Submitted to the Annals of Probability

LOWER TAILS VIA RELATIVE ENTROPY

BY GADY KOZMA[1] AND WOJCIECH SAMOTIJ[2]

[1] *Department of Mathematics, The Weizmann Institute of Science, gady.kozma@weizmann.ac.il*

[2] *School of Mathematical Sciences, Tel Aviv University, samotij@tauex.tau.ac.il*

We show that the naive mean-field approximation correctly predicts the leading term of the logarithmic lower tail probabilities for the number of copies of a given subgraph in $G(n,p)$ and of arithmetic progressions of a given length in random subsets of the integers in the entire range of densities where the mean-field approximation is viable.

Our main technical result provides sufficient conditions on the maximum degrees of a uniform hypergraph $\mathscr{H}$ that guarantee that the logarithmic lower tail probabilities for the number of edges induced by a binomial random subset of the vertices of $\mathscr{H}$ can be well-approximated by considering only product distributions. This may be interpreted as a weak, probabilistic version of the hypergraph container lemma that is applicable to all sparser-than-average (and not only independent) sets.

1. Introduction. This paper is concerned with the phenomenon that, in many cases, conditioning on an atypical event leads to a mixture of product measures. An emblematic example is the family of n-vertex graphs with no triangles. It is clear that if one divides $[\![n]\!] := \{1,\dots,n\}$ into two parts and takes only edges with one endpoint in each part, the resulting graph has no triangles. The classical result of Erdős, Kleitman, and Rothschild [18] states that the vast majority of triangle-free graphs have such simple structure. In other words, if we *condition* the random graph $G(n,\frac{1}{2})$ to have no triangles, the resulting measure can be approximated by the following process: First, choose a random partition of the vertices into two parts (according to a measure that strongly favours partitions into approximately equal parts). Then, choose the edges randomly and independently, with edges between the parts having probability $\frac{1}{2}$ and edges inside the parts having probability 0. Since, conditioned on the partition, the measure becomes a product measure, the overall process is called a mixture of product measures.

The aim of this work is to establish sufficient conditions for such a phenomenon to occur in the context of large deviations for subgraph counts in the binomial random graph $G(n,p)$. Here, the seminal work of Chatterjee and Varadhan [9] has clarified that there are in fact two independent steps involved. The first is to show that the distribution of the random graph conditioned on a tail event can be described by a (small) mixture of product measures. The second is to describe the relevant measures, which, as it turns out, are those among all product measures (essentially) supported on the relevant tail event that have the least entropic cost. The main result of [9] completes the first of these two steps and can be summarised[1] as follows: Denote by X_n the number of copies of a given graph in the binomial random graph $G(n,p)$. If the edge probability p is fixed and n tends to infinity, then

$$-\log \mathbb{P}\big(X_n \geqslant (1+\delta)\mathbb{E}[X_n]\big) = (1+o(1))\cdot \Phi_{n,p}(\delta), \tag{1}$$

where $\Phi_{n,p}(\delta)$ is the least entropic cost of a product measure supported on the upper tail event (we will give a formal definition below); the analogous result holds for the lower tail. (Here

MSC2020 subject classifications: Primary 60F10.

Keywords and phrases: Lower tail, subgraph counts, random graph.

[1] The setting of [9] is more general, but let us not state the full generality of that paper here.

and throughout the paper, $\log$ denotes the natural logarithm.) As for the second step, the problem of calculating $\Phi_{n,p}(\delta)$ turned out to be very difficult. Even in the seemingly simple case of triangle counts, only partial results are known [26, 34]. *In this paper, we address only the first step, namely, obtaining an identity akin to* (1).

A substantial drawback of the approach taken by [9], which is based on Szemerédi's regularity lemma, is that it does not extend to sparse random graphs. (One may instead use the so-called weak regularity lemma of Frieze and Kannan [19], but this allows one to extend (1) only to the regime $p \geqslant (\log n)^{-c}$, for some small positive constant c, see [27].) This was first rectified by the breakthrough work of Chatterjee and Dembo [8], who developed a general technique for computing large deviation probabilities of nonlinear functions of independent Bernoulli random variables, such as subgraph counts in $G(n,p)$. In the context of subgraph counts in $G(n,p)$, the general result of [8] implies that (1) continues to hold as long as $p \geqslant n^{-\alpha}$ for some $\alpha > 0$ that depends only on the graph whose copies are counted.

The paper of Chatterjee and Dembo inspired a series of further developments. Their general technique was further simplified and strengthened by Eldan [17]. In the context of upper tails for subgraph counts in $G(n,p)$, the range of validity of the approximation (1) was further extended by the works of Augeri [1] (for cycles), of Cook and Dembo [11] (for arbitrary graphs), and of Cook, Dembo, and Pham [12] (for arbitrary graphs and, more generally, arbitrary uniform hypergraphs). The expression $\Phi_{n,p}(\delta)$ in the right-hand side of (1) was computed in the range $n^{-1/\Delta} \ll p \ll 1$, where Δ is the maximum degree of the graph for cliques [27] and, subsequently, for arbitrary subgraphs [6]. A very different, combinatorial technique for computing upper tail probabilities of polynomials of independent Bernoulli random variables was recently developed by Harel, Mousset, and Samotij [20]. This technique was used to resolve the upper tail problem completely for cliques [20] and, subsequently, for all regular graphs [5]. More precisely, these works showed that the approximation (1) is valid in the entire range of densities p where it was expected to hold.

Let us stress that all of the works on large deviations of subgraph counts in sparse random graphs mentioned above were primarily concerned with the upper tail. (In fact, the techniques developed in both [1] and [20] are inapplicable to the lower tail problem.) Historically, the upper tail problem is considered to be the more difficult of the two. Whereas Janson, Łuczak, and Ruciński [23] determined the logarithm of the lower tail probability up to a multiplicative constant, for every graph and all densities p, already in the late 1980s, the order of magnitude of the logarithm of the upper tail probability in the special case of triangle counts was determined only around ten years ago [7, 16].

In this paper, we offer a new, entropy-based approach to the large deviation problem that is particularly effective in estimating lower tails. The idea of using entropy estimates for studying nonlinear large deviations was first used in [25] (that paper is a few years older than the current one, unlike what one might think by examining arXiv submission dates). Ultimately it stems from Avez's entropy approach to study random walks and amenability, see [2]. A straightforward corollary of our main technical result is that the analogue of (1) holds for counts of arbitrary subgraphs in $G(n,p)$ in the entire range where such an approximation was expected to be valid.

1.1. *New results.* We start with a special case of our result for triangles and $p = \frac{1}{2}$. Of course, this case is mostly covered by [9], but we will get to values of p not covered by the literature in Theorem 2 below. We first state the minimisation problem that formalises the phrase 'least entropic cost' in this setting. Given a function $q \colon \binom{[\![n]\!]}{2} \to [0,1]$, let $G(n,q)$ denote the random graph obtained by retaining each edge e of K_n independently with probability q_e. For each $t \geqslant 0$, define

$$\mathscr{Q}_t := \left\{ q \in [0,1]^{\binom{[\![n]\!]}{2}} : \mathbb{E}\left[N_{K_3}\big(G(n,q)\big)\right] \leqslant t \right\},$$

where $N_{K_3}(G)$ is the number of triangles in G, and

$$\Phi_n(t) := \min\left\{ \sum_{e \in \binom{[\![n]\!]}{2}} \Big(q_e \log q_e + (1-q_e)\log(1-q_e) + \log 2 \Big) : q \in \mathscr{Q}_t \right\}. \tag{2}$$

Note that $q\log q + (1-q)\log(1-q) + \log 2$ is the difference in entropies of Bernoulli random variables with success probabilities $\frac{1}{2}$ and q.

THEOREM 1. *Let X_n denote the number of triangles in $G(n, \frac{1}{2})$. For every n and every $t \geqslant 0$,*

$$\log \mathbb{P}(X_n \leqslant t) \leqslant -\Phi_n(t + n^{23/8}) + 2n^{15/8}. \tag{3}$$

REMARK. Note that the two error terms in the above estimate are better than $o(n^3)$ and $o(n^2)$, respectively, whereas it is not difficult to verify that $\Phi_n(t) \geqslant \Omega(n^2)$ whenever $t \leqslant \mathbb{E}[X_n] - \Omega(n^3)$, see Lemma 21. In the remainder of this paper, we follow the literature and prove results with error terms in the corresponding estimates of the lower tail probabilities being inexplicit, but here we made an exception.

We now formulate a general result concerning the lower tail of subgraph counts in $G(n,p)$. In order to phrase the minimisation problem in the case $p \neq \frac{1}{2}$, it is convenient to first define

$$i_p(q) := q \log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$$

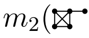
and $i_p(0) := \log\frac{1}{1-p}$. Further, given graphs H and G, let $N_H(G)$ denote the number of copies of H in G (throughout, we mean this as subgraphs and not as induced subgraphs). For every graph H, integer n, real $p \in (0,1)$, and every $\eta \in [0,1]$, let

$$\Phi^H_{n,p}(\eta) := \min\left\{ \sum_{e \in \binom{[\![n]\!]}{2}} i_p(q_e) : \mathbb{E}\left[N_H\big(G(n,q)\big)\right] \leqslant \eta \cdot \mathbb{E}\left[N_H\big(G(n,p)\big)\right] \right\},$$

where the minimum is taken over all $q \in [0,1]^{\binom{[\![n]\!]}{2}}$.

Recall that the 2-density of a graph H is the quantity $m_2(H)$ defined as follows: If H has at least two edges, then

$$m_2(H) := \max\left\{ \frac{e_F - 1}{v_F - 2} : F \subseteq H, e_F \geqslant 2 \right\};$$

otherwise, $m_2(H) := \frac{1}{2}$. The notation $F \subseteq H$ here means that F is a subgraph of H. For example, $m_2($ $) = 2$ but $m_2($$) = \frac{5}{2}$ because the maximum is attained at the subgraph .

THEOREM 2. *For every nonempty graph H, all $p_0 < 1$, and every $\varepsilon > 0$, there exists a constant L such that the following holds: Suppose that $Ln^{-1/m_2(H)} \leqslant p \leqslant p_0$ and let $X := N_H\big(G(n,p)\big)$. Then, for every $\eta \in [0,1]$,*

$$(1-\varepsilon)\cdot\Phi^H_{n,p}(\eta+\varepsilon) \leqslant -\log\mathbb{P}\big(X \leqslant \eta\mathbb{E}[X]\big) \leqslant (1+\varepsilon)\cdot\Phi^H_{n,p}\big((1-\varepsilon)\eta\big).$$

A key feature of Theorem 2 is that the lower-bound assumption on p is optimal up to constants. To see this, note first that, by Harris's inequality, for every $F \subseteq H$ and $p = o(1)$,

$$\mathbb{P}(X=0) = \mathbb{P}\big(H \nsubseteq G(n,p)\big) \geqslant \mathbb{P}\big(F \nsubseteq G(n,p)\big) \geqslant (1-p^{e_F})^{n^{v_F}} \geqslant \exp(-2n^{v_F}p^{e_F}).$$

Moreover, $m_2(H)$ is defined so that $n^{v_F}p^{e_F} = o(n^2p)$ for some $F \subseteq H$ precisely when $p \ll n^{-1/m_2(H)}$. On the other hand, for all H, n, p, and $\eta < 1$, we have $\Phi^H_{n,p}(\eta) \geqslant cn^2p$ for some positive c that depends only on H and η (see Lemma 21 below).

The boundary case $\eta = 0$ in Theorem 2, the probability that a random graph is H*-free*, has been extensively studied in the literature. In particular, Łuczak [28] computed the asymptotics of $\log \mathbb{P}\big(K_3 \nsubseteq G(n,p)\big)$ for all $p \gg n^{-1/m_2(K_3)}$ and derived an asymptotic formula for $\log \mathbb{P}\big(H \nsubseteq G(n,p)\big)$, for every nonbipartite graph H and all $p \gg n^{-1/m_2(H)}$, from the so-called KŁR conjecture [24], which was proved some fifteen years later by Balogh, Morris, and Samotij [3] and by Saxton and Thomason [32]. In fact, the hypergraph container theorems proved in [3, 32] can be used to compute the asymptotics of the logarithms of these probabilities directly, using simple, well-known results in extremal graph theory, see [3, §1.3].

Our methods allow us to generalise Theorem 2 to k-uniform hypergraphs in a straightforward way. Suppose that H is a nonempty k-uniform hypergraph. The k-density of H is the quantity $m_k(H)$ defined as follows: If H has at least two edges, then

$$m_k(H) \coloneqq \max\left\{\frac{e_F - 1}{v_F - k} : F \subseteq H, e_F \geqslant 2\right\}; \tag{4}$$

otherwise, $m_k(H) \coloneqq \frac{1}{k}$. For every integer n, real $p \in (0,1)$, and every $\eta \in [0,1]$, we define $\Phi^H_{n,p}(\eta)$ analogously to the graph case:

$$\Phi^H_{n,p}(\eta) \coloneqq \min\left\{\sum_{e \in \binom{[\![n]\!]}{k}} i_p(q_e) : \mathbb{E}\left[N_H\big(G^{(k)}(n,q)\big)\right] \leqslant \eta \cdot \mathbb{E}\left[N_H\big(G^{(k)}(n,p)\big)\right]\right\},$$

where $G^{(k)}(n,q)$ is the binomial random k-uniform hypergraph with vertex set $[\![n]\!]$.

THEOREM 3. *For every nonempty k-uniform hypergraph H, all $p_0 < 1$, and every $\varepsilon > 0$, there exists a constant L such that the following holds: Suppose that $Ln^{-1/m_k(H)} \leqslant p \leqslant p_0$ and let $X \coloneqq N_H\big(G^{(k)}(n,p)\big)$. Then, for every $\eta \in [0,1]$,*

$$(1-\varepsilon) \cdot \Phi^H_{n,p}(\eta + \varepsilon) \leqslant -\log \mathbb{P}\big(X \leqslant \eta\mathbb{E}[X]\big) \leqslant (1+\varepsilon) \cdot \Phi^H_{n,p}\big((1-\varepsilon)\eta\big).$$

As in Theorem 2, the lower-bound assumption on p in Theorem 3 is optimal and the asymptotics of $\log \mathbb{P}(X = 0)$ can be derived from the hypergraph container theorems.

The final application of our new entropy method is a solution to the lower tail problem for the number of arithmetic progressions of a given length in a binomial random subset of $[\![n]\!]$, which will demonstrate that symmetry is not crucial for our methods. Given a function $q\colon [\![n]\!] \to [0,1]$, we denote by $[\![n]\!]_q$ the random subset of $[\![n]\!]$ obtained by independently retaining each $i \in [\![n]\!]$ with probability q_i. For a positive integer k and a set $I \subseteq [\![n]\!]$, let $A_k(I)$ denote the number of k-term arithmetic progressions in I.

THEOREM 4. *For every positive integer k, all $p_0 < 1$, and every $\varepsilon > 0$, there exists a constant L such that the following holds: Suppose that $Ln^{-1/(k-1)} \leqslant p \leqslant p_0$ and let $X \coloneqq A_k\big([\![n]\!]_p\big)$. Then, for every $\eta \in [0,1]$,*

$$(1-\varepsilon) \cdot \Phi^k_{n,p}(\eta + \varepsilon) \leqslant -\log \mathbb{P}\big(X \leqslant \eta\mathbb{E}[X]\big) \leqslant (1+\varepsilon) \cdot \Phi^k_{n,p}\big((1-\varepsilon)\eta\big).$$

As before, the lower-bound assumption on p in Theorem 4 is optimal and the asymptotics of $\log \mathbb{P}(X = 0)$ can be derived from the hypergraph container theorems, see [3, Theorem 1.1].

1.2. *The main technical result.* A natural way to generalise Theorems 2, 3, and 4 is to represent the combinatorial objects we are counting as edges of an auxiliary hypergraph (no relation to the hypergraphs of Theorem 3). This way, each of the respective random variables counts the number of edges of such a hypergraph that are induced by random subset of its vertices. This idea is not new – the transference principles of Conlon and Gowers [10] and Schacht [33] and the hypergraph container theorems [3, 32] are prime examples of why taking such an abstract viewpoint may prove beneficial in our context. For example, in order to express the number of triangles in $G(n,p)$ this way, we consider the 3-uniform hypergraph with vertex set $\binom{[\![n]\!]}{2}$, the edge set of the complete graph on $[\![n]\!]$, whose hyperedges are the $\binom{n}{3}$ triples of edges that form triangles in the complete graph on $[\![n]\!]$.

We are thus led to ask the following general question: Given a hypergraph $\mathscr{H}$ and a $p \in [0,1]$, what is the probability that a random subset of the vertices of $\mathscr{H}$ formed by independently retaining each vertex with probability p contains atypically few hyperedges? For extra generality, we allow the edges of the hypergraph to have positive weights. However, we will only discuss uniform hypergraphs here; we will denote by r the common size of all hyperedges.

Suppose that an r-uniform hypergraph $\mathscr{H}$ is equipped with a weight function $d\colon \mathscr{H} \to (0,\infty)$. Here and below we do not distinguish in the notation between a hypergraph and its set of hyperedges, so d is in fact on the hyperedges of $\mathscr{H}$. The set of vertices of $\mathscr{H}$ will be denoted by $V(\mathscr{H})$. We shall denote by $e(\mathscr{H})$ the sum $\sum_{A\in\mathscr{H}} d_A$ of all edge weights and, for every set $B \subseteq V(\mathscr{H})$, we shall write

$$\deg_{\mathscr{H}} B \coloneqq \sum_{B\subseteq A\in\mathscr{H}} d_A. \tag{5}$$

Moreover, for every $s \in [\![r]\!]$, we define

$$\Delta_s(\mathscr{H}) \coloneqq \max\left\{\deg_{\mathscr{H}} B : B \subseteq V \text{ and } |B| = s\right\}.$$

Note that when $d_A = 1$ for every $A \in \mathscr{H}$, then we may simply view $\mathscr{H}$ as a hypergraph; in this case, the above definitions give the usual notions of edge counts and degrees.

Let $\mathscr{H}$ be a hypergraph and denote $V = V(\mathscr{H})$ for brevity. Let $Y = (Y_v)_{v\in V}$ be a sequence of i.i.d. Bernoulli random variables with success probability p, one for every vertex of the hypergraph $\mathscr{H}$, and let R be the corresponding random subset of V, i.e., $R \coloneqq \{v \in V : Y_v = 1\}$. For a function $q\colon V \to [0,1]$, we let $Y^{(q)} = (Y'_v)_{v\in V}$ be a sequence of independent Bernoulli random variables such that Y'_v has success probability q_v for each $v \in V$ and let $R^{(q)}$ be the corresponding random subset of V. For every nonnegative real η, define

$$\Phi_p^{\mathscr{H}}(\eta) \coloneqq \min\Big\{D_{KL}\big(Y^{(q)} \,\|\, Y\big) : q \in [0,1]^V \text{ and } \mathbb{E}[e(\mathscr{H}[R^{(q)}])] \leqslant \eta \cdot \mathbb{E}[e(\mathscr{H}[R])]\Big\}, \tag{6}$$

where D_{KL} is the Kullback–Leibler divergence, so that,

$$D_{KL}\big(Y^{(q)} \,\|\, Y\big) = \sum_{v\in V} i_p(q_v) = \sum_{v\in V} q_v \log\frac{q_v}{p} + (1-q_v)\log\frac{1-q_v}{1-p}.$$

Here and below, $\mathscr{H}[R]$ stands for the restriction of $\mathscr{H}$ to R, namely the hypergraph whose vertices are R and whose hyperedges are $\{A \in \mathscr{H} : A \subseteq R\}$; thus $e(\mathscr{H}[R]) = \sum_{A\subseteq R} d_A$. Also, for $W \subseteq V(\mathscr{H})$, we write $\mathscr{H} - W$ in place of $\mathscr{H}[V(\mathscr{H}) \setminus W]$. Finally, $v(\mathscr{H})$ denotes the number of vertices in $\mathscr{H}$, i.e., $|V|$. Note that, unlike $e(\mathscr{H})$, it does not depend on the weight function.

THEOREM 5. *For every integer r and all $p_0 < 1$, $\varepsilon > 0$, and K, there exist a positive λ and a C such that the following holds. Let V be a finite set and let $\mathscr{H}$ be a nonempty r-uniform hypergraph with vertex set V and a weight function $d\colon \mathscr{H} \to (0,\infty)$. Let $p \in (0, p_0]$ and let R be the p-random subset of V. Suppose that, for every $s \in [\![r]\!]$, the maximum degree $\Delta_s(\mathscr{H})$ satisfies*

$$\Delta_s(\mathscr{H}) \leqslant K \cdot (\lambda p)^{s-1} \cdot \frac{e(\mathscr{H})}{v(\mathscr{H})}. \tag{7}$$

Then, letting $X := e(\mathscr{H}[R])$, for every nonnegative real η,

$$-\log \mathbb{P}\big(X \leqslant \eta \mathbb{E}[X]\big) \geqslant (1-\varepsilon)\Phi_p^{\mathscr{H}}(\eta+\varepsilon) - C.$$

Let us make some remarks on the formulation of Theorem 5. We first note that the parameter K is needed only for the case $s = 1$, where (7) becomes simply $\Delta_1(\mathscr{H}) \leqslant Ke(\mathscr{H})/v(\mathscr{H})$. For all $s > 1$ it could have been removed from (7) by choosing λ slightly smaller. In applications (see §7) we choose K according to the Δ_1 of the hypergraph in question, get a λ from the theorem, and that restricts the range of applicable p, though only by a constant.

Our argument gives the following explicit dependence of λ and C on the parameters:

$$\lambda \geqslant \frac{\varepsilon^9(1-p_0)}{10^5 K^2 r^4} \qquad \text{and} \qquad C \leqslant \frac{10^6 K^2 r^5}{\varepsilon^9(1-p_0)} \log \frac{1}{1-p_0}.$$

Most importantly, the dependence on ε is polynomial.

The readers familiar with the hypergraph container method will likely notice striking similarities between the assumptions of Theorem 5 and the assumptions of the container lemmas proved in [3, 4]. This is not a coincidence – the boundary case $\eta = 0$ in Theorem 5 bounds the probability that the random set R is independent in $\mathscr{H}$ from above by $\Phi_p^{\mathscr{H}}(\varepsilon)$, a minimum over all distributions $q \in [0,1]^V$ such that $R^{(q)}$ induces at most $\varepsilon e(\mathscr{H})$ edges in $\mathscr{H}$, in expectation (cf. the combinatorial notion of containers for independent sets in [3, 32]).

While it might be tempting to replace $\Phi_p^{\mathscr{H}}(\eta+\varepsilon)$ in the assertion of Theorem 5 with $\Phi_p^{\mathscr{H}}(\eta)$, or at least $\Phi_p^{\mathscr{H}}((1+\varepsilon)\eta)$, this is not always possible for η very close to zero. To see this, examine the case $\eta = 0$. Recalling the definition (6) of Φ, we see that at $\eta = 0$ we need to consider only q such that $\mathbb{E}[e(\mathscr{H}[R^{(q)}])] \leqslant 0$. But $e(\mathscr{H}[R^{(q)}]) \geqslant 0$, and it is zero only when $\mathscr{H}[R^{(q)}]$ is empty, i.e., when $R^{(q)}$ is an independent set. Thus the expectation can be 0 only when q is supported on an independent set. Minimising $D_{KL}(Y^{(q)} \,\|\, Y)$ under this restriction, we see that the minimiser is a function q which takes the value p on some independent set of largest size and 0 elsewhere. We get $\Phi_p^{\mathscr{H}}(0) = (v(\mathscr{H}) - \alpha(\mathscr{H})) \cdot |\log(1-p)|$, where $\alpha(\mathscr{H})$ is the largest size of an independent set in $\mathscr{H}$.

Suppose now that $\mathscr{H}$ is the union of two hypergraphs with the same vertex set V: a dense hypergraph $\mathscr{H}_1$ with $\alpha(\mathscr{H}_1) \geqslant v(\mathscr{H})/2$ and a very sparse hypergraph $\mathscr{H}_2$ with $\alpha(\mathscr{H}_2) \leqslant v(\mathscr{H})/4$. (For example, if M and $v(\mathscr{H})$ are sufficiently large as a function of the uniformity r only, then a random hypergraph with $Mv(\mathscr{H})$ edges will typically have this property). Now, on the one hand, $\alpha(\mathscr{H}) \leqslant \alpha(\mathscr{H}_2)$ so

$$\Phi_p^{\mathscr{H}}(0) \geqslant \big(v(\mathscr{H}) - \alpha(\mathscr{H}_2)\big) \cdot |\log(1-p)| \geqslant \frac{3v(\mathscr{H})}{4} \cdot |\log(1-p)|.$$

But, on the other hand, by Harris's inequality, the p-random subset of some largest independent set of $\mathscr{H}_1$ has probability at least $(1-p^r)^{e(\mathscr{H}_2)}$ to be independent also in $\mathscr{H}_2$ and thus, when p is sufficiently small,

$$\mathbb{P}(X = 0) \geqslant (1-p)^{v(\mathscr{H}) - \alpha(\mathscr{H}_1)} \cdot e^{-2p^r e(\mathscr{H}_2)} \geqslant (1-p)^{2v(\mathscr{H})/3},$$

showing that $-\log \mathbb{P}(X=0)$ is not close to $\Phi_p^{\mathscr{H}}(0)$.

For easier comparison with the literature, let us reformulate Theorem 5 in the language of polynomials. We retain the notations V, Y, and $Y^{(q)}$ as above. We replace a weighted hypergraph $\mathscr{H}$ with a homogeneous polynomial f by turning each edge A of $\mathscr{H}$ into the monomial $d_A \cdot \prod_{v \in A} y_v$. The definition of Φ thus becomes

$$\Phi_p^f(\eta) = \min\left\{D_{KL}(Y^{(q)} \,\|\, Y) : q \in [0,1]^V, \mathbb{E}[f(Y^{(q)})] \leqslant \eta \cdot \mathbb{E}[f(Y)]\right\}.$$

Moreover, the assumption (7) can be now expressed in terms of partial derivatives of f. Given a $B = \{v_1, \dots, v_k\} \subseteq V$ and a polynomial f in $|V|$ variables, we denote

$$\partial_B f \coloneqq \frac{\partial}{\partial v_1} \cdots \frac{\partial}{\partial v_k} f.$$

The following statement is a reformulation of Theorem 5.

THEOREM 5'. *For every integer r and all $p_0 < 1$, $\varepsilon > 0$, and K, there exist a positive λ and a C such that the following holds. Let V be a finite set and let f be an r-homogeneous V-variate multilinear polynomial with nonnegative coefficients. Let $p \in (0, p_0]$ and let $Y = (Y_v)_{v \in V}$ be a sequence of i.i.d.* $\mathrm{Ber}(p)$ *random variables. Suppose that, for every nonempty $B \subseteq V$ with $|B| \leqslant r$,*

$$\partial_B f(\mathbf{1}) \leqslant K \cdot (\lambda p)^{|B|-1} \cdot \frac{f(\mathbf{1})}{|V|}.$$

Then, letting $X \coloneqq f(Y)$, for every nonnegative real η,

$$-\log \mathbb{P}\big(X \leqslant \eta \mathbb{E}[X]\big) \geqslant (1-\varepsilon)\Phi_p^f(\eta + \varepsilon) - C.$$

When f is a linear function, the variable X from the statement of the theorem is a sum of independent random variables. In this case, our argument can be simplified tremendously. The special case $f(y) = y_1 + \cdots + y_n$, which corresponds to the binomial distribution, is treated in §4.2, where a short, entropy-based proof of the optimal tail estimate

$$\mathbb{P}\big(\mathrm{Bin}(n,p) \leqslant nq\big) \leqslant \exp\big(-n \cdot i_p(q)\big)$$

is given.

1.3. *Lower bounds on the lower tail.* We end the results section with a lower bound on the lower tail probabilities from the statements of Theorems 5 and 5' that matches the upper bounds proved by these theorems. Since the proof of this lower bound is a relatively standard tilting argument, we relegate it to §6. Here is the exact formulation (in the language of Theorem 5').

THEOREM 6. *For every $p_0 < 1$ and $\varepsilon > 0$, there exists a C such that the following holds. Let V be a finite set, let $Y = (Y_v)_{v \in V}$ be a sequence of i.i.d.* $\mathrm{Ber}(p)$ *random variables, let $f \colon \{0,1\}^V \to [0, \infty)$ be an arbitrary increasing function, and let $X \coloneqq f(Y)$. Then, for every nonnegative real η,*

$$-\log \mathbb{P}\big(X \leqslant \eta \mathbb{E}[X]\big) \leqslant (1+\varepsilon)\Phi_p^f\big((1-\varepsilon)\eta\big) + C.$$

1.4. *Organisation of the paper.* The remainder of this paper is organised as follows. In Section 2, we outline of the proof of Theorem 1 and discuss some of the additional ideas required in the proof of Theorem 2 in the case $H = K_3$. In Section 3, which is merely three pages long, we present a complete proof of Theorem 1. In Section 4, we recall some basic properties of the Kullback–Leibler divergence and prove the key technical lemma (Lemma 15) that relates independence and conditional KL-divergence. Subsection 4.2 contains a short entropy-based proof of optimal tail bounds for binomial distributions, which might be of independent interest. Our main technical result, Theorem 5, is proved in Section 5. The matching lower bound for lower tail probabilities, Theorem 6, is proved in Section 6. Finally, Section 7 contains short derivations of Theorems 2, 3, and 4.

1.5. *Note added in proof.* After this work had been completed, we learned that an earlier work of Jain, Koehler, and Risteski [21] had independently exploited the connections between entropy and conditional independence in order to establish quantitative bounds on the tightness of the mean-field approximation to the free energy of Ising models on finite graphs (and, more generally, order-k Markov random fields on a finite set of vertices). Their use of the so-called ‘pinning lemma’ [21, Theorem 3.2] of Manurangsi and Raghavendra [29], which extends earlier work of Raghavendra and Tan [31] and of Montanari [30], and the way it is combined with Pinsker’s inequality, closely resemble our proof of Theorem 1.

2. Proof outline.

2.1. *Triangle count in $G(n, \frac{1}{2})$.* Let us first explain how to prove Theorem 1, i.e., the upper bound on the lower tail of the number of triangles. Considering only $G(n, \frac{1}{2})$, which is the *uniform* distribution on n-vertex graphs, allows us to phrase the argument in the familiar language of *entropy* rather than using the Kullback–Leibler divergence. The argument sketched here is described in full in §3 and takes no more than three pages.

Let Y be the random graph $G(n, \frac{1}{2})$ *conditioned* on having at most t triangles. Then

$$\log \mathbb{P}(X \leqslant t) = H(Y) - \binom{n}{2} \log 2, \tag{8}$$

where H is the entropy of Y. Examine the distribution of the first edge (i.e., of Y_{12}) under the conditioning.[2] For every integer $m \geqslant 0$, let

$$h_m := H(Y_{12} \mid \text{edges with at least one endpoint larger than } n - m),$$

where $H(\cdot \mid \cdot)$ is the usual conditional entropy. Since conditional entropy is nonnegative and it decreases as one increases the conditioning, we have $0 \leqslant h_{m+1} \leqslant h_m$ for every m. Since $h_0 = H(Y_{12}) \leqslant \log 2$, there must be some $m \leqslant \sqrt{n}$ such that $h_m - h_{m+1} \leqslant C/\sqrt{n}$, where C is an absolute constant.

Denote by S_m the edges from the definition of h_m, so that $h_m = H(Y_{12} \mid S_m)$. Since both $\{1, n-m\}$ and $\{2, n-m\}$ belong to $S_{m+1} \setminus S_m$, we may use the monotonicity of conditional entropy again to sandwich $H(Y_{12} \mid S_m, Y_{1,n-m}, Y_{2,n-m})$ between h_m and h_{m+1}:

$$h_{m+1} = H(Y_{12} \mid S_{m+1}) \leqslant H(Y_{12} \mid Y_{1,n-m}, Y_{2,n-m}, S_m) \leqslant H(Y_{12} \mid S_m) = h_m.$$

Hence, we also get the inequality

$$H(Y_{12} \mid S_m) - H(Y_{12} \mid Y_{1,n-m}, Y_{2,n-m}, S_m) \leqslant C/\sqrt{n}.$$

[2] We assume that the vertex set of $G(n, \frac{1}{2})$ is $[\![n]\!]$ and think of $Y \in \{0,1\}^{\binom{[\![n]\!]}{2}}$ as the characteristic vector of the edge set of the conditioned random graph.

By symmetry (every permutation of $[\![n-m]\!]$ preserves S_m), we may replace $(1,2,n-m)$ in the above inequality with any three different elements (i,j,k) of $[\![n-m]\!]$ and get

$$H(Y_{ij} \mid S_m) - H(Y_{ij} \mid Y_{ik}, Y_{jk}, S_m) \leqslant C/\sqrt{n}.$$

We now apply ***Pinsker's inequality***, which states that, for any two variables T and U, if $H(T) - H(T \mid U)$ is small, then T and U must be approximately independent. We apply this to the variables Y_{ij} conditioned on S_m to conclude that, conditioned on S_m, the three edges of every triangle are (typically) approximately independent.

Recall now the definition of $\Phi_n(t)$ given in (2). It is the minimum of $-H\big(G(n,q)\big) + \binom{n}{2}\log 2$ over all functions $q\colon \binom{[\![n]\!]}{2} \to [0,1]$ such that

$$T(q) := \mathbb{E}\big[\#\text{triangles in } G(n,q)\big] \leqslant t.$$

Consider the function $q_{ij} := \mathbb{E}[Y_{ij} \mid S_m]$. The approximate independence of the Y_{ij} gives that $T(q)$ is (typically) approximately the expected number of triangles in Y, which is at most t, by the definition of Y. Hence, $T(q) \leqslant t + o(t)$, where the $o(t)$ error term comes from the fact that the Y_{ij} are only approximately independent. We conclude that

$$H\big((Y_{ij})_{i,j\leqslant n-m} \mid S_m\big) \leqslant \sum_{i,j\leqslant n-m} H(Y_{ij} \mid S_m) \leqslant -\Phi_n(t+o(t)) + \binom{n}{2}\log 2.$$

Since S_m has only at most $n^{3/2}$ edges, its entropy is negligible and we get

$$H(Y) = H(S_m) + H\big((Y_{ij})_{i,j\leqslant n-m} \mid S_m\big) \leqslant (1-o(1))\cdot\Phi_n(t+o(t)) + \binom{n}{2}\log 2,$$

as needed.

Examining the proof above, we see that the crucial step is that of proving *conditional approximate independence*. Why was the conditioning necessary? Because Y is not close to a product measure but rather a mixture of product measures. Heuristically, the conditioning chooses one product measure from the mixture.

2.2. *Triangle count in $G(n,p)$ and beyond.* What is needed to prove Theorem 1 with $G(n,\frac{1}{2})$ replaced by $G(n,p)$? Since the latter is no longer a uniform distribution, in order to phrase a suitable analogue of (8), we certainly have to replace entropy with entropy relative to a product of p-Bernoulli variables (relative entropy is also called the Kullback–Leibler divergence, though note that the sign of the Kullback–Leibler divergence is minus that of what a straightforward analogue of entropy would have been) and we need an analogue of Pinsker's inequality for (conditional) relative entropy. These two ideas would have been enough to solve the lower tail (as well as the upper tail) problem for triangles in $G(n,p)$ for all $p \geqslant n^{-c}$, where c is an absolute positive constant.

In order to extend the argument to all $p \gg n^{-1/2}$, one needs to prove a version of Pinsker's inequality that provides a stronger upper bound on the difference of probabilities that two measures assign to *rare* events (rather than arbitrary events, as measured by the total variation distance). Furthermore, in order to use this strengthening of Pinsker's inequality, we also need to note that, when we condition our random graph on the lower tail event, the probability of every edge is at most p, even when we further condition on S_m. This follows from the Harris inequality (aka the FKG inequality). The use of Harris's inequality is the main (but not the only) reason why our methods are not as efficient for the upper tail problem.

The general setting of Theorem 5, which lacks symmetry, requires a serious overhaul of the argument. (Having said that, even in the setting of K_4 counts in $G(n,p)$, which still has a lot of symmetry, the argument sketched above does not work under the optimal assumption $p \gg n^{-2/5}$.) We no longer increase the conditioning in small steps (recall the definition of h_m above) but rather in large chunks, which are chosen randomly. The crux of the matter is relating the decrease in entropy caused by conditioning on each such random chunk to approximate independence of the remaining variables. Here, the key role is played by Lemma 15, an improvement of Pinsker's inequality that is inspired by the statement of Janson's inequality [22].

3. The lower tail of triangle count in $G(n, \frac{1}{2})$.

As explained above, our proof of Theorem 1 revolves around (information-theoretic) entropy. For convenience of the reader, we shall recall here the definitions of entropy and conditional entropy and list all of their properties required for our argument; for proofs of these properties, we refer the reader to [13, Chapter 2].

3.1. *Preliminaries.* The entropy of a random variable X taking values in a finite set $\mathscr{X}$ is the quantity $H(X)$ defined by

$$H(X) := -\sum_{x \in \mathscr{X}} \mathbb{P}(X = x) \log \mathbb{P}(X = x).$$

Further, given two random variables X and Y that take values in finite sets $\mathscr{X}$ and $\mathscr{Y}$, respectively, and have a joint distribution, the (conditional) entropy of X conditioned on Y is the quantity $H(X \mid Y)$ defined as follows:

$$H(X \mid Y) := \sum_{y \in \mathscr{Y}} \mathbb{P}(Y = y) H\big(X^{\{Y=y\}}\big),$$

where $X^{\{Y=y\}}$ denotes X conditioned on the event that $Y = y$, so that, for every $x \in \mathscr{X}$ and every $y \in \mathscr{Y}$ with $\mathbb{P}(Y = y) \neq 0$,

$$\mathbb{P}\big(X^{\{Y=y\}} = x\big) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}.$$

The above definitions ensure that entropies and conditional entropies are always nonnegative. Moreover, it is easy to verify that

$$H(X \mid Y) = H(X, Y) - H(Y). \tag{9}$$

In the remainder of this section, a discrete random variable will mean a random variable taking values in some finite set. The following elementary inequalities should be familiar to readers who have encountered the notion of entropy.

LEMMA 7. *Suppose that X, Y, and Z are discrete random variables and that X takes values in a finite set $\mathscr{X}$. We have:*

(i) $H(X) \leqslant \log |\mathscr{X}|$ *and equality holds iff X is uniform on $\mathscr{X}$;*
(ii) $H(X \mid Y) \leqslant H(X)$ *and equality holds iff X and Y are independent;*
(iii) $H(X \mid Y, Z) \leqslant H(X \mid Y)$;
(iv) $H(X, Y \mid Z) \leqslant H(X \mid Z) + H(Y \mid Z)$.

The main ingredient in our proof is Pinsker's inequality (see [15, Problem 3.18]), which, in our context, can be viewed as a 'stability' version of (ii) in Lemma 7. The statement requires the following notation: For two random variables X and Y, we denote by $X \times Y$ the random variable obtained by first letting $\tilde{X}$ and $\tilde{Y}$ be two independent copies of X and Y, respectively, and then defining $X \times Y = (\tilde{X}, \tilde{Y})$. In other words, $\mathscr{L}(X \times Y) = \mathscr{L}(X) \times \mathscr{L}(Y)$, where, as usual, $\mathscr{L}(X)$ stands for the law of X, i.e., the measure induced by X on its space of values.

LEMMA 8. *Suppose that X and Y are discrete random variables. We have*

$$d_{TV}\big((X,Y), X \times Y\big) \leqslant \sqrt{2\big(H(X) - H(X \mid Y)\big)},$$

where d_{TV} denotes the total variation distance.

3.2. *The argument.* Let Y denote the random graph $G(n, \frac{1}{2})$ conditioned on having at most t triangles. In other words, Y is a uniformly chosen random graph with vertex set $[\![n]\!] := \{1, \dots, n\}$ and at most t triangles. In particular, Lemma 7(i) implies that

$$\log \mathbb{P}(X_n \leqslant t) = H(Y) - \binom{n}{2} \log 2. \tag{10}$$

In order to bound the entropy of Y from above, it will be convenient to view Y as the random vector $(Y_e)_{e \in K_n}$, where Y_e indicates whether e is an edge of Y. For a subvector S of Y and every $e \in K_n$, we will write Y_e^S to denote the random variable whose (random) distribution is the distribution of Y_e conditioned on S, so that $\mathbb{P}(Y_e^S = 1) = \mathbb{E}[Y_e \mid S]$. The following lemma captures the notion of *conditional approximate independence* (recall the proof sketch in §2.1).

LEMMA 9. *There exists a subgraph $F \subseteq K_n$ with at most $n^{3/2}$ edges and such that, for every $\{i,j,k\} \in \binom{[\![n]\!]}{3}$, letting $S := (Y_f)_{f \in F}$ and*

$$d_{ijk}^S := d_{TV}\big((Y_{ij}^S, Y_{ik}^S, Y_{jk}^S), Y_{ij}^S \times Y_{ik}^S \times Y_{jk}^S\big),$$

we have $\mathbb{E}[d_{ijk}^S] \leqslant 2n^{-1/4}$.

PROOF. For a nonnegative integer m, let F_m be the subgraph of K_n comprising all edges $\{i,j\}$ satisfying $\max\{i,j\} > n - m$, let $S_m := (Y_f)_{f \in F_m}$ and let $h_m := H(Y_{12} \mid S_m)$. By Lemma 7(iii), the function $m \mapsto h_m$ is decreasing and hence, for some $m \leqslant \sqrt{n}$, we must have

$$h_m - h_{m+1} \leqslant \frac{h_0 - h_{\sqrt{n}}}{\sqrt{n}}.$$

Bounding the numerator is easy. On the one hand, we have

$$h_0 = H(Y_{12}) \leqslant \log 2,$$

as Y_{12} takes only two values, see Lemma 7(i); on the other hand, $h_m \geqslant 0$ for every m, as conditional entropy is always nonnegative. Thus, there must be an m with $0 \leqslant m \leqslant \sqrt{n} - 1$ such that $h_m - h_{m+1} \leqslant (\log 2)/\sqrt{n}$. Fix one such m and let $F = F_m$ and $S = S_m$; note that $e(F) \leqslant mn \leqslant n^{3/2}$. Since $F \subseteq F \cup \big\{\{1, n-m\}, \{2, n-m\}\big\} \subseteq F_{m+1}$, Lemma 7(iii) implies that

$$h_{m+1} = H(Y_{12} \mid S_{m+1}) \leqslant H(Y_{12} \mid S, Y_{1,n-m}, Y_{2,n-m}) \leqslant H(Y_{12} \mid S) = h_m$$

and, consequently,

$$H(Y_{12} \mid S) - H(Y_{12} \mid S, Y_{1,n-m}, Y_{2,n-m}) \leqslant (\log 2)/\sqrt{n}. \tag{11}$$

By symmetry (every permutation of $[\![n-m]\!]$ fixes F), we may replace the triple of indices $(1, 2, n-m)$ in (11) with any ordered triple (i,j,k) of distinct elements of $[\![n-m]\!]$. Using the definition of conditional entropy, we may rewrite this upgraded inequality as

$$\mathbb{E}\Big[\underbrace{H(Y_{ij}^S) - H(Y_{ij}^S \mid Y_{ik}^S, Y_{jk}^S)}_{\lambda_{ijk}^S}\Big] \leqslant (\log 2)/\sqrt{n},$$

where $\mathbb{E}$ averages over the values of S.

Fix an arbitrary triple $\{i,j,k\} \in \binom{[\![n]\!]}{3}$. If $\max\{i,j,k\} > n-r$, then at least two out of the three pairs ij, ik, jk belong to F; consequently, at least two out the three corresponding variables Y^S_{ij}, Y^S_{ik}, Y^S_{jk} are trivial (for every evaluation of S), which implies that $d^S_{ijk} = 0$ (the d^S_{ijk} from the statement of the lemma). Therefore, we may assume that $\{i,j,k\} \in \binom{[\![n-m]\!]}{3}$. For brevity, denote $A = Y^S_{ij}$, $B = Y^S_{ik}$, and $C = Y^S_{jk}$, so that

$$\begin{aligned} d^S_{ijk} &= d_{\mathrm{TV}}\big((A,B,C), A \times B \times C\big) \\ &\leqslant d_{\mathrm{TV}}\big((A,B,C), A \times (B,C)\big) + d_{\mathrm{TV}}\big(A \times (B,C), A \times B \times C\big) \\ &= \underbrace{d_{\mathrm{TV}}\big((A,B,C), A \times (B,C)\big)}_{d_1} + \underbrace{d_{\mathrm{TV}}\big((B,C), B \times C\big)}_{d_2}. \end{aligned}$$

Pinsker's inequality (Lemma 8) implies that

$$d_1 \leqslant \sqrt{\tfrac{1}{2}\big(H(A) - H(A \mid B,C)\big)} = \sqrt{\tfrac{1}{2}\lambda^S_{ijk}}.$$

Further,

$$d_2 \leqslant d_{\mathrm{TV}}\big((A,B,C), (A,B) \times C\big) \leqslant \sqrt{\tfrac{1}{2}\big(H(C) - H(C \mid A,B)\big)} = \sqrt{\tfrac{1}{2}\lambda^S_{jki}}$$

(the first inequality is easy to check). We conclude that

$$\begin{aligned} \mathbb{E}\big[d^S_{ijk}\big] &\leqslant \mathbb{E}\left[\sqrt{\tfrac{1}{2}\lambda^S_{ijk}}\right] + \mathbb{E}\left[\sqrt{\tfrac{1}{2}\lambda^S_{jki}}\right] \leqslant \sqrt{\tfrac{1}{2}\mathbb{E}[\lambda^S_{ijk}]} + \sqrt{\tfrac{1}{2}\mathbb{E}[\lambda^S_{ijk}]} \\ &\leqslant 2\sqrt{(\log 2)/(2\sqrt{n})} \leqslant 2n^{-1/4}, \end{aligned}$$

where the second inequality follows from the Cauchy–Schwarz inequality. □

Let F be the graph from the statement of Lemma 9 and let $S = (Y_f)_{f \in F}$, as in the claim. The chain rule for conditional entropies, identity (9) above, and Lemma 7(iv) imply that

$$H(Y) = H(S) + H\big((Y_e)_{e \in K_n \setminus F} \mid S\big) \leqslant H(S) + \sum_{e \in K_n \setminus F} H(Y_e \mid S).$$

Since S takes at most $2^{e(F)}$ different values and $e(F) \leqslant n^{3/2}$, we have, by Lemma 7(i),

$$H(Y) \leqslant n^{3/2} \log 2 + \sum_{e \in K_n} H(Y_e \mid S), \tag{12}$$

where we also used the fact that conditional entropies are nonnegative to extend the range of the sum from $K_n \setminus F$ to K_n (in fact, $H(Y_e \mid S) = 0$ for every $e \in F$).

Recall that our eventual goal is to compare the entropy of Y to $\Phi_n(t)$, which is defined as the minimum over certain functions q. Define therefore the S-measurable random function $q \colon \binom{[\![n]\!]}{2} \to [0,1]$ by letting, for each $e \in K_n$,

$$q_e := \mathbb{E}[Y_e \mid S] = \mathbb{P}(Y^S_e = 1).$$

Letting $h \colon [0,1] \to [0, \log 2]$ be the function defined by $h(x) = -x \log x - (1-x)\log(1-x)$, we may now write

$$H(Y_e \mid S) = \mathbb{E}\left[H(Y^S_e)\right] = \mathbb{E}[h(q_e)].$$

Let X^S denote the number of triangles in Y conditioned on S, that is,

$$X^S := \sum_{\{i,j,k\} \in \binom{[\![n]\!]}{3}} Y_{ij}^S Y_{ik}^S Y_{jk}^S$$

and let

$$\bar{X}^S := \sum_{\{i,j,k\} \in \binom{[\![n]\!]}{3}} q_{ij} q_{ik} q_{jk} = \mathbb{E}\left[N_{K_3}\big(G(n,q)\big)\right].$$

Recall the definition of d_{ijk}^S from the statement of Lemma 9 and observe that

$$\left|\mathbb{E}\big[Y_{ij}^S Y_{ik}^S Y_{jk}^S\big] - q_{ij} q_{ik} q_{jk}\right| \leqslant d_{ijk}^S; \tag{13}$$

indeed, the two terms in the left-hand side are the probabilities of the event that $Y_{ij}^S = Y_{ik}^S = Y_{jk}^S = 1$ under the two distributions whose total variation distance is d_{ijk}^S.

Let

$$\Delta = \sum_{\{i,j,k\} \in \binom{[\![n]\!]}{3}} d_{ijk}^S$$

and note that, by Lemma 9,

$$\mathbb{E}[\Delta] \leqslant \binom{n}{3} \cdot 2n^{-1/4} \leqslant n^{11/4}. \tag{14}$$

Summing (13) over all triples $\{i,j,k\}$, we obtain

$$\bar{X}^S \leqslant \mathbb{E}\big[X^S\big] + \Delta \leqslant t + \Delta,$$

since $X^S \leqslant t$ with probability one. In particular, the definition of Φ_n, see (2), implies that

$$\sum_{e \in K_n} \big(\log 2 - h(e_q)\big) \geqslant \Phi_n(t + \Delta).$$

We may conclude that

$$\sum_{e \in K_n} H(Y_e \mid S) = \mathbb{E}\left[\sum_{e \in K_n} h(q_e)\right] \leqslant \binom{n}{2} \log 2 - \mathbb{E}[\Phi_n(t + \Delta)].$$

Since Φ_n is decreasing and nonnegative,

$$\mathbb{E}[\Phi_n(t+\Delta)] \geqslant \mathbb{P}(\Delta \leqslant n^{23/8}) \cdot \Phi_n(t + n^{23/8}) \overset{(14)}{\geqslant} \big(1 - n^{-1/8}\big) \cdot \Phi_n(t + n^{23/8}).$$

Recalling (10) and (12), this implies that

$$\log \mathbb{P}(X_n \leqslant t) \leqslant -(1 - n^{-1/8}) \cdot \Phi_n(t + n^{23/8}) + n^{3/2} \leqslant -\Phi_n(t + n^{23/8}) + 2n^{15/8},$$

as $\Phi_n(t) \leqslant \binom{n}{2} \log 2$ for every t. This finishes the proof of Theorem 1. □

4. The Kullback–Leibler divergence.

For the proof of Theorem 5, we need the notion of Kullback–Leibler divergence, or relative entropy. Let P and Q be random variables taking values in a finite set $\mathscr{X}$ and suppose that $\mathscr{L}(P) \ll \mathscr{L}(Q)$, that is, that the distribution of P is absolutely continuous with respect to the distribution of Q. Denoting by p and q the densities of P and Q, respectively, the *Kullback–Leibler divergence* of P from Q (also known as the *relative entropy*), denoted by $D_{KL}(P \,\|\, Q)$, is defined as follows:

$$D_{KL}(P \,\|\, Q) := \sum_{x \in \mathscr{X}} p(x) \log \frac{p(x)}{q(x)},$$

where we adopt the convention that $0 \log \frac{0}{q} = 0$ for all q. The assumption that $\mathscr{L}(P) \ll \mathscr{L}(Q)$, which is a concise way of saying that $p(x) = 0$ whenever $q(x) = 0$, guarantees that $D_{KL}(P \,\|\, Q)$ is well-defined. A fundamental property of the KL-divergence is that it is always nonnegative; indeed, since $\log x \leqslant x - 1$ for all positive x, we have, letting $\mathscr{X}' = \{x \in \mathscr{X} : p(x) > 0\}$,

$$D_{KL}(P \,\|\, Q) = -\sum_{x \in \mathscr{X}'} p(x) \log \frac{q(x)}{p(x)} \geqslant \sum_{x \in \mathscr{X}'} \big(p(x) - q(x)\big) = 1 - \sum_{x \in \mathscr{X}'} q(x) \geqslant 0.$$

One easily checks that, when Q is a uniformly chosen random element of $\mathscr{X}$, then

$$D_{KL}(P \,\|\, Q) = \log|\mathscr{X}| - H(P),$$

where $H(P)$ is the entropy of P (defined in the previous section). In particular, if P is the uniformly chosen random element of a nonempty subset $\mathscr{A} \subseteq \mathscr{X}$, then, by Lemma 7(i),

$$D_{KL}(P \,\|\, Q) = \log|\mathscr{X}| - \log|\mathscr{A}| = -\log \mathbb{P}(Q \in \mathscr{A}).$$

The following property of the KL-divergence, which generalises this identity, is the starting point of our approach.

PROPOSITION 10. *Suppose that Q is a random variable taking values in a finite set $\mathscr{X}$. Suppose that $A \subseteq \mathscr{X}$ satisfies $\mathbb{P}(Q \in A) \neq 0$ and let Q^A be the random variable Q conditioned on the event $\{Q \in A\}$. Then*

$$D_{KL}(Q^A \,\|\, Q) = -\log \mathbb{P}(Q \in A).$$

PROOF. Let $q \colon \mathscr{X} \to [0,1]$ be the probability density function of Q and note that the probability density function of Q^A is the function $q^A \colon \mathscr{X} \to [0,1]$ defined by

$$q^A(x) := \begin{cases} \frac{q(x)}{\mathbb{P}(Q \in A)} & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that

$$D_{KL}(Q^A \,\|\, Q) = \sum_{x \in \mathscr{X}} q^A(x) \log \frac{q^A(x)}{q(x)} = \sum_{x \in A} \frac{q(x)}{\mathbb{P}(Q \in A)} \log \frac{1}{\mathbb{P}(Q \in A)} = \log \frac{1}{\mathbb{P}(Q \in A)},$$

as claimed. □

The next property of the KL-divergence is a generalisation of the chain rule for entropies, identity (9), and Lemma 7(ii). In fact, the equality in (15) below is a special case of an even more general identity, the chain rule for relative entropies, see [13, Theorem 2.5.3].

PROPOSITION 11. *Let Q_1 and Q_2 be random variables taking values in finite sets $\mathscr{X}_1$ and $\mathscr{X}_2$, respectively. Suppose that (P_1, P_2) is an $\mathscr{X}_1 \times \mathscr{X}_2$-valued random variable such that $\mathscr{L}(P_i) \ll \mathscr{L}(Q_i)$ for each i. Let $Q_1 \times Q_2$ denote a random variable whose independent coordinates have marginals Q_1 and Q_2, respectively; that is, $\mathscr{L}(Q_1 \times Q_2) = \mathscr{L}(Q_1) \times \mathscr{L}(Q_2)$. Then*

$$D_{KL}\big((P_1, P_2) \,\|\, Q_1 \times Q_2\big) - D_{KL}(P_2 \,\|\, Q_2) \geqslant D_{KL}(P_1 \,\|\, Q_1), \tag{15}$$

where equality holds if and only if P_1 and P_2 are independent.

The proof of the proposition employs the following elementary inequality, whose proof we include for the sake of completeness.

LEMMA 12. *Suppose that I is a finite set and, for each $i \in I$, let a_i and b_i be nonnegative reals such that $a_i = 0$ whenever $b_i = 0$. Then, letting $a = \sum_{i\in I} a_i$ and $b = \sum_{i \in I} b_i$, we have*

$$\sum_{i\in I} a_i \log \frac{a_i}{b_i} \geqslant a \log \frac{a}{b}.$$

Moreover, equality holds above if and only if $a_i b = a b_i$ for every $i \in I$.

PROOF. Without loss of generality, we may assume that $a_i > 0$ (and thus $b_i > 0$) for each $i \in I$. Since the function $x \mapsto -\log x$ is strictly convex, Jensen's inequality implies that

$$\sum_{i\in I} a_i \log \frac{a_i}{b_i} - a \log \frac{a}{b} = a \cdot \sum_{i\in I} \frac{a_i}{a} \cdot \Big(- \log \frac{ab_i}{a_i b}\Big) \geqslant -a \cdot \log \Big(\sum_{i\in I} \frac{a_i}{a} \cdot \frac{ab_i}{a_i b}\Big) = -a \cdot \log 1 = 0$$

and the inequality is strict unless $\frac{ab_i}{a_i b} = \sum_{j\in I} \frac{a_j}{a} \cdot \frac{ab_j}{a_j b} = 1$ for every $i \in I$, as claimed. □

PROOF OF PROPOSITION 11. Let $p\colon \mathscr{X}_1 \times \mathscr{X}_2 \to [0,1]$ be the probability density function of (P_1, P_2) and, for each $i \in \{1,2\}$, let $q_i \colon \mathscr{X}_i \to [0,1]$ be the probability density function of Q_i. Without loss of generality, we may assume that $q_i(x_i) > 0$ for every $i \in \{1,2\}$ and each $x_i \in \mathscr{X}_i$. The functions $p_1 \colon \mathscr{X}_1 \to [0,1]$ and $p_2 \colon \mathscr{X}_2 \to [0,1]$ defined by

$$p_1(x_1) := \sum_{x_2 \in \mathscr{X}_2} p(x_1,x_2) \qquad \text{and} \qquad p_2(x_2) := \sum_{x_1 \in \mathscr{X}_1} p(x_1,x_2)$$

are the probability density functions of P_1 and P_2, respectively. Now, denoting by L the left-hand side of (15), we have

$$\begin{aligned}
L &= \sum_{(x_1,x_2)\in\mathscr{X}_1\times\mathscr{X}_2} p(x_1,x_2) \log \frac{p(x_1,x_2)}{q_1(x_1)q_2(x_2)} - \sum_{x_2\in\mathscr{X}_2} p_2(x_2) \log \frac{p_2(x_2)}{q_2(x_2)} \\
&\overset{(*)}{=} \sum_{x_1\in\mathscr{X}_1} \sum_{x_2\in\mathscr{X}_2} p(x_1,x_2) \log \frac{p(x_1,x_2)}{q_1(x_1)p_2(x_2)} \\
&\overset{(\dagger)}{\geqslant} \sum_{x_1\in\mathscr{X}_1} \Big(\sum_{x_2\in\mathscr{X}_2} p(x_1,x_2) \Big) \log \frac{\sum_{x_2\in\mathscr{X}_2} p(x_1,x_2)}{\sum_{x_2\in\mathscr{X}_2} q_1(x_1)p_2(x_2)} \\
&= \sum_{x_1\in\mathscr{X}_1} p_1(x_1) \log \frac{p_1(x_1)}{q_1(x_1)} = D_{KL}(P_1 \,\|\, Q_1),
\end{aligned}$$

where $(*)$ follows by applying $p_2(x_2) = \sum_{x_1\in\mathscr{X}_1} p(x_1,x_2)$ to the second sum and $(\dagger)$ follows by applying Lemma 12 to the inner sum. Finally, the characterisation of equality in Lemma 12 implies that equality holds in $(\dagger)$ if and only if $p(x_1,x_2) = p_1(x_1)p_2(x_2)$ for all $x_1 \in \mathscr{X}_1$ and $x_2 \in \mathscr{X}_2$. □

4.1. *Divergence from a vector of i.i.d. Bernoulli variables.* Throughout this paper, we shall be estimating divergences of random variables from vectors of independent $\mathrm{Ber}(p)$ random variables. In view of this, it will be convenient for us to define, for a real $p \in (0,1)$, an integer $k \geqslant 1$, and a random variable X taking values in $\{0,1\}^k$, the *p-divergence* $I_p(X)$ of X by

$$I_p(X) := D_{KL}\big(X \,\|\, \mathrm{Ber}(p)^k\big) \geqslant 0. \tag{16}$$

When X is Bernoulli itself, say with parameter q, then $I_p(X)$ is a function of q which we will denote by i_p. Namely,

$$i_p(q) \coloneqq I_p\bigl(\mathrm{Ber}(q)\bigr) = D_{KL}\bigl(\mathrm{Ber}(q) \,\|\, \mathrm{Ber}(p)\bigr) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}. \tag{17}$$

Denote the first and the second derivatives of i_p by i_p' and i_p'', respectively. Let us record here, for future reference, that, for every $q \in (0,1)$,

$$i_p'(q) = \log \frac{q}{p} - \log \frac{1-q}{1-p} \qquad \text{and} \qquad i_p''(q) = \frac{1}{q} + \frac{1}{1-q}. \tag{18}$$

We also define a notion of conditional divergence. Given random variables X and Y that have a joint distribution and such that X takes values in $\{0,1\}^k$ for some integer $k \geqslant 1$, we define the *conditional p-divergence* of X conditioned on Y

$$I_p(X \mid Y) \coloneqq \mathbb{E}\left[I_p\bigl(X^Y\bigr)\right] = \mathbb{E}\left[D_{KL}\bigl(X^Y \,\|\, \mathrm{Ber}(p)^k\bigr)\right],$$

where X^Y denotes the random variable X conditioned on Y, cf. the definition of conditional entropy.

It is straightforward to verify that, when X takes values in $\{0,1\}^k$,

$$I_{1/2}(X) = k \log 2 - H(X) \qquad \text{and} \qquad I_{1/2}(X \mid Y) = k \log 2 - H(X \mid Y)$$

and therefore it should not come at a surprise that the divergence and the conditional divergence defined above satisfy similar inequalities as entropy and conditional entropy, such as the ones presented in Lemma 7, only in reverse. In particular, Proposition 11 implies that[3]

$$I_p(X \mid Y) \geqslant I_p(X) \tag{19}$$

and equality holds if and only if X and Y are independent, cf. Lemma 7(ii); moreover, if Y also takes values in $\{0,1\}^\ell$ for some integer ℓ, then

$$I_p(X,Y) = I_p(X \mid Y) + I_p(Y) \geqslant I_p(X) + I_p(Y), \tag{20}$$

where, again, equality holds if and only if X and Y are independent, cf. the chain rule for entropies (identity (9)). Generalising this further, if Z is another random variable (defined on the same probability space as X and Y), then invoking the above inequality with X and Y replaced by X^Z and Y^Z and taking the expectation of both sides yields

$$I_p(X,Y \mid Z) \geqslant I_p(X \mid Z) + I_p(Y \mid Z), \tag{21}$$

cf. Lemma 7(iv). One final property that we shall require is the following fact.

PROPOSITION 13. *Suppose that random variables X, Y, and Z have a joint distribution and that X takes values in $\{0,1\}^k$ for some integer $k \geqslant 1$. Then, for every $p \in (0,1)$,*

$$I_p(X \mid Y, Z) = \mathbb{E}\left[I_p(X^Y \mid Z^Y)\right]$$

PROOF. The assertion follows from the definition of conditional p-divergence and the fact that

$$\mathscr{L}\bigl(X^{(Y,Z)}\bigr) = \mathscr{L}\bigl((X^Y)^{Z^Y}\bigr)$$

almost surely. □

[3] In order to see this, observe first that $I_p(X \mid Y) = D_{KL}\bigl((X,Y) \,\|\, \mathrm{Ber}(p)^k \times Y\bigr)$.

4.2. *Interlude.* As an illustration of the subadditivity property of the divergence I_p, we will give a short proof of optimal tail estimates for the binomial distribution (see [14] for generalisations).

THEOREM 14. *For every positive integer n, every $p \in (0,1)$, and all $q \in [0,p]$,*

$$\mathbb{P}\big(\mathrm{Bin}(n,p) \leqslant nq\big) \leqslant \exp\big(-n \cdot i_p(q)\big) = \exp\big(-n \cdot D_{KL}\big(\mathrm{Ber}(q) \,\|\, \mathrm{Ber}(p)\big)\big).$$

PROOF. Let $Y = (Y_1, \dots, Y_n)$ be a sequence of i.i.d. $\mathrm{Ber}(p)$ random variables, let $\mathscr{A}$ denote the event that $Y_1 + \dots + Y_n \leqslant nq$, and let $Y' = (Y_1', \dots, Y_n')$ be Y conditioned on $\mathscr{A}$. By Proposition 10,

$$-\log \mathbb{P}\big(\mathrm{Bin}(n,p) \leqslant nq\big) = -\log \mathbb{P}(\mathscr{A}) = D_{KL}(Y' \,\|\, Y) = I_p(Y') \overset{(20)}{\geqslant} \sum_{k=1}^{n} I_p(Y_k').$$

By symmetry, for every $k \in [\![n]\!]$,

$$\mathbb{E}[Y_k'] = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}[Y_j'] \leqslant q.$$

In particular, since i_p is decreasing on $[0,p]$ and $q \leqslant p$, we have

$$I_p(Y_k') = i_p\big(\mathbb{E}[Y_k']\big) \geqslant i_p(q),$$

which concludes the proof of the theorem. □

4.3. *The key lemma.* The following is our key lemma. Its role in the proof of Theorem 5 will be analogous to the role that Pinsker's inequality (Lemma 8) played in the proof of Theorem 1.

LEMMA 15. *Let Y be a $\{0,1\}$-valued random variable and let $E_1, \dots, E_m$ be a sequence of Z-measurable events, for some random variable Z. Suppose that $\mathbb{E}[Y \mid Z] \leqslant p'$ for some $p' > 0$. Then, letting $\mu = \mathbb{E}[Y] = \mathbb{P}(Y = 1)$,*

$$I_p(Y \mid Z) - I_p(Y) \geqslant \frac{1}{2p'} \sum_{i=1}^{m} \big(\mathbb{P}(Y = 1 \mid E_i) - \mu\big)^2 \mathbb{P}(E_i) - \frac{p'}{2} \sum_{1 \leqslant i < j \leqslant m} \mathbb{P}(E_i \cap E_j). \tag{22}$$

REMARK. One may verify that the following identity holds for all $p \in (0,1)$ and random variables $Y \in \{0,1\}$ and Z:

$$I_p(Y \mid Z) - I_p(Y) = H(Y) - H(Y \mid Z). \tag{23}$$

This sheds some light on why the right-hand side of (22) does not depend on p. (We thank one of the referees for pointing (23) out to us.) Nevertheless, the form (22) is the one that we apply below.

Let us first show that Lemma 15 generalises Pinsker's inequality (Lemma 8) for $\{0,1\}$-valued random variables. More precisely, let $Y \in \{0,1\}$ and Z be two random variables and let $Y \times Z$ be the random variable whose independent coordinates have marginals Y and Z. Let E_1 be the Z-measurable event $\big\{\mathbb{P}(Y = 1 \mid Z) \leqslant \mathbb{P}(Y = 1)\big\}$ and let E_2 be the complementary event. As $\mathbb{P}(Y = 1 \mid Z) - \mathbb{P}(Y = 1)$ is nonpositive on E_1 (respectively, nonnegative on E_2), we have

$$d_{TV}\big((Y,Z), Y \times Z\big) = \sum_{i=1}^{2} (-1)^i \cdot \big(\mathbb{P}(Y = 1 \mid E_i) - \mathbb{P}(Y = 1)\big) \cdot \mathbb{P}(E_i).$$

In particular, the Cauchy–Schwarz Inequality gives

$$d_{TV}\big((Y,Z), Y\times Z\big)^2 \leqslant \left(\sum_{i=1}^{2}\big(\mathbb{P}(Y=1\mid E_i)-\mathbb{P}(Y=1)\big)^2\cdot\mathbb{P}(E_i)\right)\cdot\big(\mathbb{P}(E_1)+\mathbb{P}(E_2)\big).$$

It thus follows from Lemma 15, invoked with $p'=1$, that

$$d_{TV}\big((Y,Z), Y\times Z\big)^2 \leqslant 2\cdot\big(I_p(Y\mid Z)-I_p(Y)\big) \overset{(23)}{=} 2\cdot\big(H(Y)-H(Y\mid Z)\big); \tag{24}$$

this is precisely Pinsker's inequality (Lemma 8). In the proof of Theorem 5, we will use Lemma 15 with $p'=p$, which will result in a much stronger bound.

PROOF OF LEMMA 15. Observe first that the case $\mu=0$ is trivial. Indeed, by (19), the left-hand side of (22) is always nonnegative and, when $\mu=0$, each term in the first sum in the right-hand side of (22) vanishes, as $Y=0$ almost surely. We will thus assume that $\mu>0$. For the sake of brevity, let $g:=\mathbb{E}[Y\mid Z]$, so that

$$I_p(Y\mid Z)=\mathbb{E}\big[I_p(Y^Z)\big]=\mathbb{E}[i_p(g)],$$

where i_p is the function defined in (17). Expanding i_p into a Taylor series of order two around μ with Lagrange remainder gives

$$i_p(g)=i_p(\mu)+i_p'(\mu)\cdot(g-\mu)+i_p''(\xi_g)\cdot\frac{(g-\mu)^2}{2} \tag{25}$$

for some ξ_g with $0<\xi_g\leqslant\max\{\mu,g\}$. Recall from (17) and (18) that the first term $i_p(\mu)$ is $I_p\big(\mathrm{Ber}(\mu)\big)=I_p(Y)$ and that $i_p''(\xi)=\frac{1}{\xi}+\frac{1}{1-\xi}$. When we take expectations (over Z) of both sides of (25), the term $i_p'(\mu)\cdot(g-\mu)$ disappears, as $\mathbb{E}[g]=\mathbb{E}[Y]=\mu$, and thus we end up with

$$I_p(Y\mid Z)-I_p(Y)=\mathbb{E}\left[\left(\frac{1}{\xi_g}+\frac{1}{1-\xi_g}\right)\cdot\frac{(g-\mu)^2}{2}\right].$$

Since $\mu,g\leqslant p'$, we have

$$\frac{1}{\xi_g}+\frac{1}{1-\xi_g}\geqslant\frac{1}{\xi_g}\geqslant\min\left\{\frac{1}{\mu},\frac{1}{g}\right\}\geqslant\frac{1}{p'}$$

and we conclude that

$$I_p(Y\mid Z)-I_p(Y)\geqslant\frac{1}{2p'}\cdot\mathbb{E}\left[(g-\mu)^2\right]\geqslant\frac{1}{2p'}\int_{E_1\cup\cdots\cup E_m}(g-\mu)^2\,d\mathbb{P}. \tag{26}$$

It follows from Bonferroni's inequality (inclusion-exclusion) that

$$\int_{E_1\cup\cdots\cup E_m}(g-\mu)^2\,d\mathbb{P}\geqslant\sum_{i=1}^{m}\int_{E_i}(g-\mu)^2\,d\mathbb{P}-\sum_{1\leqslant i<j\leqslant m}\int_{E_i\cap E_j}(g-\mu)^2\,d\mathbb{P}.$$

Since $0\leqslant g,\mu\leqslant p'$, then $(g-\mu)^2\leqslant(p')^2$. Applying the Cauchy–Schwarz Inequality to each of the terms of the first sum above, we obtain

$$\begin{aligned}\int_{E_1\cup\cdots\cup E_m}(g-\mu)^2\,d\mathbb{P}&\geqslant\sum_{i=1}^{m}\left(\frac{1}{\mathbb{P}(E_i)}\int_{E_i}g\,d\mathbb{P}-\mu\right)^2\mathbb{P}(E_i)-\sum_{1\leqslant i<j\leqslant m}\int_{E_i\cap E_j}(p')^2\,d\mathbb{P}\\&=\sum_{i=1}^{m}\big(\mathbb{P}(Y=1\mid E_i)-\mu\big)^2\mathbb{P}(E_i)-(p')^2\sum_{1\leqslant i<j\leqslant m}\mathbb{P}(E_i\cap E_j),\end{aligned}$$

which, substituted into (26), yields the desired inequality (22). □

5. Upper bounds for the lower tail. In this section, we prove Theorem 5. Recall that we are given a hypergraph $\mathscr{H}$ on a vertex set V and that R denotes a random subset of V where every element is included independently with probability p.

5.1. *First reductions.* Let $Y = (Y_v)_{v \in V}$ be the indicator of R conditioned on the lower tail event $e(\mathscr{H}[R]) \leqslant \eta p^r e(\mathscr{H})$. Proposition 10 and the definition of I_p give

$$-\log \mathbb{P}\big(e(\mathscr{H}[R]) \leqslant \eta p^r e(\mathscr{H})\big) = I_p(Y),$$

so from now on $I_p(Y)$ will be our main focus. It will be convenient to define, for every $W \subseteq V$,

$$\mathscr{I}(W) \coloneqq \sum_{v \in V \setminus W} I_p\big(Y_v \mid (Y_w)_{w \in W}\big).$$

The point of making this definition is that

$$\begin{aligned} I_p(Y) &\overset{(20)}{=} I_p\big((Y_v)_{v \in V \setminus W} \mid (Y_w)_{w \in W}\big) + I_p\big((Y_w)_{w \in W}\big) \\ &\overset{(16)}{\geqslant} I_p\big((Y_v)_{v \in V \setminus W} \mid (Y_w)_{w \in W}\big) \overset{(21)}{\geqslant} \sum_{v \in V \setminus W} I_p\big(Y_v \mid (Y_w)_{w \in W}\big) = \mathscr{I}(W), \end{aligned} \tag{27}$$

and thus our goal becomes to find a set W such that $\mathscr{I}(W) \geqslant (1-\varepsilon)\Phi_X(\eta+\varepsilon) - C$.

We will relate $\mathscr{I}(W)$ to the quantity $\Phi(\eta+\varepsilon)$ in the following way. First, define the function $f\colon [0,1]^V \to \mathbb{R}$ by letting, for each $q \in [0,1]^V$,

$$f(q) \coloneqq \sum_{A \in \mathscr{H}} d_A \prod_{v \in A} q_v.$$

In other words, $f(q)$ is the expected number of edges of $\mathscr{H}$ induced by a random subset of V obtained by retaining each $v \in V$ independently with probability q_v. With this definition

$$\Phi(\eta+\varepsilon) = \min\left\{ \sum_{v \in V} i_p(q_v) : q \in [0,1]^V, f(q) \leqslant (\eta+\varepsilon)p^r e(\mathscr{H}) \right\}.$$

Second, given a $W \subseteq V$, we define a *random* function $q^W \colon V \to [0,1]$ by letting, for each $v \in V$,

$$q_v^W \coloneqq \begin{cases} \mathbb{E}\left[Y_v \mid (Y_w)_{w \in W}\right] & \text{if } v \notin W, \\ p & \text{otherwise.} \end{cases}$$

Finally, we write

$$\begin{aligned} \mathscr{I}(W) &\overset{(*)}{=} \sum_{v \in V \setminus W} \mathbb{E}[i_p(q_v^W)] = \mathbb{E}\left[\sum_{v \in V} i_p(q_v^W)\right] \\ &\overset{(\dagger)}{\geqslant} \mathbb{P}\left(f(q^W) \leqslant (\eta+\varepsilon)p^r e(\mathscr{H})\right) \cdot \Phi(\eta+\varepsilon), \end{aligned} \tag{28}$$

where $(*)$ follows from the definitions of H, i_p, and q^W; and where $(\dagger)$ uses $i_p \geqslant 0$ and bounds the expectation from below by the probability of the event $f(q^W) \leqslant (\eta+\varepsilon)p^r e(\mathscr{H})$ times the minimum of the sum $\sum_{v \in V} i_p(q_v^W)$ on that event. In particular, it suffices to produce a set W such that

$$\mathbb{P}\left(f(q^W) \leqslant (\eta+\varepsilon)p^r e(\mathscr{H})\right) \geqslant 1-\varepsilon. \tag{29}$$

Conditioning on $\big(Y_w\big)_{w\in W}$ for various $W \subseteq V$ will repeat so much that it is better to have a shorthand for it. Define therefore

$$\mathbb{E}_W[\cdot] \coloneqq \mathbb{E}\,[\cdot \mid (Y_w)_{w\in W}] \tag{30}$$

(so that our q^W can now be written as $q_v^W = \mathbb{E}_W[Y_v]$ for $v \notin W$). For similar reasons, given an $A \subseteq V$, define

$$Y_A \coloneqq \prod_{a\in A} Y_a.$$

Since $f(Y) = e(\mathscr{H}[R]) \leqslant \eta p^r e(\mathscr{H})$ (and, consequently, $\mathbb{E}_W[f(Y)] \leqslant \eta p^r e(\mathscr{H})$ almost surely for every $W \subseteq V$), we may obtain lower bounds on the probability in the left-hand side of (29) by bounding from above the right-hand side of the following inequality:

$$\big|f(q^W) - \mathbb{E}_W[f(Y)]\big| \leqslant \sum_{A\in\mathscr{H}} d_A \cdot \left|\prod_{a\in A} \mathbb{E}_W(Y_a) - \mathbb{E}_W\left[Y_A\right]\right|.$$

In order to do so, we will quantify the difference between $\prod_{a\in A} \mathbb{E}_W[Y_a]$ and $\mathbb{E}_W[Y_A]$ for a typical $A \in \mathscr{H}$. This is related to conditioned almost independence of the variables $\{Y_a\}_{a\in A}$. However, we are not studying full independence, but only with respect to the event that all Y_v are 1. To continue our analysis, we need a few preliminaries, which will be the topic of the next section.

5.2. *Preliminaries.* At various places we will need the following corollary of Harris's inequality:

CLAIM 16. $\mathbb{E}_W[Y_A] \leqslant p^{|A|}$ *for all* $W \subseteq V$ *and all* $A \subseteq V \setminus W$.

PROOF. Fix some possible value $y \in \{0,1\}^W$ for $(Y_w)_{w\in W}$. Writing E for the event $A \subseteq R$ and recalling that Y is the indicator of R conditioned on the lower tail event $e(\mathscr{H}[R]) \leqslant \eta p^r e(\mathscr{H})$,

$$\begin{aligned}
\mathbb{E}\big[Y_A \mid (Y_w)_{w\in W} = y\big] &= \frac{\mathbb{P}\big(Y_A = 1, (Y_w)_{w\in W} = y\big)}{\mathbb{P}\big((Y_w)_{w\in W} = y\big)} \\
&= \frac{\mathbb{P}\Big(E, (R_w)_{w\in W} = y, e(\mathscr{H}[R]) \leqslant \eta p^r e(\mathscr{H})\Big)}{\mathbb{P}\Big((R_w)_{w\in W} = y, e(\mathscr{H}[R]) \leqslant \eta p^r e(\mathscr{H})\Big)} \\
&= \frac{\mathbb{P}\Big(E, e(\mathscr{H}[R]) \leqslant \eta p^r e(\mathscr{H}) \,\Big|\, (R_w)_{w\in W} = y\Big)}{\mathbb{P}\Big(e(\mathscr{H}[R]) \leqslant \eta p^r e(\mathscr{H}) \mid (R_w)_{w\in W} = y\Big)}
\end{aligned}$$

Since the elements of V are included in R independently, conditioning on $(R_w)_{w\in W}$ gives a product measure on $(R_v)_{v\in V\setminus W}$. Moreover, under the conditioned measure, the event E is increasing and the lower tail event $e(\mathscr{H}[R]) \leqslant \eta p^r e(\mathscr{H})$ is decreasing. The claim follows from Harris's inequality. □

CLAIM 17. *For every nonempty, finite set* A *and every function* $F\colon \mathscr{P}(A) \to \mathbb{R}$,

$$F(A) - \prod_{a\in A} F(\{a\}) = \sum_{\substack{B\subseteq A\\ |B|\geqslant 2}} \sum_{b\in B} \frac{1}{\binom{|A|}{|B|}|B|} \cdot \big(F(B) - F(B\setminus\{b\})F(\{b\})\big) \prod_{a\in A\setminus B} F(\{a\}).$$

(As usual, $\mathscr{P}(A)$ denotes the power set of A.)

PROOF. The identity holds trivially when $|A| = 1$ and we may thus assume that $|A| \geqslant 2$. Observe first that the right-hand side is a linear combination of terms of the form

$$K_\emptyset := \prod_{a \in A} F(\{a\}) \qquad \text{and} \qquad K_B := F(B) \cdot \prod_{a \in A \setminus B} F(\{a\}),$$

where $B \subseteq A$ satisfies $|B| \geqslant 2$. The term $K_\emptyset$ appears only when $|B| = 2$ in the outer sum and it is easy to verify that its coefficient is

$$-\binom{|A|}{2} \cdot 2 \cdot \frac{1}{\binom{|A|}{2} \cdot 2} = -1.$$

Fix an arbitrary $B \subseteq A$ with $|B| \geqslant 2$. On the one hand, the term K_B appears with a positive sign exactly $|B|$ times (once for each $b \in B$) and the respective coefficient is

$$\frac{1}{\binom{|A|}{|B|}|B|};$$

on the other hand, it appears with a negative sign (B is then in fact $B \setminus \{b\}$) exactly $|A| - |B|$ times (once for each $b \in A \setminus B$) and the respective coefficient is (note that $|B| \leqslant |A| - 1$ in this case)

$$\frac{-1}{\binom{|A|}{|B|+1}(|B|+1)}$$

In particular, when $B \neq A$, then the positive and the negative contributions cancel, as

$$|B| \cdot \frac{1}{\binom{|A|}{|B|}|B|} = \frac{1}{\binom{|A|}{|B|}} = (|A| - |B|) \cdot \frac{1}{\binom{|A|}{|B|+1}(|B|+1)},$$

and it is easy to check that the sum of the coefficients of K_A is 1. □

5.3. *The argument.* Fix an arbitrary nonempty $A \subseteq V \setminus W$. Applying Claim 17 with $F(B) = \mathbb{E}_W[Y_B]$ yields

$$\mathbb{E}_W[Y_A] - \prod_{a \in A} \mathbb{E}_W[Y_a] = \sum_{\substack{B \subseteq A \\ |B| \geqslant 2}} \sum_{b \in B} \frac{1}{\binom{|A|}{|B|}|B|} \cdot \underbrace{\left(\mathbb{E}_W[Y_B] - \mathbb{E}_W[Y_{B \setminus \{b\}}]\mathbb{E}_W[Y_b]\right)}_{D_W(B,b)} \cdot \prod_{a \in A \setminus B} \mathbb{E}_W[Y_a]$$

(this is the definition of D_W). Consequently, by the triangle inequality,

$$\begin{aligned} \left|\mathbb{E}_W[Y_A] - \prod_{a \in A} \mathbb{E}_W[Y_a]\right| &\leqslant \sum_{\substack{B \subseteq A \\ |B| \geqslant 2}} \sum_{b \in B} \frac{1}{\binom{|A|}{|B|}|B|} \cdot |D_W(B,b)| \cdot \prod_{a \in A \setminus B} \mathbb{E}_W[Y_a] \\ &\overset{(*)}{\leqslant} \sum_{\substack{B \subseteq A \\ |B| \geqslant 2}} \sum_{b \in B} \frac{1}{\binom{|A|}{|B|}|B|} \cdot |D_W(B,b)| \cdot p^{|A|-|B|}, \end{aligned}$$

where $(*)$ follows from Claim 16. We sum this inequality over all $A \in \mathscr{H} - W = \mathscr{H}[V \setminus W]$, take expectation over $\left(Y_v\right)_{v \in W}$, and get (recall that our hypergraph is r-uniform, so $|A| = r$

for every $A \in \mathscr{H}$)

$$\mathbb{E}\left[\sum_{A\in\mathscr{H}-W} d_A \cdot \left|\mathbb{E}_W[Y_A] - \prod_{a\in A}\mathbb{E}_W[Y_a]\right|\right] \leqslant \mathbb{E}\left[\sum_{A\in\mathscr{H}-W}\sum_{\substack{B\subseteq A\\ |B|\geqslant 2}}\sum_{b\in B}\frac{d_A}{\binom{r}{|B|}|B|}\cdot |D_W(B,b)|\cdot p^{r-|B|}\right]. \tag{31}$$

We now wish to apply the Cauchy–Schwarz Inequality to the right-hand side of (31). However, since the resulting expression would be too long, we first define

$$\mathscr{E}(W) := \mathbb{E}\left[\sum_{A\in\mathscr{H}-W}\sum_{\substack{B\subseteq A\\ |B|\geqslant 2}}\sum_{b\in B}\frac{d_A\cdot D_W(B,b)^2}{\binom{r}{|B|}|B|\cdot p^{2|B|}}\right], \tag{32}$$

and then Cauchy–Schwarz yields

$$\begin{aligned}\mathbb{E}&\left[\sum_{A\in\mathscr{H}-W} d_A \cdot \left|\mathbb{E}_W[Y_A] - \prod_{a\in A}\mathbb{E}_W[Y_a]\right|\right] \\ &\leqslant \left(\sum_{A\in\mathscr{H}-W}\sum_{\substack{B\subseteq A\\ |B|\geqslant 2}}\sum_{b\in B}\frac{d_A p^{2r}}{\binom{r}{|B|}|B|}\right)^{1/2}\cdot\mathscr{E}(W)^{1/2} \\ &= p^r\big((r-1)e(\mathscr{H}-W)\big)^{1/2}\cdot\mathscr{E}(W)^{1/2},\end{aligned} \tag{33}$$

where we used the identity $\sum_{B,b}\frac{1}{\binom{r}{|B|}|B|} = r-1$, which holds because enumerating over all $B \subseteq A$ of a given size and all $b \in B$ cancels the denominator perfectly. Let us remark that most readers might be better off ignoring all these combinatorial factors. We chose to estimate them carefully in order to optimise the dependency of λ and C (from the statement of the theorem) on r. However, in most applications r will be an absolute constant.

The essence of our argument is establishing the following dichotomy: Either

(i) $\mathbb{E}[\mathscr{E}(W)]$ is quite small, or
(ii) $\mathscr{I}(W \cup W') \geqslant \mathscr{I}(W) + \Omega\big(p|V|\big)$ for some small $W' \subseteq V \setminus W$.

If (i) holds, then, by (33), we will have that $\big|\mathbb{E}_W[Y_A] - \prod_{a\in A}\mathbb{E}_W[Y_a]\big|$ is small (on average), and a few simple manipulations (done at the end of the proof of Theorem 5, page 28) will show that our candidate set W satisfies (29). Otherwise, (ii) holds and we replace W with $W \cup W'$; this can happen only $O(1)$ times since

$$\begin{aligned}\mathscr{I}(W) &\overset{(27)}{\leqslant} I_p(Y) = -\log\mathbb{P}\big(e(\mathscr{H}[R]) \leqslant \eta p^r e(\mathscr{H})\big) \\ &\leqslant -\log\mathbb{P}(R=\emptyset) = |V|\cdot\log\frac{1}{1-p} \leqslant |V|\cdot\frac{p}{1-p} \leqslant |V|\cdot\frac{p}{1-p_0}.\end{aligned} \tag{34}$$

LEMMA 18. *For all positive α, β, and K, there exist λ and V_0 such that the following holds: If $|V| \geqslant V_0$ and $\mathscr{H}$ satisfies (7) for every $s \in [\![r]\!]$, then there exists a set $W \subseteq V$ with at most $\alpha|V|$ elements that satisfies*

$$\mathscr{E}(W) \leqslant \beta \cdot e(\mathscr{H}).$$

PROOF. Without loss of generality, we may assume that $\alpha < 1/2$, $\beta < 1$, and $K > 1$. We first define a few constants:

$$\gamma := \frac{\beta^2}{300Kr}, \qquad \tau := \alpha\gamma(1-p_0), \qquad \lambda := \frac{\tau}{2r}, \qquad \text{and} \qquad V_0 := 8r^2/\tau. \tag{35}$$

A short calculation shows that the definition of V_0 guarantees that

$$\frac{\tau \cdot V_0/2 - r}{V_0/2 - r} \geqslant \frac{\tau}{2^{1/r}} \qquad \text{and} \qquad \frac{V_0/2}{V_0/2-1} \leqslant 2^{1/(2r)}. \tag{36}$$

As explained above, we shall build our set W in several rounds, starting with W being the empty set. In each round, we will use the following claim, which implements the dichotomy discussed above.

CLAIM 19. *Suppose that $W \subseteq V$ satisfies $\mathscr{E}(W) > \beta \cdot e(\mathscr{H})$ and $|V \setminus W| \geqslant V_0/2$. Then there exists a set $W' \subseteq V \setminus W$ with at most $\tau|V|$ elements such that*

$$\mathscr{I}(W \cup W') \geqslant \mathscr{I}(W) + \gamma p|V|. \tag{37}$$

PROOF OF CLAIM 19. Let W' be a uniformly chosen subset of $V \setminus W$ with density τ, that is, with exactly $\lfloor \tau \cdot |V \setminus W| \rfloor$ elements. We will show that, under the assumption that $\mathscr{E}(W) > \beta e(\mathscr{H})$ and $|V \setminus W| \geqslant V_0/2$, we have

$$\mathbb{E}\big[\mathscr{I}(W \cup W')\big] \geqslant (1-\tau) \cdot \mathscr{I}(W) + 2\gamma p|V|. \tag{38}$$

Consequently, since

$$\tau \cdot \mathscr{I}(W) \overset{(34)}{\leqslant} \frac{\tau}{1-p_0} \cdot p|V| \overset{(35)}{=} \alpha\gamma p|V| < \gamma p|V|, \tag{39}$$

the desired inequality (37) must hold for some W'.

We now write

$$\begin{aligned} \mathscr{I}(W \cup W') - \mathscr{I}(W) &= \sum_{v \in V \setminus (W \cup W')} I_p(Y_v \mid (Y_w)_{w \in W \cup W'}) - \sum_{v \in V \setminus W} I_p(Y_v \mid (Y_w)_{w \in W}) \\ &= \underbrace{\sum_{v \in V \setminus (W \cup W')} I_p(Y_v \mid (Y_w)_{w \in W \cup W'}) - I_p(Y_v \mid (Y_w)_{w \in W})}_{\mathscr{I}^+_{W'}} - \underbrace{\sum_{v \in W'} I_p(Y_v \mid (Y_w)_{w \in W})}_{\mathscr{I}^-_{W'}}. \end{aligned} \tag{40}$$

By linearity of expectation,

$$\mathbb{E}\left[\mathscr{I}^-_{W'}\right] = \frac{\lfloor \tau \cdot |V \setminus W| \rfloor}{|V \setminus W|} \cdot \mathscr{I}(W) \leqslant \tau \cdot \mathscr{I}(W),$$

and thus (38) will follow if we show that

$$\mathscr{I}^+ := \mathbb{E}\left[\mathscr{I}^+_{W'}\right] \geqslant 2\gamma p|V|. \tag{41}$$

In order to bound $\mathscr{I}^+$ from below, we will apply our main lemma (Lemma 15), conditionally on $(Y_w)_{w \in W}$, with $Z = (Y_w)_{w \in W'}$ and a careful choice of the sequence of Z-measurable events that we shall now define. To this end, for each $v \in V \setminus W$, let

$$\mathscr{H}(v) := \big\{B \subseteq V \setminus W : |B| \geqslant 2, v \in B, \text{and } B \subseteq A \text{ for some } A \in \mathscr{H} - W\big\} \tag{42}$$

and let $\mathscr{G}(v)$ be the random subset of $\mathscr{H}(v)$ formed by including each $B \in \mathscr{H}(v)$ satisfying $B \setminus \{v\} \subseteq W'$ with probability σ_B, which we will specify later, independently for each such B. We note though, already at this stage, that σ_B are independent of Y. The impatient may see their definition in (47)–(48).

Let $S \coloneqq \big(Y_w\big)_{w\in W}$ and, for every $v \in V \setminus (W \cup W')$, let Y_v^S denote Y_v conditioned on S, that is, the random variable whose (random) distribution is the distribution of Y_v conditioned on S. Define

$$J^S(v) \coloneqq I_p\big(Y_v^S \mid \big(Y_w^S\big)_{w\in W'}\big) - I_p\big(Y_v^S\big),$$

The next step is to apply Lemma 15. Recall that we need to supply the lemma with a sequence of events. The number of events in our application will also be random, but it will depend only on W' and $\mathscr{G}(v)$, so let us fix their choice for the time being. For each $B \in \mathscr{G}(v)$, let E_B^S be the event that $Y_{B\setminus\{v\}}^S = 1$; note that E_B^S is $(Y_w^S)_{w\in W'}$-measurable, as $B \setminus \{v\} \subseteq W'$. Since $\mathbb{E}\big[Y_v^S \mid (Y_w^S)_{w\in W'}\big] = E_{W\cup W'}[Y_v] \leqslant p$, by Claim 16, we may apply Lemma 15 with $Y = Y_v^S$, $Z = (Y_w^S)_{w\in W'}$, the events E_B^S, and $p' = p$ to get (recall the definition of D_W given at the start of §5.3)

$$\begin{aligned} J^S(v) &\geqslant \frac{1}{2p} \sum_{B\in\mathscr{G}(v)} \big(\mathbb{P}(Y_v^S = 1 \mid E_B^S) - \mathbb{E}[Y_v^S]\big)^2 \mathbb{P}(E_B^S) - \frac{p}{2} \sum_{\substack{B,B'\in\mathscr{G}(v)\\ B\neq B'}} \mathbb{P}(E_B^S \cap E_{B'}^S) \\ &= \frac{1}{2p} \sum_{B\in\mathscr{G}(v)} \frac{D_W(B,v)^2}{\mathbb{E}_W[Y_{B\setminus\{v\}}]} - \frac{p}{2} \sum_{\substack{B,B'\in\mathscr{G}(v)\\ B\neq B'}} \mathbb{E}_W[Y_{B\setminus\{v\}} \cdot Y_{B'\setminus\{v\}}]. \end{aligned}$$

(Note that we used here that W' and $\mathscr{G}(v)$ are independent of Y.) Since every edge of $\mathscr{G}(v)$ contains v and is disjoint from W, Claim 16 implies that $\mathbb{E}_W[Y_{B\setminus\{v\}}] \leqslant p^{|B|-1}$ and $\mathbb{E}_W[Y_{B\setminus\{v\}} \cdot Y_{B'\setminus\{v\}}] \leqslant p^{|B\cup B'|-1}$ for all $B, B' \in \mathscr{G}(v)$. This observation allows us to simplify our lower bound for $J^S(v)$ to

$$2 \cdot J^S(v) \geqslant \sum_{B\in\mathscr{G}(v)} \frac{D_W(B,v)^2}{p^{|B|}} - \sum_{\substack{B,B'\in\mathscr{G}(v)\\ B\neq B'}} p^{|B\cup B'|} \eqqcolon G(v) - L(v), \tag{43}$$

i.e., $G(v)$ is the first sum and $L(v)$ is the second.

We now return to the $\mathscr{I}^+$ from (41). It is the expectation (over W') of the sum $\mathscr{I}_{W'}^+$ defined in (40), each of whose summands is the expectation (over S) of $J^S(v)$, see Proposition 13. We wish to exchange the sum and expectation, but since the sum is over $v \notin W'$ (recall (40)) and this is an event, we need to condition on it. Hence we arrive at

$$\mathscr{I}^+ \overset{(41)}{=} \sum_{v\in V\setminus W} \mathbb{P}(v \notin W') \cdot \mathbb{E}\big[J^S(v) \mid v \notin W'\big].$$

At this point it will be convenient to switch notation slightly. From now on, we will use $\mathbb{E}'$ and $\mathbb{P}'$ to denote the expectation and the probability over the random choices of the set W' and the hypergraphs $\mathscr{G}(v)$ for all $v \in V \setminus W$. The notations $\mathbb{E}$ and $\mathbb{P}$ will be reserved to the randomness of S. With this notation

$$\begin{aligned} \mathscr{I}^+ &= \sum_{v\in V\setminus W} \mathbb{P}'(v \notin W') \cdot \mathbb{E}\left[\mathbb{E}'\big[J^S(v) \mid v \notin W'\big]\right] \\ &\overset{(43)}{\geqslant} \frac{1-\tau}{2} \cdot \mathbb{E}\left[\sum_{v\in V\setminus W} \mathbb{E}'\big[G(v) - L(v) \mid v \notin W'\big]\right]. \end{aligned} \tag{44}$$

In the remainder of the proof, we shall estimate the right-hand side of (44).

We start with the estimate of the G terms. We define

$$G'(v) \coloneqq \mathbb{E}'\big[G(v) \mid v \notin W'\big] = \sum_{B \in \mathscr{H}(v)} \mathbb{P}'\big(B \in \mathscr{G}(v) \mid v \notin W'\big) \cdot \frac{D_W(B,v)^2}{p^{|B|}}.$$

For every $B \in \mathscr{H}(v)$ we have (recall the assumption that $|V \setminus W| \geqslant V_0/2$)

$$\begin{aligned}\mathbb{P}'\big(B \in \mathscr{G}(v) \mid v \notin W'\big) &= \mathbb{P}'\big(B \setminus \{v\} \subseteq W' \mid v \notin W'\big) \cdot \sigma_B \\ &= \prod_{i=0}^{|B|-2} \frac{\lfloor \tau |V \setminus W| \rfloor - i}{|V \setminus W| - i - 1} \cdot \sigma_B \\ &\geqslant \left(\frac{\tau |V \setminus W| - r}{|V \setminus W| - r}\right)^{|B|-1} \cdot \sigma_B \overset{(36)}{\geqslant} \frac{\tau^{|B|-1}}{2} \cdot \sigma_B.\end{aligned}$$

We conclude that

$$G'(v) \geqslant \frac{1}{2\tau} \sum_{B \in \mathscr{H}(v)} \frac{\tau^{|B|} \cdot \sigma_B \cdot D_W(B,v)^2}{p^{|B|}}. \tag{45}$$

Summing (45) over all $v \in V \setminus W$ yields (recall the definitions of the hypergraphs $\mathscr{H}(v)$ and of $\deg_{\mathscr{H}-W}$ given in (42) and (5) respectively)

$$\sum_{v \in V \setminus W} G'(v) \geqslant \frac{1}{2\tau} \sum_{A \in \mathscr{H}-W} \sum_{\substack{B \subseteq A \\ |B| \geqslant 2}} \sum_{v \in B} \frac{d_A}{\deg_{\mathscr{H}-W} B} \cdot \frac{\tau^{|B|} \cdot \sigma_B \cdot D_W(B,v)^2}{p^{|B|}}, \tag{46}$$

cf. the definition of $\mathscr{E}(W)$ given in (32). This is a good moment to finally define the probabilities σ_B. We let

$$\sigma_B \coloneqq \mu \cdot \frac{\deg_{\mathscr{H}-W} B}{\binom{r}{|B|} |B| (\tau p)^{|B|}}, \tag{47}$$

where

$$\mu \coloneqq \frac{\beta \tau p |V|}{16 K (r-1) e(\mathscr{H})}. \tag{48}$$

Note that $\sigma_B \leqslant 1$ as

$$\deg_{\mathscr{H}-W} B \leqslant \Delta_{|B|}(\mathscr{H}) \overset{(7)}{\leqslant} K \cdot (\lambda p)^{|B|-1} \cdot \frac{e(\mathscr{H})}{|V|} \overset{(35)}{\leqslant} K \cdot (\tau p)^{|B|-1} \cdot \frac{e(\mathscr{H})}{|V|} \overset{(48)}{\leqslant} \frac{(\tau p)^{|B|}}{\mu}.$$

Substituting (47) into (46) yields precisely

$$\mathbb{E}\Bigg[\sum_{v \in V \setminus W} G'(v)\Bigg] \geqslant \frac{\mu}{2\tau} \sum_{A \in \mathscr{H}-W} \sum_{\substack{B \subseteq A \\ |B| \geqslant 2}} \sum_{v \in B} \frac{d_A \cdot \mathbb{E}[D_W(B,v)^2]}{\binom{r}{|B|} |B| p^{2|B|}} \overset{(32)}{=} \frac{\mu}{2\tau} \cdot \mathscr{E}(W). \tag{49}$$

This concludes our estimate of the G terms.

The estimate of the L terms in (44) is similar, but somewhat more involved. We define

$$L'(v) \coloneqq \mathbb{E}'\big[L(v) \mid v \notin W'\big] = \sum_{\substack{B, B' \in \mathscr{H}(v) \\ B \neq B'}} \mathbb{P}'\big(B, B' \in \mathscr{G}(v) \mid v \notin W'\big) \cdot p^{|B \cup B'|}.$$

Thus, we need a second moment estimate for the sum of indicators of $B\in\mathscr{G}(v)$ over all $B\in\mathscr{H}(v)$. Note first that, for each $B\neq B'$,

$$\begin{aligned}\mathbb{P}'\big(B,B'\in\mathscr{G}(v)\mid v\notin W'\big)&=\mathbb{P}'\big((B\cup B')\setminus\{v\}\subseteq W'\mid v\notin W'\big)\cdot\sigma_B\sigma_{B'}\\&=\prod_{i=0}^{|B\cup B'|-2}\frac{\lfloor\tau|V\setminus W|\rfloor-i}{|V\setminus W|-i-1}\cdot\sigma_B\sigma_{B'}\\&\leqslant\left(\frac{\tau|V\setminus W|}{|V\setminus W|-1}\right)^{|B\cup B'|-1}\cdot\sigma_B\sigma_{B'}\overset{(36)}{\leqslant}2\tau^{|B\cup B'|-1}\cdot\sigma_B\sigma_{B'}.\end{aligned}$$

Hence

$$L'(v)\leqslant\frac{2}{\tau}\sum_{\substack{B,B'\in\mathscr{H}(v)\\B\neq B'}}(\tau p)^{|B\cup B'|}\cdot\sigma_B\sigma_{B'}. \tag{50}$$

Summing (50) over all $v\in V\setminus W$ gives (again, using the definitions of $\mathscr{H}(v)$ and $\deg_{\mathscr{H}-W}$)

$$\begin{aligned}\sum_{v\in V\setminus W}L'(v)&\leqslant\frac{2}{\tau}\sum_{A,A'\in\mathscr{H}-W}\sum_{\substack{B\subseteq A,B'\subseteq A'\\|B|,|B'|\geqslant 2\\B\neq B'}}\sum_{v\in B\cap B'}\frac{d_A}{\deg_{\mathscr{H}-W}B}\cdot\frac{d_{A'}}{\deg_{\mathscr{H}-W}B'}(\tau p)^{|B\cup B'|}\sigma_B\sigma'_B\\&\overset{(47)}{=}\frac{2\mu^2}{\tau^2 p}\cdot\underbrace{\sum_{A,A'\in\mathscr{H}-W}\sum_{\substack{B\subseteq A,B'\subseteq A'\\|B|,|B'|\geqslant 2\\B\neq B'}}\frac{|B\cap B'|\cdot d_Ad_{A'}}{\binom{r}{|B|}|B|\binom{r}{|B'|}|B'|(\tau p)^{|B\cap B'|-1}}}_{(*)},\end{aligned}$$

where we used the identity $|B\cup B'|+|B\cap B'|=|B|+|B'|$ for the powers of τp. Rearranging gives

$$(*)=\sum_{A\in\mathscr{H}-W}d_A\sum_{\substack{B\subseteq A\\|B|\geqslant 2}}\frac{1}{\binom{r}{|B|}|B|}\sum_{s=1}^{r-1}\frac{s}{(\tau p)^{s-1}}\underbrace{\sum_{\substack{C\subseteq B\\|C|=s}}\sum_{\substack{A'\in\mathscr{H}-W\\C\subseteq A'}}d_{A'}\sum_{\substack{B'\subseteq A'\\|B|'\geqslant 2\\B'\neq B\\B\cap B'=C}}\frac{1}{\binom{r}{|B'|}|B'|}}_{S_{B,s}}.$$

Now, for every $A'\in\mathscr{H}-W$, every $s\geqslant 1$, and every $C\subseteq A'$ with $|C|=s$,

$$\sum_{C\subseteq B'\subseteq A'}\frac{1}{\binom{r}{|B'|}|B'|}=\sum_{b'=s}^{r}\frac{\binom{r-s}{b'-s}}{\binom{r}{b'}b'}=\sum_{b'=s}^{r}\frac{\binom{b'}{s}}{\binom{r}{s}b'}=\sum_{b'=s}^{r}\frac{\binom{b'-1}{s-1}}{\binom{r}{s}s}=\frac{1}{s}\leqslant 1.$$

Hence, for every B with at most r elements and every $s\geqslant 1$,

$$\begin{aligned}S_{B,s}&\leqslant\sum_{\substack{C\subseteq B\\|C|=s}}\sum_{\substack{A'\in\mathscr{H}-W\\C\subseteq A'}}d_{A'}\leqslant\binom{|B|}{s}\cdot\Delta_s(\mathscr{H})=\frac{|B|}{s}\binom{|B|-1}{s-1}\cdot\Delta_s(\mathscr{H})\\&\overset{(7)}{\leqslant}\frac{|B|}{s}\binom{|B|-1}{s-1}\cdot(\lambda p)^{s-1}\cdot K\cdot\frac{e(\mathscr{H})}{|V|}\leqslant\frac{|B|}{s}\cdot(r\lambda p)^{s-1}\cdot K\cdot\frac{e(\mathscr{H})}{|V|}.\end{aligned}$$

Consequently,

$$(*) \leqslant \sum_{A \in \mathscr{H} - W} d_A \sum_{\substack{B \subseteq A \\ |B| \geqslant 2}} \frac{1}{\binom{r}{|B|}} \sum_{s=1}^{r-1} \left(\frac{r\lambda}{\tau}\right)^{s-1} \cdot K \cdot \frac{e(\mathscr{H})}{|V|}$$
$$= e(\mathscr{H} - W) \cdot (r-1) \cdot \sum_{s=1}^{r-1} \left(\frac{r\lambda}{\tau}\right)^{s-1} \cdot K \cdot \frac{e(\mathscr{H})}{|V|}.$$

Since $r\lambda = \tau/2$, we conclude that

$$\sum_{v \in V \setminus W} L'(v) \leqslant \frac{2\mu^2}{\tau^2 p} \cdot (r-1) \cdot 2K \cdot \frac{e(\mathscr{H})^2}{|V|} \overset{(48)}{=} \frac{\mu}{\tau} \cdot \frac{\beta e(\mathscr{H})}{4}. \tag{51}$$

Combining this with the estimate (49) gives

$$\mathscr{I}^+ \overset{(44)}{\geqslant} \frac{1-\tau}{2} \cdot \mathbb{E}\left[\sum_{v \in V \setminus W} G'(v) - L'(v)\right] \overset{(49,51)}{\geqslant} \frac{(1-\tau)}{2} \cdot \frac{\mu}{\tau} \cdot \left(\frac{\mathscr{E}(W)}{2} - \frac{\beta e(\mathscr{H})}{4}\right)$$
$$\overset{(\dagger)}{>} \frac{(1-\tau)}{2} \cdot \frac{\mu}{\tau} \cdot \frac{\beta e(\mathscr{H})}{4} \overset{(48)}{=} \frac{(1-\tau)\beta^2}{128K(r-1)} \cdot p|V| \overset{(35)}{\geqslant} 2\gamma p|V|,$$

where (†) follows from our assumption that $\mathscr{E}(W) > \beta e(\mathscr{H})$. The claim thus follows from the discussion before (41). □

Proof of Lemma 18, continued. Suppose that the assertion of the lemma is not true, that is, $\mathscr{E}(W) > \beta \cdot e(\mathscr{H})$ for every $W \subseteq V$ with at most $\alpha|V|$ elements. We will construct a sequence $W_0, \dots, W_j$ of subsets of V, where $j = \lfloor \alpha/\tau \rfloor + 1$, such that, for each $i \in \{0, \dots, j\}$,

(i) $|W_i| \leqslant i \cdot \tau|V|$ and
(ii) $\mathscr{I}(W_i) \geqslant i \cdot \gamma p|V|$.

If such a sequence existed, we would have

$$\mathscr{I}(W_j) \geqslant j \cdot \gamma p|V| > (\alpha/\tau) \cdot \gamma p|V| \overset{(35)}{=} |V| \cdot \frac{p}{1-p_0},$$

which contradicts (34).

We start by letting $W_0 = \emptyset$. Suppose that $0 \leqslant i \leqslant j-1$ and that W_i has already been defined so that (i) and (ii) hold. Since

$$|W_i| \leqslant i \cdot \tau|V| \leqslant \lfloor \alpha/\tau \rfloor \cdot \tau|V| \leqslant \alpha|V|,$$

we have $\mathscr{E}(W_i) > \beta \cdot e(\mathscr{H})$ by the contradictory assumption. We note also that $|V \setminus W_i| \geqslant (1-\alpha)|V| \geqslant |V|/2 \geqslant V_0/2$. In particular, Claim 19, invoked with $W = W_i$, supplies a $W' \subseteq V \setminus W_i$ with at most $\tau|V|$ elements that satisfies (37). We let $W_{i+1} = W_i \cup W'$ and note that

$$|W_{i+1}| = |W_i| + |W'| \overset{\text{(i)}}{\leqslant} i \cdot \tau|V| + \tau|V| = (i+1) \cdot \tau|V|$$

and

$$\mathscr{I}(W_{i+1}) = \mathscr{I}(W_i \cup W') \overset{(37)}{\geqslant} \mathscr{I}(W_i) + \gamma p|V| \overset{\text{(ii)}}{\geqslant} (i+1) \cdot \gamma p|V|,$$

so (i) and (ii) continue to hold with i replaced by $i+1$. This completes the proof of the existence of the sequence of $W_0, \dots, W_j$, which yields the desired contradiction. □

PROOF OF THEOREM 5. Let λ and V_0 be the constants supplied by Lemma 18 invoked with

$$\alpha \coloneqq \frac{\varepsilon}{2K} \qquad \text{and} \qquad \beta \coloneqq \frac{\varepsilon^4}{4r}.$$

(We note here that $\lambda \geqslant 2^{-15}K^{-2}r^{-4}\varepsilon^9(1-p_0)$ and $V_0 = 4r/\lambda$.) We first handle the uninteresting case $|V| < V_0$. Considering, in the definition of $\Phi(\eta)$, the function $q\colon V \to [0,1]$ that assigns zero to all elements of V shows that

$$\Phi(\eta+\varepsilon) \leqslant \Phi(0) \leqslant |V| \cdot i_p(0) = -|V| \cdot \log(1-p) \leqslant -V_0 \cdot \log(1-p_0).$$

In particular, setting $C \coloneqq -V_0 \cdot \log(1-p_0)$ makes the assertion of the theorem hold vacuously.

We may thus assume that $|V| \geqslant V_0$, so that Lemma 18 supplies a set $W \subseteq V$ with at most $\varepsilon/(2K) \cdot |V|$ elements such that $\mathscr{E}(W) \leqslant \varepsilon^4/(4r) \cdot e(\mathscr{H})$. Let $q^W\colon V \to [0,1]$ be the random function defined in the proof outline, that is, $q_v^W \coloneqq \mathbb{E}_W[Y_v]$ for $v \in V \setminus W$ and $q_v^W \coloneqq p$ for $v \in W$. We have

$$\mathbb{E}\left[\sum_{A \in \mathscr{H}-W} d_A \cdot \left|\mathbb{E}_W[Y_A] - \prod_{a \in A} q_a^W\right|\right] \overset{(33)}{\leqslant} p^r \big(re(\mathscr{H})\big)^{1/2} \cdot \mathscr{E}(W)^{1/2} \leqslant \frac{\varepsilon^2}{2} \cdot p^r e(\mathscr{H}).$$

In particular, it follows from Markov's inequality that, with probability at least $1-\varepsilon$,

$$\sum_{A \in \mathscr{H}-W} d_A \prod_{a \in A} q_a^W \leqslant \sum_{A \in \mathscr{H}-W} d_A \cdot \mathbb{E}_W[Y_A] + \frac{\varepsilon}{2} \cdot p^r e(\mathscr{H}).$$

However, the definition of Y implies that, deterministically,

$$\sum_{A \in \mathscr{H}-W} d_A Y_A \leqslant \sum_{A \in \mathscr{H}} d_A Y_A = e(\mathscr{H}[R]) \leqslant \eta p^r e(\mathscr{H})$$

and thus, with probability at least $1-\varepsilon$,

$$\sum_{A \in \mathscr{H}-W} d_A \prod_{a \in A} q_a^W \leqslant (\eta + \varepsilon/2) \cdot p^r e(\mathscr{H}).$$

The definition of q^W and Claim 16 guarantee that $q_v^W \leqslant p$ for every $v \in V$ and, therefore,

$$\sum_{A \in \mathscr{H} \setminus (\mathscr{H}-W)} d_A \prod_{a \in A} q_a^W \leqslant p^r \cdot \big(e(\mathscr{H}) - e(\mathscr{H}-W)\big) \leqslant p^r \cdot |W| \cdot \Delta_1(\mathscr{H})$$

$$\overset{(7)}{\leqslant} p^r \cdot \frac{\varepsilon|V|}{2K} \cdot K \cdot \frac{e(\mathscr{H})}{v(\mathscr{H})} = \frac{\varepsilon}{2} \cdot p^r e(\mathscr{H}).$$

Summarising, with probability at least $1-\varepsilon$, we have

$$f(q^W) = \sum_{A \in \mathscr{H}} d_A \prod_{a \in A} q_a^W \leqslant (\eta+\varepsilon) \cdot p^r e(\mathscr{H}).$$

Hence, we may conclude that

$$I_p(Y) \overset{(27)}{\geqslant} \mathscr{I}(W) \overset{(28)}{\geqslant} \mathbb{P}\big(f(q^W) \leqslant (\eta+\varepsilon) p^r e(\mathscr{H})\big) \cdot \Phi(\eta+\varepsilon) \geqslant (1-\varepsilon) \cdot \Phi(\eta+\varepsilon),$$

as needed. □

6. Lower bounds for the lower tail.

In this section, we prove Theorem 6. We will need the following technical lemma.

LEMMA 20. *For every $p_0 < 1$, there exists a constant K such that the following holds. Suppose that $0 < p \leqslant p_0$ and $0 \leqslant q \leqslant p$, let $Y \sim \mathrm{Ber}(q)$, and let*

$$X := Y \log \frac{q}{p} + (1 - Y) \log \frac{1-q}{1-p}.$$

Then,

$$\mathrm{Var}(X) \leqslant K\mathbb{E}[X] = K i_p(q).$$

PROOF. This is nothing but a calculus exercise, but let us do it in details anyway. Observe first that the case $q = 0$ is trivial. Indeed, i_p is nonnegative and $\mathrm{Var}(X) = 0$ when $q = 0$. We will thus assume that $q > 0$. A direct computation shows that

$$\mathrm{Var}(X) = q(1-q)\left(\log \frac{q}{p} - \log \frac{1-q}{1-p}\right)^2 \overset{(18)}{=} q(1-q)\big(i_p'(q)\big)^2 \leqslant q \cdot \big(i_p'(q)\big)^2.$$

Since $i_p(p) = i_p'(p) = 0$, expanding both $i_p(q)$ and $i_p'(q)$ in Taylor series around p with Lagrange remainder gives $q_1, q_2 \in (q, p)$ such that

$$i_p(q) = \frac{(q-p)^2}{2} \cdot \left(\frac{1}{q_1} + \frac{1}{1-q_1}\right) \geqslant \frac{(q-p)^2}{2p}, \tag{52}$$

$$i_p'(q) = (q-p) \cdot \left(\frac{1}{q_2} + \frac{1}{1-q_2}\right).$$

Suppose first that $q \geqslant p/2$. Our assumption that $p \leqslant p_0$ implies that

$$\frac{1}{q_2} + \frac{1}{1-q_2} \leqslant \frac{1}{q} + \frac{1}{1-p} \leqslant \frac{2}{p} + \frac{1}{1-p_0} \leqslant \left(2 + \frac{p_0}{1-p_0}\right) \cdot \frac{1}{p} = \frac{2-p_0}{1-p_0} \cdot \frac{1}{p}$$

and, consequently,

$$\mathrm{Var}(X) \leqslant q \cdot \big(i_p'(q)\big)^2 \leqslant \left(\frac{2-p_0}{1-p_0}\right)^2 \cdot \frac{q \cdot (q-p)^2}{p^2} \leqslant \left(\frac{2-p_0}{1-p_0}\right)^2 \cdot 2 i_p(q).$$

If, on the other hand, $q < p/2$, then, using the inequality $(a-b)^2 \leqslant 2a^2 + 2b^2$, we get

$$\begin{aligned} \frac{\mathrm{Var}(X)}{p} &\leqslant \frac{q}{p}\left(\log \frac{q}{p} - \log \frac{1-q}{1-p}\right)^2 \leqslant \frac{2q}{p}\left(\log \frac{q}{p}\right)^2 + \frac{2q}{p}\left(\log \frac{1-q}{1-p}\right)^2 \\ &\leqslant \sup_{x \in (0,1/2)} 2x(\log x)^2 + \left(\log \frac{1}{1-p}\right)^2 \leqslant \frac{8}{e^2} + \left(\log \frac{1}{1-p_0}\right)^2, \end{aligned}$$

whereas, since i_p is decreasing in the interval $[0, p]$,

$$\frac{i_p(q)}{p} \geqslant \frac{i_p(p/2)}{p} \geqslant \frac{(p/2)^2}{2p^2} = \frac{1}{8}.$$

□

PROOF OF THEOREM 6. We may assume without loss of generality that $\varepsilon < 1$. Let $q \colon V \to [0,1]$ be the minimiser in the definition of $\Phi\big((1-\varepsilon)\eta\big)$ and let $Y' = (Y_v')_{v \in V}$ be a sequence of independent Bernoulli random variables with $\mathbb{E}[Y_v'] = q_v$ for each $v \in V$, so that

$$\mathbb{E}[f(Y')] \leqslant (1-\varepsilon)\eta\mathbb{E}[f(Y)] \qquad \text{and} \qquad \sum_{v \in V} i_p(q_v) = \Phi\big((1-\varepsilon)\eta\big).$$

We claim that $q_v \leqslant p$ for every $v \in V$. Indeed, otherwise $i_p(q_v) > 0 = i_p(p)$ and changing q_v to p can only decrease $\mathbb{E}[f(Y')]$. Let $\mathscr{Y} \subseteq \{0,1\}^V$ be arbitrary and note that

$$\begin{aligned}\mathbb{P}\big(Y' \in \mathscr{Y}\big) &= \sum_{y\in\mathscr{Y}} \frac{\mathbb{P}(Y'=y)}{\mathbb{P}(Y=y)} \cdot \mathbb{P}(Y=y) = \sum_{y\in\mathscr{Y}} \prod_{v:y_v=1} \frac{q_v}{p} \prod_{v:y_v=0} \frac{1-q_v}{1-p} \cdot \mathbb{P}(Y=y) \\ &\leqslant \max\left\{ \exp\left(\sum_{v:y_v=1} \log\frac{q_v}{p} + \sum_{v:y_v=0} \log\frac{1-q_v}{1-p} \right) : y \in \mathscr{Y} \right\} \cdot \mathbb{P}(Y \in \mathscr{Y}).\end{aligned}$$

In view of this, define, for each $y \in \{0,1\}^V$,

$$J(y) \coloneqq \sum_{v:y_v=1} \log\frac{q_v}{p} + \sum_{v:y_v=0} \log\frac{1-q_v}{1-p},$$

so that the above inequality may be rewritten as

$$\mathbb{P}(Y' \in \mathscr{Y}) \leqslant \max_{y\in\mathscr{Y}} \exp\big(J(y)\big) \cdot \mathbb{P}(Y \in \mathscr{Y}). \tag{53}$$

Now, let K be the constant given by Lemma 20, let $C' \coloneqq K/(2\varepsilon^2)$, and define

$$\begin{aligned}\mathscr{Y}_1 &\coloneqq \big\{ y \in \{0,1\}^V : f(y) \leqslant \eta \mathbb{E}[f(Y)] \big\}, \\ \mathscr{Y}_2 &\coloneqq \big\{ y \in \{0,1\}^V : J(y) \leqslant (1+\varepsilon)\Phi\big((1-\varepsilon)\eta\big) + C' \big\}, \end{aligned} \tag{54}$$

It is immediate from these definitions that

$$\begin{aligned}\mathbb{P}\big(X \leqslant \eta\mathbb{E}[X]\big) = \mathbb{P}(Y \in \mathscr{Y}_1) \geqslant \mathbb{P}(Y \in \mathscr{Y}_1 \cap \mathscr{Y}_2) &\overset{(53)}{\geqslant} \mathbb{P}(Y' \in \mathscr{Y}_1 \cap \mathscr{Y}_2) \cdot \exp\Big(-\max_{y\in\mathscr{Y}_2} J(y) \Big) \\ &\overset{(54)}{\geqslant} \mathbb{P}(Y' \in \mathscr{Y}_1 \cap \mathscr{Y}_2) \cdot \exp\big(-(1+\varepsilon)\Phi\big((1-\varepsilon)\eta\big) - C' \big).\end{aligned}$$

We will show that $\mathbb{P}(Y' \in \mathscr{Y}_1 \cap \mathscr{Y}_2) \geqslant \varepsilon/2$, which will yield the assertion of the theorem with $C \coloneqq C' + \log(2/\varepsilon)$.

Since f is nonnegative, Markov's inequality gives

$$\mathbb{P}\big(f(Y') > \eta\mathbb{E}[f(Y)]\big) \leqslant 1-\varepsilon$$

and thus

$$\mathbb{P}(Y' \in \mathscr{Y}_1) = \mathbb{P}\big(f(Y') \leqslant \eta\mathbb{E}[f(Y)]\big) \geqslant \varepsilon;$$

in particular, it is enough to show that $\mathbb{P}(Y' \notin \mathscr{Y}_2) \leqslant \varepsilon/2$. To this end, examine $J(Y')$. It is a sum of independent variables $(X_v)_{v\in V}$, where each X_v is distributed exactly like the X of Lemma 20, only with q replaced by q_v. In particular,

$$\mathbb{E}[J(Y')] = \sum_{v\in V} \mathbb{E}[X_v] = \sum_{v\in V} i_p(q_v) = \Phi\big((1-\varepsilon)\eta\big)$$

and

$$\operatorname{Var}(J(Y')) = \sum_{v\in V} \operatorname{Var}(X_v) \leqslant K \sum_{v\in V} \mathbb{E}[X_v] = K\mathbb{E}[J(Y')].$$

Therefore, writing $\mu \coloneqq \mathbb{E}[J(Y')]$, Chebyshev's inequality gives

$$\begin{aligned}\mathbb{P}(Y' \notin \mathscr{Y}_2) = \mathbb{P}\big(J(Y') > (1+\varepsilon)\mu + C'\big) &\leqslant \frac{\operatorname{Var}(J(Y'))}{(\varepsilon\mu + C')^2} \leqslant \frac{K\mu}{(\varepsilon\mu + C')^2} \\ &\leqslant \max_{x\geqslant 0} \frac{Kx}{(\varepsilon x + C')^2} = \max_{y>0} \frac{K}{(\varepsilon y + C'/y)^2} = \frac{K}{4C'\varepsilon} = \frac{\varepsilon}{2},\end{aligned}$$

as desired. □

7. Applications. In this section, we derive Theorems 2, 3, and 4 from our main technical result, Theorem 5, and the general lower-bound estimate for lower tail probabilities, Theorem 6. In order to do so, we just need to represent the number of copies of a given (hyper)graph H in subgraphs of the complete (hyper)graph (resp. the number of arithmetic progressions of a given length in subsets of positive integers) as the number of edges in some auxiliary hypergraph $\mathscr{H}$ and verify that $\mathscr{H}$ satisfies the assumptions of Theorem 5 when $p \gg n^{-1/m_r(H)}$. This is pretty straightforward, but we present the full details for the reader's convenience.

The following easy lemma, which states that $\Phi_p^{\mathscr{H}}$, defined in (6) above the statement of Theorem 5, satisfies $\Phi_p^{\mathscr{H}}(\eta) = \Theta\big(v(\mathscr{H})p\big)$ for every uniform hypergraph $\mathscr{H}$ whose maximum degree is comparable to its average degree, will be used to absorb the additive constant C from the assertions of Theorems 5 and 6 into the main term.

LEMMA 21. *Suppose that $\mathscr{H}$ is an r-uniform hypergraph that satisfies*

$$\Delta_1(\mathscr{H}) \leqslant K \cdot \frac{e(\mathscr{H})}{v(\mathscr{H})}$$

for some K. Then, for all positive reals p and ε,

$$\Phi_p^{\mathscr{H}}(1-\varepsilon) \geqslant \frac{\varepsilon^2}{2K^2} \cdot |V|p.$$

PROOF. Let $q\colon V \to [0,1]$ be a function achieving the minimum in the definition of $\Phi_p^{\mathscr{H}}$. As in the proof of Theorem 6, we may assume that $q_v \leqslant p$ for every $v \in V$. From this it is easy to conclude that

$$p^{|A|} - \prod_{v\in A} q_v \leqslant \sum_{v \in A}(p-q_v)p^{|A|-1} \tag{55}$$

for every $A \subseteq V$. We may thus conclude that

$$\begin{aligned}\varepsilon p^r e(\mathscr{H}) &\overset{(\star)}{\leqslant} p^r e(\mathscr{H}) - \mathbb{E}[e(\mathscr{H}[R^{(q)}])] = \sum_{A\in\mathscr{H}} d_A \cdot \left(p^{|A|} - \prod_{v\in A} q_v\right) \\ &\overset{(55)}{\leqslant} \sum_{A\in\mathscr{H}} d_A \cdot \sum_{v\in A}(p-q_v)p^{r-1} = \sum_{v\in V}(p-q_v)p^{r-1}\cdot \deg_{\mathscr{H}} v \\ &\leqslant \Delta_1(\mathscr{H}) \cdot \sum_{v\in V}(p-q_v)p^{r-1} = p^{r-1}\Delta_1(\mathscr{H}) \cdot \left(p|V| - \sum_{v\in V} q_v\right),\end{aligned}$$

where $(\star)$ follows because q is the minimiser of $\Phi(1-\varepsilon)$. Consequently,

$$\bar{q} \coloneqq \frac{1}{|V|}\sum_{v\in V} q_v \leqslant p \cdot \left(1 - \frac{\varepsilon e(\mathscr{H})}{|V| \cdot \Delta_1(\mathscr{H})}\right) \leqslant p \cdot \left(1 - \frac{\varepsilon}{K}\right). \tag{56}$$

Since the function i_p is convex and $i_p(q) \geqslant \frac{(q-p)^2}{2p}$ when $q \leqslant p$, see (52), we may conclude that

$$\Phi_p^{\mathscr{H}}(1-\varepsilon) = \sum_{v\in V} i_p(q_v) \geqslant |V| \cdot i_p(\bar{q}) \geqslant |V| \cdot \frac{(\bar{q}-p)^2}{2p} \overset{(56)}{\geqslant} \frac{\varepsilon^2}{2K^2} \cdot |V|p,$$

as claimed. □

PROOF OF THEOREMS 2 AND 3. Theorem 2 is merely the special case $k = 2$ in Theorem 3, so we focus on Theorem 3. Suppose that H is a nonempty k-uniform hypergraph and let $\mathscr{H}$ be the e_H-uniform hypergraph with vertex set $V := \binom{[\![n]\!]}{k}$ whose hyperedges are the edge sets of all $\frac{v_H!}{|\mathrm{Aut}(H)|} \cdot \binom{n}{v_H}$ copies of H in the complete e_H-uniform hypergraph on $[\![n]\!]$ (we take $d_A = 1$ for all A). By symmetry,

$$\Delta_1(\mathscr{H}) = \frac{e_H \cdot e(\mathscr{H})}{v(\mathscr{H})}.$$

Suppose now that $B \subseteq V$ has at least two elements and nonzero degree in $\mathscr{H}$. Then B must be the edge set of some copy of a subhypergraph $F \subseteq H$, with $e_F = |B| \geqslant 2$ and $v_F = |\bigcup B|$, in the complete k-uniform hypergraph on $[\![n]\!]$. Since $m_k(H) \geqslant \frac{e_F - 1}{v_F - k}$ (recall the definition of m_k given in (4)), we have

$$\begin{aligned} \deg_{\mathscr{H}} B &\leqslant v_H^{v_F} \cdot n^{v_H - v_F} = v_H^{v_F} \cdot n^{-(v_F - k)} \cdot n^{v_H - k} \\ &\leqslant v_H^{v_F} \cdot n^{-\frac{e_F - 1}{m_k(H)}} \cdot n^{v_H - k} = v_H^{v_F} \cdot \left(n^{\frac{-1}{m_k(H)}}\right)^{|B|-1} \cdot n^{v_H - k}. \end{aligned}$$

Since B was arbitrary, we may conclude that

$$\Delta_s(\mathscr{H}) \leqslant v_H^{v_H} \cdot \left(n^{-\frac{1}{m_k(H)}}\right)^{s-1} \cdot n^{v_H - k} \tag{57}$$

for every $s \geqslant 2$.

Let λ be the constant given by Theorem 5 invoked with the p_0 from Theorem 3, $K = e_H$ and $\varepsilon_{\text{Thm 5}} = \varepsilon/2$ and let C be the larger of the constants given by Theorems 5 and 6, also with $\varepsilon_{\text{Thm 6}} = \varepsilon/2$. Lastly, let $L = L(\varepsilon, \lambda, C, H)$ be a sufficiently large constant and suppose that $Ln^{-1/m_k(H)} \leqslant p \leqslant p_0$.

By choosing L large, we guarantee that n is large as well and, consequently,

$$\frac{e(\mathscr{H})}{v(\mathscr{H})} \geqslant \frac{\binom{n}{v_H}}{\binom{n}{k}} \geqslant \frac{n^{v_H - k}}{v_H^{v_H}}.$$

Together with (57), this estimate implies that, for every $s \geqslant 2$,

$$\Delta_s(\mathscr{H}) \leqslant v_H^{2v_H} \cdot \left(\frac{p}{L}\right)^{s-1} \cdot \frac{e(\mathscr{H})}{v(\mathscr{H})} \leqslant (\lambda p)^{s-1} \cdot \frac{e(\mathscr{H})}{v(\mathscr{H})},$$

where in the second inequality we used that L is sufficiently large. By Theorems 5 and 6, for every $\eta \in [0, 1]$,

$$(1 - \varepsilon/2) \cdot \Phi_{n,p}^H(\eta + \varepsilon/2) - C \leqslant -\log \mathbb{P}\big(X \leqslant \eta \mathbb{E}[X]\big) \leqslant (1 + \varepsilon/2) \cdot \Phi_{n,p}^H\big((1 - \varepsilon/2)\eta\big) + C.$$

Finally, we show that we may absorb the additive constant C on both sides of the above inequality. To this end, we first invoke Lemma 21 to get the following inequality:

$$\Phi_{n,p}^H(1 - \varepsilon/2) \geqslant \frac{\varepsilon^2}{8e_H^2} \cdot \binom{n}{k} p \geqslant \frac{L\varepsilon^2}{16e_H^2 k!} \geqslant \frac{2C}{\varepsilon}, \tag{58}$$

where we used the assumptions that $p \geqslant Ln^{-1/m_k(H)} \geqslant Ln^{-k}$ and that L is sufficiently large. To derive the claimed the upper bound on $-\log \mathbb{P}(X \leqslant \eta \mathbb{E}[X])$, note that, since $\eta \leqslant 1$ and the function $\eta \mapsto \Phi_{n,p}(\eta)$ is decreasing, we have

$$C \overset{(58)}{\leqslant} (\varepsilon/2) \cdot \Phi_{n,p}^H\big((1 - \varepsilon/2)\eta\big) \qquad \text{and} \qquad \Phi_{n,p}^H\big((1 - \varepsilon/2)\eta\big) \leqslant \Phi_{n,p}^H\big((1 - \varepsilon)\eta\big).$$

To derive the claimed lower bound, we may assume that $\eta + \varepsilon \leqslant 1$, since otherwise $\Phi^H_{n,p}(\eta + \varepsilon) = 0$. Therefore,

$$C \overset{(58)}{\leqslant} (\varepsilon/2) \cdot \Phi^H_{n,p}(\eta + \varepsilon/2) \quad \text{and} \quad \Phi^H_{n,p}(\eta + \varepsilon/2) \leqslant \Phi^H_{n,p}(\eta + \varepsilon).$$

This completes the proof of Theorems 2 and 3. □

PROOF OF THEOREM 4. Let k be a positive integer and let $\mathscr{H}$ be the k-uniform hypergraph with vertex set $V := [\![n]\!]$ whose hyperedges are the k-term arithmetic progressions in $[\![n]\!]$, that is,

$$\mathscr{H} := \big\{\{x, x+d, \dots, x+(k-1)d\} : x, d \in [\![n]\!], x + (k-1)d \leqslant n\big\}.$$

Since every number in $[\![n]\!]$ belongs to at most kn many k-term arithmetic progressions and every pair of numbers belongs to at most $\binom{k}{2}$ such progressions, we have

$$\Delta_1(\mathscr{H}) = kn \qquad \text{and} \qquad \Delta_k(\mathscr{H}) \leqslant \cdots \leqslant \Delta_2(\mathscr{H}) \leqslant \binom{k}{2}.$$

Moreover, since $[\![n]\!]$ contains at least $c_k n^2$ many k-term progressions, for some constant $c_k > 0$, provided that $n \geqslant k$, we conclude that $e(\mathscr{H}) \geqslant c_k n^2$ and hence

$$\Delta_s(\mathscr{H}) \leqslant K \cdot \left(n^{-\frac{1}{k-1}}\right)^{s-1} \cdot \frac{e(\mathscr{H})}{v(\mathscr{H})} \qquad \forall s \in \{1, \dots, k\}$$

for some constant K that depends only on k. Therefore, when $Ln^{-1/(k-1)} \leqslant p \leqslant p_0$ for a sufficiently large constant L, we may apply Theorems 5 and 6 to derive (with a little help from Lemma 21) the claimed estimate on $-\log \mathbb{P}(X \leqslant \eta\mathbb{E}[X])$ for every $\eta \in [0,1]$, as in the previous proof. We leave the details to the reader. □

Acknowledgments. We thank the two anonymous referees for their extremely careful reading of the paper an many helpful comments and suggestions. We also thank Vishesh Jain for bringing [21, 29, 31] to our attention.

Funding. The first author was supported in part by the Jesselson Foundation and by Paul and Tina Gardner. The second author was supported in part by the Israel Science Foundation grant 1145/18.

REFERENCES

[1] AUGERI, F. (2020). Nonlinear large deviation bounds with applications to Wigner matrices and sparse Erdős-Rényi graphs. *Ann. Probab.* **48** 2404–2448.

[2] AVEZ, A. (1976). Harmonic functions on groups. In *Differential geometry and relativity* 27–32. Mathematical Phys. and Appl. Math., Vol. 3.

[3] BALOGH, J., MORRIS, R. and SAMOTIJ, W. (2015). Independent sets in hypergraphs. *J. Amer. Math. Soc.* **28** 669–709.

[4] BALOGH, J. and SAMOTIJ, W. (2020). An efficient container lemma. *Discrete Anal.* Paper No. 17, 56.

[5] BASAK, A. and BASU, R. Upper tail large deviations of regular subgraph counts in Erdős–Rényi graphs in the full localized regime. to appear in Comm. Pure Appl. Math.

[6] BHATTACHARYA, B. B., GANGULY, S., LUBETZKY, E. and ZHAO, Y. (2017). Upper tails and independence polynomials in random graphs. *Adv. Math.* **319** 313–347.

[7] CHATTERJEE, S. (2012). The missing log in large deviations for triangle counts. *Random Structures Algorithms* **40** 437–451.

[8] CHATTERJEE, S. and DEMBO, A. (2016). Nonlinear large deviations. *Adv. Math.* **299** 396–450.

[9] CHATTERJEE, S. and VARADHAN, S. R. S. (2011). The large deviation principle for the Erdős-Rényi random graph. *European J. Combin.* **32** 1000–1017.

[10] CONLON, D. and GOWERS, W. T. (2016). Combinatorial theorems in sparse random sets. *Ann. of Math. (2)* **184** 367–454.

[11] COOK, N. and DEMBO, A. (2020). Large deviations of subgraph counts for sparse Erdős-Rényi graphs. *Adv. Math.* **373** 107289, 53.

[12] COOK, N., DEMBO, A. and PHAM, H. T. Regularity method and large deviation principles for the Erdős–Rényi hypergraph. arXiv:2102.09100 [math.PR].

[13] COVER, T. M. and THOMAS, J. A. (2006). *Elements of information theory*, Second ed. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

[14] CSISZÁR, I. (1984). Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.* **12** 768–793.

[15] CSISZÁR, I. and KÖRNER, J. (2011). *Information theory*, Second ed. Cambridge University Press, Cambridge Coding theorems for discrete memoryless systems.

[16] DEMARCO, B. and KAHN, J. (2012). Upper tails for triangles. *Random Structures Algorithms* **40** 452–459.

[17] ELDAN, R. (2018). Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations. *Geom. Funct. Anal.* **28** 1548–1596.

[18] ERDŐS, P., KLEITMAN, D. J. and ROTHSCHILD, B. L. (1976). Asymptotic enumeration of K_n-free graphs. In *Colloquio Internazionale sulle Teorie Combinatorie (Rome, 1973), Tomo II* 19–27. Atti dei Convegni Lincei, No. 17.

[19] FRIEZE, A. and KANNAN, R. (1999). Quick approximation to matrices and applications. *Combinatorica* **19** 175–220.

[20] HAREL, M., MOUSSET, F. and SAMOTIJ, W. Upper tails via high moments and entropic stability. to appear in Duke Math. J.

[21] JAIN, V., KOEHLER, F. and RISTESKI, A. (2019). Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: a unified perspective. In *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* 1226–1236. ACM, New York.

[22] JANSON, S. (1990). Poisson approximation for large deviations. *Random Structures Algorithms* **1** 221–229.

[23] JANSON, S., ŁUCZAK, T. and RUCIŃSKI, A. (1990). An exponential bound for the probability of nonexistence of a specified subgraph in a random graph. In *Random graphs '87 (Poznań, 1987)* 73–87. Wiley, Chichester.

[24] KOHAYAKAWA, Y., ŁUCZAK, T. and RÖDL, V. (1997). On K^4-free subgraphs of random graphs. *Combinatorica* **17** 173–213.

[25] KOZMA, G., MEYEROVITCH, T., PELED, R. and SAMOTIJ, W. What does a typical metric space look like? to appear in Ann. Inst. Henri Poincaré Probab. Stat.

[26] LUBETZKY, E. and ZHAO, Y. (2015). On replica symmetry of large deviations in random graphs. *Random Structures Algorithms* **47** 109–146.

[27] LUBETZKY, E. and ZHAO, Y. (2017). On the variational problem for upper tails in sparse random graphs. *Random Structures Algorithms* **50** 420–436.

[28] ŁUCZAK, T. (2000). On triangle-free random graphs. *Random Structures Algorithms* **16** 260–276.

[29] MANURANGSI, P. and RAGHAVENDRA, P. (2017). A birthday repetition theorem and complexity of approximating dense CSPs. In *44th International Colloquium on Automata, Languages, and Programming. LIPIcs. Leibniz Int. Proc. Inform.* **80** Art. No. 78, 15. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern.

[30] MONTANARI, A. (2008). Estimating random variables from random sparse observations. *European Transactions on Telecommunications* **19** 385–403.

[31] RAGHAVENDRA, P. and TAN, N. (2012). Approximating CSPs with global cardinality constraints using SDP hierarchies. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms* 373–384. ACM, New York.

[32] SAXTON, D. and THOMASON, A. (2015). Hypergraph containers. *Invent. Math.* **201** 925–992.

[33] SCHACHT, M. (2016). Extremal results for random discrete structures. *Ann. of Math. (2)* **184** 333–365.

[34] ZHAO, Y. (2017). On the lower tail variational problem for random graphs. *Combin. Probab. Comput.* **26** 301–320.