Geometric combinatorics: supplementary notes 3*

If $(X, \mathcal{F})$ is a set family then the *signed discrepancy* of $S \in \mathcal{F}$ with respect to the colouring $\chi\colon X \to \{-1, +1\}$ is $\chi(S) = \sum_{x \in S} \chi(x)$. The function χ is thought of as a colouring into two colours, normally called red and blue. The value $\chi(S)$ is then a measure of disbalance between two colours. The *combinatorial discrepancy* (or simply discrepancy) of $(X, \mathcal{F})$ with respect to χ is then $\operatorname{disc}(\mathcal{F}, \chi) = \max_{S \in \mathcal{F}} |\chi(S)|$, and the combinatorial discrepancy of $(X, \mathcal{F})$ is $\operatorname{disc}(\mathcal{F}) = \min_\chi \operatorname{disc}(\mathcal{F}, \chi)$. Finally, since we are primarily interested in families $\mathcal{F}$ that are infinite (ellipsoids, convex sets, etc), we introduce the notation $\operatorname{disc}(n, \mathcal{F}) = \max_{|A|=n} \operatorname{disc}(\mathcal{F}|_A)$. One can bound the geometric discrepancy by combinatorial discrepancy:

Theorem 1. *Let $\mathcal{F}$ be a family of Lebesgue-measurable sets in $[0,1]^d$. Suppose $\operatorname{disc}(n, \mathcal{F}) \le f(n)$ for all n. Assume that the following conditions are satisfied:*

1. *$f(2n) \le (2 - \delta) f(n)$ for some constant $\delta > 0$,*
2. *$D(n, \mathcal{F}) = o(n)$,*
3. *$[0,1]^d \in \mathcal{F}$.*

Then

$$D(n, \mathcal{F}) = O(f(n)).$$

The conditions (1)-(3) above are purely technical. Among the conditions (1) and (2), the condition (1) is stronger: morally it says that $\operatorname{disc}(n, \mathcal{F}) = O(n^{1-\delta})$ for some fixed $\delta > 0$. The bounds on discrepancy of interesting geometrically-defined set families all satisfy this condition. The condition (3) is for convenience only, and can be relaxed.

Lemma 2. *Let $(X, \mathcal{F})$ be a set system.on $|X| = 2n$ points, and $X \in \mathcal{F}$. Then there is an n-point subset $Y \subset X$ such that*

$$\left| \frac{|Y \cap S|}{|Y|} - \frac{|S|}{|X|} \right| \le \frac{\operatorname{disc}(\mathcal{F})}{|X|}$$

for every $S \in \mathcal{F}$.

*These notes are from `http://www.borisbukh.org/GeoCombEaster10/suppnotes3.pdf`.

Proof. Pick a colouring $\chi\colon X \to \{-1,+1\}$ satisfying $\operatorname{disc}(\mathcal{F}) = \operatorname{disc}(\chi,\mathcal{F})$. Let Y' be the larger of the two colour classes $\chi^{-1}(-1)$ and $\chi^{-1}(+1)$. Let $Y \subset Y'$ be any subset of exactly n points. If $S \in \mathcal{F}$, then

$$\big|2|Y' \cap S| - |S|\big| = \big||Y' \cap S| - |S \setminus Y'|\big| = |\chi(S)| \le \operatorname{disc}(\mathcal{F}).$$

Since $X \in \mathcal{F}$, this in particular implies that $|Y'| - n/2 \le \frac{1}{2}\operatorname{disc}(\mathcal{F})$. Thus, for every $S \in \mathcal{F}$

$$\begin{aligned}\big||Y \cap S| - \tfrac{1}{2}|S|\big| &\le |Y' \setminus Y| + \big||Y' \cap S| - \tfrac{1}{2}|S|\big| \\ &\le \tfrac{1}{2}\operatorname{disc}(\mathcal{F}) + \tfrac{1}{2}\operatorname{disc}(\mathcal{F}).\end{aligned}$$

□

Proof. By condition (2) there is a sufficiently large integer k and a set P_0 of $2^k n$ points such that

$$\frac{D(P_0,\mathcal{F})}{|P_0|} \le \frac{f(n)}{n}.$$

Starting with P_0 we shall build a sequence of sets $P_0, P_1, \ldots, P_k$ that are of size $|P_i| = 2^{k-i}n$, where each subsequent set P_{i+1} is a good approximation to the preceding set P_i with respect to $\mathcal{F}$. Namely, the preceding lemma applied to the set system $(P_i, \mathcal{F}|P_i)$ yields existence of set P_{i+1} that satisfies

$$\left|\frac{|P_i \cap S|}{|P_i|} - \frac{|P_{i+1} \cap S|}{|P_{i+1}|}\right| \le \frac{f(2^{k-i}n)}{2^{k-i}n}.$$

Therefore, the condition (1) of the theorem then implies

$$\begin{aligned}\left|\frac{|P_k \cap S|}{|P_k|} - \frac{|P_0 \cap S|}{|P_0|}\right| &\le \sum_i \frac{f(2^{k-i}n)}{2^{k-i}n} \le \frac{f(n)}{n}\left(1 + \frac{2-\delta}{2} + \frac{(2-\delta)^2}{2^2} + \cdots\right) \\ &= O\left(\frac{f(n)}{n}\right).\end{aligned}$$

By the choice of P_0 the geometric discrepancy of P with respect to any $S \in \mathcal{F}$ is

$$\left|\frac{|P_0 \cap S|}{|P_0|} - \operatorname{vol}(S)|P_0|\right| \le \frac{f(n)}{n}.$$

The theorem then follows from the triangle inequality. □

If there is a trivial upper bound for combinatorial discrepancy, the following bound has the best claim to that title, for it is the bound that holds for almost every choice of the colouring function χ.

Theorem 3. *Let $(X,\mathcal{F})$ be a set system. Let $s = \max_{S\in\mathcal{F}}|S|$. Then with probability at least $1/2$ a colouring $\chi\colon S \to \{-1,+1\}$ that is chosen uniformly at random from the set of all $2^{|S|}$ two-colourings of S has discrepancy*

$$\operatorname{disc}(\mathcal{F},\chi) \le \sqrt{2s\log(4|\mathcal{F}|)}.$$

Proof. For a set $S \in \mathcal{F}$ the random variable $\chi(S)$ is distributed according to the binomial distribution. By Chernoff's large deviation inequality

$$\Pr[|\chi(S)| > \lambda\sqrt{|S|}] < 2\exp(-\lambda^2/2).$$

Let $\lambda = \sqrt{2\log(4|\mathcal{F}|)}$. Then the union bound implies

$$\begin{aligned}\Pr[\operatorname{disc}(\mathcal{F},\chi) > \lambda\sqrt{s}] &= \Pr[\exists S \in \mathcal{F},\ |\chi(S)| > \lambda\sqrt{|S|}] \le \sum_{S\in\mathcal{F}} \Pr[|S| > \lambda\sqrt{|S|}] \\ &\le \sum_{S\in\mathcal{F}} \Pr[|S| > \lambda\sqrt{s}] < |\mathcal{F}| \cdot 2\exp(-\lambda^2/2) = \frac{1}{2}. \qquad \square\end{aligned}$$

As many families of geometric origin have bounded VC-dimension, the following result, when taken jointly with Theorem 1 above, supplies a non-trivial upper bound for their geometric discrepancy:

Theorem 4. *Let $(X,\mathcal{F})$ be a set family of VC-dimension d on the ground set of size $|X| = n$. Then*

$$\operatorname{disc}(\mathcal{F}) \le n^{\frac{1}{2}-\frac{1}{2d}} \log^{c_d} n.$$

Let $S_1 \Delta S_2 = (S_1 \setminus S_2) \cup (S_1 \setminus S_2)$ be the symmetric difference of S_1 and S_2. A family $\mathcal{F}$ of sets is said to be δ-separated if for every pair of distinct sets $S_1, S_2 \in \mathcal{F}$ the size of their symmetric difference is $|S_1 \Delta S_2| \ge \delta$. The size of a δ-separated family need to be small in general. Indeed, it is not hard to derive from Chernoff's inequality that a random family of $2^{n/1000}$ subsets of an n-element set is $(n/1000)$-separated. For sets of bounded VC-dimension the situation is much different.

Lemma 5. *Let $(X,\mathcal{F})$ be a δ-separated set family of VC-dimension d on the ground set of size $|X| = n$. Then*

$$|\mathcal{F}| \le c_d (n/\delta)^d \log^d(n/\delta).$$

Proof. Let $r = n/\delta$. By problem #1 from the third example sheet the VC-dimension of the family $\mathcal{F}' = \{S_1 \Delta S_2 : S_1, S_2 \in \mathcal{F}\}$ is bounded solely in terms of d. Thus there is an $\frac{1}{r}$-net N for $\mathcal{F}'$ of size $|N| = c'_d r \log r$. Since for every $S_1, S_2 \in \mathcal{F}$, the set $S_1 \Delta S_2$ contains a point of N, that point of N is in precisely one of S_1 and S_2. In particular, $S_1 \cap N \ne S_2 \cap N$. Therefore,

$$|\mathcal{F}| \le |\mathcal{F}|_N| \le g(d,|N|) = O(|N|^d) = O(r^d \log^d r). \qquad \square$$

A *partial colouring* of a set X is a function $\chi\colon X \to \{-1,0,+1\}$. In addition to the familiar colours -1 and $+1$, the zero signifies an uncoloured element. The definition of $\chi(S) = \sum_{x\in S} \chi(x)$ remains unchanged, and is still called discrepancy of S with respect to χ. The advantage of partial colourings is that they can be used to build a complete colouring in stages. First, a partial colouring with small discrepancy is found. Then task becomes of finding a small-discrepancy colouring of the uncoloured elements. The procedure can

be repeated with several partial colourings combined together in a single total colouring. Thus, the partial colourings permit us to break the difficult task of finding a total colouring into several easier tasks of finding partial colourings.

Lemma 6. *Let X be an n-point set that is a common ground set for the set systems $(X, \mathcal{F})$ and $(X, \mathcal{M})$. Let $s = \max_{M \in \mathcal{M}} |M|$. Suppose*

$$\prod \big(2|F| + 1\big) \leq 2^{n/5-1}.$$

Then there exists a partial colouring $\chi \colon X \to \{-1, 0, +1\}$ that leaves at most $(9/10)n$ points uncoloured such that $\operatorname{disc}(\mathcal{F}, \chi) = 0$ and $\operatorname{disc}(\mathcal{M}, \chi) \leq \sqrt{2s \log(4|\mathcal{M}|)}$.

Proof. Consider the family $\mathcal{C}$ of all total colourings $\chi \colon X \to \{-1, +1\}$ for which $\operatorname{disc}(\mathcal{M}, \chi) \leq \sqrt{2s \log(4|\mathcal{M}|)}$. By theorem 3 the number of colourings in $\mathcal{C}$ is at least $\frac{1}{2} \cdot 2^{|X|} = 2^{n-1}$. For each $\chi \in \mathcal{C}$ consider the $|\mathcal{F}|$-dimensional vector $d(\chi) = (\chi(F))_{F \in \mathcal{F}}$. Since $|\chi(F)| \leq |F|$, the range of d is a set with at most

$$\prod \big(2|F| + 1\big) \leq 2^{n/5-1}$$

elements. By the pigeonhole principle, there is a d_0 such that $\mathcal{C}' = \{\chi \in \mathcal{C} : d(\chi) = d_0\}$ contains at least $2^{4n/5}$ colourings. Fix such a d_0, and an arbitrary $\chi_0 \in \mathcal{C}'$. There are at most

$$N = \sum_{i \leq n/10} \binom{n}{i}$$

colourings $\chi \in \mathcal{C}'$ that differ from χ_0 in fewer than $n/10$ positions. Since $N < 2^{4n/5} \leq |\mathcal{C}'|$ there is a $\chi_1 \in \mathcal{C}'$ that differs from χ_0 in more than $n/10$ positions. The colouring $\chi = \frac{1}{2}(\chi_0 - \chi_1)$ then fulfills the conclusion of the theorem. □

Proof of theorem 4. Let δ be a parameter to be chosen later. Let $\mathcal{F}' \subset \mathcal{F}$ be a maximal δ-separated subfamily of $\mathcal{F}$. By lemma 5 the size of $\mathcal{F}'$ is only $|\mathcal{F}'| \leq (n/\delta)^d \log^d(n/\delta)$. Let

$$\begin{aligned} \mathcal{M} = \{&F \setminus F' : F \in \mathcal{F},\ F' \in \mathcal{F}',\ |F \Delta F'| \leq \delta\} \\ &\cup \{F' \setminus F : F \in \mathcal{F},\ F' \in \mathcal{F}',\ |F \Delta F'| \leq \delta\}. \end{aligned}$$

Since $\mathcal{F}'$ is a *maximal* δ-separated family every $F \in \mathcal{F}$ can be expressed as $F = (F' \cup A) \setminus B$ where $F' \in \mathcal{F}'$ and $A, B \in \mathcal{M}$, with three sets A, B and $F' \setminus B = F \setminus A$ being disjoint. If $\delta = n^{1-1/d} \log^{c_d'} n$ for an appropriate c_d' then

$$\prod_{F \in \mathcal{F}'} (2|F| + 1) \leq (2n + 1)^{|\mathcal{F}'|} \leq 2^{n/5-1},$$

and by the preceding lemma, there is a partial colouring $\chi_1 \colon X \to \{-1, 0, +1\}$ of zero discrepancy on $\mathcal{F}$ and discrepancy at most $\sqrt{2\delta \log(4|\mathcal{M}|)}$ on $\mathcal{M}$. Since $|\mathcal{M}| \leq g(\text{VC-dim}(\mathcal{M}), n) = O(n^d)$, it follows that

$$\operatorname{disc}(\mathcal{M}, \chi_1) = O(\sqrt{\delta \log n}) = O(n^{\frac{1}{2} - \frac{1}{2d}} \log^{c_d} n).$$

Note that for every $F \in \mathcal{F}$ that is written as $F = (F' \cup A) \setminus B$ we have $\chi_1(F) = \chi_1(F') + \chi_1(A) - \chi_1(B)$. Thus

$$\mathrm{disc}(\mathcal{F}, \chi_1) \leq cn^{\frac{1}{2}-\frac{1}{2d}} \log^{c_d} n.$$

The colouring χ_1 leaves at most $(9/10)n$ points uncoloured. One can then find a partial colouring χ_2 that colours $(1/10)n$ of these remaining points and has discrepancy only

$$\mathrm{disc}(\mathcal{F}, \chi_2) \leq c(\tfrac{9}{10}n)^{\frac{1}{2}-\frac{1}{2d}} \log^{c_d} n.$$

leaving only $(9/10)^2 n$ points uncoloured. Repeating this process, yields a total colouring that colours every point of X and has discrepancy at most

$$\sum_i c(\tfrac{9^i}{10^i}n)^{\frac{1}{2}-\frac{1}{2d}} \log^{c_d} n,$$

as claimed. □