Johnson-Lindenstrauss Lemma Beyond Euclidean Geometry

Chengyuan Deng, Jie Gao, Kevin Lu, Feng Luo, Cheng Xin[*]

Abstract

The Johnson-Lindenstrauss (JL) lemma is a cornerstone of dimensionality reduction in Euclidean space, but its applicability to non-Euclidean data has remained limited. This paper extends the JL lemma beyond Euclidean geometry to handle general dissimilarity matrices that are prevalent in real-world applications. We present two complementary approaches: First, we show the JL transform can be applied to vectors in pseudo-Euclidean space with signature (p, q), providing theoretical guarantees that depend on the ratio of the (p, q) norm and Euclidean norm of two vectors, measuring the deviation from Euclidean geometry. Second, we prove that any symmetric hollow dissimilarity matrix can be represented as a matrix of generalized power distances, with an additional parameter representing the uncertainty level within the data. In this representation, applying the JL transform yields multiplicative approximation with a controlled additive error term proportional to the deviation from Euclidean geometry. Our theoretical results provide fine-grained performance analysis based on the degree to which the input data deviates from Euclidean geometry, making practical and meaningful reduction in dimensionality accessible to a wider class of data. We validate our approaches on both synthetic and real-world datasets, demonstrating the effectiveness of extending the JL lemma to non-Euclidean settings.

1 Introduction

The Johnson-Lindenstrauss (JL) lemma [1] stands as a cornerstone result in dimensionality reduction. It states that random linear projection can reduce the dimensionality of datasets in Euclidean space, while approximately preserving pairwise distances. Formally,

Proposition 1.1 (Johnson-Lindenstrauss Lemma)**.** *For any set of n points $x_1, x_2, \ldots x_n$ in $\mathbb{R}^d$ and $\varepsilon \in (0, 1)$, there exists a map $f : \mathbb{R}^d \to \mathbb{R}^m$, where $m = O(\log n/\varepsilon^2)$ such that for any $i, j \in [n]$,*

$$(1-\varepsilon)\|x_i - x_j\|_2 \le \|f(x_i) - f(x_j)\|_2 \le (1+\varepsilon)\|x_i - x_j\|_2 \tag{1}$$

In modern algorithm design, the JL lemma is widely used as a key component or pre-processing step for high-dimensional data analysis. In addition to effectiveness, another crucial reason for its significant impact is that the JL lemma can be achieved by random linear maps, which are data-independent and easy to implement. Johnson-Lindenstrauss Lemma has found numerous applications in machine learning, from the most immediate applications in approximating all pairs distances (when the input data is high-dimensional), to approximate nearest neighbor search [2], approximate linear regression [3], clustering [4, 5, 6, 7], functional analysis [8] and compressed sensing [9].

Note that two conditions must be met to apply the JL lemma: (1) the data points lie in high-dimensional Euclidean space, and (2) the coordinates of all points are available. However, many real-world applications do not satisfy these two conditions. First, the dissimilarity measures used in modern data analysis are often non-Euclidean and sometimes even non-metric. Common examples

[*] Rutgers University. `{cd751,jg1555,kll160,fluo,cx122}@rutgers.edu`

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

include Minkowski distance, cosine similarity, Hamming distance, Jaccard index, Mahalanobis distance, Chebyshev distance, and Kullback–Leibler (KL) divergence, etc. Psychological studies have long observed that human similarity judgments do not conform to metric properties [10]. Second, high-dimensional coordinates may be unavailable or costly to obtain, whereas pairwise dissimilarities are easier to access. In recommendation systems, for instance, computing user or item embeddings can be expensive, while estimating pairwise dissimilarities (e.g. from co-click or co-purchase data) is relatively efficient. To date, our understanding of the JL lemma deviating from these two classical conditions mainly revolves around ℓ_p metrics and lower bound results.

In this paper, we study how to apply the Johnson-Lindenstrauss transform (a.k.a. random linear projection) in non-Euclidean, non-metric settings. We provide theoretical analysis on the performance of the JL transform in these settings and show experiments with both synthetic and real-world data.

Our setting. We consider the input as a dissimilarity matrix D of size $n \times n$ where D_{ij} is the dissimilarity between item i and j. We make only two assumptions for the dissimilarity measure: symmetric ($D_{ij} = D_{ji}$) and reflexive ($D_{ii} = 0$). Note that we do not require the triangle inequality to be satisfied, therefore the dissimilarity can be non-metric. In short, the input is a symmetric hollow dissimilarity matrix, and the expected output is a low-dimensional embedding of the original dataset that approximately preserves pairwise distances.

The major challenge we encounter in this setting is two-fold. The first is to obtain *good* coordinates from a generic input dissimilarity matrix to represent the data, before we can even apply the JL transform. Second, we need a geometric characterization of the non-Euclidean non-metric setting, to show how different our setting stands from Euclidean geometry, which the JL Lemma is based on.

Before we explain our results, we first mention existing lower bounds for dimension reduction in non-Euclidean settings. JL Lemma considers Euclidean spaces only. Naturally, one asks whether such a result is possible for other spaces. There are several non-trivial dimension reduction results for ℓ_1 norm [11] and ℓ_p norm [12]. However, the target embedding dimension for ℓ_1, ℓ_p and nuclear norm is necessarily polynomial in n for constant distortion [13, 14, 15]. These lower bounds remind us that a worst-case guarantee on low distortion and logarithmic target dimension is not possible. Rather, our results provide error analysis which depends on parameters characterizing how the input data deviates from Euclidean geometry. In other words, we have fine-grained performance analysis.

Our Contributions. We present two approaches to recover coordinates that fit into the non-Euclidean non-metric setting. For both approaches, we generalize the Euclidean norm ℓ_2 to a new form to capture the input data geometry, together with certain geometric parameters indicating how far it deviates from Euclidean geometry. We give error analysis on the dissimilarity distortion, which may involve an additive term whose magnitude is proportional to the geometric parameters.

At the heart of our first approach is the observation that any symmetric hollow matrix can be written as the distance matrix of vectors in *pseudo-Euclidean space* [16, 17, 18]. Here the distance between two vectors $x = (x_1, x_2, ...x_n)$ and $y = (y_1, y_2, ...y_n)$ is captured by a bilinear form of signature (p, q), which is defined as $\langle x, y\rangle_{p,q} = \sum_{i=1}^{p} x_i y_i - \sum_{i=p+1}^{p+q} x_i y_i$. The squared (p, q)-distance between x and y is $\|x - y\|_{p,q}^2 = \langle x - y, x - y\rangle$. When $p = n, q = 0$, it is the squared Euclidean norm. We will give a more detailed introduction later. At this point, it suffices to interpret the parameter p resembling and q negating the Euclidean space. Our first result shows the generalization of the JL lemma to the pseudo-Euclidean geometry.

Theorem 1.2 (Fine-grained JL lemma, informal Version of Theorem 2.3)**.** *Given any symmetric hollow dissimilarity matrix D of size n, we are able to obtain embeddings X in the pseudo-Euclidean space. For any $\varepsilon \in (0, 1)$, there exists a JL transform to X with target dimension $O(\log n/\varepsilon^2)$, such that any pairwise dissimilarity D_{ij} is preserved with at most $1 \pm \varepsilon \cdot C_{ij}$ multiplicative factor, where C_{ij} is the ratio of the squared Euclidean distance and squared (p, q) distance.*

Our second result takes a different route. We prove that a symmetric hollow matrix is also a matrix of *generalized power distances*. Given two points x and y with respective radius r_x and r_y, the generalized power distance is defined as $\|x - y\|_E^2 - (r_x + r_y)^2$, where $\|x - y\|_E^2$ denotes the Euclidean norm. This is the squared length of the tangent line segments of two balls centered at x, y with radii r_x, r_y. Again, at this point, it suffices to interpret r_x, r_y as a measure of deviation from Euclidean space, with both a geometric meaning and a statistical interpretation of distance between points with uncertainties.

We show that any input symmetric hollow matrix D of size n can be written as the generalized power distance matrix of n points $\{p_i\}$ with the same radius $r = \sqrt{|e_n|}/2$, where e_n is the smallest eigenvalue of the Gram matrix of D. This linear algebra result may be of independent interest. Only when D is a Euclidean distance matrix, all eigenvalues are non-negative and $r = 0$. Our second result follows from applying the JL transform on the ball centers (i.e. $\{p_i\}$). We obtain a $(1 \pm \varepsilon)$ multiplicative approximation of the generalized power distance with an additive error of $4\varepsilon r^2$.

Theorem 1.3 (Power-distance JL Lemma, informal Version of Theorem 3.3)**.** *Given any symmetric hollow dissimilarity matrix D of size n and $\varepsilon \in (0, 1)$, there exists a JL transform with target dimension $m = O(\log n/\varepsilon^2)$, such that any pairwise distance D_{ij} is preserved with at most $1 \pm \varepsilon$ multiplicative factor and an $4\varepsilon r^2$ additive factor, where $r = \sqrt{|e_n|}/2$ with e_n as the smallest eigenvalue of the Gram matrix of D.*

To complement Theorem 1.3, we are able to extend the techniques in [19] and give a lower bound of $\Omega(\log n/\varepsilon^2)$ on the target dimension to achieve the multiplicative and additive factors as mentioned. Note that $m = \Omega(\log n/\varepsilon^2)$ matches the JL lemma lower bound.

Experiments. We implement both methods of JL transform with respect to the above results and evaluate their performances on 10 datasets. We observe that the experiment results corroborate with our theoretical results, and outperform classical JL transform consistently on non-Euclidean datasts. Our codes are on the Anonymous Github[2].

1.1 Related Work

Random Linear Projection and JL Lemma A nice survey on JL transform can be found in [20]. There was a series of work [21, 22, 23, 24] that studied whether the bound on the target embedding dimension in JL lemma is tight, starting from the original JL paper [1], and the optimality of JL lemma for Euclidean dimension reduction is finally established even among non-linear embeddings [19, 25]. In terms of algorithms, there have been developments of variants of random projections that are more friendly for implementations [26, 27, 28, 29, 30]. Empirical study of JL Lemma can be found in [31].

Last, Benjamini and Makarychev [32] considered dimension reduction using the Poincaré half-space model of hyperbolic space $\mathbb{H}^n$ where each point is represented by (x, z) with $x \in \mathbb{R}^{n-1}$ and height $z \in \mathbb{R}^+$. They considered using JL random projection f on the Euclidean part and obtain $(f(x), z)$ with $f(x)$ in Euclidean space of dimension $O(\log n/\varepsilon^2)$, achieving $1 + \varepsilon$ distortion.

Data-dependent dimension reduction. A number of dimension reduction techniques that are also popular in practice include Multidimensional scaling (MDS) [33] and Principal Component Analysis (PCA) [34]. These methods are *data-dependent* and choose linear projections based on the input data. There has also been recent work trying to apply these methods on non-Euclidean data, for example PCA in hyperbolic or spherical geometry [35, 36, 37, 38]. Manifold learning (such as Isomap [39] and LLE [40]) considers input data points that come from a low-dimensional manifold, recovers the distances between points along the manifold and applies existing dimension reduction methods. A recent work called Non-Euclidean MDS [18] considers dimension reduction of pseudo-Euclidean settings and minimizes the loss called STRESS (the sum of squared difference of pairwise embedding distances to the input dissimilarities). It does not provide error analysis or guarantees for individual dissimilarity distortion.

2 Johnson-Lindenstrauss Lemma in Pseudo-Euclidean space

2.1 Pseudo-Euclidean space of signature (p, q)

We consider the d-dimensional vector space $\mathbb{R}^d$ equipped with a (p, q) bilinear form $\langle , \rangle$ where $p + q = d$. For simplicity, we use $\mathbb{R}^{p,q}$ to denote $\mathbb{R}^d$ with the (p, q) bilinear form $\langle , \rangle$. Consider two d-dimensional (column) vectors $u, v \in \mathbb{R}^{p,q}$, with $u = (u_1, u_2, ...u_{p+q})$ and $v = (v_1, v_2, ...v_{p+q})$, we define the bilinear norm as:

$$\langle u, v\rangle = \sum_{i=1}^{p} u_i v_i - \sum_{i=p+1}^{p+q} u_i v_i$$

[2] `https://anonymous.4open.science/r/Non-Euclidean-Johnson-Lindenstrauss-1673`

Equivalently, we can write $\langle u, v\rangle = u^T \Lambda v$ with Λ as a $d \times d$ diagonal matrix[3] with the first p diagonal elements as 1 and the remaining diagonal elements as -1. For a vector u, its *scalar square* is $\langle u, u\rangle$. For two vectors u, v their *interval square* is obtained by $\langle u - v, u - v\rangle$.

Consider a given symmetric dissimilarity matrix $D \in \mathbb{R}^{n\times n}$, where D_{ij} refers to the dissimilarity measure of element i and element j. We assume that D is hallow, $D_{ii} = 0$, and symmetric, $D_{ij} = D_{ji}$. The following proposition shows that a set of vectors in the Pseudo-Euclidean space of signature (p, q) can be obtained via D.

Proposition 2.1. *For any hollow symmetric matrix D, there is a symmetric bilinear form $\langle , \rangle_{p,q}$ for integers p, q and n vectors $x_1, \ldots, x_n$ such that $D_{ij} = \langle x_i - x_j, x_i - x_j\rangle$.*

We now explain how to achieve this. With an input D, we first perform centralization to obtain the Gram matrix $B = \mathrm{Gram}(D) = -CDC/2$, where $C = I - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is the centering matrix and $\mathbf{1}_n$ is a vector of ones. Since $-CDC/2$ is a symmetric matrix, its eigenvalues are real, denoted as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. We take p as the number of non-negative eigenvalues and q to be the number of negative eigenvalues. $p + q = n$. Further, suppose $U \in \mathbb{R}^{n\times n}$ is the orthogonal vectors, we have

$$B = -CDC/2 = U\,\mathrm{Diag}(\lambda_1, \cdots, \lambda_n)U^T$$

Now we can recover the coordinates of the n elements as n-dimensional vectors. Specifically X is a matrix of dimension $n \times n$ with the columns representing the coordinate vectors of the n points.

$$X = (x_1, \cdots, x_n) = \mathrm{Diag}(\sqrt{|\lambda_1|}, \cdots, \sqrt{|\lambda_n|}) \cdot U^T$$

This way, we have $B = -CDC/2 = X^T \Lambda X$, with Λ as an $n \times n$ diagonal matrix with the first p diagonal elements as either 1 (or 0) if the corresponding eigenvalue is positive (or 0), and the remaining diagonal elements as -1. Equivalently, we have

$$D_{ij} = (x_i - x_j)^T \Lambda (x_i - x_j) = \langle x_i - x_j, x_i - x_j\rangle.$$

That is, the dissimilarity D_{ij} is precisely the (p, q) interval square of x_i, x_j.

If the input dissimilarity matrix D is the squared Euclidean distance matrix of n points, the Gram matrix is positive semi-definite and the above procedure would recover the Euclidean coordinates $\{x_i\}$. This is referred to as the classical Multidimensional scaling (MDS) [33]. If the input dissimilarity matrix D generalizes beyond the squared Euclidean distance matrix but remains a hallow symmetric matrix, we can still recover the coordinates so that they produce the entries in D using a general (p, q) bilinear form. Notice that the above procedure generates coordinate vectors in dimension n.

The (p, q)-space has deep connections to spacetime and special relativity theory. Specifically, the Lorentzian n-space is the vector space $\mathbb{R}^n$ with the n-dimensional Lorentzian inner product, where the squared norm of a vector $u = (u_0, u_1, \cdots, u_{n-1})$ has the form $|u| = -u_0^2 + u_1^2 + \cdots + u_{n-1}^2$. Thus the Loerentizan space has signature $(n - 1, 1)$. Four-dimensional Lorentzian space is called the *Minkowski space* and forms the basis of special relativity. Furthermore, the collection of vectors in $\mathbb{R}^{n+1}$ with Lorentzian inner product of -1 has imaginary Lorentzian length and is precisely the hyperboloid model of the hyperbolic n-space $\mathbb{H}^n$ [41].

2.2 Johnson-Lindenstrauss Lemma in (p, q)-space

We start with some notation. For a vector $v = (v_1, v_2, \cdots, v_{p+q}) \in \mathbb{R}^{p,q}$, we will use $\|v\|_{p,q}^2$ to denote the "(p, q)-norm" $\langle v, v\rangle$ and $\|v\|_E^2$ to denote the squared Euclidean norm of v. Further, we denote $v^{(p)} = (v_1, v_2 \cdots v_p) \in \mathbb{R}^p$ and $v^{(q)} = (v_{p+1}, v_{p+2}, \cdots, v_{p+q}) \in \mathbb{R}^q$. Then $\|v\|_{p,q}^2 = \|v^{(p)}\|_E^2 - \|v^{(q)}\|_E^2$ and $\|v\|_E^2 = \|v^{(p)}\|_E^2 + \|v^{(q)}\|_E^2$.

First, we will prove a Johnson-Lindenstrauss style dimension reduction by applying the JL-lemma on the p part and q part respectively. The proofs in this section can be found in Appendix B.

Lemma 2.2. *For any set of n points $x_1, x_2, \ldots x_n$ in $\mathbb{R}^{p,q}$ and $\varepsilon \in (0, 1)$, there exists a map $f : \mathbb{R}^{p,q} \to \mathbb{R}^{p',q'}$, where $p', q' = O(\log n/\varepsilon^2)$ such that for any $i, j \in [n]$,*

$$\|x_i - x_j\|_{p,q}^2 - \varepsilon\|x_i - x_j\|_E^2 \leq \|f(x_i) - f(x_j)\|_{p',q'}^2 \leq \|x_i - x_j\|_{p,q}^2 + \varepsilon\|x_i - x_j\|_E^2 \quad (2)$$

[3] In general, for the bilinear form Λ, there can also be additional r dimensions with zero as diagonal elements, in addition to p elements of 1 and q elements of -1. In that case $p + q + r = d$. In our writing we skipped r as they do not contribute to the similarity values for ease of exposition.

Essentially for (p, q)-space we can still have a similar JL-lemma, which, compared with the standard JL-lemma Eq (1), we do not have the $1 + \varepsilon$ multiplicative error but rather an additive error to the squared Euclidean norm. This gives us an immediate JL-style result when $\|x_i - x_j\|_{p,q}^2$ and $\|x_i - x_j\|_E^2$ are a constant factor of each other. Formally, we have the following statement.

Theorem 2.3. *For any set of n points $x_1, x_2, \ldots x_n$ in $\mathbb{R}^{p,q}$ and $\varepsilon \in (0, 1)$, there exists a map $f : \mathbb{R}^{p,q} \to \mathbb{R}^{p',q'}$, where $p', q' = O(\log n/\varepsilon^2)$ such that for any $i, j \in [n]$,*

$$(1 - \varepsilon \cdot C_{ij})\|x_i - x_j\|_{p,q}^2 \leq \|f(x_i) - f(x_j)\|_{p',q'}^2 \leq (1 + \varepsilon \cdot C_{ij})\|x_i - x_j\|_{p,q}^2, \tag{3}$$

where $C_{ij} = \left|\frac{\|x_i - x_j\|_E^2}{\|x_i - x_j\|_{p,q}^2}\right|$.

A few remarks are in place. First, one can apply any JL-style algorithm for the p and q parts, for example, by using the fast JL algorithm. Suppose f_p is implemented by a random matrix $M_{p'p}$ and f_q is implemented by a random matrix $M_{q'q}$. Then f is implemented by a matrix M of dimension $(p' + q') \times (p + q)$ with M_p and M_q at the diagonal and zero elsewhere, $f(x) = Mx$. Second, random projection for the (p, q) space setting provides an error-bound guarantee for a pair of points that gracefully degrades as the (p, q) norm deviates from the Euclidean norm. Below we show a few natural situations where the (p, q) norm and Euclidean norm are indeed bounded.

Lemma 2.4. *Let $v \in \mathbb{R}^{p,q}$. If each coordinate of v is selected from the same distribution, f with a finite second moment, then we have with high probability, $\|v\|_E^2 < C\|v\|_{p,q}^2$ if $q < \frac{C-1}{C+1}p$ or $\|v\|_E^2 < -C\|v\|_{p,q}^2$ if $p < \frac{C-1}{C+1}q$ for $C > 1$.*

Corollary 2.5. *A random vector v chosen uniformly over the unit sphere $\mathbb{S}^{p+q-1}$ has with high probability, $\|v\|_E^2 < C\|v\|_{p,q}^2$ if $q < \frac{C-1}{C+1}p$.*

Now we put together all of the previous results into a single statement on a set of random points in (p, q) space.

Theorem 2.6. *Let $X \subset \mathbb{R}^{p,q}$ be a randomly selected point set of size n, with $q < \frac{C-1}{C+1}p$ for a constant C. For all $u, v \in X$, $\frac{u-v}{\|u-v\|_E}$ has distribution equal to uniformly choosing a vector on S^{p+q-1}. Then, with high probability, there exists a map $f : \mathbb{R}^{p,q} \to \mathbb{R}^{p',q'}$, $p', q' = O(\frac{C^2 \log(n)}{\epsilon^2})$ such that for any $i, j \in [n]$,*

$$(1 - \epsilon)\|x_i - x_j\|_{p,q}^2 \leq \|f(x_i) - f(x_j)\|_{p',q'}^2 \leq (1 + \epsilon)\|x_i - x_j\|_{p,q}^2. \tag{4}$$

3 Johnson-Lindenstrauss Lemma for Generalized Power Distances

3.1 Generalized Power Distance

Given two balls centered at $p, q \in \mathbb{R}^d$, with radii r_p, r_q, we define the *generalized power distance* as:

$$\text{Pow}((p, r_p), (q, r_q)) = \|p - q\|_E^2 - (r_p + r_q)^2. \tag{5}$$

This power distance measures the distance of the internal tangents between two disjoint circles (the tangent line that keeps two circles on different sides. See the middle picture in Figure 1). Note that there is another second generalized power distance between two circles centered at p_i of radius r_i given by $||p_1 - p_2||_E^2 - (r_1 - r_2)^2$, which measures the distance of the external tangents between the circles (the tangent line that keeps two circles on the same side).

Geometric Interpretation. In elementary plane geometry, the power of a point is a real number that reflects the relative distance of a given point from a given circle, introduced by Jakob Steiner in 1826. The power of a point p with respect to a circle with center q and radius r is defined as $\|p - q\|^2 - r^2$. This value is positive when p is outside the circle, and by Pythagorean theorem it is squared tangential distance $|pt|$ where t is a tangent point of p to the circle. When p is inside the circle, the power is negative, and precisely the negative of squared distance $|ps|$ where s is a point on the circle with $\triangle psq$ to be a right triangle with $\angle spq$ as 90 degree.

The generalized power distance considers the distance between two circles and reduces to the classical power distance if one of the circles is a point. Furthermore, two circles are disjoint if and only if their generalized power distance is positive. In this case, their generalized power distance is equal to the

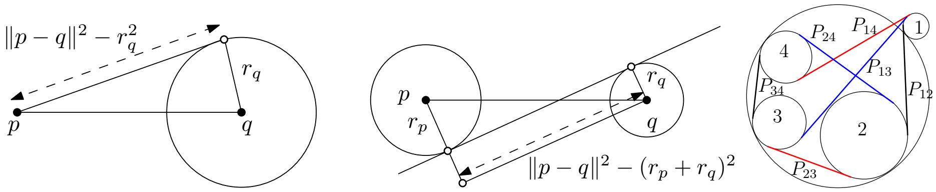


Figure 1: Left: power distance from a point p to a ball at q of radius r_q; Middle: power distance between two balls at points p, q with radius r_p and r_q respectively; Right: Casey's Theorem.

square of the distance between the two tangent points on the common internal tangent line between the two circles (with the two circles on different sides). See Figure 1 for examples. The generalized power distance can be called power distance of *weighted points* with r_p as the weight of p.

There are several interesting properties of the generalized power distance. First, the Casey's theorem [42, 43] still holds for generalized power distance. That is, if four circles in the plane are tangent to the fifth circle, then the generalized power distances P_{ij}, $i, j \in \{1, 2, 3, 4\}$, between them satisfy the Ptolemy identity $P_{12}P_{34} + P_{23}P_{41} = P_{13}P_{24}$. This Ptolemy identity plays a key role in applications of the generalized power distances. Indeed, the Ptolemy identity and the associated Ptolemy inequality $P_{12}P_{34} + P_{23}P_{41} \geq P_{13}P_{24}$ have been used in optimization and approximation for prune search space and optimize geometric networks, for approximate distance estimate, circle fitting and cyclic polygon detection; in distance geometry for ensuring distance constraints for sensor networks [44] and molecule reconstruction [45]. In metric geometry [46], a metric space is called Ptolemaic space if Ptolemy inequality holds for any quadriple points. It is well known that Euclidean space and any inner-product spaces are Ptolemaic and a normed vector space is Ptolemaic if and only if the norm comes from an inner product [47, 48]. Furthermore, a Riemannian manifold is a Ptolemaic space if and only if its sectional curvature is non-positive [49]. It shows that Ptolemy inequalities can be used for studying curvature properties of metric spaces. Recall that a length space (X, d) is Gromov δ-hyperbolic if it satisfies Gromov's Four-Point Condition: for any $x, y, z, w \in X$, $d(x, z) + d(y, w) \leq \max\{d(x, y) + d(z, w), d(x, w) + d(y, z)\} + 2\delta$. The Ptolemy inequality is another four-point condition on metric spaces for investigating curvature of general spaces [50].

Any Symmetric Dissimilarities are Power Distances. Now, we show that any symmetric hollow dissimilarity matrix can be written as the matrix of power distances between n weighted points.

Lemma 3.1. *Given any $n \times n$ symmetric hollow dissimilarity matrix, D, we can rewrite $D = E + 4r^2(I - J)$ where $r \in \mathbb{R}^+$, E is a Euclidean distance matrix, I is an $n \times n$ identity matrix, $J = \mathbf{1}_n \mathbf{1}_n^T$. More specifically E is a Euclidean distance matrix if and only if $2r^2 \geq |e_n|$ where e_n is the least eigenvalue of* $\mathrm{Gram}(D)$.

From Theorem 3.1, suppose E is an Euclidean distance matrix where the (i, j) element is $\|p_i - p_j\|^2$, with n points $\{p_i\}$. Then D is the matrix of generalized power distances $\{(p_i, r)\}$, i.e., with the (i, j) element as $\|p_i - p_j\|_E^2 - 4r^2$. Here all the weighted points have radius r which can be seen as a measure of how far the input matrix D is away from being Euclidean. When D is a Euclidean distance matrix, the Gram matrix $\mathrm{Gram}(D)$ is positive semi-definite and the smallest eigenvalue is $e_n = 0$. Thus $r = 0$ and the generalized power distance reduces to the squared Euclidean distance. When D is no longer a Euclidean distance matrix, $\mathrm{Gram}(D)$ necessarily have negative eigenvalues, i.e., $e_n < 0$. This makes r to be non-trivial.

Statistical Interpretation The generalized power distance also has a statistical interpretation using the silhouette coefficient [51], which measures how similar two clusters are. Consider two clusters X, Y, the *silhouette value* of a point $x \in X$ is the average distance from x to points in Y minus the average distance from x to other points in X. The *silhouette coefficient* of the two clusters X, Y is the average of the silhouette coefficient of all points in X, Y. In Appendix A we show that the silhouette coefficient of two Gaussian distributions $\mathcal{X} = \mathcal{N}(\mu_x, \sigma_x), \mathcal{Y} = \mathcal{N}(\mu_y, \sigma_y)$ is $SC_W(\mathcal{X}, \mathcal{Y}) = \|\mu_x - \mu_y\|^2 - (\sigma_x + \sigma_y)^2$. Notice that this is precisely the generalized power distance of weighted points (μ_x, σ_x) and (μ_y, σ_y).

Therefore, Theorem 3.1 suggests that any input dissimilarity matrix D that is hollow and symmetric is actually the silhouette coefficients of n Gaussian distributions centered at $\{p_i\}$ with variance r which is given by Theorem 3.1. Note that this statistical interpretation resonates with prior literature [52, 53] which identified that one major source of data in non-Euclidean geometry is due to measurement noises and uncertainties.

3.2 Johnson-Lindenstrauss Lemma for Power Distance

Lemma 3.2. *Given n weighted points (p_i, r_i) with $p_i \in \mathbb{R}^d$, there exists a map $f : \mathbb{R}^d \to \mathbb{R}^m$, where $m = O(\log n/\varepsilon^2)$ such that for any $i \neq j \in [n]$,*

$$(1-\varepsilon)\,\mathrm{Pow}((p_i, r_i), (p_j, r_j)) - \varepsilon(r_i + r_j)^2 \leq \mathrm{Pow}((f(p_i), r_i), (f(p_j), r_j)) \tag{6}$$

$$\leq (1+\varepsilon)\,\mathrm{Pow}((p_i, r_i), (p_j, r_j)) + \varepsilon(r_i + r_j)^2. \tag{7}$$

With Lemma 3.1 and Lemma 3.2 we immediately have the following.

Theorem 3.3. *Given any $n \times n$ symmetric hollow dissimilarity matrix, D, with power distance representation by $\{(p_i, r)\}$ where $r = \sqrt{|e_n|}/2$ and e_n is the smallest eigenvalue of $\mathrm{Gram}(D) = -\frac{1}{2}CDC$ with $C = 1 - J/n$ and J is an all 1 matrix, there exists a map $f : \mathbb{R}^n \to \mathbb{R}^m$, where $m = O(\log n/\varepsilon^2)$ such that for any $i \neq j \in [n]$,*

$$\begin{aligned}(1-\varepsilon)\,\mathrm{Pow}((p_i, r), (p_j, r)) - \varepsilon 4r^2 &\leq \mathrm{Pow}((f(p_i), r), (f(p_j), r)) \\ &\leq (1+\varepsilon)\,\mathrm{Pow}((p_i, r), (p_j, r)) + \varepsilon 4r^2.\end{aligned}$$

Last, we remark that the error bounds in Lemma 3.2 also imply a $1+\varepsilon$ distortion for dimension reduction on the Euclidean distances of the centers (as in the JL Lemma 1). Therefore, the lower bound on the target dimension being $\Omega(\log n/\varepsilon^2)$ for a dimension reduction algorithm satisfying the error bounds in Lemma 3.2 for generalized power distances holds, by the same argument in the proof of dimension optimality of standard JL-lemma [19].

4 Algorithms for Non-Euclidean Johnson-Lindenstrauss Transforms

The algorithms used for the experiments are presented as follows. In both cases we assume an input dissimilarity matrix D which is symmetric ($D_{ij} = D_{ji}$) and hollow ($D_{ii} = 0$).

Pseudo-Euclidean JL Transform

1. Find the orthogonal decomposition of the dissimilarity matrix $D = O\Lambda O^T$ where Λ is a diagonal matrix of the eigenvalues.
2. Let A be the (p, q) signature, i.e. $A = \mathrm{sign}(\Lambda)$
3. Compute the the embedding into $\mathbb{R}^{p,q}$ by finding VAV^T.
4. For each point x in the embedding, split it into $x^{(p)}$ and $x^{(q)}$. Use standard JL transform to project into $\mathbb{R}^{p'}$ and $\mathbb{R}^{q'}$ respectively where p' and q' are the specified target dimension.
5. Return the projected points $(x^{(p')}, x^{(q')})$ along with their (p', q') signature.

Power distance JL Transform

1. Find the orthogonal decomposition of the dissimilarity matrix $O\Lambda O^T$
2. Compute e_n, the smallest eigenvalue of $\mathrm{Gram}(D)$. Set $r = \sqrt{|e_n|}/2$.
3. Add $4r^2(J - I)$ to D to obtain the new Euclidean distance matrix, E.
4. Recover the Euclidean coordinates X' such that $E_{ij} = \|x'_i - x'_j\|^2$. Perform standard JL transform on X' to target dimension m.
5. Return the projected points along with their radii r.

Computational efficiency. Since both methods start with a dissimilarity matrix we necessarily need to recover the 'coordinates' first. This can be done by running the singular value decomposition

(SVD) on the Gram matrix in $O(n^3)$ time. There are several possible ways to speed up this step in practice – by using landmark MDS for example. Another approach, motivated by the generalized power distance formulation, is to consider appending a proper term $4r^2$ to the input matrix D and later use the power distance with the recovered coordinates. If it happens that the coordinates are to be learned from a representation learning module, we can skip this part and directly apply dimension reduction. When applying standard JL transform we use random projection of Gaussian vectors. One can use any JL transform for example [26, 27, 28, 29, 30] for this purpose.

5 Experiments

In this section, we present experimental results of the proposed JL transforms in non-Euclidean settings with only a dissimilarity matrix as input. First, we validate our theoretical results by showing the approximation error on datasets that are highly non-Euclidean. Next, we evaluate the algorithms on real-world datasets, with the classical JL transform as a baseline. Finally, we test the proposed JL transforms on a downstream task, k-means clustering, to show its potential wide applications.

Datasets. We use two **synthetic datasets** that are made non-Euclidean: Random-simplex and Euclidean-ball. At a high level, for the Random-simplex, given a dataset of size n, each point is constructed such that its first $n-1$ coordinates form a simplex, while the final coordinate dominates the pairwise distances. This design induces a large negative eigenvalue in the Gram matrix. The Euclidean-ball dataset, inspired by Delft's balls [53], consists of n balls with varying radii. The distance between two balls is defined as the minimal distance between any two points on their surfaces, resulting in dissimilarities that violate the triangle inequality. For **real-world data**, We consider three categories: genomics, image and graph data. The genomics data includes three cancer-related datasets from the Curated Microarray Database (CuMiDa) [54]. Following the practice in prior work [18], we obtain dissimilarities with entropic affinity. We also test two celebrated image datasets: MNIST and CIFAR-10, each with 1000 images randomly sampled. We use the measures mentioned in [55] to calculate the dissimilarities. The graph datasets are selected from the SNAP project [56]. The pairwise distances are shortest path distances. The basic statistics of all datasets are shown in Table 1. We defer more details in Appendix C.

Dataset	Simplex	Ball	Brain	Breast	Renal	MNIST	CIFAR10	Email	Facebook	Mooc
Size	1000	1000	130	151	143	1000	1000	986	4039	7047
# $\{\lambda < 0\}$	900	887	53	59	57	454	399	465	1566	268
Metric	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓

Table 1: Non-Euclidean/Non-metric Datasets used in experiments

5.1 Validation of Theoretical Results

Pseudo-Euclidean JL Transform. We start with demonstrating the multiplicative factor of pseudo-Euclidean projection indeed falls into the range of $(1 \pm \varepsilon \cdot C_{ij})$, as a corroboration of Theorem 2.3. Throughout, we set $\varepsilon = 0.5$ and the constant in the target dimension $O(\log n/\varepsilon^2))$ as 2. Figure 2 verifies the approximation ratio on the Simplex and Brain datasets. We observe that most pairwise dissimilarities behave as claimed. For the Brain dataset, even the bar range is very small we still have all green points. There are minor exceptions indicated by the red points in the Simplex dataset. However, first observe that they are still very close to the error bars. Further, we did not optimize the constant factor in $O(\log n/\varepsilon^2))$. In fact, if we double the target dimension (from 80 to 160), the percentage of violations drops significantly from 12.01% to 4.62% for the Simplex dataset. We defer the plots of other datasets to Appendix C.

Power Distance JL Transform. To validate Theorem 3.3, we show the residual additive error of the power distance JL transform. By 'residual' we refer to the absolute difference between the dissimilarities given by the JL transform and the $(1 \pm \varepsilon)$ approximation of the true dissimilarities. This term should be smaller than $4\varepsilon r^2$. As shown in Figure 3, all sampled points appear below the red line, which is taken as $4\varepsilon(r/100)^2$, a much tighter upper bound. Recall that a larger r indicates more deviation from Euclidean geometry, the fact that the residual error is small shows the effectiveness of our JL transform. We defer the plots of other datasets to Appendix C.

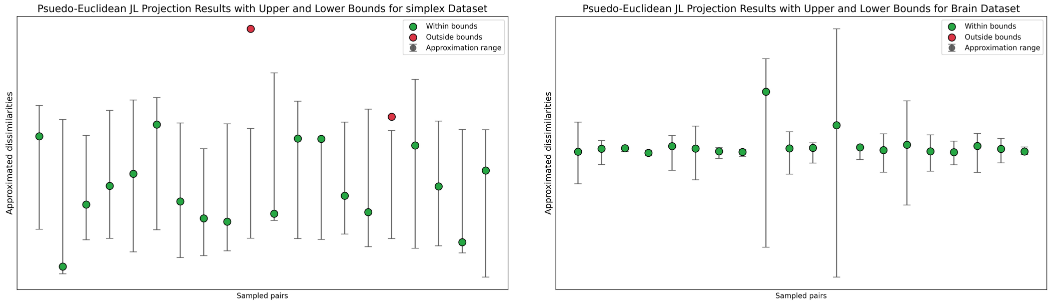


Figure 2: The left is for Simplex dataset and right for the Brain dataset. Green points indicate the approximation ratio is within the range and red the opposite. The error bars match the upper and lower bounds given by $(1 \pm \varepsilon \cdot C_{ij})$. We sample 20 pairs of dissimilarities for presentation.

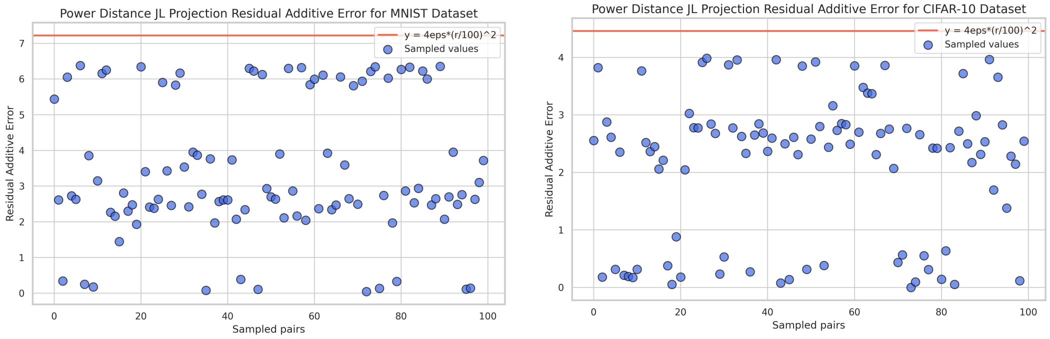


Figure 3: The left is for MNIST and right for CIFAR-10 dataset. We sampled 100 pairs of dissimilarities and adopt $r/100$ for presentation. Using the original r makes the red line at roughly $y = 40000$, which is very loose.

5.2 Performance of JL Transforms

Performance on Relative Error. We compare two proposed JL transforms with the classical JL transform on all datasets and report the relative error, which is defined as the maximum among $\frac{|D_{ij} - \hat{D}_{ij}|}{D_{ij}}$ for all (i, j) pairs. $\hat{D}_{ij}$ is the dissimilarity matrix obtained from any JL transform. It is a suitable metric because all three algorithms have different theoretical guarantees, but the end goal is always preserving pairwise dissimilarities. In Table 2 we show the worst-case (defined above) average and median relative error. A smaller value indicates better performance.

We observe JL-PE performs the best for genomics data and JL-power performs the best for the rest, on both metrics. The significant improvement implies when the JL transform matches with geometry, we can expect really good results. The image datasets report *inf* for JL because different images are projected to the same point. In a few cases Jl-PE has slightly worse relative error than JL, but note that this might be the outcome of really large factor C, which the performance of JL-PE is based on.

Method	Simplex	Ball	Brain	Breast	Renal	MNIST	CIFAR10	Email	Facebook	MOOC
JL Max	6.47e5	5.97e5	1.02e12	7.09e10	3.42e12	inf	inf	3.85e4	1.18e6	2.94e5
JL-PE Max	1.11e6	5.31e5	**8,21e4**	**262.20**	**2.37e4**	8.47e5	2.24e6	3.24e6	2.51e7	1.35e7
JL-Power Max	**109.18**	**12.102**	3.11e7	1.92e6	1.17e8	**85.76**	**55.74**	**781.45**	**82.74**	**24.22**
JL Ave	442.65	43.78	1.11e9	9.86e7	1.50e9	inf	inf	15.64	12.62	39.83
JL-PE Ave	47.59	23.71	**8.16**	**1.15**	**6.60**	15.93	18.24	13.79	52.52	63.44
JL-Power Ave	**13.52**	**1.002**	3.73e4	3.12e3	5.57e4	**1.47**	**1.61**	**1.60**	**1.32**	**2.04**
JL Median	6.36	6.68	2.69e12	1.95e11	1.18e12	1.78	2.29	7.00	5.08	6.13
JL-PE Median	**2.57**	2.78	**1284.06**	**174.74**	**3160.58**	2.11	2.04	2.06	5.83	6.03
JL-Power Median	12.79	**1.00**	1.01e7	8.20e6	7.57e7	**1.30**	**1.35**	**1.35**	**1.16**	**1.99**

Table 2: Max and Average Relative Error of All Datasets on Three JL Transforms.

Performance on Downstream k-Means Clustering. We now evaluate if the propsoed JL transforms consistently perform well on downstream tasks. The setup is straightforward: we first apply JL transform to reduce the dimension of the original data, then we compare k-Means clustering cost obtained from the processed data with the original optimal clustering cost.

Dataset	#Clusters (k)	Original	JL	JL-pq	JL-power
Brain	5	3.92×10^{-3}	728.00	**3.13**	33.26
Breast	4	0.03	827.69	28.93	**12.53**
Email	42	5614	22634	21244	**20764**
Facebook	15	20499	284142	**91644**	303203

Table 3: Performance of Three JL Transforms with k-Means Clustering.

Table 3 demonstrates that both of the proposed JL transforms outperforms the classical JL, with only one exception on Facebook where JL-power is slightly worse than JL. This observation shows better effectiveness of the proposed JL transforms in downstream tasks. However, similar as above, the relative performance varies between JL-pq and JL-power. This raises a key direction for future research: formalize the conditions under which a specific JL transform is most suitable.

6 Discussions

The Pseudo-Euclidean space we study shares conceptual foundations with recent advances in similarity learning. Notably, [57] proposed neural similarity learning (NSL) which uses bilinear matrices to generalize inner product similarity in convolutional neural networks through both static and dynamic approaches. The bilinear form structures also appear in modern neural architectures [58]. In the scaled dot-product attention mechanism used in transformers, the attention score between two tokens is computed as a bilinear form: $\text{score}(q, k) = q^{\top}Wk$, where q and k are query and key vectors and W is a learned weight matrix. This mechanism effectively learns a similarity function over the token space, which aligns with our use of bilinear forms to represent pairwise dissimilarities in non-Euclidean geometry.

Future Direction. A promising future direction is to integrate the mathematical formulation in this paper using both the pseudo-Euclidean representation and the generalized power distance representation with current machine learning pipelines for representation and similarity learning. Recently there has been increasing interest in representation learning in non-Euclidean spaces, e.g., learning modules in hyperbolic [59, 60, 61, 62, 63, 64, 65], spherical or mixed-curvature product manifolds (Cartesian products of hyperbolic, hyperspherical and Euclidean manifolds) [35, 66, 67, 68] or general manifolds [69]. Many of the prior work assume (piecewise) constant curvature in the embedded space and our setting includes these spaces and beyond. One limitation of current work is we are not able to build a theoretical connection between two proposed representations, or conditions that one outperforms the other. Further, we are not aware of any lower bound result complementing the fine-grained JL guarantee from the pseudo-Euclidean representation.

Conclusion. We explore Johnson-Lindenstrauss Transform in the non-Euclidean setting with dissimilarity matrix as the input. Via two different approaches, pseudo-Euclidean geometry and generalized power distance, we show two theoretical results with similar flavor to the classical JL lemma. The first gives a fine-grained error analysis and the second has an extra additive error term, both having parameters indicating the deviation from Euclidean geometry. The experiment results corroborate with our theoretical results, and demonstrate the superior performance of two proposed approaches for Jl-type dimension reduction with non-Euclidean data.

7 Acknowledgements

The authors would like to acknowledge funding from IIS-22298766, DMS-2220271, DMS-2311064, CRCNS-2207440, CCF-2118953 and CNS-2515159. The authors would like to thank Hongbin Sun for helpful discussions.

References

[1] William B Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[2] Piotr Indyk and Assaf Naor. Nearest-neighbor-preserving embeddings. *ACM Trans. Algorithms*, 3(3):31–es, August 2007.

[3] Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Trans. Inf. Theory*, 61(9):5096–5115, September 2015.

[4] Konstantin Makarychev, Yury Makarychev, and Ilya P. Razenshteyn. Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC*, pages 1027–1038, 2019.

[5] Zachary Izzo, Sandeep Silwal, and Samson Zhou. Dimensionality reduction for Wasserstein barycenter. *Advances in neural information processing systems*, 34:15582–15594, 2021.

[6] Shyam Narayanan, Sandeep Silwal, Piotr Indyk, and Or Zamir. Randomized dimensionality reduction for facility location and single-linkage clustering. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 7948–7957. PMLR, 2021.

[7] Luca Becchetti, Marc Bury, Vincent Cohen-Addad, Fabrizio Grandoni, and Chris Schwiegelshohn. Oblivious dimension reduction for k-means: beyond subspaces and the Johnson-Lindenstrauss lemma. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, page 1039–1050, New York, NY, USA, 2019. Association for Computing Machinery.

[8] William B Johnson and Assaf Naor. The Johnson-Lindenstrauss lemma almost characterizes Hilbert space, but not quite. *Discrete & Computational Geometry*, 43(3):542–553, 2010.

[9] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive approximation*, 28:253–263, 2008.

[10] Amos Tversky and Itamar Gati. Similarity, separability, and the triangle inequality. *Psychol. Rev.*, 89(2):123–154, March 1982.

[11] Jiří Matoušek. On variants of the Johnson–Lindenstrauss lemma. *Random Struct. Algorithms*, 33(2):142–156, September 2008.

[12] James R Lee, Manor Mendel, and Assaf Naor. Metric structures in ℓ_1: dimension, snowflakes, and average distortion. *Eur. J. Comb.*, 26(8):1180–1190, November 2005.

[13] Gideon Schechtman. Dimension reduction in l_p, $0 < p < 2$. *arXiv [math.MG]*, October 2011.

[14] Assaf Naor, Gilles Pisier, and Gideon Schechtman. Impossibility of dimension reduction in the nuclear norm. *Discrete Comput. Geom.*, 63(2):319–345, March 2020.

[15] Bo Brinkman and Moses Charikar. On the impossibility of dimension reduction in ℓ_1. *J. ACM*, 52(5):766–788, September 2005.

[16] Lev Goldfarb. A unified approach to pattern recognition. *Pattern Recognition*, 17(5):575–582, 1984.

[17] Elzbieta Pekalska and Robert P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence)*. World Scientific Publishing Co., Inc., USA, 2005.

[18] Chengyuan Deng, Jie Gao, Kevin Lu, Feng Luo, Hongbin Sun, and Cheng Xin. Neuc-MDS: Non-Euclidean multidimensional scaling through bilinear forms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, volume 37, pages 121539–121569, December 2024.

[19] Kasper Green Larsen and Jelani Nelson. Optimality of the Johnson-Lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638, 2017.

[20] Jelani Nelson. Dimensionality reduction in Euclidean space. *Not. Am. Math. Soc.*, 67(10):1, November 2020.

[21] Noga Alon. Problems and results in extremal combinatorics—i. *Discrete Math.*, 273(1-3):31–53, December 2003.

[22] Daniel Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit Johnson-Lindenstrauss families. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, Lecture notes in computer science, pages 628–639. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[23] T S Jayram and David P Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Trans. Algorithms*, 9(3):1–17, June 2013.

[24] Kasper Green Larsen and Jelani Nelson. The Johnson-Lindenstrauss Lemma is optimal for linear dimensionality reduction. In Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi, editors, *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*, volume 55 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 82:1–82:11, Dagstuhl, Germany, 2016. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[25] Noga Alon and Bo'az Klartag. Optimal compression of approximate inner products and dimension reduction. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 639–650. IEEE, October 2017.

[26] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, June 2003.

[27] Nir Ailon and Bernard Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, January 2009.

[28] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlos. A sparse Johnson-Lindenstrauss transform. STOC '10, page 341–350, New York, NY, USA, 2010. Association for Computing Machinery.

[29] Rong Yin, Yong Liu, Weiping Wang, and Dan Meng. Extremely sparse Johnson-Lindenstrauss transform: From theory to algorithm. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1376–1381. IEEE, 2020.

[30] Daniel M Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. *J. ACM*, 61(1):1–23, January 2014.

[31] Suresh Venkatasubramanian and Qiushi Wang. The Johnson-Lindenstrauss transform: An empirical study. In *2011 Proceedings of the Thirteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 164–173. Society for Industrial and Applied Mathematics, Philadelphia, PA, January 2011.

[32] Itai Benjamini and Yury Makarychev. Dimension reduction for hyperbolic space. *Proc. Am. Math. Soc.*, 137(02):695–698, September 2008.

[33] Warren S Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419, December 1952.

[34] Karl Pearson. LIII. on lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, 2(11):559–572, November 1901.

[35] Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning mixed-curvature representations in product spaces. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[36] Ines Chami, Albert Gu, Dat P Nguyen, and Christopher Re. HoroPCA: Hyperbolic dimensionality reduction via horospherical projections. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1419–1429. PMLR, 18–24 Jul 2021.

[37] Puoya Tabaghi, Michael Khanzadeh, Yusu Wang, and Siavash Mirarab. Principal component analysis in space forms. *IEEE Trans. Signal Process.*, 72:4428–4443, 2024.

[38] Xiran Fan, Chun-Hao Yang, and Baba C Vemuri. Nested hyperbolic spaces for dimensionality reduction and hyperbolic NN design. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2022:356–365, June 2022.

[39] Joshua B Tenenbaum. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[40] Sam T Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[41] John G Ratcliffe. *Foundations of Hyperbolic Manifolds*. Springer International Publishing, 2006.

[42] John Casey. On the equations and properties: (1) of the system of circles touching three circles in a plane; (2) of the system of spheres touching four spheres in space; (3) of the system of circles touching three circles on a sphere; (4) of the system of conics inscribed to a conic, and touching three inscribed conics in a plane. *Proceedings of the Royal Irish Academy (1836-1869)*, 9:396–423, 1864.

[43] Arnold Emch and Julian Lowell Coolidge. A treatise on the circle and the sphere. *Am. Math. Mon.*, 24(6):276, June 1917.

[44] Pejman Khadivi. Adaptive distance estimation and localization in wireless networks with triangle and ptolemy inequalities. In *2010 IEEE 4th International Symposium on Advanced Networks and Telecommunication Systems*, pages 88–90. IEEE, December 2010.

[45] Gordon M. Crippen and Timothy F. Havel. *Distance geometry and molecular conformation*, volume 15 of *Chemometrics Series*. Research Studies Press, Ltd., Chichester; John Wiley & Sons, Inc., New York, 1988.

[46] Dmitrij Jurévič Burago, Jurij D. Burago, and Sergej Ivanov. *A course in metric geometry*. Graduate studies in mathematics 33. American Mathem. Soc., Providence, RI, 2001.

[47] I J Schoenberg. On metric arcs of vanishing menger curvature. *Ann. Math.*, 41(4):715, October 1940.

[48] I J Schoenberg. A remark on M. M. day's characterization of inner-product spaces and a conjecture of L. M. blumenthal. *Proc. Am. Math. Soc.*, 3(6):961–964, 1952.

[49] S. M. Buckley, K. Falk, and D. J. Wraith. Ptolemaic spaces and CAT(0). *Glasg. Math. J.*, 51(2):301–314, 2009.

[50] Thomas Foertsch, Alexander Lytchak, and Viktor Schroeder. Nonpositive curvature and the Ptolemy inequality. *Int. Mathem. Res. Not.*, (9):rnm100–rnm100, July 2010.

[51] Peter Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, nov 1987.

[52] Weiping Xu, Richard C Wilson, and Edwin R Hancock. Determining the cause of negative dissimilarity eigenvalues. In *Computer Analysis of Images and Patterns*, pages 589–597. Springer Berlin Heidelberg, 2011.

[53] Robert P W Duin and Elżbieta Pękalska. Non-Euclidean dissimilarities: Causes and informativeness. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 324–333. Springer Berlin Heidelberg, 2010.

[54] Bruno Cesar Feltes, Eduardo Bassani Chandelier, Bruno Iochins Grisci, and Márcio Dorn. CuMiDa: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology*, 26(4):376–386, 2019.

[55] Rishi Sonthalia, Greg Van Buskirk, Benjamin Raichel, and Anna Gilbert. How can classical multidimensional scaling go wrong? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12304–12315, 2021.

[56] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, June 2014.

[57] Weiyang Liu, Zhen Liu, James M Rehg, and Le Song. Neural similarity learning. *Neural Inf Process Syst*, abs/1910.13003, October 2019.

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, page 6000–6010, 2017.

[59] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6341–6350, Red Hook, NY, USA, 2017. Curran Associates Inc.

[60] Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, and Nando de Freitas. Hyperbolic attention networks. In *International Conference on Learning Representations*, 2019.

[61] Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[62] Alexandru Tifrea, Gary Becigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings. In *International Conference on Learning Representations*, 2019.

[63] Christopher De Sa, Albert Gu, Christopher Ré, and Frederic Sala. Representation tradeoffs for hyperbolic embeddings. *Proc. Mach. Learn. Res.*, 80:4460–4469, 2018.

[64] Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

[65] Ines Chami, Rex Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, 32:4869–4880, December 2019.

[66] Sharon Zhang, Amit Moscovich, and Amit Singer. Product manifold learning. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3241–3249. PMLR, 13–15 Apr 2021.

[67] Daniel McNeela, Frederic Sala, and Anthony Gitter. Product manifold representations for learning on biological pathways. arXiv 2401.15478, 2025.

[68] Philippe Chlenski, Kaizhu Du, Dylan Satow, and Itsik Pe'er. Manify: A python library for learning non-euclidean representations. arXiv 2503.09576, 2025.

[69] Max Aalto and Nakul Verma. Metric learning on manifolds. *CoRR*, abs/1902.01738, 2019.

[70] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley, 1990.

[71] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 2007.

[72] Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, pages 1265–1275. ACM, 2019.

[73] Julian McAuley and Jure Leskovec. Learning to discover social circles in ego networks. In P. L. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, pages 539–547, 2012.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction reflect the contribution of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The introduction mentioned two issues defining the scope of the work. First, for non-Euclidean data there are strong lower bound on dimension reduction (even with nonlinear projections). Therefore, our work cannot bypass the theoretical lower bound. Nevertheless, we provide fine-grained analysis which does not contradict with the strong lower bound yet provide meaningful results with performance characterized by parameters measuring the deviation of the input data from Euclidean geometry. Second, the main scope of the paper is to study the conditions on performance and effectiveness of JL-style dimension reduction and we leave out the integration with existing machine learning pipelines as future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All proofs are included with the paper or in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All datasets and parameters for evaluation are included in the paper or Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All data source and code are made available either in the paper or supplemental materials.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All data and parameters for experiments were included in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The reported error is straightforward, either multiplicative or additive factors.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the information in the appendix experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We reviewed the code of ethics and the work in this paper conforms to the code.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed the potential benefit of using the dimension reduction technique proposed in this paper for speeding up machine learning for non-Euclidean data.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: N/A

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: N/A

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Relation of Generalized Power Distance to Silhouette Coefficient

The silhouette coefficient [51] is a measure of how similar an object is to its own cluster compared to other clusters. The silhouette coefficient ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

For simplicity, let us consider two clusters of finite points $X, Y \subseteq \mathbb{R}^m$ with $|X| = |Y| = n$. The classical silhouette coefficient (in two clusters) can be defined as follows: For each point $x \in X$, define "inner distance" $a(x)$ as the average distance between x and all other points in the same cluster, and "cross distance" $b(x)$ as the average distance between x and all points in the other cluster Y. That is:

$$a(x) = \frac{1}{n-1} \sum_{x' \in X,\, x' \neq x} d(x, x') \tag{8}$$

$$b(x) = \frac{1}{n} \sum_{y \in Y} d(x, y) \tag{9}$$

Then, the silhouette coefficient for a point $x \in X$ is defined as:

$$\bar{s}(x) = \frac{b(x) - a(x)}{\mathrm{Norm}(a(x), b(x))},$$

and the silhouette coefficient for the cluster X is defined as the average of the silhouette coefficients for all points in the cluster:

$$\bar{s}(X) = \frac{1}{n} \sum_{x \in X} \bar{s}(x).$$

Here the denominator Norm is a normalizing function which is usually chosen to be $\mathrm{Norm}(a(x), b(x)) = \max\{a(x), b(x)\}$ for the classical silhouette coefficient to make sure the range lies in $[-1, +1]$. The silhouette coefficient can be used to measure the quality of a cluster. For the overall clustering, the average silhouette coefficient scores over all clusters can be used to measure the quality of the whole clustering results [70].

Now we are interested in using a similar idea to study the dissimilarity (divergence) between two probability distributions, say two Gaussian distributions $\mathcal{X} = \mathcal{N}(\mu_x, \sigma_x), \mathcal{Y} = \mathcal{N}(\mu_y, \sigma_y)$. We can naturally view the two distributions as two clusters. Here we are interested in the "silhouette coefficient score" of the whole clustering.

To better analyze the "silhouette coefficient score" between two Gaussian distributions, we consider an unnormalized version of the silhouette coefficient. For any $x \in \mathcal{X}$, let

$$\begin{aligned} a(x) &= \mathbb{E}_{x' \sim \mathcal{X}}[\|x - x'\|^2], \\ b(x) &= \mathbb{E}_{y \sim \mathcal{Y}}[\|x - y\|^2], \\ s(\mathcal{X}) &= \mathbb{E}_{x \sim \mathcal{X}}[b(x) - c \cdot a(x)], \end{aligned}$$

where $c \geq 1$ is a constant parameter used as a trade-off balance between inner distance and cross distance. Now we can define an (unnormalized) silhouette score between two Gaussian distributions $\mathcal{X}$ and $\mathcal{Y}$ as:

$$\begin{aligned} SC(\mathcal{X}, \mathcal{Y}) &\triangleq \frac{1}{2}[s(\mathcal{X}) + s(\mathcal{Y})] \\ &= \frac{1}{2}[\mathbb{E}_{x \sim \mathcal{X}}[b(x) - c \cdot a(x)] + \mathbb{E}_{y \sim \mathcal{Y}}[b(y) - c \cdot a(y)]] \\ &= \mathbb{E}_{x,y \sim \mathcal{X} \times \mathcal{Y}}\left[\|x - y\|^2\right] - \frac{c}{2}\left(\mathbb{E}_{x,x' \sim \mathcal{X}}[\|x - x'\|^2] + \mathbb{E}_{y,y' \sim \mathcal{Y}}[\|y - y'\|^2]\right) \end{aligned}$$

Based on properties of Gaussian distributions, we could calculate:

$$SC(\mathcal{X}, \mathcal{Y}) = \|\mu_x - \mu_y\|^2 - (c - 1)(\sigma_x^2 + \sigma_y^2).$$

In general, we could consider a silhouette score between two Gaussian distributions composed with two parts balanced with each other:

1. Some distance $\mathrm{DIS}_{\mathbb{P}}(\mathcal{X},\mathcal{Y})$ measures the similarity between the two distributions.
2. Some negative terms Δ represent the variances of the two distributions themselves.

Here we consider the following construction of a silhouette score between two Gaussian distributions we are interested in: Let $\Delta[\mathcal{X}] = \mathbb{E}_{x,x'\sim\mathcal{X}}[\|x-x'\|^2]$.

$$\begin{aligned} SC_W(\mathcal{X},\mathcal{Y}) &\triangleq \mathrm{DIS}_W^2(\mathcal{X},\mathcal{Y}) - (\Delta[\mathcal{X}] + \Delta[\mathcal{Y}]) && (10)\\ &= \|\mu_x - \mu_y\|^2 + (\sigma_x - \sigma_y)^2 - 2(\sigma_x^2 + \sigma_y^2) && (11)\\ &= \|\mu_x - \mu_y\|^2 - (\sigma_x + \sigma_y)^2 && (12)\end{aligned}$$

where DIS_W^2 is the squared 2-Wasserstein distance. This construction is interesting because it is consistent with our construction of generalized power distance, which brings up a potential statistical interpretation.

Normalized Silhouette Score One could also define a normalized version of the Silhouette Score. In general, it could be defined as:

$$\overline{SC}(\mathcal{X},\mathcal{Y}) \triangleq \frac{\mathrm{DIS}_{\mathbb{P}}(\mathcal{X},\mathcal{Y}) - C\cdot(\Delta[\mathcal{X}] + \Delta[\mathcal{Y}])}{\mathrm{Norm}(\mathrm{DIS}_{\mathbb{P}}(\mathcal{X},\mathcal{Y}),\ \ C\cdot(\Delta[\mathcal{X}] + \Delta[\mathcal{Y}]))} \qquad (13)$$

with some normalizer function $\mathrm{Norm}()$.

For example, we could define the normalized version SC_W as:

$$\overline{SC}_W(\mathcal{X},\mathcal{Y}) \triangleq \frac{\|\mu_x - \mu_y\|^2 - (\sigma_x + \sigma_y)^2}{\|\mu_x - \mu_y\|^2 + (\sigma_x + \sigma_y)^2} \qquad (14)$$

The range of $\overline{SC}_W$ is $[-1,1]$. Intuitively, it can be used to measure how well two Gaussian distributions are separated. When these two Gaussian distributions have very close centers or very large variances, the normalized silhouette score is close to -1, which means sampling from these two distributions has a low chance of being well separated. It becomes -1 when they are identical $\mathcal{X}=\mathcal{Y}$. When these two distributions are far away or variances are very small, the normalized silhouette score is close to 1, which means sampling from these two distributions enjoys a high chance of being well separated. It becomes $+1$ when these two distributions are delta distributions on distinct center points. Zero represents borderline cases. We set it to be -1 for $\mu_x = \mu_y$ and $\sigma_x = \sigma_y = 0$.

B Missing Proofs in Section 2 and Section 3

B.1 Proofs in Section 2

Proof of Lemma 2.2. We apply the standard Johnson-Lindenstrauss lemma to the collection of points $\{x_i^{(p)}\}$ and $\{x_i^{(q)}\}$ separately. By Eq (1) there are two random projection $f_p : \mathbb{R}^p \to \mathbb{R}^{p'}$ and $f_q : \mathbb{R}^q \to \mathbb{R}^{q'}$ with $p', q' = O(\log n/\varepsilon^2)$ such that for any $i \neq j$,

$$(1-\varepsilon)\|x_i^{(p)} - x_j^{(p)}\|_E^2 \le \|f_p(x_i^{(p)}) - f_p(x_j^{(p)})\|_E^2 \le (1+\varepsilon)\|x_i^{(p)} - x_j^{(p)}\|_E^2$$
$$(1-\varepsilon)\|x_i^{(q)} - x_j^{(q)}\|_E^2 \le \|f_q(x_i^{(q)}) - f_q(x_j^{(q)})\|_E^2 \le (1+\varepsilon)\|x_i^{(q)} - x_j^{(q)}\|_E^2$$

Now we can bound the (p',q')-norm of the vector under projection f which produces $f(x_i) = (f_p(x_i), f_q(x_i))$ of dimension $p'+q'$.

$$\begin{aligned}\|f(x_i) - f(x_j)\|_{p',q'}^2 &= \|f_p(x_i)\|_E^2 - \|f_q(x_i)\|_E^2\\ &\le (1+\varepsilon)\|x_i^{(p)} - x_j^{(p)}\|_E^2 - (1-\varepsilon)\|x_i^{(q)} - x_j^{(q)}\|_E^2\\ &= \|x_i - x_j\|_{p,q}^2 + \varepsilon\|x_i - x_j\|_E^2\end{aligned}$$

The other direction is similar. □

Proof of Lemma 2.4. Consider the variables $\|v\|_E^2/(p+q)$ and $\|v\|_{p,q}^2/(p+q)$. Let $\mu = \mathbb{E}(X^2)$ where X is a random variable with distribution f. Taking $p+q\to\infty$ and using Central Limit Theorem implies that $\|v\|_E^2/(p+q)$ and $\|v\|_{p,q}^2/(p+q)$ will converge to their means which are μ and $\frac{p-q}{p+q}\mu$ respectively. Plugging in $q < \frac{C-1}{C+1}p$ and $p < \frac{C-1}{C+1}q$ give us what we want. □

Proof of Theorem 2.6. We may first write the uniform distribution over the sphere as $Y = (\frac{Y_1}{\sqrt{\sum_i^{p+q} Y_i^2}}, \frac{Y_2}{\sqrt{\sum_i^{p+q} Y_i^2}}, ... \frac{Y_{p+q}}{\sqrt{\sum_i^{p+q} Y_i^2}})$ where each $Y_i = N(0,1)$. Then the distribution of $\frac{\|u-v\|_E}{\|u-v\|_{p,q}}$ will be :

$$\frac{\sum_{i=1}^{p+q} Y_i^2}{\sum_{i=1}^{p} Y_i^2 - \sum_{i=p+1}^{p+q} Y_i^2}$$

Now $\sum_{i=1}^{p+q} Y_i^2$ is just a chi-squared random variable with $p+q$ degrees of freedom. A common known concentration inequality is:

$$\mathbb{P}(\chi^2 \in (1 \pm \delta)(p+q)) \geq 1 - 2e^{-\frac{p+q}{2}(\frac{1}{2}\delta^2 - \frac{1}{3}\delta^3)}$$

On the other hand, $\sum_{i=1}^{p} Y_i^2 - \sum_{i=p+1}^{p+q} Y_i^2$ has mean $p-q$ and we have the following concentration inequality:

$$\mathbb{P}(\chi^2 \in (p-q) \pm 2\delta(p+q)) \geq 1 - 2e^{-\frac{p+q}{2}(\frac{1}{2}\delta^2 - \frac{1}{3}\delta^3)}$$

since $\sum_{i=1}^{p+q} Y_i^2 \in (1 \pm \delta)(p+q)$ implies $\sum_{i=1}^{p} Y_i^2 - \sum_{i=p+1}^{p+q} Y_i^2 \in (p-q) \pm 2\delta(p+q)$. The greatest error ratio clearly is bounded by $q = \frac{C-1}{C+1}p$ and at:

$$\frac{(1+\delta)(p+q)}{(p-q) - 2\delta(p+q)} \frac{p-q}{p+q} - 1 = \frac{2pC(1+\delta)/(C+1)}{2p/(C+1) - 4\delta pC/(C+1)} \frac{1}{C} - 1 = \frac{\delta + 2\delta C}{1 - 2\delta C}$$

$$\mathbb{P}(\frac{\sum_{i=1}^{p+q} Y_i^2}{\sum_{i=1}^{p} Y_i^2 - \sum_{i=p+1}^{p+q} Y_i^2} \in (1 \pm \frac{\delta + 2\delta C}{1 - 2\delta C})\frac{p+q}{p-q}) \geq 1 - 2e^{-\frac{p+q}{2}(\frac{1}{2}\delta^2 - \frac{1}{3}\delta^3)}$$

Now let $\delta = \frac{\epsilon/2}{4C + 2 + \epsilon C}$:

$$\mathbb{P}(\frac{\sum_{i=1}^{p+q} Y_i^2}{\sum_{i=1}^{p} Y_i^2 - \sum_{i=p+1}^{p+q} Y_i^2} \in (1 \pm \frac{\epsilon}{2})\frac{p+q}{p-q}) \geq 1 - 2e^{-O((p+q)\epsilon^2/C^2)} \geq 1 - O(n^{-2})$$

The last inequality comes from the fact that our target dimension is $O(C^2 \log(n)/\epsilon^2)$, so we can make the assumption $p + q > O(C^2 \log(n)/\epsilon^2)$. From a union bound, then we have with high probability, $\frac{\|u-v\|_E}{\|u-v\|_{p,q}} \leq C(1 + \frac{\epsilon}{2})$ Then, we can use Theorem 2.2 with input $\epsilon' = \epsilon/(2C)$ which gives us the error bound:

$$\begin{align*}
\|f(x_i) - f(x_j)\|_{p',q'}^2 &\leq \|x_i - x_j\|_{p,q}^2 + \frac{\epsilon}{2C}\|x_i - x_j\|_E^2 \tag{15} \\
&\leq \|x_i - x_j\|_{p,q}^2 + \frac{\epsilon}{2C}C(1 + \frac{\epsilon}{2})\|x_i - x_j\|_{p,q}^2 \tag{16} \\
&= \|x_i - x_j\|_{p,q}^2 + (\frac{\epsilon}{2} + \frac{\epsilon^2}{2})\|x_i - x_j\|_{p,q}^2 \tag{17} \\
&\leq (1 + \epsilon)\|x_i - x_j\|_{p,q}^2 \tag{18} \\
& \tag{19}
\end{align*}$$

□

B.2 Proofs in Section 3

Proof of Lemma 3.1. We begin by finding the Gram Matrix of $E = D - 4r^2(I - J)$. Recall that the centering matrix $C = I - \frac{J}{n}$. The Gram matrix of E is

$$-\frac{1}{2}CEC = -\frac{1}{2}C(D - 4r^2(I - J))C = -\frac{1}{2}CDC + 2r^2C(I - J)C = -\frac{1}{2}CDC + 2r^2C.$$

Here we use the fact that $C(I - J) = C$ and $C^2 = C$. Further, $C = I - \frac{J}{n}$ is symmetric and positive semi-definite. C has a single eigenvalue of 0 with $\mathbf{1}_n$ as its eigenvector. We see that $\mathrm{Gram}(D) = -\frac{1}{2}CDC$ and C share that eigenvalue and eigenvector. The rest of the eigenvalues of C are 1. Then, the eigenvalues of $\mathrm{Gram}(E)$ are exactly $2r^2$ added to the eigenvalues of $\mathrm{Gram}(D)$. Thus, it is clear all of its eigenvalues are greater than or equal to zero as long as $2r^2 \geq |e_n|$. □

Proof of Lemma 3.2. We apply the standard Johnson-Lindenstrauss Lemma on the centers of the power distance representation $\{p_i\}$ with a random projection $f : \mathbb{R}^d \to \mathbb{R}^m$. By Eq (1) we have

$$\begin{aligned}\mathrm{Pow}((f(p_i), r_i), (f(p_j), r_j)) &= \|f(p_i) - f(p_j)\|_E^2 - (r_i + r_j)^2 \\ &\leq (1+\varepsilon)\|p_i - p_j\|_E^2 - (r_i + r_j)^2 \\ &= (1+\varepsilon)\,\mathrm{Pow}((p_i, r), (p_j, r)) + \varepsilon(r_i + r_j)^2.\end{aligned}$$

The other direction is similar. □

C Additional Experimental Results

C.1 Details of Data Sets

SNAP dataset. We tested our methods on a diverse collection of real-world network datasets sourced from the Stanford Network Analysis Project (SNAP) [56]. These datasets span various domains and exhibit a wide range of structural properties, helping to validate the robustness and generalizability of our findings.

- **Email-Eu-core (Email)** [71]: Email exchanges within a European research institution.
- **MOOC-actions (MOOC)** [72]: User actions on a Massive Open Online Course platform.
- **Facebook-ego-networks (Facebook)** [73]: Ego-networks of Facebook users.

We only used the graph structures and ignored the edge weights and directions. That means all graphs are treated as unweighted and undirected.

Synthetic datasets for examples. We also manually generated two datasets to demonstrate when projecting in Pseudo-Euclidean space is better than projecting with power distances and vice versa. In the first dataset, we had a single very large negative eigenvalue and the rest all positive eigenvalues in the diagonal matrix of the orthogonal decomposition of the distance matrix. In this situation, power distances of the points would have large radii and the error would be large while Pseudo-Euclidean projection would be on a signature of $(p, 1)$, so the error bound would be much smaller. In the second dataset, we had the ratio of positive eigenvalues to negative eigenvalues was a constant fraction, and the negative eigenvalues were all small. In this way, Pseudo-Euclidean projection would have a large error bound while power distance projection would have very small radii and a small error bound.

Other datasets. We randomly select 1000 images from MNIST and CIFAR-10. The distance measure between two images are computed by the k-neareast neighbor (i.e. Isomap) with $k = 10$. The distance measure of genomics dataset is Entopic Affinity, introduced in prior works and commonly used for this datasets. The details of implementation can be found in our Github Repo.

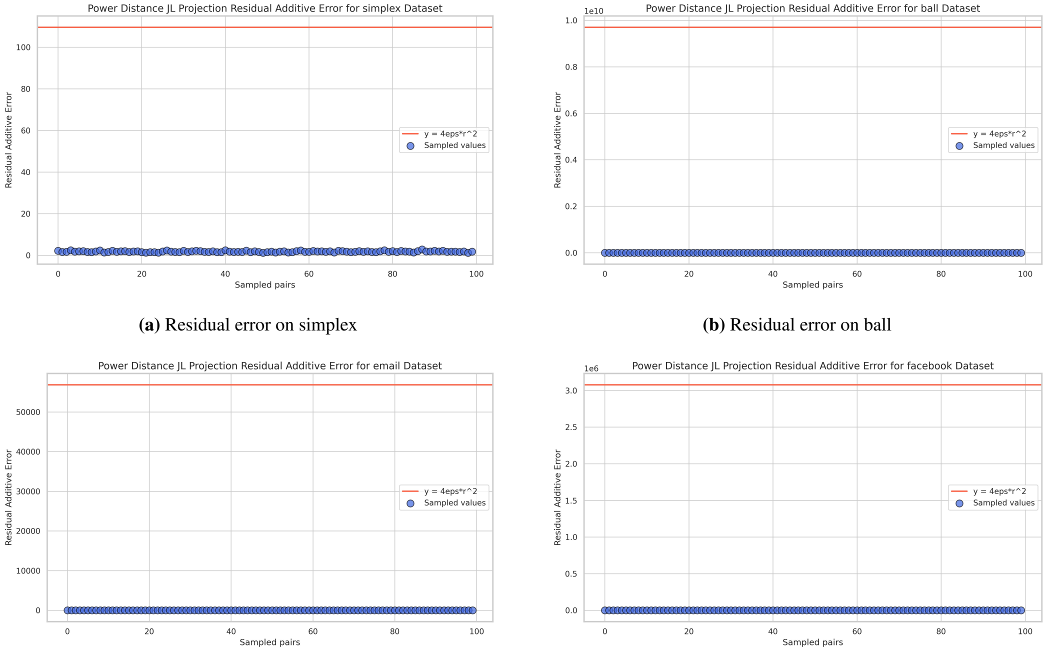
C.2 Details of Implementation

All experiments are done with a Macbook Pro with Apple M2 Chip of 16GB memory. The algorithms are fairly efficient and we did not encounter issues on computational resources.

The JL Transforms have only one parameter ε, and we set that to be 0.5 throughout the paper. The constant used in the big O notation of $O(\log n/\varepsilon^2)$ is 2. Another caveat is, our setting does not give access to data coordinates, therefore after obtaining coordinates from the JL transform, we use the original dissimilarity measure to reconstruct $\hat{D}$. If the original measure is inexplicit or unknown, we use Euclidean distance. We only use the data coordinates for the classical JL.

C.3 More Illustrations

The illustrations here extend the experiments to verify Theorem 2.3 and Theorem 3.3.



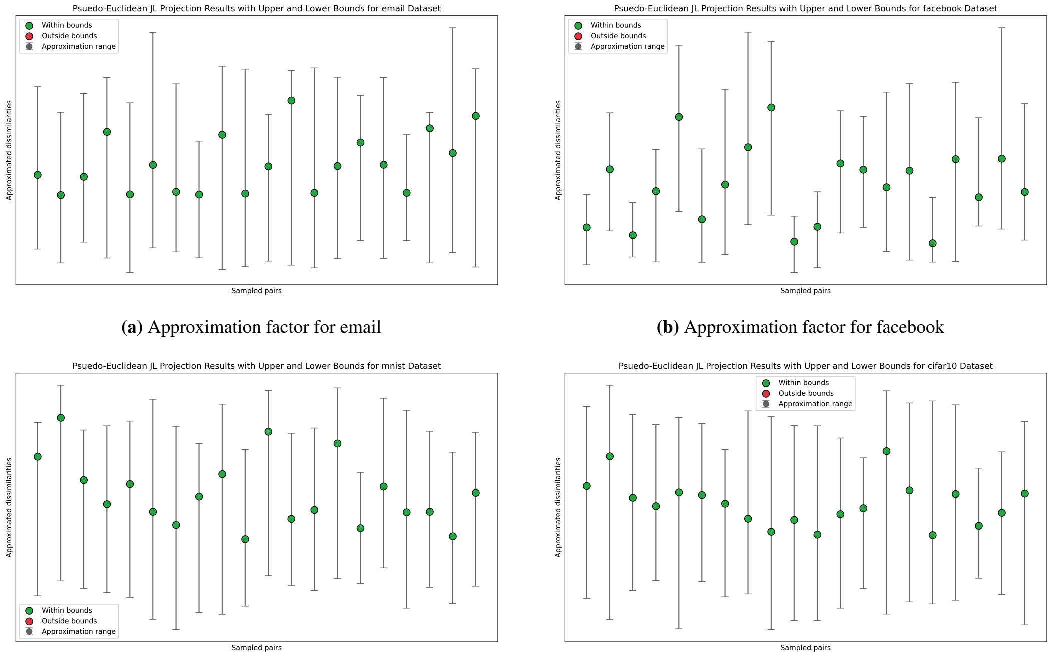
(a) Residual error on simplex

(b) Residual error on ball

(c) Residual error on email

(d) Residual error on facebook

Figure 4: Illustrations of Power Distance JL Transform Residual Error



(a) Approximation factor for email

(b) Approximation factor for facebook

(c) Approximation factor for MNIST

(d) Approximation factor for CIFAR-10

Figure 5: Illustrations of Pseudo Euclidean JL Transform Multiplicative Error