Tuesday, Feb 24

- Missed quiz policy
  1. Reach out before class
  2. Take quiz at first OH after

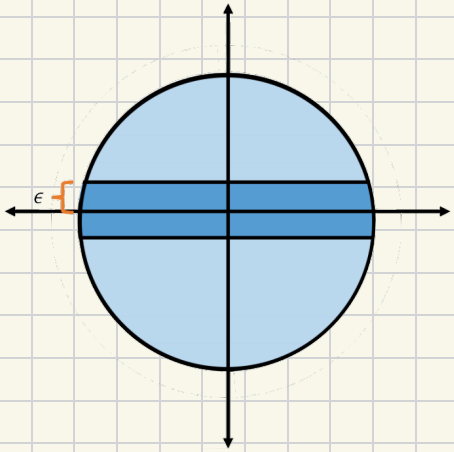
Goal: Check understanding

Incentivize class attendance

- Late assignment policy
  - ↳ 24 hours late window (life happens)
  - ↳ hard deadline after that

Plan

- High dimensional geometry is weird
- JL projection



Q: What fraction of volume is ϵ close to equator?

A: All but $\frac{1}{2^{\epsilon cd}}$ fraction of volume

ϵ-close to any equator

Draw random points from unit ball

Goal: Show that $x \sim B_d$ has $|x_i| \leq \epsilon$ wp $1 - \frac{1}{2^{c\epsilon d}}$

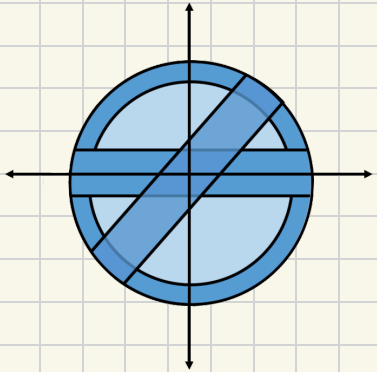
Given x from interior of unit ball, $w = \frac{x}{\|x\|_2}$ from surface

If $|w_i| \leq \epsilon \Rightarrow |x_i| < \epsilon$ since $\|x\|_2 \leq 1$

New Goal: Show that w from surface has $|x_i| \leq \epsilon$ wp $1 - \frac{1}{2^{c\epsilon d}}$

Let $g \sim N(0, I)$. $w = \frac{g}{\|g\|_2}$ from surface by rotational invariance

$\mathbb{E}\, \|g\|_2^2 = \ldots$



If:

① $\|g\|_2 \geq \sqrt{d/2}$

② $|g_1| \leq \epsilon\sqrt{d/2}$

Then

$$|w_1| = \frac{|g_1|}{\|g\|_2} \leq \frac{\epsilon\sqrt{d/2}}{\sqrt{d/2}} = \epsilon$$

$$\Pr(|w_1| \leq \epsilon) \geq \Pr(① \text{ and } ②)$$
$$\geq 1 - \Pr(①^c) - \Pr(②^c)$$

$$\Pr(① \text{ and } ②) = \Pr(①) + \Pr(②) - \Pr(① \text{ or } ②)$$
$$\geq \Pr(①) + \Pr(②) - 1$$
$$= (1 - \Pr(①^c)) + (1 - \Pr(②^c)) - 1$$
$$= 1 - \Pr(①^c) - \Pr(②^c)$$

$\Pr(①^c) = \Pr(\|g\|_2 < \sqrt{d/2}) \leq \frac{1}{2^{cd}}$ by Johnson-Lindenstrauss Lemma

$\Pr(②^c) = \Pr(|g_1| > \epsilon\sqrt{d/2}) \leq \frac{1}{2^{(c\epsilon\sqrt{d/2})^2}}$ by Gaussian tail bound

larger for small ϵ

$$\geq 1 - \frac{1}{2^{c^2\epsilon^2 d/2}} - \frac{1}{2^{cd}}$$

Despite warnings, let's play with high-dimensional data!

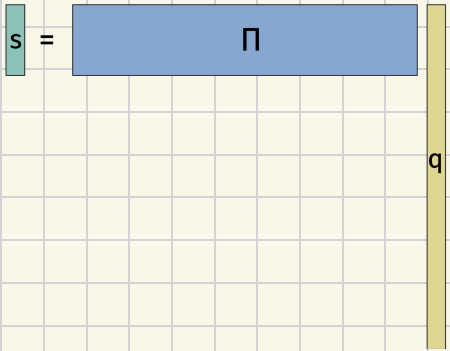
Goal: Compress to smaller dimension while preserving structure

Johnson-Lindenstrauss Lemma: $q_1, \dots, q_n \in \mathbb{R}^d$. $\exists\ \Pi: \mathbb{R}^d \to \mathbb{R}^k$ for $k = O\left(\frac{\log n}{\epsilon^2}\right)$ s.t., for all $i,j \in [n]$, wp 9/10,

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

"Lemma" as a stepping stone to another result, immensely useful

$(1 \pm \epsilon)^2 \approx (1 \pm \epsilon)$ for small ϵ



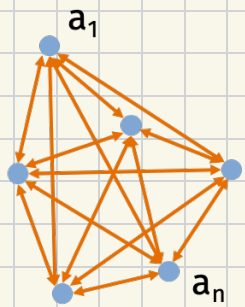
Clustering Application

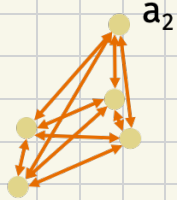
Points $a_1, \ldots, a_n \subset \mathbb{R}^d$

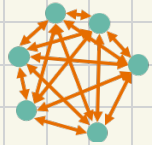
Partition $[n]$ into m clusters

$C = \{C_1, \ldots, C_m\}$

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







Exact solution is NP-hard.
Approximate efficiently with polynomial dependence on d.

Goal: Decrease d

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

is cost on projected data

By JL,

$$(1-\epsilon)\,\text{cost}(c) \le \widetilde{\text{cost}}(c) \le (1+\epsilon)\,\text{cost}(c)$$

With an approx algorithm, find ← optimal on projected

$$\widetilde{\text{cost}}(c) \le (1+\alpha)\ \widetilde{\text{cost}}(\tilde{c}^*)$$

← optimal on original

Claim: $\text{Cost}(c) \le (1 + O(\alpha+\epsilon))\,\text{cost}(c^*)$

Hint: $\frac{1}{1-\epsilon} \approx 1+\epsilon$ for small ϵ

Thursday, Feb 26

- Exam 3/5

Plan

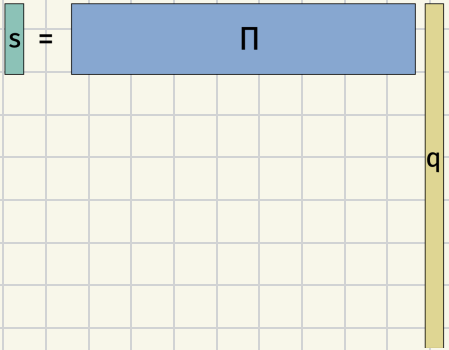
- Continue JL

Johnson-Lindenstrauss Lemma: $q_1, \dots, q_n \in \mathbb{R}^d$. $\exists\ \Pi: \mathbb{R}^d \to \mathbb{R}^k$

for $k = O\left(\frac{\log n}{\epsilon^2}\right)$ s.t., for all $i, j \in [n]$, wp 9/10,

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

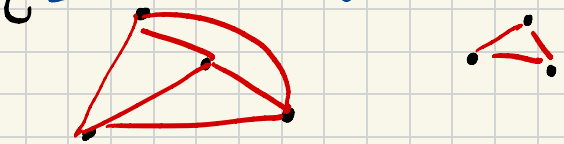
"Lemma" as a stepping stone to another result, immensely useful



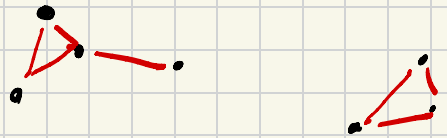
Clustering Application

$C = \{C_1, \ldots, C_m\}$

C^* ← optimal in original



$\tilde{C}^*$ ← optimal in projected



$$\text{cost}(C) = \sum_{j=1}^{m} \frac{1}{2|C_j|} \sum_{u,v \in C_j} \|a_u - a_v\|_2^2 \qquad \widetilde{\text{cost}}(C) = \sum_{j=1}^{m} \frac{1}{2|C_j|} \sum_{u,v \in C_j} \|\pi a_u - \pi a_v\|_2^2$$

JL: $(1-\epsilon)\,\text{cost}(C) \le \widetilde{\text{cost}}(C) \le (1+\epsilon)\,\text{cost}(C)$ $\quad \epsilon \in (0,1)$

Approx: Find C s.t. $\widetilde{\text{cost}}(C) \le (1+\alpha)\,\text{cost}(\tilde{C}^*)$

$$\begin{aligned}
\text{cost}(C) &\le \frac{1}{1-\epsilon}\widetilde{\text{cost}}(C) \le (1+2\epsilon)\,\widetilde{\text{cost}}(C) \le (1+2\epsilon)(1+\alpha)\,\widetilde{\text{cost}}(\tilde{C}^*) \\
&\le (1+4\epsilon+2\alpha)\,\widetilde{\text{cost}}(\tilde{C}^*) \le (1+4\epsilon+2\alpha)\,\widetilde{\text{cost}}(C^*) \\
&\le (1+4\epsilon+2\alpha)(1+\epsilon)\,\text{cost}(C^*)
\end{aligned}$$

Distributional JL Lemma

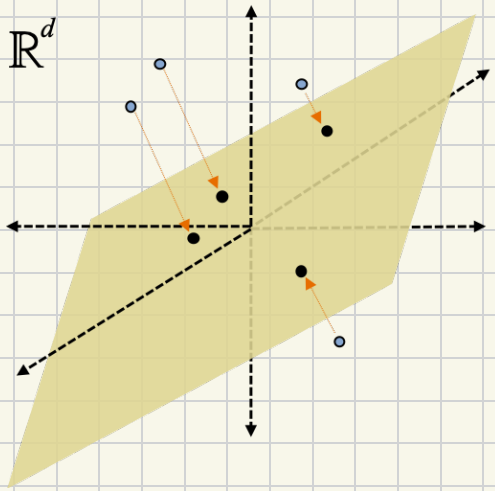
Q: Can we efficiently compute $\Pi \in \mathbb{R}^{k \times d}$?

Easiest to analyze:

$$[\Pi]_{i,j} = g/\sqrt{k} \quad \text{where} \quad g \sim N(0,1)$$

Let $k = O\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$. For fixed $x \in \mathbb{R}^d$, wp $1-\delta$,

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



close $\Leftrightarrow$ close
far $\Leftrightarrow$ far

Prove JL Lemma with $x = q_i - q_j$ and Union Bound

$g \in \mathbb{R}^k \quad g \sim N(0,1)$

Using dist. JL, prove $\Pr\left(\|g\|_2^2 \leq \frac{1}{2} \mathbb{E} \|g\|_2^2 \right) \leq \frac{1}{2^{ck}}$ for constant c

Proof of dist. JL

Goal: Show $\|\Pi x\|_2^2$ concentration

$$\mathbb{E}\|\Pi x\|_2^2 = \sum_{i=1}^{k} \frac{1}{k}\,\mathbb{E}[\langle \pi_i, x\rangle^2] = \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\Big(\overbrace{\sum_{j=1}^{d} \pi_i[j]\,x[j]}^{Z_i}\Big)^2$$

Fact: Stability of Gaussians. For $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$

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

Claim: $Z_i \sim \mathcal{N}(0, \|x\|_2^2)$.

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

$$Z = \frac{1}{k}\sum_{i=1}^{k} Z_i$$

↖ scaled chi-squared rv with k degrees of freedom

Chi-squared Concentration

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

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

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

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

JL tells us we can preserve high dimensional behavior...

But wait, isn't high-dimensional geometry weird?

Reconcile with "hard" case. Consider orthonormal $x_1, \dots, x_n \in \mathbb{R}^d$.

① JL says we can compress to $k = O(\log n / \epsilon^2)$ dimensions to retain ϵ-approx

② Probabilistic method tells us there are $2^{c\epsilon^2 k}$ nearly orthogonal vectors in k i.e., $2^{c\epsilon^2 k} = 2^{c \cdot c\epsilon^2 \log n / \epsilon^2} = 2^{c \log n} \approx n$ (between friends)

$\Rightarrow$ two sides of the same coin!

Alternate View: To have random points be "close" in d, we need $n \approx 2^d$ of them.

In this case, JL doesn't help since $\log n \approx d$