CURVATURE, RESIDUAL SCALE, AND SIGNAL PROPAGATION IN HYPERBOLIC NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

How do curvature and residual scale jointly determine signal propagation in hyperbolic neural networks? We study intrinsic input sensitivity for residual layers on hyperbolic space of sectional curvature $-\kappa^2$. Along an aligned reference trajectory, we derive an exact Jacobi-field expression for the layer differential and identify a unique transition between contraction and expansion. The transition depends on the dimensionless residual displacement κhv and the covariant derivative of the residual field, rather than on curvature alone. We establish sufficient conditions for these regimes to persist under bounded derivative perturbations and turning of the reference direction. A constrained neural parameterization realizes these conditions, and explicit smoothness estimates extend the reference analysis to contracting neighborhoods and finite-time escape from expanding cones. We further construct a nonlinear correction that preserves the reference trajectory and its entire input-Jacobian product while changing the network away from that trajectory. Numerical experiments verify the differential formulas and sufficient bounds. Controlled learning experiments reveal limited nonlinear-task performance for the original parameterization and substantial improvements after adding the correction, including when the reference Jacobian remains fixed throughout training. These results separate certified reference sensitivity from nonlinear learning behavior and provide a setting in which their relationship can be studied explicitly.

1 INTRODUCTION

The propagation of perturbations through depth is a central question in the analysis of neural networks. Mean-field approaches relate initialization and architecture to the evolution of signals and gradients (Schoenholz et al., 2017; Yang & Schoenholz, 2017). Complementary work studies the singular values of the input-output Jacobian, including conditions under which they concentrate near one (Pennington et al., 2017). These perspectives identify mechanisms that influence optimization, but their conclusions depend on the architecture, the statistical assumptions, and the sensitivity quantity being considered.

Hyperbolic neural networks introduce an additional geometric parameter. Their layers operate on a curved space, and derivatives map between tangent spaces equipped with position-dependent inner products (Alvarado & Burgos, 2024; 2026). Hyperbolic feedforward and recurrent constructions (Ganea et al., 2018; Shimizu et al., 2021), together with hyperbolic and Riemannian residual architectures (van Spengler et al., 2023; Katsman et al., 2023), provide different ways to incorporate this geometry into deep models. For such networks, an intrinsic sensitivity analysis must account for both the residual field and the differential of the exponential map. A coordinate Jacobian alone does not identify the Riemannian amplification of an input perturbation.

Existing analyses address several related aspects of hyperbolic learning. Numerical stability depends on the representation of hyperbolic space and the interaction between finite precision and optimization (Mishne et al., 2023). Recent work on hyperbolic reinforcement learning examines gradient behavior and optimization difficulties associated with hyperbolic operations (Klein et al., 2026). A separate dynamical perspective studies subhomogeneous hyperbolic layers and stochastic iterates (Alvarado & Burgos, 2026). Our focus is a deterministic, finite-dimensional question: how

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

does an explicitly controlled residual construction amplify intrinsic perturbations along a prescribed reference trajectory, and what conclusions extend to nearby inputs?

We consider layers of the form $F_\ell(x) = \exp_x^\kappa\big(hV_\ell(x)\big)$, on hyperbolic space of sectional curvature $-\kappa^2$, where $h > 0$ is the residual scale and V_ℓ is a trainable tangent vector field. Two effects enter the layer differential. Negative curvature contributes transverse amplification through the variation of geodesics, while the covariant derivative of V_ℓ can counteract or reinforce this effect. Consequently, the relevant geometric displacement is $\kappa h\|V_\ell(x)\|$. A curvature comparison must therefore specify which residual lengths and derivative scales are held fixed.

Our first result makes this interaction explicit. Along an aligned reference trajectory with constant residual speed v, let a and b denote the dimensionless transverse and longitudinal damping parameters, respectively. The transverse differential gain is $G(z, a) = \cosh z - a\frac{\sinh z}{z}, \qquad z = \kappa h v$, with the ratio defined continuously at $z = 0$. For $0 < a < 2$ and $0 < b < 1$, a unique positive threshold z_* satisfies $a = z_* \tanh(z_*/2)$. The maximal intrinsic growth rate is negative below this threshold, zero at the threshold, and positive above it. This conclusion is conditional on the specified alignment and derivative structure. It requires neither an infinite-width limit nor a random initialization model. Moreover, neutrality of the maximal growth rate does not imply dynamical isometry, the longitudinal direction remains contracting.

We then relax exact alignment by allowing bounded derivative defects and turning of the reference direction. Explicit sufficient inequalities preserve contraction or cone expansion. A normalized neural parameterization enforces the required bounds at a shared reference trajectory for every admissible parameter choice. Uniform smoothness estimates extend these first-order statements to finite neighborhoods. Nearby inputs contract under one set of conditions, whereas suitable transverse perturbations leave a specified neighborhood within a controlled number of layers under another.

Reference sensitivity does not, however, determine the network's nonlinear behavior away from that trajectory. To investigate this distinction constructively, we introduce a trainable correction whose value and first derivative vanish at each reference point. For fixed original backbone parameters, this correction preserves the reference trajectory and the complete input-Jacobian product exactly. It nevertheless changes higher-order behavior, including the smoothness constants entering the neighborhood guarantees. Our experiments use this freedom to improve a nonlinear classification task while retaining a fixed reference Jacobian throughout training.

Our main contributions are the following

1. We derive the aligned differential and identify a unique contraction-expansion threshold in terms of curvature, residual displacement, and covariant damping.
2. We give sufficient conditions under derivative perturbations and directional turning, construct a parameterization satisfying them, and obtain explicit contracting neighborhoods and finite time cone escape bounds.
3. We construct a correction preserving the reference trajectory and Jacobian product, and quantify its effect on the smoothness bounds required for neighborhood guarantees.
4. We verify the theoretical formulas and bounds, examine sensitivity across curvature and depth, and demonstrate improved nonlinear task performance with an unchanged reference Jacobian.

2 Related Work

Signal propagation and Jacobian conditioning. The propagation of signals through deep random networks has been studied through mean-field descriptions of correlations, sensitivity, and representation geometry (Poole et al., 2016; Schoenholz et al., 2017). For residual architectures, Yang & Schoenholz (2017) analyze how skip connections modify propagation across depth. A complementary approach examines the singular-value distribution of the input–output Jacobian. Pennington et al. (2017) relate dynamical isometry to the conditioning of deep networks, while Tarnowski et al. (2019) derive limiting Jacobian spectra for residual networks using random matrix and free probability methods. Our analysis concerns deterministic intrinsic differentials at finite width.

Residual scaling and stable architectures. The scale of a residual branch is an established control variable for deep-network behavior. Fixup rescales initialization to facilitate training without normalization (Zhang et al., 2019). Stable ResNet introduces depth-dependent residual scalings and analyzes gradient stability and expressivity in infinite-width and infinite-depth limits (Hayou et al., 2021). From a dynamical-systems perspective, Haber & Ruthotto (2018) connect forward propagation to discretized differential equations and use stability considerations to design architectures. Our results complement these approaches by retaining an explicit finite residual displacement on a negatively curved manifold. The dimensionless quantity $\kappa h\|V_\ell\|$ enters through geodesic variation, while $h\nabla V_\ell$ describes the local variation of the residual field.

Hyperbolic representations and neural operations. Poincaré embeddings demonstrate the usefulness of hyperbolic geometry for representing hierarchical relations (Nickel & Kiela, 2017). Learning in the Lorentz model provides another geometric and computational formulation of this representation problem (Nickel & Kiela, 2018). Hyperbolic neural networks extend feedforward and recurrent operations to hyperbolic space (Ganea et al., 2018), with subsequent developments refining the construction of neural layers (Shimizu et al., 2021). Hyperbolic graph convolutional networks incorporate feature transformations, aggregation, and trainable layerwise curvature (Chami et al., 2019). Fully Hyperbolic Neural Networks construct operations using Lorentz transformations (Chen et al., 2022). These works establish a broad architectural setting for learning with negative curvature. Our contribution concerns intrinsic perturbation propagation in a specified residual family. The theoretical assumptions and controlled synthetic experiments do not establish representation advantages for hierarchical data or guarantees for arbitrary hyperbolic layers.

Hyperbolic and Riemannian residual networks. Riemannian Residual Neural Networks formulate residual updates through tangent vector fields and exponential maps on manifolds (Katsman et al., 2023). Our layer form belongs to this existing geometric framework. Poincaré ResNet develops hyperbolic residual components and an identity-based initialization intended to preserve signal norms (van Spengler et al., 2023). Lorentzian Residual Neural Networks construct residual connections using weighted Lorentzian centroids (He et al., 2025). These architectures make the definition of a residual connection, its initialization, and its implementation central design choices. We specialize exponential-map residual updates to a controlled reference trajectory and derive explicit intrinsic differential bounds.

Geometric sensitivity and numerical stability. Intrinsic perturbation growth and numerical stability describe different aspects of a hyperbolic computation. Mishne et al. (2023) analyze representation-dependent limitations and optimization tradeoffs in hyperbolic learning. Recent work on hyperbolic reinforcement learning examines gradient behavior and optimization difficulties associated with hyperbolic operations (Klein et al., 2026). These issues motivate careful implementations, but finite precision behavior alone does not characterize the Riemannian operator norm of a layer differential. Our sensitivity measurements incorporate the tangent-space metrics at both the input and the output. The resulting quantities describe the underlying Riemannian maps, whereas agreement between numerical implementations and analytic formulas provides a separate implementation check.

Geometric and dynamical foundations. The differential of the exponential map is governed by Jacobi fields, and the hyperbolic functions appearing in our formulas are classical consequences of constant negative curvature (Lee, 2018). Likewise, contraction analysis connects infinitesimal bounds on tangent dynamics to finite-distance behavior; Forni & Sepulchre (2014) develop this principle in a differential Lyapunov framework for continuous-time systems. Related work on hyperbolic neural networks studies subhomogeneous layer maps and stochastic dynamics (Alvarado & Burgos, 2026). Our discrete, layer dependent analysis combines the Jacobi-field differential with quantitative perturbation bounds along a prescribed trajectory. The neighborhood results additionally control nonlinear remainders through explicit smoothness estimates.

3 PRELIMINARIES

Hyperbolic geometry. Let $(\mathbb{H}^d_\kappa, g_\kappa)$ denote the complete, simply connected Riemannian manifold of constant sectional curvature $-\kappa^2$, with $\kappa > 0$ and $d \geq 2$. We write d_κ for its geodesic distance

and $\|\xi\|_{g_\kappa}$ for the norm of a tangent vector. The exponential map at x is a global diffeomorphism, with inverse $\log_x^\kappa$. In particular,

$$\log_x^\kappa(\exp_x^\kappa(\xi)) = \xi, \qquad d_\kappa(x, \exp_x^\kappa(\xi)) = \|\xi\|_{g_\kappa}.$$

We denote parallel transport along the unique geodesic from x to y by $\mathcal{P}_{x\to y}$. It is a linear isometry between the corresponding tangent spaces. The Levi-Civita connection is denoted by ∇, and the covariant derivative of a vector field V at x is the linear map

$$\nabla V(x) : T_x\mathbb{H}^d_\kappa \to T_x\mathbb{H}^d_\kappa, \qquad \xi \mapsto \nabla_\xi V.$$

For explicit calculations, we use the Lorentz inner product

$$\langle x, y\rangle_{\mathrm{L}} = -x_0 y_0 + \sum_{j=1}^{d} x_j y_j$$

and the realization $\mathbb{H}^d_\kappa = \{x \in \mathbb{R}^{d+1} : \langle x, x\rangle_{\mathrm{L}} = -\kappa^{-2},\ x_0 > 0\}$. Its tangent space at x is $T_x\mathbb{H}^d = \{\xi : \langle x, \xi\rangle_{\mathrm{L}} = 0\}$, and the metric is the restriction of the Lorentz inner product to this space. For $r = \|\xi\|_{g_\kappa}$ and $c = -\kappa^2\langle x, y\rangle_{\mathrm{L}}$, the standard hyperboloid formulas give (Nagano et al., 2019)

$$\begin{aligned}
\exp_x^\kappa(\xi) &= \cosh(\kappa r)x + \frac{\sinh(\kappa r)}{\kappa r}\xi,\\
\log_x^\kappa(y) &= \frac{\operatorname{arcosh}(c)}{\sqrt{c^2-1}}(y - cx),\\
\mathcal{P}_{x\to y}\xi &= \xi + \frac{\kappa^2\langle y, \xi\rangle_{\mathrm{L}}}{1+c}(x + y).
\end{aligned}$$

The expressions are understood by continuous extension when $r = 0$ or $y = x$. All sensitivity measurements below use the Riemannian metric, rather than the ambient Euclidean norm on hyperboloid coordinates.

Residual dynamics and intrinsic sensitivity. Following the intrinsic residual formulation of Katsman et al. (2023), consider

$$F_\ell(x) = \exp_x^\kappa(hV_\ell(x)), \qquad x_{\ell+1} = F_\ell(x_\ell), \qquad \Phi_L = F_{L-1}\circ\cdots\circ F_0,$$

where $h > 0$ and each V_ℓ is a continuously differentiable tangent vector field. Higher regularity is specified where needed; the neural vector fields constructed below are smooth. For a linear map $A : T_x\mathbb{H}^d_\kappa \to T_y\mathbb{H}^d_\kappa$, define

$$\|A\|_{\mathrm{op}} = \sup_{\xi\neq 0} \frac{\|A\xi\|_{g_\kappa}}{\|\xi\|_{g_\kappa}}.$$

Singular values are computed in orthonormal tangent frames and are independent of the frames chosen. For $L \geq 1$, the finite-depth maximal growth rate is

$$\lambda_L(x_0) = \frac{1}{L}\log\|D\Phi_L(x_0)\|_{\mathrm{op}}.$$

We use the convention $\log 0 = -\infty$ and write $\lambda_{\max}(x_0) = \lim_{L\to\infty}\lambda_L(x_0)$ only when this limit exists. Otherwise, we state bounds using the limit superior or limit inferior. These quantities measure sensitivity to the input along its trajectory; they do not measure derivatives with respect to model parameters.

Residual directions and scales. Whenever $V_\ell(x_\ell) \neq 0$, let

$$e_\ell = \frac{V_\ell(x_\ell)}{\|V_\ell(x_\ell)\|_{g_\kappa}}, \qquad \Pi_{\ell,\|}\xi = \langle \xi, e_\ell \rangle_{g_\kappa} e_\ell, \qquad \Pi_{\ell,\perp} = I - \Pi_{\ell,\|}.$$

We write $P_\ell = \mathcal{P}_{x_\ell \to x_{\ell+1}}$ for parallel transport along the residual geodesic $\gamma_\ell(t) = \exp^\kappa_{x_\ell}\big(thV_\ell(x_\ell)\big)$, $t \in [0,1]$.. The dimensionless residual displacement and a useful scalar function are

$$z_\ell = \kappa h \|V_\ell(x_\ell)\|_{g_\kappa}, \qquad s(z) = \begin{cases} \sinh(z)/z, & z > 0, \\ 1, & z = 0. \end{cases}$$

When longitudinal and transverse derivative scales are denoted by β and α, respectively, we use $b = h\beta$ and $a = h\alpha$. Curvature comparisons must therefore specify the matched residual and derivative scales.

Normal coordinates. Given an orthonormal frame $U : \mathbb{R}^d \to T_p\mathbb{H}^d_\kappa$, the normal coordinate map is

$$\Psi(q) = \exp^\kappa_p(Uq), \qquad d_\kappa(p, \Psi(q)) = \|q\|.$$

Its differential has radial singular value 1 and transverse singular values $s(\kappa\|q\|)$. Thus its coordinates agree with the metric to first order at the origin. We use such charts around reference trajectories to study finite perturbations. Euclidean coordinate distances between arbitrary points are not identified with geodesic distances.

4 A CURVATURE-DEPENDENT TRANSITION IN SIGNAL PROPAGATION

Let's state the first result.

Theorem 4.1. *Suppose there are constants $v, \alpha, \beta > 0$ such that, at every point of the trajectory (i) $\|V_\ell(x_\ell)\|_{g_\kappa} = v$, (ii) $e_\ell = \dfrac{V_\ell(x_\ell)}{v}$ are preserved by parallel transport and (iii) the covariant derivative has the decomposition*

$$\nabla V_\ell(x_\ell) = -\beta\Pi_{\ell,\|} - \alpha\Pi_{\ell,\perp}, \tag{1}$$

where $\Pi_{\ell,\|}$ projects onto $\operatorname{span}\{e_\ell\}$ and $\Pi_{\ell,\perp} = I - \Pi_{\ell,\|}$. Assume $0 < h\alpha < 2, 0 < h\beta < 1$, and write $a = h\alpha, b = h\beta, z = \kappa h v$. Then the intrinsic singular values of $D\Phi_L(x_0)$ are

$$(1-b)^L \quad \textit{and} \quad |G(z,a)|^L \textit{ with multiplicity } d-1, \tag{2}$$

where $G(z,a) = \cosh z - a\dfrac{\sinh z}{z}$.

Also

$$\lambda_{\max} = \log \max\left\{1 - b, \left|\cosh z - a\frac{\sinh z}{z}\right|\right\}. \tag{3}$$

Proof. See Appendix A. □

The preceding theorem assumes that residual directions are perfectly aligned and that the covariant derivative acts as a scalar on each longitudinal and transverse subspace. We now replace these identities with quantitative error bounds.

The proof uses an elementary invariant-cone argument, a standard approach in the study of linear cocycles (Bochi & Gourmelon (2008)). The contribution here is the explicit translation from geometric deviations to bounds on network sensitivity.

Let $\Pi_{\ell,\|}$ and $\Pi_{\ell,\perp}$ be the orthogonal projections parallel and perpendicular to e_ℓ. Fix reference parameters $a \in (0,2), b \in (0,1), q = 1 - b$. The deviation of the covariant derivative from the reference structure is

$$\Delta_\ell = h\nabla V_\ell(x_\ell) + b\Pi_{\ell,\|} + a\Pi_{\ell,\perp}. \tag{4}$$

Unlike the exact model, Δ_ℓ may contain off-diagonal terms coupling the longitudinal and transverse directions.

Define the reference transverse gain by $G_\ell = G(z_\ell, a) = \cosh z_\ell - a\, s(z_\ell)$.

Theorem 4.2. *Suppose there are constants $Z < \infty$, $\eta \geq 0$, and $\delta \in [0, 2)$ such that, along the entire trajectory,*

$$0 < z_\ell \leq Z,$$
$$\|\Delta_\ell\|_{\mathrm{op}} \leq \eta,$$
$$\|e_{\ell+1} - P_\ell e_\ell\|_{g_\kappa} \leq \delta,$$

where P_ℓ is parallel transport along the ℓth residual geodesic.

Set $K = \sup_{\ell \geq 0} \max\{q, |G_\ell|\}$, $\varepsilon = \delta K + s(Z)\eta$. Then the following statements hold.

1. *If $K + \varepsilon < 1$, then, for every depth L, we have $\|D\Phi_L(x_0)\|_{\mathrm{op}} \leq (K + \varepsilon)^L$. In particular,*
$$\limsup_{L \to \infty} \frac{1}{L} \log \|D\Phi_L(x_0)\|_{\mathrm{op}} \leq \log(K + \varepsilon) < 0.$$

2. *Suppose that $g := \inf_{\ell \geq 0} |G_\ell| > 1$ and $\varepsilon < \min\left\{\frac{g-1}{\sqrt{2}}, \frac{g-q}{2\sqrt{2}}\right\}$. Then, for every depth L, $\|D\Phi_L(x_0)\|_{\mathrm{op}} \geq \left(g - \sqrt{2}\varepsilon\right)^L$. Consequently,*
$$\liminf_{L \to \infty} \frac{1}{L} \log \|D\Phi_L(x_0)\|_{\mathrm{op}} \geq \log\left(g - \sqrt{2}\varepsilon\right) > 0.$$

Proof. See Appendix A. □

This result establishes uniform contraction and expansion in parameter regions separated from criticality. Residual magnitudes may vary across layers, longitudinal and transverse perturbations may couple, and residual directions may turn. Only the turning at each individual step is bounded; the total accumulated rotation need not remain small.

The matrices J_ℓ therefore need not commute or share exact invariant eigenspaces. The expanding cone replaces the perfectly aligned transverse subspace used in the preceding theorem.

The hypotheses are conditions along the actual network trajectory. They do not follow automatically from small changes in network parameters or from approximation of a vector field on a fixed compact set. To connect this result to a neural architecture, we next construct a vector-field parameterization for which displacement, turning, and covariant-derivative errors can be controlled simultaneously.

4.1 A NEURAL REALIZATION OF THE PROPAGATION BOUNDS

The robustness theorem imposes bounds on residual displacements, direction changes, and covariant derivatives along a trajectory. We now construct a finite-width neural parameterization for which these bounds follow directly from constraints on its parameters.

The construction uses the intrinsic residual architecture of Riemannian residual networks Katsman et al. (2023). Its purpose is to realize the hypotheses of Theorem 4.2, rather than to claim a new general residual operation.

Fix a reference input $p_0 \in \mathbb{H}^d_\kappa$ and an orthonormal frame $U_0 : \mathbb{R}^d \to T_{p_0}\mathbb{H}^d_\kappa$. Let $\mathbf{e}_1$ be the first standard basis vector of $\mathbb{R}^d$. Choose constants $h > 0$, $a \in (0, 2)$, $b \in (0, 1)$, $\eta \geq 0$, and velocity bounds $0 < v_- < v_+ < \infty$. Given p_ℓ, U_ℓ, and $v_\ell \in [v_-, v_+]$, define $e_\ell = U_\ell \mathbf{e}_1$, $p_{\ell+1} = \exp^\kappa_{p_\ell}(hv_\ell e_\ell)$. Let P_ℓ denote parallel transport from p_ℓ to $p_{\ell+1}$ along this geodesic, and write $\widetilde{U}_{\ell+1} = P_\ell U_\ell$.

To allow the residual direction to turn, choose $t_\ell \in \mathbb{R}^{d-1}$ with $\|t_\ell\| \leq \tau$, where $\tau \geq 0$. Define the unit vector

$$c_\ell = \frac{(1, t_\ell)}{\sqrt{1 + \|t_\ell\|^2}} \in \mathbb{R}^d.$$

Let R_ℓ be the orthogonal rotation mapping $\mathbf{e}_1$ to c_ℓ, acting as the identity outside their span. When $t_\ell = 0$, take $R_\ell = I$.

Set $U_{\ell+1} = \widetilde{U}_{\ell+1} R_\ell$. This recursively determines all reference points and frames from the model parameters.

The reference input p_0 is fixed. The points p_ℓ and frames U_ℓ are shared by all inputs to the network; they are not recomputed by replacing p_ℓ with each example's hidden state.

Let $D = -b\mathbf{e}_1\mathbf{e}_1^\top - a\left(I - \mathbf{e}_1\mathbf{e}_1^\top\right)$. For a hidden width m, choose matrices $W_\ell \in \mathbb{R}^{d\times m}$, $B_\ell \in \mathbb{R}^{m\times d}$, satisfying $\|W_\ell\|_{\mathrm{op}} \le 1$, $\|B_\ell\|_{\mathrm{op}} \le 1$. Define the tangent-space network

$$f_\ell(q) = v_\ell \mathbf{e}_1 + \frac{1}{h} Dq + \frac{\eta}{h} W_\ell \tanh(B_\ell q), \qquad q \in \mathbb{R}^d, \tag{5}$$

where $\tanh$ acts coordinatewise.

This is an affine skip connection together with a single hidden nonlinear layer. The matrices $W_\ell B_\ell$ need not preserve longitudinal or transverse directions and may differ across layers.

To turn this network into a vector field, let $\mathcal{P}_{p_\ell \to x}$ denote parallel transport along the unique geodesic from p_ℓ to x. Define

$$V_\ell(x) = \mathcal{P}_{p_\ell \to x} U_\ell f_\ell\left(U_\ell^{-1} \log_{p_\ell}^\kappa(x)\right). \tag{6}$$

The corresponding residual layer is $F_\ell(x) = \exp_x^\kappa\big(hV_\ell(x)\big)$.

Both the logarithmic map and radial parallel transport are smooth on hyperbolic space, with their natural extensions at $x = p_\ell$. Consequently, equation 6 defines a globally smooth tangent vector field.

Proposition 4.3. *For the vector fields in equation 6, the network trajectory initialized at $x_0 = p_0$ satisfies $x_\ell = p_\ell$ for every ℓ. Along this trajectory, $V_\ell(p_\ell) = v_\ell e_\ell$ and*

$$h\nabla V_\ell(p_\ell) = U_\ell \left(D + \eta W_\ell B_\ell\right) U_\ell^{-1}.$$

In particular, with projections defined relative toe_ℓ, $\left\|h\nabla V_\ell(p_\ell) + b\Pi_{\ell,\|} + a\Pi_{\ell,\perp}\right\|_{\mathrm{op}} \le \eta$. Moreover, $\kappa h v_- \le z_\ell \le \kappa h v_+$ and

$$\|e_{\ell+1} - P_\ell e_\ell\| \le \delta_\tau, \qquad \delta_\tau = 2\sin\left(\frac{1}{2}\arctan\tau\right).$$

Thus the construction satisfies the pathwise hypotheses of Theorem 4.2, with $Z = \kappa h v_+$, $\delta = \delta_\tau$.

Proof. See Appendix A. □

The bounds can be enforced by parameterization, rather than checked after training. For arbitrary unconstrained matrices $\widehat{W}_\ell$ and $\widehat{B}_\ell$, set

$$W_\ell = \frac{\widehat{W}_\ell}{\sqrt{1 + \|\widehat{W}_\ell\|_{\mathrm{F}}^2}}, \qquad B_\ell = \frac{\widehat{B}_\ell}{\sqrt{1 + \|\widehat{B}_\ell\|_{\mathrm{F}}^2}}.$$

Their operator norms are bounded by their Frobenius norms, which are strictly below one.

For an unconstrained vector $\widehat{t}_\ell \in \mathbb{R}^{d-1}$, use $t_\ell = \tau\widehat{t}_\ell/\sqrt{1 + \|\widehat{t}_\ell\|^2}$. Finally, for an unconstrained scalar ρ_ℓ, set

$$v_\ell = v_- + (v_+ - v_-)\frac{1}{1 + \exp(-\rho_\ell)}.$$

These choices preserve the required bounds for every finite value of the unconstrained parameters. The realization proposition turns the pathwise conditions into conditions on architecture parameters. Define $I = [z_-, z_+] = [\kappa h v_-, \kappa h v_+]$, $q = 1 - b$, and

$$K_I = \max_{z\in I} \max\{q, |G(z,a)|\}, \qquad \varepsilon_I = \delta_\tau K_I + s(z_+)\eta.$$

Corollary 4.4. *For every sequence of normalized neural parameters in the construction above*

1. *If $K_I + \varepsilon_I < 1$, then $\|D\Phi_L(p_0)\|_{\mathrm{op}} \le (K_I + \varepsilon_I)^L$.*

2. *If* $g_I := \min_{z \in I} |G(z, a)| > 1$ *and*

$$\varepsilon_I < \min\left\{ \frac{g_I - 1}{\sqrt{2}}, \frac{g_I - q}{2\sqrt{2}} \right\},$$

then $\|D\Phi_L(p_0)\|_{\mathrm{op}} \geq \left(g_I - \sqrt{2}\varepsilon_I\right)^L$.

Proof. Proposition 4.3 supplies the displacement, derivative, and turning bounds. The constants K_I and ε_I uniformly bound the corresponding quantities along every reference trajectory generated by the construction. Apply Theorem 4.2. □

For further discussion see Appendix B.

5 Empirical Validation

We test four consequences of the theory. The aligned sensitivity threshold, its robustness under layer perturbations, contraction of nearby nonlinear trajectories, and finite-time escape from an expanding reference neighborhood. Figure 1 summarizes these checks. Additional experimental details and the learning experiments appear in Appendix C.

Experimental setup. We use the hyperboloid implementation of the neural construction with dimension $d = 4$, hidden width $m = 8$, residual scale $h = 1$, and damping parameters $a = 0.5$ and $b = 0.3$. Computations use double precision and centered normal coordinates at successive reference points. Reference layer Jacobians are estimated by centered finite differences with step 10^{-4} and checked against the analytic differential. These coordinates are orthonormal at their centers, so the resulting Jacobian products measure intrinsic reference sensitivity. Products are rescaled during accumulation to avoid overflow and underflow.

Aligned transition. We set $v_\ell = 1$ and remove derivative and turning perturbations. For 61 values of $\kappa \in [0.15, 1.8]$, we compare the measured depth-80 growth rate with

$$\frac{1}{L} \log \|D\Phi_L(p_0)\|_{\mathrm{op}} = \log \max\{1 - b, |G(\kappa h, a)|\}.$$

The maximum absolute discrepancy is 2.43×10^{-8}. The predicted threshold is $\kappa_* = 1.0436268956$, obtained from $a = z_* \tanh(z_*/2)$ with $h = v = 1$. The initial plateau at $\log(0.7)$ occurs because the longitudinal gain dominates the operator norm; the subsequent increase reflects the transverse gain becoming dominant. Thus the plot checks both the change of dominant direction and the predicted sign change of the maximal growth rate.

Perturbed layers. We next set $\eta = \tau = 0.01$ and $v_\ell \in [0.98, 1.02]$, generating 20 layer sequences of depth 80 for each of $\kappa = 0.6$ and $\kappa = 1.5$. The certificate constants are computed using conservative interval enclosures of G, rather than grid estimates of its extrema. Every tested depth and seed satisfies the applicable bound within numerical tolerance. The contracting products lie below their upper bound, while the expanding products lie above their lower bound. These observations support the sufficient robustness conditions without asserting that the bounds are sharp or identifying a unique threshold for perturbed networks.

Nonlinear contraction and escape. Keeping the same perturbation budgets, we use the explicit smoothness bound from Appendix B with $R_0 = 0.1$ and one fixed layer sequence generated with seed 88. At $\kappa = 0.6$, the certified radius is $R_c \approx 2.275 \times 10^{-3}$ and the contraction factor is $c \approx 0.8588$. Starting at distance $0.8R_c$ from the reference, the nonlinear trajectory satisfies

$$\frac{d(x_\ell, p_\ell)}{d(x_0, p_0)} \leq c^\ell$$

throughout the 12 tested layers.

At $\kappa = 1.5$, the certified escape radius is $R_e \approx 3.715 \times 10^{-5}$, with transverse growth factor $\mu \approx 1.3305$. For purely transverse initial perturbations with magnitudes $0.5R_e$, $0.1R_e$, $0.01R_e$, and

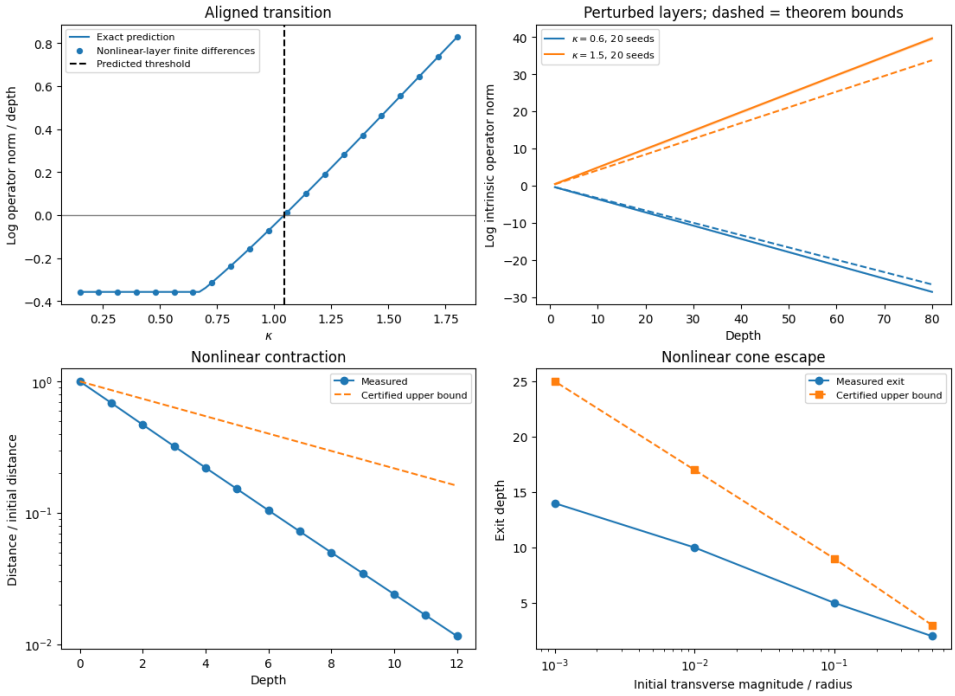


Figure 1: **Top left:** aligned depth-normalized log operator norm at depth 80, compared with the exact prediction; the vertical line marks the predicted threshold. Markers show every third point of the 61-point numerical sweep. **Top right:** log operator norms for perturbed layers; solid curves are means across 20 seeds and shading shows their min-max range. Dashed curves are a contraction upper bound for $\kappa = 0.6$ and an expansion lower bound for $\kappa = 1.5$. **Bottom left:** normalized distance of one nonlinear trajectory to the reference, compared with its certified upper bound at $\kappa = 0.6$. **Bottom right:** measured exit depths and certified upper bounds at $\kappa = 1.5$, for four initial transverse magnitudes.

$0.001R_e$, the measured exit depths are 2, 5, 10, and 14, respectively. The corresponding certified upper bounds are 3, 9, 17, and 25. Each trajectory therefore exits within its predicted depth bound. These are local finite-time checks: the certified radii need not be maximal, and the escape result does not specify the dynamics after leaving the reference neighborhood.

6 Conclusions and Future Work

In this work we established a quantitative relationship between curvature, residual scale, and intrinsic signal propagation in a structured family of hyperbolic neural networks. Along an aligned reference trajectory, the exact differential identifies a unique contraction expansion threshold governed by the dimensionless residual displacement and covariant damping. Sufficient perturbation conditions preserve these regimes, and a constrained neural parameterization realizes the required hypotheses.

The numerical experiments agree with the exact predictions and satisfy the tested sufficient bounds. The learning experiments further demonstrate that nonlinear task performance can improve while the reference trajectory and its complete Jacobian remain unchanged. This distinction clarifies the scope of reference sensitivity. It provides a controlled description of first-order propagation, while higher-order behavior remains relevant to learning.

Several directions follow from these results. First, sharper estimates of the nonlinear remainder could enlarge the certified neighborhoods and clarify the gap between sufficient bounds and observed behavior. Second, extending the analysis to more general covariant derivatives, variable curvature, and other residual constructions would broaden its architectural scope. Such extensions would require accounting for interactions between directions that are absent from the aligned setting.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing. We have not used generative AI tools to prove theoretical or empirical results, and the rest of the required disclosure tasks are not applicable to this work.

REFERENCES

Nico Alvarado and Sebastian Burgos. Convergence properties of hyperbolic neural networks on riemannian manifolds. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*, 2024. URL https://openreview.net/pdf?id=Q2c9linKfv.

Nico Alvarado and Sebastian Burgos. Ergodic and Subhomogeneous Dynamics in Hyperbolic Neural Networks. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2953–2961, 2026. URL https://proceedings.mlr.press/v300/alvarado26a.html.

Jairo Bochi and Nicolas Gourmelon. Some Characterizations of Domination, 2008. URL https://arxiv.org/abs/0808.3811.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic Graph Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 2019. URL https://papers.neurips.cc/paper/8733-hyperbolic-graph-convolutional-neural-networks.pdf.

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)*, pp. 5672–5686. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.389. URL https://aclanthology.org/2022.acl-long.389/.

Fulvio Forni and Rodolphe Sepulchre. A differential Lyapunov framework for contraction analysis. *IEEE Transactions on Automatic Control*, 59(3):614–628, 2014. doi: 10.1109/TAC.2013.2285771. URL https://arxiv.org/abs/1208.2943.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic Neural Networks. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1805.09112.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2018. doi: 10.1088/1361-6420/aa9a90. URL https://arxiv.org/abs/1705.03341.

Soufiane Hayou, Eugenio Clerico, Bobby He, George Deligiannidis, Arnaud Doucet, and Judith Rousseau. Stable ResNet. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1324–1332, 2021. URL https://proceedings.mlr.press/v130/hayou21a.html.

Neil He, Menglin Yang, and Rex Ying. Lorentzian Residual Neural Networks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025. doi: 10.1145/3690624.3709292. URL https://arxiv.org/abs/2412.14695.

Isay Katsman, Eric M. Chen, Sidhanth Holalkere, Anna Asch, Aaron Lou, Ser-Nam Lim, and Christopher De Sa. Riemannian Residual Neural Networks. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2310.10013.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Timo Klein, Thomas Lang, Andrii Shkabrii, Alexander Sturm, Kevin Sidak, Lukas Miklautz, Claudia Plant, Yllka Velaj, and Sebastian Tschiatschek. Understanding and Improving Hyperbolic Deep Reinforcement Learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/94d116f073d7eabf53f8ee0dd9ea2c66-Abstract-Conference.html.

John M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, 2 edition, 2018. doi: 10.1007/978-3-319-91755-9. URL https://doi.org/10.1007/978-3-319-91755-9.

Gal Mishne, Zhengchao Wan, Yusu Wang, and Sheng Yang. The Numerical Stability of Hyperbolic Representation Learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24925–24949, 2023. URL https://proceedings.mlr.press/v202/mishne23a.html.

Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, and Masanori Koyama. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4693–4702. PMLR, 2019. URL https://proceedings.mlr.press/v97/nagano19a.html.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1705.08039.

Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3776–3785, 2018. URL https://proceedings.mlr.press/v80/nickel18a.html.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1711.04735.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1606.05340.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep Information Propagation. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.01232.

Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic Neural Networks++. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.08210.

John W. Simpson-Porco and Francesco Bullo. Contraction Theory on Riemannian Manifolds, 2013. URL https://www.control.utoronto.ca/~jwsimpson/papers/2012za.pdf. Author-hosted preprint dated June 5, 2013.

Wojciech Tarnowski, Piotr Warchoł, Stanisław Jastrzebski, Jacek Tabor, and Maciej A. Nowak. Dynamical Isometry is Achieved in Residual Networks in a Universal Way for any Activation Function. In *International Conference on Artificial Intelligence and Statistics*, 2019. URL https://proceedings.mlr.press/v89/tarnowski19a.html.

Max van Spengler, Erwin Berkhout, and Pascal Mettes. Poincaré ResNet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2303.14027.

Greg Yang and Samuel S. Schoenholz. Mean Field Residual Networks: On the Edge of Chaos. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1712.08969.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup Initialization: Residual Learning Without Normalization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1901.09321.

A Missing Proofs

Proof of Theorem 4.1. We first compute the differential of one layer. For a perturbation $\xi \in T_{x_\ell}\mathbb{H}^d_\kappa$, the endpoint variation is a Jacobi field J along γ_ℓ with initial conditions

$$J(0) = \xi, \qquad \frac{DJ}{dt}(0) = h\nabla_\xi V_\ell.$$

Decompose $\xi = \xi_\parallel + \xi_\perp$ relative to e_ℓ. In a parallel orthonormal frame, the longitudinal and transverse Jacobi equations are $J''_\parallel = 0$, and $J''_\perp - z^2 J_\perp = 0$. By equation 1, their initial derivatives are $-b\xi_\parallel$ and $-a\xi_\perp$. Solving at $t = 1$ gives

$$P_\ell^{-1} DF_\ell(x_\ell)[\xi] = (1-b)\xi_\parallel + G(z,a)\xi_\perp.$$

Define cumulative parallel transport by $U_0 = I$, $U_{\ell+1} = P_\ell U_\ell$. The alignment condition ensures that U_ℓ maps the initial longitudinal and transverse subspaces onto the corresponding subspaces at layer ℓ. Thus, in this common frame, every layer differential has the same diagonal form

$$U_{\ell+1}^{-1} DF_\ell(x_\ell) U_\ell = \operatorname{diag}\big(1-b, G(z,a), \dots, G(z,a)\big).$$

Multiplication across layers gives

$$U_L^{-1} D\Phi_L(x_0) = \operatorname{diag}\big((1-b)^L, G(z,a)^L, \dots, G(z,a)^L\big).$$

Since U_L is an isometry, this proves equation 2 and equation 3.

Furthermore we can determine when the transverse gain has magnitude below one. For $z > 0$,

$$\begin{aligned} |G(z,a)| < 1 &\iff \frac{z(\cosh z - 1)}{\sinh z} < a < \frac{z(\cosh z + 1)}{\sinh z} \\ &\iff z\tanh(z/2) < a < z\coth(z/2). \end{aligned}$$

Because $a < 2$ and $z\coth(z/2) > 2$, the upper inequality holds automatically.

Note that the function $f(z) = z\tanh(z/2)$ is continuous, starts at zero, tends to infinity, and satisfies

$$f'(z) = \tanh(z/2) + \frac{z}{2}\operatorname{sech}^2(z/2) > 0.$$

Hence $f(z) = a$ has exactly one positive solution. Below that solution, $|G(z,a)| < 1$; at the solution, $G(z,a) = 1$; and above it, $G(z,a) > 1$. Since $0 < 1 - b < 1$, these three cases give the signs of the maximal exponent. □

Proof of Theorem 4.2. We first express all layer differentials in compatible orthonormal frames and bound their deviation from the reference diagonal operators.

Choose an isometry $U_0 : \mathbb{R}^d \to T_{x_0}\mathbb{H}^d_\kappa$ whose first basis vector is mapped to e_0. For each layer, choose an orthogonal rotation R_ℓ of $T_{x_{\ell+1}}\mathbb{H}^d_\kappa$ such that $R_\ell P_\ell e_\ell = e_{\ell+1}$.

It can be chosen to act only in the plane spanned by these two unit vectors. For this choice,

$$\|R_\ell - I\|_{\mathrm{op}} = \|e_{\ell+1} - P_\ell e_\ell\| \le \delta.$$

When the vectors coincide, take $R_\ell = I$. The assumption $\delta < 2$ excludes the antipodal case. Define the subsequent frames recursively

$$U_{\ell+1} = R_\ell P_\ell U_\ell.$$

Each U_ℓ is an isometry whose first basis vector is mapped to e_ℓ. In these frames, let

$$J_\ell = U_{\ell+1}^{-1} DF_\ell(x_\ell) U_\ell.$$

The chain rule gives

$$U_L^{-1} D\Phi_L(x_0) U_0 = J_{L-1} \cdots J_0. \tag{7}$$

Now we construct a bound on the geometric perturbation. Note that the Jacobi-field expression for one residual layer is

$$P_\ell^{-1} DF_\ell(x_\ell) = \Pi_{\ell,\|} + \cosh z_\ell \, \Pi_{\ell,\perp} + \big(\Pi_{\ell,\|} + s(z_\ell)\Pi_{\ell,\perp}\big)\, h\nabla V_\ell(x_\ell).$$

Substituting equation 4 yields

$$P_\ell^{-1} DF_\ell(x_\ell) = q\Pi_{\ell,\|} + G_\ell \Pi_{\ell,\perp} + S_\ell \Delta_\ell,$$

where $S_\ell = \Pi_{\ell,\|} + s(z_\ell)\Pi_{\ell,\perp}$. Since s is increasing on $[0,\infty)$, we have $\|S_\ell \Delta_\ell\|_{\mathrm{op}} \le s(Z)\eta$.

Now denote $D_\ell = \mathrm{diag}(q, G_\ell, \ldots, G_\ell)$. In the adapted frames, the change from parallel transport to $R_\ell P_\ell$ contributes an orthogonal factor whose distance from the identity is at most δ. Therefore

$$J_\ell = D_\ell + E_\ell, \qquad \|E_\ell\|_{\mathrm{op}} \le \delta \|D_\ell\|_{\mathrm{op}} + s(Z)\eta \le \varepsilon. \tag{8}$$

Now, equation 8 implies $\|J_\ell\|_{\mathrm{op}} \le K + \varepsilon$. Submultiplicativity and equation 7 gives

$$\|D\Phi_L(x_0)\|_{\mathrm{op}} \le (K+\varepsilon)^L.$$

This proves the contraction statement.

Finally, write a vector in $\mathbb{R}^d$ as $w = (u, v) \in \mathbb{R} \times \mathbb{R}^{d-1}$, where u is longitudinal and v is transverse. Consider the cone $\mathcal{C} = \{(u,v) : |u| \le \|v\|\}$. For $w \in \mathcal{C}$, $\|w\| \le \sqrt{2}\|v\|$. Let $(u', v') = J_\ell(u, v)$. Using equation 8, we obtain

$$|u'| \le q|u| + \varepsilon\|w\| \le (q + \sqrt{2}\varepsilon)\|v\|,$$
$$\|v'\| \ge |G_\ell|\|v\| - \varepsilon\|w\| \ge (g - \sqrt{2}\varepsilon)\|v\|.$$

The first assumption implies $q + \sqrt{2}\varepsilon < g - \sqrt{2}\varepsilon$. Hence $J_\ell \mathcal{C} \subseteq \mathcal{C}$.

The second assumption gives $\mu := g - \sqrt{2}\varepsilon > 1$. Starting from a unit vector $w_0 = (0, v_0)$, repeated cone invariance and transverse expansion yield $\|J_{L-1} \cdots J_0 w_0\| \ge \mu^L$. Taking the operator norm proves the final bound. □

Proof of Proposition 4.3. Since $\log_{p_\ell}^\kappa(p_\ell) = 0$ and $\tanh(0) = 0$, Equation equation 5 gives

$$f_\ell(0) = v_\ell \mathbf{e}_1.$$

Parallel transport from a point to itself is the identity, so

$$V_\ell(p_\ell) = v_\ell U_\ell \mathbf{e}_1 = v_\ell e_\ell.$$

Therefore

$$F_\ell(p_\ell) = \exp_{p_\ell}^\kappa(h v_\ell e_\ell) = p_{\ell+1}.$$

Induction from $x_0 = p_0$ proves that the reference sequence is an actual network trajectory.

We next compute the covariant derivative. Fix $\xi \in T_{p_\ell}\mathbb{H}_\kappa^d$ and consider

$$c(t) = \exp_{p_\ell}^\kappa(t\xi).$$

Along this geodesic,

$$U_\ell^{-1} \log_{p_\ell}^\kappa(c(t)) = t U_\ell^{-1} \xi.$$

Hence

$$V_\ell(c(t)) = \mathcal{P}_{p_\ell \to c(t)} U_\ell f_\ell(t U_\ell^{-1}\xi).$$

The transported frame is parallel along c. Differentiating covariantly at $t = 0$ therefore gives

$$\nabla_\xi V_\ell(p_\ell) = U_\ell D f_\ell(0) U_\ell^{-1} \xi.$$

Because the derivative of coordinatewise $\tanh$ at zero is the identity,

$$hDf_\ell(0) = D + \eta W_\ell B_\ell.$$

This proves the stated derivative formula.

The reference part satisfies

$$U_\ell D U_\ell^{-1} = -b\Pi_{\ell,\|} - a\Pi_{\ell,\perp}.$$

Since U_ℓ is an isometry,

$$\left\|\eta U_\ell W_\ell B_\ell U_\ell^{-1}\right\|_{\mathrm{op}} \le \eta \|W_\ell\|_{\mathrm{op}} \|B_\ell\|_{\mathrm{op}} \le \eta.$$

The displacement bounds follow from $\|V_\ell(p_\ell)\| = v_\ell$. For the turning bound, the angle between $\mathbf{e}_1$ and c_ℓ is

$$\theta_\ell = \arctan \|t_\ell\| \le \arctan \tau.$$

By construction,

$$e_{\ell+1} = \widetilde{U}_{\ell+1} c_\ell, \qquad P_\ell e_\ell = \widetilde{U}_{\ell+1} \mathbf{e}_1.$$

Because $\widetilde{U}_{\ell+1}$ is an isometry,

$$\|e_{\ell+1} - P_\ell e_\ell\| = \|c_\ell - \mathbf{e}_1\| = 2\sin(\theta_\ell/2) \le \delta_\tau.$$

□

B Nonlinear Neighborhood Guarantees and Explicit Smoothness Bounds

For fixed $a \in (0, 2)$, these conditions describe nonempty regimes with nonzero perturbation and turning budgets. Indeed, if $z_+ < z_*(a)$, then $K_I < 1$, and sufficiently small positive η and τ ensure contraction. If instead $z_- > z_*(a)$, then $G(z, a) > 1$ throughout the compact interval I. Thus $g_I > 1$, and sufficiently small positive η and τ ensure expansion.

Perfectly aligned directions and diagonal neural derivatives are therefore not required for the existence of either regime.

The reference trajectory is prescribed through learnable velocities and direction changes. The nonlinear neural correction vanishes at each reference point, but its derivative need not vanish. It can mix longitudinal and transverse directions and vary across layers.

Accordingly, the construction provides a controlled neural realization of the propagation bounds. It does not show that an unconstrained hyperbolic network automatically satisfies them.

The guarantees concern input sensitivity at p_0. They do not imply uniform behavior over a dataset or control parameter gradients through the recursive construction of the reference points.

For any fixed finite depth, strict contraction or expansion at p_0 persists in a sufficiently small input neighborhood by continuity. However, the size of that neighborhood may shrink with depth. A depth-uniform neighborhood requires additional control of the dynamics away from the reference trajectory.

B.1 Neighborhood guarantees uniform in depth

The preceding results control derivatives along the reference trajectory. We now establish what they imply for nearby inputs.

Throughout this subsection, the dimension, hidden width, curvature, step size, and architecture bounds are fixed. In particular, the constants obtained below may depend on these quantities, but not on the layer index or network depth.

The passage from differential contraction to distance contraction is standard in Riemannian stability analysis (Simpson-Porco & Bullo (2013)). Here we give a direct argument for the discrete neural architecture constructed above.

Define $\Psi_\ell(q) = \exp_{p_\ell}^\kappa(U_\ell q)$, $q \in \mathbb{R}^d$. The dynamics in coordinates centered at the reference points are

$$H_\ell(q) = U_{\ell+1}^{-1} \log_{p_{\ell+1}}^\kappa \big(F_\ell(\Psi_\ell(q))\big). \tag{9}$$

Since $F_\ell(p_\ell) = p_{\ell+1}$, we have $H_\ell(0) = 0$. For an arbitrary input trajectory, its centered coordinates satisfy

$$q_{\ell+1} = H_\ell(q_\ell), \qquad x_\ell = \Psi_\ell(q_\ell).$$

Moreover, $d_\kappa(x_\ell, p_\ell) = \|q_\ell\|$. At the origin, the exponential and logarithmic maps have identity differentials in the chosen orthonormal frames. Consequently,

$$DH_\ell(0) = U_{\ell+1}^{-1} DF_\ell(p_\ell) U_\ell.$$

The centered dynamics therefore retain exactly the reference-trajectory Jacobians analyzed previously.

Lemma B.1. *Fix $R_0 > 0$. For the neural architecture of Proposition 4.3, there is a constant $M \geq 1$ such that*

$$\sup_{\ell \geq 0} \sup_{\|q\| \leq R_0} \|D^2 H_\ell(q)\|_{\mathrm{op}} \leq M$$

for every admissible sequence of normalized parameters.

In particular, for $\|q\| \leq R_0$,

$$\|DH_\ell(q) - DH_\ell(0)\|_{\mathrm{op}} \leq M\|q\|,$$
$$\|H_\ell(q) - DH_\ell(0)q\| \leq \frac{M}{2}\|q\|^2.$$

Proof. Hyperbolic isometries preserve exponential maps, logarithmic maps, and parallel transport. An isometry sending p_ℓ and U_ℓ to a fixed origin and frame therefore removes the absolute position of the layer from Equation equation 9.

After this change of coordinates, H_ℓ depends only on the bounded parameters

$$v_\ell \in [v_-, v_+], \qquad \|t_\ell\| \leq \tau, \qquad \|W_\ell\|_{\mathrm{op}} \leq 1, \qquad \|B_\ell\|_{\mathrm{op}} \leq 1,$$

together with the fixed architecture constants.

Because the matrix dimensions are fixed, the closure of this parameter set is compact. The rotation defining the next frame depends smoothly on t_ℓ. Its target direction has strictly positive first coordinate, so no antipodal ambiguity occurs.

The activation $\tanh$ is smooth. Hyperbolic exponential maps, logarithmic maps, and radial parallel transport are also smooth, including their extensions at coincident points. Thus $D^2 H_\ell(q)$ depends continuously on the parameters and on q.

Compactness of the parameter set and the closed ball $\|q\| \leq R_0$ gives a finite common bound. Increase it to at least one and denote it by M.

The remaining inequalities follow by integrating the second derivative along the segment from 0 to q and using $H_\ell(0) = 0$. □

Strict contraction gives a fixed input neighborhood. Retain the constants K_I and ε_I from Corollary 4.4, and suppose

$$\rho := K_I + \varepsilon_I < 1.$$

Choose

$$R_{\mathrm{c}} = \min\left\{R_0, \frac{1-\rho}{2M}\right\}, \qquad c = \rho + MR_{\mathrm{c}}.$$

Then

$$0 < c \leq \frac{1+\rho}{2} < 1.$$

Theorem B.2 (A contracting neighborhood uniform in depth)**.** *For every admissible sequence of neural parameters, each input satisfying*

$$d_\kappa(x_0, p_0) \leq R_{\mathrm{c}}$$

generates a trajectory with

$$d_\kappa(x_\ell, p_\ell) \leq c^\ell d_\kappa(x_0, p_0) \qquad \textit{for every } \ell \geq 0.$$

In particular,

$$F_\ell\big(B_\kappa(p_\ell, R_\mathrm{c})\big) \subseteq B_\kappa(p_{\ell+1}, cR_\mathrm{c}).$$

For any two inputs x_0, y_0 in this neighborhood,

$$d_\kappa\big(\Phi_L(x_0), \Phi_L(y_0)\big) \le C_\mathrm{c} c^L d_\kappa(x_0, y_0),$$

where

$$C_\mathrm{c} = \frac{\sinh(\kappa R_\mathrm{c})}{\kappa R_\mathrm{c}}.$$

Both R_c and C_c are independent of depth.

Proof. The reference-trajectory bound gives

$$\|DH_\ell(0)\|_\mathrm{op} \le \rho.$$

By Lemma B.1, throughout the ball $\|q\| \le R_\mathrm{c}$,

$$\|DH_\ell(q)\|_\mathrm{op} \le \rho + MR_\mathrm{c} = c.$$

Integrating along the segment joining 0 to q, and using $H_\ell(0) = 0$, yields

$$\|H_\ell(q)\| \le c\|q\|.$$

Thus this ball maps into the ball of radius cR_c at the next layer. Induction gives

$$\|q_\ell\| \le c^\ell \|q_0\|.$$

Since $\|q_\ell\| = d_\kappa(x_\ell, p_\ell)$, this proves the first statement.

The same derivative bound implies that, for $q, \widetilde{q}$ in the ball,

$$\|H_\ell(q) - H_\ell(\widetilde{q})\| \le c\|q - \widetilde{q}\|.$$

Both coordinate trajectories remain in the ball, so

$$\|q_L - \widetilde{q}_L\| \le c^L \|q_0 - \widetilde{q}_0\|.$$

It remains to compare coordinate and intrinsic distances. On a hyperbolic normal-coordinate ball of radius R_c, the exponential map has radial singular value one and transverse singular values bounded above by C_c. Its inverse logarithmic map has operator norm at most one.

Integrating these bounds gives

$$\|q_0 - \widetilde{q}_0\| \le d_\kappa(x_0, y_0)$$

and

$$d_\kappa\big(\Psi_L(q_L), \Psi_L(\widetilde{q}_L)\big) \le C_\mathrm{c}\|q_L - \widetilde{q}_L\|.$$

Combining the inequalities proves the pairwise bound. □

This result is stronger than contraction at the reference input alone. It controls every input in a fixed initial ball and every pair of such inputs, uniformly across depth.

It also reveals a limitation for representation learning: persistent strict contraction suppresses input distinctions. The theorem establishes stability, not the usefulness of the resulting representations.

Expansion gives finite-depth escape. The expanding regime has a different implication. A fixed neighborhood can support an expanding cone, but perturbations in that cone cannot remain inside the neighborhood indefinitely.

Suppose

$$g_I := \min_{z \in I} |G(z, a)| > 1$$

and define the admissible error threshold

$$\varepsilon_* = \min\left\{ \frac{g_I - 1}{\sqrt{2}}, \frac{g_I - (1 - b)}{2\sqrt{2}} \right\}.$$

Assume $\varepsilon_I < \varepsilon_*$.

Choose

$$R_{\mathrm{e}} = \min\left\{R_0, \frac{\varepsilon_* - \varepsilon_I}{M}\right\},$$

and set

$$\widehat{\varepsilon} = \varepsilon_I + \frac{M}{2}R_{\mathrm{e}}, \qquad \mu = g_I - \sqrt{2}\widehat{\varepsilon}.$$

These choices ensure

$$\widehat{\varepsilon} < \varepsilon_*, \qquad \mu > 1.$$

Theorem B.3. *Write centered coordinates as $q(u, v) \in \mathbb{R} \times \mathbb{R}^{d-1}$ and define*

$$\mathcal{C} = \{(u, v) : |u| \le \|v\|\}.$$

For every nonzero $q_0 = (u_0, v_0) \in \mathcal{C}$ with $\|q_0\| \le R_{\mathrm{e}}$, consider the first exit time

$$T_{\mathrm{exit}} = \inf\{\ell \ge 1 : \|q_\ell\| > R_{\mathrm{e}}\}.$$

Then T_{exit} is finite and satisfies

$$T_{\mathrm{exit}} \le \left\lfloor \frac{\log(R_{\mathrm{e}}/\|v_0\|)}{\log \mu} \right\rfloor + 1.$$

Until exit, the centered trajectory remains in $\mathcal{C}$ and its transverse component satisfies

$$\|v_\ell\| \ge \mu^\ell \|v_0\|.$$

Proof. At the reference point, the centered Jacobian has the form

$$DH_\ell(0) = D_\ell + E_\ell, \qquad D_\ell = \mathrm{diag}(1 - b, G_\ell, \ldots, G_\ell),$$

with $\|E_\ell\|_{\mathrm{op}} \le \varepsilon_I$.

Lemma B.1 gives

$$H_\ell(q) = D_\ell q + E_\ell q + \mathcal{R}_\ell(q), \qquad \|\mathcal{R}_\ell(q)\| \le \frac{M}{2}\|q\|^2.$$

For $\|q\| \le R_{\mathrm{e}}$, it follows that

$$\|E_\ell q + \mathcal{R}_\ell(q)\| \le \widehat{\varepsilon}\|q\|.$$

Suppose $q = (u, v) \in \mathcal{C}$. Then $\|q\| \le \sqrt{2}\|v\|$. Writing $H_\ell(q) = (u', v')$, we obtain

$$|u'| \le \big(1 - b + \sqrt{2}\widehat{\varepsilon}\big)\|v\|,$$
$$\|v'\| \ge \big(g_I - \sqrt{2}\widehat{\varepsilon}\big)\|v\| = \mu\|v\|.$$

The choice $\widehat{\varepsilon} < \varepsilon_*$ ensures

$$1 - b + \sqrt{2}\widehat{\varepsilon} < g_I - \sqrt{2}\widehat{\varepsilon}.$$

Hence the image remains in the cone, and its transverse component grows by at least μ.

Iterating while the trajectory is inside the radius-R_{e} ball gives $\|v_\ell\| \ge \mu^\ell\|v_0\|$. Since q_0 is nonzero and belongs to $\mathcal{C}$, we have $\|v_0\| > 0$.

If the trajectory remained in the ball through depth N, then

$$\mu^N\|v_0\| \le \|q_N\| \le R_{\mathrm{e}}.$$

This is impossible once $N > \log(R_{\mathrm{e}}/\|v_0\|)/\log\mu$. The asserted exit time bound follows. □

The two regimes now have finite amplitude meanings. In the contractive regime, a depth-independent input neighborhood remains close to the reference trajectory, and pairwise distances decay exponentially. In the expansive regime, arbitrarily small nonzero cone perturbations eventually leave a fixed reference tube, with an upper bound on exit depth that depends logarithmically on their initial transverse magnitude.

The radii above are existence guarantees expressed through a uniform second-derivative bound. They may be conservative, and they are not uniform as the stability margin approaches zero.

Obtaining useful numerical radii requires an explicit estimate of M for the selected architecture. Neither these neighborhood guarantees nor the reference-trajectory results establish improved optimization or generalization.

B.2 EXPLICIT NEIGHBORHOOD CERTIFICATES

The preceding neighborhood results depend on a uniform bound for the second derivatives of the centered maps. For our neural construction, this bound can be expressed directly in terms of the architecture parameters.

Proposition B.4. *Consider the neural construction of Proposition 4.3, with*

$$\|W_\ell\|_{\mathrm{op}} \le 1, \qquad \|B_\ell\|_{\mathrm{op}} \le 1, \qquad 0 < v_\ell \le v_+.$$

Fix $\kappa > 0$ and $R_0 > 0$, and define

$$A = 1 + \max\{a, b\} + \eta, \qquad \mathcal{E} = \exp\big(\kappa\big(AR_0 + 2hv_+\big)\big).$$

Then the centered maps H_ℓ satisfy

$$\sup_{\ell \ge 0} \sup_{\|q\| \le R_0} \|D^2 H_\ell(q)\|_{\mathrm{op}} \le M_{\exp},$$

where

$$M_{\exp} = \kappa A^2 \left(3\mathcal{E}^2 + \mathcal{E}\right) + 2\eta\mathcal{E}.$$

Proof. Fix a layer and suppress its index. By an isometry, identify its reference point and frame with the origin and standard frame of the hyperboloid model. Write

$$D = -be_1e_1^\top - a(I - e_1e_1^\top).$$

The tangent displacement used by the residual layer is

$$hf(q) = hve_1 + Dq + \eta W \tanh(Bq).$$

Introduce the dimensionless displacement

$$w(q) = \kappa\big(hve_1 + Dq + \eta W \tanh(Bq)\big).$$

Since $\tanh$ acts componentwise, its first derivative has operator norm at most one. Its second derivative satisfies

$$\|D^2 \tanh(z)[u_1, u_2]\| \le 2\|u_1\|\|u_2\|.$$

Indeed, $|\tanh''(t)| \le 2$, and

$$\|u_1 \odot u_2\| \le \|u_1\|\|u_2\|,$$

where $\odot$ denotes componentwise multiplication. Consequently, for $\|q\| \le R_0$,

$$\begin{aligned} \|w(q)\| &\le \kappa\big(hv_+ + (A-1)R_0\big), \\ \|Dw(q)\|_{\mathrm{op}} &\le \kappa(A-1), \\ \|D^2 w(q)\|_{\mathrm{op}} &\le 2\kappa\eta. \end{aligned}$$

We next express the layer through Lorentz boosts. For $t \in \mathbb{R}^d$, define

$$\mathcal{A}(t) = \begin{pmatrix} 0 & t^\top \\ t & 0 \end{pmatrix}, \qquad \mathcal{B}(t) = \exp(\mathcal{A}(t)).$$

The exponential here is the matrix exponential. If $r = \|t\|$, then

$$\mathcal{B}(t) = \begin{pmatrix} \cosh r & \dfrac{\sinh r}{r} t^\top \\ \dfrac{\sinh r}{r} t & I + \dfrac{\cosh r - 1}{r^2} tt^\top \end{pmatrix},$$

with the coefficients extended continuously at $r = 0$.

Let $o = (1, 0, \ldots, 0)^\top$. In dimensionless hyperboloid coordinates, $\mathcal{B}(t)o$ represents the point with normal coordinate t. The restriction of $\mathcal{B}(t)$ to the tangent space at o gives parallel transport along the radial geodesic. Thus the layer output before recentering is represented by

$$Y(q) = \mathcal{B}(\kappa q)\mathcal{B}(w(q))o.$$

This identity includes both radial parallel transport in the definition of the vector field and the subsequent residual exponential map.

The next reference point is $\mathcal{B}(\kappa hve_1)o$. Let $\pi : \mathbb{R}^{d+1} \to \mathbb{R}^d$ denote spatial projection, and set

$$z(q) = \pi\mathcal{B}(-\kappa hve_1)Y(q).$$

Define

$$\mathcal{L}(z) = \begin{cases} \dfrac{\operatorname{arsinh}\|z\|}{\|z\|} z, & z \neq 0, \\ 0, & z = 0. \end{cases}$$

If R is the orthogonal frame rotation used in the construction, the centered map is exactly

$$H(q) = \frac{1}{\kappa} R^\top \mathcal{L}(z(q)).$$

For a fixed layer, both the recentering boost and R are constant with respect to q.

All matrix norms in the following estimates are Euclidean operator norms. Since $\|\mathcal{A}(t)\|_{\mathrm{op}} = \|t\|$, the power series for the matrix exponential gives

$$\begin{aligned} \|\mathcal{B}(t)\|_{\mathrm{op}} &\le \exp(\|t\|), \\ \|D\mathcal{B}(t)[u]\|_{\mathrm{op}} &\le \exp(\|t\|)\|u\|, \\ \|D^2\mathcal{B}(t)[u_1, u_2]\|_{\mathrm{op}} &\le \exp(\|t\|)\|u_1\|\|u_2\|. \end{aligned}$$

For example, differentiating a term of degree n twice produces $n(n-1)$ terms; their norm bounds sum to $\exp(\|t\|)\|u_1\|\|u_2\|$.

Applying the product and chain rules to z yields

$$\begin{aligned} \|Dz(q)\|_{\mathrm{op}} &\le \mathcal{E}\,\kappa A, \\ \|D^2 z(q)\|_{\mathrm{op}} &\le \mathcal{E}\left(\kappa^2 A^2 + 2\kappa\eta\right). \end{aligned}$$

The common exponential factor follows from

$$\kappa hv + \kappa\|q\| + \|w(q)\| \le \kappa\big(2hv_+ + AR_0\big).$$

It remains to bound the Euclidean derivatives of $\mathcal{L}$. Its radial and tangential derivative eigenvalues, at radius r, are respectively

$$\frac{1}{\sqrt{1+r^2}}, \qquad \frac{\operatorname{arsinh} r}{r}.$$

Both are at most one, so

$$\|D\mathcal{L}(z)\|_{\mathrm{op}} \le 1.$$

For the second derivative, use the representation

$$\mathcal{L}(z) = \int_0^1 \frac{z}{\sqrt{1+t^2\|z\|^2}}\, dt.$$

Differentiating the integrand twice and writing $r = \|z\|$ gives the bound

$$\left\| D^2 \left(\frac{z}{\sqrt{1+t^2\|z\|^2}} \right) \right\|_{\mathrm{op}} \le \frac{3t^2 r}{(1+t^2r^2)^{3/2}} + \frac{3t^4 r^3}{(1+t^2r^2)^{5/2}} \le 6t.$$

Therefore

$$\|D^2\mathcal{L}(z)\|_{\mathrm{op}} \le 3.$$

Finally, orthogonality of R and the chain rule imply

$$\begin{aligned} \|D^2 H(q)\|_{\mathrm{op}} &\le \frac{1}{\kappa}\left(3\|Dz(q)\|_{\mathrm{op}}^2 + \|D^2 z(q)\|_{\mathrm{op}}\right) \\ &\le \kappa A^2(3\mathcal{E}^2 + \mathcal{E}) + 2\eta\mathcal{E}. \end{aligned}$$

The estimate is uniform over all admissible layers. □

Now set $\overline{M} = \max\{1, M_{\exp}\}$. In the contracting regime, let $\rho = K_I + \varepsilon_I < 1$. Then Theorem B.2 applies with

$$R_{\mathrm{c}} = \min\left\{R_0, \frac{1-\rho}{2\overline{M}}\right\}, = \rho + \overline{M}R_{\mathrm{c}} < 1.$$

In the expanding regime, define

$$\varepsilon_* = \min\left\{\frac{g_I - 1}{\sqrt{2}}, \frac{g_I - (1-b)}{2\sqrt{2}}\right\}.$$

Whenever $\varepsilon_I < \varepsilon_*$, Theorem B.3 applies with

$$R_{\mathrm{e}} = \min\left\{R_0, \frac{\varepsilon_* - \varepsilon_I}{\overline{M}}\right\}, \qquad \mu = g_I - \sqrt{2}\left(\varepsilon_I + \frac{\overline{M}R_{\mathrm{e}}}{2}\right) > 1.$$

Thus both the certified neighborhood radii and the associated escape-time bound can be evaluated from the architecture parameters.

These radii are sufficient certificates and need not approximate the largest neighborhoods on which the conclusions hold. In particular, the exponential factors arise from conservative ambient matrix norm estimates. A small certified radius does not, by itself, establish a small actual basin of contraction or expansion.

C Full Numerical Experiments

C.1 Numerical verification of intrinsic propagation

We implement the layers in the hyperboloid model and express their local maps in normal coordinates centered at successive reference points. Calculations use double precision. Unless otherwise stated, the numerical checks use dimension $d = 4$, hidden width $m = 8$, residual scale $h = 1$, transverse damping $a = 0.5$, and longitudinal damping $b = 0.3$.

We first compare the analytic reference differential with centered finite differences of the nonlinear layer. For a representative perturbed layer with $\kappa = 0.8$, derivative budget $\eta = 0.03$, and turning budget $\tau = 0.04$, decreasing the finite-difference step from 10^{-2} to 10^{-4} reduces the discrepancy from approximately 1.16×10^{-6} to 1.16×10^{-10}. At step 10^{-5}, the discrepancy is approximately 2.71×10^{-11}. We also check the full singular-value formula for aligned products of depth eight.

For the aligned transition, we hold $h = v = 1$ and sweep 61 curvature values in $[0.15, 1.8]$ at depth $L = 80$. The theoretical threshold is

$$\kappa_* = z_* \approx 1.0436268956.$$

We compare the measured intrinsic growth

$$\lambda_L(p_0) = \frac{1}{L}\log\|D\Phi_L(p_0)\|_{\mathrm{op}}$$

with

$$\log\max\{1-b, |G(\kappa h v, a)|\}.$$

The maximum discrepancy across the sweep is approximately 2.43×10^{-8}. Matrix products are repeatedly rescaled during accumulation to avoid overflow and underflow.

We next test the robust sufficient bounds at $\kappa = 0.6$ and $\kappa = 1.5$ using 20 random seeds and 80 layers per seed. Here $\eta = \tau = 0.01$ and $v_\ell \in [0.98, 1.02]$. The gain bounds are obtained from conservative interval enclosures, rather than extrema estimated from a grid. Every tested depth satisfies the corresponding certified upper or lower bound within numerical tolerance.

Finally, we evaluate finite-amplitude consequences of the neighborhood results. For the contracting configuration, sampled reference and pairwise distances satisfy the predicted bounds over 12 layers. For an expanding configuration, initial transverse magnitudes equal to 0.5, 0.1, 0.01, and 0.001 times the certified escape radius produce exit depths 2, 5, 10, and 14, respectively. The corresponding theoretical upper bounds are 3, 9, 17, and 25. These checks validate the stated sufficient estimates in the tested configurations; they do not identify maximal contracting neighborhoods or sharp escape times.

C.2 Learning protocol and sensitivity measurements

The learning experiments use $d = 4$, $m = 8$, $h = 1$, $a = 0.5$, $b = 0.3$, $\eta = 0.05$, $\tau = 0.02$, and $v_\ell \in [0.99, 1.01]$. We consider

$$\kappa \in \{0.6, \kappa_*, 1.5\}, \qquad L \in \{2, 6, 12\}.$$

The middle value is the threshold of the unperturbed aligned construction. It is not asserted to remain a zero-growth configuration after perturbation or training.

Inputs have the form

$$x = \exp_{p_0}^{\kappa}(U_0 q), \qquad q \sim \mathrm{Unif}\big(B_{\mathbb{R}^4}(0, r_{\mathrm{in}})\big), \qquad r_{\mathrm{in}} = 0.02.$$

The reference centers and frames are shared by all examples. They are not recomputed separately for each input.

We generate 512 training examples and 1,024 test examples. Before independent label flips with probability 0.05, the two classification tasks have labels

$$y_{\mathrm{lin}} = \mathbf{1}\{0.4q_1 + q_2 + 0.2q_3 > 0\}, \qquad y_{\mathrm{xor}} = \mathbf{1}\{q_2 q_3 > 0\}.$$

The same input samples and label-flip mask are used across the two tasks and all experimental conditions. The first coordinate is longitudinal at the reference point; the remaining coordinates are transverse.

A linear classifier acts on the final normal coordinate

$$q_L = U_L^{-1} \log_{p_L}^{\kappa}\big(\Phi_L(x)\big).$$

Its logits are

$$\frac{w^\top q_L}{c_{\mathrm{read}}} + b_{\mathrm{read}}, \qquad c_{\mathrm{read}} = r_{\mathrm{in}} \|D\Phi_L(p_0; \theta_{\mathrm{init}})\|_{\mathrm{op}}.$$

This scalar calibration is computed at initialization and held fixed throughout training. All reported backbone sensitivities exclude the classifier and its calibration.

Each run uses 120 full-batch Adam updates with learning rate 0.03 (Kingma & Ba, 2015). We use three initialization seeds and match raw initializations across curvature values and training conditions. Networks of different depths share the same initial layer prefixes and classifier initialization. For each configuration, we compare training the original backbone and classifier with training only the classifier while freezing the backbone. This gives 108 network runs.

We measure reference sensitivity analytically and check it against automatic differentiation. Away from the reference trajectory, input and output normal-coordinate metric factors are included when computing intrinsic operator norms. Sensitivity measurements use eight fixed test inputs during training and 16 at the final checkpoint. Accuracy uses the full test split. We report the final checkpoint, without selecting an epoch using test accuracy.

Direct-input linear and 16-hidden-unit tanh classifiers provide task baselines. They receive q/r_{in} and use the same optimizer and update budget. These baselines are not parameter-matched to the hyperbolic networks.

C.3 Depth and task dependence

Table 1 reports mean test accuracy when training the original backbone and classifier. On the linear task, the preferred curvature depends on depth. For example, the largest mean accuracy among the three curvatures occurs at κ_* for depth two and at $\kappa = 0.6$ for depth twelve. The expanding configuration at $\kappa = 1.5$ decreases from 88.64% at depth two to 78.06% at depth twelve.

The original parameterization performs poorly on the XOR-type task across the tested depths and curvatures. Its largest mean accuracy is 54.17%. By comparison, the direct-input tanh baseline achieves $88.70\% \pm 0.48\%$, while the direct-input linear classifier achieves 50.10%. On the linear task, these baselines achieve $92.42\% \pm 0.34\%$ and $86.39\% \pm 1.21\%$, respectively. Here and below, the reported deviations are standard deviations across initialization seeds.

Task	Depth	$\kappa = 0.6$	$\kappa = \kappa_*$	$\kappa = 1.5$
Linear	2	86.98	89.97	88.64
Linear	6	87.60	88.80	85.16
Linear	12	89.16	88.31	78.06
XOR-type	2	50.13	50.07	49.67
XOR-type	6	50.10	50.52	53.81
XOR-type	12	50.10	54.17	52.44

Table 1: Mean final test accuracy (%) across three initialization seeds for the original trainable parameterization. The data, optimization budget, initial layer prefixes, and classifier initialization are matched across the corresponding conditions. The value κ_* denotes the aligned theoretical threshold, not a certified trained critical point.

Curvature	Readout only	Fixed reference, $\gamma = 1$	Fixed reference, $\gamma = 4$
0.6	50.00 ± 0.00	70.93 ± 0.88	83.17 ± 0.91
κ_*	50.20 ± 0.00	81.41 ± 0.95	86.39 ± 0.92
1.5	50.00 ± 0.29	83.24 ± 0.25	88.38 ± 1.95

Table 2: Final XOR-type test accuracy (%, mean $\pm$ standard deviation across three initialization seeds) at depth six. For each matched seed and curvature, all three conditions retain the same reference trajectory and reference Jacobian throughout training. Only the correction conditions can change the nonlinear backbone map away from the reference.

The normalization constraints and applicable reference-growth certificates hold at every recorded training checkpoint. Thus, the weak nonlinear-task performance occurs even when the prescribed reference conditions remain satisfied. This is an empirical limitation of the tested parameterization and optimization protocol, rather than an impossibility result for the architecture. In particular, preserving a reference contraction or expansion regime does not by itself ensure successful nonlinear classification.

C.4 Learning with a fixed reference Jacobian

We next evaluate the proposed correction at depth six on the same XOR-type task. In local coordinates, the additional term is

$$r_\ell(q) = \gamma r\, C_\ell \left[\tanh\Big(A_\ell^{(1)} q/r\Big) \odot \tanh\Big(A_\ell^{(2)} q/r\Big)\right],$$

and the local residual field receives $r_\ell(q)/h$. We use correction width eight, $r = 0.02$, and strengths $\gamma \in \{1, 4\}$. All three matrices satisfy the operator-norm constraints specified by the construction. The output matrix is initialized to zero, so the augmented and original networks initially implement the same function. The projection matrices are randomly initialized without using the label rule.

We compare two correction-training conditions. In the *augmented* condition, the original backbone, correction, and classifier are trainable. In the *fixed-reference* condition, the original backbone is frozen and only the correction and classifier are trained. A readout-only control freezes the original backbone and omits the correction. Together with the original trainable model, the three curvatures, and three seeds, these conditions give 54 network runs.

Because

$$r_\ell(0) = 0, \qquad Dr_\ell(0) = 0,$$

the fixed-reference condition preserves both the reference trajectory and the full reference Jacobian product throughout training. We check this identity using automatic differentiation at every recorded checkpoint. The original backbone parameters remain exactly unchanged in this condition.

Table 2 shows substantial improvements over the readout-only controls. With $\gamma = 4$, fixed-reference training achieves 83.17%, 86.39%, and 88.38% mean accuracy at $\kappa = 0.6$, κ_*, and $\kappa = 1.5$, respectively. Allowing the original backbone to train as well gives 84.38%, 87.04%, and 88.67% for the same correction strength. These results demonstrate improved learning within the tested family while preserving the complete reference Jacobian in the fixed-reference condition.

The off-reference sensitivities nevertheless change. For example, at $\kappa = 0.6$, the mean reference growth remains approximately -0.3566 in the fixed-reference conditions. The mean across seeds of the median sampled test-input growth changes from approximately -0.3566 for the readout-only control to -0.1922 with $\gamma = 4$. Preserving the reference differential therefore does not preserve sensitivity throughout the input distribution.

The neighborhood certificates also require updating. For $\kappa = 0.6$ and the ambient coordinate radius $R_0 = 0.1$, the explicit contracting-radius bound decreases from approximately 1.81×10^{-3} for the original model to 6.97×10^{-5} for $\gamma = 1$ and 8.75×10^{-6} for $\gamma = 4$. The reference certificate is unchanged, but the larger smoothness bound yields a smaller certified neighborhood. These are sufficient radii, not estimates of the actual contraction basin.

The numerical checks support the exact differential formulas and their stated sufficient consequences. The learning experiments establish a more specific observation: within this controlled construction, nonlinear task performance can improve substantially while the reference trajectory and Jacobian remain fixed.

The learning study uses one synthetic dataset per task, three initialization seeds, and a fixed optimization budget. The reported seed deviations do not quantify uncertainty over independently sampled datasets. Hyperparameters are not independently optimized for each model, and the direct-input baselines are not parameter-matched. Moreover, off reference sensitivity is sampled on a small set of test inputs, and the input radius exceeds the conservative certified contracting radii discussed above.

D LIMITATIONS

Our guarantees concern a structured family of hyperbolic residual networks. The exact threshold requires alignment of the reference directions and a prescribed form of the covariant derivative. The perturbation results relax these assumptions through explicit sufficient inequalities, but they do not characterize all contracting or expanding configurations. In particular, they do not establish a unique threshold for general perturbed networks. The neural parameterization enforces the required conditions by construction, which restricts the admissible residual fields.

The reference guarantees are local in their interpretation. A negative reference growth rate yields a contracting neighborhood only after controlling nonlinear remainders. Similarly, the escape theorem applies to initial perturbations in a specified cone and describes their evolution only until they leave the certified neighborhood. The explicit smoothness estimates can produce conservative radii, whose sharpness remains unresolved. Moreover, zero maximal growth at the aligned threshold does not imply dynamical isometry, since the longitudinal direction continues to contract.

Preserving the reference trajectory and its Jacobian does not preserve the network away from that trajectory. The nonlinear correction changes higher-order behavior and requires recomputing the neighborhood bounds. Our analysis does not provide guarantees for optimization, classification accuracy, or generalization, nor does it identify a universally preferred curvature.

The learning experiments use low-dimensional synthetic tasks, a single sampled dataset per task, three initialization seeds, and a fixed optimization budget. The baselines are not parameter-matched, and sensitivity away from the reference is evaluated on a limited set of inputs. The results therefore establish controlled examples of the relationship between reference sensitivity and learning, rather than a general performance advantage. Evaluation on independently sampled datasets, larger architectures, and tasks with hierarchical structure remains necessary to assess broader practical relevance.