Plan

Logistics

Review

Johnson-Lindenstrauss Lemma

Games Tonight @ 6 Tonight here

Tea Time! @ 2 Friday Bihall

Project Proposal

Gradescope

Recommend: Finish problem in class (then write up on your own)

Review

$\|x_i\|_2^2 = 1$

$x_1, \ldots, x_t \in \mathbb{R}^d$ nearly orthogonal

$|\langle x_i, x_j \rangle| \leq \epsilon \quad i \neq j$

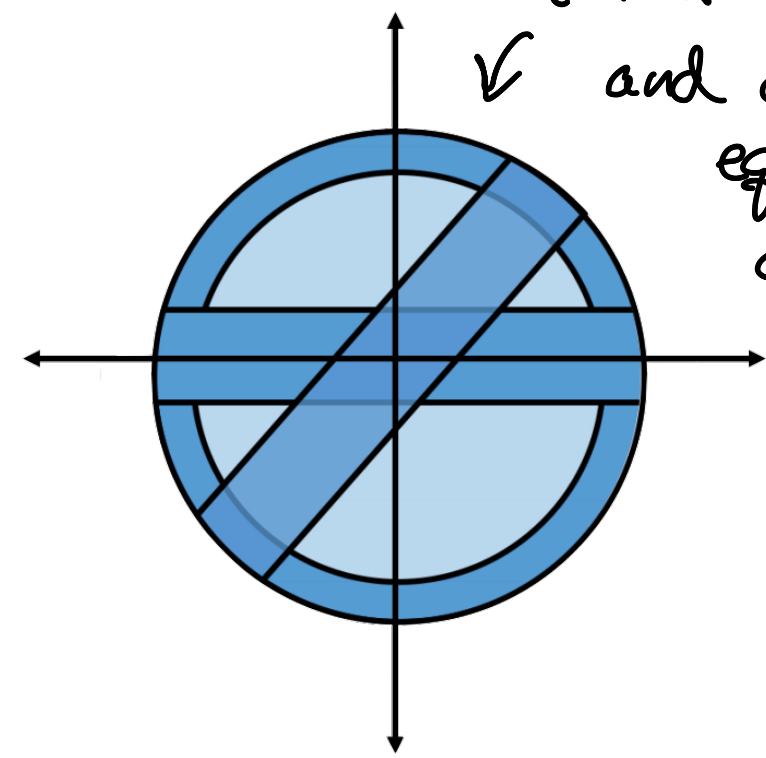
Probabilistic Method

$\Pr(x_1, \ldots, x_t$ are nearly ortho$) > 0$

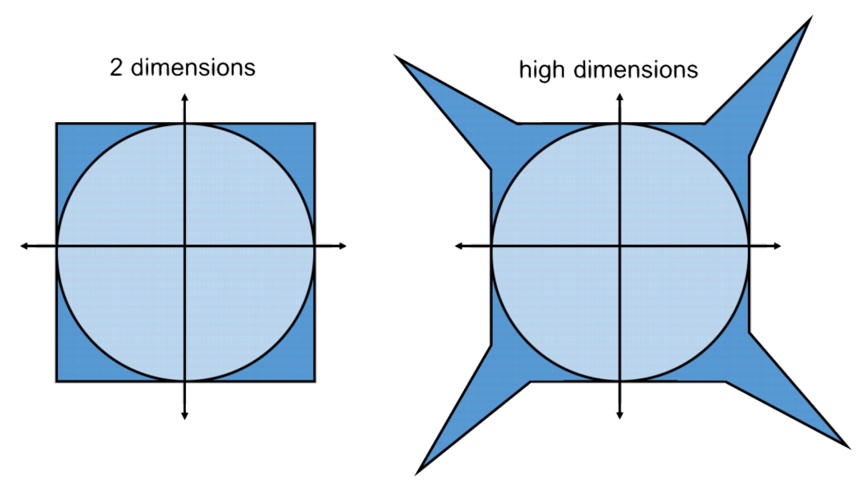
$\Rightarrow \exists$ at least one set of nearly ortho vectors

$t = 2^{c \cdot \epsilon^2 d}$

exponential in d



ϵ near surface and all equators contain almost all volume



$$\frac{\mathrm{Vol}(C_d)}{\mathrm{Vol}(B_d)} \approx d^d$$

High dimensional geometry is weird but we do want to work with it...

How do we represent data using less space while <u>approximately</u> preserving the structure?

Johnson-Lindenstrauss Lemma

↳ <u>Lemma</u> in math paper

↳ Took CS people years to find

JL Lemma

$q_1, \dots, q_n \in \mathbb{R}^d$

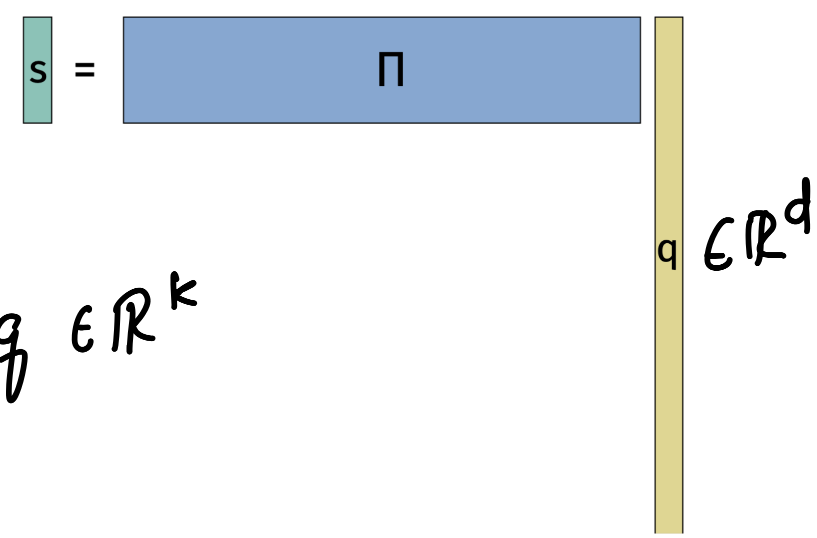
There exist (random) linear map

$\Pi : \mathbb{R}^d \to \mathbb{R}^k \qquad k = O\left(\frac{\log n}{\epsilon^2}\right)$

no dependence on d

$$(1-\epsilon)\|q_i - q_j\|_2 \leq \|\Pi q_i - \Pi q_j\|_2 \leq (1+\epsilon)\|q_i - q_j\|_2$$

for all $i \neq j$ wp $9/10$



$s = \Pi q \in \mathbb{R}^k$

$$\|x\|_2^2 = \sum_{i=1}^{d} (x[i])^2 \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{d} (x[i])^2}$$

$\epsilon < 1/2$

$\epsilon^2 < \frac{1}{2}\epsilon$

$$(1-\epsilon)\|q_i - q_j\|_2 \le \|\Pi q_i - \Pi q_j\|_2 \le (1+\epsilon)\|q_i - q_j\|_2$$

when $k = O\left(\frac{\log n}{\epsilon^2}\right)$

$$(1-\epsilon)^2\|q_i - q_j\|_2^2 \le \|\Pi q_i - \Pi q_j\|_2^2 \le (1+\epsilon)^2\|q_i - q_j\|_2^2$$

$(1-\epsilon)^2 = 1 - 2\epsilon + \epsilon^2$ $\qquad$ $(1+\epsilon)^2 = 1 + 2\epsilon + \epsilon^2$

$(1-\epsilon)^2 = 1 - \epsilon \cdot \text{constant}$ $\qquad$ $(1+\epsilon)^2 = 1 + \epsilon \cdot \text{constant}$ $\qquad$ when ϵ small

$\epsilon' = \epsilon \cdot \text{constant}$

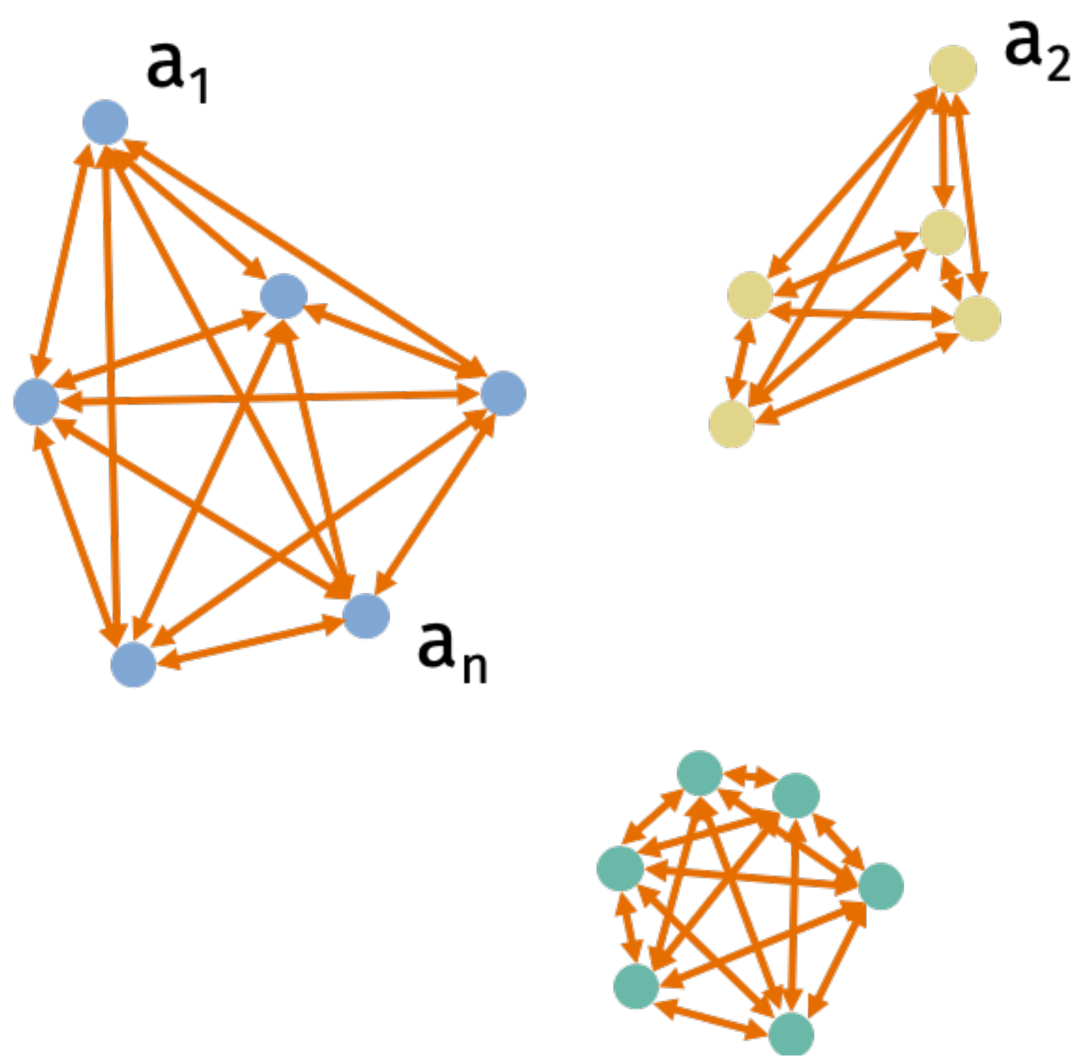
$$(1-\epsilon')\|q_i - q_j\|_2^2 \le \|\Pi q_i - \Pi q_j\|_2^2 \le (1+\epsilon')\|q_i - q_j\|_2^2$$

when $k = O\left(\frac{\log n}{\epsilon^2}\right)$

$k = O\left(\frac{\log n}{\epsilon'^2}\right)$

Clustering

Problem: Group points $a_1, \ldots, a_n \in \mathbb{R}^d$ into k clusters $C = \{C_1, \ldots, C_k\}$



$$\text{Cost}(C) = \sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{u,v \in C_j} \|a_u - a_v\|_2^2$$

NP-hard but we can approximate in time depending on d

Idea:

1. compress data (approximately preserves distance)
2. Cluster on compressed (approx solution)
3. Return cluster

JL Lemma

What is Π?

Can we efficiently compute Π?

$\Pi \in \mathbb{R}^{k \times d}$ is random matrix

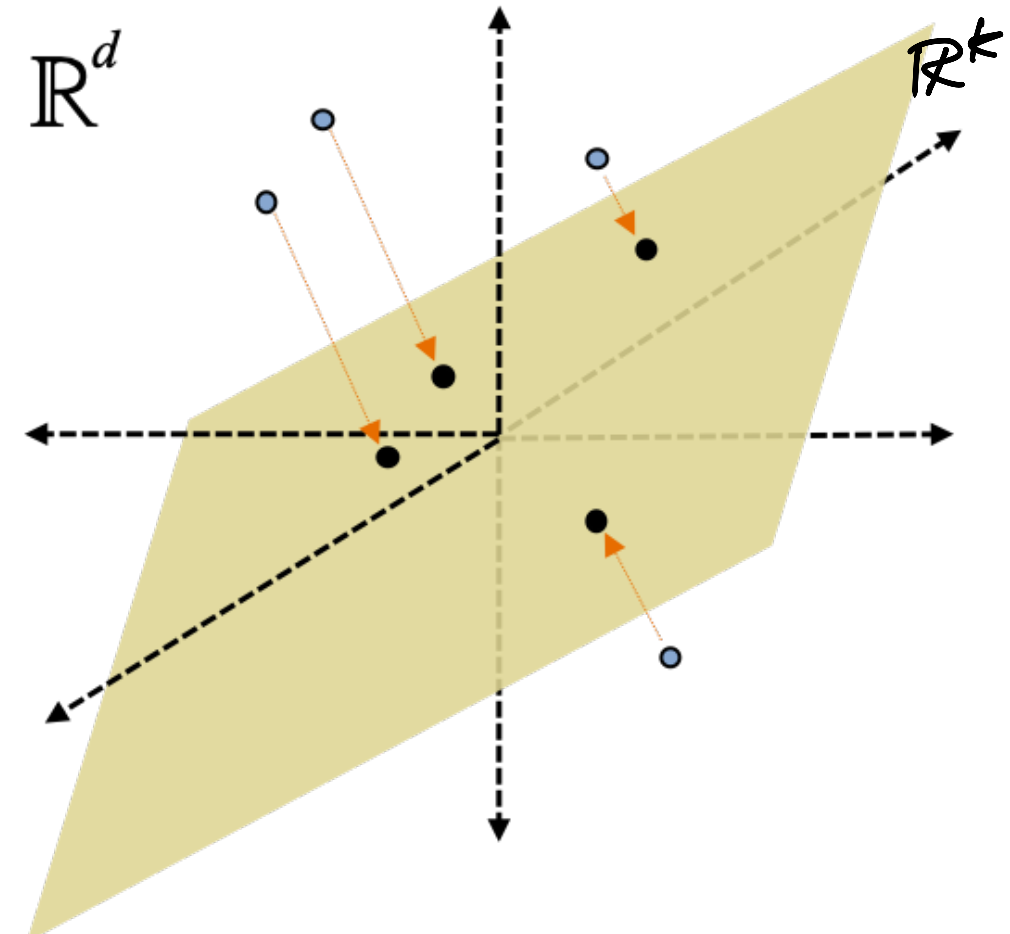
$\Pi_{i,j} \sim \mathcal{N}(0,1) \cdot \frac{1}{\sqrt{k}}$ preserving norm

Other random matrices work, too!

↳ binary

↳ sparse

↳ pseudorandom



Does projection need to be random? Why?

Distributional JL Lemma

$\Pi \in \mathbb{R}^{k \times d}$ random scaled matrix

$k = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$

Then wp $1-\delta$

$x = q_i - q_j$

$(1-\epsilon)\|x\|_2^2 \le \|\Pi x\|_2^2 \le (1+\epsilon)\|x\|_2^2$

$(1-\epsilon)\|q_i - q_j\|_2^2 \le \|\Pi q_i - \Pi q_j\|_2^2 \le (1+\epsilon)\|q_i - q_j\|_2^2$ wp $1-\delta$

Pr(fails on any)

$\le \binom{n}{2}$ Pr(fails on one)

$\le n^2 \underbrace{\text{Pr(fails on one)}}_{\delta}$ want $\le 1/10$

$\delta = \frac{1}{10n^2}$

pairs $i,j = \binom{n}{2} \le n^2$

How do we prove JL lemma using distributional JL lemma?

$O\left(\frac{\log(1/\delta)}{\epsilon^2}\right) = O\left(\frac{\log 10n^2}{\epsilon^2}\right) = O\left(\frac{\log n}{\epsilon^2}\right)$

Proving Distributional JL

$$(1-\epsilon)\|x\|_2^2 \le \|\Pi x\|_2^2 \le (1+\epsilon)\|x\|_2^2$$

$$\Longleftrightarrow$$

$$\left|\,\|\Pi x\|_2^2 - \|x\|_2^2\,\right| \le \epsilon \|x\|_2^2$$

Concentration!

$$\mathbb{E}\left[\|\Pi x\|_2^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[(\langle \Pi_i, x\rangle)^2\right]$$

$$= \frac{1}{k}\sum_{i=1}^{k} \sum_{j=1}^{d} \mathbb{E}\left[(\Pi_{i,j}\, x_j)^2\right]$$

linearity of variance

$$= \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left[Z_i^2\right] = \frac{1}{k}\sum_{i=1}^{k} \|x\|_2^2 = \|x\|_2^2$$

Stability of Gaussians

$$X_1 \sim N(\mu_1, \sigma_1^2)$$

$$X_2 \sim N(\mu_2, \sigma_2^2)$$

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

$$Z_i = \sum_{j=1}^{d} \Pi_{i,j}\, x_j$$

$$\mathbb{E}[cY] = c\,\mathbb{E}[Y]$$

$$\Pi_{i,j} \sim N(0,1)$$

$$Var(cY) = c^2\, Var(Y)$$

$$\Pi_{i,j}\, x_j \sim N(0, x_j^2)$$

$$Z_i \sim N(0, \|x\|_2^2)$$

Chernoff? :(

We actually know this distribution!

Z be chi-squared r.v. with k degrees of freedom (sum of k squared normal)

$$\Pr\left(|Z - \mathbb{E}[Z]| \geq \epsilon \cdot \mathbb{E}[Z] \right) \leq 2e^{-\epsilon^2 k/8}$$

$$Z = \frac{1}{k} \sum_{i=1}^{k} z_i^2 = \|\Pi x\|_2^2$$

$$\mathbb{E}[Z] = \|x\|_2^2$$

$$\Pr\left(\left| \|\Pi x\|_2^2 - \|x\|_2^2 \right| \geq \epsilon \|x\|_2^2 \right) \leq 2e^{-\epsilon^2 k/8} \overset{\text{want}}{=} \delta$$

$$2e^{-\epsilon^2 k/8} = \delta$$

$$\log e^{-\epsilon^2 k/8} = \log \frac{\delta}{2}$$

$$+\epsilon^2 k/8 = \log(2/\delta)$$

$$k = \frac{8 \log 2/\delta}{\epsilon^2}$$

$$k = O\left(\frac{\log(1/\delta)}{\epsilon^2} \right)$$

Is JL tight?

$x_1, \dots, x_n \in \mathbb{R}^d$ orthonormal

$$\|x_i - x_j\|_2^2 = \langle x_i - x_j, x_i - x_j \rangle = \langle x_i, x_i - x_j \rangle - \langle x_j, x_i - x_j \rangle$$
$$= \|x_i\|_2^2 - \langle x_i, x_j \rangle - \langle x_j, x_i \rangle + \|x_j\|_2^2$$
$$= \|x_i\|_2^2 - 2\langle x_i, x_j \rangle + \|x_j\|_2^2 \overset{\text{orthonormal}}{=} 2$$

JL says we can compress to $(1 \pm \epsilon)$ in $k = O\left(\frac{\log n}{\epsilon^2}\right)$ dimensions

From nearly orthogonal, we know there are $2^{O(\epsilon^2 d)}$ nearly ortho in d

When $d = k$, there $2^{O\left(\epsilon^2 \cdot \frac{\log n}{\epsilon^2}\right)} \approx n$ nearly orthogonal vectors in k