COMS 4995: Unsupervised Learning (Summer'18) Jun 5, 2018

# Lecture 5 – Dimensionality Reduction and Bourgain's Theorem

Instructor: *Nakul Verma* Scribes: *Ziyuan Zhong, Vincent Liu*

This lecture introduces the problem of embedding and talks about the proof of Bourgain's Theorem.

## 1 Embedding and dimensionality reduction

### 1.1 Overview and Motivations

Not all data people deal with has a "vector space" representation. For example, we might only have a similarity matrix, like the following:

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|
| $x_1$ | 0 | 1 | 1 | 1 |
| $x_2$ | 1 | 0 | 2 | 2 |
| $x_3$ | 1 | 2 | 0 | 2 |
| $x_4$ | 1 | 2 | 2 | 0 |

**Typical goals**:

- gain better understanding of the relationship among data points.
- embed the data into a space (typically $\mathbb{R}^d, l_2$) that we understand better. We can then apply off-the-shelf models/algorithms etc.

**Dimensionality Reduction**:

- Reduce "noise" (noise is application-specific; whatever you do not care about.)
- Increase computational efficiency

**Metric Embedding**:

- Given a metric space $(X, \rho)$, want to "embed" it into a "normed" space $\underbrace{(R^d, l_p)}_{l_\rho^p}$.
- Computational efficiency

The goal for an *embedding* is a function $f : \mathbf{X} \to \mathbb{R}^d$, where $\forall u, v \in \mathbf{X}$, $||f(u) - f(v)||_{l_p^\rho} \approx p(u, v)$. The bad news: in general, there are finite metric spaces $(\mathbf{X}, p)$, where $\mathbf{X}$ is a $n$-point metric space, that cannot be isometrically embedded into $l_2^d$ for any $d$ (in other word, no embeddings preserve distance exactly). See the following figure for an example.

**Definition 1.** *Given two metric space $(X, \rho), (Y, \sigma)$. A mapping $f : X \to Y$ is called a* $D$-embedding *of $X$ into $Y$ (where $D \geq 1$) if there exists some $r > 0$ such that $\forall x, x' \in X$,*

$$r \cdot \rho(x, x') \leq \sigma(f(x), f(x')) \leq D \cdot r \cdot \rho(x, x')$$

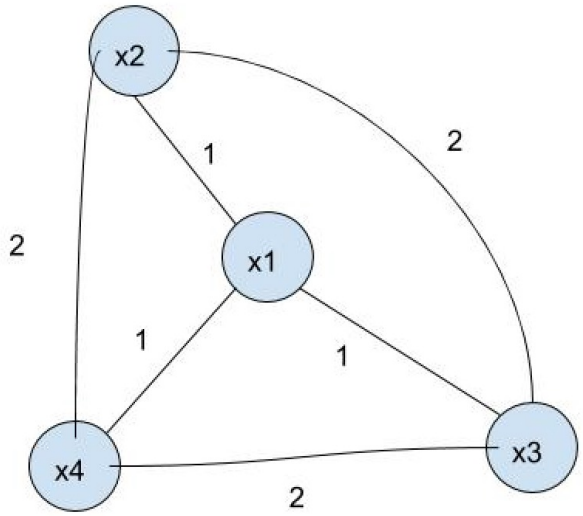



**Figure 1:** An illustration of not embeddable distance matrix

### 1.2 Embedding into $l_\infty^d$

Reminder: $||u-v||_{l_\infty^d} = \max_{1\le i\le d} |u_i - v_i|$

**Theorem 2** (Frechet)**.** *Any n-point metric space* $(\mathbf{X}, \rho)$ *with* $|\mathbf{X}| = n$ *can be isometrically embedded into* $l_\infty^d (d = n)$*.*

*Proof.* Let $x \in \mathbf{X}$, consider the function

$$f(x) = \begin{bmatrix} \rho(x, x_1) \\ \rho(x, x_2) \\ \dots \\ \rho(x, x_n) \end{bmatrix}$$

**Claim**: $f$ is a contraction. That is, $\forall u, v \in \mathbf{X}$, $||f(u) - f(v)||_{l_\infty^d} \le \rho(u, v)$.
Observation: Because $\rho$ is a metric and thus by triangle inequality,

$$\forall x_i \in \mathbf{X},\ \rho(u, x_i) - \rho(v, x_i) \le \rho(u, v)$$

It follows that

$$\max_{u,v} \rho(u, x_i) - \rho(v, x_i) \le \rho(u, v)$$

so the claim is true.
Now consider the coordinate for $u$, we have

$$\rho(u, v) \le |\rho(u, u) - \rho(u, v)|$$

Therefore,

$$\rho(u,v) \leq ||f(u) - f(v)||_{l_\infty^d} \leq \rho(u,v)$$

□

**Question**: can we do significantly better (e.g. $d = o(n)$) than $d = n$ in Frechet's Embedding?

**Theorem 3** (Incompressibility of general metric spaces)**.** *If* $\mathbf{Z}$ *is a normed space that* $D$*-embeds all* $n$*-points metric space, then,*
$dim(\mathbf{Z}) = \Omega(n)$ *for* $D < 3$*.*
$dim(\mathbf{Z}) = \Omega(n^{1/2})$ *for* $D < 5$*.*
$dim(\mathbf{Z}) = \Omega(n^{1/3})$ *for* $D < 7$*.*

If we want to compress $l_\infty^d$, we have to have more distortion.

**Theorem 4** (Construction is due to Bourgain)**.** *Let* $D = 3$ *and* $(\mathbf{X}, \rho)$ *be a* $n$*-point metric space. Then there exists a* $D$*-embedding into* $l_\infty^d$ *with* $d = \lceil 48\sqrt{n}\ln n \rceil = O(\sqrt{n}\ln n)$*.*

**Proof Sketch**
We want to have a coordinate such that $\rho(u,v) \geq [f(u) - f(v)]_i \geq \frac{1}{3}\rho(u,v)$.

$$f(u) = \begin{bmatrix} \rho(u, A_1) \\ \rho(u, A_2) \\ \cdots \\ \rho(u, A_d) \end{bmatrix}$$

where $A_i \subset \mathbf{X}$, $\rho(u, A) = \min_{x \in A} \rho(u, x)$.

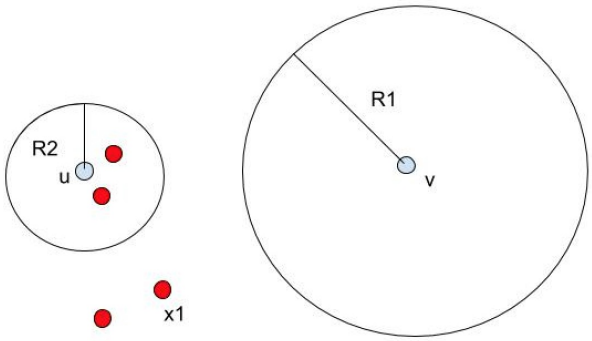



**Figure 2:** An illustration of the construction in the proof

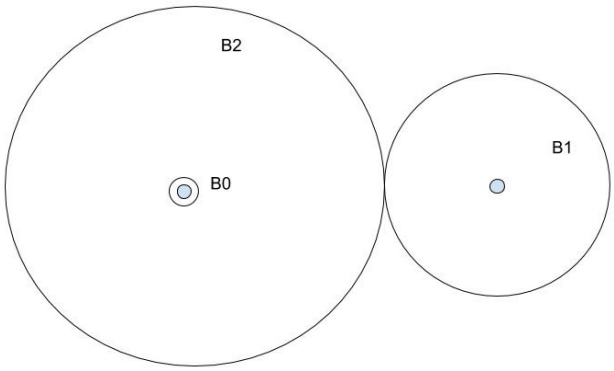
**Formal Proof**

*Proof.* For $1 \leq i \leq \lceil 24\sqrt{n}\ln n \rceil = m$:
Pick $x \in \mathbf{X}$ with probability $\min(\frac{1}{2}, \frac{1}{\sqrt{n}})$ independently and thus constitute the set $A_i$.
Pick $x \in \mathbf{X}$ with probability $\min(\frac{1}{2}, \frac{1}{n})$ independently and thus constitute the set $\bar{A}_i$.

$$\forall x \in \mathbf{X}, f(x) = \begin{bmatrix} \rho(u, A_1) \\ \rho(u, A_2) \\ ... \\ \rho(u, A_m) \\ \rho(u, \bar{A}_1) \\ \rho(u, \bar{A}_2) \\ ... \\ \rho(u, \bar{A}_m) \end{bmatrix}$$

**Claim**: Pick any $u, v \in \mathbf{X}, u \neq v$ and pick $i$, then either $|\rho(u, A) - \rho(v, A)| \geq \frac{1}{3}\rho(u, v)$ or $|\rho(u, \bar{A}) - \rho(v, \bar{A})| \geq \frac{1}{3}\rho(u, v)$ with probability $\geq \frac{1}{12\sqrt{n}}$ (over the choices of $A$ and $\bar{A}$).

**Figure 3:** An illustration of the three balls




*Proof.* (for the claim) Assume we have three balls: $B_0(u, r = 0)$, $B_1(v, r = \frac{1}{3}\rho(u, v))$, $B_2(u, r = \frac{2}{3}\rho(u, v))$.
**Idea**: either $|B_1 \cap \mathbf{X}| \leq \sqrt{n}$ (no points from $B_1$ will be picked and at least one point from $B_0$ will be picked with some probability) or $|B_1 \cap \mathbf{X}| > \sqrt{n}$ (no points from $B_2$ will be picked and at least one point from $B_1$ will be picked with probability $\geq \frac{1}{12}\sqrt{n}$).
**Case1** ($|B_1 \cap \mathbf{X}| \leq \sqrt{n}$):
Consider set $A$,
$Pr[E_1 := B_0 \cap A \neq \phi] = \min(\frac{1}{2}, \frac{1}{\sqrt{n}})$,
$Pr[E_2 := B_1 \cap A = \phi] = (1 - \min(\frac{1}{2}, \frac{1}{\sqrt{n}}))^{B_1 \cap \mathbf{X}} \geq (1 - \min(\frac{1}{2}, \frac{1}{\sqrt{n}}))^{\sqrt{n}} \geq \frac{1}{4}$,
Since $E_1$ and $E_2$ are disjoint,

$$Pr[E_1 \cap E_2] \geq \min(\frac{1}{8}, \frac{1}{4\sqrt{n}}) \geq \frac{1}{12\sqrt{n}}.$$

**Case2** ($|B_1 \cap \mathbf{X}| > \sqrt{n}$):
Consider set $\bar{A}$,
$Pr[E_3 := B_1 \cap \bar{A} \neq \phi] \geq ... \geq \frac{1}{3\sqrt{n}}$,
$Pr[E_4 := B_2 \cap \bar{A} = \phi] \geq ... \geq \frac{1}{4}$,
$Pr[E_3 \cap E_4] \geq \frac{1}{12\sqrt{n}}$.

Therefore, the claim is true. □

We have:

$$\Pr\Big[\exists u,v\in\mathbf{X}\text{ s.t. }\forall A_i,\ \bar{A}_i,\ |\rho(u,A_i)-\rho(v,A_i)|<\frac{1}{3}\rho(u,v)\text{ and }|\rho(u,\bar{A}_i)-\rho(v,\bar{A}_i)|<\frac{1}{3}\rho(u,v)\Big]$$
$$\leq \sum_{(u,v)\in\mathbf{X}\times\mathbf{X}\text{unordered pair}}(1-\frac{1}{12\sqrt{n}})^m \text{ because of the union bound}$$
$$\leq \binom{n}{2}e^{-\frac{1}{12\sqrt{n}}m}$$
$$\leq \binom{n}{2}e^{\ln\frac{1}{n^2}}$$
$$\leq \binom{n}{2}\frac{1}{n^2}$$
$$<1.$$

The proof uses the fact that $m=\lceil 24\sqrt{n}\rceil\ln n$ and $(1-x)\leq e^x$.
Therefore, the embedding $f$ exists.

□

**Open question**: Is there a deterministic construction of embedding into $l_\infty^d d=O(\sqrt{n}\ln n)$ with $D=3$?

**Theorem 5** (**Generalization**). *Let $D=2q-1\geq 3$(be odd).Then any n-point metric space can be D-embedded into $l_\infty^d$ where $d=O(qn^{1/q}\ln n)$.*

### 1.2.1 Summary

Frechet $l_\infty^d d=nd=\Omega(n), D<3$.
Bourgain $l_\infty^d, d=O(\sqrt{n}\ln n), D=3$.

## 1.3 Embedding into $l_2^d$

**Result1**: (follows from Bourgain generalization): Any $n$-point metric space can be embedded into $l_\infty^d$ with $D=O(\log^2 n)$ and $d=O(\log^2 n)$.
**Refinement**(Bourgain's $l_2$ result): Any $n$-point metric can embed in $l_2^d$ with $D=O(\log n)$.

**Theorem 6** (Johnson-Lidenstrauss "flatten" lemma(JL-lemma, 1984)). *Pick any $0<\epsilon<\frac{1}{2}$. Then for any integer n, let $d>\lceil\frac{4}{\epsilon^2}(2\ln n+\ln 3)\rceil\to d>\Omega(\frac{\ln n}{\epsilon^2})$. Then for any set $V\subset\mathbb{R}^D$, s.t. $|V|=n$, there exists a map $f:\mathbb{R}^D\to\mathbb{R}^d$ s.t. $\forall u,v\in V,\ (1-\epsilon)||u-v||_2^2\leq||f(u)-f(v)||_2^2\leq(1+\epsilon)||u-v||_2^2$.*

- *Moreover, $f$ is simply a linear map.*

- *Pick a random d-dim subspace (in D-dim), then above holds true with high probability(minor global scaling).*

For any $D$-dim $v$, define

$$f(v) = \begin{bmatrix} x_{11} & ... & x_{1D} \\ ... & & \\ x_{d1} & ... & x_{dD} \end{bmatrix} v$$

where $x_{ij}\ \forall i, j$ is drawn from a Gaussian independently. Then with high probability, $f$ satisfies the above properties.

$\exists\ n+1$ points in $\mathbb{R}^D (D \geq n)$ that cannot be isometrically embeddable in $l_2^d$ with $d < n$.

**Application of JL**:

- Fast provable clusterings(1999)
- Fast approximate nearest neighbor search
- Approximate solutions to graph problems(e.g. multi-commodity flow)
- Fast approximate linear algebra(e.g. matrix multiplication)("sketching")

**Proof Sketch**

Observation: Let $\phi$ be a random $d$-dim subspace (in $D$-dim).

**Claim**: We can show that $\mathbb{E}_\phi[||\phi(w)||^2] = \frac{d}{D}$. Pick any $0 < \epsilon < \frac{1}{2}$ and fix a unit vector $w \in \mathbb{R}^D$. Then,

$$\Pr\left[||\phi(w)||^2 < (1-\epsilon)\frac{d}{D} \text{ or } ||\phi(w)||^2 \geq (1+\epsilon)\frac{d}{D}\right] \leq 3e^{-d\epsilon^2/4}.$$

*Note:* on average, a projection of $w$ onto the random subspace $\phi$ has expected squared-norm:

$$\mathbb{E}\left[||\phi(w)||^2\right] = \frac{d}{D}.$$

Then, apply a concentration/Chernoff-type bound.

# References


[1] Reference: An Elementary Proof of a Theorem of Johnson and Lindenstrauss http://cseweb.ucsd.edu/ dasgupta/papers/jl.pdf

[2] Notes on Bourgain's theorem `http://www.cs.columbia.edu/~verma/classes/uml/ref/dim_redux_metric_bourgain.pdf`