VERI-FRIDA: GEOMETRY-AWARE IMPLICIT VARIANCE REDUCTION FOR FRÉCHET REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning on rotations, covariance matrices, and other manifold-valued data calls for methods that respect their geometry. Fréchet regression is a recent, successful approach in which predictions are minimizers of weighted squared geodesic distances. However, weights depend on the predictor query and can be negative, particularly when extrapolating. Difference-of-convex (DC) methods build on this signed structure by linearizing the negative terms, and iteratively minimizing convex models; but most existing approaches are limited to Hadamard manifolds. Positively curved spaces pose additional challenges, as squared geodesic distances on these spaces are not globally convex. Building on local geometries, recently proposed curvature-aware methods have been shown to converge, yet still require repeated full-data model evaluations, thereby limiting their scalability. We propose VariancE-Reduced Implicit FRIDA (VERI-FRIDA), a stochastic implicit framework for affine combinations of squared-distance objectives on manifolds with bounded curvature. Our central idea is to sample the positively weighted terms and add a logarithmic correction to ensure that the model gradient equals variance-reduced estimates of the full gradient. We connect inexact implicit updates to estimator error, establishing two convergence guarantees under explicit geometric and negative-weight conditions. First, when the estimator is periodically or probabilistically reset to the exact full gradient, a sufficiently small step-size yields an $O(K^{-1})$ bound on the expected squared Riemannian gradient norm. Second, for recursive estimators that use only mini-batches, a suitable diminishing step-size schedule ensures that the gradient norm converges to zero almost surely. Numerical experiments show improved performance over other methods on various manifolds.

1 INTRODUCTION

Learning with rotations, covariance matrices, or subspaces requires predictions that respect the geometry of the response space (Pennec et al., 2006; Absil et al., 2008; Boumal, 2023). Fréchet regression achieves this by extending least squares to responses for which predictions are minimizers of a query-dependent affine combination of squared geodesic distances (Petersen & Müller, 2019). Specifically, for responses $z_1, \ldots, z_m$ on a metric space $(\mathcal{M}, d)$, the empirical Fréchet regressor is the minimizer of $\sum_{i=1}^{m} w_i d^2(y, z_i)$, where $\sum_i w_i{=}1$. Its flexibility has allowed applications spanning diffusion tensor imaging and brain connectivity analysis (Yan et al., 2025; Zhou & Müller, 2022), age-related gene co-expression in single-cell genomics (Xu & Li, 2025), the composition of U.S. electricity generation (Song et al., 2026), predicting human mortality, and modeling urban transportation networks (Iao et al., 2025; Zhou et al., 2026).

When $\mathcal{M}$ is a Hilbert space with its norm distance, unique Fréchet regressors are the affine combination $\sum_{i=1}^{m} w_i z_i$, regardless of the signs of the weights (Petersen & Müller, 2019, Theorem 5). On Hadamard manifolds, a closed-form solution is generally unavailable, but squared distances remain globally geodesically convex. Consequently, positively weighted terms are geodesically convex and negatively weighted terms are geodesically concave, allowing a global difference-of-convex (DC) decomposition $f = g - h$ (Pham Dinh & Le Thi, 1997; Le Thi & Pham Dinh, 2018). This structure supports intrinsic DC methods that retain g and approximate the concave component $-h$ through logarithmic first-order models (Souza & Oliveira, 2015; Almeida et al., 2020; Bergmann et al., 2024). However, when positive sectional curvature is allowed, squared geodesic distances need not be globally geodesically convex. Recovering a DC structure, therefore, requires a carefully cho-

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.

sen neighborhood in which these terms are smooth and geodesically convex (Afsari, 2011; Zhou & Uribe, 2026). Recently, (Zhou & Uribe, 2026) proposed FRIDA, a method that allows two-sided bounded sectional curvatures, including positive curvature, by operating within controlled local regions. Its implicit update follows a DC structure, and adds proximal regularization to ensure strong geodesic convexity. Nonetheless, this structure comes at a computational cost: each step requires evaluations of full gradients, which is generally unavailable in large-scale regimes, i.e., large m.

Riemannian stochastic gradient methods reduce the need for full-data evaluations through sampling (Bonnabel, 2013). Also, variance-reduced methods (Zhang et al., 2016; Sato et al., 2019) and recursive gradient and momentum estimators (Kasai et al., 2018; Zhou et al., 2019; Han & Gao, 2021) generally improve first-order accuracy. Recently, SERENA has unified a broad family of these estimators using retractions and vector transport, including R-SRVRG, R-SVRRM, and R-Hybrid-SGD (Liu et al., 2025), enabling common analysis for families of algorithms. Others have studied stochastic proximal-gradient methods on specific manifolds, such as the Stiefel manifold (Wang et al., 2022), and stochastic formulations of Bures–Wasserstein Fréchet regression (Nguyen & Uribe, 2026). These works provide stochastic gradient estimators and problem-specific proximal constructions. Sampled implicit models still require control of curvature and its gradient error.

In Euclidean optimization, stochastic surrogate and DC methods (Mairal, 2013; Davis & Drusvyatskiy, 2019; Le Thi et al., 2017; Le Thi et al., 2022), variance-reduced proximal-point methods (Traoré et al., 2024), and gradient-corrected implicit subproblems (Shamir et al., 2014; Karagulyan et al., 2024) have successfully combined sampling with nonlinear models. But extending these constructions to positively curved spaces requires additional geometric control: squared-distance convexity holds only within suitable regions, and logarithmic first-order corrections need not be geodesically affine. Sampling the positive component introduces a center-gradient error that does not disappear merely by accurately estimating the negative gradient. The open challenge is therefore to couple an inexpensive sampled positive model with variance-reduced signed-gradient information while ensuring strong geodesic convexity for every model realization.

We introduce *VariancE-Reduced Implicit FRIDA* (VERI-FRIDA), a stochastic implicit framework for signed Fréchet regression that separates nonlinear model sampling from first-order accuracy. Our central approach is as follows: a fixed-size, weight-proportional positive batch retains the nonlinear Fréchet component; then a tangent correction ensures that the model gradient at the current iterate equals a variance-reduced estimate of the complete gradient for every batch realization. Finally, an adaptive regularizer controls the curvature of the logarithmic correction and ensures that each model is strongly geodesically convex. The resulting construction combines a small nonlinear subproblem with variance-reduced first-order information, while a common implicit-update analysis accommodates different choices of estimator. We summarize our main contributions as follows:

- **A stochastic implicit framework for Fréchet regression.** We propose a gradient-matched optimization framework for signed squared-distance objectives on Riemannian manifolds. The framework decouples nonlinear model sampling from first-order accuracy by coupling fixed-size positive models with variance-reduced gradients through an exact tangent correction.
- **Convergence guarantees for implicit variance reduction.** We quantify how gradient-estimation errors and inexact solutions of the inner problem affect the progress of the implicit updates. We then use bounds on the step length and objective decrease to establish stationarity guarantees for several DC-structured implicit variance-reduction schemes.
- **Convergence under two sampling regimes.** We prove convergence under appropriate geometric and negative-weight conditions. When the gradient estimator is periodically or randomly reset to the full gradient, a sufficiently small step-size parameter gives an $O(K^{-1})$ bound on the expected squared Riemannian gradient norm. When estimators use mini-batches, a suitable diminishing step-size schedule ensures that the gradient norm converges to zero almost surely.

2 Problem Formulation and Geometric Setting

Let $(\mathcal{M}, d)$ be a connected, complete, finite-dimensional Riemannian manifold with geodesic distance d. Given observations $(x_i, z_i) \in \mathbb{R}^p \times \mathcal{M}$, $i = 1, \dots, m$, we consider empirical global Fréchet regression at a query $x \in \mathbb{R}^p$ (Petersen & Müller, 2019). We compute a fitted response by solving

$$\widehat{m}_{\oplus}(x) \in \operatorname*{argmin}_{y \in \mathcal{M}_{\mathrm{ex}}} f_x(y),\ f_x(y) \triangleq \sum_{i=1}^{m} w_i(x) d^2(y, z_i),\ w_i(x) \triangleq \frac{1}{m}\left[1 + (x_i - \bar{x})^\top \widehat{\Sigma}^{-1}(x - \bar{x})\right]. \quad (1)$$

Here $\bar{x} \triangleq m^{-1}\sum_i x_i$ and $\widehat{\Sigma} \triangleq m^{-1}\sum_i (x_i-\bar{x})(x_i-\bar{x})^\top$, with $\widehat{\Sigma} \succ 0$. The constraint set $\mathcal{M}_{\mathrm{ex}}$ is a closed geodesic ball to be specified in Assumption 2.1. The weights satisfy $\sum_i w_i(x) = 1$, so f_x is an affine combination of squared-distance losses. At $x = \bar{x}$, all weights equal $1/m$, and the problem reduces to the empirical Fréchet mean over $\mathcal{M}_{\mathrm{ex}}$ (Grove & Karcher, 1973). For simplicity of exposition, we fix the observations and query, write $w_i \triangleq w_i(x)$ and $f \triangleq f_x$, and define $\ell_i(y) \triangleq d^2(y, z_i)$. All subsequent randomness comes from algorithmic sampling. Moreover, we denote $W_\pm \triangleq \sum_i w_i^\pm$ and $W_{\mathrm{abs}} \triangleq \sum_i |w_i| = W_+ + W_- = 1 + 2W_-$. Thus, since $W_+ - W_- = 1$, the positive component is nonempty. The quantity W_- measures the total negative weight at the chosen query. For $y \in \mathcal{M}$, let $T_y\mathcal{M}$ denote its tangent space, with norm $\|v\|_y \triangleq \sqrt{\langle v, v\rangle_y}$. We denote the Riemannian gradient, Hessian, and exponential map by grad, Hess, and Exp_y, respectively. On the normal neighborhoods used below, $\mathrm{Log}_y z$ is the initial velocity of the unique minimizing geodesic from y to z on $[0, 1]$, with $\|\mathrm{Log}_y z\|_y = d(y, z)$. Parallel transport along this geodesic is denoted by $P_{y\to z} : T_y\mathcal{M} \to T_z\mathcal{M}$. We write $\overline{B}_a(p) \triangleq \{y \in \mathcal{M} : d(p, y) \le a\}$. Moreover, separating the positive and negative weights, $w_i^+ \triangleq \max\{w_i, 0\}$ and $w_i^- \triangleq \max\{-w_i, 0\}$, gives

$$f = g - h, \qquad g(y) \triangleq \sum_{i=1}^m w_i^+ \ell_i(y), \qquad h(y) \triangleq \sum_{i=1}^m w_i^- \ell_i(y). \tag{2}$$

We introduce our geometric setting by noting that squared distances can lose smoothness along cut loci and need not be globally geodesically convex. Thus, we impose the following assumption.

Assumption 2.1 *There exist $c \in \mathcal{M}$ and $0 < r < R < R + \bar{r} < \rho < \infty$ such that $z_i \in \overline{B}_r(c)$ for all i. Define $\mathcal{M}_{\mathrm{ex}} \triangleq \overline{B}_R(c)$ and $U \triangleq \overline{B}_{R+\bar{r}}(c)$. The outer ball $\overline{B}_\rho(c)$ and each $\overline{B}_{\bar{r}}(p)$, $p \in \mathcal{M}_{\mathrm{ex}}$, are strongly convex and normal. The map $(y, z) \mapsto \mathrm{Log}_y z$ is smooth on a neighborhood of $\overline{B}_\rho(c) \times \overline{B}_\rho(c)$. On $\overline{B}_\rho(c)$, the sectional curvature satisfies $-\Lambda_- \le \sec_{\mathcal{M}} \le \Lambda_+$ for some $\Lambda_\pm \ge 0$. The distance bound $D \triangleq R + \bar{r} + r$ satisfies $\sqrt{\Lambda_+}\, D < \pi/2$.*

The ball $\overline{B}_r(c)$ contains the responses, $\mathcal{M}_{\mathrm{ex}}$ is the search region, and U contains every local ball $\overline{B}_{\bar{r}}(p)$ with $p \in \mathcal{M}_{\mathrm{ex}}$. Strong convexity ensures the existence of unique minimizing geodesics within these balls, while smooth logarithms allow differentiation. The curvature bounds and the restriction on D control the convexity and smoothness of the individual squared-distance losses. Such local conditions are standard in Riemannian mean computation (Afsari, 2011; Boumal, 2023); they allow positive curvature in sufficiently small neighborhoods of the responses. To exclude minimizers on the boundary of $\mathcal{M}_{\mathrm{ex}}$, we also control the total negative weight relative to the margin $R - r$ between the data ball and that boundary.

Assumption 2.2 *The negative weight satisfies $W_- < (R - r)/(2r)$, or equivalently $W_{\mathrm{abs}} r < R$.*

Completeness and finite dimensionality make $\mathcal{M}_{\mathrm{ex}}$ compact, so the continuous objective f attains a minimum there. Under Assumptions 2.1 and 2.2, the boundary argument in Zhou & Uribe (2026, Thm. 4.1 and Cor. 4.3) excludes boundary minimizers. Hence every minimizer $y^\star$ lies in $\mathrm{int}(\mathcal{M}_{\mathrm{ex}})$ and satisfies $\mathrm{grad}\, f(y^\star) = 0$. Thus, the region constraint is inactive at regression minimizers; uniqueness is not required. The proof of this radial argument is given in Lemma B.4; in particular, (33) localizes every regression minimizer to $\overline{B}_{W_{\mathrm{abs}} r}(c)$.

We next recall the derivative bounds used in the algorithm and its analysis. For $y \in U$, the triangle inequality gives $d(y, z_i) \le D$, and

$$\mathrm{grad}\, \ell_i(y) = -2\,\mathrm{Log}_y(z_i), \qquad \mathrm{grad}\, f(y) = \mathrm{grad}\, g(y) - \mathrm{grad}\, h(y). \tag{3}$$

Hessian comparison and smoothness of the logarithm map on the prescribed compact regions give the following uniform estimates (Zhou & Uribe, 2026, Lemmas 2.4 and 5.1–5.3).

Proposition 2.3 *Let Assumption 2.1 hold. There exist $0 < \delta \le \zeta < \infty$, $0 < \delta_p \le \zeta_p < \infty$, and $0 \le L_{\log} < \infty$, such that, for every i, $y \in U$, and $u \in T_y\mathcal{M}$,*

$$2\delta\|u\|_y^2 \le \mathrm{Hess}\, \ell_i(y)[u, u] \le 2\zeta\|u\|_y^2. \tag{4}$$

For every $p \in \mathcal{M}_{\mathrm{ex}}$, $y \in \overline{B}_{\bar{r}}(p)$, $u \in T_y\mathcal{M}$, and $\xi \in T_p\mathcal{M}$,

$$\delta_p\|u\|_y^2 \le \mathrm{Hess}_y\left(\tfrac{1}{2}d^2(p, \cdot)\right)[u, u] \le \zeta_p\|u\|_y^2, \ \left|\mathrm{Hess}_y\langle \xi, \mathrm{Log}_p(\cdot)\rangle_p[u, u]\right| \le L_{\log}\|\xi\|_p\, \|u\|_y^2.$$

Algorithm 1 VERI-FRIDA

Require: y_0; $K, b_g \geq 1$; admissible $\alpha, \mu_\star, \nu, \zeta_{\text{in}}$; estimator and its batch, refresh, and mixing parameters.

1: For each $\sigma \in \mathcal{S}$, set $\widetilde{y} \leftarrow y_0$ and $\widetilde{\mu}^\sigma \leftarrow \widetilde{v}_0^\sigma$ (for SVRG/SVRRM/SRVRG), where

$$\widetilde{v}_0^\sigma = \begin{cases} \widehat{G}^\sigma_{B_0^\sigma}(y_0), & \text{STORM/Hybrid-SGD,} \\ \operatorname{grad} F_\sigma(y_0), & \text{otherwise,} \end{cases} \qquad v_0^\sigma = \Pi^{R_\sigma}_{y_0}(\widetilde{v}_0^\sigma).$$

2: **for** $k = 0, \ldots, K-1$ **do**

3: Choose $\eta_k \in (0, 1]$ from the current history; set $v_k \leftarrow v_k^+ - v_k^-$.

4: Draw $S_k^g \sim (p^+)^{b_g}$ independently of that history. Build φ_k using (13) and (15), keeping S_k^g fixed.

5: Compute $y_{k+1} \in \mathcal{M}_k$ and $n_{k+1} \in N_{\mathcal{K}_\alpha}(y_{k+1})$ satisfying (16).

6: $j \leftarrow k+1$. Choose a_j below; use fresh batches whenever $a_j = 0$.

	PAGE $a_j \sim \operatorname{Bern}(p)$	**Periodic** $a_j \leftarrow \mathbf{1}_{\{H \mid j\}}$	**No refresh** $a_j \leftarrow 0$
$a_j = 0$	$\widetilde{v}_j^\sigma \leftarrow D_{j,B}^\sigma$	**SARAH/SPIDER:** $\widetilde{v}_j^\sigma \leftarrow D_{j,B}^\sigma$ **SVRG:** $\widetilde{v}_j^\sigma \leftarrow V_{j,C}^\sigma$ **SVRRM:** $\widetilde{v}_j^\sigma \leftarrow (1-\beta)D_{j,B}^\sigma + \beta V_{j,B}^\sigma$ **SRVRG:** $\widetilde{v}_j^\sigma \leftarrow (1-\beta)D_{j,B}^\sigma + \beta V_{j,C}^\sigma$	**STORM:** $\widetilde{v}_j^\sigma \leftarrow (1-\beta_{j-1})D_{j,B}^\sigma + \beta_{j-1}U_{j,B}^\sigma$ **Hybrid-SGD:** $\widetilde{v}_j^\sigma \leftarrow (1-\beta_{j-1})D_{j,B}^\sigma + \beta_{j-1}U_{j,C}^\sigma$
$a_j = 1$	$\widetilde{v}_j^\sigma \leftarrow \operatorname{grad} F_\sigma(y_j)$	$\widetilde{v}_j^\sigma \leftarrow \operatorname{grad} F_\sigma(y_j)$, $\widetilde{y} \leftarrow y_j$, $\widetilde{\mu}^\sigma \leftarrow \widetilde{v}_j^\sigma$	$\widetilde{v}_j^\sigma \leftarrow \operatorname{grad} F_\sigma(y_j)$

7: $v_j^\sigma \leftarrow \Pi^{R_\sigma}_{y_j}(\widetilde{v}_j^\sigma)$, $\sigma \in \mathcal{S}$.

8: **end for**

9: **return** y_K; for the finite-horizon guarantee, y_{J_K} with J_K independent and uniform on $\{0, \ldots, K-1\}$.

The first bound makes g and h geodesically convex on the local balls, giving (2) its DC structure. The second gives analogous bounds for the squared distance to a reference point. The third controls the curvature of $y \mapsto \langle \xi, \operatorname{Log}_p y \rangle_p$, which need not be affine on a manifold. Moreover, by setting $G_\ell \triangleq 2D$, $L_\ell \triangleq 2\zeta$, and $L_f \triangleq W_{\text{abs}} L_\ell$, (3) gives $\|\operatorname{grad} \ell_i(y)\|_y \leq G_\ell$ for every i and $y \in U$. Integrating the Hessian bounds along a minimizing geodesic contained in U yields, for its endpoints $y, y' \in U$,

$$\|\operatorname{grad} \ell_i(y') - P_{y \to y'} \operatorname{grad} \ell_i(y)\|_{y'} \leq L_\ell d(y, y'), \ \|\operatorname{grad} f(y') - P_{y \to y'} \operatorname{grad} f(y)\|_{y'} \leq L_f d(y, y').$$

Consequently, although f need not be geodesically convex, it holds that

$$f(y') \leq f(y) + \langle \operatorname{grad} f(y), \operatorname{Log}_y(y') \rangle_y + (L_f/2) d^2(y, y'). \tag{5}$$

The convexity of U and smoothness bounds are proved in Lemmas B.2 and B.3.

3 VERI-FRIDA AND MAIN RESULTS

In this section, we describe our proposed Algorithm 1 and explain its design. Building on the unified treatment of gradient estimators in SERENA (Liu et al., 2025), we develop VERI-FRIDA variants based on PAGE (Li et al., 2021), SARAH/SPIDER (Nguyen et al., 2017; Fang et al., 2018), SVRG (Johnson & Zhang, 2013), SVRRM (Liao et al., 2025), SRVRG (Liu et al., 2025), STORM (Cutkosky & Orabona, 2019), and Hybrid-SGD (Tran-Dinh et al., 2022).

DIRECT SAMPLING: Our starting point is the FRIDA update (Zhou & Uribe, 2026), where at a current point $y_k \in \operatorname{int}(\mathcal{M}_{\text{ex}})$, g is retained and h is replaced by its first-order approximation,

$$h(y_k) + \langle \operatorname{grad} h(y_k), \operatorname{Log}_{y_k}(y) \rangle_{y_k}.$$

However, the logarithmic term need not be geodesically affine, so the resulting model is not necessarily strongly convex. FRIDA therefore adds regularization to get a model:

$$\Phi_k(y) := g(y) - \langle \operatorname{grad} h(y_k), \operatorname{Log}_{y_k}(y) \rangle_{y_k} + (\tau_k/2) d^2(y_k, y).$$

The parameter $\tau_k > 0$ and a closed strongly convex normal ball $\mathcal{D}_k \subset \mathcal{M}_{\text{ex}}$ centered at y_k are chosen so that Φ_k is strongly convex on $\mathcal{D}_k$. The next point is then the unique minimizer of Φ_k over $\mathcal{D}_k$ where it holds that $\operatorname{grad} \Phi_k(y_k) = \operatorname{grad} f(y_k)$. The main challenge is that evaluating g and $\operatorname{grad} h(y_k)$ requires summing over m responses.

To reduce this cost, one may sample both components. Specifically, let $F_+ := g$, $F_- := h$, $\mathcal{S} := \{\sigma \in \{+, -\} : W_\sigma > 0\}$, and $p_i^\sigma := w_i^\sigma / W_\sigma$ (if W_- we ignore the negative terms), and consider independent batches S_k^g and S_k^h of sizes b_g and b_h, such that

$$\widehat{g}_k(y) := \frac{W_+}{b_g} \sum_{i \in S_k^g} \ell_i(y), \qquad \widehat{h}_k(y) := \frac{W_-}{b_h} \sum_{i \in S_k^h} \ell_i(y). \tag{6}$$

The coefficients of $\widehat{g}_k$ sum to W_+ for every batch, so the geometric bounds give $\operatorname{Hess} \widehat{g}_k \succeq 2W_+\delta I$ on the local model domains. Thus, replacing g, h gives a stochastic approximation:

$$\widehat{\Phi}_k(y) := \widehat{g}_k(y) - \langle \operatorname{grad} \widehat{h}_k(y_k), \operatorname{Log}_{y_k}(y) \rangle_{y_k} + (\tau_k/2) d^2(y_k, y). \tag{7}$$

Unfortunately, while this gradient is unbiased at the current point, it now contains sampling errors:

$$\operatorname{grad} \widehat{\Phi}_k(y_k) = \operatorname{grad} f(y_k) + \big(\operatorname{grad} \widehat{g}_k(y_k) - \operatorname{grad} \widehat{h}_k(y_k) - \operatorname{grad} f(y_k) \big). \tag{8}$$

Moreover, its variance need not vanish near a stationary point: individual loss gradients can remain nonzero when their weighted sum is zero. Direct sampling, therefore, does not by itself guarantee stationarity with fixed regularization, and decreasing $\widehat{\Phi}_k$ need not decrease f. Appendix B.2.1 gives a Euclidean example in which this variance persists at a stationary point and exact minimization of the sampled model increases the objective.

ALGORITHM DESIGN: We retain $\widehat{g}_k$ as the nonlinear part of the model and estimate both component gradients using the variance-reduction constructions following the unifying approach in SERENA (Liu et al., 2025). This allows the nonlinear model to use b_g responses even when the gradient estimates are computed in larger batches or with occasional full-gradient updates.

Let $\widetilde{v}_k^\sigma$ estimate $\operatorname{grad} F_\sigma(y_k)$. The estimator updates in Algorithm 1 use three quantities built from a common gradient oracle. For $\sigma \in \mathcal{S}$, let B be a nonempty batch of independent draws with replacement from p^σ, and define

$$\widehat{G}_B^\sigma(y) := \frac{W_\sigma}{|B|} \sum_{i \in B} \operatorname{grad} \ell_i(y), \tag{9}$$

where repeated indices are counted. For $j \geq 1$, a fresh gradient estimate is $U_{j,B}^\sigma := \widehat{G}_B^\sigma(y_j)$. To reuse the previous clipped estimate, we transport it to $T_{y_j}\mathcal{M}$ and add a gradient difference evaluated with the same batch at both points:

$$D_{j,B}^\sigma := P_{y_{j-1} \to y_j} v_{j-1}^\sigma + \widehat{G}_B^\sigma(y_j) - P_{y_{j-1} \to y_j} \widehat{G}_B^\sigma(y_{j-1}). \tag{10}$$

Alternatively, a stored point $\widetilde{y}$ and its full gradient $\widetilde{\mu}^\sigma := \operatorname{grad} F_\sigma(\widetilde{y})$ give the snapshot estimate

$$V_{j,B}^\sigma := P_{\widetilde{y} \to y_j} \widetilde{\mu}^\sigma + \widehat{G}_B^\sigma(y_j) - P_{\widetilde{y} \to y_j} \widehat{G}_B^\sigma(\widetilde{y}). \tag{11}$$

Algorithm 1 specifies how these quantities update $\widetilde{v}_j^\sigma$ and when full gradients are recomputed. Its parameters are the refresh probability $p \in (0, 1]$, the integer refresh period $H \geq 1$, and the fixed or varying mixing weights $\beta \in (0, 1)$ and $\beta_{j-1} \in (0, 1]$. Estimator batches are drawn after y_j is fixed. Within each sign, B and C denote independent batches; repeated use of B denotes the same batch. Moreover, since the component gradients satisfy $\|\operatorname{grad} F_\sigma(y_k)\|_{y_k} \leq R_\sigma := W_\sigma G_\ell$, we project their estimates onto these known bounds. For $a > 0$, let $\Pi_y^a(u) := u \min\{1, a/\|u\|_y\}$ when $u \neq 0$, with $\Pi_y^a(0) := 0$, and set $v_k^\sigma := \Pi_{y_k}^{R_\sigma}(\widetilde{v}_k^\sigma)$, $v_k := v_k^+ - v_k^-$. Because the true component gradient lies in the projection ball, clipping cannot increase its estimation error. The clipping bound is proved in Lemma B.5. Appendices B.4.1 and B.4.2 specify the sampling order and the oracle moment bounds.

Next, to make the sampled model agree with v_k at y_k, we propose

$$q_k := \operatorname{grad} \widehat{g}_k(y_k) - v_k = \big(\operatorname{grad} \widehat{g}_k(y_k) - v_k^+ \big) + v_k^- \tag{12}$$

$$\varphi_k(y) := \widehat{g}_k(y) - \langle q_k, \operatorname{Log}_{y_k}(y) \rangle_{y_k} + (\lambda_k/2) d^2(y_k, y), \qquad \lambda_k \geq 0. \tag{13}$$

The correction replaces the sampled positive gradient by v_k^+ and includes the negative component through v_k^-, giving $\operatorname{grad} \varphi_k(y_k) = v_k$. Thus, the model gradient has exactly the error of the variance-reduced estimate. Lemma B.5 proves this identity and the uniform bounds on v_k and q_k for every sampled model. The next step is to choose a domain where iterates remain in a region covered by Assumption 2.1. For $y_0 \in \operatorname{int}(\mathcal{M}_{\mathrm{ex}})$, choose

$$\max\{W_{\mathrm{abs}} r, d(c, y_0)\}/R < \alpha < 1, \qquad \mathcal{K}_\alpha := \overline{B}_{\alpha R}(c), \qquad \mathcal{M}_k := \overline{B}_{\bar{r}}(y_k) \cap \mathcal{K}_\alpha. \tag{14}$$

Such an α exists by Assumption 2.2 and the initialization condition. The ball $\mathcal{K}_\alpha$ retains all regression minimizers by Lemma B.4, and the iterates have distance at least $(1 - \alpha)R$ from $\partial \mathcal{M}_{\mathrm{ex}}$ by Proposition B.6. The radius $\bar{r}$ keeps each model domain within the region of the geometric bounds.

Now we need to select a regularizer that ensures the model is strongly convex. For $\mu_\star, \nu > 0$ and $\eta_k \in (0, 1]$, let $[a]_+ := \max\{a, 0\}$ and set

$$m_k^\star := \max\left\{\mu_\star/\eta_k, 2\|v_k\|_{y_k}/\bar{r} + \nu\right\}, \qquad \lambda_k := [L_{\log}\|q_k\|_{y_k} - 2W_+\delta + m_k^\star]_+/\delta_p. \tag{15}$$

The geometric bounds then give: $\operatorname{Hess}\varphi_k \succeq \big(2W_+\delta + \lambda_k\delta_p - L_{\log}\|q_k\|_{y_k}\big)I \succeq m_k^\star I$. The first term in $m_k^\star$ controls the strength of this convexity through η_k. The second ensures that the unique minimizer over $\mathcal{M}_k$ lies strictly inside the local ball, although it may lie on $\partial\mathcal{K}_\alpha$.

Finally, we define an approximate solution to the inner problems. Let $N_{\mathcal{K}_\alpha}(y)$ be the outward normal cone: $\{0\}$ in the interior and $\{-a\operatorname{Log}_y(c)/(\alpha R) : a \geq 0\}$ on the boundary. For $0 \leq \zeta_{\text{in}} < \mu_\star$, we accept $y_{k+1} \in \mathcal{M}_k$ if some $n_{k+1} \in N_{\mathcal{K}_\alpha}(y_{k+1})$ satisfies

$$d(y_k, y_{k+1}) < \bar{r}, \qquad \|\operatorname{grad}\varphi_k(y_{k+1}) + n_{k+1}\|_{y_{k+1}} \leq (\zeta_{\text{in}}/\eta_k)d(y_k, y_{k+1}). \tag{16}$$

Proposition B.6 shows that the unique exact minimizer lies strictly inside the local ball and satisfies (16) with zero residual. Lemma B.7 bounds the distance and objective-value errors allowed by this stopping rule. In the interior of $\mathcal{K}_\alpha$, the condition reduces to a bound on the model gradient.

CONVERGENCE GUARANTEES: Since $\operatorname{grad}\varphi_k(y_k) = v_k$, the accuracy of the model gradient is determined by the errors in the component estimates. To relate these errors to the progress of the algorithm, define

$$d_k := d(y_k, y_{k+1}), \qquad \mathcal{E}_k := \sum_{\sigma\in\mathcal{S}} \|v_k^\sigma - \operatorname{grad} F_\sigma(y_k)\|_{y_k}^2, \qquad \chi := \mu_\star - \zeta_{\text{in}} > 0. \tag{17}$$

Lemma B.8 relates each displacement to the objective decrease and the true gradient norm. The next theorem converts the accumulated error bound (18) into an expected stationarity guarantee. The subsequent corollaries verify this bound for different estimators by varying their scale, batch, refresh, and mixing parameters.

Theorem 3.1 *Let Assumptions 2.1 and 2.2 hold, $y_0 \in \operatorname{int}(\mathcal{M}_{\text{ex}})$, choose α as in* (14)*, and fix $\mu_\star, \nu > 0$ and $0 \leq \zeta_{\text{in}} < \mu_\star$. Also, let $\eta_k > 0$ be a deterministic step size with $0 < \eta_k \leq \min\{1, \chi/(2L_f)\}$. If there exist constants $0 \leq \theta < \chi^2/4$ and $0 \leq C_{\text{err}} < \infty$, such that*

$$\sum_{k=0}^{K-1} \mathbb{E}[\eta_k\mathcal{E}_k] \leq \theta \sum_{k=0}^{K-1} \mathbb{E}\left[\frac{d_k^2}{\eta_k}\right] + C_{\text{err}}, \tag{18}$$

for every horizon K, then the iterates generated by Algorithm 1 have the following properties:

(i) Feasibility and well-defined updates. *Every sampled trajectory remains in $\mathcal{K}_\alpha$, at distance at least $(1-\alpha)R$ from $\partial\mathcal{M}_{\text{ex}}$. Each model φ_k has a unique minimizer over $\mathcal{M}_k$, and this minimizer satisfies* (16) *with zero residual.*

(ii) Expected stationarity. *There exists $C < \infty$, independent of K, such that for every K,*

$$\sum_{k=0}^{K-1} \mathbb{E}[\eta_k\|\operatorname{grad} f(y_k)\|_{y_k}^2] \leq C. \tag{19}$$

Part (i) follows from the model construction and does not require the error bound (18). For part (ii), this bound limits the accumulated estimation error relative to the squared displacements, up to a finite remainder. It allows biased estimates, as arise from the recursive updates and clipping. The proof is given in Appendix B.3.2. Proposition B.9 gives the quantitative error-transfer bound, with an explicit constant C in (53). To apply the theorem, we bound the estimation error for each method in Algorithm 1. Write b_σ for the batch size in D and c_σ for the batch size in V or U, whenever these terms are used, and set

$$L_\sigma := W_\sigma L_\ell, \qquad A_\Delta := L_+^2/b_+ + L_-^2/b_-, \qquad A_{\text{sv}} := L_+^2/c_+ + L_-^2/c_- \tag{20}$$

For SVRRM and STORM, the two terms share a batch, so $c_\sigma = b_\sigma$. The constants A_Δ and A_{sv} control the variance of the gradient differences in the recursive and snapshot estimates, respectively. Lemma B.10 establishes the conditional variance bounds, including the same-batch gradient-difference estimate (63).

Corollary 3.2 (PAGE) *Let Assumptions 2.1 and 2.2 hold, with the initialization, model parameters, and fixed batch sizes specified in Theorem 3.1. If PAGE uses exact initialization, refresh probability $p \in (0, 1]$, and step-sizes satisfying $0 < \eta \leq \min\{1, \chi/(2L_f)\}$, and $\eta^2 A_\Delta(1-p)/p < \chi^2/4$, then the iterates generated by Algorithm 1 with the PAGE estimator satisfy* (19) *for every $K \geq 1$. In this case,* (18) *holds with $\theta = \eta^2(1-p)A_\Delta/p$ and $C_{\text{err}} = 0$.*

For periodic refreshes, the error also depends on the epoch length and, when recursive and snapshot estimates are combined, on the mixing weight. The next result makes these dependencies explicit.

Corollary 3.3 (Periodic refreshes) *Let Assumptions 2.1 and 2.2 hold, with initialization, model parameters, and fixed batch sizes as in Theorem 3.1. Consider SARAH/SPIDER, SVRG, SVRRM, or SRVRG with exact initialization, integer refresh period $H \geq 1$, and fixed $\beta \in (0,1)$ for SVRRM/SRVRG. For SARAH/SPIDER, SVRG, and SVRRM/SRVRG, respectively, define*

$$A_H := A_\Delta H, \quad A_H := A_{\mathrm{sv}} H^2/2, \quad A_H := \kappa\big[\beta^2 A_{\mathrm{sv}} H^2 + (1-\beta)^2 A_\Delta\big]/(\beta(2-\beta)), \tag{21}$$

where $\kappa = 2$ for SVRRM and $\kappa = 1$ for SRVRG. If the deterministic constant scale satisfies $0 < \eta \leq \min\{1, \chi/(2L_f)\}$ and $\eta^2 A_H < \chi^2/4$, then the iterates generated by Algorithm 1 satisfy (19) *for $K = H, 2H, \ldots$, with* (18) *holding for $\theta = \eta^2 A_H$ and $C_{\mathrm{err}} = 0$.*

These two corollaries give an $O(K^{-1})$ bound on the expected squared gradient norm at a uniformly selected iterate; see Appendix B.3.3. The rate counts outer model updates; their computational cost also includes model minimizations and full-gradient refreshes. Although η is constant, λ_k can vary with the sampled model and gradient estimate through (15). On the other hand, STORM and Hybrid-SGD use a sampled initialization and do not perform exact refreshes. For these methods, we decrease η_k and the mixing weight together to obtain almost-sure stationarity.

Corollary 3.4 (STORM and Hybrid-SGD) *Let Assumptions 2.1 and 2.2 hold, with initialization, model parameters, and fixed batch sizes as in Theorem 3.1. Consider STORM or Hybrid-SGD with sampled initialization, $\beta_k = c_\beta \eta_k$, $c_\beta > 0$, and $A = 2A_\Delta$ or $A = A_\Delta$, respectively. If $0 < \bar\eta \leq \min\{1, \chi/(2L_f), 1/c_\beta\}$, $A\bar\eta/c_\beta < \chi^2/4$, and the history-measurable scales $0 < \eta_k \leq \bar\eta$ satisfy $\sum_{k=0}^{\infty} \eta_k = \infty$ almost surely, and $\mathbb{E}\big[\sum_{k=0}^{\infty} \eta_k^2\big] < \infty$. Then the iterates generated by Algorithm 1 with the selected estimator satisfy*

$$\mathcal{E}_k \to 0, \qquad \|\operatorname{grad} f(y_k)\|_{y_k} \to 0 \quad \textit{almost surely}. \tag{22}$$

Moreover, (18) *holds with $\theta = A\bar\eta/c_\beta$ and finite C_{err}.*

All three corollaries allow a fixed nonlinear batch size b_g and inexact updates satisfying (16). The proofs of Corollaries 3.2, 3.3, and 3.4 are given in Appendices B.4.4, B.4.6, and B.5.4.

4 Numerical Analysis

We evaluate VERI-FRIDA using synthetic data generated on a torus and real-world data on $\mathrm{SO}(3)$ (Agricola et al., 2011) for robot motion regression. Additionally, the product manifold $S^2 \times S^1$ is also studied, and the numerical results are included in Appendix A.2. We compare its SVRG, SVRRM, SRVRG, STORM, Hybrid, PAGE, and SARAH variants with their explicit Riemannian counterparts, Riemannian SGD, and deterministic FRIDA. For both experiments, we consider total negative weights $W_- \in \{0.3, 0.9\}$ and two minibatch sizes chosen according to the dataset size, while keeping the observations fixed. Stepsizes are tuned separately for each method, negative-weight setting, and minibatch size, using seeds distinct from those used for final evaluation. We report the median full-gradient norm and interquartile range across ten random seeds. Deterministic FRIDA provides a single reference curve.

REGRESSION ON THE TORUS: We apply Algorithm 1 to an embedded Torus with angular coordinates $(\theta, \phi) \in \mathbb{S}^1 \times \mathbb{S}^1$, major radius $a = 2$, and minor radius $r = 0.7$. Its embedded and induced metrics are $\mathcal{F}(\theta,\phi) = \big((a + r\cos\phi)\cos\theta,\ (a + r\cos\phi)\sin\theta,\ r\sin\phi\big)$, and $\mathrm{d}s^2 = (a + r\cos\phi)^2 \mathrm{d}\theta^2 + r^2 \mathrm{d}\phi^2$. Its sectional curvature is $K(\phi) = \cos\phi/[r(a + r\cos\phi)]$, which is positive on the outer region and negative on the inner region. The optimization region intersects both curvature signs, and we use the uniform bounds $-100/91 \leq K \leq 100/189$, inherited from the entire torus. We compute Exp by numerical geodesic integration and Log by Newton shooting. Figure 1 illustrates this geometry, showing the signed responses and trajectories of deterministic FRIDA, VERI-FRIDA (SRVRG), and R-SRVRG from four initial points. The dotted curve separates regions of positive and negative Gaussian curvature. Appendix C.1 presents detailed

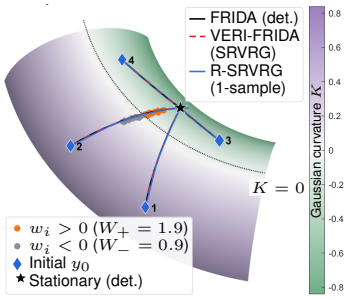

Figure 1: Fréchet regression on the Torus.

geodesic equations, the parallel-transport equations, and the intrinsic gradient formula. We draw $m = 1000$ predictors $X_i \sim \mathrm{Unif}[0, 1]$ and define

$$t(x) = 0.25 + 0.04(2x - 1), \quad c(x) = \mathcal{F}((\pi/3)t(x),\, 2\pi t(x))\,.$$

Responses are generated as $Y_i = \mathrm{Exp}_{c(X_i)}(\varepsilon_i)$ with isotropic Gaussian tangent noise of standard deviation 0.008. The data span both curvature signs, and we select an extrapolation query for each prescribed negative mass. All methods use the common initialization $y_0 = \mathrm{Exp}_{c(1)}(v_0)$, where v_0 has orthonormal tangent coordinates $(0.45, -0.80)$. This point lies approximately 0.83 geodesic units from the intrinsic data mean. We use per-sign minibatch sizes $b \in \{1, 4\}$ and run each stochastic method for 100 iterations.

Figure 1 shows optimization trajectories on the mixed-curvature torus with $W_- = 0.9$. The initializations lie in regions of both positive and negative curvature, separated by the dotted $K = 0$ curve. From each of the four initial points, deterministic FRIDA, VERI-FRIDA (SRVRG), and one-sample R-SRVRG follow similar trajectories, with all paths approaching the same stationary point. Figure 2 compares the full-gradient norm against the iteration count for all methods. The torus results in panels (a)–(d) show that the SVRG, SVRRM, SRVRG, PAGE, and SARAH variants of VERI-FRIDA reach small gradient norms in fewer iterations than their explicit Riemannian counterparts, with convergence curves close to deterministic FRIDA. STORM and Hybrid make substantially less progress within the iteration budget. Increasing the negative mass from 0.3 to 0.9 delays the rapid decrease in gradient norm, while the main qualitative trends persist across both minibatch sizes. Table 1 shows that, on the torus, VERI-FRIDA requires approximately 21–29 ms per inner solve, compared with 607–771 ms for deterministic FRIDA. These costs vary little across the tested negative masses and minibatch sizes.

Table 1: Inner minimization time (ms): median of per-seed means over 10 seeds.

	Torus ($m = 1000$)				SO(3) ($m = 100k$)			
W_-	0.3		0.9		0.3		0.9	
$\lvert B\rvert$	2	8	2	8	32	64	32	64
VERI-FRIDA								
SVRG	21.7	22.4	21.6	21.7	0.88	0.88	0.89	0.89
SVRRM	21.6	21.7	22.1	21.7	0.88	0.88	0.89	0.89
SRVRG	21.4	21.2	21.7	21.9	0.89	0.89	0.89	0.89
STORM	28.2	27.4	26.9	25.5	1.37	1.36	1.36	1.36
Hybrid-SGD	29.0	29.0	28.5	27.0	1.37	1.36	1.38	1.36
PAGE	21.3	21.2	22.7	21.8	0.89	0.88	0.90	0.90
SARAH	21.4	21.2	22.0	21.2	0.89	0.88	0.88	0.89
FRIDA (det.)	771.3		607.1		1002.9		865.4	

ROBOT MOTION REGRESSION ON $SO(3)$: We construct a manifold-valued regression problem from the TartanGround dataset (Patel et al., 2025), using the 240 ANYmal trajectories and their ground-truth poses recorded at 10 Hz. Each pose consists of a position $p_t \in \mathbb{R}^3$ and a unit quaternion, which we convert to a rotation matrix $R_t \in \mathrm{SO}(3)$. Here, Fréchet regression is used to predict a representative future change in the robot's attitude from its current translational speed. Appendix C.2 specifies the angle metric and the exponential, logarithm, and parallel-transport formulas, as well as squared-distance Hessian bounds. Specifically, the predictor and response are $X_t = 10\|p_{t+1} - p_t\|_2$ and $Y_t = R_t^\top R_{t+5} \in \mathrm{SO}(3)$, so the fitted Fréchet response at a query speed x represents the typical relative orientation change over the following 0.5 seconds. The rotation group $\mathrm{SO}(3)$ is a compact three-dimensional Lie group and, under the standard bi-invariant metric used here, has constant positive sectional curvature $1/4$. Processing yields 309,169 valid predictor-response pairs, from which we retain a certified pool containing the responses in a common normal neighborhood and draw a sample of 100,000 observations. We standardize the speed predictor by using two total minibatches $|B| \in \{32, 64\}$.

In terms of convergence in gradient norm, the $\mathrm{SO}(3)$ panels of Figure 2 show that the SVRG, SVRRM, SRVRG, PAGE, and SARAH variants of VERI-FRIDA consistently converge faster than their explicit Riemannian counterparts across all extrapolation regimes and minibatch sizes. For $W_- = 0.3$, their convergence closely matches that of deterministic FRIDA, while for the stronger extrapolation regime $W_- = 0.9$, several VERI-FRIDA variants converge faster than deterministic FRIDA. In terms of inner-solve cost, Table 1 shows that VERI-FRIDA requires only about 0.88-1.38 ms per inner problem on $\mathrm{SO}(3)$, compared with 865.4-1002.9 ms for deterministic FRIDA. This corresponds to roughly a 630-$1140\times$ reduction in time per inner minimization while retaining comparable, and in the stronger extrapolation regime sometimes faster, outer-iteration convergence. These results show that VERI-FRIDA retains the effectiveness of the implicit FRIDA update while substantially reducing the computational cost of its model subproblems.

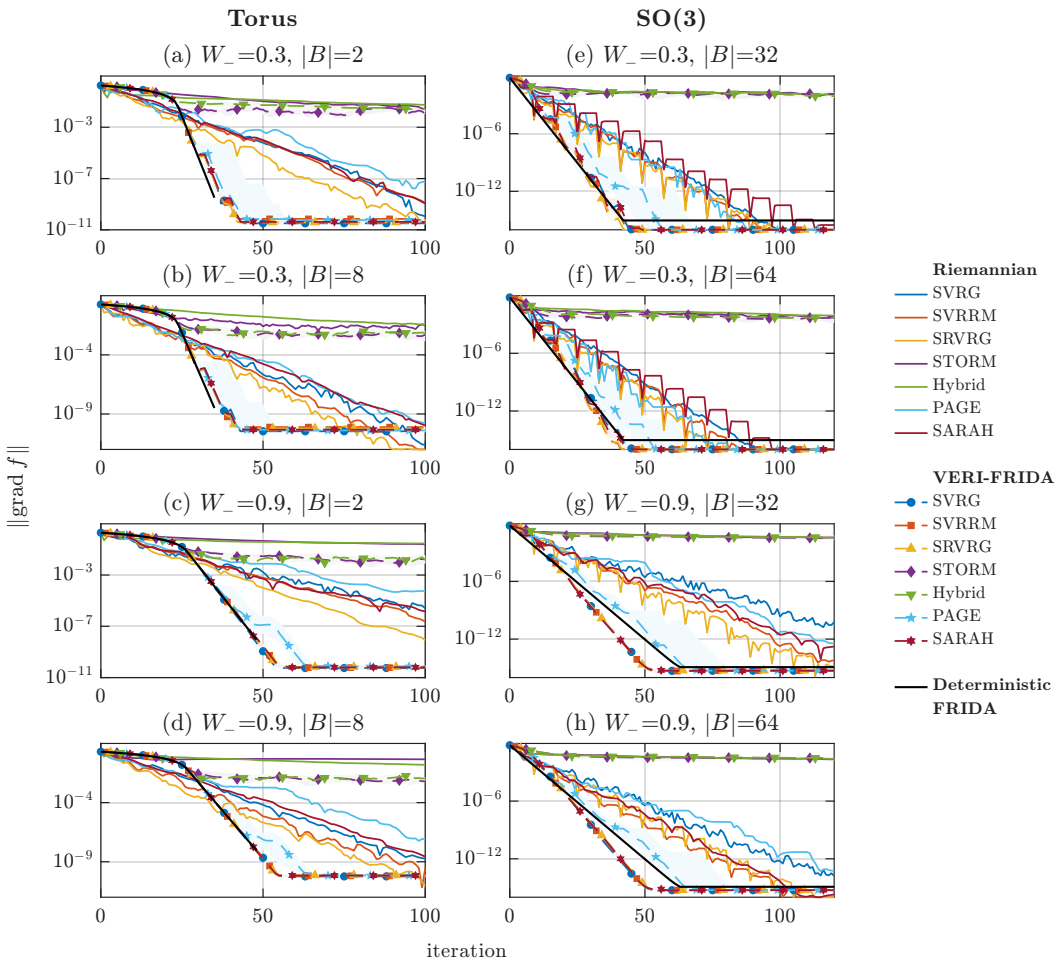

Figure 2: Convergence of VERI-FRIDA and Riemannian variance-reduced baselines on the mixed-curvature torus ((a)–(d)) and SO(3) ((e)–(h)). We compare negative masses $W_- \in \{0.3, 0.9\}$ and total estimator minibatch sizes $|B| \in \{2, 8\}$ for the torus and $|B| \in \{32, 64\}$ for SO(3). Stochastic curves show the median full-gradient norm over ten random seeds; shaded regions indicate the interquartile range for VERI-FRIDA. Dashed lines with filled markers denote VERI-FRIDA, solid lines with open markers denote the corresponding Riemannian baselines, and the black curve denotes deterministic FRIDA, indexed by outer iterations. The gradient norm is shown on a logarithmic scale, and iterations are shown on a linear scale starting at $k = 0$.

We additionally report convergence with respect to wall-clock time in Appendix A, complementing the iteration-based comparisons in Figure 2. These results in Figure 3 show that VERI-FRIDA achieves computational gains over deterministic FRIDA.

5 CONCLUSION

We introduced VERI-FRIDA, a stochastic implicit framework for signed Fréchet regression on manifolds with bounded curvature. The key idea is to combine a small sampled nonlinear model with variance-reduced gradient information through an exact tangent correction, while adaptive regularization ensures well-posed implicit updates. Building on this construction, we established a unified convergence analysis that accommodates several variance-reduced estimators, yielding an $O(K^{-1})$ stationarity guarantee for refresh-based schemes and almost-sure convergence for fully stochastic variants. Moreover, experiments on a mixed-curvature torus and real-world robot motion regression on SO(3) show that VERI-FRIDA retains the convergence behavior of deterministic FRIDA while substantially reducing the cost of its inner problems and improving over explicit Riemannian baselines. Future work will relax some of the geometric, weight, and estimator conditions considered here to obtain a more general implicit variance-reduction framework.

AI USE STATEMENT

In this work, we used generative AI tools for proofreading, editing, and typesetting; proof verification; and coding support for large-scale experiments. We have not used generative AI tools for problem formulation, algorithmic design, and conceptual ideas, proof structures, or final proofreading and editing of the text, images, and experiments. Additionally, we used generative AI tools to format tables. We have reviewed all AI-assisted work. For example, we manually verified all derivations and text. No LLM's were used for literature review. LLM-generated code was verified and tested for correctness by all authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Bijan Afsari. Riemannian l^p center of mass: Existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011. doi: 10.1090/S0002-9939-2010-10541-5.

Ilka Agricola, Julia Becker-Bender, and Thomas Friedrich. On the topology and the geometry of so (3)-manifolds. *Annals of Global Analysis and Geometry*, 40(1):67–84, 2011.

Yldenilson Torres Almeida, Jo ao Xavier da Cruz Neto, Paulo Roberto Oliveira, and Jo ao Carlos de Oliveira Souza. A modified proximal point method for dc functions on hadamard manifolds. *Computational Optimization and Applications*, 76(3):649–673, 2020. ISSN 1573-2894. doi: 10.1007/s10589-020-00173-3. URL https://doi.org/10.1007/s10589-020-00173-3.

Ronny Bergmann, Orizon P. Ferreira, Elianderson M. Santos, and João Carlos O. Souza. The difference of convex algorithm on Hadamard manifolds. *Journal of Optimization Theory and Applications*, 201(1):221–251, 2024. doi: 10.1007/s10957-024-02392-8.

Silvère Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. doi: 10.1109/TAC.2013.2254619.

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

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/b8002139cdde66b87638f7f91d169d96-Abstract.html.

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.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/1543843a4723ed2ab08e18053ae6dc5b-Abstract.html.

Karsten Grove and Hermann Karcher. How to conjugate c 1-close group actions. *Mathematische Zeitschrift*, 132(1):11–20, 1973.

Andi Han and Junbin Gao. Riemannian stochastic recursive momentum method for non-convex optimization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2505–2511, 2021. doi: 10.24963/ijcai.2021/345.

Su I Iao, Yidong Zhou, and Hans-Georg Müller. Deep Fréchet regression. *Journal of the American Statistical Association*, 120(551):1437–1448, 2025. doi: 10.1080/01621459.2025.2507982. URL https://doi.org/10.1080/01621459.2025.2507982.

Robert T. Jantzen. Geodesics on the torus and other surfaces of revolution clarified using undergraduate physics tricks with bonus: Nonrelativistic and relativistic kepler problems, 2010. Lecture notes.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/hash/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Abstract.html.

Avetik Karagulyan, Egor Shulgin, Abdurakhmon Sadiev, and Peter Richtárik. SPAM: Stochastic proximal point method with momentum variance reduction for non-convex cross-device federated learning. *arXiv preprint arXiv:2405.20127*, 2024. doi: 10.48550/arXiv.2405.20127. URL https://arxiv.org/abs/2405.20127.

Hiroyuki Kasai, Hiroyuki Sato, and Bamdev Mishra. Riemannian stochastic recursive gradient algorithm. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2516–2524, 2018.

Kwang-Rae Kim, Ian L Dryden, Huiling Le, and Katie E Severn. Smoothing splines on riemannian manifolds, with applications to 3d shape space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(1):108–132, 2021.

Hoai An Le Thi and Tao Pham Dinh. Dc programming and dca: Thirty years of developments. *Mathematical Programming*, 169(1):5–68, 2018. doi: 10.1007/s10107-018-1235-y.

Hoai An Le Thi, Hoai Minh Le, Duy Nhat Phan, and Bach Tran. Stochastic dca for the large-sum of non-convex functions problem and its application to group variable selection in classification. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3394–3403, 2017.

Hoai An Le Thi, Hoang Phuc Hau Luu, Hoai Minh Le, and Tao Pham Dinh. Stochastic dca with variance reduction and applications in machine learning. *Journal of Machine Learning Research*, 23(206):1–44, 2022. URL http://jmlr.org/papers/v23/21-1146.html.

John M. Lee. *Introduction to Riemannian Manifolds*. Springer, 2 edition, 2018. doi: 10.1007/978-3-319-91755-9.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6286–6295. PMLR, 2021. URL https://proceedings.mlr.press/v139/li21a.html.

Shichen Liao, Yan Liu, Congying Han, and Tiande Guo. Momentum-based variance-reduced stochastic Bregman proximal gradient methods for nonconvex nonsmooth optimization. *Expert Systems with Applications*, 266:125960, 2025. doi: 10.1016/j.eswa.2024.125960.

Yan Liu, Mingjie Chen, Chaojie Ji, Hao Zhang, and Ruxin Wang. SERENA: A unified stochastic recursive variance reduced gradient framework for Riemannian non-convex optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38144–38168. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25e.html.

Julien Mairal. Stochastic majorization–minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems 26*, pp. 2283–2291, 2013.

Duc Toan Nguyen and César A. Uribe. Fréchet regression on the Bures–Wasserstein manifold. *arXiv preprint arXiv:2604.03566*, 2026. doi: 10.48550/arXiv.2604.03566. URL https://arxiv.org/abs/2604.03566.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2613–2621. PMLR, 2017. URL https://proceedings.mlr.press/v70/nguyen17b.html.

Manthan Patel, Fan Yang, Yuheng Qiu, Cesar Cadena, Sebastian Scherer, Marco Hutter, and Wenshan Wang. TartanGround: A large-scale dataset for ground robot perception and navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025. URL https://arxiv.org/abs/2505.10696.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006. doi: 10.1007/s11263-005-3222-z.

Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019. doi: 10.1214/17-AOS1624.

Peter Petersen. *Riemannian geometry*. Springer, 2006.

Tao Pham Dinh and Hoai An Le Thi. Convex analysis approach to dc programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1):289–355, 1997.

Herbert Robbins and David Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. In Jagdish S. Rustagi (ed.), *Optimizing Methods in Statistics*, pp. 233–257. Academic Press, 1971.

Olinde Rodrigues. Des lois géométriques qui régissent les déplacements d'un système solide dans l'espace, et de la variation des coordonnées provenant de ces déplacements considérés indépendamment des causes qui peuvent les produire. *Journal de mathématiques pures et appliquées*, 5:380–440, 1840.

Hiroyuki Sato, Hiroyuki Kasai, and Bamdev Mishra. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472, 2019. doi: 10.1137/17M1116787.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1000–1008. PMLR, 2014. URL https://proceedings.mlr.press/v32/shamir14.html.

Wookyeong Song, Paromita Dubey, Hans-Georg Müller, and Alexander Petersen. Inference for Fréchet regression. *arXiv preprint arXiv:2605.19519*, 2026. doi: 10.48550/arXiv.2605.19519. URL https://arxiv.org/abs/2605.19519.

J. C. O. Souza and P. R. Oliveira. A proximal point algorithm for dc functions on hadamard manifolds. *Journal of Global Optimization*, 63(4):797–810, 2015. ISSN 1573-2916. doi: 10.1007/s10898-015-0282-7. URL https://doi.org/10.1007/s10898-015-0282-7.

Quoc Tran-Dinh, Nhan H. Pham, Dzung T. Phan, and Lam M. Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, 191:1005–1071, 2022. doi: 10.1007/s10107-020-01583-1.

Cheik Traoré, Vassilis Apidopoulos, Saverio Salzo, and Silvia Villa. Variance reduction techniques for stochastic proximal point algorithms. *Journal of Optimization Theory and Applications*, 203(2):1910–1939, 2024. doi: 10.1007/s10957-024-02502-6.

Jeff A. Viaclovsky. Math 865: Topics in Riemannian geometry. https://www.math.uci.edu/~jviaclov/courses/865_Fall_2011.pdf, 2011. Course notes.

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

Haoshu Xu and Hongzhe Li. Wasserstein F-tests for Fréchet regression on Bures–Wasserstein manifolds. *Journal of Machine Learning Research*, 26(77):1–123, 2025. URL https://jmlr.org/papers/v26/24-0493.html.

Lei Yan, Xin Zhang, Zhou Lan, Dipankar Bandyopadhyay, Yichao Wu, and the Alzheimer's Disease Neuroimaging Initiative. Variable screening and spatial smoothing in Fréchet regression with application to diffusion tensor imaging. *The Annals of Applied Statistics*, 19(1):655–679, 2025. doi: 10.1214/24-AOAS1978. URL https://doi.org/10.1214/24-AOAS1978.

Hongyi Zhang, Sashank J. Reddi, and Suvrit Sra. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems 29*, 2016.

Pan Zhou, Xiao-Tong Yuan, and Jiashi Feng. Faster first-order methods for stochastic non-convex optimization on riemannian manifolds. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 138–147, 2019.

Yamin Zhou and César A Uribe. Proximal DCA for Fréchet Regression on Riemannian Manifolds with Bounded Curvature. *arXiv preprint arXiv:2605.23097*, 2026.

Yidong Zhou and Hans-Georg Müller. Network regression with graph laplacians. *Journal of Machine Learning Research*, 23(320):1–41, 2022. URL https://jmlr.org/papers/v23/22-0681.html.

Yidong Zhou, Su I Iao, and Hans-Georg Müller. End-to-end deep learning for predicting metric space-valued outputs. *Journal of Machine Learning Research*, 27(132):1–38, 2026. URL https://jmlr.org/papers/v27/25-2364.html.

ORGANIZATION OF THE APPENDIX

Appendix A. Additional Numerical Results.

- Time complexity comparison between stochastic and deterministic methods 15
- Simulation Results on $S^2 \times S^1$ 16

Appendix B. Convergence theory - Proof of the main Theorem 3.1.

- Geometry of the search region 18
- Sampled model and implicit update 21
- From error budget to stationarity 24
- Exact-refresh estimators 26
- Refresh-free estimators 30
- Infinite-run consequences 33

Appendix C. Geometry of the manifolds in the numerical experiments.

- The embedded torus 35
- $SO(3)$ - Rotations with the angle metric 36
- $S^2 \times S^1$ - Product of the sphere and circle 38

A Additional Numerical Results

A.1 Time complexity comparison between stochastic and deterministic methods

To complement the iteration-based comparison, we also measure convergence of the full-gradient norm with respect to wall-clock time, accounting for both gradient-estimator updates and the cost of the implicit subproblems. The results are illustrated in Figure 3.

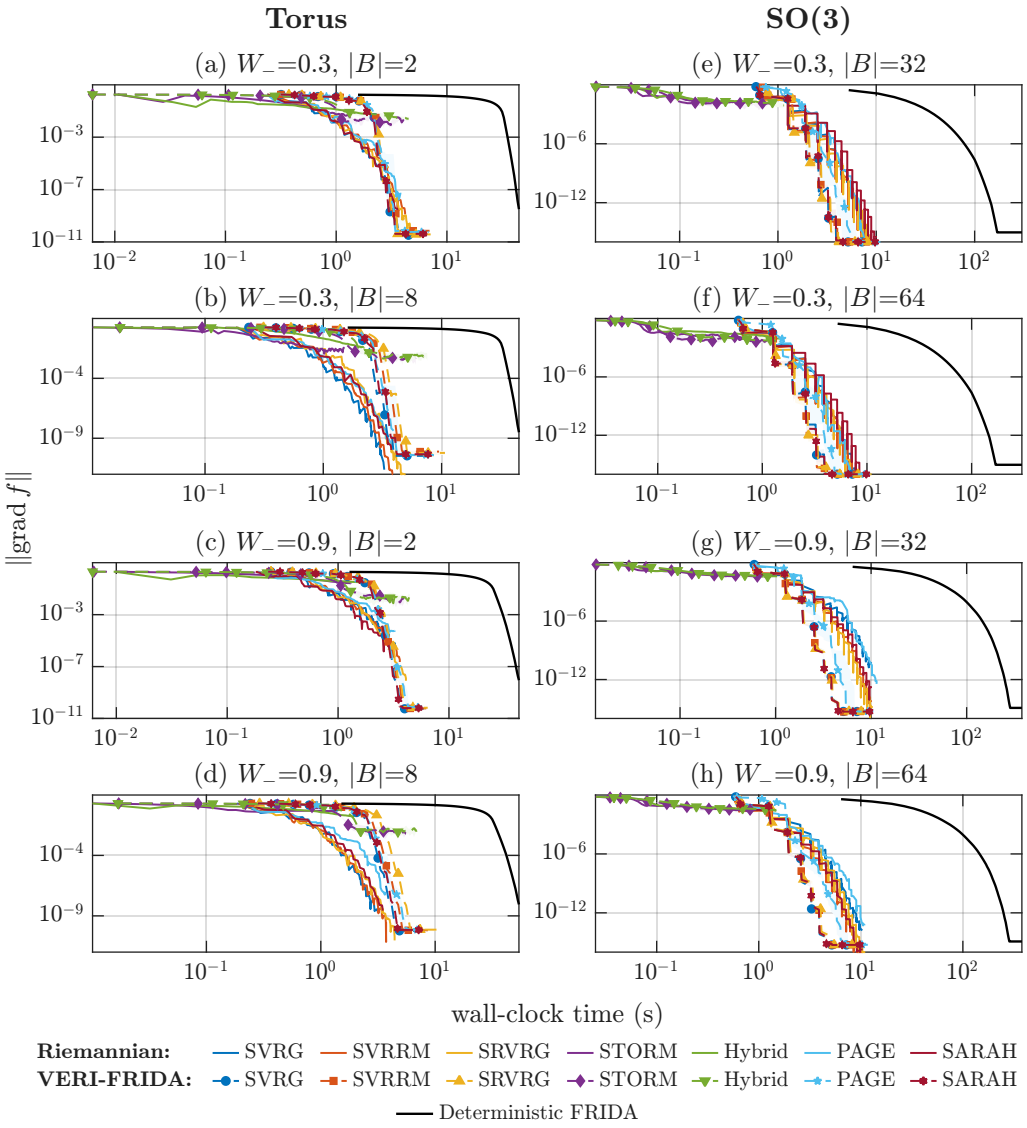

Figure 3: Wall-clock convergence of VERI-FRIDA and Riemannian variance-reduced baselines on the mixed-curvature torus ((a)-(d)) and SO(3) ((e)-(h)). We compare negative masses $W_- \in \{0.3, 0.9\}$ and total estimator minibatch sizes $|B| \in \{2, 8\}$ for the torus and $|B| \in \{32, 64\}$ for SO(3). Stochastic curves show the median full-gradient norm over ten random seeds, while the black curve denotes deterministic FRIDA. Wall-clock time includes gradient-estimator updates and the solution of the implicit subproblems. Both axes are shown on logarithmic scales.

Torus: On the torus, the SVRG, SVRRM, SRVRG, PAGE, and SARAH variants of VERI-FRIDA reach small gradient norms within a few seconds and are substantially faster than deterministic FRIDA in wall-clock time. Compared with their explicit Riemannian counterparts, however, the wall-clock performance is generally comparable, with some explicit methods slightly faster for

larger minibatches. This reflects the higher cost of the implicit model solve on the torus, while STORM and Hybrid again make limited progress.

$\mathrm{SO}(3)$**:** On $\mathrm{SO}(3)$, the wall-clock advantage of VERI-FRIDA is clearer. The successful variants achieve numerical precision in only a few seconds and generally converge earlier than their explicit Riemannian counterparts across all four settings, with this advantage becoming clearer under stronger extrapolation ($W_- = 0.9$). Deterministic FRIDA, in contrast, requires substantially more wall-clock time. Thus, the inexpensive implicit solves on $\mathrm{SO}(3)$ translate directly into end-to-end computational gains.

A.2 SIMULATION RESULTS ON $S^2 \times S^1$

We next apply Algorithm 1 to the product manifold $S^2 \times S^1$ equipped with the standard product metric. It has nonconstant sectional curvature ranging from 0 to 1: it equals 1 on planes tangent to the sphere, 0 on planes containing the circle direction, and takes intermediate values on other tangent planes. Within the safe region, the exponential map, logarithm map, and parallel transport admit closed-form, factor-wise expressions (Appendix C.3). We use $m = 1000$ equally spaced predictors $X_i = (i-1)/(m-1)$ and generate responses around

$$c(x) = \big((\sin\alpha(x), 0, \cos\alpha(x)), (\cos\theta(x), \sin\theta(x))\big), \quad \alpha(x) = 0.56(3x^2 - 2x^3), \quad \theta(x) = 0.32\pi x.$$

Specifically, $z_i = \mathrm{Exp}_{c(x_i)}(\varepsilon_i)$, where $\varepsilon_i = (0.018\omega_i e_i,\, 0.014\widetilde{\omega}_i)$ in sphere tangent and circle angular coordinates. Here $\omega_i, \widetilde{\omega}_i \sim \mathcal{N}(0,1)$, and e_i is a uniformly distributed unit tangent vector at the spherical component of $c(x_i)$, with all noise draws independent. These curve amplitudes and noise scales include a contraction factor of 0.4.

Under the product metric, the signed objective is the sum of a sphere term and a circle term. We therefore verify the negative-mass condition and the positivity of the squared-distance Hessian lower bound separately on each factor. We center the search region at $c(1/2)$ and restrict the sphere and circle components of the implicit iterates to distances at most 0.75 and 1.40 from their respective centers. Each update is restricted to a ball of radius 0.25 around the current iterate in the product metric. To cover these local balls, we use enlarged radii $1.00 = 0.75 + 0.25$ on S^2 and $1.65 = 1.40 + 0.25$ on S^1 when checking the existence and geometric conditions. The resulting negative-mass caps are approximately 1.070 and 1.072, and the common Hessian lower factor satisfies $\delta_\ell \approx 0.340 > 0$. We use the same responses for $W_- \in \{0.3, 0.9\}$, selecting query values $x \approx 0.9753$ and $x \approx 1.4035$, respectively. All methods share an initialization y_0, whose distances to the corresponding reference stationary points are approximately 1.06 and 1.54.

We use total estimator minibatch sizes $|B| \in \{2, 8\}$: VERI-FRIDA draws $b = |B|/2$ observations per sign, while the Riemannian baselines draw $|B|$ observations with probabilities proportional to $|w_i|$. The implicit positive-model batch contains 8 observations (4 for STORM), independently of $|B|$. Each stochastic method runs for 200 iterations over 10 seeds.

Figure 4(a)–(d) shows that the SVRG, SVRRM, SRVRG, PAGE, and SARAH variants of VERI-FRIDA reach gradient norms near 10^{-12} in approximately 50 iterations at $W_- = 0.3$ and 60 at $W_- = 0.9$. Their Riemannian counterparts require roughly 95–160 iterations at $W_- = 0.3$ and 125–200 or more at $W_- = 0.9$. The increase in negative mass, therefore, has a smaller effect on the iteration counts of VERI-FRIDA, while several Riemannian baselines benefit more visibly from increasing $|B|$ from 2 to 8. STORM and Hybrid converge much more slowly in both families within the iteration budget.

Panels (e)–(h) show that the iteration advantage translates into comparable wall-clock performance despite the implicit inner solves. The five faster variants in both families approach 10^{-12} in roughly 0.007–0.015 s across the four settings; the Riemannian baselines are faster at $W_- = 0.9$ and $|B| = 8$. Deterministic FRIDA takes approximately 0.03 s to reach its stopping level near 10^{-10}.

Consistent with the torus and $\mathrm{SO}(3)$ experiments, the refresh-based variants of VERI-FRIDA converge in fewer iterations while maintaining competitive wall-clock performance.

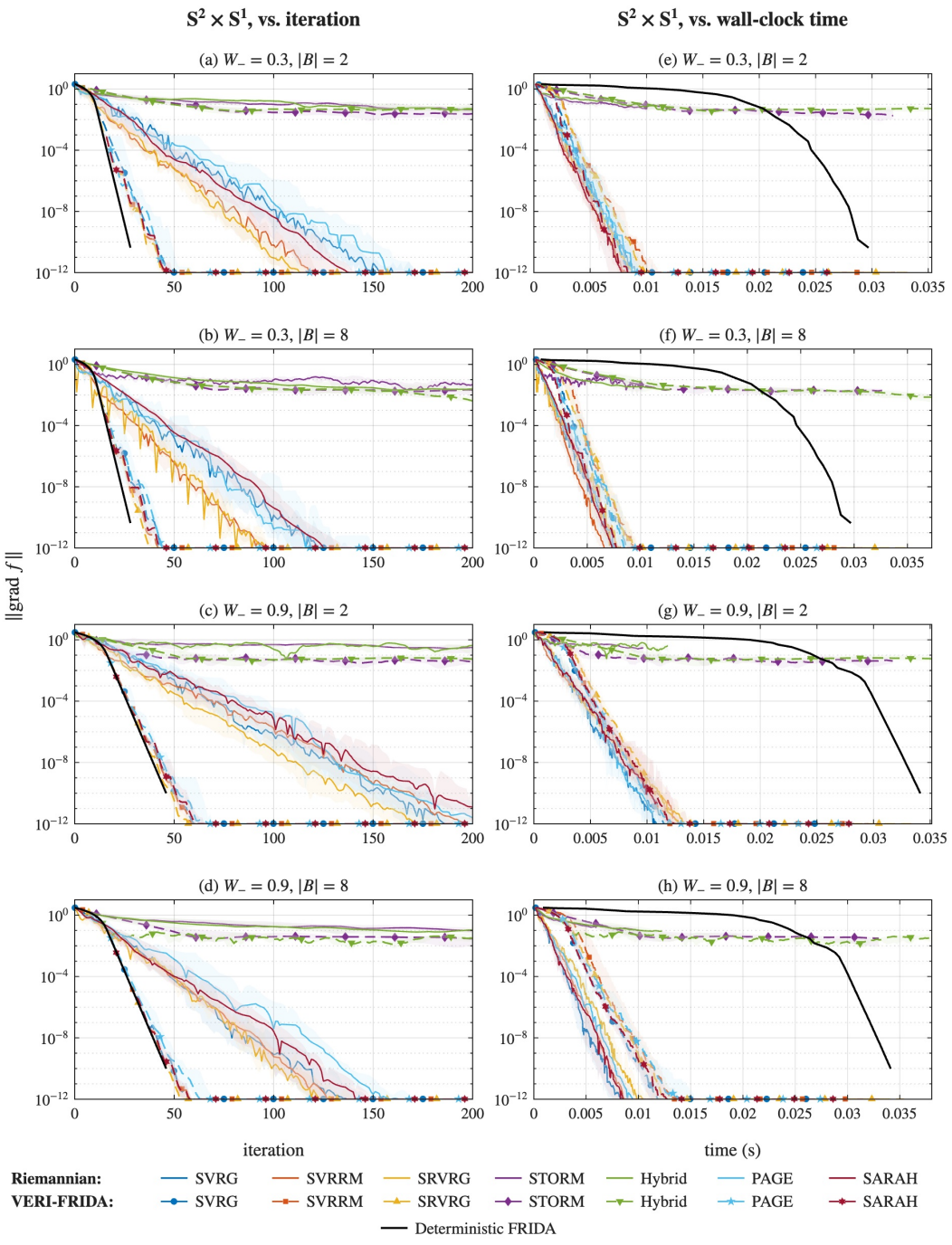

Figure 4: Signed Fréchet regression on $\mathbb{S}^2 \times \mathbb{S}^1$: full-gradient norm against iteration (a)–(d) and wall-clock time (e)–(h), for $W_- \in \{0.3, 0.9\}$ and total estimator batch sizes $|B| \in \{2, 8\}$. Solid coloured lines show the Riemannian baselines; dashed lines with markers show VERI-FRIDA. Curves and shaded regions indicate medians and interquartile ranges over 10 seeds. The black line shows one deterministic FRIDA run, counted by outer iterations. Times exclude post-run full-gradient diagnostics.

B CONVERGENCE THEORY - PROOF OF THE MAIN THEOREM 3.1.

We first establish the geometric estimates and analyze the implicit updates. We then show how a bound on the accumulated gradient-estimation error implies stationarity, and verify this bound for the different estimators. Finally, we study the behavior of infinite runs.

B.1 GEOMETRY OF THE SEARCH REGION

As in Section 2, we fix the observations and the query x, so the weights w_i in (1) remain fixed throughout the analysis. All randomness comes from the sampling steps in Algorithm 1. This subsection collects the notation and establishes the geometric estimates needed for the subsequent proofs, including those stated in Section 2.

B.1.1 WEIGHTS AND NOTATION

We begin with the weight identities, notation for successive iterates, and constants that will be used throughout the proofs.

The empirical weights in (1) satisfy

$$\sum_{i=1}^{m} w_i = 1 + \left(\frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})\right)^{\top} \widehat{\Sigma}^{-1}(x - \bar{x}) = 1.$$

Consequently,

$$W_+ - W_- = 1, \qquad W_+ = 1 + W_-, \qquad W_{\mathrm{abs}} = W_+ + W_- = 1 + 2W_-. \tag{23}$$

In particular, the positive sampling distribution is always defined. We sum only over the active signs $\mathcal{S} = \{\sigma \in \{+,-\} : W_\sigma > 0\}$. If $W_- = 0$, we set $F_- = 0$, $v_k^- = 0$, and $R_- = L_- = 0$; no negative-sign batch is drawn.

We use $\|\cdot\|$ for the norm in the tangent space containing its argument when that space is unambiguous. Parallel transport is always along the unique minimizing geodesic in the prescribed normal region. For successive iterates, write

$$s_k := \mathrm{Log}_{y_k}(y_{k+1}), \qquad d_k := \|s_k\| = d(y_k, y_{k+1}), \qquad P_k := P_{y_k \to y_{k+1}}. \tag{24}$$

Then $\mathrm{Log}_{y_{k+1}}(y_k) = -P_k s_k$, and P_k preserves inner products. We retain the constants from the main text and set

$$\begin{aligned} &G_\ell = 2D, \qquad L_\ell = 2\zeta, \qquad L_f = W_{\mathrm{abs}} L_\ell, \\ &R_\sigma = W_\sigma G_\ell, \quad L_\sigma = W_\sigma L_\ell, \quad V = R_+ + R_-, \qquad Q = 2R_+ + R_-. \end{aligned} \tag{25}$$

B.1.2 SQUARED-DISTANCE COMPARISON BOUNDS

The losses and proximal terms are built from squared distances. We therefore first bound the Hessian of the squared distance to a fixed point; these estimates will also help establish the convexity of the balls used by the algorithm.

For $t \geq 0$, define

$$\begin{aligned} a_+(t) &:= \begin{cases} \sqrt{\Lambda_+}\, t \cot(\sqrt{\Lambda_+}\, t), & \Lambda_+ > 0,\ t > 0, \\ 1, & \Lambda_+ = 0 \text{ or } t = 0, \end{cases} \\ a_-(t) &:= \begin{cases} \sqrt{\Lambda_-}\, t \coth(\sqrt{\Lambda_-}\, t), & \Lambda_- > 0,\ t > 0, \\ 1, & \Lambda_- = 0 \text{ or } t = 0. \end{cases} \end{aligned} \tag{26}$$

The first function is positive and nonincreasing for $\sqrt{\Lambda_+}\, t < \pi/2$; the second is nondecreasing and at least one.

Lemma B.1 (Comparison for a fixed center) *Let $z, y \in \overline{B}_\rho(c)$, let $s = d(z,y)$, and suppose $\sqrt{\Lambda_+}\, s < \pi/2$. Then $h_z = \frac{1}{2}d^2(z,\cdot)$ is smooth near y, $\operatorname{grad} h_z(y) = -\mathrm{Log}_y z$, and*

$$a_+(s)\|u\|^2 \leq \operatorname{Hess} h_z(y)[u,u] \leq a_-(s)\|u\|^2, \qquad u \in T_y\mathcal{M}. \tag{27}$$

If $y \neq z$, then $d(z,\cdot)$ has nonnegative Hessian at y.

Proof. Smoothness follows from the logarithm assumption. The first variation of the energy of the minimizing geodesic gives $\mathrm{d}h_z(y)[u] = -\langle \mathrm{Log}_y z, u\rangle_y$. At $y = z$, normal coordinates give $\mathrm{Hess}\, h_z(z) = I$.

Suppose $s > 0$, and let $\gamma : [0, s] \to \mathcal{M}$ be the unit-speed minimizing geodesic from z to y. It lies in $\overline{B}_\rho(c)$. Decompose $u = u_\parallel + u_\perp$ relative to $\dot\gamma(s)$. The radial part of $\mathrm{Hess}\, h_z$ is the identity and the mixed radial–transverse terms vanish. For a transverse vector, let J be the Jacobi field with $J(0) = 0$ and $J(s) = u_\perp$. The second variation formula gives

$$\mathrm{Hess}\, d(z,\cdot)(y)[u_\perp, u_\perp] = I_\gamma(J, J), \qquad I_\gamma(X, X) = \int_0^s \big(\|D_t X\|^2 - \langle \mathrm{Rm}(X, \dot\gamma)\dot\gamma, X\rangle\big)\, \mathrm{d}t.$$

The Jacobi field minimizes the index form among fields with these endpoints. To obtain the lower bound, use $\sec \le \Lambda_+$ in $I_\gamma(J, J)$ and minimize the resulting scalar constant-curvature quadratic form in a parallel frame. For $\Lambda_+ > 0$ its minimizer is $X(t) = \frac{\sin(\sqrt{\Lambda_+}t)}{\sin(\sqrt{\Lambda_+}s)} P_{y\to\gamma(t)} u_\perp$. For $\Lambda_+ = 0$ it is $X(t) = (t/s) P_{y\to\gamma(t)} u_\perp$. The corresponding minimum is $a_+(s)\|u_\perp\|^2/s$. The scalar quadratic form has this minimum because its zero-endpoint part is nonnegative for $s < \pi/\sqrt{\Lambda_+}$, with no restriction if $\Lambda_+ = 0$.

For the upper bound, test I_γ with the analogous hyperbolic field $X(t) = \frac{\sinh(\sqrt{\Lambda_-}t)}{\sinh(\sqrt{\Lambda_-}s)} P_{y\to\gamma(t)} u_\perp$, using the linear field when $\Lambda_- = 0$. The lower curvature bound then yields $I_\gamma(J, J) \le a_-(s)\|u_\perp\|^2/s$. Finally,

$$\mathrm{Hess}\, h_z(y)[u, u] = \|u_\parallel\|^2 + s\, \mathrm{Hess}\, d(z,\cdot)(y)[u_\perp, u_\perp].$$

Since $a_+(s) \le 1 \le a_-(s)$, the displayed bounds prove (27). The lower bound for the transverse Hessian of the distance is positive, and its radial Hessian is zero, proving the last assertion. ■

B.1.3 Convexity of the search sets and smoothness of the objective

We now apply these bounds to the balls $\mathcal{M}_{\mathrm{ex}} = \overline{B}_R(c)$, $U = \overline{B}_{R+\bar r}(c)$, and $\mathcal{K}_\alpha = \overline{B}_{\alpha R}(c)$. Here $\mathcal{M}_{\mathrm{ex}}$ is the regression search region, U contains the local model domains, and $\mathcal{K}_\alpha$ constrains the iterates. We establish their geodesic convexity, characterize the normal cone of $\mathcal{K}_\alpha$, and derive the gradient and descent estimates needed to analyze the inner updates.

Lemma B.2 (Convexity and normal cones) *Under Assumption 2.1, every ball $\overline{B}_a(c)$ with $0 < a \le R + \bar r$ is compact and geodesically convex. In particular, this holds for $\mathcal{M}_{\mathrm{ex}}$, U, and $\mathcal{K}_\alpha$. If $R_\alpha := \alpha R$ and $z \in \partial\mathcal{K}_\alpha$, its outward unit normal is $n_z = -\mathrm{Log}_z c / R_\alpha$. Moreover,*

$$\langle n, \mathrm{Log}_z x\rangle_z \le 0, \qquad x, z \in \mathcal{K}_\alpha, \quad n \in N_{\mathcal{K}_\alpha}(z), \tag{28}$$

where $N_{\mathcal{K}_\alpha}(z) = \{0\}$ in the interior and $N_{\mathcal{K}_\alpha}(z) = \{a n_z : a \ge 0\}$ on the boundary.

Proof. Compactness follows from completeness and finite dimensionality by Hopf–Rinow. We give the convexity argument because convexity of the outer ball by itself does not imply convexity of every smaller ball. Fix $a \le R + \bar r$ and distinct $x, y \in \overline{B}_a(c)$. For $t \in [0, 1]$, set $x_t = \mathrm{Exp}_c(t\,\mathrm{Log}_c x)$ and $y_t = \mathrm{Exp}_c(t\,\mathrm{Log}_c y)$, and let γ_t be their minimizing geodesic. These geodesics lie in the strongly convex outer ball and vary continuously with t. For small t they lie in $B_a(c)$; for example, the triangle inequality bounds every point on γ_t within distance $3ta$ of c.

If γ_1 left $\overline{B}_a(c)$, continuity of $t \mapsto \max_{s\in[0,1]} d(c, \gamma_t(s))$ would give a parameter $t_\star \in (0, 1)$ for which this maximum equals a. Both endpoints then lie strictly inside the ball. The function $\frac{1}{2}d^2(c, \gamma_{t_\star}(s))$ would attain an interior maximum at a point with radius a. Lemma B.1 gives a positive second derivative there, since $a \le R + \bar r < D < \pi/(2\sqrt{\Lambda_+})$ when $\Lambda_+ > 0$. This is a contradiction. Thus the smaller ball is geodesically convex.

The boundary of $\mathcal{K}_\alpha$ is a regular level set of $\frac{1}{2}d^2(c,\cdot)$: its gradient is $-\mathrm{Log}_z c$, of norm $R_\alpha > 0$. This proves the stated normal-cone formula. The geodesic from any $z \in \mathcal{K}_\alpha$ to any feasible x stays in the ball. At a boundary point its initial derivative cannot increase the defining function, so $\langle n_z, \mathrm{Log}_z x\rangle \le 0$. Multiplying by a nonnegative scalar proves (28). ■

Lemma B.3 (Gradient bounds and the descent inequality) *For all $y, y' \in U$,*

$$\|\operatorname{grad} \ell_i(y)\| \le G_\ell, \qquad \|\operatorname{grad} F_\sigma(y)\| \le R_\sigma. \tag{29}$$

Moreover,

$$\|\operatorname{grad} \ell_i(y') - P_{y\to y'} \operatorname{grad} \ell_i(y)\| \le L_\ell d(y, y'), \tag{30}$$

$$\|\operatorname{grad} f(y') - P_{y\to y'} \operatorname{grad} f(y)\| \le L_f d(y, y'), \tag{31}$$

$$f(y') \le f(y) + \langle \operatorname{grad} f(y), \operatorname{Log}_y y' \rangle + \frac{L_f}{2} d^2(y, y'). \tag{32}$$

The component functions g, h are geodesically convex on U.

Proof. The gradient formula in Lemma B.1 and $d(y, z_i) \le D$ give the first line. The Hessian bounds imply $\|\operatorname{Hess} \ell_i\|_{\mathrm{op}} \le L_\ell$ and $\|\operatorname{Hess} f\|_{\mathrm{op}} \le \sum_i |w_i| L_\ell = L_f$. Let $\gamma(t) = \operatorname{Exp}_y(t \operatorname{Log}_y y')$, $0 \le t \le 1$; it lies in U by Lemma B.2. Differentiating the parallel pullback of a gradient along γ gives

$$P_{y'\to y} \operatorname{grad} f(y') - \operatorname{grad} f(y) = \int_0^1 P_{\gamma(t)\to y} \operatorname{Hess} f(\gamma(t))[\dot\gamma(t)] \, \mathrm{d}t.$$

Taking norms proves (31); the same calculation proves (30). Integrating the derivative of $f \circ \gamma$ once more gives (32). Finally, $\operatorname{Hess} g \succeq 2W_+ \delta I$ and $\operatorname{Hess} h \succeq 2W_- \delta I$ on the convex set U, which proves the last assertion, including the case $h = 0$. ∎

B.1.4 Radial growth, interiority, and preservation of minimizers

We next justify restricting the iterates to $\mathcal{K}_\alpha$. The radial bound below shows that this smaller ball retains every regression minimizer and that its boundary normal cannot cancel the true gradient. Both facts will be needed in the proof of convergence.

Lemma B.4 (Radial growth of the signed objective) *Let Assumptions 2.1 and 2.2 hold. For $y \in \mathcal{M}_{\mathrm{ex}}$ with $t = d(c, y) > r$ and $n_y = -\operatorname{Log}_y c/t$,*

$$\langle \operatorname{grad} f(y), n_y \rangle \ge 2W_+(t - r) - 2W_-(t + r) = 2(t - W_{\mathrm{abs}} r). \tag{33}$$

Consequently,

$$\arg\min_{\mathcal{M}_{\mathrm{ex}}} f = \arg\min_{\mathcal{K}_\alpha} f \subseteq \overline{B}_{W_{\mathrm{abs}} r}(c) \subset \operatorname{int}(\mathcal{K}_\alpha), \tag{34}$$

and every such minimizer is stationary. For every $z \in \mathcal{K}_\alpha$ and $n \in N_{\mathcal{K}_\alpha}(z)$,

$$\|\operatorname{grad} f(z)\| \le \|\operatorname{grad} f(z) + n\|. \tag{35}$$

Proof. The distance $d(c, \cdot)$ is geodesically convex on $\mathcal{M}_{\mathrm{ex}}$. Indeed, its Hessian is nonnegative away from c by Lemma B.1. If a geodesic passes through c, the one-sided derivatives of the distance there are the negative and positive speeds, respectively, so convexity also holds across that point. Convexity along the geodesic from y to z_i gives

$$d(c, z_i) \ge t + \langle n_y, \operatorname{Log}_y z_i \rangle.$$

Since $d(c, z_i) \le r$, Cauchy–Schwarz and the triangle inequality yield the two-sided estimate

$$t - r \le \langle -\operatorname{Log}_y z_i, n_y \rangle \le d(y, z_i) \le t + r.$$

Multiply the lower bound by $2w_i^+$ and the upper bound by $-2w_i^-$, then sum. The identities in (23) prove (33).

The continuous objective attains its minimum on the compact set $\mathcal{M}_{\mathrm{ex}}$. At every point with $d(c, y) > W_{\mathrm{abs}} r$, the inward radial direction has strictly negative directional derivative and remains feasible for a sufficiently short step. Such a point cannot be a minimizer. This proves that all minimizers over $\mathcal{M}_{\mathrm{ex}}$ lie in $\overline{B}_{W_{\mathrm{abs}} r}(c)$. The choice of α places this ball strictly inside $\mathcal{K}_\alpha$. Restricting to $\mathcal{K}_\alpha \subset \mathcal{M}_{\mathrm{ex}}$ therefore preserves both the minimum value and its attaining points. Those points lie in an open neighborhood on which f is smooth, so their gradients vanish.

At a boundary point z of the safe ball,

$$\langle \operatorname{grad} f(z), n_z \rangle \ge 2(R_\alpha - W_{\mathrm{abs}} r) > 0. \tag{36}$$

Every outward normal is a nonnegative multiple of n_z. Thus $\langle \operatorname{grad} f(z), n\rangle \geq 0$, and expansion of $\| \operatorname{grad} f(z) + n\|^2$ proves (35). For interior points, the normal is zero. ■

B.2 Sampled model and implicit update

With these geometric estimates in place, we turn to the sampled model and the update it defines. We first explain the need for the gradient correction, then establish that the inner problem has a unique minimizer and that the stopping rule yields useful descent and gradient bounds.

B.2.1 Why direct sampling does not remove the center-gradient error

Consider the Euclidean example with responses $z_1 = -a, z_2 = a$, equal positive weights, and $a > 0$. The objective is $f(y) = \frac{1}{2}(y+a)^2 + \frac{1}{2}(y-a)^2 = y^2 + a^2$, with stationary point $y = 0$. A one-response positive batch gives $\widehat{g}(y) = (y - Z)^2$, where Z is uniform on $\{-a, a\}$. At the stationary point,

$$\mathbb{E}[\widehat{g}'(0)] = 0, \qquad \mathbb{E}[|\widehat{g}'(0)|^2] = 4a^2.$$

For any finite fixed $\tau \geq 0$, the direct sampled model $\widehat{g}(y)+(\tau/2)y^2$ has minimizer $y^+ = 2Z/(2+\tau)$, so

$$f(y^+) - f(0) = \frac{4a^2}{(2+\tau)^2} > 0.$$

This example can be placed strictly inside the balls required by the geometric assumptions. More generally, its repeated exact updates satisfy $y_{k+1} = by_k + (1-b)Z_k$, with $b = \tau/(2+\tau) < 1$ and independent Z_k. Starting at $y_0 = 0$,

$$\mathbb{E}[y_k^2] = (1-b)^2 a^2 \sum_{j=0}^{k-1} b^{2j} \longrightarrow \frac{1-b}{1+b} a^2 > 0.$$

Thus unbiased direct sampling need not yield vanishing expected squared gradient with fixed regularization, even with no negative weights and exact inner solves.

B.2.2 Clipping and exact gradient matching

The correction addresses the sampling error illustrated above by making the model gradient at y_k equal to the estimator v_k. We also show that clipping keeps the estimator bounded without increasing its error.

Lemma B.5 (Clipping and model bounds) *For every realization of the estimator and the positive model batch,*

$$\|v_k^\sigma - \operatorname{grad} F_\sigma(y_k)\| \leq \|\widetilde{v}_k^\sigma - \operatorname{grad} F_\sigma(y_k)\|, \tag{37}$$

$$\|v_k\| \leq V, \qquad \|q_k\| \leq Q, \qquad \operatorname{grad} \varphi_k(y_k) = v_k. \tag{38}$$

On the entire ball $\overline{B}_{\bar{r}}(y_k)$,

$$2W_+\delta I \preceq \operatorname{Hess} \widehat{g}_k \preceq 2W_+\zeta I. \tag{39}$$

Proof. The true component gradient belongs to the closed tangent ball of radius R_σ by (29). Orthogonal projection onto a closed convex set in an inner-product space is nonexpansive and fixes every point of that set. Applying this to $\Pi_{y_k}^{R_\sigma}$ proves (37). The same projection bounds $\|v_k^\sigma\|$ by R_σ.

The sampled coefficients are nonnegative and sum to W_+, including when indices repeat. Thus $\| \operatorname{grad} \widehat{g}_k(y_k)\| \leq W_+ G_\ell = R_+$, and the triangle inequality gives $\|v_k\| \leq V$ and $\|q_k\| \leq R_+ + V = Q$. At its center, the differential of Log_{y_k} is the identity and the gradient of $d^2(y_k, \cdot)$ vanishes. Therefore

$$\operatorname{grad} \varphi_k(y_k) = \operatorname{grad} \widehat{g}_k(y_k) - q_k = v_k.$$

Finally, multiply the loss Hessian bounds by the sampled coefficients and sum to obtain (39). ■

Define the errors of the clipped estimates by

$$e_k^\sigma := v_k^\sigma - \operatorname{grad} F_\sigma(y_k), \qquad e_k := v_k - \operatorname{grad} f(y_k), \qquad \mathcal{E}_k := \sum_{\sigma \in \mathcal{S}} \|e_k^\sigma\|^2. \tag{40}$$

Then $e_k = e_k^+ - e_k^-$ and

$$\|e_k\|^2 \le 2\mathcal{E}_k, \qquad 0 \le \mathcal{E}_k \le 4 \sum_{\sigma \in \mathcal{S}} R_\sigma^2. \tag{41}$$

No unbiasedness of the clipped estimates is used.

B.2.3 Existence and uniqueness of every inner minimizer

The bounds on the correction now allow us to control the model Hessian. We use them to show that the chosen regularization makes every inner problem strongly convex and gives a feasible exact update.

Proposition B.6 (Well-posedness of the sampled model) *Under the initialization and model conditions of Theorem 3.1, define*

$$\mu_k := 2W_+\delta - L_{\log}\|q_k\| + \lambda_k \delta_p. \tag{42}$$

For every model realization, $\mu_k \ge m_k^\star \ge \mu_\star/\eta_k$ and $\operatorname{Hess}\varphi_k \succeq \mu_k I$ on $\overline{B}_{\bar r}(y_k)$. The problem

$$\min_{y \in \mathcal{M}_k} \varphi_k(y), \qquad \mathcal{M}_k = \overline{B}_{\bar r}(y_k) \cap \mathcal{K}_\alpha,$$

has a unique minimizer $y_{k+1}^\star$, which satisfies

$$d(y_k, y_{k+1}^\star) < \bar r, \qquad 0 \in \operatorname{grad}\varphi_k(y_{k+1}^\star) + N_{\mathcal{K}_\alpha}(y_{k+1}^\star). \tag{43}$$

Every sequence of accepted updates is feasible and obeys

$$y_k \in \mathcal{K}_\alpha, \qquad \operatorname{dist}(y_k, \partial\mathcal{M}_{\mathrm{ex}}) \ge (1-\alpha)R. \tag{44}$$

Proof. The initial point lies in $\operatorname{int}(\mathcal{K}_\alpha)$. Assume inductively that $y_k \in \mathcal{K}_\alpha$. The three terms of the model have Hessian lower bounds $2W_+\delta I$, $-L_{\log}\|q_k\| I$, and $\lambda_k\delta_p I$, respectively. Hence their sum has lower bound $\mu_k I$. If $L_{\log}\|q_k\| - 2W_+\delta + m_k^\star \ge 0$, the regularization rule gives $\mu_k = m_k^\star$. Otherwise $\lambda_k = 0$ and $\mu_k > m_k^\star$. This proves strong convexity on the full local ball and on its intersection with the safe ball.

The local ball is strongly convex by Assumption 2.1; the safe ball is geodesically convex by Lemma B.2. Their intersection is nonempty, compact, and geodesically convex. Continuity gives existence of a minimizer. Strong convexity along the geodesic joining any two distinct candidates gives uniqueness.

If $y \in \mathcal{M}_k$ and $d(y_k, y) = \bar r$, strong convexity and gradient matching give

$$\begin{aligned}\varphi_k(y) - \varphi_k(y_k) &\ge \langle v_k, \operatorname{Log}_{y_k} y\rangle + \frac{\mu_k}{2}\bar r^2 \\ &\ge -\|v_k\|\bar r + \frac{1}{2}\left(\frac{2\|v_k\|}{\bar r} + \nu\right)\bar r^2 = \frac{\nu}{2}\bar r^2 > 0.\end{aligned}$$

Since y_k is feasible, the minimizer cannot lie on the boundary of the local ball. Only the safe-ball constraint can be active there. The usual first-order condition for this smooth ball constraint is therefore (43). At a boundary point the constraint gradient is nonzero, so this condition follows directly from the tangent half-space of feasible directions; at an interior point it is the vanishing of the gradient. It is sufficient as well, by strong convexity and (28).

The exact minimizer satisfies (16) with zero residual. Any accepted point belongs to $\mathcal{M}_k \subset \mathcal{K}_\alpha$, which completes the induction. For $z \in \partial\mathcal{M}_{\mathrm{ex}}$ the triangle inequality gives $d(y_k, z) \ge R - d(c, y_k) \ge (1-\alpha)R$. Taking the infimum over z proves the margin bound. ∎

B.2.4 Consequences of the inner residual condition

The algorithm can stop before reaching the exact inner minimizer. We next quantify how the residual bounds the distance to that minimizer and the remaining error in the model value.

Lemma B.7 (Distance and value error in an inner solve) *Let $y \in \mathcal{M}_k \cap B_{\bar{r}}(y_k)$ and $n \in N_{\mathcal{K}_\alpha}(y)$, and write $\varepsilon = \| \operatorname{grad} \varphi_k(y) + n \|$. Then*

$$d(y, y^\star_{k+1}) \le \frac{\varepsilon}{\mu_k}, \qquad 0 \le \varphi_k(y) - \varphi_k(y^\star_{k+1}) \le \frac{\varepsilon^2}{2\mu_k}. \tag{45}$$

In particular, an accepted point satisfies

$$d(y_{k+1}, y^\star_{k+1}) \le \frac{\zeta_{\mathrm{in}}}{\mu_\star} d_k, \qquad \varphi_k(y_{k+1}) - \varphi_k(y^\star_{k+1}) \le \frac{\zeta_{\mathrm{in}}^2}{2\mu_\star \eta_k} d_k^2.$$

Proof. Let $z = y^\star_{k+1}$, $t = d(y, z)$, and take $n^\star \in N_{\mathcal{K}_\alpha}(z)$ with $\operatorname{grad} \varphi_k(z) + n^\star = 0$. Strong convexity integrated along the geodesic from z to y gives

$$\langle \operatorname{grad} \varphi_k(y) - P_{z \to y} \operatorname{grad} \varphi_k(z), -\operatorname{Log}_y z \rangle \ge \mu_k t^2.$$

Normal support at the two endpoints gives $\langle n - P_{z \to y} n^\star, -\operatorname{Log}_y z \rangle \ge 0$. Adding these inequalities yields $\mu_k t^2 \le \varepsilon t$, which proves the distance bound, including $t = 0$.

Strong convexity based at y, and normal support there, also give

$$\begin{aligned} \varphi_k(y) - \varphi_k(z) &\le -\langle \operatorname{grad} \varphi_k(y), \operatorname{Log}_y z \rangle - \frac{\mu_k}{2} t^2 \\ &\le -\langle \operatorname{grad} \varphi_k(y) + n, \operatorname{Log}_y z \rangle - \frac{\mu_k}{2} t^2 \le \varepsilon t - \frac{\mu_k}{2} t^2 \le \frac{\varepsilon^2}{2\mu_k}. \end{aligned}$$

Substitute $\varepsilon \le \zeta_{\mathrm{in}} d_k / \eta_k$ and $\mu_k \ge \mu_\star / \eta_k$ to obtain the final assertions. ■

B.2.5 Step length, objective decrease, and the true gradient

We now connect the inner stopping rule to progress on the original objective f. The following bounds relate the step length, objective decrease, and true gradient to the gradient-estimation error.

Set $\chi = \mu_\star - \zeta_{\mathrm{in}} > 0$ and define

$$\begin{aligned} L_\varphi &:= 2W_+ \zeta + L_{\log} Q + \frac{\zeta_p}{\delta_p} \left(L_{\log} Q + \mu_\star + \frac{2V}{\bar{r}} + \nu \right), \\ C_\varphi &:= L_\varphi + \zeta_{\mathrm{in}} + 2L_f. \end{aligned} \tag{46}$$

These constants are independent of the iteration, all sampled batches, and the horizon.

Lemma B.8 (Pathwise descent and stationarity bounds) *Suppose the scale obeys $0 < \eta_k \le \min\{1, \chi/(2L_f)\}$. Every accepted update satisfies*

$$\langle v_k, s_k \rangle \le -\frac{\chi}{\eta_k} d_k^2, \qquad d_k \le \frac{\eta_k}{\chi} \|v_k\| \le \frac{V}{\chi} \eta_k, \tag{47}$$

$$f(y_{k+1}) \le f(y_k) - \frac{\chi}{2\eta_k} d_k^2 + \frac{2\eta_k}{\chi} \mathcal{E}_k, \tag{48}$$

$$\| \operatorname{grad} f(y_k) \|^2 \le \frac{2C_\varphi^2}{\eta_k^2} d_k^2 + 4\mathcal{E}_k. \tag{49}$$

Proof. Let $r_{k+1} = \operatorname{grad} \varphi_k(y_{k+1}) + n_{k+1}$ denote the accepted residual, so $\|r_{k+1}\| \le \zeta_{\mathrm{in}} d_k / \eta_k$. Integrating the model Hessian along the step geodesic gives

$$\langle v_k, s_k \rangle \le \langle \operatorname{grad} \varphi_k(y_{k+1}), P_k s_k \rangle - \frac{\mu_\star}{\eta_k} d_k^2.$$

Since $\operatorname{Log}_{y_{k+1}} y_k = -P_k s_k$, normal support implies $\langle n_{k+1}, P_k s_k \rangle \ge 0$. Hence

$$\langle v_k, s_k \rangle \le \langle r_{k+1}, P_k s_k \rangle - \frac{\mu_\star}{\eta_k} d_k^2 \le -\frac{\chi}{\eta_k} d_k^2.$$

Cauchy–Schwarz gives the step-length bound after division by d_k when $d_k > 0$. The claim is immediate for $d_k = 0$.

Apply (32) and use $\operatorname{grad} f(y_k) = v_k - e_k$:

$$f(y_{k+1}) \le f(y_k) - \frac{\chi}{\eta_k} d_k^2 + \|e_k\| d_k + \frac{L_f}{2} d_k^2.$$

The inequalities

$$\|e_k\| d_k \le \frac{\chi}{4\eta_k} d_k^2 + \frac{\eta_k}{\chi} \|e_k\|^2, \qquad \frac{L_f}{2} \le \frac{\chi}{4\eta_k}, \qquad \|e_k\|^2 \le 2\mathcal{E}_k$$

prove (48).

To bound the true gradient, we also need an upper bound on the model Hessian. The regularization rule and $\eta_k \le 1$ imply

$$\eta_k \lambda_k \le \frac{L_{\log} Q + \mu_\star + 2V/\bar{r} + \nu}{\delta_p}.$$

Indeed, dropping the nonpositive term $-2W_+\delta$ increases the positive part, and $\eta_k m_k^\star \le \mu_\star + 2V/\bar{r} + \nu$. Consequently,

$$\|\operatorname{Hess} \varphi_k(y)\|_{\mathrm{op}} \le 2W_+\zeta + L_{\log} Q + \lambda_k \zeta_p \le L_\varphi/\eta_k \quad \text{on } \overline{B}_{\bar{r}}(y_k).$$

Integration along the step gives

$$\|\operatorname{grad} \varphi_k(y_{k+1}) - P_k v_k\| \le \frac{L_\varphi}{\eta_k} d_k. \tag{50}$$

The normal cannot cancel an outward true gradient, by (35). Therefore

$$\begin{aligned} \|\operatorname{grad} f(y_{k+1})\| &\le \|\operatorname{grad} f(y_{k+1}) + n_{k+1}\| \\ &\le \|\operatorname{grad} f(y_{k+1}) - P_k \operatorname{grad} f(y_k)\| + \|e_k\| \\ &\quad + \|P_k v_k - \operatorname{grad} \varphi_k(y_{k+1})\| + \|r_{k+1}\| \\ &\le L_f d_k + \|e_k\| + \frac{L_\varphi + \zeta_{\mathrm{in}}}{\eta_k} d_k. \end{aligned}$$

A second application of gradient Lipschitz continuity yields $\|\operatorname{grad} f(y_k)\| \le C_\varphi d_k/\eta_k + \|e_k\|$. Squaring, using $(a+b)^2 \le 2a^2 + 2b^2$, and then applying (41) proves (49). ■

B.3 From error budget to stationarity

The preceding bounds describe a single update. We now sum them over iterations to show how control of the accumulated estimation error yields a stationarity guarantee. This argument will be shared by all the estimator choices considered below.

B.3.1 A quantitative transfer proposition

Let

$$f_\star := \min_{\mathcal{K}_\alpha} f = \min_{\mathcal{M}_{\mathrm{ex}}} f, \qquad \Delta_f := f(y_0) - f_\star \ge 0. \tag{51}$$

The minimum exists by compactness and continuity, and the equality of the two minima follows from Lemma B.4.

Proposition B.9 (Transfer of error control) *Assume the conditions of Lemma B.8. The scales may be deterministic or chosen from the current history. For any collection of horizons, suppose that*

$$\sum_{k=0}^{K-1} \mathbb{E}[\eta_k \mathcal{E}_k] \le \theta \sum_{k=0}^{K-1} \mathbb{E}\left[\frac{d_k^2}{\eta_k}\right] + C_{\mathrm{err}}, \qquad 0 \le \theta < \frac{\chi^2}{4}, \tag{52}$$

with finite constants independent of the horizon. Set

$$a_\theta := \frac{\chi}{2} - \frac{2\theta}{\chi} > 0, \qquad B_\theta := \frac{\Delta_f + 2C_{\mathrm{err}}/\chi}{a_\theta}, \qquad C_\theta := (2C_\varphi^2 + 4\theta) B_\theta + 4C_{\mathrm{err}}. \tag{53}$$

Then each such horizon satisfies

$$\sum_{k=0}^{K-1} \mathbb{E}\left[\frac{d_k^2}{\eta_k}\right] \le B_\theta, \tag{54}$$

$$\sum_{k=0}^{K-1} \mathbb{E}[\eta_k \mathcal{E}_k] \le \theta B_\theta + C_{\mathrm{err}}, \tag{55}$$

$$\sum_{k=0}^{K-1} \mathbb{E}[\eta_k \| \operatorname{grad} f(y_k)\|^2] \le C_\theta. \tag{56}$$

Proof. All finite-horizon quantities are integrable. Indeed, the objective and its gradient are bounded on the safe ball, the errors are bounded by clipping, and (47) implies $d_k^2/\eta_k \le V^2\eta_k/\chi^2 \le V^2/\chi^2$. Sum (48), take expectations, and use $f(y_K) \ge f_\star$ to obtain

$$\frac{\chi}{2}\sum_{k=0}^{K-1} \mathbb{E}\left[\frac{d_k^2}{\eta_k}\right] \le \Delta_f + \frac{2}{\chi}\sum_{k=0}^{K-1} \mathbb{E}[\eta_k \mathcal{E}_k].$$

Substitution of (52) gives $a_\theta \sum_{k<K} \mathbb{E}[d_k^2/\eta_k] \le \Delta_f + 2C_{\mathrm{err}}/\chi$, proving the first bound. The second follows by another application of the error budget. Finally, multiply (49) by η_k, sum, and take expectations:

$$\sum_{k=0}^{K-1} \mathbb{E}[\eta_k \| \operatorname{grad} f(y_k)\|^2] \le 2C_\varphi^2 \sum_{k=0}^{K-1} \mathbb{E}[d_k^2/\eta_k] + 4\sum_{k=0}^{K-1} \mathbb{E}[\eta_k \mathcal{E}_k].$$

The first two bounds give (56). No independence between η_k, d_k, and $\mathcal{E}_k$ was used; each product remains inside its expectation. ■

B.3.2 Proof of Theorem 3.1

Proof. For part (i), the choice of α puts y_0 strictly inside the safe ball. Proposition B.6 then applies inductively to every sampled model. It establishes the unique inner minimizer, its zero-residual optimality condition, feasibility of all accepted iterates, and their margin to $\partial\mathcal{M}_{\mathrm{ex}}$. These conclusions do not use (18).

For part (ii), the prescribed scales allow Lemma B.8 at every iteration. The error budget of the theorem is exactly the hypothesis of Proposition B.9. Hence (19) holds with the explicit constant $C = C_\theta$ in (53). ■

The theorem states its conclusion for deterministic scales. Proposition B.9 supplies the history-dependent version needed for Corollary 3.4; it is proved directly from the same pathwise inequalities.

B.3.3 Finite-horizon output and the scope of the guarantee

For deterministic positive scales, let $T_K = \sum_{k=0}^{K-1} \eta_k$ and choose an index independently of the algorithm with $\mathbb{P}(J_K = k) = \eta_k/T_K$. Then

$$\mathbb{E}\| \operatorname{grad} f(y_{J_K})\|^2 = \frac{1}{T_K}\sum_{k=0}^{K-1} \eta_k \mathbb{E}\| \operatorname{grad} f(y_k)\|^2 \le \frac{C_\theta}{T_K}. \tag{57}$$

For $\eta_k \equiv \eta$, this is the uniform output in Algorithm 1, and

$$\mathbb{E}\| \operatorname{grad} f(y_{J_K})\|^2 \le \frac{C_\theta}{\eta K}. \tag{58}$$

These are stationarity bounds for the fixed empirical objective. They do not assert global optimality of a stationary point, convergence of the last iterate at the same rate, or a bound on inner-solver work. For history-dependent scales the deterministic-index identity (57) is not invoked.

B.4 EXACT-REFRESH ESTIMATORS

We next verify the required error budget for estimators that occasionally recompute the full gradient. We first establish the sampling conventions and variance bounds, then apply them to PAGE and to methods with periodic refreshes.

B.4.1 SAMPLING ORDER AND CONDITIONAL INFORMATION

The order of sampling matters for the conditional expectations below. We distinguish the information available before the inner solve from that available before drawing the next estimator batches.

Let $\mathcal{F}_k$ contain the complete history before the positive model batch at iteration k, including y_k, the clipped estimator states, all stored snapshots, and the chosen scale η_k. Let $\mathcal{H}_k$ additionally contain the model batch and the completed inner solve, including y_{k+1} and its residual, but exclude the next estimator batches and the PAGE decision. Then

$$\mathcal{F}_k \subseteq \mathcal{H}_k \subseteq \mathcal{F}_{k+1}, \qquad \mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{H}_k]. \tag{59}$$

All inner-solver choices are understood to be measurable with respect to the data and samples available at that stage. In particular, the solver does not use the next estimator samples before returning y_{k+1}.

Conditional on $\mathcal{H}_k$, the next batch is an independent list of draws from p^σ. Repeated occurrences of a batch symbol mean that the same list is used at all indicated evaluation points. Within a sign, B and C are independent when both occur. STORM and SVRRM instead reuse B in their two sampled terms. The PAGE Bernoulli decision is independent of $\mathcal{H}_k$ and shared by all active signs. Independence between signs is permitted but is not needed for the sum of sign-wise squared errors used here.

For a periodic method, an epoch begins at $t \in \{0, H, 2H, \dots\}$. At that point every component gradient is exact, so $\mathcal{E}_t = 0$. The snapshot methods also store $\widetilde{y} = y_t$ and $\widetilde{\mu}^\sigma = \operatorname{grad} F_\sigma(y_t)$ until the next refresh.

B.4.2 CONDITIONAL MOMENTS OF THE GRADIENT ORACLE

With this sampling order fixed, we bound the noise in sampled gradients and gradient differences. These estimates will also be used for the refresh-free methods.

For each active sign define the random component vector $G_i^\sigma(y) = W_\sigma \operatorname{grad} \ell_i(y)$, with i drawn from $p_i^\sigma = w_i^\sigma / W_\sigma$. Its expectation is $\operatorname{grad} F_\sigma(y)$. Let

$$\xi_B^\sigma(y) := \widehat{G}_B^\sigma(y) - \operatorname{grad} F_\sigma(y). \tag{60}$$

We will use the averaging constants

$$A_b := \sum_{\sigma \in \mathcal{S}} \frac{L_\sigma^2}{b_\sigma}, \qquad A_c := \sum_{\sigma \in \mathcal{S}} \frac{L_\sigma^2}{c_\sigma}, \qquad S_b := \sum_{\sigma \in \mathcal{S}} \frac{R_\sigma^2}{b_\sigma}, \qquad S_c := \sum_{\sigma \in \mathcal{S}} \frac{R_\sigma^2}{c_\sigma}. \tag{61}$$

Only constants for batches actually used by the selected method are needed. In the main-text bounds it suffices that $A_\Delta \geq A_b$ and $A_{\mathrm{sv}} \geq A_c$. The natural choices are equality; the factors $1/b_\sigma$ and $1/c_\sigma$ are the variance reductions from averaging independent draws. Any larger constants give valid, more restrictive sufficient conditions.

Lemma B.10 (Fresh-gradient and same-batch difference moments) *Let $x, y \in \mathcal{K}_\alpha$ be fixed conditional on a history $\mathcal{H}$, and draw a fresh sign-σ batch B of size b. Then*

$$\mathbb{E}[\xi_B^\sigma(y) \mid \mathcal{H}] = 0, \qquad \mathbb{E}[\|\xi_B^\sigma(y)\|^2 \mid \mathcal{H}] \leq \frac{R_\sigma^2}{b}, \tag{62}$$

$$\mathbb{E}[\xi_B^\sigma(y) - P_{x \to y} \xi_B^\sigma(x) \mid \mathcal{H}] = 0, \qquad \mathbb{E}[\|\xi_B^\sigma(y) - P_{x \to y} \xi_B^\sigma(x)\|^2 \mid \mathcal{H}] \leq \frac{L_\sigma^2}{b} d^2(x, y). \tag{63}$$

Proof. All vectors in each difference belong to the same tangent space. For independent random vectors $Z_1, \dots, Z_b$ with common mean $\overline{Z}$, the cross terms in the squared norm of their centered

average vanish, giving

$$\mathbb{E}\left\|\frac{1}{b}\sum_{j=1}^{b}(Z_j-\overline{Z})\right\|^2=\frac{1}{b}\mathbb{E}\|Z_1-\overline{Z}\|^2\le\frac{1}{b}\mathbb{E}\|Z_1\|^2.$$

The same identity holds conditionally on $\mathcal{H}$. Apply it first to $Z_j=G^\sigma_{I_j}(y)$, whose norm is at most R_σ. Apply it next to $Z_j=G^\sigma_{I_j}(y)-P_{x\to y}G^\sigma_{I_j}(x)$, whose norm is at most $L_\sigma d(x,y)$ by Lemma B.3. Their conditional means give the centered quantities in the statement. The second calculation uses the same index I_j at x and y; independent batches at these points would not give this distance-dependent bound. ■

Sampled initialization followed by clipping therefore obeys

$$\mathbb{E}[\mathcal{E}_0]\le S_0:=\sum_{\sigma\in\mathcal{S}}\frac{R_\sigma^2}{|B_0^\sigma|}<\infty. \tag{64}$$

This permits any fixed nonempty initialization batches in Algorithm 1. In particular, $S_0=S_b$ if their sizes are b_σ, and $S_0=S_c$ if their sizes are c_σ. Exact initialization gives $\mathcal{E}_0=0$.

B.4.3 Errors of the recursive and snapshot estimates

At the transition from k to $k+1$, define

$$\varepsilon^\sigma_{k+1,B}:=\xi^\sigma_B(y_{k+1})-P_k\xi^\sigma_B(y_k), \tag{65}$$
$$Z^\sigma_{k+1,C}:=\xi^\sigma_C(y_{k+1})-P_{\widetilde{y}\to y_{k+1}}\xi^\sigma_C(\widetilde{y}). \tag{66}$$

Their conditional means given $\mathcal{H}_k$ are zero, and

$$\mathbb{E}_k\|\varepsilon^\sigma_{k+1,B}\|^2\le\frac{L_\sigma^2}{b_\sigma}d_k^2,\qquad\mathbb{E}_k\|Z^\sigma_{k+1,C}\|^2\le\frac{L_\sigma^2}{c_\sigma}d^2(\widetilde{y},y_{k+1}). \tag{67}$$

Subtracting the true gradient from the definitions in the main text gives the exact raw-error identities

$$\begin{aligned}D^\sigma_{k+1,B}-\operatorname{grad}F_\sigma(y_{k+1})&=P_ke^\sigma_k+\varepsilon^\sigma_{k+1,B},\\ V^\sigma_{k+1,C}-\operatorname{grad}F_\sigma(y_{k+1})&=Z^\sigma_{k+1,C},\\ U^\sigma_{k+1,C}-\operatorname{grad}F_\sigma(y_{k+1})&=\xi^\sigma_C(y_{k+1}).\end{aligned} \tag{68}$$

The snapshot identity uses its exact stored gradient. The recursive identity uses the previous *clipped* state, as specified in the algorithm. For example, $\mathbb{E}_k\langle P_ke^\sigma_k,\varepsilon^\sigma_{k+1,B}\rangle=0$, although e^σ_k need not have unconditional mean zero. Every estimate below bounds the raw error and then applies (37).

B.4.4 PAGE and the proof of Corollary 3.2

For PAGE, an exact refresh resets the estimation error to zero. Combining this reset with the recursive-update bound gives the following accumulated error estimate.

Lemma B.11 (PAGE error accumulation) *For exact initialization and refresh probability $p\in(0,1]$,*

$$\mathbb{E}_k[\mathcal{E}_{k+1}]\le(1-p)\mathcal{E}_k+(1-p)A_bd_k^2. \tag{69}$$

For every $K\ge1$,

$$\sum_{k=0}^{K-1}\mathbb{E}[\mathcal{E}_k]\le\frac{1-p}{p}A_b\sum_{k=0}^{K-1}\mathbb{E}[d_k^2]. \tag{70}$$

Proof. On refresh, the raw estimate is the true gradient, and clipping leaves it unchanged. On a nonrefresh step, the first identity in (68), its zero conditional cross term, and (67) give

$$\mathbb{E}_k[\mathcal{E}_{k+1}\mid\text{no refresh}]\le\mathcal{E}_k+A_bd_k^2.$$

Independence of the refresh decision proves the recursion. Rearrange it, take expectations, and sum to obtain

$$p\sum_{k=0}^{K-1}\mathbb{E}[\mathcal{E}_k] \le \mathbb{E}[\mathcal{E}_0] - \mathbb{E}[\mathcal{E}_K] + (1-p)A_b\sum_{k=0}^{K-1}\mathbb{E}[d_k^2].$$

The initial error is zero and the terminal error is nonnegative. Division by p gives (70), also when $p = 1$. ■

Proof of Corollary 3.2. *Proof.* Multiply (70) by the constant η. Since $A_b \le A_\Delta$, this gives (18) with

$$\theta = \eta^2\frac{1-p}{p}A_\Delta, \qquad C_{\mathrm{err}} = 0.$$

The assumed inequality is exactly $\theta < \chi^2/4$, and all other conditions of Theorem 3.1 hold. The weighted stationarity bound follows, and (58) gives the stated $O(K^{-1})$ consequence. ■

B.4.5 Periodic refreshes: recursive and snapshot bounds

For periodic refreshes, we instead accumulate the error within each epoch and then sum across epochs.

For $H \ge 1$, write $h_H = H(H-1)/2$. The next lemma also handles a final incomplete epoch, which makes the accumulated bounds available at every horizon.

Lemma B.12 (Accumulated error with periodic exact refreshes) *For the periodic methods of Algorithm 1,*

$$\sum_{k=0}^{K-1}\mathbb{E}[\mathcal{E}_k] \le B_{\mathsf{M}}\sum_{k=0}^{K-1}\mathbb{E}[d_k^2], \qquad K \ge 1, \tag{71}$$

where the following constants are valid:

$$B_{\mathsf{M}} = \begin{cases} (H-1)A_b, & \mathsf{M} = \mathrm{SARAH/SPIDER}, \\ h_H A_c, & \mathsf{M} = \mathrm{SVRG}, \\ \dfrac{\kappa\big((1-\beta)^2A_b + \beta^2 h_H A_c\big)}{\beta(2-\beta)}, & \mathsf{M} = \mathrm{SVRRM/SRVRG}. \end{cases} \tag{72}$$

Here $\kappa = 2$ for SVRRM and $\kappa = 1$ for SRVRG. For $H = 1$, every estimator is exact and $B_{\mathsf{M}} = 0$ may be used for all methods.

Proof. Fix an epoch start t and a number q with $0 \le q \le H-1$. We bound the errors at $t, t+1, \dots, t+q$; this includes a complete epoch when $q = H-1$. Exact refresh gives $\mathcal{E}_t = 0$.

SARAH/SPIDER. At each nonrefresh transition, the recursive raw-error identity and its vanishing cross term give $\mathbb{E}_k[\mathcal{E}_{k+1}] \le \mathcal{E}_k + A_b d_k^2$. Iteration of this inequality yields

$$\mathbb{E}[\mathcal{E}_{t+j}] \le A_b\sum_{\ell=0}^{j-1}\mathbb{E}[d_{t+\ell}^2], \qquad 1 \le j \le q.$$

Therefore

$$\begin{aligned}\sum_{j=0}^{q}\mathbb{E}[\mathcal{E}_{t+j}] &\le A_b\sum_{\ell=0}^{q-1}(q-\ell)\mathbb{E}[d_{t+\ell}^2] \\ &\le (H-1)A_b\sum_{\ell=0}^{q}\mathbb{E}[d_{t+\ell}^2].\end{aligned}$$

SVRG. The snapshot $\widetilde{y} = y_t$ has an exact stored gradient. The second raw-error identity and its moment bound imply

$$\mathbb{E}[\mathcal{E}_{t+j}] \le A_c\mathbb{E}[d^2(y_t, y_{t+j})].$$

The triangle inequality followed by Cauchy–Schwarz gives, pathwise,

$$d^2(y_t, y_{t+j}) \le \left(\sum_{\ell=0}^{j-1} d_{t+\ell}\right)^2 \le j \sum_{\ell=0}^{j-1} d_{t+\ell}^2.$$

Summing and exchanging the order of the finite sums yields

$$\begin{aligned}\sum_{j=1}^{q} d^2(y_t, y_{t+j}) &\le \sum_{\ell=0}^{q-1} \left(\sum_{j=\ell+1}^{q} j\right) d_{t+\ell}^2 \\ &\le \frac{q(q+1)}{2} \sum_{\ell=0}^{q} d_{t+\ell}^2 \le h_H \sum_{\ell=0}^{q} d_{t+\ell}^2.\end{aligned} \tag{73}$$

This proves the SVRG bound on the epoch segment.

SVRRM and SRVRG. Subtract the true gradient from the convex combination of D and V. Its raw error in sign σ is

$$(1-\beta)P_k e_k^\sigma + (1-\beta)\varepsilon_{k+1,B}^\sigma + \beta Z_{k+1,C}^\sigma.$$

The two noise terms have zero conditional mean, so their cross terms with the past error vanish. In SRVRG, the two noises are conditionally independent, and their mutual cross term also vanishes. In SVRRM they share a batch; use $\|u+v\|^2 \le 2\|u\|^2 + 2\|v\|^2$ for their sum. After clipping and summing over signs, both cases are covered by

$$\mathbb{E}_k[\mathcal{E}_{k+1}] \le (1-\beta)^2 \mathcal{E}_k + \kappa(1-\beta)^2 A_b d_k^2 + \kappa\beta^2 A_c d^2(y_t, y_{k+1}). \tag{74}$$

Sum the unconditional inequalities for $k = t, \dots, t+q-1$. With $S = \sum_{j=0}^{q} \mathbb{E}[\mathcal{E}_{t+j}]$ and $\mathcal{E}_t = 0$, the left-hand sum is S, while the preceding-error sum is $S - \mathbb{E}[\mathcal{E}_{t+q}] \le S$. Thus

$$\beta(2-\beta)S \le \kappa(1-\beta)^2 A_b \sum_{\ell=0}^{q-1} \mathbb{E}[d_{t+\ell}^2] + \kappa\beta^2 A_c \sum_{j=1}^{q} \mathbb{E}[d^2(y_t, y_{t+j})].$$

Use (73) and divide by $\beta(2-\beta) > 0$ to obtain the final constant in (72).

Partition $\{0, \dots, K-1\}$ into complete epochs and at most one final incomplete epoch. The preceding bounds apply to each part; summing proves (71). When $H = 1$, there are no nonrefresh iterations, so the errors vanish identically. ■

B.4.6 Proof of Corollary 3.3

Proof. For each method, the constant A_H in (21) is at least the corresponding B_{M} in (72). For SARAH/SPIDER this follows from $H - 1 \le H$ and $A_b \le A_\Delta$. For SVRG it follows from $h_H \le H^2/2$ and $A_c \le A_{\mathrm{sv}}$. For the two mixed methods use $h_H \le H^2$ and both variance bounds. Hence, for a constant scale,

$$\sum_{k=0}^{K-1} \mathbb{E}[\eta \mathcal{E}_k] \le \eta A_H \sum_{k=0}^{K-1} \mathbb{E}[d_k^2] = \eta^2 A_H \sum_{k=0}^{K-1} \mathbb{E}[d_k^2/\eta].$$

This is the required error budget with $\theta = \eta^2 A_H$ and $C_{\mathrm{err}} = 0$. The assumed strict inequality permits Theorem 3.1, proving the stated result for complete-epoch horizons. The incomplete-epoch argument in the lemma in fact gives the same conclusion for every $K \ge 1$. Equation (58) then gives the uniform output rate. ■

The main-text constants are convenient sufficient choices. Using (72) gives less restrictive choices when a sharper bound on the admissible scale is desired. For any exact-refresh method with accumulated constant B, one explicit choice with a fixed margin is

$$0 < \eta \le \min\left\{1, \frac{\chi}{2L_f}, \frac{\chi}{\sqrt{8B}}\right\},$$

where the last restriction is omitted if $B = 0$. Then $\theta = \eta^2 B \le \chi^2/8$ and

$$\mathbb{E}\|\operatorname{grad} f(y_{J_K})\|^2 \le \left(\frac{8C_\varphi^2}{\chi} + 2\chi\right)\frac{\Delta_f}{\eta K}. \tag{75}$$

B.5 Refresh-free estimators

We next consider STORM and Hybrid-SGD, which control estimation error through the mixing parameter β_k without exact refreshes. We derive their accumulated error budget and then use it to prove stationarity almost surely.

B.5.1 The STORM and Hybrid-SGD error recursions

Lemma B.13 (Conditional error recursion without exact refreshes) *For $0 < \beta_k \le 1$, chosen before the new estimator batches, STORM and Hybrid-SGD satisfy*

$$\mathbb{E}_k[\mathcal{E}_{k+1}] \le (1-\beta_k)^2\mathcal{E}_k + Ad_k^2 + \Sigma\beta_k^2, \tag{76}$$

with the choices

$$(A,\Sigma) = \begin{cases}(2A_\Delta, 2S_b), & \textit{STORM},\\ (A_\Delta, S_c), & \textit{Hybrid-SGD}.\end{cases} \tag{77}$$

The same bounds hold with A_Δ replaced by A_b.

Proof. For STORM, subtract the true gradient at y_{k+1} from the raw update to obtain

$$(1-\beta_k)P_k e_k^\sigma + (1-\beta_k)\varepsilon_{k+1,B}^\sigma + \beta_k\xi_B^\sigma(y_{k+1}).$$

The last two terms share the same batch. Their sum has zero conditional mean, so its cross term with the previous error vanishes. Bound its squared norm by twice the sum of the two squared norms, apply the moment bounds, and then clip. Summing over signs gives

$$\mathbb{E}_k[\mathcal{E}_{k+1}] \le (1-\beta_k)^2\mathcal{E}_k + 2(1-\beta_k)^2 A_b d_k^2 + 2S_b\beta_k^2.$$

For Hybrid-SGD the raw error is the same except that the last term is $\beta_k\xi_C^\sigma(y_{k+1})$ from an independent batch. The two noise terms are now conditionally independent and centered, so their cross term vanishes as well. Thus

$$\mathbb{E}_k[\mathcal{E}_{k+1}] \le (1-\beta_k)^2\mathcal{E}_k + (1-\beta_k)^2 A_b d_k^2 + S_c\beta_k^2.$$

Use $(1-\beta_k)^2 \le 1$ and $A_b \le A_\Delta$ to obtain the stated constants. ∎

B.5.2 The accumulated error budget

The recursion above controls one update of the estimator. Choosing β_k proportional to η_k lets us sum it into the error budget required by Proposition B.9.

Proposition B.14 (Error budget for predictable scales) *Let $\beta_k = c_\beta\eta_k$ with $c_\beta > 0$ and $0 < \eta_k \le \bar\eta \le 1/c_\beta$. Suppose η_k is $\mathcal{F}_k$-measurable and $T_2 := \mathbb{E}[\sum_{k=0}^\infty \eta_k^2] < \infty$. Then* (18) *holds for every K with*

$$\theta = \frac{A\bar\eta}{c_\beta}, \qquad C_{\mathrm{err}} = \frac{\mathbb{E}[\mathcal{E}_0]}{c_\beta} + \Sigma c_\beta T_2. \tag{78}$$

One may replace $\mathbb{E}[\mathcal{E}_0]$ by S_0 from (64)*.*

Proof. Since $0 \le \beta_k \le 1$, we have $(1-\beta_k)^2 \le 1-\beta_k$. Therefore (76) implies

$$c_\beta\eta_k\mathcal{E}_k \le \mathcal{E}_k - \mathbb{E}_k[\mathcal{E}_{k+1}] + Ad_k^2 + \Sigma c_\beta^2\eta_k^2.$$

Take expectations and sum. The error terms telescope even when η_k is random, because $\eta_k\mathcal{E}_k$ stays inside the expectation. Drop the nonpositive terminal error to obtain

$$\sum_{k=0}^{K-1}\mathbb{E}[\eta_k\mathcal{E}_k] \le \frac{\mathbb{E}[\mathcal{E}_0]}{c_\beta} + \frac{A}{c_\beta}\sum_{k=0}^{K-1}\mathbb{E}[d_k^2] + \Sigma c_\beta\sum_{k=0}^{K-1}\mathbb{E}[\eta_k^2].$$

The pathwise inequality $d_k^2 \le \bar\eta\, d_k^2/\eta_k$ and the definition of T_2 give the result. ∎

B.5.3 TWO CONVERGENCE LEMMAS

To pass from accumulated bounds to almost-sure convergence, we use two lemmas. The first controls sequences with summable conditional perturbations and is a special case of the almost-supermartingale principle of Robbins & Siegmund (1971). The second turns weighted summability and a bound on successive changes into convergence to zero.

Lemma B.15 (Summable conditional perturbations) *Let X_k, a_k, b_k be nonnegative, integrable, $\mathcal{F}_k$-measurable random variables such that*

$$\mathbb{E}[X_{k+1} \mid \mathcal{F}_k] \le X_k - a_k + b_k, \qquad \mathbb{E}\left[\sum_{k=0}^{\infty} b_k\right] < \infty.$$

Then X_k converges to a finite limit almost surely and $\sum_{k=0}^{\infty} a_k < \infty$ almost surely.

Proof. Let $B = \sum_{k=0}^{\infty} b_k$, an integrable random variable, and set

$$T_k = \mathbb{E}\left[\sum_{j=k}^{\infty} b_j \,\middle|\, \mathcal{F}_k\right], \qquad Y_k = X_k + T_k.$$

The tower property gives $\mathbb{E}[T_{k+1} \mid \mathcal{F}_k] = T_k - b_k$. Thus $\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] \le Y_k - a_k$; in particular, Y_k is a nonnegative supermartingale and converges almost surely to a finite limit. Summing this last inequality in expectation gives $\sum_{k=0}^{K-1} \mathbb{E}[a_k] \le \mathbb{E}[Y_0]$. Monotone convergence implies $\mathbb{E}[\sum_{k=0}^{\infty} a_k] < \infty$, so the series is finite almost surely.

Finally, $T_k = \mathbb{E}[B \mid \mathcal{F}_k] - \sum_{j=0}^{k-1} b_j$. The conditional-expectation martingale converges almost surely to B, since B is integrable and measurable with respect to the sigma-algebra generated by the increasing histories. The partial sum converges to B as well. Hence $T_k \to 0$, and $X_k = Y_k - T_k$ converges. The only martingale facts used here are the convergence of a nonnegative supermartingale and the convergence of conditional expectations of an integrable random variable along increasing sigma-algebras. ■

Lemma B.16 (Vanishing from weighted summability and controlled increments) *Let $a_k \ge 0$ and $t_k > 0$ be deterministic sequences such that*

$$\sum_{k=0}^{\infty} t_k = \infty, \qquad \sum_{k=0}^{\infty} t_k a_k^2 < \infty, \qquad |a_{k+1} - a_k| \le C t_k$$

for some finite $C \ge 0$. Then $a_k \to 0$.

Proof. Weighted summability and $\sum_k t_k = \infty$ imply $\liminf_k a_k = 0$. If $C = 0$, the sequence is constant and must be zero. Suppose $C > 0$ and $\limsup_k a_k > 0$. Choose $\varepsilon > 0$ so that $a_k \ge 2\varepsilon$ infinitely often. Since the lower limit is zero, there are infinitely many disjoint intervals $[p, q)$ with

$$a_p \ge 2\varepsilon, \qquad a_q \le \varepsilon, \qquad a_k > \varepsilon \quad (p \le k < q).$$

One constructs them successively by choosing a high point and the first subsequent point below ε. On each interval,

$$\varepsilon \le a_p - a_q \le C \sum_{k=p}^{q-1} t_k, \qquad \sum_{k=p}^{q-1} t_k a_k^2 \ge \varepsilon^2 \sum_{k=p}^{q-1} t_k \ge \frac{\varepsilon^3}{C}.$$

Infinitely many disjoint intervals contradict weighted summability. ■

B.5.4 PROOF OF COROLLARY 3.4

Proof. Use A, Σ from (77). The scale assumptions imply $0 < \beta_k = c_\beta \eta_k \le 1$ and $T_2 < \infty$. Sampled initialization is integrable by (64). Proposition B.14 therefore gives the stated error budget with $\theta = A\bar{\eta}/c_\beta < \chi^2/4$ and the finite constant in (78).

Although the scales here may be history-dependent, Proposition B.9 applies. Its uniform bounds, followed by monotone convergence, give

$$\begin{aligned} \mathbb{E}\left[\sum_{k=0}^{\infty} d_k^2/\eta_k\right] &< \infty, \\ \mathbb{E}\left[\sum_{k=0}^{\infty} \eta_k \mathcal{E}_k\right] &< \infty, \\ \mathbb{E}\left[\sum_{k=0}^{\infty} \eta_k \| \operatorname{grad} f(y_k)\|^2\right] &< \infty. \end{aligned} \tag{79}$$

Each of these nonnegative random series is consequently finite almost surely.

To prove convergence of the estimation error itself, combine (76) with the step bound $d_k \le V\eta_k/\chi$ and $(1-\beta_k)^2 \le 1 - c_\beta \eta_k$. This yields

$$\mathbb{E}[\mathcal{E}_{k+1} \mid \mathcal{H}_k] \le \mathcal{E}_k - c_\beta \eta_k \mathcal{E}_k + J\eta_k^2, \qquad J := \frac{AV^2}{\chi^2} + \Sigma c_\beta^2.$$

Since η_k and $\mathcal{E}_k$ are $\mathcal{F}_k$-measurable, conditioning once more gives

$$\mathbb{E}[\mathcal{E}_{k+1} \mid \mathcal{F}_k] \le \mathcal{E}_k - c_\beta \eta_k \mathcal{E}_k + J\eta_k^2. \tag{80}$$

Apply Lemma B.15 with $X_k = \mathcal{E}_k$, $a_k = c_\beta \eta_k \mathcal{E}_k$, and $b_k = J\eta_k^2$. Its hypotheses follow from clipping and $T_2 < \infty$. Thus $\mathcal{E}_k$ has an almost-sure limit and $\sum_k \eta_k \mathcal{E}_k < \infty$ almost surely. On any path where that limit were positive, $\sum_k \eta_k = \infty$ would force this last series to diverge. Hence $\mathcal{E}_k \to 0$ almost surely.

It remains to prove convergence of all gradient norms, rather than only a subsequence. Let $a_k = \| \operatorname{grad} f(y_k)\|$. By isometry of parallel transport and gradient Lipschitz continuity,

$$|a_{k+1} - a_k| \le \| \operatorname{grad} f(y_{k+1}) - P_k \operatorname{grad} f(y_k)\| \le L_f d_k \le \frac{L_f V}{\chi} \eta_k. \tag{81}$$

On the probability-one event where $\sum_k \eta_k = \infty$ and the series in (79) are finite, apply Lemma B.16 with $t_k = \eta_k$. It gives $\| \operatorname{grad} f(y_k)\| \to 0$. This proves both assertions in (22). ■

B.5.5 Deterministic schedules and a stronger error bound

The scale conditions used above admit simple deterministic choices. We give one such family and then refine the error bound for deterministic scales.

For $\eta_k = \eta_0 (k+1)^{-\gamma}$ with $1/2 < \gamma \le 1$, the series $\sum_k \eta_k$ diverges, while

$$\sum_{k=0}^{\infty} \eta_k^2 \le \eta_0^2 \left(1 + \frac{1}{2\gamma - 1}\right). \tag{82}$$

The latter follows by bounding $\sum_{n=1}^{\infty} n^{-2\gamma}$ by $1 + \int_1^{\infty} t^{-2\gamma}\, \mathrm{d}t$. Taking $0 < \eta_0 \le \bar{\eta}$ verifies the remaining schedule conditions in the corollary. An explicit finite remainder is

$$C_{\mathrm{err}} \le \frac{S_0}{c_\beta} + \Sigma c_\beta \eta_0^2 \left(1 + \frac{1}{2\gamma - 1}\right).$$

The displacement bound also gives a stronger, alternative error budget. Taking expectations in (80) and summing directly gives

$$\sum_{k=0}^{K-1} \mathbb{E}[\eta_k \mathcal{E}_k] \le \frac{\mathbb{E}[\mathcal{E}_0] + JT_2}{c_\beta} =: C_{\mathrm{err}}^{(0)} < \infty. \tag{83}$$

Thus one may use $\theta = 0$ in Proposition B.9. Under the other refresh-free assumptions, this alternative proves stationarity without the additional smallness condition $A\bar{\eta}/c_\beta < \chi^2/4$, at the cost of the larger remainder $C_{\mathrm{err}}^{(0)}$. The main-text condition is a valid sufficient condition and gives the smaller explicit remainder in (78).

B.6 INFINITE-RUN CONSEQUENCES

We finish the convergence analysis with consequences for infinite runs: almost-sure stationarity under exact refreshes, convergence of objective values, and eventual avoidance of the boundary of $\mathcal{K}_\alpha$.

B.6.1 EXACT REFRESHES ALSO GIVE ALMOST-SURE STATIONARITY

Proposition B.17 (Summability under exact refreshes) *Consider an infinite run under Corollary 3.2 or 3.3, with the same fixed scale and estimator parameters at every horizon. Then*

$$\sum_{k=0}^{\infty} d_k^2 < \infty, \qquad \sum_{k=0}^{\infty} \mathcal{E}_k < \infty, \qquad \sum_{k=0}^{\infty} \|\operatorname{grad} f(y_k)\|^2 < \infty \quad \textit{almost surely}. \tag{84}$$

In particular, both $\mathcal{E}_k$ and the gradient norm tend to zero almost surely.

Proof. For the admissible θ and $C_{\mathrm{err}} = 0$, Proposition B.9 gives uniform expected bounds on the three finite sums after division or multiplication by the fixed $\eta > 0$. Let $K \to \infty$, using complete epochs if desired. Monotone convergence makes each infinite series have finite expectation. A nonnegative random variable with finite expectation is finite almost surely, proving (84). The terms of each convergent nonnegative series tend to zero. ■

B.6.2 OBJECTIVE VALUES AND ACCUMULATION POINTS

Having established that the gradients vanish, we now examine the objective values and the accumulation points of the iterates.

Proposition B.18 (Objective convergence and stationary accumulation points) *In any infinite run covered by one of the three main-text corollaries, $f(y_k)$ converges almost surely to a finite random value. Every accumulation point is stationary.*

Proof. The preceding results give $\sum_k \eta_k \mathcal{E}_k < \infty$ and $\|\operatorname{grad} f(y_k)\| \to 0$ almost surely in each regime. Fix such a sample path, and put $u_k = (2/\chi)\eta_k \mathcal{E}_k$. By (48), $f(y_{k+1}) \le f(y_k) + u_k$. Therefore

$$f(y_{k+1}) + \sum_{j=k+1}^{\infty} u_j \le f(y_k) + \sum_{j=k}^{\infty} u_j.$$

This sequence is decreasing and bounded below by $f_\star$. It converges, and its tail sum tends to zero, proving convergence of $f(y_k)$. Compactness of $\mathcal{K}_\alpha$ gives accumulation points. For a convergent subsequence, continuity of the gradient norm and its convergence to zero show that the limit is stationary. ■

No uniqueness assumption has been made, so these statements do not by themselves imply convergence of the entire sequence to one point.

B.6.3 EVENTUAL INACTIVITY OF THE SAFE-BALL CONSTRAINT

Finally, we return to the radial bound. Combined with vanishing gradients, it shows that the iterates eventually stay strictly inside $\mathcal{K}_\alpha$, so the constraint no longer contributes a normal term to the inner residual.

Corollary B.19 (The safe-ball boundary is eventually avoided) *In any infinite run covered by Corollary 3.2, 3.3, or 3.4,*

$$\limsup_{k\to\infty} d(c, y_k) \le W_{\mathrm{abs}} r < R_\alpha \qquad \textit{almost surely}. \tag{85}$$

The iterates eventually lie in $\operatorname{int}(\mathcal{K}_\alpha)$. Consequently, the normal term in every sufficiently late accepted inner update is zero.

Proof. The radial bound and Cauchy–Schwarz imply, for every $y \in \mathcal{K}_\alpha$,

$$2[d(c, y) - W_{\mathrm{abs}} r]_+ \le \| \operatorname{grad} f(y) \|.$$

When $d(c, y) > W_{\mathrm{abs}} r \ge r$, this is (33); otherwise its left side is zero. Apply this to the iterates and use almost-sure gradient convergence. It gives (85). More explicitly, let $\varepsilon = (R_\alpha - W_{\mathrm{abs}} r)/2 > 0$. For all sufficiently large k on each such sample path, $d(c, y_k) \le W_{\mathrm{abs}} r + \varepsilon < R_\alpha$. The normal cone at those points is $\{0\}$. ■

This corollary uses gradient convergence in the stated estimator regimes. The weighted estimate in Theorem 3.1 alone, with arbitrary positive scales, does not imply the same conclusion: for example, a finite total scale need not force stationarity.

C GEOMETRY OF THE MANIFOLDS IN THE NUMERICAL EXPERIMENTS

C.1 THE EMBEDDED TORUS

We denote the minor radius by b in this subsection to distinguish it from the data-support radius r in Assumption 2.1. For $a > b > 0$, set

$$\mathcal{F}(\theta, \phi) = ((a + b\cos\phi)\cos\theta, (a + b\cos\phi)\sin\theta, b\sin\phi), \qquad h(\phi) = a + b\cos\phi.$$

The coordinate tangent vectors are orthogonal, with squared norms $h(\phi)^2$ and b^2. Thus, the metric and an orthonormal frame are

$$g = \begin{pmatrix} h(\phi)^2 & 0 \\ 0 & b^2 \end{pmatrix}, \qquad e_\theta = h(\phi)^{-1}\partial_\theta, \qquad e_\phi = b^{-1}\partial_\phi. \tag{86}$$

A vector with orthonormal coordinates (u_1, u_2) has coordinate components $(u_1/h(\phi), u_2/b)$.

In the implementation, tangent vectors are stored in this orthonormal frame, so their Euclidean norm equals their Riemannian norm. Geodesics on the embedded torus admit no elementary closed form, and evaluating the logarithm requires solving a two-point boundary-value problem (Jantzen, 2010). We approximate the exponential map using fixed-step fourth-order Runge–Kutta integration and compute the logarithm by Newton shooting against this numerical exponential. Parallel transport is computed by integrating the differential equation for it along the numerical geodesic. The formulas below describe the exact geometry; the computed maps, distances, and gradients approximate these quantities. The shooting residual measures endpoint agreement for the numerical exponential and excludes the integration error.

C.1.1 CURVATURE AND THE GEODESIC EQUATIONS

Starting from the induced metric, we compute the curvature bounds used in the analysis and the geodesic equations for evaluating the exponential and logarithm maps.

The unit normal $N = (\cos\phi\cos\theta, \cos\phi\sin\theta, \sin\phi)$ gives second fundamental coefficients $\langle \partial_{\theta\theta}\mathcal{F}, N\rangle = -h(\phi)\cos\phi$ and $\langle \partial_{\phi\phi}\mathcal{F}, N\rangle = -b$; the mixed coefficient is zero. The ratio of determinants of the second and first fundamental forms, therefore, gives

$$K(\phi) = \frac{\cos\phi}{b(a + b\cos\phi)}. \tag{87}$$

As a function of $u = \cos\phi$, its derivative is $a/[b(a + bu)^2] > 0$. Hence

$$-\frac{1}{b(a-b)} \le K \le \frac{1}{b(a+b)}.$$

At $a = 2$ and $b = 0.7$, these bounds are $-100/91$ and $100/189$, respectively.

The nonzero Christoffel symbols are

$$\Gamma^\theta_{\theta\phi} = \Gamma^\theta_{\phi\theta} = -\frac{b\sin\phi}{h(\phi)}, \qquad \Gamma^\phi_{\theta\theta} = \frac{h(\phi)\sin\phi}{b}.$$

Consequently, the exponential map is given by the time-one solution of

$$\ddot\theta - \frac{2b\sin\phi}{h(\phi)}\dot\theta\dot\phi = 0, \qquad \ddot\phi + \frac{h(\phi)\sin\phi}{b}\dot\theta^2 = 0, \tag{88}$$

with the prescribed point and tangent vector as initial conditions, See (Jantzen, 2010) for a derivation of these equations and their Clairaut first integral $h(\phi)^2\dot\theta$. For an initial vector with orthonormal coordinates (u_1, u_2), these velocities are $\dot\theta(0) = u_1/h(\phi(0))$ and $\dot\phi(0) = u_2/b$. The exact geodesic has constant speed, so a minimizing time-one geodesic has length $\sqrt{u_1^2 + u_2^2}$.

Within a normal neighborhood, the logarithm is the initial tangent vector whose time-one geodesic reaches the target. We approximate this vector by solving the corresponding numerical endpoint equation. We use a consistent unwrapped angular chart for the local computations. Angular differences provide an initial guess for shooting, but are not themselves Riemannian logarithms.

C.1.2 PARALLEL TRANSPORT AND GRADIENTS

The same metric determines how tangent vectors are transported along geodesics and how gradients are represented in coordinates. These operations are used in the gradient estimators and the inner solves.

Along a geodesic $(\theta(t), \phi(t))$, a tangent vector field with coordinate components (w^θ, w^ϕ) is parallel precisely when

$$\dot{w}^\theta = \frac{b \sin\phi}{h(\phi)}(\dot{\theta} w^\phi + \dot{\phi} w^\theta), \qquad \dot{w}^\phi = -\frac{h(\phi)\sin\phi}{b}\dot{\theta} w^\theta. \tag{89}$$

These equations follow by setting $D_t w^i = \dot{w}^i + \Gamma^i_{j\ell}\dot{\gamma}^j w^\ell$ equal to zero. Substitution into the derivative of $h(\phi)^2(w^\theta)^2 + b^2(w^\phi)^2$ shows that exact parallel transport preserves the norm. Integrating (89) together with the geodesic equations gives the numerical transport used by the estimators; norm preservation holds up to numerical error.

For a scalar function $F(\theta, \phi)$,

$$\operatorname{grad} F = \frac{\partial_\theta F}{h(\phi)^2}\partial_\theta + \frac{\partial_\phi F}{b^2}\partial_\phi, \qquad \|\operatorname{grad} F\|^2 = \frac{(\partial_\theta F)^2}{h(\phi)^2} + \frac{(\partial_\phi F)^2}{b^2}. \tag{90}$$

Equivalently, the orthonormal coordinates of $\operatorname{grad} F$ are $(\partial_\theta F/h(\phi), \partial_\phi F/b)$. Their Euclidean norm gives the Riemannian gradient norm. The Euclidean norm of the raw partial derivatives $(\partial_\theta F, \partial_\phi F)$ generally does not.

For squared-distance losses, the exact intrinsic formula is $\operatorname{grad} d^2(y, z) = -2\operatorname{Log}_y z$ for z in a normal neighborhood of y (Viaclovsky, 2011). In the implementation, we evaluate this expression using the numerical logarithm, so the reported gradient norms approximate the corresponding Riemannian gradient norms.

The uniform curvature bounds above provide admissible $\Lambda_\pm$. For the stated torus, the restriction on D in Assumption 2.1 becomes $D < \pi\sqrt{189}/20$. The remaining requirements concern the radii and the normality and convexity of the chosen balls; these properties must be established separately for the region used in the experiment.

C.2 $SO(3)$ - ROTATION MANIFOLD WITH THE ANGLE METRIC

C.2.1 METRIC, EXPONENTIAL MAP, AND LOGARITHM

Let $SO(3) = \{R \in \mathbb{R}^{3\times 3} : R^\top R = I, \det R = 1\}$. For $u \in \mathbb{R}^3$, define

$$\widehat{u} = \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix}, \qquad \operatorname{vee}(\widehat{u}) = u.$$

We represent $R\widehat{u} \in T_R SO(3)$ by its left-trivialized coordinate vector u. The metric is

$$\langle R\widehat{u}, R\widehat{v}\rangle_R = -\tfrac{1}{2}\operatorname{tr}(\widehat{u}\widehat{v}) = u^\top v. \tag{91}$$

For $\theta = \|a\|$, Rodrigues' formula (Rodrigues, 1840) is

$$\operatorname{Exp}_R(R\widehat{a}) = R\exp(\widehat{a}), \qquad \exp(\widehat{a}) = I + \frac{\sin\theta}{\theta}\widehat{a} + \frac{1-\cos\theta}{\theta^2}\widehat{a}^2, \tag{92}$$

with continuous values for the coefficients at zero. For $A = R^\top Q$ with rotation angle $0 < \theta < \pi$, write

$$a_R(Q) := \operatorname{vee}\left(\frac{\theta}{2\sin\theta}(A - A^\top)\right), \qquad \theta = \arccos\left(\frac{\operatorname{tr} A - 1}{2}\right). \tag{93}$$

Then $\operatorname{Log}_R Q = R\widehat{a_R(Q)}$ and $d(R, Q) = \|a_R(Q)\| = \theta$. The principal logarithm is smooth for angles less than π; angles equal to π belong to the cut locus and are excluded by the normal-neighborhood assumptions. Near zero, the coefficients in these formulas are evaluated by their continuous limits or Taylor series.

C.2.2 Curvature and exact parallel transport

For a bi-invariant metric, the Levi-Civita connection (Lee, 2018) on left-invariant fields is $\nabla_U V = \frac{1}{2}[U, V]$. The sectional curvature of an orthonormal pair is $\frac{1}{4}\|[U, V]\|^2$. Here $[\widehat{u}, \widehat{v}] = \widehat{u \times v}$, so $\sec = 1/4$ under (91). The injectivity radius is π: each rotation of angle less than π has a unique minimizing axis-angle vector, whereas rotations of angle π have the two minimizing choices a and $-a$.

Consider the geodesic $\gamma(t) = R\exp(t\widehat{a})$ from R to $Q = R\exp(\widehat{a})$, with $\|a\| < \pi$. A vector field $\gamma(t)\widehat{v(t)}$ is parallel if and only if

$$\dot{v}(t) + \tfrac{1}{2}(a \times v(t)) = 0.$$

Solving this constant-coefficient equation gives, in left-trivialized form coordinates,

$$P_{R\to Q} v = \exp(-\tfrac{1}{2}\widehat{a})v. \tag{94}$$

The matrix is orthogonal, so this is an isometry. It is the parallel transport along the minimizing geodesic, rather than the identity map between the coordinate vectors at different rotations.

C.2.3 Loss Hessians and admissible constants

In constant curvature $1/4$, the transverse Jacobi field along a unit-speed geodesic is proportional to $\sin(t/2)$. Applying the second-variation calculation of Lemma B.1 therefore gives the exact formula, for $\theta = d(R, Q) < \pi$,

$$\mathrm{Hess}\left(\tfrac{1}{2}d^2(\cdot, Q)\right)(R)[u, u] = \|u_\parallel\|^2 + \frac{\theta}{2}\cot\left(\frac{\theta}{2}\right)\|u_\perp\|^2. \tag{95}$$

The radial and transverse components are relative to the geodesic joining Q and R. At $R = Q$ the formula is interpreted by continuity. Thus, for the theoretical distance bounds D and $\bar{r}$,

$$\delta = \frac{D}{2}\cot(D/2), \quad \zeta = 1, \qquad \delta_p = \frac{\bar{r}}{2}\cot(\bar{r}/2), \quad \zeta_p = 1, \qquad G_\ell = 2D, \quad L_\ell = 2. \tag{96}$$

These choices require $D < \pi$. A sufficient outer-ball choice for the smooth-logarithm and strong-convexity requirements is $\rho < \pi/2$, together with the nested radii in Assumption 2.1. Such balls lift to convex spherical balls in the quaternion covering, and all pairwise distances in the outer ball are less than $2\rho < \pi$.

For responses within a ball of radius r_{supp} around the chosen support center C, one may use $r = r_{\mathrm{supp}}$. The reported support bound is 0.15 radians; the constants still depend on the chosen R and $\bar{r}$ through $D = R + \bar{r} + r$. The mass of the positive model is $W_+ = 1 + W_-$, so its Hessian floor is $2(1 + W_-)\delta$ for each negative-mass setting. It is not a common numerical floor for both values of W_-.

C.2.4 A computable bound for the logarithmic correction

The following construction gives a finite bound without estimating Hessians at a finite sample of rotations. Define, for $0 < t < \pi$,

$$h_0(t) := \frac{1 - (t/2)\cot(t/2)}{t^2}, \qquad s_0(t) := \frac{t}{2\sin(t/2)},$$

with $h_0(0) = 1/12$, $h_0'(0) = 0$, and $s_0(0) = 1$. Then the following is an admissible constant on local balls of radius $\bar{r} < \pi$:

$$L_{\log}^{SO(3)} := \max_{0\le t\le \bar{r}} s_0(t)\left(\frac{1}{2} + 2t|h_0(t)| + t^2|h_0'(t)|\right). \tag{97}$$

This bound is conservative, but its defining maximum is over a continuous scalar function on a compact interval.

Proof. Let $Q(a) = P\exp(\widehat{a})$ and use left-trivialized tangent coordinates at $Q(a)$. The differential of the exponential in these coordinates is the right Jacobian

$$J_r(a) = I - \frac{1 - \cos t}{t^2}\widehat{a} + \frac{t - \sin t}{t^3}\widehat{a}^2, \qquad t = \|a\|. \tag{98}$$

Indeed,

$$Q(a)^{-1}\mathrm{D}Q(a)[v] = \int_0^1 \exp(-s\widehat{a})\widehat{v}\exp(s\widehat{a})\,\mathrm{d}s;$$

integrating Rodrigues' formula (Rodrigues, 1840) gives (98). Its inverse is

$$J_r(a)^{-1} = I + \tfrac{1}{2}\widehat{a} + h_0(t)\widehat{a}^2.$$

The radial singular value of J_r is one. Its transverse singular values are $2\sin(t/2)/t$. Consequently, $\|J_r(a)^{-1}\|_{\mathrm{op}} = s_0(t)$. Differentiating its inverse formula in direction v and using $\|\widehat{a}\|_{\mathrm{op}} = t$, $\|\widehat{v}\|_{\mathrm{op}} = \|v\|$, and $|\mathrm{D}t[v]| \le \|v\|$ gives

$$\|\mathrm{D}(J_r^{-1})(a)[v]\|_{\mathrm{op}} \le \left(\tfrac{1}{2} + 2t|h_0(t)| + t^2|h_0'(t)|\right)\|v\|.$$

The formula extends continuously to $a = 0$.

Now take the geodesic $Q(\tau) = Q(a)\exp(\tau\widehat{u})$ and let $a(\tau) = a_P(Q(\tau))$. Its left-trivialized velocity is the constant u, so $a'(\tau) = J_r(a(\tau))^{-1}u$ and

$$a''(0) = \mathrm{D}(J_r^{-1})(a)[J_r(a)^{-1}u]u.$$

For the function $Q \mapsto \langle \xi, \mathrm{Log}_P Q\rangle_P$, the Hessian in direction $Q(a)\widehat{u}$ is $\xi^\top a''(0)$, where ξ is represented in the same coordinates at P. The preceding norm bounds give $|\xi^\top a''(0)| \le L_{\log}^{SO(3)}\|\xi\|\|u\|^2$, which is the required inequality. ■

C.2.5 Pullback gradients for a rotation-valued inner problem

Write the model in the coordinates $Q(a) = R_k\exp(\widehat{a})$, where $\|a\| < \bar{r} < \pi$, and let q_k denote the coordinate vector of its tangent correction. Then

$$\psi_k(a) := \varphi_k(Q(a)) = \widehat{g}_k(Q(a)) - q_k^\top a + \frac{\lambda_k}{2}\|a\|^2.$$

If $G_{\widehat{g}_k}(Q)$ is the left-trivialized Riemannian gradient of the positive model at Q, differentiation gives

$$\nabla_a\psi_k(a) = J_r(a)^\top G_{\widehat{g}_k}(Q(a)) - q_k + \lambda_k a, \qquad G_{\varphi_k}(Q(a)) = J_r(a)^{-\top}\nabla_a\psi_k(a). \tag{99}$$

The loss gradients entering the first term are $G_{\ell_i}(Q) = -2a_Q(z_i)$. Equation (99) identifies the intrinsic gradient needed by the stopping rule; a coordinate gradient should not be substituted for it without accounting for J_r.

C.3 $S^2 \times S^1$ - Product of the sphere and circle

C.3.1 Metric, exponential map, and logarithm

We consider the product manifold

$$\mathcal{M} = S^2 \times S^1,$$

where

$$S^2 = \{x \in \mathbb{R}^3 : \|x\| = 1\}, \qquad S^1 = \mathbb{R}/(2\pi\mathbb{Z}).$$

We represent a point by

$$p = (x, \theta), \qquad x \in S^2, \quad \theta \in S^1.$$

Its tangent space is the product

$$T_p\mathcal{M} = T_xS^2 \times T_\theta S^1 = \{(u, a) \in \mathbb{R}^3 \times \mathbb{R} : x^\top u = 0\}.$$

Here, the scalar $a \in \mathbb{R}$ represents the tangent component in the angular coordinate of S^1.

We equip $\mathcal{M}$ with the standard product metric

$$\langle (u, a), (v, b)\rangle_{(x,\theta)} = u^\top v + ab,$$

so that

$$\|(u, a)\|^2_{(x,\theta)} = \|u\|^2 + a^2.$$

All geometric operations, therefore, separate into the corresponding operations on S^2 and S^1.

For $u \in T_x S^2$, let $r = \|u\|$. The exponential map on the unit sphere is

$$\mathrm{Exp}_x^{S^2}(u) = \cos r\, x + \frac{\sin r}{r} u,$$

with the usual continuous interpretation at $r = 0$. Since the circle is parametrized by its angle, its exponential map is simply addition modulo 2π. Hence, for $\xi = (u, a) \in T_{(x,\theta)}\mathcal{M}$,

$$\mathrm{Exp}_{(x,\theta)}(u, a) = \left(\cos \|u\|\, x + \frac{\sin \|u\|}{\|u\|} u,\ \theta + a \pmod{2\pi} \right). \tag{100}$$

Conversely, let

$$p = (x, \theta), \qquad q = (y, \phi),$$

and suppose $y \neq -x$. Define

$$c = x^\top y, \qquad \alpha = \arccos(c) \in [0, \pi),$$

and let

$$\beta := \mathrm{wrap}_{[-\pi,\pi)}(\phi - \theta) \tag{101}$$

denote the principal angular displacement on S^1. The spherical logarithm is

$$\mathrm{Log}_x^{S^2}(y) = \frac{\alpha}{\sin \alpha}\big(y - \cos \alpha\, x\big) = \frac{\alpha}{\sqrt{1 - c^2}}(y - cx),$$

for $0 < \alpha < \pi$, with the continuous value 0 when $x = y$. Therefore, the product logarithm is

$$\mathrm{Log}_{(x,\theta)}(y, \phi) = \left(\frac{\alpha}{\sin \alpha}(y - \cos \alpha\, x),\ \beta \right). \tag{102}$$

The corresponding geodesic distance is

$$d^2\big((x, \theta), (y, \phi)\big) = \alpha^2 + \beta^2. \tag{103}$$

In numerical computations, we evaluate

$$\alpha = \mathrm{atan2}\big(\|y - (x^\top y)x\|, x^\top y\big),$$

which is equivalent to $\arccos(x^\top y)$ but is more accurate when x and y are very close.

The cut locus of (x, θ) is reached when either the spherical component is antipodal, $y = -x$, or the circle displacement has magnitude π. Thus, the injectivity radius of the product manifold is

$$\mathrm{inj}(S^2 \times S^1) = \pi.$$

C.3.2 Curvature and exact parallel transport

The Levi–Civita connection of a Riemannian product acts componentwise, and its curvature tensor is the direct sum of the curvature tensors of the two factors (Petersen, 2006). For tangent vectors

$$X = (u, a), \qquad Y = (v, b) \in T_{(x,\theta)}(S^2 \times S^1),$$

the S^1 factor contributes no intrinsic curvature, whereas the unit sphere has constant sectional curvature one. Consequently,

$$\langle R(X, Y)Y, X \rangle = \|u\|^2 \|v\|^2 - (u^\top v)^2. \tag{104}$$

For a two-plane $\Pi = \mathrm{span}\{X, Y\}$, its sectional curvature is therefore

$$\sec(\Pi) = \frac{\|u\|^2 \|v\|^2 - (u^\top v)^2}{(\|u\|^2 + a^2)(\|v\|^2 + b^2) - (u^\top v + ab)^2}. \tag{105}$$

In particular,

$$0 \leq \sec(\Pi) \leq 1.$$

A plane contained entirely in the S^2 factor has sectional curvature 1, whereas a mixed plane spanned by a spherical direction and the circle direction has sectional curvature 0. Hence, $S^2 \times S^1$ has nonnegative but nonconstant sectional curvature.

We next give the exact parallel transport used in the implementation. Let

$$p = (x, \theta), \qquad q = (y, \phi),$$

and assume that q does not lie in the cut locus of p. For $\xi = (u, a) \in T_p\mathcal{M}$, parallel transport along the unique minimizing product geodesic separates into its two factors:

$$P_{p\to q}(u, a) = \left(P^{S^2}_{x\to y}u,\ a\right). \tag{106}$$

The circle component remains unchanged because the angular coordinate has zero Christoffel symbol. On the unit sphere, exact parallel transport along the minimizing great-circle geodesic is Kim et al. (2021)

$$P^{S^2}_{x\to y}u = u - \frac{u^\top y}{1 + x^\top y}(x + y), \qquad y \neq -x. \tag{107}$$

Thus, the product transport has the explicit form

$$P_{(x,\theta)\to(y,\phi)}(u, a) = \left(u - \frac{u^\top y}{1 + x^\top y}(x + y),\ a\right). \tag{108}$$

One directly verifies that

$$y^\top P^{S^2}_{x\to y}u = 0$$

and that parallel transport preserves the product metric:

$$\|P_{p\to q}(u, a)\|_q = \|(u, a)\|_p.$$

The formula becomes singular at $y = -x$, reflecting the fact that the minimizing spherical geodesic, and hence parallel transport along such a geodesic, is no longer uniquely determined.