CPS296.2 Geometric Optimization **April 12, 2007**

Lecture 25: Johnson Lindenstrauss Lemma

Lecturer: Pankaj K. Agarwal *Scribe: Albert Yu*

The topic of this lecture is dimensionality reduction. Many problems have been efficiently solved in low dimensions, but very often the solution to low-dimensional spaces are impractical for high dimensional spaces because either space or running time is exponential in dimension. In order to address the curse of dimensionality, one technique is to map a set of points in a high dimensional space to another set of points in a low-dimensional space while all the important characterisitics of the data set are preserved. In this lecture, we will study **Johnson Lindenstrauss Lemma**. Essentially all the dimension reduction techniques via random projection rely on the Johnson Lindenstrauss Lemma.

25.1 Johnson Lindenstrauss Lemma

The goal of Johnson Lindenstrauss Lemma is to, for a point set P in $\mathbb{R}^d$, find a point set Q in R^k and a mapping from P to Q, such that the pairwise distances of P are preserved (approximately) under the mapping.

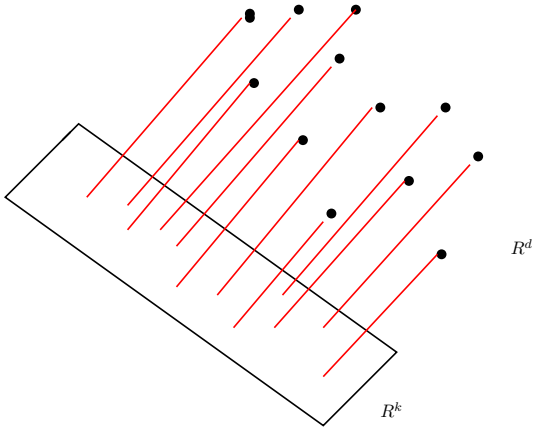

Figure 25.1: Projection

Definition 1 *Given a set of n points in $\mathbb{R}^d$ and a projection onto a random k-dimension linear subspace, a distance $||p-q||_2$ is **ϵ-preserved** if $(1-\epsilon)||p-q||_2 \leq \sqrt{n/k}||f(p)-f(q)||_2 \leq (1+\epsilon)||p-q||_2$ for any $0 < \epsilon < 1$.*

Theorem 1 (Johnson-Lindenstrauss Lemma) *Let P be a set of n points in $\mathbb{R}^d$, let $\epsilon > 0$ be a parameter,*

and let $k = (1/\epsilon^2)\log n$. Let Q be the projection of P onto a random k-dimension linear subspace. Then all pairwise distanced in P are ϵ-preserved in Q with probability at least $1/2$.

To prove Theorem 1, we only have to prove that for any random k-dimensional subspace, where $k = O((1/\delta^2)\log(1/\delta))$, a particular distance is preserved with probability $1-\delta$. Then using the simple union bound, all n^2 distances are preserved with probability $1 - n^2\delta$. If we choose δ to be $1/2n^2$, all n^2 distances will be preserved with probability $1/2$.

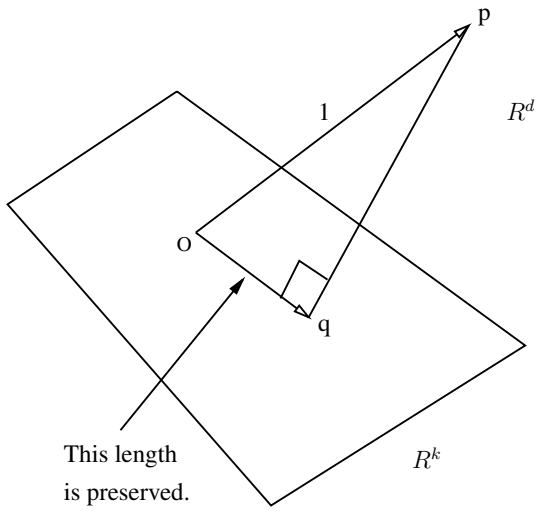

Figure 25.2: The length of p is preserved

We want to estimate the length of a unit vector in R^d when it is projected onto a random k-dimensional space. For simplicity, instead of choosing a random subspace, we fix a k-dimensional subspace to be the space spanned by the first k coordinate vectors and choose a random unit vector in R^d. Then we measure the length of the random unit vector projected down onto the subspace. We argue that projection of the random vector is highly concentrated around $\sqrt{k/d}$.

Distortion of distances

Each point p is represented in the form of $(x_1, x_2, \cdots, x_d)$, where $x_1, x_2, \cdots, x_n$ are random variables which are independently chosen from $N(0,1)$. All x_i have the same distributions which implies that all $E(x_i^2)$ are the same. By linearity of expectations and symmetry, $E[x_i^2] = 1/d$ for $1 \le i \le d$. Hence, the expected length of vector $\vec{oq}$, $E(||q||_2^2)$, is $E(\sum_{i=0}^{k} x_i^2) = \sum_{i=1}^{k} E(x_i^2) = k/d$.

The expected length of vector $\vec{op}$, $E(||p||_2^2)$, is 1 because it is a unit vector. Therefore, $\sum_{i=1}^{d} E(x_i^2) = 1$.

25.1.1 Tail bounds

Theorem 2 *For $\beta < 1$, $Pr[||q||^2 \le \beta(k/d)] \le \beta^{k/2}(1 + \frac{(1-\beta)^k}{d-k})^{(d-k)/2}$*

Theorem 3 *For $\beta > 1$, $Pr[||q||^2 \ge \beta(k/d)] \le \beta^{k/2}(1 + \frac{(1-\beta)^k}{d-k})^{(d-k)/2}$*

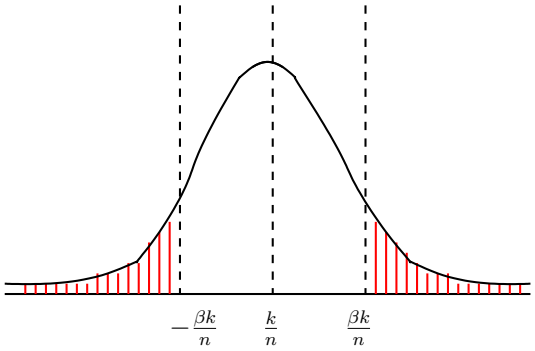

Figure 25.3: Gaussian normal distribution: The two inequalities in theorem 2 and 3 imply that the red regions are very small.

There are two ways to prove the tail bounds.

Method 1:

$$
\begin{aligned}
& \Pr[\sum_{i=1}^{k} x_i^2 \leq \beta(k/d)] \\
= \quad & \Pr[\sum_{i=1}^{k} x_i^2 \leq \beta(k/d) \sum_{i=1}^{d} x_i^2] && (\text{since } \sum_{i=1}^{d} x_i^2 = 1) \\
= \quad & \Pr[d \sum_{i=1}^{k} x_i^2 \leq \beta k \sum_{i=1}^{d} x_i^2] \\
= \quad & \Pr[\sum_{i=1}^{k} (\beta k - d) x_i^2 + \sum_{i=k+1}^{d} \beta k x_i^2 \geq 0] \\
= \quad & \Pr[\exp(t(\sum_{i=1}^{k} (\beta k - d) x_i^2 + \sum_{i=k+1}^{d} \beta k x_i^2)) \geq 1] && (\text{for } t > 0) \\
\leq \quad & E[\exp(t(\sum_{i=1}^{k} (\beta k - d) x_i^2 + \sum_{i=k+1}^{d} \beta k x_i^2))] && (\text{by Markov's inequality}) \\
= \quad & \prod_{i=1}^{k} E[\exp(t(\sum_{i=1}^{k} (\beta k - d) x_i^2))] \prod_{i=k+1}^{d} E[\exp(t\beta k x_i^2))] \\
= \quad & \prod_{i=1}^{k} 1/\sqrt{1 - 2t(\beta k - d)} \prod_{i=k+1}^{d} 1/\sqrt{1 - 2t\beta k}. && (\text{since } E[e^{s x_i^2}] = 1/\sqrt{1-2s}, \forall s < 1/2)
\end{aligned}
$$

The problem is now reformulated as a problem of choosing t to minimize:

$$f(t) = \prod_{i=1}^{k} 1/\sqrt{1 - 2t(\beta k - d)} \prod_{i=k+1}^{d} 1/\sqrt{1 - 2t\beta k}.$$

By differentiating f, we get the minimum of f to be $\beta^{k/2}(1+\frac{(1-\beta)^k}{d-k})^{(d-k)/2}$.

See [2] for details.

A more convenient form of Theorem 2:

Set $\beta = (1-\epsilon) < 1$ and assume $\Pr[||q||^2 \leq \beta(k/d)] \leq \beta^{k/2}(1+\frac{(1-\beta)^k}{d-k})^{(d-k)/2}$ is true.

$$\begin{aligned}
& \Pr[||q||^2 \leq \beta(k/d)] \\
\leq\ & \beta^{k/2}(1+\frac{(1-\beta)^k}{d-k})^{(d-k)/2} \\
\leq\ & \beta^{k/2}\exp(\frac{(1-\beta)k}{d-k}(d-k)/2) && (\text{since } 1+x \leq e^x) \\
=\ & \exp(\ln \beta k/2)\exp((1-\beta)k/2) \\
=\ & \exp(k/2(1-\beta+\ln\beta)) \\
=\ & \exp(k/2(\epsilon+\ln(1-\epsilon))) && (\text{since } \beta = 1-\epsilon) \\
\leq\ & \exp(k/2(\epsilon-\epsilon-\epsilon^2/2)) && (\text{since } ln(1-\epsilon) \leq -\epsilon-\epsilon^2/2) \\
=\ & \exp(-k\epsilon^2/2) \leq \delta.
\end{aligned}$$

Therefore, the inequalities in Theorems 2 and 3 will hold if $k = (2/\epsilon^2)\ln(1/\delta)$. If we set $\delta \approx 1/(2n^2)$, we get $k = O((1/\epsilon^2)\log(n))$.

Method 2: Geometric Proof

We measure the mass concentration on a high-dimensional unit sphere S^{d-1}. It can be shown that when d becomes large, most points on S^{d-1} concentrate around the equator. This is known as the "measure concentration phenomenon". A more general form of measure concentration result is that 1-Lipschitz functions are highly concentrated around the median.

Let function $f : S^{d-1} \to R$ be this function

$$f(x_1, x_2, \cdots, x_d) = \sqrt{x_1^2 + x_2^2 + \cdots + x_k^2},$$

i.e., it is the length of the projection of $x = (x_1, x_2, \cdots, x_d)$ on the subspace spanned by the first k coordinates.

If the orthogonal projection, p, is 1-Lipschitz, then function f is also 1-Lipschitz by triangle inequality:

$$|f(x) - f(y)| = |\|p(x)\| - \|p(y)\|| \leq \|p(x) - p(y)\| \leq 1|x-y|$$

For $\beta < 1$, there exists M_f (median of f) such that

$$Pr[f(x) < M_f - t] \leq e^{-t^2 d},$$

$$Pr[f(x) < M_f + t] \leq e^{-t^2 d}.$$

For $t \approx \epsilon\sqrt{k/d}$, the probability measure on S^{n-1} becomes $e^{-k\epsilon^2}$.

Since $E[f(x)^2] = k/d$, the median is close to $\sqrt{k/d}$.

Given $M_f \approx \sqrt{k/d}$ (M_f is approximately equal to $\sqrt{k/d}$), we have

$$Pr[f(x) < \sqrt{k/d} - \epsilon\sqrt{k/d}] \leq e^{-k\epsilon^2},$$

$$Pr[f(x) < \sqrt{k/d} + \epsilon\sqrt{k/d}] \leq e^{-k\epsilon^2}.$$

See [5] for details.

25.2 Applications of Johnson-Lendenstrauss Lemma

1) ϵ-approximation nearest neighbor (ϵ -NN) in high dimension.

Problem statement: Given a set of n points $P = \{p_1, p_2, \cdots, p_n\}$ in R^d, find the point in P which is ϵ-closest to a query point $q \in R^d$. That is, $\forall p' \in P$, $d(p, q) \leq (1 + \epsilon)d(p', q)$.

Naive solution: We simply store all n points. When a query point q is given, compute the distance from each point in P.

Space requirement:	$O(nd)$
Query time:	$O(nd)$

Solution 2: If we are allowed to use more space, then the query time can be much faster than that of the naive method.

Space requirement:	$O(n^{(1/\epsilon^2)\log(1/\epsilon)})$
Query time:	$O(d \cdot \log(n)/\epsilon^2)$

Solution 3: If we only use near-linear space, we can get sublinear query time.

Space requirement:	$O(nd)$
Query time:	$O(dn^{1/(1+\epsilon)})$

Next, we sketch solution 2.

Definition 2 *PLEB (point location among equal balls): Given a set of n points $P = \{p_1, p_2, \cdots, p_n\}$ and a query point q, return "yes" and the point $p_i \in P$ if $q \in B(p_i, r)$, else return "no".*

Definition 3 *ϵ-PLEB (ϵ-point location among equal balls): Given a set of n points $P = \{p_1, p_2, \cdots, p_n\}$ and a query point q,*

if ($\exists p_i \in P$ such that $q \in B(p_i, r)$)
return "yes" a point p_i' such that $q \in B(p_i', (1 + \epsilon)r)$,
else if ($\forall p_i \in P$, $q \notin B(p_i, (1 + \epsilon)r)$)
return "no",

else if ($r \leq d(q, p_i) \leq ((1+\epsilon)r)$ *where* p_i *is the closest point to* q)
return either "yes" or "no".

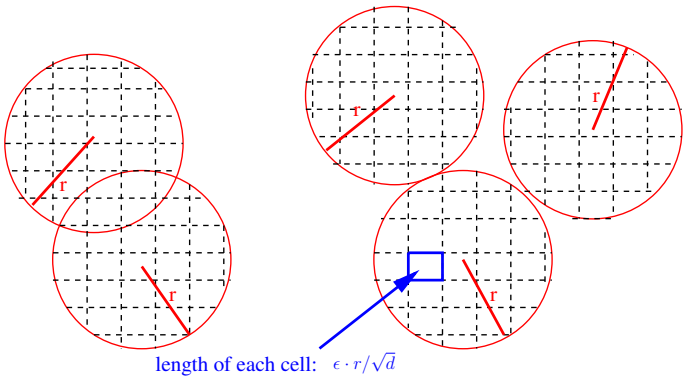

Figure 25.4: Each ball is decomposed into a bounded number of cells (cubes). Given a query point q, if q lies inside a cube, it also lies inside a ball.

Indyk and Motwani [4] showed the reduction from ϵ-NN to ϵ-PLEB using a data structure called a *ring-cover* tree. A simpler reduction is given by Har-Peled [3]. The ϵ-PLEB problem can be solved by the following algorithm.

Algorithm: Each ball is decomposed into a bounded number of cubes. We take all cubes and store them into a hash table. Given a query point q, we figure out which cube contains the point and then check if the cube that contains q is in hash table. If so, we can locate the ball. Otherwise, it does not lie in any ball. Note that a cube can be in more than one ball if the balls intersect one another.

Query time:	$O(d)$
Space requirement:	$O(n \times (1/\epsilon)^d)$

The query time is $O(d)$ because we can access a cube in the hash table in $O(1)$ time and figuring out cube that contains q takes $O(d)$ time.

For the space requirement, there are totally n balls. Each ball contains $\frac{Vol(Br)}{Vol(c_{\epsilon n}/\sqrt{d})}$ cubes. Therefore, the space required for hash table is $n \times \frac{Vol(B_r)}{Vol(c_{\epsilon r}/\sqrt{d})} \approx n \times (1/\epsilon)^d$.

By Johnson-Lindenstrauss Lemma, we can reduce P to Q which is a set of points in R^k, where $k = (1/\epsilon^2)\log(n)$. The query time is the sum of the time for projection and the query time in dimension k. Therefore, the query time is $dk + k = O(dk) = O(d \cdot \log(n)/\epsilon^2)$. The space requirement is $n \times (1/\epsilon)^k = n \times (1/\epsilon)^{(1/\epsilon^2)\log(n)} = n^{(1/\epsilon^2)\log(1/\epsilon)}$.

Query time:	$O(d \cdot \log(n)/\epsilon^2)$
Space requirement:	$O(n^{(1/\epsilon^2)\log(1/\epsilon)})$

Recently Ailon and Chazelle showed that the projection can be done in $\tilde{O}(d+k)$ time instead of $O(dk)$ time.

See [1, 4] for details.

Other applications:

Johnson Lindenstrauss Lemma has also been applied in image processing. The features of images are represented as high dimensional vectors. The dimension (number of features) can be up to thousands depending on applications. With dimension reduction techniques, we can compress the vectors while the similarity between any two vectors are preserved. Thus, we can choose to do analysis in low dimensional space.

Another application is a faster approximation for low rank matrix. In the field of data mining and machine learning, a matrix A is usually approximated by a matrix of low rank, A_r, which only contains the important and meaningful characteristics of the data. This approximation is done by the singular value decomposition (SVD). However, the computation of SVD is very expensive. An efficient way for doing low rank approximations is to first apply random projection (Johnson Lindenstrauss Lemma) to reduce dimensionality (ex: from thousands to hundreds) and then apply spectral projection (SVD) to further reduce dimension.

References

[1] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. *Proc. 38th STOC*, pages 557 - 563, 2006. 25-6

[2] S. Dasgupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. *Technical Report 99-006*, UC Berkeley, March 1999. 25-4

[3] S. Har-Peled. A replacement for Voronoi diagrams of near linear size. *Proc. 42nd Annu. IEEE Sympos. Found. Comput. Sci.*, pages 94–103, 2001. 25-6

[4] P. Indyk and R. Motwani Approximate nearest neighbors: Towards removing the curse of dimensionality. *Proc. 30th Symposium on Theory of Computing*, pages 604–613, 1998. 25-6

[5] J. Matousek. Lectures on Discrete Geometry. *Springer*, Verlag 2002, Heidelberg. 25-5