The Influence of Curvature on Deep Linear Networks on Riemannian Manifolds

Zhixuan Li[1] Taro Iyadomi[2] Luke Jones[1] Sean He[1]
Muztaba Syed[1] William Chang[1] Bo Xiong[3]

[1]University of California, Los Angeles [2]Northwestern University [3]Iowa State University

Abstract

We study how target-space curvature affects the optimization of deep linear networks on constant-curvature manifolds. We compare an intrinsic squared-geodesic loss with a tangent-space surrogate that reduces to ordinary least squares. Our main result is that curvature enters optimization as a discretization cost: gradient flow admits a curvature-free descent inequality, whereas finite-step gradient descent introduces a Taylor remainder controlled by a smoothness constant that scales with the Rauch factor $S_K(R)$. Consequently, the stable step-size ceiling satisfies $\eta_K^\star = O(1/S_K(R)^2)$, shrinking exponentially in $\sqrt{|K|}R$ under negative curvature, whereas positive curvature carries no Rauch-factor penalty. In contrast, the tangent-space surrogate has a curvature-independent Euclidean rate, showing that the dependence arises from measuring error intrinsically. We further establish linear convergence, absence of spurious critical points within the analysis tube, a curvature-adaptive step-size schedule, a curvature-dependent anisotropy in the implicit bias, and tightness of the step-size bound. Experiments confirm the predicted $S_K(R)^2$ scaling.

1 Introduction

Many kinds of data are more naturally described on a curved Riemannian manifold than in Euclidean space, and a growing body of work places inputs, embeddings, and network activations on such manifolds. Hyperbolic space is the most prominent example: its exponential volume growth allows low-distortion embeddings of hierarchical and tree-like data where Euclidean embeddings provably fail (Sala et al., 2018), which has driven work on Poincaré and Lorentz embeddings (Nickel and Kiela, 2017, 2018; Gu et al., 2018), hyperbolic neural networks (Ganea et al., 2018; Chen et al., 2022), and mixed-curvature architectures (Skopek et al., 2019; Wang et al., 2021). The empirical literature is substantial, yet the effect of target-manifold curvature on the optimization dynamics of a deep manifold-valued network has not been characterized.

The simplest setting in which to isolate the role of curvature is the deep linear network, the model whose Euclidean optimization theory is best understood (Saxe et al., 2013; Arora et al., 2018; Du and Hu, 2019). We consider such a network whose outputs live on a constant-curvature manifold and compare two natural training losses, an *intrinsic* loss measuring the squared geodesic distance on the manifold and a *tangent-space surrogate* that pulls the targets back to Euclidean space and performs ordinary least squares. In Euclidean space the two coincide, and Arora et al. (2018) proved linear convergence. On a curved manifold they come apart, and curvature must enter the convergence analysis somewhere; this paper locates where. Figure 1 summarizes the mechanism.

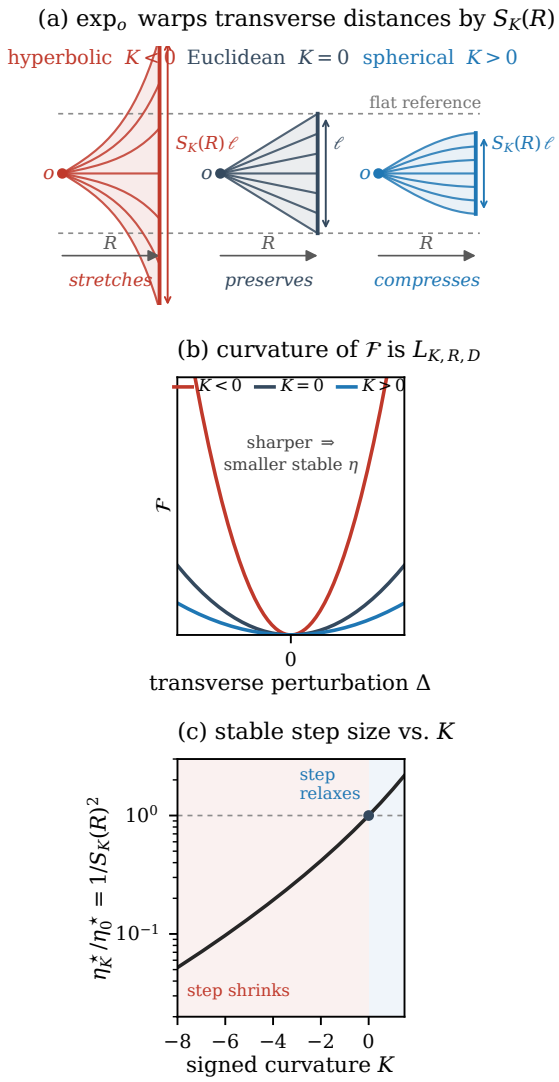


Figure 1: Why curvature enters optimization. **(a)** A transverse displacement of size ℓ at the base point o arrives at radius R with size $S_K(R)\,\ell$: stretched for $K<0$, unchanged for $K=0$, compressed for $K>0$. **(b)** Because the intrinsic loss measures error after this warping, a transverse cross-section of $\mathcal{F}$ has curvature $L_{K,R,D} \sim S_K(R)^2$, so the hyperbolic landscape is sharper. **(c)** Gradient descent is stable only for $\eta < 2/L_{K,R,D}$, so the transverse ceiling plotted here scales as $1/S_K(R)^2$, decaying exponentially in $\sqrt{\kappa}R$ for $K<0$; for $K>0$ the unstretched radial direction removes this factor from the global ceiling (Corollary 11). Panels drawn at $R=1.2$. Curvature enters Theorem 3 through the step-size bound, not through the coefficient of the guaranteed decrease.

Our Contributions. Curvature does affect the optimization dynamics, and the effect is structured. For a deep linear network on a constant-curvature manifold trained under the intrinsic squared-geodesic loss, our core result is a curvature-explicit convergence theory:

1. A **descent lemma** (Theorem 3) following Arora et al. (2018). Its gradient-flow case (Corollary 4) carries no curvature constraint; passing to discrete time introduces $L_{K,R,D}$ through the Taylor remainder, isolating the step-size penalty as a discretization cost.
2. A **linear convergence theorem** (Theorem 9) with ceiling $\eta_K^\star = O(1/S_K(R)^2)$, which decays exponentially in $\sqrt{|K|}\,R$ for $K<0$ while, by a positive-curvature extension (Corollary 11), $K>0$

carries no Rauch-factor penalty.

3. A **tangent-space control** (Proposition 12): on the same data and weights the surrogate reduces to the Euclidean setting of Arora et al. (2018) and its rate carries no curvature factor, localizing the dependence to the intrinsic loss.

4. Four **structural consequences**: no spurious critical points on the tube (Theorem 13); a curvature-adaptive schedule removing the Hessian-comparison penalty $H_K(D)$ (Proposition 15); a Rauch-weighted anisotropy that can change the limit point the flow selects (Proposition 17); and a sharpness construction showing $\eta_K^\star = \Theta(1/S_K(R)^2)$ is tight (Proposition 18).

5. An **empirical curvature sweep** (Section 5) in which the measured near-optimum sharpness tracks the predicted $S_K(R)^2$ across 4,200 configurations.

Proofs, extended related work, and three further experiments are given in the supplementary material.

2 Related Work

Non-Euclidean representation learning. Low-distortion embeddings of trees and hierarchical metrics (Sala et al., 2018) motivate Poincaré and Lorentz embeddings (Nickel and Kiela, 2017, 2018; Gu et al., 2018) and a range of manifold-valued architectures: Möbius, fully hyperbolic, and residual networks (Ganea et al., 2018; Tifrea et al., 2018; Chen et al., 2022; Bdeir et al., 2024; Van Spengler et al., 2023), hyperbolic attention (Gulcehre et al., 2018), pseudo-Riemannian (Xiong et al., 2022; Law, 2021), mixed- and adaptive-curvature models (Wang et al., 2021; Skopek et al., 2019; Cho et al., 2023; Fu et al., 2021), and hyperbolic graph networks (Zhang et al., 2021; Dai et al., 2021; Grover et al., 2025), with applications to embedding, segmentation, and generative modeling (Khrulkov et al., 2020; Atigh et al., 2022; Mathieu et al., 2019); see Yang et al. (2022) for a survey and the supplementary material for a fuller account. These architectures motivate the manifold-valued setting whose optimization we study; we treat target-manifold curvature as the object of interest and analyze its bias-free deep-linear case.

Deep linear network optimization. Trajectory analysis of gradient descent for deep linear networks began with Saxe et al. (2013). Landscape results establish the absence of spurious local minima (Lu and Kawaguchi, 2017; Laurent and Brecht, 2018; Nouiehed and Razaviyayn, 2022; Sun and Luo, 2016), and Arora et al. (2018) proved linear convergence under approximate balancedness and a deficiency margin; later work examines width (Du and Hu, 2019), identity initialization (Hardt and Ma, 2016), gradient alignment (Ji and Telgarsky, 2018), exponential convergence (Shamir, 2019), and residual architectures (Zou et al., 2020). Their factorization-induced preconditioner is the algebraic core of our descent lemma. Riemannian and convolutional treatments of the deep linear network (Bah et al., 2022; Menon, 2025; Achour et al., 2025) place the Riemannian structure on the parameter manifold induced by the factorization, whereas here it is carried by the space in which the targets live.

Optimization on manifolds. A separate tradition optimizes over manifold-valued *parameters* (Absil et al., 2008; Boumal, 2023; Hu et al., 2020; Zhang and Sra, 2016; Bonnabel, 2013), with variance-reduced (Zhang et al., 2016; Sato et al., 2019) and adaptive (Bécigneul and Ganea, 2019) extensions. In particular, Criscitiello and Boumal (2022, 2023) show that negative curvature obstructs acceleration for geodesically convex optimization, a lower bound consistent with our degradation $\eta_K^\star = O(e^{-2\sqrt{\kappa}R})$; there the curvature lives on the parameter space, whereas in ours the parameters are Euclidean and curvature enters only through the loss. Concurrently, Sarkar (2025) studies

curvature-adaptive generalization and Li et al. (2024) the Lipschitz properties of hyperbolic networks. To our knowledge, no prior work derives curvature-dependent descent lemmas, linear convergence, landscape results, adaptive schedules, or implicit-bias characterizations for deep manifold-valued networks.

3 Preliminaries

Layerwise Riemannian network. Let $N \geq 1$. For $j = 0, \ldots, N$, let $(\mathcal{M}_j, g_j)$ be a complete Riemannian manifold of dimension d_j with base point $o_j \in \mathcal{M}_j$, and let $V_j := T_{o_j}\mathcal{M}_j \simeq \mathbb{R}^{d_j}$. Data are $(x_i, y_i)_{i=1}^m$ with $x_i \in \mathcal{M}_0$, $y_i \in \mathcal{M}_N$. Layer j applies $\Phi_j(h; \theta_j) := \exp_{o_j}(F_j(\log_{o_{j-1}}(h); \theta_j))$ with $F_j : V_{j-1} \times \Theta_j \to V_j$ a tangent-space map. When every F_j is linear ($F_j(u) = W_j u$, no bias or activation), adjacent $\log_o$ and $\exp_o$ factors cancel and the network collapses to

$$h_N = \exp_{o_N}(W_N \cdots W_1 \log_{o_0}(x)). \tag{1}$$

We write $W_{a:b} := W_b \cdots W_a$ and $A := W_{1:N}$. On the Poincaré ball, (1) is the telescoping of N Möbius matrix–vector products $W \otimes_K x := \exp_o(W \log_o(x))$ (Ganea et al., 2018); the cancellation requires every layer to be bias-free and to use one base point o_j on both sides of each intermediate layer, and for $K > 0$ that every intermediate tangent vector stay inside the injectivity domain, so Möbius biases break (1) and lie outside our scope.

Model spaces and curvature conventions. Every target manifold below is a simply connected space form: the unique complete simply connected manifold of constant sectional curvature $K \in \mathbb{R}$ in each dimension, hyperbolic for $K < 0$, Euclidean for $K = 0$, and spherical for $K > 0$. We use the signed curvature K throughout, set $\kappa := |K|$, and state all theorems in terms of K alone. Where closed-form exp and log are needed we use the Poincaré ball for the hyperbolic theory and, whenever $K = 0$ must be crossed into the spherical case, the κ-stereographic model of Bachmann et al. (2020), which places sphere, plane, and hyperbolic ball in a single chart that is smooth in K. The two charts agree for $K < 0$, so no theorem depends on the choice. Throughout, the tangent-coordinate data $(\xi_i, \tilde{y}_i)$ are held fixed as K varies; the intermediate geometries cancel out of (1), so curvature acts only through the target manifold $\mathcal{M}_N$, where d_K and $\exp_{o_N}$ are evaluated. The experiments place every layer on the same model.

Two losses. Write $\xi_i := \log_{\mathbf{0}}^K(x_i)$ and $\tilde{y}_i := \log_{\mathbf{0}}^K(y_i)$. The *intrinsic loss* and *tangent-space surrogate* are

$$\mathcal{L}_K(W) := \tfrac{1}{2m} \textstyle\sum_{i=1}^m d_K\big(\exp_{\mathbf{0}}^K(A\xi_i), y_i\big)^2, \tag{2}$$

$$\mathcal{E}(W) := \tfrac{1}{2m} \textstyle\sum_{i=1}^m \|A\xi_i - \tilde{y}_i\|^2, \tag{3}$$

where d_K is the geodesic distance on the target manifold $\mathcal{M}_N$, so that all curvature dependence of the intrinsic loss is carried by the single subscript K. The intrinsic loss measures error on the manifold after applying the exponential map and so retains the distortions induced by curvature, whereas the surrogate compares tangent vectors in the Euclidean norm and reduces to ordinary least squares with target $\Phi := \frac{1}{m}\sum_i \tilde{y}_i \xi_i^*$. The two coincide in the Euclidean case. The end-to-end pullback is $\mathcal{F}(A) := \frac{1}{2m}\sum_i d_K(\exp_{o_N}(A\xi_i), y_i)^2$, so that $\mathcal{L}_K(W) = \mathcal{F}(W_{1:N})$.

Balancedness and deficiency margin. We impose two algebraic conditions on the initialization. The first controls the relative scaling of adjacent layers.

Definition 1 (δ-balancedness)**.** *Matrices $W_1, \dots, W_N$ are δ-balanced if $\|W_{j+1}^\top W_{j+1} - W_j W_j^\top\|_F \le \delta$ for all j.*

It holds with $\delta = 0$ for identity or residual initializations and, with high probability, for small random ones (Arora et al., 2018), and it makes layerwise gradient descent behave like a well-conditioned update of $A = W_{1:N}$.

Definition 2 (Deficiency margin)**.** *A matrix $W \in \mathbb{R}^{d_N \times d_0}$ has deficiency margin $c' > 0$ with respect to Φ if $\|W - \Phi\|_F \le \sigma_{\min}(\Phi) - c'$.*

Applied to $A(0)$, this says the initial end-to-end matrix is strictly closer to Φ than Φ is to the nearest rank-deficient matrix; the margin keeps the trajectory away from rank-deficient saddles and keeps $\sigma_{\min}(A(t))$ bounded below (Lemma 8). It presumes $\sigma_{\min}(\Phi) > 0$, and we assume throughout that no layer is a bottleneck, $\min_j d_j \ge \min(d_0, d_N)$, so that Φ is representable by the factorization. For $K < 0$ the margin takes the curvature-corrected form $S_K(R)(2\mathcal{E}(W(0)))^{1/2} \le \sigma_{\min}(\Phi) - c'$, with $S_K(R)$ the Rauch factor defined below. Both conditions are algebraic in the weight matrices, curvature entering them only through this factor.

Comparison-geometry notation. The model sine and cosine are

$$\mathrm{sn}_K(r) := \begin{cases} \sin(\sqrt{K}r)/\sqrt{K} & K > 0, \\ r & K = 0, \\ \sinh(\sqrt{\kappa}r)/\sqrt{\kappa} & K < 0, \end{cases} \tag{4}$$

and $\mathrm{cn}_K(r) := \mathrm{sn}_K'(r)$. The **Rauch factor** $S_K(R) := \mathrm{sn}_K(R)/R$ is the transverse stretching of $D\exp_o(v)$ for $\|v\| = R$, named after the Rauch comparison theorem, which controls how curvature makes neighboring geodesics spread or converge (Petersen, 2006). Then $S_K(R) > 1$ for $K < 0$, $S_K(R) = 1$ for $K = 0$, and $S_K(R) < 1$ for $K > 0$ within the injectivity radius; since the radial direction is unstretched, $\|D\exp_o(v)\|_{\mathrm{op}} = \widehat{S}_K(R) := \max(1, S_K(R))$. The **Hessian comparison function** for $\phi(p) = \frac{1}{2}d(p,q)^2$ is $f_K(D) := D\,\mathrm{cn}_K(D)/\mathrm{sn}_K(D)$, the eigenvalue in transverse directions, the radial one being 1 (Bridson and Haefliger, 2013); both ratios are extended by continuity, $S_K(0) = f_K(0) = 1$, and the **Hessian bound** is $H_K(D) := \max(1, f_K(D))$. Let $B_K(R) := \sup_{\|v\| \le R} \|D^2 \exp_o(v)\|_{\mathrm{op}}$. The **tube** $\mathcal{T}_{R,D}$ is the set of A with $\|A\xi_i\| \le R$ and $d(\exp_{o_N}(A\xi_i), y_i) \le D$ for all i, and the **smoothness constant** is

$$L_{K,R,D} := \tfrac{1}{m}\textstyle\sum_{i=1}^m \|\xi_i\|^2 \big[\widehat{S}_K(R)^2 H_K(D) + D\, B_K(R)\big]. \tag{5}$$

4 Main Results

4.1 Descent Lemma for the Intrinsic Loss

One gradient step decreases $\mathcal{F}$ by a definite amount provided the step size does not exceed a curvature-dependent ceiling; everything below follows by iterating this guarantee. Hypotheses (1)–(2) make A bounded and nondegenerate, (3) makes the factorized gradient a well-conditioned preconditioner, and (4) confines the update to the tube.

Theorem 3 (Descent lemma). *Let $K \le 0$ (Theorem 10 covers $K > 0$), let $A = W_{1:N} \in \mathcal{T}_{R,D}$ and $G := \nabla_A \mathcal{F}(A)$. Suppose: (1) $\|W_j\|_{\mathrm{op}} \le M$; (2) $\sigma_{\min}(A) > 0$; (3) the factors are sufficiently balanced: $\sum_j \|W^*_{j+1:N} G W^*_{1:j-1}\|^2_{\mathrm{HS}} \ge \frac{3}{4}\sigma_{\min}(A)^{2(N-1)/N}\|G\|^2_{\mathrm{HS}}$; (4) the segment $[A, A^+] \subset \mathcal{T}_{R,D}$. Then there exists $C_N > 0$ such that if*

$$\begin{aligned} &\eta C_N\big(M^{3N-4}G_{K,R,D} + L_{K,R,D}M^{4N-4}\big) \\ &\qquad \le \tfrac{1}{4}\sigma_{\min}(A)^{2(N-1)/N}, \end{aligned} \tag{6}$$

where $G_{K,R,D} := \sup_{B\in\mathcal{T}_{R,D}} \|\nabla_A \mathcal{F}(B)\|_{\mathrm{HS}}$, then gradient descent $W_j^+ = W_j - \eta\nabla_{W_j}\mathcal{L}_K$ gives

$$\mathcal{F}(A^+) - \mathcal{F}(A) \le -\tfrac{\eta}{2}\,\sigma_{\min}(A)^{2(N-1)/N}\|G\|^2_{\mathrm{HS}}. \tag{7}$$

Proof sketch. The argument has two stages. *Stage 1 (gradient flow).* Under 0-balanced factors and continuous time the end-to-end matrix evolves as $\dot{A} = -\mathcal{P}_A(G)$, where the factorization-induced preconditioner $\mathcal{P}_A(G) = \sum_j W_{j+1:N}W^*_{j+1:N}GW^*_{1:j-1}W_{1:j-1}$ is positive semidefinite with eigenvalues at least $\sigma_{\min}(A)^{2(N-1)/N}$ (Arora et al., 2018); the chain rule then gives $\frac{d}{d\tau}\mathcal{F} = -\langle G, \mathcal{P}_A(G)\rangle \le -\sigma_{\min}(A)^{2(N-1)/N}\|G\|^2$, with no Taylor remainder and no curvature dependence. *Stage 2 (discretization).* A finite step gives $A^+ - A = -\eta\mathcal{P}_A(G) + R_\eta$ with $\|R_\eta\|_{\mathrm{HS}} \le C_N\eta^2 M^{3N-4}G_{K,R,D}\|G\|_{\mathrm{HS}}$, and bounding the quadratic Taylor term needs $L_{K,R,D}$-smoothness of $\mathcal{F}$. This is where curvature enters: the Hessian of the pullback splits into a term pairing $\nabla^2\phi_i$ with $D\exp\cdot\Delta\xi_i$, bounded by $\widehat{S}_K(R)^2 H_K(D)$, and a term pairing $\nabla\phi_i$ with $D^2\exp$, bounded by $D\,B_K(R)$, which together give (5). Hypothesis (3) makes the first-order term coercive and (6) caps the collected error at a third of it, leaving (7).

Inequality (7) separates where curvature does and does not act. The guaranteed decrease is proportional to $\sigma_{\min}(A)^{2(N-1)/N}\|G\|^2_{\mathrm{HS}}$, a factor fixed by the depth-N factorization and identical to the Euclidean deep-linear case (Arora et al., 2018). Curvature constrains only the admissible step size, through $L_{K,R,D}$ in (6): a more negatively curved output manifold lowers the largest stable η, while the coefficient multiplying $\|G\|^2_{\mathrm{HS}}$ is unchanged.

Corollary 4 (Gradient flow: curvature-free descent). *For 0-balanced factors, a condition the flow preserves, the gradient flow $\dot{W}_j = -\nabla_{W_j}\mathcal{L}_K$ satisfies*

$$\frac{d}{d\tau}\mathcal{F}(A(\tau)) \le -\sigma_{\min}(A)^{2(N-1)/N}\|G\|^2_{\mathrm{HS}} \tag{8}$$

with no curvature-dependent constraint.

Remark 5 (Curvature as a discretization cost). *$L_{K,R,D}$ enters only in Stage 2. In continuous time there is no Taylor remainder, so no curvature constraint. The step-size penalty $\eta^\star_K = O(e^{-2\sqrt{\kappa}R})$ is a discrete-time phenomenon.*

The step-size constraint (6) is satisfied whenever

$$\eta \le \eta^\star_K := \frac{\sigma_{\min}(A)^{2(N-1)/N}}{4C_N(M^{3N-4}G_{K,R,D} + L_{K,R,D}M^{4N-4})}. \tag{9}$$

Since $S_K(R)$ increases with $|K|$ for $K \le 0$, $\eta^\star_K$ decreases with $|K|$ and at $K = 0$ recovers the Euclidean bound; for $\sqrt{\kappa}R \gg 1$, $\eta^\star_K = O(e^{-2\sqrt{\kappa}R})$. As $S_K(R)$ depends on K and R only through $\sqrt{\kappa}R$, the penalty is set by the trajectory radius in units of the curvature scale $1/\sqrt{\kappa}$, not by the curvature alone.

4.2 Linear Convergence

An end-to-end rate requires converting the gradient norm into the gap and showing that the hypotheses of Theorem 3 persist along the trajectory. The second geometric constant enters here: $B_K(R)$ measures the second-order deviation of exp from a linear map, and the corresponding Hessian term can be negative, so $\mu_K = 1 - D\,B_K(R)$ for $K \le 0$. Thus S_K constrains the step size and B_K the rate.

Lemma 6 (Local strong convexity of the pullback)**.** *Assume the data are whitened, meaning* $\frac{1}{m}\sum \xi_i \xi_i^* = I$*, and that the tube is narrow enough that* $\mu_K := \min(1, S_K(R)^2 f_K(D)) - D\,B_K(R) > 0$*. Then* $D^2\mathcal{F}(A)[\Delta,\Delta] \ge \mu_K \|\Delta\|_{\mathrm{HS}}^2$ *on the tube, so* $\mathcal{F}$ *is* μ_K*-strongly convex along any segment contained in it. For* $K \le 0$ *the minimum equals* 1*, giving* $\mu_K = 1 - D\,B_K(R)$*, and* $\mu_0 = 1$*.*

Proof sketch. The Hessian splits as in Theorem 3, and each term is now bounded from below. The radial eigenvalue of $\nabla^2\phi_i$ and the radial singular value of $D\exp$ (Gauss lemma) are both 1, and the transverse ones are $f_K(D)$ and $S_K(R)$, both ≥ 1 for $K \le 0$ and both ≤ 1 for $K > 0$; together they bound the first term below by $\min(1, S_K(R)^2 f_K(D))\|\Delta\xi_i\|^2$. The second term can be negative, and Cauchy–Schwarz with $\|\nabla\phi_i\| \le D$ and $\|D^2 \exp\|_{\mathrm{op}} \le B_K(R)$ bounds it below by $-D\,B_K(R)\|\Delta\xi_i\|^2$. Whitening converts $\frac{1}{m}\sum_i \|\Delta\xi_i\|^2$ into $\|\Delta\|_{\mathrm{HS}}^2$.

Lemma 7 (Gradient dominance)**.** *Under the hypotheses of Lemma 6, let* $A^\star$ *minimize* $\mathcal{F}$ *on the tube. If* $[A, A^\star] \subset \mathcal{T}_{R,D}$ *then* $\|\nabla_A \mathcal{F}(A)\|^2 \ge 2\mu_K(\mathcal{F}(A) - \mathcal{F}^\star)$*.*

Proof sketch. Integrating the Hessian lower bound of Lemma 6 along the segment joining two points of the tube gives the strong-convexity inequality. Minimizing its right-hand side over the displacement and evaluating at a minimizer yields the Polyak–Łojasiewicz bound (Nesterov et al., 2018).

Lemma 8 (Trajectory containment)**.** *Fix* Φ *and* $c' > 0$*, and let the data be whitened. There exist* $\delta_0, \eta_0 > 0$ *such that if the initialization is* δ_0*-balanced with deficiency margin* c' *and* $\eta \le \min(\eta_K^\star, \eta_0)$*, then for all* $t \ge 0$*: (a)* $A(t) \in \mathcal{T}_{R,D(0)}$ *with* $R := \max_i \max(2\|\Phi\|_F\|\xi_i\|, \|\tilde{y}_i\|)$*, which also contains the targets, and* $D(0) := (2m\,\mathcal{F}(A(0)))^{1/2}$*; (b) the factors remain* $2\delta_0$*-balanced, so hypothesis (3) of Theorem 3 holds; (c)* $\sigma_{\min}(A(t)) \ge c'$*.*

Proof sketch. Induction on t, applying Theorem 3 at each step. Part (b) is algebraic and reproduces the balancedness-preservation argument of Arora et al. (2018): the change in the balancedness defect over one step is $O(\eta^2\|G\|^2)$ and telescopes to at most δ_0 once the summed decrease of $\mathcal{F}$ is used, which is small enough to leave the coercivity hypothesis (3) intact. For (c), $\exp_{o_N}$ is expanding and $S_K(R)$-Lipschitz for $K \le 0$, so monotonicity of $\mathcal{F}$ gives $\|A(t) - \Phi\|_F \le (2\mathcal{F}(A(0)))^{1/2} \le S_K(R)(2\mathcal{E}(W(0)))^{1/2}$, and the corrected margin gives $\sigma_{\min}(A(t)) \ge c'$. For (a), $\max_i d(\exp(A(t)\xi_i), y_i)^2 \le 2m\mathcal{F}(A(t))$, so the same monotonicity keeps $A(t)$ in the tube.

Theorem 9 (Linear convergence)**.** *Let the curvature be nonpositive,* $K \le 0$*, and suppose the data are whitened. Assume the initialization is* δ*-balanced with deficiency margin* $c' > 0$*, and that it starts inside a tube narrow enough that* $A(0) \in \mathcal{T}_{R,D(0)}$ *with* $D(0)\,B_K(R) < 1$*, for* $D(0)$ *as in Lemma 8. Finally, let the step size respect the curvature-dependent ceiling,* $\eta \le \eta_K^\star$ *evaluated at* $\sigma_{\min}(A) = c'$ *and* $D = D(0)$*, with* $[A(t), A^\star] \subset \mathcal{T}_{R,D(0)}$ *as in Lemma 7. Then*

$$\mathcal{F}(A(T)) - \mathcal{F}^\star \le \big(1 - \eta\mu_K (c')^{2(N-1)/N}\big)^T \times (\mathcal{F}(A(0)) - \mathcal{F}^\star). \tag{10}$$

The iteration count to reach $\mathcal{F} - \mathcal{F}^\star \le \varepsilon$ is

$$T = O\Bigg(\frac{M^{3N-4} G_{K,R,D(0)} + M^{4N-4} L_{K,R,D(0)}}{\mu_K \, (c')^{4(N-1)/N}} \times \log \frac{\mathcal{F}(A(0)) - \mathcal{F}^\star}{\varepsilon} \Bigg). \tag{11}$$

Proof sketch. Lemma 8 certifies that the hypotheses of Theorem 3 and Lemma 6 hold at every iteration. Substituting $\|G(t)\|_{\mathrm{HS}}^2 \ge 2\mu_K(\mathcal{F}(A(t)) - \mathcal{F}^\star)$ and $\sigma_{\min}(A(t)) \ge c'$ into (7) turns the single-step decrease into the contraction (10), which iterates; solving $(1-\rho)^T \Delta\mathcal{F}(0) \le \varepsilon$ with $\log(1-\rho) \le -\rho$ and inserting $\eta = \eta_K^\star$ gives (11), whose leading curvature factor is the $S_K(R)^2$ inside $L_{K,R,D(0)}$, the remaining one being $H_K(D(0))$.

At $K = 0$ we have $\mu_0 = 1$ and (10) recovers the guarantee of Arora et al. (2018). On a curved manifold two factors appear: $\eta_K^\star = O(1/S_K(R)^2)$ shrinks with $|K|$ for negative curvature, and $\mu_K = 1 - D\,B_K(R)$ also decreases, weakening gradient dominance. The mechanism itself, the preconditioner induced by depth, is manifold-agnostic.

4.3 Extension to General Constant Curvature

The analysis above extends to arbitrary constant sectional curvature. The proof of Theorem 3 uses four comparison-geometry bounds: $\|D \exp\|_{\mathrm{op}} = \widehat{S}_K(R)$, $\|D^2 \exp\|_{\mathrm{op}} \le B_K(R)$, $\|\nabla^2 \phi_i\|_{\mathrm{op}} \le H_K(D)$, and $\|\nabla \phi_i\| \le D$, which hold globally for $K \le 0$ and within the injectivity radius for $K > 0$.

Theorem 10 (General curvature)**.** *Theorem 3 holds for constant sectional curvature $K \in \mathbb{R}$ with the constants $\widehat{S}_K(R)$, $H_K(D)$, $B_K(R)$ from Section 3, provided $R < \pi/(2\sqrt{K})$ and $D < \pi/(2\sqrt{K})$ when $K > 0$.*

Proof sketch. The four bounds enter the proof of Theorem 3 only through $L_{K,R,D}$ and $G_{K,R,D}$, and the Jacobi-field comparison theorems that yield them hold on any constant-curvature space within the injectivity radius. For $K > 0$ the model sine satisfies $\mathrm{sn}_K(r) < r$ on $(0, \pi/\sqrt{K})$, so $S_K(R) < 1$, while $B_K(R)$ and $H_K(D)$ stay finite under the stated radius constraints, which also keep $\log_{p_i}(y_i)$ single-valued. Everything else is algebraic in the weight matrices, so only the numerical values of the geometric constants change with the sign of K.

Corollary 11 (Positive curvature carries no Rauch-factor penalty)**.** *Let the curvature be positive, $K > 0$, and assume the radius and the residual stay within the injectivity radius, so that $\sqrt{K}R < \pi/2$ and $\sqrt{K}D < \pi/2$. Then: (1) $S_K(R) = \sin(\sqrt{K}R)/(\sqrt{K}R) < 1$; (2) $H_K(D) = 1$; (3) $\widehat{S}_K(R) = 1$, so $L_{K,R,D} \to L_{0,R,D}$ as $D \to 0$; (4) $\eta_K^\star \to \eta_0^\star$.*

Proof sketch. With $\theta = \sqrt{K}R \in (0, \pi/2)$, $\sin\theta < \theta$ gives (1); with $\alpha = \sqrt{K}D \in (0, \pi/2)$, $\tan\alpha > \alpha$ gives $f_K(D) = \alpha \cot\alpha \in (0,1)$ and hence (2). The contraction $S_K(R) < 1$ is felt only transversally, the radial direction being unstretched, so $\widehat{S}_K(R) = 1$ and (5) tends to its Euclidean value as $D \to 0$, giving (3) and hence (4).

4.4 The Tangent-Space Surrogate as a Control

Whether the cost above is inherent or an artifact of measuring error on the manifold is settled by training the surrogate on the same network, weights and data.

Proposition 12 (Curvature-free rate)**.** *Under δ-balancedness, deficiency margin $c' > 0$, and a step size $\eta \le \eta_0^{\mathcal{E}}(c', N, \|\Phi\|_F)$, gradient descent on $\mathcal{E}$ (3) converges linearly: $\mathcal{E}(W(T)) - \mathcal{E}^\star \le \varepsilon$ for $T \ge (\eta(c')^{2(N-1)/N})^{-1} \log((\mathcal{E}(W(0)) - \mathcal{E}^\star)/\varepsilon)$, where $\mathcal{E}^\star := \min_W \mathcal{E}(W)$. The step size, contraction factor, and iteration count are all independent of the curvature K.*

Proof sketch. Once the logarithmic maps have produced ξ_i and $\tilde{y}_i$, the surrogate is Euclidean: for whitened data $\mathcal{E}(W) = \frac{1}{2}\|A - \Phi\|_{\mathrm{HS}}^2 + \mathcal{E}^\star$, so $\nabla_A \mathcal{E} = A - \Phi$ contains no $D\exp$, no $\log_p$, and no Rauch factor. The pullback is 1-smooth and 1-strongly convex, and the objective is exactly the one analyzed by Arora et al. (2018), whose theorem applies verbatim; every constant in it depends only on c', N, and $\|\Phi\|_F$, all fixed once the tangent-space data are.

The curvature dependence of Theorem 9 therefore originates from measuring error on the manifold, not from the architecture or the data distribution. At $K = 0$ the two losses coincide ($\mathcal{L}_0 = \mathcal{E}$), and so do the two guarantees; for $K \neq 0$ the difference between them is the geometric cost quantified by $S_K(R)^2$ and μ_K.

4.5 Landscape of the Intrinsic Loss

Theorem 9 gives convergence to a global minimum but does not rule out spurious critical points elsewhere on the tube. The Euclidean landscape result of Laurent and Brecht (2018) has a local counterpart here.

Theorem 13 (No spurious critical points on the tube)**.** *Under the hypotheses of Lemma 6 (constant curvature K, whitened data, $\mu_K > 0$), every critical point A of $\mathcal{F}$ on the tube with $[A, A^\star] \subset \mathcal{T}_{R,D}$ minimizes $\mathcal{F}$ on the tube: $\nabla_A \mathcal{F}(A) = 0 \Rightarrow \mathcal{F}(A) = \mathcal{F}^\star$. For $K \le 0$ this holds on any tube with $D\,B_K(R) < 1$. For $K > 0$ it holds within the injectivity radius.*

Proof sketch. Apply Lemma 7 at a critical point: $0 = \|\nabla_A \mathcal{F}(A)\|^2 \ge 2\mu_K(\mathcal{F}(A) - \mathcal{F}^\star)$, and $\mu_K > 0$ forces $\mathcal{F}(A) = \mathcal{F}^\star$. On a Hadamard manifold the required input, $\nabla_p^2 \frac{1}{2} d^2 \succeq I$, holds globally. On a sphere the two comparison factors drop below 1 rather than failing, so μ_K takes its general form; beyond the cut locus $f_K(D)$ turns negative and ϕ_i ceases to be smooth, which is why the injectivity restriction appears.

Remark 14. *Curvature changes the convergence rate but not this conclusion, affecting only the size of the admissible tube and the value of μ_K. The statement is local, and concerns $\mathcal{F}$ rather than the factorized loss.*

4.6 Curvature-Adaptive Step-Size Schedule

The ceiling $\eta_K^\star$ uses the worst-case residual $D(0)$, so it pays the full penalty $H_K(D(0))$ throughout training; tying the step size to the current residual removes it.

Proposition 15 (Adaptive schedule)**.** *Under the hypotheses of Theorem 9, define $D(t) := \max_i d(\exp(A(t)\xi_i), y_i)$, set $\eta(t)$ via (9) with $D = D(t)$, and halve it until the step satisfies (7). Then: (1) the backtracking terminates and every accepted step obeys (7); (2) late in training, when $D(t) \ll 1/\sqrt{\kappa}$, $\eta(t) \ge c_0/S_K(R)^2$ and the contraction factor approaches $(1 - c_1/S_K(R)^2)$, independent of H_K; and (3) the resulting late-phase rate gives $T_{\text{adaptive}} \lesssim T_{\text{fixed}}/H_K(D(0))$.*

Proof sketch. Backtracking terminates because $\mathcal{F}$ is smooth near $A(t)$, and the accepted step lies in $\mathcal{T}_{R,D(t)}$, so the smoothness bound may be evaluated at $D(t)$ rather than $D(0)$, giving (1). As $D \to 0$, $H_K(D) = 1 + O(\kappa D^2) \to 1$ and $D\,B_K(R) \to 0$, so $L_{K,R,D(t)} \to S_K(R)^2 \frac{1}{m}\sum_i \|\xi_i\|^2$ and $\mu_{K,t} \to 1$, giving (2); the ratio of the two iteration counts then leaves $[S_K(R)^2 H_K(D(0)) + D(0)B_K(R)]/(\mu_K S_K(R)^2) \approx H_K(D(0))$, which is (3).

Remark 16. *For $K < 0$, $H_K(D) \gg 1$ when the initial residual is large, so the fixed schedule is overly conservative; as $D(t) \to 0$ the rate approaches the Euclidean one up to $S_K(R)^2$, an intrinsic cost no schedule can remove (Section 4.8).*

4.7 Implicit Bias under Curvature

When a model is overparameterized the loss admits many global minima, and the one gradient descent selects is fixed by the algorithm and initialization, an effect known as *implicit bias*. For shallow Euclidean matrix factorization, Gunasekar et al. (2017) and Li et al. (2018) point to the minimum-nuclear-norm solution; for deeper factorizations Arora et al. (2019) and Razin and Cohen (2020) argue that no norm captures the bias. The proposition below describes how curvature *modifies* that selection.

Proposition 17 (Curvature-dependent gradient anisotropy)**.** *Consider gradient flow on $\mathcal{L}_K$ with 0-balanced initialization $W_j(0) = \alpha^{1/N} U_j$ ($\alpha \to 0^+$), and assume the targets lie inside the curvature scale, $\sqrt{\kappa}\max_i \|\tilde{y}_i\| \ll 1$. Then:*

1. ***Early phase*** *($\|A\| \ll 1$): $\nabla^2\phi_i|_o = I + O(\kappa\|\tilde{y}_i\|^2)$, so $\nabla_A \mathcal{F} = \nabla_A \mathcal{E} + O(\kappa \max_i \|\tilde{y}_i\|^2\|A\| + \|A\|^2)$: the initial implicit bias is curvature-free to leading order and matches the Euclidean one, whatever that bias is taken to be.*
2. ***Late phase*** *($\|A\| = \Theta(1)$): the intrinsic gradient $G_K = -\frac{1}{m}\sum [D\exp(A\xi_i)]^* \log_{p_i}(y_i)\xi_i^*$ weights SVD direction k by $S_K(\sigma_k\|\xi_i\|)$ when ξ_i aligns with the k-th right singular vector, so for $K < 0$ transverse residual components are amplified and the flow reduces transverse error faster than radial error.*
3. ***Fixed-point condition****: at a limit point $A^\star$ of the flow, $G_K(A^\star) \in \ker(\mathcal{P}_{A^\star})$, which is informative whenever the residuals do not vanish. Since G_K depends on K through the Rauch-weighted adjoint, curvature can change the selected limit point; at $K = 0$ this reduces to Euclidean stationarity.*

Proof sketch. For (1), $\nabla^2\phi_i|_o$ has radial eigenvalue 1 and transverse eigenvalue $f_K(\|\tilde{y}_i\|) = 1 + O(\kappa\|\tilde{y}_i\|^2)$, so whitened data give $G_K = (A - \Phi) + O(\kappa \max_i \|\tilde{y}_i\|^2\|A\| + \|A\|^2)$, and as $\alpha \to 0^+$ the trajectory spends arbitrarily long in this neighborhood. For (2), $[D\exp_o(v_i)]^*$ scales the radial component of the residual by 1 and each transverse component by $S_K(\|v_i\|)$, and $\|v_i\| \approx \sigma_k\|\xi_i\|$ when ξ_i aligns with the k-th right singular vector of A. For (3), a fixed point of $\dot{A} = -\mathcal{P}_A(G_K)$ satisfies $\mathcal{P}_{A^\star}(G_K(A^\star)) = 0$, and $\ker(\mathcal{P}_{A^\star})$ is the orthogonal complement of the column and row spaces of $A^\star$.

4.8 Sharpness of the Step-Size Bound

Theorem 9 guarantees convergence whenever $\eta \le \eta_K^\star = O(1/S_K(R)^2)$. This ceiling cannot be improved.

Proposition 18 (Sharpness)**.** *For every $K < 0$ and $R > 0$ there is a configuration with $d_N \ge 2$, $d_0 = 1$, $N = 1$, $m = 1$ in which A_0 is a zero-loss minimizer and: (1) $D^2\mathcal{F}(A_0)[\Delta, \Delta] = L_{K,R,0}\|\Delta\|_{\mathrm{HS}}^2 = S_K(R)^2\|\Delta\|_{\mathrm{HS}}^2$ for transverse Δ; (2) gradient descent at A_0 is linearly unstable for every $\eta > 2/S_K(R)^2$; (3) hence $\eta_{\mathrm{crit}} = 2/S_K(R)^2 = \Theta(1/S_K(R)^2)$.*

Proof sketch. Take one data point with $\xi_1 = 1$ and target $y = \exp_{o_N}(Re_1)$, so $a_0 = Re_1$ has zero loss and $\nabla\mathcal{F}(a_0) = 0$. The residual vanishes, so the $D^2\exp$ term drops out and $\nabla^2\phi_1|_y = I$, leaving $D^2\mathcal{F}(a_0)[\Delta, \Delta] = \|D\exp_{o_N}(Re_1)\Delta\|^2$: this is $\|\Delta\|^2$ radially and $S_K(R)^2\|\Delta\|^2$ transversally, so the Rauch bound is attained and $L_{K,R,0} = S_K(R)^2$ for $K < 0$, giving (1). At a stationary point gradient descent linearizes to $\Delta \mapsto (I - \eta D^2\mathcal{F}(a_0))\Delta$, whose transverse multiplier $1 - \eta S_K(R)^2$ has modulus above 1 exactly when $\eta > 2/S_K(R)^2$, giving (2) and (3).

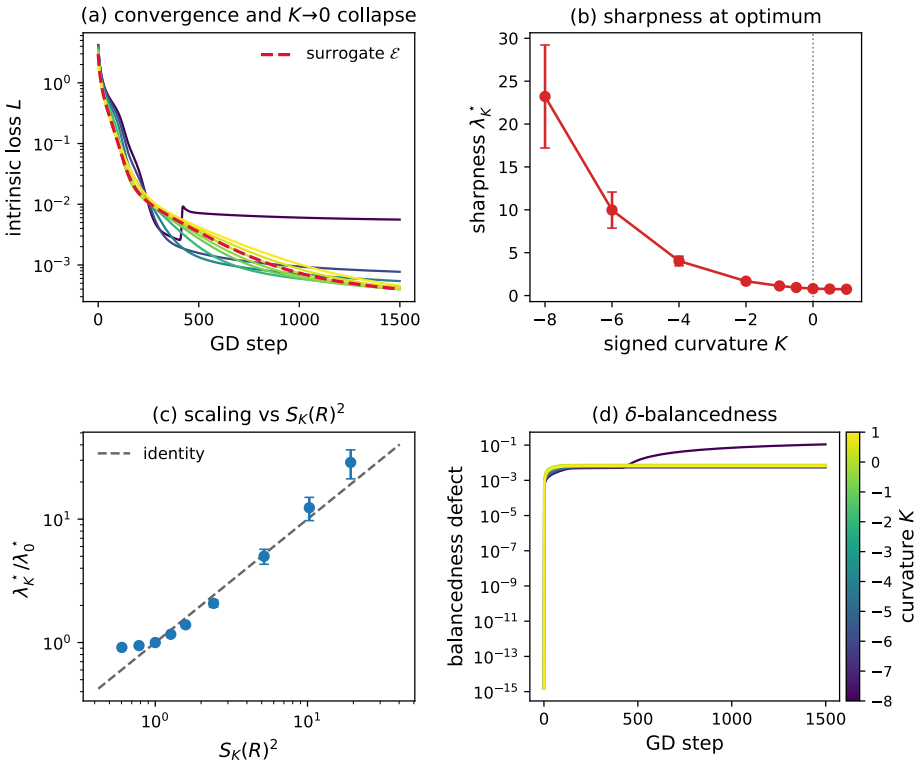


Figure 2: Optimization of the deep-linear network against the *signed* sectional curvature K at fixed radius $R = 1.2$. **(a)** Intrinsic-loss trajectories at a fixed step size, with the Euclidean tangent surrogate dashed. **(b)** Near-optimum sharpness $\lambda_K^\star$, the top eigenvalue of $D^2\mathcal{F}$ in the end-to-end matrix A at the converged solution, whose reciprocal $2/\lambda_K^\star$ is the local stability ceiling for gradient descent in A; points are means over the 50 seeds and error bars are 95% confidence intervals. **(c)** Measured ratio $\lambda_K^\star/\lambda_0^\star$ against the predicted $S_K(R)^2$ (identity dashed). **(d)** δ-balancedness defect $\max_j \|W_{j+1}^\top W_{j+1} - W_j W_j^\top\|_F$ along training, which stays $\lesssim 10^{-1}$ at every curvature.

5 Experiments

Setup. We validate the theory on the deep-linear network of Section 3 with every layer on the same constant-curvature model, using the κ-stereographic chart of Section 3 so that K can be swept continuously through $K = 0$. The experiments use (1), (2) and (3), with gradients taken with respect to the individual factors W_j. Unless stated otherwise we use depth $N = 3$, width $d = 6$, $m = 300$ samples, radius $R = 1.2$, an exactly balanced initialization ($\delta = 0$, each W_j a scaled orthogonal matrix), and 50 seeds. One tangent-space regression problem is generated once and reused at every curvature via $x^{(i)} = \exp_o(\tilde{x}^{(i)})$, $y^{(i)} = \exp_o(\tilde{y}^{(i)})$. The inputs are isotropic with $\|\tilde{x}^{(i)}\| \le R$, so their population covariance is a multiple of the identity and the sample covariance is close to it, a scale that cancels in $\lambda_K^\star/\lambda_0^\star$. All quantities are evaluated in double precision; we run full-batch gradient descent at a fixed step ($\eta = 0.02$) until convergence and obtain $\lambda_K^\star$ by power iteration (80 steps) on Hessian–vector products in A at the converged solution.

What the theory predicts. Because $L_{K,R,D}$ is a worst-case bound, the largest stable step from a high-residual start is dominated by the large-D region and is nearly curvature-flat, so it is not an informative probe. The informative quantity is the late-phase sharpness: as $D(t) \to 0$, Proposition 15 collapses the transverse part of $L_{K,R,D(t)}$ to $S_K(R)^2 \frac{1}{m}\sum_i \|\xi_i\|^2$, so the top eigenvalue $\lambda_K^\star$ of $D^2\mathcal{F}$

at the converged solution should scale as $S_K(R)^2$ and set the local ceiling $\eta^\star = 2/\lambda_K^\star$.

Results. Figure 2 reports four measurements against signed curvature $K \in \{1, \frac{1}{2}, 0, -\frac{1}{2}, -1, -2, -4, -6, -8\}$ at $R = 1.2$. **(a)** At a fixed step size every curvature converges linearly, and at $K = 0$ the intrinsic loss coincides with the Euclidean surrogate to machine precision, the collapse $\mathcal{L}_0 = \mathcal{E}$ of Proposition 12. **(b)** The near-optimum sharpness $\lambda_K^\star$ is monotone in K, rising from $\lambda_0^\star$ by a factor of $\sim$29 at $K = -8$, so the stable step $2/\lambda^\star$ tightens by that factor near the optimum, and falling for spherical $K > 0$, where exp contracts transverse directions. **(c)** Plotting $\lambda_K^\star/\lambda_0^\star$ against $S_K(R)^2$, the points closely follow the identity line, confirming the $S_K(R)^2$ scaling of the late-phase smoothness constant, with mild upward deviation at strong curvature from the H_K, B_K terms of (5). **(d)** The δ-balancedness defect stays $\lesssim 10^{-1}$ at every curvature, so the factors remain approximately balanced along the trajectory.

Robustness and further experiments. Repeating the measurement across 4 radii $R \in \{0.8, 1.0, 1.2, 1.5\}$, 3 architectures ($(N, d) \in \{(2, 4), (3, 6), (4, 10)\}$), 50 seeds and the 7 curvatures $K \in \{\frac{1}{2}, 0, -\frac{1}{2}, -1, -2, -4, -6\}$ ($4 \cdot 3 \cdot 50 \cdot 7 = 4{,}200$ configurations) gives the same picture: 22 of the 28 radius–curvature pairs satisfy $\sqrt{|K|}\, R \leq 2$, where the comparison-geometry constants are well behaved, and 94% of those 3,300 configurations fall within 1.5× of the predicted $S_K(R)^2$ (geometric mean 0.95), independently of depth and width, with agreement weakening toward the injectivity boundary (Figure S3). Three further experiments appear in the supplement: a fixed-step collapse check at $K = 0$ (Figure S1); the surrogate control of Proposition 12, whose normalized curves coincide across all K (Figure S2); and 50 balanced in-tube initializations at each of $K \in \{0, -1, -4\}$, all reaching the known global minimum numerically, consistent with Theorem 13 (Figure S4).

Scope and limitations. These experiments validate the sign and shape of the theory rather than its constants, and $\lambda_K^\star$ tracks $S_K(R)^2$ only up to the higher-order H_K, B_K terms. The setting is deep-linear and bias-free on one constant-curvature manifold; nonlinear activations and Möbius biases fall outside our scope.

6 Conclusion

For deep linear networks trained with an intrinsic loss on a constant-curvature manifold, curvature affects discrete-time optimization through the smoothness constant $L_{K,R,D}$. The differential descent inequality carries no curvature-dependent restriction, while gradient descent requires a stable step size $\eta_K^\star = O(1/S_K(R)^2)$, yielding linear convergence with iteration count $O(L_{K,R,D}/\mu_K)$. Negative curvature reduces the admissible step size, whereas positive curvature carries no Rauch-factor penalty. We also establish a benign local landscape, a curvature-adaptive schedule, a curvature-dependent anisotropy in the implicit bias, and tightness of the step-size scaling, while the tangent-space surrogate remains curvature-independent. Thus, target-space curvature governs the stable learning rate and convergence complexity.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

El Mehdi Achour, Kathlén Kohn, and Holger Rauhut. The riemannian geometry associated to gradient flows of linear convolutional networks. *arXiv preprint arXiv:2507.06367*, 2025.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4453–4462, 2022.

Gregor Bachmann, Gary Bécigneul, and Octavian Ganea. Constant curvature graph convolutional networks. In *International conference on machine learning*, pages 486–496. PMLR, 2020.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022.

Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *International Conference on Learning Representations*, volume 2024, pages 47687–47711, 2024.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods, 2019. URL https://arxiv.org/abs/1810.00760.

Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fully hyperbolic neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5672–5686, 2022.

Sungjun Cho, Seunghyuk Cho, Sungwoo Park, Hankook Lee, Honglak Lee, and Moontae Lee. Curve your attention: Mixed-curvature transformers for graph representation learning. *arXiv preprint arXiv:2309.04082*, 2023.

Christopher Criscitiello and Nicolas Boumal. Negative curvature obstructs acceleration for strongly geodesically convex optimization, even with exact first-order oracles. In *Conference on Learning Theory*, pages 496–542. PMLR, 2022.

Christopher Criscitiello and Nicolas Boumal. Curvature and complexity: Better lower bounds for geodesically convex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2969–3013. PMLR, 2023.

Jindou Dai, Yuwei Wu, Zhi Gao, and Yunde Jia. A hyperbolic-to-hyperbolic graph convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 154–163, 2021.

Simon Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pages 1655–1664. PMLR, 2019.

Xingcheng Fu, Jianxin Li, Jia Wu, Qingyun Sun, Cheng Ji, Senzhang Wang, Jiajun Tan, Hao Peng, and Philip S Yu. Ace-hgnn: Adaptive curvature exploration hyperbolic graph neural network. In *2021 IEEE international conference on data mining (ICDM)*, pages 111–120. IEEE, 2021.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *arXiv preprint arXiv:1805.09112*, 2018.

Karish Grover, Haiyang Yu, Qi Zhu, Han Xie, Vassilis Ioannidis, Christos Faloutsos, et al. Spectro-riemannian graph neural networks. In *International Conference on Learning Representations*, volume 2025, pages 57833–57859, 2025.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International conference on learning representations*, 2018.

Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, et al. Hyperbolic attention networks. *arXiv preprint arXiv:1805.09786*, 2018.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

Jiang Hu, Xin Liu, Zai-Wen Wen, and Ya-Xiang Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6418–6428, 2020.

Thomas Laurent and James Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *International conference on machine learning*, pages 2902–2907. PMLR, 2018.

Marc Law. Ultrahyperbolic neural networks. *Advances in Neural Information Processing Systems*, 34:22058–22069, 2021.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference on learning theory*, pages 2–47. PMLR, 2018.

Yuekang Li, Yidan Mao, Yifei Yang, and Dongmian Zou. Improving robustness of hyperbolic neural networks by lipschitz analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1713–1724, 2024.

Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

Emile Mathieu, Charline Le Lan, Chris J Maddison, Ryota Tomioka, and Yee Whye Teh. Continuous hierarchical representations with poincaré variational auto-encoders. *Advances in neural information processing systems*, 32, 2019.

Govind Menon. The geometry of the deep linear. In *XIV Symposium on Probability and Stochastic Processes: CIMAT, Mexico, November 20-24, 2023*, page 1. Springer Nature, 2025.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pages 3779–3788. PMLR, 2018.

Maher Nouiehed and Meisam Razaviyayn. Learning deep models: Critical points and local openness. *INFORMS Journal on Optimization*, 4(2):148–173, 2022.

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

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pages 4460–4469. PMLR, 2018.

Krisanu Sarkar. Learning beyond euclid: Curvature-adaptive generalization for neural networks on manifolds. *arXiv preprint arXiv:2507.02999*, 2025.

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.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Ohad Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In *Conference on Learning Theory*, pages 2691–2713. PMLR, 2019.

Ondrej Skopek, Octavian-Eugen Ganea, and Gary Bécigneul. Mixed-curvature variational autoencoders. *arXiv preprint arXiv:1911.08411*, 2019.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018.

Max Van Spengler, Erwin Berkhout, and Pascal Mettes. Poincare resnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5419–5428, 2023.

Shen Wang, Xiaokai Wei, Cicero Nogueira Nogueira dos Santos, Zhiguo Wang, Ramesh Nallapati, Andrew Arnold, Bing Xiang, Philip S Yu, and Isabel F Cruz. Mixed-curvature multi-relational graph neural network for knowledge graph completion. In *Proceedings of the web conference 2021*, pages 1761–1771, 2021.

Bo Xiong, Shichao Zhu, Nico Potyka, Shirui Pan, Chuan Zhou, and Steffen Staab. Pseudo-riemannian graph convolutional networks. *Advances in neural information processing systems*, 35:3488–3501, 2022.

Menglin Yang, Min Zhou, Tong Zhang, Jiahong Liu, Zhihao Li, Lujia Pan, Hui Xiong, and Irwin King. Hyperbolic graph neural networks: A review of methods and applications. *arXiv preprint arXiv:2202.13852*, 2022.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on learning theory*, pages 1617–1638. PMLR, 2016.

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

Yiding Zhang, Xiao Wang, Chuan Shi, Nian Liu, and Guojie Song. Lorentzian graph convolutional networks. In *Proceedings of the web conference 2021*, pages 1249–1261, 2021.

Difan Zou, Philip M Long, and Quanquan Gu. On the global convergence of training deep linear resnets. *arXiv preprint arXiv:2003.01094*, 2020.

Supplementary Material
The Influence of Curvature on Deep Linear Networks on Riemannian Manifolds

This document accompanies the paper *The Influence of Curvature on Deep Linear Networks on Riemannian Manifolds*. It is self-contained: Appendix A restates the setting, the notation, and every formal statement of the main paper, using the same numbering, so that no result below has to be looked up elsewhere. Appendix B gives an extended account of related work, Appendix C the coordinate models used for the constant-curvature target manifold, Appendix D the complete proofs, and Appendix E the full experimental protocol together with three further experiments.

A Setting, Notation, and Statements

A.1 The Layerwise Riemannian Network

Let $N \geq 1$. For $j = 0, \dots, N$, let $(\mathcal{M}_j, g_j)$ be a complete Riemannian manifold of dimension d_j with base point $o_j \in \mathcal{M}_j$, and let $V_j := T_{o_j}\mathcal{M}_j \simeq \mathbb{R}^{d_j}$. The identification $V_j \simeq \mathbb{R}^{d_j}$ is fixed to be a g_{o_j}-orthonormal frame, so that the Euclidean norm on $\mathbb{R}^{d_j}$ used in the surrogate (3) *is* the base-point metric norm $\|\cdot\|_{g_{o_j}}$; equivalently, the coordinates are rescaled so that the conformal factor is normalized to $\lambda_{o_j} = 1$ at the base point. With this convention $\exp_{o_j}$ is a radial isometry in the norm used throughout (its radial singular value is 1, as invoked in Appendices D.1, D.2, and D.6), and at $K = 0$ the space form is flat with $\exp_{o_N}$ an isometry, so $d_0(\exp_{o_N}(A\xi_i), \exp_{o_N}(\tilde{y}_i)) = \|A\xi_i - \tilde{y}_i\|$ and the two losses coincide, $\mathcal{L}_0 = \mathcal{E}$. The closed-form ball and stereographic formulas of Appendix C are written in the ambient chart, whose conformal factor $\lambda_x = 2/(1 \pm \kappa\|x\|^2)$ satisfies $\lambda_o = 2$; they relate to the orthonormal tangent coordinates above by the constant factor λ_o at the base point, which is absorbed once and for all by this normalization. Data are $(x_i, y_i)_{i=1}^m$ with $x_i \in \mathcal{M}_0$, $y_i \in \mathcal{M}_N$. Layer j applies $\Phi_j(h; \theta_j) := \exp_{o_j}(F_j(\log_{o_{j-1}}(h); \theta_j))$ with $F_j : V_{j-1} \times \Theta_j \to V_j$ a tangent-space map. When every F_j is linear ($F_j(u) = W_j u$, no bias or activation), adjacent $\log_o$ and $\exp_o$ factors cancel and the network collapses to

$$h_N = \exp_{o_N}(W_N \cdots W_1 \log_{o_0}(x)). \tag{1}$$

We write $W_{a:b} := W_b \cdots W_a$ and $A := W_{1:N}$. Every target manifold is a simply connected space form of constant sectional curvature $K \in \mathbb{R}$, hyperbolic for $K < 0$, Euclidean for $K = 0$, and spherical for $K > 0$; we set $\kappa := |K|$. The coordinate models are described in Appendix C.

A.2 The Two Losses

Write $\xi_i := \log_{\mathbf{0}}^K(x_i)$ and $\tilde{y}_i := \log_{\mathbf{0}}^K(y_i)$. The *intrinsic loss* and the *tangent-space surrogate* are

$$\mathcal{L}_K(W) := \tfrac{1}{2m}\sum_{i=1}^m d_K\big(\exp_{\mathbf{0}}^K(A\xi_i), y_i\big)^2, \tag{2}$$
$$\mathcal{E}(W) := \tfrac{1}{2m}\sum_{i=1}^m \|A\xi_i - \tilde{y}_i\|^2, \tag{3}$$

where d_K is the geodesic distance on the target manifold $\mathcal{M}_N$. The surrogate reduces to ordinary least squares with target $\Phi := \frac{1}{m}\sum_i \tilde{y}_i\xi_i^*$. The end-to-end pullback is $\mathcal{F}(A) := \frac{1}{2m}\sum_i d_K(\exp_{o_N}(A\xi_i), y_i)^2$, so that $\mathcal{L}_K(W) = \mathcal{F}(W_{1:N})$.

A.3 Balancedness and Deficiency Margin

Definition 1 (δ-balancedness)**.** *Matrices $W_1, \dots, W_N$ are δ-balanced if $\|W_{j+1}^\top W_{j+1} - W_j W_j^\top\|_F \le \delta$ for all j.*

Definition 2 (Deficiency margin)**.** *A matrix $W \in \mathbb{R}^{d_N \times d_0}$ has deficiency margin $c' > 0$ with respect to Φ if $\|W - \Phi\|_F \le \sigma_{\min}(\Phi) - c'$.*

Definition 2 presumes $\sigma_{\min}(\Phi) > 0$, and we assume throughout that no layer is a bottleneck, $\min_j d_j \ge \min(d_0, d_N)$, so that Φ is representable by the factorization. For $K < 0$ the margin is imposed on $A(0)$ in the curvature-corrected form $S_K(R)(2\mathcal{E}(W(0)))^{1/2} \le \sigma_{\min}(\Phi) - c'$, with $S_K(R)$ the Rauch factor of Appendix A.4; at $K = 0$ this is a loss-based sufficient strengthening of Definition 2: it reads $(2\mathcal{E}(W(0)))^{1/2} \le \sigma_{\min}(\Phi) - c'$, which since $2\mathcal{E}(W(0)) = \|A(0) - \Phi\|_F^2 + 2\mathcal{E}^\star$ reduces to Definition 2 in the realizable case $\mathcal{E}^\star = 0$ and is otherwise strictly stronger.

A.4 Comparison-Geometry Notation

The model sine and cosine are

$$\mathrm{sn}_K(r) := \begin{cases} \sin(\sqrt{K}r)/\sqrt{K} & K > 0, \\ r & K = 0, \\ \sinh(\sqrt{\kappa}r)/\sqrt{\kappa} & K < 0, \end{cases} \tag{4}$$

and $\mathrm{cn}_K(r) := \mathrm{sn}_K'(r)$. The **Rauch factor** $S_K(R) := \mathrm{sn}_K(R)/R$ is the transverse stretching of $D\exp_o(v)$ for $\|v\| = R$; it exceeds 1 for $K < 0$, equals 1 for $K = 0$, and is below 1 for $K > 0$ within the injectivity radius. Since the radial direction is unstretched, the operator norm of the differential is $\|D\exp_o(v)\|_{\mathrm{op}} = \widehat{S}_K(R) := \max(1, S_K(R))$. The **Hessian comparison function** for $\phi(p) = \frac{1}{2}d(p,q)^2$ is $f_K(D) := D\,\mathrm{cn}_K(D)/\mathrm{sn}_K(D)$, the eigenvalue in transverse directions, the radial one being 1; both ratios are extended by continuity, $S_K(0) = f_K(0) = 1$, and the **Hessian bound** is $H_K(D) := \max(1, f_K(D))$. Let $B_K(R) := \sup_{\|v\|\le R}\|D^2\exp_o(v)\|_{\mathrm{op}}$. The **tube** $\mathcal{T}_{R,D}$ is the set of A with $\|A\xi_i\| \le R$ and $d(\exp_{o_N}(A\xi_i), y_i) \le D$ for all i, and the **smoothness constant** is

$$L_{K,R,D} := \tfrac{1}{m}\sum_{i=1}^m \|\xi_i\|^2\big[\widehat{S}_K(R)^2 H_K(D) + D\,B_K(R)\big]. \tag{5}$$

A.5 Statements of the Results

The results below use the numbering of the main paper and are stated with the precise hypotheses and conclusions used in the proofs of Appendix D. Some main-text statements suppress technical conditions for readability; the formulations here specify the applicable tube, initialization, step-size, and curvature regimes. In particular, the trajectory-containment and linear-convergence results (Lemma 8, Theorem 9) are established for $K \le 0$; the positive-curvature extension (Theorem 10)

concerns the one-step descent result within the stated radius restrictions. For $K > 0$ the operator-norm Rauch factor is $\widehat{S}_K(R) := \max(1, S_K(R)) = 1$, so positive curvature carries no transverse Rauch-factor penalty in the smoothness upper bound; this alone does not imply a strictly larger worst-case step-size ceiling than in the Euclidean case. All subsequent statements and proofs should be read under the hypotheses stated here.

Theorem 3 (Descent lemma)**.** *Let $K \le 0$ (Theorem 10 covers $K > 0$), let $A = W_{1:N} \in \mathcal{T}_{R,D}$ and $G := \nabla_A \mathcal{F}(A)$. Suppose: (1) $\|W_j\|_{\mathrm{op}} \le M$; (2) $\sigma_{\min}(A) > 0$; (3) the factors are sufficiently balanced: $\sum_j \|W^*_{j+1:N} G W^*_{1:j-1}\|^2_{\mathrm{HS}} \ge \frac{3}{4}\sigma_{\min}(A)^{2(N-1)/N}\|G\|^2_{\mathrm{HS}}$; (4) the segment $[A, A^+] \subset \mathcal{T}_{R,D}$. Then there exists $C_N > 0$ such that if*

$$\begin{aligned} &\eta C_N\big(M^{3N-4} G_{K,R,D} + L_{K,R,D} M^{4N-4}\big) \\ &\qquad \le \tfrac{1}{4}\sigma_{\min}(A)^{2(N-1)/N}, \end{aligned} \tag{6}$$

where $G_{K,R,D} := \sup_{B \in \mathcal{T}_{R,D}} \|\nabla_A \mathcal{F}(B)\|_{\mathrm{HS}}$, then gradient descent $W_j^+ = W_j - \eta \nabla_{W_j} \mathcal{L}_K$ gives

$$\mathcal{F}(A^+) - \mathcal{F}(A) \le -\tfrac{\eta}{2}\,\sigma_{\min}(A)^{2(N-1)/N}\|G\|^2_{\mathrm{HS}}. \tag{7}$$

Corollary 4 (Gradient flow: curvature-free descent)**.** *For 0-balanced factors, a condition the flow preserves, the gradient flow $\dot{W}_j = -\nabla_{W_j}\mathcal{L}_K$ satisfies*

$$\frac{d}{d\tau}\mathcal{F}(A(\tau)) \le -\sigma_{\min}(A)^{2(N-1)/N}\|G\|^2_{\mathrm{HS}} \tag{8}$$

with no curvature-dependent constraint.

Remark 5 (Curvature as a discretization cost)**.** *$L_{K,R,D}$ enters only in Stage 2. In continuous time there is no Taylor remainder, so no curvature constraint. The step-size penalty $\eta_K^\star = O(e^{-2\sqrt{\kappa}R})$ is a discrete-time phenomenon.*

The step-size constraint (6) is satisfied whenever

$$\eta \le \eta_K^\star := \frac{\sigma_{\min}(A)^{2(N-1)/N}}{4C_N(M^{3N-4}G_{K,R,D} + L_{K,R,D}M^{4N-4})}. \tag{9}$$

Lemma 6 (Local strong convexity of the pullback)**.** *Assume the data are whitened, meaning $\frac{1}{m}\sum \xi_i \xi_i^* = I$, and that the tube is narrow enough that $\mu_K := \min(1, S_K(R)^2 f_K(D)) - D\,B_K(R) > 0$. Then $D^2\mathcal{F}(A)[\Delta,\Delta] \ge \mu_K\|\Delta\|^2_{\mathrm{HS}}$ on the tube, so $\mathcal{F}$ is μ_K-strongly convex along any segment contained in it. For $K \le 0$ the minimum equals 1, giving $\mu_K = 1 - D\,B_K(R)$, and $\mu_0 = 1$.*

Lemma 7 (Gradient dominance)**.** *Under the hypotheses of Lemma 6, let $A^\star$ minimize $\mathcal{F}$ on the tube. If $[A, A^\star] \subset \mathcal{T}_{R,D}$ then $\|\nabla_A \mathcal{F}(A)\|^2 \ge 2\mu_K(\mathcal{F}(A) - \mathcal{F}^\star)$.*

Lemma 8 (Trajectory containment)**.** *Let $K \le 0$. Fix Φ and $c' > 0$, and let the data be whitened. There exist $\delta_0, \eta_0 > 0$ such that if the initialization is δ_0-balanced with deficiency margin c' and $\eta \le \min(\eta_K^\star, \eta_0)$, then for all $t \ge 0$: (a) $A(t) \in \mathcal{T}_{R,\bar{D}}$ with $R := \max_i \max(2\|\Phi\|_F\|\xi_i\|, \|\tilde{y}_i\|)$, which also contains the targets, $D(0) := (2m\,\mathcal{F}(A(0)))^{1/2}$ the endpoint residual bound, and $\bar{D} := D(0) + \delta_D$ a fixed tube radius with slack $\delta_D > 0$ chosen before η_0 so that every update segment stays in the tube; (b) the factors remain $2\delta_0$-balanced, so hypothesis (3) of Theorem 3 holds; (c) $\sigma_{\min}(A(t)) \ge c'$.*

Theorem 9 (Linear convergence). *Let the curvature be nonpositive, $K \le 0$, and suppose the data are whitened. Assume the initialization is δ-balanced with $\delta \le \delta_0$ and deficiency margin $c' > 0$, and that it starts inside a tube narrow enough that $A(0) \in \mathcal{T}_{R,\bar{D}}$ with $\bar{D}\, B_K(R) < 1$, for the fixed tube radius $\bar{D} = D(0) + \delta_D$ and the constants δ_0, η_0 of Lemma 8. Finally, let the step size satisfy $\eta \le \min(\eta_K^\star, \eta_0)$, with $\eta_K^\star$ evaluated at $\sigma_{\min}(A) = c'$ and $D = \bar{D}$ and $[A(t), A^\star] \subset \mathcal{T}_{R,\bar{D}}$ as in Lemma 7. Writing $\mu_K := 1 - \bar{D}\, B_K(R)$, then*

$$\begin{aligned}\mathcal{F}(A(T)) - \mathcal{F}^\star \le \big(1 - \eta\mu_K (c')^{2(N-1)/N}\big)^T \\ \times \big(\mathcal{F}(A(0)) - \mathcal{F}^\star\big).\end{aligned} \tag{10}$$

The number of steps to reach $\mathcal{F} - \mathcal{F}^\star \le \varepsilon$ is, for the actual *step size η,*

$$T \ge \frac{1}{\eta\, \mu_K\, (c')^{2(N-1)/N}} \log \frac{\mathcal{F}(A(0)) - \mathcal{F}^\star}{\varepsilon}. \tag{11}$$

If, in addition, $\eta_K^\star$ is the active restriction ($\eta_K^\star = O(\eta_0)$) and $\eta = \Theta(\eta_K^\star)$, substituting the ceiling (9) evaluated at $\bar{D}$ gives the η-independent form

$$\begin{aligned}T = O\Bigg(\frac{M^{3N-4} G_{K,R,\bar{D}} + M^{4N-4} L_{K,R,\bar{D}}}{\mu_K\, (c')^{4(N-1)/N}} \\ \times \log \frac{\mathcal{F}(A(0)) - \mathcal{F}^\star}{\varepsilon} \Bigg).\end{aligned} \tag{12}$$

Theorem 10 (General curvature). *Theorem 3 holds for constant sectional curvature $K \in \mathbb{R}$ with the constants $\widehat{S}_K(R)$, $H_K(D)$, $B_K(R)$ from Appendix A.4, provided $R < \pi/(2\sqrt{K})$ and $D < \pi/(2\sqrt{K})$ when $K > 0$.*

Corollary 11 (Positive curvature carries no Rauch-factor penalty). *Let the curvature be positive, $K > 0$, and assume the radius and the residual stay within the injectivity radius, so that $\sqrt{K}R < \pi/2$ and $\sqrt{K}D < \pi/2$. Then: (1) $S_K(R) = \sin(\sqrt{K}R)/(\sqrt{K}R) < 1$; (2) $H_K(D) = 1$; (3) $\widehat{S}_K(R) = 1$, so $L_{K,R,D} \to L_{0,R,D}$ as $D \to 0$; (4) $\eta_K^\star \to \eta_0^\star$.*

Remark. This is a statement about the worst-case operator-norm bound. It removes the transverse Rauch-factor penalty but does not, by itself, establish $\eta_K^\star > \eta_0^\star$ for $D > 0$. Smaller empirical Hessian eigenvalues in particular datasets are compatible with, but not implied by, this bound.

Proposition 12 (Curvature-free rate). *Suppose the data are whitened ($\frac{1}{m}\sum_i \xi_i \xi_i^* = I_{d_0}$). Under δ-balancedness, deficiency margin $c' > 0$, and a step size $\eta \le \eta_0^{\mathcal{E}}(c', N, \|\Phi\|_F)$, gradient descent on $\mathcal{E}$ (3) converges linearly: $\mathcal{E}(W(T)) - \mathcal{E}^\star \le \varepsilon$ for $T \ge (\eta (c')^{2(N-1)/N})^{-1} \log((\mathcal{E}(W(0)) - \mathcal{E}^\star)/\varepsilon)$, where $\mathcal{E}^\star := \min_W \mathcal{E}(W)$. The step size, contraction factor, and iteration count are all independent of the curvature K.*

Theorem 13 (No spurious critical points on the tube). *Under the hypotheses of Lemma 6 (constant curvature K, whitened data, $\mu_K > 0$), every critical point A of $\mathcal{F}$ on the tube with $[A, A^\star] \subset \mathcal{T}_{R,D}$ minimizes $\mathcal{F}$ on the tube: $\nabla_A \mathcal{F}(A) = 0 \Rightarrow \mathcal{F}(A) = \mathcal{F}^\star$. For $K \le 0$ this holds on any tube with $D\, B_K(R) < 1$. For $K > 0$ it holds within the injectivity radius.*

Remark 14. *Curvature changes the convergence rate but not this conclusion, affecting only the size of the admissible tube and the value of μ_K. The statement is local, and concerns $\mathcal{F}$ rather than the factorized loss.*

Proposition 15 (Adaptive schedule)**.** *Assume the hypotheses of Theorem 9 and use the fixed tube $\mathcal{T}_{R,\bar{D}}$ of Lemma 8. At iteration t, form a positive trial step from (9) using the current residual $D(t) := \max_i d(\exp(A(t)\xi_i), y_i)$, cap it by η_0, and halve it until both*

$$[A(t), A(t+1)] \subset \mathcal{T}_{R,\bar{D}}$$

and the descent inequality (7) hold. Then the backtracking terminates, every accepted step obeys (7), and the variable-step trajectory retains the containment, balancedness, and singular-value conclusions of Lemma 8.

Remark 16. *Main-text Proposition 15 items (2)–(3) — the late-phase step-size lower bound and the $H_K(D(0))$ speedup — are stated here as the conditional late-phase claims of Remark 17, since the argument establishes them only asymptotically and under the additional conditions listed there, not as whole-trajectory guarantees.*

Remark 17. *For $K < 0$, the fixed ceiling can be conservative when the initial residual makes $H_K(D(0))$ large. Conditional on the trajectory entering a regime in which $D(t) \to 0$, the $G_{K,R,D(t)}$ and $D(t)B_K(R)$ terms are not dominant, accepted steps stay within a constant factor of the instantaneous trial ceiling, and the required segment containment continues to hold, the leading-order late-phase smoothness is proportional to $S_K(R)^2$. Under these additional conditions the late-phase contraction scales as $1 - c/S_K(R)^2$, and a comparison with the initial fixed ceiling suggests a speedup of order $H_K(D(0))$. This is a conditional asymptotic comparison, not a global iteration-complexity bound or a proved lower bound on every accepted step.*

Proposition 18 (Curvature-dependent gradient anisotropy)**.** *Consider gradient flow on $\mathcal{L}_K$ with 0-balanced initialization $W_j(0) = \alpha^{1/N} U_j$ ($\alpha \to 0^+$), whitened data ($\frac{1}{m}\sum_i \xi_i \xi_i^* = I_{d_0}$), and targets inside the curvature scale, $\sqrt{\kappa}\max_i \|\tilde{y}_i\| \ll 1$. Then:*

1. ***Early phase*** *($\|A\| \ll 1$): $\nabla^2 \phi_i|_o = I + O(\kappa\|\tilde{y}_i\|^2)$, so $\nabla_A \mathcal{F} = \nabla_A \mathcal{E} + O(\kappa \max_i \|\tilde{y}_i\|^2 \|A\| + \|A\|^2)$. Thus the intrinsic and Euclidean end-to-end gradients agree to leading order in the stated early-phase, small-curvature-scale regime.*
2. ***Late phase*** *($\|A\| = \Theta(1)$): the intrinsic gradient $G_K = -\frac{1}{m}\sum [D\exp(A\xi_i)]^* \log_{p_i}(y_i)\xi_i^*$ weights SVD direction k by $S_K(\sigma_k\|\xi_i\|)$ when ξ_i aligns with the k-th right singular vector, so for $K < 0$ transverse residual components are amplified and the flow reduces transverse error faster than radial error.*
3. ***Fixed-point condition****: along the flow $\dot{A} = -\mathcal{P}_A(G_K)$, and at a stationary limit point $A^\star$ (where $\dot{A} = 0$) the intrinsic gradient satisfies $G_K(A^\star) \in \ker(\mathcal{P}_{A^\star})$, informative only when the residuals do not vanish. Since G_K depends on K through the Rauch-weighted adjoint, curvature changes the stationarity equation and may affect the stationary behavior; no claim of distinct selected limits is made. At $K = 0$ this reduces to Euclidean stationarity.*

Remark 19. *Where the main text says curvature "can change the selected limit point" (Proposition 17, item 3), the precise statement is item 3 above: curvature changes the stationarity equation $G_K(A^\star) \in \ker(\mathcal{P}_{A^\star})$ and may affect the stationary behavior. No claim of distinct selected limits is made.*

Proposition 20 (Sharpness)**.** *For every $K < 0$ and $R > 0$ there is a configuration with $d_N \geq 2$, $d_0 = 1$, $N = 1$, $m = 1$ in which A_0 is a zero-loss minimizer and: (1) $D^2\mathcal{F}(A_0)[\Delta, \Delta] = L_{K,R,0}\|\Delta\|_{\mathrm{HS}}^2 = S_K(R)^2\|\Delta\|_{\mathrm{HS}}^2$ for transverse Δ; (2) gradient descent at A_0 is linearly unstable for every $\eta > 2/S_K(R)^2$; (3) hence $\eta_{\mathrm{crit}} = 2/S_K(R)^2 = \Theta(1/S_K(R)^2)$. The construction uses a zero-loss stationary point, which is necessary for an exact linear-stability threshold.*

B Extended Related Work

This appendix expands Section 2 of the main paper with the full account of the literature that space there did not permit.

Non-Euclidean representation learning. Trees and hierarchical metrics embed into hyperbolic space with far lower distortion than into $\mathbb{R}^n$ (Sala et al., 2018), motivating Poincaré and Lorentz embeddings (Nickel and Kiela, 2017, 2018; Gu et al., 2018) and a range of manifold-valued architectures: Möbius/hyperbolic networks (Ganea et al., 2018; Tifrea et al., 2018), fully hyperbolic (Chen et al., 2022; Bdeir et al., 2024), Poincaré residual networks (Van Spengler et al., 2023), hyperbolic attention (Gulcehre et al., 2018), pseudo-Riemannian (Xiong et al., 2022; Law, 2021), mixed-curvature (Wang et al., 2021; Cho et al., 2023; Skopek et al., 2019), adaptive-curvature (Fu et al., 2021), and Riemannian spiking (Sun et al., 2024) models. On the graph side, Lorentzian (Zhang et al., 2021), hyperbolic-to-hyperbolic (Dai et al., 2021), hyperbolic variational (Sun et al., 2021), and spectro-Riemannian (Grover et al., 2025) graph neural networks extend the foundational Euclidean architectures (Kipf and Welling, 2016; Hamilton et al., 2017; Veličković et al., 2017; Xu et al., 2018). Downstream applications now span image embeddings (Khrulkov et al., 2020), segmentation (Atigh et al., 2022), generative modeling (Mathieu et al., 2019), and numerically stable embedding (Yu and De Sa, 2019; Mishne et al., 2023); see Yang et al. (2022) for a survey. These architectures motivate the manifold-valued setting whose optimization we study; we treat target-manifold curvature as the object of interest and analyze its bias-free deep-linear case.

Deep linear network optimization. Trajectory analysis of gradient descent for deep linear networks began with Saxe et al. (2013). Landscape results establish the absence of spurious local minima (Lu and Kawaguchi, 2017; Laurent and Brecht, 2018; Nouiehed and Razaviyayn, 2022; Sun and Luo, 2016), and Arora et al. (2018a) proved linear convergence under approximate balancedness and a deficiency margin. Subsequent work examines the role of width (Du and Hu, 2019), identity initialization (Hardt and Ma, 2016), gradient alignment (Ji and Telgarsky, 2018), exponential convergence phenomena (Shamir, 2019), and residual architectures (Zou et al., 2020); related overparameterized convergence results appear in (Du et al., 2019; Allen-Zhu et al., 2019). The factorization-induced preconditioner of Arora et al. (2018a) is the algebraic core of our descent lemma, and Riemannian and convolutional treatments of the deep linear network appear in (Bah et al., 2022; Menon, 2025; Achour et al., 2025). Complementary to this Euclidean theory, a separate tradition optimizes over manifold-valued *parameters* (Absil et al., 2008; Boumal, 2023; Hu et al., 2020; Zhang and Sra, 2016; Bonnabel, 2013), with stochastic and variance-reduced extensions (Zhang et al., 2016; Sato et al., 2019) and adaptive methods (Bécigneul and Ganea, 2019); in particular, Criscitiello and Boumal (2022, 2023) show that negative curvature obstructs acceleration for geodesically convex optimization, a lower bound consistent with our step-size degradation $\eta_K^\star = O(e^{-2\sqrt{\kappa}R})$. In that setting the curvature lives on the parameter space, whereas in ours the parameters are Euclidean and curvature enters only through the loss. Concurrently, Sarkar (2025) studies curvature-adaptive generalization for manifold-valued networks, and Li et al. (2024) analyzes Lipschitz properties of hyperbolic neural networks, and both are complementary to our optimization-theoretic focus.

Generalization and manifold estimation. Generalization theory for graph and hyperbolic embeddings (Suzuki et al., 2021; Tang and Liu, 2023) and for neural networks more broadly (Bartlett et al., 2017; Neyshabur et al., 2017; Golowich et al., 2018; Arora et al., 2018b; Long and Sedghi, 2020; Scarselli et al., 2018; Verma and Zhang, 2019; Garg et al., 2020; Liao et al., 2020; Esser

et al., 2021; Maskey et al., 2022; Ju et al., 2023; Sun and Lin, 2024) addresses a different question, bounding test error rather than training dynamics, but motivates the manifold-valued architectures whose optimization we analyze. Classical manifold learning and estimation (Fefferman et al., 2016; Genovese et al., 2012; Kim and Zhou, 2015; Niyogi et al., 2008) provides further context for this broader program. To our knowledge, no prior work derives curvature-dependent descent lemmas, linear convergence, landscape results, adaptive schedules, or implicit-bias characterizations for deep manifold-valued networks.

C Coordinate Models for the Target Manifold

The Poincaré ball and the κ-stereographic model are two coordinate charts for the same space form, not two different geometries, and they enter only where closed-form exp and log are needed. We use the Poincaré ball for the hyperbolic theory ($K < 0$), where it is standard, and the κ-stereographic model for any result or experiment that must cross $K = 0$ into the spherical case, since the ball is undefined there. The two agree on their common domain $K < 0$, so no theorem depends on the choice. The hyperbolic literature also writes $\mathbb{D}_c$ for the ball "of curvature c," meaning $K = -c$; we avoid c as a standalone symbol, since it would collide with the deficiency margin c' (Definition 2), and translate $c = \kappa = -K$ wherever it appears.

C.1 Poincaré Ball ($K < 0$)

$\mathbb{D}_K^n := \{x \in \mathbb{R}^n : \kappa\|x\|^2 < 1\}$ carries the metric $g_x = (\lambda_x)^2 g^E$ with $\lambda_x := 2/(1 - \kappa\|x\|^2)$. At the origin $o = \mathbf{0}$, $\exp_o(v) = \tanh(\sqrt{\kappa}\|v\|)\, v/(\sqrt{\kappa}\|v\|)$ and $\log_o(y) = \tanh^{-1}(\sqrt{\kappa}\|y\|)\, y/(\sqrt{\kappa}\|y\|)$. Möbius matrix–vector multiplication is defined by $W \otimes_K x := \exp_o(W \log_o(x))$ (Ganea et al., 2018). Because $\log_o \circ \exp_o$ is the identity wherever both are defined, composing N such layers telescopes to $W_N \otimes_K \cdots \otimes_K W_1 \otimes_K x = \exp_o(W_{1:N} \log_o(x))$, the collapse (1). The telescoping is a cancellation of adjacent $\log_o$ and $\exp_o$ factors, available only because every layer is bias-free and shares the single base point o. It is not a statement about the gyroassociativity of the Möbius framework, where $\oplus_K$ is neither associative nor commutative. Inserting Möbius biases $\oplus_K b_j$ between layers breaks the telescoping and (1) fails, so biases fall outside our scope.

C.2 κ-Stereographic Model (any K)

The ball is defined only for $K < 0$ and therefore cannot express the positive-curvature results of Section 4.3 of the main paper or the continuous sweep through $K = 0$ used in the experiments. For these we use the κ-stereographic model of Bachmann et al. (2020) (see also Skopek et al., 2019), which places both the sphere and the hyperbolic ball in one coordinate chart by stereographic projection, the same construction that maps the globe onto a flat plane, so that a single formula covers all three geometries. The model is the family $\mathrm{st}_K^n := \{x \in \mathbb{R}^n : -K\|x\|^2 < 1\}$ with $\lambda_x = 2/(1 + K\|x\|^2)$, defined for every $K \in \mathbb{R}$ and smooth in K. It reduces to the Poincaré ball for $K < 0$, to $\mathbb{R}^n$ for $K = 0$, and to the stereographic sphere for $K > 0$. The single-chart property allows the experiments in Appendix E to vary K continuously through zero without switching coordinates. These formulas are written in the ambient chart, whose conformal factor satisfies $\lambda_o = 2$ at the base point; the losses and the tangent-space surrogate are expressed in the g_o-orthonormal tangent coordinates fixed in Appendix A.1 (equivalently $\lambda_o = 1$), which is what makes $\exp_o$ a radial isometry and $\mathcal{L}_0 = \mathcal{E}$ at $K = 0$.

D Proofs

Every result restated in Appendix A.5 is proved below, in the order in which it appears in the main paper.

D.1 Proof of the Descent Lemma

Proof of Theorem 3. The proof proceeds in two stages. Stage 1 establishes the descent inequality in continuous time (gradient flow), where no curvature term appears, and Stage 2 passes to discrete gradient descent, where the curvature enters through the Taylor remainder.

Stage 1: Gradient flow (continuous time).

Assume the factors $W_1, \dots, W_N$ are exactly balanced ($\delta = 0$ in Definition 1):

$$W_{j+1}^\top W_{j+1} = W_j W_j^\top, \qquad j = 1, \dots, N-1,$$

and pass to continuous time $\eta \to 0^+$, replacing discrete updates by the gradient flow

$$\dot{W}_j(\tau) = -\nabla_{W_j} \mathcal{L}_K, \qquad j = 1, \dots, N, \quad \tau \in [0, \infty).$$

The preconditioner. Under exact balancedness, Theorem 1 and Claim 1 of Arora et al. (2018a) show that the balancedness condition is preserved for all $\tau \geq 0$, and that the end-to-end matrix $A(\tau) = W_{1:N}(\tau)$ evolves according to

$$\operatorname{vec}\big(\dot{A}(\tau)\big) = -\mathcal{P}_{A(\tau)} \ \operatorname{vec}\big(G(\tau)\big), \tag{13}$$

where $G(\tau) := \nabla_A \mathcal{F}(A(\tau))$ and

$$\mathcal{P}_A(G) := \sum_{j=1}^{N} W_{j+1:N} W_{j+1:N}^* \, G \, W_{1:j-1}^* W_{1:j-1}$$

is a positive semidefinite linear operator (acting on $\operatorname{vec}(G)$) whose eigenvalues are all at least $\sigma_{\min}(A)^{2(N-1)/N}$.

To see why: write the eigenvalue decomposition $W_1^\top W_1 = UDU^\top$, $W_N W_N^\top = VEV^\top$. Exact balancedness implies that all $W_j W_j^\top$ share the same spectrum, and one can show (via Lemma 5 of Arora et al. (2018a)) that the operator $\mathcal{P}_A$ is unitarily equivalent to the diagonal operator with entries $\sum_{j=1}^{N} d_k^{j-1} e_\ell^{N-j}$ for pairs (k, ℓ), where d_k, e_ℓ are the diagonal entries of D, E. By AM–GM,

$$\begin{aligned} \sum_{j=1}^{N} d_k^{j-1} e_\ell^{N-j} \ &\geq \ N \, (d_k^{N-1} e_\ell^{N-1})^{1/N} \\ &\geq \ N \, \sigma_{\min}(A)^{2(N-1)/N}. \end{aligned}$$

(The factor N is absorbed into C_N in the discrete analysis. The relevant conclusion is the lower bound $\sigma_{\min}(A)^{2(N-1)/N}$.)

Continuous-time descent. Using the chain rule and (13):

$$\begin{aligned}\frac{d}{d\tau}\mathcal{F}(A(\tau)) &= \big\langle \mathrm{vec}(G),\ \mathrm{vec}(\dot{A})\big\rangle \\ &= -\big\langle \mathrm{vec}(G),\ \mathcal{P}_A\ \mathrm{vec}(G)\big\rangle \\ &\le -\sigma_{\min}(A)^{2(N-1)/N}\ \|G\|_{\mathrm{HS}}^2,\end{aligned}$$

where the last step uses the eigenvalue lower bound on $\mathcal{P}_A$ and $\|G\|_{\mathrm{HS}} = \|\,\mathrm{vec}(G)\|$.

No smoothness constant, Rauch factor, or curvature term appears. This proves Corollary 4 (equation (8)).

Stage 2: Discrete updates (finite $\eta > 0$).

We now work with a single gradient-descent step $W_j^+ = W_j - \eta\,\nabla_{W_j}\mathcal{L}_K$ and bound $\mathcal{F}(A^+) - \mathcal{F}(A)$ for $A^+ = W_{1:N}^+$. The argument has six steps.

Step 1 (Pullback gradient). Define $p_i := \exp_{o_N}(A\xi_i) \in \mathcal{M}_N$ and $\phi_i(p) := \frac{1}{2}\,d(p, y_i)^2$. The Riemannian gradient of ϕ_i at p_i is the standard identity $\nabla_p \phi_i\big|_{p=p_i} = -\log_{p_i}(y_i)$. The chain rule for the composition $\mathcal{F}(A) = \frac{1}{m}\sum_i \phi_i(\exp_{o_N}(A\xi_i))$, which is (2) because $\phi_i = \frac{1}{2}d(\cdot, y_i)^2$, gives

$$G \;=\; \nabla_A \mathcal{F}(A) \;=\; -\frac{1}{m}\sum_{i=1}^{m}\big[D\exp_{o_N}(A\xi_i)\big]^*\ \log_{p_i}(y_i)\ \xi_i^*. \tag{14}$$

Here $D\exp_{o_N}(v)$ is the differential of the exponential map at the tangent vector $v = A\xi_i$, and $[\cdot]^*$ denotes the adjoint (transpose in coordinates). This expression is the pullback of the Riemannian gradient $\nabla_p \phi_i$ through the exponential map and the linear map $A \mapsto A\xi_i$.

Step 2 (Curvature-dependent smoothness of $\mathcal{F}$). We show that $\mathcal{F}$ is $L_{K,R,D}$-smooth on the tube $\mathcal{T}_{R,D}$. The Hessian of the pullback in a direction $\Delta \in \mathbb{R}^{d_N \times d_0}$ is computed by differentiating (14):

$$\begin{aligned}D^2\mathcal{F}(A)[\Delta,\Delta] &= \frac{1}{m}\sum_{i=1}^{m}\big[(\mathrm{I})_i + (\mathrm{II})_i\big], \quad \text{where} \\ (\mathrm{I})_i &= \big\langle \nabla^2\phi_i\big|_{p_i}\big(D\exp(A\xi_i)\Delta\xi_i\big),\ D\exp(A\xi_i)\Delta\xi_i\big\rangle, \\ (\mathrm{II})_i &= \big\langle \nabla\phi_i\big|_{p_i},\ D^2\exp(A\xi_i)[\Delta\xi_i,\Delta\xi_i]\big\rangle.\end{aligned}$$

Term $(\mathrm{I})_i$ is the Hessian of ϕ_i and term $(\mathrm{II})_i$ the second derivative of exp.

We bound each piece using the comparison-geometry constants from Section A.4 and the tube assumption $A \in \mathcal{T}_{R,D}$ (so $\|A\xi_i\| \le R$ and $d(\exp_{o_N}(A\xi_i), y_i) \le D$).

Term (I). The Rauch comparison theorem on a constant-curvature space gives $\|D\exp_{o_N}(v)\|_{\mathrm{op}} \le \widehat{S}_K(R)$ for $\|v\| \le R$, where $S_K(R) = \mathrm{sn}_K(R)/R$ is the Rauch factor (4). Hence

$$\|D\exp(A\xi_i)\cdot\Delta\xi_i\| \le \widehat{S}_K(R)\,\|\Delta\xi_i\|.$$

The Hessian comparison theorem for $\phi_i(p) = \frac{1}{2}d(p, y_i)^2$ on a space of constant curvature K gives the eigenvalue bound $\|\nabla^2\phi_i\|_{\mathrm{op}} \le H_K(D)$, where $H_K(D) = \max(1, f_K(D))$ with $f_K(D) = D\,\mathrm{cn}_K(D)/\mathrm{sn}_K(D)$. (The radial eigenvalue is always 1, and the transverse eigenvalue $f_K(D)$ exceeds 1 when $K < 0$.) Therefore

$$|(\mathrm{I})| \le H_K(D)\cdot\widehat{S}_K(R)^2\,\|\Delta\xi_i\|^2.$$

Term (II). The gradient bound $\|\nabla\phi_i\big|_{p_i}\| = \|\log_{p_i}(y_i)\| = d(p_i, y_i) \le D$ and the second-derivative bound $\|D^2 \exp_{o_N}(v)\|_{\mathrm{op}} \le B_K(R)$ for $\|v\| \le R$ give

$$|(\mathrm{II})| \le D\, B_K(R)\, \|\Delta\xi_i\|^2.$$

Combining and summing:

$$\begin{aligned} &|D^2\mathcal{F}(A)[\Delta,\Delta]| \\ &\qquad \le \frac{1}{m}\sum_{i=1}^{m} \|\xi_i\|^2 \big[\widehat{S}_K(R)^2 H_K(D) + D\, B_K(R)\big] \\ &\qquad\qquad \times \|\Delta\|_{\mathrm{HS}}^2 \\ &\qquad = L_{K,R,D}\, \|\Delta\|_{\mathrm{HS}}^2, \end{aligned} \tag{15}$$

where $L_{K,R,D}$ is the smoothness constant (5). This establishes that $\mathcal{F}$ is $L_{K,R,D}$-smooth on $\mathcal{T}_{R,D}$.

As a consequence, the standard quadratic upper bound holds for any A, A' with the segment $[A, A'] \subset \mathcal{T}_{R,D}$:

$$\mathcal{F}(A') \le \mathcal{F}(A) + \langle G,\, A' - A\rangle + \tfrac{1}{2}\, L_{K,R,D}\, \|A' - A\|_{\mathrm{HS}}^2. \tag{16}$$

Step 3 (Factor gradients). By the chain rule for the factorized parameterization $\mathcal{L}_K(W_1, \dots, W_N) = \mathcal{F}(W_{1:N})$:

$$\nabla_{W_j}\mathcal{L}_K = W_{j+1:N}^*\, G\, W_{1:j-1}^*, \qquad j = 1, \dots, N,$$

where we set $W_{1:0} := I_{d_0}$ and $W_{N+1:N} := I_{d_N}$ for the boundary cases. The gradient step on each factor is therefore

$$W_j^+ = W_j - \eta\, W_{j+1:N}^*\, G\, W_{1:j-1}^*.$$

Step 4 (End-to-end expansion). We expand the updated end-to-end matrix $A^+ = W_N^+ \cdots W_1^+$ in powers of η. Write $W_j^+ = W_j - \eta H_j$ with $H_j := W_{j+1:N}^* G W_{1:j-1}^*$. Then

$$\begin{aligned} A^+ &= (W_N - \eta H_N)\cdots(W_1 - \eta H_1) \\ &= W_{1:N} - \eta \sum_{j=1}^{N} W_{j+1:N}\, H_j\, W_{1:j-1} \\ &\qquad + (\text{higher order in } \eta) \\ &= A - \eta\, \mathcal{P}_A(G) + R_\eta, \end{aligned}$$

where the first-order term is the preconditioned gradient

$$\mathcal{P}_A(G) = \sum_{j=1}^{N} W_{j+1:N} W_{j+1:N}^*\, G\, W_{1:j-1}^* W_{1:j-1}$$

and R_η collects all terms of order η^2 and higher.

To bound R_η, we use the multinomial expansion of the product. Each term in R_η involves at least two factors of $-\eta H_j$ and the remaining factors are W_k. Using $\|W_j\|_{\mathrm{op}} \le M$ (hypothesis (1)) and $\|H_j\|_{\mathrm{HS}} = \|W_{j+1:N}^* G W_{1:j-1}^*\|_{\mathrm{HS}} \le M^{N-1}\|G\|_{\mathrm{HS}}$, one bounds (following the combinatorial accounting in Appendix D.2 of Arora et al. (2018a)):

$$\|R_\eta\|_{\mathrm{HS}} \le C'_N\, \eta^2\, M^{3N-4}\, \|G\|_{\mathrm{HS}}^2, \tag{17}$$

where $C'_N > 0$ depends only on the depth N.

More precisely, the k-th order term ($k \geq 2$) in η involves choosing k of the N positions to replace W_j by $-\eta H_j$. Each such term is bounded in Hilbert–Schmidt norm by $\eta^k M^{k(N-1)+N-k}\|G\|_{\mathrm{HS}}^k \leq \eta^k M^{(2k-1)(N-1)}\|G\|_{\mathrm{HS}}^k$. Using the gradient bound $\|G\|_{\mathrm{HS}} \leq G_{K,R,D}$ on the tube and summing the geometric series (under the step-size constraint $\eta M^{N-2} G_{K,R,D} \leq 1/2$, which is ensured by (6)), one obtains the tighter form

$$\|R_\eta\|_{\mathrm{HS}} \leq C_N\, \eta^2\, M^{3N-4}\, G_{K,R,D}\, \|G\|_{\mathrm{HS}}, \tag{18}$$

where we have peeled off one power of $\|G\|_{\mathrm{HS}}$ from the leading $k = 2$ term and bounded the remaining $\|G\|_{\mathrm{HS}}^{k-1}$ factors using $G_{K,R,D}$.

Step 5 (First-order term: coercivity). We now bound the first-order contribution. By hypothesis (3) (the balancedness condition in the theorem statement):

$$\sum_{j=1}^{N} \|W^*_{j+1:N} G W^*_{1:j-1}\|_{\mathrm{HS}}^2 \;\geq\; \tfrac{3}{4}\, \sigma_{\min}(A)^{2(N-1)/N}\, \|G\|_{\mathrm{HS}}^2.$$

A direct computation shows

$$\begin{aligned}\langle G,\, \mathcal{P}_A(G)\rangle &= \sum_{j=1}^{N} \big\langle G,\, W_{j+1:N} W^*_{j+1:N}\, G\, W^*_{1:j-1} W_{1:j-1}\big\rangle \\ &= \sum_{j=1}^{N} \|W^*_{j+1:N} G W^*_{1:j-1}\|_{\mathrm{HS}}^2,\end{aligned}$$

where the second equality follows from the trace identity $\langle A, BCB^\top\rangle = \langle B^\top A, CB^\top\rangle = \|B^\top A\|_{\mathrm{HS}}^2$ when $C = I$ (after appropriate reshaping; here one uses $\langle G, W_{j+1:N} W^*_{j+1:N} G W^*_{1:j-1} W_{1:j-1}\rangle = \mathrm{tr}(W_{1:j-1} G^\top W_{j+1:N} W^*_{j+1:N} G W^*_{1:j-1}) = \|W^*_{j+1:N} G W^*_{1:j-1}\|_{\mathrm{HS}}^2$). Therefore

$$\langle G,\, \mathcal{P}_A(G)\rangle \;\geq\; \tfrac{3}{4}\, \sigma_{\min}(A)^{2(N-1)/N}\, \|G\|_{\mathrm{HS}}^2. \tag{19}$$

Step 6 (Combining: the smoothness bound). Write $A^+ = A - \eta\, \mathcal{P}_A(G) + R_\eta$. By hypothesis (4), the segment $[A, A^+] \subset \mathcal{T}_{R,D}$, so the smoothness bound (16) applies:

$$\begin{aligned}&\mathcal{F}(A^+) - \mathcal{F}(A) \\ &\quad\leq \langle G,\, A^+ - A\rangle + \tfrac{1}{2} L_{K,R,D} \|A^+ - A\|_{\mathrm{HS}}^2 \\ &\quad= \langle G,\, -\eta\, \mathcal{P}_A(G) + R_\eta\rangle \\ &\qquad + \tfrac{1}{2} L_{K,R,D} \|-\eta\, \mathcal{P}_A(G) + R_\eta\|_{\mathrm{HS}}^2.\end{aligned}$$

We bound each piece.

First-order terms.

$$\begin{aligned}\langle G,\; -\eta\, \mathcal{P}_A(G) + R_\eta\rangle &= -\eta\, \langle G, \mathcal{P}_A(G)\rangle + \langle G, R_\eta\rangle \\ &\leq -\tfrac{3}{4}\, \eta\, \sigma_{\min}(A)^{2(N-1)/N} \|G\|_{\mathrm{HS}}^2 \\ &\quad + \|G\|_{\mathrm{HS}}\, \|R_\eta\|_{\mathrm{HS}},\end{aligned}$$

where we used (19) and Cauchy–Schwarz.

Second-order term.

$$\tfrac{1}{2}\, L_{K,R,D}\, \|A^+ - A\|_{\mathrm{HS}}^2 \le L_{K,R,D}\big(\eta^2 \|\mathcal{P}_A(G)\|_{\mathrm{HS}}^2 + \|R_\eta\|_{\mathrm{HS}}^2\big).$$

Using $\|\mathcal{P}_A(G)\|_{\mathrm{HS}} \le NM^{2(N-1)}\|G\|_{\mathrm{HS}}$ (each summand in $\mathcal{P}_A$ is bounded by $M^{2(N-1)}\|G\|_{\mathrm{HS}}$), this is bounded by $C_N''\,\eta^2\, L_{K,R,D}\, M^{4N-4}\, \|G\|_{\mathrm{HS}}^2$ (absorbing the $\|R_\eta\|^2$ term, which is $O(\eta^4)$ and dominated).

Collecting all the error terms (the $\langle G, R_\eta\rangle$ piece from (18) and the second-order piece), we arrive at

$$\begin{aligned}\mathcal{F}(A^+) - \mathcal{F}(A) \;\le\; & -\tfrac{3}{4}\,\eta\,\sigma_{\min}(A)^{2(N-1)/N}\,\|G\|_{\mathrm{HS}}^2 \\ & + C_N\,\eta^2\big(M^{3N-4}G_{K,R,D} + L_{K,R,D}M^{4N-4}\big)\|G\|_{\mathrm{HS}}^2,\end{aligned} \tag{20}$$

where $C_N > 0$ is a (redefined) constant depending only on N, and $G_{K,R,D} = \sup_{B\in\mathcal{T}_{R,D}} \|\nabla_A \mathcal{F}(B)\|_{\mathrm{HS}}$ is the gradient norm bound on the tube.

Applying the step-size condition. The step-size assumption (6) states

$$\begin{aligned}&\eta\, C_N\big(M^{3N-4}G_{K,R,D} + L_{K,R,D}M^{4N-4}\big) \\ &\qquad \le \tfrac{1}{4}\,\sigma_{\min}(A)^{2(N-1)/N}.\end{aligned}$$

Multiplying both sides by $\eta\,\|G\|_{\mathrm{HS}}^2$, this ensures that the positive (error) term in (20) is at most

$$\tfrac{1}{4}\,\eta\,\sigma_{\min}(A)^{2(N-1)/N}\,\|G\|_{\mathrm{HS}}^2.$$

Substituting back into (20):

$$\begin{aligned}&\mathcal{F}(A^+) - \mathcal{F}(A) \\ &\quad \le -\tfrac{3}{4}\,\eta\,\sigma_{\min}(A)^{2(N-1)/N}\|G\|_{\mathrm{HS}}^2 \\ &\qquad + \tfrac{1}{4}\,\eta\,\sigma_{\min}(A)^{2(N-1)/N}\|G\|_{\mathrm{HS}}^2 \\ &\quad = -\tfrac{1}{2}\,\eta\,\sigma_{\min}(A)^{2(N-1)/N}\|G\|_{\mathrm{HS}}^2,\end{aligned}$$

which is (7).

Where curvature enters and where it does not. In Stage 1, the descent (8) is controlled by the algebraic preconditioner $\mathcal{P}_A$ and the minimum singular value $\sigma_{\min}(A)$, and no geometric quantity appears. Curvature enters only in Stage 2, through the smoothness constant $L_{K,R,D}$ (5) in the quadratic remainder of the Taylor expansion (16). Inside $L_{K,R,D}$, the three comparison-geometry bounds, namely the Rauch factor $S_K(R)^2$, the Hessian comparison $H_K(D)$, and the exponential-map curvature $B_K(R)$, all grow with $|K|$ when $K < 0$. Their growth forces η to shrink to maintain the step-size condition (6), giving the ceiling $\eta_K^\star = O(1/S_K(R)^2)$ in (9). The single-step contraction factor $\frac{1}{2}\sigma_{\min}(A)^{2(N-1)/N}\|G\|_{\mathrm{HS}}^2$ is identical to the Euclidean deep-linear guarantee of Arora et al. (2018a), so curvature constrains only the step size, not the progress at each step. □

D.2 Proof of Local Strong Convexity

Proof of Lemma 6. Recall the end-to-end pullback $\mathcal{F}(A) = \frac{1}{m}\sum_{i=1}^m \phi_i(\exp_{o_N}(A\xi_i))$ with $\phi_i(p) := \frac{1}{2}d(p, y_i)^2$, where $A \in \mathcal{T}_{R,D}$. We compute the Hessian $D^2\mathcal{F}(A)[\Delta, \Delta]$ for an arbitrary direction $\Delta \in \mathbb{R}^{d_N \times d_0}$ and establish a *lower* bound, in contrast to the *upper* bound derived in Step 2 of the proof of Theorem 3 (Appendix D.1).

Step 1: Hessian decomposition. Set $v_i := A\xi_i \in V_N \simeq \mathbb{R}^{d_N}$ and $p_i := \exp_{o_N}(v_i) \in \mathcal{M}_N$. By the chain rule applied to the composition $A \mapsto v_i = A\xi_i \mapsto p_i = \exp_{o_N}(v_i) \mapsto \phi_i(p_i)$, the second derivative of the i-th summand $f_i(A) := \phi_i(\exp_{o_N}(A\xi_i))$ in direction Δ is

$$\begin{aligned} D^2 f_i(A)[\Delta, \Delta] \\ &= \underbrace{\left\langle \nabla^2 \phi_i\big|_{p_i}\big(D\exp_{o_N}(v_i)\Delta\xi_i\big), D\exp_{o_N}(v_i)\Delta\xi_i \right\rangle}_{\text{(I)}} \\ &\quad + \underbrace{\left\langle \nabla_p \phi_i\big|_{p_i}, D^2\exp_{o_N}(v_i)[\Delta\xi_i, \Delta\xi_i] \right\rangle}_{\text{(II)}}. \end{aligned} \tag{21}$$

Term (I) arises from differentiating the outer function ϕ_i through the first derivative $D\exp$; term (II) arises from the second derivative of exp hitting the gradient $\nabla\phi_i$. We bound each term from *below*.

Step 2: Lower bound on Term (I). We need two ingredients from comparison geometry on the space form $(\mathcal{M}_N, g_N)$ of constant sectional curvature K; for $K > 0$ we work inside the injectivity radius, $R < \pi/(2\sqrt{K})$ and $D < \pi/(2\sqrt{K})$.

Ingredient (a): Hessian comparison for the squared distance. The Hessian of $\phi_i(p) = \frac{1}{2}d(p, y_i)^2$ at a point $p_i \neq y_i$ is a self-adjoint operator on $T_{p_i}\mathcal{M}_N$. In the orthogonal decomposition of $T_{p_i}\mathcal{M}_N$ into the radial direction (the unit vector $-\log_{p_i}(y_i)/d(p_i, y_i)$) and the transverse directions (its orthogonal complement), the Hessian has the following eigenvalues (see, e.g., Petersen, 2006, Ch. 4) or Bridson and Haefliger (2013, Ch. II.1):

- *Radial eigenvalue:* always equal to 1, for any K.
- *Transverse eigenvalues:* each equal to

$$f_K(D_i) \;:=\; D_i\,\frac{\mathrm{cn}_K(D_i)}{\mathrm{sn}_K(D_i)}, \qquad D_i := d(p_i, y_i),$$

  where sn_K and cn_K are the model sine and cosine from (4).

On a Hadamard manifold ($K \leq 0$), the model sine satisfies $\mathrm{sn}_K(r) \leq r\,\mathrm{cn}_K(r)$ for all $r \geq 0$ (since $\sinh(t)/t \leq \cosh(t)$ for all $t \geq 0$ when $K < 0$, and equality holds when $K = 0$), so

$$f_K(D_i) = D_i\,\frac{\mathrm{cn}_K(D_i)}{\mathrm{sn}_K(D_i)} \;\geq\; 1 \qquad \text{for all } K \leq 0 \text{ and } D_i \geq 0,$$

and $\nabla^2\phi_i\big|_{p_i} \succeq I_{T_{p_i}\mathcal{M}_N}$ as a bilinear form. For $K > 0$ with $\sqrt{K}D_i < \pi/2$ one has instead $f_K(D_i) = \sqrt{K}D_i\cot(\sqrt{K}D_i) \in (0, 1)$. In both cases the smallest eigenvalue of $\nabla^2\phi_i\big|_{p_i}$ is $\min(1, f_K(D_i)) \geq \min(1, f_K(D))$, since f_K is non-increasing. Thus

$$\begin{aligned} \text{(I)} = \left\langle \nabla^2\phi_i\big|_{p_i}(w_i),\, w_i \right\rangle \;\geq\; \min(1, f_K(D))\,\|w_i\|^2, \\ \text{where } w_i := D\exp_{o_N}(v_i)\cdot\Delta\xi_i. \end{aligned} \tag{22}$$

Ingredient (b): radial Jacobi field lower bound. The differential $D\exp_{o_N}(v_i)$ maps $T_0V_N \simeq \mathbb{R}^{d_N}$ to $T_{p_i}\mathcal{M}_N$. Its singular values are determined by the Jacobi fields along the geodesic $\gamma(t) = \exp_{o_N}(tv_i/\|v_i\|)$, $0 \leq t \leq \|v_i\|$. In the orthogonal decomposition of $\mathbb{R}^{d_N}$ into the direction of v_i (radial) and its complement (transverse):

- The *radial singular value* is always 1 (the exponential map is a radial isometry by the Gauss lemma).
- Each *transverse singular value* equals $S_K(\|v_i\|) = \mathrm{sn}_K(\|v_i\|)/\|v_i\|$, the Rauch factor.

For $K \le 0$ we have $S_K(r) \ge 1$ for all $r \ge 0$ (since $\sinh(t)/t \ge 1$ when $K < 0$, and $= 1$ when $K = 0$), so $\sigma_{\min}(D\exp_{o_N}(v_i)) = 1$; for $K > 0$ inside the injectivity radius $S_K(r) \in (0, 1)$ and $\sigma_{\min}(D\exp_{o_N}(v_i)) = S_K(\|v_i\|) \ge S_K(R)$. In both cases $\sigma_{\min}(D\exp_{o_N}(v_i)) \ge \min(1, S_K(R))$, and hence

$$\begin{aligned} \|w_i\| &= \|D\exp_{o_N}(v_i) \cdot \Delta\xi_i\| \\ &\ge \sigma_{\min}\big(D\exp_{o_N}(v_i)\big)\,\|\Delta\xi_i\| \\ &\ge \min(1, S_K(R))\,\|\Delta\xi_i\|. \end{aligned} \tag{23}$$

Combining (22) and (23), and noting that f_K and S_K are simultaneously ≥ 1 (for $K \le 0$) or ≤ 1 (for $K > 0$), so that $\min(1, f_K(D))\min(1, S_K(R))^2 = \min(1, S_K(R)^2 f_K(D)) =: m_K$:

$$\text{(I)} \;\ge\; m_K\,\|\Delta\xi_i\|^2. \tag{24}$$

Step 3: Lower bound on Term (II). We have $\nabla_p\phi_i\big|_{p_i} = -\log_{p_i}(y_i)$, so $\|\nabla_p\phi_i\big|_{p_i}\| = d(p_i, y_i) \le D$ by the tube assumption $A \in \mathcal{T}_{R,D}$. By the definition of $B_K(R)$ (Section A.4), $\|D^2\exp_{o_N}(v_i)\|_{\mathrm{op}} \le B_K(R)$ for $\|v_i\| \le R$. Therefore, by Cauchy–Schwarz:

$$\begin{aligned} \text{(II)} &= \big\langle \nabla_p\phi_i\big|_{p_i},\, D^2\exp_{o_N}(v_i)[\Delta\xi_i, \Delta\xi_i]\big\rangle \\ &\ge -\big\|\nabla_p\phi_i\big|_{p_i}\big\|\,\|D^2\exp_{o_N}(v_i)\|_{\mathrm{op}}\,\|\Delta\xi_i\|^2 \\ &\ge -D\,B_K(R)\,\|\Delta\xi_i\|^2. \end{aligned} \tag{25}$$

This term can be negative, because $D^2\exp$ can curve the image of the perturbation toward or away from y_i, so the inner product with $\nabla\phi_i$ need not have a definite sign. We use Cauchy–Schwarz to obtain the worst-case lower bound.

Step 4: Combining and using whitened data. Summing (24) and (25) over i and averaging:

$$\begin{aligned} D^2\mathcal{F}(A)[\Delta,\Delta] &= \frac{1}{m}\sum_{i=1}^m D^2 f_i(A)[\Delta,\Delta] \\ &\ge \frac{1}{m}\sum_{i=1}^m \big[m_K - D\,B_K(R)\big]\,\|\Delta\xi_i\|^2 \\ &= \mu_K\,\frac{1}{m}\sum_{i=1}^m \|\Delta\xi_i\|^2, \end{aligned} \tag{26}$$

where $\mu_K = m_K - D\,B_K(R)$ as in Lemma 6.

Now we use the whitening assumption $\frac{1}{m}\sum_{i=1}^m \xi_i\xi_i^* = I_{d_0}$. For any $\Delta \in \mathbb{R}^{d_N \times d_0}$:

$$\begin{aligned} \frac{1}{m}\sum_{i=1}^m \|\Delta\xi_i\|^2 &= \frac{1}{m}\sum_{i=1}^m \mathrm{tr}\big(\Delta\xi_i\xi_i^*\Delta^*\big) \\ &= \mathrm{tr}\Big(\Delta\,\underbrace{\Big(\tfrac{1}{m}\sum_i \xi_i\xi_i^*\Big)}_{=I_{d_0}}\Delta^*\Big) \\ &= \mathrm{tr}(\Delta\Delta^*) \;=\; \|\Delta\|_{\mathrm{HS}}^2. \end{aligned} \tag{27}$$

Substituting (27) into (26):

$$D^2\mathcal{F}(A)[\Delta,\Delta] \;\geq\; \mu_K\,\|\Delta\|_{\mathrm{HS}}^2, \tag{28}$$

with $\mu_K > 0$ by hypothesis; for $K \leq 0$ we have $m_K = 1$ and $\mu_K = 1 - D\,B_K(R)$.

Step 5: The Euclidean case $K = 0$. When $K = 0$, the manifold $\mathcal{M}_N$ is Euclidean, so $\exp_{o_N}(v) = o_N + v$ (a translation). Then:

- $D\exp_{o_N}(v) = I$ for all v, so $D^2\exp_{o_N} \equiv 0$, giving $B_0(R) = 0$.
- $\phi_i(p) = \frac{1}{2}\|p - y_i\|^2$ has Hessian $\nabla^2\phi_i = I$, the identity.

Therefore Term (II) vanishes identically, Term (I) equals $\|\Delta\xi_i\|^2$ exactly ($m_0 = 1$), and $\mu_0 = 1 - D \cdot 0 = 1$. In particular, $\mathcal{F}$ is 1-strongly convex (with respect to the Hilbert–Schmidt norm on A) in the Euclidean case, consistent with the fact that $\mathcal{F}$ reduces to $\frac{1}{2}\|A - \Phi\|_{\mathrm{HS}}^2 + \text{const}$ for whitened data when $K = 0$. □

Remark (role of each hypothesis). The three hypotheses play distinct roles:

1. $K \leq 0$ (Hadamard manifold) gives $m_K = 1$, since it ensures both $\nabla^2\phi_i \succeq I$ (Step 2, Ingredient (a)) and $\sigma_{\min}(D\exp) = 1$ (Step 2, Ingredient (b)). For $K > 0$ both factors drop below 1 and $m_K = S_K(R)^2 f_K(D)$; beyond the injectivity radius f_K turns negative and the Jacobi field can vanish, so the argument does not extend past it.
2. Whitened data ($\frac{1}{m}\sum \xi_i\xi_i^* = I$) converts the sample average $\frac{1}{m}\sum\|\Delta\xi_i\|^2$ into the squared Hilbert–Schmidt norm $\|\Delta\|_{\mathrm{HS}}^2$ in Step 4. Without whitening, the strong-convexity constant would additionally depend on the smallest eigenvalue of the data covariance.
3. $D\,B_K(R) < 1$ ensures $\mu_K > 0$, so that the negative Term (II) from the second derivative of exp does not overwhelm the positive Term (I). As $|K| \to 0$, $B_K(R) \to 0$ and $\mu_K \to 1$; as $|K|$ grows (with $K < 0$), $B_K(R)$ increases, μ_K decreases, and the strong convexity weakens.

D.3 Proof of Gradient Dominance

Proof of Lemma 7. We derive the gradient dominance inequality from the μ_K-strong convexity of $\mathcal{F}$ established in Lemma 6. The argument is the standard reduction of a Polyak–Łojasiewicz inequality from strong convexity (see, e.g., Nesterov et al., 2018, Theorem 2.1.10), given in full here to track the curvature parameter μ_K.

Step 1: The strong-convexity lower bound. Lemma 6 gives $D^2\mathcal{F}(A)[\Delta,\Delta] \geq \mu_K\|\Delta\|_{\mathrm{HS}}^2$ for every $A \in \mathcal{T}_{R,D}$ and every direction Δ. The bound holds at every point of $\mathcal{T}_{R,D}$. The radius constraint $\|A\xi_i\| \leq R$ is convex in A, but the residual constraint $d(\exp(A\xi_i), y_i) \leq D$ is not, so we do not assume that $\mathcal{T}_{R,D}$ is convex and integrate only along segments contained in it; this is exactly hypothesis (4) of Theorem 3, and it is supplied along the trajectory by Lemma 8(a). Integrating $D^2\mathcal{F}$ along a segment from A to $A' \in \mathcal{T}_{R,D}$ with $[A, A'] \subset \mathcal{T}_{R,D}$ yields the standard strong-convexity inequality:

$$\mathcal{F}(A') \;\geq\; \mathcal{F}(A) \;+\; \big\langle \nabla_A\mathcal{F}(A),\; A' - A\big\rangle \;+\; \frac{\mu_K}{2}\,\|A' - A\|_{\mathrm{HS}}^2. \tag{29}$$

To verify (29) explicitly: define $g(t) := \mathcal{F}(A + t(A' - A))$ for $t \in [0,1]$. Then $g(0) = \mathcal{F}(A)$, $g(1) = \mathcal{F}(A')$, $g'(0) = \langle\nabla_A\mathcal{F}(A), A' - A\rangle$, and $g''(t) = D^2\mathcal{F}(A + t(A' - A))[A' - A, A' - A] \geq \mu_K\|A' - A\|_{\mathrm{HS}}^2$

for all $t \in [0,1]$ by Lemma 6. Integrating twice:

$$\begin{aligned} g(1) &= g(0) + g'(0) + \int_0^1 (1-t)\, g''(t)\, dt \\ &\geq g(0) + g'(0) + \frac{\mu_K}{2}\|A' - A\|_{\mathrm{HS}}^2, \end{aligned}$$

which is (29).

Step 2: Minimize the right-hand side over A'. Write $G := \nabla_A \mathcal{F}(A)$ and set $h := A' - A$. Inequality (29) reads

$$\mathcal{F}(A') \;\geq\; \mathcal{F}(A) + \langle G,\, h\rangle + \frac{\mu_K}{2}\|h\|_{\mathrm{HS}}^2 \;=:\; q(h).$$

The right-hand side $q(h)$ is a quadratic in h with positive definite leading term $\frac{\mu_K}{2}\|\cdot\|_{\mathrm{HS}}^2$. Its (unconstrained) minimizer is found by setting the gradient to zero:

$$\nabla_h q(h) = G + \mu_K\, h = 0 \qquad \Longrightarrow \qquad h^\star = -\frac{1}{\mu_K}\, G.$$

Substituting $h^\star$ back into q:

$$\begin{aligned} q(h^\star) &= \mathcal{F}(A) + \Big\langle G,\; -\frac{1}{\mu_K} G \Big\rangle + \frac{\mu_K}{2}\Big\| -\frac{1}{\mu_K} G \Big\|_{\mathrm{HS}}^2 \\ &= \mathcal{F}(A) - \frac{1}{\mu_K}\|G\|_{\mathrm{HS}}^2 + \frac{1}{2\mu_K}\|G\|_{\mathrm{HS}}^2 \\ &= \mathcal{F}(A) - \frac{1}{2\mu_K}\|G\|_{\mathrm{HS}}^2. \end{aligned} \tag{30}$$

Step 3: Apply the bound at the global minimizer. Let $A^\star$ be a global minimizer of $\mathcal{F}$ on the tube (which exists by compactness of $\mathcal{T}_{R,D}$ and continuity of $\mathcal{F}$), with $\mathcal{F}^\star := \mathcal{F}(A^\star)$, and assume $[A, A^\star] \subset \mathcal{T}_{R,D}$ as in the statement of Lemma 7. Then (29) applies with $A' = A^\star$. Moreover, the unconstrained minimum $q(h^\star)$ is a lower bound on $q(h)$ for any h, so

$$\mathcal{F}^\star = \mathcal{F}(A^\star) \;\geq\; \inf_h q(h) \;=\; q(h^\star) \;=\; \mathcal{F}(A) - \frac{1}{2\mu_K}\|G\|_{\mathrm{HS}}^2.$$

(Strictly speaking, the infimum is over all $h \in \mathbb{R}^{d_N \times d_0}$, not only those with $A + h \in \mathcal{T}_{R,D}$, so the lower bound on $\mathcal{F}(A^\star)$ is weaker than necessary and the inequality only becomes tighter.)

Step 4: Rearrange. Moving $\mathcal{F}(A)$ to the left and $\mathcal{F}^\star$ to the right:

$$\frac{1}{2\mu_K}\|G\|_{\mathrm{HS}}^2 \;\geq\; \mathcal{F}(A) - \mathcal{F}^\star,$$

and multiplying both sides by $2\mu_K > 0$ (which holds since $D\, B_K(R) < 1$ by hypothesis):

$$\|\nabla_A \mathcal{F}(A)\|_{\mathrm{HS}}^2 \;\geq\; 2\mu_K\big(\mathcal{F}(A) - \mathcal{F}^\star\big). \tag{31}$$

This is the gradient dominance (Polyak–Łojasiewicz) inequality.

Step 5: The Euclidean case $K = 0$. When $K = 0$, Lemma 6 gives $\mu_0 = 1$, and (31) becomes $\|\nabla_A \mathcal{F}(A)\|_{\mathrm{HS}}^2 \geq 2(\mathcal{F}(A) - \mathcal{F}^\star)$. Since $\mathcal{F}$ reduces to $\frac{1}{2}\|A - \Phi\|_{\mathrm{HS}}^2 + \text{const}$ for whitened data when $K = 0$, one can verify this directly: $\nabla_A \mathcal{F} = A - \Phi$, so $\|\nabla_A \mathcal{F}\|^2 = \|A - \Phi\|^2 = 2(\mathcal{F}(A) - \mathcal{F}^\star)$. $\square$

Remark (role in the convergence proof). The gradient dominance inequality (31) is the bridge between the single-step descent of Theorem 3 and the end-to-end linear convergence of Theorem 9. Theorem 3 guarantees $\mathcal{F}(A^+) - \mathcal{F}(A) \le -\frac{\eta}{2}\sigma_{\min}(A)^{2(N-1)/N}\|G\|_{\mathrm{HS}}^2$. Substituting (31) converts this into a contraction on the suboptimality gap:

$$\mathcal{F}(A^+) - \mathcal{F}^\star \;\le\; \Big(1 - \eta\,\mu_K\,\sigma_{\min}(A)^{2(N-1)/N}\Big)\big(\mathcal{F}(A) - \mathcal{F}^\star\big),$$

which iterates to (10). The curvature dependence of $\mu_K = 1 - D\,B_K(R)$ thus enters the contraction factor, since larger $|K|$ (with $K < 0$) increases $B_K(R)$, decreases μ_K, and slows the geometric rate.

D.4 Proof of Trajectory Containment

Proof of Lemma 8. The proof proceeds by induction on the iteration counter t. At each step we verify that the hypotheses of Theorem 3 are maintained, so the descent lemma can be applied again at the next step. The three conclusions (a)–(c) are established in order; each builds on the previous ones.

Throughout, let $G(t) := \nabla_A \mathcal{F}(A(t))$, $M := (4\|\Phi\|_F)^{1/N}$ (the spectral-norm bound on each factor), and $\Phi := \frac{1}{m}\sum_i \tilde{y}_i \xi_i^*$ the tangent-space target matrix. Whitening gives $2\mathcal{E}(A) = \|A - \Phi\|_{\mathrm{HS}}^2 + 2\mathcal{E}^\star$; it does not require exact realizability.

Choice of constants δ_0 and η_0. Both are chosen after Φ and c' are fixed. Set

$$\delta_0 := \frac{(c')^2}{256\,N^3\,\|\Phi\|_F^{2(N-1)/N}}, \tag{32}$$

as in Theorem 1 of Arora et al. (2018a) (with c' playing the role of their deficiency margin c), and

$$\eta_0 := \frac{(c')^{(4N-2)/N}}{C'_N\,N^3\,\|\Phi\|_F^{(6N-4)/N}} \tag{33}$$

for a sufficiently large constant $C'_N > 0$ depending only on N. The precise value of C'_N is determined by requiring that this algebraic restriction η_{alg} simultaneously satisfies the four conditions in equation (27) of Appendix D.2 of Arora et al. (2018a) (with their $c = c'$, $M = (4\|\Phi\|_F)^{1/N}$, $B = G_{K,R,\bar{D}}$); the curvature-dependent terms enter only through $\eta_K^\star$, which is handled by the separate requirement $\eta \le \eta_K^\star$. Equation (33) defines only η_{alg}; the constant used throughout is $\eta_0 := \min(\eta_{\mathrm{alg}}, \eta_{\mathrm{seg}})$, where the segment restriction η_{seg} is fixed in Part (a) so that every update segment stays in the tube $\mathcal{T}_{R,\bar{D}}$.

Base case ($t = 0$).

(a) The endpoint bound gives $\max_i d(\exp_{o_N}(A(0)\xi_i), y_i) \le D(0) \le \bar{D}$, so $A(0) \in \mathcal{T}_{R,\bar{D}}$ and the tube condition holds at $t = 0$.

(b) The initialization is δ_0-balanced, so it is also $2\delta_0$-balanced.

(c) Since $A(0)$ has deficiency margin $c' > 0$ with respect to Φ (Definition 2), Claim 1 of Arora et al. (2018a) gives $\sigma_{\min}(A(0)) \ge c' > 0$, which is (c) at $t = 0$.

Inductive step: assume (a)–(c) hold for $0, 1, \ldots, t$; prove them at $t + 1$.

Part (c): minimum singular value. This is the deficiency-margin argument of Arora et al. (2018a) (Lemma 2, property (B) $\Rightarrow$ Claim 1). We give the full details.

Step 1: the loss is non-increasing. By the inductive hypothesis, (a)–(c) hold at every $s \in \{0, \dots, t\}$, so the hypotheses of Theorem 3 are satisfied at each step s. Applying the descent lemma (7) iteratively:

$$\begin{aligned}\mathcal{F}(A(s+1)) &\le \mathcal{F}(A(s)) \\ &\quad - \frac{\eta}{2}\,\sigma_{\min}(A(s))^{2(N-1)/N}\|G(s)\|_{\mathrm{HS}}^2 \\ &\le \mathcal{F}(A(s)),\end{aligned} \tag{34}$$

so $s \mapsto \mathcal{F}(A(s))$ is monotone non-increasing on $\{0, \dots, t+1\}$.

Step 2: the distance to Φ does not grow. Write $\tilde{y}_i = \log_{\mathbf{0}}^K(y_i)$, so that $y_i = \exp_{o_N}(\tilde{y}_i)$. For $K \le 0$ the exponential map at o_N is expanding, $d(\exp_{o_N}(u), \exp_{o_N}(v)) \ge \|u - v\|$, and by Rauch comparison it is $S_K(R)$-Lipschitz on the ball of radius R (that ball is convex, so integrate $\|D\exp\|_{\mathrm{op}} \le S_K(R)$ along the segment from u to v). Applying the expanding inequality to the pair $u = A\xi_i$, $v = \tilde{y}_i$ and then the whitening identity $2\mathcal{E}(A) = \|A - \Phi\|_{\mathrm{HS}}^2 + 2\mathcal{E}^\star$ (Appendix D.7, with $\mathcal{E}^\star = \min_W \mathcal{E} \ge 0$) gives

$$\begin{aligned}2\,\mathcal{F}(A) &= \frac{1}{m}\sum_i d\big(\exp_{o_N}(A\xi_i), \exp_{o_N}(\tilde{y}_i)\big)^2 \;\ge\; \frac{1}{m}\sum_i \|A\xi_i - \tilde{y}_i\|^2 \\ &= 2\,\mathcal{E}(A) = \|A - \Phi\|_{\mathrm{HS}}^2 + 2\mathcal{E}^\star \;\ge\; \|A - \Phi\|_{\mathrm{HS}}^2.\end{aligned}$$

No realizability is used: the inequality goes through the surrogate $\mathcal{E}$ and the nonnegative constant $\mathcal{E}^\star$, not through any assumption that $\tilde{y}_i = \Phi\xi_i$. Applying instead the $S_K(R)$-Lipschitz upper bound at $A(0)$ gives $2\,\mathcal{F}(A(0)) \le S_K(R)^2 \cdot 2\,\mathcal{E}(W(0))$. Combining these with the monotonicity of Step 1,

$$\begin{aligned}\|A(t+1) - \Phi\|_{\mathrm{HS}} &\le \big(2\,\mathcal{F}(A(t+1))\big)^{1/2} \le \big(2\,\mathcal{F}(A(0))\big)^{1/2} \\ &\le S_K(R)\,\big(2\,\mathcal{E}(W(0))\big)^{1/2}.\end{aligned} \tag{35}$$

At $K = 0$ we have $S_0(R) = 1$ and $\mathcal{F}(A) = \frac{1}{2}\|A - \Phi\|_{\mathrm{HS}}^2 + \mathcal{E}^\star$, or equivalently $\mathcal{F}(A) - \mathcal{F}^\star = \frac{1}{2}\|A - \Phi\|_{\mathrm{HS}}^2$. For $K < 0$ the factor $S_K(R) > 1$ is the price of measuring error on the manifold, and it is absorbed by imposing the deficiency margin in its curvature-corrected form, $S_K(R)(2\mathcal{E}(W(0)))^{1/2} \le \sigma_{\min}(\Phi) - c'$ (Appendix A.3). The correction cannot be dispensed with: monotonicity of $\mathcal{F}$ alone bounds $\|A(t) - \Phi\|$ by $(2\mathcal{F}(A(0)))^{1/2}$, which for $K < 0$ exceeds $\|A(0) - \Phi\|$.

Step 3: apply the deficiency-margin characterization. The initialization satisfies the curvature-corrected loss-based condition of Appendix A.3: $S_K(R)(2\mathcal{E}(W(0)))^{1/2} \le \sigma_{\min}(\Phi) - c'$. By (35), $\|A(t+1) - \Phi\|_{\mathrm{HS}} \le S_K(R)(2\mathcal{E}(W(0)))^{1/2} \le \sigma_{\min}(\Phi) - c'$. Claim 1 of Arora et al. (2018a) (proved using $\sigma_{\min}(X + Y) \ge \sigma_{\min}(X) - \sigma_{\max}(Y)$ and $\sigma_{\max}(Y) \le \|Y\|_F$) then gives:

$$\begin{aligned}\sigma_{\min}(A(t+1)) &= \sigma_{\min}\big(\Phi + (A(t+1) - \Phi)\big) \\ &\ge \sigma_{\min}(\Phi) - \sigma_{\max}(A(t+1) - \Phi) \\ &\ge \sigma_{\min}(\Phi) - \|A(t+1) - \Phi\|_{\mathrm{HS}} \\ &\ge \sigma_{\min}(\Phi) - (\sigma_{\min}(\Phi) - c') \;=\; c'.\end{aligned} \tag{36}$$

This establishes part (c) at step $t+1$, with the constant c' itself. We do not claim $\sigma_{\min}(A(t)) \ge \sigma_{\min}(A(0))/2$: the deficiency margin bounds $\sigma_{\min}(A(0))$ below by c' but not above by $2c'$, so c' is the constant to carry forward, exactly as in Claim 1 of Arora et al. (2018a).

Part (b): balancedness preservation. This part is algebraic in the weight matrices and follows identically to the proof of property (A') in Lemma 2 of Arora et al. (2018a). We reproduce the argument.

Step 1: single-step change in the balancedness defect. Define $\beta_j(t) := \|W_{j+1}(t)^\top W_{j+1}(t) - W_j(t)W_j(t)^\top\|_F$. The gradient step gives $W_j^+ = W_j - \eta\, W_{j+1:N}^* G W_{1:j-1}^*$, and expanding $W_{j+1}^{+\top} W_{j+1}^+ - W_j^+ W_j^{+\top}$ (a product of two updated factors minus a product of two updated factors) yields, after subtracting and adding the pre-update quantities:

$$\begin{aligned}\beta_j(t+1) \le \beta_j(t) \\ + 4\eta^2 \|G(t)\|_{\mathrm{HS}} \|G(t)\|_{\mathrm{op}} (2\|\Phi\|_F)^{2(N-1)/N}.\end{aligned} \tag{37}$$

The factor $4(2\|\Phi\|_F)^{2(N-1)/N}$ comes from bounding $\|W_{j+1:N}\|_{\mathrm{op}}^2 \|W_{1:j}\|_{\mathrm{op}}^2 + \|W_{1:j-1}\|_{\mathrm{op}}^2 \|W_{j+1:N}\|_{\mathrm{op}}^2$ using part (c) of the inductive hypothesis and the spectral-norm bound $\|W_j(t)\|_{\mathrm{op}} \le M = (4\|\Phi\|_F)^{1/N}$ (established below in the paragraph on operator-norm control).

No curvature term appears in (37). The cross-terms are products of weight matrices and the gradient G, whose magnitude depends on curvature through $\mathcal{F}$ but which enters only through its norms, and the algebraic identity is the same as in the Euclidean case.

Step 2: telescoping. Summing (37) from $s = 0$ to $s = t$:

$$\begin{aligned}\beta_j(t+1) \le \beta_j(0) \\ + 4(2\|\Phi\|_F)^{2(N-1)/N} \eta^2 \sum_{s=0}^{t} \|G(s)\|_{\mathrm{HS}} \|G(s)\|_{\mathrm{op}}.\end{aligned} \tag{38}$$

To bound the sum, use the descent inequality (34) summed over $s = 0, \ldots, t$:

$$\begin{aligned}\frac{\eta}{2}\,(c')^{2(N-1)/N} \sum_{s=0}^{t} \|G(s)\|_{\mathrm{HS}}^2 &\le \mathcal{F}(A(0)) - \mathcal{F}(A(t+1)) \\ &\le \mathcal{F}(A(0)) \;\le\; \tfrac{1}{2}\|\Phi\|_F^2,\end{aligned}$$

where we used $\sigma_{\min}(A(s)) \ge c'$ (part (c)). The final bound follows from $2\mathcal{F}(A(0)) \le S_K(R)^2\, 2\mathcal{E}(W(0))$ and the curvature-corrected loss-based condition $S_K(R)(2\mathcal{E}(W(0)))^{1/2} \le \sigma_{\min}(\Phi) - c' \le \|\Phi\|_F$. Since $\|G(s)\|_{\mathrm{op}} \le \|G(s)\|_{\mathrm{HS}}$, we get

$$\begin{aligned}\eta^2 \sum_{s=0}^{t} \|G(s)\|_{\mathrm{HS}} \|G(s)\|_{\mathrm{op}} &\le \eta^2 \sum_{s=0}^{t} \|G(s)\|_{\mathrm{HS}}^2 \\ &\le \frac{\eta\, \|\Phi\|_F^2}{(c')^{2(N-1)/N}}.\end{aligned}$$

Substituting into (38):

$$\beta_j(t+1) \;\le\; \delta_0 \;+\; \frac{4(2\|\Phi\|_F)^{2(N-1)/N}\, \eta\, \|\Phi\|_F^2}{(c')^{2(N-1)/N}}.$$

Step 3: the choice of η_0 absorbs the increment. The choice (33) ensures that the second term is at most δ_0 (this is a direct calculation matching the constants in (32) and (33); see equation (37) in Appendix D.2 of Arora et al. (2018a)). Therefore

$$\beta_j(t+1) \;\le\; 2\delta_0, \qquad j = 1, \ldots, N-1,$$

which is part (b) at step $t+1$.

Hypothesis (3) of Theorem 3. Part (b) is what makes the coercivity hypothesis available at every step. By the trace identity of Step 5 of Appendix D.1, $\langle G, \mathcal{P}_A(G)\rangle = \sum_j \|W^*_{j+1:N} G W^*_{1:j-1}\|^2_{\mathrm{HS}}$ exactly, and the same approximate-balancedness estimate that yields the operator-norm control below bounds the deviation of $\mathcal{P}_A$ from its exactly balanced counterpart, the one diagonalized in Appendix D.1, by $2\delta_0 N^3 M^{2(N-2)}$ in operator norm. Hence

$$\sum_j \|W^*_{j+1:N} G W^*_{1:j-1}\|^2_{\mathrm{HS}} \geq \big(\sigma_{\min}(A)^{2(N-1)/N} - 2\delta_0 N^3 M^{2(N-2)}\big)\|G\|^2_{\mathrm{HS}},$$

so hypothesis (3) holds with the constant $\frac{3}{4}$ as soon as $2\delta_0 \leq (c')^{2(N-1)/N}/(4N^3M^{2(N-2)})$, which is implied by the choice (32) of δ_0 together with $c' \leq \|\Phi\|_F$.

Operator-norm control (auxiliary). Once $2\delta_0$-balancedness and the spectral-norm bound $\|A(t+1)\|_{\mathrm{op}} \leq 2\|\Phi\|_F$ (from part (c) and the triangle inequality) are established, Lemma 6 of Arora et al. (2018a) gives the layerwise bound $\|W_j(t+1)\|_{\mathrm{op}} \leq (4\|\Phi\|_F)^{1/N} = M$, confirming hypothesis (1) of Theorem 3 at step $t+1$. The argument uses Lidskii's theorem (Lemma 3 of Arora et al. (2018a)) to pass from approximate balancedness of the Gram matrices $W_j W_j^\top$ to approximate equality of singular values, and then a contradiction argument to show no individual $\|W_j\|_{\mathrm{op}}$ can exceed $(4\|\Phi\|_F)^{1/N}$.

Part (a): tube containment. We must verify $A(t+1) \in \mathcal{T}_{R,\bar{D}}$; the endpoints obey the sharper bound $\max_i d(\exp_{o_N}(A(t+1)\xi_i), y_i) \leq D(0) \leq \bar{D}$, and the segment bound is treated below.

Residual bound. For every i and every t,

$$\begin{aligned} d(\exp_{o_N}(A(t)\xi_i), y_i)^2 &\leq \sum_{i'=1}^{m} d(\exp_{o_N}(A(t)\xi_{i'}), y_{i'})^2 \\ &= 2m\,\mathcal{F}(A(t)), \end{aligned}$$

so by the monotonicity established in Step 1 of part (c),

$$\begin{aligned} \max_i d(\exp_{o_N}(A(t)\xi_i), y_i) &\leq \big(2m\,\mathcal{F}(A(0))\big)^{1/2} \\ &=: D(0) \qquad \text{for all } t \geq 0. \end{aligned}$$

This is the *endpoint* residual bound, obtained from the monotone decrease of $\mathcal{F}$ alone with no continuity or small-step argument; the price is a factor $\sqrt{m}$ relative to the initial maximum residual. The tube radius carried through the induction is the slightly larger $\bar{D} = D(0) + \delta_D$ (the segment step below needs the extra slack δ_D), and it is this $\bar{D}$ that has to satisfy $\bar{D}\, B_K(R) < 1$ in Theorem 9.

Radius bound $\|A(t+1)\xi_i\| \leq R$. We have $\|A(t+1)\xi_i\| \leq \|A(t+1)\|_{\mathrm{op}}\|\xi_i\|$. By the operator-norm control established above, $\|A(t+1)\|_{\mathrm{op}} \leq 2\|\Phi\|_F$. By the triangle inequality:

$$\begin{aligned} \|A(t+1)\|_{\mathrm{op}} &\leq \|\Phi\|_{\mathrm{op}} + \|A(t+1) - \Phi\|_{\mathrm{op}} \\ &\leq \|\Phi\|_F + \|A(t+1) - \Phi\|_F \leq 2\|\Phi\|_F. \end{aligned}$$

The radius of Lemma 8(a), $R := \max_i \max(2\|\Phi\|_F\|\xi_i\|, \|\tilde{y}_i\|)$, therefore satisfies $\|A(t+1)\xi_i\| \leq R$ for all i, and it also contains the targets $\tilde{y}_i$, which is what the Rauch comparison of Step 2 requires. It depends only on Φ and the data, so fixing it involves no circularity with $S_K(R)$.

Segment containment. Hypothesis (4) of Theorem 3 requires $[A(t), A(t+1)] \subset \mathcal{T}_{R,\bar{D}}$, where $\bar{D} := D(0) + \delta_D$ is the tube radius fixed in the statement of Lemma 8, with $\delta_D > 0$ a fixed slack chosen

before η_0 and independent of t. The radius constraint is convex in A, so the segment satisfies it. The residual constraint is not convex, and we verify it directly: $A \mapsto d(\exp(A\xi_i), y_i)$ is $S_K(R)\|\xi_i\|$-Lipschitz and $\|A(t+1) - A(t)\|_{\mathrm{op}} \le \eta N M^{N-1} \|G(t)\|_{\mathrm{HS}}$ (from the expansion in Step 4 of Appendix D.1), so every point of $[A(t), A(t+1)]$ has residual at most $D(0) + C\,\eta M^{N-1} S_K(R) G_{K,R,\bar{D}}$ for an absolute constant C. Because δ_D is fixed in advance, we take $\eta_{\mathrm{seg}} := \delta_D / (C M^{N-1} S_K(R) G_{K,R,\bar{D}})$, so that for $\eta \le \eta_{\mathrm{seg}}$ (hence for $\eta \le \eta_0 = \min(\eta_{\mathrm{alg}}, \eta_{\mathrm{seg}})$) the whole segment has residual at most $\bar{D}$, so $[A(t), A(t+1)] \subset \mathcal{T}_{R,\bar{D}}$ for every $\eta \le \eta_0$. The order of quantification is what removes the circularity: δ_D (hence $\bar{D}$ and every tube constant $L_{K,R,\bar{D}}, G_{K,R,\bar{D}}, \mu_K = 1 - \bar{D}\,B_K(R)$) is fixed first, and only afterwards is η_0 made small enough to keep the update inside the pre-committed tube. The same containment supplies the segment hypothesis of Lemmas 6 and 7 along the trajectory. Since δ_D may be taken as small as desired, the constants at $\bar{D}$ agree with those at $D(0)$ up to a term $O(\delta_D)$, and the displayed $O(\cdot)$ bounds below, written at $D(0)$, are unchanged.

This completes part (a) and the inductive step.

Summary of what is geometric and what is algebraic. Part (b) is algebraic in the weight matrices, using only the factored structure $\nabla_{W_j} \mathcal{L} = W_{j+1:N}^* G W_{1:j-1}^*$ and spectral-norm bounds, and is identical to the argument in Arora et al. (2018a). Curvature enters part (c) through the factor $S_K(R)$ in (35), which is why the loss-based initialization condition is curvature corrected; it also enters the segment estimate in part (a) through the Lipschitz factor $S_K(R)$ and the fixed tube constants. The curvature-dependent step-size ceiling $\eta_K^\star$ from (9) handles this requirement. It ensures that the descent lemma applies, the descent lemma ensures that $\mathcal{F}$ is non-increasing, and the monotone decrease ensures tube containment. The requirement $\eta \le \min(\eta_K^\star, \eta_0)$ thus separates the curvature cost ($\eta_K^\star$) from the algebraic cost (η_0). □

D.5 Proof of Linear Convergence

Proof of Theorem 9. The argument combines the three preceding results, namely the descent lemma (Theorem 3), gradient dominance (Lemma 7), and trajectory containment (Lemma 8), into a single induction that yields a geometric contraction on the sub-optimality gap at every step.

Step 0: Verifying the hypotheses persist.

The hypotheses of the theorem are:

- $K \le 0$, the data are whitened ($\frac{1}{m}\sum \xi_i \xi_i^* = I$);
- the initialization is δ-balanced with deficiency margin $c' > 0$ (Definitions 1 and 2);
- $A(0) \in \mathcal{T}_{R,\bar{D}}$ with $\bar{D}\,B_K(R) < 1$ for the fixed tube radius $\bar{D} = D(0) + \delta_D$;
- $\eta \le \eta_K^\star$, where $\eta_K^\star$ is given by (9).

Lemma 8 (with $\delta_0 = \delta$ and η_0 chosen as in its proof, so that $\eta \le \min(\eta_K^\star, \eta_0)$) guarantees that for *every* $t \ge 0$:

(a) $A(t) \in \mathcal{T}_{R,\bar{D}}$ for the tube radius $\bar{D}$ of Lemma 8(a), so the tube condition $\bar{D}\,B_K(R) < 1$ is maintained;

(b) the factors $W_1(t), \ldots, W_N(t)$ are 2δ-balanced, so hypothesis (3) of Theorem 3 (the coercivity condition) holds at every step;

(c) $\sigma_{\min}(A(t)) \ge c' > 0$, so hypothesis (2) of Theorem 3 holds.

Since the operator-norm bound $\|W_j(t)\|_{\mathrm{op}} \le M = (4\|\Phi\|_F)^{1/N}$ is also maintained (as shown in the proof of Lemma 8), hypothesis (1) of Theorem 3 holds. The segment containment hypothesis (4) is ensured by the small step size $\eta \le \eta_0$. Likewise, since $K \le 0$, whitened data, and $D(t)\,B_K(R) < 1$ all persist, the hypotheses of Lemma 6 and hence Lemma 7 are maintained at every step.

Conclusion of Step 0: Theorem 3, Lemma 7, and Lemma 8 may all be invoked at every iteration $t \ge 0$.

Step 1: Per-step contraction on the sub-optimality gap.

Fix an arbitrary $t \ge 0$. Write $G(t) := \nabla_A \mathcal{F}(A(t))$, $\sigma_{\min}^{(t)} := \sigma_{\min}(A(t))$, and $\Delta\mathcal{F}(t) := \mathcal{F}(A(t)) - \mathcal{F}^\star$.

Ingredient 1: Descent lemma. Theorem 3, equation (7), gives

$$\mathcal{F}(A(t+1)) - \mathcal{F}(A(t)) \;\le\; -\frac{\eta}{2}\,(\sigma_{\min}^{(t)})^{2(N-1)/N}\,\|G(t)\|_{\mathrm{HS}}^2. \tag{39}$$

Ingredient 2: Gradient dominance. Lemma 7 gives

$$\|G(t)\|_{\mathrm{HS}}^2 \;\ge\; 2\mu_K\,\Delta\mathcal{F}(t), \tag{40}$$

where $\mu_K = 1 - \bar{D}\,B_K(R) > 0$ is evaluated at the tube radius $\bar{D}$ of Lemma 8(a); the instantaneous constant $\mu_{K,t} = 1 - D(t)B_K(R)$ is no smaller (as $D(t) \le D(0) \le \bar{D}$), so the inequality only becomes tighter.

Ingredient 3: Trajectory containment. Lemma 8(c) gives

$$\sigma_{\min}^{(t)} \;\ge\; c'. \tag{41}$$

Combining. Substituting (40) and (41) into (39):

$$\begin{aligned}\mathcal{F}(A(t+1)) - \mathcal{F}(A(t)) &\le -\frac{\eta}{2}\,(c')^{2(N-1)/N}\cdot 2\mu_K\,\Delta\mathcal{F}(t)\\ &= -\eta\,\mu_K\,(c')^{2(N-1)/N}\,\Delta\mathcal{F}(t).\end{aligned} \tag{42}$$

Adding $\mathcal{F}^\star$ to both sides and rearranging:

$$\Delta\mathcal{F}(t+1) = \mathcal{F}(A(t+1)) - \mathcal{F}^\star \;\le\; \big(1 - \eta\,\mu_K\,(c')^{2(N-1)/N}\big)\,\Delta\mathcal{F}(t). \tag{43}$$

Define the contraction factor

$$\rho \;:=\; \eta\,\mu_K\,(c')^{2(N-1)/N}. \tag{44}$$

Then $\rho > 0$, because $\eta > 0$, $\mu_K > 0$ and $c' > 0$; and $\rho \le 1$, because otherwise (43) would force $\Delta\mathcal{F}(t+1) < 0$, contradicting the definition of $\mathcal{F}^\star$.

Step 2: Iterating the contraction.

Inequality (43) is a one-step geometric contraction. Applying it recursively from $t = 0$ to $t = T - 1$:

$$\begin{aligned}\Delta\mathcal{F}(T) &\le (1-\rho)\,\Delta\mathcal{F}(T-1)\\ &\le (1-\rho)^2\,\Delta\mathcal{F}(T-2)\\ &\;\;\vdots\\ &\le (1-\rho)^T\,\Delta\mathcal{F}(0).\end{aligned} \tag{45}$$

Substituting ρ from (44) into (45) gives (10):

$$\mathcal{F}(A(T)) - \mathcal{F}^\star \le \big(1 - \eta\,\mu_K\,(c')^{2(N-1)/N}\big)^T \times \big(\mathcal{F}(A(0)) - \mathcal{F}^\star\big).$$

Step 3: Deriving the iteration count (12).

We want $\Delta\mathcal{F}(T) \le \varepsilon$. By (45), it suffices to have

$$(1-\rho)^T\,\Delta\mathcal{F}(0) \;\le\; \varepsilon,$$

i.e.,

$$T\,\log(1-\rho) \;\le\; \log\frac{\varepsilon}{\Delta\mathcal{F}(0)} \;=\; -\log\frac{\Delta\mathcal{F}(0)}{\varepsilon}. \tag{46}$$

Since $\rho \in (0,1)$, we have $\log(1-\rho) < 0$ and the standard inequality $\log(1-\rho) \le -\rho$ (valid for all $\rho \in (0,1)$), so (46) is satisfied whenever

$$-T\rho \;\le\; -\log\frac{\Delta\mathcal{F}(0)}{\varepsilon}, \qquad \text{i.e.,} \qquad T \;\ge\; \frac{1}{\rho}\,\log\frac{\Delta\mathcal{F}(0)}{\varepsilon}.$$

Substituting ρ from (44):

$$T \;\ge\; \frac{1}{\eta\,\mu_K\,(c')^{2(N-1)/N}}\;\log\frac{\mathcal{F}(A(0)) - \mathcal{F}^\star}{\varepsilon}. \tag{47}$$

This is the exact bound (11), valid for any admissible step η; the contraction rate scales with the actual η, so no η-independent count holds without a lower bound on η. To obtain the η-independent form (12) we now assume $\eta = \Theta(\eta_K^\star)$ and substitute the curvature-dependent step-size ceiling. By (9):

$$\eta_K^\star \;=\; \frac{\sigma_{\min}(A)^{2(N-1)/N}}{4C_N(M^{3N-4}G_{K,R,D} + L_{K,R,D}M^{4N-4})}.$$

The dominant curvature-dependent term in the denominator is $L_{K,R,D}M^{4N-4}$. From the definition of $L_{K,R,D}$ (5):

$$L_{K,R,D} = \frac{1}{m}\sum_{i=1}^{m}\|\xi_i\|^2\big[\widehat{S}_K(R)^2\,H_K(D) + D\,B_K(R)\big].$$

For whitened data, $\frac{1}{m}\sum\|\xi_i\|^2 = \mathrm{tr}(I_{d_0}) = d_0$, a curvature-independent data constant. Since $K \le 0$ here, $\widehat{S}_K(R) = S_K(R)$ and the curvature-dependent part is $S_K(R)^2 H_K(D) + D\,B_K(R)$, where:

- $S_K(R) = \sinh(\sqrt{\kappa}\,R)/(\sqrt{\kappa}\,R) \ge 1$, and $S_K(R)^2$ dominates for $\sqrt{\kappa}\,R \gg 1$;
- $H_K(D) = \sqrt{\kappa}\,D\coth(\sqrt{\kappa}\,D) \ge 1$;
- $D\,B_K(R) \ge 0$.

Therefore $L_{K,R,D} = \Theta(S_K(R)^2)$ (up to factors depending on $H_K(D)$, $B_K(R)$, and data constants), giving

$$\eta_K^\star \;=\; O\Big(\frac{1}{S_K(R)^2}\Big)$$

(with the implicit constant depending on M, N, $G_{K,R,D}$ and data quantities, but *not* on K through any other channel).

Under $\eta = \Theta(\eta_K^\star)$, substituting into (47) the numerator $(c')^{2(N-1)/N}$ of (9) contributes a *second* factor $(c')^{2(N-1)/N}$ to the denominator of T:

$$T \;=\; O\Bigg(\frac{M^{3N-4}G_{K,R,\bar{D}} + M^{4N-4}L_{K,R,\bar{D}}}{\mu_K\,(c')^{4(N-1)/N}} \times \log\frac{\mathcal{F}(A(0)) - \mathcal{F}^\star}{\varepsilon}\Bigg),$$

which is (12), whose leading curvature factor is the $S_K(R)^2$ inside $L_{K,R,\bar{D}}$ and whose remaining one is $H_K(\bar{D})$. The exponent is $4(N-1)/N$ and not $2(N-1)/N$: writing $\eta_K^\star = O(1/S_K(R)^2)$ and displaying only one of the two factors would hide the other inside the implicit constant.

Step 4: The Euclidean case $K = 0$.

When $K = 0$:

- $S_0(R) = 1$, so $L_{0,R,D} = \frac{1}{m}\sum \|\xi_i\|^2 \cdot [1 \cdot 1 + 0] = \frac{1}{m}\sum \|\xi_i\|^2$, which is data-dependent but curvature-free;
- $\mu_0 = 1 - D \cdot 0 = 1$;
- $\eta_0^\star = O(1)$ (a data-dependent constant independent of curvature).

The contraction (10) becomes

$$\mathcal{F}(A(T)) - \mathcal{F}^\star \le \big(1 - \eta\,(c')^{2(N-1)/N}\big)^T \times (\mathcal{F}(A(0)) - \mathcal{F}^\star),$$

and the iteration count (12) becomes

$$T = O\Bigg(\frac{1}{(c')^{4(N-1)/N}}\log\frac{\mathcal{F}(A(0)) - \mathcal{F}^\star}{\varepsilon}\Bigg),$$

recovering the guarantee of Theorem 1 of Arora et al. (2018a) (equation (8) therein) for the Euclidean deep linear network. The two sources of curvature cost are therefore visible, with $S_K(R)^2$ entering through the step-size ceiling (a discretization cost) and μ_K entering through the gradient-dominance constant (a strong-convexity cost). □

D.6 Proof of General Curvature and the Positive-Curvature Corollary

Proof of Theorem 10. The strategy is to show that the proof of Theorem 3 (Appendix D.1) never uses the sign of K; it uses only four comparison-geometry bounds on a constant-curvature space. For $K \le 0$ these bounds hold globally (Hadamard manifold); for $K > 0$ they hold within the injectivity radius $\mathrm{inj} = \pi/\sqrt{K}$. We verify each bound, then confirm that every algebraic step of the proof carries over unchanged.

Step 1: The four geometric bounds for general K.

Let $(\mathcal{M}_N, g_N)$ be the simply connected space form of constant sectional curvature $K \in \mathbb{R}$. When $K > 0$, this is the sphere $S^{d_N}(\frac{1}{\sqrt{K}})$ of radius $1/\sqrt{K}$, with injectivity radius $\mathrm{inj} = \pi/\sqrt{K}$. Fix a base point $o_N \in \mathcal{M}_N$.

Bound 1: Rauch comparison ($\|D\exp_{o_N}(v)\|_{\mathrm{op}} \le S_K(R)$ for $\|v\| \le R$). The differential $D\exp_{o_N}(v)$ maps $T_0V_N \simeq \mathbb{R}^{d_N}$ to $T_{\exp(v)}\mathcal{M}_N$. By the Gauss lemma, the radial singular value is 1. Each transverse singular value equals $\mathrm{sn}_K(\|v\|)/\|v\|$, the ratio of the Jacobi field on $\mathcal{M}_N$ to the Euclidean Jacobi field, evaluated at distance $\|v\|$ from the origin. For constant curvature K this ratio is

$$\frac{\mathrm{sn}_K(\|v\|)}{\|v\|} = S_K(\|v\|),$$

where

$$\mathrm{sn}_K(r) = \begin{cases} \sin(\sqrt{K}r)/\sqrt{K} & K > 0, \\ r & K = 0, \\ \sinh(\sqrt{\kappa}r)/\sqrt{\kappa} & K < 0. \end{cases}$$

Case $K \le 0$: $\sinh(t)/t \ge 1$ for all $t \ge 0$, so $S_K(r) \ge 1$ and the maximum over $\|v\| \le R$ is $S_K(R)$ (monotone increasing). The bound holds globally.

Case $K > 0$: $\sin(\theta)/\theta$ is positive and bounded for $\theta \in (0, \pi/2)$. For $\|v\| \le R < \pi/(2\sqrt{K})$, we have $\sqrt{K}\|v\| \in [0, \pi/2)$, so $\mathrm{sn}_K(\|v\|) > 0$ and $S_K(\|v\|) = \sin(\sqrt{K}\|v\|)/(\sqrt{K}\|v\|) \in (0, 1]$. The supremum over $\|v\| \le R$ is $S_K(0^+) = 1$ (since $\sin\theta/\theta \to 1$ as $\theta \to 0$), but more precisely $\sup_{\|v\|\le R} S_K(\|v\|) = 1$ with $S_K(R) < 1$ at the boundary. For the operator-norm upper bound, the relevant quantity is $\widehat{S}_K(R) = \max(1, S_K(R)) = 1$ when $K > 0$ (the radial direction gives the maximum). Accordingly the factor entering the smoothness constant (5) is $\widehat{S}_K(R)^2 = 1$, not $S_K(R)^2$: for $K > 0$ there is no transverse Rauch-factor penalty in the operator-norm smoothness bound, although the $D\,B_K(R)$ term remains.

In either case, $\|D\exp_{o_N}(v)\|_{\mathrm{op}} \le \max(1, S_K(R))$ is finite under the stated radius constraint.

Bound 2: Second derivative of the exponential map ($\|D^2\exp_{o_N}(v)\|_{\mathrm{op}} \le B_K(R)$ for $\|v\| \le R$). The second derivative $D^2\exp_{o_N}(v)$ is a bilinear map on $\mathbb{R}^{d_N}$. On a constant-curvature space, its components are expressed in terms of derivatives of the Jacobi field ratio. In the radial direction, $D^2\exp$ vanishes (the exponential map is a radial isometry, hence affine in the radial variable). In the transverse directions, the norm is controlled by $S_K'(\|v\|)\cdot\|v\|$ and cross terms.

Case $K = 0$: $\exp_{o_N}(v) = o_N + v$ (a translation), so $D^2\exp \equiv 0$ and $B_0(R) = 0$.

Case $K < 0$: $B_K(R)$ is finite and increasing in R (the hyperbolic exponential map has bounded second derivatives on any compact set).

Case $K > 0$: For $R < \pi/(2\sqrt{K})$, the Jacobi field $\mathrm{sn}_K(r) = \sin(\sqrt{K}r)/\sqrt{K}$ and all its derivatives are smooth and bounded on $[0, R]$. Therefore $B_K(R) := \sup_{\|v\|\le R}\|D^2\exp_{o_N}(v)\|_{\mathrm{op}} < \infty$. $B_K(R)$ in fact remains finite throughout the injectivity radius $R < \pi/\sqrt{K}$ (the closed form below gives $\|D^2\exp\| = O(1/R)$ even at $R = \pi/(2\sqrt{K})$, where $\cos(\sqrt{K}R) = 0$), so the exponential map does not degenerate there. The tighter restriction $R, D < \pi/(2\sqrt{K})$ imposed in Theorem 10 is not needed to bound $D^2\exp$; it is imposed to keep the transverse squared-distance Hessian positive, i.e. $f_K(D) \in (0, 1)$ and $H_K(D) = 1$ (Bound 3), so that the comparison bounds hold in a strongly controlled convexity region.

Closed form on a space form. In geodesic polar coordinates $dr^2 + \mathrm{sn}_K(r)^2 d\theta^2$ the curve $t \mapsto \exp_{o_N}(Re_1 + te_2)$ has $r(0) = R$, $r'(0) = 0$, $r''(0) = 1/R$, $\theta'(0) = 1/R$ and $\theta''(0) = 0$, so its acceleration at $t = 0$ is purely radial, while the mixed derivative is the radial derivative of the transverse Jacobi field. Hence

$$\left\|D^2\exp_{o_N}(Re_1)[e_2, e_2]\right\| = \left|\frac{1}{R} - \frac{\mathrm{sn}_K(R)\mathrm{cn}_K(R)}{R^2}\right|,$$
$$\left\|D^2\exp_{o_N}(Re_1)[e_1, e_2]\right\| = \left|\frac{\mathrm{cn}_K(R)}{R} - \frac{\mathrm{sn}_K(R)}{R^2}\right|,$$

and $D^2 \exp_{o_N}(Re_1)[e_1, e_1] = 0$ because exp is a radial isometry. Both quantities vanish identically at $K = 0$, confirming $B_0(R) = 0$. For $K < 0$, using $\mathrm{sn}_K(R)\mathrm{cn}_K(R)/R^2 = \sqrt{\kappa}\coth(\sqrt{\kappa}R)S_K(R)^2$, they give

$$B_K(R) = O\big(\sqrt{\kappa}\, S_K(R)^2\big). \tag{48}$$

Bound 3: Hessian comparison for the squared distance ($\|\nabla^2\phi_i\|_{\mathrm{op}} \le H_K(D)$ for $d(p_i, y_i) \le D$). The Hessian of $\phi_i(p) = \frac{1}{2}d(p, y_i)^2$ at p_i with $D_i := d(p_i, y_i)$ has radial eigenvalue 1 and transverse eigenvalue $f_K(D_i) = D_i\mathrm{cn}_K(D_i)/\mathrm{sn}_K(D_i)$.

Case $K \le 0$: $f_K(D_i) = D_i\sqrt{\kappa}\coth(\sqrt{\kappa}D_i) \ge 1$, so $H_K(D) = \max(1, f_K(D)) = f_K(D) \ge 1$.

Case $K > 0$: $f_K(D_i) = D_i\sqrt{K}\cot(\sqrt{K}D_i)$. For $\sqrt{K}D_i \in (0, \pi/2)$, $\cot(\theta) > 0$ and $\theta\cot\theta \in (0, 1)$ (since $\theta < \tan\theta$ for $\theta \in (0, \pi/2)$). Therefore $f_K(D_i) \in (0, 1)$ and

$$H_K(D) = \max(1, f_K(D)) = 1 \qquad \text{for } \sqrt{K}D < \pi/2.$$

The Hessian comparison therefore improves for $K > 0$, since the transverse eigenvalue drops below 1 and the overall bound is controlled by the radial eigenvalue alone.

The requirement is $D < \pi/(2\sqrt{K})$ and not merely $D < \pi/\sqrt{K}$: for $\sqrt{K}D \in (\pi/2, \pi)$ the transverse eigenvalue $f_K(D)$ is negative and unbounded below, so $H_K(D) = \max(1, f_K(D))$ would cease to bound $\|\nabla^2\phi_i\|_{\mathrm{op}}$. Under $\sqrt{K}D < \pi/2$ we have $f_K(D) \in (0, 1)$, the bound is valid, and $H_K(D) = 1$.

Bound 4: Gradient bound ($\|\nabla\phi_i\| = d(p_i, y_i) \le D$). $\nabla_p\phi_i = -\log_p(y_i)$, so $\|\nabla\phi_i\| = \|\log_p(y_i)\| = d(p, y_i) \le D$. This is a metric identity, valid on *any* Riemannian manifold whenever y_i is within the injectivity radius of p_i. For $K > 0$, the constraint $D < \pi/(2\sqrt{K}) < \mathrm{inj}$ ensures this.

Step 2: The algebraic structure is unchanged.

We now verify that every step of the proof of Theorem 3 (Appendix D.1) goes through with the general-K constants.

Stage 1 (Gradient flow). The continuous-time descent (Corollary 4, equation (8)) uses only the preconditioner $\mathcal{P}_A(G) = \sum_j W_{j+1:N}W_{j+1:N}^* G W_{1:j-1}^* W_{1:j-1}$ and its eigenvalue lower bound $\sigma_{\min}(A)^{2(N-1)/N}$. These are determined by the algebraic structure of the weight factorization and are independent of the manifold geometry. Stage 1 therefore holds for any K without modification.

Stage 2, Step 1 (Pullback gradient). The chain-rule formula (14),

$$G = -\frac{1}{m}\sum_{i=1}^{m}[D\exp_{o_N}(A\xi_i)]^* \log_{p_i}(y_i)\,\xi_i^*,$$

is valid on any Riemannian manifold where exp and log are well-defined. For $K > 0$, this requires $\|A\xi_i\| < \mathrm{inj}$ and $d(p_i, y_i) < \mathrm{inj}$, which are guaranteed by the tube constraints $R < \pi/(2\sqrt{K})$ and $D < \pi/(2\sqrt{K})$.

Stage 2, Step 2 (Smoothness). The Hessian bound

$$|D^2\mathcal{F}(A)[\Delta, \Delta]| \le L_{K,R,D}\|\Delta\|_{\mathrm{HS}}^2$$

was derived in the proof of Theorem 3 using Bounds 1–4 above. Since all four bounds remain valid for $K > 0$ under the stated radius constraints, the smoothness constant $L_{K,R,D}$ (5) is finite and the quadratic upper bound holds.

Stage 2, Steps 3–6 (Factor gradients, expansion, coercivity, combination). Steps 3–5 are algebraic. The factor-gradient formula $\nabla_{W_j}\mathcal{L} = W_{j+1:N}^* G W_{1:j-1}^*$, the multinomial expansion $A^+ = A - \eta\mathcal{P}_A(G) + R_\eta$, and the coercivity bound $\langle G, \mathcal{P}_A(G)\rangle \geq \frac{3}{4}\sigma_{\min}(A)^{2(N-1)/N}\|G\|_{\mathrm{HS}}^2$ (from hypothesis (3) of Theorem 3) involve only the weight matrices and the gradient G, not the geometry of $\mathcal{M}_N$. Step 6 combines the first-order coercivity with the $L_{K,R,D}$-smoothness bound and the step-size condition (6); since $L_{K,R,D}$ is now the general-K value and $G_{K,R,D}$ is the gradient norm bound on the general-K tube, the arithmetic is identical and yields (7).

Trajectory containment (Lemma 8). Part (b) (balancedness preservation) is purely algebraic in the weight matrices and carries over verbatim for any K. Part (c) does *not* carry over unchanged: its lower singular-value bound is derived through the distance inequality of Step 2 of Appendix D.4, which uses the nonpositive-curvature expansion $d(\exp_{o_N} u, \exp_{o_N} v) \geq \|u - v\|$. For $K > 0$ the exponential map *contracts* transverse distances, so this inequality fails in the transverse directions and the deficiency-margin argument for $\sigma_{\min}(A(t)) \geq c'$ has to be re-derived with curvature-corrected constants; we do not claim it here. Consequently Theorem 10 is stated as a one-step descent lemma for $K > 0$, and the full trajectory induction (Lemma 8) and the end-to-end convergence rate (Theorem 9) are established only for $K \leq 0$. Part (a) (tube containment), where it applies, requires $\mathcal{F}$ to be non-increasing, which follows from the descent lemma (7) together with the radius and distance constraints $R < \pi/(2\sqrt{K})$ and $D < \pi/(2\sqrt{K})$ that keep the trajectory within the injectivity radius.

Since every ingredient of the proof of the *single-step* Theorem 3 remains valid under the stated radius constraints, the conclusion (7) holds for all $K \in \mathbb{R}$. □

Proof of Corollary 11. We derive each of the four claims under $K > 0$ with $\sqrt{K}\,R < \pi/2$ and $\sqrt{K}\,D < \pi/2$.

Claim (1): $S_K(R) < 1$. The Rauch factor is

$$S_K(R) = \frac{\mathrm{sn}_K(R)}{R} = \frac{\sin(\sqrt{K}\,R)}{\sqrt{K}\,R}.$$

Set $\theta := \sqrt{K}\,R \in (0, \pi/2)$. The elementary inequality $\sin\theta < \theta$ for $\theta > 0$ gives

$$S_K(R) = \frac{\sin\theta}{\theta} < 1.$$

More precisely, $\sin\theta/\theta$ is strictly decreasing on $(0, \pi)$ with $\lim_{\theta\to 0}\sin\theta/\theta = 1$, so $S_K(R) \in (2/\pi, 1)$ for $\theta \in (0, \pi/2)$.

Claim (2): $H_K(D) = 1$. The Hessian comparison function is

$$\begin{aligned} f_K(D) &= D\,\frac{\mathrm{cn}_K(D)}{\mathrm{sn}_K(D)} = D\,\frac{\cos(\sqrt{K}\,D)\cdot\sqrt{K}}{\sin(\sqrt{K}\,D)/\sqrt{K}\cdot\sqrt{K}} \\ &= \sqrt{K}\,D\,\cot(\sqrt{K}\,D). \end{aligned}$$

Set $\alpha := \sqrt{K}\,D \in (0, \pi/2)$. Then $f_K(D) = \alpha\cot\alpha$. For $\alpha \in (0, \pi/2)$: $\tan\alpha > \alpha$ (since $\frac{d}{d\alpha}\tan\alpha = \sec^2\alpha > 1$ for $\alpha > 0$), so $\cot\alpha < 1/\alpha$, giving $\alpha\cot\alpha < 1$. Since $\alpha\cot\alpha > 0$ in this range, we have $f_K(D) \in (0, 1)$. Therefore

$$H_K(D) = \max(1,\, f_K(D)) = 1.$$

The Hessian comparison penalty equals $\sqrt{\kappa}D\coth(\sqrt{\kappa}D) \gg 1$ for large $|K|$ when $K < 0$, and it is absent for $K > 0$ within the stated range.

Claim (3): $\widehat{S}_K(R) = 1$, so $L_{K,R,D} \to L_{0,R,D}$ as $D \to 0$. From (5):

$$L_{K,R,D} = \frac{1}{m}\sum_i \|\xi_i\|^2 \big[\widehat{S}_K(R)^2\, H_K(D) + D\, B_K(R)\big].$$

For $K > 0$ claim (1) gives $S_K(R) < 1$, so $\widehat{S}_K(R) = \max(1, S_K(R)) = 1$: the contraction is felt only in the transverse directions, while the operator norm of $D\exp$ is set by the unstretched radial direction. With claim (2), $H_K(D) = 1$, so

$$\begin{aligned} L_{K,R,D} &= \frac{1}{m}\sum_i \|\xi_i\|^2 \big[1 + D\, B_K(R)\big] \\ &\xrightarrow[D\to 0]{} \frac{1}{m}\sum_i \|\xi_i\|^2 = L_{0,R,D}\big|_{D\to 0}. \end{aligned}$$

Note that the limit is approached from above: for $D > 0$ we have $B_K(R) > 0$, hence $L_{K,R,D} > L_{0,R,D}$. Positive curvature therefore removes the Rauch factor from the smoothness constant but does not push it below its Euclidean value. (For $K < 0$ the picture is different: $\widehat{S}_K(R) = S_K(R) > 1$ and $H_K(D) \geq 1$, so $L_{K,R,D} > L_{0,R,D}$ by the exponentially growing factor $S_K(R)^2$.)

Claim (4): $\eta_K^\star \to \eta_0^\star$. From (9):

$$\eta_K^\star = \frac{\sigma_{\min}(A)^{2(N-1)/N}}{4C_N(M^{3N-4}G_{K,R,D} + L_{K,R,D}M^{4N-4})}.$$

The numerator is curvature-independent, and by claim (3) $L_{K,R,D} \to L_{0,R,D}$ as $D \to 0$. The gradient bound behaves the same way: $\|G\|$ involves $\|D\exp\|_{\mathrm{op}} = \widehat{S}_K(R) = 1$ and $\|\nabla\phi_i\| \leq D$, so $G_{K,R,D} \to G_{0,R,D}$ as well. Hence $\eta_K^\star \to \eta_0^\star$: positive curvature carries no Rauch-factor penalty, in contrast with the exponential penalty $\eta_K^\star = O(e^{-2\sqrt{\kappa}R})$ incurred for $K < 0$.

Summary. Negative curvature ($K < 0$) has $S_K(R) > 1$ and $H_K(D) \geq 1$, which increases the smoothness constant $L_{K,R,D}$ and shrinks the maximum stable step size $\eta_K^\star = O(1/S_K(R)^2)$. Positive curvature ($K > 0$) removes both penalties, with $\widehat{S}_K(R) = 1$ and $H_K(D) = 1$ returning $L_{K,R,D}$ to its Euclidean value as $D \to 0$, so that $\eta_K^\star \to \eta_0^\star$ rather than exceeding it. The exponential map on a sphere compresses transverse distances, but the radial singular value remains one; consequently the worst-case operator-norm guarantee returns to the Euclidean ceiling rather than proving a larger one. On a hyperbolic space it stretches transverse directions, sharpening the loss and demanding smaller steps. □

D.7 Proof of the Tangent-Space Surrogate

Proof of Proposition 12. The argument shows that the tangent-space surrogate $\mathcal{E}$ (3) is, from the standpoint of optimization, the Euclidean deep linear network studied by Arora et al. (2018a), so their convergence theorem applies directly with no curvature-dependent modification. We make this precise in six steps.

Step 1: The surrogate is a squared Frobenius loss on whitened data.

Recall the definitions from Section A.2. The tangent-space data are $\xi_i := \log_{\mathbf{0}}^K(x_i) \in \mathbb{R}^{d_0}$ and $\tilde{y}_i := \log_{\mathbf{0}}^K(y_i) \in \mathbb{R}^{d_N}$, and the surrogate is

$$\begin{aligned} \mathcal{E}(W) \;&:=\; \frac{1}{2m}\sum_{i=1}^{m} \|A\xi_i - \tilde{y}_i\|^2, \\ A &:= W_{1:N} = W_N \cdots W_1. \end{aligned} \tag{3}$$

Once the logarithmic maps have been applied to form ξ_i and $\tilde{y}_i$, the surrogate lives in Euclidean space. Both $A\xi_i$ and $\tilde{y}_i$ are vectors in $\mathbb{R}^{d_N}$, and the norm $\|\cdot\|$ is the standard Euclidean norm. No exponential map, geodesic distance, or manifold structure appears in $\mathcal{E}$.

Define the data matrices $X := [\xi_1 \;\cdots\; \xi_m] \in \mathbb{R}^{d_0 \times m}$ and $Y := [\tilde{y}_1 \;\cdots\; \tilde{y}_m] \in \mathbb{R}^{d_N \times m}$. Then

$$\mathcal{E}(W) = \frac{1}{2m}\|AX - Y\|_F^2.$$

Since the data are whitened ($\frac{1}{m}\sum_i \xi_i\xi_i^* = \frac{1}{m}XX^\top = I_{d_0}$), the standard calculation (see equation (1) of Arora et al. (2018a)) gives

$$\mathcal{E}(W) = \frac{1}{2}\|A - \Phi\|_{\mathrm{HS}}^2 + \mathcal{E}^\star, \tag{49}$$

where $\Phi := \frac{1}{m}YX^\top \in \mathbb{R}^{d_N \times d_0}$ is the cross-covariance (the tangent-space target matrix) and $\mathcal{E}^\star := \min_W \mathcal{E}(W) \geq 0$ does not depend on A.

Equation (49) is identical in form to the loss $\mathcal{L}^1(W) = \frac{1}{2}\|W - \Phi\|_F^2$ of equation (2) in Arora et al. (2018a). The curvature parameter K entered only in the definition of the logarithmic map that produced ξ_i and $\tilde{y}_i$; once those vectors are fixed, $\mathcal{E}$ is a standard Euclidean quadratic.

Step 2: The gradient of $\mathcal{E}$ is curvature-free.

Differentiating (49) with respect to the end-to-end matrix:

$$\nabla_A \mathcal{E}(A) = A - \Phi. \tag{50}$$

The surrogate gradient is a linear function of A, independent of K. Compare with the intrinsic gradient (14):

$$\nabla_A \mathcal{F}(A) = -\frac{1}{m}\sum_i [D\exp_{o_N}(A\xi_i)]^* \log_{p_i}(y_i)\, \xi_i^*,$$

which involves the differential $D\exp$ (curvature-dependent via the Rauch factor) and the logarithmic map $\log_{p_i}(y_i)$ (curvature-dependent via the geodesic distance). The surrogate gradient contains none of these terms, with no $D\exp$, no $\log_p$, and no Rauch factor.

The factor gradients are, by the chain rule:

$$\nabla_{W_j}\mathcal{E} = W_{j+1:N}^*(A - \Phi)W_{1:j-1}^*, \qquad j = 1, \ldots, N,$$

which are again algebraic in the weights and the fixed Euclidean data.

Step 3: Smoothness and strong convexity are curvature-free.

Smoothness. The Hessian of the end-to-end pullback $\mathcal{E}^1(A) := \frac{1}{2}\|A - \Phi\|_{\mathrm{HS}}^2$ is

$$D^2\mathcal{E}^1(A)[\Delta, \Delta] = \|\Delta\|_{\mathrm{HS}}^2 \qquad \text{for all } \Delta \in \mathbb{R}^{d_N \times d_0}.$$

That is, $\mathcal{E}^1$ is exactly 1-smooth and 1-strongly convex (the generic bound (5) gives the looser constant d_0, since it passes through $\|\Delta\xi_i\| \leq \|\xi_i\|\|\Delta\|$ instead of the whitening identity) (with respect to $\|\cdot\|_{\mathrm{HS}}$) with no curvature dependence.

Comparing with the intrinsic pullback, whose smoothness constant is $L_{K,R,D}$ (5), the Hessian there had two terms, the Hessian comparison $H_K(D)$ weighted by $S_K(R)^2$ and the $D\,B_K(R)$ correction from $D^2\exp$. For the surrogate, both terms collapse, because there is no exponential map, so $D\exp = I$ ($S_K \to 1$), $D^2\exp = 0$ ($B_K \to 0$), and $\nabla^2\phi_i = I$ ($H_K \to 1$). The smoothness constant of $\mathcal{E}^1$ is thus

$$\begin{aligned} L_{\mathcal{E}} &= \frac{1}{m}\sum_i \|\xi_i\|^2 \cdot [1^2 \cdot 1 + 0] = \frac{1}{m}\sum_i \|\xi_i\|^2 \\ &= \operatorname{tr}(I_{d_0}) = d_0 \end{aligned}$$

for whitened data (or 1 per coordinate), independent of K.

Strong convexity. By the same token, $D^2\mathcal{E}^1(A)[\Delta,\Delta] = \|\Delta\|_{\mathrm{HS}}^2 \geq 1 \cdot \|\Delta\|_{\mathrm{HS}}^2$, so $\mathcal{E}^1$ is $\mu = 1$-strongly convex. Compare with Lemma 6, where the intrinsic pullback has $\mu_K = 1 - D\,B_K(R) < 1$ for $K < 0$.

Gradient dominance. The Polyak–Łojasiewicz inequality for $\mathcal{E}^1$ is

$$\begin{aligned} \|\nabla_A \mathcal{E}^1(A)\|_{\mathrm{HS}}^2 &= \|A - \Phi\|_{\mathrm{HS}}^2 = 2\,\mathcal{E}^1(A) \\ &= 2(\mathcal{E}^1(A) - \mathcal{E}^{1,\star}), \end{aligned}$$

since $\mathcal{E}^{1,\star} = 0$ when the target is realizable (or the constant term otherwise). This holds with constant $2\mu = 2$, with no curvature factor.

Step 4: Applying the Arora et al. convergence theorem.

Steps 1–3 show that gradient descent on the factorized loss $\mathcal{E}(W_1,\dots,W_N) = \mathcal{E}^1(W_{1:N})$ is the problem studied by Arora et al. (2018a), a depth-N linear network minimizing $\frac{1}{2}\|W_{1:N} - \Phi\|_F^2$ over whitened data. Their Theorem 1 (equation (8) therein) gives: if the initialization is δ-balanced (Definition 1) with deficiency margin $c' > 0$ (Definition 2) and the learning rate satisfies

$$\eta \;\leq\; \frac{(c')^{(4N-2)/N}}{C''_N\, N^3\, \|\Phi\|_F^{(6N-4)/N}} \;=:\; \eta_0^{\mathcal{E}}, \tag{51}$$

then for any $\varepsilon > 0$ and

$$T \;\geq\; \frac{1}{\eta\,(c')^{2(N-1)/N}}\,\log\frac{\mathcal{E}(W(0)) - \mathcal{E}^\star}{\varepsilon}, \tag{52}$$

the loss satisfies $\mathcal{E}(W(T)) - \mathcal{E}^\star \leq \varepsilon$, where $\mathcal{E}^\star := \min_W \mathcal{E}(W)$ is the irreducible least-squares residual. The gap $\mathcal{E} - \mathcal{E}^\star$, not $\mathcal{E}$ itself, is the quantity that contracts: by (49), $\mathcal{E}(W) - \mathcal{E}^\star = \frac{1}{2}\|A - \Phi\|_{\mathrm{HS}}^2$, which vanishes only when the tangent-space problem is realizable.

More explicitly, the proof of their Theorem 1 establishes a single-step descent for the least-squares gap:

$$\begin{aligned} \mathcal{E}(W(t+1)) - \mathcal{E}^\star \leq \mathcal{E}(W(t)) - \mathcal{E}^\star \\ - \frac{\eta}{2}\,\sigma_{\min}(A(t))^{2(N-1)/N}\|A(t) - \Phi\|_{\mathrm{HS}}^2. \end{aligned}$$

Trajectory containment gives $\sigma_{\min}(A(t)) \geq c'$ and 2δ-balancedness, while $\|A(t) - \Phi\|_{\mathrm{HS}}^2 = 2(\mathcal{E}(W(t)) - \mathcal{E}^\star)$. Hence

$$\mathcal{E}(W(t+1)) - \mathcal{E}^\star \leq \big(1 - \eta(c')^{2(N-1)/N}\big)\big(\mathcal{E}(W(t)) - \mathcal{E}^\star\big),$$

which yields (52).

Step 5: Curvature independence.

We now verify that *no quantity* in the convergence guarantee depends on K (or equivalently on $\kappa = |K|$).

(i) *Step-size ceiling $\eta_0^{\mathcal{E}}$ (equation (51)):* depends on c', N, and $\|\Phi\|_F$. The deficiency margin c' and the target Φ are properties of the initialization and the tangent-space data $(\xi_i, \tilde{y}_i)$, which are fixed once the logarithmic maps are applied. Changing K would change the values of $\xi_i, \tilde{y}_i$, but the proposition holds "on the same data and weights," meaning that we compare the two losses at the same $(\xi_i, \tilde{y}_i)$ and the same initialization. Therefore $\eta_0^{\mathcal{E}}$ is curvature-free.

(ii) *Contraction factor $\rho = \eta(c')^{2(N-1)/N}$:* depends on η and c', both curvature-free.

(iii) *Smoothness constant $L_{\mathcal{E}}$:* equals $\frac{1}{m}\sum \|\xi_i\|^2 = d_0$ for whitened data, curvature-free.

(iv) *Strong-convexity constant μ:* equals 1, curvature-free.

(v) *Iteration count (52):* $T = (\eta(c')^{2(N-1)/N})^{-1} \log((\mathcal{E}(W(0)) - \mathcal{E}^\star)/\varepsilon)$, curvature-free.

In contrast, for the intrinsic loss $\mathcal{L}_K$ (Theorem 9), the step size is $\eta_K^\star = O(1/S_K(R)^2)$ (curvature enters via $L_{K,R,D}$ (5)), the strong-convexity constant is $\mu_K = 1 - D\,B_K(R) < 1$ for $K < 0$ (Lemma 6), and the iteration count is $O(S_K(R)^2/\mu_K \cdot \ldots)$ (12).

Step 6: The $K = 0$ collapse.

When $K = 0$, the manifold $\mathcal{M}_N$ is Euclidean and the exponential map is the identity: $\exp_{o_N}(v) = o_N + v$. Then for whitened data centered at the origin ($o_N = 0$):

$$d_0(\exp_0(A\xi_i), y_i)^2 = \|A\xi_i - \tilde{y}_i\|^2,$$

so $\mathcal{L}_0(W) = \mathcal{E}(W)$ identically. The intrinsic smoothness constant collapses to $L_{0,R,D} = \frac{1}{m}\sum \|\xi_i\|^2 = L_{\mathcal{E}}$, the strong-convexity constant to $\mu_0 = 1$, and the step-size ceiling to $\eta_0^\star = \eta_0^{\mathcal{E}}$. The two convergence guarantees, Theorem 9 for $\mathcal{L}_K$ and Proposition 12 for $\mathcal{E}$, coincide at $K = 0$.

For $K \neq 0$, the difference between the two guarantees is the curvature cost. The intrinsic loss incurs $S_K(R)^2$ in its step-size ceiling and $(1 - D\,B_K(R))^{-1}$ in its strong-convexity degradation, while the surrogate incurs neither. The comparison confirms that the curvature dependence originates from measuring error on the manifold (via d^2), not from the network architecture or the factorized parameterization. □

D.8 Proof of the Landscape Theorem

Proof of Theorem 13. Let $A \in \mathcal{T}_{R,D}$ be a critical point and let $A^\star$ be a tube minimizer such that $[A, A^\star] \subset \mathcal{T}_{R,D}$. Under the hypotheses of Lemma 6,

$$\mu_K := \min(1, S_K(R)^2 f_K(D)) - DB_K(R) > 0.$$

Lemma 7 gives

$$\|\nabla_A \mathcal{F}(A)\|_{\mathrm{HS}}^2 \geq 2\mu_K \big(\mathcal{F}(A) - \mathcal{F}^\star\big).$$

At a critical point the left-hand side is zero, and therefore $\mathcal{F}(A) = \mathcal{F}^\star$.

For $K \leq 0$, $\mu_K = 1 - DB_K(R)$, so the conclusion holds on every tube with $DB_K(R) < 1$. For $K > 0$, the same argument applies within the stated radius restrictions whenever $S_K(R)^2 f_K(D) - DB_K(R) > 0$. The result is local: it says that every critical point satisfying the stated segment condition is a minimizer of the end-to-end pullback on the tube. It does not assert that curvature preserves the number or locations of critical points, and it does not concern critical points of the factorized weight-space loss.

When $K = 0$, $\mathcal{F}(A) = \frac{1}{2}\|A - \Phi\|_{\mathrm{HS}}^2 + \mathcal{E}^\star$, so the unique end-to-end minimizer is $A = \Phi$. □

D.9 Proof of the Adaptive Schedule

Proof of Proposition 15. Fix an iterate satisfying Lemma 8 and work on the fixed tube $\mathcal{T}_{R,\bar{D}}$. The segment estimate in the proof of that lemma supplies a positive threshold η_{seg} such that all smaller steps have their update segments in the fixed tube. On this tube, $\mathcal{F}$ has smoothness constant $L_{K,R,\bar{D}}$ and gradient bound $G_{K,R,\bar{D}}$.

The trial step is capped by η_0 and repeatedly halved. After finitely many halvings it is no larger than

$$\min\left\{\eta_{\mathrm{seg}}, \frac{\sigma_{\min}(A(t))^{2(N-1)/N}}{4C_N\big(M^{3N-4}G_{K,R,\bar{D}} + L_{K,R,\bar{D}}M^{4N-4}\big)}\right\}.$$

The segment and step-size hypotheses of Theorem 3 then hold, so (7) is accepted. This proves finite termination.

The induction of Lemma 8 extends to accepted variable steps $\eta_t \le \eta_0$. In the balancedness estimate, $\sum_t \eta_t^2\|G(t)\|^2 \le \eta_0 \sum_t \eta_t\|G(t)\|^2$, and the latter sum is bounded by telescoping the accepted descent inequalities. The singular-value and tube arguments use only monotone decrease, the uniform step cap, and the accepted segment test. Hence the containment, balancedness, and singular-value conclusions persist.

Remark 17 is conditional and is not used in this proof; in particular, no uniform lower bound of order $1/S_K(R)^2$ is claimed for every accepted step. □

D.10 Proof of Gradient Anisotropy

Proof of Proposition 18. Write $G_K(A) := \nabla_A\mathcal{F}(A)$ and $v_i := A\xi_i$. The exact pullback gradient is

$$G_K(A) = -\frac{1}{m}\sum_i [D\exp_o(v_i)]^* \log_{p_i}(y_i)\xi_i^*, \qquad p_i := \exp_o(v_i). \tag{53}$$

Early phase. For $f_i(v) := \frac{1}{2}d(\exp_o(v), y_i)^2$, one has $\nabla f_i(0) = -\tilde{y}_i$. Its Hessian at the base point has radial eigenvalue one and transverse eigenvalue

$$\|\tilde{y}_i\|\frac{\mathrm{cn}_K(\|\tilde{y}_i\|)}{\mathrm{sn}_K(\|\tilde{y}_i\|)} = 1 + O(\kappa\|\tilde{y}_i\|^2).$$

Taylor expansion and whitening give

$$G_K(A) = A - \Phi + O\left(\kappa \max_i \|\tilde{y}_i\|^2\|A\| + \|A\|^2\right) = \nabla_A\mathcal{E}(A) + O(\cdots). \tag{54}$$

Thus the two end-to-end gradients agree to leading order in the stated early-phase regime. This is a local vector-field comparison, not a characterization of the final implicit bias or selected solution.

Late-phase anisotropy. $D\exp_o(v_i)$ has radial singular value one and transverse singular values $S_K(\|v_i\|)$. If $A = \sum_k \sigma_k u_k\zeta_k^*$ and ξ_i aligns with ζ_k, then $\|v_i\| \simeq \sigma_k\|\xi_i\|$. Equation (53) therefore leaves radial residual components unchanged and weights transverse components by $S_K(\sigma_k\|\xi_i\|)$. They are amplified for $K < 0$ and damped for $K > 0$ within the injectivity radius.

Stationary-limit condition. Exact balancedness gives $\dot{A} = -\mathcal{P}_A(G_K)$, so any stationary limit satisfies

$$\mathcal{P}_{A^\star}(G_K(A^\star)) = 0. \tag{55}$$

For $N = 1$, $\mathcal{P}_A = I$. For $N \geq 2$, take a full SVD with orthonormal extensions and zero singular values beyond the rank. The preconditioner eigenvalue on $u_k v_\ell^*$ is

$$\sum_{j=1}^{N} (\sigma_k^\star)^{2(N-j)/N} (\sigma_\ell^\star)^{2(j-1)/N},$$

which vanishes only when both associated singular values are zero. Thus $G_K(A^\star)$ can have support only on pairs of unused left and right singular directions. Because G_K contains the curvature-dependent adjoint $[D \exp]^*$, curvature changes the stationarity equation, but this does not prove that different curvatures select different limits. At $K = 0$, the condition reduces to Euclidean stationarity with $G_0(A) = A - \Phi$. □

D.11 Proof of the Sharpness Construction

Proof of Proposition 20. We exhibit a configuration in which the Rauch factor is attained exactly as a Hessian eigenvalue *at a stationary point*, so that the linear-stability threshold of gradient descent is $2/S_K(R)^2$ and the ceiling $\eta_K^\star = O(1/S_K(R)^2)$ of Theorem 3 cannot be improved.

Step 1: The construction.

Choose:

- output dimension $d_N = n \geq 2$, input dimension $d_0 = 1$;
- depth $N = 1$, so the end-to-end matrix is a vector $a \in \mathbb{R}^n$;
- one data point, $m = 1$, with $\xi_1 = 1$;
- target $y = \exp_{o_N}(R\, e_1)$, where e_1 is the first standard basis vector;
- initial parameter $a_0 = R\, e_1$.

Then $\exp_{o_N}(a_0) = y$, so the loss

$$\begin{aligned} \mathcal{F}(a) = \tfrac{1}{2}\, d\big(\exp_{o_N}(a),\, y\big)^2 =: \phi_1\big(\exp_{o_N}(a)\big), \\ \phi_1(p) := \tfrac{1}{2} d(p, y)^2, \end{aligned} \tag{56}$$

vanishes at a_0 and a_0 is a global minimizer, hence a stationary point: $\nabla \mathcal{F}(a_0) = 0$. The residual is $D = 0$, so the smoothness constant (5) with $m = 1$ and $\|\xi_1\| = 1$ is

$$\begin{aligned} L_{K,R,0} &= \widehat{S}_K(R)^2\, H_K(0) + 0 \cdot B_K(R) \\ &= S_K(R)^2 \qquad (K < 0), \end{aligned} \tag{57}$$

using $H_K(0) = \max(1, f_K(0)) = 1$ from the continuity convention of Appendix A.4 and $\widehat{S}_K(R) = S_K(R) > 1$ for $K < 0$.

Placing the target at the image of a_0, rather than at $\exp_{o_N}(s\, e_1)$ with $s \neq R$, is what makes the argument a genuine stability statement. With $s \neq R$ the point a_0 is not stationary: gradient descent moves it along the radial geodesic, and a transverse perturbation would have to be tracked along a moving base point. Worse, by rotational symmetry about the e_1-axis a trajectory started on the axis stays on it, so the transverse coordinate is identically zero and never grows. At a stationary point both problems disappear.

Step 2: The Hessian at a_0 (Claim 1).

By the decomposition of Step 2 of Appendix D.1, the Hessian of $\mathcal{F}$ at a in a direction Δ is

$$D^2\mathcal{F}(a)[\Delta,\Delta] = \underbrace{\big\langle \nabla^2\phi_1\big|_p\big(D\exp(a)\Delta\big),\, D\exp(a)\Delta\big\rangle}_{\text{(I)}} + \underbrace{\big\langle \nabla\phi_1\big|_p,\, D^2\exp(a)[\Delta,\Delta]\big\rangle}_{\text{(II)}},$$

with $p=\exp_{o_N}(a)$. At $a=a_0$ we have $p=y$, so $\nabla\phi_1|_y = -\log_y(y)=0$ and Term (II) vanishes identically. No alignment argument and no bound on $B_K(R)$ is needed. Term (I) is evaluated with $\nabla^2\phi_1|_y = I$, the Hessian of $\frac{1}{2}d(\cdot,y)^2$ at its own minimizer, leaving

$$D^2\mathcal{F}(a_0)[\Delta,\Delta] = \big\|D\exp_{o_N}(R\,e_1)\,\Delta\big\|^2. \tag{58}$$

By the Jacobi-field decomposition of $D\exp_{o_N}(R\,e_1)$ (Bound 1 of Appendix D.6), the radial singular value is 1 and each transverse singular value is $S_K(R)$. Hence for a purely transverse perturbation $\Delta=\varepsilon e_2$,

$$D^2\mathcal{F}(a_0)[\Delta,\Delta] = S_K(R)^2\varepsilon^2 = L_{K,R,0}\,\|\Delta\|_{\mathrm{HS}}^2$$

by (57), while a radial perturbation gives $\|\Delta\|_{\mathrm{HS}}^2$. For $K<0$ we have $S_K(R)>1$, so the transverse value is the top eigenvalue, and the Rauch bound used in the smoothness proof is attained rather than merely bounding. This is Claim 1.

Step 3: Linear instability (Claims 2 and 3).

Because a_0 is stationary, gradient descent $a\mapsto a-\eta\nabla\mathcal{F}(a)$ linearizes there to

$$\Delta \;\longmapsto\; \big(I-\eta\,D^2\mathcal{F}(a_0)\big)\Delta,$$

a linear map whose eigenvalues are $1-\eta$ in the radial direction and $1-\eta S_K(R)^2$ in each transverse direction. The fixed point a_0 is linearly stable exactly when the spectral radius is at most 1, that is when $\eta\le 2/S_K(R)^2$ (the transverse directions bind, since $S_K(R)>1$). For every $\eta>2/S_K(R)^2$ the transverse multiplier has modulus greater than 1 and the perturbation grows geometrically in the linearised range, so gradient descent is linearly unstable at a_0. This is Claim 2, and it gives the exact threshold

$$\eta_{\mathrm{crit}} = \frac{2}{S_K(R)^2} = \Theta\!\left(\frac{1}{S_K(R)^2}\right), \tag{59}$$

which is Claim 3. Unlike the analysis of a non-stationary point, this is exact for every $R>0$ and needs no $\sqrt{\kappa}R\gg 1$ asymptotics; for $\sqrt{\kappa}R\gg 1$ one may substitute $S_K(R)^2\sim e^{2\sqrt{\kappa}R}/(4\kappa R^2)$ to read off the exponential decay. Comparing with $\eta_K^\star = O(1/S_K(R)^2)$ from (9), the ceiling of Theorem 3 has the correct dependence on the Rauch factor and cannot be improved.

What the construction does and does not show. It shows that the $S_K(R)^2$ factor in $L_{K,R,D}$ is attained, and hence that no analysis can remove it. It does not certify tightness of the $H_K(D)$ and $D\,B_K(R)$ terms, which vanish here because the residual is zero; those enter only away from the optimum, and Proposition 15 is what removes the first of them. □

E Experimental Details and Further Results

We illustrate the qualitative predictions of the descent theory on the deep-linear network of Appendix A.1 with every layer on the same constant-curvature model: linear convergence (Theorem 9), the curvature-explicit step-size ceiling $\eta_K^\star = O(1/S_K(R)^2)$ (Theorem 3, Corollary 11), and the $K\to 0$ collapse onto the tangent-space surrogate (Proposition 12) as a control.

E.1 Setup

We use the κ-stereographic model of Appendix C (Bachmann et al., 2020), the one model in which K can be swept continuously from spherical ($K > 0$) through Euclidean ($K = 0$) to hyperbolic ($K < 0$) without changing coordinates. At $K = 0$ the model is $\mathbb{R}^d$ and $\mathcal{L}_0 = \mathcal{E}$, so Figure 2 of the main paper can plot a single curve against signed K. The experiments use the collapse identity (1) $h = \exp_o(W_{1:N} \log_o(x))$, the intrinsic loss $\mathcal{L}_K$ (2), and the surrogate $\mathcal{E}$ (3), with gradients taken with respect to the individual factors W_j. Unless stated otherwise we use depth $N = 3$, width $d = 6$, $m = 300$ samples, bounded radius $R = 1.2$, and an initialization that is exactly balanced at $t = 0$ ($\delta = 0$, with each W_j a scaled orthogonal matrix). A single tangent-space regression problem is generated once and reused at every curvature via $x^{(i)} = \exp_o(\tilde{x}^{(i)})$, $y^{(i)} = \exp_o(\tilde{y}^{(i)})$, so that curvature enters only through the geometry and never through the data. The inputs are isotropic with $\|\tilde{x}^{(i)}\| \le R$, so their population covariance is a multiple of the identity and the sample covariance is close to it, a scale that cancels in the ratio $\lambda_K^\star/\lambda_0^\star$ reported below. Unless stated otherwise all results average over 50 seeds.

Experimental protocol. All geometric quantities are evaluated in double precision, since the stereographic model is numerically unstable at single precision. To reach the optimum we run full-batch gradient descent on the factors at a small fixed step ($\eta = 0.02$) until convergence. The near-optimum Hessian eigenvalue $\lambda_K^\star$ is then obtained by power iteration (80 steps) on Hessian–vector products in the end-to-end matrix A at the converged solution, so that the measured quantity is the top eigenvalue of $D^2\mathcal{F}$, which is what $L_{K,R,D}$ bounds. The robustness sweep (Figure S3) comprises $4 \cdot 3 \cdot 50 \cdot 7 = 4{,}200$ independent configurations.

E.2 Expectations of the Theory

Curvature enters the descent lemma only through the smoothness constant $L_{K,R,D}$ (5), and hence only through the step-size ceiling $\eta_K^\star = O(1/S_K(R)^2)$ (9). A larger $S_K(R)$ (more negative curvature) forces a smaller stable step, while Corollary 11 gives $\widehat{S}_K(R) = 1$ for $K > 0$, so positive curvature returns the ceiling to its Euclidean value rather than enlarging it. The constant $L_{K,R,D}$ is a worst-case bound, and the whole-trajectory stable step from a high-residual initialization is dominated by the large-D region and is nearly curvature-flat, so it is not an informative probe. The informative quantity is instead the late-phase sharpness. For $K \le 0$, the sharpness construction and comparison bounds identify $S_K(R)^2$ as the transverse scale that can control the local Hessian and the stability ceiling $\eta^\star = 2/\lambda_K^\star$. For $K > 0$, the worst-case operator-norm factor is instead $\widehat{S}_K(R)^2 = 1$; positive-curvature measurements are therefore empirical and are not consequences of an $S_K(R)^2$ top-eigenvalue theorem. We report the near-optimum sharpness rather than an inflated whole-trajectory step size.

E.3 Results

Figure 2 of the main paper reports four measurements against signed curvature $K \in \{1, \frac{1}{2}, 0, -\frac{1}{2}, -1, -2, -4, -6, -8\}$ at $R = 1.2$ and is discussed there: every curvature converges linearly at a fixed step size, the $K = 0$ trajectory coincides with the Euclidean surrogate to machine precision, the near-optimum sharpness $\lambda_K^\star$ is monotone in K, the negative-curvature ratio $\lambda_K^\star/\lambda_0^\star$ follows the identity line against $S_K(R)^2$ qualitatively, and the δ-balancedness defect stays $\lesssim 10^{-1}$ throughout training. As a control, the tangent surrogate $\mathcal{E}$ has a convergence exponent independent of K.

Figure S1 adds the fixed-step counterpart of that sharpness measurement: the single-step convergence exponent $|b|$ of the intrinsic loss equals the surrogate exponent at $K = 0$ and moves

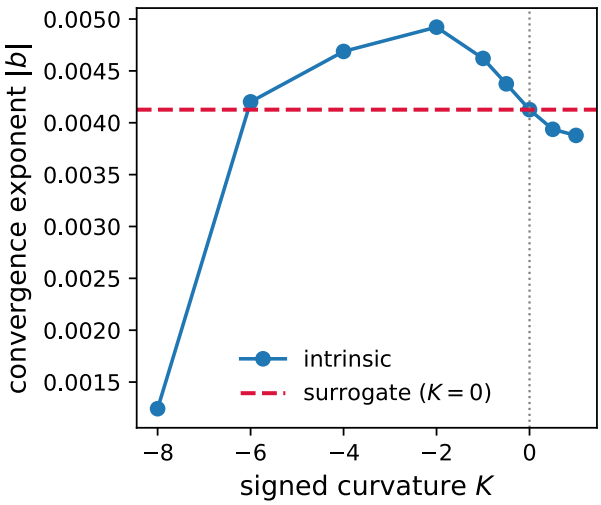


Figure S1: Curvature-collapse check. The single-step convergence exponent $|b|$, the slope of log loss against gradient-descent step fitted over the middle of the trajectory, of the intrinsic loss equals the Euclidean tangent-surrogate exponent at $K = 0$ (dashed line) and moves with curvature away from it. The coincidence at $K = 0$ is the $\mathcal{L}_0 = \mathcal{E}$ collapse of Proposition 12.

with curvature away from it, which is the collapse $\mathcal{L}_0 = \mathcal{E}$ of Proposition 12 seen along the trajectory rather than at the optimum.

Additional experiments. Three further experiments probe robustness and the remaining structural results.

Robustness of the negative-curvature $S_K(R)^2$ scale. Figure S3 repeats the sharpness measurement across four radii, three architectures, 50 seeds, and seven curvatures. For $K \leq 0$, the horizontal axis $S_K(R)^2$ is the theoretically motivated transverse scale. Positive-curvature points are shown only as an empirical extension and are excluded from any theorem-based agreement claim. The negative-curvature measurements follow the predicted trend in the controlled range $\sqrt{|K|}R \leq 2$, with increased scatter near the boundary where the H_K and B_K terms become more influential.

Surrogate as curvature-free control (Proposition 12). Figure S2 compares the intrinsic loss $\mathcal{L}_K$ under intrinsic gradient descent with the surrogate loss $\mathcal{E}$ under surrogate gradient descent at the same step size. The surrogate curves overlap to machine precision across all K, so the surrogate trajectory is curvature-free. The intrinsic curves separate. More negative curvature produces steeper initial descent (the intrinsic gradient is amplified by the metric) but is eventually limited by the tighter step-size ceiling.

No spurious critical points (Theorem 13). Figure S4 launches 50 balanced initializations at each of $K \in \{0, -1, -4\}$ on the tube. Every trajectory reaches the global minimum (final loss $< 10^{-3}$ in all cases), consistent with Theorem 13, under which every local minimum on the tube is global.

Scope and limitations. These experiments illustrate the sign and shape of the descent theory rather than verifying its constants or hypotheses (they do not check tube containment, the deficiency margin, or the exact ceiling constants). Since $L_{K,R,D}$ is a worst-case upper bound, the curvature penalty is a late-phase phenomenon visible in the near-optimum sharpness rather than in the whole-trajectory step size, and for $K \leq 0$, $\lambda^\star_K$ tracks the $S_K(R)^2$ scale only up to the higher-order H_K and B_K terms. For positive curvature, the observed stable steps may exceed the Euclidean

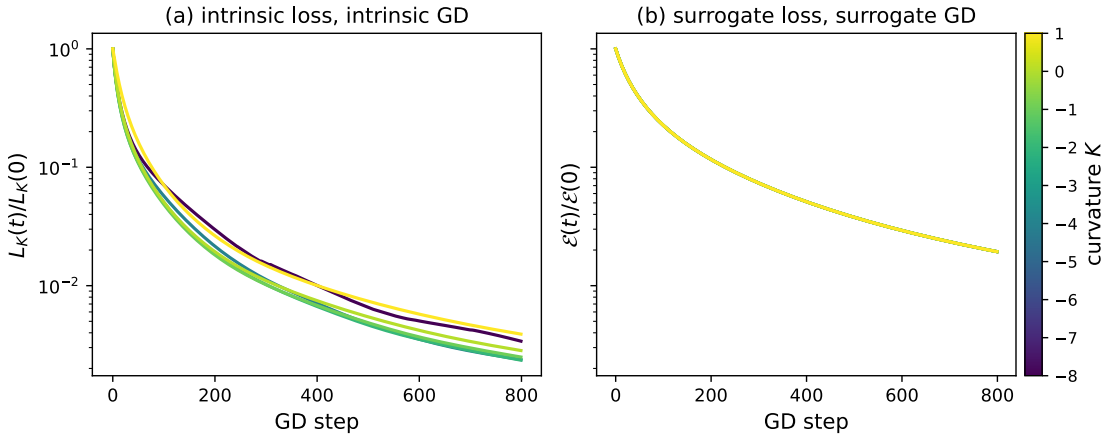


Figure S2: Surrogate as a curvature-free control (Proposition 12). **(a)** Under intrinsic gradient descent the normalized intrinsic loss $L_K(t)/L_K(0)$ separates with curvature (colored by K). **(b)** Under surrogate gradient descent the normalized surrogate loss $\mathcal{E}(t)/\mathcal{E}(0)$ is independent of K and the curves coincide, as the surrogate loss and its gradient do not depend on the curvature.

empirical threshold in these instances; this is an empirical observation, stronger than and not implied by the worst-case guarantee of Corollary 11 (see the remark following Corollary 11). The setting is deep-linear and bias-free with a single constant-curvature manifold, and nonlinear activations and Möbius biases fall outside our scope.

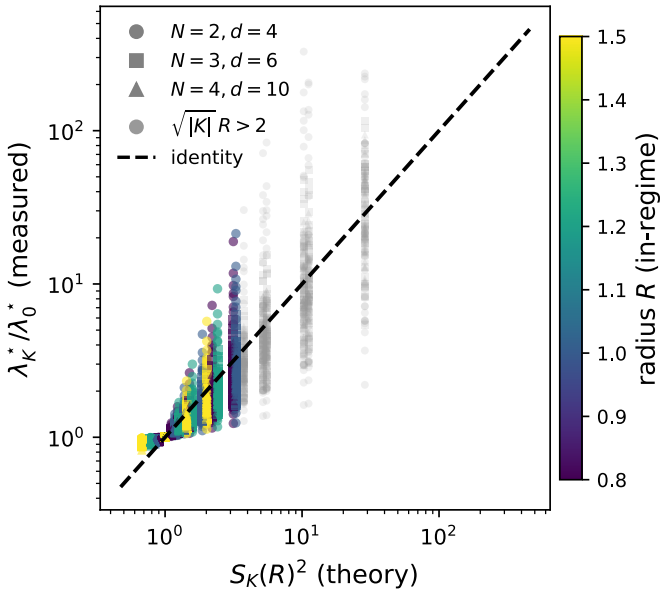


Figure S3: Robustness of the negative-curvature $S_K(R)^2$ scale. Measured sharpness ratio $\lambda_K^\star/\lambda_0^\star$ versus $S_K(R)^2$ across four radii, three architectures, and 50 seeds. For $K \le 0$, $S_K(R)^2$ is the transverse scale predicted by the negative-curvature analysis. Positive-curvature points are empirical only, because the corresponding worst-case operator-norm factor is $\widehat{S}_K(R)^2 = 1$. Increased scatter near the boundary reflects the higher-order H_K and B_K terms.

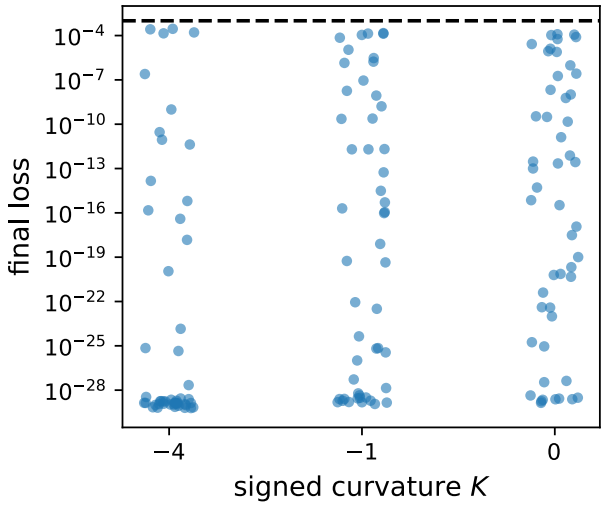


Figure S4: No spurious critical points on the tube (Theorem 13). Final loss of 50 random balanced, in-tube initializations at each curvature. Every run reaches the global minimum (below the dashed 10^{-3} threshold), which is consistent with—but does not by itself prove—the local tube theorem.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

El Mehdi Achour, Kathlén Kohn, and Holger Rauhut. The riemannian geometry associated to gradient flows of linear convolutional networks. *arXiv preprint arXiv:2507.06367*, 2025.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018a.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International conference on machine learning*, pages 254–263. PMLR, 2018b.

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4453–4462, 2022.

Gregor Bachmann, Gary Bécigneul, and Octavian Ganea. Constant curvature graph convolutional networks. In *International conference on machine learning*, pages 486–496. PMLR, 2020.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *International Conference on Learning Representations*, volume 2024, pages 47687–47711, 2024.

Silvere Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods, 2019. URL https://arxiv.org/abs/1810.00760.

Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fully hyperbolic neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5672–5686, 2022.

Sungjun Cho, Seunghyuk Cho, Sungwoo Park, Hankook Lee, Honglak Lee, and Moontae Lee. Curve your attention: Mixed-curvature transformers for graph representation learning. *arXiv preprint arXiv:2309.04082*, 2023.

Christopher Criscitiello and Nicolas Boumal. Negative curvature obstructs acceleration for strongly geodesically convex optimization, even with exact first-order oracles. In *Conference on Learning Theory*, pages 496–542. PMLR, 2022.

Christopher Criscitiello and Nicolas Boumal. Curvature and complexity: Better lower bounds for geodesically convex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2969–3013. PMLR, 2023.

Jindou Dai, Yuwei Wu, Zhi Gao, and Yunde Jia. A hyperbolic-to-hyperbolic graph convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 154–163, 2021.

Simon Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pages 1655–1664. PMLR, 2019.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

Pascal Esser, Leena Chennuru Vankadara, and Debarghya Ghoshdastidar. Learning theory can (sometimes) explain generalisation in graph neural networks. *Advances in Neural Information Processing Systems*, 34:27043–27056, 2021.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Xingcheng Fu, Jianxin Li, Jia Wu, Qingyun Sun, Cheng Ji, Senzhang Wang, Jiajun Tan, Hao Peng, and Philip S Yu. Ace-hgnn: Adaptive curvature exploration hyperbolic graph neural network. In *2021 IEEE international conference on data mining (ICDM)*, pages 111–120. IEEE, 2021.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *arXiv preprint arXiv:1805.09112*, 2018.

Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *International conference on machine learning*, pages 3419–3430. PMLR, 2020.

Christopher R Genovese, Marco Perone Pacifico, Isabella Verdinelli, Larry Wasserman, et al. Minimax manifold estimation. *Journal of machine learning research*, 13:1263–1291, 2012.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference on learning theory*, pages 297–299. PMLR, 2018.

Karish Grover, Haiyang Yu, Qi Zhu, Han Xie, Vassilis Ioannidis, Christos Faloutsos, et al. Spectro-riemannian graph neural networks. In *International Conference on Learning Representations*, volume 2025, pages 57833–57859, 2025.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *International conference on learning representations*, 2018.

Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, et al. Hyperbolic attention networks. *arXiv preprint arXiv:1805.09786*, 2018.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

Jiang Hu, Xin Liu, Zai-Wen Wen, and Ya-Xiang Yuan. A brief introduction to manifold optimization. *Journal of the Operations Research Society of China*, 8(2):199–248, 2020.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

Haotian Ju, Dongyue Li, Aneesh Sharma, and Hongyang R Zhang. Generalization in graph neural networks: Improved pac-bayesian bounds on graph diffusion. In *International conference on artificial intelligence and statistics*, pages 6314–6341. PMLR, 2023.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6418–6428, 2020.

Arlene KH Kim and Harrison H Zhou. Tight minimax rates for manifold estimation under hausdorff loss. 2015.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Thomas Laurent and James Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *International conference on machine learning*, pages 2902–2907. PMLR, 2018.

Marc Law. Ultrahyperbolic neural networks. *Advances in Neural Information Processing Systems*, 34:22058–22069, 2021.

Yuekang Li, Yidan Mao, Yifei Yang, and Dongmian Zou. Improving robustness of hyperbolic neural networks by lipschitz analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1713–1724, 2024.

Renjie Liao, Raquel Urtasun, and Richard Zemel. A pac-bayesian approach to generalization bounds for graph neural networks. *arXiv preprint arXiv:2012.07690*, 2020.

Philip M. Long and Hanie Sedghi. Generalization bounds for deep convolutional neural networks, 2020. URL https://arxiv.org/abs/1905.12600.

Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

Sohir Maskey, Ron Levie, Yunseok Lee, and Gitta Kutyniok. Generalization analysis of message passing neural networks on large random graphs. *Advances in neural information processing systems*, 35:4805–4817, 2022.

Emile Mathieu, Charline Le Lan, Chris J Maddison, Ryota Tomioka, and Yee Whye Teh. Continuous hierarchical representations with poincaré variational auto-encoders. *Advances in neural information processing systems*, 32, 2019.

Govind Menon. The geometry of the deep linear. In *XIV Symposium on Probability and Stochastic Processes: CIMAT, Mexico, November 20-24, 2023*, page 1. Springer Nature, 2025.

Gal Mishne, Zhengchao Wan, Yusu Wang, and Sheng Yang. The numerical stability of hyperbolic representation learning. In *International Conference on Machine Learning*, pages 24925–24949. PMLR, 2023.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pages 3779–3788. PMLR, 2018.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1):419–441, 2008.

Maher Nouiehed and Meisam Razaviyayn. Learning deep models: Critical points and local openness. *INFORMS Journal on Optimization*, 4(2):148–173, 2022.

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

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pages 4460–4469. PMLR, 2018.

Krisanu Sarkar. Learning beyond euclid: Curvature-adaptive generalization for neural networks on manifolds. *arXiv preprint arXiv:2507.02999*, 2025.

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.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Franco Scarselli, Ah Chung Tsoi, and Markus Hagenbuchner. The vapnik–chervonenkis dimension of graph and recursive neural networks. *Neural Networks*, 108:248–259, 2018.

Ohad Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In *Conference on Learning Theory*, pages 2691–2713. PMLR, 2019.

Ondrej Skopek, Octavian-Eugen Ganea, and Gary Bécigneul. Mixed-curvature variational autoencoders. *arXiv preprint arXiv:1911.08411*, 2019.

Li Sun, Zhongbao Zhang, Jiawei Zhang, Feiyang Wang, Hao Peng, Sen Su, and Philip S Yu. Hyperbolic variational graph neural network for modeling dynamic graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4375–4383, 2021.

Li Sun, Zhenhao Huang, Qiqi Wan, Hao Peng, and Philip S Yu. Spiking graph neural network on riemannian manifolds. *Advances in Neural Information Processing Systems*, 37:34025–34055, 2024.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Tan Sun and Junhong Lin. Pac-bayesian adversarially robust generalization bounds for graph neural network. *arXiv preprint arXiv:2402.04038*, 2024.

Atsushi Suzuki, Atsushi Nitanda, Linchuan Xu, Kenji Yamanishi, Marc Cavazza, et al. Generalization bounds for graph embedding using negative sampling: Linear vs hyperbolic. *Advances in Neural Information Processing Systems*, 34:1243–1255, 2021.

Huayi Tang and Yong Liu. Towards understanding generalization of graph neural networks. In *International Conference on Machine Learning*, pages 33674–33719. PMLR, 2023.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018.

Max Van Spengler, Erwin Berkhout, and Pascal Mettes. Poincare resnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5419–5428, 2023.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Saurabh Verma and Zhi-Li Zhang. Stability and generalization of graph convolutional neural networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1539–1548, 2019.

Shen Wang, Xiaokai Wei, Cicero Nogueira Nogueira dos Santos, Zhiguo Wang, Ramesh Nallapati, Andrew Arnold, Bing Xiang, Philip S Yu, and Isabel F Cruz. Mixed-curvature multi-relational graph neural network for knowledge graph completion. In *Proceedings of the web conference 2021*, pages 1761–1771, 2021.

Bo Xiong, Shichao Zhu, Nico Potyka, Shirui Pan, Chuan Zhou, and Steffen Staab. Pseudo-riemannian graph convolutional networks. *Advances in neural information processing systems*, 35:3488–3501, 2022.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Menglin Yang, Min Zhou, Tong Zhang, Jiahong Liu, Zhihao Li, Lujia Pan, Hui Xiong, and Irwin King. Hyperbolic graph neural networks: A review of methods and applications. *arXiv preprint arXiv:2202.13852*, 2022.

Tao Yu and Christopher M De Sa. Numerically accurate hyperbolic embeddings using tiling-based models. *Advances in Neural Information Processing Systems*, 32, 2019.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on learning theory*, pages 1617–1638. PMLR, 2016.

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

Yiding Zhang, Xiao Wang, Chuan Shi, Nian Liu, and Guojie Song. Lorentzian graph convolutional networks. In *Proceedings of the web conference 2021*, pages 1249–1261, 2021.

Difan Zou, Philip M Long, and Quanquan Gu. On the global convergence of training deep linear resnets. *arXiv preprint arXiv:2003.01094*, 2020.