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 $n \leq m$) 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 $\chi : X \to \{1, -1\}$,*

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

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

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

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 $|\chi(S)| \leq \sqrt{2|S| \ln(4|\mathcal{S}|)}$ for all $S \in \mathcal{S}$. ■

Notice that if nothing is known about the sets $S \in \mathcal{S}$ then $|S| \leq |X| = n$ 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 $|\mathcal{S}| = m \geq n$ 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 $\deg_{\mathcal{S}}(x) \leq t$ for all $x \in X$, where $\deg_{\mathcal{S}}(x) = |\{S \in \mathcal{S} : x \in S\}|$. Then $\mathsf{disc}(\mathcal{S}) \leq 2t - 1$.*

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*

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

Let coloring $\chi : X \to \{-1, 1\}$ 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. Notice here that what we have is essentially a matrix multiplication. Your resulting vector contains the chromatic number of each set.

Next we need to recall the definition of disc_p from chapter 1:

$$\mathsf{disc}_p(\chi, \mathcal{S}) = \left(\frac{1}{|\mathcal{S}|} \sum_{S \in \mathcal{S}} |\chi(S)|^p \right)^{1/p}$$

in particular we will need the instance when $p = 2$ because

$$\mathsf{disc}_2(\mathcal{S}) = \left(\frac{1}{|\mathcal{S}|} \sum_{S \in \mathcal{S}} |\chi(S)|^2 \right)^{1/2} = \frac{1}{\sqrt{m}} \cdot \min_{x \in \{-1,1\}^n} \|Ax\|.$$

The linear algebra fact that $\|Ax\|^2 = (Ax)^T(Ax) = x^T(A^TA)x$ comes in handy soon.

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.*

Side 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*

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

Proof: Ok, linear algebra incoming. Observe that

$$A^T A = \frac{1}{4}(H+J)^T (H+J) = \frac{1}{4}\left(nI + nJ + nR + nR^T\right)$$

where R is the $n \times n$ matrix whose first row is all ones and the rest of the matrix is all zeros. Now lets put back the x^T, x, and the $1/n$ (used to be a $1/m$ in the above side note but here $m = n$ since A is square):

$$\begin{aligned} \frac{1}{n} x^T (A^T A) x &= \frac{1}{4}\left(\sum_{i=1}^{n} x_i^2 + 2x_1\left(\sum_{i=1}^{n} x_i\right) + \left(\sum_{i=1}^{n} x_i\right)^2\right) \\ &= \frac{1}{4}\left(\sum_{i=2}^{n} x_i^2 + (2x_1 + x_2 + \cdots + x_n)^2\right) \\ &\geq \frac{1}{4}\left(\sum_{i=2}^{n} x_i^2\right) = \frac{n-1}{4}. \end{aligned}$$

After applying Spencer's upper bound (from) above, you get the inequality that you want. ■

Theorem 4.7 ***(Eigenvalue bound for discrepancy.)*** *Let $(\mathcal{S}, X)$ be a system of m sets on an n-point set, and let A denote its incidence matrix. Then*

$$\mathsf{disc}(\mathcal{S}) \geq \mathsf{disc}_2(\mathcal{S}) \geq \sqrt{\frac{n\lambda_{min}}{m}}$$

4.3 Hereditary and Linear Discrepancy

Definition 4.8 *The **hereditary discrepancy** of a set system $\mathcal{S}$ taken over a ground set X is*

$$\mathsf{herdisc}(\mathcal{S}) = \max_{Y \subset X} \mathsf{disc}(\mathcal{S}|_Y).$$

It should be easy to see that $\mathsf{herdisc}(\mathcal{S}) \geq \mathsf{disc}(\mathcal{S})$. *In-fact, by the way it is defined, hereditary discrepancy avoids the issue of* bad *ground sets so can be a good measure of complexity for the set system.*

Definition 4.9 *On a related note: **linear discrepancy** of $\mathcal{S}$ is a form of discrepancy arising from the rounding problem. If each point $x \in X$ has weight $w(x) \in [-1, 1]$, now we want to color the points ± 1 such that $\chi(S)$ is close to the sum of the weights of points in S for every $S \in \mathcal{S}$.*

$$\mathsf{lindisc}(A) = \max_{w \in [-1,1]^n} \min_{x \in \{-1,1\}^n} \|A(x - w)\|_\infty$$

for incidence matrix A. Again it should be easy to see that $\mathsf{lindisc}(A) \geq \mathsf{disc}(A)$ *since* $\mathsf{disc}(A) = \min_{x \in \{-1,1\}^n} \|Ax\|_\infty$.

Though there does not seem to be a large connection between $\mathsf{herdisc}$ and $\mathsf{lindisc}$ we have the following relationship:

Theorem 4.10 *For any set system $\mathcal{S}$,*

$$\mathsf{lindisc}(\mathcal{S}) \leq 2\mathsf{herdisc}(\mathcal{S}).$$

The above remains true if we replace the set system with an incidence matrix A.

Proof: Before we can go through with the proof, we actually need some geometric intuition of the three different types of discrepancies. Let A be an $m \times n$ incidence matrix and define

$$U_A = \{x \in \mathcal{R}^n : \|Ax\|_\infty \leq 1\}.$$

Observe that U_A is some symmetric convex polyhedron about the origin which contain all real colorings of the n points such that the discrepancy is less than one. Generally, each row of the incidence matrix can be though of as a pair of constraints bounding the coloring to be less than one, but also greater than -1.

Next observe that for any $x \in \mathcal{R}^n$, $\|Ax\|_\infty = \min\{t \geq: x \in tU_A\}$ since U_A is the unit bounding region for the solution and you need to scale up by $\|Ax\|_\infty$ for the bounding region to contain x. Thus $\mathsf{disc}(A)$ can be restated as the smallest value of t such that tU_A contains a vertex of the cube $[-1,1]^n$. Note a vertex of the cube represents a valid coloring. Alternatively, we could shift this whole arrangement from the origin to each vertex $a \in \{-1,1\}^n$ of the cube and ask for the smallest t such that $tU_A + a$ contains the origin.

Thus the geometric interpretation of $\mathsf{disc}(A)$, $\mathsf{herdisc}(A)$, and $\mathsf{lindisc}(A)$ are

1. $\mathsf{disc}(A)$ is the first moment when the growing polytopes first encloses the origin (the reason for this is discussed above).
2. $\mathsf{herdisc}(A)$ is the first moment that for each face F of the cube (of every dimension), the center of F is covered by some bodies centered at the vertices of F. Note that the center of F corresponds to setting some $x \in X$ to be 0 i.e. do not worry about coloring some subset of X. This corresponds precisely with coloring a subset of X.
3. $\mathsf{lindisc}(A)$ is the first moment when the whole cube is covered. Consider what it means for a point in the cube to be covered by the growing polytopes. It means that there is a ± 1 coloring a of the elements of X (each coloring is a vertex of the cube) such that there infinity norm is contain with in $tU_A + a$ (described above). Since linear discrepancy is concerned with rounding, all points are covered.

See page 111 of the book for an illustrative picture.

Lets see if we can get anything out of this proof. In light of the geometry, we show that if U is a closed convex body such that $\cup_{a\in\{-1,1\}^n}(U + a)$ covers all the points of $\{-1,0,1\}^n$ then the set $C = \cup_{a\in\{-1,1\}^n}(2U + a)$ covers the whole cube. This should be pretty reasonable if you consider the two dimensional case.

Hum... they prove this in quite a round about way. First they divide the cube into a grid and induct on the length of one edge of the grid i.e. grid size $1, 1/2, 1/2^2, ..., 1/2^k$ and induct on k. I think you can manage. ■

4.3.1 Lower Bound in Terms of Determinants

Theorem 4.11 *For any set system $\mathcal{S}$,*

$$\mathsf{herdisc}(\mathcal{S}) \geq \frac{1}{2} \max_k \max_B |\det(B)|^{1/k},$$

where B ranges over all $k \times k$ sub-matrices of the incident matrix of $\mathcal{S}$.

Proof: Use $\mathsf{lindisc}(\mathcal{S}) \leq 2\mathsf{herdisc}(A)$ and Lemma 4.12. ■

Lemma 4.12 *If A is an $n \times n$ matrix, then* $\mathsf{lindisc}(A) \geq |\det(A)|^{1/n}$.