

Dimensionality reduction:

theoretical perspective on practical measures

Nova Fandina

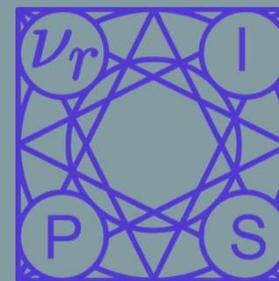
Hebrew University, Israel

fandina@cs.huji.ac.il

Joint work with

Yair Bartal, Hebrew University, Israel, yair@cs.huji.ac.il

Ofer Neiman, Ben Gurion University, Israel, neimano@cs.bgu.ac.il



NeurIPS19

8-14 December,
Vancouver, Canada

Introduction

- Measuring the Quality of Embedding
 - in theory : worst case distortion analysis
 - in practice: average case distortion measures
 - in between: theoretical analysis of practical measures
(for dimensionality reduction methods)
- Our Results
 - upper bounds
 - lower bounds
 - approximating optimal embedding
 - empirical experiments

Measuring the quality of embedding: in theory

A basic task in metric embedding theory (informally) is:

Given metric spaces X and Y , embed X into Y ,
with small error on the distances.

How *well* it can be done? How to measure an error?

In theory: “well” traditionally means to minimize **distortion** of the **worst** pair

Definition: worst case distortion

For an embedding $f: X \rightarrow Y$, for a pair of points $u \neq v \in X$

- $expans_f(u, v) = \frac{d_Y(f(u), f(v))}{d_X(u, v)}$, $contr_f(u, v) = \frac{d_X(u, v)}{d_Y(f(u), f(v))}$
- $distortion(f) = \max_{u \neq v \in X} \{expans_f(u, v)\} \cdot \max_{u \neq v \in X} \{contr_f(u, v)\}$

Measuring the quality of embedding: in practice

In practice, the demand for the worst-case guarantee is too strong: the quality of a method in practical applications is rather usually measured by its **average** performance over all pairs.

There is a reach body of research literature where the variety of average quality measurement criteria is studied and applied:

- Yuval Shavitt and Tomer Tankel. **Big-bang simulation for embedding network distances in Euclidean space.** IEEE/ACM Trans. Netw., 12(6), 2004.
- P. Sharma, Z. Xu, S. Banerjee, and S. Lee. **Estimating network proximity and latency.** Computer Communication Review, 36(3), 2006.
- P. J. F. Groenen, R. Mathar, and W. J. Heiser. **The majorization approach to multidimensional scaling for minkowski distances.** Journal of Classification, 12(1), 1995.
- J. F. Vera, W. J. Heiser, and A. Murillo. **Global optimization in any minkowski metric: A permutation-translation simulated annealing algorithm for multidimensional scaling.** Journal of Classification, 24(2), 2007.
- A. Censi and D. Scaramuzza. **Calibration by correlation using metric embedding from nonmetric similarities.** IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(10), 2013.
- C. Lumezanu and N. Spring. **Measurement manipulation and space selection in network coordinates.** The 28th International Conference on Distributed Computing Systems, 2008.
- S. Chatterjee, B. Neff, and P. Kumar. **Instant approximate 1-center on road networks via embeddings.** In Proceedings of the 19th International Conference on Advances in Geographic Information Systems, GIS '11, 2011.
- S. Lee, Z. Zhang, S. Sahu, and D. Saha. **On suitability of Euclidean embedding for host-based network coordinate systems.** IEEE/ACM Trans. Netw., 18(1), 2010.
- L. Chennuru Vankadara and U. von Luxburg. **Measures of distortion for machine learning.** Advances in Neural Information Processing Systems, Curran Associates, Inc., 2018.

Just a small sample from googolplex number of such studies.

Measuring the quality of embedding: moments of distortion and relative error measure

For $f: X \rightarrow Y$, for a pair $u \neq v \in X$, $dist_f(u, v) := \max\{expans_f(u, v), contract_f(u, v)\}$

ℓ_q -distortion [defined in ABN11]

For $f: X \rightarrow Y$, for a distribution Π over pairs of X , $q \geq 1$

$$\ell_q^{(\Pi)}-dist(f) = \left(E_{\Pi} \left[\left(dist_f(u, v) \right)^q \right] \right)^{1/q}$$

Relative Error Measure [commonly used in network applications: CDKLM04, SXBL06, ST04]

We further generalize:

$$REM_q^{(\Pi)} = \left(E_{\Pi} \left[\left(|dist_f(u, v) - 1| \right)^q \right] \right)^{1/q}$$

Measuring the quality of embedding: additive distortion measures

Initiated and studied within the Multi-Dimensional Scaling framework [CC00].
Found an enormous number of applications in visualization, clustering, indexing
and many more fields [see a long list of citations in the paper].

We further generalize the basic variants that appear in the literature:

For a pair $u \neq v \in X$, $d_{uv} = d_X(u, v)$, $\hat{d}_{uv} = d_Y(f(u), f(v))$, $q \geq 1$

$$\text{Stress}_q(f) = \left(\frac{E_{\Pi}[|d_{uv} - \hat{d}_{uv}|^q]}{E_{\Pi}[(d_{uv})^q]} \right)^{1/q}$$

$$\text{Stress}_q^*(f) = \left(\frac{E_{\Pi}[|d_{uv} - \hat{d}_{uv}|^q]}{E_{\Pi}[(\hat{d}_{uv})^q]} \right)^{1/q}$$

$$\text{Energy}_q(f) = \left(E_{\Pi} \left[\left(\frac{|\hat{d}_{uv} - d_{uv}|}{d_{uv}} \right)^q \right] \right)^{1/q}$$

$$\text{REM}_q(f) = \left(E_{\Pi} \left[\left(\frac{|\hat{d}_{uv} - d_{uv}|}{\min\{d_{uv}, \hat{d}_{uv}\}} \right)^q \right] \right)^{1/q}$$

New machine learning motivated distortion measure

σ -distortion: defined and studied in VL18 [NeurIPS18]

$$\sigma - dist^{(\Pi)}_{q,r}(f) = \left(E_{\Pi} \left[\left(\left| \frac{exapns_f(u, v)}{\ell_r^U - expans(f)} - 1 \right| \right)^q \right] \right)^{1/q}$$

- $\ell_r^{(U)} - expans(f) = E_U[(expans_f(u, v)^r)]$
- $\ell_r^{(U)} - contr(f) = E_U[(contr_f(u, v)^r)]$

Necessary properties a quality measure has to posses to be valid for the ML applications were defined and studied in [VL18]:

- translation invariance
- scale invariance
- monotonicity
- robustness (outliers, noise)
- incorporation of probability

Contributions of our paper:

the first theoretical study of the average distortion measures

- We show that all the other average distortion measures considered here can be easily adapted to satisfy similar ML motivated properties, generalizing the results of VL18.
- We show deep tight relations between these different objective functions, and further develop properties and tools for analyzing embeddings for these measures.

While these measures have been extensively studied from a practical point of view, and many heuristics are known in the literature, almost nothing is known in terms of rigorous analysis and absolute bounds. Moreover, many real-world misconceptions exist about what dimension may be necessary for good embeddings.

- We present the *first theoretical analysis* of all these measures providing absolute bounds that shed light on these questions. We exhibit approximation algorithms for optimizing these measures, and further applications.
- We validate our theoretical findings experimentally, by implementing our algorithms and running them on various randomly generated Euclidean and non-Euclidean metric spaces.

Moment analysis of dimensionality reduction: bridging the gap between theory and practice outlook

The main theoretical question we study in the paper is:

$\alpha(k, q)$ -Dimension Reduction

Given a dimension bound $\mathbf{k} \geq \mathbf{1}$ and $\mathbf{q} \geq \mathbf{1}$, what is the least $\alpha(\mathbf{k}, \mathbf{q})$ such that every finite subset of Euclidean space embeds into \mathbf{k} dim. with $\mathbf{Measure}_{\mathbf{q}} \leq \alpha(\mathbf{k}, \mathbf{q})$?

- We answer the question by providing almost tight **upper** and **lower** bounds on $\alpha(\mathbf{k}; \mathbf{q})$, for all the discussed measures.
- We prove that the Johnson-Lindenstrauss dimensionality reduction achieves bounds in terms of q and k that dramatically outperform a widely used in practice PCA algorithm.
- Moreover, in experiments, we show that the JL outperforms Isomap and PCA methods, on various randomly generated metric spaces.

Technical results: upper bounds on $\alpha(k, q)$ -dim. Reduction

Given an n -point metric space X in ℓ_2^d and $\epsilon > 0$, the JL lemma states:

[JL84] Projection of X onto a random subspace of dim. $k = \mathbf{O}(\log n / \epsilon^2)$,
with const. prob. has worst case $\mathbf{dist}(f) = \mathbf{1} + \epsilon$.

There are many implementations of the JL transform (satisfying the JL property):

[AchI03] The entries of T are uniform indep. from $\{\pm 1\}$.

[DKS10,KN10,AL10] Sparse/Fast: particular distr. from $\{\pm 1, 0\}$.

[IM98] T is a matrix of size $k \times d$ with indep. entries sampled from $N(0, 1)$.

The embedding $f: X \rightarrow \ell_2^k$ is defined by $f(x) = \mathbf{1}/\sqrt{k} \cdot T(x)$.

Technical results: the IM98 implementation of the JL transform

- The JL transform of IM98 provides **constant** upper bounds for all $Measure_q$. The bounds are almost tight. All our theorems true for that implementation.
- Other mentioned implementations **do not work** for ℓ_q -dist and for REM_q :

Observation

If a linear transformation $T: R^d \rightarrow R^k$ samples its entries from a discrete set of values of size $s \leq d^{1/k}$, then applying it on a standard basis of R^d results in ℓ_q -dist, $REM_q = \infty$.

- PCA may produce an embedding of extremely poor quality for all the measures (this does not happen to the JL). In the next slides we give an example of a family of Euclidean metric spaces, on which PCA produces provably large distortions.

On the limitations of the classical MDS (PCA) method

PCA/c-MDS For a given finite $X \in \ell_2^d$ and a given integer $k \geq 1$, computes the best rank k - approx. to X :

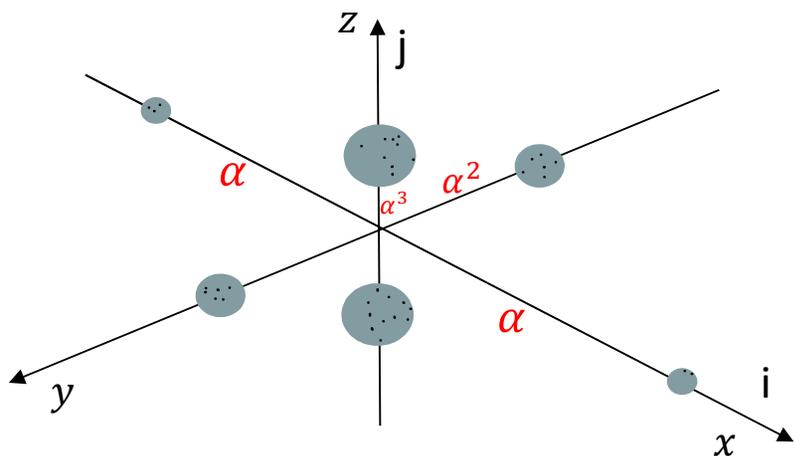
A projection P onto the k - dim subspace spanned by the largest eigenvectors of the covariance matrix, with the smallest $\sum_{u \in X} \|u - P(u)\|^2$.

- $P: X \rightarrow \ell_2^k$ has optimal $\sum_{u \neq v \in X} (d_{uv}^2 - \hat{d}_{uv}^2)$ over all projections.
- Often misused: “minimizing $Stress_2$ over all embeddings into k -dim”.
- In fact, PCA does not minimize any of the mentioned measures.

Next, we present a metric space of dimension $d \geq 1$ that can be efficiently embedded into a line (with small $Measure_q$ distortions) but such that PCA fails to produce a comparable result.

Bad metric space for the PCA method

- The metric is in d dimensional Euclidean space, for any d large enough.
- Fix some $\alpha < 1$, and $q \geq 1$.
- Consider the standard basis vectors e_1, \dots, e_d .
- For each vector e_i , let X_i be the set of $\left(\frac{1}{\alpha^i}\right)^q$ copies of vector $\alpha^i \cdot e_i$, and let Y_i be the set of the same size of the antipodal vector $-\alpha^i \cdot e_i$.



The JL embedding has bounded measures for any space: $\alpha(k) \rightarrow 0$, as k increases.

In the paper we show an embedding of this metric space into R with: **$Stress_2 \leq \alpha/d^{1/2}$** .

PCA projects this space onto $span\{e_1, \dots, e_k\}$
For $k < 0.99d$ we have:

- **$Stress_2 \geq \Omega(1)$**
- PCA is not better than a naïve algo: any non-expansive embedding has const Stress measure
- **$\ell_q\text{-dist}/REM_q = \infty$**

Technical results: upper bounds: moment analysis of JL transform

Theorem [Moment analysis of JL transform]

There is a map (JL or normalized JL) $f: X \rightarrow \ell_2^k$ s.t. for a given $q \geq 1$ with const. prob.

	$1 \leq q < \sqrt{k}$	$\sqrt{k} \leq q \leq k/4$	$k/4 \leq q \leq k$	$q = k$	$k \leq q \leq \infty$
ℓ_q -dist(f) =	$1 + O\left(\frac{1}{\sqrt{k}}\right)$	$1 + O\left(\frac{q}{k-q}\right)$	$\left(\frac{k}{k-q}\right)^{O(1/q)}$	$O\left(\sqrt{\log n}\right)^{1/k}$	$n^{O\left(\frac{1}{k-q}\right)}$

The bounds are almost tight in most of the ranges of values of q and k :

Theorem[Lower bounds for ℓ_q -distortion]. Let E_n denote an n -point equilateral space.

- Any embedding $f: E_n \rightarrow \ell_2^k$ must have ℓ_q -dist(f) = $1 + \Omega(q/k)$, for $1 \leq q \leq \sqrt{k}$.
- There is a finite Euclidean space Z such that any embedding $f: Z \rightarrow \ell_2^k$ must have

$$\ell_q$$
-dist(f) = $1 + \Omega\left(\frac{q}{k-q}\right)$, for $\sqrt{k} \leq q < k$.

For the values of $q \sim k$ we prove a **lower bound** of $\Omega\left(\sqrt{\log n}\right)^{1/k}$, which exhibits a *phase transition* phenomenon, providing a guidance on how to choose the target dimension k .

Technical results: simultaneous guarantees for all $q \geq 1$

A stronger demand is to require a single embedding to *simultaneously* achieve best possible bounds for all values of q . We show almost tight upper bounds:

Theorem (simultaneous analysis)

There is a map (JL) $f: X \rightarrow \ell_2^k$ s.t. with constant probability

	$1 \leq q \leq \sqrt{k}$	$\sqrt{k} \leq q \leq k$	$q = k$	$q \geq k$
$\ell_q\text{-dist}(f) =$	$1 + O(1/\sqrt{k})$	$1 + O(q/(k - q))$	$O(\log n)^{1/k}$	$\tilde{O}(n^{2/k} - n^{2/q})$

- **The bounds are almost tight:** Applying f on the equilateral space E_n :
 1. If $\ell_{\sqrt{k}}\text{-dist}(f) \leq 1 + 1/\sqrt{k}$, then $\ell_1\text{-dist}(f) \geq 1 + \Omega(1/\sqrt{k})$.
 2. If $\ell_1\text{-dist}(f) = O(1)$, then $\ell_q\text{-dist}(f)$ is at least as in the table, for $q \geq k$.

Technical results: REM and additive distortion measures

Theorem (REM and additive measures analysis of JL)

There is a map (JL) $f: X \rightarrow R^k$, for $k \geq 2$, s.t. with const. prob. for all $1 \leq q \leq k - 1$ (simultaneously):

$$\sigma\text{-dist}, \text{Stress}_q, \text{Stress}^*, \text{Energy}_q(f) \leq \text{REM}_q(f) = O(\sqrt{q/k}).$$

- Tight for $q \geq 2$: any embedding $f: E_n \rightarrow \ell_2^k$ must have $\text{Measure}_q = \Omega(\sqrt{q/k})$, for all the measures.
- For $1 \leq q < 2$, the lower bound is $\Omega(1/k^{1/q})$, for the space E_n as well.

We note that for the equilateral space the lower bound for all the measures are the best that can be achieved for an embedding into ℓ_2^k :

Theorem[Optimal embedding of E_n]. For every $k \geq 3$, for all $1 \leq q \leq \sqrt{k}$, for every distribution Π over pairs of E_n , there is a random map $f: E_n \rightarrow \ell_2^k$ s.t. with const. prob.:

1. $\ell_q\text{-dist}(f) = 1 + O\left(\frac{q}{k}\right)$,
2. $\text{Measure}_q(f) = O\left(\frac{1}{k^{1/q}}\right)$, for any $1 < q < 2$, for all the additive measures.

Approximating optimal embedding: applications

General Metrics: Approximating the Optimal Embedding

For a given finite X and for $k \geq 1$, compute an embedding of X into k -dim Euclidean space that *approximates* the best possible embedding, for a given $Measure_q$.

There are very few previous theoretical works on that direction:

[CD06] Optimizing is NP-hard for $Stress_q$ and $k = 1$.

[ABN11] Every finite X embeds into $\ell_p^{O_p(\log n)}$, with ℓ_q -distortion $O(q/p)$.

Gives $O(q)$ approximation to the optimum under ℓ_q -distortion.

[HIL03] Give 2-approx. to $Stress_\infty$, for embedding into $k = 1$ dim.

[Bado03] Gives $O(1)$ -approx. to $Stress_\infty$, for embedding into $k = 2$ dim, under l_1 norm.

[Dham04] Gives $O(\log^{1/q} n)$ -approx. to $Stress_q$, for embedding into $k = 1$ dim.

*We apply our bounds for the Euclidean (k, q) -dimension reduction to provide the **first** approximation algorithms for embedding general metric spaces into low dimensional Euclidean space, for all the various distortion criteria.*

Approximating optimal embedding: applications

For a given X and $q, k \geq 1$, for an objective measure Obj_q let $OPT: X \rightarrow \ell_2^k$ denote an **optimal** embedding for the measure Obj_q . Combing convex optimization with the JL transform we obtain:

Theorem: Approximating optimal embedding

For a given finite X , for a given $k \geq 3$ and $2 \leq q \leq k - 1$, there is a randomized polytime algorithm that computes an embedding $F: X \rightarrow \ell_2^k$ s.t. with const. prob.

- $l_q\text{-dist}(F) = \left(1 + O\left(\frac{1}{\sqrt{k}} + \frac{q}{k-q}\right)\right) \cdot OPT$.
- $Obj_q^{(\Pi)}(F) = O\left(Obj_q(OPT)\right) + O\left(\sqrt{q/k}\right)$, for all the rest measures.

Proof outline:

- Compute an optimal embedding f without constraining the dimension (convex optimization).
- Reduce dimension with JL (IM98 implementation) and show this works.

Approximating optimal embedding: proof outline

Compute an optimal embedding $f: X \rightarrow \ell_2$, *without constraining the dimension*.

- [LLR95]: **SDP** computes a map $f: X \rightarrow \ell_2$ with optimal *worst case* distortion:

For $X = \{x_0, \dots, x_{n-1}\}$, for each pair, variables $z_{ij} = \|f(x_i) - f(x_j)\|^2$,

$g_{ij} = 1/2(z_{0i} + z_{0j} - z_{ij})$ represent $\langle f(x_i), f(x_j) \rangle$, $f(x_0) = 0$.

Then, $\sqrt{z_{ij}}$ are Euclidean distances iff matrix $G[i, j] := g_{ij}$ is PSD.

- **Convex program** to optimize the *Energy_q*: for $q \geq 2$:

$$\min \sum_{0 < i < j < n} \left(\left(\sqrt{\frac{z_{ij}}{d_{ij}}} - 1 \right)^2 \right)^{q/2}, \quad \text{s.t. } z_{ij} \geq 0, G \text{ is PSD.}$$

The objective and the constraints are convex.

Approximating optimal embedding: proof outline

Reduce dimension with JL (IM98 implementation) and show this works:

- Apply JL: $g: f(X) \rightarrow \ell_2^k$.
- The embedding $F: X \rightarrow \ell_2^k$ is defined by $F := g \circ f$.

For all the measures we prove the composition claims. We present here the claim for Energy:

Claim[Composition of $Energy_q$]

If $f: X \rightarrow Y$ is some embedding, and $g: Y \rightarrow Z$ is a random embedding that has $E[expans_g(u, v)] = A$, and $E[contr_g(u, v)] = B$, for all pairs in X , then

$$E[Energy_q(g \circ f)] \leq 4 \cdot Energy_q(f) \cdot (E[Energy_q(g)])^{\frac{1}{q}} + 4 \cdot (E[Energy_q(g)])^{\frac{1}{q}}.$$

- Note that the JL is as in the claim, implying $E[Energy_q(F)] \leq c \cdot \mathbf{OPT} + O\left(\frac{\sqrt{q}}{k}\right)$.

In the paper we similarly prove the approximation ratios for the rest measures.

Empirical experiments: summary

We validate our theoretical findings experimentally on various randomly generated Euclidean and non-Euclidean metric spaces.

As predicted by our lower bounds, the ***phase transition*** is clearly seen in the JL, PCA and ISOMAP methods, for all the measurement criteria.

In our simulations the JL based approximation algorithm (as well as the JL itself, when applied on Euclidean metrics) has shown ***dramatically better*** performance than the PCA(c-MDS) and ISOMAP heuristics, for all distortion measures, indicating that *the JL-based approximation algorithm is a better choice when the preservation of metric properties is desirable.*

Moreover, evidence exists that there is correlation between *lower distortion measures* and *quality of machine learning algorithms* applied on the resulting space, such as in [VL18], where such correlation is experimentally shown between σ -distortion and error bounds in classification. This evidence suggests that the improvement in distortion bounds should be reflected in better bounds for machine learning applications.

Empirical experiments: Euclidean dimensionality reduction

A Euclidean X of a fixed size and dimension $n=d = 800$ is sampled from Normal distribution, with random variance. We embed X into $k \in [4,30]$ dimensions with JL/PCA/Isomap ; the value of $q = 10$.

- In **Fig. 1(a)**, the ℓ_q -distortion as a function of k of the JL embedding is shown for $q = 8, 10, 12$. The phase transitions are seen at around $k \sim q$ as predicted by our theorems.
- In **Fig. 1(b)** the bounds and the phase transitions of the PCA and Isomap methods are shown, for the same setting ($d = 800$; $q = 10$), as predicted by our lower bounds.
- In **Fig. 1(c)**, ℓ_q -dist. bounds are shown for increasing values of $k > q$. Note that the ℓ_q -dist. of the JL is a *small constant close to 1*, as predicted, compared to values significantly > 2 for the compared methods.

Fig. 1 clearly shows that JL dramatically outperforms the other methods for all the range of values of k .

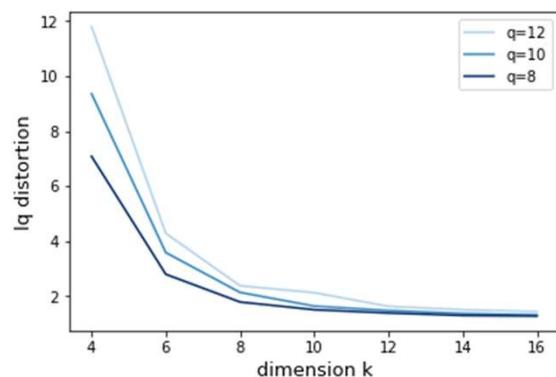


Fig. 1(a) Phase transition of JL.

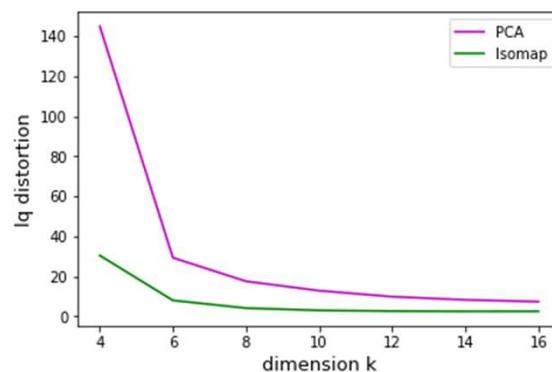


Fig. 1(b) Phase transition of PCA/ISOMAP.

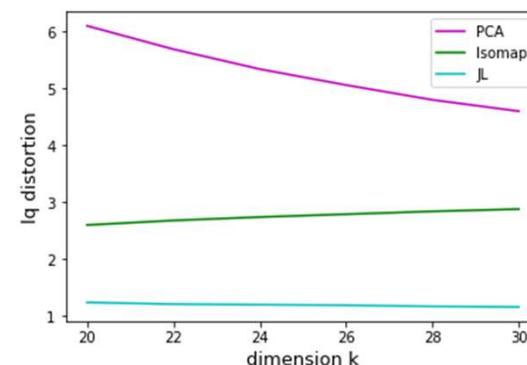


Fig. 1(c) Comparing ℓ_q -dists for $k > q$.

Empirical experiments: Euclidean dimensionality reduction

- In Fig. 2(a), the results are shown for the σ -distortion (the experiment applied on the same setting as before). Again, there is a clear advantage of what the JL achieves with comparison to the other methods.
- In Fig. 2(b), we tested the behavior of the σ -distortion as a function of d -the dimension of the input data set, similarly to that of VL[18](Fig.2). The tests are shown for target dimension $k = 20$ and $q = 2$. According to our theorems, the σ -dist of the JL transform is $O(\sqrt{q/k})$, which is bounded by constant for $q < k$. It is seen that the σ -dist is growing as d increases for both PCA/Isomap, whereas it is a constant for JL, as predicted. Moreover, JL obtains a significantly smaller value of σ -distortion.

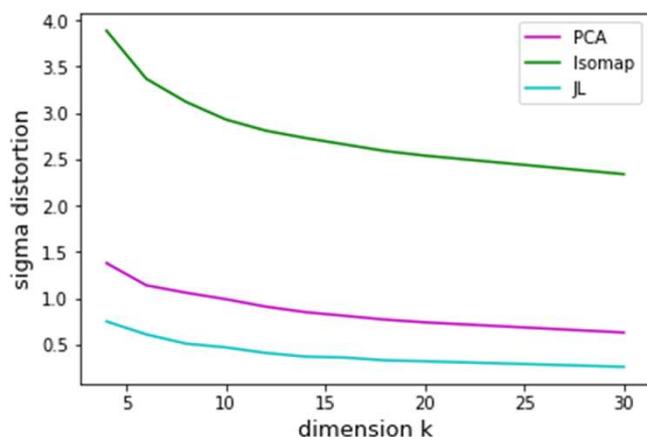


Fig. 2(a) σ -distortion of PCA/ISOMAP/JL.

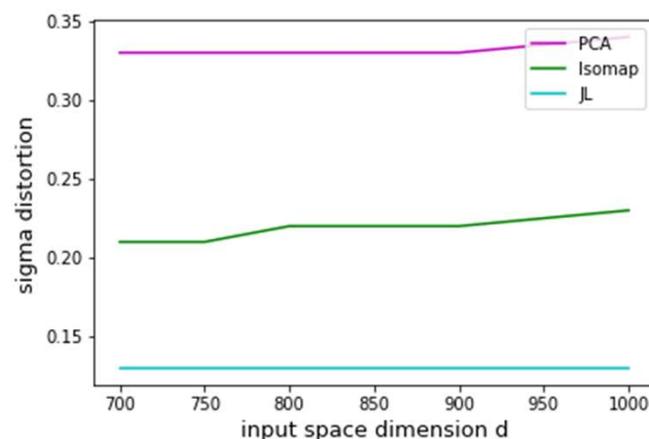


Fig. 2(b) σ -distortion as a function of original dimension d .

Empirical experiments: non - Euclidean dimensionality reduction

- In the last experiment, **Fig.3**, we tested the quality of our approximation algorithm on *non-Euclidean* input spaces versus the classical MDS and Isomap methods (adapted for non-Euclidean input spaces).
- The construction of the space is as follows:
 - a sampled Euclidean space X , of size and dimension $n = d = 100$, is generated as above;
 - the interpoint distances of X are distorted with a noise factor $1 + \epsilon$, with Normally distributed $\epsilon < 1$. We ensure that the resulting space is a valid non-Euclidean metric.
- We then embed the final space into $k \in [10,30]$ dimensions with $q = 5$. Since the non-Euclidean space is $1 + \epsilon$ far from being Euclidean, we expect a similar behavior to that shown in **Fig. 1(c)**.

The experiment clearly demonstrates the superiority of the JL-based approximation algorithm.

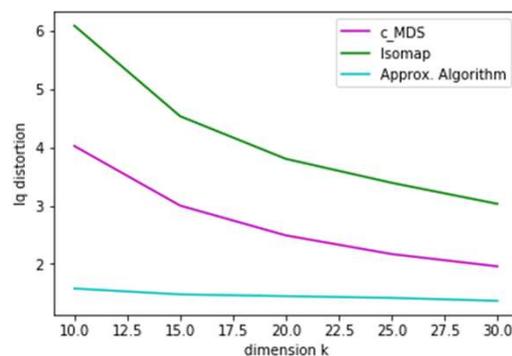


Fig. 3 non-Euclidean input metric: l_q -dist. behavior.

Discussion

- We initiate theoretical study of practically oriented *average case* quality measurement criteria. While often studied in practice, no theoretical studies have thus far attempted at providing rigorous analysis of these criteria. The vast majority of theoretical research has been devoted to analyzing the *worst case* behavior of embeddings, and therefore has little relevance to practical settings.
- We provide the first analysis of these, as well as the new distortion measure developed in [VL13] designed to possess machine learning desired properties. Moreover, we show that all considered measures can be adapted to possess similar qualities.
- We show nearly tight bounds on the absolute values of all distortion criteria, essentially showing that the JL transform is the optimal tool for dimensionality reduction.
- A phase transition, exhibited in our bounds, provides guidance on how to choose the target dimension k .
- Our bounds result in the *first* approximation algorithms for embedding any finite metric into k -dim Euclidean space, with provable approximation guarantees.
- Our theoretical findings are supported by the empirical experiments applied on various randomly generated Euclidean and non-Euclidean data sets.

References

- [ABN11] Ittai Abraham, Yair Bartal, and Ofer Neiman. Advances in metric embedding theory. *Advances in Mathematics*, 228(6):3026 – 3126, 2011.
- [Ach03] Dimitris Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671 – 687, 2003.
- [Alon09] Noga Alon. Perturbed identity matrices have high rank: Proof and applications. *Combinatorics, Probability & Computing*, 18(1-2):3–15, 2009.
- [AL10] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. *ACM Trans. Algorithms*, 9(3):21:1–21:12, June 2013.
- [Bado03] Mihai Badoiu. Approximation algorithm for embedding metrics into a two-dimensional space. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 434–443.
- [CC00] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling (Monographs on Statistics and Applied Probability)*. Chapman and Hall/CRC, 2nd edition, 2000.
- [CDKLM04] Russ Cox, Frank Dabek, Frans Kaashoek, Jinyang Li, and Robert Morris. Practical, distributed network coordinates. *SIGCOMM Comput. Commun. Rev.*, 34(1):113–118, January 2004.
- [CD06] Lawrence Cayton and Sanjoy Dasgupta. Robust Euclidean embedding. In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, pages 169–176.
- [DKS10] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson- Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 341–350. ACM, 2010.

[Dham04] Kedar Dhamdhere. Approximating additive distortion of embeddings into line metrics. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 96–104.

[HIL03] Johan Håstad, Lars Ivansson, and Jens Lagergren. Fitting points on the real line and its application to rh mapping. *J. Algorithms*, 49(1):42–62, October 2003.

[IM98] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98.

[JL84] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, pages 189–206.

[LLR95]:Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[LN16] K. G. Larsen and J. Nelson, "Optimality of the Johnson-Lindenstrauss Lemma," *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, Berkeley, CA, 2017, pp. 633–638.

[SXBL06] Puneet Sharma, Zhichen Xu, Sujata Banerjee, and Sung-Ju Lee. Estimating network proximity and latency. *Computer Communication Review*, 36(3):39–50, 2006.

[ST04] Yuval Shavitt and Tomer Tanel. Big-bang simulation for embedding network distances in euclidean space. *IEEE/ACM Trans. Netw.*, 12(6):993–1006, December 2004.

[VL18] Leena Chennuru Vankadara and Ulrike von Luxburg. Measures of distortion for machine learning. In *Advances in Neural Information Processing Systems 31*, pages 4891–4900. Curran Associates, Inc., 2018.

Appendix: Proof of Theorem [Moment analysis of JL transform]

Let $f: X \rightarrow \ell_2^k$ be the JL embedding (IM98 implementation). For every $u \neq v \in X$:

$$E_f \left[\left(\text{dist}_f(u, v) \right)^q \right] = E_f \left[\left(\max \left(\frac{\|f(u) - f(v)\|}{\|u - v\|}, \frac{\|u - v\|}{\|f(u) - f(v)\|} \right) \right)^q \right].$$

Since f is a linear map, it is enough to estimate for any $z \in \mathbb{R}^d$, with $\|z\| = 1$

$$E_f \left[\max \left(\|f(z)\|^q, \frac{1}{\|f(z)\|^q} \right) \right]. \text{ By the definition of the JL transform:}$$

- $f(z) = 1/\sqrt{k} \cdot T(z) = 1/\sqrt{k} \cdot (\langle z, T_1 \rangle, \dots, \langle z, T_k \rangle) = 1/\sqrt{k} (Y_1, \dots, Y_k)$, where $Y_i \sim N(0,1)$.
- $\|f(z)\|^q = (\|f(z)\|^2)^{q/2} = \left(\frac{X}{k}\right)^{q/2}$, where $X \sim \chi_k^2$.

Therefore,

$$E_f \left[\max \left(\|f(z)\|^q, \frac{1}{\|f(z)\|^q} \right) \right] \leq E_{X \sim \chi_k^2} [(X/k)^{q/2}] + E_{X \sim \chi_k^2} [(k/X)^{q/2}]. \text{ We estimate the expectations separately.}$$

goes to ∞ , as $q \rightarrow k$

$$\bullet E_{X \sim \chi_k^2}[(k/X)^{q/2}] = \int_0^\infty \left(\frac{k}{x}\right)^{\frac{q}{2}} \frac{x^{\frac{k}{2}-1}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right) e^{\frac{x}{2}}} dx = \frac{\left(\frac{k}{2}\right)^{\frac{q}{2}} \Gamma\left(\frac{k}{2} - \frac{q}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}$$

Gamma function: $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$, $\Gamma(t+1) = t \Gamma(t)$. We estimate:

- $E_f \left[\left(\ell_q - \text{contr}(f) \right)^q \right] = \left(1 + O\left(\frac{q}{k-q}\right) \right)^q$, for all $q < k$.
- $E_f \left[\left(\ell_q - \text{expans}(f) \right)^q \right] = \left(1 + O\left(\frac{q}{k}\right) \right)^q$, for all $q \geq 1$.

Therefore,

$$E_f[\ell_q - \text{dist}(f)] \leq 2^{\frac{1}{q}} \left(1 + O\left(\frac{q}{k-q}\right) \right) \leq \left(1 + \frac{1}{q} \right) \left(1 + O\left(\frac{q}{k-q}\right) \right) = 1 + O(1/\sqrt{k}),$$

taking $q = \sqrt{k}$.

■

Appendix: Lower bounds for ℓ_q -distortion for $q < k$

Theorem:

There is a finite Euclidean space Z , such that any $f: Z \rightarrow \ell_2^k$ must have

$$\ell_q\text{-dist}(f) = 1 + \Omega(q/(k - q)), \text{ for } \sqrt{k} \leq q < k.$$

Proof outline: Z is based on a lower bound example for worst case distortion of [LN16] + our claim that derives lower bound on ℓ_q -dist from a lower bound on w.c. distortion.

Claim[From worst case to average case distortion lower bound]:

Let X and Y be any metric spaces, such that for every embedding $f: X \rightarrow Y$ has $\text{dist}(f) \geq \alpha$.

Then, for every $N \geq n$ ($|X| = n$), there is a metric space $|Z| = N$, such that every embedding

$$F: Z \rightarrow Y, \text{ has } \ell_q\text{-dist}(F) \geq \left(1 + \frac{2(\alpha^{\frac{q}{2}}) - 1}{n^2}\right)^{1/q}.$$

Appendix: Lower bounds for additive measures, for $q < k$

Additive measures and REM: follow from lower bounds on Energy. We prove the following theorem in our paper:

Theorem (Energy is tight for any $q \geq 2$)

For any integer $k \geq 2$, for any $2 \leq q \leq k$, for $n(q, k)$ large enough, every embedding $f: E_n \rightarrow \ell_2^k$ has $\text{Energy}_q(f) = \Omega(\sqrt{q/k})$.

For $1 \leq q < 2$, every embedding $f: E_n \rightarrow \ell_2^k$ has $\text{Energy}_q(f) = \Omega((1/k)^{1/q})$.

Proof outline: Based on [Alon09], lower bound for the w.c. distortion.

[Alon09] develops lower bounds on the rank of a Gram matrix of the image vectors, assuming the distortion of each pair is at most $1 + \epsilon$.

We use these to bound rank of a Gram matrix raised to the power of **integer $q/2$** .

Appendix: Phase transition of ℓ_q -distortion at $q = k$

Theorem

- For any $k \geq 1$, any embedding $f: E_n \rightarrow \ell_2^k$ has ℓ_k -dist(f) = $\frac{\Omega(\sqrt{\log})^{\frac{1}{k}}}{k^{1/4}}$.
- For any $q > k \geq 1$, any embedding $f: E_n \rightarrow \ell_2^k$ has ℓ_q -dist(f) = $\Omega(n^{\frac{1}{2k} - \frac{1}{2q}})$.

Proof outline: It is enough to show that for any **non-expansive** $f: E_n \rightarrow \ell_2^k$, it holds that ℓ_k -dist(f) $\geq \frac{\Omega(\log n)^{\frac{1}{k}}}{\sqrt{k}}$, by the claim we prove in the paper:

Claim: If for any non-expansive $F: E_n \rightarrow Y$, ℓ_k -dist(F) $\geq D(k, n)$, then for any $f: E_n \rightarrow Y$,
 ℓ_k -dist(f) $\geq \text{const} \cdot \sqrt{D(k, n)}$.

Since $\ell_2^k \sim \ell_\infty^k$ with distortion \sqrt{k} , it is enough to prove that any non-expansive embedding $f: E_n \rightarrow \ell_\infty^k$ has ℓ_k -dist(f) $\geq \Omega(\log n)^{1/k}$.

Appendix: Phase transition of ℓ_q -distortion at $q = k$

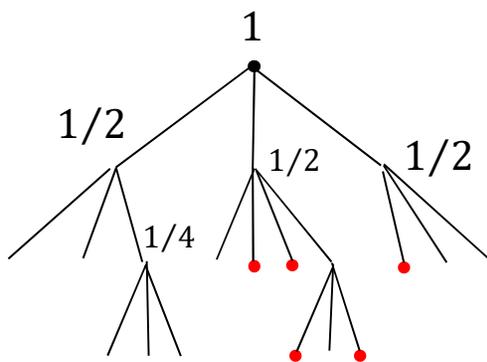
Claim: For every non-expansive $f: E_n \rightarrow \ell_\infty^k$, $\ell_k\text{-dist}(f) \geq \Omega(\log n)^{1/k}$.

Proof outline: Basically, embedding (non-expansively) E_n into ℓ_∞^k is as embedding it (non-expansively) into a family of certain tree metrics.

2-HST metrics of degree k – a family of all rooted trees on n leaves, with each node having at most 2^k children.

The nodes have labels, decreasing by a factor of 2 along the paths from the root to each leaf.

The root's label is 1.

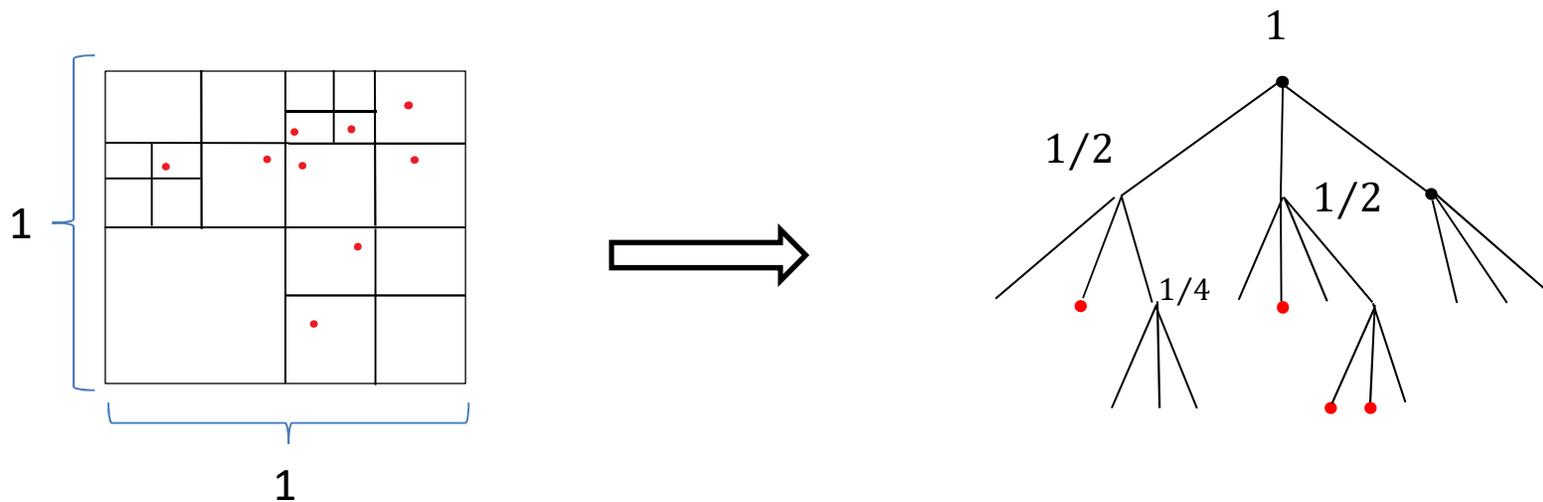


Each such tree defines a metric over the set of its **leaves**: $\text{dist}(u, v) = \text{label}(\text{lca}(u, v))$.

Appendix: Phase transition of ℓ_q -distortion at $q = k$

We show that every **non-expansive** embedding $f: E_n \rightarrow \ell_\infty^k$ can be modified to the one that embeds E_n into a 2-HST tree from the family of degree k , with better ℓ_k -distortion.

By induction on $k \geq 2$. For $k = 2$, recursively construct 2-HST tree, of degree 4.



Appendix: Phase transition of ℓ_q -distortion at $q = k$

So, it is enough to prove that:

Claim

Any **non-expansive** embedding f of E_n into a family of 2-HST's of degree k , has $(\ell_k - \text{dist}(f))^k \geq \Omega(\log n)$.

Proof outline: By induction on n , showing that the best tree is the perfectly balanced (each node has exactly 2^k children). Computing its weight completes the proof.



The complete proofs of all the theorems in claims are presented in the full version of the paper.