COMBINATORICA
Bolyai Society – Springer-Verlag

COMBINATORICA **30** (2) (2010) 175–200
DOI: 10.1007/s00493-010-2401-x

SET SYSTEMS WITHOUT A SIMPLEX OR A CLUSTER

PETER KEEVASH*, DHRUV MUBAYI†

Received April 25, 2007

A d-dimensional simplex is a collection of $d+1$ sets with empty intersection, every d of which have nonempty intersection. A k-uniform d-cluster is a collection of $d+1$ sets of size k with empty intersection and union of size at most $2k$.

We prove the following result which simultaneously addresses an old conjecture of Chvátal [6] and a recent conjecture of the second author [28]. For $d \geq 2$ and $\zeta > 0$ there is a number T such that the following holds for sufficiently large n. Let $\mathcal{G}$ be a k-uniform set system on $[n] = \{1, \ldots, n\}$ with $\zeta n < k < n/2 - T$, and suppose either that $\mathcal{G}$ contains no d-dimensional simplex or that $\mathcal{G}$ contains no d-cluster. Then $|\mathcal{G}| \leq \binom{n-1}{k-1}$ with equality only for the family of all k-sets containing a specific element.

In the non-uniform setting we obtain the following exact result that generalises a question of Erdős and a result of Milner, who proved the case $d = 2$. Suppose $d \geq 2$ and $\mathcal{G}$ is a set system on $[n]$ that does not contain a d-dimensional simplex, with n sufficiently large. Then $|\mathcal{G}| \leq 2^{n-1} + \sum_{i=0}^{d-1} \binom{n-1}{i}$, with equality only for the family consisting of all sets that either contain some specific element or have size at most $d-1$.

Each of these results is proved via the corresponding stability result, which gives structural information on any $\mathcal{G}$ whose size is close to maximum. These in turn rely on a stability theorem that we obtain using an earlier result of Frankl.

1. Introduction

Extremal problems for set systems are fundamental in Combinatorics. We will be concerned with maximising the size of a system that does not contain

Mathematics Subject Classification (2000): 05C35, 05C65, 05D05

* Research supported in part by NSF grant DMS-0555755.

† Research partially supported by National Science Foundation Grant DMS-0400812, and an Alfred P. Sloan Research Fellowship.

a configuration of sets called a d-dimensional simplex (or d-simplex), which is a collection of $d+1$ sets with empty intersection, every d of which have nonempty intersection.

Various cases of this basic question can be traced back to some of the oldest theorems and conjectures in extremal combinatorics. First, there is the theorem of Erdős, Ko and Rado [11] which is one of the fundamental results in extremal set theory. It states that for $n \geq 2k$ an intersecting k-uniform set system on $[n] = \{1, \ldots, n\}$ can have size at most $\binom{n-1}{k-1}$, and if $n > 2k$, then equality holds only for a star, i.e., a family consisting of all sets that contain some specific element $x \in [n]$. Since a 1-simplex is a pair of nonempty disjoint sets this can be interpreted as solving the uniform extremal problem for 1-simplices.

Second, there is the $(6,3)$-theorem of Ruzsa and Szemerédi, which states that a nearly disjoint triple system (meaning that every two triples have at most one common element) on $[n]$ containing no 2-simplex has size at most $o(n^2)$. The correct growth rate of this maximum is still a major open problem. This has nontrivial consequences in number theory, as it implies Roth's Theorem on 3-term arithmetic progressions (a special case of Szemerédi's Theorem).

Third, there is the Turán problem for hypergraphs, which asks for the maximum size of a k-uniform hypergraph on $[n]$ which contains no complete k-uniform hypergraph on $d+1$ elements. This problem is open for all $d+1 > k > 2$. When $d = k$, the forbidden configuration is a d-simplex, and determining this maximum even for $d = k = 3$ is a famous conjecture of Turán from the 1940's (Erdős offered \$1000 for its solution).

The case $d = k = 2$ of the Turán problem for hypergraphs is quite easy, and is a special case of Turán's theorem (proved by Mantel in 1907), which is the starting point of extremal graph theory. It states that a graph on $[n]$ with no triangle has at most $\lfloor n^2/4 \rfloor$ edges and equality holds only for the complete bipartite graph $K_{\lfloor n/2 \rfloor, \lceil n/2 \rceil}$. Motivated by this, Erdős [10] posed the more general problem of determining the largest k-uniform set system on $[n]$ with no triangle (i.e., 2-simplex). Many cases of this were solved by various authors ([5,7,12,13]) over the years, until finally the problem was completely solved by the second author and Verstraëte [30], who showed that for $k \geq 3$ and $n \geq 3k/2$ a k-uniform set system on $[n]$ with no triangle can have size at most $\binom{n-1}{k-1}$, with equality only for a star.

Chvátal [6] generalized Erdős' conjecture as follows.

Conjecture 1.1 (Chvátal [6]). Suppose $k \geq d+1 \geq 2$, $n > k(d+1)/d$ and $\mathcal{G}$ is a k-uniform set system on $[n]$ with no d-simplex. Then $|\mathcal{G}| \leq \binom{n-1}{k-1}$, with equality only for a star.

We have already mentioned the solutions of this conjecture for $d=1$ and $d=2$, but it is open for larger d. A significant breakthrough was achieved by Frankl and Füredi [16], who proved Conjecture 1.1 for sufficiently large n.

Many similar problems in extremal set theory are easier to solve for n large compared to k. The best example of this is the Erdős–Ko–Rado theorem for t-intersecting families. It is quite easy to determine the maximum size of a k-uniform t-intersecting family on $[n]$ for large n (indeed this was known to Erdős–Ko–Rado), but the solution for all n was open for over thirty years until it was finally settled by Ahlswede and Khachatrian [1]. The case $t = 2$ of the Ahlswede–Khachatrian result was recently used in complexity theory. Indeed, it was a key component in the work of Dinur and Safra [9], who showed that approximating the Minimum Vertex Cover problem to within a factor of 1.3606 is NP-hard. For the simplex problem, even the solution for large n by Frankl and Füredi was far from trivial. Our main result settles Conjecture 1.1 when k/n and $n/2-k$ are both bounded away from zero.

Another potential generalisation of the Erdős–Ko–Rado theorem was suggested by Katona (see [15]) and extended by the second author [28]. A k-uniform d-cluster is a collection of $d+1$ sets of size k with empty intersection and union of size at most $2k$. Note that a 1-cluster is the same as a 1-simplex (which consists of two disjoint sets), for which the Erdős–Ko–Rado theorem solves the extremal problem. Katona posed the case $d=2$ and Frankl and Füredi [15] obtained partial results and made a conjecture for the extremal problem. This was settled by the second author [28], who showed that if $k\geq 3$ and $n\geq 3k/2$ a k-uniform set system on $[n]$ with no 2-cluster can have size at most $\binom{n-1}{k-1}$, with equality only for a star. The same question with $2k$ replaced by a smaller number, say $2k-1$, leads to many interesting and unsolved questions (see [15,23] for results in the case $k=3$). The following conjecture, which generalizes the Frankl–Füredi conjecture, was posed in [28].

Conjecture 1.2 ([28]). Let $k\geq d+1\geq 2$ and $n>k(d+1)/d$. Suppose that $\mathcal{G}$ is a k-uniform set system on $[n]$ with no d-cluster. Then $|\mathcal{G}|\leq\binom{n-1}{k-1}$, with equality only if $\mathcal{G}$ is a star.

As mentioned above, Conjecture 1.2 holds for $d = 1$ and $d = 2$. The second author [27] recently proved that for fixed $k \geq d+1 \geq 2$ we have $|\mathcal{G}|\leq(1+o(1))\binom{n-1}{k-1}$ as $n\to\infty$, and that Conjecture 1.2 holds for $d=3$ and large n. It was also recently observed by Chen and Liu that for $d=k-1$, Conjecture 1.2 reduces to Conjecture 1.1, and since the latter was solved by Chvátal [6], Conjecture 1.2 holds for $d = k-1$. Our result addresses

Conjecture 1.2 for all $d \geq 2$ but in a different range, namely when k/n and $n/2-k$ are both bounded away from zero.

Erdős also posed his extremal question for triangles in the non-uniform setting. It was solved by Milner (unpublished), who showed that a triangle-free set system on $[n]$ can have size at most $2^{n-1}+n$. The second author and Verstraëte [30] gave a short proof and showed that equality holds only for the family consisting of all sets that either contain some specific element or have size at most 1. We will generalise this and determine the maximum size of a set system on $[n]$ with no d-simplex. Although we initially believed that this generalisation would not be too difficult, our methods for this problem more or less use the full machinery for the problem in the uniform setting. It would be interesting to obtain a new and shorter argument.

A key tool in our proofs is the idea of stability, or approximate structure, which can be traced back to work of Erdős and Simonovits in the 60's in extremal graph theory. Informally stated, a stability result tells us about the structure of configurations that are close to optimal in an extremal problem: for example, a triangle-free graph with $n^2/4-o(n^2)$ edges differs from a complete bipartite graph by $o(n^2)$ edges. Such a result is interesting in its own right, but somewhat surprisingly it is often a useful stepping stone in proving an exact result. Indeed, it was developed by Erdős and Simonovits to determine the exact Turán number for k-critical graphs. This approach has been recently used with great success in hypergraph Turán theory (see [18–20,24,25,29,32,33]), enumeration of discrete structures [4] and extremal set theory (see [2,26,27]).

Recently, Friedgut [17] (see also Dinur and Friedgut [8]) has proved some stability results for intersecting families using spectral methods and discrete fourier analysis. We will strengthen his result, using a purely combinatorial theorem of Frankl [14], and this stability result will enable us to derive others for simplices and clusters.

The rest of this paper is organised as follows. In the next section we will state our results and describe the strategy of the proof. We have chosen to postpone this as we need to introduce another more complicated configuration which behaves nicely in induction arguments. Section 3 begins with a summary of our notation and contains other preliminary material, namely, our basic inductive lemmas and estimates on hypergeometric and binomial random variables. Sections 4–6 contain the main proofs, firstly in the uniform and secondly in the non-uniform setting. The final section contains some concluding remarks and conjectures.

2. Results

Our main result for the problems about uniform families settles Conjectures 1.1 and 1.2 when k/n and $n/2-k$ are both bounded away from 0. One of the main new ideas to solve the simplex problem is to prove a result that guarantees a structure more complicated than a simplex.

Definition. Fix $d \geq 1$. A collection of $d+2$ sets $A, A_1, \dots, A_{d+1}$ is a *strong* d-simplex if $\{A_1, \dots, A_{d+1}\}$ is a d-simplex, and A contains an element of $\bigcap_{i \neq j} A_i$ for each $j \in [d+1]$.

For example, a strong 1-simplex is a path of length 3, i.e., three sets A, B, C such that $A \cap B$ and $B \cap C$ are nonempty and $A \cap C = \emptyset$.

Theorem 2.1 (Main Result). *For all $\zeta > 0$ and $d \geq 2$ there exists $\delta > 0$ and integers T and N so that the following holds for all $n > N$. Suppose $\mathcal{G}$ is a k-uniform set system on $[n]$ where $k = n/2 - t$, with $T < t < (1/2 - \zeta)n$, and $|\mathcal{G}| > (1 - \delta t/n)\binom{n-1}{k-1}$. Suppose also either that $\mathcal{G}$ does not contain a strong d-simplex or that $\mathcal{G}$ does not contain a d-cluster. Then $\mathcal{G}$ is a star, and so $|\mathcal{G}| \leq \binom{n-1}{k-1}$.*

Theorem 2.1 certainly implies the result stated in the abstract, and actually carries some stronger structural information about set systems that are near to maximum size, which is independently interesting and also facilitates our inductive argument. It is noteworthy that the theorem does not hold when $d = 1$. For example, one can take the construction from the Hilton–Milner theorem [21], where $\mathcal{G}$ comprises all sets containing $\{1\}$ and intersecting $\{2, \dots, k+1\}$, together with the set $\{2, \dots, k+1\}$. Clearly $\mathcal{G}$ is intersecting, its size is very close to $\binom{n-1}{k-1}$ (as long as $n < 3k$ for example), and it is not a star. Nevertheless, we need a statement similar to Theorem 2.1 for the base case in our inductive proof. For intersecting families, such a stability result follows from a result of Frankl. The theorem below will be proved in Section 4.

Theorem 2.2 (Stability Result for intersecting families). *For all $\epsilon, \zeta > 0$ there exists $\gamma > 0$ and M so that the following holds for all $n > M$. Suppose $\mathcal{G}$ is a k-uniform set system on $[n]$ where $k = n/2 - t$, with $0 < t < (1/2 - \zeta)n$, and $|\mathcal{G}| > (1 - \gamma t/n)\binom{n-1}{k-1}$. Suppose also that $\mathcal{G}$ does not contain a strong 1-simplex. Then there is some x in $[n]$ so that all but at most $\epsilon\binom{n-1}{k-1}$ sets of $\mathcal{G}$ contain x.*

Remark. Since a strong 1-simplex contains a 1-cluster, Theorem 2.2 clearly holds if we replace 'strong 1-simplex' by '1-cluster'. This will be used in the course of the paper.

As we mentioned earlier, Friedgut [17] has proved some stability results for intersecting families using spectral methods and discrete fourier analysis. Specifically, he proves Theorem 2.2 but only when $t > \zeta n$. Also, in [8], with Dinur, they take care of the case $t = [1/2 - o(1)]n$. However, since in much of the current work we must analyze the situation when k is very close to $n/2$ (especially for the nonuniform case), we need the full generality of Theorem 2.2.

In the non-uniform setting we prove the following theorem.

Theorem 2.3 (Exact Result – nonuniform case). *Suppose $d \geq 2$ and $\mathcal{G}$ is a set system on $[n]$ that does not contain a d-simplex, with n sufficiently large. Then $|\mathcal{G}| \leq 2^{n-1} + \sum_{i=0}^{d-1} \binom{n-1}{i}$, with equality iff $\mathcal{G}$ consists of all sets that either contain some fixed element x or have size at most $d-1$.*

Theorem 2.3 is also proved via the stability approach. The difficulty in applying this method is that intersecting families on $[n]$ of size 2^{n-1} are abundant (indeed, any intersecting family can be augmented to one of size 2^{n-1}), and no reasonable stability result for non-uniform intersecting families holds, i.e., no stability result holds for Theorem 2.3 when $d = 1$. Nevertheless, we are able to prove the following stability theorem, from which Theorem 2.3 can be deduced with relative ease. Our result also applies to d-clusters, where we say that a (not necessarily uniform) collection of sets $A_1, \ldots, A_{d+1}$ is a d-cluster if $\bigcap_{i=1}^{d+1} A_i = \emptyset$ and $|\bigcup_{i=1}^{d+1} A_i| \leq 2 \max_{i=1}^{d+1} |A_i|$.

Theorem 2.4 (Stability Result – nonuniform case). *Suppose $d \geq 2$, n is sufficiently large, $\mathcal{G}$ is a set system on $[n]$ and $|\mathcal{G}| > (1 - n^{-5/8})2^{n-1}$. Suppose also that either $\mathcal{G}$ does not contain a d-simplex or $\mathcal{G}$ does not contain a d-cluster. Then there is some x in $[n]$ so that every set $A \in \mathcal{G}$ with $||A| - n/2| < n^{2/3}$ contains x. In particular $|\mathcal{G}| < 2^{n-1} + 2^{n-n^c}$ for any constant $c < 1/3$.*

The basic idea behind the proof of Theorem 2.4 is to analyse those sets of $\mathcal{G}$ whose size is close to $n/2$. Since $|\mathcal{G}|$ is close to 2^{n-1}, most of its sets fall in this range. We then use Theorem 2.1 to obtain structural information about these sets, and finally we use this information to deduce the structure of all of $\mathcal{G}$. Many of the bounds used in the proof involve somewhat delicate estimates of sums and products of binomial coefficients $\binom{n}{k}$, where $k \sim n/2$. Our tools for these estimates are the binomial and hypergeometric distributions, and our results on these are collected in Section 3.3, and proved in the Appendix.

3. Preliminaries

3.1. Notation

We consider set systems on a ground set $[n]=\{1,\dots,n\}$, mostly denoted by $\mathcal{G}$ (calligraphic). Subsets are generally denoted by upper case Roman letters, integers by lower case Roman letters, reals by Greek letters. If $\mathcal{G}$ is non-uniform we use $\mathcal{G}^k$ to denote its sets of size k and $\mathcal{G}^I$ to denote $\bigcup_{k\in I}\mathcal{G}^k$ for a set of sizes I.

Suppose $\mathcal{G}$ is a set system on $[n]$ and $x\in[n]$. The degree $d_{\mathcal{G}}(x)$ is the number of sets of $\mathcal{G}$ that contain x. The sets $A\subset[n]\setminus\{x\}$ with $A\cup\{x\}\in\mathcal{G}$ fall into two families: $L_x(\mathcal{G})$ consists of those A for which there is some $y\neq x$ for which $A\cup\{y\}$ is also in $\mathcal{G}$; $S_x(\mathcal{G})$ consists of those A for which $A\cup\{y\}\in\mathcal{G}$ implies that $y=x$. Note that $d_{\mathcal{G}}(x)=|L_x(\mathcal{G})|+|S_x(\mathcal{G})|$.

Say that $x,y\in[n]$ are in the same connected component of $\mathcal{G}$ if there is a sequence $x=x_1,x_2,\dots,x_t=y$ for some t such that for every $1\le i\le t-1$ there is a set $A_i\in\mathcal{G}$ with $\{x_i,x_{i+1}\}\in A_i$.

3.2. Inductive lemmas

In this subsection we give two lemmas that are the cornerstone of our inductive approach.

Lemma 3.1. *Suppose $n>k\ge d+1\ge 3$, $\mathcal{G}$ is a k-uniform set system on $[n]$ and $x\in[n]$.*
(1) If $L_x(\mathcal{G})$ contains a strong $(d-1)$-simplex then $\mathcal{G}$ contains a strong d-simplex.
(2) If $L_x(\mathcal{G})$ contains a $(d-1)$-cluster then $\mathcal{G}$ contains a d-cluster.

Proof.
(1) Suppose that $L_x(\mathcal{G})$ contains the strong $(d-1)$-simplex $A,A_1,\dots,A_d$, where $A_1,\dots,A_d$ is a $(d-1)$-simplex. By definition of $L_x(\mathcal{G})$ there exists $y\neq x$ such that $B_{d+1}=A\cup\{y\}\in\mathcal{G}$. Let $B=A\cup\{x\}$ and $B_i=A_i\cup\{x\}$ for $i\in[d]$. We will argue that $B,B_1,\dots,B_{d+1}$ is a strong d-simplex in $\mathcal{G}$. For each $j\in[d]$, let

$$C_j=\bigcap_{\substack{i=1\\ i\neq j}}^{d}A_i \quad\text{and}\quad D_j=\bigcap_{\substack{i=1\\ i\neq j}}^{d}B_i.$$

Since $A_1,\dots,A_d$ forms a $(d-1)$-simplex, $\bigcap_{i=1}^{d}A_i=\emptyset$, and so $\bigcap_{i=1}^{d}B_i=\{x\}$. As $\{x\}\cap B_{d+1}=\emptyset$, we conclude that $\bigcap_{i=1}^{d+1}B_i=\emptyset$.

By definition of strong $(d–1)$-simplex, $C_j \cap A \neq \emptyset$ for each $j \in [d]$. Therefore $D_j \cap B_{d+1} \neq \emptyset$ for each j. Also, $\{x\} \subset \bigcap_{i=1}^d B_i$, so the intersection of every d of the B_is is nonempty. We conclude that $B_1, \ldots, B_{d+1}$ is a d-simplex. To see that $B, B_1, \ldots, B_{d+1}$ is a strong d-simplex, observe that $\{x\} \subset \bigcap_{i=1}^d B_i \cap B$ and $B \cap D_j \cap B_{d+1} \supset A \cap C_j \neq \emptyset$.

(2) Suppose that $L_x(\mathcal{G})$ contains the $(d-1)$-cluster $A_1, \ldots, A_d$. There exists $y \neq x$ such that $B_{d+1} = A_1 \cup \{y\} \in \mathcal{G}$. Let $B_i = A_i \cup \{x\}$ for $i \in [d]$. Since $A_1, \ldots, A_d$ forms a $(d-1)$-cluster, $\bigcap_{i=1}^d A_i = \emptyset$, and so $\bigcap_{i=1}^d B_i = \{x\}$. As $x \notin B_{d+1}$, we conclude that $\bigcap_{i=1}^{d+1} B_i = \emptyset$. Also, $|\bigcup_{i=1}^{d+1} B_i| \leq |\bigcup_{i=1}^d A_i| + |\{x,y\}| \leq 2(k-1)+2 = 2k$. Consequently, $B_1, \ldots, B_{d+1}$ is a d-cluster in $\mathcal{G}$. ∎

Lemma 3.2. *Suppose $1 < l < k$, $\mathcal{G}$ is a k-uniform set system on $[n]$, $S \in \mathcal{G}$, $E \subset [n] \setminus S$ with $|E| = k-l$ and $\mathcal{H}$ is an l-uniform set system on S with the property that $A \cup E \in \mathcal{G}$ for every $A \in \mathcal{H}$.*
(1) If $\mathcal{H}$ contains a strong $(d-1)$-simplex then $\mathcal{G}$ contains a strong d-simplex.
(2) If $\mathcal{H}$ contains a $(d-1)$-cluster then $\mathcal{G}$ contains a d-cluster.

Proof.
(1) Suppose that $\mathcal{H}$ contains the strong $(d-1)$-simplex $A, A_1, \ldots, A_d$, where $A_1, \ldots, A_d$ is a $(d-1)$-simplex. Let $B = A \cup E$, $B_i = A_i \cup E$ for all $i \in [d]$, and $B_{d+1} = S$. By the proof of Lemma 3.1 part (1), replacing $\{x\}$ by E, we conclude that $B, B_1, \ldots, B_{d+1}$ is a strong d-simplex in $\mathcal{G}$.

(2) Suppose that $\mathcal{H}$ contains the $(d-1)$-cluster $A_1, \ldots, A_d$. Let $B_i = A_i \cup E$ for $i \in [d]$ and $B_{d+1} = S$. Since $A_1, \ldots, A_d$ forms a $(d-1)$-cluster, $\bigcap_{i=1}^d A_i = \emptyset$, and so $\bigcap_{i=1}^d B_i = E$. As $E \cap B_{d+1} = \emptyset$, we conclude that $\bigcap_{i=1}^{d+1} B_i = \emptyset$. Also, $|\bigcup_{i=1}^{d+1} B_i| \leq |S| + |E| = k + (k-l) \leq 2k$. Consequently, $B_1, \ldots, B_{d+1}$ is a d-cluster in $\mathcal{G}$. ∎

3.3. Binomial and hypergeometric estimates

In this subsection, we describe some estimates on hypergeometric and binomial distributions that will take some work out of our later calculations. We postpone the proofs to Appendix A.

The hypergeometric random variable X with parameters (n, m, k) is defined as follows. Fix a set $S \subset [n]$, of size $|S| = m = rn$. Pick a random $T \subset [n]$, of size $|T| = k = pn$. Define $X = |T \cap S|$. Note that $\mathbb{E}X = km/n = prn$. Write $q = 1-p$, $s = 1-r$. In the estimates below the hidden constants in the $O(\cdot)$ terms depend on p, q, r, s in such a way that they are bounded uniformly in n whenever p, q, r, s are uniformly bounded away from 0, which will always be the case in our applications.

Firstly, we have an asymptotic formula for the probabilities of individual values:

$$\mathbb{P}(X = \mathbb{E}X + t) = (2\pi pqrsn)^{-1/2} e^{-t^2/2pqrsn + O(t/n + t^3/n^2)}. \tag{1}$$

For larger deviations the following 'Chernoff bound' approximation is useful (see [22, pp. 27–29]). Suppose either $Y = X$ or $Y = B(n,p)$ is a binomial variable (equal to the number of heads in n independent tosses of a coin that comes up heads with probability p). Suppose $0 < a < 3/2$. Then

$$\mathbb{P}(|Y - \mathbb{E}Y| > a\mathbb{E}Y) < 2e^{-\frac{a^2}{3}\mathbb{E}Y}. \tag{2}$$

We will also need the fact that the median of X is close to its mean: for any $\epsilon > 0$ we have

$$\mathbb{P}(X \geq \mathbb{E}X) = 1/2 + O(n^{-1/2+\epsilon}). \tag{3}$$

Finally we record the following estimate for later use (for fixed $d \geq 2$ and large n):

$$\sum_{t=n^{1/10}-d}^{n^{3/5}} \frac{t}{n} \binom{n-1}{n/2-t-1} > \frac{1}{10} n^{-1/2} 2^{n-1}. \tag{4}$$

4. Stability for intersecting families – Proof of Theorem 2.2

In this section we prove Theorem 2.2. We need the following result of Frankl [14]. Given k and $3 \leq i \leq k+1$ we define a k-uniform intersecting set system $\mathcal{F}_i$ on $[n]$ by

$$\mathcal{F}_i = \{A \subset [n] \colon |A| = k, 1 \in A, A \cap \{2, \dots, i\} \neq \emptyset\} \cup \{A \subset [n] \colon |A| = k, 1 \notin A, \{2, \dots, i\} \subset A\}.$$

Given a set system $\mathcal{G}$ on $[n]$ the degree of a set $S \subset [n]$ is $d_{\mathcal{G}}(S) = |\{A \colon A \in \mathcal{G}, S \subset A\}|$. Let $\Delta(\mathcal{G}) = \max_{x \in [n]} d_{\mathcal{G}}(x)$ denote the maximum degree of a singleton.

Theorem 4.1 (Frankl [14]). *Suppose $n > 2k$, $3 \leq i \leq k+1$, $\mathcal{G}$ is a k-uniform intersecting set system on $[n]$ and $|\mathcal{G}| > |\mathcal{F}_i|$. Then $\Delta(\mathcal{G}) > \Delta(\mathcal{F}_i)$.*

Now we use this to deduce Theorem 2.2. We recall the statement: For all $\epsilon, \zeta > 0$ there exists $\gamma > 0$ and M so that the following holds for all $n > M$. Suppose $\mathcal{G}$ is a k-uniform set system on $[n]$ where $k = n/2 - t$, with $0 < t < (1/2 - \zeta)n$, and $|\mathcal{G}| > (1 - \gamma t/n)\binom{n-1}{k-1}$. Suppose also that $\mathcal{G}$ does not contain a strong 1-simplex. Then there is some x in $[n]$ so that all but at most $\epsilon\binom{n-1}{k-1}$ sets of $\mathcal{G}$ contain x.

Proof of Theorem 2.2. Set $i = \lceil 2\zeta^{-1}\log\epsilon^{-1}\rceil$ and $\gamma = \zeta^i$. Choose M so that all calculations below requiring large n hold.

First of all we show that $\mathcal{G}$ is intersecting. Partition $\mathcal{G}$ into components $C_1,\dots,C_l$. If $l>1$ then $|\mathcal{G}| \le \sum_{i=1}^{l}\binom{|C_i|}{k} \le \binom{n-k}{k}+1$, using convexity of binomial coefficients and the fact that all components contain at least k points. This contradicts our assumed lower bound on $\mathcal{G}$, so $\mathcal{G}$ must be connected. Now if there are two disjoint sets A and B then we can find a walk starting with A and ending with B, and some 3 consecutive edges on this walk will form a strong 1-simplex, contradiction. We deduce that $\mathcal{G}$ is intersecting.

Now

$$\begin{aligned}\binom{n-1}{k-1} - |\mathcal{F}_i| &= \binom{n-i}{k-1} - \binom{n-i}{k-i}\\ &= \binom{n-1}{k-1}\binom{n-1}{i-1}^{-1}\left(\binom{n-k}{i-1} - \binom{k-1}{i-1}\right)\\ &= \binom{n-1}{k-1}\binom{n-1}{i-1}^{-1}\sum_{m=n/2-t-1}^{n/2+t-1}\binom{m}{i-2}\\ &> \binom{n-1}{k-1}\binom{n-1}{i-1}^{-1}(2t-1)\binom{n/2-t-1}{i-2}.\end{aligned}$$

Since

$$\begin{aligned}(2t-1)\binom{n-1}{i-1}^{-1}\binom{n/2-t-1}{i-2} &= (2t-1)\frac{i-1}{n-1}\binom{n-2}{i-2}^{-1}\binom{n/2-t-1}{i-2}\\ &> \zeta^{i-2}\frac{t}{n} > \gamma\frac{t}{n},\end{aligned}$$

we have $|\mathcal{G}|>|\mathcal{F}_i|$, and so by Theorem 4.1 $\Delta(\mathcal{G})>\Delta(\mathcal{F}_i)$. But we have

$$\begin{aligned}1-\binom{n-1}{k-1}^{-1}\Delta(\mathcal{F}_i) &= \binom{n-1}{k-1}^{-1}\binom{n-i}{k-1} = \prod_{j=1}^{i-1}\frac{n-k-j+1}{n-j}\\ &< \left(\frac{n-k}{n-1}\right)^{i-1} < (1-\zeta/2)^i < e^{-\zeta i/2} \le \epsilon,\end{aligned}$$

and so there is some x in $[n]$ so that all but at most $\epsilon\binom{n-1}{k-1}$ sets of $\mathcal{G}$ contain x. ∎

5. Main Result – Proof of Theorem 2.1

In this section we prove Theorem 2.1. We will need the following result of the second author and Verstraëte [31].

Theorem 5.1 ([31]). *Suppose $n \geq 2k > 2$, and $\mathcal{G}$ is a k-uniform set system on $[n]$ with no strong 1-simplex. Then $|\mathcal{G}| \leq \binom{n-1}{k-1}$.*

Now we are ready to prove the main result in the uniform setting. We restate it and state a lemma, and will prove both results simultaneously by induction.

Theorem 2.1 (Main Result). *For all $\zeta > 0$ and $d \geq 2$ there exists $\delta > 0$ and integers T, N so that the following holds for $n > N$. Suppose $\mathcal{G}$ is a k-uniform set system on $[n]$ where $k = n/2 - t$, with $T < t < (1/2 - \zeta)n$, and $|\mathcal{G}| > (1 - \delta t/n)\binom{n-1}{k-1}$. Suppose also either that $\mathcal{G}$ does not contain a strong d-simplex or that $\mathcal{G}$ does not contain a d-cluster. Then $\mathcal{G}$ is a star, and so $|\mathcal{G}| \leq \binom{n-1}{k-1}$.*

Lemma 5.2. *For all $\zeta' > 0$ and $d \geq 2$ there exist integers T', N' so that the following holds for $n' > N'$. Suppose $\mathcal{G}'$ is a k'-uniform set system on $[n']$, where $k' = n'/2 - t'$, with $T' < t' < (1/2 - \zeta')n'$, $C \subset [n']$ is some set with constant size $|C| = c$, $C \subset A$ for every $A \in \mathcal{G}'$. Suppose also that $S \subset [n'] \setminus C$, $|S| = k''$ with $|k'' - k| < 2n'^{2/3}$, and either $\mathcal{G}' \cup \{S\}$ does not contain a strong d-simplex or $\mathcal{G}' \cup \{S\}$ does not contain a d-cluster (possibly non-uniform). Then $|\mathcal{G}'| < \frac{4}{5}\binom{n'-c}{k'-c}$.*

Proof of Theorem 2.1 and Lemma 5.2. We prove Theorem 2.1 and Lemma 5.2 together by induction on d, and simultaneously present the argument for the base case $d = 2$ and the induction step. We will indicate how to find values for the constants δ, T, T', N, N' in the beginning of each proof.

Proof of Lemma 5.2. Let $\zeta' > 0$ and $d \geq 2$ be given. Let T, N be the outputs of Theorem 2.1 with inputs $\zeta = \zeta'/3$ and $d - 1$. Set $T' = T$ and $N' > 2N/\zeta'$. We also assume that N' is sufficiently large that all calculations involving $O(\cdot)$ estimates in the following proof are valid. Now suppose that $n' > N'$.

Let X be a hypergeometric random variable with parameters $(n' - c, k'', k' - c)$. Then $\mathbb{E}X = k''(k' - c)/(n' - c)$. We estimate $|\mathcal{G}'|$ in four parts according to size of intersection with S. Write $\mathcal{G}'_I = \{A \in \mathcal{G}' : |A \cap S| \in I\}$. We use the decomposition $\mathcal{G}' = \mathcal{G}'_{I_1} \cup \mathcal{G}'_{I_2} \cup \mathcal{G}'_{I_3} \cup \mathcal{G}'_{I_4}$ where

$$I_1 = [0, \mathbb{E}X/2], \qquad I_2 = [\mathbb{E}X/2, \mathbb{E}X - T'],$$
$$I_3 = [\mathbb{E}X - T', \mathbb{E}X], \qquad I_4 = [\mathbb{E}X, k''].$$

Using equation (3) for X we have

$$\binom{n'-c}{k'-c}^{-1}|\mathcal{G}'_{I_4}| \le \mathbb{P}(X \ge \mathbb{E}X) = 1/2 + O(n'^{-0.49}).$$

Next, equation (2) gives

$$\binom{n'-c}{k'-c}^{-1}|\mathcal{G}'_{I_1}| \le \mathbb{P}(X \le \mathbb{E}X/2) < 2e^{-\frac{k''(k'-c)}{12(n'-c)}} = O(1/n').$$

Also, equation (1) gives

$$\binom{n'-c}{k'-c}^{-1}|\mathcal{G}'_{I_3}| = \mathbb{P}(\mathbb{E}X - T' \le X \le \mathbb{E}X) = T' \cdot O(n'^{-1/2}) = O(n'^{-1/2}).$$

To estimate $\mathcal{G}'_{I_2}$ we consider the following sets:

$$\mathcal{G}'(E) = \{F\colon F \subset S, F \cup E \cup C \in \mathcal{G}'_{I_2}\}$$

where $E \subset [n'] \setminus (C \cup S)$. Then $\mathcal{G}'(E)$ is an l-uniform set system on S, with $|S| = k''$, and $l = k' - c - |E|$ for some $l \in I_2$. By Lemma 3.2 either $\mathcal{G}'(E)$ contains no strong $(d-1)$-simplex or $\mathcal{G}'(E)$ contains no $(d-1)$-cluster. Also, $l/k'' \ge \mathbb{E}X/(2k'') = (k'-c)/2(n'-c) > \zeta'/3$ and $l \le k/2 - T'$ so by the choice of T' we can apply the induction hypothesis of Theorem 2.1 for $d \ge 3$ or Theorem 5.1 for $d = 2$ to obtain $|\mathcal{G}'(E)| \le \binom{k''-1}{l-1}$. Let Y be a hypergeometric random variable with parameters $(n'-c-1, k''-1, k'-c-1)$. Then

$$\begin{aligned}\mathcal{G}'_{I_2} &= \sum_{E\colon l=k'-c-|E|\in I_2} |\mathcal{G}'(E)| \le \sum_{l\in I_2} \binom{n'-k''-c}{k'-c-l}\binom{k''-1}{l-1} \\ &\le \mathbb{P}(Y \le \mathbb{E}X - T')\binom{n'-c-1}{k'-c-1} = \big(1/2 + O(n'^{-0.49})\big)\frac{k'-c}{n'-c}\binom{n'-c}{k'-c} \\ &\le \big(1/4 + O(n'^{-0.49})\big)\binom{n'-c}{k'-c}.\end{aligned}$$

Here we used equation (3) for Y and the fact that $|\mathbb{E}X - \mathbb{E}Y| = O(1)$. In total we have

$$|\mathcal{G}'| = |\mathcal{G}'_{I_1}| + |\mathcal{G}'_{I_2}| + |\mathcal{G}'_{I_3}| + |\mathcal{G}'_{I_4}| \le \big(3/4 + O(n'^{-0.49})\big)\binom{n'-c}{k'-c} < \frac{4}{5}\binom{n'-c}{k'-c},$$

for large n', which proves the lemma. ∎

Proof of Theorem 2.1. Suppose $\zeta > 0$ and $d \geq 2$ are given. Let γ, M be the outputs of Theorem 2.2 with inputs $\epsilon = 1/10$ and $\zeta/2$. If $d \geq 3$, then let $\delta_{d-1}, T_{d-1}, N_{d-1}$ be the outputs of Theorem 2.1 (via induction) with inputs $\zeta/2$ and $d-1$. Let T'_{d-1}, N'_{d-1} be the outputs of Lemma 5.2 (via induction) with inputs $\zeta' = \zeta/2$ and $d-1$. Now choose

$$\delta = \frac{1}{3}\min\{\gamma, \delta_{d-1}\}, \qquad T > 1 + \max\{3/\delta, T_{d-1}, T'_{d-1}\}.$$

Finally, choose N such that $N > 1+\max\{M, N_{d-1}, N'_{d-1}\}$ and N is sufficiently large that the calculations below showing that (5) holds are valid.

We start with the following claim.

Claim. There exists $y \in [n]$ such that $|L_y(\mathcal{G})| > \left(1-3\delta\frac{t}{n}\right)\binom{n-2}{k-2}$ and $|S_y(\mathcal{G})| < \frac{1}{10}\binom{n-1}{k-1}$.

Proof of Claim. Consider the following double-counting equation:

$$\begin{aligned}(1-\delta t/n)k\binom{n-1}{k-1} < k|\mathcal{G}| = \sum_{x\in[n]} d(x) &= \sum_{x\in[n]} (|S_x(\mathcal{G})| + |L_x(\mathcal{G})|) \\ &= \sum_{x\in[n]} |S_x(\mathcal{G})| + \sum_{x\in[n]} |L_x(\mathcal{G})|.\end{aligned}$$

Since every $(k-1)$-set is counted by at most one $S_x(\mathcal{G})$, we have $\sum_x |S_x(\mathcal{G})| \leq \binom{n}{k-1}$, and so

$$\sum_x |L_x(\mathcal{G})| > (1-\delta t/n)k\binom{n-1}{k-1} - \binom{n}{k-1} > (1-2\delta t/n)n\binom{n-2}{k-2}. \tag{5}$$

Note that

$$\begin{aligned}&(n-1)(n-k+1)\binom{n-1}{k-1}^{-1}\left[(1-\delta t/n)k\binom{n-1}{k-1} - \binom{n}{k-1}\right. \\ &\qquad\qquad \left. - (1-2\delta t/n)n\binom{n-2}{k-2}\right] \\ &= (1-\delta t/n)k(n-1)(n-k+1) - n(n-1) - (1-2\delta t/n)(k-1)n(n-k+1) \\ &\qquad = \left\{\begin{array}{ll} \frac{1}{4}(t\delta-3)n^2 + o(n^2), & \text{if } t = o(n); \\ \delta(a/4-a^3)n^3 + O(n^2), & \text{if } t = an,\ a > 0. \end{array}\right\}\end{aligned}$$

Since $t > T > 3/\delta + 1$, we see that (5) holds for $n > N$.

Let $V_0 = \{x\colon |L_x(\mathcal{G})| \le (1-3\delta t/n)\binom{n-2}{k-2}\}$. For every x, we see from Lemma 3.1 that either $L_x(\mathcal{G})$ contains no strong $(d-1)$-simplex or $L_x(\mathcal{G})$ contains no $(d-1)$-cluster. Therefore $|L_x(\mathcal{G})| \le \binom{n-2}{k-2}$ for every x, using the induction hypothesis for $d \ge 3$ (by the choice of T and N) or Theorem 5.1 for $d=2$. So

$$\begin{aligned}(1-2\delta t/n)n\binom{n-2}{k-2} &\le \sum_{x\in V_0}|L_x(\mathcal{G})| + \sum_{x\in[n]\setminus V_0}|L_x(\mathcal{G})| \\ &\le |V_0|(1-3\delta t/n)\binom{n-2}{k-2} + (n-|V_0|)\binom{n-2}{k-2},\end{aligned}$$

which simplifies to $|V_0| \le 2n/3$. Finally there must be some $y \in [n]\setminus V_0$ with $|S_y(\mathcal{G})| < \frac{1}{10}\binom{n-1}{k-1}$, otherwise we would have $\sum_x |S_x(\mathcal{G})| \ge (n/30)\binom{n-1}{k-1} > \binom{n}{k-1}$, contradiction. This y has the properties required to prove the Claim. ∎

From this claim we can deduce that there is some $w \ne y$ contained in at least $\frac{9}{10}\binom{n-2}{k-2}$ sets of $L_y(\mathcal{G})$. To see this, observe that by Lemma 3.1 either $L_y(\mathcal{G})$ contains no strong $(d-1)$-simplex or $L_y(\mathcal{G})$ contains no $(d-1)$-cluster. Also $|L_y(\mathcal{G})| > \left(1-3\delta\frac{t}{n}\right)\binom{n-2}{k-2} > \left(1-3\delta\frac{k-1}{n-1}\right)\binom{n-2}{k-2}$. If $d=2$ then by the choice of δ and N, we can apply Theorem 2.2 to $L_y(\mathcal{G})$ and obtain the desired w. If $d \ge 3$, then since $T > 1+T_{d-1}$ and $N > 1+N_{d-1}$ we can apply the induction hypothesis to give some w that is contained in *every* set of $L_y(\mathcal{G})$.

Now we can finish the proof by two applications of Lemma 5.2, which applies due to the choice of T, N. First we apply it with $\mathcal{G}' = L_y(\mathcal{G})$ on the ground set $[n]\setminus\{y\}$ and $C = \{w\}$ and see that there cannot be $S \in \mathcal{G}$ with $S \cap C = \emptyset$: otherwise we would have $|L_y(\mathcal{G})| < \frac{4}{5}\binom{n-2}{k-2}$, which contradicts our choice in the Claim. Therefore every set contains one of w or y.

Since $(k-1)/(n-1) < k/n < 1/2$, the number of sets that contain y but not w is at most

$$|L_y(\mathcal{G})| - \frac{9}{10}\binom{n-2}{k-2} + |S_y(\mathcal{G})| \le \frac{1}{10}\binom{n-2}{k-2} + \frac{1}{10}\binom{n-1}{k-1} \le \frac{3}{20}\binom{n-1}{k-1}.$$

Therefore the number of sets that contain w is at least $|\mathcal{G}| - \frac{3}{20}\binom{n-1}{k-1} > \frac{4}{5}\binom{n-1}{k-1}$. Now applying Lemma 5.2 with $\mathcal{G}' = \mathcal{G}$ and $C = \{w\}$ we see that every set in $\mathcal{G}$ contains w, as required. ∎

6. Non-uniform systems

In this section we prove our results on non-uniform systems without a simplex or a cluster.

6.1. Lemmas for uniform families

In this subsection, we state and prove some results on the uniform problem. Some of our estimates are not exact, but suffice for later purposes. For 2-clusters we use the following result of the second author [28].

Theorem 6.1 ([28]). *Suppose $k \geq 3$, $n \geq 3k/2$ and $\mathcal{G}$ is a k-uniform set system on $[n]$ that does not contain a 2-cluster. Then $|\mathcal{G}| \leq \binom{n-1}{k-1}$.*

We can use this to derive a (non-exact) bound for d-clusters.

Lemma 6.2. *If $\mathcal{G}$ is a k-uniform system on $[n]$ then $k|\mathcal{G}| \leq \sum_{x\in[n]} |L_x(\mathcal{G})| + \binom{n}{k-1}$.*

Proof. $k|\mathcal{G}| = \sum_{x\in[n]} d(x) = \sum_{x\in[n]} (|L_x(\mathcal{G})| + |S_x(\mathcal{G})|) \leq \sum_{x\in[n]} |L_x(\mathcal{G})| + \binom{n}{k-1}$, the last inequality since each $(k-1)$-set is counted by at most one S_x. ∎

Lemma 6.3. *Suppose $d \geq 2$, $k \geq d+1$, $|k - n/2| < n^{3/5}$, n is large and $\mathcal{G}$ is a k-uniform set system on $[n]$ that does not contain a d-cluster. Then $|\mathcal{G}| \leq (1+5d/n)\binom{n-1}{k-1}$.*

Proof. We argue by induction on d. The base case $d = 2$ follows from Theorem 6.1. For the induction step, we apply the induction hypothesis to $L_x(\mathcal{G})$ for each $x \in [n]$, which has no $(d-1)$-cluster by Lemma 3.1 to get $|L_x(\mathcal{G})| \leq \left(1 + \frac{5(d-1)}{n-1}\right)\binom{n-2}{k-2}$. Then by Lemma 6.2 we have

$$\begin{aligned}
|\mathcal{G}| &\leq \frac{n}{k}\left(1 + \frac{5(d-1)}{n-1}\right)\binom{n-2}{k-2} + k^{-1}\binom{n}{k-1} \\
&= \left(\frac{kn-n}{kn-k}\left(1 + \frac{5(d-1)}{n-1}\right) + \frac{n}{k(n-k+1)}\right)\binom{n-1}{k-1} \\
&< \left(1 + \frac{5(d-1)}{n-1} + \frac{4}{n} + O(n^{-4/5})\right)\binom{n-1}{k-1} \leq (1+5d/n)\binom{n-1}{k-1}
\end{aligned}$$

as required. ∎

Next we give analogous bounds for strong simplices.

Lemma 6.4. *Suppose $n/2+2 \leq k < 2n/3$ and $\mathcal{G}$ is a k-uniform set system on $[n]$ that does not contain a strong 2-simplex. Then $|\mathcal{G}| \leq \binom{n-1}{k-1}$.*

This follows quickly from the following result of Frankl (see also [30]).

Theorem 6.5 (Frankl [13]). *Suppose $n/2 < k < 2n/3$ and $\mathcal{G}$ is a k-uniform set system on $[n]$ that does not contain a 2-simplex. Then $|\mathcal{G}| \leq \binom{n-1}{k-1}$.*

Proof of Lemma 6.4. Suppose $|\mathcal{G}| > \binom{n-1}{k-1}$. By Theorem 6.5 $\mathcal{G}$ contains a triangle A,B,C. Since there is no strong 2-simplex every set in $\mathcal{G}$ misses one of $A\cap B$, $B\cap C$, $C\cap A$. Each of these intersections has size ≥ 4, so we have $|\mathcal{G}| \leq 3\binom{n-4}{k} < \binom{n-1}{k-1}$, contradiction. ∎

The next lemma follows from Lemma 6.4 in the same way that Lemma 6.3 follows from Theorem 6.1.

Lemma 6.6. *Suppose $d \geq 2$, $n/2 + d \leq k < 2n/3$ and $\mathcal{G}$ is a k-uniform set system on $[n]$ that does not contain a strong d-simplex. Then $|\mathcal{G}| \leq (1+5d/n)\binom{n-1}{k-1}$.* ∎

Now we need some estimates when k is quite close to $n/2$.

Lemma 6.7. *Suppose n is sufficiently large, $n/2-n^{1/10} < k < n/2+n^{1/10}$ and $\mathcal{G}$ is a k-uniform set system on $[n]$ that does not contain a strong 2-simplex. Then $|\mathcal{G}| \leq (1+3n^{-1/4})\binom{n-1}{k-1}$.*

Our proof of Lemma 6.7 will use the following results of Frankl and Ahlswede–Khachatrian.

Lemma 6.8 (Frankl [14] Proposition 1.3). *Suppose we have a p-uniform system P and a q-uniform system Q on $[m]$ with $m > p+q$, $p \geq q$ and P, Q are cross-intersecting. Then $|P|+|Q| \leq \binom{m}{p}$.*

Theorem 6.9 (Ahlswede–Khachatrian [1]). *Suppose $n \geq k \geq t$. For each $0 \leq r \leq k-t$ define a family $\mathcal{F}(n,k,t,r) = \{A \subset [n]\colon |A| = k, |A \cap [t+2r]| \geq t+r\}$. There is some r for which $\mathcal{F}(n,k,t,r)$ is a maximum size k-uniform t-intersecting family on $[n]$.*

Proof of Lemma 6.7. Suppose $|\mathcal{G}| > (1+3n^{-1/4})\binom{n-1}{k-1}$. We will show that there $\mathcal{G}$ contains a 2-simplex A,B,C with each pairwise intersection size at least 4. The existence of a strong 2-simplex will then follow as in the Proof of Lemma 6.4.

Suppose that every two sets of $\mathcal{G}$ have more than $2n^{1/10}$ elements in common. Then by Theorem 6.9, we obtain $|\mathcal{G}| \leq |\mathcal{F}(n,k,t,r)|$ for some r where $t = 2n^{1/10}$. To estimate $|\mathcal{F}(n,k,t,r)|$, consider the hypergeometric random variable Y with parameters $(n,t+2r,k)$. Since $n/2-n^{1/10} < k < n/2+n^{1/10}$, we deduce that $\mathbb{E}Y < t+r$, so by equation (3),

$$|\mathcal{G}| < \mathbb{P}(Y > \mathbb{E}Y)\binom{n}{k} < \big(1/2 + O(n^{-0.49})\big)\binom{n}{k} < \big(1+3n^{-1/4}\big)\binom{n-1}{k-1},$$

where the last inequality follows by a short calculation using $k > n/2-n^{1/10}$. This contradiction implies that there are sets A and B in $\mathcal{G}$ so that $I = A\cap B$

satisfies $|I| \le 2n^{1/10}$. For the purpose of estimation let X be a hypergeometric random variable with parameters (n,k,k). Then $\mathbb{E}X = k^2/n$, so $|\mathbb{E}X - k/2| < n^{1/10}$. Let

$$\mathcal{G}_0 = \{C \in \mathcal{G} \colon ||C \cap A| - \mathbb{E}X| > n^{3/5}\} \cup \{C \in \mathcal{G} \colon ||C \cap B| - \mathbb{E}X| > n^{3/5}\},$$
$$\mathcal{G}_1 = \{C \in \mathcal{G} \colon ||C \cap A| - \mathbb{E}X| < 6n^{1/10}\} \cup \{C \in \mathcal{G} \colon ||C \cap B| - \mathbb{E}X| < 6n^{1/10}\},$$
$$\mathcal{G}_2 = \mathcal{G} \setminus (\mathcal{G}_0 \cup \mathcal{G}_1).$$

By equation (2) we have

$$\binom{n}{k}^{-1} |\mathcal{G}_0| < 2\mathbb{P}\big(|X - \mathbb{E}X| > n^{3/5}\big) < 4e^{-n^{1/5}/3},$$

and from equation (1) we have

$$\binom{n}{k}^{-1} |\mathcal{G}_1| < 2\mathbb{P}\big(|X - \mathbb{E}X| < 6n^{1/10}\big) < 24n^{1/10} \cdot O\big(n^{-1/2}\big) = O\big(n^{-2/5}\big),$$

so $|\mathcal{G}_2| > (1 + 2n^{-1/4})\binom{n-1}{k-1}$. Without loss of generality

$$\mathcal{G}_3 = \{C \in \mathcal{G}_2 \colon \mathbb{E}X + 6n^{1/10} < |C \cap A| < \mathbb{E}X + n^{3/5}\}$$

has size $|\mathcal{G}_3| \ge |\mathcal{G}_2|/2$.

Consider the families $\mathcal{G}_3(D) = \{C \setminus D \colon C \in \mathcal{G}_3, C \cap (A \setminus I) = D\}$ obtained by taking intersections of sets in $\mathcal{G}_3$ with $A \setminus I$. By definition of $\mathcal{G}_3$ we only consider sets D of size $|D| = d > \mathbb{E}X + 6n^{1/10} - |I| > \mathbb{E}X + 4n^{1/10}$. We claim that there must be some D for which

$$|\mathcal{G}_3(D)| > \big(1 + n^{-1/4}\big)\binom{n - 1 - k + |I|}{k - 1 - |D|}. \tag{6}$$

Otherwise, considering a hypergeometric random variable Y with parameters $(n-1, k-|I|, k-1)$ we get the contradiction

$$\begin{aligned}
\frac{1}{2}\big(1 + 2n^{-1/4}\big) &< \binom{n-1}{k-1}^{-1} |\mathcal{G}_3| \\
&\le \binom{n-1}{k-1}^{-1} \sum_{d > \mathbb{E}X + 4n^{1/10}} \binom{k - |I|}{d} \big(1 + n^{-1/4}\big) \binom{n - 1 - k + |I|}{k - 1 - |D|} \\
&= \big(1 + n^{-1/4}\big)\mathbb{P}\big(Y > \mathbb{E}X + 4n^{1/10}\big) \\
&< \big(1 + n^{-1/4}\big)\mathbb{P}(Y > \mathbb{E}Y) = \big(1 + n^{-1/4}\big)\big(1/2 + O(n^{-0.49})\big).
\end{aligned}$$

Fix a set D satisfying equation (6) and for each $J \subset I$ consider $\mathcal{F}_J = \{E \subset [n] \setminus A\colon E \cup J \cup D \in \mathcal{G}_3\}$. If we can find $E \in \mathcal{F}_J$, $E' \in \mathcal{F}_{I\setminus J}$ with $E \cap E' = \emptyset$ for some J, then B, $E \cup J \cup D$, $E' \cup (I \setminus J) \cup D$ is a 2-simplex, in which the pairwise intersection sizes are at least 4 (by far!), and then we are done, as noted at the beginning of the proof. Otherwise, since $|D| > \mathbb{E}X + 4n^{1/10} > k/2 + 3n^{1/10}$ we have $(n-k) - 2(k-|D|) > n - 2k + 6n^{1/10} > 0$ and we can apply Lemma 6.8 to see that

$$|\mathcal{F}_J| + |\mathcal{F}_{I\setminus J}| < \binom{n-k}{k-|D|}.$$

Since this holds for each $J \subset I$ we have

$$\begin{aligned}
|\mathcal{G}_3(D)| &< \frac{1}{2}\sum_{J\subset I}\binom{n-k}{k-|D|} \\
&= 2^{|I|-1}\binom{n-1-k+|I|}{k-1-|D|}\frac{n-2k+|D|+1}{k-|D|}\prod_{j=1}^{|I|-1}\frac{n-2k+|I|+|D|+1-j}{n-k+|I|-j} \\
&= 2^{|I|-1}\binom{n-1-k+|I|}{k-1-|D|}\big(1+O(n^{-9/10})\big)\big(1/2+O(n^{-9/10})\big)^{|I|-1} \\
&= \big(1+O(n^{-4/5})\big)\binom{n-1-k+|I|}{k-1-|D|}.
\end{aligned}$$

This contradiction with equation (6) completes the proof. ∎

Lemma 6.10. *Suppose $d \geq 2$, n is sufficiently large, $n/2 - n^{1/10} + d < k < n/2 + n^{1/10} - d$ and $\mathcal{G}$ is a k-uniform set system on $[n]$ that does not contain a strong d-simplex. Then $|\mathcal{G}| \leq (1+4n^{-1/4})\binom{n-1}{k-1}$.*

Proof. We show by induction on d that $|\mathcal{G}| \leq (1+3n^{-1/4}+5d/n)\binom{n-1}{k-1}$, from which the stated bound follows. The base case $d=2$ follows from Lemma 6.7. For the induction step, we apply the induction hypothesis to $L_x(\mathcal{G})$ to get $|L_x(\mathcal{G})| \leq (1+3n^{-1/4}+5(d-1)/(n-1))\binom{n-2}{k-2}$ for each $x \in [n]$. Then by Lemma 6.2 we have

$$\begin{aligned}
|\mathcal{G}| &\leq \frac{n}{k}\left(1+3n^{-1/4}+\frac{5(d-1)}{(n-1)}\right)\binom{n-2}{k-2} + k^{-1}\binom{n}{k-1} \\
&= \left(\frac{kn-n}{kn-k}\left(1+3n^{-1/4}+\frac{5(d-1)}{(n-1)}\right)+\frac{n}{k(n-k+1)}\right)\binom{n-1}{k-1} \\
&\leq \big(1+3n^{-1/4}+5d/n\big)\binom{n-1}{k-1},
\end{aligned}$$

as required. ∎

6.2. Stability for non-uniform systems

In this subsection, we prove the following stability result for non-uniform set systems without strong simplices or clusters.

Theorem 6.11. *Suppose $d \geq 2$, n is sufficiently large, $\mathcal{G}$ is a set system on $[n]$, $|\mathcal{G}| > (1-2n^{-5/8})2^{n-1}$, and either $\mathcal{G}$ does not contain a strong d-simplex or $\mathcal{G}$ does not contain a d-cluster. Then there is some x in $[n]$ so that every set $A \in \mathcal{G}$ with $||A|-n/2| < n^{2/3}$ contains x. In particular $|\mathcal{G}| < 2^{n-1}+2^{n-n^c}$ for any constant $c<1/3$ (by the Chernoff bound).*

Proof. Let δ and N be the outputs of Theorem 2.1 with inputs $\zeta = 1/4$ and d. Let N' be the output of Lemma 5.2 with inputs $\zeta' = 1/4$ and d. Choose n sufficiently large that all calculations involving $O(\cdot)$ estimates in the following proof hold, and $n > \max\{N, N', (30/\delta)^8\}$.

We partition $\mathcal{G}$ according to various intervals of set sizes. Recall that for an interval I we write $\mathcal{G}^I = \{A \in \mathcal{G} \colon |A| \in I\}$. We use the intervals

$$\begin{gathered}\big[0, n/2-n^{3/5}\big], \quad \big[n/2-n^{3/5}, n/2-n^{1/10}+d\big], \quad \big[n/2-n^{1/10}+d, n/2+d\big], \\ \big[n/2+d, n/2+n^{3/5}\big], \quad \big[n/2+n^{3/5}, n\big].\end{gathered}$$

By Chernoff bounds we have $|\mathcal{G}^{[0,n/2-n^{3/5}]}| + |\mathcal{G}^{[n/2+n^{3/5},n]}| < 2^{0.99n}$. Also, applying Lemma 6.6 if $\mathcal{G}$ has no strong d-simplex or Lemma 6.3 if $\mathcal{G}$ has no d-cluster we have $|\mathcal{G}^{[n/2+2,n/2+n^{3/5}]}| \leq \sum_{k=n/2+d}^{n/2+n^{3/5}} (1+5d/n)\binom{n-1}{k-1} < (1+5d/n)2^{n-2}$. Next, applying Lemma 6.10 if $\mathcal{G}$ has no strong d-simplex or Lemma 6.3 if $\mathcal{G}$ has no d-cluster we have

$$\big|\mathcal{G}^{[n/2-n^{1/10}+d,n/2+d]}\big| < \big(1+4n^{-1/4}\big) \sum_{k=n/2-n^{1/10}+d}^{n/2+d} \binom{n-1}{k-1}.$$

We deduce that

$$\begin{aligned}\big|\mathcal{G}^{[n/2-n^{3/5},n/2-n^{1/10}+d]}\big| &> \big(1-2n^{-5/8}\big)2^{n-1} - 2^{0.99n} - (1+5d/n)2^{n-2} \\ &\quad - \big(1+4n^{-1/4}\big) \sum_{k=n/2-n^{1/10}+d}^{n/2+d} \binom{n-1}{k-1} \\ &> \big(1-5n^{-5/8}\big) \sum_{k<n/2-n^{1/10}+d} \binom{n-1}{k-1}.\end{aligned}$$

This last inequality is rather delicate, and perhaps the reader will find it helpful if we point out that in the final term the factor $4n^{-1/4}$ can be neglected

as it belongs to a contribution of order $O(n^{-1/4+1/10-1/2}) = O(n^{-13/20}) < O(n^{-5/8})$.

Note that $2^{1-n}\sum_{k<n/2-n^{1/10}+d}\binom{n-1}{k-1} = 1/2 - O(n^{-2/5})$. Now there must be some $k = n/2 - t$ satisfying $n^{1/10} - d < t < n^{3/5}$ with $|\mathcal{G}^k| > (1-\delta t/n)\binom{n-1}{k-1}$. Otherwise, using equation (4) and $n > (30/\delta)^8$, we would have the contradiction

$$\begin{aligned}\left(1-3n^{-5/8}\right)\sum_{k<n/2-n^{1/10}+d}\binom{n-1}{k-1} &< \left|\mathcal{G}^{[n/2-n^{3/5},n/2-n^{1/10}+d]}\right| \\ &< \sum_{t=n^{1/10}-d}^{n^{3/5}}(1-\delta t/n)\binom{n-1}{n/2-t-1} \\ &< \sum_{k<n/2-n^{1/10}+d}\binom{n-1}{k-1} - \frac{\delta}{10}n^{-1/2}\sum_{k<n/2-n^{1/10}+d}\binom{n-1}{k-1}.\end{aligned}$$

By the choice of n, we may apply Theorem 2.1 with inputs $\zeta = 1/4$ and d and conclude that there is some point x contained in every set of $\mathcal{G}^k$. Now we see that x belongs to every set in $\mathcal{G}^{[n/2-n^{2/3},n/2+n^{2/3}]}$. For otherwise (again by the choice of n) we can apply Lemma 5.2 with $C = \{x\}$ to see that $|\mathcal{G}^k| < \frac{4}{5}\binom{n-1}{k-1}$, which contradicts the choice of k above. (Note that this is the only place in the proof where we need to consider non-uniform clusters or simplices.) This completes the proof. ∎

Note that Theorem 2.4 is an immediate consequence, as a system with no d-simplex certainly contains no strong d-simplex.

6.3. Non-uniform systems: the exact result

Now we can use our stability result to deduce an exact result.

Proof of Theorem 2.3. We argue by induction on d. The case $d = 2$ is an unpublished theorem of Milner (see discussion and proof in [30]), and actually our argument will also prove this base case.

Suppose $d \geq 2$ and n_{d-1} is such that, for $n \geq n_{d-1}$, any set system on $[n]$ that does not contain a $(d-1)$-simplex has at most $2^{n-1} + \sum_{i=0}^{d-2}\binom{n-1}{i}$ sets. Fix a large enough number n_d so that all following inequalities are true. Now suppose $n \geq n_d$, $\mathcal{G}$ is a set system on $[n]$ that does not contain a d-simplex and $|\mathcal{G}| \geq 2^{n-1}$. By Theorem 6.11 there is a point x so that all but at most $2^{n-n_{d-1}-1}$ sets of $\mathcal{G}$ contain x.

It is enough to show that any set in $\mathcal{G}$ that does not contain x has size at most $d-1$. Suppose for a contradiction that $x \notin A \in \mathcal{G}$ and $|A| \geq d$. First consider the case when $|A| < n_{d-1}$. Fix an arbitrary partition of A into d non-empty parts $B_1, \ldots, B_d$. If we could find sets $A_1, \ldots, A_d$ in $\mathcal{G}$ so that $x \in A_i$ and $A_i \cap A = A \setminus B_i$ for $1 \leq i \leq d$ then $A, A_1, \ldots, A_d$ would be a d-simplex. Since $\mathcal{G}$ does not contain a d-simplex there must be some $1 \leq i \leq d$ such that $\mathcal{G}$ does not contain any of the $2^{n-1-|A|}$ sets with $x \in A_i$ and $A_i \cap A = A \setminus B_i$. But there are at most $2^{n-n_{d-1}-1}$ sets of $\mathcal{G}$ that do not contain x, so $|\mathcal{G}| \leq 2^{n-1} - 2^{n-1-|A|} + 2^{n-n_{d-1}-1} < 2^{n-1}$, contradiction.

Next consider the case $|A| > n_{d-1}$. We break up the sets in $\mathcal{G}$ containing x according to their intersection with $[n] \setminus A$: for each B with $x \in B$ and $B \cap A = \emptyset$ we let $\mathcal{G}(B) = \{C \subset A \colon C \cup B \in \mathcal{G}\}$. Now $\mathcal{G}(B)$ cannot contain a $(d-1)$-simplex $C_1, \ldots, C_d$, for then $A, C_1 \cup B, \ldots, C_d \cup B$ would be a d-simplex in $\mathcal{G}$. Applying the induction hypothesis for $d \geq 3$, or the fact that an intersecting family of subsets of A has size $\leq 2^{|A|-1}$ for $d = 2$, we have $|\mathcal{G}(B)| \leq 2^{|A|-1} + \sum_{i=0}^{d-2} \binom{|A|-1}{i}$ for each B. Therefore

$$|\mathcal{G}| \leq 2^{n-1-|A|} \left(2^{|A|-1} + \sum_{i=0}^{d-2} \binom{|A|-1}{i} \right) + 2^{n-n_{d-1}-1} < 2^{n-1}.$$

This contradiction shows that any set in $\mathcal{G}$ that does not contain x has size at most $d-1$, so we are done. ∎

7. Concluding remarks

We make the following conjecture, which would substantially strengthen Theorem 2.1.

Conjecture 7.1. Fix $d \geq 2$ and $\zeta > 0$. Suppose $\mathcal{G}$ is a k-uniform set system on $[n]$, where $\zeta n < k < n/2$, n is sufficiently large, and either $\mathcal{G}$ contains no strong d-simplex or $\mathcal{G}$ contains no d-cluster. If $|\mathcal{G}| > (1+\zeta)\binom{n-2}{k-2}$, then $\mathcal{G}$ is a star.

If true, Conjecture 7.1 would be essentially sharp for both problems. For the strong simplex problem, we can take $\mathcal{G}$ to be all sets containing two specified elements a, b, together with two disjoint sets A, B with $a \in A$ and $b \in B$. Then $|\mathcal{G}| = \binom{n-2}{k-2} + 2$, it contains no strong d-simplex for $d > 1$, and it is not a star. For the d-cluster problem, we can let $\mathcal{G}' = \mathcal{G} - \mathcal{H}$, where $\mathcal{H}$ consists of all sets of $\mathcal{G}$ containing a, b and lying within $A \cup B$. Since $n > (2+\zeta')k$, we conclude that $|\mathcal{H}|$ is exponentially (in k) smaller than $|\mathcal{G}|$. Therefore, $|\mathcal{G}'| > (1-\zeta)\binom{n-2}{k-2}$, it contains no d-cluster for $d > 1$, and it is not a star.

We end with the following ambitious conjecture which simultaneously strengthens Conjectures 1.1 and 1.2. Call a k-uniform collection of $d+1$ sets a d-cluster-simplex if it is both a d-cluster and a d-simplex.

Conjecture 7.2. Suppose $k \geq d+1 > 2$, $n > k(d+1)/d$ and $\mathcal{G}$ is a k-uniform set system on $[n]$ with no d-cluster-simplex. Then $|\mathcal{G}| \leq \binom{n-1}{k-1}$, with equality only for a star.

References

[1] R. Ahlswede and L. H. Khachatrian: The complete intersection theorem for systems of finite sets, *European J. Combin.* **18** (1997), 125–136.

[2] R. P. Anstee and P. Keevash: Pairwise intersections and forbidden configurations, *European J. Combin.* **27** (2006), 1235–1248.

[3] T. M. Apostol: An elementary view of Euler's summation formula, *Amer. Math. Monthly* **106** (1999), 409–418.

[4] J. Balogh, B. Bollobás and M. Simonovits: The number of graphs without forbidden subgraphs, *J. Combin. Theory Ser. B* **91** (2004), 1–24.

[5] J. C. Bermond and P. Frankl: On a conjecture of Chvátal on m-intersecting hypergraphs, *Bull. London Math. Soc.* **9** (1977), 310–312.

[6] V. Chvátal: An extremal set-intersection theorem, *J. London Math. Soc.* **9** (1974/75), 355–359.

[7] R. Csákány and J. Kahn: A homological approach to two problems on finite sets, *J. Algebraic Combin.* **9** (1999), 141–149.

[8] I. Dinur and E. Friedgut: Intersecting families are essentially contained in juntas, *Combin. Probab. Comput.* **18(1-2)** (2009), 107–122.

[9] I. Dinur and S. Safra: On the hardness of approximating minimum vertex cover, *Ann. of Math. (2)* **162** (2005), 439–485.

[10] P. Erdős: Topics in combinatorial analysis, in: *Proc. Second Louisiana Conf. on Comb., Graph Theory and Computing (R. C. Mullin et al., eds.)*, pp. 2–20, Louisiana State Univ., Baton Rouge, 1971.

[11] P. Erdős, H. Ko and R. Rado: Intersection theorems for systems of finite sets, *Quart. J. Math. Oxford Ser.* **12** (1961), 313–320.

[12] P. Frankl: On Sperner families satisfying an additional condition, *J. Combin. Theory Ser. A* **20** (1976), 1–11.

[13] P. Frankl: On a problem of Chvátal and Erdős on hypergraphs containing no generalized simplex, *J. Combin. Theory Ser. A* **30** (1981), 169–182.

[14] P. Frankl: Erdős–Ko–Rado theorem with conditions on the maximal degree, *J. Combin. Theory Ser. A* **46** (1987), 252–263.

[15] P. Frankl and Z. Füredi: A new generalization of the Erdős–Ko–Rado theorem, *Combinatorica* **3(3-4)** (1983), 341–349.

[16] P. Frankl and Z. Füredi: Exact solution of some Turán-type problems, *J. Combin. Theory Ser. A* **45** (1987), 226–262.

[17] E. Friedgut: On the measure of intersecting families, uniqueness and stability; *Combinatorica* **28(5)** (2008), 503–528.

[18] Z. Füredi, O. Pikhurko and M. Simonovits: On Triple Systems with Independent Neighborhoods, *Combin. Probab. Comput.* **14** (2005), 795–813.
[19] Z. Füredi, O. Pikhurko and M. Simonovits: 4-Books of three pages, *J. Combin. Theory Ser. A* **113** (2006), 882–891.
[20] Z. Füredi and M. Simonovits: Triple systems not containing a Fano configuration, *Combin. Probab. Comput.* **14** (2005), 467–484.
[21] A. J. W. Hilton and E. C. Milner: Some intersection theorems for systems of finite sets, *Quart. J. Math. Oxford Ser.* **18** (1967), 369–384.
[22] S. Janson, T. Łuczak and A. Ruciński: *Random Graphs*, Wiley, (2000).
[23] P. Keevash and D. Mubayi: Stability results for cancellative hypergraphs, *J. Comb. Theory Ser. B* **92** (2004), 163–175.
[24] P. Keevash and B. Sudakov: The Turán number of the Fano plane, *Combinatorica* **25(5)** (2005), 561–574.
[25] P. Keevash and B. Sudakov: On a hypergraph Turán problem of Frankl, *Combinatorica* **25(6)** (2005), 673–706.
[26] D. Mubayi: Structure and Stability of Triangle-free set systems, *Trans. Amer. Math. Soc.* **359** (2007), 275–291.
[27] D. Mubayi: An intersection theorem for four sets, *Advances in Mathematics* **215(2)** (2007), 601–615.
[28] D. Mubayi: Erdős–Ko–Rado for three sets, *J. Combin. Theory Ser. A* **113** (2006), 547–550.
[29] D. Mubayi and O. Pikhurko: A new generalization of Mantel's theorem to k-graphs, *J. Combin. Theory Ser. B* **97(4)** (2007), 669–678.
[30] D. Mubayi and J. Verstraëte: Proof of a conjecture of Erdős on triangles in set-systems, *Combinatorica* **25(5)** (2005), 599–614.
[31] D. Mubayi and J. Verstraëte: Minimal paths and cycles in set systems, *European J. Combin.* **28(6)** (2007), 1681–1693.
[32] O. Pikhurko: Exact computation of the hypergraph Turán function for expanded complete 2-graphs, accepted, *J. Combin. Theory Ser. B* (publication suspended for an indefinite time, see http://www.math.cmu.edu/~pikhurko/Copyright.html).
[33] O. Pikhurko: An exact Turán result for the generalized triangle, *Combinatorica* **28(2)** (2008), 187–208.
[34] H. Robbins: A remark on Stirling's formula, *Amer. Math. Monthly* **62** (1955), 26–29.
[35] M. Simonovits: A method for solving extremal problems in graph theory, stability problems; in: *1968 Theory of Graphs (Proc. Colloq., Tihany, 1966)*, pp. 279–319, Academic Press, New York.

A. Proofs of hypergeometric estimates

Here we derive the hypergeometric estimates quoted in subsection 3.3. Perhaps they already appear in the literature, but we could not find a reference, so we will deduce them from the following 'Stirling Formula' inequality of Robbins [34]:

$$n\log n - n + \frac{1}{2}\log(2\pi n) - 1/(12n+1) < \log n! < n\log n - n + \frac{1}{2}\log(2\pi n) - 1/(12n).$$

Proof of equation (1)**.** Recall that we must prove

$$\mathbb{P}(X = \mathbb{E}X + t) = (2\pi pqrsn)^{-1/2} e^{-t^2/2pqrsn + O(t/n + t^3/n^2)}. \tag{7}$$

We have

$$\begin{aligned}\mathbb{P}(X = \mathbb{E}X + t) &= \frac{\binom{rn}{rpn+t}\binom{sn}{spn-t}}{\binom{n}{pn}} \\ &= \frac{(rn)!}{(prn+t)!(qrn-t)!}\frac{(sn)!}{(psn-t)!(qsn+t)!}\frac{(pn)!(qn)!}{n!}.\end{aligned}$$

To estimate this we take logs and group all terms according to four parts from the Robbins inequality: (I) $n\log n$, (II) $-n$, (III) $\frac{1}{2}\log(2\pi n)$ and (IV) $1/(12n+1)$ or $1/(12n)$ for lower/upper bounds respectively.

(I) The $n\log n$ contribution to $\log\mathbb{P}(X=\mathbb{E}X+t)$ is

$$\begin{aligned}&rn(\log r + \log n) - (prn+t)\big(\log p + \log r + \log n + \log(1 + t/prn)\big) \\ &- (qrn - t)\big(\log q + \log r + \log n + \log(1 - t/qrn)\big) + sn(\log s + \log n) \\ &- (psn - t)\big(\log p + \log s + \log n + \log(1 - t/psn)\big) \\ &- (qsn + t)\big(\log q + \log s + \log n + \log(1 + t/qsn)\big) \\ &+ pn(\log p + \log n) + qn(\log q + \log n) - n\log n.\end{aligned}$$

All terms cancel apart from

$$\begin{aligned}&-(prn+t)\log(1+t/prn) - (qrn-t)\log(1-t/qrn) \\ &-(psn-t)\log(1-t/psn) - (qsn+t)\log(1+t/qsn).\end{aligned}$$

Expanding the logs using the series $\log(1+x) = x - x^2/2 + O(x^3)$ we get a contribution

$$-\frac{t^2}{2n}\left(\frac{1}{pr} + \frac{1}{qr} + \frac{1}{ps} + \frac{1}{qs}\right) + O(t^3/n^2) = -t^2/2pqrsn + O(t^3/n^2).$$

Here, and throughout all subsequent estimates, the constant in the $O(\cdot)$ term is uniformly bounded provided that p, q, r, s are bounded away from 0.

(II) The $-n$ contribution to $\log\mathbb{P}(X=\mathbb{E}X+t)$ is 0.

(III) The $\frac{1}{2}\log(2\pi n)$ contribution to $\log\mathbb{P}(X=\mathbb{E}X+t)$ is $1/2$ times

$$\begin{aligned}&\log 2\pi + \log r + \log n - \big(\log 2\pi + \log p + \log r + \log n + \log(1 + t/prn)\big) \\ &- \big(\log 2\pi + \log q + \log r + \log n + \log(1 - t/qrn)\big) + \log 2\pi + \log s + \log n \\ &- \big(\log 2\pi + \log p + \log s + \log n + \log(1 - t/psn)\big) \\ &- \big(\log 2\pi + \log q + \log s + \log n + \log(1 + t/qsn)\big) \\ &+ \log 2\pi + \log p + \log n + \log 2\pi + \log q + \log n - (\log 2\pi + \log n).\end{aligned}$$

Simplifying and expanding the logs in series we get a contribution

$$-\frac{1}{2}(\log 2\pi + \log p + \log q + \log r + \log s + \log n) + O(t/n).$$

(IV) The $1/12n$ contribution to $\log \mathbb{P}(X = \mathbb{E}X + t)$ is $O(1/n)$.

Putting together the estimates (I) to (IV) we obtain equation (1). ∎

Before continuing we recall some well-known integrals pertaining to the normal distribution:

$$\int_0^\infty e^{-x^2/2} = \sqrt{\pi/2}, \qquad \int_0^\infty x e^{-x^2/2} = 1.$$

Proof of equation (3)**.** We use the Euler–Maclaurin summation formula (see [3]), which is as follows. Suppose $f(t)$ is a smooth function and a is a natural number. Then $I = \int_0^a f(t)$ can be approximated by $S = \frac{1}{2}f(0) + f(1) + \cdots + f(a-1) + \frac{1}{2}f(a)$ with error $|S - I| < \int_0^a |f'(t)|\, dt$. Write

$$\mathbb{P}(X \geq \mathbb{E}X) = \sum_{t \geq 0} \mathbb{P}(X = \mathbb{E}X + t).$$

By equation (2) we can truncate the sum at $t = n^{1/2+\epsilon/3}$ with an error $\exp(-O(n^{2\epsilon/3}))$. We will apply the Euler–Maclaurin formula with

$$f(t) = (2\pi pqrsn)^{-1/2} e^{-t^2/2pqrsn},$$

and $a = n^{1/2+\epsilon/3}$. Halving the first and last terms incurs $O(n^{-1/2})$ error by equation (1). Also by equation (1), in using $f(t)$ to approximate $\mathbb{P}(X = \mathbb{E}X + t)$ we incur relative error $O(t/n + t^3/n^2) = O(n^{-1/2+\epsilon})$. (By relative error we mean that the absolute error is obtained by multiplying by the final estimate. This will turn out to be $1/2$, so the absolute error is also $O(n^{-1/2+\epsilon})$.) Therefore $|\mathbb{P}(X \geq \mathbb{E}X) - S| = O(n^{-1/2+\epsilon})$.

To approximate by I we also incur an error

$$\begin{aligned} |S - I| < \int_0^a |f'(t)|\, dt &\leq (2pqrsn)^{-3/2} \int_0^\infty 2te^{-t^2/2pqrsn}\, dt \\ &\leq (pqrsn)^{-1/2} \int_0^\infty 2xe^{-x^2/2}\, dx = O\big(n^{-1/2}\big), \end{aligned}$$

where we substitute $t = (pqrsn)^{1/2}x$. Finally we can extend the range of integration to infinity with error $\exp(-O(n^{2\epsilon/3}))$ by applying the estimate

of equation (2), which is also valid for the normal distribution. We have thus succeeded by approximating $\mathbb{P}(X \geq \mathbb{E}X)$ to error $O(n^{-1/2+\epsilon})$ by

$$\int_0^\infty (2\pi pqrsn)^{-1/2} e^{-t^2/2pqrsn}\, dt = \int_0^\infty (2\pi)^{-1/2} e^{-x^2/2}\, dx = 1/2.$$

This gives the required estimate. ∎

Proof of equation (4). By similar (and much simpler) calculations to those in the proof of equation (1) we can estimate

$$2^{-n}\binom{n}{n/2+t} = (\pi n/2)^{-1/2} e^{-2t^2/n + O(t/n + t^3/n^2)}. \tag{8}$$

Consider the expression which it is required to estimate:

$$E = 2^{1-n} \sum_{t=n^{1/10}-d}^{n^{3/5}} \frac{t}{n}\binom{n-1}{n/2-t-1}.$$

First we extend the sum down to $t=0$ with an error $\sum_{t=0}^{n^{1/10}} \frac{t}{n} O(n^{-1/2}) < O(1/n)$ using equation (8). Next we use equation (8) to replace the sum by

$$\sum_{t=0}^{n^{3/5}} (\pi n/2)^{-1/2} \frac{t}{n} e^{-2t^2/n},$$

with relative error $O(t/n + t^3/n^2) = O(n^{-1/5})$ (an absolute error of $O(n^{-7/10})$). Applying the Euler–Maclaurin formula (checking the error estimates as in the proof of equation (3)) and substituting $t = xn^{1/2}$ we approximate by

$$\int_0^{n^{3/5}} (\pi n/2)^{-1/2} \frac{t}{n} e^{-2t^2/n}\, dt = \int_0^{n^{1/10}} (\pi n/2)^{-1/2} x e^{-2x^2}\, dx.$$

Then extending the range of integration to ∞ we approximate by $(\pi n/2)^{-1/2} \cdot 1/4$. In particular we have a lower bound of $\frac{1}{10} n^{-1/2}$ for large n. ∎

Peter Keevash

School of Mathematical Sciences
Queen Mary, University of London
Mile End Road, Room B14
London E1 4NS
UK
p.keevash@qmul.ac.uk

Dhruv Mubayi

Department of Mathematics, Statistics, & Computer Science
University of Illinois at Chicago
851 S. Morgan Street
Chicago, IL 60607-7045
USA
mubayi@math.uic.edu