Lecture 11 (10/06/2017): Discrepency, Congestion minimization

Lecturer: Shi Li *Scribe: Jiayi Xian*

11.1 Discrepency problem

Given n subsets $S_1, S_2, \cdots, S_n$ of n. The goal is to find a coloring $\chi : [n] \to \{-1, 1\}$ so as to minimize $\max_{i\in[n]} |\sum_{j\in S_i} \chi_j|$. Here $|\sum_{j\in S_i} \chi_j|$ is also referred to as the *discrepancy* of S_i, denoted as $\text{disc}(S_i)$. We don't have such algorithm but we can provide an upper bound using Chernoff bounds technique:

Theorem 11.1. *There exists a coloring χ with maximum discrepancy $d = O(\sqrt{n \lg n})$*

Proof: Let x_j be a $\{0, 1\}$-random variable indicate whether we color the element j by -1 or 1. We choose x_j independently and identically distributed:

$$x_j = \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2 \end{cases}$$

We are going to show that with constant probability the discrepancy of such random coloring will be of $O(\sqrt{n \ln n})$.

Fix a subset S_i, since $x_j = (\chi_j + 1)/2$, $\text{disc}(S_i)$ can be written as $\sum_{j\in S_i} x_j - |S_i|/2$. We can use Chernoff bounds now for $\sum_{j\in S_i} x_j$ since $\mu = \mathbb{E}[\sum_{j\in S_i} x_j] = |S_i|/2$ and each x_j is independently chosen. Let $\delta = \frac{d}{|S_i|}$:

$$\Pr[\sum_{j\in S_i} x_j \geq \frac{|S_i|}{2} + \frac{d}{2}] = \Pr[\sum_{j\in S_i} x_j \geq \frac{|S_i|}{2}(1 + \frac{d}{|S_i|})] \leq \Big(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\Big)^{\mu} \leq e^{-\frac{\delta^2}{3}\cdot\mu} \leq e^{-\frac{\delta^2|S_i|}{6}}$$

if we let $d = \sqrt{6n \ln 3n}$, then $\delta = \sqrt{\frac{6n \ln 3n}{|S_i|^2}} \geq \sqrt{\frac{6 \ln 3n}{|S_i|}}$. We'll have

$$\Pr[\sum_{j\in S_i} x_j \geq \frac{|S_i|}{2} + \frac{d}{2}] \leq e^{-\frac{\delta^2|S_i|}{6}} \leq \frac{1}{3n} \tag{11.1}$$

Similar analysis can show:

$$\Pr[\sum_{j\in S_i} (1 - x_j) \geq \frac{|S_i|}{2} + \frac{d}{2}] \leq \frac{1}{3n} \tag{11.2}$$

The probability of no such bad events for any $i \in [n]$ can be bounded using union bound:

$$\Pr[\forall i, \text{ scenarios in equation 11.1 and 11.2 does not happen}] \geq 1 - 2n \cdot \frac{1}{3n} = \frac{1}{3}$$

This is the same as saying:

$$\Pr[\chi \text{ is a coloring with max discrepency } d] \geq \frac{1}{3}$$

Since the probability space is finite, constant probability implies existence. ■

11.1.1 Congesiton Minimization

Given a graph $G = (V, E)$, k source-sink pairs $(s_1, t_1), (s_2, t_2), \cdots, (s_k, t_k)$. The goal is to choose a path p_i from s_i to t_i for each $i \in [k]$, so as to minimize

$$\max_{e \in E} |\{i \in [k]\} : e \in p_i|$$

The problem can be expressed as an exponential-sized Integer Programming(IP) problem as follows (but we will keep in mind that this IP is equivalent to a polynomial-sized IP, as we will show in Homework 2):
For each $i \in [k]$, define P_i be the set of all paths from s_i to t_i. Define $P = \bigcup_{i \in [k]} P_i$. Then for each $i \in [k]$, for every $p \in P_i$, we use $x_p \in \{0, 1\}$ to indicate whether we use the path p to satisfy the demand pair (s_i, t_i) in our final solution. The IP is:

$$\begin{aligned} \min \quad & C & \text{s.t.} \\ \sum_{p \in P_i} x_p &= 1 & \forall i \in [k] \\ \sum_{p \in P, p \ni e} x_p &\leq C & \forall e \in E \\ x_p &\in \{0, 1\} & \forall p \in P \end{aligned}$$

The relaxed LP can have $x_p \geq 0$ for the last constraint. Now the algorithm is:
(1). Solve LP to obtain the set of $\{x_p\}_{p \in P}$ values.
(2). Rounding: For each $i \in [k]$, independently choose a path $p_i \in P_i$ follows the distribution given by LP, i.e.. $\Pr[p_i = p] = x_p$
(3). Rounding: Let

$$\widetilde{x}_p = \begin{cases} 0 & \text{if } p \text{ is not selected by (2)} \\ 1 & \text{if } p \text{ is selected by (2)} \end{cases}$$

Fix an $e \in E$, now we want to bound $\Pr[\sum_{p \in P, p \ni e} \widetilde{x}_p \geq (1 + \delta)C]$ for some δ we will choose later. Notice that $\{\widetilde{x}_p\}_{p \in P_i}$ are definitely dependent. But such dependency won't hurt here. In fact, let $y_{i,e} = \sum_{p \in P_i, p \ni e} \widetilde{x}_p$, then $\{y_{i,e}\}_{i \in [k]}$ is independent. And we have $\mathbb{E}[\sum_{i \in [k]} y_{i,e}] = \mathbb{E}[\sum_{p \in P, p \ni e} \widetilde{x}_p] = \sum_{p \in P, p \ni e} x_p \leq C$. Then:

$$\Pr[\sum_{p \in P, p \ni e} \widetilde{x}_p \geq (1 + \delta)C] = \Pr[\sum_{i \in [k]} y_{i,e} \geq (1 + \delta)C] \leq \left(\frac{e^\delta}{(1 + \delta)^{1+\delta}}\right)^C \leq \frac{e^\delta}{(1 + \delta)^{1+\delta}}$$

since C is at least 1 in order to make the problem meaningful.
To use the union bound for every edge, δ should be selected as small as possible but satisfy

$$\frac{e^\delta}{(1 + \delta)^{1+\delta}} \leq \frac{1}{2n^2}$$

or

$$\delta - (1 + \delta) \ln (1 + \delta) \leq - \ln 2n^2$$

The equality holds if we let $\delta = \Theta(\ln n / \ln \ln n)$.
Therefore, by union bound, $\Pr[\{\widetilde{x}_p\}$ has congestion at most $(1 + \delta)C] \geq \frac{1}{2}$ for such δ. Put it in another way, there exist an $O(\frac{\ln n}{\ln \ln n})$-approximation for congestion minimization problem.