Threshold Functions for Systems of Equations on Random Sets

Juanjo Rué, ANA ZUMALACÁRREGUI

Instituto de Ciencias Matemáticas (CSIC-UAM-UC3M-UCM), Madrid

Seminaire Modèles Combinatoires - LIX

Instituto de Ciencias Matemáticas

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

UNIVERSIDAD AUTONOMA DE MADRID

Introduction

Two Examples

$$n = 100;\ |\mathcal{A}| = 5$$

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Not Probable to get a 3-AP

Two Examples

$n = 100;\ |\mathcal{A}| = 20$

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Probable to get a 3-AP

A General Principle

- In discrete structures, there exists a **TRANSITION** between the non-existence and the existence of certain patterns.
- Furthermore this transition is, in general, **ABRUPT**.

$$\Downarrow$$

Threshold Phenomena

In This Talk...

1.– **Definitions**

2.– **Linear Systems of Equation. Our Results**

3.– **Trivial and Degenerated Solutions**

4.– **The Probabilistic Method**

5.– **Further Research**

Definitions

Random sets in $[n]$

Two models:

- $\mathcal{A} \subseteq [n]$ a subset chosen **UNIFORMLY** at random among all the subsets with the same size.
- $\mathcal{A} \subseteq [n]$ a subset of elements chosen **INDEPENDENTLY** at random in $[n]$:

$$p(k \in \mathcal{A}) = p = p(n)$$

EQUIVALENCE: these two models are "*equivalent*" iff

$$p = \frac{|\mathcal{A}|}{n}$$

What is a threshold?

Let P a combinatorial property.

$$\mathcal{A} \in P \text{ iff } \mathcal{A} \text{ satisfies the property } P.$$

$$t(n) \text{ is a } \textbf{threshold} \begin{cases} p = o(t(n)), \text{then } \lim p(\mathcal{A} \in P) \to 0 \\ t(n) = o(p), \text{then } \lim p(\mathcal{A} \in P) \to 1 \end{cases}$$

Threshold=**abrupt transition**

Observations and results

Thresholds are **NOT** defined uniquely.

A property P is **monotone increasing** iff

$$\mathcal{A} \subseteq \mathcal{B},\, \mathcal{A} \in P \Rightarrow \mathcal{B} \in P.$$

THEOREM (Bollobás, Thomason):

A monotone increasing property **ALWAYS** has a threshold.

Linear Systems of Equation.
Our Results

Some definitions and codification

Many natural conditions in additive combinatorics can be codified via linear systems of equations:

- Free **SET** of k-AP: avoids

$$\begin{cases} x_1 + x_3 = 2x_2 \\ \dots \\ x_{k-2} + x_k = 2x_{k-1} \end{cases}$$

- Sidon **SET**: avoids

$$x_1 + x_2 = x_3 + x_4$$

- $B_h[g]$ **SET**: avoids

$$\begin{cases} x_{1,1} + \cdots + x_{h,1} = x_{1,2} + \cdots + x_{h,2} \\ \dots \\ x_{1,g-1} + \cdots + x_{h,g-1} = x_{1,g} + \cdots + x_{h,g} \\ x_{1,g} + \cdots + x_{h,g} = x_{1,g+1} + \cdots + x_{h,g+1} \end{cases}$$

TRIVIAL solutions are **NOT** allowed!

The General Problem

Constructing *dense* subsets which exclude and arithmetical condition is a very involved problem, which requires *ad hoc* arguments.

$$\Downarrow$$

We study *common* properties instead of *extremal* properties.

Let $M \cdot \mathbf{x} = 0$ be a linear system of r equations and m variables and let $\mathcal{A}$ be a random set in $[n]$.

$$\Downarrow$$

$$P_M : M \cdot \mathbf{x} = 0 \textbf{ has NON TRIVIAL solutions in } \mathcal{A}^m$$

What do we study?

Questions we study:

- Location of the position of the threshold.
- Nature of the threshold.

And...how do we do this?

By means of GENERAL arguments

Our Results (I)

- **Location** of the threshold

(R., Zumalacárregui) Let $r < m$ and $M \cdot \mathbf{x} = 0$ be a linear system of r equations and m variables:

- maximum rank.
- with a solution with pairwise different positive coordinates.

Then $p = n^{\frac{r}{m}-1}$ is a threshold for the property P_M.

Our Results (and II)

- **Nature** of the threshold

(R., Zumalacárregui) If $p = cn^{\frac{r}{m}-1}$, then

$$\lim_{n\to\infty} p\left(\mathcal{A} \in P_M\right) = 1 - e^{-\frac{\mathrm{Vol}(\mathcal{P}_\mathrm{M})}{\mu_M}c^m},$$

- $M \cdot \mathbf{x} = 0$, $\mathbf{x} \in [0,1]^m$ defines a polytope $\mathcal{P}_M$ with volume $\mathrm{Vol}(\mathcal{P}_\mathrm{M})$.
- μ_M is a symmetry factor of the matrix M.

(The distribution of the number of solutions is a **Poisson**...)

Examples: k-AP

The system under study is the following:

$$\begin{cases} x_1 + x_3 = 2x_2 \\ \dots \\ x_{k-2} + x_k = 2x_{k-1} \end{cases}$$

	r	m	p	$\mathbb{E}[\lvert\mathcal{A}\rvert]$	$\mathrm{Vol}(\mathcal{P}_M)$	μ_M
$k-\mathrm{AP}$	$k-2$	k	$n^{-2/k}$	$n^{1-2/k}$	$\frac{1}{2(k-1)}$	1

Let us compare with the extremal values:

$$n \cdot \frac{(\log n)^{1/4}}{e^{c\sqrt{\log n}}} \ll \max_{\mathcal{A}\subset[n]} \{|\mathcal{A}| : \mathcal{A} \text{ avoiding } 3-\mathrm{AP}\} \ll n \cdot \frac{(\log\log n)^5}{\log n}$$

$$n \cdot \frac{(\log n)^{(2\log k)^{-1}}}{e^{c(k)(\log n)^{\log^{-1} k}}} \ll \max_{\mathcal{A}\subset[n]} \{|\mathcal{A}| : \mathcal{A} \text{ avoiding } k-\mathrm{AP}\} \ll n \cdot (\log\log n)^{-2^{-2^{(k+9)}}}$$

The common behavior approximates the extremal one when $k \to \infty$

Examples: Sidon Sets

The system under study is the following:

$$x_1 + x_2 = x_3 + x_4$$

	r	m	p	$\mathbb{E}[\|\mathcal{A}\|]$	$\mathrm{Vol}(\mathcal{P}_M)$	μ_M
Sidon	1	4	$n^{-3/4}$	$n^{1/4}$	$\frac{2}{3}$	8

There exist Sidon sets of cardinality of order $n^{1/2}$.

Examples: $B_h[g]$ sets

The system under study is the following:

$$\begin{cases} x_{1,1} + \cdots + x_{h,1} = x_{1,2} + \cdots + x_{h,2} \\ \cdots \\ x_{1,g-1} + \cdots + x_{h,g-1} = x_{1,g} + \cdots + x_{h,g} \\ x_{1,g} + \cdots + x_{h,g} = x_{1,g+1} + \cdots + x_{h,g+1} \end{cases}$$

	r	m	p	$\mathbb{E}[\lvert\mathcal{A}\rvert]$	$\mathrm{Vol}(\mathcal{P}_M)$	μ_M
$B_h[g]$	g	$h(g+1)$	$n^{\frac{g}{h(g+1)}-1}$	$n^{\frac{g}{h(g+1)}}$	♣	$(g+1)!(h!)^{g+1}$

In the extremal case we have $\asymp n^{1/h}$, and the difficult point is to compute the constant.

$$\Updownarrow\Updownarrow\Updownarrow$$

Doing $g \to \infty$, we recover the extremal estimate.

We **GENERALIZE** the result of Godbole, Janson, Loncatore and Rapoport for $B_h[1]$.

Trivial and Degenerated Solutions

Two Examples

- 3-AP.

TRIVIAL solutions are the ones with difference 0.

- Sidon Sets. The solutions are:

1.- 4 different components.
$\rightarrow$ **NO TRIVIAL, NO DEGENERATED.**

2.- $x_1 = x_2$, but $x_3 \neq x_4$: $2x_1 = x_3 + x_4$.
$\rightarrow$ **NO TRIVIAL, DEGENERATED.**

3.- $x_1 = x_3$ i $x_2 = x_4$: with two elements we have enough.
$\rightarrow$ **TRIVIAL, DEGENERATED.**

We need to define carefully *degenerated* and *trivial*.

The Partition associated to a Solution

- Let $\mathbf{x} = (x_1, \dots, x_m)$ be a solution of the system $M \cdot \mathbf{x} = 0$. This solution induces a partition of $[m]$ in terms of equality of components: $\mathfrak{p}(\mathbf{x})$.
- This solution comes from a *subordinate* system to $M \cdot \mathbf{x} = 0$ by equaling variables in $\mathbf{x}$ in terms of the partition.

Many situations may happen in a subordinate system:

- The rank of the system do **NOT** decrease:
  NO TRIVIAL DEGENERATED solution.
- The rank of the system decrease:
  TRIVIAL DEGENERATED solution.

This definition generalizes the one posted by Ruzsa in *Solving a linear equation in a set of integers I, II.*

The dynamics of the solutions

By increasin from 0 the density of the random set we observe:

- The first solutions are trivial ones.
- The first **NON TRIVIAL** solutions are **NON DEGENERATED** (pairwise different components).
- **NON TRIVIAL DEGENERATED** solutions appear later.

RESUMING:

The threshold is a consequence of NON TRIVIAL NON DEGENERATED solutions

The probabilistic method

The Ideas (I)

We want to count the (expected) number of solutions of the system with coordinates in $\mathcal{A}$:

$$\text{Solution } \mathbf{x} \leftrightarrow \text{Event } E_{\mathbf{x}}$$

The events must be considered up to symmetry

$$\mathbf{x} = (1,4,2,3),\ \mathbf{y} = (4,1,3,2),\ \text{and } E_{\mathbf{x}} = E_{\mathbf{y}}.$$

Each event has the following probability:

$$p(E_{\mathbf{x}}) = p^{\sharp\,\text{different components}} \rightarrow \mathbf{X} = \sum_{\mathbf{x} \in S_M} \mathbb{I}_{\mathbf{x}}$$

We need to estimate the number of solutions of a linear system of equations, where components are bounded by n

The Ideas (II)

The number of solutions of $M \cdot \mathbf{x} = 0$ with coordinates in $[n] \cup \{0\}$ is given by Ehrhart's theory on polytopes:

Teorema d'Ehrhart (Simplificat)
Let $\mathcal{P}$ be a d-dimensional convex polytope defined by a linear system of equations. Then:

$$\left| n \cdot \mathcal{P} \cap \mathbb{Z}^d \right| = \mathrm{Vol}(\mathcal{P}) n^d (1 + o(1)).$$

$$\mathbb{E}[\mathbf{X}] = \sum_{\mathbf{x} \in S_M} p(E_{\mathbf{x}}) = \frac{\mathrm{Vol}(\mathcal{P}_M)}{\mu_M} n^{m-r} p^m (1 + o(1)),$$

where $o(1)$ encapsulates both lower order terms and **NON TRIVIAL DEGENERATED** solutions.

The Ideas (III)

If $p = o(n^{\frac{r}{m}-1})$, then $\mathbb{E}[\mathbf{X}] = o(1)$, and $\mathbf{X} = 0$ a.a.s.!

$$\Downarrow\downarrow\Downarrow$$

And if $n^{\frac{r}{m}-1} = o(p)$ ...**NOT** as simple ($\mathbf{X} > 0$ a.a.s.)...

PHILOSOPHY: Is the r.v. $\mathbf{X}$ concentrated around $\mathbb{E}[\mathbf{X}]$?

$$\Updownarrow$$

Study of the second moment of $\mathbf{X}$

The Ideas (IV)

Just with the information coming from the first moment and the second moment...we have enough!

(SECOND MOMENT) Let $\mathbf{X} = \mathbb{I}_1 + \cdots + \mathbb{I}_s$ be a sum of indicator r.v. , where $\mathbb{I}_i$ is associated to the event E_i.
Let $i \sim j$ if $i \neq j$ and the events E_i, E_j are dependent.

$$\Delta = \sum_{i \sim j} p\left(E_i \wedge E_j\right)$$

If $\mathbb{E}[\mathbf{X}] \to \infty$ and $\Delta = o\left(\mathbb{E}[\mathbf{X}]^2\right)$, $\mathbf{X} \sim \mathbb{E}[\mathbf{X}]$ a.a.s.
In particular, $X > 0$ a.a.s.

We show that the dominant contribution in Δ arises from solutions with pairwise different components.

The Ideas (and V)

- For $p = cn^{\frac{r}{m}-1}$ we study

$$p\left(\bigwedge_{\mathbf{x}\in S_M} \overline{E_{\mathbf{x}}}\right)$$

- The events are not independent...but almost!

(JANSON's INEQUALITY) Let $\{E_i\}_{i\in I}$ be a set of events. Let $\varepsilon > 0$ such that for all $i \in I$, $p(E_i) \leq \varepsilon$. Then

$$\prod_{i\in I} p(\overline{E_i}) \leq p(\bigwedge_{i\in I} \overline{E_i}) \leq e^{\frac{\Delta}{2(1-\varepsilon)}} \prod_{i\in I} p(\overline{E_i}),$$

As before, the main contribution arises from solutions with pairwise different components.

Further Research

Far beyond Janson's Inequality

Using the *Brun's Sieve* we obtain the limiting distribution of $\mathbf{X}$ around the threshold:

$$\lim_{n\to\infty} p(\mathbf{X}=k) = \frac{1}{k!}\left(\frac{\mathrm{Vol}(\mathcal{P}_{\mathrm{M}})}{\mu_M}c^m\right)^k e^{-\frac{\mathrm{Vol}(\mathcal{P}_{\mathrm{M}})}{\mu_M}c^m}$$

Obtaining this limiting distribution is based on the fact that around the threshold the *dependence is very weak*.

$\Downarrow\Updownarrow\Downarrow$

It is not common to get a solution, and if it happens, it is very sparse.

$\Downarrow\Updownarrow\Downarrow$

We could try to erase some elements in the set in order to kill these solutions to increase the density!

The Alteration Method: a new Frontier

Once we have a probabilistic construction one has to apply the *Alteration Method*, which gives *for free* better density results.

$$\updownarrow\Downarrow\updownarrow$$

We fix a probability p bigger than the threshold .

- Number of expected elements in $\mathcal{A}$: pn
- Number of expected solutions: $p^m n^{m-r}$

Equaling:

$$pn = p^m n^{m-r} \to 1 = p^{m-1} n^{m-r-1} \to p = n^{\frac{r}{m-1}-1}.$$

This is what we call the **"weak threshold"**.

Far beyond the “ weak threshold”

Once we have a density bigger that the one given by the “weak threshold”, every element in $\mathcal{A}$ contributes to several solutions of the system. Consequently, in this point the dependence is very important.

$\downarrow\Downarrow\downarrow$

The arguments we used for the threshold (Second moment, Janson) do not work here.

$\downarrow\Downarrow$ ...**BUT**... $\Downarrow\downarrow$

The r.v. $\mathbf{X}$ is a polynomial of bounded degree of independent indicator r.v.: **Kim-Vu concentration result**.

PRINCIPAL QUESTION: Could we find a limiting distribution for the number of solutions in this regime?

Merci

Threshold Functions for Systems of Equations on Random Sets

Juanjo Rué, Ana Zumalacárregui

Instituto de Ciencias Matemáticas (CSIC-UAM-UC3M-UCM), Madrid

Seminaire Modèles Combinatoires - LIX