L5: Locality Sensitive Hashing

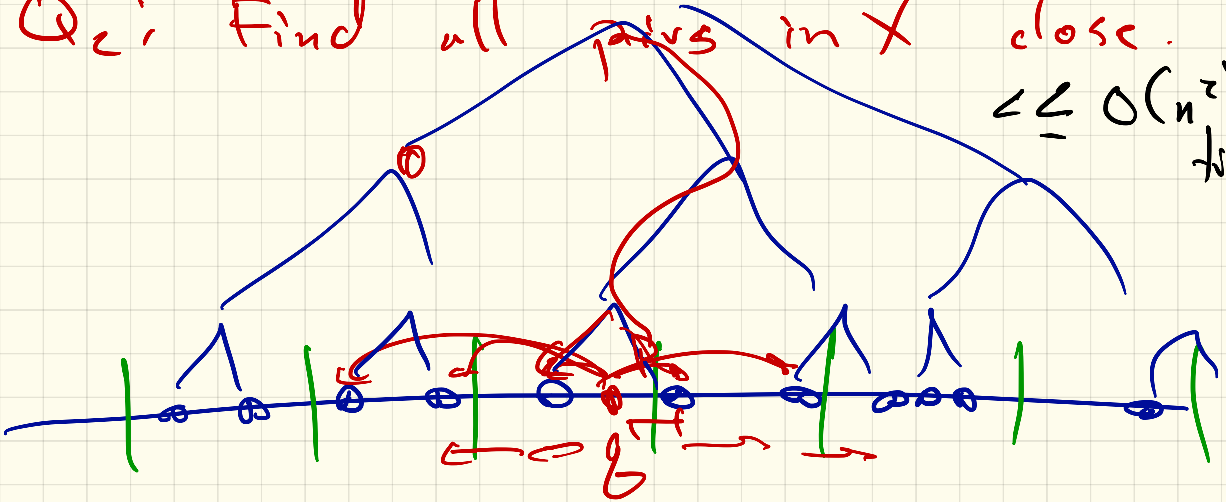
Jeff M. Phillips

January 24, 2018

Set $X = \{P_1, P_2, \ldots P_n\}$

Q_1: Given query doc q
Find all in X close $<\leq O(n)$ time

Q_2: Find all pairs in X close. $<\leq O(n^2)$ time.

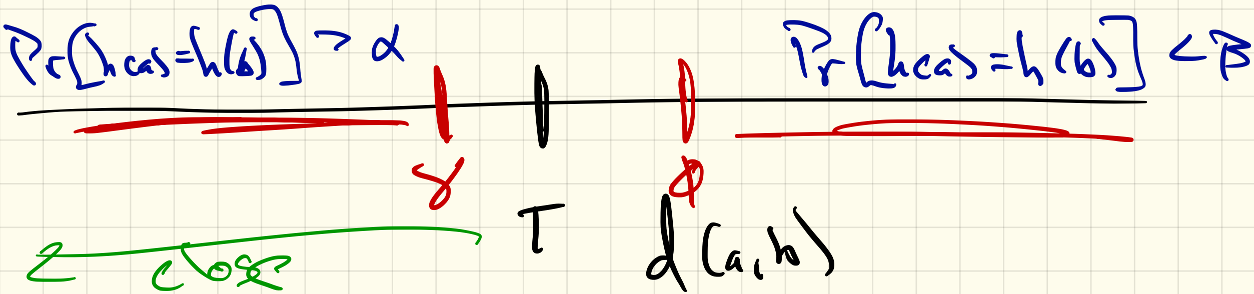


LSH

hash fxns $\{h_1, h_2, \dots h_k\} \in \mathcal{H}$

$h \in \mathcal{H}$ is $(\gamma, \phi, \alpha, \beta)$-sensitive

- $\Pr[h(a)=h(b)] > \alpha$ if $d(a,b) < \gamma$
- $\Pr[h(a)=h(b)] < \beta$ if $d(a,b) > \phi$



Min-Hashing $\Rightarrow (T, T, 1-T, T)$-sensitive

Documents	D_1	D_2	...	D_n
$h_1 \rightarrow m_1$	1	7		1
$h_2 \rightarrow m_2$	2	2		8
		V_2		

Choose threshold $T = \phi_T = \gamma$

$\Pr[h(a) = h(b)] > \alpha = 1-T$

T

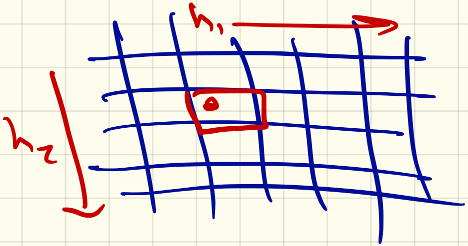
$\Pr[h(a) = h(b)] < \beta = T$

$\leftarrow S(a,b)$

$\boxed{d(a,b) \rightarrow}$

JS threshold $= 1-T$

1 hash fxn



for query doc q

Return all $D_i \in X$ s.t. $h(q) = h(D_i)$

Papa

- Apply b hash fxns, and return $D_i \in X$ s.t. all collide w/ $h(q)$

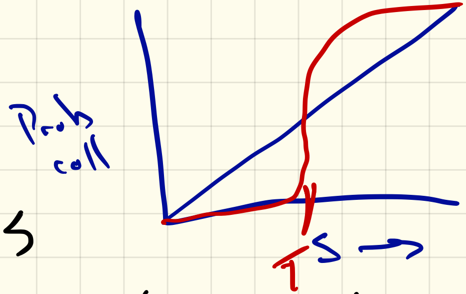
  ↳ super hash table $\vec{h} = h_1 \times h_2 \times \dots \times h_b$

Mama

- Apply r hash fxns → r hash tables
  return union of collisions

Banding

Use $b \cdot r$ hash fxns



r superhash tables, each w/ b hash fxns

$s = JS(a,b)$

$\boxed{\text{Prob return } b \text{ on query } a}$

s^b = Prob collide | superh hash

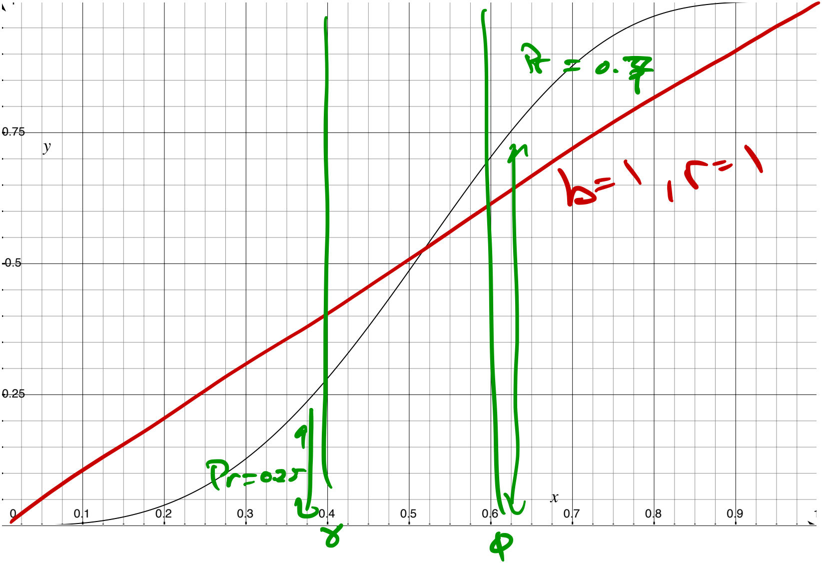
$(1-s^b)$ = Prob don't collide, | s.h.

$(1-s^b)^r$ = Prob no s.h.t. collides

$f(s) = 1-(1-s^b)^r$ = Prob at least 1 s.h.t. collides

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

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



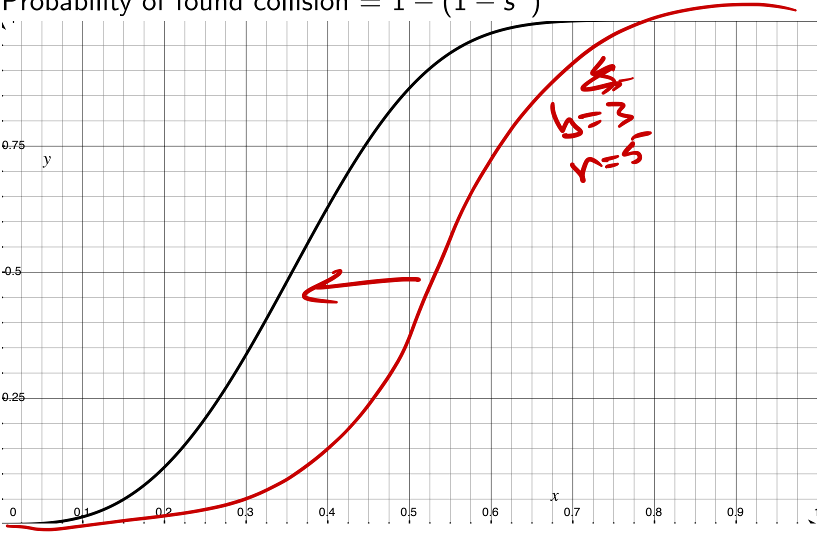
LSH $b = 3$ and $r = 15$

increase # s.h.f.

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

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

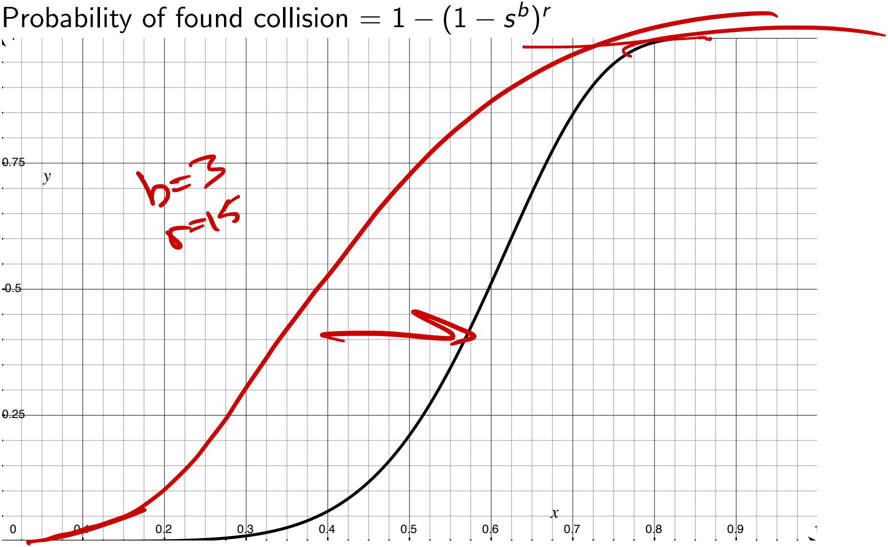
Probability of found collision $= 1 - (1 - s^b)^r$



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

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

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

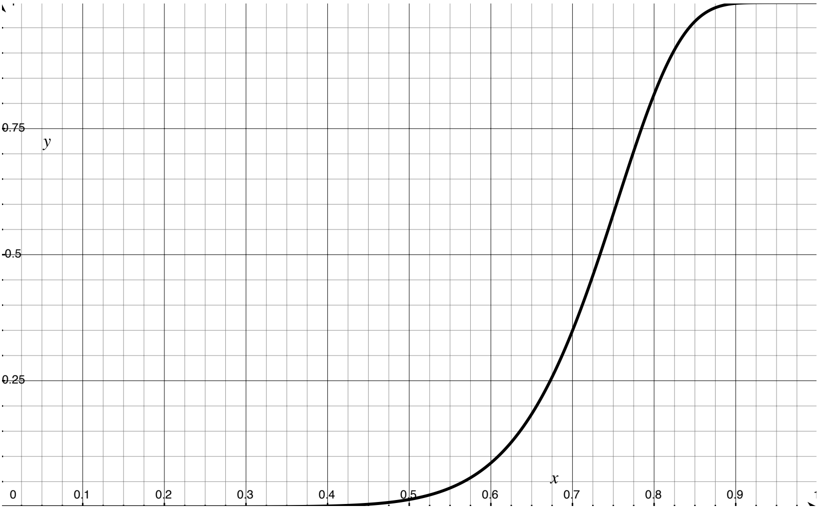


LSH $b = 10$ and $r = 15$

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 $r = 100$

$b, r \rightarrow$ steeper

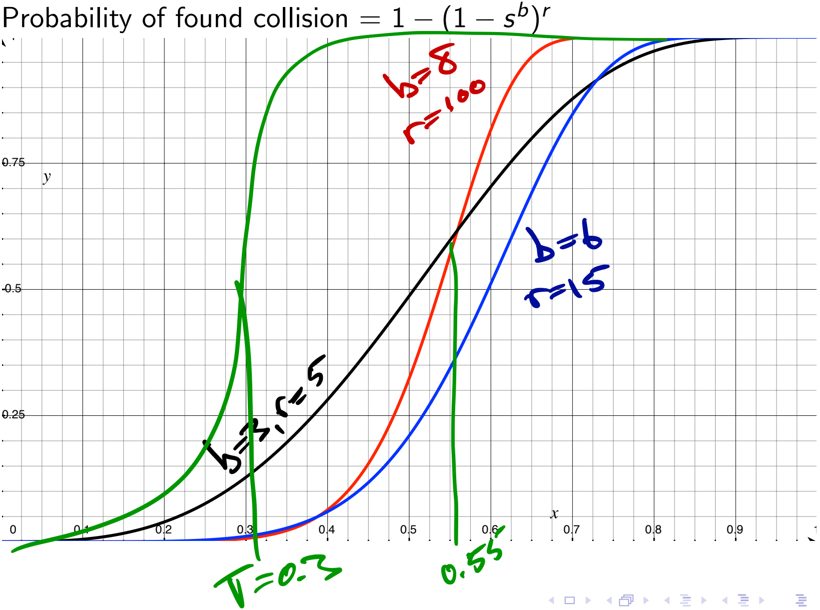
Probability of found collision $= 1 - (1 - s^b)^r$

$\tau = 0.52$

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

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

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



Choosing r, b so curve is steepest at T
treshold

$t = r \cdot b$

steepest $T \lessapprox (1/r)^{(1/b)}$

$b = -\log_r(t)$

"rule of thumb"

$r = t/b$

$\to$ make integers

$\hookrightarrow$ then experiment

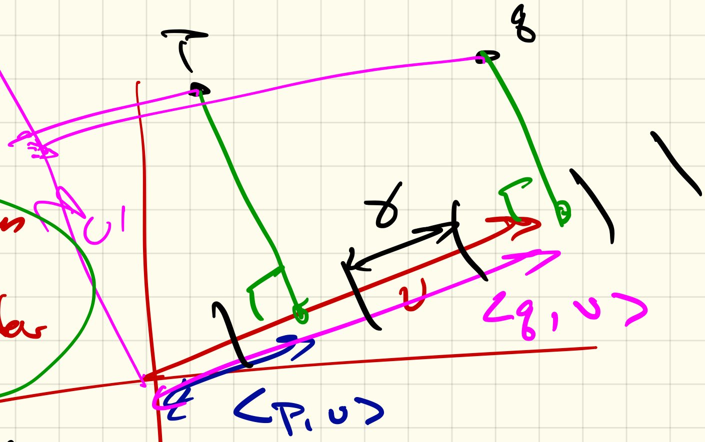
$(\gamma/2, 2\gamma, 1/2, 1/3)$-sensitive

LSH for Euclidean Dist.

$p = (p_1 \dots p_d)$

$q \in (q_1 \dots q_d)$

$$\|p-q\| = \sqrt{\sum_{i=1}^{d} (p_i - q_i)^2}$$



hash fxn

1. project random direction

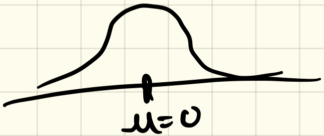
   u unit vector

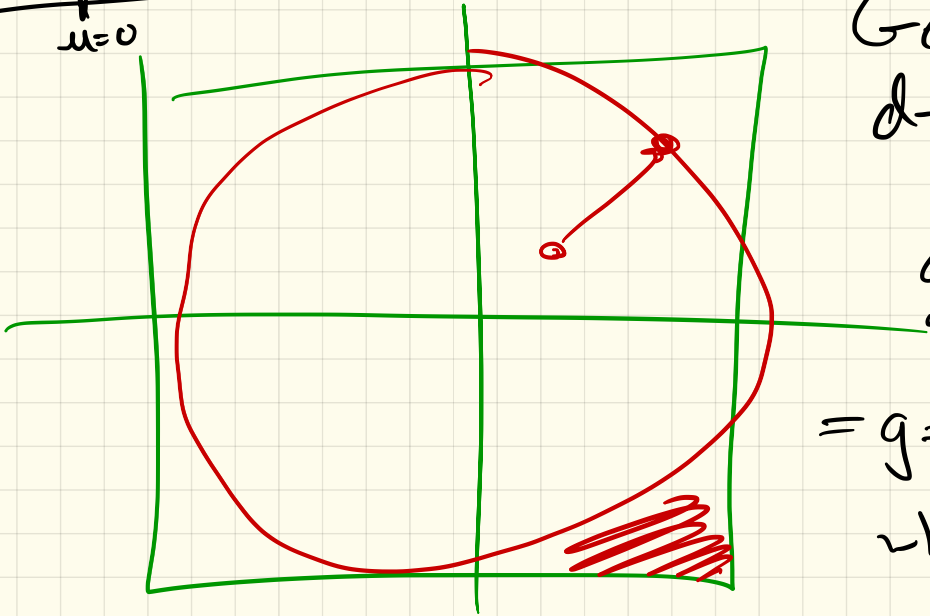
how?

2. bin distances

$h(p) = \lceil c + \langle u, p \rangle \cdot \gamma \pmod{n} \rceil \rightarrow$ 1 | 1 | ... | n

↑ random





Generate d-dim Gaussian R.V.

$g \sim G_d$

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

where each

$g_i \sim G_1$

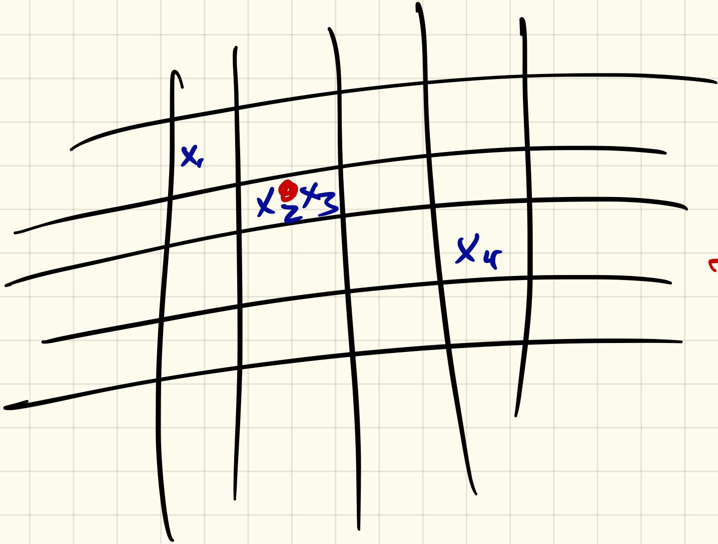
$= \exp(-x^2$

2 uniform $u_1, u_2 \in \text{Unif}(0,1)$

$\Rightarrow g_1, g_2 \in G_1$ Box-Mueller Transform

$$g_1 = \sqrt{-2\ln(u_1)}\,\cos(2\pi u_2)$$
$$g_2 = \sqrt{-2\ln(u_1)}\,\sin(2\pi u_2)$$

query
q



$\Rightarrow$ Union $\{x_2, x_3\}$
$\{x_3\}$