CSC2414 - Metric Embeddings*
Lecture 5: Dimension Reduction

Notes taken by Igor Naverniouk and Ilya Sutskever
revised by Hamed Hatami

Summary: In this lecture, we prove the Johnson-Lindenstrauss lemma [JL84], which shows that it is possible to embed any ℓ_2 metric space X on n points into ℓ_2^k with distortion $(1+\epsilon)$, where $k = O(\log(n)/\epsilon^2)$.

1 Required Dimension in ℓ_1

Recall that in ℓ_2, n points can be isometrically embedded into ℓ_2^{n-1}. Does a similar statement hold for ℓ_1? Given n points in ℓ_1^N, what is the dimension that we really need?

First let us show that it is possible to decrease the dimension to a function of N.

Claim 1.1. *if (X,d) embeds isometrically into ℓ_1, then it embeds isometrically into $\ell_1^{n!}$*

Proof. We say $i \sim j$ if for all points $x, y \in X$, $x_i \le y_i$ if and only if $x_j \le y_j$. Note that $\sim$ is easily seen to be an equivalence relation. Trivially there are at most $n!$ equivalence classes defined by $\sim$. Denote by $\mathcal{F}$ the set of equivalence classes. For every x and every equivalence class $F \in \mathcal{F}$ let $x_F = \sum_{i \in F} x_i$. Now define $f : X \to \ell_1^{n!}$ as

$$f(x) = \sum_{F \in \mathcal{F}} x_F e_F,$$

where e_F are the natural basis of $\ell_1^{|\mathcal{F}|}$. So here x_F is a coordinate of x in ℓ_1^N. To see that f is an isometry, note that

$$f(x) - f(y) = \sum_{F \in \mathcal{F}} \left| \sum_{i \in F} (x_i - y_i) \right| = \sum_{F \in \mathcal{F}} \sum_{i \in F} |x_i - y_i| = \|x - y\|_1,$$

where the second inequality is valid because $\operatorname{sign}(x_i - y_i)$ is constant for all $i \in F$. This shows that f is an isometry. $\square$

However, $N = \binom{n}{2}$ also suffices.

* Lecture Notes for a course given by Avner Magen, Dept. of Computer Sciecne, University of Toronto.

Claim 1.2. *Assume that (X, d) embeds isometrically into ℓ_1. Then it also embeds isometrically into ℓ_1^M, where $M = \binom{n}{2}$.*

Proof. By Claim 1.1, we can assume that (X, d) is in ℓ_1^N. Let $c_i : X^2 \to [0, \infty)$ be the distance between the points on the ith coordinate, where $1 \leq i \leq N$. Therefore,

$$d(x, y) = \sum_{1 \leq i \leq N} c_i(x, y).$$

In addition, we may view d and c_i as elements of $\mathbb{R}^M$ where $M = \binom{n}{2}$ with non-negative entries. Since $d \in \operatorname{span}(c_i)_{i \leq N}$, then $d = \sum_i \alpha_i c_i$, and $\alpha_i \in \mathbb{R}$ may be chosen so that all but at most $\binom{n}{2}$ of them are 0. But all the components of c_i and d are positive. Does it mean that we may choose the α_i's to be positive as well? If the α_i's are positive, then we immediately get an $\ell_1^{\binom{n}{2}}$ metric for the points by

$$x \mapsto (\alpha_i x_i)_{i : \alpha_i \neq 0} \in \ell_1^{\binom{n}{2}}.$$

The answer is affirmative and can be obtained by an application of the Carathéodory's Fundamental Theorem (see e.g. [Eck93]), which we state here without the proof:

Theorem 1.3. *Each point in the convex hull of a set S in $\mathbb{R}^n$ is in the convex combination of $n + 1$ or fewer points of S.*

Next consider the following sets in $\mathbb{R}^{\binom{n}{2}}$:

$$C = \{ \sum_{1 \leq j \leq N} \beta_j c_j : \beta_j \geq 0 \},$$

and

$$D = \{ \lambda d : \lambda \geq 0 \},$$

and the plane

$$E = \{ x \in \mathbb{R}^{\binom{n}{2}} : x_i \geq 0, \sum_i x_i = 1 \}.$$

Trivially $C \cap E$ is an $\binom{n}{2} - 1$ dimensional convex set, and $D \cap E$ is a point $\lambda d \in C \cap E$. To apply Caratheodory's theorem, note that $C \cap E$ is the convex hull of $\lambda_i c_i$, where λ_i are chosen in such a way that $\lambda_i c_i \in E$. Applying the theorem on $C \cap E$ with the point $\lambda d \in C \cap E \cap D$, we get that

$$\lambda d = \sum_i w_i \lambda_i c_i$$

with $\binom{n}{2}$ terms in the summation, and with all the constants $w_i, \lambda_i \geq 0$. Dividing the above expression by λ gives us the desired expression of d as a linear combination of at most $\binom{n}{2}$ of the c_i's with positive coefficients. $\square$

2 Johnson-Lindenstrauss lemma

In the previous section we considered isometric dimension reductions: if $X \subset \ell_1$, $|X| = n$, then we can isometrically embed X into $\ell_1^{\binom{n}{2}}$. If $X \subset \ell_2$, then X can be isometrically embedded into ℓ_2^{n-1}. What can be said about the dimension if we relax the isometry condition to having distortion $1+\epsilon$? Johnson and Lindenstrauss [JL84] answered this question for ℓ_2:

Theorem 2.1. *(Johnson-Lindenstrauss) If $X \subset \ell_2$, $|X| = n$, then for every $\epsilon > 0$,*

$$X \overset{1+\epsilon}{\hookrightarrow} \ell_2^{O\left(\frac{\log n}{\epsilon^2}\right)}.$$

To prove the theorem we will give a linear embedding $T : \ell_2^{n-1} \to \ell_2^{O\left(\frac{\log n}{\epsilon^2}\right)}$. It is sufficient for T to satisfy

$$(1-\epsilon/4)M\|x-y\|_2 \leq \|T(x)-T(y)\|_2 \leq (1+\epsilon/4)M\|x-y\|_2,$$

for some $M > 0$ and every $x, y \in X$ (The distortion will be at most $\frac{1+\epsilon/4}{1-\epsilon/4} \leq 1+\epsilon$). Since T is linear $T(x) - T(y) = T(x-y)$, and so the above is equivalent to

$$(1-\epsilon/4)M\|v\|_2 \leq \|T(v)\|_2 \leq (1+\epsilon/4)M\|v\|_2, \tag{1}$$

for all the possible $\binom{n}{2}$ vectors v between the points of X.

The linear transformation T will be in fact an orthogonal projection into a random[1] subspace E of ℓ_2^n with $\dim(E) = O\left(\frac{\log n}{\epsilon^2}\right)$:

Remember that a finite dimensional linear subspace of ℓ_2 is the span of exactly $k := \dim(E)$ orthonormal vectors, $u_1, \ldots, u_k$, where orthonormality means:

- $\langle u_i, u_i \rangle = 1$ for $i = 1, \ldots, k$.
- $\langle u_i, u_j \rangle = 0$ for $i \neq j$.

Note that the subspace E is indeed isometrically isomorphic to ℓ_2^k with the embedding $u \mapsto \sum_{i=1}^k \langle u_i, u \rangle e_i$, where u is a vector in E, and e_i are the natural basis of ℓ_2^k.

Now the orthogonal projection P from ℓ_2 onto E is defined by $P(v) := \sum_{i=1}^k \langle v, u_i \rangle u_i$ for every $v \in \ell_2$. Note that

(i) $\langle P(v), w \rangle = \langle v, w \rangle$ for every $w \in E$;

(ii) $\langle v - P(v), w \rangle = 0$ for every $w \in E$.

Exercise 2.2. Prove the properties (**i**) and (**ii**) of orthogonal projections and show that these two properties define the projection uniquely.

Exercise 2.3. Prove that every projection P satisfies $\|P(v)\|_2 \leq \|v\|_2$ for all $v \in \ell_2$.

Next we need to specify that what do we mean by a random subspace E. First we need to define the Haar measure on the Sphere.

[1] will be defined formally below

2.1 Haar measure on the Sphere

If we want to pick a point at random on the surface of a sphere in $\mathbb{R}^n$, we need to define a probability distribution μ on S^{n-1}. If we want our point to be "uniform", then the measure μ must be invariant under rotations, i.e. $\mu(A) = \mu(TA)$ where T is an arbitrary rotation. First we need to show that such a measure exists.

Definition 2.4. Let $A \subseteq S^{n-1}$. Define A_ϵ as the set of points of distance ϵ from A, where the distance is the geodesic distance on the sphere. More formally

$$A_\epsilon := \{x \in S^{n-1} : \rho(x, A) \leq \epsilon\},$$

where ρ is the geodesic distance.

Remark 2.5. In Definition 2.4 we used geodesic distance and this will appear repeatedly in this lecture note. The geodesic distance between two points on the sphere is the length of the smallest arc between them. Note that for $x, y \in S^{n-1}$, we have

$$\frac{2}{\pi}\rho(x, y) \leq \|x - y\|_2 \leq \rho(x, y),$$

and so the geodesic distance and the Euclidian distance are within a constant factor of each other.

When $A = \{x\}$, A_ϵ is called a cap centered at $x \in S^{n-1}$. Define the measure of a cap C as

$$\mu(C) = \frac{\text{area}(C)}{\text{area}(S^{n-1})}.$$

Note that $\mu(S^{n-1}) = 1$. The extension of this measure to all Borel[2] sets on the sphere is called the Haar measure on the sphere (intuitively $\mu(A)$ is the normalized area of $A \subseteq S^{n-1}$). Note that this measure is invariant under the rotation, and in fact this is the unique probability measure on Borel sets that satisfies this property.

Remark 2.6. To choose a point uniformly at random (according to Haar measure) on S^{n-1}, one might consider a sequence $(X_1, X_2, \dots, X_n)$ of normally distributed random variables. Then, by the property of Gaussians, the vector $(X_1, X_2, \dots, X_n) \in \mathbb{R}^n$ is a rotationally invariant n-dimensional Gaussian. Now normalize $(X_1, X_2, \dots, X_n)$, and we get a uniformly random point on S^{n-1}.

2.2 An example: Embedding into one dimension

The simple case is embedding into one dimension. Let $\phi(x) = \langle u, x\rangle$, where $u \in S^{n-1}$ is chosen uniformly at random. What do we expect to happen? The value of $\phi(x)$ might range from 0 (when x is orthogonal to u) to $\|x\|_2$ (when x and u are parallel). But u is chosen randomly, and to understand the behavior of $\phi(x)$ we need to see that what is a "typical" u with respect to all the possible $\binom{n}{2}$ vectors v between the points of X.

[2]Borel sets are the sets that can be constructed from open or closed sets by repeatedly taking countable unions and intersections

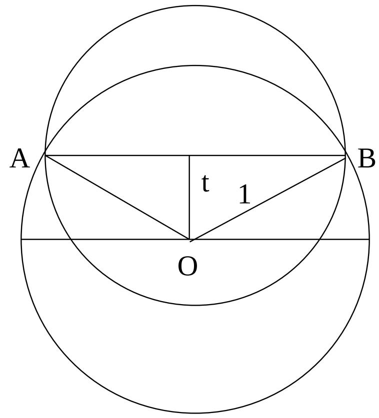

Figure 1: Enclosing a cone inside a sphere

2.3 Isoperimetric Inequality

Consider a fixed vector x. We may assume that $\|x\|_2 = 1$, as $\phi(x) = \|x\|_2\langle u, y\rangle$ where $y = \frac{x}{\|x\|_2}$ and so $\|y\|_2 = 1$. Since everything is rotationally invariant, let $x = (0, 0, \ldots, 0, 1)$. What is $\langle u, x\rangle$ in this case? Think of the Earth (S^2), with x being the North Pole and u being a random point on the surface. Then $\langle u, x\rangle$ is the latitude of u, normalized to the range $[-1, 1]$.

What percentage of the Earth's area is farther than latitude t from the equator? In other words, what is the probability, under the uniform measure, that $|\langle u, x\rangle| > t$? What is the expected value $\mathbb{E}[\langle u, x\rangle^2]$, the expected value of the square of the last coordinate of u? The answer is

$$\mathbb{E}[\langle u, x\rangle^2] = \mathbb{E}[u_n^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[u_i^2] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} u_i^2\right] = \frac{1}{n},$$

by symmetry. Now we want to say that $\langle u, x\rangle$ is well-concentrated around its expected value.

Lemma 2.7. *If u is a uniformly random unit vector in $\mathbb{R}^n$, then*

$$\Pr[|u_n| \geq t] \leq 2e^{-\frac{t^2 n}{2}}.$$

Proof. Note that $\Pr[|u_n| \geq t]$ is simply the surface area of that part of the unit sphere in n dimensions that is above the horizontal plane $u_n = t$ or below the horizontal plane $u_n = -t$. The two caps are symmetric, so let us find the measure (normalized area) of the top cap.

Since we have a unit sphere, the measure of the top cap is equal to the volume of the cone, whose point is at the origin and whose base is the top cap. From the triangle AOB, we can get that the radius of the base of the cone is half of AB and is equal to $\sqrt{1-t^2}$. Then we can enclose the whole cone inside a sphere of that radius. That sphere will be centered at the midpoint of AB and will have volume at most $(\sqrt{1-t^2})^n$. Since we have to take care of the bottom cap as well, the total volume of the two required spheres is at most

$$2(\sqrt{1-t^2})^n \leq 2(e^{-t^2})^{n/2} = 2e^{\frac{-t^2 n}{2}}.$$

$\square$

Recall the definition of A_ϵ, Definition 2.4. How small can $\mu(A_\epsilon)$ be with respect to $\mu(A)$. Intuitively $\mu(A_\epsilon)$ minimizes when A is a cap with measure $\mu(A)$, and in fact this intuition is true although the proof is not easy.

Lemma 2.8. *For each* $0 < a < 1$ *and* $\epsilon > 0$*,*

$$\min\{\mu(A_\epsilon) : A \subseteq S^{n-1}, \mu(A) = a\}$$

exists and is attained on A_o *is a cap with* $\mu(A_o) = a$*.*

Consider the hemisphere $C = \{x \in S^{n-1} : x_n \leq 0\}$ which is a cap of measure $1/2$. Lemma 2.7 shows that

$$\mu(\{x : d(x, C) \leq \epsilon\}) \geq 1 - e^{\frac{\epsilon^2 n}{2}},$$

where $d(x, C)$ is the Euclidian distance. Note that on the other hand for every $A \subseteq S^{n-1}$, we have that $\{x : d(x, A) \leq \epsilon\} = A_\delta$ for some δ that can be calculated according to ϵ. Now from Lemma 2.8 we know that among all sets $A \cap S^{n-1}$ of measure $1/2$, the measure of A_δ minimizes when A is a hemisphere. So the above bound holds for every set A of measure at least $1/2$.

Lemma 2.9 (Isoperimetric inequality I). *Let* $A \subseteq S^{n-1}$ *with* $\mu(A) \geq 1/2$*. Then*

$$\mu(\{x : d(x, A) \leq \epsilon\}) \geq 1 - 2e^{-\frac{\epsilon^2 n}{2}},$$

where $d(x, A)$ *is the Euclidian distance.*

Remark 2.10. A similar result holds for hypercubes $\{0,1\}^n$. If $|A| \geq \frac{1}{2}2^n$ and $A \subseteq \{0,1\}^n$, then

$$|\{x \in \{0,1\}^n : d(x, A) > t\}| \leq 2^n \cdot 4e^{-\frac{t^2}{2n}},$$

where $d(x, A)$ is the Hamming distance from x to the set A.

The isoperimetric inequality has an important application, namely the Levy's lemma [Lév51]. Consider a continuous function $f : S^{n-1} \to \mathbb{R}$. The *median* of f is $M_f \in \mathbb{R}$ such that $\mu(\{x : f(x) \leq M_f\}) \geq 1/2$ and $\mu(\{x : f(x) \geq M_f\}) \geq 1/2$.

Lemma 2.11 (Levy's Lemma I). *If* $f : S^{n-1} \to \mathbb{R}$ *is* a*-Lipschitz*[3]*, then*

$$\Pr_{x \in S^{n-1}} [|f(x) - M_f| > t] \leq 4e^{\frac{-t^2 n}{2a^2}}.$$

[3] $f : (X, d) \to (X', d')$ is called a-Lipschitz if $d'(f(x), f(y)) \leq a \times d(x, y)$. So here we have $|f(x) - f(y)| \leq a\|x - y\|_2$.

Proof. Let $A := \{x : f(x) \leq M_f\}$ and $B := \{x : f(x) \geq M_f\}$ and note that

$$\{x : |f(x) - M_f| > t\} \subseteq \{x : d(x, A) > at\} \cup \{x : d(x, B) > at\},$$

and use Lemma 2.9. $\square$

Sometimes it is desirable to substitute the median in Lemma 2.11 with the expected value. Next corollary shows that under some conditions the expected value is close to the median.

Corollary 2.12. *If $f : S^{n-1} \to \mathbb{R}^{\geq 0}$ is a-Lipschitz, then*

$$|\mathbb{E}[f] - M_f| \leq \frac{2\sqrt{\pi}a}{\sqrt{n}}.$$

Proof. By Lemma 2.11 we have

$$\mu(\{x : |f(x) - M_f| > t\}) \leq 4e^{\frac{-t^2 n}{2a^2}}.$$

So

$$|\mathbb{E}[f] - M_f| \leq \mathbb{E}|f(x) - M_f| \leq 4\int_0^\infty e^{\frac{-t^2 n}{2a^2}} dt \leq \frac{4a}{\sqrt{n}} \int_0^\infty e^{\frac{-s^2}{2}} ds = \frac{2\sqrt{\pi}a}{\sqrt{n}}.$$

$\square$

2.3.1 Isoperimetric inequality for geodesics distance.

One can actually calculate the area of caps to obtain:

Lemma 2.13 (Isoperimetric inequality II[4]). *Let $A \subseteq S^{n+1}$ with $\mu(A) \geq 1/2$. Then*

$$\mu(A_\epsilon) \geq 1 - \sqrt{\pi/8}e^{-\frac{\epsilon^2 n}{2}}.$$

Similar to Lemma 2.11.

Lemma 2.14 (Levy's Lemma II). *Let $f : S^{n+1} \to \mathbb{R}$ be continuous and let $A = \{x : f(x) = M_f\}$. Then*

$$\mu(A_\epsilon) \geq 1 - \sqrt{\pi/2}e^{-\epsilon^2 n/2}.$$

Proof. Note that $A_\epsilon = \{x : f(x) \leq M_f\}_\epsilon \cap \{x : f(x) \geq M_f\}_\epsilon$ and use Lemma 2.13. $\square$

[4]Note that the lemma is stated for S^{n+1} instead of S^{n-1}.

3 Proof of the Johnson-Lindestrauss Theorem

Now the goal is to move to k dimensions. To construct the Johnson-Lindenstrauss embedding, project X onto a raondom k-dimensional linear subspace. How do we pick a "random k-dimensional subspace"? Consider rotating $\mathbb{R}^n$ randomly. Then pick the first k dimensions (coordinates). If π_k is the projection onto the first k dimensions, then the random projection we use is $T := \pi_k \circ R$, where R is a random rotation[5].

Exercise 3.1. Show that the above is equivalent to choosing a k-dimensional subspace and then projecting the point on it.

Since we rotated the sphere randomly, $R(x)$ is a uniformly random point on S^{n-1} when $\|x\|_2 = 1$. So the question is: what is the norm of the first k dimensions of a random unit vector in $\mathbb{R}^n$?

Let $M := \mathrm{Median}(\sqrt{\sum_{i=1}^k u_i^2})$. Then we need to show that

$$\Pr\left[\left|\sqrt{\sum_{i=1}^k u_i^2} - M\right| > \frac{\epsilon}{4} M\right] < \frac{1}{n^3}. \tag{2}$$

If this is true, then

$$(1-\epsilon/4)M\|x\|_2 \le \|T(x)\| \le (1+\epsilon/4)M\|x\|_2$$

with probability $\ge 1 - n^{-3}$. So, with probability $\ge 1 - n^{-1}$, this is true for all the $\binom{n}{2}$ difference vectors. Let

$$f(u) = \sqrt{\sum_{i=1}^k u_i^2}.$$

Then f is 1-Lipschitz (see Exercise 2.3) and Lemma 2.11 implies that

$$\Pr[|f(x) - M_f| > t] \le 4e^{\frac{-t^2 n}{2}},$$

or equivalently,

$$\Pr\left[|f(x) - M_f| > \frac{\epsilon}{4} M_f\right] \le 4e^{\frac{-\epsilon^2 n}{8} M_f^2}. \tag{3}$$

where $\epsilon = t/M_f$.

What we would like is to replace M_f^2 with the mean of f^2. Firstly, note that, by symmetry,

$$\mathbb{E}[f^2] = \mathbb{E}\left[\sum_{i=1}^k u_i^2\right] = \frac{k}{n}.$$

Since the maximum value that f^2 can get is k, for every $\Delta > 0$, we have

$$\mathbb{E}[f^2] \le (M_f + \Delta)^2 + \Pr[f(x) \ge M_f + \Delta]k \le (M_f + \Delta)^2 + 4ke^{\frac{-\Delta^2 n}{2}}.$$

[5] R can be expressed as a random orthonormal matrix with determinant 1.

So

$$\mathbb{E}[f^2] \leq 2M_f^2 + 2\Delta^2 + 4ke^{\frac{-\Delta^2 n}{2}}.$$

Now set $\Delta = \sqrt{k/(8n)}$. We get

$$\frac{k}{n} \leq 2M_f^2 + 2\frac{k}{8n} + 4ke^{-k/16},$$

which shows that $M_f^2 \geq \frac{k}{4n}$ for $k \geq 32\log(n)$. Now substituting this in (3) completes the proof.

4 Remarks

There is an elementary proof of the Johnson-Lindenstrauss lemma by Dasgupta and Gupta [DG99].

The embedding used in the proof of the lemma is randomized, but it is possible to derandomize it.

References

[DG99] S. Dasgupta and A. Gupta. An elementary proof of the johnson-lindenstrauss lemma. *ICSI Technical Report TR 99-006*, 1999.

[Eck93] Jürgen Eckhoff. Helly, Radon, and Carathéodory type theorems. In *Handbook of convex geometry, Vol. A, B*, pages 389–448. North-Holland, Amsterdam, 1993.

[JL84] W. Johnson and J. Lindenstrauss. Extensions of lipschitz mats into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[Lév51] Paul Lévy. *Problèmes concrets d'analyse fonctionnelle. Avec un complément sur les fonctionnelles analytiques par F. Pellegrino*. Gauthier-Villars, Paris, 1951. 2d ed.