Generalizations of the Johnson-Lindenstrauss embedding lemma

Participant:

Moriya Bitton

Yuval Ben Yaakov

Itay Yosef

Meytar Gil-Ron

Introduction:

Johnson-Lindenstrauss lemma grew out of the intuition that high dimensional information can be represented in a much lower dimensional.

Problem: ***We have n vectors***: $\{x_1 \dots x_n\}$, ***which n is extremely higher.***

The lead hypothesis is that the data is usually not uniformly distributed; rather the data concentrate on some structure, which means that our problem is much less complex.

Practically speaking the original storing $P(n\ vectors\ from\ R^n)$ is usually $O(n^2)$ but after the J-L lemma we can store P in $O(N \log n)$, this improvement is and can be a great asset for many geometric algorithms.

Lemma:

Let $\varepsilon \in \mathbb{R}\ in\ range\ (0, 0.5)$ and let $P = \{p_1, p_2, \dots, p_n\}$ be a set of n points in R_n.

Let $k \in \mathbb{N}$ with $k \geq c\varepsilon^{-2} \log n$, were c is a sufficiently large absolute constant.

Then there exists a mapping

$$f: R_n \to R_k\ s.t.\ (1 - \varepsilon)\|p_i - p_j\| \leq \|f(p_i) - f(p_j)\| \leq (1 + \varepsilon)\|p_i - p_j\|\ for\ \forall i, j \in [1, n],$$

$\|.\|$**denotes the Euclidean norm**.

In this survey, we will examine five different articles which all relate and each one improved the JL lemma in their way. Another thing that is done in this survey is to examine the improvements of the articles over one another.

The articles:

- *Database-friendly random projections: Johnson-Lindenstrauss with binary coins, by D. Achioptas.*
- *Near-optimal dimensionality reductions that preserve volume, by A. Magen and A. Zouzias.*
- *The Johnson-Lindenstrauss lemma and the sphericity of some graphs*, by A. Frankl and H. Maehara.
- On variants of the Johnson–Lindenstrauss lemma, by J. Matoušek
- A Sparse Johnson–Lindenstrauss Transform, by A. Dasgupta, R. Kumar, and T. Sarlós.

The Johnson-Lindenstrauss lemma and the sphericity of some graphs

By A. Frankl and H. Maehara.

1986

In my account, I will represent the result of the article "The Johnson-Lindenstrauss lemma and the sphericity of some graphs", by A. Frankl and H. Maehara.

Motivation:

We would like to reduce the dimensional with minimum distortion of our data, to be more specific we want to preserve the essence of the data. Despite the multiple features that can be considered as the essence of the data (pairwise distance, norm, correlation, and so), the J-L lemma attributes the highest importance to the pairwise distance (the reason is that many of the local systems rely on it, like linear regression, nearest neighbors, KNN, and so).

This result of the J-L lemma applied to show us:

1. ***if G is a graph on n vertices and with smallest eigenvalue λ, then its sphericity $sph(G) \leq c\lambda^2 \log(n)$.***
2. ***if $G \vee \bar{G}$ is a forest, then $sph(G) \leq c \log(n)$ holds.***

To prove the J-L lemma the article represents the following result in the above order:

1. **Some Upper bounds on the Sphericity of graph**

   Thanks to our motivation we can state a slightly improved version of the J-L lemma.

   Lemma:

   $for\ \varepsilon\ \left(0 < \varepsilon < \frac{1}{2}\right)\ and\ n\ (n \in \mathrm{N})\ let\ k(n,\varepsilon) = \lceil 9\ \left(\varepsilon^2 - 2\varepsilon^3/_3\right)^{-1} \log n \rceil + 1.$
   $if\ n > k(n,\varepsilon)^2, then\ for\ any\ n - points\ set\ S\ in\ R^n,$ there exists a map $f: S \longrightarrow R^{k(n,\varepsilon)}$
   $s.t\ (1 - \varepsilon)||u - v||^2 < ||f(u) - f(v)||^2 < (1 + \varepsilon)||u - v||^2\ for\ all\ u, v \in S.$

   - We are going to apply this lemma to the sphericity problem.
     The sphericity of $graph\ G(V,E), sph(G)$, is the smallest n s.t. there is an embedding
     f: V $\longrightarrow$ R^n s.t $0 < ||f(u) - f(v)||^2 < 1$ iff $(u, v) \in E$.
     An eigenvalue of graph G is the eigenvalue of the adjacency matrix A(G).

   1.1. **Theorem 1**

   Let G be a graph with minimum eigenvalue $\lambda_{min} \geq -c$ $(c \geq 2)$ and suppose that $|G| > [12(2c-1)^2 \log|G|]^2$, then $sph(G) < 12(2c-1)^2 \log|G|$.

   1.1.1. **Corollary 1**

   Let G be a graph with maximum dgree d and suppose $|G| > [12(2d-1)^2 \log|G|$, then $sph(G) < 12(2d-1)^2 \log|G|$.

   1.1.2. **Corollary 2**

   Let G be a graph with m edges, then $sph(L(G)) < 108 \log(m)$ for $m > (108 \log(m))^2$.

1.2. Theorem 2

Let T be a tree with sufficiently large order, then $sph(T) < 105\log|T|$.

1.3. Theorem 3

for any forest F $sph(\bar{F}) \leq 8\lceil \log|F| \rceil$.

2. A simple short proof of the Johnson-Lindenstrauss lemma

Let $\boldsymbol{V}$ *be a vector in* R^n, *and* $\boldsymbol{H}$ *a "random k-dimensional subspace" through the origin, and let us define the random variable* $\boldsymbol{X}$ *as the square lengh of the projection of* V *onto* H.

2.1. Proposition

Suppose $\frac{1}{2} > \varepsilon > 0, n > k^2, k > 24\log(n) + 1,$

then $P_\varepsilon = P\left(\left|X - \frac{k}{n}\right| > \varepsilon\frac{k}{n}\right) < 2\sqrt{k}\exp\left(-(k-1)\left(\frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}\right)\right).$

2.1.1. Formula (1)

$$V_n = \int_0^{\frac{\pi}{2}} V_k\,(\cos\theta)^{k-1} V_{n-k}(\sin\theta)^{n-k-1}\,d\theta \quad (valide\ for\ all\ 1 \leq k \leq n).$$

θ is the angle between v and H (which means $X = \cos^2\theta$).

- We use this formula to prove level 2.1.

2.2. Proof of the Lemma

- After we embed our vectors in the new dimension, we claim that the pairwise distance between them stays in a range of $[1-\varepsilon, 1+\varepsilon]$, to obtain the preserve the essence of our data.
- In level 2.1, we calculated the Upper bound for probability to the deviation of the expectation for one pair of Projection vectors embedding.
- Hence, to complete the proof we would like to obtain our embedding by showing the Upper bounds of the probability that exists a pair of vectors (from all possible pairs $\binom{n}{2}$) same as above.
- We obtain desired embedding of *S in k-dimensional*. ■

Conclusion:

We modeled a subspace from a high-dimensional problem, and claim about any embedding data that the distortion, by pairwise distance, will be in a low defined range $[1-\varepsilon, 1+\varepsilon]$.
Hence, the probability for deviation out of our range, for any possible pair of vectors, blocking with a very small number on top.

Database-friendly random projections: Johnson-Lindenstrauss with binary coins, by D. Achiloptas. 2001

In my account, I will represent the result of the article "Database-friendly random projections Johnson-Linden Strauss with binary coins".

As a result of Johnson and Lindenstrauss, we know that any set of n points in d-dimensional Euclidean space can be embedded into k-dimensional Euclidean space, where k is logarithmic in n and independent of d so that all pairwise distances are maintained within an arbitrarily small factor.

for producing JL-embeddings, including ours, the heart of the matter is showing that for any vector, the squared length of its projection is sharply concentrated around its expected value.

A proof for that was greatly simplified and sharpened by Frankl and Maehara-
" The Johnson-Lindenstrauss lemma and the sphericity of some graphs".

Problem:

All known constructions of such embeddings involve projecting the n points onto a spherically random k-dimensional hyperplane through the origin. While this is conceptually simple, in practical terms it amounts to multiplying the input matrix A with a dense matrix of real numbers. This can be a non-trivial task in many practical computational environments.

Motivation:

We want to find constructions of such embeddings with the property that all elements of the projection matrix belong in {-1,0,1}. Such constructions are particularly well suited for database environments, as the computation of the embedding reduces to evaluating a single aggregate over k random partitions of the attributes.

The research contribution:

Theorem 1.1. *Let P be an arbitrary set of n points in $\mathbb{R}^d$, represented as an $n \times d$ matrix A. Given $\varepsilon, \beta > 0$ let*

$$k_0 = \frac{4+2\beta}{\varepsilon^2/2 - \varepsilon^3/3} \log n.$$

let R be a d x k random matrix with $R(i,j) = rij$;
where $\{rij\}$ are independent random variables from the following probability distribution:

$$r_{ij} = \sqrt{3} \times \begin{cases} +1 & \text{with probability } 1/6, \\ 0 & \text{with probability } 2/3, \\ -1 & \text{with probability } 1/6. \end{cases}$$

Let

$$E = \frac{1}{\sqrt{k}} AR$$

and let $f : \mathbb{R}^d \to \mathbb{R}^k$ map the ith row of A to the ith row of E.
With probability at least $1 - n^{-\beta}$, for all $u, v \in P$

$$(1-\varepsilon)||u-v||^2 \leqslant ||f(u) - f(v)||^2 \leqslant (1+\varepsilon)||u-v||^2.$$

This probability distribution gives an additional threefold speedup as we only need to process a third of all attributes for each of the k coordinates.

Conclusion:

the best known bound for k is: $k \geqslant k_0 = (4+2\beta)(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \log n$

Near-optimal dimensionality reductions that preserve volume,
by A. Magen and A. Zouzias.
2008

Motivation:

The motivation of this paper is to construct such an embedding of the Johnson Linderstrauss lemma that would preserve the volume for any sub-group, $where\ |S| < K$.
In other terms Create a projection that even after the dimensionality reduction it would preserve the volume of any sub-group S of points from the dataset X up to an ε.

Mathematically speaking this would entail this inequality $1 - \varepsilon < \left(\frac{vol(f(S))}{vol(S)}\right)^{\frac{1}{|S|-1}} < 1 + \varepsilon$
where f is the embedding and vol is the volume function.

The Results:

In this paper, they show that it is indeed sufficient to do the reduction onto a Dimension of $(max\{\frac{k}{\varepsilon}, \varepsilon^{-2}\log n\})$.
This is an improvement to previous work that guaranteed the same thing but with $O(k\varepsilon^{-2}\log n)$ reduction.

The importance of the results:

One thing that these results show is that using this paper one could construct a projection to a lower dimension and preserve the rich volume that any subset of points S has.
This means that if before we could say that the relationship of every pair of vectors would stay the same up to a factor of ε, now we can say that for any subset S.

The intuition behind the results:

From this, we can conclude that using such an embedding for solving problems in higher dimensions would yield better and more accurate results since more information about the complex relationship of the original dataset was saved using this technique.

On variants of the Johnson–Lindenstrauss lemma,
by J. Matoušek
2008

the on variants of J-L Lemma discusses the on variants that the Matrix A of the linear mapping T, that suits that T will be suitable for the lemma.

the different variants:

1. Johnson and Lindenstrauss - choosing $A(kxn)$ $k\ rows\ as\ orthonormal\ vectors\ in\ R^n$.
2. Indyk and Motwani - said that you can drop off the orthogonality of the A entries, just random variables.
3. Dasgupta and Gupta - same as Indyk and Motwani but each vector is a unit vector.
4. Achlioptas - showed that the variants of A can be chosen to be -1 or 1 (that gives us a much better way to compute A (3 times faster).
1. he also showed you can pick the variants of A to be 0 with 2/3 probability and -1 or 1 with 1/6 probability each.
5. Ailon and Chazelle - improved Achlioptas idea by constructing A as the product of MHD (Matrix M, H, D) one obstacle they overcame with this idea was that A could be too sparse (almost every variant of A was 0)

main proof results of each stage:

He merged achiloptas and Ailon and Chazelle concepts.

The article results for this section were similar to Ailon and Chazelle's results but in this article, the distribution for the matrix variants is easier to compute (1,-1).

A Sparse Johnson–Lindenstrauss Transform, by A. Dasgupta, R. Kumar, and T. Sarlós. 2010

All the results cited by the article:

The team obtained a sparse random projection matrix of the size $K \times D$.
The number of zeros in that alleged matrix is at most
O ($\frac{1}{\varepsilon} log^2$ (k /δ) log(1 /δ)) , and K is $O\left(\varepsilon^{-2} log\left(\frac{1}{\delta}\right)\right)$.
For the team's results to be an improvement they must assume $log^2\left(\frac{k}{\delta}\right) = o\left(\frac{1}{\varepsilon}\right)$.
Another thing worth mentioning about this construction is that they didn't use random variables as traditional, instead, they used hash function.
The use of a hash function means that there is a little dependency between the matrix values, something that they had to adapt to and solve.
The main reason to use a hash function is that they could claim that the number of non-zeros entries is fixed in each column.
Using the block-Hadamard based preconditioner can help achieve a running time of $\tilde{O}\left(min\left(d, \frac{n_{nz}}{\varepsilon}\right)\right)$, where n_{nz} is the number of non-zero entries in a vector.
3 main theorems were proven in the article.
The first one proves that in the probability of $1 - 4\delta$ the resulting matrix $\boldsymbol{\varphi}$ doesn't change the $2 - norm$ of the vector x up to an ε of the original value.
The second theorem promises the same thing but with probability $1 - 3\delta$ for a matrix H.
The last theorem gives that exists a preconditioner G such that the product of H from the second theorem and G satisfies the condition that is promised by the first and second theorems, and states that the time for computing HGx is $O(min(\frac{n_{nz}(x)}{\epsilon} log^4(\frac{1}{\epsilon\delta}), d) log(\frac{1}{\delta\epsilon}))$.

Note that picking a value too small for lambda can result in slowing down the run time of theorem 3 for dense vectors, but it is still comparable to all the best existing rival methods, but for sparse vectors, it makes the computation time faster.

The importance of the results encountered:

The importance is that with this construction of a sparse projection matrix the running time of the method of Johnson and Linderstrauss can be bounded with something better than previously thought. Meaning we can take a dataset X and project it to a lower dimension much faster than we could before. These results were already proven to help speed up the nearest-neighbor computation time for sparse vectors.

The intuition underlying the results:

They are two main intuitions, the first is that the projection matrix could be made sparse without it hurting the special conditions of the lemma.
The second is that the construction of the projection matrix doesn't have to be made of standard gaussian variables, it could be also a random hash function causing a little dependence between the entries of the matrix.

The Relation to other articles:

This article improved Matousek in "On variants of the Johnson–Lindenstrauss lemma".
They improved time achieved by Matousek for projecting a vector x with $n_{nz}(x)\ non\ zero\ enteries$ by a factor of $\frac{1}{\varepsilon}$.
The results in this article are $\tilde{O}\left(\frac{n_{nz}(x)}{\varepsilon}\right) for\ all\ vectors$ while Matousek achieved $\frac{n_{nz}(x)}{d}\ for\ sparse\ vectors\ only$.