Stanford University	**Spring 2016**
Math 233: Non-constructive methods in combinatorics	**Instructor: Jan Vondrák**
Lecture date: May 4, 2016	**Scribe: Fan Wei**

Lecture 11. Discrepancy of Set Systems

Here we turn to another problem whose existential solution was known for a long time but an algorithmic solution was missing. This is the problem of discrepancy of set systems, and Spencer's famous result of "six standard deviations".

Setup: $A_1, \ldots, A_m \subseteq [n]$ are given sets. We want a labeling $l : [n] \to \{-1, +1\}$ so that $\chi = \max_{1 \le i \le m} |\sum_{x \in A_i} l(i)|$ is minimized. We are primarily interested in the case $m = n$.

11.1 Natural approach: random labeling

Recall the Chernoff bound:

Lemma 11.1 (Chernoff bound) *If $X_1, \ldots, X_n$ are independent random variables with values being $-1, +1$ with probability 1/2, then*

$$Pr[\sum_{t=1}^{n} X_i > \lambda] \le e^{-\lambda^2/(2n)}.$$

Consider $m = n$ and let l be a random labeling: for each $i \in [n]$ independently, $l(i) = +1$ with probability 1/2 and -1 with probability 1/2. By definition, $l(A_i) = \sum_{x \in A_i} l(x)$. Thus by the Chernoff bound, we have

$$Pr[|l(A_i)| > \lambda] \le e^{-\lambda^2/(2|A_i|)} \le e^{-\lambda^2/(2n)},$$

where the last inequality holds because $|A_i| \le n$. By the union bound, and by setting $\lambda = 2\sqrt{n \log n}$, we have $2ne^{-\lambda^2/(2n)} \ll 1$, and thus

$$Pr[|l(A_i)| > \lambda \text{ for some } i] \le 2/n.$$

We thus have an upper bound on the discrepancy of n subsets of $[n]$: $\chi \le 2\sqrt{n \log n}$.

11.2 Lower bound: $\chi = \Omega(\sqrt{n})$

This is achieved by a set system related to the Hadamard matrix. Definition: $H_0 = \begin{bmatrix} 1 \end{bmatrix}$, $H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, and

$$H_{k+1} = \begin{bmatrix} H_k & H_k \\ H_k & -H_k \end{bmatrix}.$$

By induction it is easy to see that H_k has dimension $n = 2^k$, and these matrices have orthogonal rows and columns. Let $\mathbf{h}_i \in \{-1, 1\}^{2^k}$ be the row vectors of H_k, $1 \le i \le 2^k$.

We define a set system of $m = 2 \cdot 2^k$ sets on 2^k elements, arranged in complementary pairs where $A_i = \{+1 \text{ coordinates of } \mathbf{h}_i\}$, and $A_i' = \overline{A}_i = \{-1 \text{ coordinates of } \mathbf{h}_i\}$, $1 \le i \le 2^k$. Consider a labeling $l : [2^k] \to \{-1, +1\}$. Let $\mathbf{l}$ be the vector of dimension 2^k with coordinates $l(i) \in \{-1, +1\}$. By the definition of $\mathbf{h}_i, \mathbf{l}$, it is easy to see that

$$l(A_i) - l(\overline{A}_i) = \mathbf{l} \cdot \mathbf{h}_i.$$

We thus have

$$|\mathbf{l} \cdot \mathbf{h}_i| = |l(A_i) - l(\overline{A}_i)| \le |l(A_i)| + |l(\overline{A}_i)|.$$

Since the $\mathbf{h}_i$ form an orthogonal basis, and $\|\mathbf{h}_i\|^2 = 2^k$ we have

$$\frac{1}{2^{2k}} \sum_{i=1}^{2^k} (\mathbf{l} \cdot \mathbf{h}_i)^2 = \sum_{i=1}^{2^k} \frac{(\mathbf{l} \cdot \mathbf{h}_i)^2}{\|\mathbf{h}_i\|^2} = \|\mathbf{l}\|^2 = 2^k.$$

Hence, there exists an i with $(\mathbf{l} \cdot \mathbf{h}_i)^2 \ge 2^k$. Thus it implies that

$$2^{k/2} \le |\mathbf{l} \cdot \mathbf{h}_i| = |l(A_i) - l(\overline{A}_i)| \le |l(A_i)| + |l(\overline{A}_i)|.$$

Therefore either $|l(A_i)| \ge \frac{1}{2} 2^{k/2}$ or $|l(\overline{A}_i)| \ge \frac{1}{2} 2^{k/2} = \frac{1}{2}\sqrt{n}$.

11.3 Spencer's Theorem

Spencer proved that $\Theta(\sqrt{n})$ is indeed the right answer for n sets on n elements.

Theorem 11.2 (Spencer '85) *For any set system $A_1, \ldots, A_n \subseteq [n]$, there is a labeling $l : [n] \to \{-1, +1\}$ such that for any $1 \le i \le n$, we have $|l(A_i)| \le 6\sqrt{n}$.*

We will prove this result with a somewhat weaker constant. The proof uses three ingredients which we state without proof:

1. Chernoff bound (above).
2. $\sum_{i=0}^{\alpha n} \binom{n}{i} \le 2^{nH(\alpha)}$ where $H(\alpha) = \alpha \log_2 \frac{1}{\alpha} + (1-\alpha) \log_2 \frac{1}{1-\alpha}$ is the binary entropy function.
3. Kleitman's inequality:

   Lemma 11.3 (Kleitman's Inequality) *If $S \subset \{-1, +1\}^n$ and $|S| > \sum_{i=0}^{r} \binom{n}{i}$, then the diameter of S is greater than $2r$, i.e., there are two points $x, y \in S$ such that x, y differ in more than $2r$ coordinates.*

Now we proceed to the proof of Spencer's theorem.

Step 1: Many "realistic labelings".

Definition 11.4 *A labeling l is called* realistic *if at most $2^{s+2}e^{-50(2s-1)^2}n$ sets have discrepancy greater than $10(2s-1)\sqrt{n}$ for any integer $s \geq 1$.*

We consider a random labeling $l : [n] \to \{-1, +1\}$. We know by Chernoff bound again, that for any positive integer s,

$$Pr[|l(A_i)| > 10(2s-1)\sqrt{n}] < 2e^{-50(2s-1)^2}.$$

Thus by Markov inequality,

$$Pr[\text{ more than } 2^{s+2}e^{-50(2s-1)^2}n \text{ sets with discrepancy greater than } 10(2s-1)\sqrt{n}] < 2^{-s-1}.$$

By union bound over all s, we have that the probability l is not realistic is at most $1/2$. It is equivalent to say that there are at least 2^{n-1} realistic labelings.

Step 2: Not many choices of "signatures".

Definition 11.5 *For a labeling $l : [n] \to \{-1, +1\}$, we define a* signature $T(l) \in \mathbb{Z}^n$ *where $(T(l))_i$ is the integer closest to $l(A_i)/(20\sqrt{n})$.*

Ideally, we would like to find a labeling l such that $T(l)$ is the all-zero vector. This is not easy directly, but we will prove that there are two labelings l', l'' with the same signature and many different coordinates. This allows us to find a "partial labeling" $l = (l' - l'')/2$ of low discrepancy, and then we can iterate to find a full labeling.

First, we prove the following.

Lemma 11.6 *If $\mathcal{R}$ denotes the set of all realistic labelings for $A_1, \ldots, A_n \subseteq [n]$, then*

$$|T(\mathcal{R})| \leq 2^{10^{-12}n}.$$

Proof: (a) By the definition of being realistic, there are at most $8e^{-50}n$ sets having discrepancy greater than $10\sqrt{n}$. This means for realistic l, that $T(l)$ has at most $8e^{-50}n$ coordinates being non-zero. The number of ways to choose at most $8e^{-50}n$ coordinates out of n coordinates is at most $\sum_{i=0}^{8e^{-50}n} \binom{n}{i} \leq 2^{nH(8e^{-50})}$ by the second ingredient.[1]

(b) For any choice of the non-zero coordinates, there are $2^{8e^{-50}n}$ choices of signs ($\pm$) for these coordinates.

(c) We bound the number of choices for the values of these non-zero coordinates. The coordinates such that $|T(l)| > 1$ corresponds to discrepancy at least $30\sqrt{n}$, which by the definition of being realistic, there are at most $16e^{-450}n$ such coordinates. Similarly, there are at most $2^{s+2}e^{-50(2s-1)^2}n$ coordinates with $|T(l)| > s-1$. Thus at most $2^{nH(2^{s+2}e^{-50(2s-1)^2})}$ choices for coordinates with value greater than $s-1$. [2]

Combining (a), (b), (c), the total number of choices for $T(l)$ is at most

$$2^{8e^{-50}n} \prod_{s=1}^{\infty} 2^{nH(2^{s+2}e^{-50(2s-1)^2})} \leq 2^{10^{-12}n}$$

(using some simple crude bounds on $H(\alpha)$). □

[1] We ignore rounding issues here; $8e^{-50}n$ might not be an integer but for $n \to \infty$ the rounding errors become negligible.

[2] (a) is a special case of (c).

Step 3: Find two labellings with similar signature but far from each other. Since there are at least 2^{n-1} realistic labellings, and there are at most $2^{10^{-12}n}$ signatures for realistic labellings, we know by the pigeon-hole principle that there must be a signature $b \in \mathbb{Z}^n$ such that at least $2^{n-1}/2^{10^{-12}n} = 2^{(1-10^{-12})n-1}$ realistic labelings have $T(l) = b$.

Now by Kleitman's inequality, there exists two realistic labelings l', l'' such that $T(l') = T(l'') = b$, while their Hamming distance is at least $(1 - 10^{-6})n$ by applying S to be the set of realistic labellings with signature b.

Step 4: Construct a satisfactory labeling. We have obtained two labellings l', l'' with similar discrepancy for each A_i (since their signatures are the same), but they are far away from each other. Let $l = (l' - l'')/2$. Thus l has values $0, \pm 1$ for each coordinate. However, the number of zero coordinates for l is at most $10^{-6}n$ by the fact that the hamming distance between l', l'' is at least $(1 - 10^{-6})n$. Also, for every set A_i, $|l(A_i)| = |l'(A_i) - l''(A_i)|/2 \leq 10\sqrt{n}$ as l', l'' have the same signature.

Step 5: Iterate... We have almost achieved what we wanted: a labeling with small discrepancy. However, l is only a partial labeling; it has a small number of zero coordinates (at most $10^{-6}n$) for which we still have to decide between ± 1. We recurse the process on this sets of coordinates.

We have to repeat our analysis in a slightly more general setting: with m sets on n elements, $m \geq n$. One can prove the following statement.

Lemma 11.7 *For $m \geq n$ and any system of m sets on n elements, there is a labeling $l : 2^{[n]} \to \{-1, 0, +1\}$ such that at most $10^{-6}n$ elements are labeled 0 and for each $i \in [m]$,*

$$|l(A_i)| \leq 10\sqrt{n \log \frac{2m}{n}}.$$

We do not repeat all the steps here. The key modifications of the above proof are that a realistic labeling is one for which the discrepancy of at most $2^{s+2}(\frac{n}{2m})^{50(2s-1)^2}m$ sets is more than $10(2s-1)\sqrt{n \log \frac{2m}{n}}$. Then we define a signature $T(l)$ whose i-th coordinate is the integer closest to $l(A_i)/(10\sqrt{n \log(2m/n)})$. A similar analysis as above implies that there are two labelings l', l'' of the same signature, which differ in at least $(1 - 10^{-6})n$ coordinates. Then $l = (l' - l'')/2$ is the desired labeling.

By recursing on the elements that are labeled 0, and composing the resulting labelings in a natural way, we obtain a labeling ℓ of discrepancy

$$\chi \leq 10\sqrt{n} + 10\sqrt{10^{-6}n \log(2 \cdot 10^6)} + 10\sqrt{10^{-12}n \log(2 \cdot 10^{12})} + \ldots \leq 11\sqrt{n}.$$