arXiv:2305.05565v1 [math.PR] 9 May 2023

随机击中集问题的贪心启发式与线性松弛

GABRIEL ARPINO[1], DANIIL DMITRIEV[2], 和 NICOLO GROMETTO[3]

摘要。考虑击中集问题：给定一个全集 $\mathcal{X} = \{1 \dots, n\}$ 和一个子集族 $\mathcal{S}_1, \dots, \mathcal{S}_m$，目标是找出 $\mathcal{X}$ 的最小子集，使其与该子集族中的每个元素都有非空交集。我们研究这一问题的一种概率表述，其中底层子集通过以概率 p 独立地纳入全集中的每个元素而形成。对于足够大的 n，我们严格分析 Lovász 著名的贪心算法 [10]相对于所选输入分布的平均情形性能。此外，我们还研究该问题的线性规划解与整数规划解之间的整数间隙。

1. 引言

击中集是组合优化中的一个经典问题。给定一个由元素构成的基集$\mathcal{X}:= \{1, ..., n\}$ 以及由 $\mathcal{X}$的子集构成的集合族 $\mathcal{C}:= \{\mathcal{S}_1, ..., \mathcal{S}_m\}$，该问题要求找出与 $\mathcal{C}$中的每个子集都相交的最小集合 $\mathcal{S} \subseteq \mathcal{X}$。研究超图最小顶点覆盖 （MVCH）时，可以自然地引出击中集问题：将超边视为子集，将顶点视为基集中的元素即可。此外，击中集问题与集合覆盖 问题 [12], 互为对偶；集合覆盖在计算复杂性理论中历史悠久，也是卡普的 21 个 NP 完全问题之一。本研究旨在通过分析一大类随机输入下的平均情形复杂度，深化我们对典型击中集实例难度的理解。这类随机输入通过以固定概率 p将基集中的每个元素独立地分配给任意子集来生成。为了进一步阐明本研究的目的，我们先回顾击中集问题的如下整数规划（IP）表述：

$$\text{val}_{\text{IP}} := \begin{cases} \underset{\boldsymbol{x}}{\text{minimize}} & \|\boldsymbol{x}\|_1 \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} \geq \mathbf{1},\ \boldsymbol{x} \in \{0,1\}^n, \end{cases} \tag{1.1}$$

其中， $\boldsymbol{A} \in \{0,1\}^{m \times n}$ 的第 i 行以二进制编码表示集合 $\mathcal{S}_i$ 中元素 $\mathcal{X}$的隶属关系，且 $\mathbf{1} := (1, \dots, 1) \in \mathbb{R}^m$。对于当前问题的顶点覆盖表述，我们注意到 $\boldsymbol{A}$ 由底层超图的关联矩阵构成。具体而言，(1.1) 中的第一个约束确保候选解向量击中 $\mathcal{C}$ 中的每个集合。通过允许分数解，可以得到一种自然的凸松弛，其可表示为以下线性规划（LP）：

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

尽管显然 $\text{val}_{\text{LP}} \leq \text{val}_{\text{IP}}$，但一般而言，紧性未必成立。事实上，对于满足每行和每列恰好包含 k 个 1 的 $m = n$ 和 $\boldsymbol{A} \in \{0,1\}^{n \times n}$（其中 $1 < k < n$ 为某个固定值），一个最优

[1]剑桥大学，CB2 1PZ，剑桥，英国[2]苏黎世联邦理工学院和ETH人工智能中心，8092 苏黎世，瑞士[3]普林斯顿大学，谢勒德楼，普林斯顿，新泽西州 08540，美国电子邮箱地址：ga442@cam.ac.uk, daniil.dmitriev@ai.ethz.ch, ng1069@princeton.edu。日期：2023年5月10日。关键词与短语。 击中集，随机超图，整数间隙，贪心算法。

解由 $\boldsymbol{x}^*_{\mathrm{LP}} = (1/k, ..., 1/k)$ 给出，该解不是整数解，因此当 n/k 不是整数时，目标值会严格更小。这表明存在显著的整数间隙，其大小由比值 $\mathrm{val}_{\mathrm{IP}}/\mathrm{val}_{\mathrm{LP}}$ 量化。在 [10], , Lovász 证明，击中集乘法整数间隙的最坏情况上界基本上是最优的，为 $1+\log\, d_{\max}$，其中 $d_{\max}$ 对应于底层超图中的最大度。该上界通过分析贪心算法（算法1）得到：该算法通过依次将未覆盖超边中度数最高的顶点加入顶点覆盖来构造解，下一节将对此进行更详细的讨论。然而，在许多自然实例中，最大度 $d_{\max}$ 会随着超图中顶点数量的增加而增长，因此超图规模越大，上界就会逐渐变差。贪心算法不仅可以说是解决击中集问题最自然的候选方法，而且已被证明，对于这一经典问题的最坏情况实例，它基本上是可能达到的最佳多项式时间近似算法 [13] 。具体而言，贪心算法的近似比为$O(\log\, m)$，因此对于足够大的常数 $C > 0$，它找到的覆盖规模至多是最小覆盖的 $C \log m$ 倍。

尽管过去几十年来人们对击中集开展了大量研究，但对于线性规划和贪心算法在随机问题实例上的典型性能，我们仍缺乏充分理解。因此，我们提出以下问题：

(1) 对于哪些 m、n、p 的情形，击中集的随机实例中会出现（乘法意义下的）整数间隙？(2) 针对选定的输入分布，贪心算法的平均情形性能如何?

在本文中，对于上述问题，我们将在基集 $\mathcal{X}$ 的基数 n 较大但有限的情形下，以非渐近的意义给出高概率（w.h.p.）成立的答案。通过对一种输出与分数覆盖具有相同规模的整数覆盖的算法进行平均情形分析，我们将证明在 n、m、p 的广泛情形下不存在整数间隙。此外，通过直接归约，我们还将对贪心例程进行严格分析。以下结果在下列条件下成立，这些条件将在全文中始终假定成立。

假设 1. 我们假设

(1) 每个元素 $j \in \mathcal{X}$ 以概率 $p \equiv p(n)$ 被分配到任意子集 $\mathcal{S}_i$, $i \in [m]$ ，且相互独立。也就是说，$\boldsymbol{A} \in \{0,1\}^{m\times n}$ 满足 $A_{ij} \overset{\text{独立同分布}}{\sim}$ 伯努利(p)；(2) n 意在取较大但有限的值；(3) $m \equiv m(n) = \mathrm{poly}(n)$，即存在 $\exists c, C > 0$，使得当 n 足够大时， $cn^c \le m \le Cn^C$ ；(4) 存在 $\delta \in (0,1)$，使得 $p \equiv p(n)$ 满足 $1/n^\delta \le p \le 1/2$，对所有足够大的 n 均成立。

请注意，在假设 1.3 中，上界的选取是为了避免以高概率出现平凡解，例如当集合的数量随 $\mathcal{X}$ 的基数呈指数增长时就会出现这种情况。此外，假设 1.4 绝非限制性条件，因为可以证明，当 $m = \mathrm{poly}(n)$ 且 $np \ll \log n$ 时，以高概率 $\boldsymbol{A}$ 包含一行全零行，即整数规划没有可行解。要求 $p \le 1/2$ 是出于技术上的便利，可以放宽为任意常数 p，从而涵盖 [8] 中的情形。

我们的贡献源于对包含集大小 $I_j := \{i \in [m] : \ j \in \mathcal{S}_i\}$（$j \in [n]$）的研究。在当前问题的 MVCH 表述中，这些大小对应于底层超图中顶点的度。我们重点考察的关键量是平均包含集大小，即在当前分布假设下，对所有 j 而言的 $\mathbb{E}|I_j| = mp$。该量呈现出两个不同且值得关注的区间，分别称为稀疏区间（ $mp \ll \log\, n$）和稠密区间（ $mp \gg \log n$）。这两个区间进而决定了最大包含集的大小，也就是最大度 $d_{\max} := \max_{j\in[n]} |I_j|$。我们刻画了整数间隙的行为（精确到乘法常数），并分析了 Lovász 贪心算法 [10] 在这两个区间中的表现，且当 $n \to \infty$ 时该结论以高概率成立。为此，我们证明了一种简单贪心启发式方法 BlockGreedy 算法（算法 2）的成功性。全文中，我们用符号 $\mathrm{val}_{\mathrm{Gr}}$、$\mathrm{val}^{\mathrm{Alg}}$ 分别表示贪心算法和 BlockGreedy 返回的击中集大小。

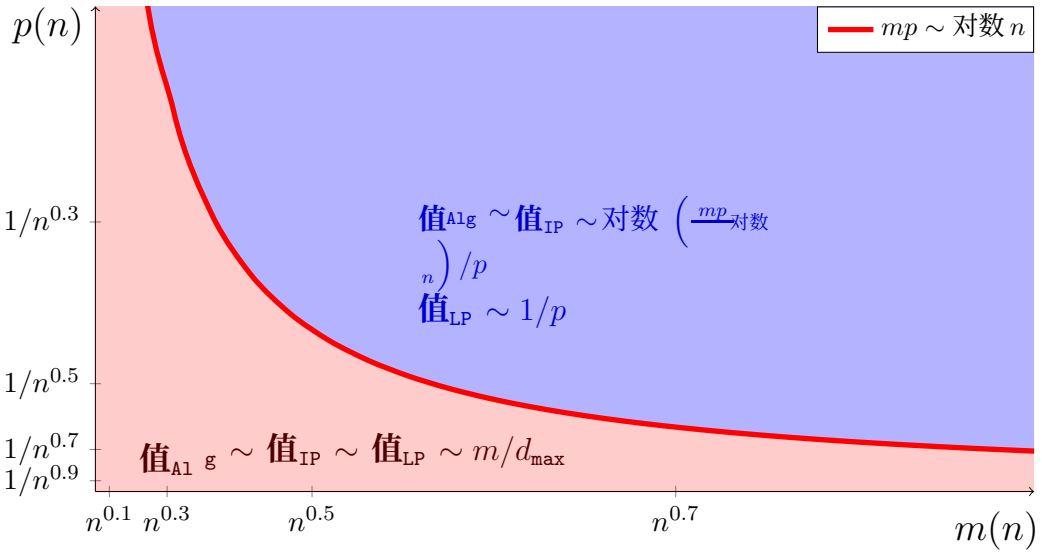

图1. 当平均包含集大小为 mp 时，稀疏区间与稠密区间之间的转变。

下面我们对主要结果作非正式说明，这些结果也见图1。正式表述见定理 1 和定理 2。

稀疏区间 ($mp \ll \log n$)：我们证明，在稀疏区间内，BlockGreedy 算法能够达到 LP 下界 $m\overline{d_{\max}}$，因此不存在整数间隙 ($\text{val}_{\text{LP}}/\text{val}_{\text{IP}} \sim 1$)。

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

稠密区间 ($mp \gg \log n$)：我们通过一阶矩方法证明，在稠密区间内，整数规划与 LP 之间存在紧且不消失的整数间隙。我们证明，BlockGreedy 算法在此区间内的表现与整数规划一样好，即

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

阈值区间 ($mp \sim \log n$)：该区间在稀疏区间与稠密区间之间平滑过渡，且不存在整数间隙。所有关注量的缩放均为 $m/d_{\max} \sim 1/p$。

贪心：我们证明，对于 $\delta < 1/2$，$\text{val}_{\text{Gr}} \sim \text{val}_{\text{IP}}$，其中 δ 是假设 1.4 中的参数 .

本文其余部分安排如下。第 2 节介绍相关记号。第 3 节概述并讨论相关文献。第 4 节证明若干初步结果，这些结果将为后续核心论证奠定基础。第 5 节研究 $d_{\max}$ 的值。随后，第 6 节探讨该问题的算法方面：首先给出简单算法 BlockGreedy 的性能保证，然后通过归约分析 Greedy。最后，我们在第 7 节总结研究结果，并对未来工作提出展望。为使行文更加简洁、方便读者阅读，技术性较强的结果的证明放在附录中。

2. 符号与约定

对于整数 $k \in \mathbb{N}$，我们记作 $[k] := \{1, ..., k\}$。我们分别用粗体罗马字母 x $\boldsymbol{A} \in \mathbb{R}^k, \mathbb{R}^{k\times k}$ 表示向量和矩阵，其中 $k \in \mathbb{N}$。定义元素或节点 $j \in [n]$ 的包含集 为 $I_j = \{i \in [m] : j \in \mathcal{S}_i\}$。我们用 X_j, $j \in [n]$ 表示 $\boldsymbol{A}$ 的第 j 列的 ℓ_1 范数，并注意到 $X_j = |I_j|$ 和 $X_1, \dots X_n$ 独立同分布$\sim$，服从二项分布 Binomial(m, p)。此外，令 $d_{\max} \equiv d_{\max}(X_1, \dots, X_n) := \max_{i\in[n]} X_i$。我们分别用 $\mathbb{E}$ 和 Var 表示期望和方差。用 $\lesssim$ 表示在常数因子范围内成立的不等式。我们用 $A \sim B$ 表示当 n 足够大时 $A \lesssim B \lesssim A$ 成立。我们用 log 表示自然对数。
对于可能的随机函数 $f(n)$ $g(n)$，我们用 $\{f \lesssim g\}$ 表示事件序列 $\{f(n) \le Ag(n)\}$，其中常数 $A > 0$ 与 n无关。因此， $\mathbb{P}(f \lesssim g)$被视为 n的函数。其他不等式的记号定义类似。我们称事件序列 $\{A_n\}$ 相对于概率测度 $\mathbb{P}$ 以高概率 (w.h.p.) 成立，如果存在一个与 n无关的常数 $c > 0$，使得对于足够大的 n， $\mathbb{P}(A_n) \ge 1 - n^{-c}$。

3. 相关工作

或许，求解击中集（或等价地，最大顶点覆盖超图）最为知名的算法，是 Lovász 的贪心算法，运行时间复杂度为 $O(mn^2)$。该算法通过依次添加基集中能够击中最多剩余子集的元素来构造覆盖。Lovász [10] 和 Johnson [9] 最初分别针对确定性超图研究了这一算法。Lovász 分析了贪心算法，并得出了击中集整数间隙的上界 $1 + \log d_{\max}$。然而，对于许多自然实例，最大度会随着超图中顶点数量的增加而增长，导致随着规模扩大，上界也逐渐变差。Slavik [13] 提出了迄今最紧的分析，表明对于集合覆盖而言，贪心算法本质上是最优的多项式时间近似算法，其近似比为 $O(\log m)$。

然而，对于击中集的随机实例，关于多项式时间算法典型表现的了解要少得多。从理论角度以及组合推断的应用角度来看，弥补这一差距都至关重要。群组测试就是一个典型例子：这是一类经典的推断问题，目标是通过尽可能少的混合测试，从大量物品中识别出一小部分缺陷物品，其应用涵盖通信协议分析 [6] 、DNA 测序 [4] 和搜索问题 [3]。在 [8], Iliopulos 和 Zadik 将最小击中集视为群组测试问题中的一种估计量，并称之为最小满足集 估计量。具体而言，他们提供了大量实验证据，支持这样一种观点：由非自适应群组测试所产生的随机击中集问题实例类别，可以由计算机高效求解。

Mézard 和 Tarzia 的研究分析了击中集的随机实例，并依赖统计物理中的非严格技术 [11]。该研究考察了正则均匀超图，其中顶点的度和超边的大小固定且假设为常数。根据这些取值，他们展示了副本对称性、一副本对称性破缺和完全副本对称性破缺各相之间的相变，这些相刻画了该问题在平均情形下优化景观的复杂性。

Telelis 和 Zissimopoulos [14] 还研究了随机伯努利超图的情形，其中元素以独立的方式归属于子集，且概率为固定 $p \in (0, 1)$。他们分析的是基集大小 n 趋于无穷大的渐近区间。在这种情形下，他们研究了一种简单确定性算法的平均情形性能。该算法几乎处处能以至多为 $\log m$ 阶的加性误差项近似随机击中集。由于在此情形下 $d_{\max}$ 至多为 m 阶，这一结果改进了 Lovász 在 [10], 中的论证。然而， [14] 中的分析未能涵盖如下情形：当 $n \to \infty$ 且 $p \to 0$ 时得到的稀疏超图；更广泛地说，也未涵盖任何 p 随基集大小增长的区间。 [14] 中的分析也未能在所选参数区间内证明贪心算法的保证。

为求完整，我们还提请读者关注一系列较新的研究 [1, 2], 。作者针对随机整数规划 $\max \boldsymbol{c}^T\boldsymbol{x}$, $\boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}, \boldsymbol{x} \in \{0,1\}^n$ 与其线性规划松弛的值之间的（加性）整数间隙，在 $(\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c})$ 上广泛的分布情形下给出了界；当 $n \to \infty$ 时，这些界以高概率成立。其涵盖的情形包括： $\boldsymbol{A}$ 的元素均匀分布于一个至少含有三个元素的整数区间，以及 $\boldsymbol{A}$ 的列服从各向同性对数凹分布。然而，这些结果未能涵盖 $\boldsymbol{A}$ 稀疏且元素取自 $\{0,1\}$ 的情形，而这正是击中集问题所关注的。

4. Bounds on $\mathrm{val}_{\mathtt{LP}}$ and $\mathrm{val}_{\mathtt{IP}}$

本节旨在获得 $\mathrm{val}_{\mathtt{LP}}$、$\mathrm{val}_{\mathtt{IP}}$ 的初步界，出发点是 $\mathrm{val}_{\mathtt{LP}}$ 的以下确定性下界；该下界适用于 m、n、p 的所有区间。

引理 1. 我们有

$$val_{LP} \ge \frac{m}{d_{\max}}.$$

证明. 设 $\boldsymbol{x}^*_{\mathtt{LP}} = (\boldsymbol{x}^*_1, \boldsymbol{x}^*_2, \ldots, \boldsymbol{x}^*_m)$ 是 (1.2) 的一个最优解。由于 $\boldsymbol{A}\boldsymbol{x}^*_{\mathtt{LP}} \ge \mathbf{1}$ 逐项成立，对所有条目求和可得

$$m \le \sum_i \boldsymbol{x}^*_i X_i \le d_{\max} \sum_i \boldsymbol{x}^*_i = d_{\max}\mathrm{val}_{\mathtt{LP}}.$$

整理后即可得到所需结果。 □

除此之外，我们还有 $\mathrm{val}_{\mathtt{LP}}$ 的以下基本上界，该上界在稀疏区间和稠密区间均成立。

引理 2. 存在 $c > 0$，且与 n 无关，使得

$$\mathbb{P}\left(val_{LP} \lesssim \frac{1}{p}\right) \ge 1 - \exp\left(-cn^{1-\delta}\right).$$

这也意味着 $\mathbb{P}(\text{整数规划 可行}) \ge 1 - \exp\left(-cn^{1-\delta}\right)$.

证明。 考虑候选可行解 $\hat{\boldsymbol{x}} := \frac{1}{\tilde{C}np}\mathbf{1}$，其中 $0 < \tilde{C} < 1$ 为某个常数。对约束应用并集界和标准的切尔诺夫界即可得到以下结果。

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

考虑上述事件的补事件，并注意到 $\|\hat{\boldsymbol{x}}\|_1 \sim 1/p$，即可得到所需结论。注意，事件 $\{\hat{\boldsymbol{x}}$ 对 LP$\}$ 可行，意味着事件 {整数规划可行}。□

引理 1 显然意味着 $\mathrm{val}_{\mathtt{IP}} \ge m/d_{\max}$。然而，当 $mp \gg \log n$ 时，这个下界并不精确。事实上，我们将应用一阶矩方法，证明 $\mathrm{val}_{\mathtt{IP}}$ 的一个更紧的下界。我们先回顾 Lambert W 函数的以下性质，该函数是方程 $ye^y = x$ 在 $x \ge 0$ 下的解。

引理 3 (Lambert W 函数， [7])。对于任意 $x \ge e$，均有

$$\log x - \log\log x + \frac{\log\log x}{2\log x} \le W_0(x) \le \log x - \log\log x + \frac{e}{e-1}\frac{\log\log x}{\log x}. \tag{4.1}$$

特别是，

$$W_0(x) = \log x - \log\log x + o(1), \quad as \quad x \to \infty. \tag{4.2}$$

此外，对于任意 $x \geq 1/e$，以下恒等式成立

$$W_0(x) = \log \frac{x}{W_0(x)}. \tag{4.3}$$

现在，我们通过一阶矩方法，在 $mp \gg \log n$ 的范围内，为 $\mathsf{val}_{\mathtt{IP}}$ 提供一个更紧的下界。

引理 4。如果 $mp \gg \log n$，那么对于任意 $D \geq 1$，当 n 足够大时，

$$\mathbb{P}\left(val_{IP} \gtrsim \frac{1}{p} \log\left(\frac{mp}{\log n}\right)\right) \geq 1 - n^{-D}.$$

引理 *4* 的证明。固定 $D \geq 1$。令 $Z_k := |\{\boldsymbol{x} \in \{0,1\}^m : \boldsymbol{A}\boldsymbol{x} \geq \mathbf{1}, \|\boldsymbol{x}\|_1 = k\}|$ 表示范数恰好为 k 的可行解的数量。显然，对于任意 $k \geq 0$，都有 $Z_k \leq Z_{k+1}$。此外，我们还有

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

现在我们将证明，当 $k \ll \frac{1}{p}\log\left(\frac{mp}{\log n}\right)$ 时，有 $\mathbb{E}Z_k \leq n^{-D}$。利用假设 **1.4** 中的 $p \leq 1/2$，以及对于 $x \in (0, 1/2)$ 有 $(1-x)^y \geq e^{-2xy}$，我们可以界定

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

因此，$\mathbb{E}Z_k \leq n^{-D}$ 可由

$$2pke^{2pk} \leq -2Dpe^{2pk} + \frac{2mp}{\log n}. \tag{4.4}$$

由于 $k \ll \frac{1}{p}\log\left(\frac{mp}{\log n}\right)$，我们还有 $k \leq k_* := \frac{1}{2p}W_0\left(D^{\frac{mp}{\log n}}\right)$，当 n 足够大时成立。对于 $k = k_*$，(4.4) 的左侧等于 $D^{\frac{mp}{\log n}}$，而右侧的下界为 $\frac{mp}{\log n}$。由于 $D \geq 1$，我们得到 $\mathbb{E}Z_k \leq n^{-D}$。注意，当 n 足够大时，$\mathsf{val}_{\mathtt{IP}} \ll \frac{1}{p}\log\frac{mp}{\log n}$ 意味着 $Z_{k_*} > 0$。因此，应用马尔可夫不等式可得

$$\mathbb{P}\left(\mathrm{val}_{\mathtt{IP}} \ll \frac{1}{p}\log\left(\frac{mp}{\log n}\right)\right) \leq \mathbb{P}\left(Z_{k_*} > 0\right) \leq \mathbb{E}Z_{k_*} \leq n^{-D}, \tag{4.5}$$

通过考虑互补事件即可完成证明。注意，采用类似的推导还可以证明，对于 $k^* := \frac{1}{p}\log\left(\frac{1}{\delta}\frac{mp}{\log n}\right)$，其中 δ 在假设 **1.4** 中定义，我们有 $\mathbb{E}Z_{k^*} \geq 1$。□

5. Analysis of $d_{\mathtt{max}}$

本节将首先估计 $\mathbb{E}d_{\mathtt{max}}$ 的大小，随后证明以高概率有 $d_{\mathtt{max}} \lesssim \mathbb{E}d_{\mathtt{max}}$，这足以满足本文的需要。关于 $d_{\mathtt{max}}$ 集中于其均值附近的更多论证，请参阅引理 **11**、**10**。

为了处理更微妙的稀疏区间，即 $mp \ll \log n$，我们陈述以下技术引理，其证明见附录 B。

引理 5。对于足够大的 $mp \ll$ 对数 n、$\varepsilon > 0$ 和 n，我们有

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

我们通过区分平均包含集大小的稀疏区间和稠密区间，来估计 $\mathbb{E}d_{\max}$。以下引理刻画了最大包含集大小 $\max_{j\in[n]}|I_j|$ 的两个不同区间。

引理 6（二项分布的最大值）。令 $X_1,\ldots,X_n \overset{\text{独立同分布}}{\sim} Bin(m,p)$。在假设 *1* 的条件下，有如下结论：

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

证明。为简化记号，定义 $b_n:=\frac{\log n}{mp}$，$b_n^*:=\frac{1}{e}(b_n-1)$，$g_n:=\frac{\log n}{\log(\log n/mp)}$。我们首先证明所需的上界。根据詹森不等式，并用正值之和来界定其最大值，对于任意 $\lambda>0$，可得

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

最后，计算二项随机变量的矩生成函数，结合不等式 $1-x\le e^{-x}$ 得出

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

在 $mp\gtrsim\log n$ 的情形下，我们可以任意选择 $\lambda>0$，且不依赖于 n，由此立即可得 $\mathbb{E}\max_{i\in[n]}X_i\lesssim mp$。

对于 $mp\ll\log n$，我们对上述公式的最后一行求导，并令所得表达式为零。由此，我们可以选择 λ 作为以下方程的解。

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

在当前假设下，这可用 LambertW 函数表示为 $\lambda=1+W_0(b_n^*)$，因此由 (4.3) 可得

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

在稠密 $mp\gtrsim\log n$ 区间，通过注意到 $\mathbb{E}\max_{i\in[n]}X_i\ge\mathbb{E}X_1=mp$，很容易得到匹配的下界。

为了处理稀疏区间，令 $\tau=1/16$。由马尔可夫不等式，

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

因此，应用引理 5，当 n 足够大时，

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

由此为稀疏区间给出了匹配的下界。 □

注 1. 我们注意到，在稀疏区间 $mp\ll\log n$ 中，有 $mp\ll\mathbb{E}d_{\max}\ll\log n$；而在稠密区间 $mp\gg\log n$ 中，有 $mp\sim\mathbb{E}d_{\max}\gg\log n$。在阈值区间 $mp\sim\log n$ 中，X_i 的平均值和最大值处于同一数量级，即 $mp\sim\mathbb{E}d_{\max}\sim\log n$。从证明中可以看出平滑过渡：注意到在此区间内，$b_n,b_n^*,W_0(b_n^*)\sim 1$。

本节最后，我们证明 $d_{\max}$ 以其期望为上界，允许相差一个乘法常数，且该结论以高概率成立。虽然这一单侧结果足以支撑后续分析，但我们预计同样存在匹配的下界。关于 $d_{\max}$ 集中性的更多见解，可参见附录 A 中的引理 10、11。

引理 7. 设 $X_1, \ldots, X_n$ 独立同分布$\sim Bin(m,p)$。那么，存在常数 $c, \tilde{c} > 0$，它们与 n 无关，使得

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

也就是说，以高概率有 最大化$_{i\in[n]} X_i \lesssim$ 𝔼最大化$_{i\in[n]} X_i$ *w.h.p.*.

证明。 让我们分别考虑稀疏区间和稠密区间。
对于 $mp \gtrsim \log n$ 的稠密区间，存在常数 c_1、 c_2、 $c_3 > 0$，使得 $c_1 mp \leq \mathbb{E}\max_{i\in[n]} X_i \leq c_2 mp$（如引理 6 中所论证），且 $mp \geq c_3 \log n$。我们按照引理 9 中的方法应用并集界和切尔诺夫界，得到对任意 $t \geq 1/c_1$,

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

现在只需选择 t 作为 c_1、 c_3 的函数，使得 $\frac{c_3(tc_1-1)^2}{1+tc_1} > 1$)。通过整理并求解所得的二次方程，可以立即得出任意 $t > \frac{1}{c_1} + \frac{1+\sqrt{1+8c_3}}{2c_3 c_1} > \frac{1}{c_1}$ 均满足条件。因此，存在通用常数 c、 $\tilde{c}$，使得所需结论成立。

现在我们考虑稀疏区间 $mp \ll \log n$。根据引理 6，存在 $c_4 > 0$，使得$mp \leq c_4 \log n / \log\left(\frac{\log n}{mp}\right)$。
注意，对任意 $\lambda > 0$, $\max_{i\in[n]} X_i \leq \frac{1}{\lambda}\log \sum_{i=1}^n e^{\lambda X_i}$。对任意 $t > 0$，应用马尔可夫不等式可得,

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

这里我们利用了 $1 + x < e^x$，从而得到最后一个不等式。最后，取 $t = 3/c_4$ 和 $\lambda = \log(\log n/mp)$，即可得到

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

□

6. 算法解法

本节旨在针对预设的伯努利随机情形，对击中集问题的标准贪心算法进行严格分析。特别是，我们证明该算法能够以高概率构造出最优规模的击中集，与第 4 节的结果相同，至多相差一个乘法常数。为此，我们首先分析贪心启发式算法的一种变体，随后通过归约论证得出结论。

贪心算法的核心原则是分步构造一个可行解：每次向候选解中依次添加一个能够命中最多剩余集合的元素。在所选设定下，元素以相同概率且相互独立地被添加到集合中，我们

算法1贪心

$\mathcal{I} \leftarrow \{I_1, \ldots, I_n\}$ ▷ 包含集
$U \leftarrow [m]$
$t \leftarrow 0$
当 $|U| > 0$ **时执行**
  $P \leftarrow \text{argmax}_{I \in \mathcal{I}} |I \cap U|$ ▷ 贪心步骤
  $\mathcal{I} \leftarrow \mathcal{I} \setminus \{P\}$
  $U \leftarrow U \setminus P$
  $t \leftarrow t + 1$
值$_{\texttt{Gr}} \leftarrow t$
返回$\text{val}_{\texttt{Gr}}$。

可以精确估计一个最先选取的元素所命中的子集数量。事实上，这个集合的大小由相互独立的二项分布随机变量的最大值给出，而第 5 节对此进行了分析。然而，这第一步便在剩余矩阵列之间引入了非平凡的依赖关系，使得追踪随后每次向候选解中添加元素所产生的边际增益变得十分复杂。

为了解决这一问题，我们引入了一种改进的贪心例程，称为 BlockGreedy 算法。在该算法中，基集 $[n]$ 中的元素被划分为若干个独立的集合，我们称之为块。在第 t 次迭代时，算法只从前 t 个块中选取最大的列。通过适当选择块的数量，可以保证每次迭代都能找到足够多的独立列，同时找到规模最优的解。此外，额外的块约束使我们能够分析每个选定元素命中了多少个子集，而不仅仅是第一个子集，同时仍然依赖独立二项分布随机变量的基本性质。

BlockGreedy 的详细描述见算法 2，非正式地说，其工作方式如下。

(1) 令 K 为解的大小（由理论分析给出建议）；(2) 将 n 列均匀随机地划分为 K 个块，每块包含 n/K 列；(3) 从一个空的候选列集合开始；(4) 在第 t次迭代中，先加入第 t个块中的列（步骤 6），然后在当前候选集合上执行一次贪心步骤（步骤 7）；(5) 如果算法经过 K 次迭代后仍有子集未覆盖，则采用平凡覆盖，即用单独一列覆盖每个子集。高概率下可以做到这一点。

令 v_t 为 BlockGreedy 在第 t步选取的元素， b_t 为被 v_t命中的新子集数， $B_t := \sum_{i=1}^{t} b_i$ 为经过 t 步后命中的子集总数。为了分析 BlockGreedy 选取了多少个元素，我们将引入序列f_1、 f_2、……、 f_s，其中 $F_t := \sum_{i=1}^{t} f_i$，使其满足以下条件：

(1) 对每个 $t \le s$，以高概率有 $F_t \le B_t$；(2) $F_s = m$；
(3) 若 $mp \lesssim \log n$，则 $s \lesssim \text{val}_{\text{LP}}$，否则， $s \lesssim \text{val}_{\text{IP}}$。

前两个性质确保 BlockGreedy 至多选取 s 元素，而最后一个性质给出了 s的最优界。保证 BlockGreedy 成功的一种方法是证明，在 BlockGreedy 的每一步 t的选择中，以高概率存在一个元素 $\tilde{v}_t$ ，它至少命中f_t 个新子集。我们将证明，只需在新列块$\mathcal{B}_t$中寻找 $\tilde{v}_t$ 即可，该列块在步骤 t中加入。注意，除非 $F_t = m$，否则我们有 $f_t \ge 1$，因为以高概率每个子集都至少被一个元素命中。因此，只需找到一个序列 $\{f_1, f_2, \ldots, f_v\}$使得 $F_v \ge m - v$，因为这意味着 $F_{2v} = m$。这使我们能够将证明 BlockGreedy 有效性的问题归约为一个关键的技术引理。该引理假设在步骤t之前，恰好有 F_{t-1} 个子集已被命中，并从下方界定新桶中的某个顶点至少命中 f_t 个新子集的概率。这归结为计算 $\mathbb{P}(\text{Bin}(m - F_{t-1}, p) \ge f_t)$。

算法 2 BlockGreedy

令 $\mathcal{B}_t \subset \{I_1, ..., I_n\}$ 表示第 t 个块，即在步骤 t 变得可用的包含集。
$\mathcal{I} \leftarrow \varnothing$
$U \leftarrow [m]$
$t \leftarrow 0$
当 $|U| > 0$ 且 $\mathcal{B}_t \neq \varnothing$ **时**
  $\mathcal{I} \leftarrow \mathcal{I} \cup \mathcal{B}_t$ ▷ 从新块中添加元素
  $P \leftarrow \operatorname{argmax}_{I \in \mathcal{I}} |I \cap U|$ ▷ 贪心步骤
  $\mathcal{I} \leftarrow \mathcal{I} \setminus \{P\}$
  $U \leftarrow U \setminus P$
  $t \leftarrow t + 1$
值$_{\mathtt{Alg}} \leftarrow t$
如果 $|U| > 0$ **then** 用一个简单算法覆盖 U 的其余部分，$\mathrm{val}_{\mathtt{Alg}} \leftarrow \mathrm{val}_{\mathtt{Alg}} + |U|$
返回$\mathrm{val}_{\mathtt{Alg}}$。

引理 8. 令 $\varepsilon > 0$。考虑以下选择：f_1，f_2，.....：

(i) 如果 $mp \lesssim$ 对数 n，对于某些常数 $\tau > 0$ 和 $1 < \alpha < \beta$，

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

(ii) 如果 $mp \gg$ 对数 n，并且 对数 $mp \ll$ 对数 n，

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

$$f_t = \tilde{f}_{t-t^*}, \quad \text{否则，其中} \quad \tilde{f}_t \text{ 是来自情形 } mp \lesssim \log n \text{;的序列}$$

(iii) 否则，也就是说，当 $\log mp \gtrsim \log n$,时

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

那么，存在 K，使得

$$\begin{array}{ll} (i) & F_K \geq m - K; \\ (ii) & \text{如果 } mp \lesssim \log n, \text{，那么 } K \sim \mathit{val}_{LP}; \\ & \text{如果 } mp \gg \log n, \text{，那么 } K \sim \mathit{val}_{IP}. \end{array} \tag{6.1}$$

此外，对于这个序列 f_t（它取决于 ε），对于任意 $t \leq K$，

$$\mathbb{P}(\mathrm{Bin}(m - F_{t-1}, p) \geq f_t) \geq n^{-\varepsilon}. \tag{6.2}$$

请注意，陈述 $K \sim \mathit{val}_{LP}$ 或 $K \sim \mathit{val}_{IP}$ 中的隐含常数取决于 ε .

这个引理突显了该问题对平均度 mp 与 $\log n$ 之间关系的关键依赖。我们将阐述证明背后的直觉，证明见附录 C。正如第 5 节所示，当期望值较大（$mp \gtrsim \log n$）时，$\mathbb{E} d_{\max}$ 与期望值同步增长；否则则与期望值相去甚远。这是证明的核心性质。

当 $mp \sim n^\gamma$ 满足某个 $\gamma > 0$时，分析很直接，因为只需随机选择列就足够了（即选择 $f_t = \lceil mp(1-p)^{t-1} \rceil$）。更具挑战性的情况是平均度接近 $\log n$，例如 $mp = \log^2 n$（稠密）、$mp = \log n$，或 $mp = \log n / \log\log n$（稀疏）。在这里，我们需要仔细追踪最大度的变化。我们要寻找一个元素，它 i）覆盖大量子集，即接近最大数量的期望值 $\mathbb{E} d_{\max}$；且 ii）能够以足够大的概率找到。第二个性质对于归约到标准贪心算法很重要，因为直接分析该算法有很大困难；我们将在本节后文进行分析。量 $\mathbb{E} d_{\max}$ 对 mp 很敏感，只要后者接近 $\log n$。因此，我们需要调整要寻找的对象

以与之相适应。这通过设置 $f_t = \left\lceil (\alpha/\beta)^k \tau \mathbb{E} d_{\texttt{max}} \right\rceil$ 来实现，并随着剩余行数 $m - F_t$减少而增大参数 k。

例如，考虑 $mp = \log^2 n$ 的情形。首先，我们只能随机选择一个元素，因为它的效果与最大元素一样好。然而，在算法执行过程中，问题会变得更加稀疏；如果我们继续随机选择元素，就会构造出一个次优解。因此，一旦进入 $mp \sim \log n$ 的范围，我们就开始逐渐提高新选元素所能覆盖的数量，使其相对于随机元素更有优势。
现在很容易证明以下定理，它使第 1 节中的陈述更加严谨。

定理 1. 在假设 *1* 下，我们有

$$\begin{aligned} &(i)\ \textit{if}\ \ mp \lesssim \log n\ \textit{then, for any}\ \varepsilon > 0\ \textit{and}\ n\ \textit{large enough,} \\ &\qquad \mathbb{P}\left(val_{\texttt{Alg}} \lesssim \frac{m}{\mathbb{E} d_{\texttt{max}}} \right) \geq 1 - \exp\left(-n^{1-\delta-\varepsilon}\right); \\ &(ii)\ \textit{if}\ \ mp \gg \log n,\ \textit{then, for any}\ \varepsilon > 0\ \textit{and}\ n\ \textit{large enough,} \\ &\qquad \mathbb{P}\left(val_{\texttt{Alg}} \lesssim \frac{1}{p} \log\left(\frac{mp}{\log n} \right) \right) \geq 1 - \exp\left(-n^{1-\delta-\varepsilon}\right). \end{aligned} \tag{6.3}$$

注意，如果 $mp \gtrsim n^\gamma$ 对某个 $\gamma > 0$ 成立，那么对数 mp 对数 $n \sim$ 对数 n，且 (ii) 中的界可以简化。

证明。 证明的主要思路是分析每一步所添加列的分布 t。这些列彼此独立，并且对于每个新添加的列，它所覆盖的额外子集数量服从 $\mathrm{Bin}(m - F_{t-1}, p)$ 分布，其中 F_{t-1} 是已经覆盖的子集数量。引理 8 允许我们对 F_t 给出下界，下面我们将证明，高概率下可以做到这一点。

固定 $\varepsilon > 0$，并令 $\varepsilon' := \varepsilon/4$。令 $f_1, f_2, \ldots$ 为针对 ε' 的引理 8 中给出的序列，并令 K 为使 (6.1) 成立的值，即 $F_K \geq m - K$。注意，对于足够大的 n，存在某个常数 $C > 0$，使得 $K \leq C \max\left\{ \frac{m}{\mathbb{E} d_{\texttt{max}}}, \frac{1}{p} \log(mp \log n) \right\}$。我们将 n 个元素（列）均匀随机地划分为 K 组，每组包含 n/K 个元素（不失一般性地假设 K 整除 n；否则，我们考虑每组大小为 $\lfloor n/K \rfloor$ 的分组），使得在每次迭代 $t \leq K$ 中，$\mathcal{B}_t$ 都会产生一个包含 n/K 个元素的新集合，并且对于 $t > K$，$\mathcal{B}_t = \varnothing$。如果在第 t 步之前，至少有 $m - F_{t-1}$ 个子集被覆盖，但在第 t 步之后，被覆盖的集合少于 $m - F_t$ 个，则称算法在第 t 步失败。利用以下事实：当 n 足够大时，(i) 每个新添加的块中的列相互独立；(ii) $\mathbb{P}(\mathrm{Bin}(m - F_{t-1}, p) \geq f_t) \geq n^{-\varepsilon'}$；以及 (iii) $n/K \geq n^{1-\delta-\varepsilon'}$，我们得到

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

接下来，我们通过应用并集界得到该结果，

$$\begin{aligned} &K\mathbb{P}(\text{BlockGreedy 在最初的 } K \text{ 步中失败}) \leq \sum \mathbb{P}(\text{BlockGreedy 在第 t 步失败})_{t=1} \\ &\qquad \leq K \cdot \exp\left(-n^{1-\delta-2\varepsilon'} \right) \\ &\qquad \leq \exp\left(-n^{1-\delta-3\varepsilon'} \right), \end{aligned}$$

其中，第二个不等式成立是因为根据定义，算法运行 K 次迭代；对于足够大的 n，第三个不等式成立。我们证明了 BlockGreedy 能够成功找到至多 K 个元素，使得至多 $m - F_K$ 个集合仍未覆盖。由于根据构造，$m - F_K \leq K$，我们

可以轻松覆盖剩余的行，因为根据引理 2，整数规划以高概率可行。这证明了

$$\mathbb{P}\left(\mathrm{val}_{\mathtt{Alg}} \le 2K\right) \ge 1 - \exp\left(-n^{1-\delta-4\varepsilon'}\right) = 1 - \exp\left(-n^{1-\delta-\varepsilon}\right),$$

对于足够大的 n。回顾一下， $K \lesssim \mathrm{val}_{\mathrm{LP}}$对于 $mp \lesssim \log n$成立，而 $K \lesssim \mathrm{val}_{\mathrm{IP}}$对于 $mp \gg \log n$成立，证毕。□

有了上述结果，我们现在通过一种合适的归约来分析贪心算法。回顾一下，我们分别用 $\mathrm{val}_{\mathtt{Alg}}$ 和 $\mathrm{val}_{\mathtt{Gr}}$表示 BlockGreedy 和贪心算法的输出。

定理 2. 在假设 *1* 下，若 $\delta < 1/2$，则对于足够大的 n，有

$$\mathbb{P}\left(val_{\mathtt{Gr}} \sim val_{IP}\right) \ge 1 - \exp\left(-\sqrt{n}\right).$$

证明. 我们对 $\varepsilon = 1/8 - \delta/4$应用定理 1，并令 K、$\mathcal{B}_t$ 如定理 1 的证明中所定义。对于足够大的 n，有

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

其中 $\Delta := 3/4 - \delta/2 > 1/2$。
给定矩阵 $\boldsymbol{A}$，考虑对 $J := \exp(\sqrt{n})$ 运行上述 BlockGreedy 定义，每次都重新打乱列的顺序。$\mathrm{val}_{\mathtt{Alg}}$ 和 $\mathrm{val}_{\mathtt{Gr}}$ 都是随机变量，但在给定 $\boldsymbol{A}$ 的条件下，$\mathrm{val}_{\mathtt{Gr}}$ 是确定的，而 $\mathrm{val}_{\mathtt{Alg}}$ 仍取决于将列划分为块的随机性。利用并集界，我们得到

$$\begin{aligned}\mathbb{P}\left(\mathrm{val}_{\mathtt{Gr}} > 2K\right) \le\ & \mathbb{P}\left(\exists \text{ a failed copy of } \texttt{BlockGreedy}\right) \\ & + \mathbb{P}\left(\mathrm{val}_{\mathtt{Alg}} < \mathrm{val}_{\mathtt{Gr}} \text{ over all } J \text{ copies}\right).\end{aligned} \tag{6.4}$$

再次应用并集界，我们可以对（6.4）中的第一项给出上界：

$$\mathbb{P}\left(\exists \text{ a failed copy of } \texttt{BlockGreedy}\right) \le J\exp\left(-n^{\Delta}\right) = \exp\left(-n^{\Delta} + n^{1/2}\right). \tag{6.5}$$

现在我们聚焦于（6.4）中的第二项。令 v_1、v_2、…、v_g 为 Greedy 选取的元素的有序序列。令 $M_t := \{v_1 \in \mathcal{B}_1, v_2 \in \mathcal{B}_1 \cup \mathcal{B}_2, \ldots, v_t \in \mathcal{B}_1 \cup \ldots \cup \mathcal{B}_t\}$。事件 $\{\mathrm{val}_{\mathtt{Alg}} \ge \mathrm{val}_{\mathtt{Gr}}\}$contains the eve nt M_g，因为在这种情况下，BlockGreedy 必然会选取完全相同的列v_1、v_2、…、v_g。鉴于每次重新排列列都会使可能的 n 列划分上的 $\mathcal{B}_i$'s 生成均匀分布，我们得到，即

$$\mathbb{P}\left(M_g\right) = \mathbb{P}\left(v_1 \in \mathcal{B}_1\right)\mathbb{P}\left(v_2 \in \mathcal{B}_1 \cup \mathcal{B}_2 \mid M_1\right)\ldots\mathbb{P}\left(\boldsymbol{v}_g \in \mathcal{B}_1 \cup \ldots \cup \mathcal{B}_g \mid M_{g-1}\right).$$

上述乘积中的第 t 项等于

$$\mathbb{P}(\boldsymbol{v}_t \in \mathcal{B}_1 \cup \ldots \cup \mathcal{B}_t \mid M_{t-1}) = \frac{t\frac{n}{K} - (t-1)}{n - (t-1)} \ge \frac{t}{K} - \frac{t-1}{n} \ge \frac{t}{2(K-1)},$$

其中，最后一个不等式对 $n \ge 4K$ 成立（回想一下， $n \gg K$）。由于 $M_g \subset \{\mathrm{val}_{\mathtt{Alg}} \ge \mathrm{val}_{\mathtt{Gr}}\}$，我们可以如下界定后一事件的概率（注意，当 $g < K$ 时，乘积中的项数会更少，因此 $\mathbb{P}(M_g)$ 会更大），

$$\mathbb{P}\left(\mathrm{val}_{\mathtt{Alg}} \ge \mathrm{val}_{\mathtt{Gr}} \text{ for 1 copy}\right) \ge \mathbb{P}(M_g) \ge \prod_{t=1}^{K-1}\mathbb{P}(\boldsymbol{v}_t \in \mathcal{B}_1 \cup \ldots \cup \mathcal{B}_t \mid M_{t-1}) \ge \prod_{t=1}^{K-1}\frac{t}{2(K-1)} \ge e^{-2K},$$

其中我们在最后一个不等式中用到了 $k! \ge (k/e)^k$。由于 $K \le C\max\left\{\frac{m}{\mathbb{E}d_{\max}}, \frac{1}{p}\log(\underline{\ ^{mp}\log n})\right\}$ and$1/p \le n^{\delta}$，存在一个足够大的常数 $\tilde{C} > 0$，使得 $K \le \tilde{C}n^{\delta}\log n$。因此，利用

各副本之间重排的独立性，我们可以计算

$$\begin{aligned}\mathbb{P}\left(\mathrm{val}_{\mathtt{Alg}} < \mathrm{val}_{\mathtt{Gr}} \text{ over all } J \text{ copies}\right) &= \left(1 - \mathbb{P}\left(\mathrm{val}_{\mathtt{Alg}} \geq \mathrm{val}_{\mathtt{Gr}} \text{ for 1 copy}\right)\right)^J \\ &\leq (1 - e^{-2K})^J \\ &\leq \exp\left(-e^{\sqrt{n} - 2\tilde{C} n^{\delta} \log n}\right).\end{aligned} \tag{6.6}$$

结合 (6.4)、(6.5) 和 (6.6)，我们证明了对于足够大的 n， $\mathbb{P}(\mathrm{val}_{\mathtt{Gr}} > 2K) \leq \exp(-\sqrt{n})$，从而完成了证明。□

备注 2。 我们注意到，定理 2 中的 $\delta < 1/2$ 条件可能并非最优，并且可以通过采用经过更谨慎选择的集合 $\mathcal{B}_t$ 将问题归约到 BlockGreedy 来放宽该条件。特别是，合适的集合大小 $|\mathcal{B}_t|$ 在各个 $t \leq K$ 中可能并不相同。这种情况下的分析会更加复杂，我们将其视为一个值得探讨的开放方向。

7. 讨论与开放问题

我们的工作刻画了随机击中集问题中的乘法整数间隙。本节将讨论主要结果背后的直觉，以及开放问题和猜想。

7.1. 我们的结果与证明技巧总结。 我们发现，整数间隙的性质取决于包含集的大小，也可以将其视为底层超图的稀疏程度。具体而言，当顶点的平均度较小时，也就是说，每个元素属于的子集数量较少时，我们证明了线性规划与整数规划的解之间仅存在常数级间隙，并给出了一种简单的算法解法。超图变得稠密时，情况则有所不同：我们证明整数间隙会不断增大。这种差距主要源于二项分布的性质：当期望值较大时，随机变量的最大值与期望值以相同的方式增长；但如果 $mp \ll \log n$，最大值则会偏离期望值。

在分析 BlockGreedy 时，我们用几何级数来追踪这种行为变化。这意味着算法执行得越深入，我们选取的元素与平均元素之间的比值就越大。这与二项分布在平均度降低时的表现完全吻合：对于较大的实例，它近似呈高斯分布；但当平均度较小时，泊松近似开始占据主导地位，右尾变得更重， $d_{\mathtt{max}}$ 与 mp 之间的差距也随之扩大。我们的分析追踪了高斯行为向类泊松行为的转变。

7.2. 乘法整数间隙与加法整数间隙。 我们的结果只涉及乘法间隙，但分析中得到的常数可能很大。这或许是所研究问题具有一般性的结果。例如，如果只关注常数 p的情形（这立即意味着按照我们的刻画，该实例非常稠密），那么 [14] 证明，一个简单算法在以较小的加法误差近似整数规划方面是最优的。在更一般的情形下证明类似的常数上界，是一个有趣的开放问题。基于数值实验，我们提出以下猜想。

猜想 1 （极稀疏）。对于 $mp \ll 1$,

$$\frac{val_{Gr}}{val_{LP}} \to 1. \tag{7.1}$$

猜想 2 （稀疏）。对于 $1 \lesssim mp \ll$ 对数 n,

$$\frac{val_{Gr}}{val_{IP}} \to 1, \qquad \frac{val_{IP}}{val_{LP}} \to C_1, \tag{7.2}$$

其中 $1 < C_1 < 1.5$。

猜想 3（稠密）。对于 $mp \gg$ 对数 n,

$$\frac{val_{Gr}}{val_{IP}} \to C_2, \tag{7.3}$$

其中 $1 \le C_2 < 1.5$。

7.3. 线性规划解的分析。研究整数规划与线性规划之间的差距，以及线性规划解本身，一个动机是构造一种舍入方案，将分数解转换为整数解。我们认为，这也是未来值得探索的一个方向。具体而言，数值实验表明，分数解中取值较大的项，很可能对应于整数解中被选取的元素。这支持了这样一种观点：将贪心方法与线性规划方法相结合，或许有助于高效求解击中集问题。进一步研究的一种方法是先求解线性规划，用线性规划解中最大的元素初始化 $\boldsymbol{x}$，再通过贪心算法覆盖其余子集。

8. 致谢

作者感谢 Dylan J. Altschuler、Afonso S. Bandeira、Raphaël Barboni 和 Anastasia Kireeva 提供的宝贵讨论。DD 获得 ETH人工智能中心博士奖学金和 ETH 数据科学基础计划的支持。GA 获得剑桥信托基金和 Invenia Labs 的支持。NG 感谢 Elizaveta Rebrova 提供的资助。

参考文献

[1] S. Borst、D. Dadush、S. Huiberts 和 S. Tiwari。关于具有高斯数据的二元整数规划的整数间隙。数学规划，2022。[2] S. Borst、D. Dadush 和 D. Mikulincer。通过差异性分析随机整数规划的整数间隙。载于*2023 年 ACM-SIAM 离散算法研讨会论文集（SODA）*，2023。[3] D. Du、F. K. Hwang 和 F. Hwang。组合群组测试及其应用。World Scientific，2000。[4] Y. Erlich、A. Gilbert、H. Ngo、A. Rudra、N. Thierry-Mieg、M. Wootters、D. Zielinski 和 O. Zuk。线性码的生物筛查：理论与工具。*BioRxiv*，2015。[5] W. Feller。概率论及其应用导论。1957。[6] A. Fernández Anta、M. A. Mosteiro 和 J. Ramón Muñoz。多址信道中的无界竞争消解。*Algorithmica*，2013。[7] A. Hoorfar 和 M. Hassani。关于朗伯 W 函数和超幂函数的不等式。*J. Inequal. Pure and Appl. Math*，2008。[8] F. Iliopoulos 和 I. Zadik。群组测试与局部搜索：是否存在计算—统计差距？载于学习理论会议，2021。[9] D. S. Johnson。组合问题的近似算法。载于第五届 *ACM* 计算理论年度研讨会论文集，1973。[10] L. Lovász。关于最优整数覆盖与分数覆盖之比。离散数学，1975。[11] M. Mézard 和 M. Tarzia。击中集问题的统计力学。物理评论 *E*，2007。[12] V. T. Paschos。若干覆盖与装箱问题的近似最优解综述。*ACM* 计算调查（*CSUR*），1997。[13] P. Slavík。集合覆盖贪心算法的紧确分析。载于第二十八届 *ACM* 计算理论年度研讨会论文集，1996。[14] O. A. Telelis 和 V. Zissimopoulos。随机集合覆盖近似中的绝对 O(log m) 误差：平均情形分析。信息处理快报，2005。[15] R. Van Handel。高维概率。讲义，2014。

附录 A. 辅助引理

引理 9。（切尔诺夫界——上尾）设 $X_1,...,X_n$ 是取值于 $\{0,1\}$ 的相互独立随机变量， X 表示它们的和，且 $\mu = \mathbb{E}X$。那么对于任意 $\delta > 0$,

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

为了处理 $d_{\max}$ 围绕其期望的集中性，我们陈述一个关于方差张量化的有用结果。我们引入记号 Var_i 和 $\mathbb{E}_i$，其中下标 i 表示在底层随机向量除第 i 个分量以外的每个分量上进行条件化。

引理 10（定理 2。3, [15]）。设 X_1，...,X_n 为相互独立的随机变量，对于每个函数f： $\mathbb{R}^n \to \mathbb{R}$，定义

$$Var_i f(x_1,...,x_n) := Var\left(x_1,...,x_{i-1},X_i,x_{i+1},...,x_n\right).$$

那么，有如下结论:

$$Var\left(f\left(X_1,...,X_n\right)\right) \leq \mathbb{E}\sum_{i=1}^{n} Var_i f\left(X_1,...,X_n\right)$$

Le引理 11 (关于 $d_{\max}$ 的集中性)。设 $X_1,\ldots,X_n \overset{\text{独立同分布}}{\sim} Bin(m,p)$。那么，对于任意 $t > 0$,

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

备注 3。注意，在满足假设 1 的 m,p 的所有情形下，取 $t \sim \mathbb{E}d_{\max}$ 即足以根据前一个引理推出 $d_{\max} \sim \mathbb{E}d_{\max}$ 以高概率成立。

证明。由切比雪夫不等式，只需证明 Var(*d*max ≤ *mp*。根据引理 10，我们有

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

这正是所需结论。 □

引理 12（二项分布概率质量函数的渐近表达式）。设 $a \equiv a(n)$ 和 $b \equiv b(n)$ 满足

(1) $1 \ll b \ll \sqrt{a}$,

(2) $p \ll 1$。

如果对于 $C > 1$ 有 $b \geq Cap$，那么

$$\log \mathbb{P}(\mathrm{Bin}(\lceil a \rceil, p) = \lceil b \rceil) \geq -\left(b\log\frac{b}{ap} - b + ap\right)(1+o(1)), \tag{8.1}$$

如果还满足 $b \gg ap$，则有

$$\log \mathbb{P}(\mathrm{Bin}(\lceil a \rceil, p) = \lceil b \rceil) \geq -\left(b\log\frac{b}{ap}\right)(1+o(1)), \tag{8.2}$$

此外，将 $\lceil a \rceil$ 替换为 $\lfloor a \rfloor$ 后，所有界仍然成立。

证明。

$$\begin{aligned}\mathbb{P}(\mathrm{Bin}(\lceil a\rceil, p)=\lceil b\rceil) &= \binom{\lceil a\rceil}{\lceil b\rceil} p^{\lceil b\rceil}(1-p)^{\lceil a\rceil-\lceil b\rceil}\\ &\overset{(i)}{\geq} \frac{(\lceil a\rceil\, p)^{\lceil b\rceil}}{4(\lceil b\rceil)!}(1-p)^{\lceil a\rceil-\lceil b\rceil}\\ &\overset{(ii)}{\geq} \frac{1}{\lceil b\rceil}\left(\frac{\lceil a\rceil\, ep}{\lceil b\rceil}\right)^{\lceil b\rceil}(1-p)^{\lceil a\rceil-\lceil b\rceil}\\ &\overset{(iii)}{\geq} \frac{1}{\lceil b\rceil}\left(\frac{\lceil a\rceil\, ep}{\lceil b\rceil}\right)^{\lceil b\rceil} e^{-\lceil a\rceil p}(1-p)^{\lceil a\rceil p-\lceil b\rceil},\end{aligned}$$

其中，(i) 由 $\binom{n}{k}\geq\frac{n^k}{4k!}$ 对于 $0\leq k\leq\sqrt{n}$ 得出，(ii) 由 $n!\leq\frac{n}{4}(n/e)^n$ 对于足够大的 n 得出，而 (iii) 由 $(1+x/n)^n\geq e^x(1-x^2/n)$ 对于 $|x|\leq n$ 得出。取对数后，我们得到

$$\log\mathbb{P}(\mathrm{Bin}(\lceil a\rceil, p)=\lceil b\rceil)\geq -\left(\lceil b\rceil\log\frac{\lceil b\rceil}{\lceil a\rceil\, p}-\lceil b\rceil+\lceil a\rceil\, p+\log\lceil b\rceil-(\lceil a\rceil\, p-\lceil b\rceil)\log(1-p)\right).$$

若 $b\geq C\lceil a\rceil\, p$ 对于 $C>1$，则有

$$b\log\frac{b}{ap}-b+ap\geq\gamma b\gg 1,\qquad \text{for } \gamma:=\frac{1}{C}+\log C-1>0.$$

由于 $b\gg 1$，我们有

$$\frac{\lceil b\rceil\log\frac{\lceil b\rceil}{\lceil a\rceil p}-\lceil b\rceil+\lceil a\rceil\, p}{b\log\frac{b}{ap}-b+ap}=1+o(1).$$

现在，由于 $\log\lceil b\rceil\ll b$ 且 $(\lceil a\rceil\, p-\lceil b\rceil)\log(1-p)\ll b$ 对于 $p\ll 1$，因此有

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

此外，若 $b\gg ap$，则

$$\frac{b\log\frac{b}{ap}-b+ap}{b\log\frac{b}{ap}}=1+o(1),$$

最后，

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

Un在我们的假设 $1\ll b\ll\sqrt{a}$ 下，对于 $\log\mathbb{P}(\mathrm{Bin}(\lfloor a\rfloor, p)=\lfloor b\rfloor)$，同样的界也成立。 □

引理 13 (二项分布单调性)。设 $S_m\sim\mathrm{Bin}(m,p)$。那么对于 $r\geq mp$，有 $\mathbb{P}(S_m=r+1)\leq\mathbb{P}(S_m=r)$ 以及 $\mathbb{P}(S_{m-1}=r)\leq\mathbb{P}(S_m=r)$。

证明。 证明过程与 [5] 中给出的论证类似。

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

类似的论证表明，$\mathbb{P}(S_{m-1}=r)\leq\mathbb{P}(S_m=r)$。 □

附录 B. $mp \lesssim \log n$ 情况下的主要工具及引理 5 的证明

引理 14。 如果 $mp \lesssim \log n$，那么，对于任意 $\varepsilon > 0$，都存在常数 $\tau > 0$ 和 $1 < \alpha < \beta$，使得对于 $k \lesssim \log n$ 以及任意满足 $\beta^{-k-1}m \le \tilde{m} \le \beta^{-k}m$ 的 $\tilde{m}$，当 n 足够大时，

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

证明。 该证明本质上是对引理 12 的谨慎应用。设 τ、α、β 为稍后确定的常数，并设 $\tilde{m} = \lfloor \beta^{-k-1}m \rfloor$。根据我们满足 $mp \ll \log n$ 还是 $mp \sim \log n$ 的情况，引理 12 中的渐近表达式将由不同的项占主导。

我们从 $mp \ll \log n$ 的情形开始。由引理 6 可知，这意味着 $mp \ll \mathbb{E}d_{\max} \ll \log n$。这里我们可以固定 $\alpha \equiv 2$ and $\beta \equiv 3$。对 $a = 3^{-k-1}m$ and $b = (2/3)^k \tau \mathbb{E}d_{\max}$ 应用 (8.2)，得到：

$$\log \mathbb{P}\left(\mathrm{Bin}(\tilde{m}, p) = \left\lceil (2/3)^k \tau \mathbb{E}d_{\max} \right\rceil\right) \ge -(2/3)^k \tau \mathbb{E}d_{\max} \log\left(\frac{2^k 3 \tau \mathbb{E}d_{\max}}{mp}\right)(1 + o(1)) \tag{8.3}$$

回想一下，我们的目标是证明 $\log \mathbb{P}\left(\mathrm{Bin}(\tilde{m}, p) = \left\lceil (2/3)^k \tau \mathbb{E}d_{\max} \right\rceil\right) \ge -\varepsilon \log n$。我们首先证明存在满足以下两个不等式的 $\tau > 0$：

$$\begin{aligned} &(i) \qquad (2/3)^k \tau (\log 3 + k \log 2) \frac{\mathbb{E}d_{\max}}{\log n} \;\le\; \frac{\varepsilon}{4}, \\ &(ii) \qquad (2/3)^k \tau \frac{\mathbb{E}d_{\max}}{\log n} \log\left(\frac{\mathbb{E}d_{\max}}{mp}\right) \;\le\; \frac{\varepsilon}{4}. \end{aligned} \tag{8.4}$$

事实上，由于 $\mathbb{E}d_{\max} \ll \log n$ and $k \ll (3/2)^k$，对于足够大的 $\tau > 0$，不等式 (i) 对任意 n 都成立。对于 (ii)，我们需要使用 $\mathbb{E}d_{\max}$ 的显式界。特别地，根据引理 6，我们知道存在 $C > 0$，使得当 n 足够大时，$\mathbb{E}d_{\max} \le C \log n/(\log\log n - \log mp)$。将其代入 (ii)，对于 $k = 0$，我们得到：

$$\tau \frac{\mathbb{E}d_{\max}}{\log n} \log\left(\frac{\mathbb{E}d_{\max}}{mp}\right) \le \frac{\tau C(\log C + \log\log n - \log(\log\log n - \log mp) - \log mp)}{\log\log n - \log mp} = \tau C + o(1). \tag{8.5}$$

对于 $\tau = \varepsilon/(8C)$，当 n 足够大时，(ii) 对 $k = 0$ 成立。增大 k 只会减小 (ii) 的左边，因此，同一个 τ 对任意 $k \ge 0$ 都适用。

最后，通过将（i）与（ii）相加，我们证明了当 n 足够大时，

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

从而完成了对 $mp \ll \log n$ 情形的证明。

现在我们关注 $mp \sim \log n$ 的情形。在这里，我们对 $a = \beta^{-k-1}m$ 和 $b = (\alpha/\beta)$ 的取值应用（8.1），同时注意条件 $b \ge Cap$ 及 $C > 1$。于是有

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

我们取 $\tau = \gamma mp/\mathbb{E}d_{\max}$，其中 $\gamma > 1$ 是稍后指定的某个常数。注意，这样一来，应用（8.1）的条件 $\frac{b}{ap} \ge C > 1$ 得以满足，因为 $\frac{b}{ap} \ge \frac{\tau \mathbb{E}d_{\max}}{mp} = \gamma > 1$。这将后一个表达式化简为：

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

由于在此范围内，对于某个 $D > 0$，我们有 $mp \le D \log n$，当 n 足够大时，只需证明

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

我们首先证明存在常数 $1<\alpha<\beta$ 和 $\gamma>1$，它们取决于 ε 和 D，并且对任意 $k\geq 0$ 都满足以下两个不等式:

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

注意，(i) 的左侧会随着 k 的增加而减小，因此只需考察 $k=0$。我们需要证明，存在取决于 ε、D 的 β、$\gamma>1$，使得

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

注意，只要满足 $\beta\gamma>1$，$\frac{\partial f}{\partial\beta}=\gamma/\beta-1/\beta^2>0$ 和 $\frac{\partial f}{\partial\gamma}=\log\beta\gamma>0$。由于 $f(1,1)=0$，我们可以找到足够接近 1 的 β、$\gamma>1$，使得 $f(\beta,\gamma)\leq\varepsilon/(4D)$。我们采用 β 和 γ 的这些取值（或等价地，采用 τ）。由于 $k\ll(\beta/\alpha)^k$，存在 $\alpha\in(1,\beta)$，使得 (ii) 成立。将 (i) 与 (ii) 相加可知，当 n 足够大时，

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

我们证明了，对于 $mp\lesssim\log n$，对于任意 $\varepsilon>0$，当 n 足够大时，存在 τ,α,β，使得

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

由于 $\beta^{-k-1}mp<\beta^{-k}mp<\lceil(\alpha/\beta)^k\tau\mathbb{E}d_{\max}\rceil$，根据二项分布单调性（引理 13），对于任意满足 $\beta^{-k-1}m\leq\tilde{m}\leq\beta^{-k}m$ 的 $\tilde{m}$，我们有，

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

□

引理 *5* 的证明。我们沿用引理 14 中的论证，将 $k=0$ 和 $\mathbb{E}d_{\max}$ 替换为 $\log n/(\log\log n-\log mp)$。注意，在引理 14 的证明中，当 $mp\ll\log n$ 时，我们只用到了 $mp\ll\mathbb{E}d_{\max}\ll\log n$，以及对于某个 $C>0$，$\mathbb{E}d_{\max}\leq C\log n/(\log\log n-\log mp)$。由于将 $\mathbb{E}d_{\max}$ 替换为 $\log n/(\log\log n-\log mp)$ 后，这两个性质仍然成立，因此证明得证。由于在引理 5 的设定下 $\tau=\varepsilon/(8C)$，且在此论证中 $C=1$，我们取 $\tau=\varepsilon/8$。□

附录 C. 引理 8 的证明

$\tilde{}$ $\tilde{}$我们首先证明存在 K，使得 $m-F_{\tilde{K}}\lesssim K$；然后，$\tilde{}$将 K 乘以一个乘法因子，我们便可找到 K，使得 $m-F_K\leq K$。

情形 $mp\lesssim\log n$**.** 由引理 14，存在常数 $\tau>0$、α、β，满足 $1<\alpha<\beta$，使得对于任意满足 $\beta^{-k-1}m\leq\tilde{m}\leq\beta^{-k}m$ 的 $\tilde{m}$，当 n 足够大时，

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

回顾一下，在这种情况下 $f_t=\lceil(\alpha/\beta)^k\tau\mathbb{E}d_{\max}\rceil$，其中 k 满足 $\beta^{-k-1}m\leq m-F_{t-1}\leq\beta^{-k}m$ 和 $F_t=\sum_{s=1}^{t}f_s$。由引理 14 可知 $\mathbb{P}(\mathrm{Bin}(m-F_{t-1},p)=f_t)\geq n^{-\varepsilon}$。我们的目标是证明存在 $s\lesssim\mathrm{val}_{\mathrm{LP}}\sim m/\mathbb{E}d_{\max}$，使得 $m-F_s\lesssim s$。

引理 15。设 $t^{(k)}:=\frac{\beta-1}{\beta\tau}\frac{m}{\mathbb{E}d_{\max}}\alpha^{-k}$。

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

通俗地说，如果在执行 $t-1$ 步`BlockGreedy`后，至多有 $\beta^{-k}m$ 个子集未覆盖，那么在执行 $t+t^{(k)}-1$ 步后，至多仍有 $\beta^{-k-1}m$ 个子集未覆盖。

证明。令 $s \geq t$。只要 $m - F_{s-1} > \beta^{-k-1}m$，我们就始终有 $f_s = \lceil(\alpha/\beta)^k \tau \mathbb{E}d_{\max}\rceil$。下面通过反证法进行证明。假设 $m - F_{t+t^{(k)}-1} > \beta^{-k-1}m$。这意味着对于所有$s \in [t-1, t+t^{(k)}-1]$，都有 $f_s = f := \lceil(\alpha/\beta)^k \tau \mathbb{E}d_{\max}\rceil$。因此，

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

and

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

因此，我们必须有 $m - F_{t+t^{(k)}-1} \leq \beta^{-k-1}m$。 □

注意，我们始终有 $\beta^{-1}m \leq m - F_0 = m$。若从 $k=0$ 开始连续应用引理 15，则对于 $v(k) := \sum_{s=0}^{k} t^{(s)}$，我们有 $m - F_{v(k)-1} \leq \beta^{-k-1}m$。因此，对于 $k := \text{对数}\,\mathbb{E}d_{\max}\text{对数}\,\beta$，我们有 $m - F_{v(k)-1} \leq \frac{m}{\mathbb{E}d_{\max}}$。我们可以界定

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

由引理 7，我们以高概率有 $d_{\max} \lesssim \mathbb{E}d_{\max}$。结合引理 1，可推出$\tilde{}$$\mathsf{val}_{\mathtt{LP}} \geq \frac{m}{d_{\max}} \gtrsim \frac{m}{\mathbb{E}d_{\max}}$。现在，若我们选取 $K := v(k) \lesssim \frac{m}{\mathbb{E}d_{\max}}$，则有 $\mathsf{val}_{\mathtt{Alg}} \lesssim \frac{m}{\mathbb{E}d_{\max}}$。由于 $\mathsf{val}_{\mathtt{LP}} \leq \tilde{}\,\mathsf{val}_{\mathtt{Alg}}$，因此我们有 $K \sim \mathsf{val}_{\mathtt{LP}}$ 以及 $m - F_{\tilde{K}} \lesssim K$。

情况 $mp \gg \log n$**.** 这里有 $\mathbb{E}d_{\max} = mp(1+o(1))$，因此，选择一个命中平均数量子集的元素，与选择一个命中接近最大数量子集的元素大致相同。根据二项分布的均值和中位数的性质可知，对于任意 $\tilde{m}$，$\mathbb{P}(\text{二项分布}(\tilde{m}, p) \geq \lceil \tilde{m}p \rceil) \geq 1/3$。

我们先考虑 $\log mp \ll \log n$ 的情况。这意味着 mp 无法随 n 多项式增长，但例如，$mp \sim \log^2 n$ 是可能的。在此范围内，$\mathsf{val}_{\mathtt{IP}} \sim \frac{1}{p}\log\left(\frac{mp}{\log n}\right)$。设 $K_1 = \left\lceil \frac{1}{p}\log\left(\frac{mp}{\log n}\right)\right\rceil$，且 $f_1, \ldots, f_{K_1}$ 是一个满足 $f_s = \lceil mp(1-p)^s \rceil$ 的序列。那么，我们有 $m - F_{K_1} \leq m(1-p)^{K_1} \leq \frac{1}{p}\log n$。$\tilde{}$因此，$(m - F_{K_1})p \sim \log n$，并且我们可以沿用上一节中的 f_t，其中 $mp \sim \log n$，且 $\tilde{F}_t := \sum_{s=1}^{t} \tilde{f}_t$。对于这个序列 $\tilde{f}_1, \ldots, \tilde{f}_{K_2}$，我们有 $K_2 \lesssim \frac{1}{p}$，且 $m - F_{K_1} - \tilde{F}_{K_2} \lesssim \frac{1}{p} \ll \frac{1}{p}\log\left(\frac{mp}{\log n}\right)$。所需结论对组合序列 f_t 、 $\tilde{f}_t$ 和 $\tilde{K} := K_1 + K_2$ 成立。最后，我们研究 $\log mp \gtrsim \log n$ 的情况，这意味着 $\mathsf{val}_{\mathtt{IP}} \sim \frac{1}{p}\log n$。这种情况显然成立，因为可以选取 $\tilde{K} = \left\lceil \frac{1}{p}\log\left(\frac{mp}{\log n}\right)\right\rceil \lesssim \mathsf{val}_{\mathtt{IP}}$，以及 $f_1, \ldots, f_{\tilde{K}}$ 作为一个满足 $f_s = \lceil mp(1-p)^s \rceil$ 的序列。$\tilde{}$接着，我们有 $m - F_{\tilde{K}} \leq m(1-p)^K \leq \frac{1}{p}\log n \lesssim \mathsf{val}_{\mathtt{IP}}$。

$\tilde{}$**从** $m - F_{\tilde{K}} \lesssim K$ **到** $m - F_K \leq K$。最后，利用引理 2 可知，除非 $F_t = m$，否则 $f_t \geq 1$，因此$\tilde{}$存在某个常数 $C > 0$，使得对于 $K := CK$、$F_K \geq m - K$，从而完成证明。