$$JS(A,B) = \Pr\left[v_i(A) = v_i(B)\right]$$

$$JS(A,B) = E\left[\frac{1}{k}\sum_{i=1}^{k} \mathbb{1}\left(v_i(A)=v_i(B)\right)\right]$$

L5: Locality Sensitive Hashing

Jeff M. Phillips

January 23, 2019

Raw Data

$\downarrow$

text $\longrightarrow$ k-grams

Abstract Set

$\downarrow$ min hashing

vector $[m]^k$

$(v_1, v_2, \ldots v_k)$

$n = 1$ million documents $D_1, D_2 \dots D_n$

k-grams: sets $A_1, A_2 \dots A_n$

minhash $v_1, v_2, \dots v_n$

Q1: which pairs of objects are similar?

$JS(A_i, A_j) > T$ (eg: $= 0.85$)

n^2 distance calculations

Q2: Given a query $D_q \to A_q \to v_q$,

which objects are similar?

n distance calculations.

Have n numbers $S=\{s_1, s_2, \dots, s_n\} \subset \mathbb{R}$

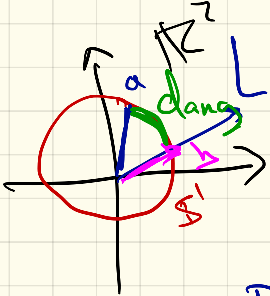
Store S in a data structure

so given query $q \in \mathbb{R}$

return $\arg\min_{s_i \in S} |q-s|$

- Sort in list.
- Build binary tree

Q2: $O(\log n)$ time



Locality Sensitive Hashing

$$h \underset{iid}{\in} \mathcal{H}$$

$d_{ang}(a,b)$ $\Pr[h(a)=h(b)] \approx S(a,b)$

$= \arccos(\langle a,b\rangle)$

Jaccard $\Pr[h(A)=h(B)] = JS(A,B)$

Angular Sim $\Pr[h(A)=h(B)] = S_{ang}(A,B)$

$a,b \in \mathbb{S}^{d-1} = \{a,b \in \mathbb{R}^d \mid \|a\|=1, \|b\|=1\}$

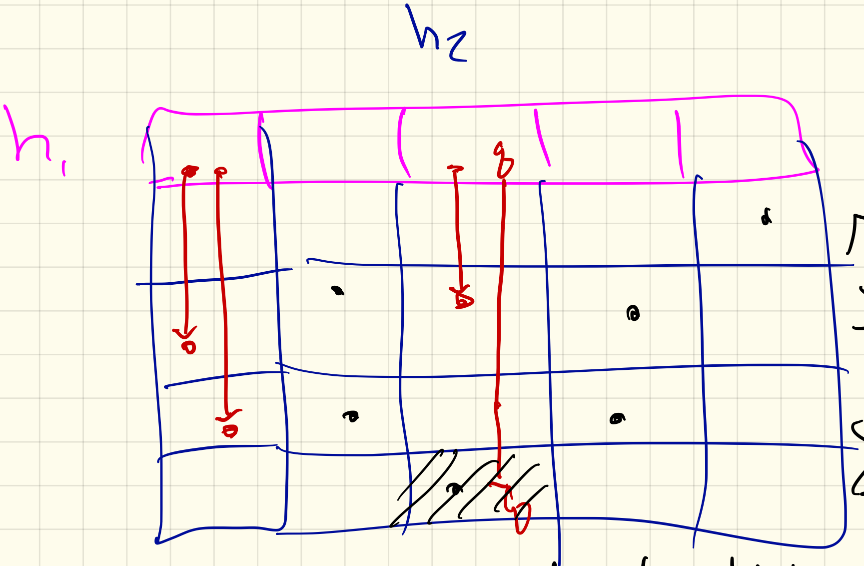
$\bar{a} \leftarrow \frac{v \in \mathbb{R}^d}{\|v\|}$

$S_{ang}(a,b) = 1 - \frac{1}{\pi} \underline{\arccos(\langle a,b\rangle)} \in [0,1]$

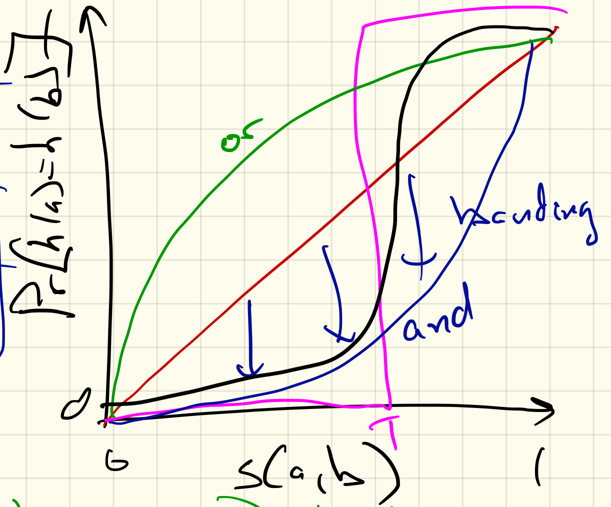
Euclidean $\Pr[h(A)=h(B)] \approx \langle A,B\rangle$

$\hookrightarrow$ dot sim

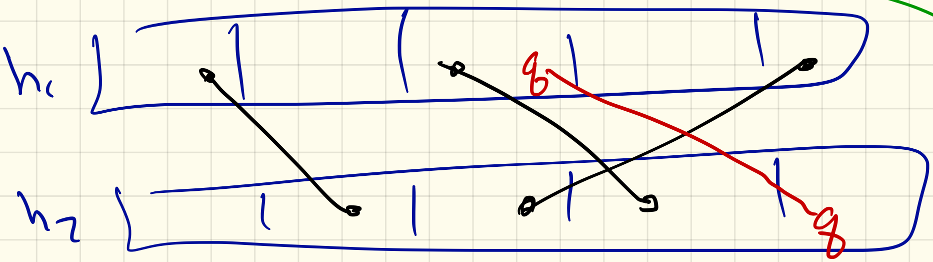
$$\Pr[h(a)=h(b)] = s(a,b)$$



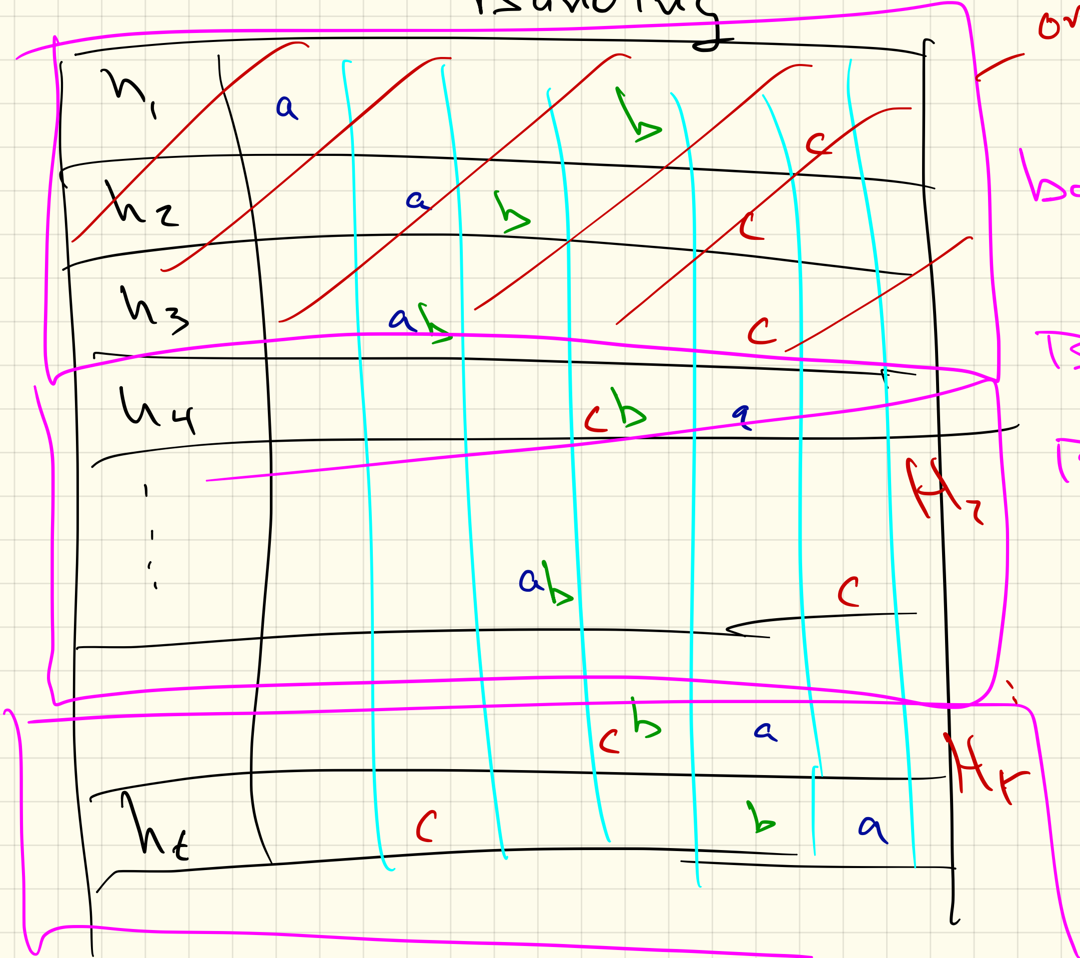
band ≡ one big hash table



papa bear + mama bear ⟹ baby bear



banding



one hash table H_1

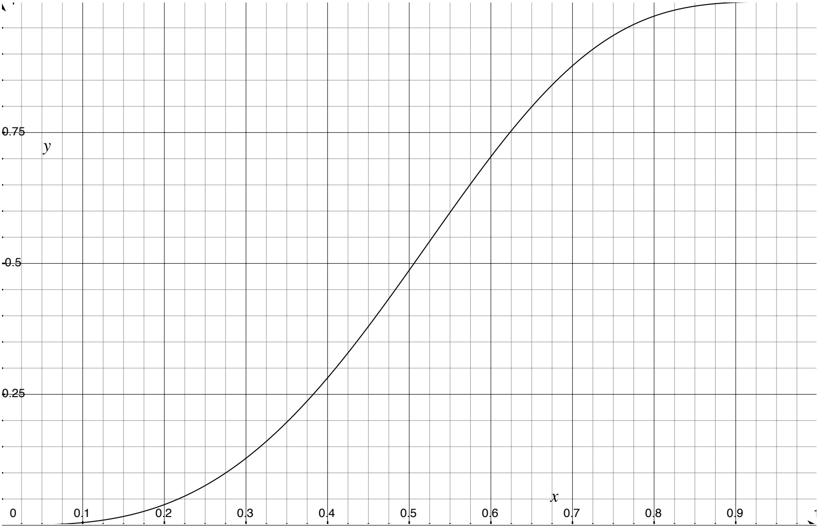
band of b hashes AND

$Band_1$
$Band_2$
⋮
$Band_r$

$Band_1$ or $Band_2$ or ⋮ or $Band_r$

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

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



Banding $\quad b$ hashes in band

r bands (meta Hash fxn)

$s = S(A,B) \leftarrow \Pr[h(a) = h(b)] \quad h \in \mathcal{H}$

s^b = prob all hash fxn in one band collide

$(1-s^b)$ = prob not all hash " "

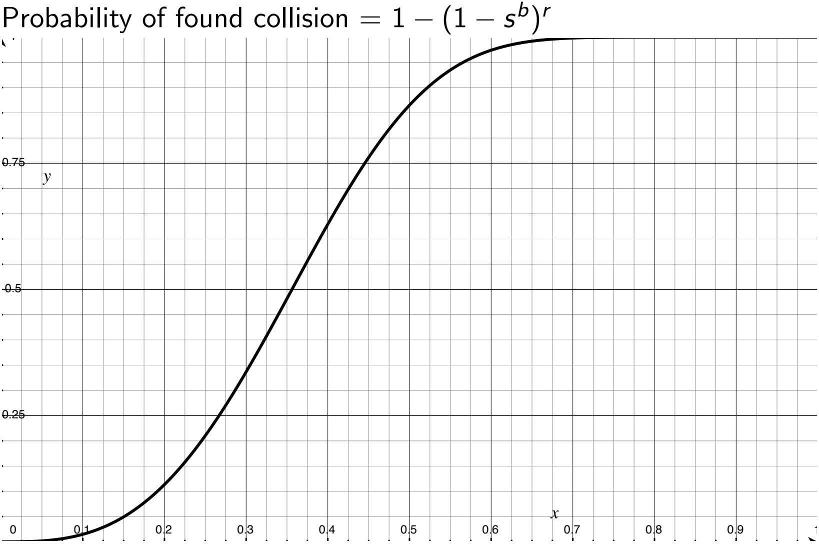
$(1-s^b)^r$ = probability that in no band, do all hash collide

$f(s) = 1-(1-s^b)^r$ = prob of at least one band finds collision.

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

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

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

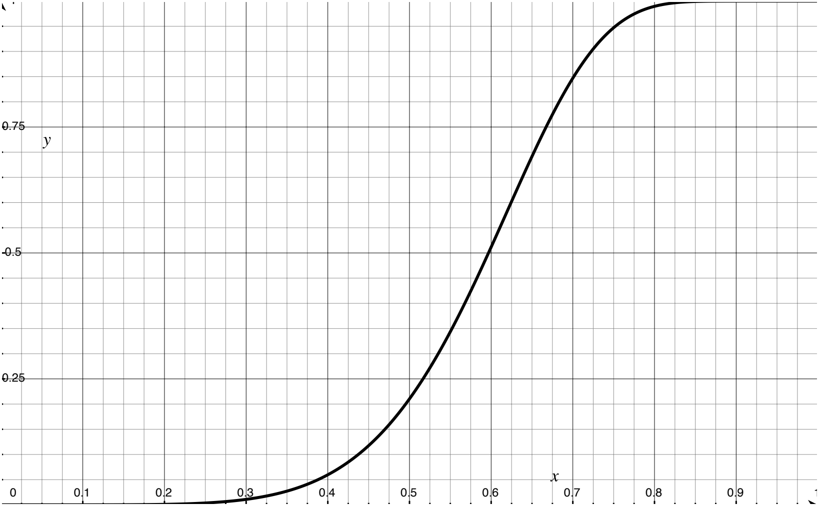


LSH $b = 6$ and $r = 15$

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

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

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

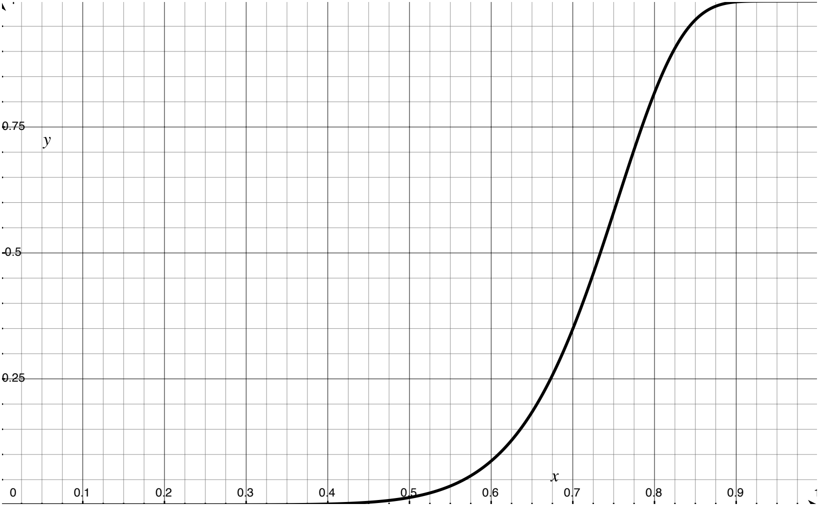


LSH $b = 10$ and $r = 15$

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

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

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



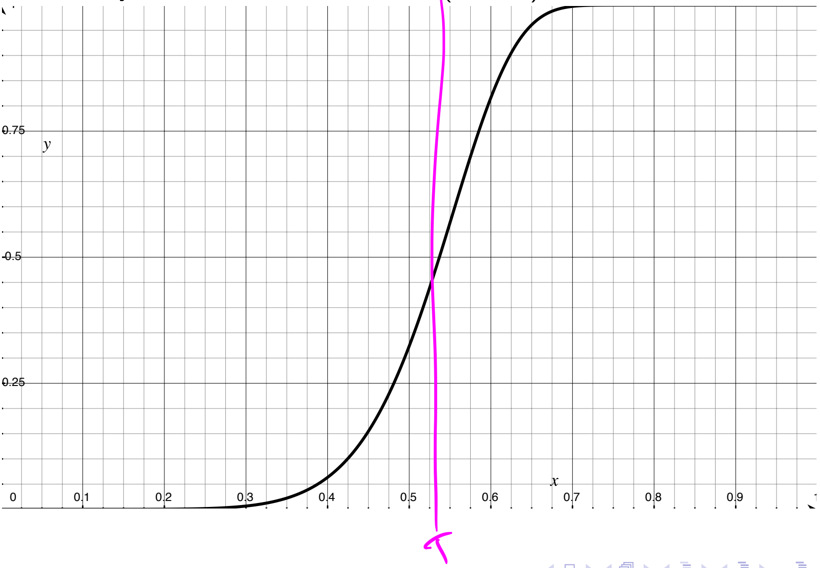
LSH $b = 8$ and $r = 100$

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

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

$t = r \cdot b = 100 \cdot 8 = 800$

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



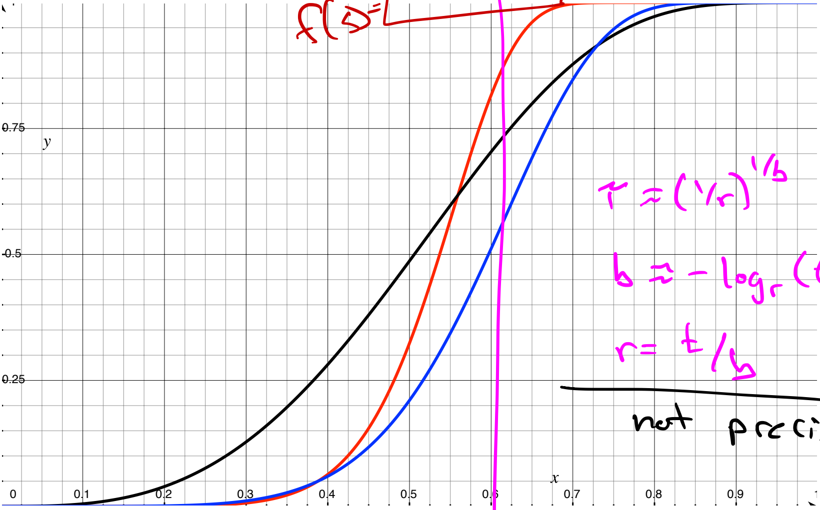
LSH $(b = 3, r = 5)$ & $(b = 6, r = 15)$ & $(b = 8, r = 100)$

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

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

$t = 1000$

Probability of found collision $= \boxed{1 - (1 - s^b)^r} = f(s)$



$f(s)$

$\tau \approx (1/r)^{1/b}$

$b \approx -\log_r(t)$

$r = t/b$

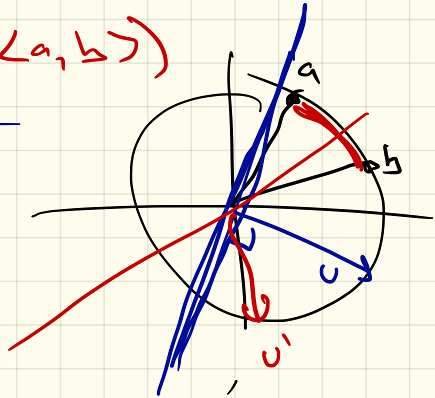
not precise

s

$T = 0.6$

$$S_{ang} : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to [0,1]$$

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



$h_v \in \mathcal{H}_{ang}$

$$\boxed{v \sim \mathrm{Unif}(\mathbb{S}^{d-1})}$$

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

$$h_v(a) = \mathrm{sign}(\langle v, a \rangle)$$

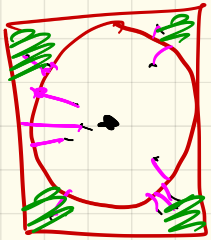
$h_v(a) = +1 \qquad h_v(b) = +1$

$h_{v'}(a) = -1 \qquad h_{v'}(b) = +1$

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

<u>Guess</u> $\quad p \sim \left(\mathrm{Unif}[-1,1]\right)^d$

$v = \frac{p}{\|p\|}$



$\mathbb{R}^d \quad d=2$: rejection sampling

$p \sim \mathrm{Unif}[-1,1]^2$

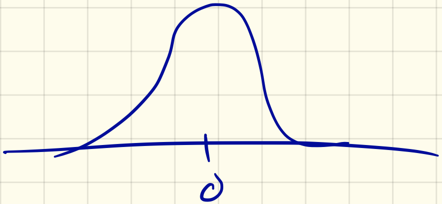
if $\|p\| > 1 \rightarrow$ start over

o.w $\quad v = \frac{p}{\|p\|}$

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

1. $g \sim G_d(x) = \frac{1}{(2\pi)^{d/2}} e^{-\|x\|^2/2}$

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



$g = (g_1, g_2, \ldots g_d)$

$g_i \underset{iid}{\sim} G_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\|x\|^2/2}$

Box-Moeller Transform

$u_1, u_2 \sim \mathrm{Unif}(0,1) \Longrightarrow \begin{array}{l} g_1 \leftarrow \sqrt{-2\ln(u_1)}\,\cos(2\pi u_2) \\ g_2 \leftarrow \sqrt{-2\ln(u_1)}\,\sin(2\pi u_2) \end{array}$