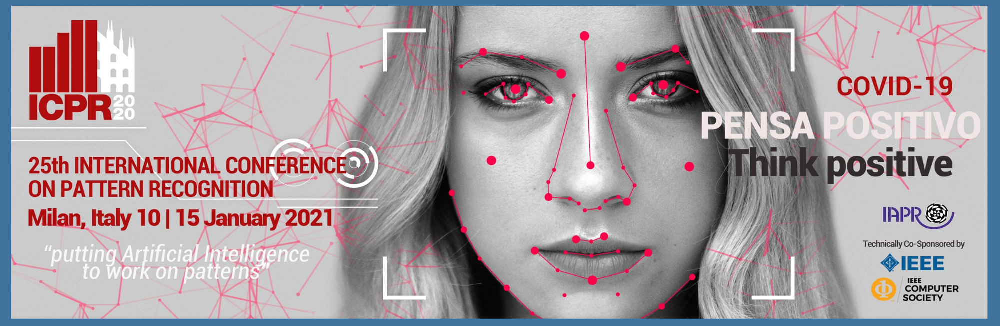


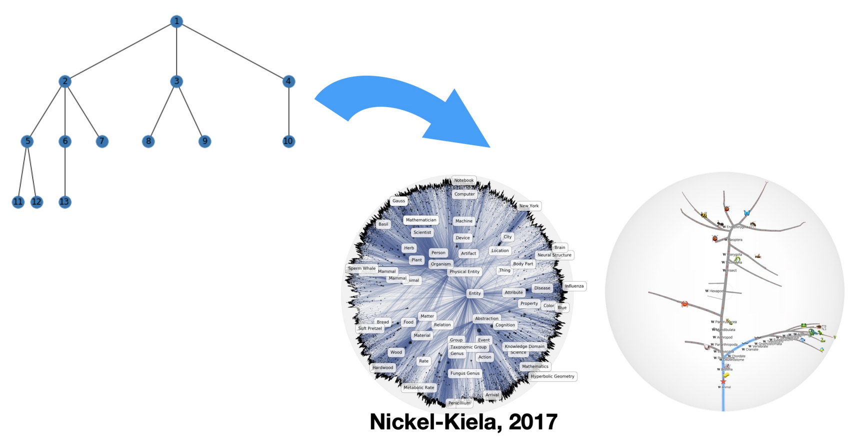
Probabilistic Word Embeddings in Kinematic Space

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

[1]KLE Technological University, Hubli, India.
[2]Mercedes Benz R and D Centre, Bangalore, India.

Motivation

1. Hierarchical data benefits from Hyperbolic space.



2. Can probabilistic inference problems benefit from appropriate geometric biases? Yes!

Wrapped Normal Distribution

1. Wrapped Normal Distribution - Gaussian-like distribution but constructed on the Lorentz Model of hyperbolic geometry [2].
2. Construct Gaussian-like distribution on the tangent space at $\mu_0 = 0$. Use Parallel Transport and Exponential Map to map to a Riemannian manifold.

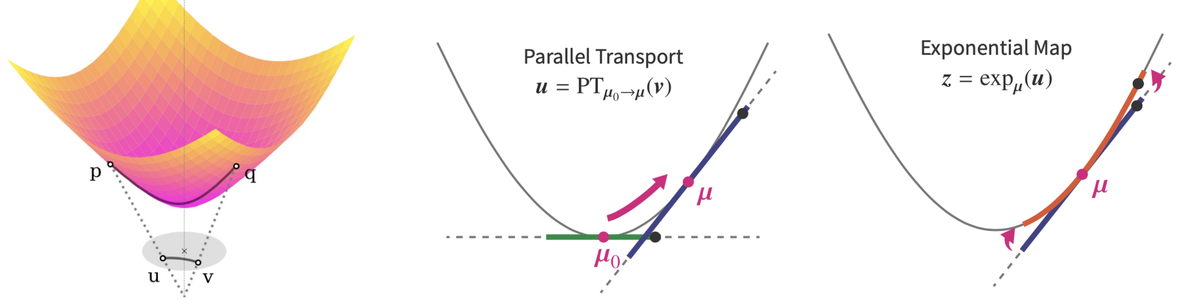
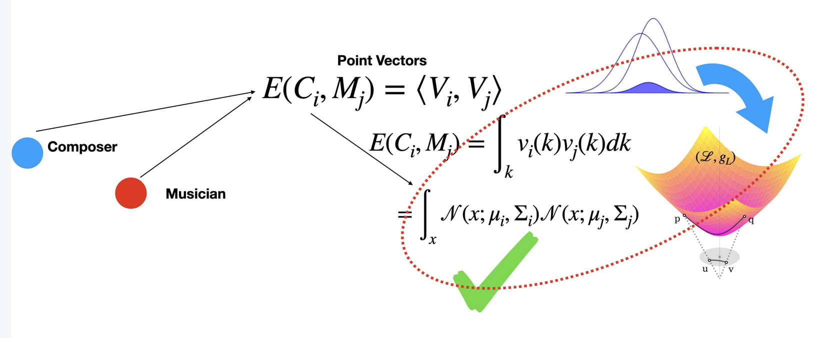


Fig Credits: [2], [Nickel-Kiela, 2017].

Application- Probabilistic Word Embeddings

1. Map lexically distributed representations to density, instead of point vectors.
2. Gaussian-like distribution constructed on the Lorentz Model.

Probailistic Word Embeddings



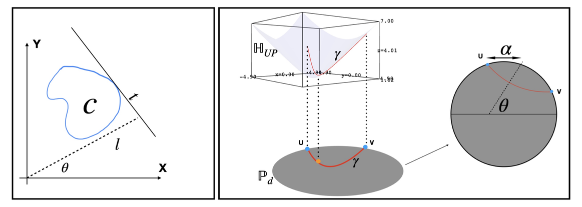
Can we go beyond hyperbolic spaces for more powerful representations?

Our Contributions

1. Kinematic space - auxiliary Lorentzian space for Deep Representation Learning.
2. Using Kinematic space, we show $\mathbb{H}_{UP} \rightleftharpoons_{\mathcal{K}_s}$ de Sitter space.
3. Map word representations to Gaussian-like distribution constructed on de Sitter space. Better MAP and Rank.

Kinematic Space

1. An auxiliary Lorentzian geometry inspired by Theoretical Physics [1] and Integral Geometry [4].
2. Powerful mathematical formalism that can transform geometrical information such as geodesic distance and exponential map from one space to another.



3. How? Using -

Crofton's Formula

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

Length of a curve can be re-interpreted as volume of intersecting lines (geodesics). The space of oriented geodesics - Kinematic space.

$de\mathbb{S}_2$

The de Sitter space is a maximally symmetric, Lorentzian manifold with constant positive curvature. 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 $-z_0^2 + z_1^2 + z_3^2 + ... + z_n^2 = \lambda^2 = \frac{1}{K}$. The induced distance function is given by

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

de Sitter space ($de\mathbb{S}_2$) as the Kinematic Space

1. We propose to use Poincaré upper half plane model ($\mathbb{H}_{UP}$) of hyperbolic geometry to construct Gaussian-like distribution.
2. Rarely considered in literature - Computationally Intractable.
3. $\mathbb{H}_{UP} \rightleftharpoons_{\mathcal{K}_s}$ de Sitter space.

Learning in $\mathbb{H}_{UP}$ is equivalent to learning in $de\mathbb{S}_2$!

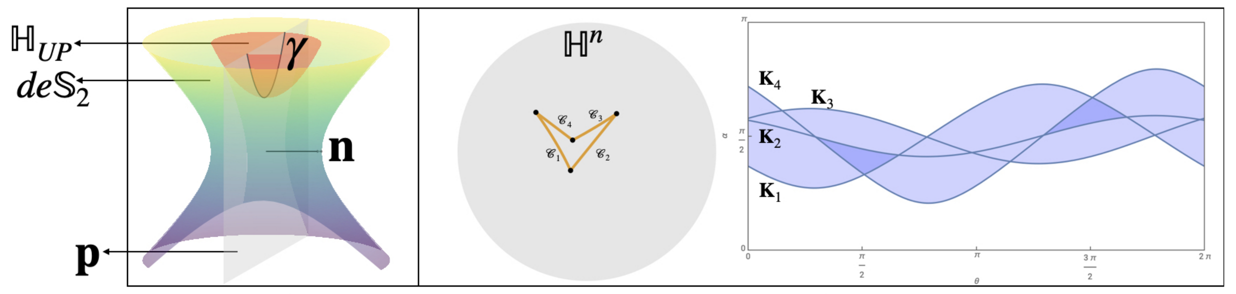


Figure 1. (Left) The de Sitter space can be visualized as a single sheeted hyperboloid in Minkowski space. A geodesic γ drawn on the Upper half plane model can be interpreted as a single point in de Sitter space.(Right) Curves drawn in hyperbolic space $\mathbb{H}^2$ and their corresponding Kinematic space.

Wrapped Normal Distribution in $de\mathbb{S}_2$

1. Sampling a vector $\mathbf{v}$ from the Gaussian distribution $\mathcal{N}(0, \Sigma)$ defined over $\mathbb{R}^n$.
2. 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,

$$PT_{\mathbf{o}\to\mathbf{u}}(\mathbf{v}) = \mathbf{v} + \frac{K\langle y, u \rangle_{\mathcal{L}}}{1 + K\langle o, u \rangle_{\mathcal{L}}}(o + u) \tag{3}$$

3. Map the point $\mathbf{j}$ to the manifold using the exponential map at $\mathbf{u}$ given by

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

To calculate the probability density of $\mathcal{G}_{de\mathbb{S}_2}(\mu, \Sigma)$,

$$\log g(\mathbf{z}) = \log g(v) - (n-1)\log\left(\frac{\sinh||\mathbf{j}||}{||\mathbf{j}||}\right) \tag{5}$$

where, $\log g(\mathbf{z})$ is the wrapped normal distribution and $\log g(v)$ is the normal distribution in tangent space of $\mathbf{o}$.

Application - Probabilistic Word Embeddings in Kinematic Space

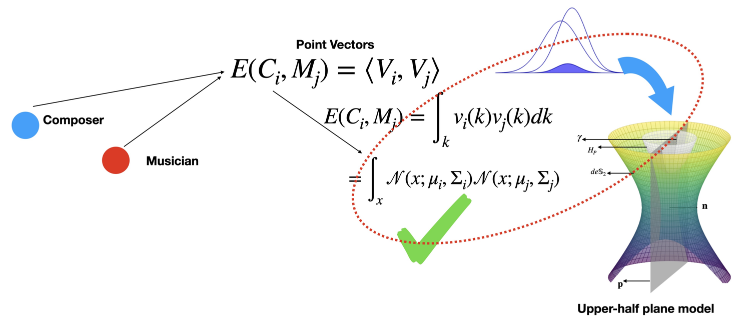


Table 1. We compare our probabilistic word embedding framework with [5] and [2] which also maps word representations to a density on the WordNet-Noun hierarchy dataset.

	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	**4.23 ± 2.98**	**0.53 ± 0.13**
10	24.06 ± 8.85	0.43 ± 0.02	15.67 ± 4.78	0.53 ± 0.07	**1.43 ± 0.01**	**0.86 ± 0.12**
20	13.63 ± 1.69	0.65 ± 0.04	8.27 ± 2.59	0.71 ± 0.06	**2.05 ± 1.33**	**0.94 ± 0.06**
50	6.43 ± 2.17	0.75 ± 0.05	4.84 ± 0.95	0.74 ± 0.01	**1.50 ± 0.23**	**0.97 ± 0.00**

Table 2. We compare our proposed method and the hyperbolic version [2] with the deterministic embeddings framework proposed by authors in [3] on the WordNet-Noun dataset.

	Poincaré [3]		Hyperbolic		**Ours**	
Dimension	Rank	MAP	Rank	MAP	Rank	MAP
5	4.9 ± 0.00	**0.823 ± 0.00**	90.81 ± 8.01	0.20 ± 0.01	**4.23 ± 2.98**	0.53 ± 0.13
10	4.02 ± 0.00	0.851 ± 0.00	15.67 ± 4.78	0.53 ± 0.07	**1.43 ± 0.01**	**0.86 ± 0.12**
20	3.84 ± 0.00	0.855 ± 0.00	8.27 ± 2.59	0.71 ± 0.06	**2.05 ± 1.33**	**0.94 ± 0.06**
50	3.98 ± 0.00	0.86 ± 0.00	4.84 ± 0.95	0.74 ± 0.01	**1.50 ± 0.23**	**0.97 ± 0.00**

References

[1] Bartlomiej et.al Czech.
Integral geometry and holography.
JHEP, 2015.

[2] Yoshihiro Nagano et al.
A wrapped normal distribution on hyperbolic space for gradient-based learning.
In *ICML*, 2019.

[3] Maximillian Nickel and Douwe Kiela.
Poincaré embeddings for learning hierarchical representations.
In *NeurIPS*, 2017.

[4] Luis A. Santaló and Mark Kac.
Integral Geometry and Geometric Probability.
Cambridge University Press, 2004.

[5] L. Vilnis and A McCallum.
Word representations via gaussian embedding.
In *ICLR*, 2015.