L7 : Locality Sensitive Hashing

+ Distances for Distributions

Data Mining

Sep 15, 2026

Jeff M. Phillips

Min Hashing

$h_a \sim \mathcal{H}$ two sets S, S'

a ← salt

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

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

indicator 1 if true, 0 if false

many hashes $h_1, h_2, \dots h_t \overset{iid}{\sim} \mathcal{H}$

$t = 100, \dots 1000$

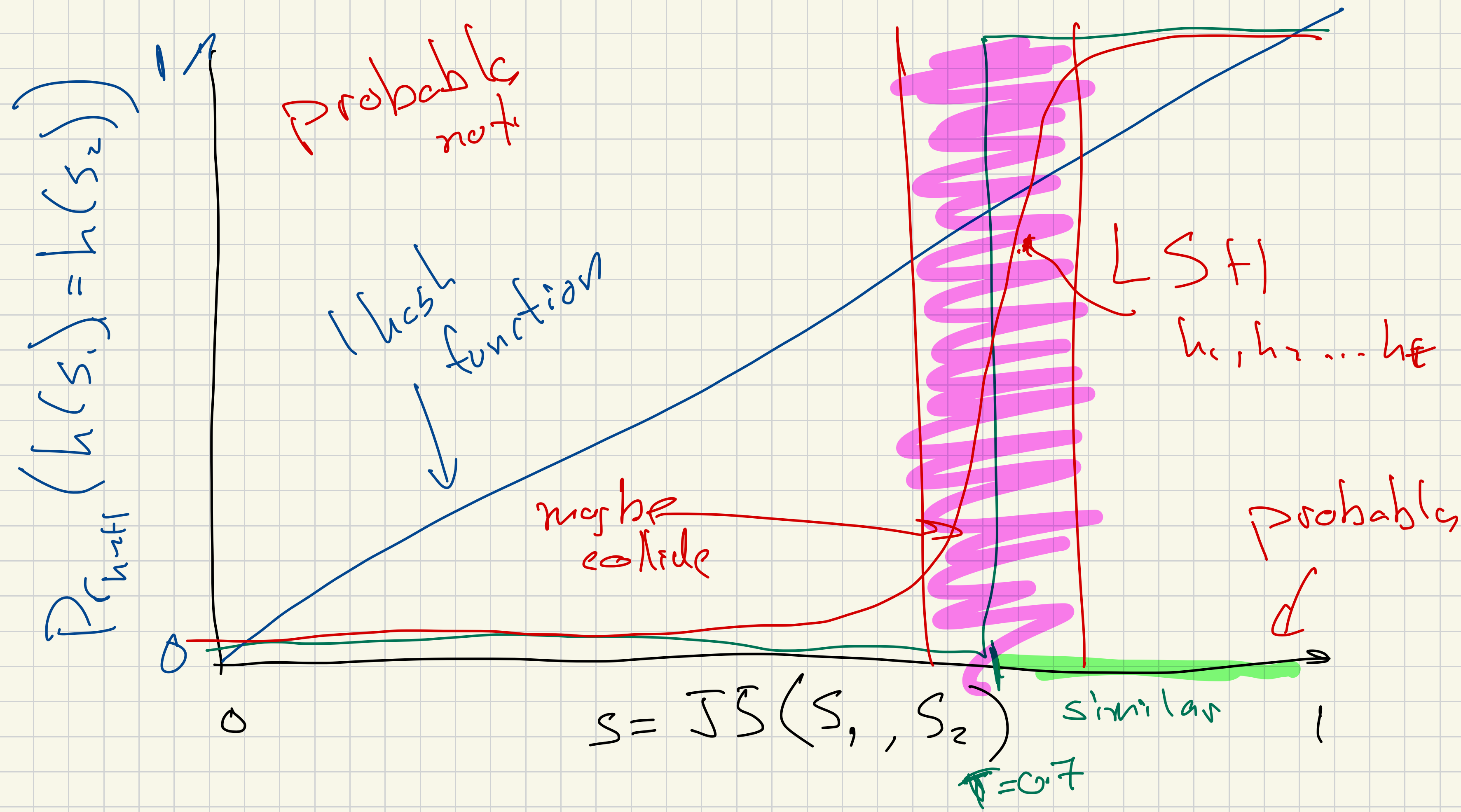
$$\hat{JS}(S, S') = \frac{1}{t} \sum_{i=1}^{t} \mathbb{1}(h_i(S) = h_i(S'))$$

$$\mathbb{E}[\hat{JS}(S, S')] = JS(S, S')$$

CLT

①

② as $t \to \infty$ better estimate.



probable not
Hash function
maybe collide
LSH
h1, h2 ... ht
probable
similar
0
1
s = JS(S1, S2)
t=0.7

① **Aggressive** (few false negatives)

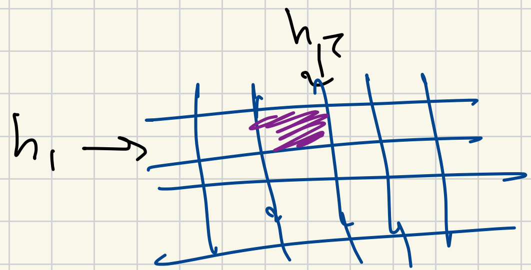
query $q \longrightarrow h_1(q), \; h_2(q), \; \dots \; h_t(q)$

$$\begin{bmatrix} x_1 \\ x_3 \\ x_7 \end{bmatrix} \cup \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} \cup \dots \cup \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} \Rightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_7 \end{bmatrix}$$

② **Conservative** (few false positives)

concatenation $[h_1(q), \; h_2(q), \; \dots, \; h_t(q)] \longrightarrow \mathbb{R}^t$

$$\begin{bmatrix} x_1 \end{bmatrix}$$



Banding

$$t = \underbrace{b}_{\text{size of band}} \cdot r$$

bands

h_1
h_2
$\vdots$
h_b

$$H_1 = \text{concat}(h_1, \ldots h_b)$$

h_{b+1}
$\vdots$
h_{2b}
h_{2b+1}

$$H_2 = \text{concat}(h_{b+1} \ldots h_{2b})$$

$\vdots$

h_t

$$H_r = \text{concat} \ldots h_t)$$

Return union collisions on

$$H_1(q) \cup H_2(q) \ldots \cup H_r(q)$$

LSH b (size of band) and r (number of bands)

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

$Pr(D_1 \equiv D_2)$

$JS(D_1, D_2) = s$

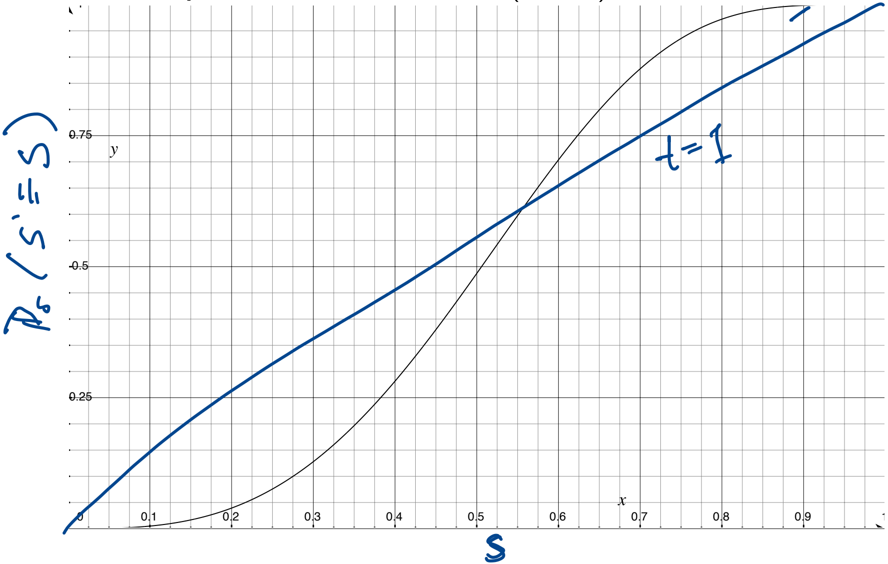
$s = \text{JS}(D_1, D_2)$ is probability $h(D_1) = h(D_2)$:

$$
\begin{aligned}
s^b &= \text{probability all hashes collide in 1 band} \\
(1 - s^b) &= \text{probability not all collide in 1 band} \\
(1 - s^b)^r &= \text{probability that in no band, do all hashes collide} \\
1 - (1 - s^b)^r &= \text{probability all hashes collide in at least 1 band}
\end{aligned}
$$

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

$r \cdot b = t = 15$

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 $r = 15$

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

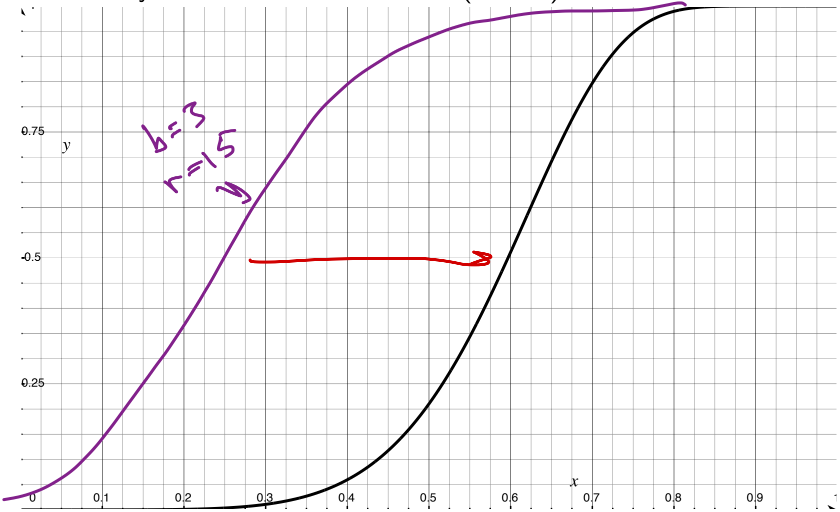
LSH $b = 6$ and $r = 15$

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

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

$t = r \cdot b = 90$

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



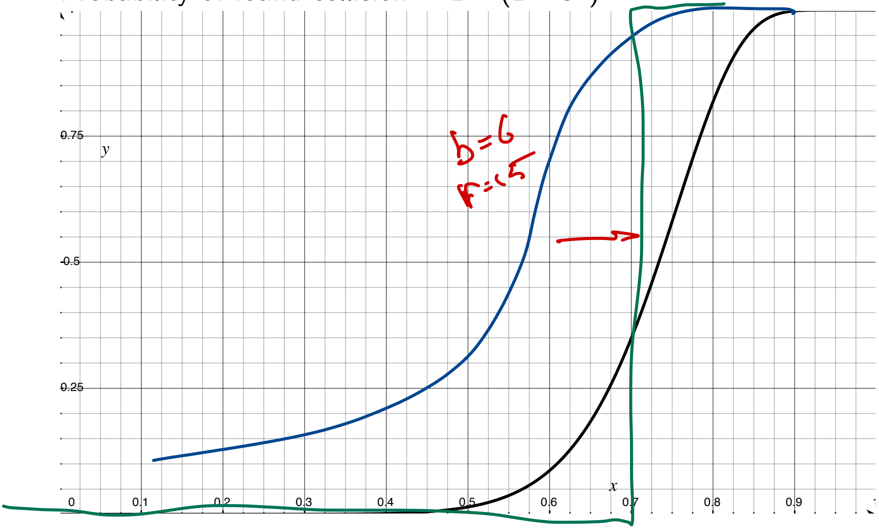
LSH $b = 10$ and $r = 15$

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

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

$t = 150$

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



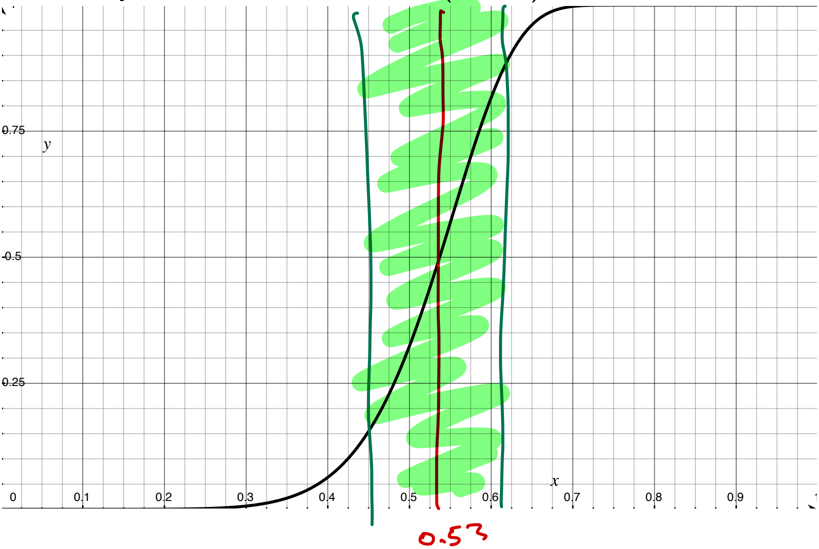
LSH $b = 8$ and $r = 100$

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

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

1-800

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

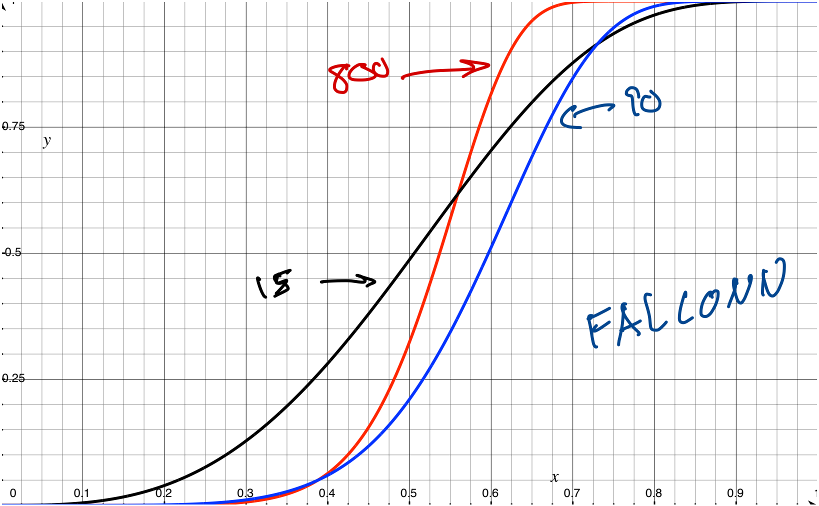


LSH $(b = 3, r = 5)$ & $(b = 6, r = 15)$ & $(b = 8, r = 100)$

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

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

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



Cosine Similarity

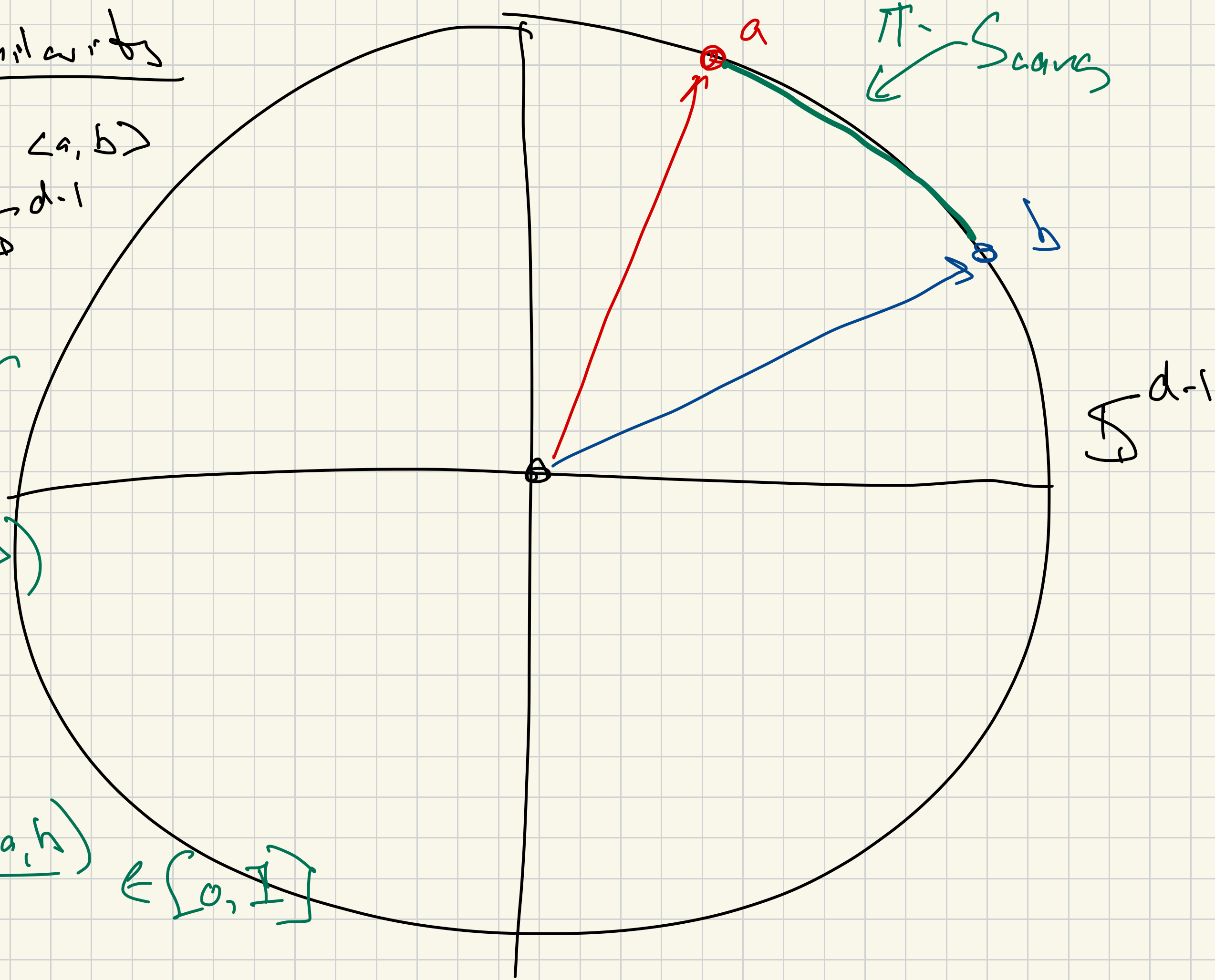
$$S_{cos}(a,b) = \langle a,b \rangle$$

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

Angular Sim

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

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



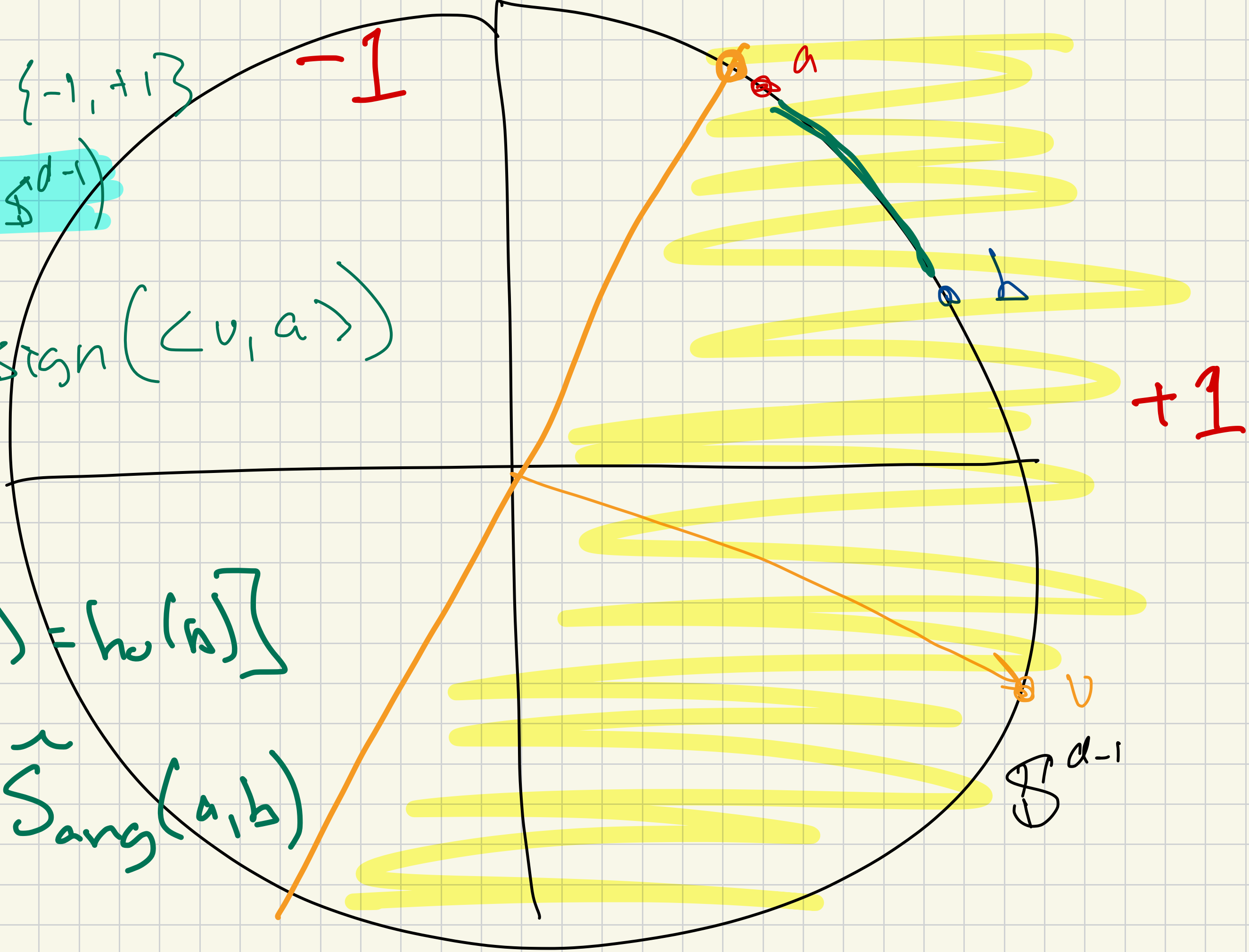
$$h_v : \mathbb{S}^{d-1} \to \{-1, +1\}$$

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

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

$$\Pr_{v \sim \mathbb{S}}[h_v(a) = h_v(b)]$$

$$= \tilde{S}_{ang}(a, b)$$

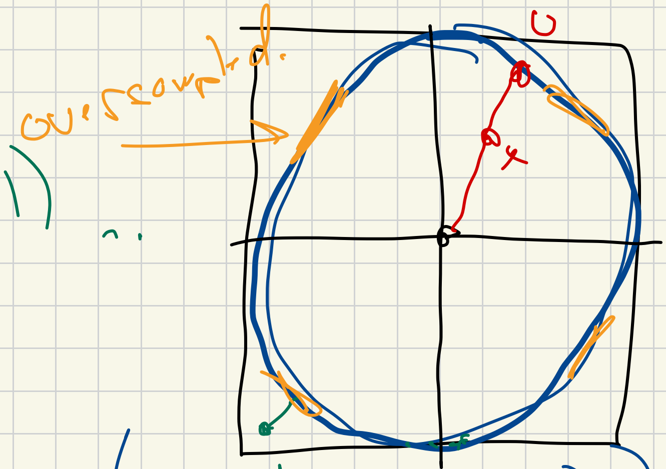


$u \sim \text{Unif}(S^{d-1})$

proposal

$$x \in \left(\overset{1}{\text{Unif}[-1,+1]}, \overset{2}{\text{Unif}[-1,+1]}, \ldots \overset{d}{\text{Unif}[-1,+1]}\right)$$

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



correct

$$g \in \left(\text{Norm}(0,1), \text{Norm}(0,1) \ldots\right.$$

$$g_i \sim N(0,1)$$

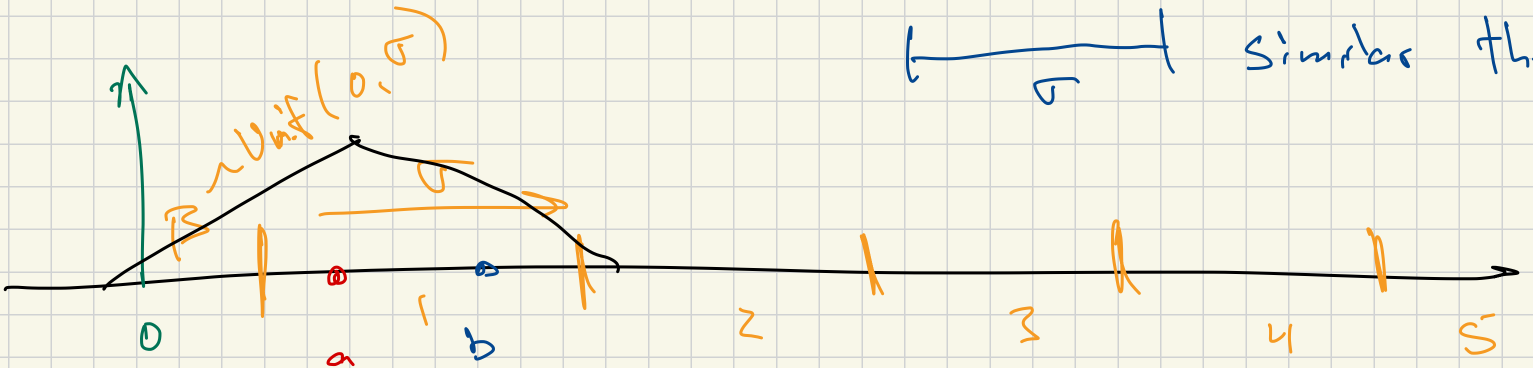
$$u = \frac{g}{\|g\|}$$

BoxMueller $(x_1, x_2 \sim \text{Unif}(0,1))$

$g_1, g_2 \sim N(0,1)$

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

$\vdash\!\!\!-\!\!\!-\!\!\!\dashv$ similar threshold
σ

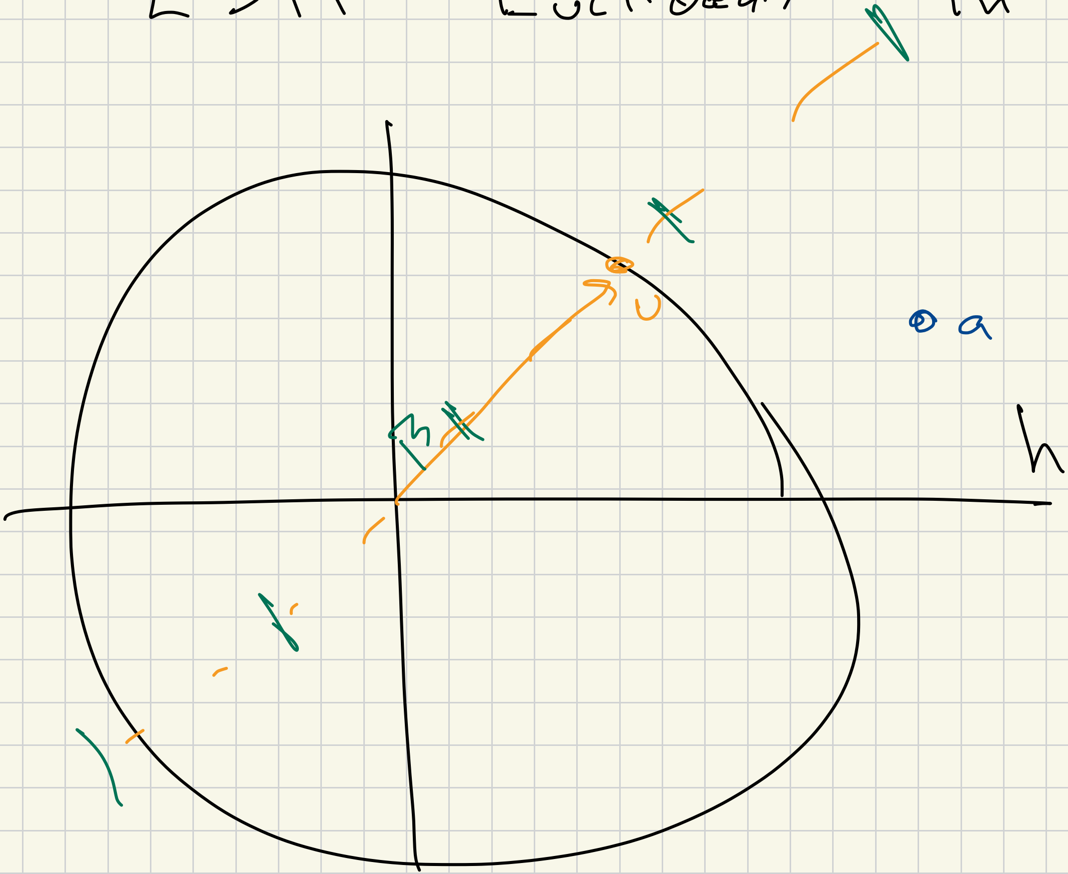


$$h(a) = (a - \beta) \bmod \sigma$$

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

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

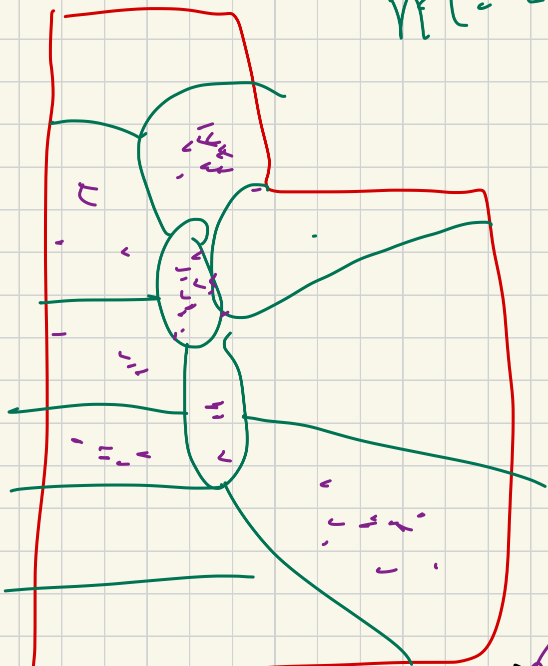
LSH Euclidean in $\mathbb{R}^d$



$$h_{u,\beta}(a) = \lfloor -\beta + \langle a, u \rangle \rfloor$$

Distances for Distribution

$m = 29$ contours



Map counters to coordinates

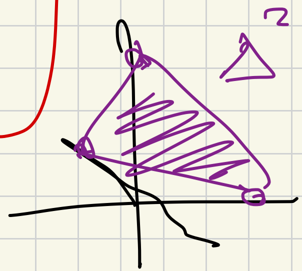
vector $x \in \mathbb{R}^{29 = m}$

$c_1, c_2 \dots c_{29}$ counts

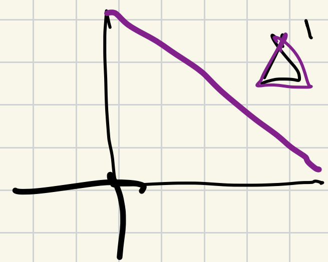
$x_i = \frac{c_i}{N}$ $\quad$ N total number

$x \in \Delta^{m-1} = \Delta^{28}$

$x_i \in [0,1]$ $\quad$ $\sum_i x_i = 1$

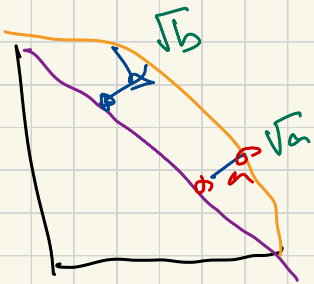


$m = 2$



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

Kullbach-Leibler Divergence

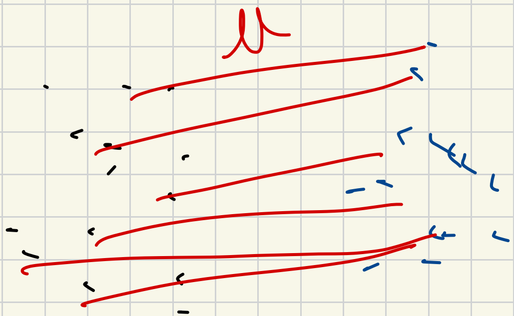


$$D_{KL}(a,b) = D_{KL}(a \| b) = \sum_{i=1}^{m} a_i \ln\left(\frac{a_i}{b_i}\right)$$

Hellinger Dist

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

Data X_1 X_2



$X_1 \subset \mathbb{R}^d$ $X_2 \subset \mathbb{R}^d$

EMD

Wassserstein

W_1

$$\min_{\mu} \frac{1}{n} \sum_{x_i \in X_1} \| x_i - \mu(x_i) \|_1 = W_1(X_1, X_2)$$

$$W_2 = \min_{\mu} \frac{1}{n} \sqrt{\sum_{x_i} \| x_i - \mu(x_i) \|^2}$$