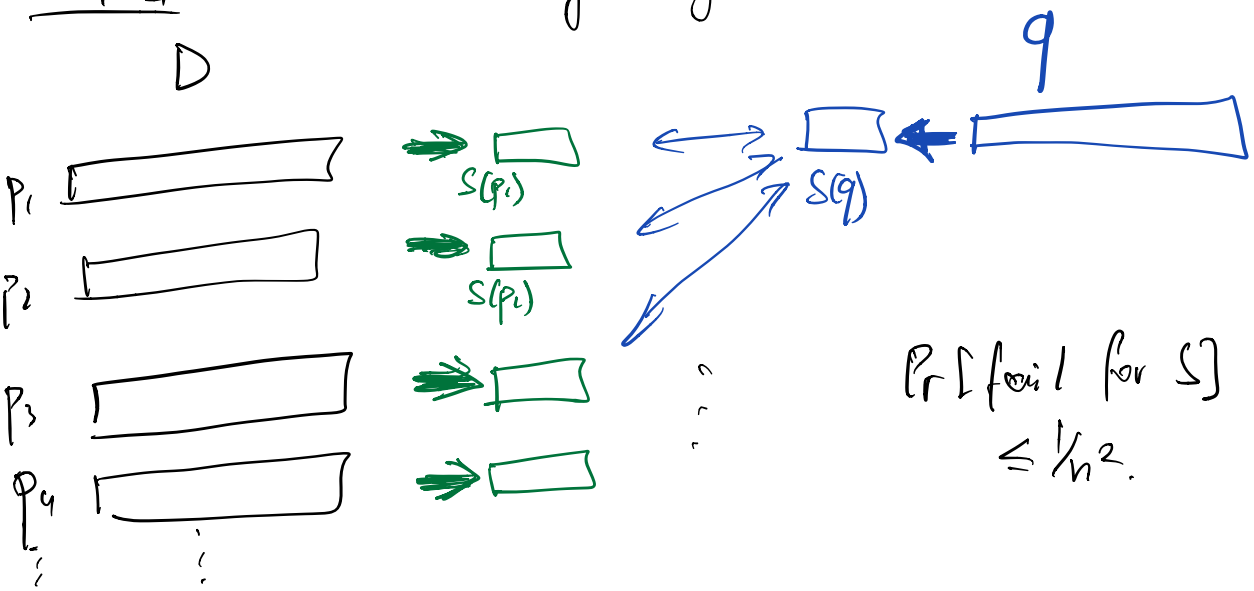
AA Lecture 9 (2/9/21)

Nearest Neighbor Search

c-ANN: for any $r>0$, given $D \subset \mathbb{R}^d$ (or $\{0,1\}^d$), preprocess D s.t. for given query q
- if $\exists p^* \in D$, $\|p^*-q\| \le r$, output p with $\|p-q\| \le cr$.

Sol 1: via sketching (generalization of DR)



Thm [Kushilevitz-Ostrovsky-Rabani '98]
sketch for $\{0,1\}^d$: $\exists \theta>0$ s.t. $\forall r \ge 1$, $\exists$ distribution over $\varphi: \{0,1\}^d \to \{0,1\}^k$, $k = O\left(\frac{\log n}{\varepsilon^2}\right)$
s.t.: - if $\|p-q\| \le r \Rightarrow \Pr[\|\varphi(p)-\varphi(q)\|_1 \le \theta k] \ge 1-\frac{1}{n^3}$
- if $\|p-q\| \ge cr \Rightarrow \Pr[\|\varphi(p)-\varphi(q)\|_1 \ge \theta k] \ge 1-\frac{1}{n^3}$

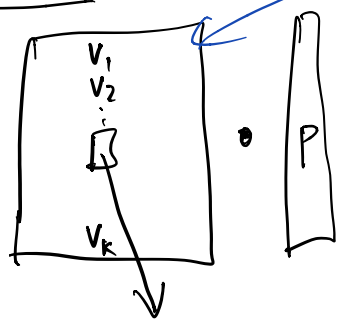
where $c = 1 + O(\varepsilon)$.

like DR for scale r.

size $= d \cdot k$

$t \Rightarrow \oplus$ $\mathbb{F}_2$.

proof [sketch]:

$\varphi(p) =$ 

$= 1$ with prob. $\alpha = \frac{1}{2r}$.

$v_i \in \{0,1\}^d$, $i = 1 \ldots k$.

$\varphi(p) = (v_1 \cdot p,\ v_2 \cdot p, \ldots, v_k \cdot p)$.

$\|\varphi(p) - \varphi(q)\|_1 = \sum_{i=1}^{k} \mathbb{1}[v_i \cdot p \neq v_i \cdot q]$.

$\mathbb{E}[\ \nearrow\ ] = \mathbb{E}[\ \nearrow\ ] = k \cdot \Pr[v_i \cdot p \neq v_i \cdot q]$.

$\Pr_{v}[v \cdot p \neq v \cdot q] = \Pr_{v}\left[\bigoplus_{i=1}^{d} v_i \cdot p_i \neq \bigoplus_{i=1}^{d} v_i \cdot q_i\right]$

each coord $= 1$ w/ Pr $= \alpha$

$= \Pr_{v}\left[\bigoplus v_i \cdot (p_i \oplus q_i) \neq 0\right]$

$v_i = 1$, Pr $= \alpha$.

$\#i$ s.t. $p_i \neq q_i$ is $= \|p - q\|_1$.

$= \text{Prob that } \|p-q\|_1 \text{ random } q_i\text{'s}$
where $Pr=1$ is α, xor to $\neq 0$.

$$= 1 - \Pr_{u_i}\left[\bigoplus_{i=1}^{d} u_i \cdot (p_i \oplus q_i) = 0\right]$$

u_i: with Prob $= 2\alpha$, $u_i \in \{0,1\}$ ← "randomizing coordinates"
$= 0$ otherwise.

$$= 1 - (1-2\alpha)^{\|p-q\|} - \left(1 - (1-2\alpha)^{\|p-q\|}\right)\cdot\frac{1}{2}$$

$$= \frac{1}{2}\left(1 - (1-2\alpha)^{\|p-q\|}\right).$$

$\alpha = \frac{1}{2r}$

1) if $\|p-q\| < r \Rightarrow \Pr[u_p \neq u_q] \leq \frac{1}{2}\left(1 - \left(1-\frac{2}{2r}\right)^{r}\right)$

$$\approx \frac{1}{2}\left(1 - e^{-1}\right) = \frac{1}{2} - \frac{1}{2e}.$$

$(1-x) \approx e^{-x}$ for x small.

2) if $\|p-q\| \geq (1+\epsilon)r \Rightarrow$

$$\Pr[u_p \neq u_q] \geq \frac{1}{2}\left(1 - \left(1-\frac{2}{2r}\right)^{r\cdot(1+\epsilon)}\right)$$

$$\approx \frac{1}{2}\left(1 - e^{-1-\epsilon}\right)$$

$$\approx \frac{1}{2}\left(1 - e^{-1}\cdot(1-\varepsilon)\right)$$

$$\approx \frac{1}{2} - \frac{1}{2e} + \varepsilon\cdot\frac{1}{2e}.$$

Using Chernoff bounds (like HW2/PS hint), can prove that: 1) $\|\underbrace{\varphi(p)-\varphi(q)}_{k}\| \le \frac{1}{2} - \frac{1}{2e} + \frac{\varepsilon}{4e}$

2) $-\!/\!/- \ge \frac{1}{2} - \frac{2}{2e} + \frac{\varepsilon}{2e} - \frac{\varepsilon}{5e}$.

(with prob. $\ge 1 - 1/n^3$). ☒

Corollary: can solve $(1\pm\varepsilon)$-ANN in $\{0,1\}^d$

using spaces: $O(n\cdot k) = O\left(n\cdot\frac{\lg n}{\varepsilon^2}\right)$.

q.t.: $O(d\cdot k) + O\left(n\cdot\frac{\lg n}{\varepsilon^2}\right)$.

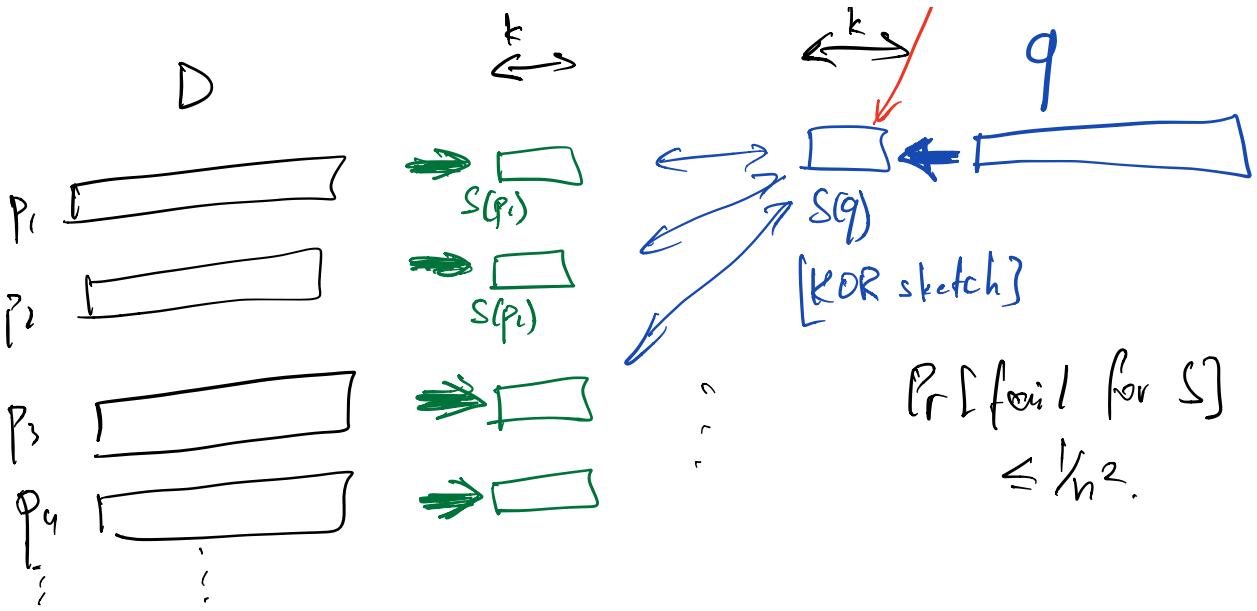
[compute] $\varphi(q)$

Sol 2: space: $n^{O(1/\varepsilon^2)}$.

q.t.: $O\left(d\cdot\frac{\lg n}{\varepsilon^2}\right)$.

in $\{0,1\}^q$!

binary strings of length k.

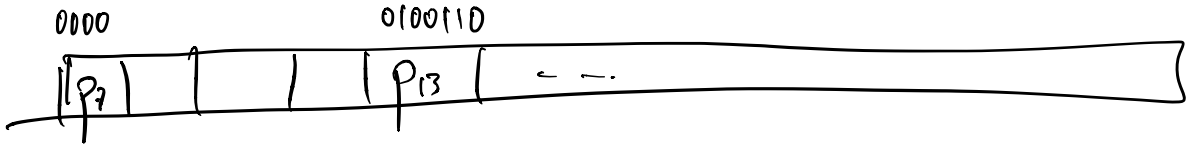


$\Pr[\text{fail for } S] \leq 1/n^2.$

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

Idea: $S(q) = \varphi(q) \in \{0,1\}^k$, $k =$ ____ is all the info we need from q to determine sol to ANN.

Algorithm: for each possible string $\sigma \in \{0,1\}^k$, prepare what would be the ANN answer if $S(q) = \sigma$.



space: $2^k = 2^{O(\frac{\lg n}{\varepsilon^2})} = n^{O(1/\varepsilon^2)}$ (#cells).

q.t.: $O(dk) + O(1) = O\left(\frac{\lg n}{\varepsilon^2}\right)$.

compute $s(q)$

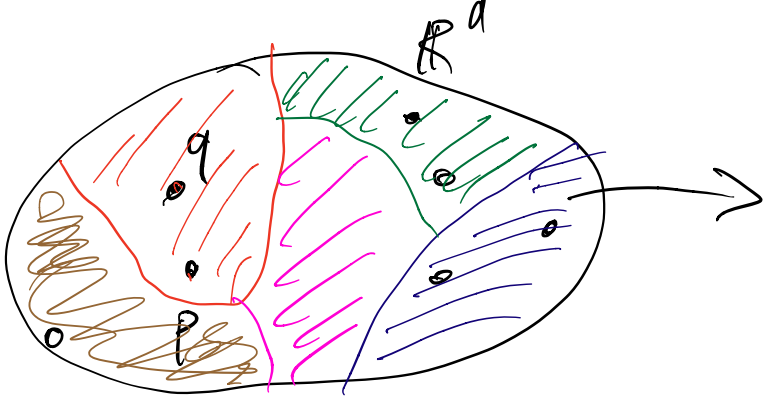
for $\varepsilon < 1/2$.

Sol 3: space: closer to lin. in n

q.t.: n^ρ, $\rho < 1$.

still sub-lin. in n.

Idea: hashing: Locality-Sensitive Hashing (LSH).



Def: fix $r > 0$, approx c. (r, cr, P_1, P_2)-Family $\mathcal{H}$ of $h: \mathbb{R}^d \to U$ is LSH if: $\|p-q\| \le r$, $\Pr_{h \in \mathcal{H}}[h(p) = h(q)] \ge P_1$

- $\|p-q\| > cr$, Pr[ ] $\leq P_2$.

Super-opt: $\exists\ \mathcal{H}$ is $(r, cr, 1, \frac{1}{n})$-LSH.

Algo based on LSH:

Preproc: - compute $h(p)$, $p \in D$
- store in a table (dictionary data struct.)

@ query q: - compute $h(q)$.
- look-up all pts $p \in D$ s.t. $h(q) = h(p)$
- iterate through these p's until we find one that is $\|p-q\| \leq cr$.

Proof of corr/analysis:

$\mathcal{E}$: the bucket $h(q)$ must contain the point p^* (p^* s.t. $\|q-p^*\| \leq r$)

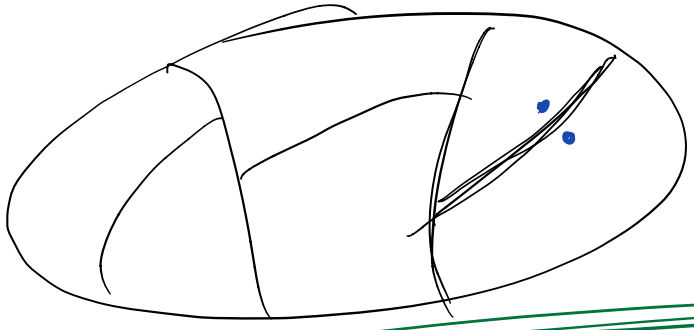
q.t.: $\mathbb{E}[\text{q.t.}] \leq O(1) + \mathbb{E}[\#\text{pts } p \in D$

$\|p-q\| > cr$,
$h(p) = h(q)]$ d

$\leq O(1) + n \cdot P_2 \cdot d$

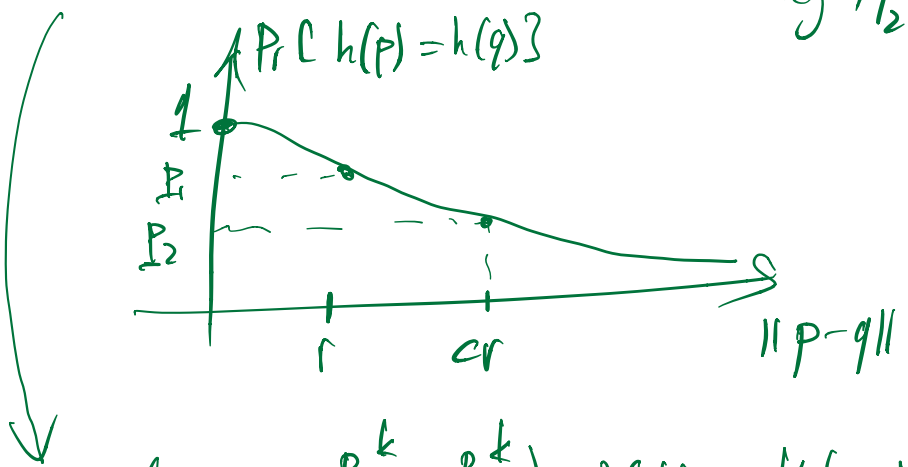
$= O(d n \cdot P_2) = O(d)$.

Caveat: $(r, cr, 1, \frac{1}{n})$-LSH is not possible.



P_1 must be < 1.

(r, cr, P_1, P_2) $\quad \rho = \frac{\lg 1/P_1}{\lg 1/P_2}$.



$\Rightarrow (r, cr, P_1^k, P_2^k)$-LSH $\forall k \geq 1$.

for $\ell_1 / \{0,1\}^\ell$: opt. $\rho = 1/c$ UB: [IM '98]

$$\left[\begin{array}{ll} \text{Sp:} & n^{1+\delta} \\ \text{q.f.} & n^{\delta} \end{array}\right]$$

LB: [MNP '06]
[OWZ '10].

for ℓ_2: $\rho = 1/c^2$. UB: [AI '06].