Probabilistic Word Embeddings in Kinematic Space

Adarsh Jamadandi [1] Rishabh Tigadoli[2] Ramesh Tabib [1]
Uma Mudenagudi [1]

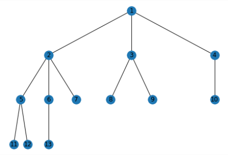
December 10, 2020

[1]KLE Technological University
Hubli, India.

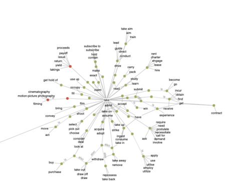
[2]Mercedes-Benz R and D Centre,
Bangalore India.

How to learn symbolic data?

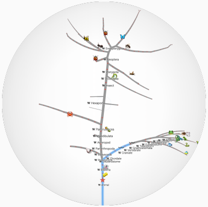
1. Symbolic data often exhibits hierarchical anatomy. For example



Tree-like structure of Graphs.



WordNet-Lexical database.



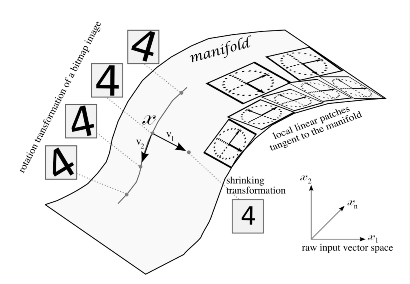
Phylogenetic Tree.

2. How to learn symbolic data with deep learning algorithms? It is important to preserve the semantic/functional relationship between entities in the data - Nickel and Kiela [2018].

WordNet Visualization - `http://wordvis.com/about.html`
Phylogenetic Tree - `https://github.com/glouwa/d3-hypertree`

Manifold Hypothesis

1. Data lies on a low-dimensional manifold embedded in input space.
2. Resurgence of Manifold hypothesis - with explicit assumption of the underlying geometry - Spherical/Hyperbolic.

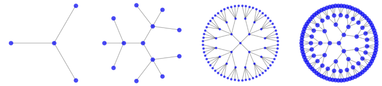


Pixels of images lie on a natural image manifold.

Figure : Bengio [2012]

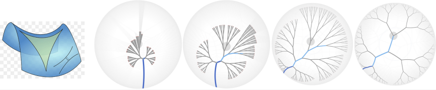
Hyperbolic Space $\mathbb{H}^n$

1. Euclidean space introduces massive distortions when modelling hierarchical data.



Trying to embed a binary tree in Euclidean space, we quickly run out of space. Notice how unrelated nodes are forced together.

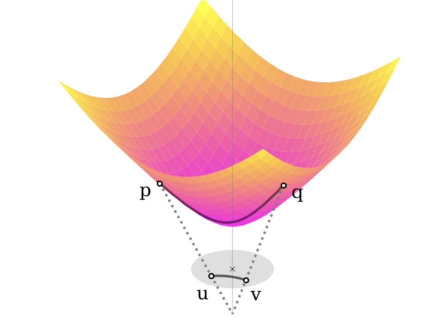
2. Hyperbolic space provides an exciting alternative - Non-Euclidean geometry with constant negative curvature - Space grows exponentially!



Probabilistic Inference in Hyperbolic Space

Wrapped Normal Distribution in $\mathbb{H}^n$

1. Authors Nagano et al. [2019] propose a Wrapped Normal Distribution on the Lorentz model of the Hyperbolic space.
2. Lorentz model - a Riemannian manifold $(\mathcal{L}, g_L)$, where $\mathcal{L} = \{\mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -1, x_o > 0\}$ and the Minkowski inner product defined by,



Inner product

$$\langle x, y \rangle_{\mathcal{L}} := -x_0 y_0 + x_1 y_1 + ... + x_n y_n$$

Lorentz Distance Metric

$$d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) = \text{arcosh}(-\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}})$$

Figure Lorentz Model : Nickel and Kiela [2018]

Constructing Wrapped Normal Distribution

1. Use a combination of Parallel Transport and Exponential Map to construct a Normal distribution on a Riemannian manifold.
2. Sample from a Gaussian distribution defined in the tangent space at $\mu_0 = 0$. Use Parallel transport and Exponential map to map the point onto the manifold.
3. How does this help probabilistic inference problems?

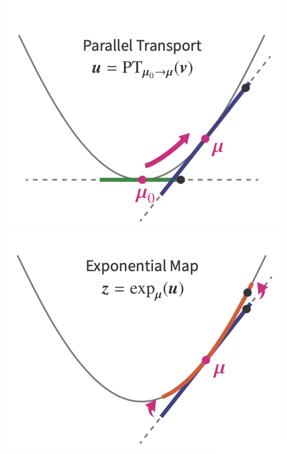


Figure:Nagano et al. [2019].

Gaussian Word Embeddings (Vilnis and McCallum [2015])

1. Map lexically distributed representations to a density instead of point vectors.
2. Advantages - Better expression of Asymmetry and Uncertainty.

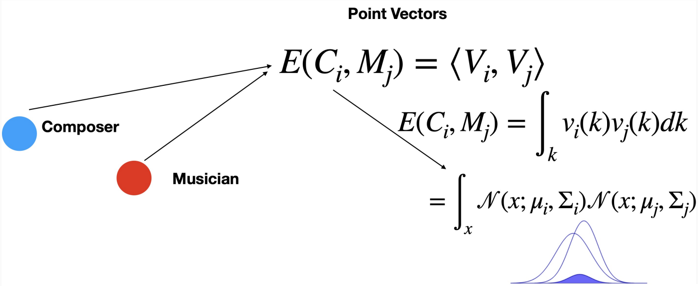
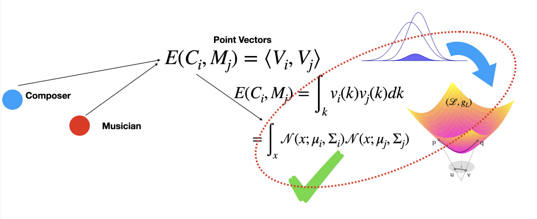


Figure Lorentz Model : Nickel and Kiela [2018]

Word Embeddings in Hyperbolic Space

1. Probabilistic Word Embeddings in hyperbolic space - (Nagano et al. [2019])

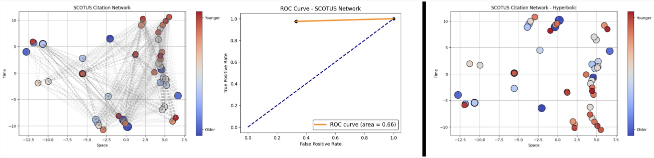


	Euclid		Hyperbolic	
n	MAP	Rank	MAP	Rank
5	0.296$\pm$.006	25.09$\pm$.80	**0.506$\pm$.017**	**20.55$\pm$1.34**
10	0.778$\pm$.007	**4.70$\pm$.05**	**0.795$\pm$.007**	5.07$\pm$.12
20	0.894$\pm$.002	**2.23$\pm$.03**	**0.897$\pm$.005**	2.54$\pm$.20
50	0.942$\pm$.003	1.51$\pm$.04	**0.975$\pm$.001**	**1.19$\pm$.01**
100	0.953$\pm$.002	1.34$\pm$.02	**0.978$\pm$.002**	**1.15$\pm$.01**

Going beyond Hyperbolic space

Going Beyond Hyperbolic Space - Motivation

1. Can we obtain more powerful representations, if we go beyond hyperbolic space?
2. Yes! Symbolic data also exhibit properties such as Causality, in addition to hierarchy. Hyperbolic space fails to account for this property.
3. Pseudo-Riemannian manifolds such as Lorentzian manifolds are more natural embedding spaces - Clough and Evans [2017].



Citation networks exhibit property such as Causality in addition to Hierarchy.

Embedding the networks in Lorentzian manifolds preserves the causal structure.

Kinematic Space

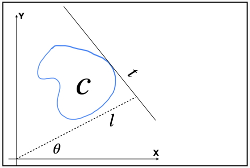
1. We propose an auxiliary Lorentzian space called Kinematic Space - a space of oriented geodesics.
2. Inspired from Integral Geometry [Santaló and Kac [2004]] and Theoretical Physics [Czech et al. [2015]].
3. Powerful mathematical framework that can transform geometrical information from one space to another.

Crofton's Formula

1. Suppose we are interested in measuring the length of a curve $\mathcal{C}$ on an Euclidean plane, we can draw a number of tangents $\mathbf{t}$ to the curve. The equation of the straight line given by -

$$x \cos\theta + y \sin\theta - l = 0 \tag{1}$$

2. where θ is the polar angle and l is the distance of straight line from the origin. We can estimate the length of the curve $\mathcal{C}$ by the Crofton's formula [Santaló and Kac [2004]].



Crofton's Formula

Crofton's Formula

$$\text{Length} = \frac{1}{4}\int_0^{2\pi} d\theta \int_{-\infty}^{+\infty} \eta(\theta, l) dl$$

1. $\eta(\theta, l)$ gives the intersection number of the tangents and the curve.
2. The space of oriented geodesics panning $\theta \in [0, 2\pi]$ and $l \in (-\infty, \infty)$ is called the Kinematic Space of the Euclidean plane.
3. We can extend this formulation to hyperbolic space - Straight lines replaced by Geodesics.

Crofton's Formula in $\mathbb{H}^n$

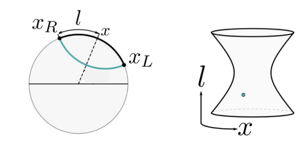
1. Equation of Geodesic -

$$\tanh \rho \cos(\hat{\theta} - \theta) = \cos \alpha \tag{2}$$

2. α is the opening angle of the geodesic and θ is the angular coordinate of the center of the geodesic. The Kinematic Space of the hyperbolic plane is now the space of geodesics panned by $\theta \in [0, 2\pi]$ and $\alpha \in [0, \pi]$.
3. The length of the curve can be interpreted as volume of lines intersecting the curve!

Kinematic Space ($\mathcal{K}_s$)

1. Kinematic space - a Lorentzian geometrical space of oriented geodesics.
2. Geodesics γ are represented as Points in Kinematic space.
3. Can transform geometrical information from one space to another.



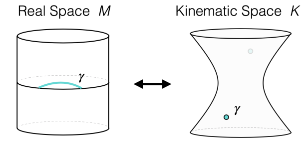
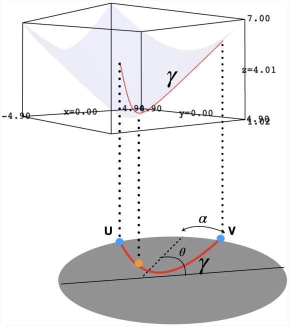


Figure: Czech et al. [2015]

Kinematic space as a Geometric Inductive

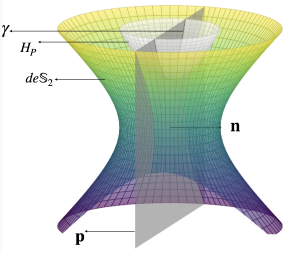
1. Lorentz model is the preferred model of hyperbolic geometry, computationally tractable - Mathieu et al. [2019], Nickel and Kiela [2017], Bose et al. [2020].
2. We propose to use Poincaré upper half plane model $\mathbb{H}_{UP}$ as a geometrical inductive for deep representation learning - Rarely considered in literature.
3. $\mathbb{H}_{UP} \rightleftharpoons_{\mathcal{K}_s}$ Some Computationally Tractable Manifold



de Sitter space

What is de Sitter space?

Let $de\mathbb{S}_2$ be the $(d+1)$ dimensional de Sitter space in the $(d+2)$ dimensional Minkowski space $\mathbb{M}$ visualized as a single sheeted hyperboloid with pseudo-radius λ given by
$-z_0^2 + z_1^2 + z_3^2 + ... + z_n^2 = \lambda^2 = \frac{1}{K}$.



A maximally symmetric, positive curvature, Lorentzian manifold, visualized as a single sheeted hyperboloid.

de Sitter space ($de\mathbb{S}_2$) as the Kinematic space for $\mathbb{H}_{UP}$

Proposition 1

The Kinematic space ($\mathcal{K}_s$) of the upper half-plane model ($\mathbb{H}_{UP}$) is the $(d+1)$ dimensional de Sitter space ($de\mathbb{S}_2$) visualized as a single sheeted hyperboloid in the Minkowski space $\mathbb{M}^{d+2}$, and there is a canonical identification between the geodesics $\gamma \in \mathbb{H}_{UP}$ and the points in $\mathcal{K}_s$.

Proposition 2

For every geodesic γ that can be drawn on $\mathbb{H}_{UP}$, we can find a unique plane $\mathbf{p}$ intersecting $\mathbb{H}_{UP}$, whose normal $\mathbf{n}$ at the origin corresponds to a point in $de\mathbb{S}_2$.

de Sitter space

Induced Distance Metric

$$d_{de\mathbb{S}_2}(\mathbf{x},\mathbf{y}) = \lambda \operatorname{arcosh}\left(\frac{-\langle \mathbf{x},\mathbf{y}\rangle_{\mathcal{L}}}{\lambda^2}\right) \tag{3}$$

Exponential Map and Log Map

Exponential Map $\exp_p : \mathcal{T}_p de\mathbb{S}_2 \to de\mathbb{S}_2$

$$\exp_p(\mathbf{v}) = \cosh(\sqrt{K}||\mathbf{v}||_{\mathcal{L}})\mathbf{p} + \mathbf{v}\frac{\sinh(\sqrt{K}||\mathbf{v}||_{\mathcal{L}})}{\sqrt{K}||\mathbf{v}||_{\mathcal{L}}} \tag{4}$$

Log Map $\log_p : \mathcal{T}_p de\mathbb{S}_2 \to de\mathbb{S}_2$

$$\log_p(\mathbf{y}) = \frac{\operatorname{arcosh}(-K\langle \mathbf{x},\mathbf{y}\rangle_{\mathcal{L}})}{\sinh(\operatorname{arcosh}(-K\langle \mathbf{x},\mathbf{y}\rangle_{\mathcal{L}}))}(\mathbf{y} - K\langle \mathbf{x},\mathbf{y}\rangle_{\mathcal{L}}\mathbf{p}) \tag{5}$$

ICPR 2020 25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION Milan, Italy 10 | 15 January 2021

Wrapped Normal Distribution in de Sitter space

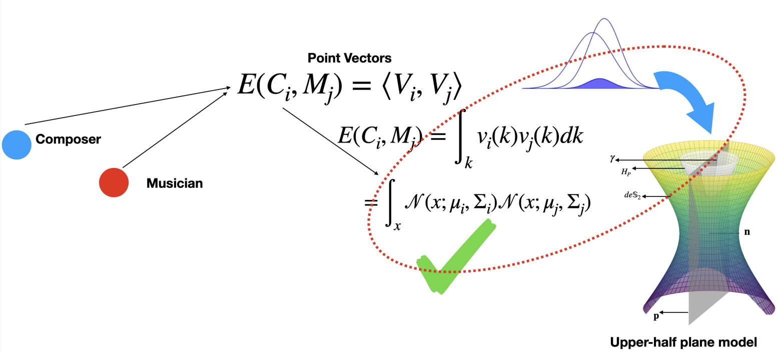
1. We extend the formulation of Wrapped Normal Distribution in Lorentz model by authors [Nagano et al. [2019]] to de Sitter space.
2. Sampling a vector $\mathbf{v}$ from the Gaussian distribution $\mathcal{N}(0, \Sigma)$ defined over $\mathbb{R}^n$.
3. Parallel transporting $\mathbf{v}$ from the tangent space $\mathbf{o}$ to the tangent space of new point $\mathbf{u}$ to obtain $\mathbf{j}$ by using the formula,

$$PT_{\mathbf{o}\to\mathbf{u}}(\mathbf{v}) = \mathbf{v} + \frac{K\langle y, u\rangle_{\mathcal{L}}}{1 + K\langle o, u\rangle_{\mathcal{L}}}(o + u) \tag{6}$$

4. Map the point $\mathbf{j}$ to the manifold using the exponential map at $\mathbf{u}$ given by Equation (Bose et al. [2020]).

$$\log g(\mathbf{z}) = \log g(v) - (n-1)\log\left(\frac{\sinh ||\mathbf{j}||}{||\mathbf{j}||}\right) \tag{7}$$

Probabilistic Word Embeddings in Kinematic Space



ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

We compare our probabilistic word embedding framework with (Vilnis and McCallum [2015]) and (Nagano et al. [2019]).

	Euclid		Hyperbolic		**Ours**	
Dimension	Rank	MAP	Rank	MAP	Rank	MAP
5	70.15 ± 3.76	0.15 ± 0.01	90.81 ± 8.01	0.20 ± 0.01	$\mathbf{4.23 \pm 2.98}$	$\mathbf{0.53 \pm 0.13}$
10	24.06 ± 8.85	0.43 ± 0.02	15.67 ± 4.78	0.53 ± 0.07	$\mathbf{1.43 \pm 0.01}$	$\mathbf{0.86 \pm 0.12}$
20	13.63 ± 1.69	0.65 ± 0.04	8.27 ± 2.59	0.71 ± 0.06	$\mathbf{2.05 \pm 1.33}$	$\mathbf{0.94 \pm 0.06}$
50	6.43 ± 2.17	0.75 ± 0.05	4.84 ± 0.95	0.74 ± 0.01	$\mathbf{1.50 \pm 0.23}$	$\mathbf{0.97 \pm 0.00}$

We compare our proposed method and the hyperbolic version Nagano et al. [2019] with the deterministic embeddings framework proposed by authors in Nickel and Kiela [2017] on the WordNet-Noun dataset.

	Poincaré Nickel and Kiela [2017]		Hyperbolic		**Ours**	
Dimension	Rank	MAP	Rank	MAP	Rank	MAP
5	4.9 ± 0.00	$\mathbf{0.823 \pm 0.00}$	90.81 ± 8.01	0.20 ± 0.01	$\mathbf{4.23 \pm 2.98}$	0.53 ± 0.13
10	4.02 ± 0.00	0.851 ± 0.00	15.67 ± 4.78	0.53 ± 0.07	$\mathbf{1.43 \pm 0.01}$	$\mathbf{0.86 \pm 0.12}$
20	3.84 ± 0.00	0.855 ± 0.00	8.27 ± 2.59	0.71 ± 0.06	$\mathbf{2.05 \pm 1.33}$	$\mathbf{0.94 \pm 0.06}$
50	3.98 ± 0.00	0.86 ± 0.00	4.84 ± 0.95	0.74 ± 0.01	$\mathbf{1.50 \pm 0.23}$	$\mathbf{0.97 \pm 0.00}$

Effect of Curvature and $\delta-$ hyperbolicity

1. $\delta-$ hyperbolicity measures the Tree-likeliness of the data.
2. The smaller δ value $\implies$ the data can be isometrically embedded in hyperbolic space.
3. Changing curvature, effectively makes space more hyperbolic $\implies$ better MAP and Rank values.

Effect of curvature on learning word embeddings.

Curvature	K=2		K=3		K=5	
Dimension	Rank	MAP	Rank	MAP	Rank	MAP
5	24.77± 18.69	0.24 ± 0.12	4.33± 2.51	0.53± 0.01	4.22±2.98	0.53± 0.13
10	2.22± 0.10	0.79± 0.04	1.66± 0.62	0.85± 0.17	1.44± 0.00	0.86 ± 0.12
20	2.05± 1.33	0.94 ± 0.06	1.61± 0.23	0.80± 0.21	13.00± 15.39	0.60± 0.33
50	1.50± 0.23	0.96± 0.01	3.00 ± 0.31	0.78 ± 0.11	0.58 ± 0.29	8.94 ± 10.92

Summary

1. We introduce Kinematic space, an auxiliary Lorentzian geometry in the context of deep representation learning for hierarchical data.
2. Leveraging this formulation, we show that learning representations in the upper half-plane model is equivalent to learning in a maximally symmetric pseudo-Riemannian manifold called de Sitter space, where Riemannian optimization methods are applicable.
3. We formulate Wrapped Normal Distribution in Kinematic Space and use it for probabilistic word embeddings.

Questions?

References

Y. Bengio. Evolving culture vs local minima. *Studies in Computational Intelligence*, 557, 03 2012. doi: 10.1007/978-3-642-55337-0_3.

A. J. Bose, A. Smofsky, R. Liao, P. Panangaden, and W. L. Hamilton. Latent variable modelling with hyperbolic normalizing flows. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

J. R. Clough and T. S. Evans. Embedding graphs in lorentzian spacetime. *PLOS ONE*, 12(11):1–14, 11 2017. doi: 10.1371/journal.pone.0187301. URL https://doi.org/10.1371/journal.pone.0187301.

References ii

B. Czech, L. Lamprou, S. McCandlish, and J. Sully. Integral Geometry and Holography. *JHEP*, 2015. doi: 10.1007/JHEP10(2015)175.

E. Mathieu, C. Le Lan, C. J. Maddison, R. Tomioka, and Y. W. Teh. Continuous hierarchical representations with poincaré variational auto-encoders. In *Advances in Neural Information Processing Systems 32*, pages 12565–12576. Curran Associates, Inc., 2019.

Y. Nagano, S. Yamaguchi, Y. Fujita, and M. Koyama. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *ICML*, 2019.

M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6338–6347. Curran Associates, Inc., 2017.

M. Nickel and D. Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. *CoRR*, abs/1806.03417, 2018. URL http://arxiv.org/abs/1806.03417.

L. A. Santaló and M. Kac. *Integral Geometry and Geometric Probability*. Cambridge University Press, 2004.

L. Vilnis and A. McCallum. Word representations via gaussian embedding. In *International Conference on Learning Representations*, 2015.