},p}(G + \sigma)^p \eta_0^{p-1} T^{-(p-1)/2} \right.$$
$$\left. + \bar{\rho} D_{\mathcal{M}} \sqrt{2C_{\mathcal{R},p}(G + \sigma)^p \eta_0^{p-2}}\, T^{-p/4}\sqrt{\ell} + \frac{\bar{\rho} D_{\mathcal{M}}^2}{3\eta_0} T^{-1/2}\ell\right].$$

Replacing $\log(1/\delta)$ by the larger ℓ only enlarges this bound. Since $(p-1)/2 \le p/4 \le 1/2$ and $T \ge 1$, this proves the claimed rate for RSSG.

For GeoClip, $\eta_T \Lambda_T = \eta_0 \le r_{\mathcal{R}}$ and $\Lambda_T \ge 2G$, so Theorem D.2 applies. Set $K = \max\{2G, 1\}$, so that $\sqrt{T} \le \Lambda_T \le K\sqrt{T}$. The terms in (32) satisfy

$$\frac{\Delta}{\eta_T T} \le \frac{\Delta K}{\eta_0} T^{-1/2}, \qquad \sigma^p \Lambda_T^{1-p} \le \sigma^p T^{-(p-1)/2},$$

$$\frac{\Lambda_T \ell}{T} \le K T^{-1/2}\ell, \qquad \sqrt{\frac{\sigma^p \Lambda_T^{2-p}\ell}{T}} \le \sigma^{p/2} K^{(2-p)/2} T^{-p/4}\sqrt{\ell},$$

$$\eta_T M_2 \le 2\eta_0 G^2 T^{-1/2} + c_2 \eta_0 \sigma^p T^{-(p-1)/2}, \qquad \eta_T \frac{\Lambda_T^2 \ell}{T} \le \eta_0 K T^{-1/2}\ell.$$

Substitution into (32) and the same comparison of exponents give the stated GeoClip bound. For either method, requiring its right-hand side to be at most ε^2 yields the oracle complexity because $Q = T$. □

K Retraction distance examples

Proposition K.1 (Local-radius exceedance under finite moments)**.** *On a positive-dimensional manifold, fix $p \in (1, 2]$, $\sigma > 0$, a stepsize $\eta > 0$, a radius $r > 0$, and a horizon T. For the constant objective $f \equiv 0$ and any $B > \max\{\sigma, r/\eta\}$, there is a conditionally unbiased tangent oracle with independent spike magnitudes and signs and conditional p-th central moment σ^p such that*

$$\mathbb{P}\left(\max_{1 \le t \le T} \eta \,\|\widetilde{v}_t\| > r\right) = 1 - \left(1 - \frac{\sigma^p}{B^p}\right)^T. \tag{58}$$

A finite moment assumption alone therefore cannot ensure that all steps stay within the radius with a failure probability below the right-hand side.

To see this, condition on the history and choose a unit tangent direction. Multiply this direction by a scalar taking the values $+B$ and $-B$, each with probability $\sigma^p/(2B^p)$, and zero otherwise, with independent scalar draws across iterations. Then $\mathbb{E}_t\widetilde{v}_t = 0 \in \partial_{\mathcal{M}} f(x_t)$ and $\mathbb{E}_t \|\widetilde{v}_t\|^p = \sigma^p$. Since $\eta B > r$, a radius exceedance occurs exactly when the oracle is nonzero. Independence of these events gives (58). Clipping then serves two purposes: it controls the oracle and keeps the update within the retraction's local validity radius.

K.1 Distance errors for QR and second-order retractions

The defect already occurs for standard retractions. For a full-rank matrix Z, let $\mathrm{qf}(Z)$ denote the thin-QR factor whose triangular factor has positive diagonal; the standard QR retraction is $\mathcal{R}^{\mathrm{qr}}_X(\Xi) = \mathrm{qf}(X + \Xi)$.

Proposition K.2 (The QR retraction has quadratic positive defect)**.** *For the standard thin-QR retraction on $\mathrm{St}(3,2)$, there are $X, Y_t \in \mathrm{St}(3,2)$ and tangent steps $\Xi_t \in \mathrm{T}_X\mathrm{St}(3,2)$ with $\|\Xi_t\|_F \to 0$ such that*

$$\|\mathcal{R}^{\mathrm{qr}}_X(\Xi_t) - Y_t\|_F^2 - \|X + \Xi_t - Y_t\|_F^2 = 2(\sqrt{2} - 1)t^2 + O(t^3) > 0. \tag{59}$$

In particular, the positive defect is $\Theta(\|\Xi_t\|_F^2)$.

The same retraction can change descent into ascent for an objective in our problem class.

Proposition K.3 (QR can reverse an affine descent step)**.** *Let $X = [e_1, e_2] \in \mathrm{St}(3,2)$ and $\Xi_t = te_3[1,1]$ for $t > 0$. There is an ambient linear function $f_t(Z) = \langle C_t, Z\rangle_F$, satisfying Assumption 3.1 with $\rho = 0$ and $G = \|C_t\|_F$, such that*

$$f_t(X + \Xi_t) < f_t(X) < f_t(\mathcal{R}^{\mathrm{qr}}_X(\Xi_t)). \tag{60}$$

Moreover, for every $\eta > 0$, $\pi \in (0,1)$, and $p \in (1,2]$, there is an unbiased finite-support tangent oracle for f_t at X whose QR update equals $\mathcal{R}^{\mathrm{qr}}_X(\Xi_t)$ with probability π and therefore increases f_t on that event.

Proposition K.3 shows that retraction can reverse tangent-space descent even at arbitrarily small step norms. The quadratic positive distance defect in Proposition K.2 distinguishes QR from the global polar comparison in Lemma B.1. Positive distance error also occurs for a second-order retraction, as the following circle construction shows.

Proposition K.4 (Positive distance error on the circle)**.** *Let $\mathcal{M} = \mathbb{S}^1$, let J be the ninety-degree rotation, and write a tangent vector as sJx. The map*

$$\mathcal{R}_x(sJx) = \mathrm{Rot}(s + s^3)x \tag{61}$$

is a smooth second-order retraction. Nevertheless, for $y = x$ and every sufficiently small nonzero s,

$$\|\mathcal{R}_x(sJx) - x\|^2 > \|sJx\|^2 .$$

A generic second-order retraction therefore need not satisfy (10) *with $\kappa_{\mathcal{R}} = 0$.*

K.2 Proof of Proposition K.2

Proof. Let $X = [e_1, e_2] \in \mathrm{St}(3,2)$ and $\Xi_t = te_3[1,1]$. Since $X^\top \Xi_t = 0$, the perturbation is tangent. Put $Z_t = X + \Xi_t$, $a = (1 + t^2)^{1/2}$, and $b = (1 + 2t^2)^{1/2}$. Gram–Schmidt with positive diagonal gives

$$R_t := \mathcal{R}^{\mathrm{qr}}_X(\Xi_t) = \begin{bmatrix} a^{-1} & -t^2/(ab) \\ 0 & a/b \\ t/a & t/(ab) \end{bmatrix}.$$

For $A_t = R_t - Z_t$, Taylor expansion yields

$$A_t = t^2 A_0 + O(t^3), \qquad A_0 = \begin{bmatrix} -1/2 & -1 \\ 0 & -1/2 \\ 0 & 0 \end{bmatrix}.$$

The nonzero 2×2 block of A_0 has squared Frobenius norm $3/2$ and determinant $1/4$. Therefore, $\|A_0\|_*^2 = \|A_0\|_F^2 + 2|\det(A_0[1\!:\!2, :])| = 2$, so $\|A_t\|_* = \sqrt{2}t^2 + O(t^3)$.

The rectangular orthogonal Procrustes identity gives

$$\min_{Y \in \mathrm{St}(3,2)} \langle Y, A_t \rangle_F = -\|A_t\|_* .$$

Choose a minimizer Y_t. Since $\|R_t\|_F^2 = 2$ and $\|Z_t\|_F^2 = 2 + 2t^2$,

$$\begin{aligned} \|R_t - Y_t\|_F^2 - \|Z_t - Y_t\|_F^2 &= -2t^2 - 2\langle Y_t, A_t \rangle_F \\ &= 2(\sqrt{2} - 1)t^2 + O(t^3), \end{aligned}$$

which is positive for all sufficiently small $t > 0$. □

K.3 Proof of Proposition K.3

Proof. Fix $t > 0$, write $U = \Xi_t$ and $W = \mathcal{R}_X^{\mathrm{qr}}(U) - X$. The QR factor displayed in the preceding proof gives

$$\langle W, U \rangle_F = \frac{t^2}{\sqrt{1+t^2}} \left(1 + \frac{1}{\sqrt{1+2t^2}}\right) > 0. \tag{62}$$

Let

$$\gamma = \frac{\langle W, U \rangle_F}{\|U\|_F^2}, \qquad E = W - \gamma U, \qquad C = E - \frac{\|E\|_F^2}{2\langle W, U \rangle_F} U.$$

Then $\langle E, U \rangle_F = 0$, and $E \neq 0$: the $(1,1)$ entry of W is $(1+t^2)^{-1/2} - 1 < 0$, whereas every entry in the first row of U is zero. For $f_t(Z) = \langle C, Z \rangle_F$,

$$\begin{aligned} f_t(X+U) - f_t(X) &= \langle C, U \rangle_F = -\frac{\|E\|_F^2 \|U\|_F^2}{2\langle W, U \rangle_F} < 0, \\ f_t(\mathcal{R}_X^{\mathrm{qr}}(U)) - f_t(X) &= \langle C, W \rangle_F = \|E\|_F^2 - \frac{\|E\|_F^2}{2} = \frac{1}{2}\|E\|_F^2 > 0. \end{aligned}$$

This proves (60).

For the oracle construction, let $v = P_{\mathrm{T}_X \mathrm{St}(3,2)} C$, which is the Riemannian gradient of f_t at X. Set $g_1 = -U/\eta$ and $g_0 = (v - \pi g_1)/(1 - \pi)$. Both vectors are tangent and the random vector equal to g_1 with probability π and to g_0 otherwise has mean v. If $D = \|g_1 - v\|$, then for every $p \in (1, 2]$,

$$\mathbb{E}\|g - v\|^p = \left(\pi + \frac{\pi^p}{(1-\pi)^{p-1}}\right) D^p < \infty.$$

On the event $g = g_1$, the update is $\mathcal{R}_X^{\mathrm{qr}}(-\eta g) = \mathcal{R}_X^{\mathrm{qr}}(U)$, so the second strict inequality above proves the claim. □

K.4 Proof of Proposition K.4

Proof. For the curve $c(s) = \mathrm{Rot}(s + s^3)x$, we have $c(0) = x$, $c'(0) = Jx$, and $c''(0) = -x$. The last identity is the normal acceleration of the circle's exponential map, so (61) is second order. For $y = x$,

$$\begin{aligned} \|\mathcal{R}_x(sJx) - x\|^2 &= 2 - 2\cos(s + s^3) \\ &= s^2 + \frac{23}{12}s^4 + O(s^6), \end{aligned}$$

whereas $\|x + sJx - x\|^2 = s^2$. The strict inequality follows for all sufficiently small nonzero s. □

L ALGORITHM DETAILS

Algorithm 1 RSSG, GeoClip-RSSG, and GeoNorm-RSSG

Require: Deterministic initial point $x_1 \in \mathcal{M}$, integer horizon $T \geq 1$, tangent oracle satisfying Assumption 3.2, and method RSSG, GeoClip, or GeoNorm.
Require: A smooth retraction satisfying Definition 3.4 with constants $(r_{\mathcal{R}}, \kappa_{\mathcal{R}})$; it is globally defined for RSSG.
For RSSG, choose $\eta > 0$.
For GeoClip, use the bound $G > 0$ in Assumption 3.1 and choose $\Lambda_{\mathrm{stat}} \geq 2G$, $0 < \eta \leq r_{\mathcal{R}}/(2G)$. Set $\Lambda = \min\{\Lambda_{\mathrm{stat}}, r_{\mathcal{R}}/\eta\}$.
For GeoNorm, choose integer $B \geq 1$ and $0 < \alpha \leq r_{\mathcal{R}}$.
for $t = 1, \ldots, T$ **do**
  if method is RSSG **then**
    Query $\widetilde{v}_t \in \mathrm{T}_{x_t}\mathcal{M}$ and set $d_t = \eta\widetilde{v}_t$.
  else if method is GeoClip **then**
    Query $\widetilde{v}_t \in \mathrm{T}_{x_t}\mathcal{M}$.
    Set $d_t = \eta \min\{1, \Lambda/\left\|\widetilde{v}_t\right\|\}\widetilde{v}_t$ if $\widetilde{v}_t \neq 0$, and $d_t = 0$ otherwise.
  else
    Query B conditionally independent samples $\widetilde{v}_{t,1}, \ldots, \widetilde{v}_{t,B} \in \mathrm{T}_{x_t}\mathcal{M}$ with the same conditional mean.
    Compute $\bar{v}_t = B^{-1}\sum_{j=1}^{B}\widetilde{v}_{t,j}$. Set $d_t = \alpha\bar{v}_t/\left\|\bar{v}_t\right\|$ if $\bar{v}_t \neq 0$, and $d_t = 0$ otherwise.
  end if
  Set $x_{t+1} = \mathcal{R}_{x_t}(-d_t)$.
end for
return the trajectory $(x_t)_{t=1}^{T}$; optionally x_R for an independent uniform $R \in \{1, \ldots, T\}$.

The algorithm uses $Q = T$ oracle calls for RSSG and GeoClip and $Q = TB$ for GeoNorm. Parameters are constant during a run and may depend on T. Theorem 5.2 gives the RSSG choice using p, σ. For GeoClip, take $\Lambda_{\mathrm{stat}} = \Lambda_T$ and $\eta = \eta_T$ from Theorem 5.3; these satisfy the input conditions and give $\Lambda = \Lambda_T$. Corollary J.1 gives the single-sample choices independent of p, σ. For GeoNorm, Theorem 5.4 uses $B = T$ and $\alpha = \alpha_0/\sqrt{T}$, with $0 < \alpha_0 \leq r_{\mathcal{R}}$. The finite-horizon bound in Theorem D.3 applies to any fixed $B \geq 1$.

The high-probability theorems control $T^{-1}\sum_{t=1}^{T}\Theta_{\bar{\rho}}(x_t)^2$ along the random trajectory. For the optional random output, this average is $\mathbb{E}_R[\Theta_{\bar{\rho}}(x_R)^2 \mid x_1, \ldots, x_T]$.

M ADDITIONAL EXPERIMENTAL DETAILS

M.1 IMPLEMENTATION AND STATISTICAL PROTOCOL

The reported statistics comprise 14,110 confirmatory runs and 200 paired QR-ablation runs. The supplementary ledger records the complete run inventory, including eight incomplete runs outside the final comparison scope after the timing-based protocol adjustment described below.

The implementation uses PyTorch 2.4 with explicit tangent projections and QR, polar, sphere exponential, and circle retractions. Geometry and rate experiments use float64; DPCP and Hopkins155 use float32, with the final proximal-stationarity computation promoted to float64. The correctness suite checks tangency, Stiefel feasibility, CPU/GPU agreement, deterministic replay, the signs and asymptotic scales of the counterexamples, and the absence of a positive polar defect in 100 random trials. The largest float64 orthogonality error is 4.82×10^{-15} and the CPU/GPU relative discrepancy is below 1.53×10^{-15}.

The primary DPCP stepsizes were selected on separate validation seeds. Five screening seeds reduced each method to two candidates, and 20 further validation seeds selected between those candidates. Both stages use the Pareto $p = 1.5$, 50%-outlier setting at $(n, r, N) = (100, 5, 5000)$; the selected stepsizes transfer to all six primary settings. Confirmatory seeds start at 1000 and do not overlap either tuning stage. The reduction from the original primary matrix to the six settings in

Table 3: Experimental settings used for the reported results.

Study	Settings
Online rates	$\mathrm{St}(32,1)$ and $\mathrm{St}(32,4)$, $\|A\|_F = 1$, $\bar{\rho} = 4$, $T = 64, 128, 256, 512, 1024, 2048$; 50 seeds. Radial noise MAD $s = 0.25$.
QR ablation	$\mathrm{St}(32,4)$, $Q = 1024$, $\bar{\rho} = 10$, radial MAD 4, threshold 10, radius 0.3, $\eta\Lambda \in \{0.2, 1\}$; 50 seeds.
Noise parameters	Student-t degrees of freedom and Pareto exponent $p + 0.2$; stable index $p + 0.1$ for $p \leq 1.8$; Gaussian at $p = 2$.
Primary DPCP	$(n, r, N) = (100, 5, 5000)$; $Q = 262144$; 20 seeds; GeoNorm $B = T = 512$. Step parameters: RSSG/Clip 0.0075, $B = 1$: 0.0025, $B = 64$: 0.075, $B = T$: 0.15, MoM 0.25, full batch 2.4019223; clipping threshold 25.
Hopkins155	Motion rank 4; $Q = 16384$; GeoNorm $B = T = 128$. Step parameters: RSSG/Clip 0.03, $B = 1$: 0.01, $B = 64, T$: 0.3, MoM 1, full batch 10; clipping threshold 10.

Table 2 was made after pilot timing, before the confirmatory outcome metrics were inspected. The run ledger includes all partial results from the original matrix.

Every inferential comparison is paired. We use the median of the paired projector-error differences and a percentile interval from 10,000 bootstrap resamples. These 95% intervals are pointwise. Wilcoxon signed-rank tests are corrected by Holm's method over the full family of 70 method–setting comparisons, including the dimension-scaling settings. Table 5 displays 12 of these comparisons. The phase diagram uses Wilson intervals. Rate exponents are Theil–Sen slopes fitted to log median stationarity at all six horizons; their 95% slope intervals use the rank-based interval returned by `scipy.stats.theilslopes`. A run that produces a nonfinite value, times out, or exhausts its single retry is counted as a failure.

The online oracle adds noise formed by multiplying a unit tangent direction by a symmetric scalar amplitude. Directions come from 64 Gaussian matrices, projected at the current iterate, and normalized. The amplitude has median absolute deviation $s = 0.25$. The theoretical scale σ in Assumption 3.2 is the upper bound on the p-th moment raised to the power $1/p$. For example, the Gaussian amplitude standard deviation is $0.25/\Phi_{\mathrm{N}}^{-1}(0.75)$, where Φ_{N} is the standard normal distribution function. The rate runs use s as the clipping scale input. Their slope comparison evaluates the predicted powers of T. Finite-horizon constants and coverage as a function of δ are separate from this rate comparison. For DPCP, stored gradient norms are labeled as proxies. The additional smoothed proximal solves give numerical residual estimates; the exact proximal-stationarity plots use the linear objective.

M.2 Geometry and rate diagnostics

For a Stiefel tangent step, the largest distance error over the comparator in Definition 3.4 is available in closed form. Writing $Z = X + \Xi$ and $R = \mathcal{R}_X(\Xi)$,

$$\max_{Y \in \mathrm{St}(n,r)} \mathfrak{D}_{\mathcal{R}}(X, Y; \Xi) = 2 \|Z - R\|_* - \|\Xi\|_F^2 , \tag{63}$$

where $\|\cdot\|_*$ is the nuclear norm. Indeed, tangency gives $\|Z\|_F^2 = r + \|\Xi\|_F^2$, so expansion leaves $-\|\Xi\|_F^2 + 2\langle Z - R, Y\rangle$; rectangular orthogonal Procrustes maximizes the last inner product at $\|Z - R\|_*$. We use (63) to measure the uniform geometric error along actual optimizer trajectories.

The retraction comparison along optimization trajectories uses the orientation-sensitive linear Stiefel objective from the rate experiment. Its minimizer and proximal stationarity are available in closed form. Orientation sensitivity distinguishes QR and polar orthogonalizations of the same trial matrix: they span the same subspace, so the DPCP identity $F(XQ) = F(X)$ for $Q \in \mathsf{O}(r)$ gives them the same objective value.

Table 4 gives the large-step summaries. The column headed "step > 0.3" records exceedance of the controlled QR radius; polar has global distance control. QR-GeoClip reduces the median maximum objective rise by about 80% relative to QR-StatClip and attains lower final stationarity. Upward

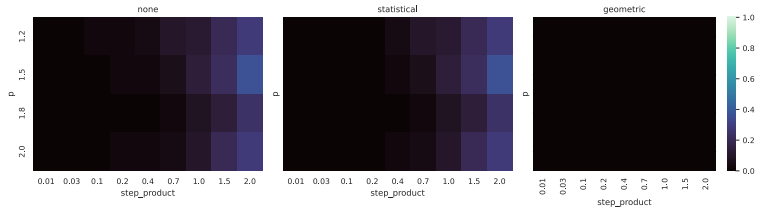

Figure 4: Probability that a Pareto step exceeds the prescribed QR radius in the one-step sampling experiment (50 trials per cell). Statistical clipping allows radius exceedance; the geometric cap keeps every observed QR step within the radius. Polar satisfies the distance bound globally, so its corresponding violation grid is identically zero.

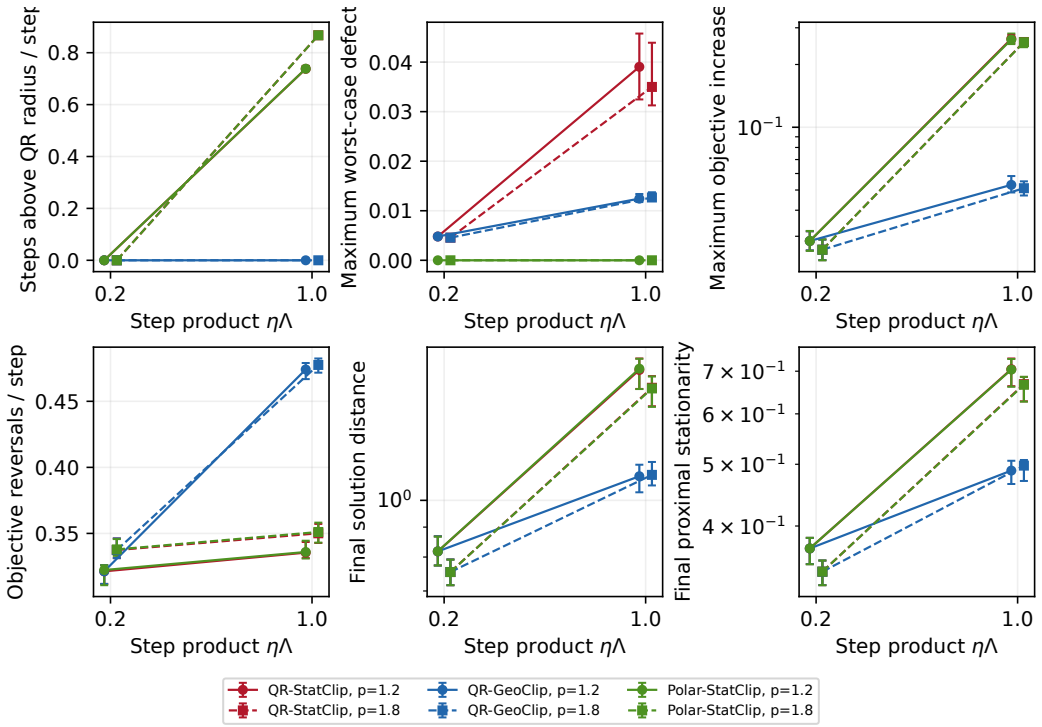

Figure 5: QR ablation along optimization trajectories. Curves show medians and IQRs over 50 paired seeds. Objective increases are evaluated at every iteration. The cap reduces the maximum upward fluctuation and improves final exact stationarity, with more frequent small upward moves.

moves become smaller and more frequent, identifying step-amplitude control as the mechanism captured by this ablation.

Figure 6 plots the positive defect. The QR example has quadratic growth, while the circle example with comparator $y = x$ has quartic growth, as shown in the proof of Proposition K.4. The polar curve stays at the plotting floor because every measured defect is nonpositive, in agreement with Lemma B.1.

Figure 1 reports the Pareto rate curves. The full rate table also contains Student-t, stable, and Gaussian noise. Each plotted point is the median of 50 seeds. The GeoNorm family uses $B \in \{1, 64, T\}$; the tangent geometric-median baseline uses eight blocks of eight samples. The $B = T$ experiment evaluates tail-parameter-free iteration rates with T^2 oracle calls; the DPCP comparison fixes the oracle budget across batch sizes. In Figure 1, RSSG and GeoClip share the stepsize $\eta_T = 0.5 \min\{\Lambda_T^{-1}, T^{-1/2}\}$, with $\Lambda_T = \max\{2, 0.25T^{1/p}\}$; all GeoNorm variants use $\alpha_T = 0.25T^{-1/2}$.

Table 4: Median trajectory results at $\eta\Lambda = 1$ (50 paired seeds).

p	Method	Step > 0.3	Max defect	Reversal	Max rise	$\|X - X^*\|_F$	Θ
1.2	QR-StatClip	.738	.0390	.335	.266	1.671	.705
1.2	QR-GeoClip	.000	.0124	.474	.0529	1.100	.489
1.2	Polar-StatClip	.738	.0000	.336	.263	1.679	.704
1.8	QR-StatClip	.867	.0350	.350	.256	1.557	.668
1.8	QR-GeoClip	.000	.0127	.478	.0510	1.105	.498
1.8	Polar-StatClip	.867	.0000	.351	.256	1.553	.667

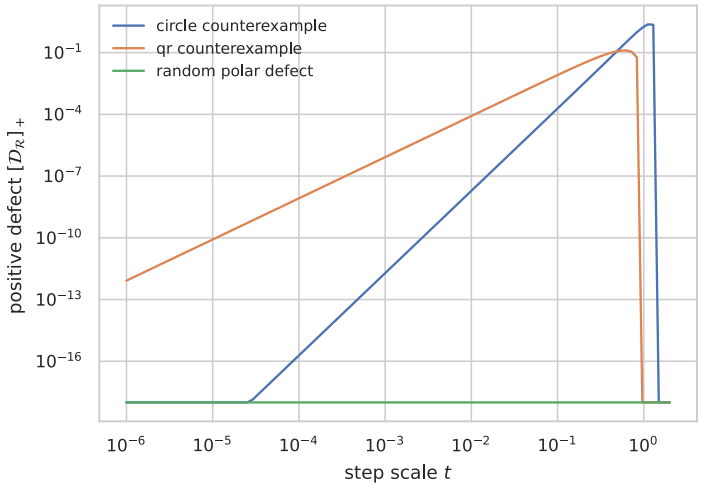

Figure 6: Positive retraction defect over step scale t. The plotting floor is 10^{-18}. QR has positive quadratic defect and the second-order circle example has positive quartic defect; polar has no positive value in the float64 trials.

The fitted rates evaluate GeoClip and $B = T$ GeoNorm. The primary DPCP comparisons use independently validation-selected stepsizes for each method.

M.3 DPCP COMPARISONS

Table 5 reports the paired difference $\mathrm{err}(\mathrm{RSSG}) - \mathrm{err}(\mathrm{method})$. Here RSSG is the basic update in (11); positive values favor the listed method. Four of the five $p < 2$ settings have a strictly positive interval for GeoClip or GeoNorm ($B = T$), and each method has three heavy-tail improvements significant after Holm correction. For Pareto $p = 1.2$, GeoClip's positive interval accompanies a Holm-adjusted p-value of 0.216. At Pareto $p = 1.8$, RSSG remains competitive with both methods. The $p = 2$ row evaluates the finite-variance control. The displayed inferential comparisons pair GeoClip and GeoNorm with $B = T$ against RSSG; their adjusted p-values use the full 70-comparison family described above. Median errors and IQRs characterize the $B = 1$ normalized update descriptively in Table 2.

Figure 2 fixes $(n, r, N) = (100, 5, 5000)$, Pareto $p = 1.5$, and 50% outliers. It complements the final-error table with recovery trajectories at matched oracle counts. Large batches aggregate more samples per update; single-sample methods use the same budget for more frequent updates along noisier paths.

The GeoClip schedule in (15) uses a supplied moment order and noise-scale input. The misspecification experiment sets $\widehat{p} = p \pm 0.3$ or $\widehat{s} \in \{s/4, 4s\}$, where s denotes the configured radial noise scale. Figure 8 shows that underestimating the scale can substantially increase recovery error, while the effect of overestimation depends on the fixed budget and stepsize coupling. This comparison measures sensitivity to the supplied moment order and scale. Corollary J.1 and Theorem 5.4 provide parameter choices independent of p, σ.

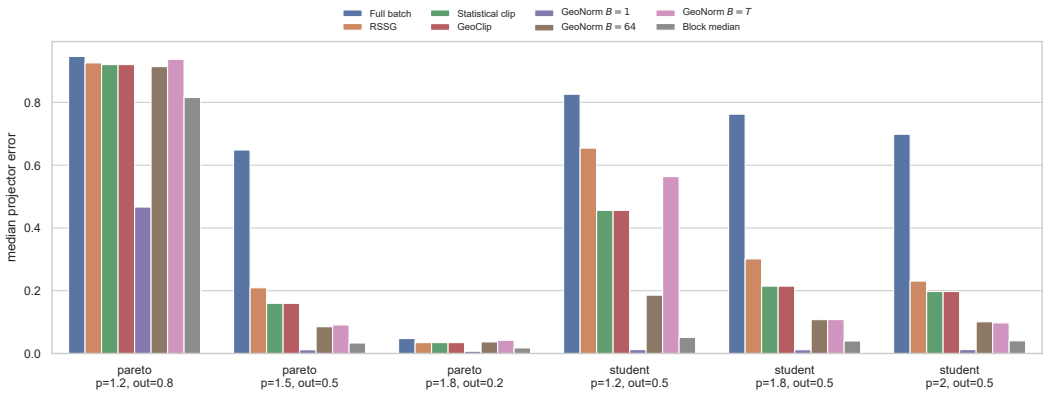

Figure 7: Median final projector error in the six primary synthetic DPCP settings (20 paired seeds, 2^{18} oracle calls). Lower is better.

Table 5: Paired primary-DPCP improvement relative to RSSG. Each row uses 20 seeds. CI is the 95% paired bootstrap interval; p_{H} is Holm-adjusted.

Family	p	Outliers	Method	Median	95% CI	p_{H}
Pareto	1.2	0.8	GeoClip	0.008	[0.003, 0.012]	0.216
Pareto	1.2	0.8	GeoNorm ($B = T$)	-0.003	[-0.016, 0.008]	1.000
Pareto	1.5	0.5	GeoClip	0.047	[0.013, 0.102]	2.0×10^{-4}
Pareto	1.5	0.5	GeoNorm ($B = T$)	0.103	[0.073, 0.191]	9.2×10^{-5}
Pareto	1.8	0.2	GeoClip	0.000	$[0, 2.7 \times 10^{-6}]$	0.415
Pareto	1.8	0.2	GeoNorm ($B = T$)	-0.006	[-0.013, 0.000]	1.000
Student-t	1.2	0.5	GeoClip	0.179	[0.149, 0.220]	9.2×10^{-5}
Student-t	1.2	0.5	GeoNorm ($B = T$)	0.047	[0.026, 0.135]	0.0059
Student-t	1.8	0.5	GeoClip	0.055	[0.017, 0.088]	9.2×10^{-5}
Student-t	1.8	0.5	GeoNorm ($B = T$)	0.187	[0.128, 0.207]	9.2×10^{-5}
Student-t	2.0	0.5	GeoClip	0.025	[0.009, 0.040]	0.0024
Student-t	2.0	0.5	GeoNorm ($B = T$)	0.125	[0.078, 0.157]	9.2×10^{-5}

M.4 Hopkins155 and wall-clock measurements

Hopkins155 contains 104 checkerboard, 38 traffic, and 13 articulated sequences (Tron & Vidal, 2007). We run all 155 sequences once for each $p \in \{1.2, 1.5, 1.8\}$ and repeat a stratified subset of 30 sequences with ten run seeds. Each seed controls corruption, initialization, and oracle sampling jointly. The all-sequence comparison uses seed zero and gives each sequence equal weight. The repeated subset uses seeds zero through nine; we first compute the median and IQR within each sequence, then report their category medians. Thus the repeated subset measures run-to-run variability with equal sequence weights. The motion rank is four. Labels are not used for centering or initialization; they identify the largest-motion evaluation target and the tracks that receive the radial corruption. All trajectories are centered and globally median-scaled before positive Pareto multipliers are applied to outlier tracks. At a fixed seed, changing these multipliers leaves the $B = 1$ direction in (18) unchanged. This explains the identical rounded $B = 1$ errors across p in the Hopkins tables.

For an estimated normal basis X and reference $X_\star$, we report

$$\mathrm{err}(X, X_\star) = \frac{\left\| XX^\top - X_\star X_\star^\top \right\|_F}{\sqrt{2r}}. \tag{64}$$

With F frames, the ambient dimension is $2F$ and the normal rank is $r = 2F - 4$. The error is bounded by $\sqrt{\min\{r, 4\}/r}$: the squared Frobenius difference of the complementary rank-four

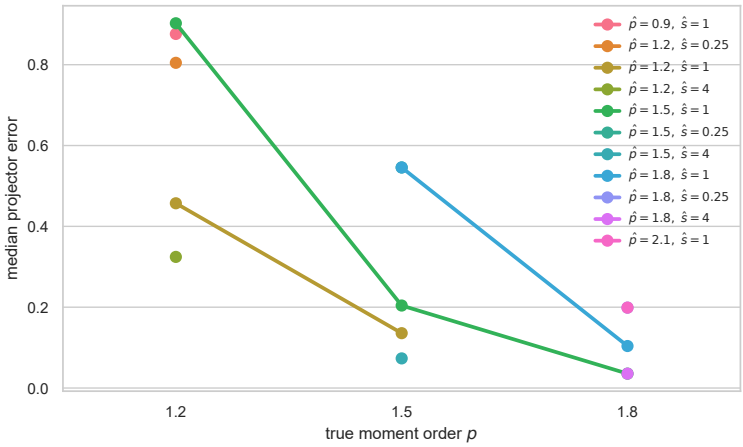

Figure 8: Median GeoClip projector error under misspecified tail order or scale. A connected line holds $(\widehat{p}, \widehat{s})$ fixed while the true moment order varies; isolated points correspond to settings evaluated at only one true moment order.

Table 6: Hopkins155 median projector error over all 155 sequences, with one observation per sequence (seed zero): 13 articulated, 104 checker, and 38 traffic sequences. Lower is better.

Category	p	RSSG	GeoClip	GeoNorm ($B = T$)	GeoNorm ($B = 1$)
Articulated	1.2	0.1191	0.1187	0.1718	0.1290
Articulated	1.5	0.1189	0.1190	0.1778	0.1290
Articulated	1.8	0.1211	0.1211	0.1756	0.1290
Checker	1.2	0.1274	0.1249	0.1873	0.1359
Checker	1.5	0.1273	0.1272	0.1893	0.1359
Checker	1.8	0.1216	0.1214	0.1881	0.1359
Traffic	1.2	0.1324	0.1322	0.1639	0.1322
Traffic	1.5	0.1311	0.1311	0.1647	0.1322
Traffic	1.8	0.1312	0.1312	0.1639	0.1322

projectors is at most $2\min\{r, 4\}$. Method comparisons are therefore paired within each sequence; category summaries retain this normal-rank normalization.

Table 7 reports repeated-run variability separately from the all-sequence comparison. For articulated sequences at $p = 1.2$, GeoClip's median within-sequence IQR is 0.0031, compared with 0.0063 for RSSG. Across the traffic subset, single-sample GeoNorm has median within-sequence IQRs near 0.0012 for all three moment orders.

The synchronized timing subset ran each process sequentially on one NVIDIA A800 80GB PCIe. In the representative Pareto DPCP setting, RSSG took 142.6 s, statistical clipping 147.8 s, GeoNorm ($B = 1$) 150.2 s, and GeoClip 152.6 s. The vectorized batch methods took 1.1 s for GeoNorm ($B = T$), 3.2 s for GeoNorm ($B = 64$), and 9.2 s for the block geometric median. These synchronized times complement oracle counts by quantifying the implementation's use of vectorized batch computation on the A800.

N Further Related Work

Table 8 separates the domain, noise model, and guarantee.

Finite-moment weakly convex analysis. Zhu et al. (2026, Theorem 4.2) assume a finite p-th central moment and a bounded convex domain, and give an expectation guarantee for unclipped SGD. Theorem 4.4 gives a high-probability guarantee for unclipped SGD under sub-Weibull noise; Theorem 5.1 instead uses clipping to obtain a finite-moment high-probability guarantee on possibly

Table 7: Hopkins155 repeated-run variability over 30 sequences and ten seeds controlling corruption, initialization, and sampling. Each cell gives the category median of per-sequence median error, followed by the category median of within-sequence IQR in parentheses.

Category	p	RSSG	GeoClip	GeoNorm ($B = T$)	GeoNorm ($B = 1$)
Articulated	1.2	.0698 (.0063)	.0698 (.0031)	.1744 (.0032)	.0712 (.0030)
Articulated	1.5	.0689 (.0014)	.0689 (.0016)	.1709 (.0033)	.0712 (.0030)
Articulated	1.8	.0701 (.0039)	.0701 (.0039)	.1711 (.0031)	.0712 (.0030)
Checker	1.2	.1330 (.0087)	.1323 (.0081)	.1881 (.0062)	.1368 (.0095)
Checker	1.5	.1313 (.0096)	.1327 (.0092)	.1879 (.0070)	.1368 (.0095)
Checker	1.8	.1333 (.0088)	.1333 (.0088)	.1874 (.0066)	.1368 (.0095)
Traffic	1.2	.1313 (.0035)	.1291 (.0027)	.1648 (.0033)	.1306 (.0012)
Traffic	1.5	.1299 (.0030)	.1295 (.0027)	.1644 (.0038)	.1306 (.0012)
Traffic	1.8	.1290 (.0035)	.1290 (.0035)	.1650 (.0041)	.1306 (.0012)

Table 8: Positioning of the present theory. "HT" denotes a finite p-th-moment heavy-tail oracle and "HP" a high-probability guarantee.

Work	Domain	Objective	Noise	Geometry / guarantee
Li et al. (2021)	Stiefel/manifold	weakly convex	second moment	proximal stationarity
Chen et al. (2021)	Stiefel	smooth	classical	polar global
Davis et al. (2025)	prox-smooth set	weakly convex	classical	set models + retraction
Zhu et al. (2026)	Euclidean	weakly convex	HT	raw: expectation; clipped: HP
Hübler et al. (2025)	Euclidean	smooth nonconvex	HT	normalized: HP
Liu (2026b)	bounded Euclidean	nonsmooth	HT	expectation
This work	compact manifold	weakly convex	HT	expectation + HP

unbounded convex domains. Our RSSG theorem combines an unclipped update, a finite p-th central moment, and a high-probability bound on a compact manifold. Its proof bounds the variance of proximal-value increments through the retracted displacement. The clipping threshold balance in (15) follows the Euclidean analysis; the additional argument handles curvature and retraction distance error.

Unclipped methods and stationarity criteria. For unclipped Euclidean methods, Fatkhullin et al. (2025) analyze Hölder-smooth objectives using a p-order Moreau envelope and forward–backward stationarity, and consider mini-batches under standard smoothness. Liu (2026a) study differentiable upper-smooth objectives without bounded domains, with a general nonconvex guarantee on a $p/2$ moment of averaged squared gradient norms. The bounded-domain results of Liu (2026b) are also in expectation. These results use different objective assumptions or stationarity criteria from our high-probability bound for the quadratic constrained proximal residual.

Clipping, normalization, and update geometry. Refined clipping analysis can reveal effective-dimension gains and new lower bounds (Liu, 2026c). Other recent results study convergence of vanilla momentum and tight lower bounds for smooth nonconvex heavy-tail models (Yamada et al., 2026; Fradin et al., 2026). Our analysis combines retraction distance comparisons with finite-moment stochastic estimates to control the constrained proximal residual. The update geometry affects the remainder and admissible step lengths. RSSG uses compactness to concentrate proximal-value increments. Clipping and normalization bound the tangent steps and introduce controlled bias in their directions.

Structured weakly convex objectives and submanifold updates. The randomized submanifold method of Cheung et al. (2025) samples blocks and evaluates exact partial subgradients. Assumption 3.2 describes data-sampling noise in a tangent subgradient. For structured composite objectives, Riemannian smoothing yields $O(\varepsilon^{-3})$ deterministic and $O(\varepsilon^{-5})$ stochastic iteration bounds (Peng et al., 2023); momentum-based variance reduction gives $O(\varepsilon^{-3})$ stochastic complexity under the corresponding assumptions (Deng et al., 2025). These results use generalized stationarity and access to the smooth component and the proximal map of the nonsmooth component.

Heavy-tailed observations and other oracle models. Heavy-tailed observations and heavy-tailed stochastic subgradients define different oracle models. Robust low-rank regression with absolute, Huber, or quantile losses admits Riemannian subgradient recovery guarantees under statistical assumptions and suitable initialization (Shen et al., 2023). We study proximal stationarity for a general weakly convex population objective under a conditional moment bound on the tangent oracle. The zeroth-order method of Ma et al. (2026) assumes a smooth objective and noisy function values; its main convergence results require additional second-moment conditions on the oracle or aggregate error. Heavy-tail analyses of matrix update rules such as Muon (Iiduka, 2026) concern a different optimization setting from manifold-constrained weakly convex minimization.