Ergodic and Subhomogeneous Dynamics in Hyperbolic Neural Networks

Nico Alvarado
Centro Nacional de
Inteligencia Artificial

Sebastian Burgos
The Pennsylvania State University

Abstract

We analyze the long term behavior of hyperbolic neural networks through sub-homogeneous layer maps, focusing on stability, growth control, and robustness under stochastic perturbations. This work unifies the standard hyperbolic models via explicit isometries and Möbius operations, allowing statements to be transported across representations without loss of geometric meaning. Within this model invariant view, we study iterated, noise perturbed transformations and develop an ergodic theoretic framework that characterizes their asymptotic behavior, including conditions that promote stability and convergence of averaged iterates. Beyond theory, these insights inform practical design choices for training procedures that remain well behaved in the presence of noise and avoid unbounded parameter growth, thereby supporting more reliable use of hyperbolic representations in hierarchical and graph-structured learning tasks.

1 INTRODUCTION

Hyperbolic Neural Networks (HNNs) have garnered significant attention in recent years due to their remarkable ability to represent and process data with hierarchical structures. The curvature and geometry of hyperbolic spaces naturally capture tree-like or hierarchical relationships more efficiently than their Euclidean counterparts [Alvarado et al., 2023, Hamann, 2018, Chami et al., 2020, Nickel and Kiela, 2017]. This geometric advantage has facilitated a wide range of applications, including natural language processing with syntactic parse trees, hierarchical social network analysis, and knowledge graph embeddings [Chami et al., 2019, Ganea et al., 2018b, Liu et al., 2019, Yang et al., 2022, Sala et al., 2018]. Additionally, hyperbolic spaces have been shown to enable compact representations while preserving meaningful distance relationships at various scales, leading to improved interpretability and performance in tasks involving hierarchical structures [Nickel and Kiela, 2018, Peng et al., 2021, Rodríguez-Flores and Papadopoulos, 2020].

Despite their empirical success, the theoretical foundations of HNNs remain only partially understood. Most existing analyses focus on the geometry of the feature space and practical optimization strategies, offering insights into why hyperbolic embeddings are effective [Ganea et al., 2018a, Gu et al., 2019]. However, fundamental questions regarding the dynamical behavior of layer transformations, particularly in the presence of noise or under sub-homogeneous constraints, remain largely unexplored. In deep learning settings, the iterative nature of layer transformations combined with the non-Euclidean geometry introduces significant analytical challenges, particularly when investigating long-term stability, convergence, and ergodic properties [Khrulkov et al., 2020, Alvarado and Lobel, 2024].

One of the central themes of this work is the study of sub-homogeneous layer maps, a mathematical formalism that provides a way to control the growth of transformations by imposing scaling constraints [Sittoni and Tudisco, 2024]. In the context of neural networks, sub-homogeneous layers ensure that weight updates remain stable over time, preventing unbounded parameter growth that could lead to numerical instabilities or poor generalization in large-scale architectures. This stability is particularly relevant in hyperbolic neural networks, where exponential distance distortions due to negative curvature necessitate careful control over transformation magnitudes.

When random noise or stochastic perturbations are introduced, the analysis becomes even more intricate. Stochasticity in neural networks arises naturally from various sources, including gradient noise in stochastic

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

optimization algorithms, weight initialization variability, and environmental fluctuations [Zhou et al., 2017]. While noise can aid in escaping sharp local minima, it also raises important theoretical questions regarding the stability of network states and their asymptotic behavior under repeated transformations [Zhou et al., 2019].

The theory of dynamical systems and ergodic theory provides powerful tools to analyze the long-term behavior of iterated transformations. Recent works have extended these ideas to non-Euclidean geometries, leading to subadditive and multiplicative ergodic theorems on manifolds [Avelin and Karlsson, 2022, Alvarado and Burgos, 2024]. These results offer a rigorous framework for understanding how repeated applications of transformations affect points in curved spaces, with potential implications for convergence to fixed points, periodic orbits, or more complex dynamical regimes [Zhang, 2023].

In hyperbolic spaces, curvature plays a crucial role in ensuring contraction properties and imposing subadditive bounds on transformations. By leveraging these geometric features—often through Möbius operations—one can derive refined results on the almost sure convergence of stochastic processes [Ganea et al., 2018b, Liu et al., 2019, Ungar, 2008]. Our work builds upon these foundational studies and extends them to the context of HNNs by investigating the asymptotic behavior of sub-homogeneous and noise-perturbed layer maps. Specifically, we aim to establish conditions under which these transformations exhibit stability, ergodicity, or convergence, thereby contributing to the theoretical understanding of hyperbolic neural networks and their broader implications in machine learning.

1.1 Contributions

Our main contribution is a representation invariant theory of stability for noisy, sub-homogeneous dynamics across all standard hyperbolic models (Lorentz, Poincaré, Klein, upper half-plane). We prove unconditional convergence guarantees under bounded noise and time-varying scalings and show that three simple diagnostics—stepwise geodesic distance, sliding-window diameter, and averaged tangent norm—collapse to the same decay profile under isometric conjugations, yielding a model-agnostic certification pipeline. Unlike prior hyperbolic learning work that fixes one model [Alvarado and Burgos, 2024] and reports mode-specific behavior, our results and tests are basepoint agnostic and representation invariant by construction. On the theory side, we extend sub-homogeneous dynamical analyses from cone metrics to hyperbolic geometry with explicit noise and scaling regimes. Specifically

- We formalize isometries linking standard hyperbolic models and prove that asymptotic growth descriptors for iterated sub-homogeneous maps are representation-invariant.
- We establish almost-sure convergence of ergodic averages for sequences of layer maps with bounded perturbations and bounded time varying scalings, yielding well-behaved long run dynamics.
- We propose noise-handling regimes and empirically verify stabilization and cross-model invariance on learned contractive surrogates.
- We cast common HNN layers as order-preserving, sub-homogeneous maps in tangent charts, unifying theory and practice and providing concrete guidance for stable training in hyperbolic spaces.

Overall, this work contributes a rigorous theoretical framework for understanding the dynamics of HNNs under realistic settings, laying the groundwork for more robust and efficient hyperbolic architectures. The results not only deepen our insight into convergence properties but also have practical implications for the design of training algorithms that are resilient to noise and parameter growth.

2 PRELIMINARIES

Hyperbolic spaces. Hyperbolic geometry exhibits distinct geometric features that require a deep understanding of its fundamental principles, particularly the hyperbolic parallel postulate and the concept of negative curvature [Gromov, 1987]. Unlike in Euclidean space, where parallel lines remain equidistant, hyperbolic geometry allows multiple distinct lines to pass through a given point without intersecting a reference line, leading to a radically different notion of parallelism. This unique behavior stems from the intrinsic curvature of hyperbolic space, which profoundly affects distances, angles, and geodesic structures.

Furthermore, the representation and visualization of hyperbolic space play a crucial role in both theoretical and applied settings. Various models, such as the Poincaré disk, the Poincaré half-space, and the Lorentz hyperboloid model, provide different perspectives on hyperbolic geometry, each with its own advantages depending on the context. The Lorentz hyperboloid model, in particular, is essential for understanding hyperbolic embeddings in machine learning and theoretical physics, as it naturally arises from the Minkowski space formulation and provides a convenient framework for computations involving isometries and distance preservation. Mastering these representations is key to developing an intuitive and rigorous understanding of hyperbolic spaces and their applications [Shavitt and Tankel, 2008, Billera et al., 2001].

Isometries between Hyperbolic manifolds.

Definition 2.1. A map $\varphi\colon \mathcal{H}_1^n \to \mathcal{H}_2^n$, is called a hyperbolic isometry if it preserves the hyperbolic distances $d_{\mathcal{H}_1}$ and $d_{\mathcal{H}_2}$ between any two points in the hyperbolic manifold, i.e.,

$$d_{\mathcal{H}_2}(\varphi(x),\varphi(y)) = d_{\mathcal{H}_1}(x,y), \forall x, y \in \mathcal{H}_1^n.$$

In the literature, there exist three widely studied isometric models of hyperbolic space: the Poincaré ball model, the Lorentz hyperboloid model, and the upper-half plane model. In this work, we consider the following four models of n-dimensional hyperbolic space, each equipped with its respective metric.

Definition 2.2. We define the Lorentz hyperboloid model $\mathbb{L}^n$, the Poincaré ball model $\mathbb{B}^n$, the Klein model $\mathbb{K}^n$ and the upper half plane model $\mathbb{H}^n$ as

$$\begin{aligned}
\mathbb{L}^n &= \{x \in \mathbb{R}^{n+1}\colon x_0 > 0,\ -x_0^2 + \sum_{i=1}^{n} x_i^2 = -1\},\\
\mathbb{B}^n &= \{x \in \mathbb{R}^n\colon \|x\| < 1\},\\
\mathbb{K}^n &= \{x \in \mathbb{R}^n\colon \|x\| < 1\},\\
\mathbb{H}^n &= \{x \in \mathbb{R}^n\colon x_n > 0\}.
\end{aligned}$$

Each model is equipped with the corresponding hyperbolic distance function

$$\begin{aligned}
d_{\mathbb{L}}(x,y) &= \operatorname{arccosh}\left(x_0 y_0 - \sum_{i=1}^{n} x_i y_i\right),\\
d_{\mathbb{B}}(x,y) &= \operatorname{arccosh}\left(1 + 2\frac{\|x-y\|^2}{(1-\|x\|^2)(1-\|y\|^2)}\right),\\
d_{\mathbb{K}}(x,y) &= \operatorname{arccosh}\left(\frac{1 - x\cdot y}{\sqrt{1-\|x\|^2}\sqrt{1-\|y\|^2}}\right),\\
d_{\mathbb{H}}(x,y) &= \operatorname{arccosh}\left(1 + \frac{\|x-y\|^2}{x_n y_n}\right).
\end{aligned}$$

The hyperbolic distance functions previously defined play a crucial role in this context. They provide a way to measure distances in different hyperbolic models, ensuring consistency across representations of hyperbolic space. Also, these metrics enable efficient geometric computations, making them essential in fields such as differential geometry, geometric group theory, and neural networks.

Next, we introduce direct relations (isometries) between the models. Each of the four hyperbolic models represents the same hyperbolic space, and there exist smooth isometries between them. These isometries allow transferring geometric properties and computations from one model to another.

The hyperboloid model serves as a universal covering space, from which all other models can be derived via suitable projections. These isometries preserve hyperbolic distances and angles, ensuring that results obtained in one model remain valid in another, making them valuable tools in both theoretical and applied contexts.

Specifically we will use

- $\varphi_1\colon \mathbb{L}^n \to \mathbb{B}^n$, $\varphi_1(y) = \left(\frac{y_1}{1+y_0}, \ldots, \frac{y_n}{1+y_0}\right)$
- $\varphi_2\colon \mathbb{L}^n \to \mathbb{K}^n$, $\varphi_2(y) = \left(\frac{y_1}{y_0}, \ldots, \frac{y_n}{y_0}\right)$
- $\varphi_3\colon \mathbb{L}^n \to \mathbb{H}^n$

$$\varphi_3(y) = \left(\frac{y_1}{y_0+y_n}, \ldots, \frac{y_{n-1}}{y_0+y_n}, \frac{1}{y_0+y_n}\right).$$

Also we will use their inverses.

As an example of the geometric relevance about the isometries, in Figure 1 and Figure 2 we illustrate the isometry between $\mathbb{B}^2$ and $\mathbb{H}^2$.

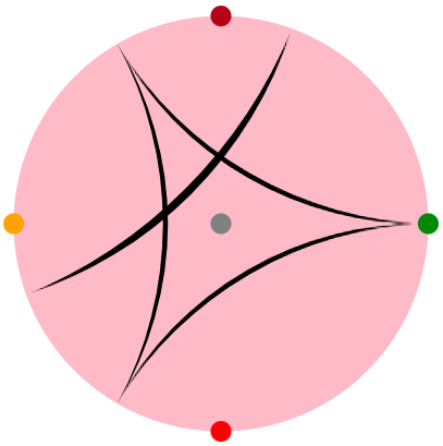

Figure 1: Hyperbolic triangle and a geodesic in the Poincaré ball model.

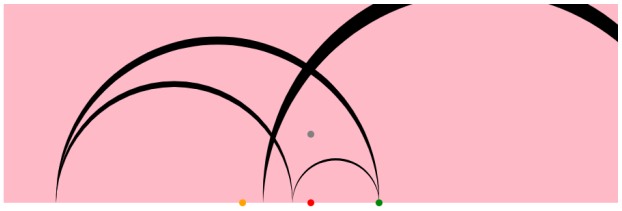

Figure 2: Hyperbolic triangle and a geodesic in the half-upper plane model.

Möbius operations. In order to state the results in $\mathcal{H}^n$, we need a well-defined way to perform addition and scalar multiplication while preserving the geometric structure of the space. Since $\mathcal{H}^n$ lacks a natural vector space structure, we rely on Möbius operations, which extend standard Euclidean operations using the exponential and logarithm maps. Specifically, addition and scalar multiplication are defined in terms of their counterparts in $\mathbb{R}^n$, adapted to the hyperbolic setting via these maps. This ensures consistency with the underlying Riemannian structure and allows for smooth transformations while maintaining key geometric properties, such as distance distortions induced by negative curvature.

In the following definitions, $\mathcal{H}$ represents a hyperbolic manifold isometric to $\mathbb{L}$.

Definition 2.3. For $a, b, x \in \mathcal{H}^n$ and $\alpha \in \mathbb{R}$, we define the Möbius addition $\oplus$ and scalar multiplication $\otimes$ by $a \oplus b = \exp_y(\log_y a + \log_y b)$ and $\alpha \otimes x = \exp_y(\alpha \log_y(x))$, where $y \in \mathcal{H}^n$ is a fixed point.

The maps $\exp_y \colon \mathcal{T}_y\mathcal{H}^n \to \mathcal{H}^n$ and $\log_y \colon \mathcal{H}^n \to \mathcal{T}_y\mathcal{H}^n$ are the exponential and logarithmic map (respectively) at y of the manifold $\mathcal{H}^n$.

Definition 2.4. A subset $X \subset \mathcal{H}^n$ is called a *cone* if it is the image of a cone in $T_y\mathcal{H}^n$ under the exponential map.

Hyperbolic Neural Networks. Hyperbolic neural networks extend the framework of Euclidean neural networks by utilizing the properties of hyperbolic geometry, which differs significantly from the flat Euclidean space typically employed in conventional neural networks. Hyperbolic spaces are particularly adept at representing hierarchical and tree-like structures, making them especially suitable for data with inherent hierarchical relationships, such as syntactic trees in natural language or social networks.

Formally we have the following

Definition 2.5. We define a Deep Neural Network as

$$\begin{aligned} f(x) &= f_1 \circ f_2 \circ \cdots \circ f_k(x) \\ f_i(x) &= \sigma_i(W_i x + b_i) \quad \text{or} \\ f_i(x) &= W_i^T \sigma_i(W_i x + b_i), \quad 1 \le i \le k \end{aligned}$$

where $W_i \in \mathbb{R}^{n\times n}$, $b_i \in \mathbb{R}^n$ and σ is the activation function.

To define a neural network in $\mathcal{H}^n$, we transfer a neural network defined on $\mathbb{R}^n$ to $\mathcal{H}^n$ using the exponential map and its inverse.

Definition 2.6. Given a function $T \colon \mathbb{R}^n \to \mathbb{R}^n$, we define the Möbius version of T by

$$\begin{aligned} T^{\otimes} \colon \mathcal{H}^n &\to \mathcal{H}^n \\ x &\mapsto \exp_y(T(\log_y(x))). \end{aligned}$$

Definition 2.7. We define a Hyperbolic Neural Network as

$$\begin{aligned} f(x) &= f_1 \circ f_2 \circ \cdots \circ f_k(x) \\ f_i(x) &= \sigma_i^{\otimes}(W_i^{\otimes} x \oplus b_i) \quad \text{or} \\ f_i(x) &= (W_i^T)^{\otimes} \sigma_i^{\otimes}(W_i^{\otimes} x \oplus b_i), \quad 1 \le i \le k. \end{aligned}$$

where $W_i \in \mathbb{R}^{n\times n}$, $b_i \in \mathcal{H}^n$ and σ is the activation function.

Sub-homogeneous functions. sub-homogeneous functions play a crucial role in understanding the behavior and stability of dynamical systems.

Definition 2.8. Let $\mathcal{H}^n$ be a hyperbolic model isometric to $\mathbb{L}^n$. Define a partial order in $\mathcal{H}^n$ as follows. For $x, x' \in \mathcal{H}^n$,

$$x \le x' \quad \iff \quad \log_y(x) \le \log_y(x'),$$

where the order in the right hand side is the partial order in $\mathcal{T}_y\mathcal{H}^n \cong \mathbb{R}^n$.

Definition 2.9. Let $X \subset \mathcal{H}^n$ be a cone. A map $f \colon X \to X$ is called sub-homogeneous if for every $x \in X$ and $\lambda \in (0,1)$ we have $f(\lambda \otimes x) \le \lambda \otimes f(x)$, whenever the order is possible.

Is easy to see that if $f \colon \mathcal{T}_y\mathcal{H}^n \to \mathcal{T}_y\mathcal{H}^n$ is sub-homogeneous then $f^{\otimes} \colon \mathcal{H}^n \to \mathcal{H}^n$ is also sub-homogeneus.

3 THEORETICAL RESULTS

In [Alvarado and Burgos, 2024] the authors proved Theorem C.2. The following result extends that Theorem and states that, despite the nonlinear geometry and the order-preserving assumption, the iterated sequence an iterated sequence $\{z_m\}$ still exhibits an effectively constant exponential rate of growth. Only the coordinate-dependent factor in the limit changes. In effect, we get a hyperbolic counterpart of the growth description, packaged into each of the main models.

Theorem 3.1. *Choose* $\exp_y \colon \mathcal{T}_y\mathcal{H}^n \to \mathcal{H}^n$. *Let* $Y = \exp_y(X)$, *where* X *is the positive cone in* $\mathbb{R}^n$. *Consider* $\varphi \colon \mathbb{L}^n \to \mathcal{H}^n$ *an isometry and let* $f_i \colon Y \to Y$ *be a sequence of order preserving and sub-homogeneous maps such that* $T_m := \log_y \circ \varphi \circ f_m \circ \varphi^{-1} \circ \exp_y$ *is a stationary sequence of maps in* X.

Let $z_m = f_1 f_2 \cdots f_m(z_0)$ for a fixed $z_0 \in Y$. Then, if $\varphi_j^{-1} \colon \mathbb{L}^n \to \mathcal{H}^n$ we have

$$\lim_{m\to\infty} \sup_{1\le i\le n} \left(\tilde{z}_m^{(j)}(i)\right)^{1/m} = e^{\lambda}$$

Here, for $j = 1, 2, 3$, $\tilde{z}_m^{(j)}(i)$ is a term depending on the isometry φ_j and the corresponding logarithmic map.

Sketch of the proof. Fix $z_0 \in \mathbb{L}^n$. We have that $T_m \colon X \to X$ is a stationary sequence by assumption. It follows directly from Definitions 2.8 and 2.9 that this sequence is also order preserving and sub-homogeneous. Choose $x_0 := \log_y(\varphi(z_0)) \in X$. Note that $z_m = \exp_{\varphi^{-1}(y)}(D_y\varphi^{-1}(x_m))$, where $x_m = T_1 \cdots T_m x_0$ and the exponential map is defined in $\mathbb{L}^n$. Computing x_m coordinate-wise we have:

- For $\varphi_1^{-1} \colon \mathbb{B}^n \to \mathbb{L}^n$,

$$\begin{aligned} x_m(i) = {} & \frac{\sqrt{2}\,\operatorname{arccosh} z_m(0)}{2\sqrt{\|z_m\|^2 - 1}} z_m(i)(1 - \|y\|^2) \\ & - \frac{2y_i \log_y(\varphi_1(z_m))(0)}{(1 - \|y\|^2)} \\ & - \frac{2y_i \sum_{j=1}^n y_j \log_y(\varphi_1(z_m))(j)}{(1 - \|y\|^2)}. \end{aligned}$$

- For $\varphi_2^{-1} \colon \mathbb{K}^n \to \mathbb{L}^n$,

$$\begin{aligned} x_m(i) = {} & \frac{\sqrt{2}\,\operatorname{arccosh} z_m(0)}{\sqrt{\|z_m\|^2 - 1}} z_m(i)(1 - \|y\|^2)^{1/2} \\ & - \frac{2y_0 y_i \log_y(\varphi_2(z_m))(0)}{(1 - \|y\|^2)} \\ & - \frac{2y_i \sum_{j=1}^n y_j \log_y(\varphi_2(z_m))(j)}{(1 - \|y\|^2)}. \end{aligned}$$

- For $\varphi_3^{-1} \colon \mathbb{H}^n \to \mathbb{L}^n$,

$$\begin{aligned} x_m(i) = {} & \frac{\sqrt{2}\,\operatorname{arccosh} z_m(0)}{\sqrt{\|z_m\|^2 - 1}} z_m(i) y_n \\ & + \frac{y_i}{y_n} \log_y(\varphi_3(z_m))(n). \end{aligned}$$

□

Example 3.2. Let $z \in \mathbb{R}^2$ be the 2D latent of a small autoencoder trained with MSE on MNIST. The decoder-encoder composition $z' = \mathrm{Enc}(\mathrm{Dec}(z))$ is well-approximated by a linear map $z' \approx Az + c$. Take entrywise absolute values and spectrally clamp to ensure A is order-preserving and $\|A\|_2 < 1$

$$A = \begin{pmatrix} 0.92 & 0.08 \\ 0.03 & 0.88 \end{pmatrix}, c = \begin{pmatrix} 0.01 \\ 0.00 \end{pmatrix}$$

$\|A\|_2 \le 0.95$.

Identify the Lorentz chart at a basepoint $y \in \mathbb{L}^2$ with $\mathbb{R}^2$ and set

$$T(v) = Av + c, \quad X = \mathbb{R}^2_{\ge 0}, \quad Y = \exp_y(X) \subset \mathbb{L}^2,$$

$$f = \exp_y \circ T \circ \log_y : Y \to Y.$$

Then T is order-preserving on the positive cone X and sub-homogeneous; hence f satisfies the assumptions of Theorem 3.1. Consider the stationary sequence $f_m \equiv f$ and iterate $z_{m+1} = f(z_m)$ with $z_0 \in Y$. For each model obtained by conjugation with an isometry $\phi_j : \mathbb{L}^2 \to \mathcal{H}^2$, write $z_m^{(j)} = \phi_j(z_m)$ and let $\tilde{z}_m^{(j)}(i)$ be the model specific coordinate factor. There exists a deterministic $\lambda \in \mathbb{R}$ such that

$$\lim_{m\to\infty} \sup_{i=1,2} \left(\tilde{z}_m^{(j)}(i)\right)^{1/m} = e^{\lambda} \quad \text{for } j \in \{1, 2, 3\},$$

with the same e^{λ} across models; only the coordinate weights $\tilde{z}_m^{(j)}(i)$ differ. In this concrete case, e^{λ} equals the spectral radius of A (which is < 1 here, so the trajectory contracts in chart).

Between different models, the weights varies, but the overall convergence behavior remains consistent. This suggests that despite differences in parameterization, the fundamental properties governing the convergence process remain intact. The stability of convergence across models highlights the underlying geometric invariance (due to the isometries), ensuring that changes in representation do not affect the asymptotic behavior of the learning process.

Furthermore, when using the model directly rather than applying an explicit isometry transformation—the computed convergence rate remains unchanged. This is a direct consequence of the isometry preserving distances and inner products, which guarantees that the optimization dynamics remain equivalent. Since isometries do not alter the structure of the underlying space, the learning process exhibits the same rate of convergence as in the original formulation. This invariance reinforces the robustness of the approach, showing that convergence properties are dictated more by the intrinsic geometry of the problem than by the specific choice of representation.

sub-homogeneity means that repeated application of certain maps shrinks variations over many iterations, ensuring the sequence cannot diverge. Consequently, in the presence of only bounded perturbations, one expects the iteration to stabilize (a.s.) at a single point.

Proposition 3.3 leads to a stable outcome even when small random variations are present at each step. Meaning that no matter how these transformations keep acting on data over many iterations, their combined effect settles down to a single point, rather than behaving chaotically.

The bounded noise in each step does not accumulate endlessly, so the system does not wander off or explode in value. This is encouraging for using these transformations in applications where data can be messy or uncertain.

The result considers a sequence of layer maps that are sub-homogeneous and 1-Lipschitz. These properties mean that applying the transformations multiple times will not blow up the values; instead, they have a form of built-in non-expansiveness. Even though, there is a sequence of real coefficients that can change from step to step but the effect of repeatedly applying these transformations still converges to a stable point.

This result is a consequence of Corollary C.3.

Proposition 3.3. *For both results we assume that Möbius operations are measure preserving.*

1. *Consider a sequence of sub-homogeneous layer maps of a HNN $f_i\colon \mathbb{L}^n \to \mathbb{L}^n$ of the form $f_i(x) = \exp_y(W_i^T\sigma_i(W_i(\log_y x)+b_i))$, where W_i are invertible matrices with $\|W_i\| \le 1$ and σ_i is a 1-Lipschitz linear activation function. Let $(\epsilon_i)_i \subset \mathbb{L}^n$ be a sequence of bounded noise, i.e. there is $C > 0$ such that $\|\log_y \epsilon_i\| \le C$ for all i. Then there exists a point (a.s.) $z \in \mathbb{L}^n$ such that*

$$\lim_{m\to\infty} \frac{1}{m} \otimes (f_1 \oplus \epsilon_1) \cdots (f_m \oplus \epsilon_m)\,(z_0) \to z.$$

2. *Consider a sequence of sub-homogeneous layer maps of a HNN $f_i : \mathbb{L}^n \to \mathbb{L}^n$ of the form $f_i(x) = \exp_y(W_i^T\sigma_i(W_i(\log_y x)+b_i))$, where $\|W_i\| \le 1$ and σ_i is 1-Lipschitz. Let η_i be a sequence of real numbers with $|\eta_i| \le 1$. Then, almost surely, there exists $z \in \mathbb{L}^n$ such that*

$$\lim_{m\to\infty} \frac{1}{m} \otimes ((\eta_1 \otimes f_1) \cdots (\eta_n \otimes f_n))\,(z_0) = z.$$

Proof. See Appendix B □

Corollary 3.4. *Proposition 3.3 remain valid for any isometric hyperbolic model.*

Proof. See Appendix B □

The previous result shows that despite variations in the learning rate, the repeated transformations retain enough stable behavior to settle on a single point. This means that the sequence will not spiral out of control or oscillate indefinitely.

Example 3.5. Work on $\mathbb{L}^2$ with basepoint y. For each layer $i \ge 1$, choose

$$W_i = \begin{pmatrix} 0.85 & 0.10 \\ 0.00 & 0.90 \end{pmatrix}, \qquad b_i = \begin{pmatrix} 0.02 \\ 0.00 \end{pmatrix},$$

and a componentwise 1-Lipschitz activation $\sigma_i(t) = \tanh(t)$ (or $\sigma_i(t) = a_i t$ with $a_i \in [0,1]$). Define the noiseless layer map in chart and lift to hyperbolic space as

$$f_i(x) = \exp_y\Big(W_i^\top\, \sigma_i\big(W_i\, \log_y x + b_i\big)\Big).$$

Add bounded hyperbolic noise ε_i and a bounded time-varying scalar schedule η_i, with

$$\|\log_y \varepsilon_i\| \le C \quad \text{for a fixed } C = 0.05, \qquad |\eta_i| \le 1.$$

Consider the perturbed scaled layers

$$x \mapsto \exp_y\Big(W_i^\top\big(\eta_i\sigma_i(W_i \log_y x + b_i)\big)\Big) \oplus \varepsilon_i.$$

- If $\sigma_i(t) = a_i t$, we have $f_i(x) \oplus \varepsilon_i$ equal to

$$\exp_y\Big(W_i^\top\sigma_i\big(W_i \log_y x + \underbrace{b_i + (W_i^\top)^{-1}\tfrac{\log_y \varepsilon_i}{a_i}}_{=\tilde{b}_i}\big)\Big),$$

  so the perturbation is absorbed into a shifted bias $\tilde{b}_i$ and the map remains in the required class.

- With $|\eta_i| \le 1$, define $\tilde{\sigma}_i = \eta_i\sigma_i$, which is again 1-Lipschitz componentwise; hence

$$\eta_i \otimes f_i(x) = \exp_y\Big(W_i^\top\tilde{\sigma}_i(W_i \log_y x + b_i)\Big)$$

  has the same form as the noiseless layer.

Let $F_m = \big((\eta_1 \otimes f_1) \oplus \varepsilon_1\big) \cdots \big((\eta_m \otimes f_m) \oplus \varepsilon_m\big)$. Under the 1-Lipschitz and bounded-noise assumptions above, Proposition 3.3 yields

$$\lim_{m\to\infty} \frac{1}{m} \otimes F_m(z_0) = z \qquad \text{a.s. for some } z \in \mathbb{L}^2,$$

independently of z_0. The same conclusion holds in any other hyperbolic model by conjugation via isometries.

We close this section with the following idea. Consider a random sequence of smooth maps T_ω on the Lorentz model, which generates a trajectory $\varphi(n,\omega)$. Assume these maps are sufficiently regular and that, on average, their cumulative composition is contracting. Then such an on-average contracting random system admits a unique moving reference point $z(\omega)$ that evolves with the randomness, and every trajectory regardless of its initialization converges to this reference point at an exponential rate. In simple terms, despite step to

step randomness, the dynamics consistently pull states toward the same random, time varying anchor, yielding a hyperbolic analogue of a unique, exponentially stable random fixed point.

For more details see Appendix A.

4 NUMERICAL ANALYSIS

In this section we develop three experiments. For further details and more experiments see Appendix E.

4.1 Representation-Invariant Stability in Hyperbolic Classification

First, we evaluate on CIFAR–100. Distances between fine labels reflect tree proximity, and coarse–level matches are semantically meaningful.

All models use a ResNet–18 backbone adjusted for 32×32 inputs (first 3×3 conv, no initial maxpool). We compare

1. a Euclidean softmax head; and
2. a hyperbolic prototype head with one prototype per class and classification by (negative) geodesic distance

(see Table E.3.4).

We instantiate the hyperbolic head in two isometric charts, the Poincaré ball $\mathbb{B}^d$ and the Lorentz hyperboloid $\mathbb{L}^d$. The embedding dimension is $d = 16$ and curvature is fixed.

To verify model invariance, we train the same architecture in Lorentz and Poincaré charts and compare the stability diagnostics below. Because the two models are isometric, the decay profiles of our diagnostics must coincide up to numerical noise, independently of the chosen chart or basepoint.

In the Poincaré ball mode, hierarchical accuracy rises (Figure 3) while all three diagnostics (window diameter, averaged-tangent norm, step distance (Figures 12, 13, 14) respectively) peak mid-training and then contract, indicating stabilization.

In the Lorentz hyperboloid model hierarchical accuracy also rises (Figure 4).

4.2 Noisy sub-homogeneous dynamics with ergodic averaging.

We first study bounded noise: $\|\varepsilon_m\| \le C$. Across seeds, $d(u_{m+1}, u_m)$ and the window diameter decay geometrically toward numerical zero, while $\|\bar{v}_m\|$ shrinks in lockstep. Crucially, the curves overlap across Lorentz, Poincaré, Klein, and upper half–plane realizations, confirming that the growth/decay profile is *representation invariant* under isometries.

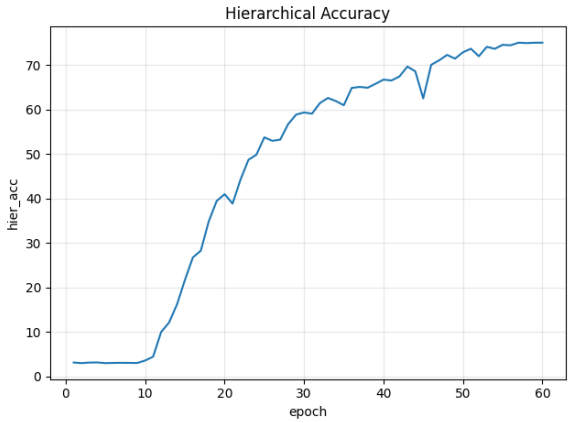

Figure 3: Hierarchical score improves to ∼75%, showing the hyperbolic head captures the label tree: coarse-level consistency strengthens as training stabilizes.

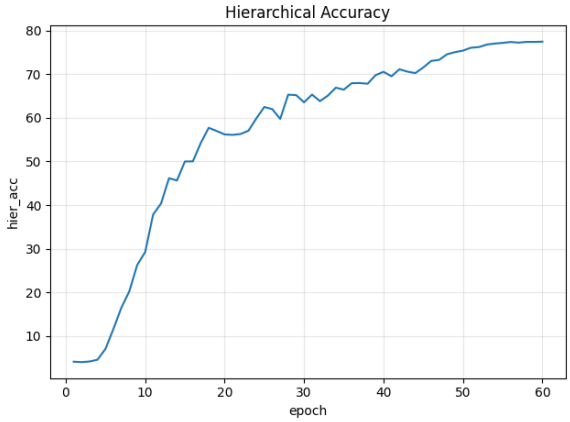

Figure 4: Hierarchical score improves to ∼80%, showing the hyperbolic head captures the label tree: coarse-level consistency strengthens as training stabilizes.

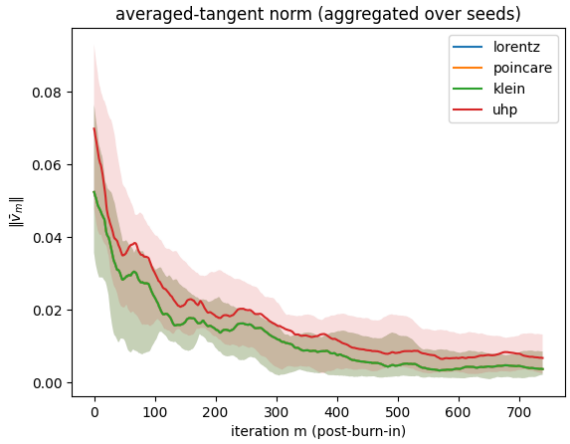

Figure 5: Curves overlap across isometric realizations, confirming representation invariance. Shaded regions show 10–90% seed bands after burn–in.

By burn-in we mean the initial portion of an iterative run that we discard from analysis because it's dominated by transient behavior (e.g., initialization effects) rather than the steady-state dynamics you care about. In our plots, we start computing metrics only after burn-in steps so convergence curves aren't skewed by those early transients.

4.3 Stabilization with time varying scalings.

Next we insert a bounded modulation η_m with $|\eta_m| \leq 1$ at each step, $x_{m+1} = \exp_y\left(\eta_m A_m \log_y(x_m) + \varepsilon_m\right)$. Median trajectories (shaded 10–90% bands over seeds) again contract, with larger average $|\eta_m|$ slowing but not preventing convergence. Final window diameters match across models to within numerical tolerance, reinforcing invariance (Figure 6).

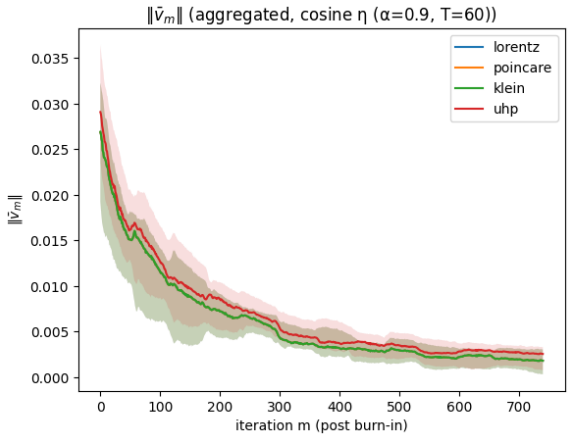

Figure 6: Trajectories coincide across Lorentz, Poincaré, Klein, and upper half–plane within numerical tolerance.

4.4 Heavy–tailed perturbations and clipping schedules.

To prove robustness beyond the bounded noise setting, we consider noisy sub-homogeneous dynamics with heavy–tailed perturbations in the tangent chart at a fixed basepoint y on $\mathbb{L}^2$ and we follow the 3 regimes: *(A) Bounded* ($\|\varepsilon_m\| \leq C$), *(B) Heavy–tailed (clipped)*: ε_m has i.i.d. t-Student components (df ν) and is clipped to $\|\varepsilon_m\| \leq C$, and *(C) Heavy–tailed (growing cap)*: the same but clipped at a slowly increasing radius $C_m = C_0(1+\log(1+m))^\beta$. Again we form the ergodic averages $\bar{v}_m$ (Figure 7).

Across all settings, the ergodic averages exhibit stable contraction, and the decay profiles coincide across hyperbolic models, empirically validating our representation–invariance theory. This matters in practice, it licenses model-agnostic diagnostics and training rules that transfer across hyperbolic models without retuning.

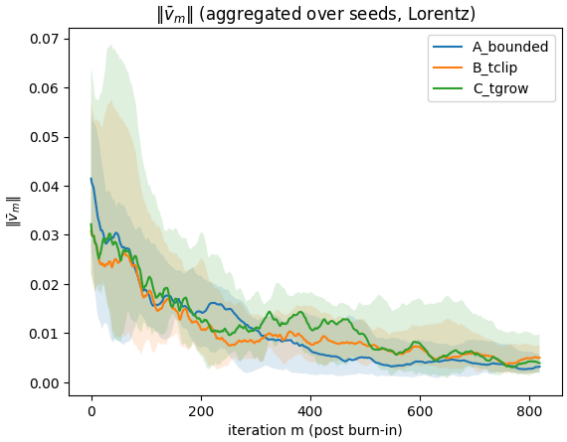

Figure 7: t-Student noise is injected in the tangent chart under (A) bounded, (B) heavy–tailed with hard clip, and (C) heavy–tailed with growing cap $C_m = C_0(1+\log(1+m))^\beta$.

4.5 Link prediction

This experiment has two complementary objectives. First, we assess whether a hyperbolic representation is advantageous on a prototypically hierarchical dataset (WordNet hypernym graph) when compared to a Euclidean baseline under matched embedding dimension. Second, we empirically verify the isometric model invariance predicted by Theorem 3.1 by evaluating the same learned hyperbolic embedding under multiple isometric realizations of $\mathbb{H}^d$, and checking that performance is unchanged up to numerical tolerance.

Let $G = (V, E)$ be the undirected graph obtained from WordNet noun synsets by connecting each synset to its hypernyms (and symmetrizing the relation), restricted to a finite induced subgraph of size $|V| = N$. We randomly split the undirected edge set E into disjoint subsets $E = E_{\text{train}} \dot{\cup} E_{\text{val}} \dot{\cup} E_{\text{test}}$, and sample an equal size set of pairs $(u, v) \notin E$ with $u \neq v$.

Now, given an embedding map $\Phi : V \to \mathcal{M}$, where $\mathcal{M}$ is either Euclidean space $\mathbb{R}^d$ or a model of hyperbolic space $\mathbb{H}^d$, we score a candidate edge (u, v) by a monotonically decreasing function of distance $s(u, v) = -d_{\mathcal{M}}\big(\Phi(u), \Phi(v)\big)^2$

In the Euclidean baseline, $d_{\mathcal{M}}$ is the standard Euclidean distance. In the hyperbolic setting, $d_{\mathcal{M}}$ is the hyperbolic distance. We train embeddings by minimizing a logistic objective that promotes larger scores for positive edges than for sampled negative pairs. We report standard link prediction metrics on E_{test} using Area Under the ROC Curve (AUC) and Average Precision (AP) (see Figures 10 and 11, respectively).

WordNet hypernym structure exhibits pronounced hi-

erarchical organization. Hyperbolic geometry supports exponential volume growth with radius, which enables low-distortion representations of tree-like/hierarchical data in low dimension. In contrast, Euclidean space requires substantially higher dimension to represent similar hierarchical branching with comparable distortion. At matched dimension d, the hyperbolic (Poincaré) embedding trained with $s(u, v)$ and with loss

$$\mathcal{L}(\Phi) = \mathbb{E}_{(u,v)\sim E_{\text{train}}}\big[\text{softplus}(-s(u, v))\big] + \mathbb{E}_{(u,v)\sim\nu}\big[\text{softplus}(s(u, v))\big],$$

where ν is a negative sampling distribution over non-edges and $\text{softplus}(t) = \log(1 + e^t)$, achieve higher AUC/AP than the Euclidean baseline (at largest dimensions), with the largest gap at small d.

In table 1, we report mean ± std AUC and AP across 5 seeds for each dimension d for both Euclidean and hyperbolic embeddings. Additionally, for each trained hyperbolic embedding we report AUC computed in the four isometric models, together with the maximum absolute pairwise discrepancy in computed distances as a numerical consistency diagnostic.

5 CONCLUSIONS AND FUTURE WORK

This paper develops a representation invariant view of stability for noisy, sub–homogeneous dynamics in HHNs and turns it into a practical certification pipeline. Theoretically, we show almost-sure convergence of ergodic averages for sequences of layer maps with bounded perturbations and bounded time varying scalings, yielding well-behaved long run dynamics. Also, we propose noise-handling regimes and empirically verify stabilization and cross-model invariance on learned contractive surrogates. Empirically, CIFAR–100 confirms the theory. Diagnostic curves coincide across charts and correlate with improved hierarchical accuracy and robust, well–behaved training. Together, these results shift the focus from model–specific coordinates to chart–agnostic behavior, offering a tool to monitor and enforce stability independent of representation. While our analysis assumes sub–homogeneity and fixed curvature.

Looking ahead, we will test sharpness beyond sub–homogeneity by probing borderline non–contractive layers and identifying minimal conditions that still guarantee convergence. Also, we will replace synthetic schedules with optimizer–induced cocycles to study stability under realistic training noise, curriculum shifts, and data–order randomness, quantifying how momentum, adaptive steps, and weight decay affect the diagnostics. And finally we will scale experiments to hierarchical image and graph benchmarks to stress test chart invariance under richer label taxonomies.

Acknowledgments

The first author acknowledge funding support from the National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

References

[Alvarado and Burgos, 2024] Alvarado, N. and Burgos, S. (2024). Convergence properties of hyperbolic neural networks on riemannian manifolds. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*.

[Alvarado and Lobel, 2024] Alvarado, N. and Lobel, H. (2024). Geometric deep learning with quasiconformal neural networks: An introduction. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*.

[Alvarado et al., 2023] Alvarado, N., Lobel, H., and Petrache, M. (2023). Curvature-dimension tradeoff for generalization in hyperbolic space. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*.

[Avelin and Karlsson, 2022] Avelin and Karlsson (2022). Deep limits and a cut-off phenomenon for neural networks. *Journal of Machine Learning Research*.

[Billera et al., 2001] Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767.

[Chami et al., 2020] Chami, I., Gu, A., Chatziafratis, V., and Ré, C. (2020). From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. *Advances in Neural Information Processing Systems*, 33:15065–15076.

[Chami et al., 2019] Chami, I., Ying, Z., Ré, C., and Leskovec, J. (2019). Hyperbolic Graph Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[Ganea et al., 2018a] Ganea, O., Bécigneul, G., and Hofmann, T. (2018a). Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655. PMLR.

[Ganea et al., 2018b] Ganea, O., Bécigneul, G., and Hofmann, T. (2018b). Hyperbolic neural networks. *Advances in neural information processing systems*, 31.

[Gromov, 1987] Gromov, M. (1987). *Hyperbolic groups*. Springer.

[Gu et al., 2019] Gu, A., Sala, F., Gunel, B., and Ré, C. (2019). Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*.

[Hamann, 2018] Hamann, M. (2018). On the tree-likeness of hyperbolic spaces. In *Mathematical proceedings of the cambridge philosophical society*, volume 164, pages 345–361. Cambridge University Press.

[Khrulkov et al., 2020] Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I., and Lempitsky, V. (2020). Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6418–6428.

[Liu et al., 2019] Liu, Q., Nickel, M., and Kiela, D. (2019). Hyperbolic graph neural networks. *Advances in neural information processing systems*, 32.

[Nickel and Kiela, 2017] Nickel, M. and Kiela, D. (2017). Poincaré Embeddings for Learning Hierarchical Representations. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Nickel and Kiela, 2018] Nickel, M. and Kiela, D. (2018). Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pages 3779–3788. PMLR.

[Peng et al., 2021] Peng, W., Varanka, T., Mostafa, A., Shi, H., and Zhao, G. (2021). Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10023–10044.

[Rodríguez-Flores and Papadopoulos, 2020] Rodríguez-Flores, M. A. and Papadopoulos, F. (2020). Hyperbolic mapping of human proximity networks. *Scientific Reports*, 10(1):20244.

[Sala et al., 2018] Sala, F., De Sa, C., Gu, A., and Ré, C. (2018). Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pages 4460–4469. PMLR.

[Shavitt and Tankel, 2008] Shavitt, Y. and Tankel, T. (2008). Hyperbolic embedding of internet graph for distance estimation and overlay construction. *IEEE/ACM Transactions on Networking*, 16(1):25–36.

[Sittoni and Tudisco, 2024] Sittoni, P. and Tudisco, F. (2024). Subhomogeneous deep equilibrium models. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45794–45812. PMLR.

[Ungar, 2008] Ungar, A. A. (2008). *Analytic hyperbolic geometry and Albert Einstein's special theory of relativity*. World Scientific.

[Yang et al., 2022] Yang, M., Zhou, M., Li, Z., Liu, J., Pan, L., Xiong, H., and King, I. (2022). Hyperbolic graph neural networks: A review of methods and applications. *arXiv preprint arXiv:2202.13852*.

[Zhang, 2023] Zhang, F. (2023). Deep neural networks from the perspective of ergodic theory. *arXiv preprint arXiv:2308.03888*.

[Zhou et al., 2019] Zhou, M., Liu, T., Li, Y., Lin, D., Zhou, E., and Zhao, T. (2019). Toward understanding the importance of noise in training neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7594–7602. PMLR.

[Zhou et al., 2017] Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S., and Glynn, P. W. (2017). Stochastic mirror descent in variationally coherent optimization problems. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A An Important Conjecture

Conjecture A.1. Let $(\Omega, \mathcal{F}, \mathbb{P}, \sigma)$ be an ergodic dynamical system. Let $\{T_\omega\}_{\omega\in\Omega}$ be a family of C^1-diffeomorphisms on the Lorentz model $\mathbb{L}^n$, inducing the random cocycle

$$\varphi(n,\omega) = T_{\theta^{n-1}\omega} \circ \cdots \circ T_\omega, \quad \varphi(0,\omega) = \omega.$$

Assume

1. (H1) $\int_\Omega \log\|DT_\omega\|\, d\mathbb{P}(\omega) < \infty$; and
2. (H2) $\lambda := \lim_{n\to\infty} \frac{1}{n} \int_\Omega \log\|D\varphi(n,\omega)\|\, d\mathbb{P}(\omega) < 0.$

Then, there exists a unique measurable random point $z\colon \Omega \to \mathbb{L}^n$ with the invariance property

$$T_\omega\big(z(\omega)\big) = z\big(\theta\omega\big) \quad \text{for a.e. } \omega,$$

and constants $C(\omega) > 0$, $\gamma < 0$, such that for all initial $z_0 \in \mathbb{L}^n$,

$$d_{\mathbb{L}}\big(\varphi(n,\omega)(z_0), z(\theta^n\omega)\big) \le C(\omega)e^{\gamma n} \quad \text{for all } n \ge 0,$$

where $d_{\mathbb{L}}$ is the Lorentz metric.

Throughout, assume Conjecture A.1 to be true, then the following holds.

1. If the base noise is i.i.d., the induced Markov kernel $P(x, A) = \mathbb{P}\{T_\omega(x) \in A\}$ on $(\mathbb{L}^n, d_{\mathbb{L}})$ admits a unique invariant law π, and there exist $a \in (0,1)$, $b < \infty$ such that

$$W_1\left(P^k(x,\cdot), \pi\right) \;\le\; ba^k \quad \text{for all } x \in \mathbb{L}^n, k \ge 0,$$

   where W_1 is the 1–Wasserstein distance for the ground metric $d_{\mathbb{L}}$.

2. For any Lipschitz observable $f : \mathbb{L}^n \to \mathbb{R}$ with $\mathbb{E}|f(z(\omega))| < \infty$, the time averages along any trajectory satisfy a law of large numbers

$$\frac{1}{N}\sum_{k=0}^{N-1} f\left(\varphi(k,\omega)x\right) \xrightarrow[N\to\infty]{a.s.} \mathbb{E}f\left(z(\omega)\right),$$

   and a central limit theorem holds under standard mixing and moment conditions.

3. If $\{T^\varepsilon_\omega\}$ is a small $\mathcal{C}^1$ perturbation preserving (H1)–(H2) with negative top exponent, then the random equilibrium $z^\varepsilon(\cdot)$ and the rate γ^ε vary continuously with ε.

4. Let $\theta_{k+1} = \exp_{\theta_k}\big(-\eta\,\xi(\theta_k, \omega_k)\big)$ be a Riemannian SGD step on $\mathbb{L}^n$ with i.i.d. ω_k, where the one-step map coincides with T_{ω_k} up to $o(\eta)$ and moments satisfy (H1). If the cocycle of the exact maps satisfies (H2) with $\lambda < 0$, then there exist random $C(\omega) > 0$ and $\gamma < 0$ such that, for any initialization, $d_{\mathbb{L}}\big(\theta_k, z(\theta^k\omega)\big) \le C(\omega)e^{\gamma k} + o(1)$, and in particular $\theta_k \to z(\theta^k\omega)$ almost surely.

5. The above statements transfer to any isometric model of $\mathcal{H}^n$ via the canonical isometries.

B Missing Proofs

Proof of Theorem 3.1. Fix $z_0 \in \mathbb{L}^n$. We have that $T_m \colon X \to X$ is a stationary sequence by assumption. It follows directly from Definitions 2.8 and 2.9 that this sequence is also order preserving and sub-homogeneous. Choose $x_0 := \log_y(\varphi(z_0)) \in X$. Note the following

$$\begin{aligned} z_m = f_1 \cdots f_m(z_0) &= f_1 \cdots f_m(\varphi^{-1}(\exp_y x_0)) \\ &= \varphi^{-1}(\exp_y(T_1 \cdots T_m x_0)) \\ &= \varphi^{-1}(\exp_y x_m) \\ &= \exp_{\varphi^{-1}(y)}(D_y\varphi^{-1}(x_m)), \end{aligned}$$

where $x_m = T_1 \cdots T_m x_0$ and the last exponential map is the associated to the hyperboloid model $\mathbb{L}^n$.

Now,

$$\begin{aligned} z_m = (z_m(0), \dots, z_m(n))) &= \cosh\|D_y\varphi^{-1}(x_m)\|\varphi^{-1}(y) + \frac{\sinh\|D_y\varphi^{-1}(x_m)\|}{\|D_y\varphi_y^{-1}(x_m)\|} D_y\varphi^{-1}(x_m) \\ &= \Big(\cosh\|D_y\varphi^{-1}(x_m)\|, \frac{\sinh\|D_y\varphi^{-1}(x_m)\|}{\|D_y\varphi^{-1}(x_m)\|}[D_y\varphi^{-1}(x_m)](1), \dots, \frac{\sinh\|d\varphi_y^{-1}(x_m)\|}{\|d\varphi_y^{-1}(x_m)\|}[D_y\varphi^{-1}(x_m)](n)\Big). \end{aligned}$$

For $1 \le i \le n$ we have

$$[D_y\varphi^{-1}(x_m)](i) = \frac{\|D_y\varphi^{-1}(x_m)\|}{\sinh\|D_y\varphi^{-1}(x_m)\|} z_m(i). \tag{1}$$

Let's examine the first case. Choose

$$\begin{aligned} \varphi_1^{-1} \colon \mathbb{B}^n &\to \mathbb{L}^n \\ y &\mapsto \left(\frac{1+\|y\|^2}{1-\|y\|^2}, \frac{2y_1}{1-\|y\|^2}, \dots, \frac{2y_n}{1-\|y\|^2}\right). \end{aligned}$$

Is easy to see that

$$D_y\varphi_1^{-1} = \frac{1}{(1-\|y\|^2)^2} \begin{pmatrix} 4y_1 & 4y_2 & \cdots & 4y_n \\ 2(1-\|y\|^2)+4y_1^2 & 4y_1y_2 & \cdots & 4y_1y_n \\ 4y_2y_1 & 2(1-\|y\|^2)+4y_2^2 & \cdots & 4y_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ 4y_ny_1 & 4y_ny_2 & \cdots & 2(1-\|y\|^2)+4y_n^2 \end{pmatrix}$$

$$D_y\varphi_1^{-1}(x_m) = \frac{1}{(1-\|y\|^2)^2} \begin{pmatrix} 4y_1x_m(0) + 2(1-\|y\|^2)x_m(1) + 4y_1\sum_{j=1}^n y_jx_m(j) \\ 4y_2x_m(0) + 2(1-\|y\|^2)x_m(2) + 4y_2\sum_{j=1}^n y_jx_m(j) \\ \vdots \\ 4y_nx_m(0) + 2(1-\|y\|^2)x_m(n) + 4y_n\sum_{j=1}^n y_jx_m(j) \end{pmatrix}.$$

Then, replacing in Equation 1 we obtain

$$
\begin{aligned}
[D_y\varphi_1^{-1}(x_m)](i) &= \frac{\|D_y\varphi_1^{-1}(x_m)\|}{\sinh\|D_y\varphi_1^{-1}(x_m)\|}z_m(i)\\
\frac{4y_ix_m(0)+2(1-\|y\|^2)x_m(i)+4y_i\sum_{j=1}^n y_jx_m(j)}{(1-\|y\|^2)^2} &= \frac{\sqrt{2}\operatorname{arccosh}z_m(0)}{\sqrt{\|z_m\|^2-1}}z_m(i)\\
4y_i\log_y(\varphi_1(z_m))(0)+2(1-\|y\|^2)x_m(i)+4y_i\sum_{j=1}^n y_j\log_y(\varphi(z_m))(j) &= \frac{\sqrt{2}\operatorname{arccosh}z_m(0)}{\sqrt{\|z_m\|^2-1}}z_m(i)(1-\|y\|^2)^2\\
2(1-\|y\|^2)x_m(i) &= \frac{\sqrt{2}\operatorname{arccosh}z_m(0)}{\sqrt{\|z_m\|^2-1}}z_m(i)(1-\|y\|^2)^2\\
&\quad -4y_i\log_y(\varphi_1(z_m))(0)\\
&\quad -4y_i\sum_{j=1}^n y_j\log_y(\varphi_1(z_m))(j)\\
x_m(i) &= \frac{\sqrt{2}\operatorname{arccosh}z_m(0)}{2\sqrt{\|z_m\|^2-1}}z_m(i)(1-\|y\|^2)\\
&\quad -\frac{2y_i\log_y(\varphi_1(z_m))(0)}{(1-\|y\|^2)}\\
&\quad -\frac{2y_i\sum_{j=1}^n y_j\log_y(\varphi_1(z_m))(j)}{(1-\|y\|^2)}
\end{aligned}
$$

since $\cosh\|D_y\varphi_1^{-1}(x_m)\| = z_m(0)$ and $x_m = T_1\cdots T_m(x_0) = \log_y(\varphi_1(z_m))$. Here $\log_y$ is the logarithmic map of the Poincaré ball model at y.

For the second case let

$$
\begin{aligned}
\varphi^{-1}\colon \mathbb{K}^n &\to \mathbb{L}^n\\
y &\mapsto \left(\frac{1}{\sqrt{1-\|y\|^2}}, \frac{y_1}{\sqrt{1-\|y\|^2}}, \ldots, \frac{y_n}{\sqrt{1-\|y\|^2}}\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
D_y\varphi^{-1} &= \frac{1}{(1-\|y\|^2)^{3/2}}\begin{pmatrix} 2y_0 & 2y_1 & \cdots & 2y_n\\ 2y_0y_1 & (1-\|y\|^2)+2y_1^2 & \cdots & 2y_ny_1\\ \vdots & \vdots & \ddots & \vdots\\ 2y_0y_n & 2y_1y_n & \cdots & (1-\|y\|^2)+2y_n^2\end{pmatrix}\\
D_y\varphi^{-1}(x_m) &= \frac{1}{(1-\|y\|^2)^{3/2}}\begin{pmatrix} 2y_0x_m(0)+2y_1x_m(1)+2y_2x_m(2)+\cdots+2y_nx_m(n)\\ 2y_0y_1x_m(0)+(1-\|y\|^2+2y_1^2)x_m(1)++\cdots+2y_ny_1x_m(n)\\ \vdots\\ 2y_0y_nx_m(0)+2y_1y_nx_m(1)+\cdots+(1-\|y\|^2+2y_n^2)x_m(n)\end{pmatrix}
\end{aligned}
$$

Then, replacing in Equation 1 we obtain

$$\begin{aligned}
[D_y\varphi^{-1}(x_m)](i) &= \frac{\|D_y\varphi^{-1}(x_m)\|}{\sinh\|D_y\varphi^{-1}(x_m)\|}z_m(i)\\
\frac{2y_0y_ix_m(0)+(1-\|y\|^2)x_m(i)+2y_i\sum_{j=1}^n y_jx_m(j)}{(1-\|y\|^2)^{3/2}} &= \frac{\sqrt{2}\operatorname{arccosh}z_m(0)}{\sqrt{\|z_m\|^2-1}}z_m(i)\\
2y_0y_i\log_y(\varphi(z_m))(0)+(1-\|y\|^2)x_m(i)+2y_i\sum_{j=1}^n y_j\log_y(\varphi(z_m))(j) &= \frac{\sqrt{2}\operatorname{arccosh}z_m(0)}{\sqrt{\|z_m\|^2-1}}z_m(i)(1-\|y\|^2)^{3/2}\\
x_m(i) &= \frac{\sqrt{2}\operatorname{arccosh}z_m(0)}{\sqrt{\|z_m\|^2-1}}z_m(i)(1-\|y\|^2)^{1/2}\\
&\quad-\frac{2y_0y_i\log_y(\varphi(z_m))(0)}{(1-\|y\|^2)}\\
&\quad-\frac{2y_i\sum_{j=1}^n y_j\log_y(\varphi(z_m))(j)}{(1-\|y\|^2)}.
\end{aligned}$$

Finally, consider

$$\begin{aligned}
\varphi^{-1}\colon \mathbb{H}^n &\to \mathbb{L}^n\\
y &\mapsto \left(\frac{1+\|y\|^2}{2y_n},\frac{y_1}{y_n},\dots,\frac{y_{n-1}}{y_n},\frac{1-\|y\|^2}{2y_n}\right).
\end{aligned}$$

Hence

$$D_y\varphi^{-1} = \begin{pmatrix} y_0/y_n & y_1/y_n & \cdots & y_{n-1}/y_n & 1-\frac{1+\|y\|^2}{2y_n^2}\\ 0 & 1/y_n & \cdots & 0 & -y_1/y_n^2\\ 0 & 0 & \cdots & 0 & -y_2/y_n^2\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & 1/y_n & -y_{n-1}/y_n^2\\ -y_0/y_n & -y_1/y_n & \cdots & -y_{n-1}/y_n & -1-\frac{1-\|y\|^2}{2y_n^2}\end{pmatrix}$$

$$D_y\varphi^{-1}(x_m) = \begin{pmatrix} \sum_{j=0}^{n-1}\frac{y_i}{y_n}x_m(j)+x_m(n)-\frac{1+\|y\|^2}{2y_n}x_m(n)\\ \frac{x_m(1)}{y_n}-\frac{y_1}{y_n^2}x_m(n)\\ \vdots\\ \frac{x_m(n-1)}{y_n}-\frac{y_{n-1}}{y_n^2}x_m(n)\\ -\left(\sum_{j=0}^{n-1}\frac{y_i}{y_n}x_m(j)\right)-x_m(n)-\frac{1-\|y\|^2}{2y_n}x_m(n)\end{pmatrix}$$

Thus, replacing in Equation 1 for $1\le i\le n-1$ we obtain

$$\begin{aligned}
[D_y\varphi^{-1}(x_m)](i) &= \frac{\|D_y\varphi^{-1}(x_m)\|}{\sinh\|D_y\varphi^{-1}(x_m)\|}z_m(i)\\
\frac{1}{y_n}x_m(i)-\frac{y_i}{y_n^2}x_m(n) &= \frac{\sqrt{2}\operatorname{arccosh}z_m(0)}{\sqrt{\|z_m\|^2-1}}z_m(i)\\
x_m(i) &= \frac{\sqrt{2}\operatorname{arccosh}z_m(0)}{\sqrt{\|z_m\|^2-1}}z_m(i)y_n+\frac{y_i}{y_n}\log_y(\varphi(z_m))(n).
\end{aligned}$$

□

Proof of Proposition 3.3. 1. Let $\sigma(t)$ be a linear activation. Specifically $\sigma(t)=a_it$. Observe that for each i,

$$\begin{aligned}
f_i(x)\oplus\epsilon_i &= \exp_y(W_i^Ta_iW_i(\log_y x)+a_iW_i^Tb_i+\log_y\epsilon_i)\\
&= \exp_y\left(W_i^T\sigma_i\left(W_i(\log_y x)+b_i+(W^T)^{-1}\frac{\log_y\epsilon_i}{a_i}\right)\right)
\end{aligned}$$

Define $g_i(x) := \exp_y(W_i^T \sigma_i(W_i(\log_y x) + \tilde{b_i}))$, where $\tilde{b_i} = b_i + (W^T)^{-1}\dfrac{\log_y \epsilon_i}{a_i}$. Observe that $g_i(x) = f_i(x) \oplus \epsilon_i$, and has the form required to apply Corollary C.3, giving the desired result.

2. Observe that for each i, we have $\eta_i \otimes f_i(x) = \exp_y(W_i^T \eta_i \sigma(W_i(\log_y x) + b_i))$. Now, define $g_i(x) := \exp_y(W_i^T \tilde{\sigma_i}(W_i(\log_y x) + b_i))$, where $\tilde{\sigma_i}(t) = \eta_i \sigma_i(t)$ is also 1-Lipschitz as $|\eta_i| \leq 1$. Observe that $g_i(x) = \eta_i \otimes f_i(x)$, and has the form required to apply Corollary C.3.

□

Proof of Corollary 3.4. For the first part, let $\mathcal{H}^n$ be any n-dimensional hyperbolic (isometric to $\mathbb{L}^n$) model. Suppose $\{g_i\}_{i\geq 1}$ is a sequence of maps on $\mathcal{H}^n$ that, under a fixed isometry $\varphi\colon \mathbb{L}^n \to \mathcal{H}^n$, correspond exactly to the maps $\{f_i\}_{i\geq 1}$ from Proposition 3.3 via

$$g_i = \varphi \circ f_i \circ \varphi^{-1}.$$

If $G_m = g_1 \cdot g_2 \cdot \ldots \cdot g_m$, then $\dfrac{1}{m} \otimes G_m(x)$ on $\mathcal{H}^n$ also converges almost surely to a point in $\mathcal{H}^n$, under the same sub-homogeneity and bounded-noise assumptions.

The second part of the result follows under the same argument. □

C Auxiliary Results

The following result is due to [Avelin and Karlsson, 2022], but we write our proof.

Theorem C.1. *Let $X = \{x \in \mathbb{R}^N\colon x_i \geq 0 \quad \forall i\}$ be the standard positive cone, equipped with the Thompson metric*

$$d(x,y) = \log\Big(\max\{\max_i \tfrac{x_i}{y_i}, \max_i \tfrac{y_i}{x_i}\}\Big).$$

Let $\{T_i\}_{i=1}^{\infty}$ be a stationary sequence of maps $T_i\colon X \to X$ which are order-preserving and sub-homogeneous. Assume the integrability condition

$$\int d\big(T_i(x_0), x_0\big) d\mathbb{P} < \infty$$

for some $x_0 \in X$. Define $x_n = T_1 T_2 \cdots T_n(x_0)$. Then there exists a deterministic $\lambda \in \mathbb{R}$ such that almost surely

$$\lim_{n\to\infty} \sup_{1\leq i\leq N} \big(x_n(i)\big)^{1/n} = e^{\lambda},$$

and moreover there is a (random) coordinate $i_0 \in \{1, \ldots, N\}$ for which

$$\lim_{n\to\infty} \big(x_n(i_0)\big)^{1/n} = e^{\lambda}.$$

Proof. By order-preserving and sub-homogeneity for all $x, y \in X$ we have $x \leq \lambda y$, Then $T_i(x) \leq T_i(\lambda y) \leq \lambda T_i(y)$, and symmetrically $y \leq \lambda x$ implies $T_i(y) \leq \lambda T_i(x)$. Hence

$$d\big(T_i(x), T_i(y)\big) \leq d(x,y),$$

so each T_i is non-expansive in (X, d).

Define the subadditive cocycle $a(n) = d\big(x_0, x_n\big) = d\big(x_0, T_1 \cdots T_n(x_0)\big)$. To prove that $a(n)$ is a subbadditive cocycle we fix $m, n \geq 1$ and ω. Define $S_k(\omega) := T_1(\omega) T_2(\omega) \cdots T_k(\omega)$,

so that

$$x_k = S_k(\omega)(x_0), \quad x_{m+n} = S_{m+n}(\omega)(x_0).$$

By the triangle inequality for the Thompson metric,

$$d\big(x_0, x_{m+n}\big) = d\big(x_0, S_{m+n}(x_0)\big) \leq d\big(x_0, S_m(x_0)\big) + d\big(S_m(x_0), S_{m+n}(x_0)\big).$$

Since each T_i is order-preserving and sub-homogeneous, it is 1-Lipschitz in d, so

$$d\big(S_m(x_0), S_{m+n}(x_0)\big) = d\big(S_m(x_0), S_m\big(T_{m+1} \cdots T_{m+n}(x_0)\big)\big) \leq d\big(x_0, T_{m+1}(\omega) \cdots T_{m+n}(\omega)(x_0)\big).$$

Combining these inequalities we obtain

$$d\big(x_0, x_{m+n}\big) \leq d\big(x_0, x_m\big) + d\big(x_0, T_{m+1} \cdots T_{m+n}(x_0)\big).$$

By stationarity we have $d\big(x_0, T_{m+1} \cdots T_{m+n}(x_0)\big) = a\big(n, \theta^m \omega\big)$, therefore

$$a(m+n, \omega) = d\big(x_0, x_{m+n}\big) \leq a(m, \omega) + a\big(n, \theta^m \omega\big),$$

proving the subadditive cocycle property.

Now, by the subadditive ergodic theorem there exists a.s. a limit

$$\lim_{n \to \infty} \frac{a(n)}{n} = \lambda.$$

Moreover by Theorem 1 ([Avelin and Karlsson, 2022]) we have

$$\lim_{n \to \infty} -\frac{1}{n} h(x_n) = \lambda$$

for some metric functional h.

By Proposition 9 ([Avelin and Karlsson, 2022]) we have

$$h(x) = \max\Big\{\sup_i (\log x_i + \log u_i), \sup_i (-\log x_i + \log v_i)\Big\},$$

where $u, v \geq 0$ and $\max_i \max\{u_i, v_i\} = 1$. From

$$\lim_{n \to \infty} -\frac{1}{n} h(x_n) = \lambda$$

we have

$$\lim_{n \to \infty} \frac{1}{n} \sup_i \log x_n(i) = \lambda,$$

since the negative term cannot dominate in the limit. Now,

$$\lim_{n \to \infty} \sup_i \big(x_n(i)\big)^{1/n} = e^{\lambda}.$$

Finally, by choosing a coordinate i_0 where the supremum is (asymptotically) attained we have the desired result. □

Theorem C.2 ([Alvarado and Burgos, 2024])**.** *Let $Y = \exp_y(X)$, where X is the positive cone in $\mathbb{R}^n$. Let $f_i \colon Y \to Y$ be a sequence of order preserving and sub-homogeneous maps such that $T_m := \log_y \circ f_m \circ \exp_y$ is a stationary sequence of maps in X. Let $z_m = f_1 f_2 \cdots f_m(z_0)$ for a fixed $z_0 \in Y$. Then, we have*

$$\lim_{m \to \infty} \sup_{1 \leq i \leq n} \left(\frac{\sqrt{2} \operatorname{arccosh}(z_m(0))}{\sqrt{\|z_m\|^2 - 1}} z_m(i) \right)^{1/m} = e^{\lambda}.$$

Corollary C.3 ([Alvarado and Burgos, 2024])**.** *Let $f_m \colon \mathbb{L}^n \to \mathbb{L}^n$ be a sequence of maps of the form $f_m(x) = T_m^{\otimes}(x)$, where $T_m \colon \mathbb{R}^n \to \mathbb{R}^n$ defined by $T_m(v) = W^{\top} \sigma(Wv + b)$, is a stationary sequence of layer maps, $\|W\| \leq 1$, $b \in \mathbb{R}^n$ and σ is $1-$Lipschitz componentwise. Then, as $m \to \infty$, almost surely there exist $z \in \mathbb{L}^n$ such that*

$$\frac{1}{m} \otimes f_1 f_2 \ldots f_m(z_0) \to z.$$

The point $z \in \mathbb{L}^n$ is independent of the initial data $z_0 \in \mathbb{L}^n$.

D Examples

Example D.1. Let $G = (V, E)$ be a directed acyclic taxonomy (e.g. WordNet noun hypernym graph) and let $\psi : V \to \mathcal{H}^2$ be a 2D hyperbolic embedding trained to preserve hierarchy. Fix a basepoint $y \in \mathbb{L}^2$ and identify $T_y\mathbb{L}^2 \simeq \mathbb{R}^2$. Suppose that $\mathrm{Upd}(x) := \exp_y\Big(\alpha \log_y(x) + (1-\alpha)\dfrac{1}{\deg(u)} \displaystyle\sum_{v \in \mathcal{N}(u)} \log_y(\psi(v))\Big)$, $\alpha \in (0,1)$, is applied nodewise ($\mathcal{N}(u)$ denotes the set of neighbors of u in G and $x = \psi(u)$). The induced tangent update $z' = \log_y(\mathrm{Upd}(\exp_y(z)))$ is well-approximated by an affine map $z' \approx Az + c$ on a positive cone X.

$$A = \begin{pmatrix} 0.90 & 0.07 \\ 0.02 & 0.86 \end{pmatrix}, \quad c = \begin{pmatrix} 0.015 \\ 0.005 \end{pmatrix}, \quad \|A\|_2 \le 0.94.$$

Define $T(z) = Az + c$, $X = \mathbb{R}^2_{\ge 0}$, $Y = \exp_y(X) \subset \mathbb{L}^2$, $f = \exp_y \circ T \circ \log_y : Y \to Y$.

Then T is order-preserving on X and sub-homogeneous; hence f satisfies the assumptions of Theorem 3.1 and there exists a deterministic $\lambda \in \mathbb{R}$ given by Theorem 3.1. In this concrete instance, e^λ coincides with the spectral radius $\rho(A)$, and since $\rho(A) < 1$, the induced orbit contracts in the tangent chart while remaining representation-invariant across isometric realizations of $\mathcal{H}^2$.

Example D.2. Choose $y \in \mathbb{L}^n$ a fixed basepoint y. Let $p > 1$ and define power $v^{\odot p} := (v_1^p, \dots, v_n^p)$. Fix $\alpha \in (0,1)$ and consider a stationary sequence of nonnegative matrices $\{A_m\}_{m \ge 1}$ such that $\|A_m\|_2 \le L$ for some deterministic $L < \infty$. Define the nonlinear residual type update on the cone $X = \mathbb{R}^n_{\ge 0}$ by

$$T_m(v) := \alpha v + (1-\alpha) A_m\big(v^{\odot p}\big), \qquad v \in X.$$

Now, if $v \le w$ then $v^{\odot p} \le w^{\odot p}$ and $A_m \ge 0$ gives $A_m(v^{\odot p}) \le A_m(w^{\odot p})$, hence T_m is order preserving. Finally, for $\lambda \in (0,1)$ we have,

$$T_m(\lambda v) = \alpha\lambda v + (1-\alpha)\lambda^p A_m(v^{\odot p}) \le \lambda\Big(\alpha v + (1-\alpha) A_m(v^{\odot p})\Big) = \lambda\, T_m(v),$$

since $\lambda^p \le \lambda$ for $p > 1$. Now lift to hyperbolic space on the cone image $Y = \exp_y(X) \subset \mathbb{L}^n$ by

$$f_m := \exp_y \circ T_m \circ \log_y : Y \to Y.$$

Because $\{T_m\}$ is stationary and each T_m is order-preserving and sub-homogeneous on X, the sequence $\{f_m\}$ satisfies the assumptions of Theorem 3.1, hence the (model-invariant) asymptotic growth descriptor e^λ exists a.s. for the iterates $z_m = f_1 \cdots f_m(z_0)$.

Example D.3. Let $X \subset \mathbb{R}^n_{\ge 0}$ be the positive cone with the coordinatewise order. If $T : X \to X$ is order-preserving and positively q-homogeneous for some $q \ge 1$, i.e.

$$T(\alpha v) = \alpha^q T(v) \qquad \forall \alpha > 0, v \in X,$$

then T is sub-homogeneous. Let $p \ge 1$ and let $\{A_m\}_{m \ge 1}$ be a stationary sequence of nonnegative matrices in $\mathbb{R}^{n \times n}$. Define $T_m : X \to X$ on $X = \mathbb{R}^n_{\ge 0}$ by the coordinatewise p-pooling rule

$$T_m(v) := \Big(A_m\big(v^{\odot p}\big)\Big)^{\odot(1/p)}, \qquad \text{where } u^{\odot(1/p)} = (u_1^{1/p}, \dots, u_n^{1/p}).$$

This map is *nonlinear* for $p \ne 1$. It is order-preserving because each operation is monotone on $\mathbb{R}^n_{\ge 0}$. Moreover, it is positively 1-homogeneous,

$$T_m(\lambda v) = \Big(A_m((\lambda v)^{\odot p})\Big)^{\odot(1/p)} = \Big(\lambda^p A_m(v^{\odot p})\Big)^{\odot(1/p)} = \lambda\, T_m(v),$$

so it is sub-homogeneous by the sufficient condition above. Lifting to $Y = \exp_y(X) \subset \mathbb{L}^n$ via $f_m = \exp_y \circ T_m \circ \log_y$, Theorem 3.1 applies to the iterates $z_m = f_1 \cdots f_m(z_0)$, yielding a deterministic λ and representation-invariant growth descriptor e^λ.

E Full Numerical Experiments

E.1 DeepWalk

Let $G = (V, E)$ be a graph with $|V| = N$. DeepWalk constructs node embeddings by sampling short random walks on G and training a Skip-gram model on the resulting sequences. Concretely, for each node $v \in V$ we sample R random walks $\mathcal{W}_v = (w_0 = v, w_1, \ldots, w_{L-1})$ of length L using a simple random walk transition rule. Given a window size ω, each walk yields training pairs (w_i, w_j) for all $0 \le i < L$ and $\max(0, i-\omega) \le j \le \min(L-1, i+\omega), j \ne i$.

Then, DeepWalk learns two embedding tables $\mathbf{z}_v \in \mathbb{R}^d$ and $\mathbf{c}_v \in \mathbb{R}^d$ by maximizing the Skip-gram log-likelihood with negative sampling

$$\max_{\{\mathbf{z}_v, \mathbf{c}_v\}} \sum_{(u,v)\in\mathcal{P}} \log \sigma(\langle \mathbf{z}_u, \mathbf{c}_v \rangle) + \sum_{k=1}^{K} \mathbb{E}_{v_k \sim \nu} \big[\log \sigma(-\langle \mathbf{z}_u, \mathbf{c}_{v_k} \rangle)\big],$$

where $\mathcal{P}$ is the multi set of observed center-context pairs from the random walks, $\sigma(t) = (1 + e^{-t})^{-1}$, ν is a noise distribution over nodes, and K is the number of negative samples.

Given DeepWalk embeddings $\{\mathbf{z}_v\}_{v\in V} \subset \mathbb{R}^d$, we construct a data-driven linear map $A \in \mathbb{R}^{d\times d}$ as follows.

First, we enforce a positive-cone representation by setting $\mathbf{z}_v \leftarrow |\mathbf{z}_v|$ entrywise. Next we define a graph smoothing target

$$\mathbf{z}'_v := \frac{1}{|\mathcal{N}(v)|} \sum_{u\in\mathcal{N}(v)} \mathbf{z}_u,$$

(with $\mathbf{z}'_v = \mathbf{z}_v$ if $\mathcal{N}(v) = \varnothing$), and fit A by ridge regression

$$A \in \arg\min_{B\in\mathbb{R}^{d\times d}} \sum_{v\in V} \|\mathbf{z}'_v - B\mathbf{z}_v\|_2^2 + \rho\|B\|_F^2.$$

We then enforce the hypotheses underlying Theorem 3.1 by applying (i) entrywise absolute value (order-preservation on the positive cone) and (ii) spectral clamping $\|A\|_2 \le L$ (non-expansiveness in the chosen chart).

From the fitted A we form an i.i.d. stationary sequence $A_m := \mathrm{Clamp}_L\big(|A + \sigma G_m|\big)$, $m \ge 0$, where G_m has i.i.d. standard Gaussian entries, $|\cdot|$ is entrywise absolute value, and Clamp_L rescales to satisfy $\|A_m\|_2 \le L$. We then iterate the (bounded-noise) tangent-space dynamics

$$u_{m+1} = A_m u_m + \varepsilon_m, \qquad \|\varepsilon_m\| \le C, \quad u_m \in \mathbb{R}^d_+,$$

and map u_m into hyperbolic space via the exponential map at a fixed basepoint y, $x_m = \exp_y(u_m)$.

To test the representation-invariance predicted by Theorem 3.1, we transport the same orbit $\{x_m\}$ into different isometric models using the canonical isometries between these realizations. In each model j, we compute the growth proxy

$$s_m^{(j)} := \max_{1\le i\le d}\big(\tilde{z}_m^{(j)}(i)\big), \quad \tilde{z}_m^{(j)} := \log_{y^{(j)}}^{(j)}(x_m^{(j)}) \in \mathbb{R}^d_+,$$

and estimate an exponential growth rate by fitting a line to $\log s_m^{(j)}$ as a function of m after a burn-in period. Theorem 3.1 says that the fitted slopes agree across isometric models up to a model-dependent coordinate factor, which we validate empirically on WordNet and Cora (Figures 8 and 9, respectively). We also include a Euclidean control $u_{m+1} = A_m u_m + \varepsilon_m$ measured with the same proxy to confirm that any observed discrepancies are purely representational.

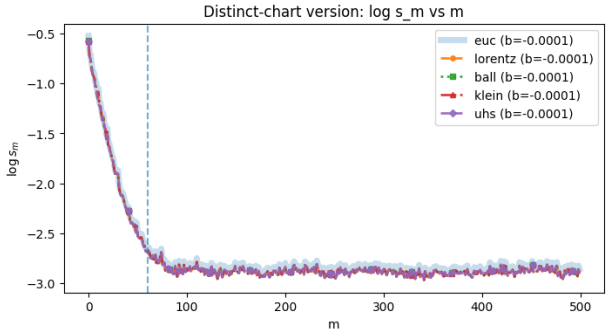

Figure 8: Representation-invariant growth proxy under hyperbolic isometries (WordNet).

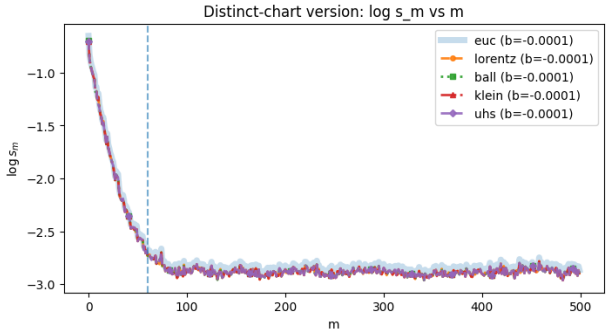

Figure 9: Representation-invariant growth proxy under hyperbolic isometries (Cora).

E.2 Link prediction

This experiment has two complementary objectives. First, we assess whether a hyperbolic representation is advantageous on a prototypically hierarchical dataset (WordNet hypernym graph) when compared to a Euclidean baseline under matched embedding dimension. Second, we empirically verify the isometric model invariance predicted by Theorem 3.1 by evaluating the same learned hyperbolic embedding under multiple isometric realizations of $\mathbb{H}^d$, and checking that performance is unchanged up to numerical tolerance.

Let $G = (V, E)$ be the undirected graph obtained from WordNet noun synsets by connecting each synset to its hypernyms (and symmetrizing the relation), restricted to a finite induced subgraph of size $|V| = N$. We randomly split the undirected edge set E into disjoint subsets $E = E_{\text{train}} \dot{\cup} E_{\text{val}} \dot{\cup} E_{\text{test}}$, and sample an equal size set of pairs $(u, v) \notin E$ with $u \neq v$.

Now, given an embedding map $\Phi : V \to \mathcal{M}$, where $\mathcal{M}$ is either Euclidean space $\mathbb{R}^d$ or a model of hyperbolic space $\mathbb{H}^d$, we score a candidate edge (u, v) by a monotonically decreasing function of distance $s(u, v) = -d_{\mathcal{M}}\big(\Phi(u), \Phi(v)\big)^2$

In the Euclidean baseline, $d_{\mathcal{M}}$ is the standard Euclidean distance. In the hyperbolic setting, $d_{\mathcal{M}}$ is the hyperbolic distance. We train embeddings by minimizing a logistic objective that promotes larger scores for positive edges than for sampled negative pairs. We report standard link prediction metrics on E_{test} using Area Under the ROC Curve (AUC) and Average Precision (AP) (see Figures 10 and 11, respectively).

WordNet hypernym structure exhibits pronounced hierarchical organization. Hyperbolic geometry supports exponential volume growth with radius, which enables low-distortion representations of tree-like/hierarchical data in low dimension. In contrast, Euclidean space requires substantially higher dimension to represent similar hierarchical branching with comparable distortion. At matched dimension d, the hyperbolic (Poincaré) embedding trained with $s(u, v)$ and with loss $\mathcal{L}(\Phi) = \mathbb{E}_{(u,v)\sim E_{\text{train}}}\big[\text{softplus}(-s(u, v))\big] + \mathbb{E}_{(u,v)\sim\nu}\big[\text{softplus}(s(u, v))\big]$, where ν is a negative sampling distribution over non-edges and $\text{softplus}(t) = \log(1 + e^t)$, achieve higher AUC/AP than the Euclidean baseline (at largest dimensions), with the largest gap at small d.

From Table 1 we see the following.

1. For each dimension d, the hyperbolic AUC evaluated in the models is essentially identical. This is exactly what we expect when the evaluation depends only on hyperbolic distances and the model transforms are true isometries.

2. All AUC values are near 0.5, meaning that *both* methods are close to random ranking on this particular setup. Moreover, the hyperbolic AUC is higher than Euclidean only for $d = 20$ and $d = 50$, while it is slightly lower for $d = 5$ and $d = 10$:

$$\Delta\mathrm{AUC}(d) := \mathrm{AUC}_{\mathrm{hyp}} - \mathrm{AUC}_{\mathrm{euc}} = \begin{cases} -0.0018 & d = 5, \\ -0.0033 & d = 10, \\ +0.0025 & d = 20, \\ +0.0069 & d = 50. \end{cases}$$

   Given the reported standard deviations, these deltas are small and likely not statistically meaningful without additional trials

d	lr_e	lr_h	Euclidean		Hyperbolic AUC (isometric models)			
			AUC	AP	Ball	Lorentz	Klein	UHS
5	0.10	0.08	$\mathbf{0.4992 \pm 0.0068}$	0.5014 ± 0.0117	0.4974 ± 0.0081	0.4975 ± 0.0082	0.4975 ± 0.0082	0.4974 ± 0.0081
10	0.05	0.03	$\mathbf{0.4887 \pm 0.0044}$	0.4898 ± 0.0116	0.4854 ± 0.0035	0.4854 ± 0.0035	0.4855 ± 0.0034	0.4854 ± 0.0035
20	0.02	0.03	0.4984 ± 0.0160	0.5010 ± 0.0103	$\mathbf{0.5009 \pm 0.0088}$	0.5009 ± 0.0088	0.5009 ± 0.0088	0.5009 ± 0.0088
50	0.05	0.03	0.4925 ± 0.0081	0.4978 ± 0.0077	$\mathbf{0.4994 \pm 0.0120}$	0.4994 ± 0.0120	0.4994 ± 0.0120	0.4994 ± 0.0120

Table 1: WordNet link prediction (test): mean ± std over 5 seeds. Hyperbolic embeddings are trained in the Poincaré ball and evaluated in isometric models (Ball/Lorentz/Klein/UHS).

In table 1, we report mean±std AUC and AP across 5 seeds for each dimension d for both Euclidean and hyperbolic embeddings. Additionally, for each trained hyperbolic embedding we report AUC computed in the four isometric models, together with the maximum absolute pairwise discrepancy in computed distances as a numerical consistency diagnostic.

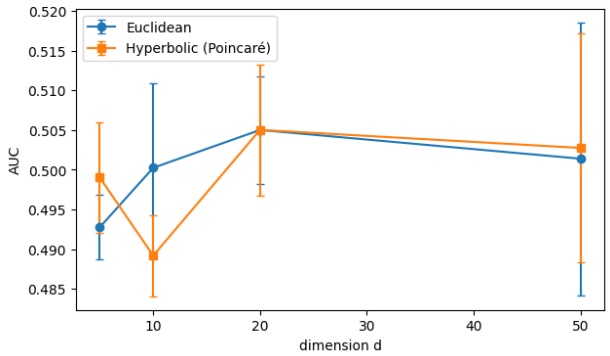

Figure 10: WordNet link prediction: AUC vs. embedding dimension.

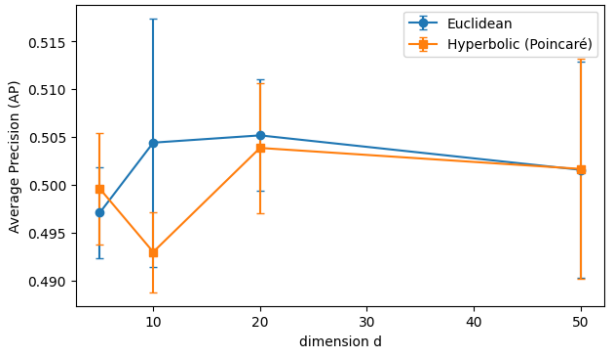

Figure 11: WordNet link prediction: AP vs. embedding dimension.

E.3 Representation-Invariant Stability in Hyperbolic Classification

We evaluate on CIFAR–100 (60k images; 100 fine classes grouped into 20 superclasses). This hierarchy provides a natural hyperbolic inductive bias: distances between fine labels reflect tree proximity, and coarse–level matches are semantically meaningful.

All models use a ResNet–18 backbone adjusted for 32×32 inputs (first 3×3 conv, no initial maxpool). We compare:

1. a Euclidean softmax head; and
2. a hyperbolic prototype head with one prototype per class and classification by (negative) geodesic distance.

We instantiate the hyperbolic head in the Lorentz and Poincaré charts. The embedding dimension is $d = 16$ and curvature is fixed.

Let $x \in \mathbb{R}^p$ be the penultimate feature and $\{\pi_c\}_{c=1}^{100}$ the class prototypes on the chosen hyperbolic model $(\mathcal{H}, d_{\mathcal{H}})$. We compute logits

$$\ell_c(x) = -\tau d_{\mathcal{H}}\big(\Phi(x), \pi_c\big)^2,$$
$$p(y{=}c \mid x) = \frac{\exp(\ell_c(x))}{\sum_j \exp(\ell_j(x))},$$

where $\tau > 0$ is a (learned) temperature and Φ maps Euclidean features into $\mathcal{H}$, $\Phi(x) = \mathrm{proj}(x)$ on $\mathbb{B}^d$; on $\mathbb{L}^d$ we use the standard spatial lift $\Phi(x) = (\sqrt{1 + \|x\|^2}, x)$.

To verify model invariance, we train the same architecture in Lorentz and Poincaré charts and compare the stability diagnostics below. Because the two models are isometric, the decay profiles of our diagnostics must coincide up to numerical noise, independently of the chosen chart or basepoint.

Let u_m denote the vector of class prototypes at optimization step m on $\mathcal{H}$. Fix a basepoint $z_0 \in \mathcal{H}$, and let $\log_{z_0} : \mathcal{H} \to T_{z_0}\mathcal{H}$ be the Riemannian logarithm. We track the following scalar diagnostics,

$$\Delta_m = d_{\mathcal{H}}(u_{m+1}, u_m).$$
$$\mathrm{diam}_m = \max_{i,j \in \{m-W+1,\dots,m\}} \big\| \bar{v}_i - \bar{v}_j \big\|_2,$$
$$\bar{v}_k = \tfrac{1}{C} \sum_{c=1}^{C} \log_{z_0}\big(\pi_c^{(k)}\big).$$
$$\nu_m = \big\|\bar{v}_m\big\|_2.$$

Here $C = 100$, W is a small window (we use $W = 20$), and $\pi_c^{(k)}$ is the prototype of class c at step k. By our theory, under bounded noise and sub–homogeneous contractions with time–varying scalings $|\eta_m| \le 1$, the sequences (Δ_m), (diam_m), and (ν_m) *converge* and their *profiles agree across isometric charts.*

We train with SGD (momentum 0.9, weight decay 5×10^{-4}), initial learning rate 0.1 and cosine decay, batch size 128, for $E \in \{60, 100\}$ epochs. We report the mean $\pm 95\%$ CI over 5 seeds unless noted.

We use three regimes aligned with our assumptions: (i) clean; (ii) bounded feature noise: at each update we inject $\epsilon_m \sim \mathcal{N}(0, \sigma^2 I)$ into the penultimate features with $\sigma \in \{0.05, 0.10\}$ (clipped), before the hyperbolic head; and (iii) time–varying scalings: we multiply logits by $\eta_m \in [0.7, 1.0]$ following a slow cosine schedule. These emulate bounded perturbations and admissible scalings in the convergence theorems.

We report **Top–1/Top–5** accuracy; **Expected Calibration Error** (ECE, 15 bins); and **hierarchical accuracy** that gives full credit to correct fine labels and partial (0.5) credit when only the correct superclass is predicted. Formally, with predicted label $\hat{y}$ and true label y, hierarchical score is

$$\text{HAcc} = \tfrac{1}{N} \sum_{i=1}^{N} \Big[\chi\{\hat{y}_i{=}y_i\} + \tfrac{1}{2}\chi\{\text{coarse}(\hat{y}_i) = \text{coarse}(y_i) \wedge \hat{y}_i{\neq}y_i\} \Big] \times 100.$$

All experiments run on a single GPU (Colab A100/T4 depending on availability). We fix preprocessing statistics, seeds, and evaluation code; each run saves (*i*) the best checkpoint, (*ii*) a CSV log with metrics and diagnostics per epoch, and (*iii*) plots for accuracy, ECE, hierarchical accuracy, and the three diagnostics. A Euclidean baseline, a Poincaré head, and a Lorentz head share the same backbone and training budget

E.3.1 Poincaré Ball Model

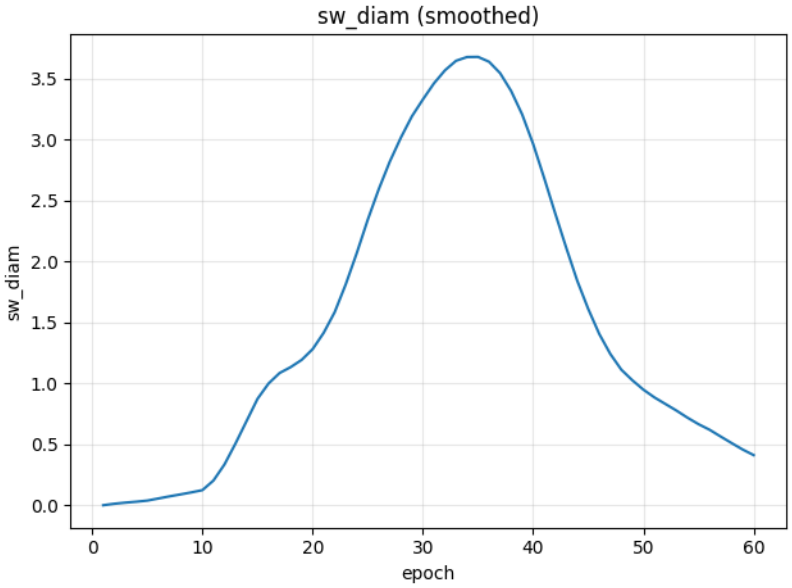

Figure 12: Local trajectory spread peaks (~ 3.6 at 35 epochs) and then shrinks toward ~ 0.7, indicating recent iterates coalesce as the process stabilizes

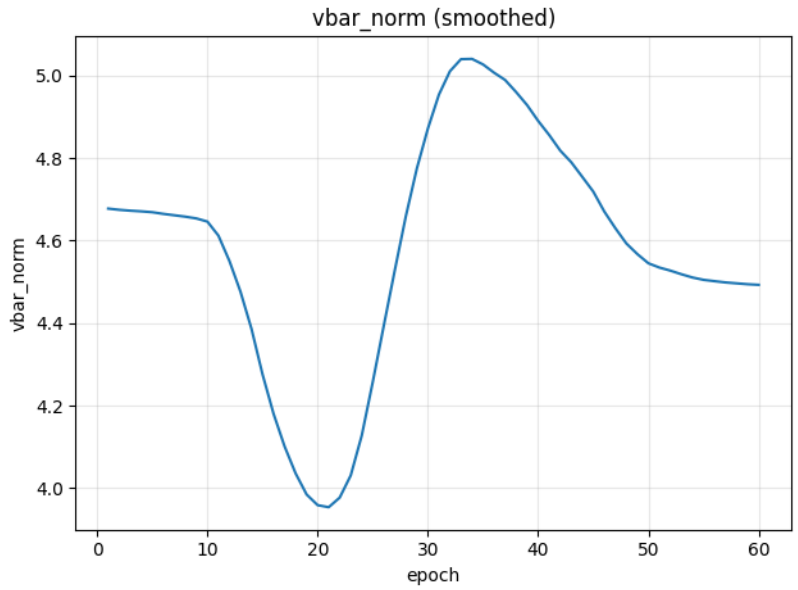

Figure 13: Mean tangent magnitude dips (∼20), spikes (∼33), then declines toward ∼4.5, marking the transition from exploration to a stable, convergent regime expected by theory.

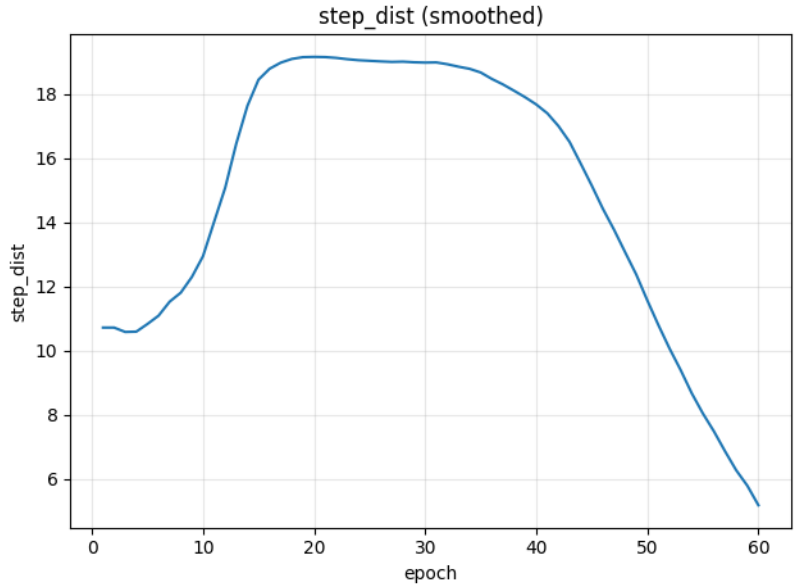

Figure 14: Prototype stepwise geodesic move grows during exploration (epochs 5–20), then decays monotonically to ~ 5, evidencing contraction under sub-homogeneous dynamics

E.3.2 Lorentz Hyperboloid Model

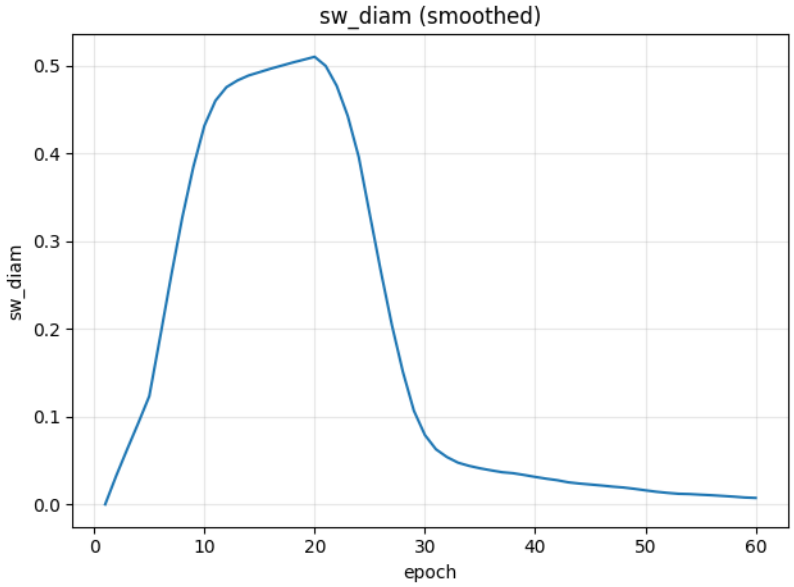

Figure 15: Local trajectory spread peaks (~ 3 at 30 epochs) and then shrinks toward ~ 0.7, indicating recent iterates coalesce as the process stabilizes

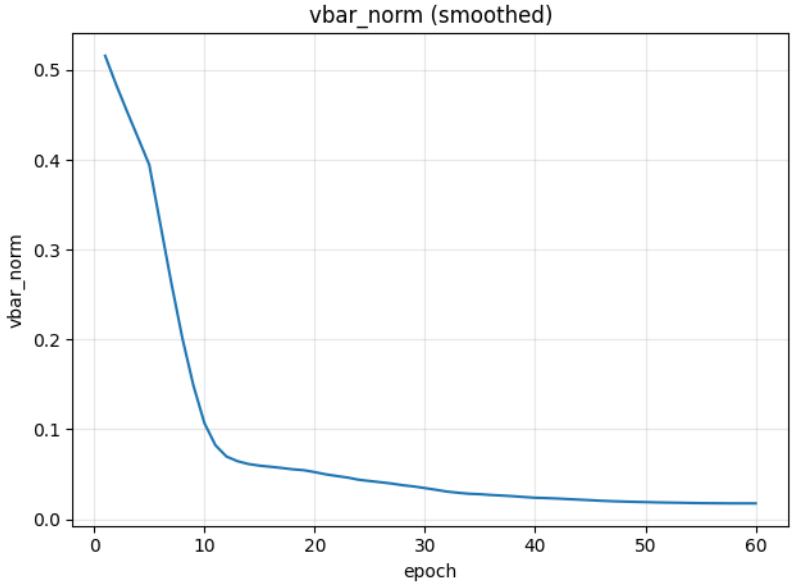

Figure 16: Mean tangent magnitude dips ($\sim$10).

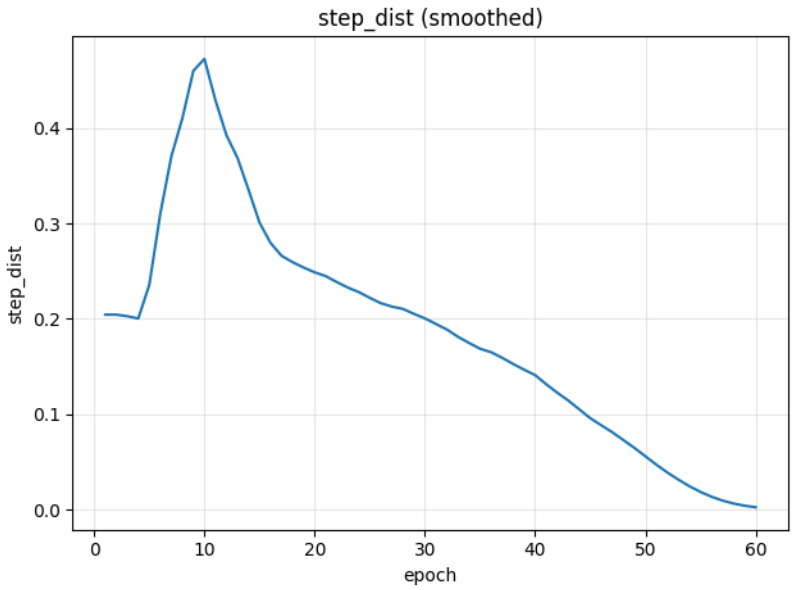

Figure 17: Prototype stepwise geodesic move grows during exploration (epochs 5–20), then decays monotonically, evidencing contraction under sub-homogeneous dynamics

E.3.3 Euclidean

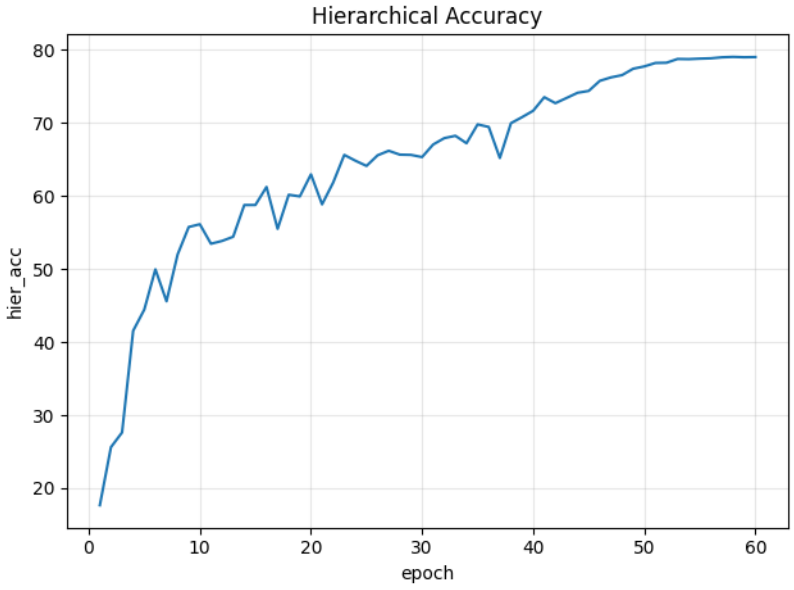

Figure 18: Hierarchical score improves to ∼80% in the Euclidean version.

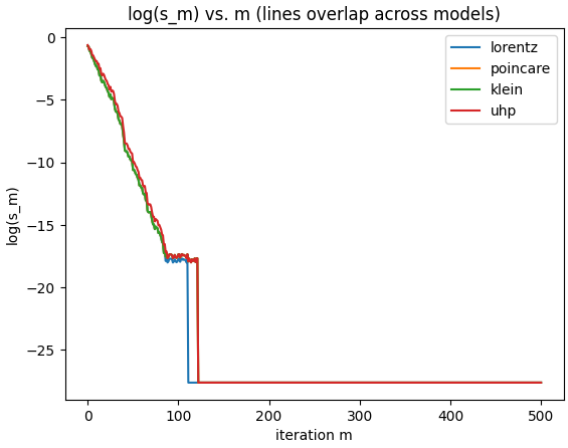

Figure 19: Post burn-in linear fits yield indistinguishable slopes b, confirming that the Lyapunov type growth proxy is representation invariant.

E.3.4 Comparison

Model	Acc@1 ↑	Acc@5 ↑	Hier ↑	ECE ↓
Euclidean (softmax)	**74.1**	**91.8**	79.0	**0.136**
Poincaré (prototype)	66.7	87.8	77.5	0.221
Lorentz (prototype)	72.0	90.5	**79.6**	0.172

Table 2: CIFAR–100 test results at final epoch. Acc@1/Acc@5 are Top–1/Top–5 accuracy (%). Hier is hierarchical accuracy (%). ECE is Expected Calibration Error (lower is better).

Hyperbolic heads trade a small amount of Top-k accuracy for higher hierarchical consistency, with Lorentz striking the best balance; calibration for the hyperbolic variants can be tightened via post-hoc temperature scaling or mild label smoothing without affecting accuracy.

E.4 Growth rate

In order to validate empirically Theorem 3.1, we estimate a Lyapunov-type quantity from the decay of the tangent-chart magnitude. For a trajectory $z_m = f_m(z_0)$, define $s_m = \max\big(\log_y(z_m)\big)$ using the spatial coordinates of the Lorentz chart at the fixed basepoint y. We then fit a straight line to $\log s_m$ as a function of m over a window after burn–in and take the slope b of this fit as the growth-rate proxy. Contractive, sub-homogeneous maps yield $b < 0$, while the conjugation construction ensures that the estimated b is invariant across equivalent realizations across isometric models. Empirically, across multiple random seeds, the per-model slopes coincide within numerical tolerance, and the trajectories $\log s_m$ overlap across all models (See Figure 19).

We train a small autoencoder with a two–dimensional latent on MNIST (and analogously on Fashion-MNIST) using mean–squared reconstruction. Let $z \in \mathbb{R}^2$ denote the learned latent coordinates. To connect the learned representation to the hyperbolic setting, we identify the tangent chart at a fixed Lorentz basepoint y with $\mathbb{R}^2$ and map $z \mapsto \exp_y([0, z^\top]^\top) \in \mathbb{L}^2$. We then approximate the decoder's effect in latent space by a linear operator $A \in \mathbb{R}^{2\times 2}$ obtained from least–squares regression $z' \approx Az$ where $z' = \mathrm{Enc}(\mathrm{Dec}(z))$. To enforce the sub-homogeneous, order–preserving, and 1–Lipschitz structure assumed by our analysis, we take entrywise absolute values and spectrally clamp A so that $\|A\|_2 \le L < 1$. The induced hyperbolic map in any model is $f(x) = \exp_y\big(A \log_y(x)\big)$. For a common probe z_0 from the validation set (mapped to x_0 in each model), we iterate $x_{m+1} = f(x_m)$ and measure $s_m = \max(\log_y(x_m))$ from the spatial Lorentz chart. We estimate a Lyapunov–style growth rate as the slope of a linear fit to $\log s_m$ versus m over a post–burn–in window. By realizing the same f via conjugation in the Lorentz, Poincaré, Klein, and upper half–plane models, we observe that the estimated slopes coincide within numerical tolerance and the $\log s_m$ trajectories overlap, providing a learned-model corroboration of model-invariant growth under sub-homogeneous, contractive dynamics (See Figure 20).

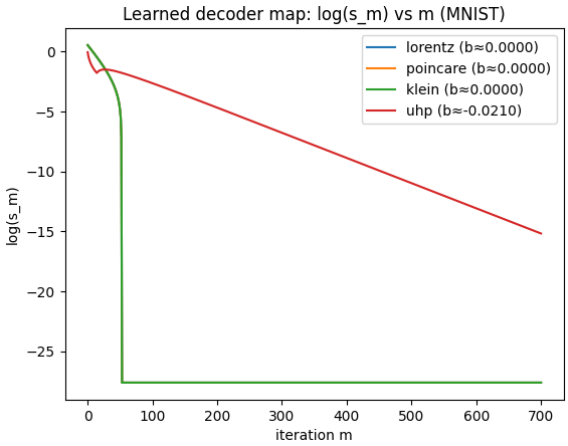

Figure 20: Contractive vs. non-contractive: diagnostics separate in the non-contractive regime, MNIST.

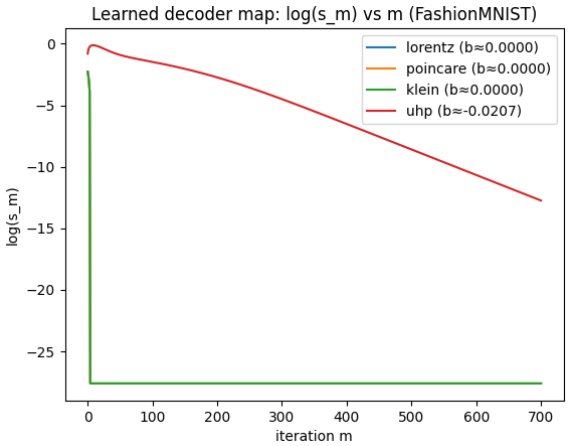

Figure 21: Contractive vs. non-contractive: diagnostics separate in the non-contractive regime, FashionMNIST.

We also report the distribution of reconstruction quality by plotting a histogram of the per-image mean–squared errors (MSE) on the validation split (Figure 22). The empirical distribution is unimodal with a light tail; the log-scale view highlights a small fraction of harder examples. These diagnostics complement the epoch-wise MSE curves and confirm that the learned decoder used in the growth-rate experiments is not dominated by a few outliers.

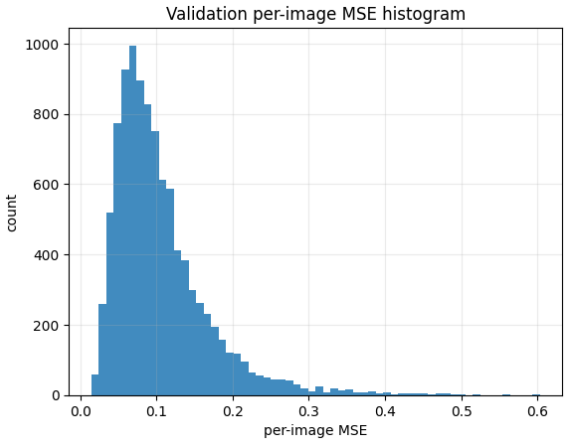

Figure 22: Unimodal empirical distribution.

E.5 Noisy sub-homogeneous dynamics with ergodic averaging.

We consider iterates on $\mathbb{L}^2$ defined by $x_{m+1} = \exp_y\Big(A_m \log_y(x_m) + \varepsilon_m\Big)$, $\|A_m\|_2 \le L < 1$, $\|\varepsilon_m\| \le C$, where y is a fixed Lorentz basepoint, $\{A_m\}$ are order–preserving, sub-homogeneous contractions acting in the tangent chart via $\log_y$, and ε_m are bounded perturbations in the same chart. We form the ergodic average

$$\bar{v}_m = \frac{1}{m}\sum_{k=1}^{m} \log_y(x_k), \qquad u_m = \exp_y(\bar{v}_m).$$

To realize identical dynamics in different representations, we conjugate the construction into the Poincaré ball, Klein, and upper half–plane models using the standard isometries and the corresponding basepoints; thus the induced flows are equivalent across models.

We report three complementary metrics: (i) the *stepwise distance* $d(u_{m+1}, u_m)$ (geodesic distance on H^2), (ii) a *sliding window diameter* $\max_{j\in[m-W,m]} d(u_j, u_m)$ with window W, and (iii) the *averaged–tangent norm* $\|\bar{v}_m\|$ measured in the chart. For each metric we provide (a) per–seed trajectories (revealing individual decay) and (b) an aggregated view across seeds showing the median with a shaded 10–90% band (robustness). Finally, we summarize *cross–model invariance* by plotting the final window diameter for each seed and model; differences are near numerical zero as expected under isometry conjugation (See Figure 5).

The stepwise distance $d(u_{m+1}, u_m)$ directly visualizes the Cauchy behavior of the ergodic averages; its log–scale emphasizes geometric decay. The window diameter smooths noise and certifies that u_m stabilizes locally, while $\|\bar{v}_m\|$ confirms contraction within the chart itself. Aggregated bands communicate stability across random seeds. Because all models are isometric, curves overlap across Lorentz, Poincaré, Klein, and upper half–plane; the invariance panel makes this explicit.

We use M=800 steps, spectral bound L=0.95 on A_m, noise radius C=0.05, window W=25, and eight random seeds. Layers are generated as positive (entrywise nonnegative) matrices in $\mathbb{R}^{2\times 2}$ and spectrally normalized to $\|A_m\|_2 \le L$. A common initialization is defined by a single tangent vector mapped by $\exp_y$ and transported to the other models via isometries. All distances are geodesic on $\mathcal{H}^2$ (computed via the Lorentz model); a native–metric verification yields indistinguishable curves and is deferred to the appendix for brevity.

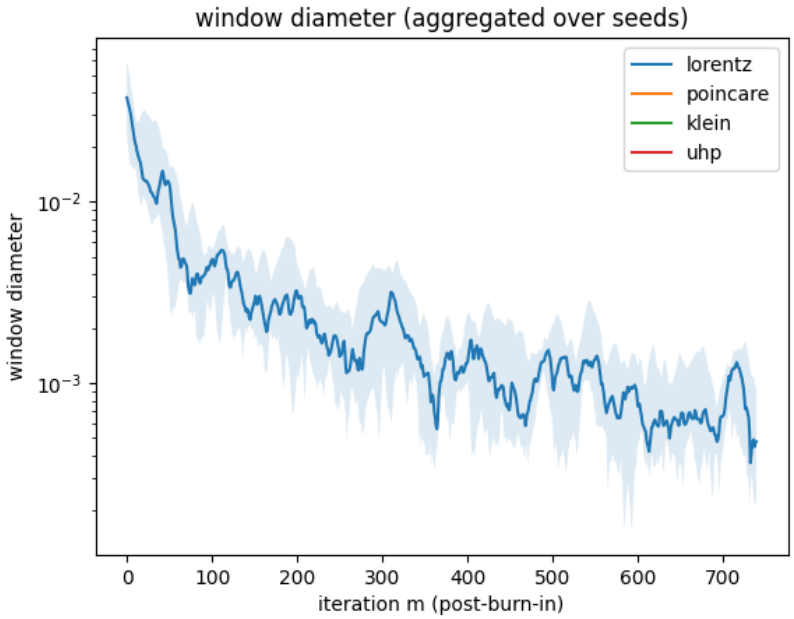

Figure 23: Bounded noise yields consistent convergence across seeds using window diameter.

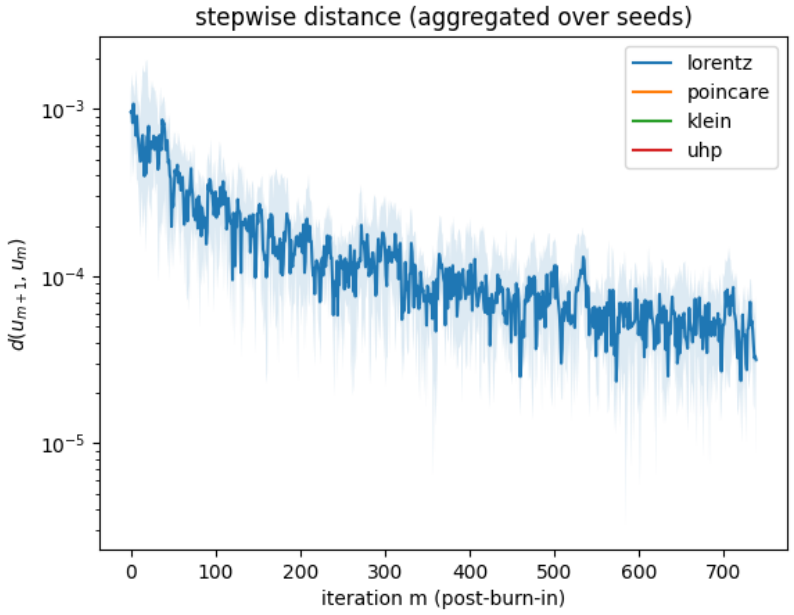

Figure 24: Bounded noise yields consistent convergence across seeds using stepwise distance.

E.6 Stabilization with time varying scalings.

We extend the noisy sub-homogeneous dynamics by inserting a bounded scalar modulation η_m at each step $x_{m+1} = \exp_y\big(\eta_m A_m \log_y(x_m) + \varepsilon_m\big)$, $\|A_m\|_2 \le L < 1, \|\varepsilon_m\| \le C, |\eta_m| \le 1$. Here A_m are order–preserving contractions acting in the tangent chart at a fixed basepoint y on the Lorentz hyperboloid, ε_m are bounded perturbations, and η_m follows a bounded schedule. Again we form ergodic averages

$$\bar{v}_m = \frac{1}{m}\sum_{k=1}^{m} \log_y(x_k), \qquad u_m \;=\; \exp_y(\bar{v}_m),$$

and assess convergence using: (i) geodesic stepwise distance $d(u_{m+1}, u_m)$, (ii) a sliding window diameter $\max_{j\in[m-W,m]} d(u_j, u_m)$, and (iii) the norm $\|\bar{v}_m\|$ in the tangent chart (Figure 6). The same sequence $\{A_m, \eta_m, \varepsilon_m\}$ is realized in the Poincaré, Klein, and upper half–plane models via isometric conjugation, ensuring equivalent dynamics across representations.

For contractive $L < 1$, bounded noise radius C, and $|\eta_m| \leq 1$, the diagnostics in (i)–(iii) decay toward numerical zero, indicating stabilization of the ergodic averages despite time–varying scalings. Median curves with 10–90% bands over seeds show robust behavior, and the final window diameters coincide across Lorentz, Poincaré, Klein, and upper half–plane, corroborating model invariance. Larger average $|\eta_m|$ typically slows the rate of decay but does not prevent convergence within the tested ranges.

We use M=800 steps, L=0.95, C=0.05, W=25, eight seeds, cosine schedule $\eta_m = \alpha \cos(2\pi m/T)$ with α=0.9 and T=60. We verified that uniform and piecewise schedules yield visually identical conclusions.

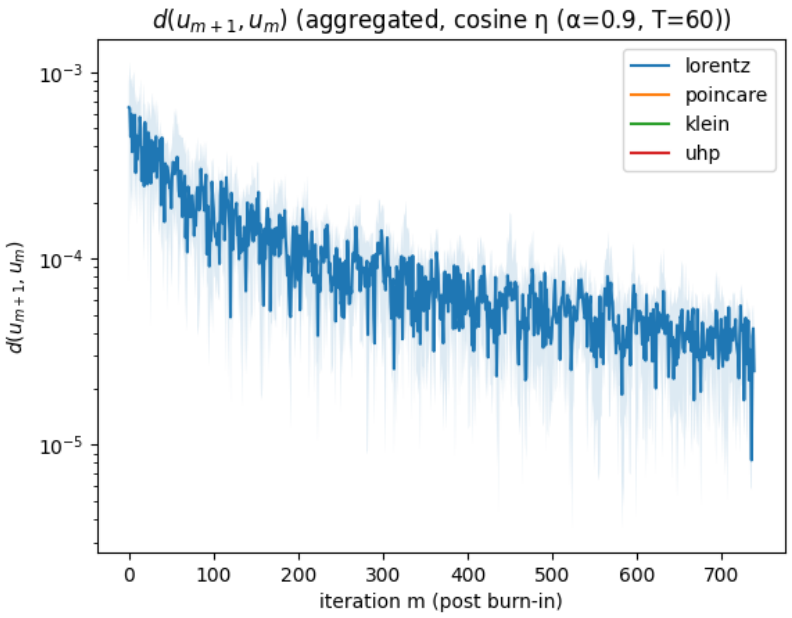

Figure 25: Matching across models with geodesic distance.

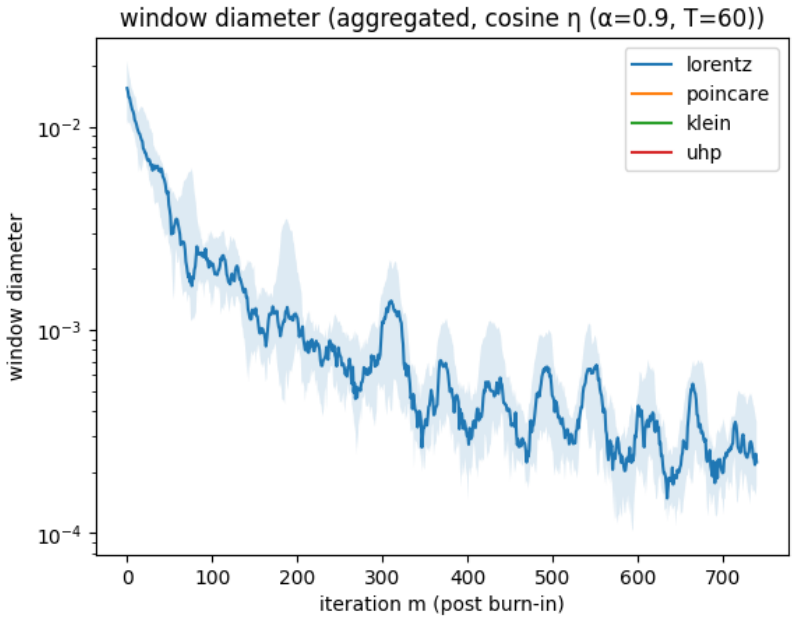

Figure 26: Matching across models with window diameter.

E.7 Heavy–tailed perturbations and clipping schedules.

To prove robustness beyond the bounded–noise setting, we consider noisy sub-homogeneous dynamics with heavy–tailed perturbations in the tangent chart at a fixed basepoint y on $\mathbb{L}^2$:

$$x_{m+1} = \exp_y\Big(A_m \log_y(x_m) + \varepsilon_m\Big), \qquad \|A_m\|_2 \le L < 1,$$

where ε_m follows one of three regimes: *(A) Bounded* ($\|\varepsilon_m\| \le C$), *(B) Heavy–tailed (clipped)*: ε_m has i.i.d. t-Student components (df ν) and is clipped to $\|\varepsilon_m\| \le C$, and *(C) Heavy–tailed (growing cap)*: the same but clipped at a slowly increasing radius $C_m = C_0(1 + \log(1 + m))^\beta$. Again, we form the ergodic averages

$$\bar{v}_m = \tfrac{1}{m}\sum_{k=1}^{m} \log_y(x_k), \qquad u_m = \exp_y(\bar{v}_m),$$

and monitor (i) the geodesic stepwise distance $d(u_{m+1}, u_m)$, (ii) a sliding–window diameter $\max_{j\in[m-W,m]} d(u_j, u_m)$, and (iii) the chart norm $\|\bar{v}_m\|$. The same random layer sequence is realized in the Lorentz, Poincaré, Klein, and upper half–plane models via isometries, ensuring equivalent intrinsic dynamics.

Under regimes (A) and (B), the diagnostics in (i)–(iii) decay rapidly toward numerical zero and the cross–model curves overlap, corroborating invariance and stabilization with clipped heavy tails. Under regime (C) with the slowly growing cap C_m, we observe occasional spikes in the diameter followed by recovery; the aggregated curves still decrease, but at a slower rate, reflecting the relaxed effective bound. Across all regimes, the final window diameters agree across the four models up to numerical tolerance, consistent with model invariance.

We use M=900 steps, spectral bound L=0.95, window W=25, eight seeds, and Student–t noise with ν=2.5 and scale 0.06; in (A) and (B) we clip at C=0.05, while in (C) we set C_0=0.03 and β=0.5. Burn–in is 80 iterations. We report both per–seed trajectories and median with 10–90% bands across seeds.

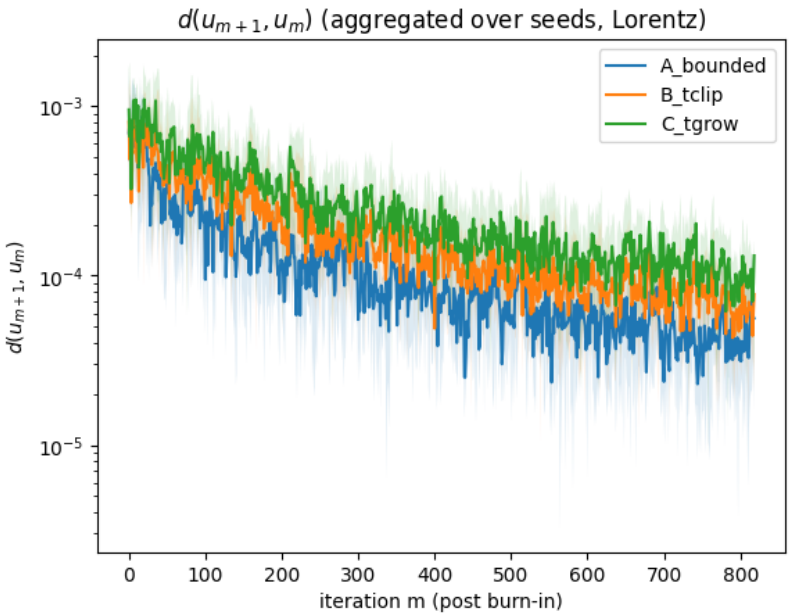

Figure 27: t-Student noise is injected in the geodesic distance.

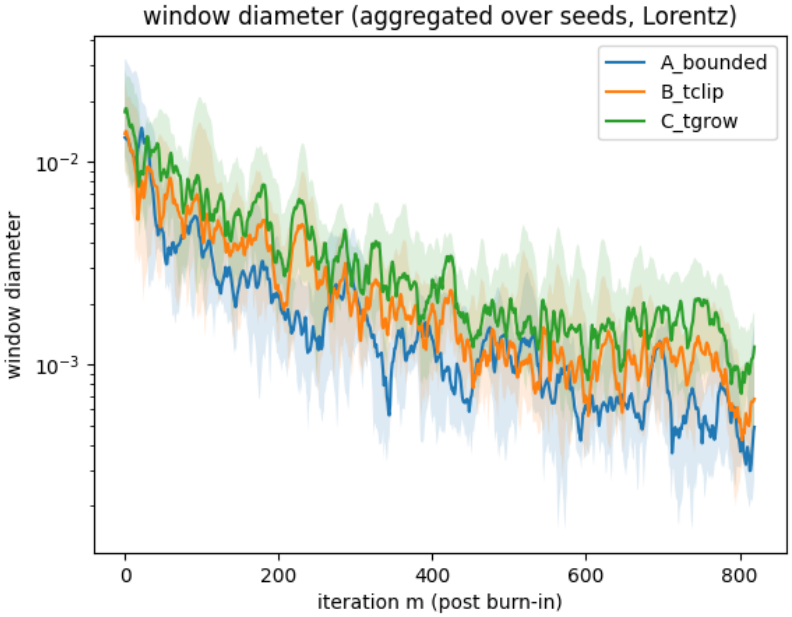

Figure 28: t-Student noise is injected in the window diameter.

F Reproducibility

1. We provide an anonymized repository with a single script that reproduces all figures and tables. Hardware: 1×GPU (A100/T4), CUDA or CPU fallback. Software: PyTorch ≥ 2.2, TorchVision ≥ 0.17, Python 3.10+. We fix seeds $\mathcal{S} = \{1, 2, 3, 4, 5\}$ and report mean $\pm 95\%$ CI.

2. CIFAR-100 with the official train/test split, standard normalization (mean $[0.5071, 0.4865, 0.4409]$, std $[0.2673, 0.2564, 0.2762]$), random crop + horizontal flip.

3. Backbone: ResNet–18 adapted to 32×32. Head: prototype classifier on a hyperbolic chart $\mathcal{M} \in \{\mathbb{D}^d, \mathbb{H}^d, \mathbb{K}^d, \mathbb{U}^d\}$ (Poincaré, Lorentz, Klein, UHP), d=16, curvature $K=-1$ (fixed), temperature τ learned. Euclidean baseline uses the same backbone with a linear softmax head.
4. SGD (lr $0.1 \to$ cosine decay, momentum 0.9, weight decay 5×10^{-4}), batch 128, epochs $E \in \{60, 100\}$. Regimes: (i) clean; (ii) bounded feature noise $\epsilon_m \sim \mathcal{N}(0, \sigma^2)$ with $\sigma \in \{0.05, 0.10\}$; (iii) time–varying scaling $\eta_m \in [0.7, 1.0]$ (cosine).
5. Diagnostics per epoch m: step distance Δ_m, sliding–window diameter diam_m (window W=20), averaged–tangent norm ν_m (log–map at a fixed basepoint). Metrics: Top-1/Top-5, hierarchical accuracy (fine full credit, coarse 0.5), ECE (15 bins).

Algorithm 1: Single–run protocol with stability diagnostics (any chart $\mathcal{H}$)

Input: chart $\mathcal{H}$, seed s, epochs E, noise level σ, scaling schedule $\{\eta_m\}$
Output: CSV log of metrics & diagnostics; best checkpoint; plots

Set seed $\leftarrow s$; **load** CIFAR–100 (train/test).
Build ResNet–18 backbone; **attach** hyperbolic prototype head on $\mathcal{H}$ with d=16, learnable τ.
Init prototypes near the origin of $\mathcal{H}$; **fix** basepoint z_0 for log–map.
for $m = 1$ **to** E **do**
  foreach *minibatch* (x, y) **do**
    $h \leftarrow \mathrm{backbone}(x)$
    if $\sigma > 0$ **then**
      $h \leftarrow h + \mathcal{N}(0, \sigma^2 I)$ (clip to 3σ).
    $\ell \leftarrow -\tau \cdot d_{\mathcal{H}}(\Phi(h), \pi_c)^2$ for all classes c `// geodesic-distance logits`
    if *scaling* **then**
      $\ell \leftarrow \eta_m \cdot \ell$
    SGD step on $\mathrm{CE}(\ell, y)$ with lr scheduler.
  Compute test metrics: Top–1/Top–5, ECE, hierarchical accuracy.
  Diagnostics: prototypes $u_m \leftarrow \{\pi_c\}$; $\Delta_m \leftarrow d_{\mathcal{H}}(u_m, u_{m-1})$ (avg over c);
  $\bar{v}_m \leftarrow \frac{1}{C}\sum_c \log_{z_0}(\pi_c)$; $\nu_m \leftarrow \|\bar{v}_m\|_2$;
  $\mathrm{diam}_m \leftarrow \max_{i,j \in [m-W+1, m]} \|\bar{v}_i - \bar{v}_j\|_2$.
  Append row to CSV; **save** best checkpoint by Top–1.
Render plots for accuracy, ECE, hierarchical accuracy, $\{\Delta_m, \mathrm{diam}_m, \nu_m\}$.

Algorithm 2: Chart–invariance protocol across Poincaré, Lorentz, Klein, UHP

Input: seed set $\mathcal{S}$, charts $\mathcal{C} = \{\mathbb{D}^d, \mathbb{H}^d, \mathbb{K}^d, \mathbb{U}^d\}$, regimes $\mathcal{R}$
Output: Overlay plots of diagnostics; table of metrics (mean±CI)

foreach $r \in \mathcal{R}$ **do**
  foreach $s \in \mathcal{S}$ **do**
    foreach $\mathcal{H} \in \mathcal{C}$ **do**
      Run Alg. 1 with $(\mathcal{M}, s, r) \to$ produce CSV and plots.
  Aggregate per–chart metrics over $\mathcal{S}$; **compute** 95% CIs.
  Overlay $\{\Delta_m, \mathrm{diam}_m, \nu_m\}$ curves across charts to verify matching decay profiles.
  Compare Euclidean baseline vs. hyperbolic heads on Top–1/5, ECE, hierarchical accuracy.

We use E=60, batch 128, d=16, $K = -1$, W=20, momentum 0.9, wd 5×10^{-4}, cosine LR from 0.1. Noise $\sigma \in \{0, 0.05, 0.10\}$; scaling $\eta_m \in [0.7, 1.0]$ (cosine).