THE PHASE TRANSITION OF DISCREPANCY IN RANDOM HYPERGRAPHS

CALUM MACRURY, TOMÁŠ MASAŘÍK, LEILANI PAI, AND XAVIER PÉREZ-GIMÉNEZ

Abstract. Motivated by the Beck-Fiala conjecture, we study the discrepancy problem in two related models of random hypergraphs on n vertices and m edges. In the first (*edge-independent*) model, a random hypergraph H_1 is constructed by fixing a parameter p and allowing each of the n vertices to join each of the m edges independently with probability p. In the parameter range in which $pn \to \infty$ and $pm \to \infty$, we show that with high probability (*w.h.p.*) H_1 has discrepancy at least $\Omega(2^{-n/m}\sqrt{pn})$ when $m = O(n)$, and at least $\Omega(\sqrt{pn \log \gamma})$ when $m \gg n$, where $\gamma = \min\{m/n, pn\}$. In the second (*edge-dependent*) model, d is fixed and each vertex of H_2 independently joins exactly d edges uniformly at random. We obtain analogous results for this model by generalizing the techniques used for the edge-independent model with $p = d/m$. Namely, for $d \to \infty$ and $dn/m \to \infty$, we prove that *w.h.p.* H_2 has discrepancy at least $\Omega(2^{-n/m}\sqrt{dn/m})$ when $m = O(n)$, and at least $\Omega(\sqrt{(dn/m)\log \gamma})$ when $m \gg n$, where $\gamma = \min\{m/n, dn/m\}$. Furthermore, we obtain nearly matching asymptotic upper bounds on the discrepancy in both models (when $p = d/m$), in the dense regime of $m \gg n$. Specifically, we apply the partial colouring lemma of Lovett and Meka to show that *w.h.p.* H_1 and H_2 each have discrepancy $O(\sqrt{dn/m}\log(m/n))$, provided $d \to \infty$, $dn/m \to \infty$ and $m \gg n$. This result is algorithmic, and together with the work of Bansal and Meka characterizes how the discrepancy of each random hypergraph model transitions from $\Theta(\sqrt{d})$ to $o(\sqrt{d})$ as m varies from $m = \Theta(n)$ to $m \gg n$.

1. Introduction

A **hypergraph**[1] $H = (V, E)$ consists of a set $V = \{v_1, \dots, v_n\}$ of n vertices together with a multiset $E = \{e_1, \dots, e_m\}$ of m edges, where each edge e_i is a subset of V. We denote the size of an edge e as $|e|$. Note that H is allowed to have duplicate edges (i.e. we may have $e_i = e_{i'}$ for some $i \neq i'$). We can bijectively represent H by an $m \times n$ $\{0,1\}$-matrix $\boldsymbol{A} = \boldsymbol{A}(H) = (A_{i,j})_{1 \leqslant i \leqslant m,\, 1 \leqslant j \leqslant n}$, where $A_{i,j} = 1$ if $v_j \in e_i$ and $A_{i,j} = 0$ if $v_j \notin e_i$. We call $\boldsymbol{A}$ the **incidence matrix** of H. In particular, each pair of duplicate edges in H corresponds to a pair of identical rows in $\boldsymbol{A}$. Moreover, we define the **degree** of a vertex v_j of H to be the number of edges containing that vertex, i.e. the number of 1's in the jth column of $\boldsymbol{A}$. A classical problem in discrepancy theory is to find a 2-colouring of the vertices of H so that every edge is as "balanced" as possible. To make this more precise, we define a **colouring** of H to be a function $\psi : V \to \{-1, 1\}$. This can be extended to a map $\psi : E \to \mathbb{Z}$, by defining $\psi(e) := \sum_{v \in e} \psi(v)$ for each $e \in E$. We call $|\psi(e)|$ the **discrepancy** of edge e, and note that it measures how unbalanced the colouring ψ is on that edge. Further, the **discrepancy** of colouring ψ, denoted $\operatorname{disc}(\psi)$, is the discrepancy of the least balanced edge of H. That is,

$$\operatorname{disc}(\psi) := \max_{e \in E} |\psi(e)|$$

Finally, we define the **discrepancy** of hypergraph H as

$$\operatorname{disc}(H) := \min_{\psi} \operatorname{disc}(\psi),$$

where the minimization is over all colourings of V.

[1]The definition of a hypergraph is equivalent to that of a set system, though we exclusively use the former terminology in this work.

This and other related notions of combinatorial discrepancy have been studied from various angles and in different contexts. (For a more detailed introduction to the subject, we refer to books [18], [7] and [8].) One of the central problems in discrepancy theory in the above setting is to bound the discrepancy of a hypergraph H in terms of its maximum degree d. In [5], it was proven by Beck and Fiala that the discrepancy of H is no larger than $2d-1$. Moreover, they conjectured that the correct upper bound is of the order $O(d^{1/2})$. There has been much work in trying to improve on the original bound of [5]. Most recently, it was proven by Bukh [6] that $\operatorname{disc}(H) \leqslant 2d - \lg^*(d)$, where $\lg^*$ is the binary iterated logarithm. This of course yields no asymptotic improvement in terms of d, but to this date is the strongest upper bound known which solely depends on d. If the upper bound is allowed dependence on the multiple parameters of the hypergraph, then there are results yielding improvements for hypergraphs in the correct range of parameters. For instance, Banaszczyk [3] showed that if $n := |V|$, then $\operatorname{disc}(H) = O(\sqrt{d \log n})$—a bound which was later made algorithmic by Bansal and Meka [4]. Recently, Potukuchi [20] proved that, for d-regular H, $\operatorname{disc}(H) = O(\sqrt{d} + \lambda)$, where $\lambda := \max_{v \perp \mathbf{1}, \|v\|_2 = 1} \|\boldsymbol{A}\, v\|_2$ and $\boldsymbol{A}$ is the incidence matrix of H.

In order to find upper bounds which depend solely on the maximum degree, restricted classes of hypergraphs have instead been studied. For example, if $H = (V, E)$ is assumed to be both d-regular and d-uniform (that is, the incidence matrix $\boldsymbol{A}$ has exactly d 1's in each column and row), then a folklore application of the Lovász local lemma can be used to show that there exists a colouring which achieves discrepancy $O(\sqrt{d \log d})$ (see [10, 19] for details).

Another approach is to restrict one's attention to hypergraphs which are generated randomly. In this work, we focus on two specific random hypergraph models. Both of these models are defined as distributions over the set of hypergraphs with $n \geqslant 1$ vertices and $m \geqslant 1$ edges.

1.1. **The Edge-Independent Model.** In [15], Hoberg and Rothvoss introduced a random hypergraph model, denoted $\mathbb{H}(n, m, p)$, in which a probability parameter $0 \leqslant p \leqslant 1$ is given (in addition to n and m). Their model, which we refer to as the **edge-independent model**, is a distribution on hypergraphs which we describe through the following randomized procedure:

- Fix the vertex set $V = \{v_1, \ldots, v_n\}$.
- For each $1 \leqslant i \leqslant m$, construct edge e_i by placing each $v \in V$ in e_i independently with probability p.

We denote $E = \{e_1, \ldots, e_m\}$ and define $\mathbb{H}(n, m, p)$ to be the distribution of the random hypergraph $H = (V, E)$. In other words, the entries of the incidence matrix $\boldsymbol{A}$ of H are independent Bernoulli random variables of parameter p. We write $H \sim \mathbb{H}(n, m, p)$ to indicate that H is drawn from $\mathbb{H}(n, m, p)$.

If $m = m(n)$ and $p = p(n)$ are functions which depend on n, then we say that $\mathbb{H}(n, m, p)$ satisfies a property $Q = Q(n)$ *w.h.p.*, provided that $\mathbb{P}[H(n) \text{ satisfies } Q(n)] \to 1$ as $n \to \infty$, where $H = H(n)$ is drawn from $\mathbb{H}(n, m, p)$. Often, we abuse terminology slightly and say that the random hypergraph H satisfies Q *w.h.p.*

Hoberg and Rothvoss showed that, if $n \geqslant C_1 m^2 \log m$ and $C_1 \log n/m \leqslant p \leqslant 1/2$ for some sufficiently large constant $C_1 > 0$, then $\operatorname{disc}(H) \leqslant 1$ *w.h.p.* for $H \sim \mathbb{H}(n, m, p)$. A natural question left open by their work is whether or not H continues to have constant discrepancy when n transitions from $C_1 m^2 \log m$ to $\Theta(m \log m)$. Potukuchi [19] provided a positive answer to this question for the special case when $p = 1/2$ by showing that if $n \geqslant C_2 m \log m$ for $C_2 = (2 \log 2)^{-1}$, then *w.h.p.* $\operatorname{disc}(H) \leqslant 1$. Very recently, Altschuler and Niles-Weed [2] used Stein's method [9] in conjunction with the second moment method to substantially generalize this result to hold when $p = p(n)$ depends on n. This includes the challenging case when $p(n) \to 0$.

1.2. **The Edge-Dependent Model.** A related model was introduced by Ezra and Lovett in [10]. As before, we fix $n \geqslant 1$ and $m \geqslant 1$, yet we now also consider a parameter $d \geqslant 1$ which satisfies

$d \leqslant m$. The **edge-dependent model**, denoted $\mathcal{H}(n, m, d)$, is again a distribution on hypergraphs, though we describe it through a different randomized procedure:

- Fix vertex set $V = \{v_1, \dots, v_n\}$.
- For each vertex $v \in V$, independently and *u.a.r.* (uniformly at random) draw $I_v \subseteq [m]$ with $|I_v| = d$.
- For each $1 \leqslant i \leqslant m$, construct edge e_i by defining $e_i := \{u \in V : i \in I_u\}$.

By setting $E := \{e_1, \dots, e_m\}$, we define $\mathcal{H}(n, m, d)$ to be the distribution of the random hypergraph $H = (V, E)$. In other words, the incidence matrix $\boldsymbol{A}$ of $H \sim \mathcal{H}(n, m, d)$ is a random $m \times n$ matrix where each column has d ones and $m - d$ zeros. Note that the columns of $\boldsymbol{A}$ are independent, but the rows are not. We define what it means for $\mathcal{H}(n, m, d)$ to satisfy a property *w.h.p.* in the same way as in the edge-independent model.

Ezra and Lovett showed that, if $m \geqslant n \geqslant d \to \infty$, then *w.h.p.* $\operatorname{disc}(H) = O(\sqrt{d \log d})$. Bansal and Meka [4] later showed that the factor of $\sqrt{\log d}$ is redundant, thereby matching the bound claimed in the Beck-Fiala conjecture. Specifically, they showed that, for the entire range of n and m, $\operatorname{disc}(H) = O(\sqrt{d})$ *w.h.p.*, provided $d = \Omega((\log \log m)^2)$. This result can be easily modified to also apply to the edge-independent model, provided the analogous condition $pm = \Omega((\log \log m)^2)$ holds.

In [12], Franks and Saks considered the more general problem of **vector balancing**. Their main result concerns a collection of random matrix models in which the columns are generated independently. In particular, their results apply to the sparse regime ($m \ll n$) of both the random hypergraph models we have discussed. Specifically, they show that if $n = \Omega(m^3 \log^3 m)$, then *w.h.p.* $\operatorname{disc}(H) \leqslant 2$, provided H is drawn from $\mathbb{H}(n, m, p)$ or $\mathcal{H}(n, m, d)$ for $p = d/m$. Finally, in a very recent work, Turner et al. [**TurnerMR20colt**, 22] considered the problem of vector balancing when the entries of the random matrix $\boldsymbol{A}$ are each distributed as standard Gaussian random variables which are independent and identically distributed (*i.i.d.*). Amongst other results, they showed that the discrepancy of $\boldsymbol{A}$ is $\Theta(2^{-n/m}\sqrt{n})$ *w.h.p.*, provided $m \ll n$.

1.3. **An Overview of Our Results.** All results proven in this paper are asymptotic with respect to the parameter n, the number of vertices of the model. Thus, we hereby assume that $m = m(n)$, $p = p(n)$ and $d = d(n)$ are functions which depend on n, with $p = d/m$.

While previous results have successfully matched (or improved upon) the conjectured Beck-Fiala bound of $\sqrt{d}$ in the random hypergraph setting, they either apply to a restricted parameter range [10, 12, 15, 19], or do not provide asymptotically tight results for the full parameter range [4, 10]. In particular, when $n/\log n \ll m \ll n$ or $m \gg n$, the correct order of the discrepancy is unknown in either model. In this paper, we obtain (almost) matching upper and lower bounds in the dense regime of $m \gg n$. Moreover, our upper bounds are algorithmic. In the sparse regime of $m \ll n$, we provide the first lower bounds which apply to the full parameter range under the mild assumption that both the average edge size $dn/m = pn$ and degree $d = pm$ tend to infinity with n. Proving the existence of a colouring whose discrepancy matches our lower bound in the regime $n/\log n \ll m \ll n$ remains open. This problem is particularly challenging from an algorithmic perspective, as the partial colouring lemma [17] does not appear to be useful in this range of parameters, and this is the main tool used in the literature.

We now formally state our main results:

Theorem 1.1. *Suppose that H is generated from $\mathbb{H}(n, m, p)$ with $pn \to \infty$, $pm \to \infty$ and p bounded away from 1. If $m = O(n)$, then w.h.p.*

$$\operatorname{disc}(H) = \Omega\left(\max\{2^{-n/m}\sqrt{pn}, 1\}\right),$$

Moreover, if $m \gg n$, then w.h.p.

$$\operatorname{disc}(H) = \Omega\left(\sqrt{pn \log \gamma}\right),$$

where $\gamma := \min\{m/n, pn\}$.

Remark 1. This bound complements the upper bound of 1 in [2] for $m \leqslant Cn/\log n$ where $C = 2\log 2$, but also implies that if $\varepsilon > 0$ is a constant, then H has non-constant discrepancy for $m \geqslant (C+\varepsilon)n/\log(np)$. Thus, H exhibits a sharp phase transition at $Cn/\log(np)$ for constant p.

We obtain analogous results regarding the edge-dependent model $\mathcal{H}(n, m, d)$:

Theorem 1.2. *Suppose that H is generated from $\mathcal{H}(n, m, d)$ with $dn/m \to \infty$, $d \to \infty$, and d/m bounded away from 1. If $m = O(n)$, then w.h.p.*

$$\operatorname{disc}(H) = \Omega\left(\max\{2^{-n/m}\sqrt{\frac{dn}{m}}, 1\}\right),$$

Moreover, if $m \gg n$, then w.h.p.

$$\operatorname{disc}(H) = \Omega\left(\sqrt{\frac{dn}{m}\log \gamma}\right),$$

where $\gamma := \min\{m/n, dn/m\}$.

Remark 2. The techniques used in [2] do not seem to apply to the edge dependent model. Thus, if $m = O(n^{1/3}/\log n)$, then [12] implies $\operatorname{disc}(H) \leqslant 2$, however when $n^{1/3}/\log n \ll m \ll n$, $O(\sqrt{d})$ remains the best known upper bound for $\operatorname{disc}(H)$ [4].

Finally, we prove an upper bound in the dense regime of $m \gg n$ which holds for both models:

Theorem 1.3. *Assume that H is drawn from $\mathcal{H}(n, m, d)$ or $\mathbb{H}(n, m, p)$ with $p = d/m$, and pick any $\beta = \beta(n) \geqslant 1$ satisfying*[2]

$$\beta\frac{dn}{m} \geqslant \log(m/n)\left(\log\left(\frac{dn}{m}\right) + 2\right)^5. \tag{1}$$

If $m \gg n$, and $dn/m \to \infty$, then w.h.p.

$$\operatorname{disc}(H) = O\left(\sqrt{\frac{dn}{m}\log\left(\frac{m}{n}\right)\beta}\right).$$

Moreover, whenever this holds, we can find a colouring ψ of H with such a discrepancy in expected polynomial time.

Remark 3. Observe that the smaller we are able to take β, the better upper bound we get. In particular, if $\beta := \log(m/n)$, then (1) and $\beta(n) \geqslant 1$ are satisfied (for large enough n). Therefore, at worst the upper bound is $O\left(\sqrt{\frac{dn}{m}\log\left(\frac{m}{n}\right)}\right)$, which is significantly smaller than the upper bound of $O(\sqrt{d})$ of [4], as $m \gg n$.

Theorem 1.3 provides asymptotically matching bounds for the lower bounds of Theorems 1.1 and 1.2 in a broad range of the dense regime $m \gg n$. For instance, this happens when $d = pm \geqslant (m/n)^{1+\varepsilon}$, where $\varepsilon > 0$ is a constant, since in that case $\beta := 1$ clearly satisfies (1), and $\log(dn/m) = \Omega(\log(m/n))$, so $\gamma = \Theta(\log(m/n))$.

[2]Let us remark at this place that by log we always mean the natural logarithm. In the proof of this theorem it is convenient to use lg to denote logarithm of base 2.

Corollary 1.4. *Assume that H is drawn from $\mathcal{H}(n,m,d)$ or $\mathbb{H}(n,m,p)$, where $p = d/m$. If $m \gg n$ and $d = pm \geqslant (m/n)^{1+\varepsilon}$ for constant $\varepsilon > 0$, then w.h.p.*

$$\operatorname{disc}(H) = \Theta\left(\sqrt{\frac{dn}{m}\log\left(\frac{m}{n}\right)}\right).$$

Remark 4. This shows that in the dense regime, the main parameter of interest describing the discrepancy of each random hypergraph model is the average edge size, dn/m, opposed to d, the average/maximum degree (depending on the model).

2. Preliminaries

In this section, we first provide some central-limit-type results for sums of independent random variables, which will be used in Section 3 to obtain lower bounds on the discrepancy. In the sequel, we write $N(0,1)$ to denote a generic random variable distributed as a standard Gaussian with cumulative distribution function

$$\Phi(x) := \mathbb{P}[N(0,1) \leqslant x] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \exp(-t^2/2)\,dt \qquad \text{for each } x \in \mathbb{R}.$$

The Berry-Esseen Theorem (see section XVI.5 in [11]), which we state below for convenience, yields a quantitative form of the Central Limit Theorem for sums of independent random variables with finite third moment.

Theorem 2.1 (Berry-Esseen)**.** *The following holds for some universal constant $c_{uni} > 0$. Let $Y_1,\dots,Y_n$ be independent random variables with $\mathbb{E}[Y_i] = 0$, $\mathbb{E}[Y_i^2] = \sigma_i^2$ and $\mathbb{E}[|Y_i|^3] = \rho_i < \infty$ for all $i \in [n]$. Consider the sum $Y = \sum_{i=1}^n Y_i$, with standard deviation $\sigma = \sqrt{\sum_{i=1}^n \sigma_i^2}$, and let $\rho = \sum_{i=1}^n \rho_i$. Assume $\sigma > 0$. Then,*

$$\sup_{x\in\mathbb{R}} |\mathbb{P}[Y/\sigma \leqslant x] - \Phi(x)| \leqslant \frac{(c_{uni}/2)\rho}{\sigma^3}.$$

Remark 5. There has been a series of works improving upon the constant c_{uni}, the latest of which by Shevtsova [21] shows that $c_{uni}/2 \leqslant 0.560$ in our setting. However, since we are only concerned with the asymptotic growth of discrepancy, the precise value of c_{uni} is not important.

Theorem 2.1 and the triangle inequality immediately yield the following corollary:

Corollary 2.2. *For any interval $I \subseteq (-\infty,\infty)$,*

$$|\mathbb{P}[Y/\sigma \in I] - \mathbb{P}[N(0,1) \in I]| \leqslant \frac{c_{uni}\rho}{\sigma^3}.$$

We will apply this result to linear combinations of independent Bernoulli's with coefficients in $\{-1,1\}$. More precisely, let $X_1,\dots,X_n$ be independent random variables with $X_i \sim \mathrm{Ber}(p_i)$ for some $\boldsymbol{p} = (p_1,\dots,p_n) \in [0,1]^n$ (where $\mathrm{Ber}(p_i)$ is a Bernoulli of parameter p_i). Given a vector $\boldsymbol{a} = (a_1,\dots,a_n) \in \{-1,1\}^n$, consider the sum $S_{\boldsymbol{a},\boldsymbol{p}} := \sum_{k=1}^n a_i X_i$, whose standard deviation we denote by σ. Under these assumptions we obtain the following bound.

Lemma 2.3. *For any bounded interval $[L,R] \subseteq (-\infty,\infty)$,*

$$\mathbb{P}[S_{\boldsymbol{a},\boldsymbol{p}} \in [L,R]] \leqslant \frac{c_{uni}}{\sigma} + \left(1 - \exp\left(-\frac{(R-L)^2}{2\pi\sigma^2}\right)\right)^{1/2}.$$

(When $\sigma = 0$, the right-hand side of the bound above is simply interpreted as $+\infty$.)

Proof. Let $\mu = \mathbb{E}[S_{\boldsymbol{a},\boldsymbol{p}}] = \sum_{i=1}^{n} a_i p_i$. Then we centre $S_{\boldsymbol{a},\boldsymbol{p}}$ by defining the random variables $Y_i = a_i(X_i - \mathbb{E}[X_i])$ and setting

$$Y := \sum_{k=1}^{n} Y_i = S_{\boldsymbol{a},\boldsymbol{p}} - \mu.$$

Observe that Y has the same standard deviation σ as $S_{\boldsymbol{a},\boldsymbol{p}}$, which we assume is non-zero (otherwise the lemma holds trivially). Moreover, $S_{\boldsymbol{a},\boldsymbol{p}} \in [L, R]$ if and only if $Y/\sigma \in [\widetilde{L}, \widetilde{R}]$, where $\widetilde{L} := (L-\mu)/\sigma$ and $\widetilde{R} := (R-\mu)/\sigma$. Further,

$$\mathbb{E}[|Y_i|^3] = (1-p_i)p_i^3 + p_i(1-p_i)^3 = p_i(1-p_i)\left(p_i^2 + (1-p_i)^2\right) \leqslant p_i(1-p_i),$$

and hence

$$\rho = \sum_{i=1}^{n} \mathbb{E}[|Y_i|^3] \leqslant \sum_{i=1}^{n} p_i(1-p_i) = \sigma^2.$$

Then, Corollary 2.2 immediately yields

$$\mathbb{P}[S_{\boldsymbol{a},\boldsymbol{p}} \in [L, R]] = \mathbb{P}[Y/\sigma \in [\widetilde{L}, \widetilde{R}]] \leqslant \frac{c_{uni}}{\sigma} + \mathbb{P}[N(0,1) \in [\widetilde{L}, \widetilde{R}]].$$

To finalize the proof, it only remains to bound $\mathbb{P}[N(0,1) \in [\widetilde{L}, \widetilde{R}]]$. In order to do so, we will use the inequality

$$\int_{-t}^{t} \exp(-x^2/2)\,dx \leqslant \sqrt{2\pi(1-\exp(-2t^2/\pi))} \qquad \text{for all } t \in \mathbb{R}, \tag{2}$$

which can be found in [23]. Furthermore, note that $\exp(-x^2/2)$ is an even function, decreasing with x^2. Combining this fact with (2), we get

$$\begin{aligned}
\mathbb{P}[N(0,1) \in [\widetilde{L}, \widetilde{R}]] &= \frac{1}{\sqrt{2\pi}} \int_{\widetilde{L}}^{\widetilde{R}} \exp(-x^2/2)\,dx \\
&\leqslant \frac{1}{\sqrt{2\pi}} \int_{-(\widetilde{R}-\widetilde{L})/2}^{(\widetilde{R}-\widetilde{L})/2} \exp(-x^2/2)\,dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-(R-L)/2\sigma}^{(R-L)/2\sigma} \exp(-x^2/2)\,dx \\
&\leqslant \sqrt{1 - \exp\left(-\frac{(R-L)^2}{2\pi\sigma^2}\right)},
\end{aligned}$$

which concludes the proof of the lemma. □

Now suppose there exist $0 < p < 1$, $0 < \zeta < 1$ and $0 \leqslant \varepsilon < 1$ such that

$$\sum_{i=1}^{n} p_i \geqslant (1-\varepsilon)pn \qquad \text{and} \qquad p_i \leqslant \zeta \quad \text{for each } i \in [n]. \tag{3}$$

Then we can restate the upper bound of Lemma 2.3 in the following convenient way, which we use as our key tool in proving Theorems 1.1 and 1.2.

Lemma 2.4. *Suppose $p_1, \ldots, p_n$ satisfy (3) for some $0 \leqslant \varepsilon < 1$, $0 < p < 1$ and $0 < \zeta < 1$. Then, for any bounded interval $[L, R] \subseteq (-\infty, \infty)$,*

$$\mathbb{P}[S_{\boldsymbol{a},\boldsymbol{p}} \in [L, R]] \leqslant \frac{c_{uni}}{\sqrt{(1-\zeta)(1-\varepsilon)np}} + \left(1 - \exp\left(-\frac{(R-L)^2}{2\pi(1-\zeta)(1-\varepsilon)np}\right)\right)^{1/2} \tag{4}$$

$$\leqslant \frac{c_{uni} + |R-L|/\sqrt{2\pi}}{\sqrt{(1-\zeta)(1-\varepsilon)np}}. \tag{5}$$

Proof. First note that the variance σ^2 of $S_{\boldsymbol{a},\boldsymbol{p}}$ satisfies

$$\sigma^2 = \sum_{i=1}^{n} p_i(1-p_i) \geqslant (1-\zeta)\sum_{i=1}^{n} p_i \geqslant (1-\zeta)(1-\varepsilon)pn.$$

This bound, used with Lemma 2.3, immediately gives (4). Then (5) follows by applying the inequality $1-\exp(-x) \leqslant x$, which holds for every $x \in \mathbb{R}$. □

In Theorem 1.3, we prove an upper bound on discrepancy in the dense regime ($m \gg n$). In this parameter range, we make use of the **algorithmic partial colouring lemma**, a seminal result of Lovett and Meka [17] later made deterministic by Levy, Ramadas, and Rothvoss [16]. We defer the statement of this result to Lemma 4.1 of Section 4, as it will not be needed until then.

3. Lower Bounding Discrepancy

3.1. **The Edge-Independent Model.** We now return to the setting of hypergraph discrepancy in the context of the edge-independent model $\mathbb{H}(n,m,p)$. Throughout this section, $m = m(n)$, $p = p(n)$ and asymptotic statements are with respect to $n \to \infty$. We first observe that *w.h.p.* there are some edges containing an odd number of vertices and thus the discrepancy cannot be zero.

Proposition 3.1. *Suppose $H \sim \mathbb{H}(n,m,p)$ with $m \to \infty$, $pn \to \infty$ and p bounded away from 1 as $n \to \infty$. Then w.h.p.* $\operatorname{disc}(H) \geqslant 1$.

Proof of Proposition 3.1. By hypothesis, $p \leqslant 1-\varepsilon$ for some sufficiently small constant $\varepsilon > 0$. Let $e_1, \dots, e_m$ be the edges of H, and observe that the number of vertices contained in each edge is distributed as $\mathrm{Bin}(n,p)$. Then the probability that e_i has an even number of vertices is

$$\sum_{j \text{ even}} \binom{n}{j} p^j (1-p)^{n-j} = \frac{1}{2}(1+(1-2p)^n) = 1/2 + o(1), \tag{6}$$

where we used the fact that $|1-2p|^n \leqslant \max\{(1-2\varepsilon)^n, e^{-2pn}\} = o(1)$ as $n \to \infty$. Hence, the probability all the edges of H contain an even number of vertices is $(1/2+o(1))^m = o(1)$. Therefore, *w.h.p.* H has an edge with an odd number of vertices, and thus has discrepancy at least 1. □

Proposition 3.1 trivially implies Theorem 1.1 in the regime in which $2^{-n/m}\sqrt{pn} = O(1)$. We now use Lemma 2.4 to prove the remaining cases of Theorem 1.1 via a simple first moment argument:

Proof of Theorem 1.1. Suppose that $H = (V,E)$ is generated from $\mathbb{H}(n,m,p)$ with $pn \to \infty$, $pm \to \infty$ and p bounded away from 1, as $n \to \infty$. We define

$$\hat{f} = \hat{f}(n) = \begin{cases} 2^{-n/m}\sqrt{p(1-p)n} & \text{if } m = O(n), \\ \sqrt{p(1-p)n\log\gamma} & \text{if } m \gg n, \end{cases}$$

where $\gamma = \min\{pn, m/n\}$, and choose a sufficiently small constant $\kappa > 0$. To prove the theorem, it suffices to show that *w.h.p.* $\operatorname{disc}(H) \geqslant \max\{\kappa\hat{f}, 1\}$. Note that this is trivially true when $\hat{f} \leqslant 1/\kappa$, in view of Proposition 3.1. So we will assume that $\hat{f} > 1/\kappa$, and show that *w.h.p.* $\operatorname{disc}(H) \geqslant \kappa\hat{f}$. Let Ψ be the set of all colourings $\psi : V \to \{-1,1\}$, and let Z denote the number of colourings $\psi \in \Psi$ with discrepancy $\operatorname{disc}(\psi) \leqslant \kappa\hat{f}$. Since the random edges $e_1, \dots, e_m$ of H are *i.i.d.*,

$$\mathbb{E}[Z] = \sum_{\psi\in\Psi} \mathbb{P}[\operatorname{disc}(\psi) \leqslant \kappa\hat{f}] = \sum_{\psi\in\Psi} \mathbb{P}[|\psi(e_1)| \leqslant \kappa\hat{f}]^m. \tag{7}$$

Note that $\psi(e_1) = \sum_{i=1}^{n} \psi(v_i)\mathbf{1}_{v_i\in e_1}$, where $\mathbf{1}_{v_i\in e_1}$ denotes the indicator random variable of the event that edge e_1 contains vertex v_i, so $\psi(e_1)$ is distributed as $S_{\boldsymbol{a},\boldsymbol{p}}$ in Section 2 (with $a_i = \psi(v_i)$ and

$p_i = \mathbb{P}[v_i \in e_1]$). Hence, by applying (5) in Lemma 2.4 (with $\epsilon = 0$, $\zeta = p$ and $[L, R] = [-\kappa\hat{f}, \kappa\hat{f}]$) to each one of the 2^n terms of the last sum in (7), it follows that

$$\mathbb{E}[Z] \leqslant 2^n \left(\frac{c_{uni} + 2\kappa\hat{f}/\sqrt{2\pi}}{\sqrt{p(1-p)n}} \right)^m \leqslant 2^n \left(\frac{\kappa\hat{f}(c_{uni} + \sqrt{2/\pi})}{\sqrt{p(1-p)n}} \right)^m,$$

where we also used that $\kappa\hat{f} > 1$. Let us consider first the case that $m = O(n)$. Then, from the definition of $\hat{f}$ and assuming $\kappa < 1/(c_{uni} + \sqrt{2/\pi})$,

$$\mathbb{E}[Z] \leqslant \left(\kappa(c_{uni} + \sqrt{2/\pi})\right)^m = o(1).$$

Now suppose that $m \gg n$. In this case, we bound the factor $\mathbb{P}[|\psi(e_1)| \leqslant \kappa\hat{f}]$ on the right-hand side of (7) using the tighter inequality (4) in Lemma 2.4 instead of (5). Then, assuming that $C := 2\kappa^2/\pi < 1/2$, we get

$$\begin{aligned}
\mathbb{E}[Z] &\leqslant 2^n \left(\frac{c_{uni}}{\sqrt{p(1-p)n}} + \left(1 - \exp\left(-\frac{(2\kappa\hat{f})^2}{2\pi p(1-p)n}\right)\right)^{1/2} \right)^m \\
&= 2^n \left(\frac{c_{uni}}{\sqrt{p(1-p)n}} + \left(1 - \gamma^{-2\kappa^2/\pi}\right)^{1/2} \right)^m \\
&= \left(1 + O\left(n/m\right)\right)^m \left(O(1/\sqrt{pn}) + \left(1 - \gamma^{-C}\right)^{1/2} \right)^m \\
&= \left(1 - \frac{1}{2}\gamma^{-C}(1 + o(1))\right)^m \\
&= \exp\left(-\frac{m}{2}\gamma^{-C}(1 + o(1))\right) = o(1),
\end{aligned}$$

where we used the facts $1/\sqrt{pn} = o(\gamma^{-C})$, $n/m = o(\gamma^{-C})$ and $m/\gamma^C \to \infty$. In either case, $\mathbb{E}[Z] = o(1)$, and therefore *w.h.p.* $\operatorname{disc}(H) \geqslant \kappa\hat{f}$. □

3.2. **The Edge-Dependent Model.** In this section, we derive a lower bound on the discrepancy of a hypergraph generated from the edge-dependent model and prove Theorem 1.2. In view of our previous results for the edge-independent model, one natural approach is to compare both models $\mathcal{H}(n, m, d)$ and $\mathbb{H}(n, m, p)$ via a **coupling procedure**. For instance, let $m = n$ and $d \gg \log n$ for simplicity, and suppose that we can generate (H_1, H_2) with $H_1 \sim \mathcal{H}(n, n, d)$ and $H_2 \sim \mathbb{H}(n, n, p)$, with edge sets $E(H_1) = \{e_1^1, \ldots, e_n^1\}$ and $E(H_2) = \{e_1^2, \ldots, e_n^2\}$, in such a way that *w.h.p.* for every $i = 1, \ldots, n$ we have $|e_i^1 \Delta e_i^2| \leqslant \eta$, for some suitable $\eta = \eta(n)$. In particular, this implies that *w.h.p.* $|\operatorname{disc}(H_1) - \operatorname{disc}(H_2)| \leqslant \eta$, and thus $\operatorname{disc}(H_1) = \Omega(\sqrt{d} + \eta)$ by Theorem 1.1. Unfortunately, since the standard deviation of the size of an edge is $\Theta(\sqrt{d})$ in either model, most *naïve* attempts to build such a coupling require $\eta \gg \sqrt{d}$, which is too large for our purposes. (In fact, it is not hard to build such a coupling with any $\eta \gg \sqrt{d \log n}$.) As a result, while it is conceivable that a more delicate coupling argument works, we abandon this approach. Instead, we handle the dependencies of the edges of H_2 by applying a careful conditioning argument, while generalizing how we apply Lemma 2.4.

As in the edge-independent model, we first prove a constant lower bound on the discrepancy of $H \sim \mathcal{H}(n, m, d)$.

Proposition 3.2. *Suppose $H \sim \mathcal{H}(n, m, d)$ with $m \to \infty$, $dn/m \to \infty$ and d/m bounded away from 1 as $n \to \infty$. Then w.h.p.* $\operatorname{disc}(H) \geqslant 1$.

Remark 6. It is conceivable that in the proof of the proposition above one could obtain an upper bound on $\mathbb{P}[W = 0]$ that is exponentially small in m, in the same spirit as in the proof of

Proposition 3.1. However, this would require some additional work due to the fact that the edges of $H \sim \mathcal{H}(n,m,d)$ are *not* formed independently.

Proposition 3.2 trivially implies Theorem 1.2 in the regime in which $2^{-n/m}\sqrt{dn/m} = O(1)$. To prove the remaining cases, we will generalize the ideas we used in the proof of Theorem 1.1. However, the dependencies among the edges make the argument much more delicate.

Proof of Theorem 1.2. Suppose that $H = (V,E)$ is generated from $\mathcal{H}(n,m,d)$ with $m = m(n)$ and $d = d(n)$ satisfying $dn/m \to \infty$ and $d \to \infty$ (as $n \to \infty$) and with $p = d/m \leqslant c$ for some constant $0 < c < 1$. For short, we use $\mathcal{H}$ to denote the sample space of the distribution, i.e. the set of all possible outcomes of H. Fix a constant $0 \leqslant \varepsilon < \min\{1, 1/c - 1\}$, and define

$$\hat{f} = \hat{f}(n) := \begin{cases} 2^{-n/m}\sqrt{pn(1-\varepsilon)(1-c(1+\varepsilon))} = \Omega(2^{-n/m}\sqrt{dn/m}) & \text{if } m = O(n), \\ \sqrt{pn\log\gamma(1-\varepsilon)(1-c(1+\varepsilon))} = \Omega(\sqrt{pn\log\gamma}) & \text{if } m \gg n, \end{cases}$$

where $\gamma = \min\{pn, m/n\}$. Let $\kappa > 0$ be a sufficiently small constant. To prove Theorem 1.2, it suffices to show that *w.h.p.* $\operatorname{disc}(H) \geqslant \max\{\kappa\hat{f}, 1\}$. Note that this is trivially true when $\hat{f} \leqslant 1/\kappa$, in view of Proposition 3.2. So we will assume that $\hat{f} > 1/\kappa$, and show that *w.h.p.* $\operatorname{disc}(H) \geqslant \kappa\hat{f}$. Note that this assumption ensures that $n = O(m\log d)$, i.e. there are not *too* many more vertices than edges.

Let Z be the number of colourings $\psi : V \to \{-1,1\}$ with discrepancy $\operatorname{disc}(\psi) \leqslant \kappa\hat{f}$. We would like to prove an analogue of (7) in order to bound $\mathbb{E}Z$, but unfortunately the random edges $e_1, \dots, e_m$ of H are no longer *i.i.d.* In order to overcome this obstacle, we introduce some random variables that will play an essential role in the analysis of Z. For each $j = 1, \dots, m$ and $k = 1, \dots, n$, let $A_{j,k} := \mathbf{1}_{[v_k \in e_j]}$ be the $\{0,1\}$ value of the (j,k) entry of the incidence matrix $\boldsymbol{A}$ of H, and let $\boldsymbol{A}_j = (A_{j,1}, \dots, A_{j,n})$ denote the j-th row of $\boldsymbol{A}$. Moreover, for each $k = 1, \dots, n$, we define

$$B_{0,k} := d \qquad \text{and} \qquad B_{i,k} := d - \sum_{j=1}^{i} A_{j,k} \quad \text{for } i = 1, \dots, m. \tag{8}$$

In other words, $B_{i,k}$ counts the number of ones that appear in the k-th column of $\boldsymbol{A}$ below the i-th row (recall that each column of A has exactly d ones). Note that each $B_{i,k}$ can be expressed as a function of the first i rows of $\boldsymbol{A}$, and moreover the distribution of $A_{i+1,k}$ conditional on the outcome of $A_{1,k}, \dots, A_{i,k}$ can be described solely in terms of $B_{i,k}$. More formally, for each $i = 1, \dots, m$, let $\mathcal{F}_i$ be the sigma algebra generated by $\boldsymbol{A}_1, \dots, \boldsymbol{A}_i$, and let $\mathcal{F}_0 := \{\emptyset, \mathcal{H}\}$ denote the trivial sigma algebra. Then, for each $i = 0, \dots, m$ and $k = 1, \dots, n$, $B_{i,k}$ is measurable with respect to $\mathcal{F}_i$, and (for $i < m$)

$$P_{i,k} := \mathbb{P}[A_{i+1,k} = 1 \mid \mathcal{F}_i] = \frac{B_{i,k}}{m-i}.$$

Intuitively speaking, we would like that the above conditional probabilities remain close to p (on average) as we reveal new rows of $\boldsymbol{A}$, at least for a large number of rows. In view of that, for each $i = 0, \dots, m-1$, we consider the event Q_i that for every $0 \leqslant j \leqslant i$

$$\sum_{k=1}^{n} P_{j,k} \geqslant (1-\varepsilon)pn \qquad \text{and} \qquad P_{j,k} \leqslant (1+\varepsilon)c \quad \text{for } k = 1, \dots, n.$$

(Here $(1+\varepsilon)c < 1$ from our choice of ε.) Observe that $\mathcal{H} = Q_0 \supseteq \cdots \supseteq Q_{m-1}$ is a decreasing sequence of events, and each Q_i is $\mathcal{F}_i$-measurable by construction. Let $\alpha := \max\{n/(n+m), 1/2\}$. We need the following technical result, which we prove in

Proposition 3.3. *Under the assumptions in the proof of Theorem 1.2, event $Q_{\lfloor \alpha m \rfloor}$ holds w.h.p.*

Now let Ψ be the set of all colourings $\psi : V \to \{-1, 1\}$, and pick an arbitrary $\psi \in \Psi$. For each $i = 1, \dots, m$, let R_i^ψ denote the event that $|\psi(e_i)| \leqslant \kappa \hat{f}$, and let $R_{\leqslant i}^\psi := \bigcap_{j=1}^{i} R_j^\psi$. (By convention, $R_{\leqslant 0}^\psi = \mathcal{H}$.) Clearly, R_i^ψ and $R_{\leqslant i}^\psi$ are $\mathcal{F}_i$-measurable. Note that, conditional upon any outcome of $\boldsymbol{A}_1, \dots, \boldsymbol{A}_{i-1}$ satisfying Q_{i-1}, the random variable $\psi(e_i)$ is distributed as $S_{\boldsymbol{a},\boldsymbol{p}}$ in Section 2 (with $a_k = \psi(v_k)$ and $p_k = \mathbb{P}[v_k \in e_i]$) and it satisfies the conditions of Lemma 2.4 (with $\zeta = (1+\varepsilon)c < 1$ and $[L, R] = [-\kappa\hat{f}, \kappa\hat{f}]$). Hence, we can use that lemma to bound the conditional probability of R_i^ψ. We first consider the sparse regime of $m = O(n)$. By Lemma 2.4, assuming $\kappa < 1/\left(3(c_{uni} + \sqrt{2/\pi})\right)$ and since $\kappa\hat{f} > 1$,

$$\mathbb{P}[R_i^\psi \mid \mathcal{F}_{i-1}]\mathbf{1}_{[Q_{i-1}]} \leqslant \frac{c_{uni} + 2\kappa\hat{f}/\sqrt{2\pi}}{\sqrt{(1-\zeta)(1-\varepsilon)pn}} \leqslant \frac{\kappa\hat{f}(c_{uni} + \sqrt{2/\pi})}{\sqrt{(1-\zeta)(1-\varepsilon)pn}} < 2^{-n/m}/3. \tag{9}$$

In particular, since $R_{\leqslant i-1}^\psi \cap Q_{i-1}$ is $\mathcal{F}_{i-1}$-measurable and is contained in Q_{i-1},

$$\mathbb{P}[R_1^\psi] \leqslant 2^{-n/m}/3 \qquad \text{and} \qquad \mathbb{P}[R_i^\psi \mid R_{\leqslant i-1}^\psi \cap Q_{i-1}] \leqslant 2^{-n/m}/3 \quad \text{for } i = 2, \dots, m.$$

Thus, for each $i = 2, \dots, m$,

$$\mathbb{P}[R_{\leqslant i}^\psi \cap Q_{i-1}] = \mathbb{P}[R_i^\psi \mid R_{\leqslant i-1}^\psi \cap Q_{i-1}] \cdot \mathbb{P}[R_{\leqslant i-1}^\psi \cap Q_{i-1}] \leqslant \left(2^{-n/m}/3\right) \mathbb{P}[R_{\leqslant i-1}^\psi \cap Q_{i-2}],$$

and inductively

$$\mathbb{P}[R_{\leqslant i}^\psi \cap Q_{i-1}] \leqslant \left(2^{-n/m}/3\right)^i.$$

Let $t := \lceil \alpha m \rceil$. Next, we will bound $\operatorname{disc}(H)$ from below based on the discrepancies of the first t rows of $\boldsymbol{A}$ when Q_{t-1} holds. First note that, since $t \geqslant nm/(n+m)$,

$$\mathbb{P}[R_{\leqslant t}^\psi \cap Q_{t-1}] \leqslant \left(2^{-n/m}/3\right)^t \leqslant \left(2^{-n/m}/3\right)^{nm/(n+m)} = 2^{-n}(2/3)^{nm/(n+m)} = o(2^{-n}), \tag{10}$$

and then, by applying Markov's inequality to the random variable $Z\mathbf{1}_{Q_{t-1}}$,

$$\mathbb{P}[\operatorname{disc}(H) \leqslant \kappa\hat{f} \text{ and } Q_{t-1}] \leqslant \mathbb{E}[Z\mathbf{1}_{Q_{t-1}}] = \sum_{\psi \in \Psi} \mathbb{P}[R_{\leqslant m}^\psi \cap Q_{t-1}] \leqslant \sum_{\psi \in \Psi} \mathbb{P}[R_{\leqslant t}^\psi \cap Q_{t-1}] = o(1). \tag{11}$$

Before proceeding with the proof, we consider the dense regime of $m \gg n$. In that case, we obtain an analogue of (9) by using the tighter inequality (4) in Lemma 2.4 instead of (5). With $\zeta = (1+\varepsilon)c < 1$ and assuming that $C := 2\kappa^2/\pi < 1/2$, we get

$$\begin{aligned}\mathbb{P}[R_i^\psi \mid \mathcal{F}_{i-1}]\mathbf{1}_{[Q_{i-1}]} &\leqslant \frac{c_{uni}}{\sqrt{(1-\zeta)(1-\varepsilon)pn}} + \left(1 - \exp\left(-\frac{(2\kappa\hat{f})^2}{2\pi(1-\zeta)(1-\varepsilon)pn}\right)\right)^{1/2} \\ &= O(1/\sqrt{pn}) + \left(1 - \gamma^{-2\kappa^2/\pi}\right)^{1/2} = 1 - \Theta\left(\gamma^{-C}\right),\end{aligned}$$

where we used the fact that $1/\sqrt{pn} = o(\gamma^{-C})$. Reasoning as before, we obtain the following analogue of (10):

$$\mathbb{P}[R_{\leqslant t}^\psi \cap Q_{t-1}] \leqslant \left(1 - \Theta\left(\gamma^{-C}\right)\right)^t \leqslant \left(1 - \Theta\left(\gamma^{-C}\right)\right)^{m/2} = e^{-n\Theta\left(\gamma^{-C} m/n\right)} = o(2^{-n}),$$

where we used the facts that $t \geqslant m/2$ and $n/m = o(\gamma^{-C})$. As a result, our bound in (11) is also valid when $m \gg n$ as well. Then, in either regime ($m = O(n)$ or $m \gg n$),

$$\mathbb{P}[\operatorname{disc}(H) \leqslant \kappa\hat{f}] \leqslant \mathbb{P}[\operatorname{disc}(H) \leqslant \kappa\hat{f} \text{ and } Q_{t-1}] + \mathbb{P}[\neg Q_{t-1}] = o(1), \tag{12}$$

by (11), Proposition 3.3 and the fact $Q_{\lceil \alpha m \rceil - 1} \supseteq Q_{\lfloor \alpha m \rfloor}$. This shows that *w.h.p.* $\operatorname{disc}(H) \geqslant \kappa\hat{f}$, and concludes the proof of Theorem 1.2. □

3.3. **Proof of Proposition 3.3.** In this section, we prove Proposition 3.3. For any $m \in \mathbb{N}$, let $[m] := \{1, 2, \dots, m\}$ and $[0] := \emptyset$. Suppose that $J \subseteq [m]$ is a fixed subset of size $0 \leqslant j \leqslant m$. If $S \subseteq [m]$ is a random subset of size d, then the distribution of the random variable $|S \cap J|$ is said to be **hypergeometric** with parameters m, d and j. We denote this distribution by $\mathrm{Hyper}(m, d, j)$ in what follows. Now, $\mathrm{Hyper}(m, d, j)$ is at least as concentrated about its expectation as the binomial distribution, $\mathrm{Bin}(d, j/m)$ (see Chapter 21 in [13] for details). As such, standard Chernoff bounds ensure the following:

Theorem 3.4. *Suppose that $X \sim$ Hyper(m, d, j), and $\mu := \mathbb{E}[X] = dj/m$. In this case, for every $0 < \lambda < 1$,*

$$\mathbb{P}(|X - \mu| \geqslant \lambda\mu) \leqslant 2\exp\left(\frac{-\lambda^2\mu}{3}\right).$$

Let $\boldsymbol{A}$ be the adjacency matrix of $H \sim \mathcal{H}(n, m, d)$, which has exactly d ones in each column at random positions. Let $p = d/m$. Recall that, for each $i = 0, \dots, m$ and $k = 1, \dots, n$, the random variable $B_{i,k}$ counts the number of ones in the k-th column and below the i-th row of $\boldsymbol{A}$ (cf. (8)). Also recall $P_{i,k} := B_{i,k}/(m-i)$ (for $i < m$). Clearly, $B_{i,k} \sim \mathrm{Hyper}(m, d, m-i)$, so we may apply Theorem 3.4 to control the value of $B_{i,k}$, and thus of $P_{i,k}$. Let $\alpha := \max\{n/(n+m), 1/2\}$ and $t := \lceil \alpha m \rceil$. For a fixed column k, our goal is to show that *w.h.p.* $B_{i,k}$ remains "close" to $\mathbb{E}[B_{i,k}] = d(m-i)/m$ for all $i = 1, \dots, t$. By combining the error term in Theorem 3.4 with a naïve union bound, we can bound the probability of failure by something of the order of $m\exp(\Theta(-d))$, which does not tend to 0 unless $d = \Omega(\log m)$. To overcome this, we need a more subtle argument in which we take the union bound over a smaller set of indices i and take into account that $B_{i,k}$ does not change too much between two consecutive values of i. This is made more precise in the following claim:

Proposition 3.5. *Assume $0 < \alpha, \lambda, \xi < 1$ with $\xi \geqslant 1/m$ and $\alpha + \xi < 1$. Fix $1 \leqslant k \leqslant n$. Then, with probability at least*

$$1 - 8\xi^{-1}\exp\left(\frac{-d\lambda^2(1-\alpha-\xi)^2}{3}\right),$$

it holds that, for all $i = 0, \dots, \lfloor \alpha m \rfloor$,

$$(1-\lambda)\left(1 + \frac{\xi}{1-\alpha-\xi}\right)^{-1} p \leqslant P_{i,k} \leqslant (1+\lambda)\left(1 + \frac{\xi}{1-\alpha-\xi}\right) p. \tag{13}$$

Proof. As the columns of $\boldsymbol{A}$ are identically distributed, we may assume that $k = 1$ in what follows. We thus drop the index k from the notation of $A_{i,k}, B_{i,k}, P_{i,k}$ for simplicity. Recall $B_i \sim \mathrm{Hyper}(m, d, m-i)$ with $\mathbb{E}B_i = p(m-i)$ for each $i = 0, \dots, m$.

Let $r := \lceil (m-1)/\lfloor \xi m \rfloor \rceil$, which satisfies $1 \leqslant r \leqslant m-1$ by assumption. Our first goal is to partition the set $[m-1]$ into r intervals, each of size at most ξm. For each $q = 0, \dots, r-1$ let $I_q := [q\lfloor \xi m \rfloor]$, and let $I_r := [m-1]$. Then, setting $\widetilde{I}_q := I_q \setminus I_{q-1}$ for $q = 1, \dots, r$, gives us the desired partition $\widetilde{I}_1, \dots, \widetilde{I}_r$ of $[m-1]$. Now, let $r_0 := \lceil \lfloor \alpha m \rfloor / \lfloor \xi m \rfloor \rceil$. Since $r_0 \lfloor \xi m \rfloor \geqslant \lfloor \alpha m \rfloor$, the set $[\lfloor \alpha m \rfloor]$ is contained in $\bigcup_{q=1}^{r_0} \widetilde{I}_q$. Clearly, $r_0 \leqslant r$ since $\lfloor \alpha m \rfloor \leqslant m-1$. If $r_0 = r$, then $(m-1) - \lfloor \alpha m \rfloor < \lfloor \xi m \rfloor$, which implies $\lfloor \alpha m \rfloor + \lfloor \xi m \rfloor \geqslant m$ by integrality. This contradicts the fact that $\lfloor \alpha m \rfloor + \lfloor \xi m \rfloor \leqslant (\alpha + \xi)m < m$. As a result, $r_0 \leqslant r - 1$.

For each $0 \leqslant q \leqslant r-1$, define $Y_q := B_{q\lfloor \xi m \rfloor}$ and let $Y_r := B_{m-1}$. In other words, each Y_q counts the number of ones in the k-th column of $\boldsymbol{A}$ below all the rows indexed by I_q. We will prove that the variables $Y_0, \dots, Y_{r_0}$ are concentrated around their mean, and from that derive a concentration result for $B_0, \dots, B_{\lfloor \alpha m \rfloor}$. Observe that Y_r must be defined slightly differently, due to divisibility issues. Fortunately, our argument will only require the analysis of $Y_0, \dots, Y_{r_0}$, where $r_0 \leqslant r-1$, so this fact will cause no trouble. For $q = 0, \dots, r-1$,

$$\mathbb{E}Y_q = p(m - |I_q|) = p(m - q\lfloor \xi m \rfloor),$$

and therefore, for every $q = 1, \dots, r_0$,

$$\frac{\mathbb{E}Y_{q-1}}{\mathbb{E}Y_q} = 1 + \frac{\lfloor \xi m \rfloor}{m - q\lfloor \xi m \rfloor} \leqslant 1 + \frac{\lfloor \xi m \rfloor}{m - \lfloor \alpha m \rfloor - \lfloor \xi m \rfloor} \leqslant 1 + \frac{\xi}{1 - \alpha - \xi}, \tag{14}$$

where we also used the fact that $r_0\lfloor \xi m \rfloor \leqslant \lfloor \alpha m \rfloor + \lfloor \xi m \rfloor$. Now, let E be the event that

$$|Y_q - \mathbb{E}Y_q| \leqslant \lambda \mathbb{E}Y_q \qquad \text{for all } q = 0, \dots, r_0.$$

A direct application of Theorem 3.4 yields

$$\mathbb{P}(\neg E) \leqslant \sum_{q=0}^{r_0} 2\exp\left(\frac{-\lambda^2 p^2 (m - q\lfloor \xi m \rfloor)^2}{3d}\right) \leqslant 2(r_0+1)\exp\left(\frac{-\lambda^2 p^2 (m - r_0\lfloor \xi m \rfloor)^2}{3d}\right).$$

Using the fact that $r_0\lfloor \xi m \rfloor \leqslant (\alpha + \xi)m$ and the rough bound $r_0 + 1 \leqslant \frac{\alpha m}{\xi m/2} + 2 \leqslant \frac{4}{\xi}$, we conclude that

$$\mathbb{P}(\neg E) \leqslant (8/\xi)\exp\left(\frac{-d\lambda^2(1-\alpha-\xi)^2}{3}\right). \tag{15}$$

Finally, we turn our attention to $B_1, \dots, B_{\lfloor \alpha m \rfloor}$. For each $i \in [\lfloor \alpha m \rfloor]$, we pick $q \in [r_0]$ such that $i \in \widetilde{I}_q$. Then, by monotonicity,

$$Y_q \leqslant B_i \leqslant Y_{q-1} \qquad \text{and} \qquad \mathbb{E}Y_q \leqslant \mathbb{E}B_i \leqslant \mathbb{E}Y_{q-1}.$$

Combining this with (14) yields

$$\mathbb{E}Y_{q-1} \leqslant \left(1 + \frac{\xi}{1-\alpha-\xi}\right)\mathbb{E}B_i \qquad \text{and} \qquad \mathbb{E}Y_q \geqslant \left(1 + \frac{\xi}{1-\alpha-\xi}\right)^{-1}\mathbb{E}B_i.$$

As a result, event E implies that for every $1 \leqslant i \leqslant \lfloor \alpha m \rfloor$,

$$(1-\lambda)\mathbb{E}Y_q \leqslant Y_q \leqslant B_i \leqslant Y_{q-1} \leqslant (1+\lambda)\mathbb{E}Y_{q-1},$$

and hence

$$(1-\lambda)\left(1 + \frac{\xi}{1-\alpha-\xi}\right)^{-1}\mathbb{E}B_i \leqslant B_i \leqslant (1+\lambda)\left(1 + \frac{\xi}{1-\alpha-\xi}\right)\mathbb{E}B_i. \tag{16}$$

(Note that the equation above is also valid for $i = 0$, since $B_0 = d = \mathbb{E}B_0$.) Dividing (16) by $m - i$, we conclude that event E implies that, for every $0 \leqslant i \leqslant \lfloor \alpha m \rfloor$,

$$(1-\lambda)\left(1 + \frac{\xi}{1-\alpha-\xi}\right)^{-1} p \leqslant P_i \leqslant (1+\lambda)\left(1 + \frac{\xi}{1-\alpha-\xi}\right) p.$$

Our bound on $\mathbb{P}(\neg E)$ in (15) completes the proof of the proposition. □

Corollary 3.6. *Suppose $m \geqslant d \to \infty$ and $n = O(m \log d)$ as $n \to \infty$. Set $p = d/m$ and $\alpha = \max\{n/(n+m), 1/2\}$. Given any fixed constant $0 < \varepsilon < 1$ and any $1 \leqslant k \leqslant n$, the following holds with probability at least $1 - \exp\left(-\Omega(d/\log^3 d)\right)$. For every $i = 0, \dots, \lfloor \alpha m \rfloor$,*

$$|P_{i,k} - p| \leqslant \varepsilon p. \tag{17}$$

Proof. Since the probability bound in the statement is asymptotic as $n \to \infty$, we will implicitly assume throughout the proof that n is sufficiently large for all the inequalities therein to be valid. First, define $\lambda := \xi := 1/\log^{3/2} d$. Clearly, $\xi \geqslant 1/d \geqslant 1/m$. Observe that, since $n = O(m \log d)$, we have

$$1 - \alpha = \min\{m/(n+m), 1/2\} = \Omega(1/\log d). \tag{18}$$

In particular $\alpha + \xi < 1$, and thus all the assumptions in Proposition 3.5 are satisfied. Moreover,

$$\frac{\xi}{1-\alpha-\xi} = O\left(\frac{1/\log^{3/2} d}{1/\log d}\right) = O(\log^{-1/2} d).$$

As a result, we can relax the inequalities in (13) to

$$P_{i,k} = \left(1 + O(\log^{-3/2} d)\right)\left(1 + O(\log^{-1/2} d)\right) p = (1 + o(1))p,$$

which implies that $|P_{i,k} - p| \leqslant \varepsilon p$ (eventually for n sufficiently large). In view of Proposition 3.5, this fails for some $i = 0, \dots, \lfloor \alpha m \rfloor$ with probability at most

$$8\xi^{-1} \exp\left(\frac{-d\lambda^2(1-\alpha-\xi)^2}{3}\right) \leqslant 8(\log^{3/2} d) \exp\left(\frac{-d(1/2 - \log^{-3/2} d)^2}{3\log^3 d}\right)$$
$$= \exp\left(-\Omega(d/\log^3 d)\right).$$

This finishes the proof of the corollary. □

Now we are ready to prove Proposition 3.3, which we restate below in a more explicit form for convenience.

Proposition 3.7. *Let $0 < \varepsilon, c < 1$ be fixed constants with $(1+\varepsilon)c < 1$. Assume that $d \to \infty$, $dn/m \to \infty$ and $n = O(m \log d)$ as $n \to \infty$, and suppose that $p = d/m \leqslant c$. Let $\alpha = \max\{n/(n+m), 1/2\}$. Then w.h.p., for every $i = 0, \dots, \lfloor \alpha m \rfloor$,*

$$\sum_{k=1}^{n} P_{i,k} \geqslant (1-\varepsilon)pn, \tag{19}$$

and

$$P_{i,k} \leqslant (1+\varepsilon)c \quad \text{for } k = 1, \dots, n. \tag{20}$$

Proof. For $k = 1, \dots, n$, we say that column k of $\boldsymbol{A}$ is *controllable* if, for every $i = 0, \dots, \lfloor \alpha m \rfloor$, it holds that

$$|P_{i,k} - p| \leqslant \varepsilon_0 p,$$

where $\varepsilon_0 := \varepsilon/2$. Let $U \subseteq [n]$ denote the set of indices of the *uncontrollable* columns. Then, by Corollary 3.6 (with ε replaced by ε_0),

$$\mathbb{E}|U| \leqslant n \exp\left(-\Omega(d/\log^3 d)\right) = o(n).$$

Hence, we can apply Markov's inequality to ensure that $|U|/n \leqslant \varepsilon_0$ *w.h.p.* On the other hand, by applying the trivial lower bound to $P_{i,k}$ for each controllable column $k \in [n]$,

$$\sum_{k=1}^{n} P_{i,k} \geqslant (n - |U|)(1-\varepsilon_0)p \geqslant (1-\varepsilon_0)^2 pn \geqslant (1 - 2\varepsilon_0)pn,$$

w.h.p., thereby proving (19) (as $2\varepsilon_0 = \varepsilon$).

In order to verify that (20) holds, we first consider the regime in which $d \leqslant \log^2 n$. Observe then that *deterministically*

$$P_{i,k} \leqslant \frac{d}{m-i} \leqslant \frac{d}{(1-\alpha)m},$$

for each $i = 1, \dots, \lfloor \alpha m \rfloor$ and $k = 1, \dots, n$. In particular, since $1 - \alpha = \Omega(\log^{-1} d)$ (in view of (18)) and $n = O(m \log d)$, it holds that

$$P_{i,k} = O(d(\log^2 d)/n) = o(1).$$

Thus, (20) holds in this regime, as $c(1+\varepsilon) > 0$ is a fixed constant. On the other hand, if $d \geqslant \log^2 n$, we can apply Corollary 3.6 again, which ensures that with probability at least

$$1 - n \exp\left(-\Omega(d/\log^3 d)\right) = 1 - o(1)$$

we have

$$P_{i,k} \leqslant (1+\varepsilon)p \leqslant (1+\varepsilon)c$$

for all $i = 0, \dots, \lfloor \alpha m \rfloor$ and $k = 1, \dots, n$. The proof is therefore complete. □

4. Upper Bounding Discrepancy—Proof of Theorem 1.3

The main tool we make use of is the algorithmic partial colouring lemma [17], as done in [4, 19, 20]. For convenience, we restate this lemma in the relevant hypergraph terminology, where we define a **fractional colouring** to be a relaxation of a (hypergraph) colouring, whose values are allowed to lie in the interval $[-1, 1]$:

Lemma 4.1 (Partial Colouring Lemma [17])**.** *Suppose that $H = (V, E)$ is a hypergraph with m edges and n vertices which are coloured by some fractional colouring $\rho : V \to [-1, 1]$. Moreover, assume that $\delta > 0$ and $(\lambda_e)_{e\in E}$ are non-negative values such that*

$$\sum_{e\in E} \exp(-\lambda_e^2/16) \leqslant \frac{n}{16}. \tag{21}$$

Under these assumptions, there exists some fractional colouring $\psi : V \to [-1, 1]$ for which

1. *$|\psi(e) - \rho(e)| \leqslant \lambda_e |e|^{1/2}$ for all $e \in E$, and*
2. *$|\psi(v)| \geqslant 1 - \delta$ for at least $n/2$ vertices of V.*

Moreover, ψ can be computed by a randomized algorithm in expected time

$$O((n + m)^3 \delta^{-2} \log(mn/\delta)).$$

In what follows, it will be convenient to refer to ρ as the **target (fractional) colouring** and δ as the **rounding parameter**. We make use of Lemma 4.1 in the same way as in [4, 17, 19, 20]. In fact, we analyze the same *two-phase* algorithm considered by both Bansal and Meka [4] and Potukuchi [19, 20], though we must tune the relevant asymptotics carefully in order to achieve the bound claimed in Theorem 1.3. In particular, we shorten phase one and change the target discrepancy in each application of Lemma 4.1. These modifications allow us to derive more precise asymptotics in the studied range of parameters.

Let us suppose that $H = (V, E)$ is a hypergraph drawn from $\mathcal{H}(n, m, d)$ or $\mathbb{H}(n, m, p)$, where $p = d/m$. From now on, we assume that n is a power of 2 for convenience. This follows *w.l.o.g.* as we can always add extra vertices to V which do not lie in any of the edges of G. Given the lower bounds of Theorems 1.1 and 1.2, ideally we would like to compute an **output colouring**, $\phi : V \to \{-1, 1\}$ with matching discrepancy. However, without finer proof techniques, this does not seem fully attainable for the full parameter range of $m \gg n$. Let us fix $\mu := dn/m$. We recall the definition of $\beta = \beta(n)$ as given in the statement of Theorem 1.3: For all n sufficiently large, $\beta(n) \geqslant 1$ and

$$\beta(n)\mu \geqslant \log(m/n) \left(\log \mu + 2\right)^5. \tag{22}$$

Let $\hat{f} := \sqrt{\mu \log(m/n)\beta}$ be the target upper bound on the discrepancy that we are aiming to prove. Our argument is based on analyzing the Iterated-Colouring-Algorithm, which we now describe. Fix $t_1 := \lg \mu$, $0 \leqslant i \leqslant t_1$, and let $H_0 := H$, and $\delta := 1/n$. For convenience, we refer to the below procedure as **round** i. We define $\hat{f}_i := \hat{f}(i+2)^{-2}$ as the desired discrepancy bound to be attained in round i.

1. Remove all edges of $H_i = (V_i, E_i)$ of size less than or equal to $\hat{f}$.
2. Update λ_e to be $\hat{f}_i/|e|^{1/2}$ for each $e \in E_i$. Update the target colouring, which we denote by ρ_i, to be the previously computed colouring $\psi_{i-1} : V_{i-1} \to [-1, 1]$(where ψ_{-1} is the identically zero function by convention), restricted to V_i.
3. If $\sum_{e\in E_i} \exp(-\lambda_e^2/16) \leqslant |V_i|/16$, then apply Lemma 4.1 to H_i with the above values, yielding the fractional colouring $\psi_i : V_i \to [-1, -1]$. Otherwise, abort the round and declare an error.
4. Assuming an error did not occur, compute $S_i \subseteq V_i$, such that $|\psi_i(v)| \geqslant 1-1/n$ for all $v \in S_i$, where $|S_i| := |V_i|/2$. Afterwards, construct $H_{i+1} = (V_{i+1}, E_{i+1})$ by restricting the edges of E_i to V_{i+1}, where $V_{i+1} := V_i \setminus S_i$.

We refer to the rounds $i = 0, \dots, t_1$ as **phase one** of the algorithm's execution. Note that we refer to the vertices of S_i as **inactive** after round i, as the value ϕ assigns to them will not change at any point onwards. We refer to the remaining vertices as being **active**. Observe then the following proposition:

Proposition 4.2. *For each $0 \leqslant i \leqslant t_1$ and $e \in E_i$, it holds that*

$$|\psi_i(e) - \rho_i(e)| \leqslant \frac{\hat{f}}{(i+2)^2}.$$

Assuming that none of the rounds yielded an error, there are exactly $2^{-t_1}n = n\mu^{-1} = m/d$ active vertices at the end of phase one, as $t_1 = \lg \mu$. Heuristically, this means that we expect H_{t_1+1} to have edges of roughly constant size. As such, we can easily complete the colouring ϕ by executing the **phase two** procedure for rounds $t_1 + 1 \leqslant i \leqslant t_2$, where $t_2 := \lg(10n/\hat{f}) + 1$. The phase two procedure is identical to that of phase one, with the exception that in step 2, we update λ_e to be 0 (rather than $\hat{f}_i/|e|^{1/2}$) for each $e \in E_i$. Analogously, we observe the following.

Proposition 4.3. *For each $t_1 + 1 \leqslant i \leqslant t_2$ and $e \in E_i$, it holds that*

$$|\psi_i(e) - \rho_i(e)| = 0.$$

Assuming that none of the rounds yielded an error, there are exactly $2^{-t_2}n = n\hat{f}/20n = \hat{f}/20$ active vertices at the end of phase two. In order to complete the construction of ϕ, we conclude with a **post-processing phase**. That is, we arbitrarily assign -1 or 1 to any of the vertices which remain active at the end of phase two. Finally, we round each remaining fractional value assigned by ϕ to the nearest integer within $\{-1, 1\}$.

Let us assume that the above procedure succeeds in its execution on H; that is, it does *not* abort during any iteration in either phase one or two. In this case, we conclude the proof by showing the next lemma.

Lemma 4.4. *If neither phase one nor phase two fails then Theorem 1.3 holds.*

Proof. For each $0 \leqslant i \leqslant t_2$, let us formally extend ψ_i to all of V. That is, define $\psi_i(v) := 0$ for each $v \in V \setminus V_i$, and keep ψ_i unchanged on V_i. Moreover, do the same for the target colouring ρ_i. Observe then that once phase two ends, ϕ can be expressed as a sum of differences involving the partial colourings $(\psi_i)_{i=0}^{t_2}$ and $(\rho_i)_{i=0}^{t_2-1}$ Specifically,

$$\phi(v) = \sum_{i=0}^{t_2} (\psi_i(v) - \rho_i(v)).$$

Let t_e be the time when edge e becomes smaller than $\hat{f}$ or $t_e = t_2$ if it never happens. After applying Propositions 4.2 and 4.3 we get that

$$\begin{aligned}
|\phi(e)| &\leqslant \hat{f} + \sum_{i=0}^{t_e} |\psi_i(e) - \rho_i(e)| \\
&\leqslant \hat{f} + \sum_{i=0}^{t_1} |\psi_i(e) - \rho_i(e)| + \sum_{i=t_1+1}^{t_2} |\psi_i(e) - \rho_i(e)| \\
&\leqslant \hat{f} + \sum_{i=0}^{\infty} \frac{\hat{f}}{(i+2)^2} = O(\hat{f}).
\end{aligned}$$

The post-processing phase cannot increase the discrepancy that ϕ attains on any edge of E by more than $\hat{f}$ for the remaining active vertices; as we already observed, there are at most $\hat{f}/20$ of them. The rounding of inactive vertices increases the discrepancy by at most 1. □

Bounding the Failure Probability. First, we recall the following lemma proven in [4] by Bansal and Meka:

Lemma 4.5 (Lemma 6 in [4])**.** *Suppose that H is generated from $\mathcal{H}(n,m,d)$ and $\boldsymbol{M}$ is a fixed $r \times \ell$ sub-matrix of the $m \times \ell$ incidence matrix $\boldsymbol{A}$ of H. If $s \geqslant 10d\ell/m$, and $B(r,\ell,s)$ corresponds to the event in which each row of $\boldsymbol{M}$ has at least s 1's, then*

$$\mathbb{P}[B(r,\ell,s)] \leqslant \exp\left(-\frac{rs\log((sm)/(d\ell))}{2}\right).$$

While this lemma is stated for the case when H is generated from $\mathcal{H}(n,m,d)$, the upper bound on $\mathbb{P}[B(r,\ell,s)]$ is proven by instead bounding the probability of the analogous event when H is generated from $\mathbb{H}(n,m,p)$ for $p = d/m$. As such, this lemma extends to the edge independent model. We now restate it in a form which will be more convenient for our purposes.

Lemma 4.6. *Suppose that H is generated from $\mathcal{H}(n,m,d)$ or $\mathbb{H}(n,m,p)$ for $p = d/m$, whose incidence matrix we denote by $\boldsymbol{A}$. If $s \geqslant 10d\ell/m$, then define $Q(r,\ell,s)$ as the event in which there exists an $r \times \ell$ sub-matrix of $\boldsymbol{A}$ in which each row has at least s 1's. In this case,*

$$\mathbb{P}[Q(r,\ell,s)] \leqslant \binom{m}{r}\binom{n}{\ell}\exp\left(-\frac{rs\log((sm)/(d\ell))}{2}\right).$$

Using this lemma, we can ensure that *w.h.p.* Iterated-Colouring-Algorithm will not abort during phase one (Proposition 4.7) or two (Proposition 4.8) and thus conclude the proof of Theorem 1.3.

Proposition 4.7. *If* Iterated-Colouring-Algorithm *inputs a hypergraph drawn from $\mathcal{H}(n,m,d)$ or $\mathbb{H}(n,m,p)$ where $p = d/m$, then w.h.p. it does not abort during phase one, provided we assume that $\mu = dn/m \to \infty$ and $m \gg n$.*

Proof. Given $0 \leqslant i \leqslant t_1$, we say that round i is **good**, provided there are at most $n_i/17$ rows of H_i whose size is greater than $s_i := \beta\mu/16(i+2)^5$, where β satisfies (22). Otherwise, we say that the round is **bad**. Recall that $t_1 = \lg\mu$.

Now, if round i is good, then we claim that Iterated-Colouring-Algorithm does *not* abort in iteration i. To see this, it suffices to show that for n sufficiently large

$$\sum_{e\in E_i}\exp\left(-\lambda_e^2/16\right) \leqslant n_i/16,$$

where $n_i := |V_i| = n/2^i$, $\hat{f}_i := \hat{f}(i+2)^{-2}$ and $\lambda_e := \hat{f}_i/|e|^{1/2}$ for $e \in E_i$. Observe now that since the round is good, we get that

$$\begin{aligned}\sum_{e\in E_i}\exp\left(-\lambda_e^2/16\right) &= \sum_{\substack{e\in E_i:\\ |e|\leqslant s_i}}\exp\left(-\lambda_e^2/16\right) + \sum_{\substack{e\in E_i:\\ |e|> s_i}}\exp\left(-\lambda_e^2/16\right)\\ &\leqslant m\exp\left(-\frac{\hat{f}_i^2}{16s_i}\right) + n_i/17.\end{aligned}$$

On the other hand, since $\hat{f} := \sqrt{\mu\log(m/n)\beta}$,

$$\begin{aligned}m\exp\left(-\hat{f}_i^2/16s_i\right) &= m\exp(-\log(m/n)(i+2))\\ &= m\left(\frac{n}{m}\right)^{i+2}\\ &= n\left(\frac{n}{m}\right)^{i+1} = o(n_i)\end{aligned}$$

where the last line follows since $(n/m)^{i+1} \ll 2^{-i}$, as $n \ll m$. Thus,

$$\sum_{e\in E_i} \exp\left(-\lambda_e^2/16\right) \leqslant (1+o(1))\frac{n_i}{17} \leqslant \frac{n_i}{16}.$$

We now must show that *w.h.p.*, all of the rounds are good. Now, observe that if some round $1 \leqslant i \leqslant t_1$ is *bad*, then there exists an $(n_i/17) \times n_i$ sub-matrix of $\boldsymbol{A}$, say $\boldsymbol{M}$, in which each row of $\boldsymbol{M}$ has greater than s_i 1's. In fact, since $n_{t_1} \leqslant n_i$ and $s_{t_1} \leqslant s_i$, we can take a sub-matrix of $\boldsymbol{M}$ (and thus of $\boldsymbol{A}$) in which each row has at least s_{t_1} 1's, and whose size is $(n_{t_1}/17) \times n_{t_1}$. Thus, we observe the following claim:

(1) If a bad round occurs, then there exists a $(n_{t_1}/17) \times n_{t_1}$ sub-matrix of $\boldsymbol{A}$ in which each row has more than s_{t_1} 1's.

Let us define $Q(n_{t_1}/17, n_{t_1}, s_{t_1})$ as this latter event; namely, that there exists an $(n_{t_1}/17) \times n_{t_1}$ sub-matrix of $\boldsymbol{A}$ in which each row has more than s_{t_1} 1's. In order to complete the proof, it suffices to show that *w.h.p.*, $Q(n_{t_1}/17, n_{t_1}, s_{t_1})$ does not occur. Recall $n_{t_1} = n/\mu = m/d.$ as the number of active vertices drops by exactly half in each round. It follows that

$$s_{t_1} = \frac{\beta\mu}{16(t_1+2)^5} \geqslant \frac{\log\frac{m}{n}(\log\mu+2)^5}{16(\lg\mu+2)^5} \gg 5 = 10\frac{dn}{2\mu m} \geqslant 10\frac{dn_{t_1}}{m}.$$

Thus, we can apply Lemma 4.6 to ensure that

$$\begin{aligned}
\mathbb{P}[Q(n_{t_1}/17, n_{t_1}, s_{t_1})] &\leqslant \binom{m}{n_{t_1}/17}\binom{n}{n_{t_1}} \exp\left(-\frac{n_{t_1}s_{t_1}}{34}\log\left(\frac{s_{t_1}m}{dn_{t_1}}\right)\right) \\
&\leqslant \binom{m}{n_{t_1}}^2 \exp\left(-\frac{n_{t_1}s_{t_1}}{34}\log s_{t_1}\right) \\
&\leqslant \left(\frac{me}{n_{t_1}}\right)^{2n_{t_1}} \exp\left(-\frac{n_{t_1}s_{t_1}\log s_{t_1}}{34}\right)
\end{aligned}$$

where the inequalities follow since $m \gg n$, $\binom{m}{n_{t_1}} \leqslant (me/n_{t_1})^{n_{t_1}}$. Now,

$$\begin{aligned}
\left(\frac{me}{n_{t_1}}\right)^{2n_{t_1}} &= \exp\left(2n_{t_1}(\log(m/n) + \log(ne/n_{t_1}))\right) \\
&= \exp\left(2n_{t_1}(\log(m/n) + \log(e\mu))\right),
\end{aligned}$$

so

$$\mathbb{P}[Q(n_{t_1}/17, n_{t_1}, s_{t_1})] \leqslant \exp\left(-2n_{t_1}\left(\frac{s_{t_1}\log s_{t_1}}{68} - \log(m/n) - \log(e\mu)\right)\right).$$

Thus, by our assumption (22) on β, we get that

$$s_{t_1} = \frac{\beta\mu}{16\left(\lg\mu+2\right)^5} \geqslant \frac{\log\frac{m}{n}(\log\mu+2)^5}{16(\lg\mu+2)^5} \geqslant \frac{1}{16}\log(m/n),$$

and

$$s_{t_1} = \frac{\beta\mu}{16\left(\lg\mu+2\right)^5} \geqslant \frac{\mu}{16\left(\lg\mu+2\right)^5},$$

as $\beta \geqslant 1$. The proposition follows as

$$\frac{\log(m/n)\log(\log(m/n)/16)}{16\cdot 68} \gg \log(m/n),$$

and

$$\frac{\mu}{68\cdot 16\left(\lg\mu+2\right)^5} \gg \log(\mu e).$$

□

Proposition 4.8. *If* Iterated-Colouring-Algorithm *inputs a hypergraph drawn from* $\mathcal{H}(n,m,d)$ *or* $\mathbb{H}(n,m,p)$ *where* $p = d/m$*, then w.h.p. it does not abort in phase two, provided* $\mu = dn/m \to \infty$ *as* $n \to \infty$ *and* $m \gg n$*.*

Proof. Suppose that $t_1 + 1 \leqslant i \leqslant t_2$ for $t_2 := \lg(10n/\hat{f}) + 1$. Recall $n_{t_2} = n/2(10n/\hat{f}) = \hat{f}/20$ and that during phase two, $\lambda_e = 0$ for each $e \in E_i$. Thus, we get that

$$\sum_{e \in E_i} \exp\left(\lambda_e^2/16\right) = |E_i|.$$

On the other hand, in order for an edge of H to remain in E_i, it must have greater than $\hat{f}$ vertices which lie in V_i. As a result, if Iterated-Colouring-Algorithm aborts in round i, then H_i must have at least $n_i/16$ edges of size greater than $\hat{f}$. In particular, since $n_{t_2} \leqslant n_i$, this implies that the incidence matrix $\boldsymbol{A}$ of H has an $(n_{t_2}/16) \times n_{t_2}$ sub-matrix in which each row has greater than $\hat{f}$ 1's. Thus, if $Q(n_{t_2}/16, n_{t_2}, \hat{f})$ corresponds to the event in which $\boldsymbol{A}$ has an $(n_{t_2}/16) \times n_{t_2}$ sub-matrix in which each row has greater than $\hat{f}$ 1's, then we get the following claim:

(1) If Iterated-Colouring-Algorithm aborts in some round $t_1+1 \leqslant i \leqslant t_2$, then $Q(n_{t_2}/16, n_{t_2}, \hat{f})$ must occur.

As a result, in order to show that *w.h.p.* Iterated-Colouring-Algorithm does *not* abort in any round it suffices to prove that $Q(n_{t_2}/16, n_{t_2}, \hat{f})$ does not occur *w.h.p.* Now, it follows that

$$\hat{f} \geqslant 10\frac{d\hat{f}}{20m} = 10\frac{dn_{t_2}}{m},$$

as $d \leqslant m$. Thus, we can apply Lemma 4.6 to ensure that

$$\begin{aligned}\mathbb{P}[Q(n_{t_2}/16, n_{t_2}, \hat{f})] &\leqslant \binom{m}{n_{t_2}/16}\binom{n}{n_{t_2}} \exp\left(-\frac{n_{t_2}\hat{f}}{32}\log\left(\frac{\hat{f}m}{dn_{t_2}}\right)\right) \\ &\leqslant \binom{m}{n_{t_2}/16}\binom{n}{n_{t_2}} \exp\left(-\frac{n_{t_2}\hat{f}}{32}\log\left(\frac{20m}{d}\right)\right),\end{aligned}$$

and so $\mathbb{P}[Q(n_{t_2}/16, n_{t_2}, \hat{f})]$ is upper bounded by

$$\exp\left(-2n_{t_2}\left(\frac{\hat{f}}{64} - \log(m/n) - \log(\mu e)\right)\right), \tag{23}$$

after applying the same simplifications as in Proposition 4.7. The proposition then follows by assumption (22) on β, as

$$\hat{f}/64 = \sqrt{\beta\mu\log(m/n)}/64 \geqslant \frac{\log(m/n)\log^5(\mu+2)}{64} \gg \log m/n,$$

and

$$\frac{\log(m/n)\log^5(\mu+2)}{64} \gg \log(\mu e).$$ □

5. Conclusion and Open Problems

We have lower bounded the discrepancy of the random hypergraph models $\mathbb{H}(n,m,p)$ and $\mathcal{H}(n,m,d)$ for the full parameter range in which $d \to \infty$ and $dn/m \to \infty$ where $p = d/m$. In the dense regime of $m \gg n$, we have provided asymptotically matching upper bounds, under the assumption that $d = pm \geqslant (m/n)^{1+\varepsilon}$ for some constant $\varepsilon > 0$. These upper bounds are algorithmic, and so the main question left open by our work is whether analogous upper bounds can be proven in the sparse regime of $n/\log n \ll m \ll n$. Our lower bounds suggest that the discrepancy is $\Theta\left(2^{-n/m}\sqrt{pn}\right)$, and while we believe that a second moment argument could be used to prove the existence of such

a colouring—particularly, in the edge-independent model $\mathbb{H}(n, m, p)$—the partial colouring lemma does not seem to be of much use here. This leaves open whether such a colouring can be computed efficiently in this parameter range. If this is not possible, then ideally one could find a reduction to a problem which is believed to be hard on average. One candidate may be the random-lattice problem of Ajtai [1] and Goldreich et al. [14], in which a random m by n matrix $\boldsymbol{M}$ with *i.i.d.* entries from $\mathbb{Z}_q$ is generated, and one wishes to compute a vector $\boldsymbol{x} \in \{0, 1\}^n$ such that $\boldsymbol{M}\boldsymbol{x} = 0$.

Acknowledgements

This work was initiated at the 2019 Graduate Research Workshop in Combinatorics, which was supported in part by NSF grant #1923238, NSA grant #H98230-18-1-0017, a generous award from the Combinatorics Foundation, and Simons Foundation Collaboration Grants #426971 (to M. Ferrara), #316262 (to S. Hartke) and #315347 (to J. Martin). We thank Aleksandar Nikolov for suggesting the problem, and Puck Rombach and Paul Horn for discussions and encouragements in the early stages of the project. Moreover, T. Masařík received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme Grant Agreement 714704. He completed a part of this work while he was a postdoc at Simon Fraser University in Canada, where he was supported through NSERC grants R611450 and R611368. X. Pérez-Giménez was supported in part by Simons Foundation Grant #587019.

References

[1] Miklós Ajtai. "Generating Hard Instances of Lattice Problems (Extended Abstract)". In: *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*. STOC '96. Philadelphia, Pennsylvania, USA: Association for Computing Machinery, 1996, pp. 99–108. ISBN: 0897917855. DOI: 10.1145/237814.237838. URL: https://doi.org/10.1145/237814.237838.

[2] Dylan J. Altschuler and Jonathan Niles-Weed. "The Discrepancy of Random Rectangular Matrices". In: *CoRR* abs/2101.04036 (2021). arXiv: 2101.04036.

[3] Wojciech Banaszczyk. "Balancing vectors and Gaussian measures of n-dimensional convex bodies". In: *Random Structures & Algorithms* 12.4 (1998), pp. 351–360. DOI: 10.1002/(SICI)1098-2418(199807)12:4<351::AID-RSA3>3.0.CO;2-S.

[4] Nikhil Bansal and Raghu Meka. "On the discrepancy of random low degree set systems". In: *Random Structures & Algorithms* 57.3 (June 2020), pp. 695–705. DOI: 10.1002/rsa.20935.

[5] József Beck and Tibor Fiala. ""Integer-making" theorems". In: *Discrete Applied Mathematics* 3.1 (Feb. 1981), pp. 1–8. DOI: 10.1016/0166-218x(81)90022-6.

[6] Boris Bukh. "An Improvement of the Beck–Fiala Theorem". In: *Combinatorics, Probability and Computing* 25.3 (2016), pp. 380–398. DOI: 10.1017/S0963548315000140.

[7] Bernard Chazelle. *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, July 2000. DOI: 10.1017/cbo9780511626371.

[8] William Chen, Anand Srivastav, and Giancarlo Travaglini, eds. *A Panorama of Discrepancy Theory*. Springer International Publishing, 2014. DOI: 10.1007/978-3-319-04696-9.

[9] Persi Diaconis and Susan Holmes. *Stein's Method: Expository Lectures and Applications*. Institute of Mathematical Statistics. Institute of Mathematical Statistics, 2004. ISBN: 978-0-940600-62-1. URL: https://books.google.ca/books?id=3n-iQIU9LNEC.

[10] Esther Ezra and Shachar Lovett. "On the Beck-Fiala conjecture for random set systems". In: *Random Structures & Algorithms* 54.4 (Nov. 2018), pp. 665–675. DOI: 10.1002/rsa.20810.

[11] William Feller. *An introduction to probability theory and its applications. Vol. II*. Second edition. John Wiley & Sons Inc., 1971.

[12] Cole Franks and Michael Saks. "On the discrepancy of random matrices with many columns". In: *Random Structures & Algorithms* 57.1 (2020), pp. 64–96. DOI: 10.1002/rsa.20909.

[13] Alan Frieze and Michał Karoński. *Introduction to Random Graphs*. Cambridge University Press, 2015. DOI: 10.1017/CBO9781316339831.

[14] Oded Goldreich, Shafi Goldwasser, and Shai Halevi. "Collision-Free Hashing from Lattice Problems". In: *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation: In Collaboration with Lidor Avigad, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman*. Ed. by Oded Goldreich. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 30–39. ISBN: 978-3-642-22670-0. DOI: 10.1007/978-3-642-22670-0_5.

[15] Rebecca Hoberg and Thomas Rothvoss. "A fourier-analytic approach for the discrepancy of random set systems". In: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2019, pp. 2547–2556. DOI: 10.1137/1.9781611975482.156.

[16] Avi Levy, Harishchandra Ramadas, and T. Rothvoss. "Deterministic Discrepancy Minimization via the Multiplicative Weight Update Method". In: *Integer Programming and Combinatorial Optimization*. 2017, pp. 380–391. DOI: 10.1007/978-3-319-59250-3_31.

[17] Shachar Lovett and Raghu Meka. "Constructive Discrepancy Minimization by Walking on the Edges". In: *SIAM Journal on Computing* 44.5 (Jan. 2015), pp. 1573–1582. ISSN: 1095-7111. DOI: 10.1137/130929400.

[18] Jiří Matoušek. *Geometric Discrepancy: An Illustrated Guide*. 1st ed. Algorithms and Combinatorics 18. Springer-Verlag Berlin Heidelberg, 1999. DOI: 10.1007/978-3-642-03942-3.

[19] Aditya Potukuchi. *Discrepancy in random hypergraph models*. 2018. arXiv: 1811.01491 [math.CO].

[20] Aditya Potukuchi. "A Spectral Bound on Hypergraph Discrepancy". In: *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*. Ed. by Artur Czumaj, Anuj Dawar, and Emanuela Merelli. Vol. 168. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020, 93:1–93:14. ISBN: 978-3-95977-138-2. DOI: 10.4230/LIPIcs.ICALP.2020.93.

[21] Irina Shevtsova. "An Improvement of Convergence Rate Estimates in the Lyapunov Theorem". In: *Doklady Mathematics* 82 (Dec. 2010), pp. 862–864. DOI: 10.1134/S1064562410060062.

[22] Paxton Turner, Raghu Meka, and Philippe Rigollet. "Balancing Gaussian vectors in high dimension". In: *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*. Ed. by Jacob D. Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 3455–3486. URL: http://proceedings.mlr.press/v125/turner20a.html.

[23] J. D. Williams. "An Approximation to the Probability Integral". In: *Annals of Mathematical Statistics* 17.3 (Sept. 1946), pp. 363–365. DOI: 10.1214/aoms/1177730951.

Department of Computer Science, University of Toronto, Toronto, ON, Canada
Email address: cmacrury@cs.toronto.edu

Department of Applied Mathematics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic & Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland & Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada
Email address: masarik@kam.mff.cuni.cz

Department of Mathematics, University of Nebraska-Lincoln, Lincoln NE, USA
Email address: lpai@huskers.unl.edu

Department of Mathematics, University of Nebraska-Lincoln, Lincoln NE, USA
Email address: xperez@unl.edu