Introduction to Geometric Discrepancy **Fall 2017**

Section 4: Combinatorial Discrepancy

Lecturer: Jiri Matousek *Scribe: Lily Li*

4.1 Upper Bounds for General Set Systems

The question we want to consider here is: Given an n-point set X and a set system § on X where $|\mathcal{S}| = m$ (typically with $n \le m$) what is the largest value of $\mathsf{disc}(\mathcal{S})$? We show that random coloring actually produces pretty good discrepancy.

Lemma 4.1 *(Random coloring lemma). Let $\mathcal{S}$ be a set system on an n-point set X. For random coloring $\chi : X \to \{1, -1\}$,*

$$|\chi(S)| \le \sqrt{2|S| \ln(4|\mathcal{S}|)}$$

for all sets $S \in \mathcal{S}$ simultaneously with probability at least $1/2$.

Proof: *Warning:* random probability in coming!Consider any $S \subset X$. For a random coloring χ, $\chi(S) = \sum_{x \in S} \chi(x)$ which is the sum of $s = |S|$ random ± 1 random variables. This sum has a binomial distribution with standard deviation $\sqrt{2}$ (why?). Use the Chernoff tail estimate to get

$$\Pr[|\chi(S) \ge \gamma\sqrt{2}] < 2e^{-\gamma^2/2}.$$

By setting $\gamma = \sqrt{2\ln(4|\mathcal{S}|)}$ in the above we have the probability bounded by $(2|\mathcal{S}|)^{-1}$. Thus over all $|\mathcal{S}|$ sets there is at least a $1/2$ probability that $|\chi(S)| \le \sqrt{2|S| \ln(4|\mathcal{S}|)}$ for all $S \in \mathcal{S}$. ■

Notice that if nothing is known about the sets $S \in \mathcal{S}$ then $|S| \le |X| = n$ and the bound reduces to $\mathsf{disc}(\mathcal{S}) = O(\sqrt{n \log m})$. Thus the discrepancy varies with both the sizes of each set in $\mathcal{S}$ as well and the number of sets in $\mathcal{S}$ though the latter is of greater importance (small sets are better). Next we show a small improvement on the above lemma:

Theorem 4.2 *(Spencer's upper bound) Let $\mathcal{S}$ be a set system on a n-point set X with $|\mathcal{S}| = m \ge n$ then*

$$\mathsf{disc}(\mathcal{S}) = O\left(\sqrt{n \log(\frac{2m}{n})}\right)$$

in particular if $m = O(n)$ then $\mathsf{disc}(\mathcal{S}) = O(\sqrt{n})$.

We will be ready for this proof after reading section 4.6. But first a nice theorem (they are presenting it not to be used but to be marveled at):

Theorem 4.3 *(Beck-Fiala theorem) Let $\mathcal{S}$ be a set system on an arbitrary finite set X such that $\deg_{\mathcal{S}}(x) \le t$ for all $x \in X$, where $\deg_{\mathcal{S}}(x) = |\{S \in \mathcal{S} : x \in S\}|$. Then $\mathsf{disc}(\mathcal{S}) \le 2t - 1$.*

Proof: Remember, to show that $\mathsf{disc}(\mathcal{S}) \leq 2t-1$ we need to find some coloring χ such that for all n, a point set $|P| = n$ and all $S \in \mathcal{S}$, $\mathsf{disc}(\chi, P, S) \leq 2t-1$. Let $X = \{1, ..., n\}$ and to each $j \in X$ assign the real variable $x_j \in [-1,1]$. We will fix a coloring one variable at a time by changing the value of the x_js. Defined *fixed* variable x_j to a variable with value ±1 otherwise the variable is *floating.* Set all v_j to be 0 a the beginning of the algorithm (thus making them all float). Further call a set $S \in \mathcal{S}$ *dangerous* if it contains more than t elements j with x_j currently floating. Otherwise S is *safe.* We maintain invariant (1):

$$\sum_{j \in S} x_j = 0 \quad \text{for all dangerous } S \in \mathcal{S}.$$

For each $S \in \mathcal{S}$ we can build a linear equation of the form of invariant (1) with floating variable as variables of out equation. The system of equations has a solution (namely the current value of the floating variables) and the set of all possible solutions lives inside of the cube $[-1,1]^{|F|}$. We want to show that some solution exists on the boundary (picking this boundary point will fix some floating variables). This is because the number of dangerous sets is smaller than the number of floating variables. Thus this system has more variables than unknowns an there exists a line in the solution space (this line hits the boundary of the boundary of the hyper cube). When all sets become safe there can be at most t floating variables per set. Each variable can change by less than 2 thus the discrepancy is at most $2t-1$. ■

4.2 Matrices, Lower Bounds, and Eigenvalues

Definition 4.4 *Let $(X, \mathcal{S})$ be a set system on a finite set. Let the elements of X be $x_1, x_2, ..., x_n$ and the elements of $\mathcal{S}$ be $S_1, S_2, ..., S_m$. The* ***the incidence matrix*** *of $(X, \mathcal{S})$ is the $m \times n$ matrix X with columns corresponding to the points of X and rows corresponding to sets of $\mathcal{S}$, whose element a_{ij} are*

$$a_{ij} = \begin{cases} 1 & \text{if } x_j \in S_i \\ 0 & \text{otherwise.} \end{cases}$$

Let coloring $\chi : X \rightarrow \{-1, 1\}$ be regarded as the column vector $(\chi(x_1), \chi(x_2), ..., \chi(x_n))^T \in \mathcal{R}^n$. The product A_χ is the row vector $(\chi(S_1), \chi(S_2), ..., \chi(S_m)) \in \mathcal{R}^m$. We can rewrite the definition of discrepancy of $\mathcal{S}$ to be

$$\mathsf{disc}(\mathcal{S}) = \min_{x \in \{-1,1\}^n} \|Ax\|_\infty$$

where the norm $\|y\|_\infty$ of a vector y is defined to be its maximum element.

Definition 4.5 *A* ***Hadamard Matrix*** *is an $n \times n$ matrix H with entries ±1 such that any two distinct columns are orthogonal i.e. $H^T H = nI$ where I is the $n \times n$ identity matrix. One convention is to assume that the first column of H consists of all ones.*

Note H does not exist for every n. Simple requirements include n must be even but the existence problem for H is not yet solved.

Proposition 4.6 ***(Hadamard set system).*** *Let H be an $n \times n$ Hadamard matrix and $\mathcal{S}$ be the set system with incident matrix $A = \frac{1}{2}(H + J)$ where J is the $n \times n$ all ones matrix. Then*

$$\mathsf{disc}(\mathcal{S}) \geq \mathsf{disc}_2(\mathcal{S}) \geq \frac{\sqrt{n-1}}{2}.$$