Acta Math. Univ. Comenianae
Vol. LXXXVIII, 3 (2019), pp. 679–683

DEVIATION PROBABILITIES FOR ARITHMETIC PROGRESSIONS AND OTHER REGULAR DISCRETE STRUCTURES

G. FIZ PONTIVEROS, S. GRIFFITHS, M. SECCO AND O. SERRA

ABSTRACT. Let $\mathcal{H}$ be a k-uniform hypergraph on a vertex set V and B_m be a uniformly sampled m-set from V. Set X to be the random variable given by the number of edges induced by the set B_m. We provide tight upperbounds (up to a constant in the exponent) for the tail distribution of $X - \mathbb{E}(X)$ for a wide range of deviations, provided some near regularity conditions are satisfied by the hypergraph $\mathcal{H}$. In particular, the bounds may be applied to the setting of arithmetic progressions and more generally to solutions of linear systems.

1. INTRODUCTION

Determining how well a random variable X is concentrated around its expectation $\mathbb{E}[X]$ has a long history and is of great interest in many areas of mathematics. There is today a plethora of methods to prove concentration of measure inequalities but more often than not these general bounds are not optimal in specific applications.

In probabilistic combinatorics, the random variables of interest are typically counts of some fixed combinatorial objects in a random structure. Notable examples are subgraph counts in the Erdős-Rényi random graph model $G(n,p)$ and arithmetic progressions in a random set, i.e., given $k \geq 3$, let X be the number of arithmetic progressions of length k in $[N]_p$, the random subset of $[N] = \{1, \dots, N\}$ where each element is included independently with probability p.

In the first instance, Janson, Oleszkiewicz and Ruciński [**9**] provided a moment-based method that, for subgraph counts in random graphs, gives estimates for $\mathbb{P}(X \geq (1+\epsilon)\mathbb{E}[X])$ which are best possible up to logarithmic factors in the exponent. The problem of closing this gap remained open for several years, with breakthroughs by Chatterjee, Chatterjee and Varadhan, Lubetzky and Zhao, and DeMarco and Kahn for particular subgraph counts (see [**5**] for a detailed account)

Received June 7, 2019.
2010 *Mathematics Subject Classification.* Primary 60C05, 05C80, 05C65.
G. F. P. is supported by BGSMath Postdoctoral Grant and the Spanish Research Agency under projects MDM-2014-0445 and MTM2017-82166.
S. G. is supported by PUC-Rio, CNPq Proc. 310656/2016-8 and FAPERJ Proc. 202.713/2018.
M. S. is supported in part by the Coordenação de Aperfeiçoamento de Pessoal de NÃ-vel Superior-Brasil (CAPES) – Finance Code 001.

and there has been further progress this year on the upper tail by Harel, Mousset and Samotij [**16**]. Tight bounds are also known in parts of the moderate deviations range, see [**6, 7, 14, 12**]. For the second instance: Janson and Ruciński [**11**] extended their technique to the setting of arithmetic progressions of length k in random subsets and obtained analogous bounds. More precisely it was shown that for fixed $\epsilon > 0$ and $k \geq 3$

$$p^{C\sqrt{\varepsilon \mathbb{E}[X]}} \leq \mathbb{P}\left(X \geq (1+\epsilon)\mathbb{E}\left[X\right]\right) \leq p^{c\sqrt{\mathbb{E}[X]}}, \tag{1}$$

where the constant $C = C(k)$ and $c = c(k, \epsilon)$. Subsequently, Warnke [**20**] improved the upperbound in (1) to match the lowerbound, i.e., for fixed $\epsilon > 0$ and $k \geq 3$

$$p^{C\sqrt{\varepsilon \mathbb{E}[X]}} \leq \mathbb{P}\left(X \geq (1+\epsilon)\mathbb{E}\left[X\right]\right) \leq p^{c\sqrt{\epsilon \mathbb{E}[X]}}, \tag{2}$$

where the both constants C and c now only depend on k. Recently Bhattacharya, Ganguly, Shao and Zhao got precise asymptotics for a more restricted range of p. They showed in [**4**] that

$$\mathbb{P}\left(X \geq (1+\epsilon)\mathbb{E}\left[X\right]\right) = p^{(1+o(1))\sqrt{\epsilon \mathbb{E}[X]}}.$$

Most counting problems in probabilistic combinatorics can be formulated under the following general framework: let $\mathcal{H}$ be a k-uniform hypergraph on a vertex set V and let V_p denote a random subset of V where each $v \in V_p$ is chosen independently with probability p. Let X be the random variable given by $|E(V_p)|$, the number of edges induced by the random subset V_p. How well can we bound the probability that X deviates from its expectation? Note that in this setting the random variable X we wish to understand has a very special structure: it is a polynomial of degree at most k of independent Bernoulli random variables. Even so, despite the advances made by Kim and Vu [**13**] and others, there is no concentration inequality that systematically gives a sharp bounds to the upper tail. Indeed, the results in [**11**] and [**20**] tackle the problem of large deviations in this general framework.

Very recently Goldschmidt, Griffiths and Scott [**14**] introduced a new approach using martingale methods to analyze moderate deviations in subgraph counts in the $G(n, m)$ model instead of the $G(n, p)$ model. They argue that, in the case of moderate deviations, the $G(n, m)$ model is more natural, and in any case show that very sharp bounds in the $G(n, p)$ model follow from knowledge of the asymptotic rate in the $G(n, m)$ model.

The present work was motivated by the desire to generalise the methods developed in [**14**] to the setting of arithmetic structures given by solutions of a linear system in an abelian group. In the course of the project we found that the approach generalised naturally to a class of hypergraphs which have a high degree of regularity, which include hypergraphs whose edges represent the solutions to linear systems, see e.g. [**17**].

2. The set up and results

Let $\mathcal{H}$ be a k-uniform hypergraph with vertex set V and consider the following random process of ordered subsets $B_m \subseteq V$: let $v_1, \dots, v_N$ be a uniformly chosen permutation of the elements of V and define

$$B_m = (v_1, \dots, v_m),\ m = 0, 1, \dots, N.$$

Our goal is to bound the probability of deviations from the mean in the number of edges of $\mathcal{H}$ induced by the set B_m. For our analysis, we shall also need to consider partially filled edges along the process. This motivates the following definitions:

Definition 1. A pair (e, x) is a $(\mathcal{H}, j)$-sequence if $e \in E(\mathcal{H})$ is an edge of $\mathcal{H}$ and $x = (x_1, \dots, x_j)$ is a j-tuple of distinct vertices of e. Let $N_j(B_m)$ denote the number of $(\mathcal{H}, j)$-sequences induced by B_m and observe that

$$N_j(B_m) = \sum_{e \in E(H)} (|e \cap B_m|)_j,$$

where for a set S we denote by $(S)_j$ the set of all j-tuples of distinct elements of S, which has cardinality $(|S|)_j = |S|(|S|-1)\cdots(|S|-j+1)$. Let $L_j(m) = \mathbb{E}(N_j(B_m))$ be the mean and set

$$D_j(B_m) = N_j(B_m) - L_j(m).$$

We now state our main results. We call a hypergraph is *r-tuple-regular* if every r-subset belongs to the same number d_r of edges.

Theorem 1. *Let $1 \le r \le k$. Let $\mathcal{H}$ be a k-uniform hypergraph on $[N]$. Suppose that $\mathcal{H}$ is $(r-1)$-tuple-regular and has maximum r-degree Δ_r. Then*

$$\mathbb{P}\left(D_j(B_m) > a\right) \le N^{O_k(1)} \exp\left(\frac{-\Omega_k(1)a^{2/r}}{m\Delta_r^{2/r}}\right),$$

for all $r \le j \le k$ and for all $a > 0$.

Furthermore, the same bounds apply to the corresponding negative deviations.

In applications, such as to hypergraphs arising from arithmetic configurations, the regularity hypothesis is slightly too restrictive. It is natural to consider a weaker notion of regularity:

Definition 2. A k-uniform hypergraph $\mathcal{H} = (V, E)$ is (r, η)-near-regular, $1 \le r \le k$, if every r-subset of vertices is contained in $(1 \pm \eta)\bar{d}_r$ edges, where $\bar{d}_r = \bar{d}_r(\mathcal{H})$ denotes the average degree of r-sets in H.

In this setting, we obtain an analogous result at the expense of restricting the range of deviations covered by the result.

Theorem 2. *Let $1 \le r \le k$ and let $\eta \in [0, 3^{-r+1}]$. Let $\mathcal{H}$ be a k-uniform hypergraph on $[N]$. Suppose that $\mathcal{H}$ is $(r-1, \eta)$-near-regular with maximum r-degree Δ_r. Then*

$$\mathbb{P}\left(D_j(B_m) > a\right) \le N^{O_k(1)} \exp\left(\frac{-\Omega_k(1)a^{2/r}}{m\Delta_r^{2/r}}\right),$$

for all $r \leq j \leq k$ *and for all* $a \geq C_r \eta^{r/(r-1)} h(m/N)^{(j-1)r/(r-1)}$, *where* $C_r = (10k!)^{10^r}$.

Furthermore, the same bounds apply to the corresponding negative deviations.

In fact the bound in Theorem 2 is best possible, up to a constant in the exponent. We will provide details of the construction in the upcomming journal version of the paper.

Theorem 2 is flexible enough to handle the hypergraph corresponding to counts of arithmetic configurations arising from solutions of linear systems in the integers or in cyclic groups (usually with a very small value of η such as $1/N$). In particular, one can obtain the following bound for deviations in the counts of k term arithmetic progressions:

$$\mathbb{P}\left(D_k(B_m) > a\right) \leq N^{O_k(1)} \exp\left(-\Omega_k(1)\frac{a}{m}\right).$$

3. The approach

The proofs of Theorems 1 and Theorem 2 follow a hybrid strategy from [**14**] and [**13**]. We start by expressing the deviation $D_j(B_m)$ as a sum of martingale increments and also find good bounds for the maximum step size. The martingale decomposition of $D_j(B_m)$ is a generalization of the subgraph count decomposition given in [**14**] to general k-uniform hypergraphs.

In this setting, our main tool to bound the probability of deviations is a simple modification to the classical Azuma-Hoeffding inequality: if a martingale $(M)_{i=0}^n$ is obtained from a random process with at most n possibilities at each step and is such that $\|M_i - M_{i-1}\|_\infty < c_i$ for all $i = 1, \ldots, n$ except probability at most $\exp(-b)$, then the probability of a deviation a is at most

$$\exp\left(\frac{-a^2}{2\sum_1^n c_i^2}\right) + n\exp(-b).$$

This sets up the basic structure of our proof. In order to obtain strong bounds on deviation probabilities for $(\mathcal{H}, j)$-sequences we first require good bounds on the related martingale increments. We show that the martingale increments correspond to deviations of auxiliary $(k-1)$-uniform hypergraphs $\mathcal{H}(x)$ (sometimes called the link hypergraphs of $\mathcal{H}$) which inherit some regularity from $\mathcal{H}$. We therefore prove our main results by a double induction on r, in which $P(r)$ is the statement of the theorem and $Q(r)$ is a statement about the size of the increments.

References

1. Azuma K., *Weighted sums of certain dependent random variables*, Tohoku Math. J. **19** (1967), 357–367.
2. Boucheron S., Lugosi G. and Massart P., *Concentration Inequalities*, Oxford University Press, Oxford, 2013.
3. Bhattacharya B. B. and Mukherjee S., *A note on neplica symmetry in upper tails of mean-field hypergraphs*, `arXiv:1812.09841`.
4. Bhattacharya B. B., Ganguly S., Shao X. and Zhao Y., *Upper tails for arithmetic progressions in a random set*, `arXiv:1605.02994`.

5. Chatterjee S., *An introduction to large deviations for random graphs*, Bull. Amer. Math. Soc. (N. S.) **53** (2016), 617–642.
6. Döring H. and Eichelsbacher P., *Moderate deviations in a random graph and for the spectrum of Bernoulli random matrices*, Electron. J. Probab. **14** (2009), 2636–2656.
7. Feray V., Meliot P. L. and Nikeghbali A., *Mod-ϕ Convergence I: Normality Zones and Precise Deviations*, Springer Briefs in Probability and Mathematical Statistics, Springer, 2016.
8. Hoeffding W., *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc. **58** (1963), 13–30.
9. Janson S., Oleszkiewicz K. and Ruciński A., *Upper tails for subgraph counts in random graphs*, Israel J. Math. **142** (2004), 61–92.
10. Janson S. and Ruciński A., *When are small subgraphs of a random graph normally distributed?*, Probab. Theory Related Fields **78** (1988), 1–10.
11. Janson S. and Ruciński A., *Upper tails for counting objects in randomly induced subhypergraphs and rooted random graphs*, Ark. Mat. **49** (2011), 79–96.
12. Janson S. and Warnke L., *The lower tail: Poisson approximation revisited*, Random Structures Algorithms **48** (2015), 219–246.
13. Kim J. H. and Vu V. H., *Concentration of multivariate polynomials and its applications*, Combinatorica **20** (2000), 417–434.
14. Goldschmidt C., Griffiths S. and Scott A., *Moderate deviations of subgraph counts in the Erdős-Rényi random graphs $G(n,m)$ and $G(n,p)$*, `arXiv:1902.06830`.
15. Graham R., Rödl V. and Ruciński A., *On Schur properties of random subsets of integers*, J. Number Theory **61** (1996), 388–408.
16. Harel M., Mousset F. and Samotij W., *Upper tails via high moments and entropic stability*, `arXiv:1904.08212`.
17. Rué J., Serra O. and Vena L., *Counting configuration-free sets in groups*, European J. Combin. **66** (2017), 281–307.
18. Schacht M., *Extremal results for random discrete structures*, Ann. of Math. (2) **184** (2016), 333–365.
19. Warnke L., *On the method of typical bounded differences*, Combin. Probab. Comput. **25** (2016), 269–299.
20. Warnke L., *Upper tails for arithmetic progressions in random subsets*, Israel J. Math. (2017), 221–317.

G. Fiz Pontiveros, Departament de Matemàtiques, Universitat Politècnica de Catalunya, Barcelona, Spain,
e-mail: `gonzalo.fiz@upc.edu`

S. Griffiths, Departamento de Matemàtica, PUC-Rio, Gávea, Rio de Janeiro, Brazil,
e-mail: `simon@mat.puc-rio.br`

M. Secco, Departamento de Matemàtica, PUC-Rio, Gávea, Rio de Janeiro, Brazil,
e-mail: `matheussecco@mat.puc-rio.br`

O. Serra, Departament de Matemàtiques, Universitat Politècnica de Catalunya, Barcelona, Spain,
e-mail: `oriol.serra@upc.edu`