9. Discrepancy

- Discrepancy theory ~ study of "balanced representation" or "good [simultaneous] approximations" or "[ir]regularities of distributions"
- Geometric discrepancy: Distribute n points in the unit square so that each axis-parallel rectangle R contains approximately $\text{Area}(R) \cdot n$ points.
- Rounding problems: For a matrix M and a vector x, find an integer vector u so that $\|Mx - Mu\|_\infty$ is small.
- Combinatorial discrepancy: For a given family of sets (hypergraph) $\mathcal{A} \subseteq \mathcal{P}([n])$, find a two-colouring of $[n]$ so that all $A \in \mathcal{A}$ contain a balanced number of points of both colours.

- We shall focus on combinatorial discrepancy with occasional reformulations in terms of vector norms.

9.1 Low-discrepancy Colourings

- Let thus $\mathcal{A} \subseteq \mathcal{P}([n])$ be a set system. Represent a given two-colouring of $[n]$ as a map

$$\chi : [n] \to \{-1, +1\}.$$

For any $A \subseteq [n]$, denote $\chi(A) = \sum_{j \in A} \chi(j)$.

Then the discrepancy of $\mathcal{A}$ with respect to χ is

$$\text{disc}(\mathcal{A}, \chi) = \max_{A \in \mathcal{A}} |\chi(A)|$$

and the discrepancy of $\mathcal{A}$ is

$$\text{disc}(\mathcal{A}) = \min_{\chi} \text{disc}(\mathcal{A}, \chi).$$

- A reformulation of this notion in terms of vector norms is as follows: Let $\mathcal{A} = \{A_1, \ldots, A_m\}$, and let $M = (m_{ij})$ be the incidence matrix:

$$m_{ij} = \begin{cases} 1 & \text{if } j \in A_i, \\ 0 & \text{if } j \notin A_i. \end{cases}$$

If a colouring $\chi : [n] \to \{-1, +1\}$ is now represented as a vector $u = (\chi(1), \ldots, \chi(n)) \in \{-1, +1\}^n$, then

$$Mu = (\chi(A_1), \ldots, \chi(A_m)),$$

and so

$$\operatorname{disc}(\mathcal{A}) = \min_{u \in \{+1,-1\}^n} |Mu|_\infty$$

Alternately, if $v_j \in \{0,1\}^m$ denotes the j^{th} column vector of M, then

$$\operatorname{disc}(\mathcal{A}) = \min_{\varepsilon \in \{\pm 1\}^n} \| \varepsilon_1 v_1 + \cdots + \varepsilon_n v_n \|_\infty .$$

- <u>Theorem</u> 9.1 Let $\mathcal{A} \subseteq \mathcal{P}([n])$ consist of m sets. Then

$$\operatorname{disc}(\mathcal{A}) \leq \sqrt{2n \ln(2m)}.$$

<u>Proof</u> Let $\chi : [n] \to \{-1, +1\}$ be random. For each $A \in \mathcal{A}$, let X_A be an indicator random variable for the event

"$|\chi(A)| > \alpha$", where $\alpha = \sqrt{2n \ln(2m)}$.

If $|A| = k$, then $\chi(A)$ is a sum of k independent ± 1 -valued random variables, as in Thm 8.1 (Chernoff bound). Thus

$$\Pr(|\chi(A)| > \alpha) < 2e^{-\alpha^2/2k} \leq 2e^{-\alpha^2/2n} = \frac{1}{m}.$$

Let then X be the number of $A \in \mathcal{A}$ with $|\chi(A)| > a$, so that

$$X = \sum_{A \in \mathcal{A}} X_A .$$

By linearity of expectation,

$$E[X] = \sum_{A \in \mathcal{A}} E[X_A] < m \cdot \frac{1}{m} = 1 .$$

Hence for some colouring χ, $X(\chi) < 1$, i.e. $X(\chi) = 0$. For this colouring, then,

$$\chi(A) \le a \quad \text{for all } A \in \mathcal{A},$$

and so $\operatorname{disc}(\mathcal{A}) \le a = \sqrt{2n \ln(2m)}$. □

- Thm 9.1 implies, e.g., that if a set system $\mathcal{A}$ on n points contains $m = 2^{\varepsilon n}$ sets, $0 < \varepsilon < 1$, then there is a two-colouring of the points so that no $A \in \mathcal{A}$ has colour imbalance greater than

$$\sqrt{2n \ln 2^{\varepsilon n}} = \sqrt{2n (\varepsilon \ln 2 \cdot n)} \approx 1{,}18 \sqrt{\varepsilon}\, n .$$

- In the case $m = n$, Thm 9.1 implies that $\operatorname{disc}(\mathcal{A}) \le \sqrt{2n \ln n}$. There is a nontrivial strengthening of this result:

Theorem 9.2 (Spencer 1985). Let $\mathcal{A} \subseteq \mathcal{P}([n])$ consist of n sets. Then

$$\operatorname{disc}(\mathcal{A}) \le 6\sqrt{n} .$$

- Since for a fixed $A \in \mathcal{A}$ and random $\chi : [n] \to \{-1, +1\}$, the variable $\chi(A)$ has zero mean and standard deviation $\le \sqrt{n}$, the result tells that in this case "six standard deviations suffice" to guarantee a colouring which is balanced for all $A \in \mathcal{A}$. — In fact, one can show that the size of the base set is nonessential, so that for any family $\mathcal{A}$ of m sets,

$$\operatorname{disc}(\mathcal{A}) \le 12\sqrt{m} .$$

9.2 High-discrepancy Systems

- Let us then show that there exist set systems $\mathcal{A}$ with m sets such that $\mathrm{disc}(\mathcal{A}) = \Omega(\sqrt{m})$.

- A <u>Hadamard matrix of order</u> n is a matrix $H \in \{-1,+1\}^{n \times n}$ all of whose row (and also column) vectors are mutually orthogonal. (Thus $H^T H = H H^T = nI$.)

  Hadamard matrices of order n are known to exist e.g. when $n = 2^k$, $k = 1, 2, \ldots$ and are conjectured to exist whenever $n = 4k$, $k = 1, 2, \ldots$

- Let H be a Hadamard matrix of order n, with row vectors $r_1, \ldots, r_n$ and column vectors $c_1, \ldots, c_n$. Let $u \in (u_1, \ldots, u_n) \in \{-1,+1\}^n$ be arbitrary. Then

  $$Hu = (L_1, \ldots, L_n), \text{ where each } L_i = r_i u, \; i = 1, \ldots, n.$$

  Thus

  $$L_1^2 + \cdots + L_n^2 = \|Hu\|^2 = (u^T H^T)(Hu) = n\|u\|^2 = n^2,$$

  (Euclidean norm; orthogonality)

  Hence some $L_i^2 \geq n$, and $\|Hu\|_\infty = \max_i |L_i| \geq \sqrt{n}$.

- Now let us reformulate this in terms of set systems. For a given Hadamard matrix $H \in \{-1,+1\}^{n \times n}$, consider the matrix

  $$M = \tfrac{1}{2}(H + J) \in \{0,1\}^{n \times n} \quad (J = \text{all-ones matrix}),$$

  and interpret it as the incidence matrix of a set system $\mathcal{A} = \{A_1, \ldots, A_n\} \subseteq \mathcal{P}([n])$. As before, any colouring $\chi : [n] \to \{-1,+1\}$ can be represented as a vector $u = (\chi(1), \ldots, \chi(n))$, in which case

  $$\mathrm{disc}(\mathcal{A}, \chi) = \|Mu\|_\infty.$$

- For a given $u \in \{-1,+1\}^n$, denote $\lambda = u_1 + \cdots + u_n$, so that $Ju = (\lambda, \ldots, \lambda)$. Then

$$Mu = \tfrac{1}{2}(Hu + Ju) = \tfrac{1}{2}(L_1 + \lambda, \ldots, L_n + \lambda).$$

- We may assume H to be normalised so that its first row and column are all-1's. Then

$$L_1 + \cdots + L_n = \sum_{i=1}^n r_i u = \sum_{i,j=1}^n h_{ij} u_j$$
$$= \sum_{j=1}^n u_j \sum_{i=1}^n h_{ij} = u_1 n = \pm n$$
(orthogonality to 1^n)

- Thus

$$\|(H+J)u\|^2 = \sum_{i=1}^n (L_i + \lambda)^2 = \sum_{i=1}^n (L_i^2 + 2\lambda L_i + \lambda^2)$$
$$= n^2 \pm 2n\lambda + n\lambda^2.$$

Assuming n is even (H-matrices do not even exist for odd $n > 1$), one can deduce from this that λ is an even integer. The quadratic above has a minimum at $\lambda = \mp 1$, and so for even integer λ the minima are at $\lambda = 0, \mp 2$, showing that

$$\|(H+J)u\|^2 \geq n^2.$$

Thus some coordinate of $(H+J)u$ must be $\sqrt{n}$, showing that

$$\|Mu\|_\infty = \tfrac{1}{2}\|(H+J)u\|_\infty \geq \frac{\sqrt{n}}{2}.$$

- <u>Theorem</u> 9.3 If a Hadamard matrix of order $n > 1$ exists, then there is a set system $\mathcal{A} \subseteq \mathcal{P}([n])$ containing n sets, such that

$$\operatorname{disc}(\mathcal{A}) \geq \frac{\sqrt{n}}{2}. \quad \square$$

9.3 The Beck-Fiala Theorem

- For a set system $\mathcal{A}$, denote by $\deg(\mathcal{A})$ the maximal number of sets in $\mathcal{A}$ containing any given point. (I.e. the "max vertex degree" of $\mathcal{A}$ as a hypergraph.)

- **Theorem** 9.4 (Beck & Fiala 1981) Let $\mathcal{A}$ be a finite family of finite sets, with $\deg(\mathcal{A}) \leq t$. Then

$$\operatorname{disc}(\mathcal{A}) \leq 2t - 1.$$

Proof Say $\mathcal{A} = \{A_1, \dots, A_m\}$, with each $A_i \subseteq [n] = \{1, \dots, n\}$. Associate to each $j \in [n]$ a "greyscale" variable $x_j \in [-1, 1]$ whose value initially is $x_j = 0$ and at the end of the proof $x_j = \pm 1$. (The eventual colouring is determined as $\chi(j) = x_j$.)

Each set A_i has associated tentative discrepancy value $S_i = \sum_{j \in A_i} x_j$. Initially of course each $S_i = 0$, and we maintain the invariant that throughout the proof $S_i < 2t$.

At any time, point j is **fixed** if $x_j = \pm 1$, otherwise **floating**. A set A_i is **safe** if it contains $\leq t$ floating points, otherwise **active**.

Note that as each point is contained in $\leq t$ sets, and active sets contain $> t$ floating points, there are always fewer active sets than floating points.

We maintain the condition that at all times $S_i = 0$ for all active sets A_i. This is trivially true initially, so assume that it holds at some stage. Consider x_j a variable for each floating j and a constant for each fixed j. The condition

"$S_i = 0$ for all active sets i"

is then an **underdetermined linear system** in the floating variables x_j.

Hence there is a line, parametrized as

$$x'_j = x_j + \lambda y_j \quad , \text{ for all } j \text{ floating}$$

along which all the active sets retain value zero.

Let λ be the smallest value making some x'_k become ± 1. Replace each x_j by the corresponding x'_j and repeat the process.

Note that at each iteration of the above process, the safe sets remain safe, and at least one formerly floating variable becomes fixed. (Thus possibly making some formerly active sets safe.)

Now consider any discrepancy value S_i associated to a set A_i. Initially $S_i = 0$, and this stays true as long as A_i stays active. However, at the moment A_i becomes safe, only $\leq t$ of its points remain floating, i.e. with values $x_j \in (-1, +1)$. Since each of these can change by less that 2 before the process terminates in the final colouring χ, it is guaranteed that

$$\chi(A_i) = \sum_{j \in A_i} x_j < 0 + t \cdot 2 = 2t.$$

Since the final discrepancy values are integers, this means that

$$\chi(A_i) \leq 2t - 1. \quad \square$$