# Discrepancy theory
# Or: How much balance is possible?

Thomas Rothvoß

**Abstract**

Discrepancy theory is a subfield of combinatorics in which one asks the following question: given a finite set system $S_1, \dots, S_m \subseteq \{1, \dots, n\}$; color the elements $\{1, \dots, n\}$ with two colors, say *red* and *blue*. What is the difference between red and blue elements in the most unbalanced set for the best coloring?

The two main results are the Beck-Fiala Theorem and Spencer's Theorem, which both have very elegant proofs as we will see in this lecture.

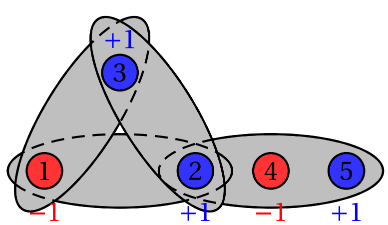
## 1 Preliminaries

For this lecture, we consider a finite *ground set* of elements $\{1, \dots, n\}$ and a family of sets $S_1, \dots, S_m \subseteq \{1, \dots, n\}$. We abbreviate $\mathscr{S} := \{S_1, \dots, S_m\}$. A *coloring* is a map $\chi : \{1, \dots, n\} \to \{-1, +1\}$ and the *discrepancy* of a coloring is the maximum inbalance $\max_{i=1,\dots,m} |\chi(S_i)|$ where we write $\chi(S_i) := \sum_{j \in S_i} \chi(j)$. The discrepancy of the whole set system is the discrepancy of the best coloring, i.e.

$$\mathrm{disc}(\mathscr{S}) = \min_{\chi:\{1,\dots,n\} \to \{\pm 1\}} \max_{S \in \mathscr{S}} |\chi(S)|$$

Of course, this bound will heavily depend on the structure of the set system as well as on the number $n$ of elements and the number $m$ of sets.

Applications of discrepancy theory can be found e.g. in computer science. For example in the set system $\mathscr{S} = \{\{1,2\}, \{1,3\}, \{2,3\}, \{2,4,5\}\}$ the best coloring is depicted below and the discrepancy is 2.



## 2 Random colorings

A simple way to obtain fairly balanced colorings is to choose a *random coloring*. More precisely, we choose each color $\chi(j)$ independently from $\{-1,1\}$ with $\Pr[\chi(j)=1]=\Pr[\chi(j)=-1]=\frac{1}{2}$. Then for each set we expect that $\chi(S)$ is closely concentrated around the mean – which is 0. Recall the *Chernov bound*:

**Theorem 1** (Chernov bound)**.** *Take independent random variables $X_1,\dots,X_k$ with $\Pr[X_i=1]=\Pr[X_i=-1]=\frac{1}{2}$. Then for any $\lambda\geq 0$, $\Pr[|\sum_{i=1}^k X_k|>\lambda\sqrt{k}]<2e^{-\lambda^2/2}$.*

This provides a first, simple way to obtain good colorings:

**Theorem 2.** *Take a random coloring $\chi:\{1,\dots,n\}\to\{\pm 1\}$. With prob at least $\frac{1}{2}$,*

$$|\chi(S_i)|\leq\sqrt{2|S_i|\cdot\ln(4m)}\quad\forall i=1,\dots,m$$

*Proof.* First, fix a set $S_i$. Then by Chernov bound

$$\Pr\Big[\Big|\sum_{j\in S_i}\chi(j)\Big|>\underbrace{\sqrt{2\cdot\ln(4m)}}_{:=\lambda}\cdot\sqrt{|S_i|}\Big]\leq 2e^{-(\sqrt{2\cdot\ln(4m)})^2/2}\leq\frac{1}{2m}$$

Then

$$\Pr[\exists i\in\{1,\dots,m\}:|\chi(S_i)|>\sqrt{2|S_i|\cdot\ln(4m)}]\overset{\text{Union bound}}{\leq} m\cdot\frac{1}{2m}=\frac{1}{2}$$

$\square$

## 3 The Beck Fiala Theorem

In many applications, the set system is *sparse*, that means it contains a huge number of sets and elements, but the number of incidences is far smaller than the worst case $n\cdot m$. In this case, the Beck Fiala Theorem can provide fairly good colorings.

**Theorem 3** (Beck-Fiala Theorem '81 [2])**.** *Let $\mathcal{S}$ be any set system where no element is in more than $t$ sets. Then disc($\mathcal{S}$) $<2t$.*

*Proof.* We introduce variables $y_j\in[-1,1]$. Consider the following system

$$\begin{aligned}\overbrace{\sum_{j\in S}y_j}^{=Ay,A\in\{0,1\}^{m\times n}} &= 0 \quad \forall S\in\mathcal{S} \qquad (1)\\ -1\leq y_j &\leq 1 \quad \forall j\in[n]\end{aligned}$$

We make 2 claims.

**Claim 4.** *If $n > m$, then there is a solution $y$ for* (1) *with $y_j \in \{-1, 1\}$ for at least one $j$.*

*Proof of claim.* Take $y \in \ker(A)$. Scale $y$ s.t. $\|y\|_\infty = 1$. $\Diamond$

**Claim 5.** *Suppose we delete all sets with $|S| \leq t$. Then $n > m$.*

*Proof of claim.* Suppose $|S_i| > t \; \forall i$. Then

$$m \cdot t < \#\text{ ones in } A \leq n \cdot t$$

because no element appears in more than $t$ sets. $\Diamond$

This suggests the following method to find a coloring:

(1) Set $\chi(j) :=$ undef

(2) WHILE not yet all elements defined DO

(3) Compute a solution $y$ to

$$\begin{aligned} \sum_{j \in S} y_j &= 0 \quad \forall S \in \mathscr{S} : S \text{ contains } > t \text{ undefined elements} \\ y_j &= \chi(j) \quad \text{if } \chi(j) \text{ defined} \\ -1 \leq y_j &\leq 1 \end{aligned}$$

with maximal number of $j$'s with $y_j \in \{\pm 1\}$.

(4) IF $y_j \in \{\pm 1\}$ THEN $\chi(j) := y_j$

Each time the algorithm runs (3), we have the invariant #{undefined elements} > #{sets with $> t$ undefined elements}, thus the solution $y$ is never unique and we can move choose the $y$ such that it satisfies one more inequality $-1 \leq y_j \leq 1$ with equality.

Now, let $\chi \in \{-1, 1\}^n$ be the final coloring. Now consider, how the discrepancy $\sum_{j \in S} y_j$ of set $S$ behaves. Until the moment in which the constraint for $S$ was removed, we had $y(S) = 0$. But only $t$ elements where fractional at that point. In the worst case, they can switch from $-0.999..$ to $+1$, but in any case at the very end $|y(S)| < 2t$. $\square$

In fact, a much stronger bound is conjectured:

**Conjecture 6** (Beck-Fiala)**.** *For any $t$ and set system with no element in more than $t$ sets, one has disc$(\mathscr{S}) \leq O(\sqrt{t})$.*

But not even $O(t^{0.999})$ is known!!

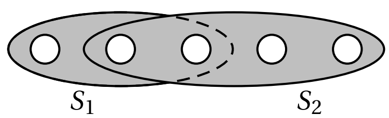
## 4 Spencer's Theorem

Now we come to the best possible bound in the case of arbitrary *dense* set systems, which is a celebrated result of Joel Spencer.

**Theorem 7** (Spencer '85 [4])**.** *For any set system with $m \geq n$ sets on $n$ elements, one has disc($\mathcal{S}$) $\leq O(\sqrt{n \cdot \ln(\frac{2m}{n})})$. In particular, if $m \leq O(n)$, then disc($\mathcal{S}$) $\leq O(\sqrt{n})$.*

The proof idea is that we have an enourmous number $2^n$ of potential colorings. If the number of sets is not too large, by a simple counting argument, we must have some colorings $\chi_A, \chi_B$ such that $\chi_A(S_i) \approx \chi_B(S_i)$ for all sets. Even if both colorings are bad, their *difference* $\frac{1}{2}(\chi_A - \chi_B)$ is not.

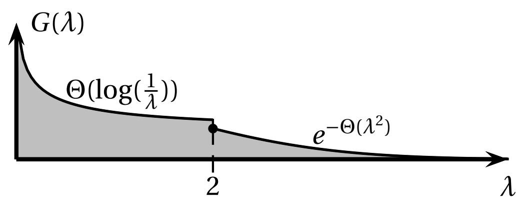
| | | | | | |
|---|---|---|---|---|---|
| coloring A: | $+1$ | $-1$ | $+1$ | $+1$ | $+1$ |
| coloring B: | $-1$ | $+1$ | $+1$ | $-1$ | $+1$ |
| difference | 1 | $-1$ | 0 | 1 | 0 |




This can be formalized in the following very useful lemma:

**Lemma 8** (Partial coloring lemma)**.** *Define*

$$G(\lambda) := \begin{cases} 10 \cdot e^{-\lambda^2/10} & \lambda \geq 2 \\ 10 \cdot \log(\frac{10}{\lambda}) & \lambda < 2 \end{cases}$$




*and choose parameters $\Delta_1, \ldots, \Delta_m > 0$ such that*

$$\sum_{i=1}^{m} G\left(\frac{\Delta_i}{\sqrt{|S_i|}}\right) \leq \frac{n}{10} \tag{2}$$

*Then there is a partial coloring $\chi : [n] \to \{0, \pm 1\}$ with $|\chi(S_i)| \leq \Delta_i$ for all $i = 1, \ldots, m$ and $|supp(\chi)| \geq \frac{n}{10}$.*[1]

Let's first check, why this quickly implies Spencer's Theorem:

[1] Here $supp(\chi) := \{j \in [n] \mid \chi(j) \neq 0\}$ is the *support* of $\chi$.

*Proof of Spencer's Thm.* We first claim that we can find a partial coloring satisfying the claimed bound of $\Delta := C\sqrt{n \cdot \log \frac{2m}{n}}$. Then we bound (2) by

$$\sum_{i=1}^{m} G\left(C\sqrt{\log \frac{2m}{n}}\right) \leq m \cdot 10e^{-(C\sqrt{\log \frac{2m}{n}})^2/10} \leq \frac{n}{10}$$

for $C > 0$ large enough, hence there is at least a partial coloring $\chi$ with $|\chi(S)| \leq O\left(\sqrt{n \cdot \ln(\frac{2m}{n})}\right)$. We color the elements in $\mathrm{supp}(\chi)$ and remove them from the set system. We iterate this until all elements are colored. Then

$$\mathrm{disc}(\mathscr{S}) \leq \sum_{i \geq 0} O\left(\sqrt{0.9^i n \cdot \ln(\frac{2m}{0.9^i n})}\right) = O\left(\sqrt{n \cdot \ln(\frac{2m}{n})}\right)$$

In other words: the discrepancy bound is decreasing geometrically in $i$, hence the error is dominated by the first term. $\square$

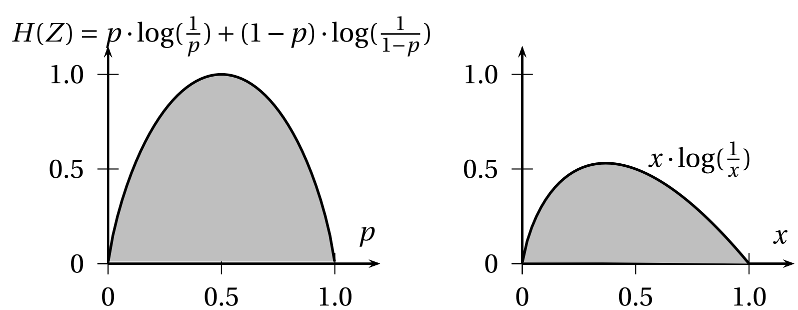
In the remaining lecture, we proof the partial coloring lemma.

### 4.1 Entropy

The *entropy* of an arbitrary discrete random variable $Z$ (the domain does not matter – it could be $Z \in \mathbb{Z}$ or $Z \in \mathbb{Z}^n$) is defined as

$$H(Z) = \sum_x \Pr[Z = x] \cdot \log_2\left(\frac{1}{\Pr[Z = x]}\right)$$

Here the sum runs over all values that $Z$ can attain. Imagine that a data source generates a string of $n$ symbols according to distribution $Z$. Then intuitively, an optimum compression needs asymptotically for $n \to \infty$ an expected number of $n \cdot H(Z)$ many bits to encode the string. If $Z$ attains only two values, say $\Pr[Z = a] = p$ and $\Pr[Z = b] = 1 - p$ then the entropy looks as follows:




Two useful facts on entropy are:

- *Uniform distribution maximizes entropy:* If $Z$ attains $k$ distinct values, then $H(Z)$ is maximal if $Z$ is the uniform distribution. In that case $H(Z) = \log_2(k)$. Conversely, if $H(Z) \leq \delta$, then there must be at least one event $x$ with $\Pr[Z = x] \geq (\frac{1}{2})^{\delta}$.

- *Subadditivity:* If $Z, Z'$ are random variables and $f$ is any function, then $H(f(Z, Z')) \leq H(Z) + H(Z')$.

First of all, let us show where the "magic" function $G$ is coming from. Recall that $\lceil \cdot \rfloor$ rounds to be nearest integer, i.e. $\lceil 0.7 \rfloor = 1$ and $\lceil 0.2 \rfloor = 0$.

**Lemma 9.** *Suppose $\chi$ is a random coloring and $\Delta = \lambda\sqrt{|S|}$, $\lambda > 0$. Then*

$$H\left(\left\lceil \frac{\chi(S)}{2\Delta} \right\rfloor\right) \leq G(\lambda).$$

*Proof.* We show only the case $\lambda \geq 2$ and save the case $\lambda < 2$ for the exercises. We are also loose with the constants. Define indicator variables

$$X_j = \begin{cases} 1 & \left\lceil \frac{\chi(S)}{2\Delta} \right\rfloor = j \\ 0 & \text{otherwise} \end{cases}$$

Note that for $\lambda \gg 2$, we have $\Pr[X_0 = 1] \approx 1$ and $\Pr[X_j = 1] \approx 0$, thus the entropy of those random variables must be small.

We can use $H(X_j) \leq 2\Pr[X_j = 1] \cdot \log \frac{1}{\Pr[X_j = 1]}$ as long as $\Pr[X_j = 1] \leq \frac{1}{2}$. For $|j| > 0$ we can use the Chernov bound

$$\begin{aligned} & \Pr[X_j = 1] \leq \Pr[\chi(S) \geq (2j-1)\lambda\sqrt{|S|}] \leq e^{-\Omega((\lambda j)^2)} \\ \Rightarrow \quad & H(X_j) \leq 2\Pr[X_j = 1] \cdot \log\left(\frac{1}{\Pr[X_j = 1]}\right) \leq O(1) \cdot e^{-\Omega((\lambda j)^2)} \cdot (\lambda j)^2 \end{aligned}$$

Moreover

$$\Pr[X_0 = 1] \geq 1 - e^{-\lambda^2/4} \Rightarrow H(X_0) \leq 2\Pr[X_j = 0] \cdot \log\left(\frac{1}{\Pr[X_j = 0]}\right) \leq O(1) \cdot e^{-\lambda^2/4} \cdot \lambda^2$$

Then

$$H\left(\left\lceil \frac{\chi(S)}{2\lambda\Delta} \right\rfloor\right) = H((X_j)_{j \in \mathbb{Z}}) \overset{\text{subadditivity}}{\leq} \sum_{j \in \mathbb{Z}} H(X_j) \leq O(1) \cdot e^{-\Omega(\lambda^2)}.$$

$\square$

*Proof of partial coloring lemma.* Denote

$$\underbrace{Z}_{\in \mathbb{Z}^m} := Z(\chi) := \left( \left\lceil \frac{\chi(S_1)}{2\Delta_1} \right\rfloor, \dots, \left\lceil \frac{\chi(S_m)}{2\Delta_m} \right\rfloor \right)$$

Then

$$H(Z) \overset{\text{subaddivity}}{\leq} \sum_{i=1}^{m} H(Z_i) \overset{\text{Lemma 9}}{\leq} \sum_{i=1}^{m} G\left( \frac{\Delta_i}{\sqrt{|S_i|}} \right) \overset{\text{assumption}}{\leq} \frac{n}{10}$$

Thus there is a vector $b \in \mathbb{Z}^n$ s.t. $\Pr[Z = b] \geq (\frac{1}{2})^{n/10}$. In other words, there are $2^n \cdot (\frac{1}{2})^{n/10}$ many colorings $\chi$ s.t.

$$Z(\chi) = b \quad \Longrightarrow \quad \left\lceil \frac{\chi(S_i)}{2\Delta_i} \right\rfloor = b_i \ \forall i \in [m] \quad \Longrightarrow \quad |\chi(S_i) - 2\Delta_i b_i| \leq \Delta_i \ \forall i \in [m]$$

In other words: all those colorings might be very bad, but at least they are very similar. We use the following fact (and defer its proof to the exercises):

> **Fact:** For any $X \subseteq \{0,1\}^n$ of size $|X| \geq 2^{0.9n}$, there are $x, y \in X$ with $\|x - y\|_1 \geq n/10$.

Now, take two colorings $\chi_A, \chi_B \in \{\pm 1\}^m$ with $Z(\chi_A) = Z(\chi_B) = b$ that differ in at least $\frac{n}{10}$ entries and define

$$\chi(j) := \frac{1}{2}(\chi_A(j) - \chi_B(j)) \in \{-1, 0, +1\}$$

Finally[2]

$$|\chi(S_i)| = \frac{1}{2}(\underbrace{|\chi_A(S_i) - \chi_B(S_i)|}_{\leq 2\Delta_i}) \leq \Delta_i.$$

$\square$

## 5 Further material

A very readable source for more details on discrepancy theory is Chapter 4 in the book of Matousek [3].

Observe that the Beck-Fiala Theorem uses simple linear algebra and gives immediately a polynomial time algorithm. On the other hand, the Entropy method

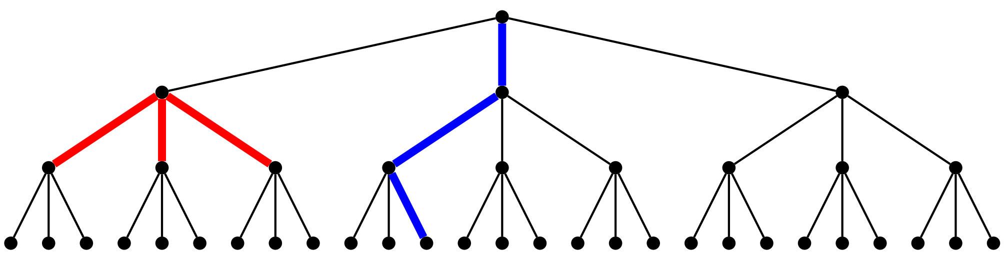
[2] Note that $\frac{1}{2}\left(\begin{pmatrix}1\\1\\-1\\-1\end{pmatrix} - \begin{pmatrix}1\\-1\\1\\-1\end{pmatrix}\right) = \begin{pmatrix}0\\1\\-1\\0\end{pmatrix}$

is based on the pigeonhole principle — with exponentially many pigeons and pigeonholes. But Lovett and Meka provided a simple and elegant algorithm based on random walks that can find the coloring provided by Spencer's Theorem (this simplifies a more complex algorithm of Bansal [1]).

## Exercises

**Exercise (Hypergraph splitting)**
Let $G = (V, E)$ be a 3-uniform, 6-regular hypergraph (i.e. $V = \{1, \dots, n\}$ is a finite set of vertices and $E = \{e_1, \dots, e_m\}$ is a finite set of hyperedges with $e_i \subseteq V$ and $|e_i| = 3$. Moreover every node is contained in exactly 6 hyperedges, i.e. $\forall j \in V : |\{i \in [m] : j \in e_i\}| = 6$). Show that one can partition the set of hyperedges into $E = E_1 \dot\cup E_2$ such that $E_1$ and $E_2$ both still cover all the nodes (i.e. $\bigcup_{e \in E_1} e = \bigcup_{e \in E_2} e = V$).

**Exercise ($k$ary trees)**
Let $k \in \mathbb{N}$ with $k \geq 2$. Consider a $k$-ary tree of depth $k$ (below, you can find one for $k = 3$).



We consider all its edges $E$ as elements (i.e. $n := |E| = k + k^2 + \ldots + k^k$) and we define two set systems

$$\begin{aligned} \mathcal{S}_1 &:= \{S \subseteq E \mid S \text{ is a path from the root to a leaf}\} \\ \mathcal{S}_2 &:= \{\text{outgoing edges of } v \mid v \text{ is interiour node}\} \end{aligned}$$

(in other words, $\mathcal{S}_2$ is a partition of the edge set; one set in $\mathcal{S}_1$ is drawn in bold-blue, one set in $\mathcal{S}_2$ is drawn in bold-red). Show the following:

i) $\text{disc}(\mathcal{S}_1) \leq 1$ and $\text{disc}(\mathcal{S}_2) \leq 1$

ii) $\text{disc}(\mathcal{S}_1 \cup \mathcal{S}_2) = k$

iii) There is a partial coloring $\chi$ with $|\text{supp}(\chi)| \geq \Omega(n)$ such that $|\chi(S)| \leq O(1)$ for all $S \in \mathcal{S}_1 \cup \mathcal{S}_2$.

**Exercise (The Beck Fiala setting)**
Consider a set system $\mathcal{S} = \{S_1, \dots, S_m\}$ with $n$ elements and suppose that every element is in at most $t$ sets and each set has size $|S_i| \leq t$. First show that there is

a partial coloring $\chi : [n] \to \{0, \pm 1\}$ with $|\text{supp}(\chi)| \geq \frac{n}{10}$ and $|\chi(S)| \leq O(\sqrt{t})$ for each $S \in \mathscr{S}$. Then conclude that $\text{disc}(\mathscr{S}) \leq O(\sqrt{t} \cdot \log n)$.
**Hint:** If you have difficulties in getting the bound right, suppose that the sets all have the same size.

**Exercise (Many elements and few sets)**
Suppose $S_1, \ldots, S_m$ is a set system over $n$ elements with $n \geq 1000m \cdot \log(n)$. Show that there is a partial coloring with $|\chi(S_i)| = 0$ for all $i = 1, \ldots, m$ and $|\text{supp}(\chi)| \geq \frac{n}{10}$.

**Exercise (Missing case of Lemma 9)**
Let $S \subseteq [n]$ be a set and let $\chi : [n] \to \{\pm 1\}$ be a random coloring. Let $k \in \mathbb{Z}_{\geq 2}$ and $\Delta := \frac{\sqrt{|S|}}{k}$. Show that $H\left(\left\lceil \frac{\chi(S)}{2\Delta} \right\rfloor\right) \leq c \cdot \log(k)$ for a large enough constant $c > 0$.
**Hint:** Write $\left\lceil \frac{\chi(S)}{2\Delta} \right\rfloor = \left\lceil \frac{\chi(S)}{2\sqrt{|S|}} \cdot k \right\rfloor = \left\lceil \frac{\chi(S)}{2\sqrt{|S|}} \right\rfloor \cdot k + f(\chi)$ for some function $f(\chi) \in \{-k, \ldots, k\}$.

# References


[1] N. Bansal. Constructive algorithms for discrepancy minimization. *CoRR*, abs/1002.2259, 2010. informal publication.

[2] J. Beck and T. Fiala. "Integer-making" theorems. *Discrete Appl. Math.*, 3(1):1–8, 1981.

[3] J. Matoušek. *Geometric discrepancy*, volume 18 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 1999. An illustrated guide.

[4] J. Spencer. Six standard deviations suffice. *Transactions of the American Mathematical Society*, 289(2):679–706, 1985.