Hyperbolic Entailment Cones for Learning Hierarchical Embeddings

Octavian-Eugen Ganea [1] **Gary Bécigneul** [1] **Thomas Hofmann** [1]

Abstract

Learning graph representations via low-dimensional embeddings that preserve relevant network properties is an important class of problems in machine learning. We here present a novel method to embed directed acyclic graphs. Following prior work, we first advocate for using hyperbolic spaces which provably model tree-like structures better than Euclidean geometry. Second, we view hierarchical relations as partial orders defined using a family of nested geodesically convex cones. We prove that these entailment cones admit an optimal shape with a closed form expression both in the Euclidean and hyperbolic spaces, and they canonically define the embedding learning process. Experiments show significant improvements of our method over strong recent baselines both in terms of representational capacity and generalization.

1. Introduction

Producing high quality feature representations of data such as text or images is a central point of interest in artificial intelligence. A large line of research focuses on embedding discrete data such as graphs (Grover & Leskovec, 2016; Goyal & Ferrara, 2017) or linguistic instances (Mikolov et al., 2013; Pennington et al., 2014; Kiros et al., 2015) into continuous spaces that exhibit certain desirable geometric properties. This class of models has reached state-of-the-art results for various tasks and applications, such as link prediction in knowledge bases (Nickel et al., 2011; Bordes et al., 2013) or in social networks (Hoff et al., 2002), text disambiguation (Ganea & Hofmann, 2017), word hypernymy (Shwartz et al., 2016), textual entailment (Rocktäschel et al., 2015) or taxonomy induction (Fu et al., 2014).

[1]Department of Computer Science, ETH Zurich, Switzerland. Correspondence to: Octavian-Eugen Ganea <octavian.ganea@inf.ethz.ch>, Gary Bécigneul <gary.becigneul@inf.ethz.ch>, Thomas Hofmann <thomas.hofmann@inf.ethz.ch>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Popular methods typically embed symbolic objects in low dimensional Euclidean vector spaces using a strategy that aims to capture semantic information such as functional similarity. Symmetric distance functions are usually minimized between representations of correlated items during the learning process. Popular examples are word embedding algorithms trained on corpora co-occurrence statistics which have shown to strongly relate semantically close words and their topics (Mikolov et al., 2013; Pennington et al., 2014).

However, in many fields (e.g. Recommender Systems, Genomics (Billera et al., 2001), Social Networks), one has to deal with data whose latent anatomy is best defined by non-Euclidean spaces such as Riemannian manifolds (Bronstein et al., 2017). Here, the Euclidean symmetric models suffer from not properly reflecting complex data patterns such as the latent hierarchical structure inherent in taxonomic data. To address this issue, the emerging trend of geometric deep learning[1] is concerned with non-Euclidean manifold representation learning.

In this work, we are interested in geometrical modeling of hierarchical structures, directed acyclic graphs (DAGs) and entailment relations via low dimensional embeddings. Starting from the same motivation, the order embeddings method (Vendrov et al., 2015) explicitly models the partial order induced by entailment relations between embedded objects. Formally, a vector $x \in \mathbb{R}^n$ represents a more general concept than any other embedding from the Euclidean entailment region $\mathcal{O}_x := \{y \mid y_i \geq x_i, \forall 1 \leq i \leq n\}$. A first concern is that the capacity of order embeddings grows only linearly with the embedding space dimension. Moreover, the regions $\mathcal{O}_x$ suffer from heavy intersections, implying that their disjoint volumes rapidly become bounded[2]. As a consequence, representing wide (with high branching factor) and deep hierarchical structures in a bounded region of the Euclidean space would cause many points to end up undesirably close to each other. This also implies that Euclidean distances would no longer be capable of reflecting the original tree metric.

Fortunately, the hyperbolic space does not suffer from the aforementioned capacity problem because the volume of

[1]http://geometricdeeplearning.com/

[2]For example, in n dimensions, no $n+1$ distinct regions $\mathcal{O}_x$ can simultaneously have unbounded disjoint sub-volumes.

any ball grows exponentially with its radius, instead of polynomially as in the Euclidean space. This exponential growth property enables hyperbolic spaces to embed any weighted tree while almost preserving their metric[3] (Gromov, 1987; Bowditch, 2006; Sarkar, 2011). The tree-likeness of hyperbolic spaces has been extensively studied (Hamann, 2017). Moreover, hyperbolic spaces are used to visualize large hierarchies (Lamping et al., 1995), to efficiently forward information in complex networks (Krioukov et al., 2009; Cvetkovski & Crovella, 2009) or to embed heterogeneous, scale-free graphs (Shavitt & Tankel, 2008; Krioukov et al., 2010; Bläsius et al., 2016).

From a machine learning perspective, recently, hyperbolic spaces have been observed to provide powerful representations of entailment relations (Nickel & Kiela, 2017). The latent hierarchical structure surprisingly emerges as a simple reflection of the space's negative curvature. However, the approach of (Nickel & Kiela, 2017) suffers from a few drawbacks: first, their loss function causes most points to collapse on the border of the Poincaré ball, as exemplified in Figure 3. Second, the hyperbolic distance alone (being symmetric) is not capable of encoding asymmetric relations needed for entailment detection, thus a heuristic score is chosen to account for concept generality or specificity encoded in the embedding norm.

We here inspire ourselves from hyperbolic embeddings (Nickel & Kiela, 2017) and order embeddings (Vendrov et al., 2015). Our contributions are as follows:

- We address the aforementioned issues of (Nickel & Kiela, 2017) and (Vendrov et al., 2015). We propose to replace the entailment regions $\mathcal{O}_x$ of order-embeddings by a more efficient and generic class of objects, namely *geodesically convex entailment cones*. These cones are defined on a large class of Riemannian manifolds and induce a partial ordering relation in the embedding space.
- The optimal entailment cones satisfying four natural properties surprisingly exhibit canonical closed-form expressions in both Euclidean and hyperbolic geometry that we rigorously derive.
- An efficient algorithm for learning hierarchical embeddings of directed acyclic graphs is presented. This learning process is driven by our entailment cones.
- Experimentally, we learn high quality embeddings and improve over experimental results in (Nickel & Kiela, 2017) and (Vendrov et al., 2015) on hypernymy link prediction for word embeddings, both in terms of capacity of the model and generalization performance.

[3] See end of Section 2.2 for a rigorous formulation.

We also compute an analytic closed-form expression for the exponential map in the n-dimensional Poincaré ball, allowing us to perform full Riemannian optimization (Bonnabel, 2013) in the Poincaré ball, as opposed to the approximate optimization method used by (Nickel & Kiela, 2017).

2. Mathematical preliminaries

We now briefly visit some key concepts needed in our work.

Notations. We always use $\|\cdot\|$ to denote the Euclidean norm of a point (in both hyperbolic or Euclidean spaces). We also use $\langle\cdot,\cdot\rangle$ to denote the Euclidean scalar product.

2.1. Differential geometry

For a rigorous reasoning about hyperbolic spaces, one needs to use concepts in differential geometry, some of which we highlight here. For an in-depth introduction, we refer the reader to (Spivak, 1979) and (Hopper & Andrews, 2010).

Manifold. A *manifold* $\mathcal{M}$ of dimension n is a set that can be locally approximated by the Euclidean space $\mathbb{R}^n$. For instance, the sphere $\mathbb{S}^2$ and the torus $\mathbb{T}^2$ embedded in $\mathbb{R}^3$ are 2-dimensional manifolds, also called surfaces, as they can locally be approximated by $\mathbb{R}^2$. The notion of manifold is a generalization of the notion of surface.

Tangent space. For $x \in \mathcal{M}$, the *tangent space* $T_x\mathcal{M}$ of $\mathcal{M}$ at x is defined as the n-dimensional vector-space approximating $\mathcal{M}$ around x at a first order. It can be defined as the set of vectors v that can be obtained as $v := c'(0)$, where $c : (-\varepsilon, \varepsilon) \to \mathcal{M}$ is a smooth path in $\mathcal{M}$ such that $c(0) = x$.

Riemannian metric. A *Riemannian metric* g on $\mathcal{M}$ is a collection $(g_x)_x$ of inner-products $g_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ on each tangent space $T_x\mathcal{M}$, depending smoothly on x. Although it defines the geometry of $\mathcal{M}$ locally, it induces a global distance function $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$ by setting $d(x, y)$ to be the infimum of all lengths of smooth curves joining x to y in $\mathcal{M}$, where the length ℓ of a curve $\gamma : [0, 1] \to \mathcal{M}$ is defined as

$$\ell(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}(\gamma'(t), \gamma'(t))}dt. \tag{1}$$

Riemannian manifold. A smooth manifold equipped with a Riemannian metric is called a *Riemannian manifold*. Subsequently, due to their metric properties, we will only consider such manifolds.

Geodesics. A *geodesic* (straight line) between two points $x, y \in \mathcal{M}$ is a smooth curve of minimal length joining x to

y in $\mathcal{M}$. Geodesics define shortest paths on the manifold. They are a generalization of lines in the Euclidean space.

Exponential map. The *exponential map* $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ around x, when well-defined, maps a small perturbation of x by a vector $v \in T_x\mathcal{M}$ to a point $\exp_x(v) \in \mathcal{M}$, such that $t \in [0,1] \mapsto \exp_x(tv)$ is a geodesic joining x to $\exp_x(v)$. In Euclidean space, we simply have $\exp_x(v) = x + v$. The exponential map is important, for instance, when performing gradient-descent over parameters lying in a manifold (Bonnabel, 2013).

Conformality. A metric $\tilde{g}$ on $\mathcal{M}$ is said to be *conformal* to g if it defines the same angles, *i.e.* for all $x \in \mathcal{M}$ and $u, v \in T_x\mathcal{M} \setminus \{0\}$,

$$\frac{\tilde{g}_x(u,v)}{\sqrt{\tilde{g}_x(u,u)}\sqrt{\tilde{g}_x(v,v)}} = \frac{g_x(u,v)}{\sqrt{g_x(u,u)}\sqrt{g_x(v,v)}}. \tag{2}$$

This is equivalent to the existence of a smooth function $\lambda : \mathcal{M} \to (0,\infty)$ such that $\tilde{g}_x = \lambda_x^2 g_x$, which is called the *conformal factor* of $\tilde{g}$ (w.r.t. g).

2.2. Hyperbolic geometry

The hyperbolic space of dimension $n \geq 2$ is a fundamental object in Riemannian geometry. It is (up to isometry) uniquely characterized as a complete, simply connected Riemannian manifold with constant negative curvature (Cannon et al., 1997). The other two model spaces of constant sectional curvature are the flat Euclidean space $\mathbb{R}^n$ (zero curvature) and the hyper-sphere $\mathbb{S}^n$ (positive curvature).

The hyperbolic space has five models which are often insightful to work in. They are isometric to each other and conformal to the Euclidean space (Cannon et al., 1997; Parkkonen, 2013)[4]. We prefer to work in the Poincaré ball model $\mathbb{D}^n$ for the same reasons as (Nickel & Kiela, 2017) and, additionally, because we can derive a closed form expression of geodesics and exponential map.

Poincaré metric tensor. The Poincaré ball model $(\mathbb{D}^n, g^{\mathbb{D}})$ is defined by the manifold $\mathbb{D}^n = \{x \in \mathbb{R}^n : \|x\| < 1\}$ equipped with the following Riemannian metric

$$g_x^{\mathbb{D}} = \lambda_x^2 g^E, \quad \text{where } \lambda_x := \frac{2}{1 - \|x\|^2}, \tag{3}$$

and g^E is the Euclidean metric tensor with components $\mathbf{I}_n$ of the standard space $\mathbb{R}^n$ with the usual Cartesian coordinates.

As the above model is a Riemannian manifold, its metric tensor is fundamental in order to uniquely define most of its geometric properties like distances, inner products (in tangent spaces), straight lines (geodesics), curve lengths or volume elements. In the Poincaré ball model, the Euclidean metric is changed by a simple scalar field, hence the model is *conformal* (*i.e.* angle preserving), yet distorts distances.

Induced distance and norm. It is known (Nickel & Kiela, 2017) that the induced distance between 2 points $x, y \in \mathbb{D}^n$ is given by

$$d_{\mathbb{D}}(x,y) = \cosh^{-1}\left(1 + 2\frac{\|x-y\|^2}{(1-\|x\|^2)\cdot(1-\|y\|^2)}\right). \tag{4}$$

The Poincare norm is then defined as:

$$\|x\|_{\mathbb{D}} := d_{\mathbb{D}}(0,x) = 2\tanh^{-1}(\|x\|) \tag{5}$$

Geodesics and exponential map. We derive parametric expressions of unit-speed geodesics and exponential map in the Poincaré ball. Geodesics in $\mathbb{D}^n$ are all intersections of the Euclidean unit ball $\mathbb{D}^n$ with (degenerated) Euclidean circles orthogonal to the unit sphere $\partial\mathbb{D}^n$ (equations are derived below). We know from the Hopf-Rinow theorem that the hyperbolic space is complete as a metric space. This guarantees that $\mathbb{D}^n$ is geodesically complete. Thus, the exponential map is defined for each point $x \in \mathbb{D}^n$ and any $v \in \mathbb{R}^n (= T_x\mathbb{D}^n)$. To derive its closed form expression, we first prove the following.

Theorem 1. *(Unit-speed geodesics) Let $x \in \mathbb{D}^n$ and $v \in T_x\mathbb{D}^n (= \mathbb{R}^n)$ such that $g_x^{\mathbb{D}}(v,v) = 1$. The unit-speed geodesic $\gamma_{x,v} : \mathbb{R}_+ \to \mathbb{D}^n$ with $\gamma_{x,v}(0) = x$ and $\dot{\gamma}_{x,v}(0) = v$ is given by*

$$\gamma_{x,v}(t) = \frac{\left(\lambda_x \cosh(t) + \lambda_x^2\langle x, v\rangle \sinh(t)\right)x + \lambda_x \sinh(t) v}{1 + (\lambda_x - 1)\cosh(t) + \lambda_x^2\langle x, v\rangle \sinh(t)} \tag{6}$$

Proof. See appendix B. □

Corollary 1.1. *(Exponential map) The exponential map at a point $x \in \mathbb{D}^n$, namely $\exp_x : T_x\mathbb{D}^n \to \mathbb{D}^n$, is given by*

$$\exp_x(v) = \frac{\lambda_x\left(\cosh(\lambda_x\|v\|) + \langle x, \frac{v}{\|v\|}\rangle \sinh(\lambda_x\|v\|)\right)}{1 + (\lambda_x - 1)\cosh(\lambda_x\|v\|) + \lambda_x\langle x, \frac{v}{\|v\|}\rangle \sinh(\lambda_x\|v\|)} x + \frac{\frac{1}{\|v\|}\sinh(\lambda_x\|v\|)}{1 + (\lambda_x - 1)\cosh(\lambda_x\|v\|) + \lambda_x\langle x, \frac{v}{\|v\|}\rangle \sinh(\lambda_x\|v\|)} v \tag{7}$$

Proof. See appendix C. □

We also derive the following fact (useful for future proofs).

[4] https://en.wikipedia.org/wiki/Hyperbolic_space

Corollary 1.2. *Given any arbitrary geodesic in $\mathbb{D}^n$, all its points are coplanar with the origin O.*

Proof. See appendix D. □

Angles in hyperbolic space. It is natural to extend the Euclidean notion of an angle to any geodesically complete Riemannian manifold. For any points A, B, C on such a manifold, the angle $\angle ABC$ is the angle between the initial tangent vectors of the geodesics connecting B with A, and B with C, respectively. In the Poincaré ball, the angle between two tangent vectors $u, v \in T_x\mathbb{D}^n$ is given by

$$\cos(\angle(u,v)) = \frac{g_x^{\mathbb{D}}(u,v)}{\sqrt{g_x^{\mathbb{D}}(u,u)}\sqrt{g_x^{\mathbb{D}}(v,v)}} = \frac{\langle u, v\rangle}{\|u\|\|v\|} \tag{8}$$

The second equality happens since $g^{\mathbb{D}}$ is conformal to g^E.

Hyperbolic trigonometry. The notion of angles and geodesics allow definition of the notion of a triangle in the Poincaré ball. Then, the classic theorems in Euclidean geometry have hyperbolic formulations (Parkkonen, 2013). In the next section, we will use the following theorems.

Let $A, B, C \in \mathbb{D}^n$. Denote by $\angle B := \angle ABC$ and by $c = d_{\mathbb{D}}(B, A)$ the length of the hyperbolic segment BA (and others). Then, the hyperbolic laws of cosines and sines hold respectively

$$\cos(\angle B) = \frac{\cosh(a)\cosh(c) - \cosh(b)}{\sinh(a)\sinh(c)} \tag{9}$$

$$\frac{\sin(\angle A)}{\sinh(a)} = \frac{\sin(\angle B)}{\sinh(b)} = \frac{\sin(\angle C)}{\sinh(c)} \tag{10}$$

Embedding trees in hyperbolic vs Euclidean space. Finally, we briefly explain why hyperbolic spaces are better suited than Euclidean spaces for embedding trees. However, note that our method is applicable to any DAG.

(Gromov, 1987) introduces a notion of δ-hyperbolicity in order to characterize how 'hyperbolic' a metric space is. For instance, the Euclidean space $\mathbb{R}^n$ for $n \geq 2$ is not δ-hyperbolic for any $\delta \geq 0$, while the Poincaré ball $\mathbb{D}^n$ is $\log(1+\sqrt{2})$-hyperbolic. This is formalized in the following theorem[5] (section 6.2 of (Gromov, 1987), proposition 6.7 of (Bowditch, 2006)):

Theorem: For any $\delta > 0$, any δ-hyperbolic metric space (X, d_X) and any set of points $x_1, ..., x_n \in X$, there exists a finite weighted tree (T, d_T) and an embedding $f : T \to X$ such that for all i, j,

$$|d_T(f^{-1}(x_i), f^{-1}(x_j)) - d_X(x_i, x_j)| = \mathcal{O}(\delta \log(n)). \tag{11}$$

[5] https://en.wikipedia.org/wiki/Hyperbolic_metric_space

Conversely, any tree can be embedded with arbitrary low distortion into the Poincaré disk (with only 2 dimensions), whereas this is not true for Euclidean spaces even when an unbounded number of dimensions is allowed (Sarkar, 2011; De Sa et al., 2018).

The difficulty in embedding trees having a branching factor at least 2 in a quasi-isometric manner comes from the fact that they have an exponentially increasing number of nodes with depth. The exponential volume growth of hyperbolic metric spaces confers them enough capacity to embed trees quasi-isometrically, unlike the Euclidean space.

3. Entailment Cones in the Poincaré Ball

In this section, we define "entailment" cones that will be used to embed hierarchical structures in the Poincaré ball. They generalize and improve over the idea of order embeddings (Vendrov et al., 2015).

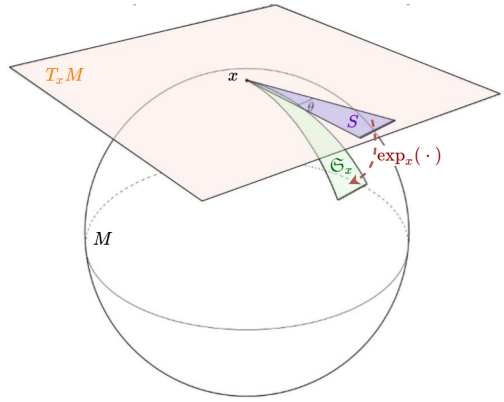

Figure 1. Convex cones in a complete Riemannian manifold.

Convex cones in a complete Riemannian manifold. We are interested in generalizing the notion of a convex cone to any geodesically complete Riemannian manifold $\mathcal{M}$ (such as hyperbolic models). In a vector space, a convex cone S (at the origin) is a set that is closed under non-negative linear combinations

$$v_1, v_2 \in S \implies \alpha v_1 + \beta v_2 \in S \quad (\forall \alpha, \beta \geq 0)\,. \tag{12}$$

The key idea for generalizing this concept is to make use of the exponential map at a point $x \in \mathcal{M}$.

$$\exp_x : T_x\mathcal{M} \to \mathcal{M}, \quad T_x\mathcal{M} = \text{ tangent space at } x \tag{13}$$

We can now take any cone in the tangent space $S \subseteq T_x\mathcal{M}$ at a fixed point x and map it into a set $\mathfrak{S}_x \subset \mathcal{M}$, which we call the S-cone at x, via

$$\mathfrak{S}_x := \exp_x(S)\,, \quad S \subseteq T_x\mathcal{M}\,. \tag{14}$$

Note that, in the above definition, we desire that the exponential map be injective. We already know that it is a local diffeomorphism. Thus, we restrict the tangent space in Eq. 14 to the ball $\mathcal{B}^n(O, r)$, where r is the injectivity radius of M at x. Note that for hyperbolic space models the injectivity radius of the tangent space at any point is infinite, thus no restriction is needed.

Angular cones in the Poincaré ball. We are interested in special types of cones in $\mathbb{D}^n$ that can extend in all space directions. We want to avoid heavy cone intersections and to have capacity that scales exponentially with the space dimension. To achieve this, we want the definition of cones to exhibit the following four intuitive properties detailed below. Subsequently, solely based on these necessary conditions, we formally prove that the optimal cones in the Poincaré ball have a closed form expression.

1) Axial symmetry. For any $x \in \mathbb{D}^n \setminus \{0\}$, we require circular symmetry with respect to a central axis of the cone $\mathfrak{S}_x$. We define this axis to be the spoke through x from x:

$$A_x := \{x' \in \mathbb{D}^n : x' = \alpha x,\ \frac{1}{\|x\|} > \alpha \geq 1\} \tag{15}$$

Then, we fix any tangent vector with the same direction as x, e.g. $\bar{x} = \exp_x^{-1}\left(\frac{1+\|x\|}{2\|x\|}x\right) \in T_x\mathbb{D}^n$. One can verify using Corollary 1.1 that $\bar{x}$ generates the axis-oriented geodesic as:

$$A_x = \exp_x\left(\{y \in \mathbb{R}^n : y = \alpha\bar{x},\ \alpha > 0\}\right). \tag{16}$$

We next define the angle $\angle(v, \bar{x})$ for any tangent vector $v \in T_x\mathbb{D}^n$ as in Eq. 8. Then, the axial symmetry property is satisfied if we define the angular cone at x to have a non-negative aperture $2\psi(x) \geq 0$ as follows:

$$\begin{aligned} S_x^{\psi(x)} &:= \{v \in T_x\mathbb{D}^n : \angle(v, \bar{x}) \leq \psi(x)\} \\ \mathfrak{S}_x^{\psi(x)} &:= \exp_x(S_x^{\psi(x)}). \end{aligned} \tag{17}$$

We further define the conic border (face):

$$\partial S^{\psi} := \{v : \angle(v, \bar{x}) = \psi(x)\}, \quad \partial\mathfrak{S}_x^{\psi} := \exp_x(\partial S_x^{\psi}). \tag{18}$$

2) Rotation invariance. We want the definition of cones $\mathfrak{S}_x^{\psi(x)}$ to be independent of the angular coordinate of the apex x, *i.e.* to only depend on the (Euclidean) norm of x:

$$\psi(x) = \psi(x') \quad (\forall x, x' \in \mathbb{D}^n \setminus \{0\},\ \text{s.t. } \|x\| = \|x'\|). \tag{19}$$

This implies that there exists $\tilde{\psi} : (0, 1) \to [0, \pi)$ s. t. for all $x \in \mathbb{D}^n \setminus \{0\}$ we have $\psi(x) = \tilde{\psi}(\|x\|)$.

3) Continuous cone aperture functions. We require the aperture ψ of our cones to be a continuous function. Using Eq. 19, it is equivalent to the continuity of $\tilde{\psi}$. This requirement seems reasonable and will be helpful in order to prove uniqueness of the optimal entailment cones. When optimization-based training is employed, it is also necessary that this function be differentiable. Surprisingly, we will show below that the optimal functions ψ are actually smooth, even when only requiring continuity.

4) Transitivity of nested angular cones. We want cones to determine a partial order in the embedding space. The difficult property is transitivity. We are interested in defining a cone width function $\psi(x)$ such that the resulting angular cones satisfy the *transitivity* property of partial order relations, *i.e.* they form a nested structure as follows

$$\forall x, x' \in \mathbb{D}^n \setminus \{0\} : \quad x' \in \mathfrak{S}_x^{\psi(x)} \implies \mathfrak{S}_{x'}^{\psi(x')} \subseteq \mathfrak{S}_x^{\psi(x)}. \tag{20}$$

Closed form expression of the optimal ψ. We now analyze the implications of the above necessary properties. Surprisingly, the optimal form of the function ψ admits an interesting closed-form expression. We will see below that mathematically ψ cannot be defined on the entire open ball $\mathbb{D}^n$. Towards these goals, we first prove the following.

Lemma 2. *If transitivity holds, then*

$$\forall x \in \mathit{Dom}(\psi) : \quad \psi(x) \leq \frac{\pi}{2}. \tag{21}$$

Proof. See appendix E. □

Note that so far we removed the origin 0 of $\mathbb{D}^n$ from our definitions. However, the above surprising lemma implies that we cannot define a useful cone at the origin. To see this, we first note that the origin should "entail" the entire space $\mathbb{D}^n$, i.e. $\mathfrak{S}_0 = \mathbb{D}^n$. Second, similar with property 3, we desire the cone at 0 be a continuous deformation of the cones of any sequence of points $(x_n)_{n\geq 0}$ in $\mathbb{D}^n \setminus \{0\}$ that converges to 0. Formally, $\lim_{n\to\infty} \mathfrak{S}_{x_n} = \mathfrak{S}_0$ when $\lim_{n\to\infty} x_n = 0$. However, this is impossible because Lemma 2 implies that the cone at each point x_n can only cover at most half of $\mathbb{D}^n$. We further prove the following:

Theorem 3. *If transitivity holds, then the function*

$$h : (0, 1) \cap \mathit{Dom}(\tilde{\psi}) \to \mathbb{R}_+, \quad h(r) := \frac{r}{1 - r^2}\sin(\tilde{\psi}(r)), \tag{22}$$

is non-increasing.

Proof. See appendix F. □

The above theorem implies that a non-zero $\tilde{\psi}$ cannot be defined on the entire $(0, 1)$ because $\lim_{r\to 0} h(r) = 0$, for

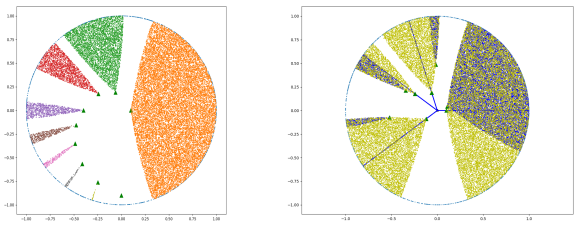

Figure 2. Poincaré angular cones satisfying Eq. 26 for K = 0.1. Left: examples of cones for points with Euclidean norm varying from 0.1 to 0.9. Right: transitivity for various points on the border of their parent cones.

any function $\tilde{\psi}$. As a consequence, we are forced to restrict $\text{Dom}(\tilde{\psi})$ to some $[\epsilon, 1)$, *i.e.* to leave the open ball $\mathcal{B}^n(O, \epsilon)$ outside of the domain of ψ. Then, theorem 3 implies that

$$\forall r \in [\epsilon, 1): \quad \sin(\tilde{\psi}(r))\frac{r}{1-r^2} \leq \sin(\tilde{\psi}(\epsilon))\frac{\epsilon}{1-\epsilon^2}. \tag{23}$$

Since we are interested in cones with an aperture as large as possible (to maximize model capacity), it is natural to set all terms $h(r)$ equal to $K := h(\epsilon)$, *i.e.* to make h constant:

$$\forall r \in [\epsilon, 1): \quad \sin(\tilde{\psi}(r))\frac{r}{1-r^2} = K, \tag{24}$$

which gives both a restriction on ϵ (in terms of K):

$$K \leq \frac{\epsilon}{1-\epsilon^2} \quad \Longleftrightarrow \quad \epsilon \in \left[\frac{2K}{1+\sqrt{1+4K^2}}, 1\right), \tag{25}$$

as well as a closed form expression for ψ

$$\begin{aligned} \psi : \mathbb{D}^n \setminus \mathcal{B}^n(O, \epsilon) &\to (0, \pi/2) \\ x &\mapsto \arcsin(K(1-\|x\|^2)/\|x\|), \end{aligned} \tag{26}$$

which is also a sufficient condition for transitivity to hold:

Theorem 4. *If ψ is defined as in Eqs.25-26, then transitivity holds.*

The above theorem has a proof similar to that of Thm. 3.

So far, we have obtained a closed form expression for hyperbolic entailment cones. However, we still need to understand how they can be used during embedding learning. For this goal, we derive an equivalent (and more practical) definition of the cone $\mathfrak{S}_x^{\psi(x)}$:

Theorem 5. *For any $x, y \in \mathbb{D}^n \setminus \mathcal{B}^n(O, \epsilon)$, we denote the angle between the half-lines $(xy$ and $(0x$ as*

$$\Xi(x, y) := \pi - \angle Oxy, \tag{27}$$

Then, this angle equals

$$\arccos\left(\frac{\langle x, y\rangle(1+\|x\|^2) - \|x\|^2(1+\|y\|^2)}{\|x\| \cdot \|x-y\|\sqrt{1+\|x\|^2\|y\|^2 - 2\langle x, y\rangle}}\right), \tag{28}$$

Moreover, we have the following equivalent expression of the Poincaré entailment cones satisfying Eq. 26:

$$\mathfrak{S}_x^{\psi(x)} = \left\{ y \in \mathbb{D}^n \;\middle|\; \Xi(x, y) \leq \arcsin\left(K\frac{1-\|x\|^2}{\|x\|}\right) \right\}. \tag{29}$$

Proof. See appendix G. □

Examples of 2-dimensional Poincaré cones corresponding to apex points located at different radii from the origin are shown in Figure 2. This figure also shows that transitivity is satisfied for some points on the border of the hypercones.

Euclidean entailment cones. One can easily adapt the above proofs to derive entailment cones in the Euclidean space $(\mathbb{R}^n, g^E)$. The only adaptations are: i) replace the hyperbolic cosine law by usual Euclidean cosine law, ii) geodesics are straight lines, and iii) the exponential map is given by $\exp_x(v) = x + v$. Thus, one similarly obtains that $h(r) = r\sin(\psi(r))$ is non-decreasing, the optimal values of ψ are obtained for constant h being equal to $K \leq \varepsilon$ and

$$\mathfrak{S}_x^{\psi(x)} = \{y \in \mathbb{R}^n \mid \Xi(x, y) \leq \psi(x)\}, \tag{30}$$

where $\Xi(x, y)$ now becomes

$$\Xi(x, y) = \arccos\left(\frac{\|y\|^2 - \|x\|^2 - \|x-y\|^2}{2\|x\| \cdot \|x-y\|}\right), \tag{31}$$

for all $x, y \in \mathbb{R}^n \setminus \mathcal{B}(O, \varepsilon)$. From a learning perspective, there is no need to be concerned about the Riemannian optimization described in Section 4.2, as the usual Euclidean gradient-step is used in this case.

4. Learning with entailment cones

We now describe how embedding learning is performed.

4.1. Max-margin training on angles

We learn hierarchical word embeddings from a dataset $\mathcal{X}$ of entailment relations $(u, v) \in \mathcal{X}$, also called hypernym links, defining that u entails v, or, equivalently, that v is a subconcept of u[6].

We choose to model the embedding entailment relation (u, v) as v belonging to the entailment cone $\mathfrak{S}_u^{\psi(u)}$.

[6]We prefer this notation over the one in (Nickel & Kiela, 2017)

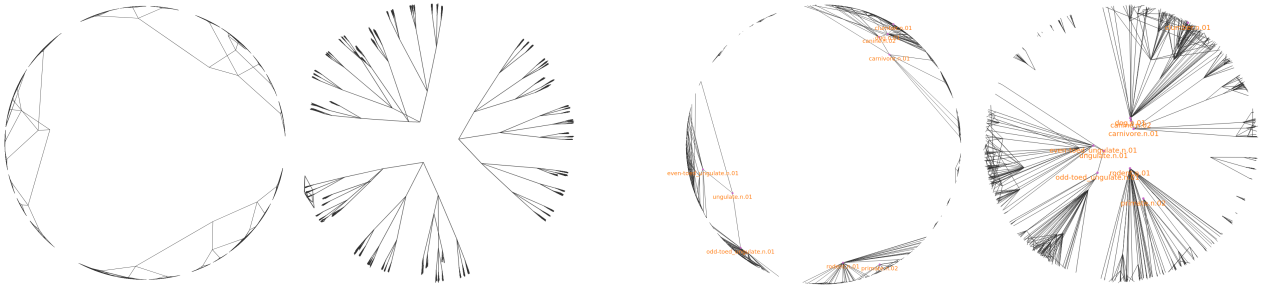

Figure 3. Two dimensional embeddings of two datasets: a toy uniform tree of depth 7 and branching factor 3, with root removed (left); the mammal subtree of WordNet with 4230 relations, 1165 nodes and top 2 nodes removed (right). (Nickel & Kiela, 2017) (each left side) has most of the nodes and edges collapsed on the space border, while our hyperbolic cones (each right side) nicely reveal the data structure.

Our model is trained with a max-margin loss function similar to the one in (Vendrov et al., 2015):

$$\mathcal{L} = \sum_{(u,v)\in P} E(u,v) + \sum_{(u',v')\in N} \max(0, \gamma - E(u',v')), \tag{32}$$

for some margin $\gamma > 0$, where P and N define samples of positive and negative edges respectively. The energy $E(u,v)$ measures the penalty of a wrongly classified pair (u,v), which in our case measures how far is point v from belonging to $\mathfrak{S}_u^{\psi(u)}$ expressed as the smallest angle of a rotation of center u bringing v into $\mathfrak{S}_u^{\psi(u)}$:

$$E(u,v) := \max(0, \Xi(u,v) - \psi(u)), \tag{33}$$

where $\Xi(u,v)$ is defined in Eqs. 28 and 31. Note that (Vendrov et al., 2015) use $\|\max(0, v-u)\|^2$. This loss function encourages positive samples to satisfy $E(u,v) = 0$ and negative ones to satisfy $E(u,v) \geq \gamma$. The same loss is used both in the hyperbolic and Euclidean cases.

4.2. Full Riemannian optimization

As the parameters of the model live in the hyperbolic space, the back-propagated gradient is a Riemannian gradient. Indeed, if u is in the Poincaré ball, and if we compute the usual (Euclidean) gradient $\nabla_u \mathcal{L}$ of our loss, then

$$u \leftarrow u - \eta \nabla_u \mathcal{L} \tag{34}$$

makes no sense as an operation in the Poincaré ball, since the substraction operation is not defined in this manifold. Instead, one should compute the Riemannian gradient $\nabla_u^R \mathcal{L}$ indicating a direction in the tangent space $T_u\mathbb{D}^n$, and should move u along the corresponding geodesic in $\mathbb{D}^n$ (Bonnabel, 2013):

$$u \leftarrow \exp_u(-\eta \nabla_u^R \mathcal{L}), \tag{35}$$

where the Riemannian gradient is obtained by rescaling the Euclidean gradient by the inverse of the metric tensor. As our metric is conformal, *i.e.* $g^{\mathbb{D}} = \lambda^2 g^E$ where $g^E = \mathbf{I}$ is the Euclidean metric (see Eq 3), this leads to a simple formulation

$$\nabla_u^R \mathcal{L} = (1/\lambda_u)^2 \nabla_u \mathcal{L}. \tag{36}$$

Previous work (Nickel & Kiela, 2017) optimizing word embeddings in the Poincaré ball used the retraction map $\mathcal{R}_x(v) := x + v$ as a first order approximation of $\exp_x(v)$. Note that since we derived a closed-form expression of the exponential map in the Poincaré ball (Corollary 1.1), we are able to perform full Riemannian optimization in this model of the hyperbolic space.

5. Experiments

We evaluate the representational and generalization power of hyperbolic entailment cones and of other baselines using data that exhibits a latent hierarchical structure. We follow previous work (Nickel & Kiela, 2017; Vendrov et al., 2015) and use the full transitive closure of the WordNet noun hierarchy (Miller et al., 1990). Our binary classification task is link prediction for unseen edges in this directed acyclic graph.

Dataset splitting. Train and evaluation settings. We remove the tree root since it carries little information and only has trivial edges to predict. Note that this implies that we co-embed the resulting subgraphs together to prevent overlapping embeddings (see smaller examples in Figure 3). The remaining WordNet dataset contains 82,114 nodes and 661,127 edges in the full transitive closure. We split it into train - validation - test sets as follows. We first compute the transitive reduction[7] of this directed acyclic graph, *i.e.* *"basic" edges* that form the minimal edge set for which the original transitive closure can be fully recovered. These edges are hard to predict, so we will always include them in the training set. The remaining *"non-basic" edges* (578,477)

[7] https://en.wikipedia.org/wiki/Transitive_reduction

	Embedding Dimension = 5				Embedding Dimension = 10			
	Percentage of Transitive Closure (Non-basic) Edges in Training							
	0%	10%	25%	50%	0%	10%	25%	50%
Simple Euclidean Emb	26.8%	71.3%	73.8%	72.8%	29.4%	75.4%	78.4%	78.1%
Poincaré Emb	29.4%	70.2%	78.2%	83.6%	28.9%	71.4%	82.0%	85.3%
Order Emb	**34.4**%	70.2%	75.9%	81.7%	**43.0**%	69.7%	79.4%	84.1%
Our Euclidean Cones	28.5%	69.7%	75.0%	77.4%	31.3%	81.5%	84.5%	81.6%
Our Hyperbolic Cones	29.2%	**80.1**%	**86.0**%	**92.8**%	32.2%	**85.9**%	**91.0**%	**94.4**%

Table 1. Test F1 results for various models. Simple Euclidean Emb and Poincaré Emb are the Euclidean and hyperbolic methods proposed by (Nickel & Kiela, 2017), Order Emb is proposed by (Vendrov et al., 2015).

are split into validation (5%), test (5%) and train (fraction of the rest).

We augment both the validation and the test parts with sets of negative pairs as follows: for each true (positive) edge (u, v), we randomly sample five (u', v) and five (u, v') negative corrupted pairs that are not edges in the full transitive closure. These are then added to the respective negative set. Thus, ten times as many negative pairs as positive pairs are used. They are used to compute standard classification metrics associated with these datasets: precision, recall, F1. For the training set, negative pairs are dynamically generated as explained below.

We make the task harder in order to understand the generalization ability of various models when differing amounts of transitive closure edges are available during training. We generate four training sets that include 0%, 10%, 25%, or 50% of the non-basic edges, selected randomly. We then train separate models using each of these four sets after being augmented with the basic edges.

Baselines. We compare against the strong hierarchical embedding methods of *Order embeddings* (Vendrov et al., 2015) and *Poincaré embeddings* (Nickel & Kiela, 2017). Additionally, we also use *Simple Euclidean embeddings*, *i.e.* the Euclidean version of the method presented in (Nickel & Kiela, 2017) (one of their baselines). We note that Poincaré and Simple Euclidean embeddings were trained using a symmetric distance function, and thus cannot be directly used to evaluate asymmetric entailment relations. Thus, for these baselines we use the heuristic scoring function proposed in (Nickel & Kiela, 2017):

$$score(u, v) = (1 + \alpha(\|u\| - \|v\|))d(u, v) \tag{37}$$

and tune the parameter α on the validation set. For all the other methods (our proposed cones and order embeddings), we use the energy penalty $E(u, v)$, e.g. Eq. 33 for hyperbolic cones. This scoring function is then used at test time for binary classification as follows: if it is lower than a threshold, we predict an edge; otherwise, we predict a non-edge. The optimal threshold is chosen to achieve maximum F1 on the validation set by passing over the sorted array of scores of positive and negative validation pairs.

Training details. We provide training details in Sec. H.

Results and discussion. Table 1 shows the obtained results. For a fair comparison, we use models with the same number of dimensions. We focus on the low dimensional setting (5 and 10 dimensions) which is more informative. It can be seen that our hyperbolic cones are better than all the baselines in all settings, except in the 0% setting for which order embeddings are better. However, once a small percentage of the transitive closure edges becomes available during training, we observe significant improvements of our method, sometimes by more than 8% F1 score. Moreover, hyperbolic cones have the largest growth when transitive closure edges are added at train time. We further note that, while mathematically not justified[8], if embeddings of our proposed Euclidean cones model are initialized with the Poincaré embeddings instead of the Simple Euclidean ones, then they perform on par with the hyperbolic cones.

6. Conclusion

Learning meaningful graph embeddings is relevant for many important applications. Hyperbolic geometry has proven to be powerful for embedding hierarchical structures. We here take one step forward and propose a novel model based on geodesically convex entailment cones and show its theoretical and practical benefits. We empirically discover that strong embedding methods can vary a lot with the percentage of the taxonomy observable during training and demonstrate that our proposed method benefits the most from increasing size of the training data. As future work, it would be interesting to understand if the proposed entailment cones can be used to embed more complex data such as sentences or images.

Our code is publicly available[9].

[8] Indeed, mathematically, hyperbolic embeddings cannot be considered as Euclidean points.

[9] https://github.com/dalab/hyperbolic_cones.

Acknowledgements

We would like to thank Maximilian Nickel, Colin Evans, Chris Waterson, Marius Pasca, Xiang Li and Vered Shwartz for helpful discussions about related work and evaluation settings.

This research is funded by the Swiss National Science Foundation (SNSF) under grant agreement number 167176. Gary Bécigneul is also funded by the Max Planck ETH Center for Learning Systems.

References

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

Bläsius, T., Friedrich, T., Krohmer, A., and Laue, S. Efficient embedding of scale-free graphs in the hyperbolic plane. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 57. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

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

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pp. 2787–2795, 2013.

Bowditch, B. H. A course on geometric group theory. 2006.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017.

Cannon, J. W., Floyd, W. J., Kenyon, R., Parry, W. R., et al. Hyperbolic geometry. *Flavors of geometry*, 31:59–115, 1997.

Cvetkovski, A. and Crovella, M. Hyperbolic embedding and routing for dynamic graphs. In *INFOCOM 2009, IEEE*, pp. 1647–1655. IEEE, 2009.

De Sa, C., Gu, A., Ré, C., and Sala, F. Representation tradeoffs for hyperbolic embeddings. *arXiv preprint arXiv:1804.03329*, 2018.

Fu, R., Guo, J., Qin, B., Che, W., Wang, H., and Liu, T. Learning semantic hierarchies via word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 1199–1209, 2014.

Ganea, O.-E. and Hofmann, T. Deep joint entity disambiguation with local neural attention. *arXiv preprint arXiv:1704.04920*, 2017.

Goyal, P. and Ferrara, E. Graph embedding techniques, applications, and performance: A survey. *arXiv preprint arXiv:1705.02801*, 2017.

Gromov, M. Hyperbolic groups. In *Essays in group theory*, pp. 75–263. Springer, 1987.

Grover, A. and Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864. ACM, 2016.

Hamann, M. On the tree-likeness of hyperbolic spaces. *Mathematical Proceedings of the Cambridge Philosophical Society*, pp. 117, 2017. doi: 10.1017/S0305004117000238.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

Hopper, C. and Andrews, B. The ricci flow in riemannian geometry, 2010.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. Skip-thought vectors. In *Advances in neural information processing systems*, pp. 3294–3302, 2015.

Krioukov, D., Papadopoulos, F., Boguñá, M., and Vahdat, A. Greedy forwarding in scale-free networks embedded in hyperbolic metric spaces. *ACM SIGMETRICS Performance Evaluation Review*, 37(2):15–17, 2009.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguná, M. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

Lamping, J., Rao, R., and Pirolli, P. A focus+ context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 401–408. ACM Press/Addison-Wesley Publishing Co., 1995.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4): 235–244, 1990.

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, pp. 6341–6350, 2017.

Nickel, M., Tresp, V., and Kriegel, H.-P. A three-way model for collective learning on multi-relational data. 2011.

Parkkonen, J. Hyperbolic geometry. 2013.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pp. 1532–43, 2014.

Robbin, J. W. and Salamon, D. A. Introduction to differential geometry. *ETH, Lecture Notes, preliminary version, January*, 2011.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiskỳ, T., and Blunsom, P. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*, 2015.

Sarkar, R. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pp. 355–366. Springer, 2011.

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

Shwartz, V., Goldberg, Y., and Dagan, I. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 2389–2398, 2016.

Spivak, M. A comprehensive introduction to differential geometry. volume four. 1979.

Vendrov, I., Kiros, R., Fidler, S., and Urtasun, R. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.