Tuesday, March 3

- Exam on Thursday
- Self-grade due Friday

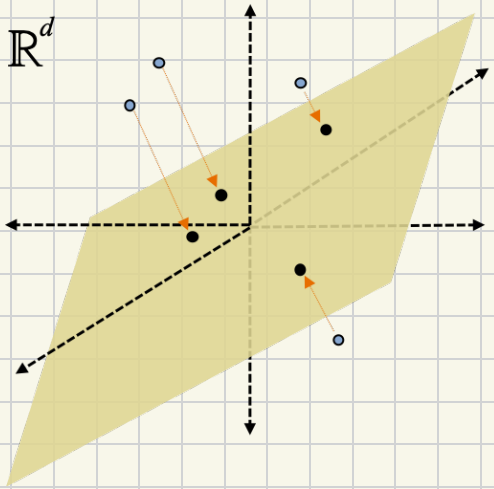
Plan

- Proof of distributional JL
- Questions
- Similarity Estimation

Distributional JL Lemma

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

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



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

Today:

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

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

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

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 $\mathbb{E}\,\|\Pi x\|_2^2 =$

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

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

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

$k =$

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.

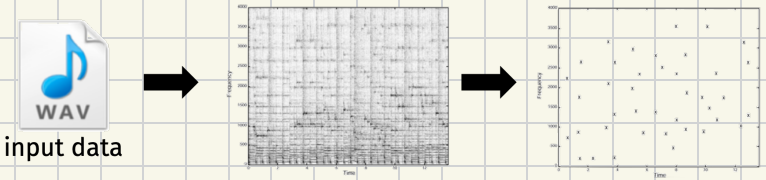
In this case, JL doesn't help since $\log n \approx d$

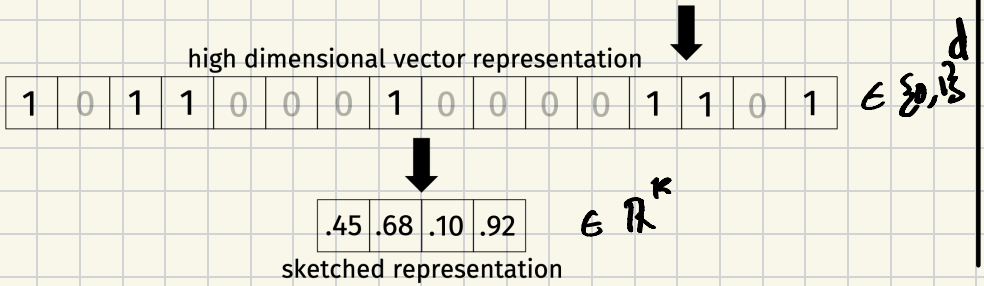
Similarity Estimation

Q: How does Shazam match song clips to a library of n = tens of millions of songs in a fraction of a second?

↑ background noise

↑ song clips offset





Query song $q \in \{0,1\}^d$

Goal: Find "nearby" $y \in \{0,1\}^d$

Challenge:

- $O(nd$ bits$)$
- $O(d)$ time to compare q, y

Building Block:

Sketch y into $C(y) \in \mathbb{R}^k$

to preserve Jaccard similarity

Jaccard Similarity

$$J(x,y) = \frac{|x \cap y|}{|x \cup y|} = \frac{\#\text{ non-zero entries in common}}{\#\text{ non-zero entries total}}$$

	X	Y
1	0	1
2	1	0
3	1	1
4	0	0
5	1	1

↑ vector representation

↙ set representation

$\{2,3,5\}$ $\{1,3,5\}$

Applications

- "bag of words"
- earthquakes
- cached webpages

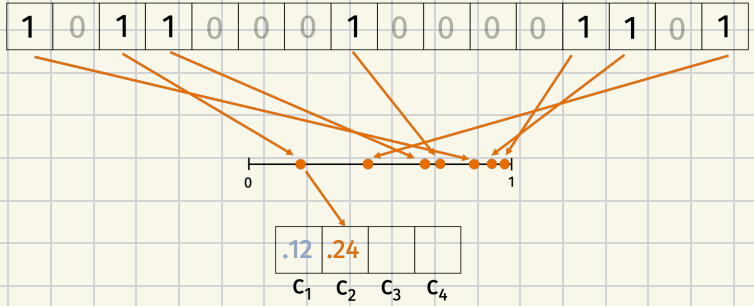
Min Hash Algorithm

$X \in \{0,1\}^d$

Compression function $C: \{0,1\}^d \to \mathbb{R}^k$

k Hash functions

$$C_i(x) = \min_{j \in [d]: x_j = 1} h_i(j)$$



Claim: $\Pr(c_i(x) = c_i(y)) = J(x,y) = \frac{|x \cap y|}{|x \cup y|}$

$$\Rightarrow \hat{J}(x,y) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}[c_i(x) = c_i(y)]$$

Claim: When $k = O(\frac{1}{\epsilon^2 \delta})$, w.p. $1-\delta$,

$$J(x,y) - \epsilon \leq \hat{J}(x,y) \leq J(x,y) + \epsilon$$