High-dimensional random arrays. Structural decompositions and concentration. Part I

Pandelis Dodos

Athens, 15 January 2021

Functional Analysis and Operator Algebras Seminar

Joint work with Kostas Tyros and Petros Valettas

1.a. Motivation/Overview

Concentration: *a function which depends smoothly on its variables is essentially constant, as long as the number of the variables is large enough.*

• (Gaussian concentration) Let $\boldsymbol{G} = (G_1, \dots, G_n)$ be a random vector with independent standard normal entries. If $f\colon (\mathbb{R}^n, \|\cdot\|_2) \to \mathbb{R}$ is 1-Lipschitz, then for any $t > 0$

$$\mathbb{P}\big(\big|f(\boldsymbol{G}) - \mathbb{E}[f(\boldsymbol{G})]\big| > t\big) \leqslant C\exp(-ct^2).$$

• (Bounded differences inequality) Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random vector with independent entries which take values in a Polish space $\mathcal{X}$. Let $f\colon \mathcal{X}^n \to \mathbb{R}$ be measurable, and for every $i \in [n]$ let $c_i > 0$ be such that $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leqslant c_i$ if $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}^n$ differ only in the i-th coordinate. Then for any $t > 0$

$$\mathbb{P}\big(\big|f(\boldsymbol{X}) - \mathbb{E}[f(\boldsymbol{X})]\big| > t\big) \leqslant C\exp\Big(\frac{-ct^2}{c_1^2 + \cdots + c_n^2}\Big).$$

1.b. Motivation/Overview

It is easy to see that this phenomenon is no longer valid if we drop the smoothness assumption. Nevertheless:

- (D, Kanellopoulos, Tyros–2016) For every $p > 1$ and every $0 < \varepsilon \leqslant 1$, there exists a constant $c > 0$ with the following property. If $n \geqslant 2/c$ is an integer, $\boldsymbol{X} = (X_1, \dots, X_n)$ is a random vector with independent entries which take values in a measurable space $\mathcal{X}$, and $f \colon \mathcal{X}^n \to \mathbb{R}$ is a measurable function with $\mathbb{E}[f(\boldsymbol{X})] = 0$ and $\|f(\boldsymbol{X})\|_{L_p} = 1$, then there exists an interval I of $[n]$ with $|I| \geqslant cn$ such that

$$\mathbb{P}\big(\big|\mathbb{E}[f(\boldsymbol{X}) \,|\, \mathcal{F}_I]\big| \leqslant \varepsilon\big) \geqslant 1 - \varepsilon$$

where $\mathbb{E}[f(\boldsymbol{X}) \,|\, \mathcal{F}_I]$ denotes the conditional expectation of $f(\boldsymbol{X})$ with respect to the σ-algebra $\mathcal{F}_I := \sigma(\{X_i : i \in I\})$.

1.c. Motivation/Overview

- Roughly speaking, this result asserts that if a function of several variables is sufficiently integrable, then, by integrating out some coordinates, it becomes essentially constant.

- It was motivated by—and it has found several applications in—problems in combinatorics. Most notably, it was used to give a new proof of the density Hales–Jewett theorem.

2. Main goal

- In a nutshell, our main goal is to extend the previous concentration estimate to functions of random vectors with not necessarily independent entries.

- We will focus on high-dimensional random arrays whose distribution is invariant under certain symmetries. The motivation to study functions of symmetric random arrays is related to an important combinatorial conjecture of Bergelson.

3. Random arrays

Definition (Random arrays, and their subarrays)

Let d be a positive integer, and let I be a set with $|I| \geqslant d$. A *d-dimensional random array on I* is a stochastic process $\boldsymbol{X} = \langle X_s : s \in \binom{I}{d} \rangle$ indexed by the set $\binom{I}{d}$ of all d-element subsets of I. If J is a subset of I with $|J| \geqslant d$, then the *subarray of $\boldsymbol{X}$ determined by J* is the d-dimensional random array $\boldsymbol{X}_J := \langle X_s : s \in \binom{J}{d} \rangle$; moreover, by $\mathcal{F}_J$ we shall denote the σ-algebra $\sigma(\{X_s : s \in \binom{J}{d}\})$ generated by $\boldsymbol{X}_J$.

One-dimensional random arrays are just random vectors; two-dimensional random arrays are essentially the same as random symmetric matrices, and their subarrays correspond to principal submatrices. More generally, higher-dimensional random arrays correspond to random symmetric tensors.

4.a. Notions of symmetry

Random arrays with a sufficiently symmetric distribution are a classical object of study in probability: de Finetti, Diaconis/Freedman, Aldous, Hoover, Kallenberg, Fremlin/Talagrand, Austin/Tao,. . .

• A d-dimensional random array $\boldsymbol{X} = \langle X_s : s \in \binom{I}{d} \rangle$ on I is called *exchangeable* if for every (finite) permutation π of I, the random arrays $\boldsymbol{X}$ and $\boldsymbol{X}_\pi := \langle X_{\pi(s)} : s \in \binom{I}{d} \rangle$ have the same distribution.

• A d-dimensional random array $\boldsymbol{X}$ on I is called *spreadable* if for every pair J, K of finite subsets of I with $|J| = |K| \geqslant d$, the subarrays $\boldsymbol{X}_J$ and $\boldsymbol{X}_K$ have the same distribution.

$$\text{exhangeability} \;\Rightarrow\; \text{spreadability}$$

4.b. Notions of symmetry

Definition (Approximate spreadability)

Let $\boldsymbol{X}$ be a d-dimensional random array on a (possibly infinite) set I, and let $\eta \geqslant 0$. We say that $\boldsymbol{X}$ is *η-spreadable* provided that for every pair J, K of finite subsets of I with $|J| = |K| \geqslant d$ we have

$$\rho_{\mathrm{TV}}(P_J, P_K) \leqslant \eta$$

where P_J and P_K denote the laws of the random subarrays $\boldsymbol{X}_J$ and $\boldsymbol{X}_K$ respectively, and ρ_{TV} stands for the total variation distance.

4.c. Notions of symmetry

The following result—whose proof is a fairly straightforward application of Ramsey's theorem—shows that finite-valued, approximately spreadable random arrays are ubiquitous.

Fact

For every triple m, n, d of positive integers with $n \geqslant d$, and every $\eta > 0$, there exists an integer $N \geqslant n$ with the following property. If $\mathcal{X}$ is a set with $|\mathcal{X}| = m$ and $\boldsymbol{X}$ is an $\mathcal{X}$-valued, d-dimensional random array on a set I with $|I| \geqslant N$, then there exists a subset J of I with $|J| = n$ such that the random array $\boldsymbol{X}_J$ is η-spreadable.

5.a. A basic example

Example

Let $n \geqslant d$ be positive integers, let $\xi_1, \dots, \xi_n$ be i.i.d. random variables, and define a d-dimensional random array $\boldsymbol{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$ on $[n]$ by setting

$$X_s := \prod_{i \in s} \xi_i .$$

- The random array $\boldsymbol{X}$ is always exchangeable and *dissociated*, that is, for every pair J, K of disjoint subsets of $[n]$ with $|J|, |K| \geqslant d$, the subarrays $\boldsymbol{X}_J$ and $\boldsymbol{X}_K$ are independent. (But of course, the entries of $\boldsymbol{X}$ are not independent.)

- Concentration estimates for linear (and, more generally, smooth) functions of random arrays of this form, have been studied by several authors (Latala, Adamczak/Wolff, Götze/Sambale/Sinulis, Vershynin).

5.b. A basic example

- The previous example can be easily generalized. Specifically, let $n \geqslant d$ be positive integers, let $\xi_1, \dots, \xi_n$ be i.i.d. random variables, let $h\colon \mathbb{R}^d \to \mathbb{R}$ be a Borel function, and define a d-dimensional random array $\boldsymbol{X} = \langle X_s : s \in \binom{[n]}{d} \rangle$ on $[n]$ by setting for every $s = \{i_1 < \dots < i_d\} \in \binom{[n]}{d}$

$$X_s := h(\xi_{i_1}, \dots, \xi_{i_d}).$$

- These random arrays are spreadable and dissociated.

- As we shall see, the distribution of an arbitrary **finite**, finite-valued, approximately spreadable, random array is a *mixture* of distributions of random arrays of this form.

6.a. The main problem (in more detail)

Problem

Let $n \geqslant d$ be positive integers, let $\boldsymbol{X}$ be a d-dimensional random array on $[n]$ whose entries take values in a measurable space $\mathcal{X}$, let $f\colon \mathcal{X}^{\binom{[n]}{d}} \to \mathbb{R}$ be a measurable function, and assume that $\mathbb{E}[f(\boldsymbol{X})] = 0$ and $\|f(\boldsymbol{X})\|_{L_p} = 1$ for some $p > 1$. Under what condition on $\boldsymbol{X}$ can we find a large subset I of $[n]$ such that the random variable $\mathbb{E}[f(\boldsymbol{X}) \,|\, \mathcal{F}_I]$ is concentrated around its mean?

(Recall that $\mathcal{F}_I$ denotes the σ-algebra generated by $\boldsymbol{X}_I$.)

6.b. The main problem (in more detail)

Two comments are in order here.

• The condition we are referring to should be fairly concrete, in the sense that even its *negation* provides useful information on the random array $\boldsymbol{X}$.

• Secondly, note that we demand that the random variable $f(\boldsymbol{X})$ becomes concentrated after conditioning it on a *subarray* of $\boldsymbol{X}$. This is a fairly natural requirement in this context, and it is essential for combinatorial applications.

7.a. The main result (two-dimensional, boolean case)

Theorem (D, Tyros, Valettas–2020)

Let $1 < p \leqslant 2$, let $0 < \varepsilon \leqslant 1$, let $k \geqslant 2$ be an integer, and set

$$C = C(p,\varepsilon,k) := \exp\Big(\frac{34}{\varepsilon^8 (p-1)^2} \cdot k^2\Big).$$

Also let $n \geqslant C$ be an integer, let $\boldsymbol{X} = \langle X_s : s \in \binom{[n]}{2} \rangle$ be a $\{0,1\}$-valued, $(1/C)$-spreadable, two-dimensional random array on $[n]$, and assume that

$$(*) \qquad \big|\mathbb{E}[X_{\{1,3\}} X_{\{1,4\}} X_{\{2,3\}} X_{\{2,4\}}] - \\ - \mathbb{E}[X_{\{1,3\}}]\,\mathbb{E}[X_{\{1,4\}}]\,\mathbb{E}[X_{\{2,3\}}]\,\mathbb{E}[X_{\{2,4\}}]\big| \leqslant \frac{1}{C}.$$

7.b. The main result (two-dimensional, boolean case)

Theorem (cont'd)

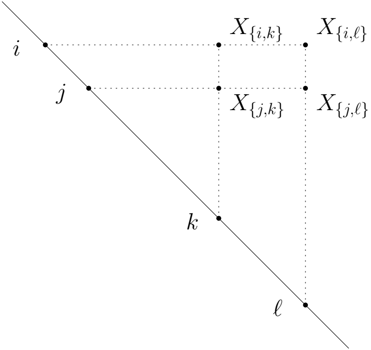
Then for every function $f\colon \{0,1\}^{\binom{[n]}{2}} \to \mathbb{R}$ *with* $\mathbb{E}[f(\boldsymbol{X})] = 0$ *and* $\|f(\boldsymbol{X})\|_{L_p} = 1$ *there exists an interval* I *of* $[n]$ *with* $|I| = k$ *and such that*

$$\mathbb{P}\Big(\big|\mathbb{E}[f(\boldsymbol{X}) \,|\, \mathcal{F}_I]\big| \leqslant \varepsilon\Big) \geqslant 1 - \varepsilon.$$

8.a. The box independence condition

Condition $(*)$ together with the $(1/C)$-spreadability of $\boldsymbol{X}$ imply that for every $i, j, k, \ell \in [n]$ with $i < j < k < \ell$ we have

$$(*)' \qquad \big|\mathbb{E}[X_{\{i,k\}} X_{\{i,\ell\}} X_{\{j,k\}} X_{\{j,\ell\}}] - \\ - \mathbb{E}[X_{\{i,k\}}]\, \mathbb{E}[X_{\{i,\ell\}}]\, \mathbb{E}[X_{\{j,k\}}]\, \mathbb{E}[X_{\{j,\ell\}}]\big| \leqslant \frac{6}{C}.$$



8.b. The box independence condition

• Though not obvious at first sight, as the parameter C gets bigger, condition $(*)'$ forces the random variables $X_{\{i,k\}}, X_{\{i,\ell\}}, X_{\{j,k\}}, X_{\{j,\ell\}}$ to behave independently. (It also implies that the correlation matrix of $\boldsymbol{X}$ is close to the identity.)

• We also note that $(*)'$ is, essentially, an optimal condition. Specifically, for every integer $n \geqslant 4$ there exist:

- a boolean, exchangeable, two-dimensional random array $\boldsymbol{X}$ on $[n]$, and
- a translated multilinear polynomial $f\colon \mathbb{R}^{\binom{[n]}{2}} \to \mathbb{R}$ of degree 4 with $\mathbb{E}[f(\boldsymbol{X})] = 0$ and $\|f(\boldsymbol{X})\|_{L_\infty} \leqslant 1$,

such that the correlation matrix of $\boldsymbol{X}$ is the identity and the random variable $f(\boldsymbol{X})$ is not conditionally concentrated. (And, of course, $\boldsymbol{X}$ does not satisfy condition $(*)'$.)

9. Higher-dimensional extensions

• Analogous concentration estimates hold true for d-dimensional, finite-valued, approximately spreadable, random arrays for any positive integer d.

• In the higher-dimensional case, we can find an interval I of $[n]$ of size

$$|I| \approx \sqrt[d]{\log n}.$$

• As expected, the higher-dimensional version of the "box independence condition" is also optimal.

10.a. Ideas of the proof

The proof proceeds in two steps.

Step 1. It is based on an *energy increment strategy*, and it uses estimates for martingale difference sequences in L_p spaces. It applies to random arrays with arbitrary distributions (in particular, not necessarily approximately spreadable), and it shows that the conditional concentration of $f(\boldsymbol{X})$ is equivalent to an approximate form of the dissociativity of $\boldsymbol{X}$.

The main advantage of this step is that it enables us to forget about the function f and focus exclusively on the random array $\boldsymbol{X}$.

10.b. Ideas of the proof

Step 2. We show that the "box independence condition" propagates and forces all, not too large, subarrays of $\boldsymbol{X}$ to behave independently.

This is analogous to the phenomenon, discovered in the theory of quasi-random graphs (Thomason, Chung/Graham/Wilson, Rödl, Gowers,...), that a graph G which contains (roughly) the expected number of 4-cycles must also contain the expected number of any other, not too large, graph H.

In fact, this is more than an analogy; this step easily yields the aforementioned property of quasi-random graphs. We shall discuss further the relation between the "box independence condition" and quasi-randomness of graphs and hypergraphs next week.

10.c. Ideas of the proof

The proof of the second step proceeds by induction on the dimension d. The argument is based on repeated averaging and an appropriate version of the weak law of large numbers in order to gradually upgrade the box independence condition. The combinatorial heart of the matter lies in the selection of this averaging. (Looks like playing bricks for kids.)

Thanks for listening!