Representations

Representations

What?

- Given data (in certain representation), produce a representation which provides a better understanding of the data

Why?

- Several ML models require data in a specific representation to work well

  Usually R^d, sometimes as a similarity function,

  occasionally graphs, rarely as curved spaces

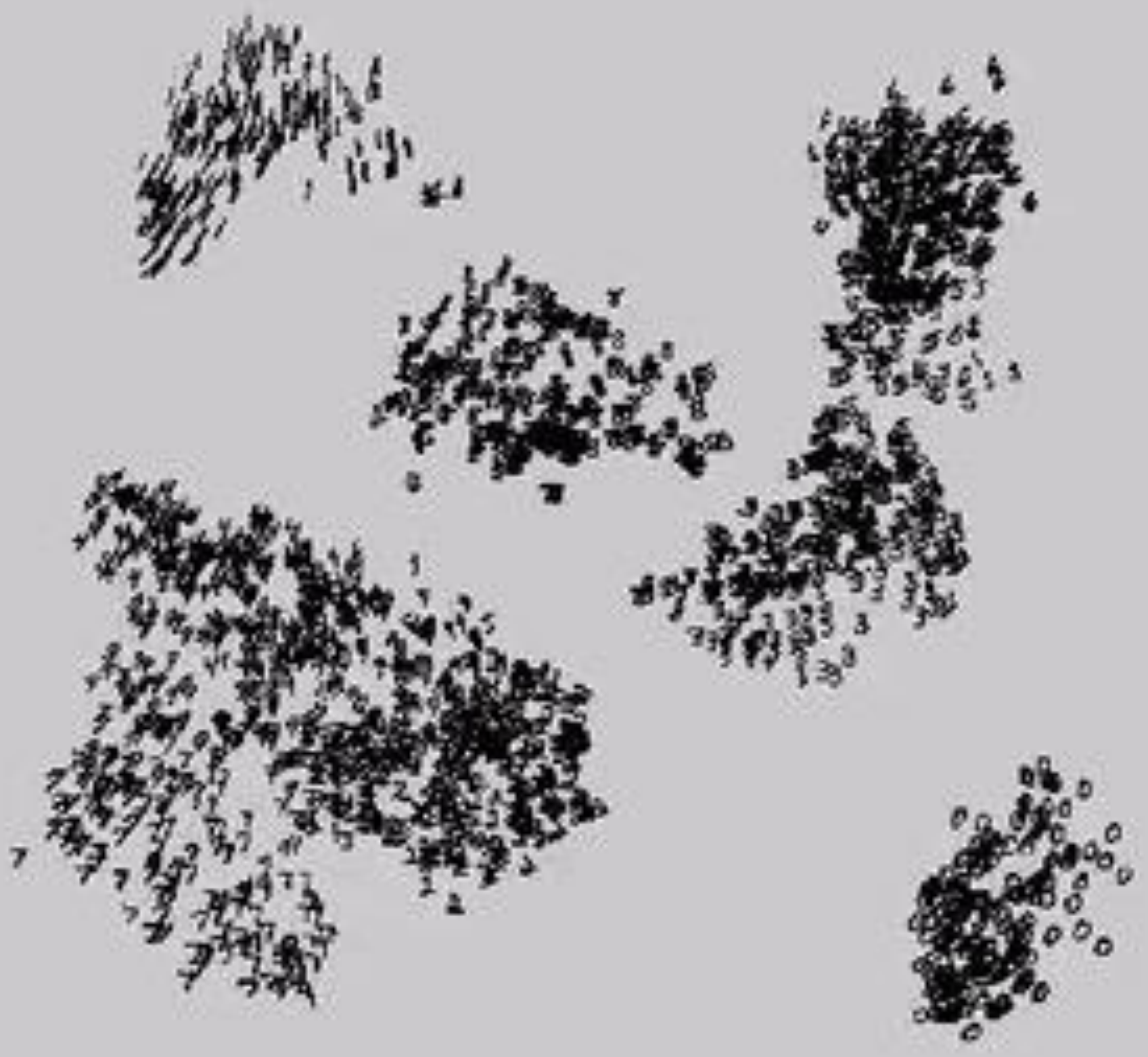
- Enhance the signal in data

  Discover underlying structure, suppress noise

- Improve computational efficiency and decrease space usage

  Dimensionality reduction, can use simpler models

Dimension Reduction: A Successful Example



Re-Representation Results in Information Loss

Any kind of data processing results in information loss.

Underlying process

Any processing

Theorem (Data Processing Inequality): Suppose $X \rightarrow Y \rightarrow Z$, then

$$I(X;Y) \geq I(X;Z)$$

Original measurements

No clever manipulation of the data (deterministic or randomized) can improve inference or provide more information about the underlying process X than Y itself

Data Processing Inequality

Data Processing Inequality: Suppose $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$

Recall:

- $I(A;B) = H(A)-H(A|B) = H(B)-H(B|A)$
- $I(A;B) \geq 0$

Proof:

Consider $I(X;(Y,Z)) = H(X) - H(X|YZ)$

$$= H(X) - H(X|Z) + H(X|Z) - H(X|YZ)$$

$$= I(X;Z) + I(X;Y|Z) \quad \geq 0$$

$$= I(X;Y) + I(X;Z|Y) \quad = 0$$

[b/c of the Markovian property $X \perp Z \mid Y$]

The theorem follows.

$\blacksquare$

Data Processing Inequality: Implications

Data Processing Inequality: If $X \rightarrow Y \rightarrow Z$, then $I(X;Y) \geq I(X;Z)$

This seems like bad news:

Any processing/re-representation of data can only result in information loss about the underlying process.

Catch:

If we are smart about our processing, we can ensure that we retain important aspects of data that are useful for our understanding of the underlying process, and loose all the frivolous/uninteresting information.

Example:

Suppose we want a representation for **effective nearest neighbors**, then we only need to **retain ordinal information** (*a* is closer to *b* than *c*)

Metric Embeddings

A Motivating Example

Given a data in a 'dissimilarity between objects' form

	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

How can we come up a vectoral representation, which respects the relations?

- To gain better understanding of the relationships between data
- If we embed the data in (R^d, L_2) we can apply off-the-shelf models.

Metric Embeddings

Goal: Given a metric space (X,ρ) want to embed it in a normed space (R^d, L_p)

L_p^d

Why normed spaces?

- easier to deal with
- we have a better understanding

Measuring the quality of an embedding:

Given two metric spaces (X,ρ) and (Y,σ). A mapping f: $X \rightarrow Y$ is called a **D-embedding** of X into Y (for $D \geq 1$) if there exists an $r > 0$ s.t. for all $x, x' \in X$,

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

- D is called the *distortion* of the embedding f
- If $D = 1$, then f is *distance-preserving* and thus called *isometric* (typically r=1)
- If $D \geq 1$, and $r \leq (1/D)$, then f is a contraction

Embeddings into L_∞

Theorem (Fréchet): An *n*-point metric space (X,ρ) can be isometrically embedded into L_∞^n

Reminder: $\| u - v \|_\infty = \max_i(|u_i - v_i|)$

Proof:

Consider the mapping $f(x) = \begin{bmatrix} \rho(x,x_1) \\ \rho(x,x_2) \\ \cdots \\ \rho(x,x_n) \end{bmatrix}$

Observation:

- *f* is a contraction, ie $\forall$ u, v $\in$ X, $\|f(u) - f(v)\|_{L_\infty^n} \le \rho(u,v)$

  Why? By triangle inequality $\forall$ u,v, $x_i \in$ X, $\rho(u,x_i) - \rho(v,x_i) \le \rho(u,v)$

  in particular, $\max_i |\rho(u,x_i) - \rho(v,x_i)| \le \rho(u,v)$

  $$\|f(u) - f(v)\|_{L_\infty^n} \le \rho(u,v)$$

- $\forall$ u, v $\in$ X, $\exists$i s.t. $\rho(u,v) = (f(u) - f(v))_i$

  Why? $f(u) = \begin{bmatrix} \rho(u,x_1) \\ \cdots \\ \rho(u,v) \\ \rho(u,x_n) \end{bmatrix}$ $f(v) = \begin{bmatrix} \rho(v,x_1) \\ \cdots \\ \rho(v,v) \\ \rho(v,x_n) \end{bmatrix}$

  For row *i* corresponding to v

  $$(f(u) - f(v))_i = \rho(u,v)$$

Fréchet Embedding

Theorem (Fréchet): An *n*-point metric space (X,ρ) can be isometrically embedded into L_∞^n

Reminder: $\| u - v \|_\infty = \max_i(|u_i - v_i|)$

Proof:

Consider the mapping $f(x) = \begin{bmatrix} \rho(x, x_1) \\ \rho(x, x_2) \\ \cdots \\ \rho(x, x_n) \end{bmatrix}$

Observation:

- $\forall$ u, v $\in$ X, $\|f(u) - f(v)\|_{L_\infty^n} \leq \rho(u, v)$
- $\forall$ u, v $\in$ X, $\exists$i s.t. $\rho(u, v) = (f(u) - f(v))_i$

$$\rho(u, v) = (f(u) - f(v))_i \leq |f(u) - f(v)|_i$$
$$\leq \max_i |f(u) - f(v)| = \|f(u) - f(v)\|_{L_\infty^n} \leq \rho(u, v)$$

■

L_∞ is a universal space!

Embeddings into L_∞

Good news: L_∞ is a universal space… for finite metric spaces

Some issues:

- The target dimension is huge ($d = n$). Can it be reduced?

  well… we can drop it down to $n - 1$ 

  (second observation can be refined by one coordinate)

Is it possible to get a significant improvement?

Incompressibility result

Theorem (Incompressibility of general metric spaces):

If Z is a normed space that D-embeds all *n* points metric space, then

- dim(Z) = $\Omega(n)$ for D < 3
- dim(Z) = $\Omega(n^{1/2})$ for D < 5
- dim(Z) = $\Omega(n^{1/3})$ for D < 7

More compression requires bigger distortion

Proof (very crude sketch):

Consider a collection of *n*-vertex graphs with large girth (length of the shortest cycle). The goal is to embed all such graphs in a normed space Z with some dimension dim(Z) without incurring too much distortion.

Using volume argument, ie each graph must occupy some part of the space, by embedding all such graphs we will eventually run out of 'room' and would require more space (ie dimensions)

Embeddings into L_∞ (with some distortion)

In light of the incompressibility result, can we get a good low-dimensional embedding incurring some distortion?

Theorem (Bourgain): Let D = 3 and (X,ρ) an *n*-point metric space. There exists a D-embedding of X into L_∞^d with $d = \lceil 48\sqrt{n}\ln n \rceil = O(\sqrt{n}\ln n)$.

Compare with the lowerbound (Incompressibility result),

we are almost tight (there is a ln(n) gap)

Bourgain's Embedding (into L_∞)

Theorem (Bourgain): Let D = 3 and (X,ρ) an *n*-point metric space. There exists a D-embedding of X into L_∞^d with $d = \lceil 48\sqrt{n}\ln n \rceil = O(\sqrt{n}\ln n)$.

Proof Idea:

Since we are happy with distortion D = 3, for any pair of point u,v ∈ X, we want $(1/3)\rho(u,v) \leq \|f(u) - f(v)\|_{L_\infty^d} \leq \rho(u,v)$

Equivalently, ∃ coordinate *i*, s.t. $(1/3)\rho(u,v) \leq [f(u) - f(v)]_i \leq \rho(u,v)$

Using ideas from Fréchet's construction, we will have a mapping *f* as

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

each $A_i \subset$ X is a going to be a ***set*** of points

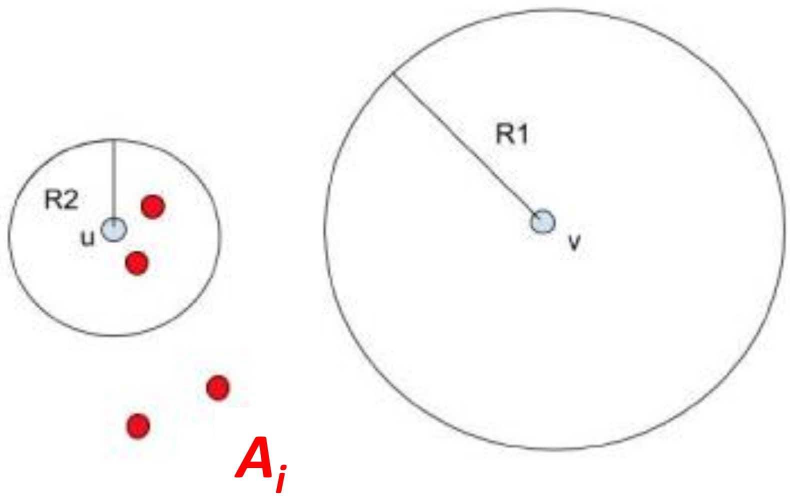
instead of using just a single point as A_i to create the necessary distance, using a set will be more economical thus reducing the number of coordinates

Bourgain's Embedding (into L_∞)

Theorem (Bourgain): Let D = 3 and (X,ρ) an *n*-point metric space. There exists a D-embedding of X into L_∞^d with $d = \lceil 48\sqrt{n}\ln n \rceil = O(\sqrt{n}\ln n)$.

Proof sketch cont.:

a picture to keep in mind…



The goal of the coordinates A_i is to "create" distance between pairs of points u,v

Need to ensure that A_i chosen in a way that there is some point in A_i close to u and all points in A_i are far away from v

$$[f(u) - f(v)]_i = |\rho(v, A_i) - \rho(u, A_i)|$$

$$\rho(u, A_i) = \inf_{x \in A_i} \rho(u, x)$$

$$\left(\begin{array}{l} \rho(v, A_i) \geq R_1 \\ \rho(u, A_i) \leq R_2 \end{array}\right) \qquad \geq R_1 - R_2$$

Bourgain's Embedding (into L_∞)

Theorem (Bourgain): Let D = 3 and (X,ρ) an *n*-point metric space. There exists a D-embedding of X into L_∞^d with $d = \lceil 48\sqrt{n}\ln n \rceil = O(\sqrt{n}\ln n)$.

Proof:

Let $m = \lceil 24\sqrt{n}\ln n \rceil$, for each *i* = 1,…,*m*

- Construct sets A_i as: for each x ∈ X, include it in A_i w.p min(1/2, $1/\sqrt{n}$)

high prob set Each choice is independent and all A_i are constructed independently

- Construct sets $\bar{A}_i$ as: for each x ∈ X, include it in $\bar{A}_i$ w.p. min(1/2, 1/*n*)

low prob set Each choice is independent and all $\bar{A}_i$ are constructed independently

We'll use the map (for each x)

$$f(x) = \begin{bmatrix} \rho(x, A_1) \\ \rho(x, A_2) \\ \dots \\ \rho(x, A_m) \\ \rho(x, \bar{A}_1) \\ \rho(x, \bar{A}_2) \\ \dots \\ \rho(x, \bar{A}_m) \end{bmatrix}$$

Proof of Bourgain's Embedding (into L_∞)

Claim: Pick any distinct u,v $\in$ X and any *i*, then

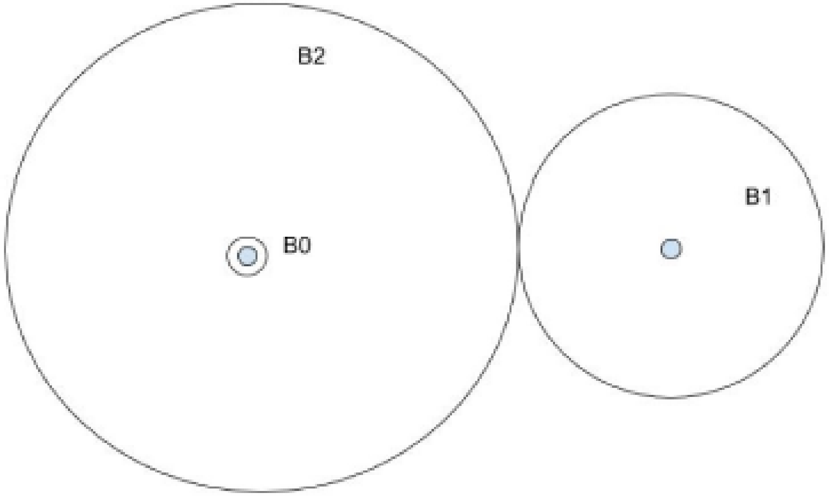
- Either $|\rho(u, A_i) - \rho(v, A_i)| \geq (1/3)\rho(u, v)$
- or $|\rho(u, \bar{A}_i) - \rho(v, \bar{A}_i)| \geq (1/3)\rho(u, v)$

with probability $\geq 1/(12\sqrt{n})$ over the choice of A_i and $\bar{A}_i$

Proof:

Consider three balls

- B_0(u, r = 0)
- B_1(v, r = 1/3 ρ(u,v))
- B_2(u, r = 2/3 ρ(u,v))



The process of randomly creating the sets can be analyzed in two disjoint cases…

- $|B_1 \cap X| \leq \sqrt{n}$ *we'll analyze the effect of A_i*
- $|B_1 \cap X| > \sqrt{n}$ *we'll analyze the effect of $\bar{A}_i$*

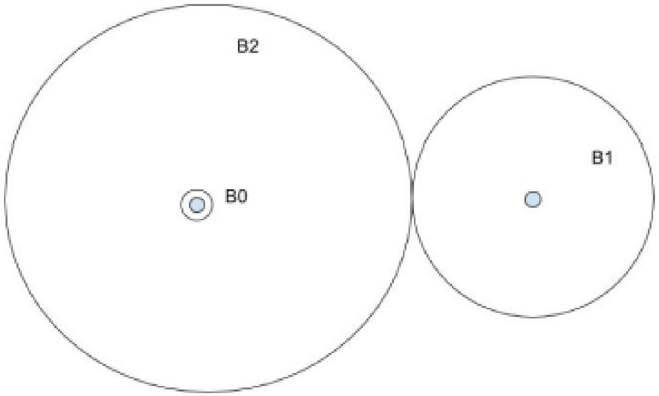
Proof of Bourgain's Embedding (into L_∞)

Claim: Pick any distinct u,v $\in$ X and any *i*, then

- Either $|\rho(u, A_i) - \rho(v, A_i)| \geq (1/3)\rho(u, v)$
- or $|\rho(u, \bar{A}_i) - \rho(v, \bar{A}_i)| \geq (1/3)\rho(u, v)$

with probability $\geq 1/(12\sqrt{n})$ over the choice of A_i and $\bar{A}_i$

Proof:



If $|B_1 \cap X| \leq \sqrt{n}$, for a given *i*

Consider the events

- $E_1 := |B_0 \cap A_i| \neq \phi$
- $E_2 := |B_1 \cap A_i| = \phi$

(at least one point of A_i is in B_0 and no points in B_1)

- B_0(u, r = 0)
- B_1(v, r = 1/3 ρ(u,v))
- B_2(u, r = 2/3 ρ(u,v))

So $|\rho(u, A_i) - \rho(v, A_i)| \geq (1/3)\rho(u, v)$

$P[E_1] = P[u \text{ is in } A_i] = \min(1/2, 1/\sqrt{n})$

$P[E_2] = [\, 1 - \min(1/2, 1/\sqrt{n})\,]^{|B1 \cap X|} \geq [1 - \min(\frac{1}{2}, \frac{1}{\sqrt{n}})]^{\sqrt{n}} \geq \frac{1}{4}$

$P[E_1 \text{ and } E_2] \geq \min(\frac{1}{8}, \frac{1}{4\sqrt{n}}) \geq \frac{1}{12\sqrt{n}}$

Proof of Bourgain's Embedding (into L_∞)

Claim: Pick any distinct u,v $\in$ X and any *i*, then

- Either $|\rho(u, A_i) - \rho(v, A_i)| \geq (1/3)\rho(u, v)$
- or $|\rho(u, \bar{A}_i) - \rho(v, \bar{A}_i)| \geq (1/3)\rho(u, v)$

with probability $\geq 1/(12\sqrt{n})$ over the choice of A_i and $\bar{A}_i$

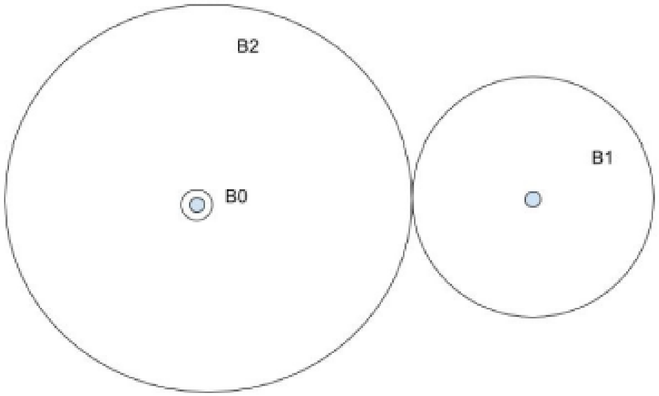
Proof:

If $|B_1 \cap X| > \sqrt{n}$, for a given *i*

Consider the events

- $E_3 := |B_1 \cap \bar{A}_i| \neq \phi$
- $E_4 := |B_2 \cap \bar{A}_i| = \phi$

(at least one point of $\bar{A}_i$ is in B_1 and no points in B_2)



- B_0(u, r = 0)
- B_1(v, r = 1/3 ρ(u,v))
- B_2(u, r = 2/3 ρ(u,v))

So $|\rho(u, \bar{A}_i) - \rho(v, \bar{A}_i)| \geq (1/3)\rho(u, v)$

$P[E_3] = 1 - [1 - \min(\frac{1}{2}, \frac{1}{n})]^{|B_1 \cap X|} \geq \ldots \geq \frac{1}{3\sqrt{n}},$

$P[E_4] = [1 - \min(\frac{1}{2}, \frac{1}{n})]^{|B_2 \cap X|} \geq \ldots \geq \frac{1}{4}$

$P[E_3 \text{ and } E_4] \geq \dfrac{1}{12\sqrt{n}}$ $\blacksquare$

Proof of Bourgain's Embedding (into L_∞)

Claim: Pick any distinct u,v $\in$ X and any *i*, then

- Either $|\rho(u, A_i) - \rho(v, A_i)| \geq (1/3)\rho(u, v)$
- Or $|\rho(u, \bar{A}_i) - \rho(v, \bar{A}_i)| \geq (1/3)\rho(u, v)$

with probability $\geq 1/(12\sqrt{n})$ over the choice of A_i and $\bar{A}_i$

So, the claim is true… BUT only for a fixed u, v

We can now use a union bound over various u, v pairs!

P[$\exists$u,v $\forall$ A_i *and* $\bar{A}_i$, $\begin{array}{l} |\rho(u, A_i) - \rho(v, A_i)| < (1/3)\rho(u, v) \\ |\rho(u, \bar{A}_i) - \rho(v, \bar{A}_i)| < (1/3)\rho(u, v) \end{array}$] *(Bad event)*

$$\leq \sum_{(u,v)\in \mathbf{X}\times\mathbf{X}\,\text{unordered pair}} \left(1 - \frac{1}{12\sqrt{n}}\right)^m \leq \binom{n}{2} e^{-\frac{1}{12\sqrt{n}}m} \leq \binom{n}{2} e^{\ln\frac{1}{n^2}}$$

< 1

*(this means with non-zero probability the complement is true, so such an **embedding exists**!)* ∎

Bourgain's Embeddings

Theorem (Bourgain's L_∞): Let D = 3 and (X,ρ) an *n*-point metric space. There exists a D-embedding of X into L_∞^d with $d = \lceil 48\sqrt{n}\ln n \rceil = O(\sqrt{n}\ln n)$.

Theorem (Bourgain's L_∞ Generalization): For any integer q ≥ 2, let D := 2q – 1. Then any *n*-point metric space (X,ρ) can be D-embedded into L_∞^d with $d = O(qn^{1/q}\ln n)$.

Need to have multiple sets and multiple balls to create this refinement... $A_i, \bar{A}_i, \bar{\bar{A}}_i$

Corollary (Bourgain's L_2 Weak form): Any *n*-point metric space (X,ρ) can be D-embedded into L_2^d with D = O($\log^2$ n) and d = O($\log^2$ n).

Follows immediately from the L_∞ result!

Theorem (Bourgain's L_2): Any *n*-point metric space (X,ρ) can be D-embedded into L_2^d with D = O(log n) and d = O(log n).

A few thoughts on Bourgain's Embedding

- There is a log(n) gap between the (normed) incompressibility result and Bourgain's L_∞ result. Can this be closed?

- Bourgain's embedding is an existential result with a probabilistic construction. Can we get a deterministic analog of Bourgain's embedding?

Other Metric Embedding results

Several positive and negative normed space embeddings (L_1, L_2, L_∞) exist for n-point metric spaces (far too many to list here)

One important result (Negative result in L_2)…

Theorem: For all n, $\exists$ n-point metric spaces that cannot be embedded in L_2 for any dimension d with distortion less than ($c \log n / \log \log n$) for an appropriate constant $c > 0$.

Dimension Reduction in L_2

Bad news… in general, it is not possible to isometric embed points in L_2^n in any lower dimensional space

Fact: Consider n+1 points V = {0,e_1,…,e_n} in L_2^n . Then V cannot be isometrically embedded in L_2^{n-1}. (*ie no compression is possible!*)

Proof: Let *f* : $L_2^n \to L_2^{n-1}$ is an isometry. WLOG assume *f*(0) = 0.

$$\|f(u) - f(v)\|^2 = \|f(u)\|^2 - 2\langle f(u), f(v)\rangle + \|f(v)\|^2$$

$$\|f(u)\|^2 = \|f(u) - f(0)\|^2 = \|u\|^2 = 1$$

Now, For any non-zero distinct u, v

$$2 = \|u - v\|^2 = \|f(u) - f(v)\|^2 = 2 - 2\langle f(u), f(v)\rangle$$

Thus, $\langle f(u), f(v)\rangle = 0$

We found *n* vectors *f*(e_1),…, *f*(e_n) that are all mutually orthogonal in R^{n-1} !! ■

Dimension Reduction in L_2

So isometric low-dimensional embeddings are not possible

BUT, if we are allowed to even a small amount of distortion, we can significantly compress the space!!!

Theorem (Johnson and Lindenstrauss '84):

For any n and $0 < \varepsilon < ½$, let $d \geq \left\lceil \frac{4}{\varepsilon^2}(2\ln n + \ln 3) \right\rceil = \Omega\left(\frac{1}{\varepsilon^2} \ln n\right)$. Then for any $X \subset R^D$ s.t. $|X| = n$, there exists a mapping $f : R^D \to R^d$, s.t. for all $x, y \in X$

$$(1-\epsilon)\|x-y\|^2 \leq \|f(x)-f(y)\|^2 \leq (1+\epsilon)\|x-y\|^2$$

approx. isometry of sq. distances implies approx. isometry of distances

- log(n) *compressibility with just* ($1 \pm \varepsilon$) *distortion !*
- *a* **LINEAR** *f (ie a linear projection) can achieve this !!*
- *a* **RANDOM** *d-dim subspace can achieve this w.h.p. !!!*

Johnson-Lindenstrauss "flattening" Lemma

Theorem (Johnson and Lindenstrauss '84):

For any n and 0 < ε < ½ , let $d \geq \left\lceil \frac{4}{\varepsilon^2}(2\ln n + \ln 3) \right\rceil = \Omega\left(\frac{1}{\varepsilon^2}\ln n\right)$. Then for any X ⊂ R^D s.t. |X| = n, there exists a mapping f : $R^D \rightarrow R^d$, s.t. for all x,y ∈ X

$$(1-\epsilon)\|x-y\|^2 \leq \|f(x)-f(y)\|^2 \leq (1+\epsilon)\|x-y\|^2$$

Proof idea:

Consider a fixed vector v ∈ R^D;

for any d x D matrix L (entries mean 0, variance 1/d iid)

$$\mathbb{E}\left[\|Lv\|^2\right] = \mathbb{E}\Big[\sum_{i=1}^{d}(L^i \cdot v)^2\Big] = \sum_{i=1}^{d}\mathbb{E}\Big[\sum_{j,j'}(L^i_j v_j L^i_{j'} v_{j'})\Big]$$

$$= \sum_{i=1}^{d}\sum_{j=1}^{D} v_j^2 \cdot \mathbb{E}(L^i_j)^2 = \|v\|^2$$

so, by smartly selecting L we can ensure (sq) length preservation in expectation

JL Proof Cont.

We can preserve lengths on avg., what about for a *particular realization* of L?

Idea:

- choose a distribution that is concentrated around its mean value (ie 0).
- Then we expect that a specific realization of L acts like the average case!

Suppose we have a d x D matrix L such that:

$$\mathbb{P}_L\left[\|Lv\|^2 \leq (1-\epsilon)\|v\|^2 \text{ or } \|Lv\|^2 \geq (1+\epsilon)\|v\|^2\right] \leq e^{-\Omega(d\epsilon^2)}$$

bad event (L distorts v too much)

ie the chance of bad event occurring is exponentially bounded

Then: for any fixed pair x, y of points in R^D we have (set v = x – y above)

$$\mathbb{P}_L\left[\|Lx-Ly\|^2 \leq (1-\epsilon)\|x-y\|^2 \text{ or } \|Lx-Ly\|^2 \geq (1+\epsilon)\|x-y\|^2\right] \leq e^{-\Omega(d\epsilon^2)}$$

ie the chance that L distorts the interpoint distance between x and y by more than (1±ε) is exponentially small!

JL Proof Cont.

Now suppose we want to approx. preserve lengths of all points in $X \subset R^D$

Then want:

P_L [$\exists$ pair x,y$\in$ X s.t. $\begin{matrix} \|Lx - Ly\|^2 \leq (1-\epsilon)\|x-y\|^2 \\ \|Lx - Ly\|^2 \geq (1+\epsilon)\|x-y\|^2 \end{matrix}$]

(Bad event, needs to be bounded)

$$\leq \binom{|X|}{2} e^{-\Omega(d\epsilon^2)} \leq |X|^2 e^{-\Omega(d\epsilon^2)}$$

So, we can pick d = Ω ((log |X|) / ε^2) to ensure that the bad event above < 1

*(this means by picking d sufficiently large, with non-zero probability the complement is true, so such a **linear map exists**!)*

All what is left to show…

there are distributions which generate L with the property

$$\mathbb{P}_L\left[\|Lv\|^2 \leq (1-\epsilon)\|v\|^2 \text{ or } \|Lv\|^2 \geq (1+\epsilon)\|v\|^2\right] \leq e^{-\Omega(d\epsilon^2)}$$

JL Proof Cont.

Let each entry of dxD of L be i.i.d N(0,1/d), then:

Recall: choose a distribution that is concentrated around its mean value (ie 0).

$$\mathbb{P}_L\left[\|Lv\|^2 \leq (1-\epsilon)\|v\|^2 \text{ or } \|Lv\|^2 \geq (1+\epsilon)\|v\|^2\right] \leq e^{-\Omega(d\epsilon^2)}$$

This can be proved by applying an exponential concentration inequality, such as the Chernoff bound!

$\blacksquare$

Significance and Applications of JL Lemma

- Dimensionality Reduction

  Random Projections

- Data compression (Compressed Sensing)

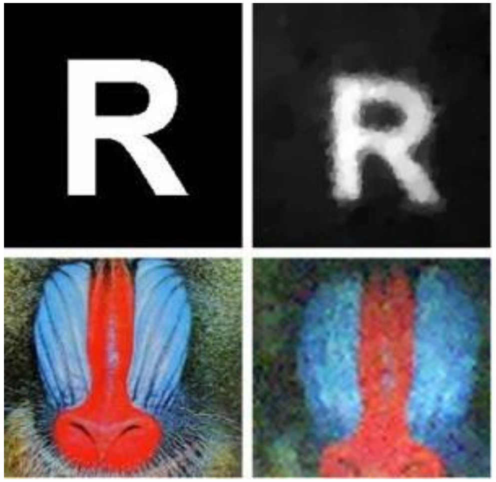
  Single pixel camera

- Approximate Linear algebra (Sketching methods)

- Fast Approximate nearest neighbors

  Locality Sensitive Hashing (LSH), Random projection trees (RPTrees)

- Fast Provable Clusterings



Thoughts on JL Lemma

JL Lemma is a very powerful result that has found numerous applications throughout CS

- Can you use other ‘concentrated’ distributions?

  Yes, any subguassian distribution work do

- What about the amount of randomness?

  Derandomized versions exist

- What about optimality, can we compress the space further?

  Negative results exists stating you cannot compress down further for linear and even non-linear maps (Alon ’03, Larson and Nelson ’16, ’17)

Extensions of JL Lemma

JL only guarantees approximate preservation of **finite pairs** of interpoint distances.

What if we want to preserve an infinite set of pairs?

In general it cannot be done… unless the set is structured in some way

- Subspace extension
- Sparse extension (Devenport et al. ’08)
- Manifolds extension (Baraniuk/Wakin ’ 08, Clarkson ’09 and Verma ’11)

The key idea is to do a clever application of the covering argument