# L7: Locality Sensitive Hashing & Distribution Distances

Jeff M. Phillips

September 10, 2025

Min Hashing $\quad h \sim \mathcal{H} \quad S, S'$ sets

$$\Pr_{h \sim \mathcal{H}}[h(S) = h(S')] = JS(S, S')$$

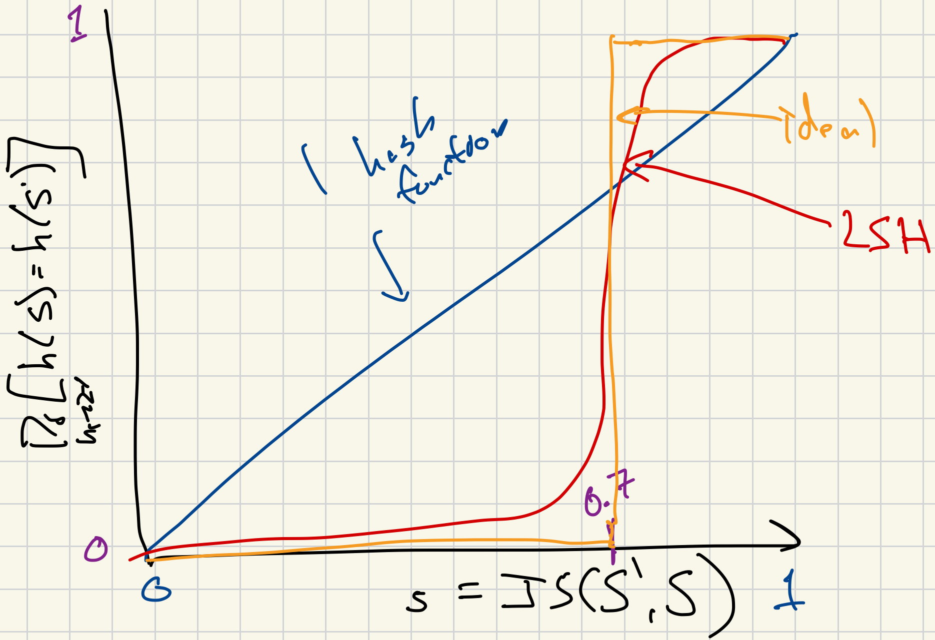
$$\mathbb{E}_{h \sim \mathcal{H}}\left[\mathbb{1}(h(S) = h(S'))\right] =$$

$$h_1, h_2, \ldots h_t \underset{iid}{\sim} \mathcal{H}$$

$$\begin{cases} 1 & \text{if True} \\ 0 & \text{if False} \end{cases}$$

$$\widehat{JS}(S, S') = \frac{1}{t} \sum_{j=1}^{t} \mathbb{1}\left(h_j(S) = h_j(S')\right)$$

$$\mathbb{E}_{h_1 \ldots h_t \sim \mathcal{H}}\left[\widehat{JS}(S, S')\right] = JS(S, S')$$

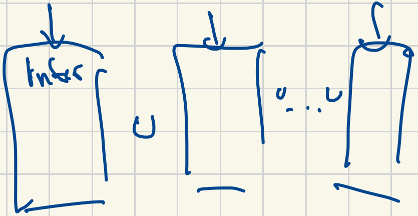


$\Pr_{h \sim H}[h(s) = h(s')]$
1
0
0
1
$s = JS(S', S)$
0.7
1 hash function
ideal
LSH

## Aggresive (few false negatives)

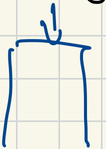
Put guess $g \to h_1(g), h_2(g) \, .. \, h_t(g)$

Union of all intersections!




## Conservative (few false positives)

Concatenate $[h_1(g) h_2(g) \, .. \, h_t(g)] \to$ his hash table



# Banding

$$\left.\begin{array}{l} h_1 \\ \vdots \\ h_b \end{array}\right\} \quad \left.\begin{array}{l} h_1 \\ h_2 \end{array}\right\} \quad H_1$$

$$\left.\begin{array}{l} h_{b+1} \\ \vdots \\ h_{2b} \end{array}\right\} \quad \left.\begin{array}{l} h_3 \\ \vdots \\ h_4 \end{array}\right\} \quad H_2$$

$$\vdots$$

$$\left.\begin{array}{l} h_{t-b} \\ h_t \end{array}\right\} \quad \left.\begin{array}{l} h_{t-1} \\ h_t \end{array}\right\} \quad H_{t/b = r}$$

$$t = b \cdot r$$

\# new hashes

size of band

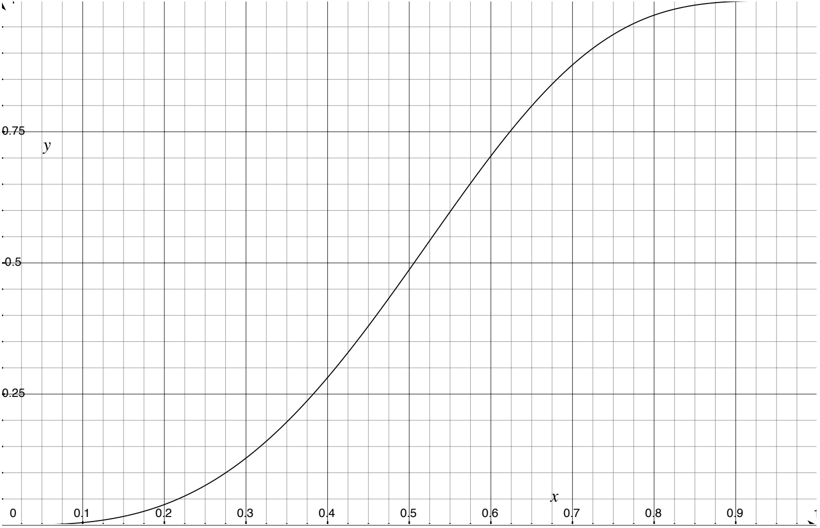
\# of bands

Return Union of collisions on $H_1, H_2, \dots H_{t/b}$

# LSH $b = 3$ and $r = 5$

$t = b \cdot r = 3 \cdot 5 = 15$

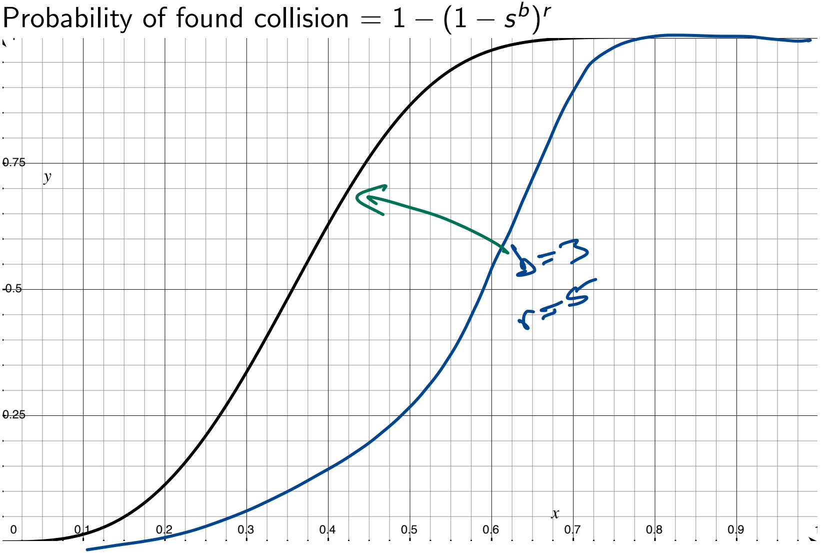
Probability of found collision $= 1 - (1 - s^b)^r$



# LSH $b = 3$ and $r = 15$

Probability of found collision $= 1 - (1 - s^b)^r$

# LSH $b = 3$ and $r = 15$



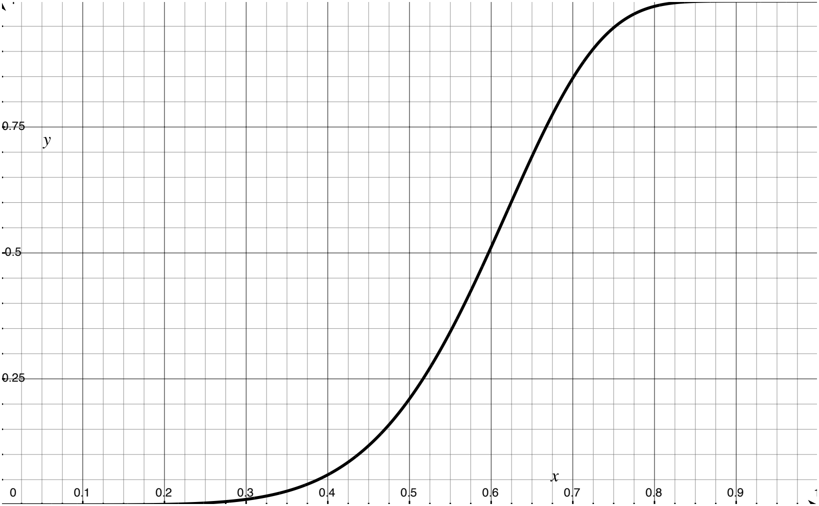
# LSH $b = 6$ and $r = 15$

Probability of found collision $= 1 - (1 - s^b)^r$

# LSH $b = 6$ and $r = 15$

$t = 6 \cdot 15 = 90$

Probability of found collision $= 1 - (1 - s^b)^r$

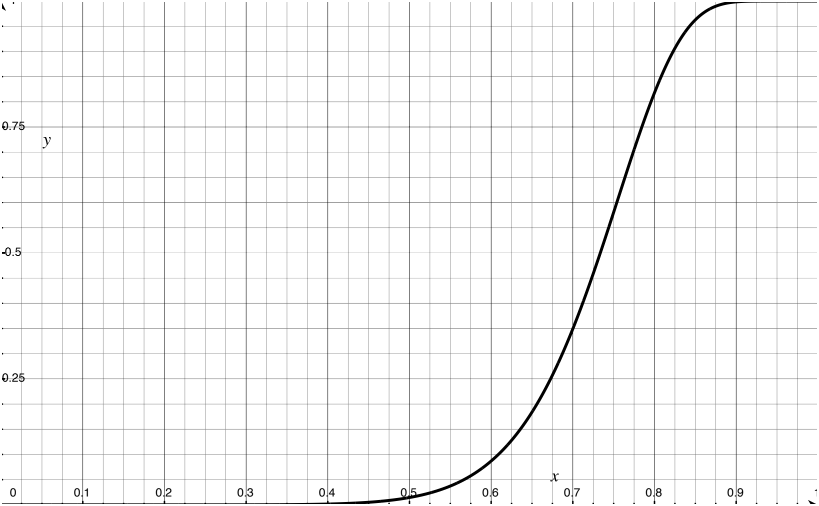


# LSH $b = 10$ and $r = 15$

Probability of found collision $= 1 - (1 - s^b)^r$

# LSH $b = 10$ and $r = 15$

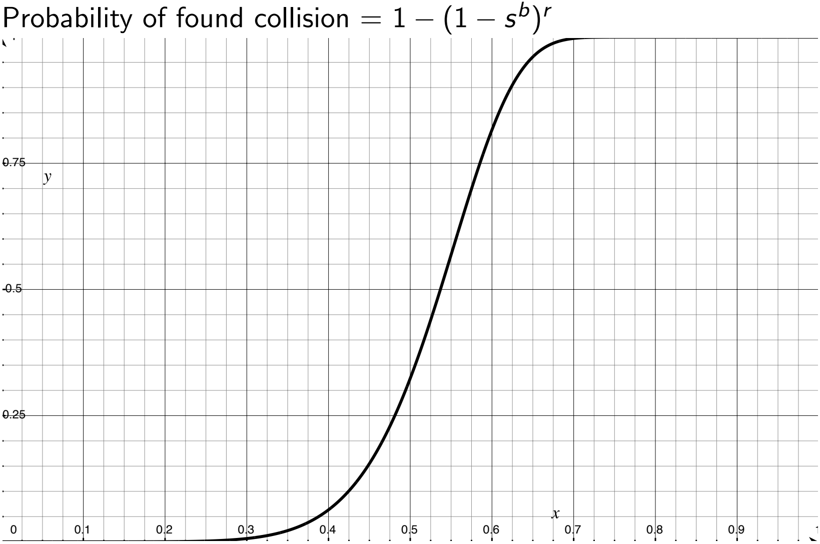
Probability of found collision $= 1 - (1 - s^b)^r$



## LSH $b = 8$ and $r = 100$

Probability of found collision $= 1 - (1 - s^b)^r$

# LSH $b = 8$ and $r = 100$

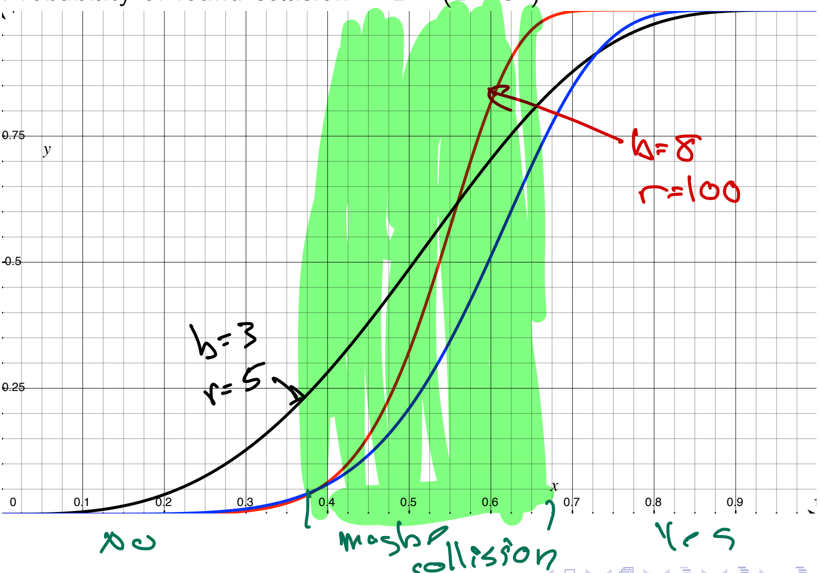


# LSH $(b = 3, r = 5)$ & $(b = 6, r = 15)$ & $(b = 8, r = 100)$

Probability of found collision $= 1 - (1 - s^b)^r$

# LSH $(b = 3, r = 5)$ & $(b = 6, r = 15)$ & $(b = 8, r = 100)$

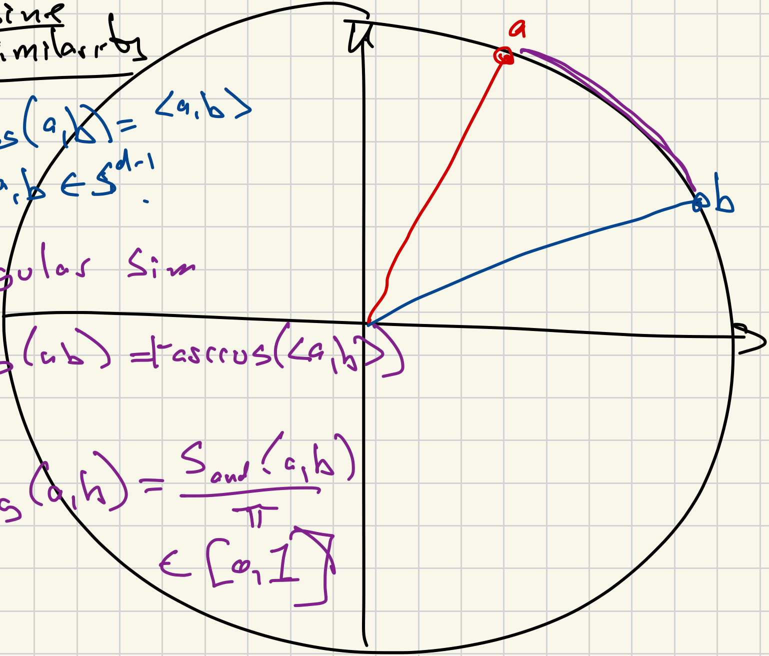
Probability of found collision $= 1 - (1 - s^b)^r$



## Cosine similarity

$$S_{cos}(a,b) = \langle a,b \rangle$$

$$a,b \in S^{d-1}$$




Angular Sim

$$S_{ang}(a,b) = 1 - \arccos(\langle a,b \rangle)$$

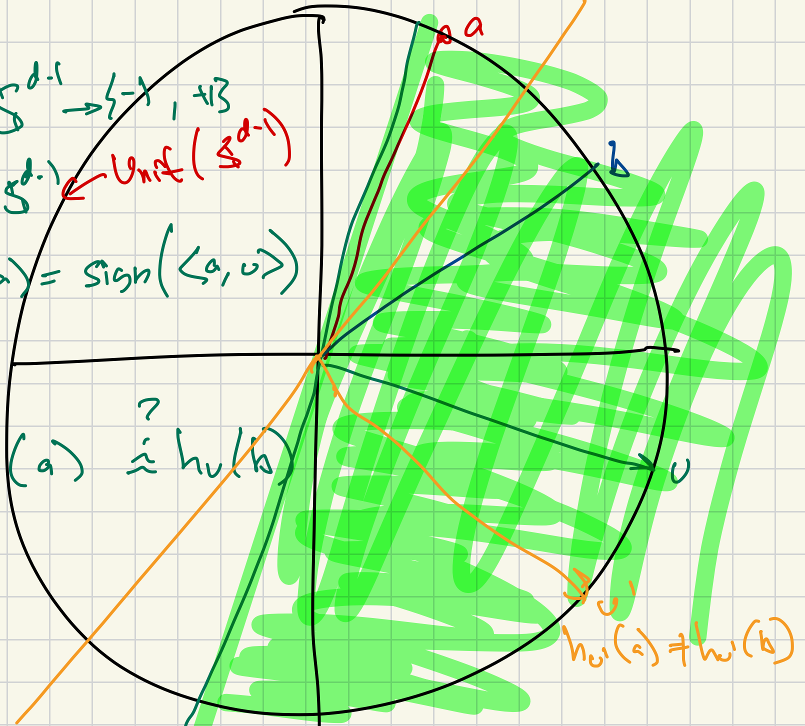
$$\bar{S}_{ang}(a,b) = \frac{S_{ang}(a,b)}{\pi} \in [0,1]$$

$h_u : S^{d-1} \to \{-1, +1\}$

$u \in S^{d-1} \leftarrow \mathrm{Unif}(S^{d-1})$

$h_u(a) = \mathrm{sign}(\langle a, u \rangle)$

$h_u(a) \overset{?}{=} h_u(b)$



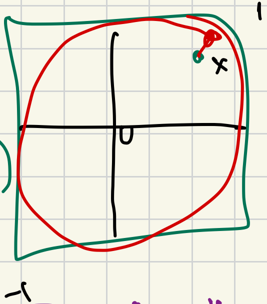

$h_{u'}(a) \neq h_{u'}(b)$

$u \sim \mathrm{Unif}(S^{d-1})$

proposal

$x \in \mathrm{Unif}[-1,1] \times \mathrm{Unif}[-1,1] \times \dots \mathrm{Unif}[-1,1]$

$u = \frac{x}{\|x\|}$



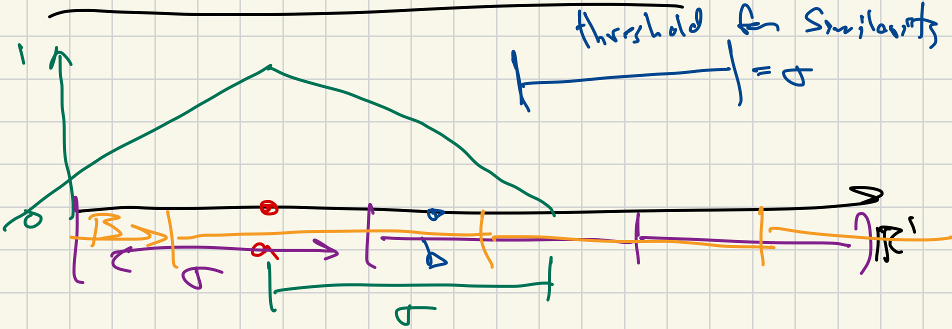

this works

$g \sim N(0,1) \times N(0,1) \times \dots \times N(0,1)$

$u = \frac{g}{\|g\|}$

Box-Mueller Transf

# LSH for Euclidean in $\mathbb{R}^1$

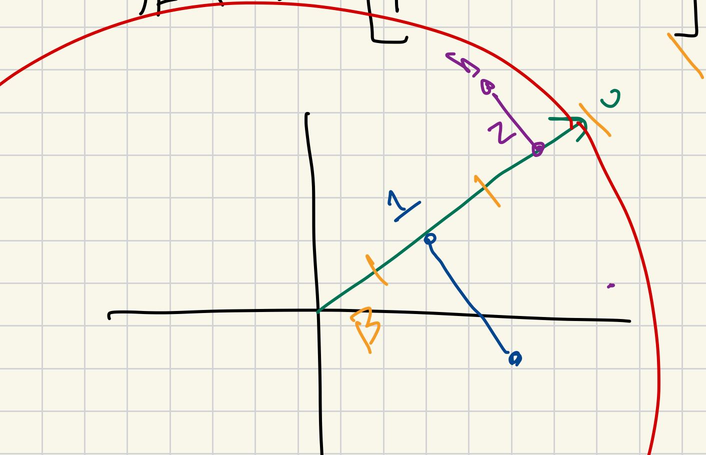



$$\Pr[h(a) = h(b)] = 1 - \frac{\|a-b\|}{\sigma} \quad \text{if } \|a-b\| < \sigma$$

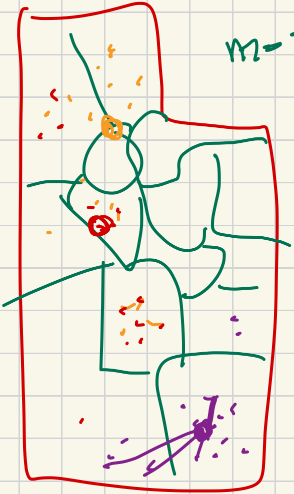
$$= 0 \quad \text{otherwise}$$

## LSH for $\mathbb{R}^d$ Euclidean

$$h_{v,\beta}(a) = \left\lfloor \beta + \langle a, v \rangle \right\rfloor$$



# Distances for Distributions



m = 29 countries

Map to counts counts

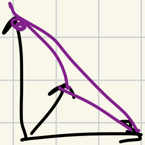
vector $= x \in \mathbb{R}^m$

$$x_j = \frac{\#\text{ strikes in counts } i}{N = \text{tot }\#_{\text{strike}}}$$

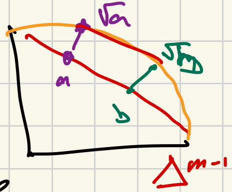
$$x, x', x'' \in \Delta^{m-1} \subset \mathbb{R}^d$$

$$x_j \in [0,1]$$

$$\sum_j x_j = 1$$



$x, x' \in \Delta^{m-1}$



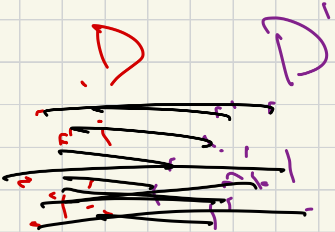

Kullback-Leibler Divergence

$$D_{KL}(a,b) = D(a \| b)$$

$$= \sum_{j=1}^{m} a_j \ln\left(\frac{a_j}{b_j}\right)$$

Hellinger Dist

$$D_H(a,b) = \frac{1}{\sqrt{2}} \sqrt{\sum_{j=1}^{m} \left(\sqrt{a_j} - \sqrt{b_j}\right)^2}$$




Wasserstein Distance

Earth Mover Dist

Find best matching $\gamma : D \to D'$

$$W_1(D, D') = \frac{1}{N} \sum_{x \in D} \| x - \gamma(x) \|$$

Runtime $N^3$