

# Complexity and Characterization of Set Splitting



Peter Bernstein  
Tufts University

Cashous Bortner  
University of Nebraska

Samuel Coskey  
Boise State University

Shuni Li  
Macalester College

Connor Simpson  
Cornell University

## Discrepancy theory

Combinatorial discrepancy theory is a well-studied area of mathematics with applications to computational geometry, machine learning, probabilistic algorithm design, and other fields concerned with the regularity of distributions.

- Let  $X$  be a finite set and let  $\mathcal{B} = \{B_1, B_2, \dots, B_n\} \subseteq 2^X$  be a collection of subsets. The **discrepancy of  $\mathcal{B}$**  is

$$\text{disc}(\mathcal{B}) := \min_{S \subseteq X} \max_{B_i \in \mathcal{B}} \left| |B_i \cap S| - |B_i \setminus S| \right|.$$

- In a celebrated 1985 result, Joel Spencer gave a tight upper bound of

$$\text{disc}(\mathcal{B}) \leq K\sqrt{n}$$

where  $K$  is an absolute constant [5]. He later conjectured that no efficient algorithm exists to find a set  $S$  witnessing that discrepancy is within his bound [1].

- In 2010, Bansal and others disproved the conjecture by giving an efficient algorithm to find such an  $S$  [2].
- However,  $\text{disc}(\mathcal{B})$  can be much smaller than Spencer's bound, so Bansal's work prompts the following question:

Is it efficient to determine whether  $\text{disc}(\mathcal{B}) \leq 1$  and to find a witness  $S$  when this is the case?

Our work provides an answer.

## Splittability

We use  $[x]$  to denote the nearest integer to  $x$ , with free rounding if  $x$  is an odd multiple of  $\frac{1}{2}$ . Let  $\mathcal{B} = \{B_1, \dots, B_n\} \subseteq 2^X$  be a collection of subsets of a set  $X$ , and fix  $0 < p < 1$ .

### $p$ -Splittable

$\mathcal{B}$  is  **$p$ -splittable** if there exists a  $S \subseteq X$  such that for each  $B_i \in \mathcal{B}$ ,

$$|B_i \cap S| = [p|B_i|].$$

Note that when  $p = \frac{1}{2}$ , being  $p$ -splittable is equivalent to having  $\text{disc}(\mathcal{B}) \leq 1$ .

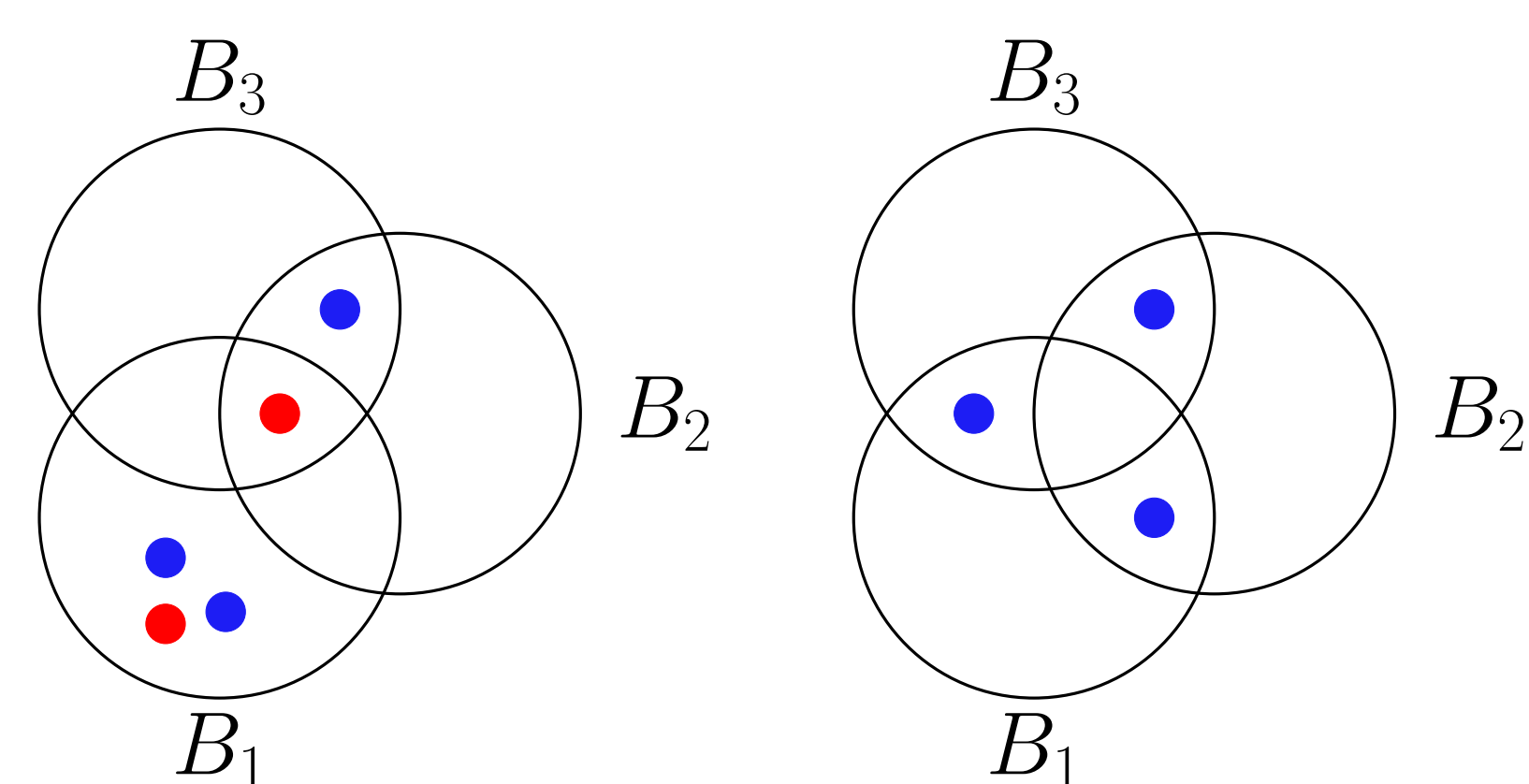


Fig. Splittable and unsplittable collections for  $p = 1/2$

## Complexity of $p$ -splitting

### Main Theorem

Determining whether a collection is  $p$ -splittable is NP-Complete for any  $0 < p < 1$ .

### Selected proof techniques:

Here, we outline some parts of our reduction from the NP-complete problem ZERO-ONE EQUATIONS (ZOE) [4]. ZOE is stated as follows: Given a 0,1-matrix  $A$ , does there exist a 0,1-vector  $\vec{y}$  such that  $A\vec{y} = \vec{1}$ , where  $\vec{1}$  is the vector of ones?

- We can encode  $\mathcal{B} = \{B_1, \dots, B_n\} \subseteq 2^X$  in the form of a 0,1-matrix  $M$ , which has a 1 in its  $(i, j)$  position precisely when element  $j$  of the collection is in  $B_i$ .

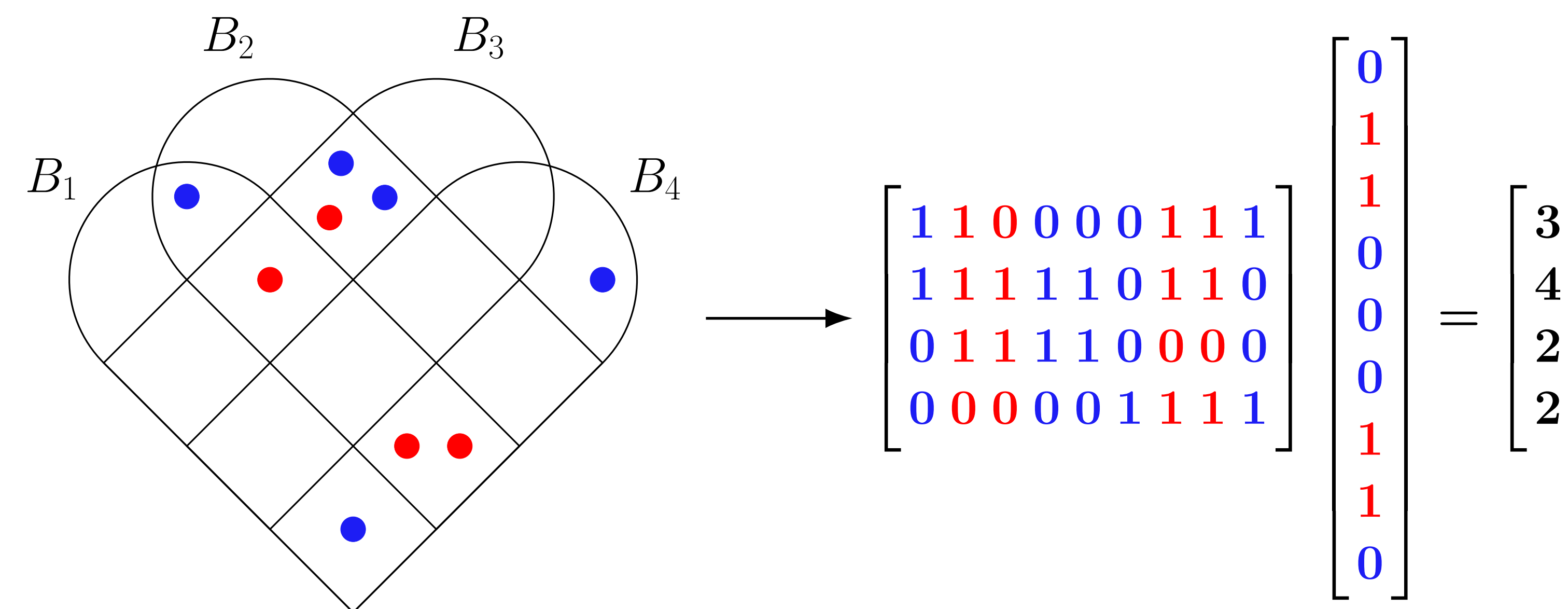


Fig. A split collection and its corresponding matrix equation for  $p = 1/2$

- To represent splitting  $\mathcal{B}$ , we apply  $M$  to a 0,1-vector  $\vec{x}$  encoding a potential solution  $S \subseteq X$  to the splitting problem. The set  $S$  corresponding to  $\vec{x}$  is a valid solution if the  $i$ th entry of  $M\vec{x}$  is equal to  $[p|B_i|]$ .
- We make use of this encoding by applying a polynomial-time construction that turns an arbitrary input to ZOE into a  $p$ -splitting problem in the form described above.

### Corollary

Given a collection  $\mathcal{B}$ , determining whether there exists a set  $S$  witnessing  $\text{disc}(\mathcal{B}) \leq 1$  is NP-complete.

## When is a collection $p$ -splittable?

For general  $p$ , finding simple rules to tell when a collection of sets is  $p$ -splittable is very difficult, as one would expect given the theorem above. However, we do have criteria for some special collections.

### Lemma

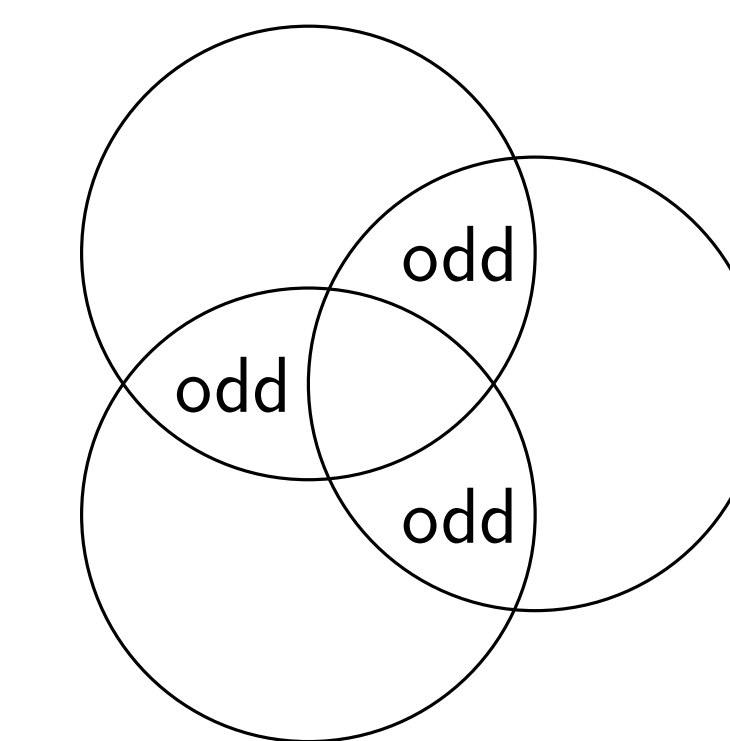
Let  $\mathcal{B} = \{B_1, \dots, B_n\}$ , a collection of sets whose elements all lie in exactly  $m$  sets.

- If  $\mathcal{B}$  is  $p$ -splittable, then  $\sum_{i=1}^n [p|B_i|]$  is divisible by  $m$ .
- If  $m = 1$  or  $m = n - 1$ , the converse holds: if  $\sum_{i=1}^n [p|B_i|]$  is divisible by  $m$  then  $\mathcal{B}$  is  $p$ -splittable.

## Criteria for $\frac{1}{2}$ -splittability

While the main theorem implies it is hard to find splittability criteria in general, we have had some success with small  $n$  and  $p = \frac{1}{2}$ . Some known results in this case are:

- Every collection of one or two sets is splittable.
- A collection of three sets is splittable if and only if it is not of the form [3]:



The situation becomes much more complex for four or more sets.

### 4-Set Classification Theorem

Every unsplittable collection of four sets falls into one of eleven simple patterns.

To prove this theorem we used a supercomputer to check all cases with a small number of elements, manually sorted the output, and generalized the conclusion using the lemma below.

### Lemma

If  $\mathcal{B}$  is splittable, then  $\mathcal{B}$  remains splittable when an even number of elements are added to any of its Venn regions.

Computer experiments also lead us to the following:

### Conjecture

Any collection of sets with no empty Venn regions is splittable.

## References & Acknowledgements

- N. Alon and J. Spencer. *The Probabilistic Method*. Wiley, 2000, p. 240.
- N. Bansal. "Constructive Algorithms for Discrepancy Minimization". In: *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. FOCS '10. IEEE Computer Society, 2010, pp. 3–10.
- D. Condon et al. "On Generalizations of Separating and Splitting Families". In: *submitted* (2015). arXiv: 1412.4683 [math.CO].
- S. Dasgupta, C. Papadimitriou, and U. Vazirani. *Algorithms*. McGraw Hill, 2006, p. 254.
- J. Spencer. "Six Standard Deviations Suffice". In: *Transactions of the AMS* 289.2 (1985), pp. 679–706.

This work was supported by the NSF under grant No. DMS 1359425 and Boise State University.

