AMLDS Project: JL Lemma in Hyperbolic Spaces

Siddarth Sagar, Ola El Khatib

Fall 2021

1 Introduction

Classical machine learning models represent all types of data sets in Euclidean embeddings due to their vectorial structure and convenience of working with it [1]. However, lately, many research areas in machine learning are exploring embedding certain data sets in non-Euclidean spaces such as hyperbolic spaces. Research has shown that many types of data sets (especially hierarchical data sets) are better represented in hyperbolic spaces than Euclidean ones due to their properties and structure.

Embeddings are highly used today in machine learning to analyse data in high dimensions by doing dimensionality reduction into a much smaller dimensional space without losing information of the original dataset especially the pair wise distances between points. One of the most powerful tools in this domain is the Johnson-Lindenstraus Lemma which shows that distances between points in an n-dimensional space can be embedded with $(1+\epsilon)$ distortion into $O(\frac{logn}{\epsilon^2})$-dimensional space.

In this paper, we will present an overview of what a hyperbolic space is and what advantages it has over Euclidean spaces. In addition, we will study the possibilities of applying JL lemma as a dimensionality reduction technique on points embedded in the hyperbolic space.

2 Background on Hyperbolic Spaces

Hyperbolic geometry is the geometry obtained when the fifth postulate of Euclid is negated. Euclid's fifth postulate states that given a line and a point not on the line ,there is exactly one line through the given point that is parallel to the given line. Negating this postulate to get the hyperbolic geometry, states that from a line and a point not on the line, there are infinitely many lines through that point that do not intersect that line.

Hyperbolic geometry is a Riemannian geometry with constant negative curvature where curvature measures deviation from flat Euclidean geometry.

In order to study the hyperbolic space on a 2D-model, the negative curvature is extended into a larger area and projected on a unit disk as shown in Figure 1. The mathematical model is obtained is the Poincare disk.

The characteristics of the Poincare model:

- The circular disk has no boundary.
- Straight lines in hyperbolic spaces consist of arcs of Euclidean circles contained in the disk and are orthogonal with the circumference of a disk.
- The distance is measure towards the edge of the circle in the plane.

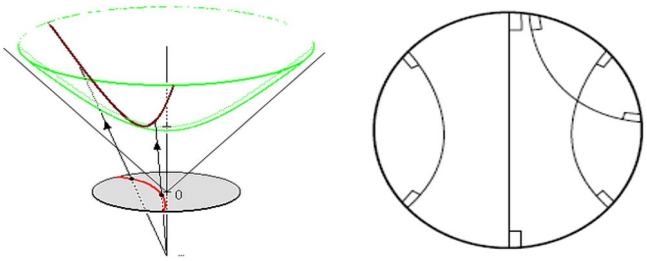

Figure 1: Poincare model

- The space expands exponentially as the circumference of the disk is approached.

3 Motivation to look for non-Euclidean Spaces for Embeddings

Embedding is a mapping of a discrete, categorical variable to a vector of continuous numbers. The accuracy of a graph embedding is measured when the the close distance of nodes in the embedding space reflect the similarity between nodes in the graph, i.e: nodes are close in the embedding space if and only if there exist an edge between the nodes in the graph mode. However, Euclidean embeddings can fail to preserve this property and thus lead researchers to look for another non-Euclidean spaces to use as embeddings. We mention below some of the main motivations look for non-Euclidean embeddings [3].

- In Euclidean spaces, similarity dot-product functions constrained to be PSD (Positive Semi Definite Gram Matrices).
- Euclidean spaces cannot embed large classes of graphs without low distortion or without loss of information (ex: tree, cycles)
- Euclidean spaces can only model intrinsically "flat" manifolds/surfaces.
- Euclidean space is too "Narrow" for Tree-like graphs since the volume of the Euclidean ball only grows polynomially with the radius $V_d^E(r) = \theta(r^d)$, so we quickly run out of space when we try to embed a large graph.
- The two leaf nodes of are very close in Euclidean space but are very far in true distance so the embedding space has already lost the property of capturing the graph as shown in Figure 2.

4 Our work

Given the above, it is noticed that hyperbolic embeddings provide a better fit for tree-like structures. So domain of hyperbolic embeddings encapsulates a lot of potential still to be discovered and leveraged upon in order to enhance machine learning and deep learning algorithms. In this section, we will present some analysis that we did on the hyperbolic and Euclidean distances in relation with the Johnson–Lindenstrauss (JL) lemma. These preliminary thoughts provide the intuition to pursue more rigorous mathematical proofs that might lead to useful tools to be used in hyperbolic embeddings.

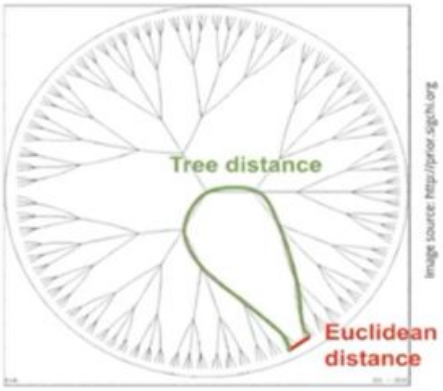

Figure 2: Random graph model [3]

4.1 Relationship between Hyperbolic Distance and Euclidean Distance

Let x, y be two d-dimensional vectors, the Euclidean distance between x and y is given by,

$$d_{\mathbf{E}}(x, y) = \|x - y\|_2$$

The d-dimensional Poincare-model of hyperbolic spaces: $H^d = \{x \in R : \|x\| < 1\}$, where $\|.\|$ is the Euclidean norm. In this model,the Riemannian distance can be computed in cartesian coordinates by:

$$d_{\mathbf{H}}(x, y) = cosh^{-1}\left(1 + 2\frac{\|x - y\|^2}{(1 - \|x\|^2)(1 - \|y\|^2)}\right)$$

It is known that the hyperbolic space behaves like Euclidean space near the center of the Poincare disk. For example, it is noticed that on the Circle Limit diagram, which resembles the tiling of the hyperbolic space, the images are scaled to their real size near the center while they become smaller and smaller towards the edge of the disk. Given this analysis, it is worth studying the relationship between the Euclidean distance and hyperbolic distance with respect to the distance from the center. This analysis would provide us insights on which tools from the Euclidean space can be utilized in the hyperbolic space especially near the center of the Poincare disk. Figure 3 clearly shows how the Euclidean and hyperbolic distances grow further from each other as we move away from the center.

To dig deeper into these observations, we plot the ratio between hyperbolic distance and Euclidean Distance on a 2-dimension Poincare disk and study how the ratio varies as we move away from the center towards the edge.

Figure 4 shows that the ratio of hyperbolic distance with Euclidean distance is a smooth curve that increases exponentially as we move towards the edge of the Poincare disc. This means that there is a function of r such that,

$$\frac{\textit{Hyperbolic distance of point } x}{\textit{Euclidean distance of point } x} = f(r(\textit{distance from the center}))$$

$$\frac{d_H(0, x)}{d_E(0, x)} = f(r)$$

$$d_H(0, x) = f(r) d_E(0, x)$$

$$d_H(0, x) = f(\|x\|) d_E(0, x)$$

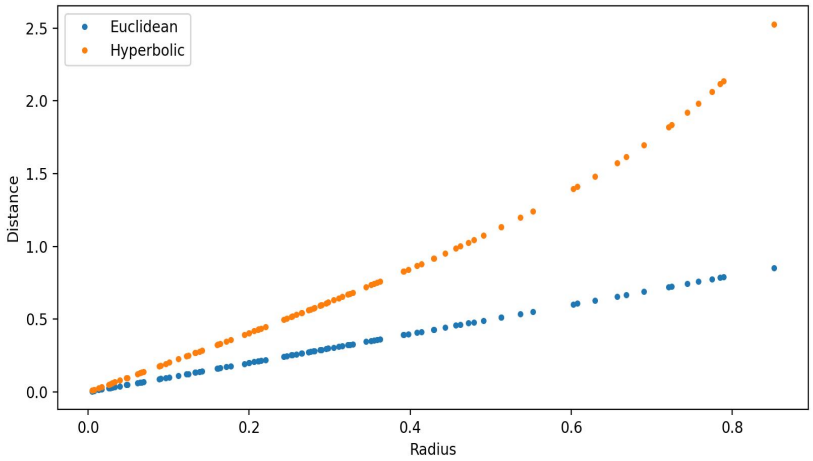

Figure 3: Plot of hyperbolic distance and Euclidean distance vs the distance from center

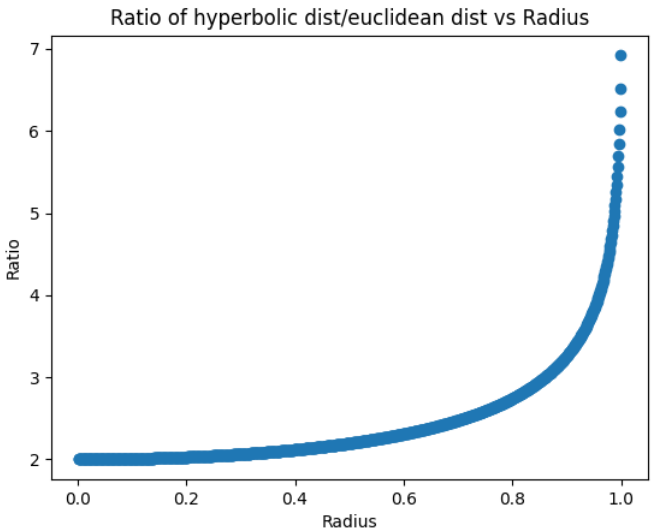

Figure 4: Plot of $\frac{hyperbolic\ distance}{Euclidean\ distance}$ versus the distance of point from center

Although, we did not infer the exact formula for the function, we were able to verify it by testing the function on distances from a point y not on the center and checking if the ratio remains an increasing convex function as we move away from the center. The function $f(r)$ could possibly be a complex function that takes some complex algebra to understand. The exact form of this function could be a goal for papers in the future to explore deeply.

We plotted the ratio graph for about 300 points starting points which are not the origin. Figure 5 describes how $f(r)$ behaves. We take points along the 45° angle in the first quadrant and we study how the ratio varies with respect to y set as points in second, third and fourth quadrant.
Figure 5 shows that $f(r)$ is a convex function. When we take y as a point other than (0,0) we can see that the ratio is not just a function radius but a function of distances of both points from the center.

$$\frac{d_H(x,y)}{d_E(x,y)} = f(\|x\|_2, \|y\|_2)$$

or

$$\frac{d_H(x,y)}{d_E(x,y)} = f(\|x-y\|_2)$$

Looking at the curve for the point $(-0.7,-0.7)$, we get a U shaped curve giving us more reason to explore the effect of choice of the point y on the $\frac{d_H(x,y)}{d_E(x,y)}$. We noticed that the choice of point y changes the shape of the curve.
The dotted lines on each curve are exponential functions of degree 5 we fit for those points. Although not completely accurate, these empirically generated functions give us a better idea of $f(\|x\|_2, \|y\|_2)$, and show that it is a more complex function that cannot be captured by a simple degree 5 exponential function and would require more work to discover. The choice of an exponential function was inspired by the fact that the distance in Poincare disk increases exponentially as we move towards the edge of the disc.

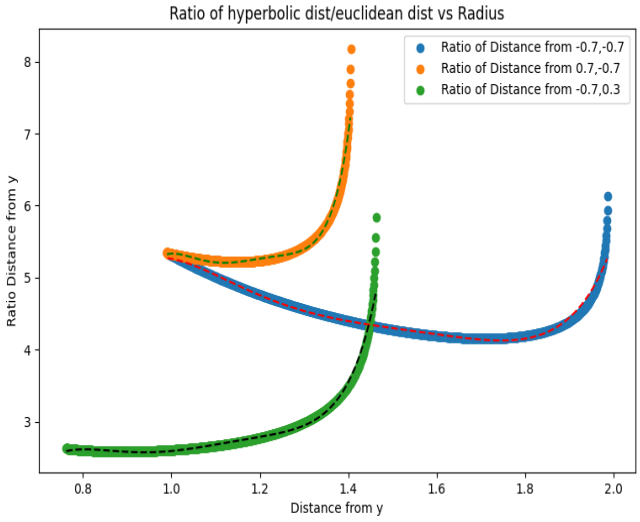

Figure 5: Plot of $\frac{Hyperbolic\ distance}{Euclidean\ distance}$ versus the distance of point from point y

4.2 JL Lemma and Perturbation Analysis

It is worth studying the possibility of applying the JL lemma on parts of the Poincare disk in the sense that for any set of n data points $q_1, q_2, ...q_n \in R$ there exists a linear map $\Pi : R^d \to R^k$ where $k = O(\frac{logn}{\epsilon^2})$ such that for all i, j:

$$(1-\epsilon)\|\Pi q_i - \Pi q_j\|_2^2 \leq \|q_i - q_j\|_2^2 \leq (1+\epsilon)\|\Pi q_i - \Pi q_j\|_2^2$$

Figure 3 shows that the distance between points grows exponentially when moving away from the center towards the boundary. Since the space behaves similar to Euclidean near the center and since JL lemma is specific to Euclidean spaces, it is worth investigating the possibility of applying JL lemma near the center.

To study the effect of applying the sketching on the Euclidean on the hyperbolic distance on JL lemma, we will do a perturbation analysis to get an insight about the relationship of ϵ with the hyperbolic distance.

$$d_{\mathbf{H}}(x,y) = cosh^{-1}\left(1 + 2\frac{\|\Pi x - \Pi y\|^2}{(1-\|x\|^2)(1-\|y\|^2)}\right) = cosh^{-1}\left(1 + 2\frac{(1+\epsilon)\|x-y\|^2}{(1-\|x\|^2)(1-\|y\|^2)}\right)$$

As expected, Figure 6 shows that the effect of increasing epsilon grows as we move away from the center.

From 4.1 we found that,

$$\frac{d_H(x,y)}{d_E(x,y)} = f(\|x\|_2, \|y\|_2)$$

$$d_H(x,y) = f(\|x\|_2, \|y\|_2) d_E(x,y)$$

Since we use $\|x\|$ and $\|y\|$ in the function, we can save them in the sketch allowing us to use it instead of $\|\Pi x\|$ and $\|\Pi y\|$. Applying JL-Lemma to the above function we get,

$$d_H(\Pi x, \Pi y) = f(\|x\|_2, \|y\|_2) d_E(\Pi x, \Pi y)$$

We know from JL-Lemma in Euclidean space that,

$$(1-\epsilon)\|x-y\|_2 \leq \|\Pi x - \Pi y\|_2 \leq (1+\epsilon)\|x-y\|_2$$

Hence the function becomes,

$$d_H(\Pi x, \Pi y) = f(\|x\|_2, \|y\|_2)\|\Pi x - \Pi y\|$$

$$f(\|x\|_2, \|y\|_2)(1-\epsilon)\|x-y\| \leq d_H(\Pi x, \Pi y) \leq f(\|x\|_2, \|y\|_2)(1+\epsilon)\|x-y\|$$

While in Euclidean space, applying JL-lemma perturbs the point with a small constant ϵ, in hyperbolic spaces this is not the case. The perturbation is multiplied by a factor of $f(\|x\|_2, \|y\|_2)$ which can make the small ϵ very large when the points are near the corners.

4.3 Using Triangle Inequality

Since JL lemma fails when we move away from the center, in this section we will analyze the possibility of using triangle inequality in hyperbolic spaces in order to approximate distances at the edge of the Poincare where the effect of epsilon is very large on the hyperbolic distance. Here, the hyperbolic distance between two points can be approximated by the triangle inequality by choosing the third point to be near the center, ie: $d_H(a,b) \leq d_H(a,z) + d_H(b,z)$ where z is chosen to be the center and a and b belong to a circle centered at the

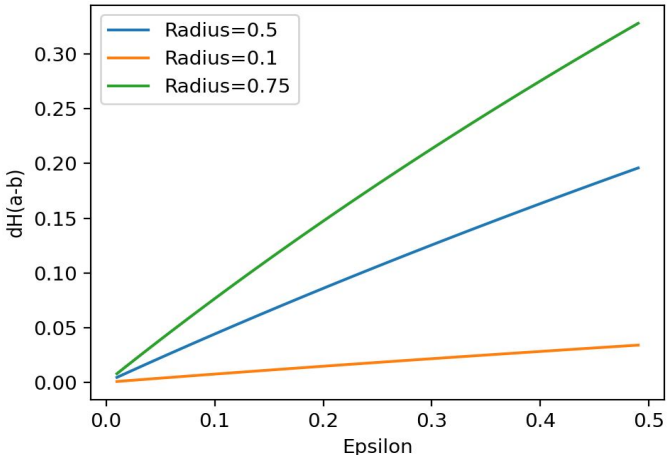

Figure 6: Plot of hyperbolic distance vs epsilon for different Radii.

origin and radius r. To validate this intuition, we plot the Euclidean, hyperbolic and hyperbolic via triangle inequality distances.

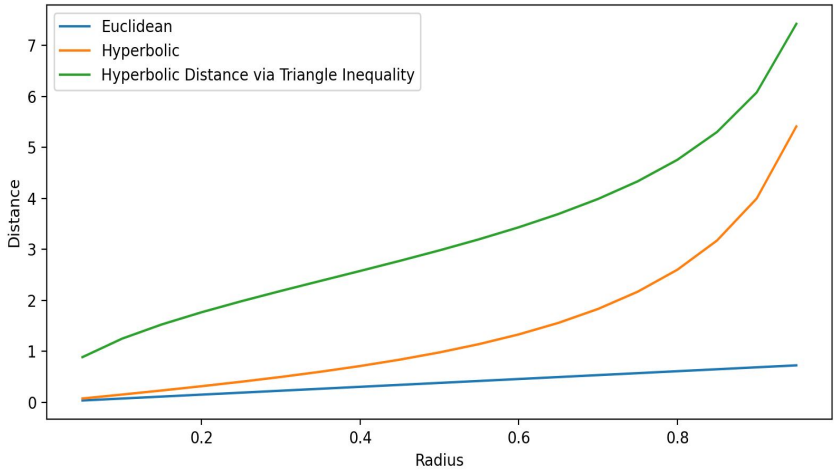

Figure 7: Plot of hyperbolic distance and Euclidean distance and hyperbolic via triangle inequality vs the distance from center

Figure 7 shows that the triangle inequality tends to approximate the hyperbolic distance near the edge. From the fitting equation shown before,

$$d_H(0,a) = f(\|a\|).d_E(0,a)$$

$$d_H(a,b) \approx \delta d_H(a,z) + \delta d_H(b,z), \textit{ as } a \textit{ and } b \textit{ approach the edge}$$

Where δ is some positive constant < 1 that changes the $\leq$ of triangle inequality to $\approx$.

$$d_H(a,b) \approx \delta f(\|a\|,\|z\|).d_E(a,z) + \delta f(\|b\|,\|z\|).d_E(b,z)$$

Applying JL lemma,

$$d_H(a,b) \approx \delta(1+\epsilon) f(\|a\|, \|z\|).d_E(\Pi a, \Pi z) + \delta(1+\epsilon) f(\|b\|, \|z\|).d_E(\Pi b, \Pi z)$$

We can put δ inside the $1+\epsilon$ term since both are constants. We finally get,

$$d_H(a,b) \approx (1+\epsilon) f(\|a\|, \|z\|).d_E(\Pi a, \Pi z) + (1+\epsilon) f(\|b\|, \|z\|).d_E(\Pi b, \Pi z)$$

Therefore, the JL lemma can be possibly applied using the triangle inequality for points near the edge of the Poincare.

5 Conclusion and Future Work

This study proposes a new way to calculate the hyperbolic distance that does not involve the Reimannian distance. We use this inference to explore the possibility of applying the JL lemma on regions of the Poincare disk in hyperbolic spaces. Our analysis shows empirically that the JL lemma can be better applied near the center of the disc and fails as we approach the edge. With this conclusion, we propose a better way to estimate distances near the edge using triangle inequality with a third point on the center and thus apply JL lemma there.

Future work in this area would be to mathematically find the function that describes the ratio of the hyperbolic and Euclidean distance. Furthermore, the JL lemma can mathematically be proved to fail by multiplying two points near the edge by a random Gaussian matrix and checking if they get mapped closer together. This can be done since the distances under the linear map are Chi-squared distributed and thus we can better analyse how likely a distance is to be contracted or expanded.

References

[1] Ganea, Octavian-Eugen, Gary Bécigneul, and Thomas Hofmann. "Hyperbolic neural networks." arXiv preprint arXiv:1805.09112 (2018).

[2] Chami, Ines, et al. "HoroPCA: Hyperbolic Dimensionality Reduction via Horospherical Projections." International Conference on Machine Learning. PMLR, 2021.

[3] Ganea, Octavian-Eugen. Non-Euclidean Neural Representation Learning of Words, Entities and Hierarchies. Diss. ETH Zurich, 2019.