RIEMANNIAN EMBEDDED TRANSFORMERS

Hilal Kılıç
Department of Physics
Middle East Technical University
Ankara, Türkiye
hilal.kilic@metu.edu.tr

February 8, 2026

ABSTRACT

The capacity of Euclidean Neural Networks to represent complex hierarchical data is fundamentally limited by the polynomial growth of Euclidean space. In this work, we present a robust implementation of a Transformer architecture grounded in the Poincaré Ball model of Hyperbolic geometry. By defining the manifold's structural integrity through its metric tensor $g = \lambda^2\delta$, we derive and implement a suite of Riemannian operations, including the Möbius addition, exponential mapping, and geodesic distance functions. Unlike standard architectures that collapse hierarchical relationships into flat embeddings, our Riemannian Embedding layer utilizes the exponential map to project tangent vectors into a curved manifold where space grows exponentially with the radius. We provide a rigorous translation of the geodesic equations into computationally efficient PyTorch operations, ensuring numerical stability through precision clamping and conformal scaling. This framework lays the foundation for Hyperbolic Attention mechanisms, offering a geometrically principled approach to modeling the latent hierarchies inherent in natural language and relational datasets.

1 Introduction

In recent years, Transformer architectures have revolutionized the field of Deep Learning, primarily due to their ability to capture long-range dependencies through self-attention mechanisms. However, standard Transformers operate within a Euclidean framework, which assumes a flat geometry. This assumption is fundamentally suboptimal for representing data with latent hierarchical structures, such as biological taxonomies, citation networks, or organizational trees.

The primary challenge in Euclidean embeddings is the "dimensionality bottleneck": the number of nodes in a hierarchy grows exponentially with depth, while the volume of Euclidean space grows only polynomially. This discrepancy leads to significant distortion and a high requirement for embedding dimensions to maintain structural integrity.

To address this, we propose the **Poincaré Transformer**, a manifold-aware architecture that integrates the representational power of Transformers with the inductive bias of Hyperbolic geometry. By mapping the self-attention mechanism and linear transformations into the Poincaré Ball ($\mathbb{B}^n$), we leverage a space whose volume grows exponentially with its radius—perfectly mirroring the growth of hierarchical data.

Our contributions are summarized as follows:

- **Riemannian Manifold Integration**: We define a Transformer block where Query, Key, and Value projections are computed within the tangent space of the Poincaré manifold, ensuring the transformations respect the underlying curvature.
- **Hyperbolic Attention**: We introduce a self-attention mechanism that replaces Euclidean dot-products with geodesic distances, allowing the model to naturally prioritize hierarchical proximity.
- **Geometric Robustness**: We implement an optimized training pipeline utilizing Riemannian Stochastic Gradient Descent (RSGD) with a retraction-based projection method to ensure numerical stability at the manifold's boundary.

2 Mathematical Framework for the Poincaré Manifold

The Poincaré Ball model $(\mathbb{B}^n, g_x)$ is defined on the open unit ball $\mathbb{B}^n = \{x \in \mathbb{R}^n : \sqrt{c}\|x\| < 1\}$ with constant negative curvature $-c$ ($c > 0$).

2.1 The Riemannian Metric and Conformal Factor

The metric tensor g_x is a conformal scaling of the Euclidean metric δ. The local geometry is governed by the conformal factor λ_x, which accounts for the exponential expansion of the manifold's volume toward the boundary:

$$g_x = \lambda_x^2 \delta, \quad \text{where} \quad \lambda_x = \frac{2}{1 - c\|x\|^2} \tag{1}$$

This factor is utilized in the optimization stage to rescale Euclidean gradients into their Riemannian counterparts.

2.2 Möbius Addition

To perform translations within the manifold while maintaining the constraint $x \oplus_c y \in \mathbb{B}^n$, we utilize the Möbius addition operation:

$$x \oplus_c y = \frac{(1 + 2c\langle x, y\rangle + c\|y\|^2)x + (1 - c\|x\|^2)y}{1 + 2c\langle x, y\rangle + c^2\|x\|^2\|y\|^2} \tag{2}$$

2.3 Exponential and Logarithmic Maps

The **Exponential Map** $\exp_p(v)$ projects a tangent vector v from the tangent space $T_p\mathbb{B}^n$ onto the manifold:

$$\exp_p(v) = p \oplus_c \left(\tanh\left(\frac{\sqrt{c}\lambda_p\|v\|}{2}\right)\frac{v}{\sqrt{c}\|v\|}\right) \tag{3}$$

The **Logarithmic Map** $\log_p(x)$ maps points from the manifold back to the tangent space, facilitating the linear transformations required for multi-head attention projections:

$$\log_p(x) = \frac{2}{\sqrt{c}\lambda_p}\text{arctanh}(\sqrt{c}\| - p \oplus_c x\|)\frac{-p \oplus_c x}{\| - p \oplus_c x\|} \tag{4}$$

2.4 Geodesic Distance and Attention

The attention mechanism is driven by the geodesic distance $d_c(x, y)$, which represents the shortest curved path between two points in hyperbolic space:

$$d_c(x, y) = \frac{1}{\sqrt{c}}\text{acosh}\left(1 + 2c\frac{\|x - y\|^2}{(1 - c\|x\|^2)(1 - c\|y\|^2)}\right) \tag{5}$$

Attention scores are computed as a function of the negative squared distance to prioritize local hierarchical dependencies:

$$\text{Attention}(Q, K) = \text{softmax}\left(\frac{-d_c(Q, K)^2}{\sqrt{d_k}}\right) \tag{6}$$

2.5 Riemannian Stochastic Gradient Descent (RSGD)

Optimization is performed directly on the manifold. The update rule for parameters θ involves rescaling the Euclidean gradient ∇_E by the squared inverse of the conformal factor:

$$\theta_{t+1} = \exp_{\theta_t}\left(-\eta\frac{(1 - c\|\theta_t\|^2)^2}{4}\nabla_E\mathcal{L}(\theta_t)\right) \tag{7}$$

where η is the learning rate.