




Greedy Heuristics and Linear Relaxations for the Random Hitting Set Problem

Gabriel Arpino   

University of Cambridge, UK

Daniil Dmitriev   

ETH Zürich and ETH AI Center, Switzerland

Nicolo Grometto 

Princeton University, USA

Abstract

Consider the **Hitting Set** problem where, for a given universe $\mathcal{X} = \{1, \dots, n\}$ and a collection of subsets $\mathcal{S}_1, \dots, \mathcal{S}_m$, one seeks to identify the smallest subset of \mathcal{X} which has a nonempty intersection with every element in the collection. We study a probabilistic formulation of this problem, where the underlying subsets are formed by including each element of the universe independently with probability p . We rigorously analyze integrality gaps between linear programming and integer programming solutions to the problem. In particular, we prove the absence of an integrality gap in the sparse regime $mp \lesssim \log n$ and the presence of a non-vanishing integrality gap in the dense regime $mp \gg \log n$. Moreover, for large enough values of n , we look at the performance of Lovász's celebrated **Greedy** algorithm [12] with respect to the chosen input distribution, and prove that it finds optimal solutions up to multiplicative constants. This highlights separation of **Greedy** performance between average-case and worst-case settings.

2012 ACM Subject Classification Mathematics of computing \rightarrow Combinatorial optimization; Theory of computation \rightarrow Approximation algorithms analysis; Theory of computation \rightarrow Randomness, geometry and discrete structures

Keywords and phrases Hitting Set, Random Hypergraph, Integrality Gap, Greedy Algorithm

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2024.30

Category APPROX

Related Version *Extended Version*: <https://arxiv.org/abs/2305.05565> [1]

Acknowledgements The authors thank Dylan J. Altschuler, Afonso S. Bandeira, Raphaël Barboni, and Anastasia Kireeva for helpful discussions. DD is supported by ETH AI Center doctoral fellowship and ETH Foundations of Data Science initiative. GA is supported by the Cambridge Trust. NG is grateful for the funding received from Elizaveta Rebrova.

1 Introduction

Hitting Set is a classical problem in combinatorial optimization which, for a given ground set $\mathcal{X} := \{1, \dots, n\}$ of elements and a collection $\mathcal{C} := \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$ of subsets of \mathcal{X} , asks to identify the smallest set $\mathcal{S} \subseteq \mathcal{X}$ that intersects every subset in \mathcal{C} . **Hitting Set** arises naturally from the study of *Minimum Vertex Covers on Hypergraphs* (MVCH), upon viewing hyperedges as subsets and vertices as elements of the ground set. This is also known as the *Set Cover* problem [14], which has a rich history in worst-case computational complexity theory, including appearing as one of Karp's 21 NP-complete problems. An important question regards the behaviour of natural random instances of **Hitting Set** where each element of the ground set is independently assigned to any subset with probability p , motivated, among others, by applications such as group testing [10]. A classical theorem of Lovász [12] gives



© Gabriel Arpino, Daniil Dmitriev, and Nicolo Grometto;
licensed under Creative Commons License CC-BY 4.0

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2024).

Editors: Amit Kumar and Noga Ron-Zewi; Article No. 30; pp. 30:1–30:22



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

an upper bound on the integrality gap in this problem which grows with the degree of the underlying hypergraph, i.e., the maximum number of subsets intersecting any one element. This bound was shown to be tight in the worst-case, but leaves much to be desired from an average-case perspective.

In this paper, we characterize the average-case integrality gap present in random **Hitting Set** and prove that, with high probability, Lovász's greedy algorithm [12] finds the minimal hitting set in polynomial time. Namely, we consider the following integer programming (IP) formulation of the problem,

$$\text{val}_{\text{IP}} := \begin{cases} \text{minimize} & \|\mathbf{x}\|_1 \\ \text{subject to} & \mathbf{Ax} \geq \mathbf{1}, \mathbf{x} \in \{0, 1\}^n, \end{cases} \quad (1.1)$$

where the i -th row of $\mathbf{A} \in \{0, 1\}^{m \times n}$ provides a binary encoding of the membership of the elements of \mathcal{X} in the set \mathcal{S}_i and $\mathbf{1} := (1, \dots, 1) \in \mathbb{R}^m$. With the vertex cover formulation of the problem at hand, we note that \mathbf{A} consists of the incidence matrix of the underlying hypergraph. In particular, the constraint $\mathbf{Ax} \geq \mathbf{1}$ ensures that each set in \mathcal{C} is hit by a prescribed candidate solution vector. A natural convex relaxation is obtained by allowing fractional solutions, and may be expressed as the following linear program (LP),

$$\text{val}_{\text{LP}} := \begin{cases} \text{minimize} & \|\mathbf{x}\|_1 \\ \text{subject to} & \mathbf{Ax} \geq \mathbf{1}, \mathbf{x} \in [0, 1]^n. \end{cases} \quad (1.2)$$

Whilst clearly $\text{val}_{\text{LP}} \leq \text{val}_{\text{IP}}$, tightness need not hold in general. In fact, for $m = n$ and $\mathbf{A} \in \{0, 1\}^{n \times n}$ chosen such that each row and column contains exactly k ones, for some fixed $1 < k < n$, an optimal solution is provided by $\mathbf{x}_{\text{LP}}^* = (1/k, \dots, 1/k)$, which is not integral, thus leading to a strictly smaller objective whenever n/k is not an integer. This evidences the existence of a multiplicative *integrality gap*, as we define next.

► **Definition 1.** *Given solutions val_{IP} and val_{LP} to Equation (1.1) and Equation (1.2) respectively, we define multiplicative integrality gap as follows:*

$$\text{IPGAP} := \frac{\text{val}_{\text{IP}}}{\text{val}_{\text{LP}}}. \quad (1.3)$$

In [12], Lovász proved an essentially optimal worst-case upper bound on the **Hitting Set** multiplicative integrality gap: $\text{IPGAP} \leq 1 + \log d_{\max}$, where d_{\max} corresponds to the maximum degree in the underlying hypergraph. This is obtained by analysing the **Greedy** algorithm (Algorithm 1), which constructs a vertex cover by sequentially adding vertices with the highest degree amongst the uncovered edges, and will be discussed in more detail in the next sections. However, in many natural examples, the maximum degree d_{\max} grows with the number of vertices in the hypergraph, thus leading to progressively worse bounds for increasingly large hypergraphs. Besides being arguably the most natural candidate for solving **Hitting Set**, the greedy algorithm has been shown to be the best possible polynomial time approximation algorithm [15] for the worst-case instances of this classical problem.

Despite extensive work conducted on **Hitting Set** in the last decades, a gap remains in our understanding of the typical performance of linear programming and the greedy algorithm on random problem instances. We hence pose the following questions:

1. Are there integrality gaps in random instances of **Hitting Set**?
2. Can near-optimal solutions be found efficiently?

In the present work, we provide answers to the above questions *with high probability* (w.h.p.) in a non-asymptotic sense, in the setting where the cardinality n of the ground set \mathcal{X} is large but finite. We will prove the absence of integrality gaps up to constants in a wide regime of n, m, p , by conducting an average case analysis of an algorithm that outputs integral covers of matching size to the fractional ones. In addition, a rigorous analysis of the greedy routine will follow by a straightforward reduction. The forthcoming results are valid under the conditions listed below, which will be assumed to hold throughout.

► **Assumption 2.** *We assume that*

1. *Each element $j \in \mathcal{X}$ is assigned to any subset \mathcal{S}_i , $i \in [m]$ with probability $p \equiv p(n)$, independently. That is, $\mathbf{A} \in \{0, 1\}^{m \times n}$ is such that $A_{ij} \stackrel{iid}{\sim} \text{Bernoulli}(p)$;*
2. *n is intended to be large but finite;*
3. *$m \equiv m(n) = \text{poly}(n)$, i.e. $\exists c, C > 0$, such that $cn^c \leq m \leq Cn^C$ for n large enough;*
4. *There exist $\delta \in (0, 1)$, such that $p \equiv p(n)$ satisfies $1/n^\delta \leq p \leq 1/2$, for all n large enough.*

Note that in Assumption 2.3, the upper bound is chosen to avoid trivial solutions w.h.p. which arise, for example, in the setting where the number of sets grows exponentially in the cardinality of \mathcal{X} . In addition, Assumption 2.4 is by no means restrictive, since one can show that for $m = \text{poly}(n)$ and $np \ll \log n$, we have that \mathbf{A} contains an all-zero row w.h.p., yielding an infeasible solution for IP. The requirement $p \leq 1/2$ is chosen for technical convenience and can be relaxed to any constant p , encompassing the regime in [10].

Our contributions stem from the study of the size of the inclusion sets $I_j := \{i \in [m] : j \in \mathcal{S}_i\}$, for $j \in [n]$, which in the MVCH formulation of the problem at hand correspond to the set of hyperedges incident to any given vertex. The key quantity under study is the average inclusion set size, that is $\mathbb{E}|I_j| = mp$, for all j , under the present distributional assumptions. This quantity exhibits two separate regimes of interest, referred to as the *sparse*, $mp \ll \log n$, and *dense*, $mp \gg \log n$, regimes. These, in turn, determine the size of the maximum inclusion set, or maximum degree, $d_{\max} := \max_{j \in [n]} |I_j|$. We characterize the integrality gap behaviour up to multiplicative constants and analyse Lovász's Greedy algorithm [12] in these two regimes w.h.p as $n \rightarrow \infty$. We do this by proving the success of a simple greedy heuristic, the BlockGreedy algorithm (Algorithm 2). Throughout, we use the notation val_{Gr} , val_{BGr} to denote the size of the hitting set returned by Greedy and BlockGreedy respectively. Below we provide an informal description of the main results which hold with high probability, where $A(n) \sim B(n)$ denotes that $cA(n) \leq B(n) \leq CA(n)$ for large enough n and for some constants $c, C > 0$:

Sparse Regime ($mp \ll \log n$)

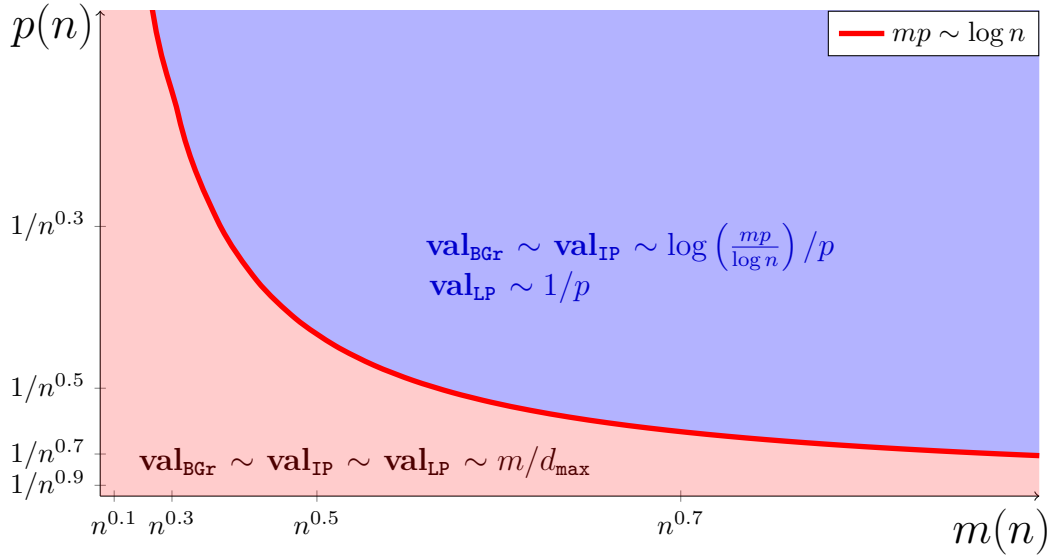
We show that $\text{IPGAP} \sim 1$ in the sparse regime by proving that the BlockGreedy algorithm succeeds in reaching the LP lower bound of $\frac{m}{d_{\max}}$.

$$\text{val}_{\text{BGr}} \sim \text{val}_{\text{IP}} \sim \text{val}_{\text{LP}} \sim \frac{m}{d_{\max}}.$$

Dense Regime ($mp \gg \log n$)

We prove that $\text{IPGAP} \sim \log \frac{mp}{\log n}$ in the dense regime. We show that the BlockGreedy algorithm performs as well as IP in this regime, i.e.

$$\frac{1}{p} \log \left(\frac{mp}{\log n} \right) \sim \text{val}_{\text{BGr}} \sim \text{val}_{\text{IP}} \gg \text{val}_{\text{LP}} \sim \frac{1}{p} \sim \frac{m}{d_{\max}}.$$



■ **Figure 1** Transition between the sparse and the dense regime for different values of the average inclusion set size mp .

Threshold Regime ($mp \sim \log n$)

This regime smoothly interpolates between the sparse and dense ones, with $\text{IPGAP} \sim 1$. The scaling for all quantities of interest is $m/d_{\max} \sim 1/p$.

Greedy

We prove that $\text{val}_{\text{Gr}} \sim \text{val}_{\text{IP}}$ when $\delta < 1/2$, where δ is the parameter from Assumption 2.4.

The results above are also depicted in Figure 1, and the formal statements are given in Corollary 9 and Theorem 10. The rest of the paper is organized as follows. In Section 2, we present relevant notation. In Section 3, we outline and discuss related literature. In Section 4, we prove a number of preliminary results that will be instrumental in developing the core arguments. Subsequently, in Section 5, we delve into the algorithmic aspects of the problem at hand by first providing guarantees for a simple algorithm, **BlockGreedy**. We then analyse **Greedy** by means of a reduction. We conclude in Section 6 by summarizing the results and offering indications for future work. We defer the proofs of more technical results to the appendix, in order to streamline the presentation for the reader's convenience.

2 Notation and conventions

For integers $k \in \mathbb{N}$, we write $[k] := \{1, \dots, k\}$. We denote vectors, matrices by bold-faced Roman letters $\mathbf{x}, \mathbf{A} \in \mathbb{R}^k, \mathbb{R}^{k \times k}$, respectively, for some $k \in \mathbb{N}$. Define the *inclusion set* of an element, or node, $j \in [n]$ as $I_j = \{i \in [m] : j \in \mathcal{S}_i\}$. We denote the ℓ_1 norm of the j -th column of \mathbf{A} by X_j , $j \in [n]$, noting that $X_j = |I_j|$ and $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Binomial}(m, p)$. In addition, we let $d_{\max} \equiv d_{\max}(X_1, \dots, X_n) := \max_{i \in [n]} X_i$. We use \mathbb{E}, Var to denote expectation and variance, respectively. By \lesssim, \gtrsim we denote inequalities up to multiplicative constants. We let $A \sim B$ denote that $A \lesssim B \lesssim A$ for large enough n . We let \log denote the natural logarithm. For possibly random functions $f(n), g(n)$, we let $\{f \lesssim g\}$ denote a sequence of

events $\{f(n) \leq Ag(n)\}$ for some constant $A > 0$ independent of n . Consequently, $\mathbb{P}(f \lesssim g)$ is viewed as a function of n . For deterministic functions $h(n), w(n)$, we let $h \ll w, h \gg w$ denote that $h/w \rightarrow 0, w/h \rightarrow 0$ respectively, as $n \rightarrow \infty$. The notation for other inequalities is defined analogously. We say that a sequence of events $\{A_n\}$ holds *with high probability* (w.h.p.) with respect to a probability measure \mathbb{P} if there exists a constant $c > 0$, independent of n , such that $\mathbb{P}(A_n) \geq 1 - n^{-c}$, for large enough values of n .

3 Related Work

Worst-case analysis of Greedy

Perhaps the most well-known algorithm for solving **Hitting Set**, or equivalently **MVCH**, is the greedy algorithm of Lovász [12], with runtime complexity $O(mn^2)$. This algorithm, which constructs a cover by sequentially adding elements of the ground set which hit the largest number of remaining subsets, was initially studied by Lovász [12] and Johnson [11] independently, for deterministic hypergraphs. Lovász analyses the greedy algorithm to obtain an upper bound on the **Hitting Set** integrality gap of $1 + \log d_{\max}$. Slavik [15] developed the tightest known approximation lower bound for **Greedy**, constructing an instance where **Greedy** finds coverings at least $\log m$ times as large as the minimum one. Importantly, Feige [6] proved that an approximation ratio of $(1 - \epsilon) \log m$ is not achievable in polynomial time for any $\epsilon > 0$ unless $NP \subset TIME[n^{O(\log \log n)}]$, certifying **Greedy** as the best possible polynomial-time approximation algorithm for set cover in the worst-case.

Random Hitting Set

Little is known about the typical performance of polynomial-time algorithms on random instances of **Hitting Set**. Closing this gap is important from a theoretical standpoint and for applications in combinatorial inference. A prime example of this is found in *group testing*, a classical inference problem where one aims to identify a small subset of defective items within a large population by conducting the smallest number of pooled tests, with applications ranging from the analysis of communication protocols [8] to DNA sequencing [5] and search problems [4]. In [10], Iliopoulos and Zadik consider the smallest hitting set as an estimator in the setting of the group testing problem, referring to it as the *Smallest Satisfying Set* estimator. In particular, they provide extensive empirical evidence supporting the claim that the class of instances of the random hitting set problem induced by non-adaptive group testing is tractably solvable by computers.

Insights from Statistical Physics

The analysis of a random instance of **Hitting Set** appears in the work of Mézard and Tarzia and relies on nonrigorous techniques from statistical physics [13]. This work considers regular uniform hypergraphs, where the degree of vertices and the size of edges are fixed and assumed to be constant. Depending on these values, they evidence sharp transitions between three different phases, the so-called replica symmetry, 1-replica symmetry breaking, and full replica symmetry breaking phases, which characterize the complexity of the optimization landscape for this problem in the average case setting.

Fixed p regime

Another instance was studied by Telelis and Zissimopoulos [16] in the setting of random Bernoulli hypergraphs, where elements belong to subsets independently with *fixed* probability $p \in (0, 1)$. Their analysis concerns the asymptotic regime where the size n of the ground set scales to infinity. In this setting, they study the average-case performance of a simple deterministic algorithm which approximates random **Hitting Set** within an *additive* error term at most $o(\log m)$ almost everywhere. This gives an improvement over Lovász's argument in [12] which provides a multiplicative bound. However, the analysis in [16] does not capture the case of sparse hypergraphs, i.e., when $p \rightarrow 0$ as $n \rightarrow \infty$. The analysis in [16] also does not prove guarantees for the **Greedy** algorithm in the chosen parameter regime.

Related problem formulations

We bring to the reader's attention a more recent line of work [2, 3], where the authors obtain bounds on (additive) integrality gaps between the value of a random integer program $\max \mathbf{c}^T \mathbf{x}, \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \{0, 1\}^n$ with m constraints and that of its linear programming relaxation for a wide range of distributions on $(\mathbf{A}, \mathbf{b}, \mathbf{c})$, holding w.h.p. as $n \rightarrow \infty$. These include the case where the entries of \mathbf{A} are uniformly distributed on an integer interval consisting of at least three elements and where the columns of \mathbf{A} are distributed according to an isotropic logconcave distribution. However, these fail to capture the setting where \mathbf{A} is sparse with entries in $\{0, 1\}$, which is of interest for **Hitting Set**.

4 Preliminary Bounds

In this section, we outline preliminary bounds on $\text{val}_{\text{LP}}, \text{val}_{\text{IP}}, d_{\text{max}}$ which will prove crucial to analysing IPGAP and **Greedy**. We begin by characterizing the value of the linear program:

► **Lemma 3.** *There exists $c > 0$, independent of n , such that with probability at least $1 - \exp(-cn^{1-\delta})$, we have that*

$$\frac{m}{d_{\text{max}}} \leq \text{val}_{\text{LP}} \lesssim \frac{1}{p}.$$

The proof is included in Appendix A, and follows from a maximum argument and a standard Chernoff bound. We note that the proof also implies $\mathbb{P}(\text{IP is feasible}) \geq 1 - \exp(-cn^{1-\delta})$. Although Lemma 3 readily yields $\text{val}_{\text{IP}} \geq m/d_{\text{max}}$, we highlight that this lower bound is not tight whenever $mp \gg \log n$. Indeed, we apply the first moment method to obtain a tighter lower bound on val_{IP} in this regime:

► **Lemma 4.** *Let $mp \gg \log n$. For any $D \geq 1$ and n large enough, with probability at least $1 - n^{-D}$ we have that*

$$\frac{1}{p} \log \left(\frac{mp}{\log n} \right) \lesssim \text{val}_{\text{IP}}$$

The proof of Lemma 4 is provided in Appendix A. Lemmas 3 and 4 come short of providing a full characterization of IPGAP, namely lacking an upper bound on val_{IP} . In this light, we turn our attention to the **Greedy** algorithm, and utilize it to construct a feasible integral solution and hence an upper bound on the value of IP. The analysis of **Greedy** crucially relies on characterizing the maximum inclusion set size, $d_{\text{max}} := \max_{j \in [m]} |I_j|$. The following lemma offers such a characterization in expectation, and evidences a key difference between the sparse and dense regimes of our problem:

Algorithm 1 Greedy.

```

1:  $\mathcal{I} \leftarrow \{I_1, \dots, I_n\}$  ▷ Inclusion sets
2:  $U \leftarrow [m]$ 
3:  $t \leftarrow 0$ 
4: while  $|U| > 0$  do
5:    $P \leftarrow \operatorname{argmax}_{I \in \mathcal{I}} |I \cap U|$  ▷ Greedy step
6:    $\mathcal{I} \leftarrow \mathcal{I} \setminus \{P\}$ 
7:    $U \leftarrow U \setminus P$ 
8:    $t \leftarrow t + 1$ 
9:  $\operatorname{val}_{\text{Gr}} \leftarrow t$ 
10: return  $\operatorname{val}_{\text{Gr}}$ .
```

► **Lemma 5** (Maximum of Binomials). *Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bin}(m, p)$. Under the conditions in Assumption 2, it holds that*

$$\mathbb{E}d_{\max} = \mathbb{E} \max_{i \in [n]} X_i \sim \begin{cases} \frac{\log n}{\log(\log n / mp)} & , \text{ if } mp \ll \log n, \\ mp & , \text{ if } mp \gtrsim \log n. \end{cases}$$

The proof of Lemma 5 is provided in Appendix A, and involves a straight forward application of Markov's and Jensen's inequalities. Lemma 5 indicates a sharp transition between two regimes: the sparse regime $mp \ll \log n$, where binomial random variables are known to be well approximated by Poisson random variables, and the dense regime $mp \gg \log n$, where binomial random variables are known to be well approximated by Gaussian random variables. Importantly, in the sparse (Poisson-like) regime, the expected maximum of binomial random variables exceeds their individual expectations: $\mathbb{E}X_1 \ll \mathbb{E}d_{\max}$. Meanwhile in the dense (Gaussian-like) regime, the expected maximum and individual expectations are asymptotically equivalent up to multiplicative constants: $\mathbb{E}X_1 \sim \mathbb{E}d_{\max}$. This fine-grained characterization of the maxima of binomial random variables will prove essential to analysing the behaviour of **BlockGreedy** in Section 5. Finally, we characterize the asymptotic behaviour of d_{\max} and prove that $d_{\max} \lesssim \mathbb{E}d_{\max}$ with high probability. Whilst this one sided result suffices for the forthcoming analysis, we expect a matching lower bound to hold as well. Additional insights into the concentration of d_{\max} may be found in Lemmas 19, 20, in Appendix A.

► **Lemma 6.** *Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bin}(m, p)$. Then, there exist constants $c, \tilde{c} > 0$, independent of n , such that*

$$\mathbb{P} \left(\max_{i \in [n]} X_i \geq c \cdot \mathbb{E} \max_{i \in [n]} X_i \right) \leq \frac{1}{n^{\tilde{c}}}.$$

The proof of Lemma 6 is provided in Appendix A.

5 Algorithmic solutions

5.1 Challenges of Greedy analysis

The aim of the present section is to conduct a rigorous analysis of the standard **Greedy** algorithm for the **Hitting Set** problem, within the prescribed Bernoulli random setting. In particular, we show that this routine succeeds at constructing hitting sets of optimal size w.h.p., as in the results of Section 4, up to multiplicative constants. This is done by first analysing a variation of the greedy heuristic, and subsequently proceeding by a reduction argument.

Algorithm 2 BlockGreedy.

```

1: Let  $\mathcal{B}_t \subset \{I_1, \dots, I_n\}$  denote the  $t$ -th block, i.e. inclusion sets that become available at
   step  $t$ .
2:  $\mathcal{I} \leftarrow \emptyset$ 
3:  $U \leftarrow [m]$ 
4:  $t \leftarrow 0$ 
5: while  $|U| > 0$  and  $\mathcal{B}_t \neq \emptyset$  do
6:    $\mathcal{I} \leftarrow \mathcal{I} \cup \mathcal{B}_t$  ▷ Adding elements from the new block
7:    $P \leftarrow \operatorname{argmax}_{I \in \mathcal{I}} |I \cap U|$  ▷ Greedy step
8:    $\mathcal{I} \leftarrow \mathcal{I} \setminus \{P\}$ 
9:    $U \leftarrow U \setminus P$ 
10:   $t \leftarrow t + 1$ 
11:  $\text{val}_{\text{BGr}} \leftarrow t$ 
12: if  $|U| > 0$  then cover the rest of  $U$  with a trivial algorithm,  $\text{val}_{\text{BGr}} \leftarrow \text{val}_{\text{BGr}} + |U|$ 
13: return  $\text{val}_{\text{BGr}}$ .

```

The core principle of **Greedy** is to construct a feasible solution in steps, by sequentially adding to the candidate solution an element which hits the largest number of remaining sets. In the chosen setting, where elements are added to sets with equal probability and independently of each other, we have precise estimates on the number of subsets hit by an element which is *picked first*. In fact, the size of this set is given by the maximum of independent Binomial random variables, which was analysed in Section 4. However, this very first step introduces nontrivial dependencies amongst the remaining matrix columns and significantly complicates keeping track of the marginal gains of each subsequent element addition to the candidate solution.

5.2 BlockGreedy algorithm

In order to circumvent this issue, we introduce a modified greedy routine, which we refer to as the **BlockGreedy** algorithm, where the elements of the ground set $[n]$ are split into separate sets of a given size, which we call blocks. At the t -th iteration, the algorithm picks the element hitting the largest number of remaining sets across *the first t blocks only*. By choosing the size of the blocks appropriately, we have that at each iteration t one is guaranteed to find a solution of near-optimal size at least within the set of newly-included independent columns.

BlockGreedy is detailed in Algorithm 2, whilst informally, it works as follows.

1. Let K be the size of the solution (suggested by theoretical analysis);
2. Uniformly at random split n columns into K blocks with n/K columns per block;
3. Start with an empty set of possible choices of columns;
4. At the t -th iteration, first add the columns from the t -th block (Step 6). Then, perform one greedy step on the current set of possible choices (Step 7);
5. If after K iterations of the algorithm, some subsets remain uncovered, we use a trivial covering, i.e., covering each subset by a separate column.

Note that the first selection of the element which hits the most number of subsets again introduces dependencies. However, the columns that are in the newly added block are independent of everything else at time t . Let v_t be the element which is picked at the t -th step

of **BlockGreedy**, f_t be the number of new subsets that are hit by v_t^1 , and $F_t := \sum_{i=1}^t f_i$ be the total number of subsets which are hit after t steps. In order to analyse how many elements **BlockGreedy** has picked, we will consider the sequence f_1, f_2, \dots, f_s , with $F_t := \sum_{i=1}^t f_i$, such that the following holds:

1. $F_s = m$;
2. if $mp \lesssim \log n$, then $s \lesssim \text{val}_{\text{LP}}$, otherwise, $s \lesssim \text{val}_{\text{IP}}$.

The first property ensures that **BlockGreedy** picks at most s elements, and the second property gives optimal bounds on s . One way to guarantee that **BlockGreedy** succeeds is to prove that among the choices of **BlockGreedy** at each step t , there was an element \tilde{v}_t which hits at least f_t new subsets w.h.p. We will prove that it is enough to look for \tilde{v}_t in the new block of columns \mathcal{B}_t , which are added at step t . Note that unless $F_t = m$, we have that $f_t \geq 1$, since each subset is hit by at least one element w.h.p.. Therefore, it will be enough to find a sequence $\{f_1, f_2, \dots, f_v\}$ such that $F_v \geq m - v$, since it implies $F_{2v} = m$. This allows us to reduce the problem of proving the effectiveness of **BlockGreedy** to a key technical lemma. This lemma assumes that before step t , exactly F_{t-1} subsets are hit, and bounds from below the probability that some vertex in the new block will hit at least f_t new subsets. This boils down to computing $\mathbb{P}(\text{Bin}(m - F_{t-1}, p) \geq f_t)$.

► **Lemma 7** (Informal, see Lemma 26). *Let $\varepsilon > 0$ and $mp \lesssim \log n$. For some constants $\tau > 0$, $1 < \alpha < \beta$, and for $t \in \mathbb{N}$, let:*

$$f_t = \left\lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \right\rceil \quad \text{where } k \text{ is such that } \beta^{-k-1}m < m - F_{t-1} \leq \beta^{-k}m;$$

Then there exists a choice of τ, α, β and K , such that $F_K \geq m - K$ and $K \sim \text{val}_{\text{LP}}$. Furthermore, for this sequence f_t (which depends on ε), for any $t \leq K$,

$$\mathbb{P}(\text{Bin}(m - F_{t-1}, p) \geq f_t) \geq n^{-\varepsilon}. \tag{5.1}$$

Note that the implicit constants in the statements $K \sim \text{val}_{\text{LP}}$ depend on ε .

This lemma highlights the crucial dependency of the problem on the relationship between the average degree, mp , and $\log n$. For clarity of exposition, we only state the lemma for the case $mp \lesssim \log n$ and refer to the Lemma 26 in the Appendix for the full version and corresponding proof. Here we comment on the intuition behind the proof.

When $mp \lesssim \log n$, we need to carefully track how the maximum degree changes. We look for an element which (i) covers a large number of subsets, i.e., close to the expected maximum number, $\mathbb{E}d_{\max}$ and (ii) can be found with large enough probability. The second property is important for the reduction to the standard **Greedy** algorithm, whose direct analysis presents substantial difficulties, and is done later in this section. The quantity $\mathbb{E}d_{\max}$ is sensitive to mp whenever the latter is close to $\log n$. Hence, we need to adjust which element we look for accordingly. This is done by setting $f_t = \left\lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \right\rceil$ and increasing the parameter k as the number of remaining rows, $m - F_t$, decreases.

For example, consider the case $mp = \log n$. First, we can only pick a random element, since it will be as good (up to a multiplicative constant) as the maximal element. However, during the execution of the algorithm, the problem becomes more sparse, and if we continue to

¹ It may happen that v_t hits *more* than f_t new subsets. In this case, we still only count that exactly f_t are covered, and several extra sets will be covered multiple times in subsequent rounds. This overcounting simplifies the analysis and does not result in suboptimal solution.

30:10 Greedy Heuristics and Linear Relaxations for the Random Hitting Set Problem

pick random elements, we will construct a suboptimal solution. Therefore, we gradually increase how much the newly picked element will cover, *with respect to a random element*. This corresponds to the transition between Gaussian-like and Poisson-like behaviour of $\text{Bin}(m - F_{t-1}, p)$.

It is now straightforward to prove the following theorem, which makes rigorous the statements in Section 1.

► **Theorem 8.** *Under Assumption 2, we have that*

- (i) if $mp \lesssim \log n$ then, for any $\varepsilon > 0$ and n large enough,
- $$\mathbb{P}\left(\text{val}_{\text{BGr}} \lesssim \frac{m}{\mathbb{E}d_{\max}}\right) \geq 1 - \exp(-n^{1-\delta-\varepsilon});$$
- (ii) if $mp \gg \log n$, then, for any $\varepsilon > 0$ and n large enough,
- $$\mathbb{P}\left(\text{val}_{\text{BGr}} \lesssim \frac{1}{p} \log\left(\frac{mp}{\log n}\right)\right) \geq 1 - \exp(-n^{1-\delta-\varepsilon}).$$
- (5.2)

Note that if $mp \gtrsim n^\gamma$ for some $\gamma > 0$, then $\log \frac{mp}{\log n} \sim \log n$, and the bound in (ii) can be simplified.

Proof. The main idea of the proof is to analyse the distribution of the columns that are added at each step t . These columns are independent, and for each newly added column, the number of additional subsets which it covers is distributed according to $\text{Bin}(m - F_{t-1}, p)$, where F_{t-1} is the number of subsets which are already covered. Lemma 26 (see Lemma 7 above for an informal version) allows us to lower bound F_t , and we show now that we can do this with high probability.

Fix $\varepsilon > 0$ and let $\varepsilon' := \varepsilon/4$. Let f_1, f_2, \dots be the sequence from Lemma 26 for ε' and let K be the value for which (C.1) is satisfied, i.e. $F_K \geq m - K$. Notice that $K \leq C \max\left\{\frac{m}{\mathbb{E}d_{\max}}, \frac{1}{p} \log\left(\frac{mp}{\log n}\right)\right\}$ for some constant $C > 0$, for n large enough. We uniformly at random split n elements (columns) into K groups of size n/K each (assuming without loss of generality that K divides n , otherwise we consider groups of size $\lfloor n/K \rfloor$), so that \mathcal{B}_t yields a new set of n/K elements at each iteration $t \leq K$ and $\mathcal{B}_t = \emptyset$ for $t > K$. We say that the algorithm fails at step t if before step t , at least F_{t-1} subsets are covered, but after step t less than F_t sets are covered. Using that, for n large enough, (i) columns in each newly added block are independent, (ii) $\mathbb{P}(\text{Bin}(m - F_{t-1}, p) \geq f_t) \geq n^{-\varepsilon'}$, and (iii) $n/K \geq n^{1-\delta-\varepsilon'}$, we get

$$\begin{aligned} \mathbb{P}(\text{BlockGreedy fails at step } t) &\stackrel{(i)}{\leq} (\mathbb{P}(\text{Bin}(m - F_{t-1}, p) < f_t))^{n/K} \\ &\stackrel{(ii)}{\leq} (1 - n^{-\varepsilon'})^{n/K} \\ &\stackrel{(iii)}{\leq} \exp(-n^{1-\delta-2\varepsilon'}). \end{aligned}$$

We then proceed by applying a union bound to obtain the result,

$$\begin{aligned} &\mathbb{P}(\text{BlockGreedy fails during first } K \text{ steps}) \\ &\leq \sum_{t=1}^K \mathbb{P}(\text{BlockGreedy fails at step } t) \leq K \cdot \exp(-n^{1-\delta-2\varepsilon'}) \leq \exp(-n^{1-\delta-3\varepsilon'}), \end{aligned}$$

where the second inequality holds since, by definition, the algorithm runs for K iterations, and the third one holds for n large enough. We proved that **BlockGreedy** succeeds in finding

at most K elements such that at most $m - F_K$ sets remain uncovered. Since by construction, $m - F_K \leq K$, we can cover the remaining rows trivially using that IP is feasible by Lemma 16 with high probability, which proves that

$$\mathbb{P}(\text{val}_{\text{BGr}} \leq 2K) \geq 1 - \exp\left(-n^{1-\delta-4\varepsilon'}\right) = 1 - \exp\left(-n^{1-\delta-\varepsilon}\right),$$

for n large enough. Recalling that $K \lesssim \text{val}_{\text{LP}}$ for $mp \lesssim \log n$, and that $K \lesssim \text{val}_{\text{IP}}$ for $mp \gg \log n$, finishes the proof. \blacktriangleleft

► **Corollary 9.** *Under Assumption 2, we have that for any $D > 0$,*

- (i) *for any n large enough,*

$$\mathbb{P}(\text{val}_{\text{BGr}} \sim \text{val}_{\text{IP}}) \geq 1 - n^{-D};$$
- (ii) *if $mp \lesssim \log n$, then, for any n large enough,*

$$\mathbb{P}(\text{IPGAP} \sim 1) \geq 1 - n^{-D}; \tag{5.3}$$
- (iii) *if $mp \gg \log n$, then, for any n large enough,*

$$\mathbb{P}\left(\text{IPGAP} \sim \log\left(\frac{mp}{\log n}\right)\right) \geq 1 - n^{-D}.$$

Proof. Proof follows from Lemma 3, Lemma 4, and Theorem 8. \blacktriangleleft

5.3 Reduction from BlockGreedy to Greedy

With the above results at hand, we now proceed to analyse the **Greedy** algorithm by means of a suitable reduction. Recall that we denote outputs of **BlockGreedy** and **Greedy** as val_{BGr} and val_{Gr} respectively.

► **Theorem 10.** *Under Assumption 2 with $\delta < 1/2$, we have that, for n large enough,*

$$\mathbb{P}(\text{val}_{\text{Gr}} \sim \text{val}_{\text{IP}}) \geq 1 - \exp\left(-\sqrt{n}\right).$$

Proof. We use Theorem 8 with $\varepsilon = 1/8 - \delta/4$, and let K, \mathcal{B}_t be as defined in the proof of Theorem 8. We have that, for n large enough,

$$\mathbb{P}(\text{BlockGreedy fails at any step}) \leq \exp\left(-n^\Delta\right),$$

where $\Delta := 3/4 - \delta/2 > 1/2$.

Given a matrix \mathbf{A} , consider running the above definition of **BlockGreedy** for $J := \exp(\sqrt{n})$ times, each time reshuffling the columns. In what follows, we address **BlockGreedy** and **Greedy** defined with the same tie-breaking strategy when it comes to a number of elements hitting the same number of sets, i.e., selecting the left-most column in the associated matrix \mathbf{A} . Both val_{BGr} and val_{Gr} are random variables, but conditioned on \mathbf{A} , val_{Gr} is deterministic, while val_{BGr} still depends on the randomness of separating columns into blocks. Using the union bound, we have that

$$\begin{aligned} \mathbb{P}(\text{val}_{\text{Gr}} > 2K) &\leq \mathbb{P}(\exists \text{ a failed copy of BlockGreedy}) \\ &\quad + \mathbb{P}(\text{val}_{\text{BGr}} < \text{val}_{\text{Gr}} \text{ over all } J \text{ copies}). \end{aligned} \tag{5.4}$$

Applying the union bound again, we can upper bound the first term in (5.4):

$$\mathbb{P}(\exists \text{ a failed copy of BlockGreedy}) \leq J \exp\left(-n^\Delta\right) = \exp\left(-n^\Delta + n^{1/2}\right). \tag{5.5}$$

30:12 Greedy Heuristics and Linear Relaxations for the Random Hitting Set Problem

Now we focus on the second term in (5.4). Let v_1, v_2, \dots, v_g be the ordered sequence of elements picked by **Greedy**. Let $M_t := \{v_1 \in \mathcal{B}_1, v_2 \in \mathcal{B}_1 \cup \mathcal{B}_2, \dots, v_t \in \mathcal{B}_1 \cup \dots \cup \mathcal{B}_t\}$. The event $\{\text{val}_{\text{BGr}} \geq \text{val}_{\text{Gr}}\}$ contains the event M_g , since in this case **BlockGreedy** will necessarily pick exactly the same columns v_1, v_2, \dots, v_g . Given that each reshuffling of the columns generates a uniform distribution of \mathcal{B}_i 's over possible partitions of n columns, we get that

$$\mathbb{P}(M_g) = \mathbb{P}(v_1 \in \mathcal{B}_1) \mathbb{P}(v_2 \in \mathcal{B}_1 \cup \mathcal{B}_2 \mid M_1) \dots \mathbb{P}(v_g \in \mathcal{B}_1 \cup \dots \cup \mathcal{B}_g \mid M_{g-1}).$$

The t -th term in the product above is equal to

$$\mathbb{P}(v_t \in \mathcal{B}_1 \cup \dots \cup \mathcal{B}_t \mid M_{t-1}) = \frac{t \frac{n}{K} - (t-1)}{n - (t-1)} \geq \frac{t}{K} - \frac{t-1}{n} \geq \frac{t}{2(K-1)},$$

where the last inequality holds for $n \geq 4K$ (recall that $n \gg K$). Since $M_g \subset \{\text{val}_{\text{BGr}} \geq \text{val}_{\text{Gr}}\}$, we can lower bound the probability of the latter event as follows (note that when $g < K$ there will be less terms in the product, hence, $\mathbb{P}(M_g)$ will be even larger),

$$\begin{aligned} \mathbb{P}(\text{val}_{\text{BGr}} \geq \text{val}_{\text{Gr}} \text{ for 1 copy}) &\geq \mathbb{P}(M_g) \\ &\geq \prod_{t=1}^{K-1} \mathbb{P}(v_t \in \mathcal{B}_1 \cup \dots \cup \mathcal{B}_t \mid M_{t-1}) \geq \prod_{t=1}^{K-1} \frac{t}{2(K-1)} \geq e^{-2K}, \end{aligned}$$

where we used that $k! \geq (k/e)^k$ in the last inequality. Since $K \leq C \max\left\{\frac{m}{\mathbb{E}d_{\max}}, \frac{1}{p} \log\left(\frac{mp}{\log n}\right)\right\}$ and $1/p \leq n^\delta$, there exists a constant $\tilde{C} > 0$ large enough, such that $K \leq \tilde{C} n^\delta \log n$. Therefore, using independence of the reshuffling between the copies, we can compute

$$\begin{aligned} \mathbb{P}(\text{val}_{\text{BGr}} < \text{val}_{\text{Gr}} \text{ over all } J \text{ copies}) &= (1 - \mathbb{P}(\text{val}_{\text{BGr}} \geq \text{val}_{\text{Gr}} \text{ for 1 copy}))^J \\ &\leq (1 - e^{-2K})^J \\ &\leq \exp\left(-e^{\sqrt{n} - 2\tilde{C}n^\delta \log n}\right). \end{aligned} \tag{5.6}$$

Combining (5.4), (5.5) and (5.6), we showed that $\mathbb{P}(\text{val}_{\text{Gr}} > 2K) \leq \exp(-\sqrt{n})$ for n large enough, which finishes the proof. \blacktriangleleft

► **Remark 11.** We note that the $\delta < 1/2$ condition in Theorem 10 is likely not optimal, and could be relaxed by reducing to **BlockGreedy** with more carefully chosen sets \mathcal{B}_t . In particular, the appropriate set sizes $|\mathcal{B}_t|$ may not be identical across $t \leq K$. The analysis becomes more technical in this case, and we highlight this as an interesting open direction.

6 Discussion and Open Questions

Our work characterises multiplicative integrality gaps for the random hitting set problem. In this section, we discuss the intuition behind our main results, together with open questions and conjectures.

6.1 Summary of our results and proof techniques

We identified that the nature of integrality gaps depends on the size of the inclusion set, also viewed as the sparsity of the underlying hypergraph. In particular, when the average degree of a vertex is small, i.e., when each element belongs to a small number of subsets, we proved that there exists only a constant gap between linear and integer program solutions, together with a simple algorithmic solution. The situation changes when the hypergraph

becomes dense, where we show an increasing integrality gap. This separation stems mostly from the property of the binomial distribution, where the maximum of random variables grows identically to the expected value whenever the expected value is large, but is away from it if $mp \ll \log n$.

In our analysis of **BlockGreedy**, we track this change of behaviour using a geometric series, which means that the further we are in the execution of the algorithm, the larger the ratio between the element we pick and the average element will be. This picture coincides exactly with how the binomial distribution will behave if we decrease the average degree: for large instances, it will look approximately as a Gaussian, but when the average degree is small, Poisson approximation starts to dominate, the right tail becomes heavier, and the difference between d_{\max} and mp increases. Our analysis tracks the transition between Gaussian and Poisson-like behavior.

6.2 Multiplicative vs. additive integrality gaps

Our result only concerns multiplicative gaps, but the constants in our analysis can be large. This might be a consequence of the generality of the studied problem. For example, if one focuses only on the case of constant p , which immediately implies a very dense instance in our characterization, [16] proves that a simple algorithm is optimal for approximating the integer program up to a small additive error. Proving similar upper bounds on the constant in more general cases is an interesting open problem. Based on numerical experiments, we formulate the following conjectures.

► **Conjecture 12** (Very sparse). For $mp \ll 1$, $\frac{\text{val}_{\text{Gr}}}{\text{val}_{\text{LP}}} \rightarrow 1$.

► **Conjecture 13** (Sparse). For $1 \lesssim mp \ll \log n$, $\frac{\text{val}_{\text{Gr}}}{\text{val}_{\text{IP}}} \rightarrow 1$, and $\frac{\text{val}_{\text{IP}}}{\text{val}_{\text{LP}}} \rightarrow C_1 \in (1, 1.5)$.

► **Conjecture 14** (Dense). For $mp \gg \log n$, $\frac{\text{val}_{\text{Gr}}}{\text{val}_{\text{IP}}} \rightarrow C_2 \in (1, 1.5)$.

6.3 Analysis of a linear program solution.

One motivation for studying the gaps between the integer and linear programs together with the solutions of linear programs themselves is to construct a rounding scheme which converts a fractional solution to an integer one. We believe this is another interesting direction for future work. In particular, numerical experiments show that entries which have large value in the fractional solution have a strong tendency to correspond to elements that are picked for the integer solution. This supports the claim that a combination of the greedy and linear programming approach might be fruitful in efficiently solving **Hitting Set**. One approach for further study consists of first solving a linear program, initializing \mathbf{x} with the largest elements in the linear solution, and greedily covering the remaining subsets.

References

- 1 Gabriel Arpino, Daniil Dmitriev, and Nicolo Grometto. Greedy heuristics and linear relaxations for the random hitting set problem, 2023. [arXiv:2305.05565](https://arxiv.org/abs/2305.05565).
- 2 Sander Borst, Daniel Dadush, Sophie Huiberts, and Samarth Tiwari. On the integrality gap of binary integer programs with gaussian data. *Mathematical Programming*, 2022. doi:10.1007/s10107-022-01828-1.
- 3 Sander Borst, Daniel Dadush, and Dan Mikulincer. Integrality gaps for random integer programs via discrepancy. In *Proceedings of the 2023 ACM-SIAM Symposium on Discrete Algorithms, SODA*, 2023. doi:10.1137/1.9781611977554.ch65.

- 4 Dingzhu Du, Frank K Hwang, and Frank Hwang. *Combinatorial group testing and its applications*. World Scientific, 2000. doi:10.1142/4252.
- 5 Yaniv Erlich, Anna Gilbert, Hung Ngo, Atri Rudra, Nicolas Thierry-Mieg, Mary Wootters, Dina Zielinski, and Or Zuk. Biological screens from linear codes: theory and tools. *BioRxiv*, 2015. doi:10.1101/035352.
- 6 Uriel Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM (JACM)*, 1998. doi:10.1145/285055.285059.
- 7 William Feller and Philip M Morse. *An introduction to probability theory and its applications*. American Institute of Physics, 1958. doi:10.1063/1.3062516.
- 8 Antonio Fernández Anta, Miguel A Mosteiro, and Jorge Ramón Muñoz. Unbounded contention resolution in multiple-access channels. *Algorithmica*, 2013. doi:10.1007/s00453-013-9816-x.
- 9 Abdolhossein Hoorfar and Mehdi Hassani. Inequalities on the lambert w function and hyperpower function. *Journal of Inequalities in Pure and Applied Mathematics*, 2008. URL: <https://arxiv.org/abs/2305.05565>.
- 10 Fotis Iliopoulos and Ilias Zadik. Group testing and local search: is there a computational-statistical gap? In *Conference on Learning Theory*. PMLR, 2021. URL: <https://proceedings.mlr.press/v134/iliopoulos21a.html>.
- 11 David S Johnson. Approximation algorithms for combinatorial problems. In *Proceedings of the fifth annual ACM symposium on Theory of computing*, 1973. doi:10.1145/800125.804034.
- 12 László Lovász. On the ratio of optimal integral and fractional covers. *Discrete mathematics*, 1975. doi:10.1016/0012-365X(75)90058-8.
- 13 Marc Mézard and Marco Tarzia. Statistical mechanics of the hitting set problem. *Physical Review E*, 2007. doi:10.1103/PhysRevE.76.041124.
- 14 Vangelis T Paschos. A survey of approximately optimal solutions to some covering and packing problems. *ACM Computing Surveys (CSUR)*, 1997. doi:10.1145/254180.254190.
- 15 Petr Slavík. A tight analysis of the greedy algorithm for set cover. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, 1996. doi:10.1145/237814.237991.
- 16 Orestis A Telelis and Vassilis Zissimopoulos. Absolute $O(\log m)$ error in approximating random set covering: an average case analysis. *Information Processing Letters*, 2005. doi:10.1016/j.ipl.2005.02.009.
- 17 Ramon Van Handel. Probability in high dimension. Lecture notes, 2014. URL: <https://api.semanticscholar.org/CorpusID:124828412>.

A Auxiliary lemmas

► **Lemma 15** (Lower Bound in Lemma 3). *We have that*

$$\text{val}_{LP} \geq \frac{m}{d_{\max}}.$$

Proof. Let $\mathbf{x}_{LP}^* = (\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_m^*)$ be an optimal solution for (1.2). Since $A\mathbf{x}_{LP}^* \geq \mathbf{1}$ entrywise, by summing all entries we obtain that

$$m \leq \sum_i \mathbf{x}_i^* X_i \leq d_{\max} \sum_i \mathbf{x}_i^* = d_{\max} \text{val}_{LP}.$$

which upon rearranging yields the desired result. ◀

In addition to the above, we have the following elementary upper bound on val_{LP} , which holds both in the sparse and dense regime.

► **Lemma 16** (Upper Bound in Lemma 3). *There exists $c > 0$, independent of n , such that*

$$\mathbb{P}\left(\text{val}_{LP} \lesssim \frac{1}{p}\right) \geq 1 - \exp(-cn^{1-\delta}).$$

This also implies that $\mathbb{P}(\text{IP is feasible}) \geq 1 - \exp(-cn^{1-\delta})$.

Proof. Consider the candidate feasible solution $\hat{\mathbf{x}} := \frac{1}{\tilde{C}np} \mathbf{1}$, for some constant $0 < \tilde{C} < 1$. The following results from applying a union bound over constraints and the standard Chernoff bound.

$$\begin{aligned} \mathbb{P}(\hat{\mathbf{x}} \text{ not feasible}) &= \mathbb{P}(\exists i \in [m] : (\mathbf{A}\hat{\mathbf{x}})_i < 1) \\ &\leq m\mathbb{P}(\text{Bin}(n, p) < \tilde{C}np) \\ &\leq n^C \exp\left(-\frac{(1-\tilde{C})^2 np}{2}\right) \\ &\leq \exp(-cn^{1-\delta}). \end{aligned}$$

The desired conclusion follows by considering the complementary event to the one above and noting that $\|\hat{\mathbf{x}}\|_1 \sim 1/p$. Note that the event $\{\hat{\mathbf{x}} \text{ is feasible for LP}\}$ implies the event $\{\text{IP is feasible}\}$. \blacktriangleleft

► **Lemma 17** (Lambert W function, [9]). *For any $x \geq e$, there holds that*

$$\log x - \log \log x + \frac{\log \log x}{2 \log x} \leq W_0(x) \leq \log x - \log \log x + \frac{e}{e-1} \frac{\log \log x}{\log x}. \quad (\text{A.1})$$

In particular,

$$W_0(x) = \log x - \log \log x + o(1), \quad \text{as } x \rightarrow \infty. \quad (\text{A.2})$$

In addition, for any $x \geq 1/e$, the following identity is satisfied

$$W_0(x) = \log \frac{x}{W_0(x)}. \quad (\text{A.3})$$

Proof of Lemma 4. Fix $D \geq 1$. Let $Z_k := |\{\mathbf{x} \in \{0, 1\}^m : \mathbf{A}\mathbf{x} \geq \mathbf{1}, \|\mathbf{x}\|_1 = k\}|$ be the number of feasible solutions of norm exactly k . Clearly, $Z_k \leq Z_{k+1}$ for any $k \geq 0$. We also have that

$$\mathbb{E}Z_k = \sum_{\|\mathbf{x}\|=k} \mathbb{P}((\mathbf{A}\mathbf{x})_i \geq 1, \forall i \in [m]) = \binom{n}{k} (1 - (1-p)^k)^m.$$

We will now show that for $k \ll \frac{1}{p} \log\left(\frac{mp}{\log n}\right)$, we have $\mathbb{E}Z_k \leq n^{-D}$. Using that $p \leq 1/2$ from Assumption 4 and that for $x \in (0, \frac{1}{2})$, we have $(1-x)^y \geq e^{-2xy}$, we can bound

$$\begin{aligned} \mathbb{E}Z_k &= \binom{n}{k} (1 - (1-p)^k)^m \leq n^k (1 - e^{-2pk})^m \\ &\leq n^k e^{-me^{-2pk}} = \exp\{k \log n - me^{-2pk}\}. \end{aligned}$$

Therefore, $\mathbb{E}Z_k \leq n^{-D}$ will follow from

$$2pke^{2pk} \leq -2Dpe^{2pk} + \frac{2mp}{\log n}. \quad (\text{A.4})$$

Since $k \ll \frac{1}{p} \log\left(\frac{mp}{\log n}\right)$, we also have that $k \leq k_* := \frac{1}{2p} W_0\left(\frac{mp}{D \log n}\right)$ for n large enough. For $k = k_*$, the left hand side of (A.4) is equal to $\frac{mp}{D \log n}$, while the right hand side is lower bounded by $\frac{mp}{\log n}$. Since $D \geq 1$, we recover that $\mathbb{E}Z_k \leq n^{-D}$. Note that for n large enough, $\text{val}_{\text{IP}} \ll \frac{1}{p} \log \frac{mp}{\log n}$ implies that $Z_{k_*} > 0$. Therefore, applying Markov's inequality, we get that

$$\mathbb{P}\left(\text{val}_{\text{IP}} \ll \frac{1}{p} \log\left(\frac{mp}{\log n}\right)\right) \leq \mathbb{P}(Z_{k_*} > 0) \leq \mathbb{E}Z_{k_*} \leq n^{-D}, \quad (\text{A.5})$$

30:16 Greedy Heuristics and Linear Relaxations for the Random Hitting Set Problem

and the proof follows by considering the complementary events. Note that using similar derivations, one can also show that for $k^* := \frac{1}{p} \log \left(\frac{1}{\delta} \frac{mp}{\log n} \right)$, where δ is defined in Assumption 4, we have $\mathbb{E}Z_{k^*} \geq 1$. \blacktriangleleft

Proof of Lemma 5. For ease of notation, let us define $b_n := \frac{\log n}{mp}$, $b_n^* := \frac{1}{e} (b_n - 1)$, $g_n := \frac{\log n}{\log(\log n/mp)}$. We begin by proving the desired upper bound. By Jensen's inequality and bounding the maximum of positive values by their sum, for any $\lambda > 0$, we obtain

$$\mathbb{E} \max_{i \in [n]} X_i \leq \frac{1}{\lambda} \log \mathbb{E} \exp \left(\lambda \max_{i \in [n]} X_i \right) = \frac{1}{\lambda} \log \mathbb{E} \left(\max_{i \in [n]} \exp(\lambda X_i) \right) \leq \frac{1}{\lambda} \log \sum_{i \in [n]} \mathbb{E} \exp(\lambda X_i).$$

Finally, computing the moment generating function of binomial random variables, together with the inequality $1 - x \leq e^{-x}$ yields

$$\mathbb{E} \max_{i \in [n]} X_i = \frac{\log n + m \log(1 - p(1 - e^\lambda))}{\lambda} \leq \frac{\log n - mp(1 - e^\lambda)}{\lambda}.$$

In the regime where $mp \gtrsim \log n$, we may choose $\lambda > 0$ arbitrary, independent of n , from which it immediately follows that $\mathbb{E} \max_{i \in [n]} X_i \lesssim mp$.

For $mp \ll \log n$, we proceed by differentiating the last line in the above display and setting the resulting expression to zero. From this, we may choose λ as the solution of the following.

$$e^{\lambda-1} (\lambda - 1) = b_n^*$$

Under the present assumptions, this is expressed in terms of the Lambert W function as $\lambda = 1 + W_0(b_n^*)$, so that by (A.3), we obtain

$$\mathbb{E} \max_{i \in [n]} X_i \leq \frac{\log n \left(1 - \frac{1}{b_n} + \frac{b_n^*}{b_n} \frac{e}{W_0(b_n^*)} \right)}{1 + W_0(b_n^*)} \sim g_n.$$

In the dense $mp \gtrsim \log n$ regime, a matching lower bound is easily obtained by noting that $\mathbb{E} \max_{i \in [n]} X_i \geq \mathbb{E} X_1 = mp$.

To deal with the sparse regime, let $\tau = 1/16$. From Markov's inequality,

$$\mathbb{E} \max_{i \in [n]} X_i \geq \tau g_n \mathbb{P} \left(\max_{i \in [n]} X_i = \lceil \tau g_n \rceil \right) = \tau g_n (1 - (1 - \mathbb{P}(X_1 = \lceil \tau g_n \rceil))^n).$$

Hence, applying Lemma 25, for n large enough,

$$\mathbb{E} \max_{i \in [n]} X_i \geq \tau g_n \left(1 - \left(1 - n^{-1/2} \right)^n \right) \geq (\tau/2) g_n,$$

thus providing a matching lower bound for the sparse regime.

In the intermediate threshold regime $mp \sim \log n$, the average and maximum of X_i 's become of the same order, that is $mp \sim \mathbb{E} d_{\max} \sim \log n$. The smooth transition follows by noting that in this regime, $b_n, b_n^*, W_0(b_n^*) \sim 1$. \blacktriangleleft

► Lemma 18 (Chernoff Bound - upper tail). *Let X_1, \dots, X_n be independent random variables taking values in $\{0, 1\}$, X denote their sum and $\mu = \mathbb{E}X$. Then for any $\delta > 0$,*

$$\mathbb{P}(X \geq (1 + \delta)\mu) \leq e^{-\delta^2 \mu / (2 + \delta)}.$$

In order to deal with concentration of d_{\max} around its expectation, we state the following useful result on tensorization of variance. We introduce notation Var_i and \mathbb{E}_i , where subscript i indicates conditioning on each component of an underlying random vector, except for the i -th one.

► **Lemma 19** (Theorem 2.3, [17]). *Let X_1, \dots, X_n be independent random variables and for each function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, define*

$$\text{Var}_i f(x_1, \dots, x_n) := \text{Var}(x_1, \dots, x_{i-1}, X_i, x_{i+1}, \dots, x_n).$$

Then, there holds that

$$\text{Var}(f(X_1, \dots, X_n)) \leq \mathbb{E} \sum_{i=1}^n \text{Var}_i f(X_1, \dots, X_n)$$

► **Lemma 20** (Concentration for d_{\max}). *Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bin}(m, p)$. Then, for any $t > 0$,*

$$\mathbb{P}(|d_{\max} - \mathbb{E}d_{\max}| > t) \leq \frac{mp}{t^2}.$$

► **Remark 21.** Note that in all regimes of m, p satisfying Assumption 2, choosing $t \sim \mathbb{E}d_{\max}$ is sufficient to deduce from the previous lemma that $d_{\max} \sim \mathbb{E}d_{\max}$ w.h.p..

Proof. Proceeding by Chebyschev's inequality, it suffices to show that $\text{Var}(d_{\max}) \leq mp$. By Lemma 19, we have that

$$\begin{aligned} \text{Var}(d_{\max}) &\leq \mathbb{E} \sum_{i=1}^n \mathbb{E}_i (d_{\max} - \mathbb{E}_i d_{\max})^2 \\ &= \mathbb{E} \sum_{i=1}^n \mathbb{E}_i \left[(d_{\max} - \mathbb{E}_i d_{\max})^2 \mid d_{\max} = X_i \right] \mathbb{P}d_{\max} = X_i \\ &\quad + \mathbb{E} \sum_{i=1}^n \mathbb{E}_i \left[(d_{\max} - \mathbb{E}_i d_{\max})^2 \mid d_{\max} \neq X_i \right] \mathbb{P}d_{\max} \neq X_i \\ &= \frac{1}{n} \mathbb{E} \sum_{i=1}^n \text{Var} X_i \\ &\leq mp, \end{aligned}$$

which is as required. ◀

Proof of Lemma 6. Let us consider the sparse and dense regimes separately.

In the dense regime for $mp \gtrsim \log n$, there exist constants $c_1, c_2, c_3 > 0$ such that $c_1 mp \leq \mathbb{E} \max_{i \in [n]} X_i \leq c_2 mp$, as argued in Lemma 5, and $mp \geq c_3 \log n$. We apply the union and Chernoff bounds as in Lemma 18 to obtain, for any $t \geq 1/c_1$,

$$\begin{aligned} \mathbb{P} \left(\max_{i \in [n]} X_i \geq t \cdot \mathbb{E} \max_{i \in [n]} X_i \right) &\leq n \mathbb{P}(X_1 \geq tc_1 mp) \\ &\leq n \exp \left(- \frac{(tc_1 - 1)^2 mp}{1 + tc_1} \right) \\ &\leq n \exp \left(- \frac{c_3 (tc_1 - 1)^2 \log n}{1 + tc_1} \right). \end{aligned}$$

It now suffices to choose t as a function of c_1, c_3 such that $\frac{c_3 (tc_1 - 1)^2}{1 + tc_1} > 1$. By rearranging and solving the resulting quadratic equation, it follows immediately that any $t > \frac{1}{c_1} + \frac{1 + \sqrt{1 + 8c_3}}{2c_3 c_1} > \frac{1}{c_1}$ suffices. Hence, there exist universal constants c, \bar{c} , such that the desired conclusion holds. We now consider the sparse regime $mp \ll \log n$, where by Lemma 5 there exists $c_4 > 0$ such that $mp \leq c_4 \log n / \log \left(\frac{\log n}{\log mp} \right)$. Notice that for any $\lambda > 0$, $\max_{i \in [n]} X_i \leq \frac{1}{\lambda} \log \sum_{i=1}^n e^{\lambda X_i}$. We apply Markov's inequality to obtain, for any $t > 0$,

$$\begin{aligned}
 \mathbb{P}\left(\max_{i \in [n]} X_i \geq t \cdot \mathbb{E} \max_{i \in [n]} X_i\right) &\leq \mathbb{P}\left(\sum_{i=1}^n e^{\lambda X_i} \geq e^{\lambda t \mathbb{E} \max_{i \in [n]} X_i}\right) \\
 &\leq \frac{n \mathbb{E} e^{\lambda X_1}}{\exp(\lambda t \mathbb{E} \max_{i \in [n]} X_i)} \\
 &= \frac{n(1-p+pe^\lambda)^m}{\exp(\lambda t \mathbb{E} \max_{i \in [n]} X_i)} \\
 &\leq \exp\left(\log n + mp(e^\lambda - 1) - \frac{\lambda t c_4 \log n}{\log\left(\frac{\log n}{mp}\right)}\right),
 \end{aligned}$$

where we used that $1+x < e^x$ to obtain the last inequality. Finally, by choosing $t = 3/c_4$ and $\lambda = \log(\log n/mp)$, we obtain

$$\mathbb{P}\left(\max_{i \in [n]} X_i \geq \frac{3}{c_4} \cdot \mathbb{E} \max_{i \in [n]} X_i\right) \leq \frac{1}{n}. \quad \blacktriangleleft$$

► **Lemma 22** (Asymptotic expression for binomial probability mass function).

Let $a \equiv a(n)$ and $b \equiv b(n)$ be such that

1. $1 \ll b \ll \sqrt{a}$,
2. $p \ll 1$.

If $b \geq Cap$ for $C > 1$, then

$$\log \mathbb{P}(\text{Bin}(\lceil a \rceil, p) = \lceil b \rceil) \geq -\left(b \log \frac{b}{ap} - b + ap\right)(1 + o(1)), \quad (\text{A.6})$$

If also $b \gg ap$, we have that

$$\log \mathbb{P}(\text{Bin}(\lceil a \rceil, p) = \lceil b \rceil) \geq -\left(b \log \frac{b}{ap}\right)(1 + o(1)), \quad (\text{A.7})$$

Furthermore, all bounds remain valid upon replacing $\lceil a \rceil$ to $\lfloor a \rfloor$.

Proof. We defer the proof of Lemma 22 to the extended version of this work found in [1]. ◀

► **Lemma 23** (Binomial Monotonicity). Let $S_m \sim \text{Bin}(m, p)$. Then for $r \geq mp$, we have that $\mathbb{P}(S_m = r+1) \leq \mathbb{P}(S_m = r)$ and $\mathbb{P}(S_{m-1} = r) \leq \mathbb{P}(S_m = r)$.

Proof. The proof follows a similar argument as that presented in [7].

$$\begin{aligned}
 \frac{\mathbb{P}(S_m = r+1)}{\mathbb{P}(S_m = r)} &= \frac{\binom{m}{r+1} p^{r+1} (1-p)^{m-r-1}}{\binom{m}{r} p^r (1-p)^{m-r}} \\
 &= \frac{\frac{m!}{(r+1)!(m-r-1)!} p^{r+1} (1-p)^{m-r-1}}{\frac{m!}{r!(m-r)!} p^r (1-p)^{m-r}} \\
 &= \frac{(m-r)p}{(r+1)(1-p)} \leq 1.
 \end{aligned}$$

Similar arguments show that $\mathbb{P}(S_{m-1} = r) \leq \mathbb{P}(S_m = r)$. ◀

B Main tool for the case $mp \lesssim \log n$ and Proof of Lemma 25

► **Lemma 24.** *If $mp \lesssim \log n$, then, for any $\varepsilon > 0$, there exist constants $\tau > 0$ and $1 < \alpha < \beta$, such that, for $k \lesssim \log n$ and for any \tilde{m} , satisfying $\beta^{-k-1}m \leq \tilde{m} \leq \beta^{-k}m$, for all n large enough,*

$$\mathbb{P}\left(\text{Bin}(\tilde{m}, p) = \lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \rceil\right) \geq n^{-\varepsilon}.$$

Proof. The proof is essentially a careful application of Lemma 22. Let τ, α, β be constants to be fixed later and $\tilde{m} = \lfloor \beta^{-k-1}m \rfloor$. Depending on whether we have $mp \ll \log n$ or $mp \sim \log n$, different terms will dominate the asymptotic expression from Lemma 22.

We start with the case $mp \ll \log n$. From Lemma 5, this implies that $mp \ll \mathbb{E}d_{\max} \ll \log n$. Here we can fix $\alpha \equiv 2$ and $\beta \equiv 3$. Applying (A.7) for $a = 3^{-k-1}m$ and $b = (2/3)^k \tau \mathbb{E}d_{\max}$, we have:

$$\log \mathbb{P}\left(\text{Bin}(\tilde{m}, p) = \lceil (2/3)^k \tau \mathbb{E}d_{\max} \rceil\right) \geq -(2/3)^k \tau \mathbb{E}d_{\max} \log\left(\frac{2^k 3 \tau \mathbb{E}d_{\max}}{mp}\right) (1 + o(1)) \quad (\text{B.1})$$

Recall that our goal is to show $\log \mathbb{P}\left(\text{Bin}(\tilde{m}, p) = \lceil (2/3)^k \tau \mathbb{E}d_{\max} \rceil\right) \geq -\varepsilon \log n$. We first show that there exists $\tau > 0$ satisfying the following two inequalities:

$$\begin{aligned} \text{(i)} \quad & (2/3)^k \tau (\log 3 + k \log 2) \frac{\mathbb{E}d_{\max}}{\log n} \leq \frac{\varepsilon}{4}, \\ \text{(ii)} \quad & (2/3)^k \tau \frac{\mathbb{E}d_{\max}}{\log n} \log\left(\frac{\mathbb{E}d_{\max}}{mp}\right) \leq \frac{\varepsilon}{4}. \end{aligned} \quad (\text{B.2})$$

Indeed, since $\mathbb{E}d_{\max} \ll \log n$ and $k \ll (3/2)^k$, inequality (i) will be satisfied for any $\tau > 0$ for n large enough. For (ii) we need to use explicit bound for $\mathbb{E}d_{\max}$, in particular from Lemma 5 we know that there exists $C > 0$, such that $\mathbb{E}d_{\max} \leq C \log n / (\log \log n - \log mp)$ for n large enough. Plugging this into (ii), we get for $k = 0$,

$$\tau \frac{\mathbb{E}d_{\max}}{\log n} \log\left(\frac{\mathbb{E}d_{\max}}{mp}\right) \leq \frac{\tau C (\log C + \log \log n - \log(\log \log n - \log mp) - \log mp)}{\log \log n - \log mp} = \tau C + o(1). \quad (\text{B.3})$$

For $\tau = \varepsilon/(8C)$, (ii) holds for $k = 0$ for n large enough. By increasing k we only decrease left hand side of (ii), therefore, the same value of τ works for any $k \geq 0$.

Finally, by adding (i) and (ii) we showed that, for n large enough,

$$\log \mathbb{P}\left(\text{Bin}(\tilde{m}, p) = \lceil (\alpha/2)^k \tau \mathbb{E}d_{\max} \rceil\right) \geq -\frac{\varepsilon}{2} \log n (1 + o(1)) > -\varepsilon \log n,$$

which finishes the proof for the case $mp \ll \log n$.

Now we focus on the case $mp \sim \log n$. Here we apply (A.6) for the values $a = \beta^{-k-1}m$ and $b = (\alpha/\beta)^k \tau \mathbb{E}d_{\max}$ keeping in mind the condition $b \geq Cap$ with $C > 1$. We have

$$\begin{aligned} & \log \mathbb{P}\left(\text{Bin}(\tilde{m}, p) = \lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \rceil\right) \\ & \geq -\left((\alpha/\beta)^k \tau \mathbb{E}d_{\max} \log\left(\frac{\beta \alpha^k \tau \mathbb{E}d_{\max}}{mp}\right) - (\alpha/\beta)^k \tau \mathbb{E}d_{\max} + \beta^{-k-1}mp\right) (1 + o(1)) \end{aligned}$$

We pick $\tau = \gamma mp / \mathbb{E}d_{\max}$, for some constant $\gamma > 1$ to be specified later. Note that this way condition for applying (A.6), $\frac{b}{ap} \geq C > 1$, is satisfied since $\frac{b}{ap} \geq \frac{\tau \mathbb{E}d_{\max}}{mp} = \gamma > 1$. This simplifies the latter expression to the following:

$$\begin{aligned} & \log \mathbb{P} \left(\text{Bin}(\tilde{m}, p) = \lceil (\alpha/\beta)^k \gamma mp \rceil \right) \\ & \geq -mp \left((\alpha/\beta)^k \gamma \log(\beta \gamma \alpha^k) - (\alpha/\beta)^k \gamma + \beta^{-k-1} \right) (1 + o(1)) \end{aligned}$$

Since in this regime we have $mp \leq D \log n$ for some $D > 0$, for n large enough, it is enough to show

$$(\alpha/\beta)^k \gamma \log(\beta \gamma \alpha^k) - (\alpha/\beta)^k \gamma + \beta^{-k-1} \leq \varepsilon/(2D).$$

We first show that there exist constants $1 < \alpha < \beta$ and $\gamma > 1$, depending on ε and D , satisfying the following two inequalities for any $k \geq 0$:

$$\begin{aligned} \text{(i)} \quad & (\alpha/\beta)^k \left(\gamma \log \beta \gamma - \gamma + \frac{1}{\alpha^k \beta} \right) \leq \frac{\varepsilon}{4D}, \\ \text{(ii)} \quad & (\alpha/\beta)^k k \log \alpha \leq \frac{\varepsilon}{4D}. \end{aligned}$$

Note that left hand side of (i) decreases as k increases, therefore, it is enough to look at $k = 0$. We need to show that there exist $\beta, \gamma > 1$, depending on ε, D such that

$$f(\beta, \gamma) := \gamma \log \beta \gamma - \gamma + \frac{1}{\beta} \leq \frac{\varepsilon}{4D}.$$

Note that $\frac{\partial f}{\partial \beta} = \gamma/\beta - 1/\beta^2 > 0$ and $\frac{\partial f}{\partial \gamma} = \log \beta \gamma > 0$ as long as $\beta \gamma > 1$. Since $f(1, 1) = 0$, we can find $\beta, \gamma > 1$, close enough to 1, such that $f(\beta, \gamma) \leq \varepsilon/(4D)$. We use these values of β and γ (or, equivalently, τ). Since $k \ll (\beta/\alpha)^k$, there exists $\alpha \in (1, \beta)$, such that (ii) holds. Summing (i) and (ii) shows that, for n large enough,

$$\log \mathbb{P} \left(\text{Bin}(\tilde{m}, p) = \lceil (\alpha/\beta)^k \gamma mp \rceil \right) \geq -\frac{\varepsilon mp}{2D} (1 + o(1)) \geq -\frac{\varepsilon \log n}{2} (1 + o(1)) \geq -\varepsilon \log n.$$

We proved that for $mp \lesssim \log n$, for any $\varepsilon > 0$, for n large enough, there exists τ, α, β , such that

$$\Pr \left(\text{Bin}(\lfloor \beta^{-k-1} m \rfloor, p) = \lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \rceil \right) \geq n^{-\varepsilon}.$$

Since $\beta^{-k-1} mp < \beta^{-k} mp < \lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \rceil$, from binomial monotonicity, Lemma 23, we have that for any \tilde{m} such that $\beta^{-k-1} m \leq \tilde{m} \leq \beta^{-k} m$,

$$\mathbb{P} \left(\text{Bin}(\tilde{m}, p) = \lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \rceil \right) \geq n^{-\varepsilon}.$$

In order to deal with the more delicate sparse regime throughout the paper where $mp \ll \log n$, we apply the following technical lemma.

► **Lemma 25.** For $mp \ll \log n$, $\varepsilon > 0$, and n large enough, we have

$$\mathbb{P} \left(\text{Bin}(m, p) = \left\lceil \frac{\varepsilon}{8} \frac{\log n}{\log(\log n / mp)} \right\rceil \right) \geq n^{-\varepsilon}. \quad \blacktriangleleft$$

Proof of Lemma 25. We follow the argument in Lemma 24 with $k = 0$ and $\mathbb{E}d_{\max}$ replaced by $\log n / (\log \log n - \log mp)$. Note that in the proof of Lemma 24, in the case $mp \ll \log n$, we only used that $mp \ll \mathbb{E}d_{\max} \ll \log n$ and $\mathbb{E}d_{\max} \leq C \log n / (\log \log n - \log mp)$ for some $C > 0$. Since both these properties remain true upon replacing $\mathbb{E}d_{\max}$ with $\log n / (\log \log n - \log mp)$, the proof follows. Since $\tau = \varepsilon/(8C)$, in the setting of Lemma 25, and $C = 1$ in this argument, we pick $\tau = \varepsilon/8$. ◀

C

 Lemma 26, formal version of Lemma 7

► **Lemma 26.** Let $\varepsilon > 0$. Consider the following choices of f_1, f_2, \dots :

(i) if $mp \lesssim \log n$, for some constants $\tau > 0$ and $1 < \alpha < \beta$,

$$f_t = \left\lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \right\rceil \quad \text{where } k \text{ is such that } \beta^{-k-1}m < m - F_{t-1} \leq \beta^{-k}m;$$

(ii) if $mp \gg \log n$, and $\log mp \ll \log n$,

$$f_t = \lceil mp(1-p)^{t-1} \rceil \quad \text{if } t \leq t^* := \left\lceil \frac{1}{p} \log \left(\frac{mp}{\log n} \right) \right\rceil,$$

$$f_t = \tilde{f}_{t-t^*}, \quad \text{otherwise, where } \tilde{f}_t \text{ is the sequence from the case } mp \lesssim \log n;$$

(iii) otherwise, i.e., when $\log mp \gtrsim \log n$,

$$f_t = \lceil mp(1-p)^{t-1} \rceil.$$

Then, there exists K , such that

- (i) $F_K \geq m - K$;
 - (ii) if $mp \lesssim \log n$, then $K \sim \text{val}_{LP}$;
 - if $mp \gg \log n$, then $K \sim \text{val}_{IP}$.
- (C.1)

Furthermore, for this sequence f_t (which depends on ε), for any $t \leq K$,

$$\mathbb{P}(\text{Bin}(m - F_{t-1}, p) \geq f_t) \geq n^{-\varepsilon}. \quad (\text{C.2})$$

Note that the implicit constants in the statements $K \sim \text{val}_{LP}$ or $K \sim \text{val}_{IP}$ depend on ε .

Proof. We proceed in the proof by first showing that there exists \tilde{K} , such that $m - F_{\tilde{K}} \lesssim \tilde{K}$, and then, by increasing \tilde{K} by a multiplicative factor, we find K such that $m - F_K \leq K$.

Case $mp \lesssim \log n$. From Lemma 24, there exist constants $\tau > 0$, α, β with $1 < \alpha < \beta$, such that, for any \tilde{m} , satisfying $\beta^{-k-1}m \leq \tilde{m} \leq \beta^{-k}m$, for all n large enough,

$$\mathbb{P}\left(\text{Bin}(\tilde{m}, p) = \lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \rceil\right) \geq n^{-\varepsilon}.$$

Recall that in this case $f_t = \lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \rceil$, where k is such that $\beta^{-k-1}m \leq m - F_{t-1} \leq \beta^{-k}m$ and $F_t = \sum_{s=1}^t f_s$. From Lemma 24 we have that $\mathbb{P}(\text{Bin}(m - F_{t-1}, p) = f_t) \geq n^{-\varepsilon}$. Our goal is to prove that there exists $s \lesssim \text{val}_{LP} \sim m/\mathbb{E}d_{\max}$, such that $m - F_s \lesssim s$.

► **Lemma 27.** Let $t^{(k)} := \frac{\beta-1}{\beta\tau} \frac{m}{\mathbb{E}d_{\max}} \alpha^{-k}$.

$$\begin{aligned} \text{If } & m - F_{t-1} \leq \beta^{-k}m \\ \text{then } & m - F_{t+t^{(k)}-1} \leq \beta^{-k-1}m. \end{aligned}$$

Informally, if after $t - 1$ steps of *BlockGreedy*, at most $\beta^{-k}m$ subsets are uncovered, then after $t + t^{(k)} - 1$ steps, at most $\beta^{-k-1}m$ subsets remain uncovered.

Proof. Let $s \geq t$. As long as $m - F_{s-1} > \beta^{-k-1}m$, we will always have $f_s = \lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \rceil$. We proceed by contradiction. Assume that $m - F_{t+t^{(k)}-1} > \beta^{-k-1}m$. This implies that for all $s \in [t - 1, t + t^{(k)} - 1]$, we have $f_s = f := \lceil (\alpha/\beta)^k \tau \mathbb{E}d_{\max} \rceil$. Therefore,

$$F_{t+t^{(k)}-1} - F_{t-1} = t^{(k)} f \geq \frac{m(\beta-1)}{\beta^{k+1}} = \beta^{-k}m - \beta^{-k-1}m,$$

30:22 Greedy Heuristics and Linear Relaxations for the Random Hitting Set Problem

and

$$\begin{aligned} m - F_{t+t^{(k)}-1} &= m - F_{t-1} - (F_{t+t^{(k)}-1} - F_{t-1}) \\ &\leq \beta^{-k}m - (\beta^{-k}m - \beta^{-k-1}m) = \beta^{-k-1}m. \end{aligned}$$

Therefore, we must have $m - F_{t+t^{(k)}-1} \leq \beta^{-k-1}m$. \blacktriangleleft

Note that we always have $\beta^{-1}m \leq m - F_0 = m$. If we consecutively apply Lemma 27 starting with $k = 0$, then, for $v(k) := \sum_{s=0}^k t^{(s)}$ we have $m - F_{v(k)-1} \leq \beta^{-k-1}m$. Therefore, for $k := \frac{\log \mathbb{E}d_{\max}}{\log \beta}$, we have $m - F_{v(k)-1} \leq \frac{m}{\mathbb{E}d_{\max}}$. We can bound

$$v(k) \leq \sum_{s=0}^{\infty} t^{(s)} = \frac{\beta - 1}{\beta\tau(\alpha - 1)} \frac{m}{\mathbb{E}d_{\max}} \sim \frac{m}{\mathbb{E}d_{\max}}.$$

From Lemma 6 we have $d_{\max} \lesssim \mathbb{E}d_{\max}$ with high probability. Together with Lemma 15 this implies $\text{val}_{\text{LP}} \geq \frac{m}{d_{\max}} \gtrsim \frac{m}{\mathbb{E}d_{\max}}$. Now, if we pick $\tilde{K} := v(k) \lesssim \frac{m}{\mathbb{E}d_{\max}}$, we have that $\text{val}_{\text{BGr}} \lesssim \frac{m}{\mathbb{E}d_{\max}}$. Since $\text{val}_{\text{LP}} \leq \text{val}_{\text{BGr}}$, we have that $\tilde{K} \sim \text{val}_{\text{LP}}$ and $m - F_{\tilde{K}} \lesssim \tilde{K}$.

Case $mp \gg \log n$. Here, we have that $\mathbb{E}d_{\max} = mp(1 + o(1))$, therefore, picking an element that hits an average number of subsets is approximately the same as picking an element that hits close to maximum number of subsets. From the properties of the mean and the median of the binomial distribution, it follows that $\mathbb{P}(\text{Bin}(\tilde{m}, p) \geq \lceil \tilde{m}p \rceil) \geq 1/3$, for any \tilde{m} .

We begin with the case $\log mp \ll \log n$. This means that mp cannot grow polynomially in n , but e.g. $mp \sim \log^2 n$ is possible. In this regime, $\text{val}_{\text{IP}} \sim \frac{1}{p} \log \left(\frac{mp}{\log n} \right)$. Let $K_1 = \left\lceil \frac{1}{p} \log \left(\frac{mp}{\log n} \right) \right\rceil$ and f_1, \dots, f_{K_1} be a sequence such that $f_s = \lceil mp(1-p)^s \rceil$. Then, we have that $m - F_{K_1} \leq m(1-p)^{K_1} \leq \frac{1}{p} \log n$. Therefore, $(m - F_{K_1})p \sim \log n$, and we can continue with \tilde{f}_t from the previous section $mp \sim \log n$, with $\tilde{F}_t := \sum_{s=1}^t \tilde{f}_s$. For this sequence $\tilde{f}_1, \dots, \tilde{f}_{K_2}$, we have $K_2 \lesssim \frac{1}{p}$, and $m - F_{K_1} - \tilde{F}_{K_2} \lesssim \frac{1}{p} \ll \frac{1}{p} \log \left(\frac{mp}{\log n} \right)$. The required statement holds for combined sequences f_t and \tilde{f}_t and $\tilde{K} := K_1 + K_2$.

Finally, we study the case $\log mp \gtrsim \log n$, which implies that $\text{val}_{\text{IP}} \sim \frac{1}{p} \log n$. This case is trivial, as one can pick $\tilde{K} = \left\lceil \frac{1}{p} \log \left(\frac{mp}{\log n} \right) \right\rceil \lesssim \text{val}_{\text{IP}}$ and $f_1, \dots, f_{\tilde{K}}$ a sequence such that $f_s = \lceil mp(1-p)^s \rceil$. Then, we have that $m - F_{\tilde{K}} \leq m(1-p)^{\tilde{K}} \leq \frac{1}{p} \log n \lesssim \text{val}_{\text{IP}}$.

From $m - F_{\tilde{K}} \lesssim \tilde{K}$ to $m - F_K \leq K$. Finally, using that $f_t \geq 1$ by Lemma 16 unless $F_t = m$, there exists some constant $C > 0$, such that for $K := C\tilde{K}$, $F_K \geq m - K$, which finishes the proof. \blacktriangleleft