AA Lecture 10 2/11/21

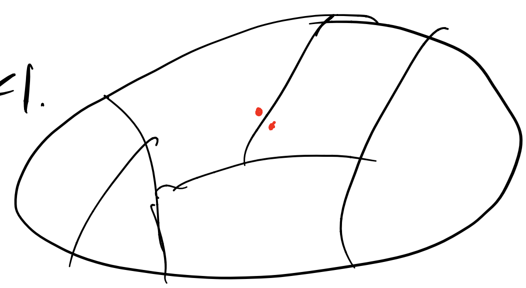
c-approx. near neighbor search in $\{0,1\}^d$.

Def: family $\mathcal{H}$ of $h: \{0,1\}^d \to U$ is (r, cr, P_1, P_2)-LSH if: $\forall p, q \in \{0,1\}^d$

- $\|p-q\|_1 \le r \Rightarrow \Pr_h[h(p)=h(q)] \ge P_1$
- $\|p-q\|_1 > cr \Rightarrow \Pr_h[h(p)=h(q)] < P_2$

Fact: if $P_2 < 1$, then $P_1 < 1$.

(related to isoperimetry questions)

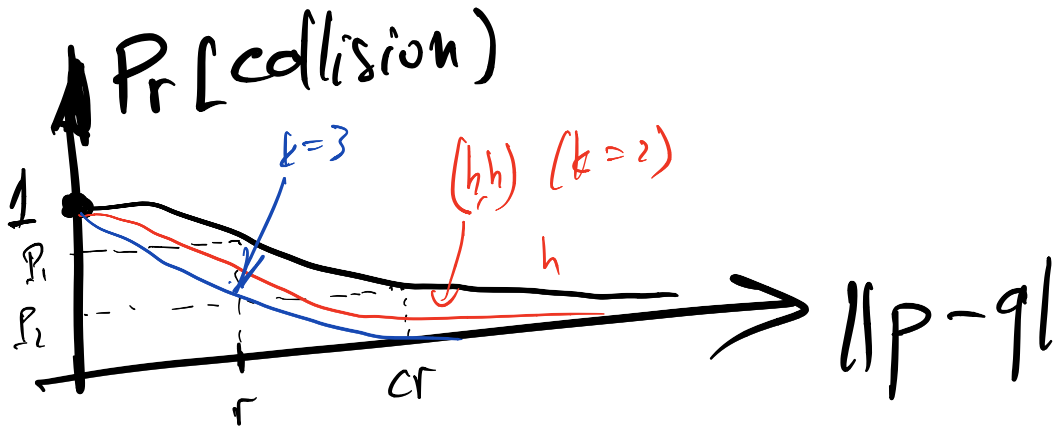


Thm [Indyk-Motwani '98]: if $\exists$ (r, cr, P_1, P_2) LSH, then can solve c-ANN: space: $O(nd + n^{1+\rho}/P_1)$

q.t.: $O(n^\rho/P_1) \cdot d \cdot$ [time to compute $h(q)$].

where $\rho = \frac{\lg 1/P_1}{\lg 1/P_2} \le 1$

proof: $h \in \mathcal{H} \to (r, cr, P_1, P_2)$.



Obs: fix $k \geq 1$. $g(p) = (h_1(p), h_2(p), \dots, h_k(p))$

$h_1, \dots h_k \in \mathcal{H}$ iid

$\Rightarrow$ distrib over $g : \{0,1\}^d \to U^k$

is LSH $\boxed{(r, cr, P_1^k, P_2^k)}$

$$\Pr[g(p) = g(q)] = \prod_{i=1}^{k} \Pr[h_i(p) = h_i(q)]$$

Let's design c-ANN data structure:

Preproc: build a dictionary data structure on points $g(p)$, $p \in D$.

$\hookrightarrow$ all pts $p \in D$ s.t. $g(p) = z$.

@query q: - compute $g(q)$

- retrieve all pts $p \in D$ s.t. $g(p) = g(q)$. (using dict.)
- enumerate through them until find a point p with $\|q-p\| \le cr$.

Analysis: <u>space:</u> just store dict. on n pts
$\Rightarrow O(n)$ space
$(+\ O(nd)$ to store orig. pts$)$

<u>query time:</u> time to compute $g(q)$.
$+$ $\boxed{\text{distance calculations in the bucket.}}$

$$\mathbb{E}[\#\text{ dist. calc.}] \le 1 + \mathbb{E}[\#\text{ dist. calc to pts } p \in D, \text{ s.t. } \|p-q\| > cr, \text{ and } g(p) = g(q)]$$

$$\le 1 + n \cdot \Pr[g(p) = g(q) \mid \|p-q\| > cr]$$

$$\le 1 + n \cdot \underbrace{P_2^k}_{\text{last lecture, assumed } = 1/n}.$$

Correctness: assuming $\exists p^*$ at dist $\leq r$ from q

$\Pr$[algo outputs a cr-near neighbor]

$\geq \Pr[g(p^*) = g(q)] \geq P_1^k$.

To improve $\Pr$[success], we just repeat above $L = 10/P_1^k$ times. (each with fresh hash func. g).

Space: $O(L \cdot n + nd)$

q.t.: $O(L \cdot (1 + nP_2^k + \text{time to comp } g(q)))$

$\Pr[\text{succ.}] \geq 1 - (1-P_1^k)^L \approx 1 - e^{-10/P_1^k \cdot P_1^k} \geq 0.9$.

Set k to minimize q.t.: k^* s.t. $P_2^k = \frac{1}{n}$.

$$L = O\left(\frac{1}{P_1^k}\right) = O\left(\frac{1}{(P_2)^{k_0 \boxed{\frac{\lg 1/P_1}{\lg 1/P_2}}}}\right) = O(n^\rho).$$

$\hookrightarrow \rho$.

Note: factor $1/P_1$ appears in thm statement since k has to be integer.

$$k = \left\lceil \frac{\lg n}{\lg 1/p_2} \right\rceil.$$

LSH family for $\{0,1\}^d$ [IM'98]

$\mathcal{H} = \{ h_i \mid i = 1..d \}$ $\qquad h_i(p) = p_i \in \{0,1\}.$

$g(p) =$ concatenation of k hash func. h

$g(p) =$ projection of p onto k coord. (random).

$$P_1 = \Pr_h [\, h(p) = h(q) \mid \|p-q\| \le r \,] \geq 1 - r/d \approx e^{-r/d}$$

$$P_2 \le 1 - \frac{cr}{d} \approx e^{-cr/d}.$$

$$\rho = \frac{\lg 1/p_1}{\lg 1/p_2} = \frac{r/d}{cr/d} \leq 1/c.$$

Corollary: c-ANN for $\{0,1\}^d$ with:

space: $O(nd + n^{1+1/c})$

q.t.: $O(n^{1/c} \cdot d)$.

$c = 2$

$n^{1.5}$

$\sqrt{n} \cdot d$.

- Thm [MNP'06, OWZ'10]: for $\{0,1\}^d$, $\rho \geq 1/c$.
- for ℓ_2: can get $\rho \leq 1/c^2$, best possible.
- it is possible to beat these bounds by considering data-dep. LSH. ($\sim$ perfect hashing).

- can trade-off sp vs q.t.

  sp: $n^{1+\rho_u}$

  q.t: n^{ρ_q}

  ρ_u, ρ_q satisfy constraints.

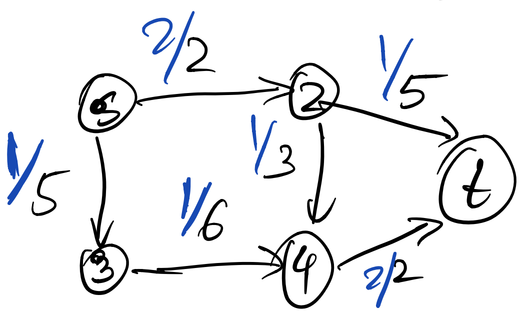
  e.g. $\rho_u \approx 0$, $\rho_q < 1$.

Graphs: max-flow problem.

Consider $G = (V, E, c)$ directed.

nodes, edges, capacities

$c_e \geq 0$.



$n = \#$ nodes

$m = \#$ edges

flow: fix s = start node
t = destination.

f is flow vector $f \in \mathbb{R}_+^m$ s.t.:

1) $f_e \geq 0$, $\forall e \in E$.

2) $f_e \leq c_e$, $\forall e \in E$

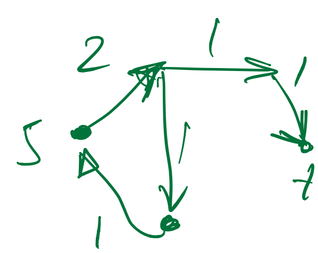
3) flow conservations $\forall v \neq s, t$:

$$\sum_{(u,v)\in E} f_{u,v} - \sum_{(v,u)\in E} f_{v,u} = 0$$

Eg: f = all 0's is valid flow.

Problem of max-flow: find f maximizing flow shipped from s to t:

$$|f| = \sum_{(s,u)\in E} f_{su} - \sum_{(u,s)\in E} f_{us}$$

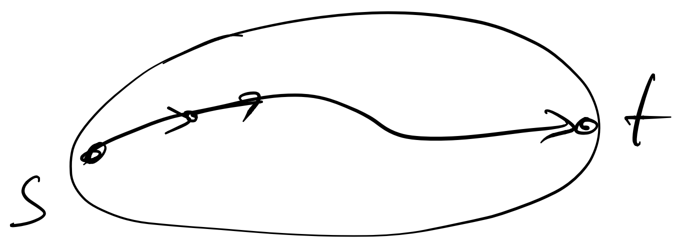


$$\overset{(*)}{=} \sum_{(u,t)\in E} f_{u,t} - \sum_{(t,u)\in E} f_{t,u}.$$

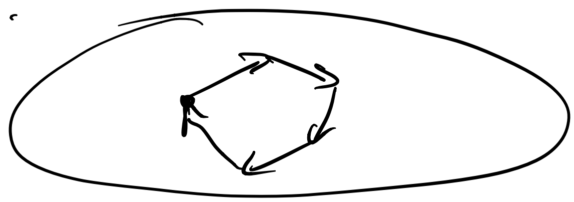
Can prove ⊛ for any flow by summing up all flow cons constraints for $v \neq s,t$.

$\Rightarrow$ $\forall$ edge (u,v) not incident with s,t will cancel out.

<u>**Def:**</u> <u>path flow</u> is a particular type of flow:

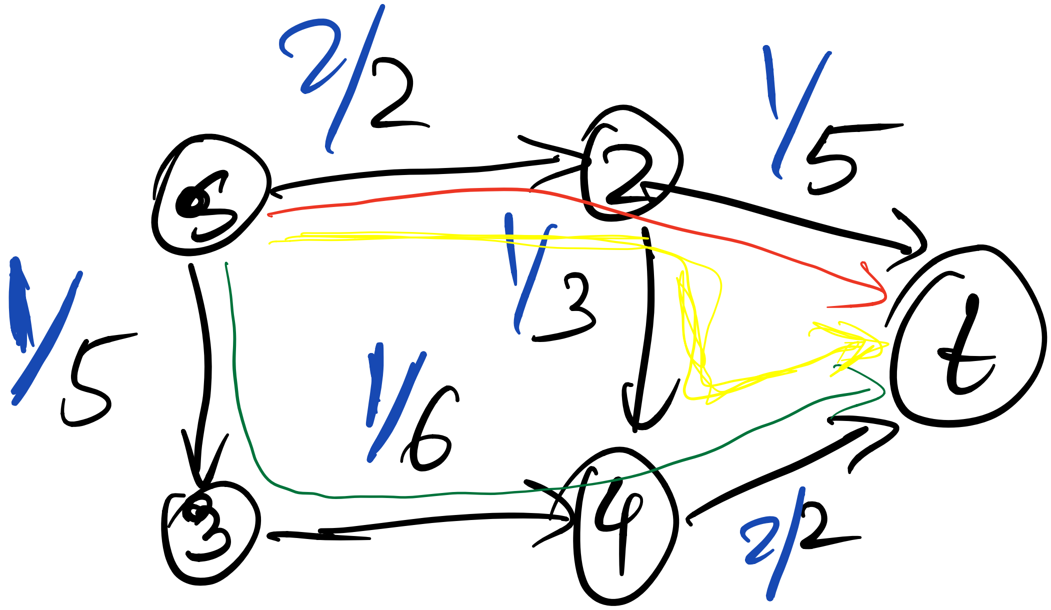


<u>cycle flow</u>:



<u>**Thm:**</u> any valid flow f can be decomposed into a set of path flows $f_{P_1}, f_{P_2}, \dots, f_{P_k}$ and cycle flows $f_{C_1}, f_{C_2}, \dots, f_{C_\ell}$ s.t.:

$$f = f_{P_1} + \dots + f_{P_2} + f_{C_1} + \dots + f_{C_\ell}.$$



$k + \ell \le m.$