Plan

Logistics

Review

Johnson-Lindenstrauss Lemma

Games!

Tea!

Proposal

Review

$\|x_i\|_2^2 = 1 \quad \forall i$

$x_1, \dots, x_t$ nearly orthogonal

if $|\langle x_i, x_j \rangle| < \epsilon$ for $i \neq j$

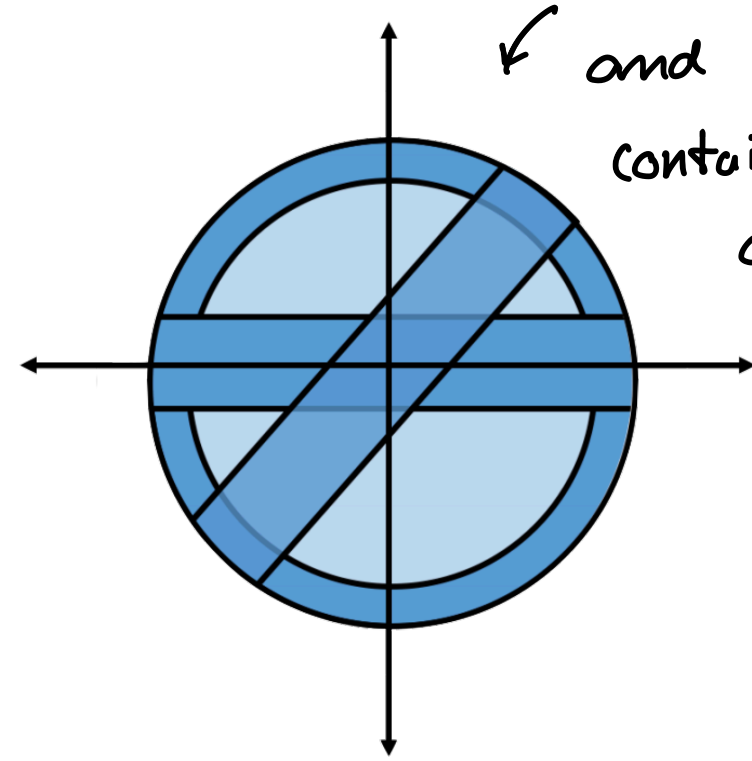
Probabilistic method:

$\Pr(x_1, \dots, x_t \text{ nearly ortho}) > 0$

$\Rightarrow \exists$ nearly ortho $x_1, \dots, x_t$

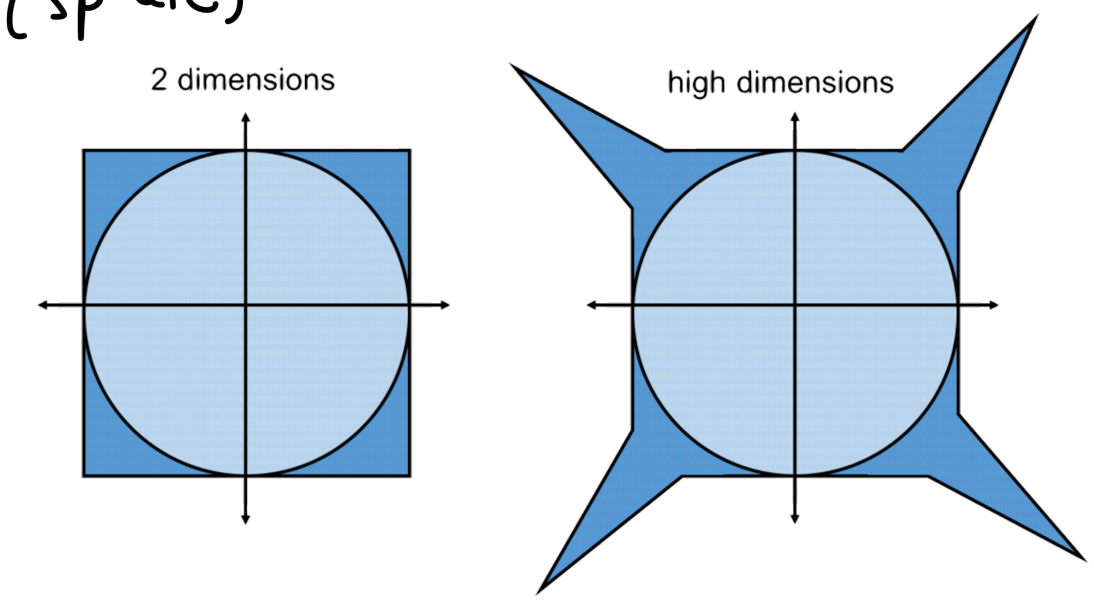
Proved when $t = 2^{c\epsilon^2 d}$

exponential!



ϵ-near surface and all equators contain almost all volume!

$$\frac{Vol(cube)}{Vol(sphere)} \gtrsim d^d$$



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

How do we represent data using less space while approximately preserving structure?

Johnson-Lindenstrauss Lemma

↳ Lemma in math paper

↳ Years for CS community to find

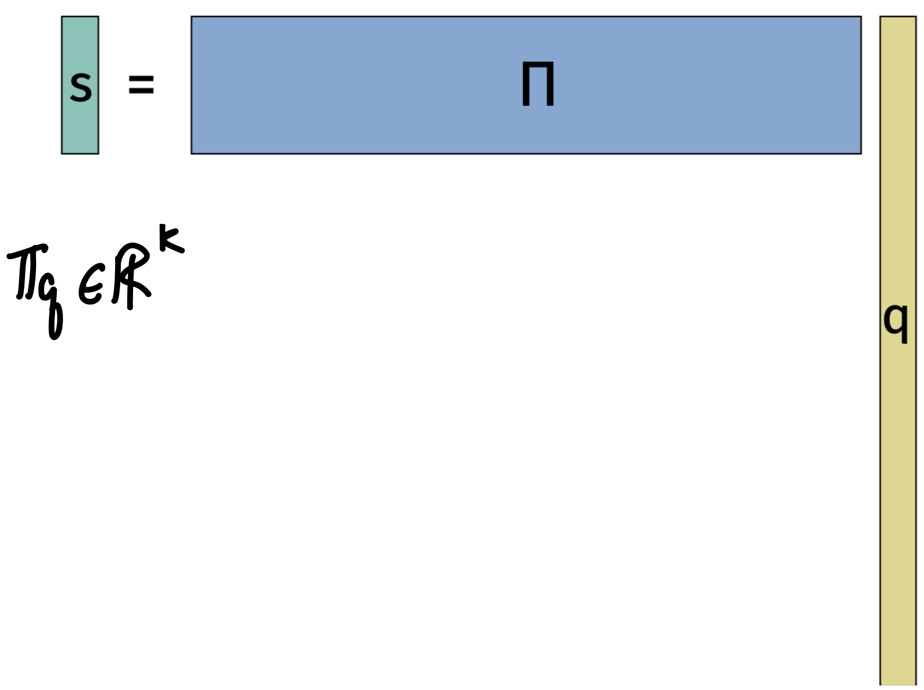
JL Lemma

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

There exists $\Pi: \mathbb{R}^d \to \mathbb{R}^k$

for $k = O\left(\frac{\log n}{\epsilon^2}\right)$ independent of d so that

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



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

$\in \mathbb{R}^d$

What about squared norm?

$$(1-\epsilon)\|q_i - q_j\|_2 \le \|\Pi q_i - \Pi q_j\|_2 \le (1+\epsilon)\|q_i - q_j\|_2 \quad \text{with} \quad k = O\left(\frac{\log n}{\epsilon^2}\right)$$

$$(1+\epsilon)^2 = 1 + 2\epsilon + \epsilon^2 = 1 + \epsilon \cdot \text{constant} \quad \text{for small } \epsilon$$

$$(1-\epsilon)^2\|q_i - q_j\|_2^2 \le \|\Pi q_i - \Pi q_j\|_2^2 \le (1+\epsilon)^2\|q_i - q_j\|_2^2 \quad \text{with} \quad k = O\left(\frac{\log n}{\epsilon^2}\right)$$

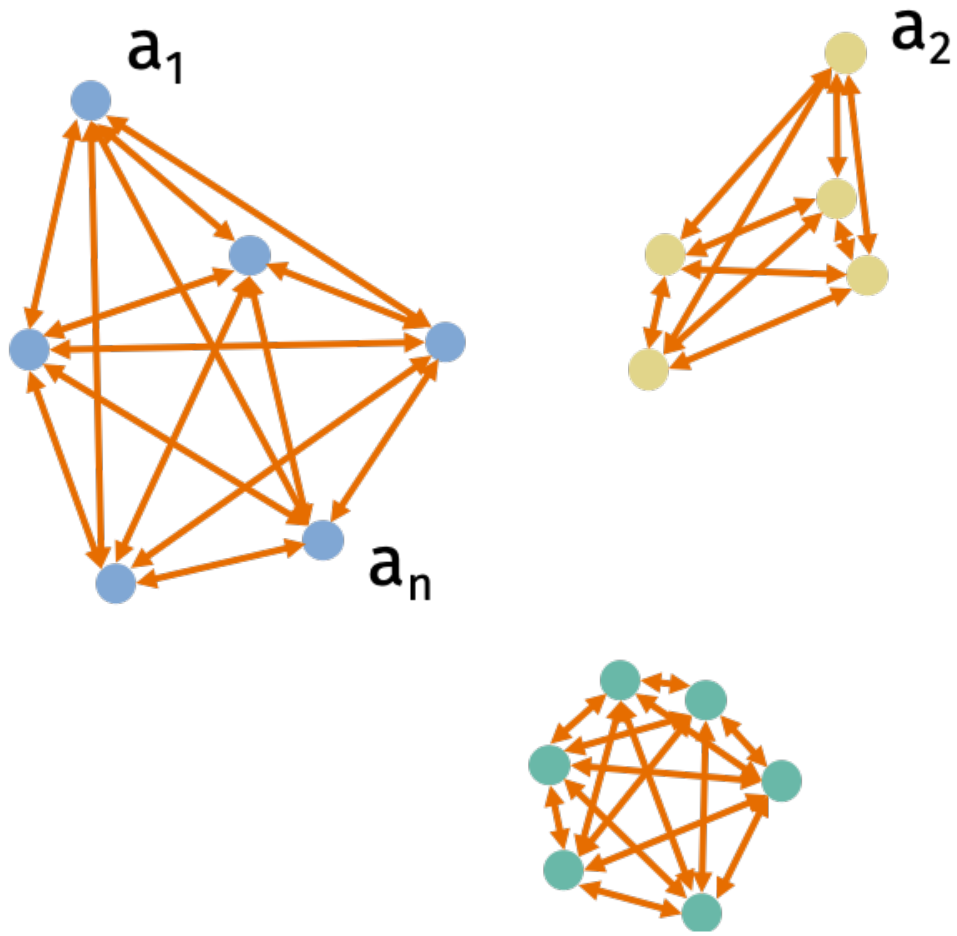
$$(1-\epsilon')\|q_i - q_j\|_2^2 \le \|\Pi q_i - \Pi q_j\|_2^2 \le (1+\epsilon)\|q_i - q_j\|_2^2 \quad \text{with} \quad k = O\left(\frac{\log n}{(\epsilon')^2}\right)$$

$$\epsilon' = 2\epsilon + \epsilon^2$$
$$= \epsilon \cdot \text{constant}$$

Clustering with k-means

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

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



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

Idea:

1. Compress data (Approximately preserve distance)
2. Cluster on compressed (Approximate solution) — faster!
3. Return cluster

JL Lemma

what is Π?

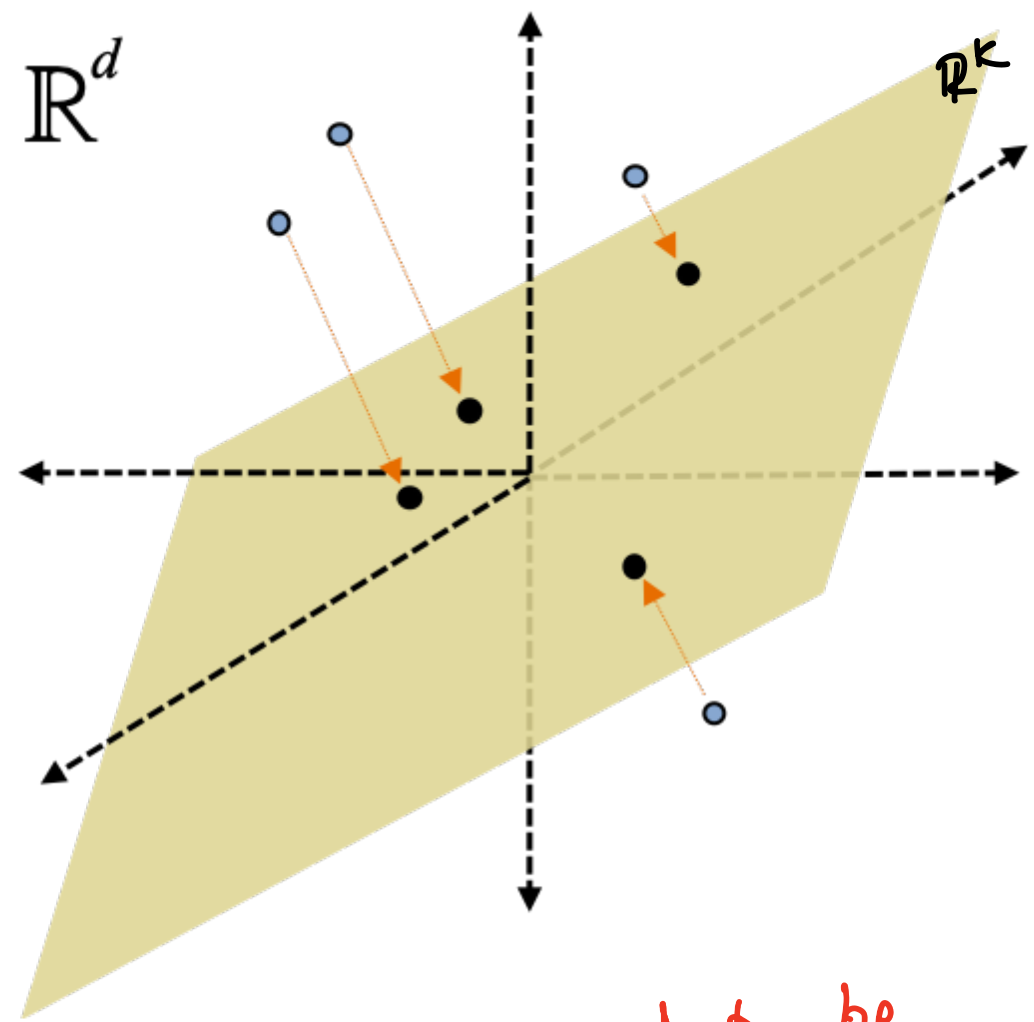
Can we efficiently compute Π?

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

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

Other random matrices work, too!

↳ binary
↳ sparse
↳ pseudorandom



↑ Does projection need to be random?

Distributional JL Lemma

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

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

Then wp $1-\delta$

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

How do we prove JL with distributional JL?

Proving Distributional JL

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

$$\Longleftrightarrow$$

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

Concentration!

ith row of Π

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

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

linearity of variance

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

Stability of Gaussians

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

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

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

$$Z_i = \sum_{j=1}^{d} \Pi_i[j]\, X[j]$$

$$\Pi_i[j] \sim \mathcal{N}(0,1) \qquad Var(cX) = c^2 Var(X)$$

$$\Pi_i[j] \cdot x[j] \sim \mathcal{N}(0, (x[j])^2)$$

$$Z_i \sim \mathcal{N}(0, \|x[j]\|_2^2)$$

Chernoff? $\ddot{\frown}$

We actually know this distribution!

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

$$\Pr(|Z - \mathbb{E}[Z]| > \epsilon\,\mathbb{E}[Z]) \leq 2e^{-\epsilon^2 k/8}$$

$$Z = \frac{1}{k}\sum_{i=1}^{k} z_i^2 = \frac{1}{k}\sum_{i=1}^{k} (\langle \pi_i, x \rangle)^2 = \|\pi x\|_2^2$$

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

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

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

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

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

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

Can we hope for fewer dimensions?

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

Then $\|x_i - x_j\|_2^2 = \langle x_i - x_j, x_i - x_j \rangle$

$$= \langle x_i, x_i - x_j \rangle - \langle x_j, x_i - x_j \rangle$$

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

$$= \underbrace{\|x_i\|_2^2}_{1} - \underbrace{\langle x_i, x_j \rangle}_{0} - \underbrace{\langle x_j, x_i \rangle}_{0} + \underbrace{\|x_j\|_2^2}_{1}$$

$$\overset{\text{orthonormal}}{=} 2$$

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

In k, there are $2^{O(\epsilon^2 \cdot \frac{\log n}{\epsilon^2})} \approx n$ nearly orthogonal vectors so we can't put any more in (up to constants)!