Plan

Logistics

Review

Similarity Search

Problem set

5pm Tomorrow Gradescope

↳ Please select which question you're answering

Proposal

5pm Monday Gradescope

JL Lemma

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

$\Pi \in \mathbb{R}^{k \times d}$ $\quad \Pi_{i,j} =$ random variable

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

wp 9/10

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

← no dimension dependence

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

$x = x_i - x_j$

Distributional JL Lemma

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

wp $1-\delta$

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

what if we want to preserve inner products?

$$|\langle x_i, x_j \rangle - \langle \Pi x_i, \Pi x_j \rangle| \le \frac{\epsilon}{2}\left(\|x_i\|_2^2 + \|x_j\|_2^2\right)$$

↳ using JL ↳ $\|x-y\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2\langle x, y\rangle$

Application: Fast Set Size Estimation

$X =$ people in class

$Y =$ people who climb

$$x = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{matrix} \leftarrow \text{Sujay} \\ \\ \leftarrow \text{Iris} \\ \\ \end{matrix} \qquad y = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \leftarrow \text{Aidan} \end{matrix}$$

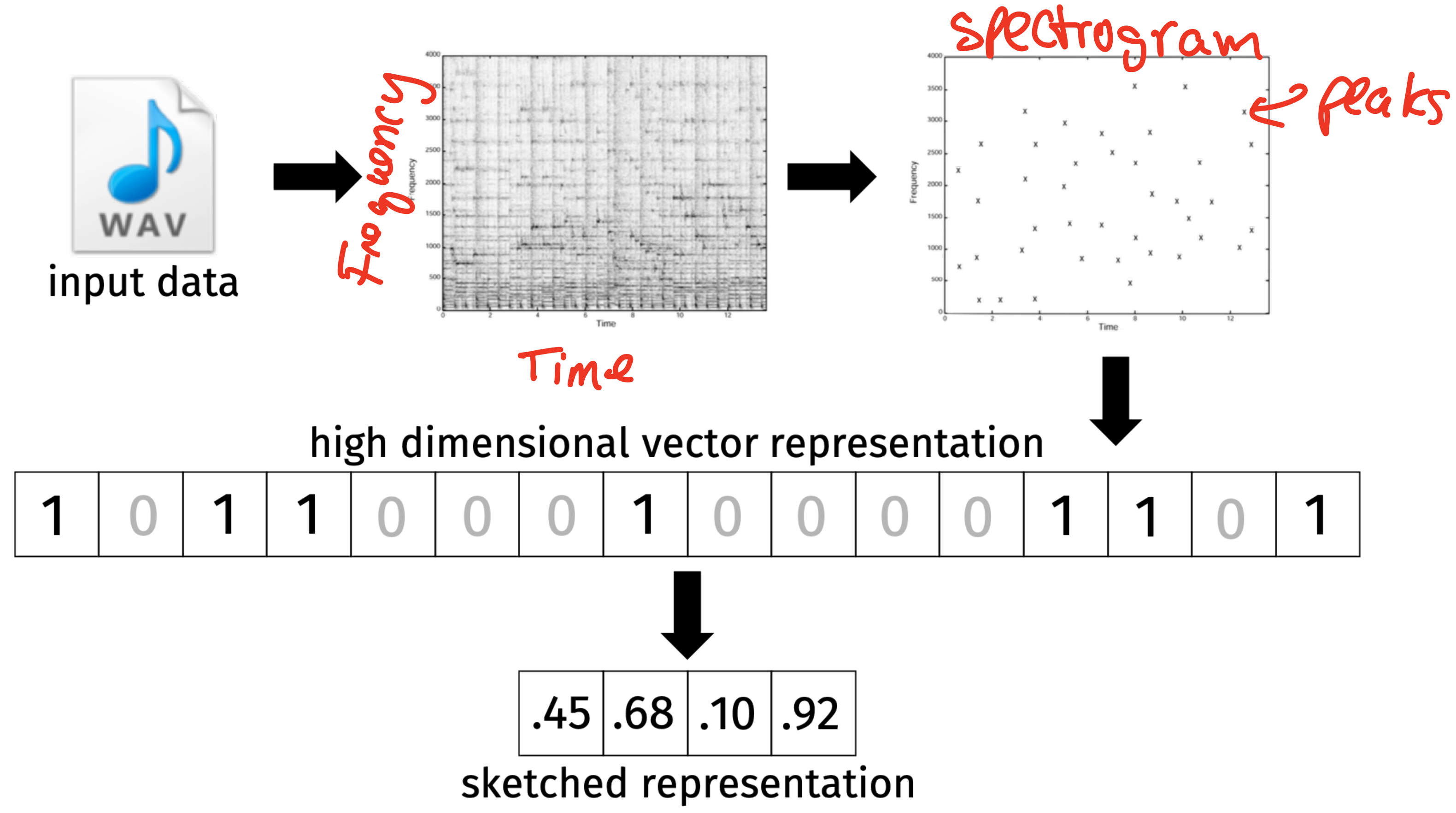
$\langle x, y \rangle = |X \cap Y|$

$\|x\|_2^2 = |X|$

Similarity Estimation

JL preserves distance.

How about "similarity"?



Shazam matches short, noisy clips against huge database

Problem: Given query $q \in \{0,1\}^d$, find similar song $y \in \{0,1\}^d$

Naive: $O(nd)$ space to store, $O(nd)$ time to search

"Sketch" into lower dimension

$$c: \{0,1\}^d \to \mathbb{R}^k \quad k \ll d$$

$$c(x) \approx c(y) \quad \text{if } x \approx y$$

similarity

Jaccard Similarity

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

$$= \frac{\#\text{ non-zero in common}}{\#\text{ non-zero total}}$$

e.g.:

$$x = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$J(x,y) = \frac{2}{5}$$

Also useful for

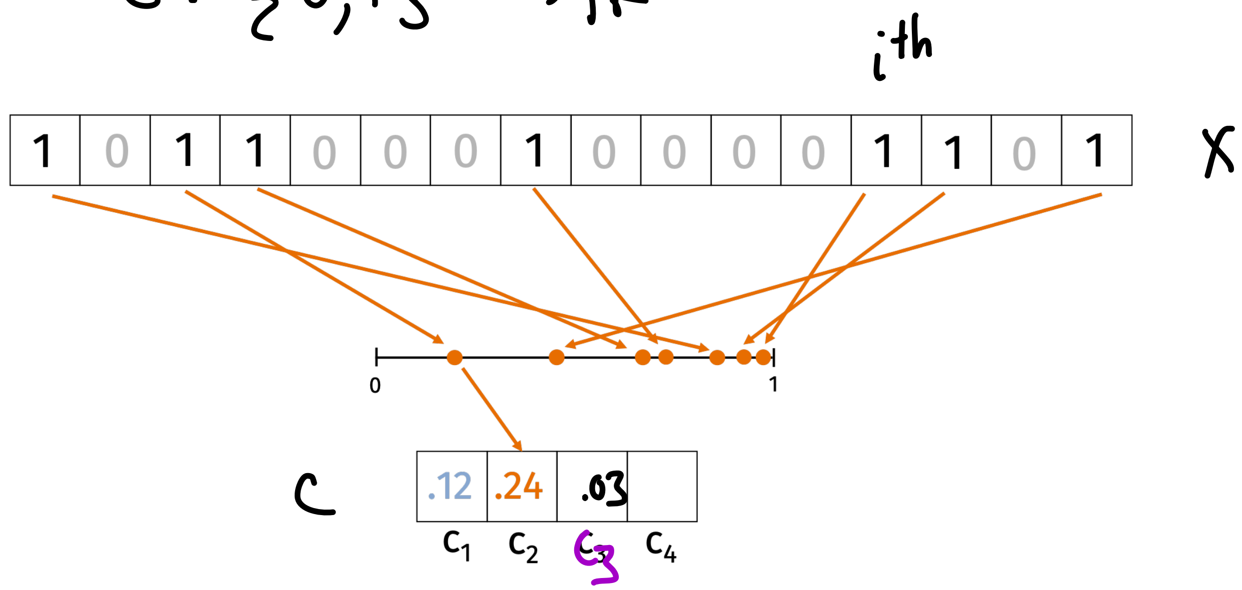
↳ "bag of words" document

↳ cached webpages

↳ earthquake detection

Minhash

$$c : \{0,1\}^d \to \mathbb{R}^k$$



$$h_i : \{1, \ldots, d\} \to [0,1]$$

$$c_i = \min_{j \,:\, x_j = 1} h_i(j)$$

$$\Pr\big(c_i(x) = c_i(y)\big) = \frac{|X \cap Y|}{|X \cup Y|} = J(x,y)$$

Estimate $J(x,y)$ using c

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

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

$$= J(x,y)$$

$$\mathrm{Var}(\hat{J}(x,y)) = \frac{1}{k^2} \sum_{i=1}^{k} \mathrm{Var}(\mathbb{1}(c_i(x) = c_i(y)))$$

$$\leq \frac{1}{k^2} \sum_{i=1}^{k} J(x,y) = \frac{J(x,y)}{k}$$

$$\mathrm{Var}(\mathbb{1}) = \mathbb{E}[\mathbb{1}^2] - \mathbb{E}[\mathbb{1}]^2$$

$$= J(x,y) - J(x,y)^2$$

Chebyshev's

$$\Pr\left(|\hat{J} - J| \geq \alpha \cdot \sigma\right) \leq \frac{1}{\alpha^2}$$

$$\sigma = \sqrt{\mathrm{Var}(\hat{J})} \approx \sqrt{\frac{J}{k}}$$

$$\Pr\left(|\hat{J} - J| \geq \alpha \cdot \frac{\sqrt{J}}{\sqrt{k}}\right) \leq \frac{1}{\alpha^2} \overset{\text{want}}{=} \delta$$

$$\Rightarrow \Pr\left(|\hat{J} - J| \geq \overbrace{\alpha \cdot \frac{1}{\sqrt{k}}}^{\epsilon}\right) \leq \frac{1}{\alpha^2}$$

$$\Pr\left(|\hat{J} - J| \geq \epsilon\right) \leq \delta$$

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

$$\epsilon = \alpha \frac{1}{\sqrt{k}} \qquad \delta = \frac{1}{\alpha^2}$$

$$\epsilon = \frac{1}{\sqrt{\delta}} \cdot \frac{1}{\sqrt{k}} \qquad \alpha = \frac{1}{\sqrt{\delta}}$$

$$\frac{1}{\epsilon^2 \delta} = k$$

using biased coin theorem

↳ heads if $c_i(x) = c_i(y)$

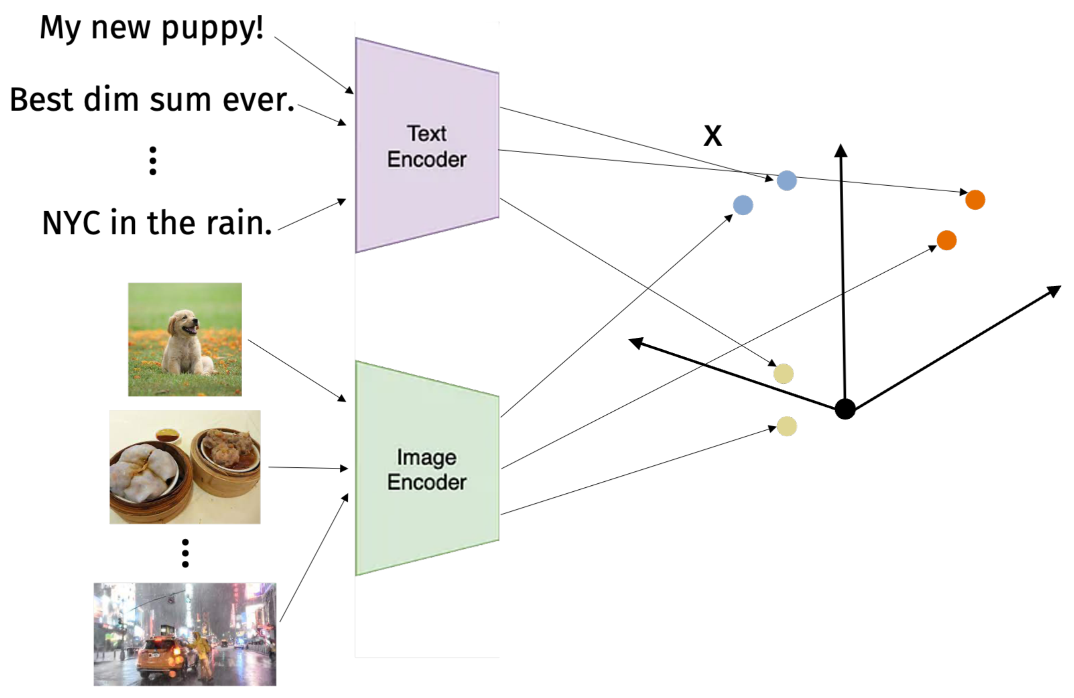
↳ bias $b = J(x,y)$

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

$$O(d) \xrightarrow{\text{compression}} O(k) \quad \text{compute (approx) similarity}$$

$$O(dn) \rightarrow O(k \cdot n) \quad \text{naive search}$$

How do we find similar points faster?



Locality Sensitive Hashing

↳ Hash function $h : \{0,1\}^d \to \{1, \dots, m\}$

↳ similarity function c e.g. Jaccard

↳ h is *locally sensitive* if

$$\Pr(h(x) = h(y)) = \begin{cases} \text{large} & \text{when } x \approx y \\ \text{small} & \text{when } x \not\approx y \end{cases}$$

Our approach:

$c : \{0,1\}^d \to [0,1]$ single MinHash

$g : \mathbb{R} \to \{1, \dots, m\}$

$h(x) = g(c(x))$

$h(x) = h(y)$ when

(1) $c(x) = c(y)$ or

(2) $c(x), c(y)$ happen to hash to the same cell

$$\Pr(h(x) = h(y))$$

$$= \Pr(c(x) = c(y)) \cdot 1$$

$$+ (1 - \Pr(c(x) = c(y))) \cdot \frac{1}{m}$$

negligible

$$\approx \Pr(c(x) = c(y))$$

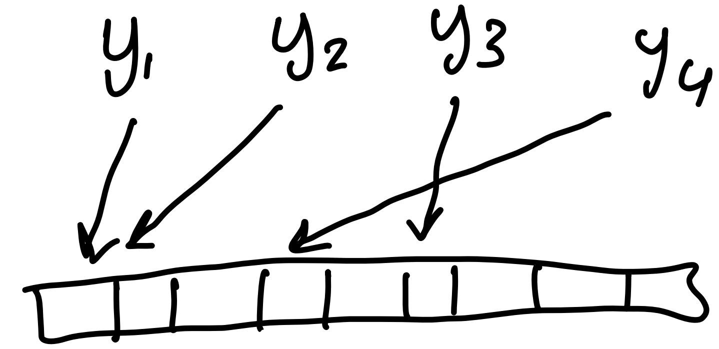
$$= J(x, y)$$

Preprocessing

Choose h by choosing g, c

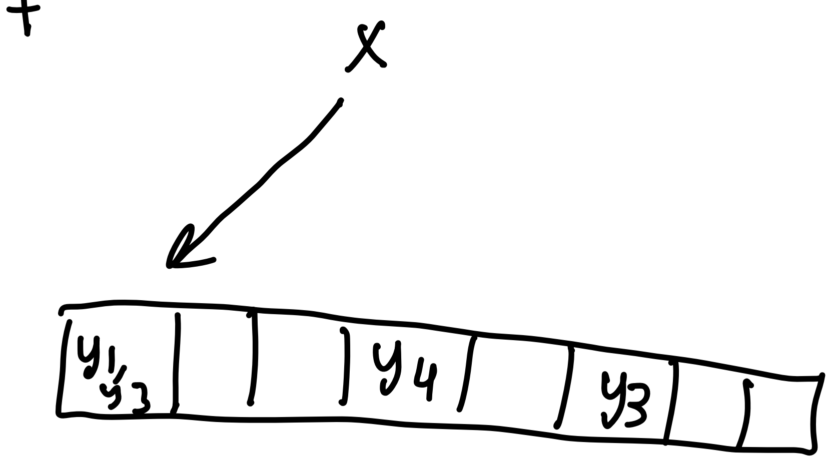
Create a table with m slots

For each vector, we compute $h(y)$ and store in corresponding slot



Query

Compute $h(x)$ and look in corresponding slot



Repeat with t tables

Two questions:

↳ False negative: what's the probability we <u>don't</u> find similar vector?

↳ False positive: what's the probability we do find a non-similar vector?

$$Pr(\text{find } y) = 1 - Pr(y \text{ not in slot of table})^t$$
$$= 1 - Pr(h_i(x) \neq h_i(y))^t = 1 - (1 - J(x,y))^t$$

When $J(x,y) = .4$, $t = 10$, $Pr(\text{find } y) = 1 - (1 - .4)^{10} \approx .99$:)

When $J(x,y) = .2$, $t = 10$, $Pr(\text{find } y) = 1 - (1 - .2)^{10} \approx .89$:(

Our Approach

$c_1, \ldots, c_r : \{0,1\}^d \to [0,1]$

$g : [0,1]^r \to \{1, \ldots, m\}$

$h(x) = g(c_1(x), \ldots, c_r(x))$

negligible

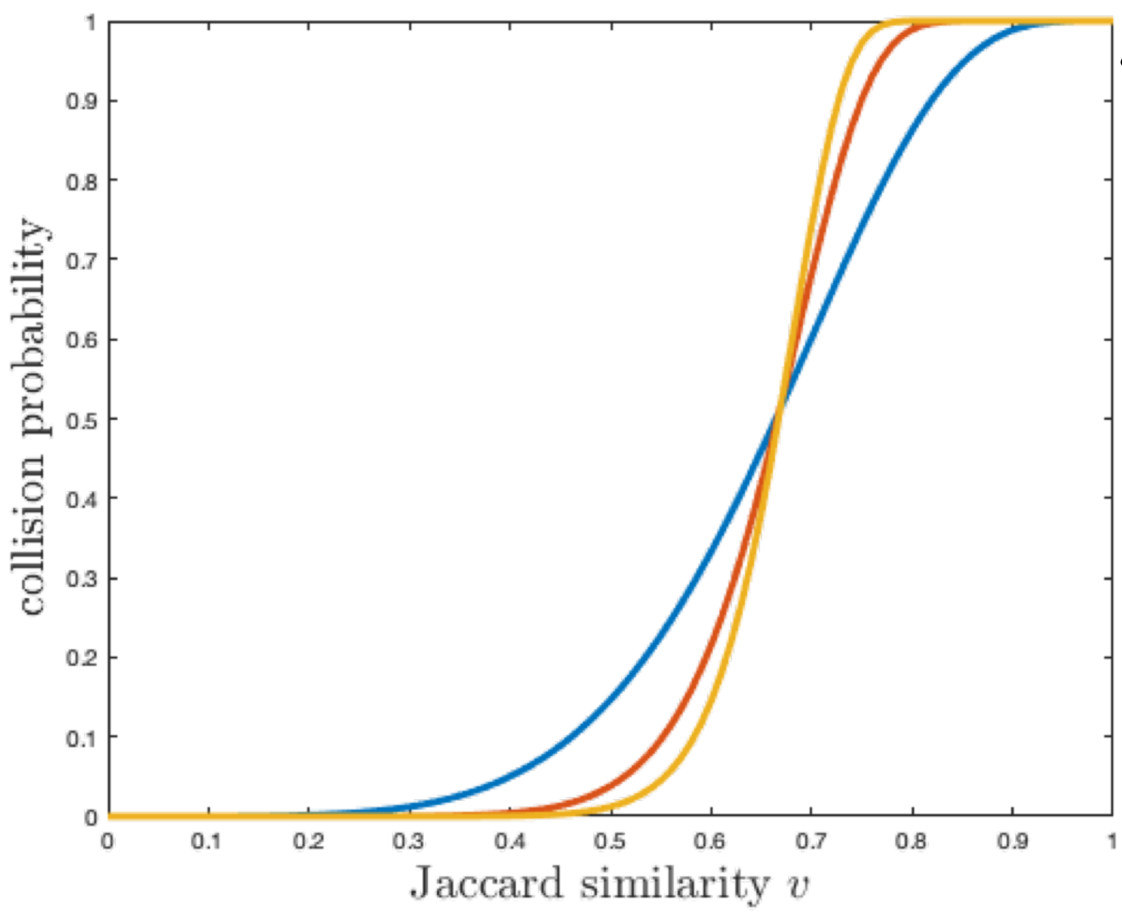
$Pr(h(x) = h(y)) = Pr(c_i(x) = c_i(y) \ \forall i) + \frac{1}{m}$

$\overset{\text{inden}}{=} Pr(c_1(x) = c_1(y)) \, Pr(c_2(x) = c_2(y)) \cdots Pr(c_r(x) = c_r(y))$

$= J(x,y)^r$

$Pr(\text{find } y) = 1 - Pr(\text{not find } y \text{ in table})^t$

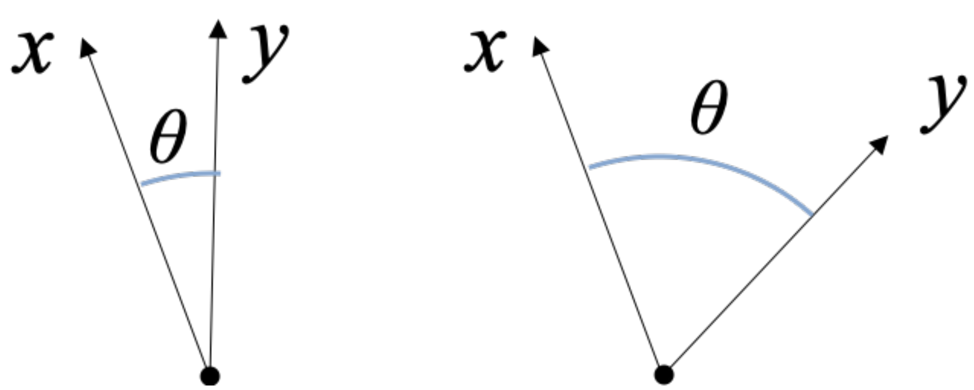
$= 1 - (1 - J(x,y)^r)^t$



control with r, t

Cosine similarity

$$\cos(\theta(x,y)) = \frac{\langle x,y\rangle}{\|x\|_2\|y\|_2}$$



"inverse to distance"

$$\|x-y\|_2^2 = \|x\|_2^2 - 2\langle x,y\rangle + \|y\|_2^2$$

$$= \|x\|_2^2 + \|y\|_2^2 - 2\|x\|_2\|y\|_2 \cos\theta(x,y)$$

Let $g_1, \dots, g_r \in \mathbb{R}^d$ random vectors with $\mathcal{N}(0,1)$

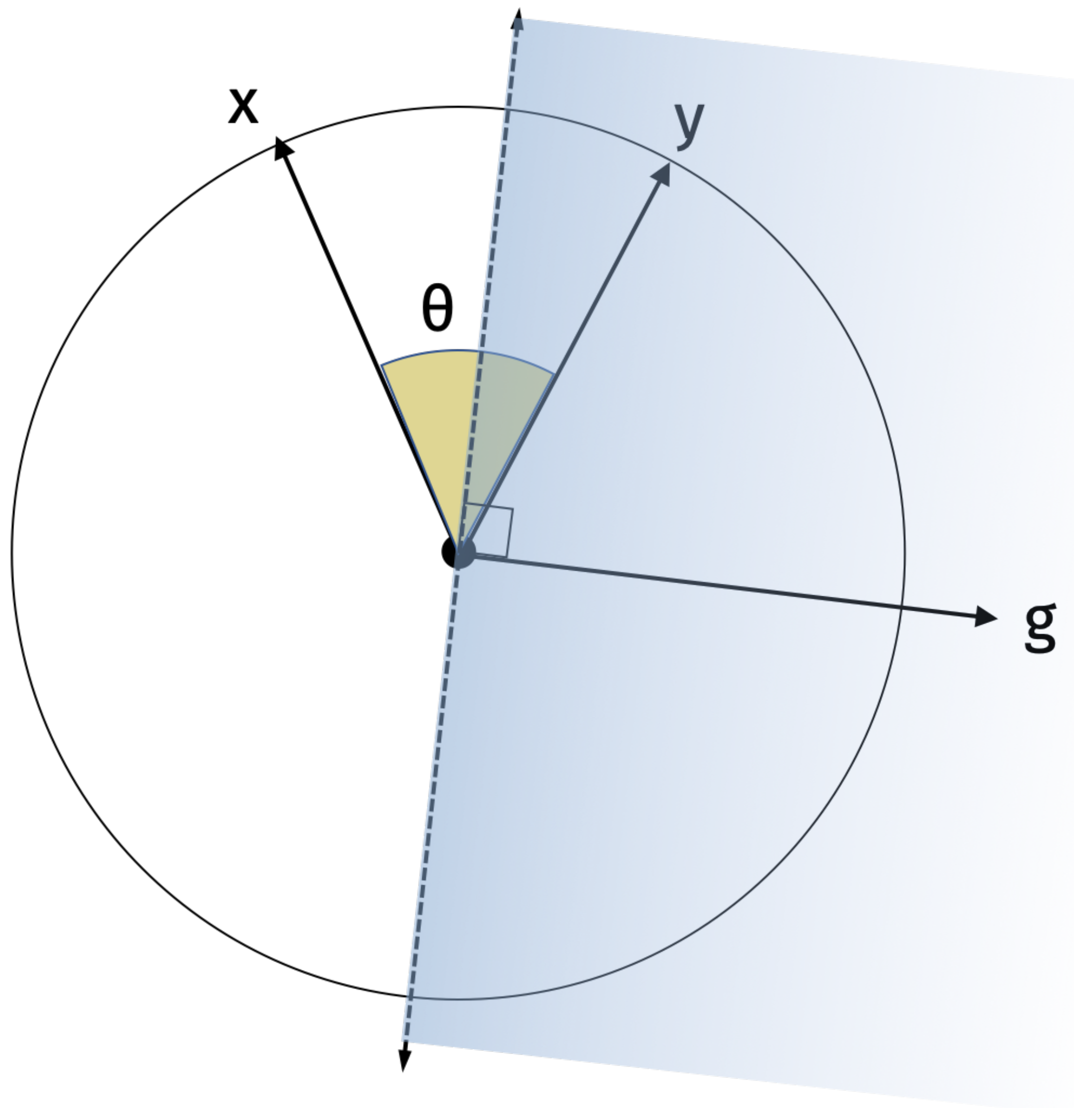
$$f : \{-1,1\}^r \to \{1, \dots, m\}$$

$$h : \mathbb{R}^d \to \{1, \dots, m\}$$

$$h(x) = f(\text{sign}(\langle g_1, x\rangle), \dots, \text{sign}(\langle g_r, x\rangle))$$

$$\Pr(\text{sign}(\langle g_i, x\rangle) = \text{sign}(\langle g_i, y\rangle))$$

$$=$$



$\mathrm{sign}(\langle g, x\rangle)$ = which side of hyperplane

$\mathrm{sign}(\langle g, y\rangle)$ = which side of hyperplane

$$Pr(\text{different sides}) = \frac{2\theta}{2\pi} = \frac{\theta}{\pi}$$

$$Pr(\text{same}) = 1 - \frac{\theta}{\pi}$$

$$= Pr[\mathrm{sign}(\langle g, x\rangle) = \mathrm{sign}(\langle g, y\rangle))$$

$$Pr(\text{find } y) = 1 - Pr(y \text{ not in table}]^t = 1 - Pr(h_i(x) \neq h_i(y))^t$$

$$= 1 - (1 - Pr(h(x) = h(y))^t = 1 - (1 - Pr(\text{same})^r)^t = 1 - (1 - (1 - \frac{\theta}{\pi})^r)^t$$