Dimension reduction

Roie Salama

January 14, 2016

The concept of this chapter is the idea of reducing the dimension of the points in a given set $P \in \mathbb{R}^n$ such that the distances are preserved approximately. In this chapter, we will go through a series of theorems toward proving a theorem called "The Johnson-Lindenstrauss Lemma" which tells us we can do so (reducing the dimension while approximately preserving the distances) by simply projecting the set into random subspaces.

We shall start with 2 simple definitions :

Definition 1. (translation) : Let $A \in \mathbb{R}^d$ and $p \in \mathbb{R}^d$. So the following set:

$$A + q = \{a + q \mid a \in A\}$$

is the translation of A by q.

Definition 2. (Minkowsky sum) : Given $A, B \in \mathbb{R}^d$, the minkowski sum of A and B is given by:

$$A + B := \{a + b \mid a \in A, b \in B\}$$

Theorem 3. *(Brunn-Minkowsky) : Let $A, B \neq \phi$ be compact sets of $\mathbb{R}^d$. So :*

$$Vol(A + B)^{\frac{1}{n}} \geq Vol(A)^{\frac{1}{n}} + Vol(B)^{\frac{1}{n}}$$

before proving the theorem, we shall define the following:

Definition 4. $A \subset \mathbb{R}^n$ is a brick set if $A = \cup_{i=1}^{c} B_i$ where B_i are parallel to the axis boxes with disjoint interiors :

Proof. (for theorem 3) :

Sufficient to prove for brick sets by the definition of volume as a limit of brick sets approximation.

The proof is by induction on the total number of bricks in A and B denoted by k.

if $k = 2$ than $A = \alpha$ and $B = \beta$ where α and β are bricks with dimensions $\alpha_1, .., \alpha_n$ and $\beta_1, .., \beta_n$. So $A + B = \alpha + \beta$ is a brick with the dimensions $\alpha_1 + \beta_1, .., \alpha_n + \beta_n$.

So we need to prove:

$$(\prod_{i=1}^{n} \alpha_i)^{\frac{1}{n}} + (\prod_{i=1}^{n} \beta_i)^{\frac{1}{n}} \leq (\prod_{i=1}^{n} \alpha_i + \beta_i)^{\frac{1}{n}}$$

equivalently we will prove:

$$\frac{(\prod_{i=1}^{n} \alpha_i)^{\frac{1}{n}} + (\prod_{i=1}^{n} \beta_i)^{\frac{1}{n}}}{\prod_{i=1}^{n} \alpha_i + \beta_i)^{\frac{1}{n}}} \leq 1$$

Now:

$$\frac{(\prod_{i=1}^{n} \alpha_i)^{\frac{1}{n}} + (\prod_{i=1}^{n} \beta_i)^{\frac{1}{n}}}{\prod_{i=1}^{n} \alpha_i + \beta_i)^{\frac{1}{n}}} = (\prod_{i=1}^{n} \frac{\alpha_i}{\alpha_i + \beta_i})^{\frac{1}{n}} + (\prod_{i=1}^{n} \frac{\beta_i}{\alpha_i + \beta_i})^{\frac{1}{n}}$$

by the arithmetic-geometric mean inequality :

$$\leq \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_i + \beta_i}{\alpha_i + \beta_i} = 1$$

as desired.

For $k > 2$ Assume A has at least 2 bricks. So there is a hyperplane h which separates the interiors of two bricks of A

.We can assume WLOG the following assumptions: □

1. $h = \{x_1 = 0\}$
2. B is split in the same relation as A. That is, if we define $A^+ = A \cap h^+$ and $A^- = A \cap h^-$ and similarly define B^+, B^- then $r := \frac{Vol(A^+)}{Vol(A)} = \frac{Vol(B^+)}{Vol(B)}$.

The reason we can do it, is that volume is invariant to translations.

Now since $A^+ + B^+$ and $A^- + B^-$ are interior disjoint subsets of $A + B$, and by the induction hypothesis we get:

$$Vol(A+B) \geq Vol(A^+ + B^+) + Vol(A^- + B^-) \geq (Vol(A^+)^{\frac{1}{n}} + Vol(B^+)^{\frac{1}{n}})^n + (Vol(A^-)^{\frac{1}{n}} + Vol(B^-)^{\frac{1}{n}})^n$$

$$= (r^{\frac{1}{n}}Vol(A) + r^{\frac{1}{n}}Vol(B))^n + ((1-r)Vol(A)^{\frac{1}{n}}) + ((1-r)Vol(B)^{\frac{1}{n}})^n =$$

$$(Vol(A)^{\frac{1}{n}} + Vol(B)^{\frac{1}{n}})^n$$

As desired.

Theorem 5. *Let P be a convex set in R^{n+1},and let $A = P \cap (x_1 = a)$, $B = P \cap (x_1 = b)$, $C = P \cap (x_1 = c)$ be three slices of A where $a < b < c$. So :*

$$Vol(B) \geq min(Vol(A), Vol(C))$$

Proof. sufficient to show that v given by :

$$v(t) = Vol(P \cap (x_1 = t))^{\frac{1}{n}}$$

is concave on its support.

Let $\alpha = \frac{b-a}{c-a}$. So $b = (1-\alpha) * a + a * c$. By the convexity of P , $(1-\alpha)A + \alpha C \subset B$. By theorem 3 :

$$v(b) = Vol(B)^{\frac{1}{n}} \geq Vol((1-\alpha)A + \alpha C)^{\frac{1}{n}} \geq Vol((1-\alpha)A)^{\frac{1}{n}} + Vol(\alpha C)^{\frac{1}{n}} =$$

$$((1-\alpha)^n Vol(A))^{\frac{1}{n}} + (\alpha^n Vol(C))^{\frac{1}{n}} = (1-\alpha)Vol(A)^{\frac{1}{n}} + \alpha Vol(C)^{\frac{1}{n}}$$

$$= (1-\alpha)v(a) + v(c)$$

So the function is indeed concave. □

Corollary 6. *For compact sets $A, B \subset \mathbb{R}^n$,$Vol(\frac{A+B}{2}) \geq \sqrt{Vol(A)Vol(B)}$*

Proof. $Vol((A+B)/2)^{\frac{1}{n}} = Vol(\frac{A}{2} + \frac{B}{2})^{\frac{1}{n}} \geq Vol(\frac{A}{2})^{\frac{1}{n}} + Vol(\frac{B}{2})^{\frac{1}{n}} = \frac{Vol(A)^{\frac{1}{n}} + Vol(B)^{\frac{1}{n}}}{2} \geq \sqrt{Vol(A)^{\frac{1}{n}} + Vol(B)^{\frac{1}{n}}}$ □

1 The n- dimensional sphere/ball

Before continuing to our next theorem (about measure concentration on the sphere) we would like to get some intuition about the properties of the n-dimensional sphere/ball in high dimensions.

Consider the ball of radius r in R^n where b_n is the unit ball. Then $Vol(rb_n) = r^n Vol(b_n)$. If $r = 1 - \delta$ (think of it as shrinking the radius of the unit ball by a little) . Then $r^n \leq e^{-\delta n}$ which decreases to 0 rapidly for δ which is asymptotically smaller than $\frac{1}{n}$. We conclude that in high dimensions, almost all of the ball's volume is concentrated close to the surface.

Another interesting fact is that $Vol(b_n) \rightarrow_{n\rightarrow\infty} 0$. The method to analyze the properties of the volume of b_n and surface area of $\mathbb{S}^{n-1}$ is by using Cavalieri's principle which states that the volume of an object in $\mathbb{R}^n$ can be calculated by one dimensional integration through the volumes of $n-1$ dimensional slices of the object. Using this principle we can obtain the recursive formula:

$$Vol(b_n) = \int_{x_n=-1}^{1} Vol(\sqrt{1-x_n^2}b_{n-1})dx_n = Vol(b_{n-1})\int_{x_n=-1}^{1}(1-x_n^2)^{\frac{n-1}{2}}dx_n$$

Now, when n is large, the term $(1-x_n)^{\frac{n-1}{2}}$ is very close to 0 except for a small interval around 0. Which means that in higher dimensions , most of the contribution of the volumes lays in a small strip around an equator.

Measure concentration on the sphere

Theorem 7. *Let $A \subset \mathbb{S}^{n-1}$ be a measurable set with $Pr(A) \geq \frac{1}{2}$, and let A_t denote the set of $\mathbb{S}^{n-1}$ with distance at most t from A for $t \leq 2$. So 1-$Pr(A_t) \leq 2exp(\frac{-nt^2}{2})$*

Proof. We will prove a weaker bound , with $\frac{-nt^2}{4}$ in the exponent. Let $A' = T(A)$ where

$$T(X) = \{ax \mid x \in X, \alpha \in [0,1]\} \subset b_n$$

where b_n is the unit ball in R^n. So $Pr[A] = \mu(A')$, where $\mu(A') = \frac{Vol(A')}{Vol(b_n)}$. That is true, since $\mu(A') = Pr(A')$ by definition, and each point in A corresponds to it's normalized point in on the sphere. So $Pr(A) = Pr(A')$. Define $B = \mathbb{S}^{n-1}\backslash A_t$ and $B' = T(B)$, so for all $a \in A$ and $b \in B$ we have $||a-b|| \geq t$. It can be shown that $\frac{(A'+B')}{2} \subset rb_n$ where $r = 1 - \frac{t^2}{8}$ so we get :

$$\mu(rb_n) = \frac{Vol(rb_n)}{Vol(b_n)} = r^n = (1 - \frac{t^2}{8})^n$$

By corollary 6:

$$(1 - \frac{t^2}{8})^n = \mu(rb_n) \geq \mu(\frac{(A' + B')}{2}) \geq \sqrt{\mu(A')\mu(B')} = \sqrt{Pr(A)Pr(B)} \geq \sqrt{\frac{Pr(B)}{2}}$$

Thus:

$$Pr(B) \leq 2(1 - \frac{t^2}{8})^{2n} \leq 2exp(\frac{-2nt^2}{8})$$

As desired. □

2 Concentration of Lipschitz functions

Consider a function $f : \mathbb{S}^{n-1} \to \mathbb{R}$ and assume we have a probability density function over the sphere.

Let $Pr(f \leq t) = Pr[\{x \in \mathbb{S}^{n-1} \mid f(x) \leq t\}]$

Definition 8. We define the median of f as $Sup(t)$ such that $Pr(f \leq t) \leq \frac{1}{2}$.

Lemma 9. *$Pr(f < med(f)) \leq \frac{1}{2}$ and $Pr(f > med(f)) \leq \frac{1}{2}$*

The proof is trivial and we will skip it.

Theorem 10. *Let $f : \mathbb{S}^{n-1} \to \mathbb{R}$ be 1-Lipschitz . Then for all $t \in [0, 1]$ we have:*

$$Pr[f > med(f) + t] \leq 2exp(\frac{-t^2 n}{2})$$

$$Pr[f < med(f) - t] \leq 2exp(\frac{-t^2 n}{2})$$

Proof. We prove only the first inequality , since the second one follows by symmetry .

Define :

$$A = \{x \in \mathbb{S}^{n-1} \mid f(x) \leq med(f)\}$$

By the lemma we mentioned earlier , we get $Pr(A) \geq \frac{1}{2}$. Now let $x \in A_t$ where A_t is defined like in theorem 7. Let $y \in A$. By definition of A_t we get

$$||x - y|| < t$$

So , since f is 1-Lipschitz we get :

$$f(x) \leq f(y) + ||y - x|| < med(f) + t$$

Now we get by theorem 7:

$$Pr[f > med(f) + t] \leq Pr(A_t^c) \leq 2exp(\frac{-t^2 n}{2})$$

□

3 The Johnson -Lindenstrauss Lemma

Lemma 11. *Define* $f_k : \mathbb{S}^{n-1} \to \mathbb{R}$ *by:*

$$f_k(x) = \sqrt{\sum_{i=1}^{k} x_i^2}$$

The length of the projection into the first k coordinates. Then f is sharply concentrated . That is , there exists $m = m(n,k)$ such that:

$$Pr(f \geq m + t) \leq 2exp(\frac{-t^2 n}{2})$$

$$Pr(f \leq m - t) \leq 2exp(\frac{-t^2 n}{2})$$

for all $t \in [0,1]$.Furthermore, for $k \geq 10ln(n)$, we have $m \geq \frac{1}{2} \cdot \sqrt{\frac{k}{n}}$.

Proof. It isn't too hard to verify that f is $1 - Lipschitz$. So by theorem 10 we get the first part of the claim with $m = med(f)$.

Now, what is left to prove is the lower bound for m. Now, for any $x \in \mathbb{S}^{n-1}$ we get :

$$1 = E(||x||^2) = \sum_{i=1}^{n} E(x_i^2) = nE(x_j^2)$$

for any $1 \leq j \leq n$ (by symmetry) .Thus $E(x_j^2) = \frac{1}{n}$

conclusion:

$$E[f(x)^2] = \frac{k}{n}$$

$$\frac{k}{n} = E(f^2) \leq Pr[f \leq m+t](m+t)^2 + Pr[f \geq m+t] \leq (m+t)^2 + 2exp(\frac{-t^2 n}{2})$$

Now let $t = \sqrt{\frac{k}{5n}}$. By assumption $k \geq 10ln(n)$, so we get $2exp(\frac{-t^2n}{2}) \leq \frac{2}{n}$, and by the previous inequality :

Define $u = x - y$. Since the projection is a linear operator: $P(u) = P(x) - P(y)$. So the so condition becomes :

$$(1 - \frac{\epsilon}{3})m||u|| \leq ||P_{\mathcal{F}}(u)|| \leq (1 + \frac{\epsilon}{3})m||u||$$

Also, since P is a linear operator, then for any $\alpha > 0$ the condition is equivalent to :

$$(1 - \frac{\epsilon}{3})m\alpha||u|| \leq ||P_{\mathcal{F}}(\alpha u)|| \leq (1 + \frac{\epsilon}{3})m\alpha||u||$$

So by picking $\alpha = \frac{1}{||u||}$ we can assume $||u|| = 1$. Namely we need to show:

$$|\ ||P(u)|| - m\ | \leq \frac{\epsilon}{3} m$$

Let $f(u) = ||P(u)||$.By lemma 11 , for $t = \frac{\epsilon m}{3}$ we get that the probability this not hold is bounded by:

$$Pr[|\ f(u) - m\ | \geq t] \leq 4exp(\frac{-\epsilon^2 m^2 n}{18}) \leq 4exp(-\frac{\epsilon^2 k}{72}) < n^{-2}$$

since $m \geq \frac{1}{2}\sqrt{\frac{k}{n}}$ and $k = 200\epsilon^{-2}ln(n)$

$$\frac{k}{n} \leq (m + \sqrt{\frac{k}{5n}})^2 + \frac{2}{n}$$

$$\Rightarrow m \geq \sqrt{\frac{k-2}{n}} - \sqrt{\frac{k}{5n}} \geq \frac{1}{2}\sqrt{\frac{k}{n}}$$

as desired. □

The last result tells us that by picking a random point on the sphere, we get that the length of its projection is highly concentrated. Next thing we want to do is to flip this result around, and argue that given a fixed point x, we can project it into a random k dimensional subspace , such that its length is highly concentrated.

The method by which we create a random unit vector is by sampling from a multi-dimensional distribution. In a somewhat similar way we also create a random orthogonal/rotation transform matrix . And to create a random projection we simply use the transform on the given vector and then projecting the result into the first k coordinates.

Lemma 12. *Let $x \in \mathbb{S}^{n-1}$be an arbitrary unit vector. Let $\mathcal{F}$ be a random $k-dimensional$ subspace $\mathcal{F}$, and let $f(x)$ be the length of the projection of x into $\mathcal{F}$. So there exists m such which satisfy the conclusion Lemma 11.*

Proof. Let v_i be the ith unit vector , and let M be a random translation of space (rotation). Clearly Mx is distributed uniformly on the sphere. Denote by the e_i the ith vector of the random matrix M.

so $e_i = M^T v_i$. Now:

$$< Mx, v_i >= (Mx)^T v_i = x^T M^T v_i = x^T e_i =< x, e_i >$$

So by projecting Mx into the first k coordinates we get:

$$f(Mx) = \sqrt{\sum_{i=1}^{k} < Mx, v_i >^2} = \sqrt{\sum_{i=1}^{k} < x, e_i >^2}$$

Note that the right side is exactly a projection into a random $k - dimensional$ space. And we see it distributes exactly like f on random vector. So by the previous lemma we get the result. □

Before proving the Johnson Lindenstrauss lemma we need to consider one more definition :

Definition 13. The mapping $f : \mathbb{R}^n \to \mathbb{R}^k$ is called a $K-$bi Lipschitz for a subset $X \subset \mathbb{R}^n$ if there exist a constant $c > 0$ such that :

$$cK^{-1}||x - y|| \leq ||f(x) - f(y)|| \leq c||x - y||$$

for all $x, y \in X$.

if K_0 is the least K for which f is bi-Lipschitz , we refer to f as a K_0-embedding of X.

Theorem 14. *(The Johnson Lindenstrauss lemma) : Let X be an n-point set in the Euclidean space , and let $\epsilon \in (0,1]$. Then There exists a $(1+\epsilon)$ embedding of X into $\mathbb{R}^k$, where $k = O(\epsilon^{-2}log(n))$*

Proof. Assume $X \subset \mathbb{R}^n$. Let $k = 200\epsilon^{-2}ln(n)$. Assume $k < n$ and let $\mathcal{F}$ be a random $k-$*dimensional* linear subspace of $\mathbb{R}^n$. Let $P_{\mathcal{F}} : \mathbb{R}^n \to \mathcal{F}$ be the orthogonal projection to the subspace $\mathcal{F}$.

We prove that:

$$(1-\frac{\epsilon}{3})m||x-y|| \leq ||P_{\mathcal{F}}(x) - P_{\mathcal{F}}(y)|| \leq (1+\frac{\epsilon}{3})m||x-y||$$

holds with probability $\geq 1 - n^{-2}$. Since there are $O(n^2)$ pairs, we get that this holds for all pairs with some constant probability, say $\geq \frac{1}{3}$. In such case, the mapping P is an $D-$embedding of X into $\mathbb{R}^k$ with

$$D \leq \frac{1+\frac{\epsilon}{3}}{1-\frac{\epsilon}{3}} \leq 1+\epsilon$$

for $\epsilon \leq 1$.

Define $u = x - y$. Since the projection is a linear operator: $P(u) = P(x) - P(y)$. So the so condition becomes :

$$(1-\frac{\epsilon}{3})m||u|| \leq ||P_{\mathcal{F}}(u)|| \leq (1+\frac{\epsilon}{3})m||u||$$

Also, since P is a linear operator, then for any $\alpha > 0$ the condition is equivalent to :

$$(1-\frac{\epsilon}{3})m\alpha||u|| \leq ||P_{\mathcal{F}}(\alpha u)|| \leq (1+\frac{\epsilon}{3})m\alpha||u||$$

So by picking $\alpha = \frac{1}{||u||}$ we can assume $||u|| = 1$. Namely we need to show:

$$| \; ||P(u)|| - m \; | \leq \frac{\epsilon}{3}m$$

Let $f(u) = ||P(u)||$.By lemma 11 , for $t = \frac{\epsilon m}{3}$ we get that the probability this not hold is bounded by:

$$Pr[| \; f(u) - m \; | \geq t] \leq 4exp(\frac{-\epsilon^2 m^2 n}{18}) \leq 4exp(-\frac{\epsilon^2 k}{72}) < n^{-2}$$

since $m \geq \frac{1}{2}\sqrt{\frac{k}{n}}$ and $k = 200\epsilon^{-2}ln(n)$ □