Convergence Properties of Hyperbolic Neural Networks on Riemannian Manifolds

Nico Alvarado
nfalvarado@mat.uc.cl

Sebastian Burgos
smb7512@psu.edu

Abstract

Hyperbolic neural networks have attracted increasing attention within the community in recent years, with various empirical studies on the subject standing out. However, there is little theoretical research on this topic. In this work, we use results from Avelin and Karlsson to ensure convergence of hyperbolic neural networks defined in the Lorentz hyperboloid model. Also, we extend this result to any Riemannian manifold.

1 Introduction

Hyperbolic neural networks (HNNs) represent an emerging field within machine learning and artificial intelligence, characterized by their use of hyperbolic geometry to enhance the performance and representation capabilities of neural networks ([10, 9, 15]). Despite their promising potential, the domain is still in early stages, and several critical aspects require further exploration and improvement.

Theoretical foundations and practical applications of hyperbolic neural networks are not yet fully understood. There is a need for more rigorous mathematical formulations and proofs to establish the fundamental principles governing these networks ([4, 16, 14, 1])

Applying concepts from dynamical systems and ergodic theory to the convergence of neural networks can lead to significant improvements. By treating neural networks as dynamical systems, we can better understand the stability of their training processes. This approach helps identify stable solutions and avoid unstable ones, resulting in more reliable convergence to optimal solutions. Ergodic theory, which deals with the statistical behavior of systems over time, can improve our understanding of convergence patterns ([2]).

Additionally, dynamical systems techniques can lead to new regularization methods that enhance the generalization ability of neural networks ([8, 12]). By understanding parameter trajectories, we can design regularization methods to prevent overfitting, ensuring the network generalizes better to unseen data. Ergodic theory also helps mitigate chaotic behavior during training, leading to more stable and predictable training dynamics.

Convergence of neural networks can be understood through the lens of optimization theory. Training a neural network typically involves minimizing a loss function using gradient-based methods. Theoretical results shows that gradient descent methods converge to minimizers under certain conditions ([6, 3, 13]).

On the other hand, in [2] the authors proves several important results about convergence of neural networks. First, the growth is realized at one fixed coordinate, avoiding spiralling in a cone and showing that more complicated fluctuating behaviour is not possible.

Theorem 1.1 ([2])**.** *Let X be the positive cone in $\mathbb{R}^N$ and let $T_i : X \to X$ be a stationary sequence of maps that is order preserving and subhomogeneous. Let $x_n = T_1 T_2 \ldots T_n x_0$, for a fixed $x_0 \in X$. Then*

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

Mathematics of Modern Machine Learning Workshop at NeurIPS 2024.

and there is a (random) coordinate $1 \leq i_0 \leq d$ such that

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

Here λ is the average logarithmic rate of contraction (or expansion) of distances in the space under the transformations.

Another important result is that no matter where you start in your (Euclidean) space and despite any randomness in your sequence of transformations, the result will lead you to a specific, predictable outcome. The process "forgets" the starting point and converges to a consistent average behavior.

Theorem 1.2 ([2])**.** *Let $(X = \mathbb{R}^N, \|\cdot\|_N)$ be a normed vector space which has the above monotonicity property and strictly convex unit ball. Consider a stationary sequence of layer maps T_n of the form $T(x) = W^T\sigma(Wx + b)$, $\|W\|_N \leq 1$, $b \in X$, and σ is 1-Lipschitz when applied componentwise in $(X = \mathbb{R}^N, \|\cdot\|_N)$. Then as $n \to \infty$ it holds that a.s. there exists a vector v such that*

$$\frac{1}{n} T_1 T_2 \ldots T_n x_0 \to v.$$

The vector v is a priori random but independent of the initial data x_0. The norm of v is deterministic.

And finally, there's a consistent, average rate at which these transformations stretch distances between points. If this rate is positive, even tiny differences can grow exponentially large after many transformations. This behavior is predictable and quantifiable, allowing us to understand and anticipate how the system evolves over time.

Theorem 1.3 ([2])**.** *Under certain assumptions there is a number λ so that*

$$\lim_{n\to\infty} \left(\sup_{x\neq y} \frac{\|T_n T_{n-1} \ldots T_1 x - T_n T_{n-1} \ldots T_1 y\|}{\|x - y\|} \right)^{1/n} = e^{\lambda}, \qquad a.s.$$

Moreover, in case $\lambda > 0$ there exists a point $x \in \Omega$ and a sequence $z_i = (x_i, y_i) \in \{(x, y) : x, y \in \Omega, x \neq y\}$ such that $z_i \to (x, x)$ and for any $\epsilon > 0$ there is a number p so that for $n > p$

$$\frac{\|T_n \ldots T_1 x_i - T_n \ldots T_1 y_i\|}{\|x_i - y_i\|} \geq e^{(\lambda-\epsilon)n},$$

for all sufficiently large i.

The main contributions of this work are:

1. Using Theorem 1.2, we prove convergence of Neural Networks defined in a Riemannian manifolds under certain conditions (Theorem 3.2). This result suggests that under the given conditions, the system stabilizes in a well-defined manner, providing predictability and stability in applications such as iterative algorithms and deep learning models on manifolds.
2. We found specific parameters of the initial data that allows convergence of the HNN defined in the Lorentz hyperboloid model (Theorem 3.1).
3. We found convergence in reverse dynamics of layers maps, specifically in compact metric spaces (Theorem 3.4).

2 Preliminaries

Riemannian geometry basics (see [7]). A d-dimensional differentiable manifold M is a topological space that is locally parameterized by open sets of $\mathbb{R}^d$, such that every change of parametrization is a differentiable map. It is possible to define infinitesimal directions at each point $p \in M$, forming the tangent space T_pM of M at p. If a differentiable manifold M has an inner product $g_p(\cdot, \cdot)$ defined on each tangent space $T_pM \simeq \mathbb{R}^d$ (called a Riemannian metric), it is called a Riemannian manifold and it is denoted by (M, g). By integrating the $g_{\gamma(t)}$-norm of the tangent vectors along a curve $\gamma(t)$, we can define the Riemannian length of a curve and the minimum length required to connect two points gives a Riemannian distance on X. A geodesic is a curve on a Riemannian manifold that locally minimizes the length between its endpoints.

We need to relate the manifold M with its tangent space T_pM for some $p \in M$. This can be done using the so called *exponential map*. It is well known that for every $v \in T_pM$ (with sufficiently small norm) there is a unique geodesic $\gamma_v \colon (-2, 2) \to M$ such that $\gamma_v(0) = p$ and $\gamma_v'(0) = v$. Then, the exponential map at $p \in M$ is defined for $v \in T_pM$ by $\exp_p(v) := \gamma_v(1)$. Here γ' is the derivative of γ.

Proposition 2.1. *Given $p \in M$, there is $\varepsilon > 0$ such that $\exp_p \colon B_\varepsilon(0) \subset T_pM \to M$ is a diffeomorphism onto an open set of M. Here $B_\varepsilon(0)$ is the ball of radius ε centered at 0.*

Hyperbolic spaces (see Figure 1). In contrast to Euclidean geometry, hyperbolic geometry possesses unique geometric properties. A comprehensive understanding of this geometry requires familiarity with the hyperbolic parallel postulate and the concept of curvature. Furthermore, it is imperative to comprehend the representation and visualization of hyperbolic space using models such as the Lorentz hyperboloid model (see Figure 2).

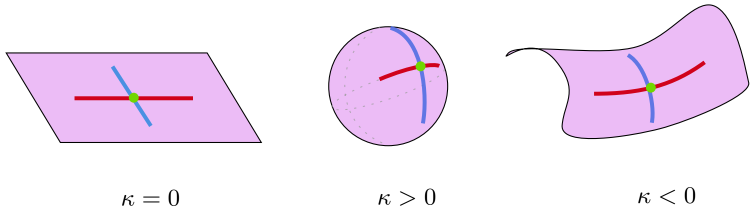

Figure 1: Surfaces with different curvatures κ. Examples of plane geometry, spherical geometry and hyperbolic geometry respectively from left to right.

Hyperboloid Model The hyperboloid model is a classical representation of hyperbolic geometry, a non-Euclidean geometry characterized by a constant negative curvature. This model is named for its use of a hyperboloid surface to describe the geometry's structure (see Figure 2). It offers an intuitive and algebraically convenient way to understand and work with hyperbolic spaces.

We consider the set $\mathbb{L}^n := \left\{x \in \mathbb{R}^{n+1} : -x_0^2 + \sum_{i=1}^n x_i^2 = -1,\, x_0 > 0\right\}$

and we fix its origin $y = (1, 0, \ldots, 0) \in \mathbb{L}^n$. We identify the tangent space of $\mathbb{L}^n$ at y with $\mathbb{R}^n$ by setting $T_y\mathbb{L}^n = \{v \in \mathbb{R}^{n+1} : v_0 = 0\}$.

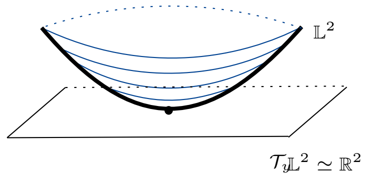

Figure 2: The tangent space $T_y\mathbb{L}^2$.

In this setting the exponential map $\exp_y \colon T_y\mathbb{L}^n \to \mathbb{L}^n$ is invertible and its inverse is denoted by $\log_y \colon \mathbb{L}^n \to T_y\mathbb{L}^n$. For $v \in T_y\mathbb{L}^n$ and $x \in \mathbb{L}^n$, these maps satisfy

$$\exp_y(v) = \cosh(\|v\|)y + \sinh(\|v\|)\frac{v}{\|v\|}, \qquad \log_y(x) = d(x, y)\frac{x + g(x, y)y}{\|x + g(x, y)y\|},$$

where $d(x, y) = \operatorname{arccosh}(x_0y_0 - \sum_{i=1}^n x_iy_i)$ and $g(x, y) = -x_0y_0 + \sum_{i=1}^n x_iy_i$ (for more details see [5]).

Definition 2.2. A subset $X \subset \mathbb{L}^n$ is called a *cone* if it is the image of a cone in $T_y\mathbb{L}^n$ under the exponential map.

Möbius operations In order to state the results in $\mathbb{L}^n$, we need to be able to add elements and multiply them by a scalar. We do this by using the usual addition and multiplication by scalar in $\mathbb{R}^n$, together with the exponential and logarithm maps.

Definition 2.3. For $a, b, x \in \mathbb{L}^n$ 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 $\alpha \otimes x = \exp_y(\alpha \log_y(x))$.

This definition can be extended to a Riemannian manifold.

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 definitions.

Definition 2.4. We define a Deep Neural Network as

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

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

In order to define a neural network on the hyperboloid model, a good idea is to "transfer" a neural network defined on $\mathbb{R}^n$ to $\mathbb{L}^n$ using the exponential map and its inverse.

Definition 2.5. Given a function $T\colon \mathbb{R}^n \to \mathbb{R}^n$, we define the Möbius version of T by

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

Note: On the other hand, any map $f\colon \mathbb{L}^n \to \mathbb{L}^n$ has its "Euclidean version" by applying the reverse process. That is, the map $\tilde{f}(v) = \log_y(f(\exp_y(v)))$ is such that $(\tilde{f})^{\otimes} = f$.

Definition 2.6. We define a Hyperbolic Neural Network as

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

where $W_i \in \mathbb{R}^{n \times n}$, $b_i \in \mathbb{L}^n$ and σ is the activation function. Recall that we are always identifying $T_y\mathbb{L}^n \simeq \mathbb{R}^n$.

Remark 2.7. Both deep neural networks and hyperbolic neural networks can be defined between spaces of different dimensions, that is, the matrices W_i do not need to be square matrices and we then should use the corresponding exponential and logarithm of the "correct dimension". However, in our work we will restrict to networks between spaces of the same dimension n.

Ergodic theorems and cocycles Let $(M, \mathcal{B}, \mu, T)$ be an ergodic dynamical system, that is, $T\colon M \to M$ is a measurable transformation, $\mu(T^{-1}A) = \mu(A)$ and all T-invariant sets have measure either 0 or 1. A *subadditive cocycle* over T is a measurable function $\phi : M \times \mathbb{N}_0 \to \mathbb{R}$ satisfying

$$\phi(\omega, n+m) \leq \phi(\omega, n) + \phi(T^n\omega, m) \quad \text{for all } \omega \in M \text{ and } n, m > 0.$$

For convenience set $\phi(\omega, 0) \equiv 0$. For a subadditive cocycle coming from a sequence of maps (e.g.Furstenberg-Kesten theorem), it is more convenient to write the subadditive cocycle property as $a(1, n+m) \leq a(1, n) + a(n, m)$ (see [2] for more details).

Denote by $SC(X)$ the set of all non-expansive maps on the metric space X. Consider a map $\varphi\colon M \to SC(X)$. The so called *ergodic cocycle* $u(\omega, n) = \varphi(\omega)\varphi(T\omega)\cdots\varphi(T^{n-1}\omega)$ is called *integrable* if $\int_M d(\varphi(\omega)x, x)d\mu(\omega) < \infty$.

This condition does not depend on the point x due to the maps $\varphi(\omega)$ being non-expansive.

These ergodic cocycles when used with sequences of layer maps will be called *stationary sequences*. For more details see [11] and [2].

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

Definition 2.8. Define a partial order in $\mathbb{L}^n$ as follows. For $x, x' \in \mathbb{L}^n$,

$$x \leq x' \quad \Longleftrightarrow \quad \log_y(x) \leq \log_y(x'),$$

where the order in the right hand side is the partial order in $T_y\mathbb{L}^n \cong \mathbb{R}^n$.

Definition 2.9. Let $X \subset \mathbb{L}^n$ 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 $f(\lambda \otimes x) \leq \lambda \otimes f(x)$, whenever the order is possible.

Proposition 2.10. *Let $f\colon T_y\mathbb{L}^n \to T_y\mathbb{L}^n$ be subhomogeneous. That is, for every $v \in T_y\mathbb{L}^n$ and $\lambda \in (0,1)$ we have $f(\lambda v) \leq \lambda f(v)$ whenever the order is possible. Then, the induced map on the hyperboloid $f^{\otimes}\colon \mathbb{L}^n \to \mathbb{L}^n$ is also subhomogeneous.*

Proof. Is direct from the definition of subhomogeneity and $f^{\otimes}$. □

3 Main results

Our results usually apply for general sequence of maps $f_m\colon \mathbb{L}^n \to \mathbb{L}^n$ satisfying the corresponding assumptions of each theorem, in particular they apply for hyperbolic neural networks from Definition 2.6, as these will come from deep neural networks on the tangent space $\mathbb{R}^n$. This results are relevant because they broaden the scope of the results presented in [2], demonstrating that the behaviors of convergence, and exponential growth rates are pervasive across various mathematical structures and types of transformations.

The following theorem is a weighted version of Theorem 1.1 in the hyperboloid model, where we obtain a similar convergence of the coordinates.

Theorem 3.1. *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 subhomogeneous 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*

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

Proof. Fix $z_0 \in Y$. 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 subhomogeneous. Thus, we can apply Theorem 1.1 using the point $x_0 := \log_y(z_0) \in X$. Observe that

$$z_m = f_1 f_2 \cdots f_m(z_0) = \exp_y(T_1 T_2 \cdots T_m(x_0)) = \exp_y(x_m),$$

where we use the notation of Theorem 1.1. Using the specific expression we have for the exponential map, we obtain

$$\begin{aligned} z_m = (z_m(0), \ldots, z_m(n))) &= \cosh\|x_m\| y + \frac{\sinh\|x_m\|}{\|x_m\|} x_m \\ &= \left(\cosh\|x_m\|, \frac{\sinh\|x_m\|}{\|x_m\|} x_m(1), \ldots, \frac{\sinh\|x_m\|}{\|x_m\|} x_m(n) \right). \end{aligned}$$

For $1 \leq i \leq n$ we have $x_m(i) = \frac{\|x_m\|}{\sinh\|x_m\|} z_m(i)$. Observe that $\|x_m\| = \operatorname{arccosh}(z_m(0))$ and $\|z_m\|^2 = \cosh^2\|x_m\| + \sinh^2\|x_m\| = 1 + 2\sinh^2\|x_m\|$. The result follows as

$$x_m(i) = \frac{\sqrt{2}\,\operatorname{arccosh}(z_m(0))}{\sqrt{\|z_m\|^2 - 1}} z_m(i).$$

□

The following theorem extends neural network layer dynamics to Riemannian manifolds, showing that iterated transformations in the tangent space, via the exponential map, converge to a stable point.

Theorem 3.2. *Let (M, g) be a Riemannian manifold. Fix $y \in M$ and $r > 0$ such that $\varphi := \exp_y : B_r(0) \subset T_y M \to V := \exp_y(B_r(0))$ is a diffeomorphism (up to reducing r we can assume that φ extends continuously to the closure of the ball of radius r centered at 0, $\overline{B_r(0)}$). Consider a sequence $f_n : V \to V$ consisting of maps of the form*

$$f(x) = \varphi(W^\top \sigma(W\varphi^{-1}(x) + b)),$$

where $\|W\| \leq 1$, σ is 1-Lipschitz componentwise and $b \in T_y M$ satisfy $f_n(V) \subset V$, and such that $\tilde{f}_n(v) = W_n^\top \sigma(W_n v + b_n))$ is a stationary sequence of layer maps in $\mathbb{R}^n$. Then, as $m \to \infty$, almost surely there exist $z \in V$ such that

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

The point $z \in V$ is independent of the initial data $z_0 \in V$.

Proof. Let $U := B_r(0)$, and consider the sequence $T_n : U \to U$ defined by $T_n(v) := \tilde{f}_n(v)$. This sequence is stationary and satisfies the hypotheses of Theorem 1.2 (we use the Euclidean norm in $T_y M \simeq \mathbb{R}^n$). Fix an initial data $z_0 \in V$, which gives us an initial data $x_0 := \varphi^{-1}(z_0) \in U$. By Theorem 1.2, as $m \to \infty$ almost surely there is $x \in \overline{U}$ such that

$$\frac{1}{m} T_1 \cdots T_m(x_0) \to x.$$

Since

$$\frac{1}{m} \otimes f_1 \cdots f_m(z_0) = \varphi\left(\frac{1}{m}\varphi^{-1}(f_1 \cdots f_m(z_0))\right) = \varphi\left(\frac{1}{m} T_1 \cdots T_m(x_0)\right),$$

we obtain the desired limit with $z := \varphi(x)$ as φ is continuous. □

This result demonstrate a strong convergence property for a specific type of functions defined on a Riemannian manifold. Despite the complexity and nonlinearity of the maps f_n, the sequence of compositions of these maps, when averaged properly, converges almost surely to a single point.

The following result is an immediate application of Theorem 3.2 to the hyperboloid model. Recall that in this situation the map $\varphi := \exp_y$ is invertible on the whole tangent space at the origin $T_y \mathbb{L}^n$.

Corollary 3.3. *Let $f_m : \mathbb{L}^n \to \mathbb{L}^n$ be a sequence of maps of the form $f_m(x) = T_m^\otimes(x)$, where $T_m : \mathbb{R}^n \to \mathbb{R}^n$ 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 $z_0 \in \mathbb{L}^n$.

It is easy to see that Corollary 3.3 applies to any isometric hyperbolic manifold (e.g. Poincaré ball model).

We now extend the example of the reverse order in [2] to a compact metric space setting.

Theorem 3.4. *Let (Ω, d_0) be a compact metric space and consider a stationary sequence of homeomorphisms $T_m : \Omega \to \Omega$. Then, almost surely there is a number λ such that*

$$\lim_{m \to \infty} \left(\sup_{x \neq y} \frac{d_0(T_m T_{m-1} \cdots T_1 x, T_m T_{m-1} \cdots T_1 y)}{d_0(x, y)} \right)^{1/m} = e^\lambda.$$

Proof. See Appendix A. □

4 Conclusions and future work

In this work we extend some convergence results of neural networks from the Euclidean setting to Riemannian manifolds, focusing on the hyperbolic model. By leveraging the exponential map, it follows that under certain conditions, deep neural networks defined on these manifolds exhibit convergence to a stable point.

This work successfully proves the convergence of hyperbolic neural networks under specific conditions using results from Avelin and Karlsson. The convergence is guaranteed for networks defined in the Lorentz hyperboloid model, which is significant for ensuring stability and predictability in these networks.

By applying concepts from dynamical systems and ergodic theory, the study enhances the understanding of the training stability and convergence patterns of HNNs. This approach helps in identifying stable solutions and avoiding unstable ones, leading to more reliable outcomes. Also this work suggests that understanding parameter trajectories can lead to new regularization methods, which can prevent overfitting and improve the generalization abilities of neural networks.

For future work, empirical validation of the theorems is necessary to confirm their practical applicability and effectiveness in real-world scenarios. Also, by using the exponential map and its inverse (when defined), it would be interesting to study neural networks in specific manifolds, e.g. the sphere, the torus, etc.

Acknowledgments and Disclosure of Funding

The first author acknowledge funding support from the the National Center of Artificial Intelligence, Chile (CENIA FB210017, Basal ANID) Fondecyt Regular grant number 1241772 entitled "Embed the city: An artificial intelligence-based approach to learn transferable representations from multi-source and multi-modal urban data."

References

[1] Nico Alvarado and Hans Lobel. Hyperbolic optimizer as a dynamical system. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 1243–1260. PMLR, 21–27 Jul 2024.

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

[3] A. Jentzen B. Fehrman, B. Gess. onvergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 2022.

[4] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic Graph Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[5] Chen and et. al. Fully hyperbolic neural networks. *Annual Meeting of the Association for Computational Linguistics*, 2022.

[6] M. I. Jordan B. Recht D. Lee, M. Simchowitz. Gradient descent only converges to minimizers. *Proceedings of the 29th Annual Conference on Learning Theory*, 2016.

[7] M. Do Carmo. *Riemannian Geometry*. Springer, 1992.

[8] Guilherme França, Daniel Robinson, and René Vidal. Admm and accelerated admm as continuous dynamical systems. 05 2018.

[9] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International Conference on Machine Learning*, pages 1646–1655. PMLR, 2018.

[10] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

[11] Gouëzel and Karlsson. Subadditive and multiplicative ergodic theorems. *J. Eur. Math. Soc.*, 2020.

[12] Bing-Sheng He, Min Tao, and Xiaoming Yuan. Alternating direction method with gaussian back substitution for separable convex programming. *SIAM Journal on Optimization*, 22, 05 2012.

[13] K.S. Narendra and K. Parthasarathy. Gradient methods for the optimization of dynamical systems containing neural networks. *IEEE Transactions on Neural Networks*, 2(2):252–262, 1991.

[14] Maximillian Nickel and Douwe Kiela. Poincaré Embeddings for Learning Hierarchical Representations. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[15] Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10023–10044, 2021.

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

A Proof of Theorem 3.4

Proof. We follow the line of arguments in [2]. Let X be the space of metrics on Ω that are bi-Lipschitz equivalent to d_0. Define the following Thompson metric for $d_1, d_2 \in X$:

$$D(d_1, d_2) = \log\left(\max\left\{\sup_{x\neq y}\frac{d_1(x,y)}{d_2(x,y)}, \sup_{x\neq y}\frac{d_2(x,y)}{d_1(x,y)}\right\}\right).$$

Let $\mathcal{F}\colon X \to Z := L^\infty(\Omega\times\Omega)\cap C(\Omega\times\Omega\setminus\{x=y\})$ be the map defined by

$$\mathcal{F}(d)(x,y) = \log\frac{d(x,y)}{d_0(x,y)},$$

and set $Y = \mathcal{F}(X)$. Observe that $\|\mathcal{F}(d_1)-\mathcal{F}(d_2)\|_Z = \sup_{x\neq y}\Big|\log\dfrac{d_1(x,y)}{d_2(x,y)}\Big| = D(d_1,d_2)$, so $\mathcal{F}$ is an isometry between X and Y (here $\|\cdot\|_Z$ is the L^∞ norm on the space Z of bounded and continuous functions on $\Omega\times\Omega\setminus\{x=y\}$).

Now given a non-expansive map $U\colon X\to X$, it induces a non-expansive map $\tilde{U}\colon Y\to Y$ given by $\tilde{U}(\mathcal{F}(d)) := \mathcal{F}\circ U(d)$. Indeed, observe that

$$\|\tilde{U}(\mathcal{F}(d)) - \tilde{U}(\mathcal{F}(d'))\|_Z = \|\mathcal{F}(U(d)) - \mathcal{F}(U(d'))\|_Z = D(U(d),U(d')) \leq D(d,d').$$

Consider as usual a stationary sequence of maps T_n, which induce non-expansive maps in the space (X,D) by $T^*d(x,y) := d(Tx,Ty)$. Set $f_n := \mathcal{F}((T_n\cdots T_1)^*d_0)$. We claim that $a(1,n) = \|f_n\|_Z$ is a subadditive cocycle. Indeed, since the T_n's are homeomorphisms, we have

$$\begin{aligned}
a(1,n+m) = \|f_{n+m}\|_Z &= \sup_{x\neq y}\Big|\log\frac{d_0(T_{n+m}\cdots T_1x, T_{n+m}\cdots T_1y)}{d_0(x,y)}\Big| \\
&= \sup_{x\neq y}\Big|\log\frac{d_0(T_{n+m}\cdots T_1x, T_{n+m}\cdots T_1y)}{d_0(T_m\cdots T_1x, T_m\cdots T_1y)}\cdot\frac{d_0(T_m\cdots T_1x, T_m\cdots T_1y)}{d_0(x,y)}\Big| \\
&\leq \sup_{x'\neq y'}\Big|\log\frac{d_0(T_{n+m}\cdots T_{m+1}x', T_{n+m}\cdots T_{m+1}y')}{d_0(x',y')}\Big| + \sup_{x\neq y}\Big|\log\frac{d_0(T_m\cdots T_1x, T_m\cdots T_1y)}{d_0(x,y)}\Big| \\
&= a(m,n) + a(1,m).
\end{aligned}$$

By the subadditive ergodic theorem, there is λ such that

$$\lambda = \lim_{m\to\infty}\frac{1}{m}\|f_n\|_Z = \lim_{m\to\infty}\frac{1}{m}\sup_{x\neq y}\Big|\log\frac{d_0(T_mT_{m-1}\cdots T_1x, T_mT_{m-1}\cdots y)}{d_0(x,y)}\Big|,$$

giving the desired result. □

B Useful results

For the sake of completion, we show that $\exp_y$ and $\log_y$ are inverses of each other.

Proposition B.1. *For every $x \in \mathbb{L}^n$, we have $\exp_y(\log_y(x)) = x$.*

Proof. Observe that $\|\log_y(x)\| = d(y,x)$ and recall that $y = (1, 0, \dots, 0)$. Since $\sinh(t) = \sqrt{\cosh^2(t) - 1}$, we obtain

$$\begin{aligned}
\exp_y(\log_y(x)) &= \cosh(\|\log_y(x)\|)y + \sinh(\|\log_y(x)\|)\frac{\log_y(x)}{\|\log_y(x)\|} \\
&= \cosh(\operatorname{arccosh}(x_0))y + \sinh(\operatorname{arccosh}(x_0))\frac{d(y,x)}{d(y,x)}\frac{x + g(y,x)y}{\|x + g(y,x)y\|} \\
&= x_0 y + \sqrt{x_0^2 - 1}\frac{(0, x_1, x_2, \dots, x_n)}{\sqrt{\sum_{i=1}^n x_i^2}} \\
&= (x_0, 0, \dots, 0) + (0, x_1, \dots, x_n) \\
&= x.
\end{aligned}$$

□

Proposition B.2. *For every $v \in T_y\mathbb{L}^n$, we have $\log_y(\exp_y(v)) = v$.*

Proof. We can identify $T_y\mathbb{L}^n = \{v \in \mathbb{R}^{n+1} : v_0 = 0\}$. By definition, we have

$$\log_y(\exp_y(v)) = d(y, \exp_y(v))\frac{\exp_y(v) + g(y, \exp_y(v))y}{\|\exp_y(v) + g(y, \exp_y(v))y\|}.$$

Since

$$\begin{aligned}
\exp_y(v) &= \cosh(\|v\|)y + \sinh(\|v\|)\frac{v}{\|v\|} \\
&= (\cosh\|v\|, 0, \dots, 0) + \frac{\sinh\|v\|}{\|v\|}(0, v_1, v_2, \dots, v_n) \\
&= \left(\cosh\|v\|, \frac{\sinh\|v\|}{\|v\|}v_1, \dots, \frac{\sinh\|v\|}{\|v\|}v_n\right),
\end{aligned}$$

we have

$$\exp_y(v) + g(y, \exp_y(v))y = \left(0, \frac{\sinh\|v\|}{\|v\|}v_1, \dots, \frac{\sinh\|v\|}{\|v\|}v_n\right) = \frac{\sinh\|v\|}{\|v\|}(0, v_1, \dots v_n) = \frac{\sinh\|v\|}{\|v\|}v.$$

On the other hand, we have

$$d(y, \exp_y(v)) = \operatorname{arccosh}(\cosh\|v\|) = \|v\|.$$

Putting everything together, we obtain

$$\log_y(\exp_y(v)) = \|v\| \cdot \frac{\frac{\sinh\|v\|}{\|v\|}v}{\left\|\frac{\sinh\|v\|}{\|v\|}v\right\|} = v.$$

□