arXiv:2301.05658v1 [cs.CC] 13 Jan 2023

流式下界与非对称集合不相交问题

Shachar Lovett * Jiapeng Zhang +计算机科学系 加利福尼亚大学 圣迭戈分校计算机科学系 南加利福尼亚大学 slovett@ucsd.edu jiapengz@usc.edu

2023年1月16日

摘要

数据流中的频率估计是流式算法中的经典问题之一。经过大量研究，对于数据流由对手控制的情形，样本数量与算法空间复杂度之间所需权衡的上界和下界如今几乎匹配。然而，对于数据流按随机顺序给定或具有随机性的情形，目前只有较弱的下界。在本文中，我们将这一差距缩小到对数因子以内。

为此，我们考虑针问题。这是一个自然的频率估计困难问题，曾在（Andoni 等，2008年；Crouch等，2016年）中研究。这里的目标是区分两种由 t 个样本组成的数据流上的分布。第一种是在足够大的定义域上均匀分布。第二种是植入模型：均匀随机选取一根秘密“针”，随后数据流中的每个元素以 p 的概率等于这根针，否则就在定义域中均匀随机选取。设计流式算法，利用 $s \approx 1/(p^2 t)$ 的空间区分这两种分布并不困难。由于已有下界较弱，这一界是否紧致尚不明确。我们弥合了这一差距，并证明这种权衡在相差一个对数因子的范围内近乎最优。

我们的证明构建并拓展了流式算法与通信复杂度之间的经典联系，具体而言，是多方唯一集合不相交问题。我们引入了两个新要素，从而能够证明紧确的界。第一项是多方唯一集合不相交问题非对称版本的下界：参与者收到大小不同的输入集合，且每位参与者的通信量都按其输入长度进行归一化。第二项是一种组合技术，通过先采样区间，再在每个区间内采样一根均匀针，从而能够在植入模型中对针进行采样。

*本研究由美国国家科学基金会资助，项目编号为 1953928 和 2006443。

+本研究由美国国家科学基金会职业奖资助，项目编号为 2141536。

1 引言

这个针问题是在随机流的流式算法研究中探讨的一个基本问题[AMS99，AMOP08，GH09，CMVW16，BVWY18]。目标是使用一种节省空间的单遍流式算法，区分从两种可能的底层分布中采样得到的数据流。

设定记号：令 t 表示样本数，s 表示流式算法的空间，n 表示域大小，$p \in \in$ (0,1) 表示针概率。两个底层分布如下：

- 均匀分布：</style}从 $[n]$ 中抽取 t 个均匀分布的元素。
- Planted：</style}令 $x \in [n]$ 为均匀随机选取的元素（即"针"）。抽取 t 个元素，其中每个元素独立地以 p = x 的概率取该元素，否则从 $[n]$ 中均匀抽取。

我们假设 $n = \Omega(t^2)$，因此以高概率，数据流中的所有元素（植入模型中的针除外）均互不相同。问题在于，需要多少空间才能以高概率区分这两个模型。

针问题的样本-空间权衡。我们先介绍针问题的一些基本流式算法，以建立直观认识。首先注意，我们需要 $p = \Omega(1/t)$；否则，由于植入模型中以高概率不会出现针，因此这两个分布在统计上很接近。

一种可能的算法是检查数据流中是否存在两个相邻且相等的元素。这需要 $t = \Theta(1/p^2)$ 个样本和 $s = \Theta(\log n)$ 的空间。另一种可能的算法是将整个数据流存储在内存中，并检查是否有重复元素。这种算法所需的样本较少，为 $t = \Theta(1/p)$，但所需空间较多，为 $s = t\log\ n$。注意，在这两种情况下，样本与空间的权衡都是 $st = \Theta((\log n)/p^2)$。我们可以在这两种基本算法之间进行插值，但它们的乘积 st 始终保持不变。这启发了以下猜想，该猜想在 [CMVW16] 中被明确提出，并在[AMOP08]中被隐含提出。

猜想 1.1 (针问题的样本—空间权衡)。任何单遍流式算法，若能以高概率区分均匀分布模型和植入模型，其中 p 为针概率，t 为样本数，s 为空间，则满足 $p^2st = \Omega(1)$。

迄今为止，关于猜想 1.1 的最佳结果由 Andoni 等人 [AMOP08] 给出，他们证明了 $p^{2.5}st^{1.5} = \Omega(1)$（这个界确实较弱，因为 $p = \Omega(1/t)$）。Guha 等人 [GH09] 声称证明了猜想 1.1，但后来发现证明中存在一个错误，详见 [CMVW16]。本文证明了猜想 1.1，误差仅相差对数因子。我们还可以处理多次遍历数据流的流式算法，其复杂度随遍历次数线性增长。

定理 1.2 (主定理)。任何 ℓ 遍流式算法，若能以高概率区分均匀分布模型与植入模型，其中 p 为针概率，t 为样本数，s 为空间，则满足 $\ell p^2 st\log(t) = \Omega(1)$。

1.1 应用：随机流中的频率估计下界

对于许多流式问题，在对抗模型中，当前最先进的流式算法所需的空间已知是紧的（相差至多多对数项），其中数据流以对抗顺序到达。在一系列关于随机顺序模型的研究之后 [MP80、DLOM02、GM07、CCM08、CJP08、AMOP08、GM09],Crouch 等人 [CMVW16] 开创了对随机流的研究，其中数据流从某个底层分布中抽样得到。问题在于，在这种模型中，利用数据流的随机性质，能否获得优于对抗模型的流式算法；或者，现有下界能否也加强到这一模型。我们所描述的针问题就是随机模型中的一个问题示例。

流式处理文献中的一个基本问题，最早由以下开创性工作提出：[AMS99],，即估计数据流的频率矩。给定一个由 x_1，……，x_t中的元素组成的数据流 $[n]$，令 f_x 表示元素 x 在数据流中出现的次数。数据流的第 k阶频率矩为

$$F_k = \sum_{x \in [n]} f_x^k.$$

在对抗模型中，流式算法近似 F_k [CKS03, IW05] 所需空间的上界和下界相匹配，均为 $\tilde{\Theta}(n^{1-2/k})$ [1] 。 [CMVW16] 猜想，在随机模型中同样存在这一下界。他们证明，[AMOP08] 的结果给出了一个稍弱的下界，即需要 $\tilde{\Omega}(n^{1-2.5/k})$ 空间；并且猜想 1.1 若成立，则意味着紧确下界为 $\tilde{\Omega}(n^{1-2/k})$。因此，定理 1.2 验证了他们的猜想（相差对数因子），这仍然意味着下界为 $\tilde{\Omega}(n^{1-2/k})$。更多细节请参阅 [CMVW16] 。

我们注意到另一个相关应用，这是 David Woodruff 告知我们的。Mc- Gregor 等人 [MPTW12] 研究了基于对数据流进行子采样的流式算法。他们研究的问题之一尤其是频率估计。他们设计了基于子采样且空间高效的流式算法，并基于 Guha 等人的结果给出了匹配的下界。[GH09]然而，由于后来发现后续工作中存在一个错误，McGregor 等人的期刊版本 [MPTW16]删除了这些下界。利用定理 1.2 ，所声称的下界在相差一个对数因子的范围内成立。

1.2 证明方法

我们通过归约到通信复杂性中的唯一集合不相交问题，证明定理 1.2 。这是证明流式算法下界的常用技术 [CKS03、BYJKS04、AMOP08、GH09、BVWY18、KPW21]。

[1] 我们使用 $\tilde{\Theta}\,\tilde{\Omega}$ 来忽略多对数项。

基本思路是将流样本划分为若干区间 I_1，……，I_k，并考虑这样一种流分布：在每个区间中均匀放置一根针，而流中的其他元素则从均匀分布中采样。显然，任何能够使用 s 空间区分该分布与均匀分布的流式算法，都可以用来构造一个通信协议，以解决 k方唯一集合不相交问题，其中参与方 i 获得一个大小为 $|I_i|$的集合，并且每个参与方发送 s 比特。如果我们例如令各区间大小相等 $|I_1| = \ldots = |I_k| = t/k$，那么利用多方唯一集合不相交问题现有的紧下界，就能证明对抗模型中的紧样本空间下界[2]。该领域许多先前研究采用了这一方法 [CKS03、BYJKS04、AMOP08、GH09、BVWY18、KPW21]。我们的计划是将这一方法推广到随机模型中。然而，这会带来两个新的挑战。

首先，通过简单计算可知，针的数量以高概率为 $k \approx pt$ ，但针之间的间隔并非均匀分布；例如，最近的两根针之间的间隔为 $\approx p^2t$。因此，如果仍计划在每个区间内放置一根针，就必须采用长度差异很大的区间。相应地，我们需要证明多方唯一集合不相交问题的下界，其中参与者收到的输入长度各不相同。在这种模型下，再衡量协议的总通信量便没有意义了。为此，我们提出了一种新的度量方式，将每位参与者的通信量按其输入长度进行归一化。我们将在第1.3节进一步讨论这一点。

第二个挑战是，仅用一个区间划分数据流，再在每个区间内植入一根均匀针，无法得到植入针分布。相反，我们需要精心构造一个区间集合上的分布，使得在每个区间内放置一根均匀针后，得到的数据流分布恰好与植入分布一致。我们将在第1.4节进一步讨论这一点。

1.3 不同集合大小下的多方唯一集合不相交问题

我们首先定义标准多方唯一集合不相交问题。设 $k \geq 2$表示参与者人数。参与者的输入是集合 S_1、……、$S_k \subset [n]$。已知以下两种情况之一成立：

- **不相交**：集合 S_1、……、S_k 两两不相交。
- **唯一交集**：存在一个公共元素 $x \in S_1 \cap \ldots \cap S_k$，且集合$S_1 \setminus \{x\}$、……、$S_k \setminus \{x\}$ 两两不相交。

他们的目标是在尽量减少通信量[3]的同时，判断属于哪种情况。

注意，在这两种承诺情形中的任意一种下，其中一位参与方的输入大小为$|S_i| \leq n/k + 1$。一种简单的协议是由这位参与方发送其输入，这样其他

[2]具体而言，该协议的总通信量为 ks，而 k方唯一集合不相交问题的下界为 $\Omega(t/k)$。因此 $ks = \Omega(t/k)$。取 $k = pt$ 可得 $p^2st = \Omega(1)$。[3]形式化地说，我们考虑黑板模型中的随机多方协议：每一轮中，一位参与者在所有参与者都能看到的公共黑板上写下一条消息。

参与者就可以自行解决问题。这种简单的协议会发送 $O(n/k \cdot$对数 $n)$比特。利用 $O(n/k)$ 比特的通信量，还可以通过 [HW07] 的技术进一步改进。研究下界的一系列工作包括 [AMS99、BYJKS04、CKS03、Gro09、Jay09、YZ22] 。紧下界最早由[Gro09]证明。

定理1.3 ([Gro09，Jay09])。 任何能够解决 k方唯一集合不相交问题的随机通信协议，都必须发送$\Omega(n/k)$ 比特。

如第 1.2节所述，我们需要一种更精细的唯一集合不相交问题变体，其中集合大小固定，但参与者之间可以不同。

定义1.4 (固定大小多方唯一集合不相交问题)。 令 s_1，.....， $s_k \geq 1$。$[s_1, \ldots, s_k]$大小的 k方唯一集合不相交问题，是将 k方唯一集合不相交问题限制为输入集合大小为 $|S_i| = s_i$的情形。

考虑 $[s_1, \ldots, s_k]$大小的 k方唯一集合不相交问题的协议。对于任意 $i \in [k]$，一种做法是由第 i位参与方将其输入发送给其他参与方，这需要发送 $c_i = \Omega(s_i)$比特。如果输入大小 s_1，.....， s_k 差异很大，那么再考虑参与方发送的比特总数就没有意义了。相反，我们应当将第 i位参与方 c_i 发送的比特数按其输入长度 s_i进行归一化。我们证明，经过这种归一化后，简单协议确实是最优的。

为此，我们作如下定义：如果对于 Π 的任意通信记录，第 i 个参与方发送的比特数至多为 c_i，则称 k 方协议 Π 为 $[c_1, \ldots, c_k]$-有界协议。

定理 1.5 （固定大小多方唯一集合不相交问题的下界）。 设 Π 是一个随机化的 k 方 $[c_1, \ldots, c_k]$-有界协议，它以高概率解决 $[s_1, \ldots, s_k]$ 大小的 k 方唯一集合不相交问题，其中 $\sum s_i \leq n/2$。那么

$$\sum_{i \in [k]} \frac{c_i}{s_i} = \Omega(1).$$

我们以三点评论结束本小节。首先，条件 $\sum s_i \leq n/2$ 是证明技术中出现的一个技术性条件；它足以满足我们的应用需求，而且我们认为未来的工作可以将其去除。

其次，已知标准多方唯一集合不相交问题的困难情形是所有集合的大小大致相同，即 $s_1 = \ldots = s_k = \Theta(n/k)$。在这种情况下，定理 1.5 推出 $\sum c_i = \Omega(n/k)$，从而得到定理 1.3。

最后，我们通过构造输入上的一个困难分布，并证明在该分布下确定性协议的下界，来证明定理 1.5。这个困难分布很自然，即输入大小为 s_1、.....、 s_k 的均匀分布。详情请参见定理 2.13。此外，我们证明（命题2.15）定理 1.5 与定理 2.13实际上等价。

1.4 将针问题高效归约到多方唯一集合不相交问题

我们通过将针问题的下界归约为唯一集合不相交问题的下界，再应用定理 1.2 来证明定理 1.5 （更准确地说，是定理 2.13 ）。为此，我们需要一种方法，将唯一集合不相交问题的输入映射为流式算法的输入。例如，[AMOP08], 采用的一种自然方法是将数据流划分为若干区间，并为每个参与方分配一个区间。我们沿用这一思路并加以推广，从而能够用唯一集合不相交问题的随机输入来模拟针问题的植入分布。

回顾一下， n 表示域大小， t 表示样本数， p 表示针概率。我们的目标是使用多方唯一集合不相交问题的输入来模拟植入分布。为此，我们定义区间系统。

定义 1.6 (区间系统)。 区间系统 F 是由两两不相交的非空区间 $F = \{I_1, \ldots, I_k\}$ 组成的族，其中 $I_1, \ldots, I_k \subset [t]$。

给定区间系统 F，我们定义数据流上的植入分布 Planted$[F]$ $X \in [n]^t$ ，如下所示：

1. 采样均匀针 $x \in [n]$;
2. 在每个区间 $I \in F$ 中均匀采样索引 $i \in I$ 并设置 $X_i = x$;
3. 从 $[n]$ 中均匀采样所有其他数据流元素。

利用定理 1.5，我们证明了能够区分 F 的均匀分布和植入分布的流式算法的空间下界。在这里，我们利用了这样一个事实：即使参与者的集合大小各不相同，我们也能证明唯一集合不相交问题的下界。我们使用以下记号：给定区间系统 F，其值记为 val(F) $= \sum_{I \in F} \frac{1}{|I|}$。

引理 1.7。 设 F 是一个区间系统。任何能够以高概率区分 *Planted*$[F]$ 和均匀分布的流式算法，都必须使用如下大小的空间

$$s = \Omega\left(\frac{1}{val(F)}\right).$$

为了完成归约，我们需要使用针对区间系统 F 的植入分布来模拟植入分布。显然，无法仅用一个区间系统做到这一点，因此我们需要考虑随机化的区间系统。

随机区间系统 $\mathcal{F}$ 是定义在区间系统上的一个分布 F。植入分布 Planted$[\mathcal{F}]$ 对于 $\mathcal{F}$ 的定义如下：首先采样 $F \sim \mathcal{F}$，然后 $X \sim$ Planted$[F]$。 $\mathcal{F}$的值为 val$(\mathcal{F}) = \mathbb{E}_{F \sim \mathcal{F}}[val(F)$]。我们可以将引理 1.7 推广到随机区间系统。

引理 1.8。 设 $\mathcal{F}$ 是一个随机区间系统。任何能够以高概率区分 $Planted[\mathcal{F}]$ 与均匀分布的流式算法，都必须使用空间

$$s = \Omega\left(\frac{1}{val(\mathcal{F})}\right).$$

为了证明针问题的下界，我们需要让 Planted$[\mathcal{F}]$ 精确模拟植入分布；我们将这类随机区间系统称为完美。

定义 1.9（完美随机区间系统）。 若随机区间系统 $\mathcal{F}$ 满足 $Planted[\mathcal{F}]$ 的分布与植入分布完全相同，则称其为完美 。

根据引理 1.8，我们需要一个值尽可能低的完美随机区间系统 $\mathcal{F}$。相对容易证明，若 $\mathcal{F}$ 是完美的，则 $\text{val}(\mathcal{F}) = \Omega(p^2t)$。下述定理给出了一种几乎达到该下界的构造。

定理 1.10。 存在一个完美随机区间系统 $\mathcal{F}$，其 $val(\mathcal{F}) = O\left(p^2t\log(t)\right)$。

定理 1.2 可由引理 1.8 和定理 1.10直接结合得出。

1.5 相关工作

在一项开创性工作中，Miltersen 等人 [MNSW95]首次发现了非对称通信复杂度与其在单元探测模型中应用于数据结构之间的联系。此后，若干工作[BR00、JKKR03、PT06、BIPW10、CKLM18] 通过与非对称通信复杂度下界的联系，证明了数据结构下界和流式下界。据我们所知，这些工作都建立在双方通信问题之上。与之不同，我们在本文中研究多方通信复杂度。多方通信能否为数据结构和流式下界带来更多应用，是一个有趣的问题。

除了与数据结构下界和流式下界的联系之外，Dinur 等人 [DDKS16] 还研究了密码学中的针问题。探索我们的工作与密码学之间更多的联系，将会很有意思。

致谢。我们感谢 David Woodruff 就流式算法展开的有益讨论，并感谢他对本文早期版本提出的宝贵意见。

论文结构。 我们在第 1.5 节证明了多方唯一集合不相交问题（2 定理）的固定集合大小下界。我们在第 1.7 和 1.8 节设计了一种利用区间系统的高效归约（引理 3）。我们结合这两者，在第 1.2 节证明了针问题的下界（定理 4）。我们在第 5 节讨论开放问题。

2 Low非对称唯一集合不相交问题的下界 s

我们在本节证明定理 1.5 。首先，我们回顾一些定义，并固定一些记号。

记号。为方便起见，我们将集合与其指示向量等同起来；因此，我们将 $X \in \{0,1\}^n$ 与集合 $\{i : X_i = 1\} \subset [n]$ 等同。令 $k \geq 2$ 表示参与者的数量。参与者的输入为 $X = (X_1, \ldots, X_k)$，其中 $X_i = (X_i(1) \ldots, X_i(n)) \in \{0,1\}^n$。我们还将方便地定义 $X^j = (X_1(j) \ldots, X_k(j)) \in \{0,1\}^k$，表示所有参与者的第 j 个坐标，其中 $j \in [n]$。本节中使用粗体表示随机变量（例如 $\boldsymbol{X}$、$\boldsymbol{W}$），以便将它们与非随机变量区分开来。

协议。设 Π 是一个协议。给定输入 X，我们用 $\Pi(X)$表示在 X上运行 Π 所得的通信记录。我们假设每条通信记录也有一个由其确定的比特输出值（例如，发送的最后一个比特）。如果某个协议在输入分布 ν 下以错误率 δ解决一个判定问题，即当输入从 ν中抽取时，它以至少 $1-\delta$的概率输出正确答案。我们将证明，对于若干种输入分布，解决唯一集合不相交问题的协议存在下界。因此，除非另有说明，我们可以假设这些协议都是确定性的。

最后，回顾一下，我们称 k方协议 Π 是 $[c_1, \ldots, c_k]$有界的，如果在 Π的任何通信记录中，第 i个参与方发送的比特数至多为 c_i。

多方唯一集合不相交问题。 k方唯一集合不相交问题定义在来自两个承诺集合的输入上：

- **不相交**： $\mathcal{F}^0 = \{X \in (\{0,1\}^n)^k : \ \forall j \in [n], \ |X^j| \leq 1\}$,
- **唯一交集 相交**： $\mathcal{F}^1 = \{X \in (\{0,1\}^n)^k : \ \exists j \in [n], \ |X^j| = k, \ \forall j' \neq j, \ |X^{j'}| \leq 1\}$.

为了证明定理1.5，我们的第一步是考虑乘积分布下的唯一集合不相交问题，其中参与者之间的权重分配是不对称的。

2.1 非对称乘积分布的下界 s

设 ν 是 $[k]$上的一个分布。我们将 ν^n 记为 $\boldsymbol{W} \in [k]^n$上的分布，其中对所有 $j \in [n]$，我们独立地采样 $\boldsymbol{W}_j \sim \nu$。我们定义两个分布， $\mu^0_{\text{prob}}[\nu]$支撑在 $\mathcal{F}^0$上， $\mu^1_{\text{prob}}[\nu]$支撑在 $\mathcal{F}^1$上。

定义 2.1（不相交非对称分布）。设 $\boldsymbol{X} \in (\{0,1\}^n)^k$ 按如下方式采样：

1. 对 $\boldsymbol{W} \sim \nu^n$进行采样。
2. 对每个 $j \in [n]$，如果 $\boldsymbol{W}_j = i$，则从 $\boldsymbol{X}_i(j) \in \{0,1\}$ 中均匀采样，并对所有 $i' \neq i$设置 $\boldsymbol{X}_{i'}(j) = 0$。

我们将 $\boldsymbol{X}$的边缘分布记为 $\mu^0_{prob}[\nu]$，并注意到其支撑集为 $\mathcal{F}^0$。

定义 2.2（唯一交集非对称分布）。令 $\boldsymbol{Y} \in (\{0,1\}^n)^k$ 按如下方式采样：

1. 采样 $\boldsymbol{X} \sim \mu^0_{prob}[\nu]$。

2. 均匀采样 $\boldsymbol{j} \in [n]$ 。

3. 如果 $\boldsymbol{j} = j$ ，那么我们令 $\boldsymbol{Y}^j = 1^k$ 和 $\boldsymbol{Y}^{j'} = \boldsymbol{X}^{j'}$ 对所有 $j' \neq j$成立。

我们用$\mu^1_{prob}[\nu]$ 表示 $\boldsymbol{Y}$的边缘分布，并注意到它的支撑集包含于 $\mathcal{F}^1$。

我们用 $\mu_{\text{prob}}[\nu]$ 表示混合分布：先均匀采样 $\boldsymbol{b} \in \{0,1\}$ ，然后采样 $\boldsymbol{X} \sim \mu^{\boldsymbol{b}}_{\text{prob}}[\nu]$。我们的主要技术结果是针对在输入分布 $\mu_{\text{prob}}[\nu]$下求解唯一集合不相交问题的协议给出的通信复杂度下界。稍后我们会将固定集合大小的情形归约到这一模型。

定理 2.3. 固定 n， $k \geq 1$。设 ν 是 $[k]$上的一个分布。设 Π 是一个 $[c_1, \ldots, c_k]$有界的 k方确定性协议，它在输入分布$\mu_{prob}[\nu]$ 下以2%的错误率求解唯一集合不相交问题。那么

$$\sum_{i \in [k]} \frac{c_i}{\nu(i)} = \Omega(n).$$

我们注意到，定理 2.3 是对称情形下下界的推广[Gro09、Jay09], 其中 $\nu(i) = 1/k$ 对所有 $i \in [k]$成立。在这种情形下，定理 2.3 给出$\sum_i c_i = \Omega(n/k)$。

2.1.1 信息论框架

我们将使用信息论来证明定理 2.3。尽管我们假设 Π 相对于 $\mu^0_{\text{prob}}[\nu]$ 和 $\mu^1_{\text{prob}}[\nu]$的错误率都很低，但我们只研究其相对于 $\mu^0_{\text{prob}}[\nu]$的信息复杂度。下文中，令 $\boldsymbol{W} \in [k]^n$， $\boldsymbol{X} \in (\{0,1\})^n$ 为定义 2.1中所述的联合样本。下面的观察将发挥重要作用。

观察 2.4。 给定 $\boldsymbol{W} = W$，随机变量 $(\boldsymbol{X}_i(j) : i \in [k], \ j \in [n])$ 相互独立。

我们首先基于信息论给出单方通信的一般界限，该界限仅假设存在这样的 $\boldsymbol{W}$，使得 $\boldsymbol{X}_1$、 ..... 、 $\boldsymbol{X}_k$ 相互独立。

引理 2.5. 设 Π 是一个 k 方参与的 $[c_1, \ldots, c_k]$-有界协议。假设联合随机变量 $(\boldsymbol{W}, \boldsymbol{X})$，其中 $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$ 是参与者的输入，并且对于 $\boldsymbol{W}$ 的每个取值 W，随机变量 $\boldsymbol{X}_1|\boldsymbol{W} = W, \ldots, \boldsymbol{X}_k|\boldsymbol{W} = W$ 相互独立。那么对于每个 $i \in [k]$ ，我们有

$$c_i \geq I(\boldsymbol{X}_i : \Pi(\boldsymbol{X})|\boldsymbol{W}).$$

证明。 我们先引入一些记号。我们用 π 表示 Π 的一个可能通信记录，并令 $\pi_{<t} = (\pi_1, \ldots, \pi_{t-1})$ 为一个部分通信记录。我们用 $\boldsymbol{\pi} = \Pi(\boldsymbol{X})$ 表示协议在 $\boldsymbol{X}$ 上运行时的通信记录。

固定协议中的一个时间步 t，以及一段部分通信记录 $\pi_{<t}$。下一位发言的参与方由目前的通信记录决定，因此记作 $\text{next}(\pi_{<t}) \in [k]$。我们还用 $\text{locs}(\pi, i) = \{t \text{ 表示：参与方 } \text{next}(\pi_{<t}) = i\}$ 在通信记录 π 中发送比特的位置。根据我们的假设，对于任意通信记录 π，都有 $|\text{locs}(\pi, i)| \le c_i$。

考虑 $\boldsymbol{W}$ 的任意取值 W。注意，在给定 $\boldsymbol{\pi}_{<t} = \pi_{<t}$ 的条件下，发送的下一个比特 $\boldsymbol{\pi}_t$ 是参与方 $i =$ 的一个函数，其中该参与方为 $\text{next}(\pi_{<t})$。如果 $i' \neq i$，那么由于 $\boldsymbol{X}_i|\boldsymbol{W} = W$ 和 $\boldsymbol{X}_{i'}|\boldsymbol{W} = W$ 相互独立，我们有

$$I(\boldsymbol{X}_{i'} : \boldsymbol{\pi}_t | \boldsymbol{W} = W, \boldsymbol{\pi}_{<t} = \pi_{<t}) = 0.$$

由于 $\boldsymbol{\pi}_t \in \{0, 1\}$，我们也可以直接得到如下界

$$I(\boldsymbol{X}_i : \boldsymbol{\pi}_t | \boldsymbol{W} = W, \boldsymbol{\pi}_{<t} = \pi_{<t}) \le 1.$$

对 $\pi_{<t}$ 和 W 取平均，得到

$$I(\boldsymbol{X}_i : \boldsymbol{\pi}_t | \boldsymbol{W}, \boldsymbol{\pi}_{<t}) \le \Pr[\text{next}(\boldsymbol{\pi}_{<t}) = i].$$

对 t 求和，即得结果：

$$I(\boldsymbol{X}_i : \boldsymbol{\pi} | \boldsymbol{W}) = \sum_t I(\boldsymbol{X}_i : \boldsymbol{\pi}_t | \boldsymbol{W}, \boldsymbol{\pi}_{<t}) = \mathbb{E}|\text{locs}(\boldsymbol{\pi}, i)| \le c_i.$$

□

下文将 $\boldsymbol{\pi} = \Pi(\boldsymbol{X})$ 简记为此。利用引理 2.5、观察 2.4以及数据处理不等式[4] 可得

$$c_i \ge I(\boldsymbol{X}_i : \boldsymbol{\pi} | \boldsymbol{W}) \ge \sum_{j \in [n]} I(\boldsymbol{X}_i(j) : \boldsymbol{\pi} | \boldsymbol{W}).$$

为了证明定理 2.3，考虑以下表达式

$$\sum_{i \in [k]} \frac{c_i}{\nu(i)} \ge \sum_{i \in [k]} \frac{1}{\nu(i)} I(\boldsymbol{X}_i : \boldsymbol{\pi} | \boldsymbol{W}) \ge \sum_{i \in [k]} \frac{1}{\nu(i)} \sum_{j \in [n]} I(\boldsymbol{X}_i(j) : \boldsymbol{\pi} | \boldsymbol{W})$$

我们定义如下

$$L := \frac{1}{n} \sum_{i \in [k]} \frac{1}{\nu(i)} \sum_{j \in [n]} I(\boldsymbol{X}_i(j) : \boldsymbol{\pi} | \boldsymbol{W})$$

因此，以下引理证明了定理 2.3。

引理 2.6。 $L = \Omega(1)$。

我们将在下一小节通过归约到 2.6 比特与函数的协议来证明引理 k-bit AND function。

[4] 若 x, y, z a 关于随机变量，若 x、y 相互独立，则 $I(xy : z) \ge I(x : z) + I(y : z)$。

2.1.2 归约到 AND 函数的信息复杂度

本节中，我们考察 k 比特的 AND 函数及其信息复杂度。设 Λ是一个用于计算该函数的 k 方协议： k 位参与者各自接收一个比特作为输入，他们的目标是计算这些比特的 AND。也就是说，判断它们是否全都等于 1。

设 $\boldsymbol{b} \in \{0,1\}$ 是一个随机比特。对于 $i \in [k]$，令 $e_i[\boldsymbol{b}] \in \{0,1\}^k$ 表示一个向量，其在坐标 $\boldsymbol{b}$ 处为 i，并且 在其他位置均为 0 。以下引理将证明引理 2.6 归约为分析 k 位 AND 的协议的信息复杂度；这类协议仅在两个输入上（即全零输入和全一输入）具有较低错误率。

引理2.7。 存在一个公开随机性的 k 方协议 Λ，用于计算 k 位 *AND* 函数，使用公开随机性 $\boldsymbol{R}$，并满足以下保证：

1. Λ 对于输入 0^k 和 1^k的错误率至多为8% 。
2. $L = \sum_{i\in[k]} I(\boldsymbol{b}, \Lambda(e_i[\boldsymbol{b}], \boldsymbol{R})|\boldsymbol{R})$.

我们将在本小节余下部分证明引理 2.7。首先，将 $\boldsymbol{d} \in [k]$、$\boldsymbol{U} \in \{0,1\}^k$按如下方式联合采样：

1. 根据 ν 对 $\boldsymbol{d} \in [k]$ 进行采样。
2. 给定 $\boldsymbol{d} = d$，对 $\boldsymbol{U}_d \in \{0,1\}$ 进行均匀采样，并对所有 $i \neq d$设定 $\boldsymbol{U}_i = 0$ 。

令 $\sigma = \sigma(\nu)$ 表示 $\boldsymbol{U}$的边缘分布，并注意到对于任意 $j \in [n]$，它与 $\boldsymbol{X}^j$ 的边缘分布相同。事实上，对于任意 j，$(\boldsymbol{d}, \boldsymbol{U})$ 的联合分布与 $(\boldsymbol{W}_j, \boldsymbol{X}^j)$ 相同。下一个命题利用这一点，从 Π 中提取一个用于计算 k比特 AND 的协议 Λ，使其信息复杂度度量相关，且对于输入0^k和 1^k的错误率很低。

命题 2.8. 存在一个使用公共随机性 $\boldsymbol{R}$的（公共随机性） k方协议 Λ，用于计算 k比特 *AND* 函数，并具有以下性质：

1. 对于输入 0^k 和 1^k， Λ 的错误率至多为8% 。
2. $I(\boldsymbol{U}_i : \Lambda(\boldsymbol{U}, \boldsymbol{R})|\boldsymbol{d}, \boldsymbol{R}) = \frac{1}{n}\sum_{j=1}^{n} I(\boldsymbol{X}_i(j) : \boldsymbol{\pi}|\boldsymbol{W})$ *for all* $i \in [k]$.

证明。 我们首先定义协议 Λ。令 $U \in \{0,1\}^k$ 表示 AND 函数的输入。首先，使用公共随机性对 $\boldsymbol{j} \in [n]$ 进行均匀采样；然后采样 $\boldsymbol{W}_{-\boldsymbol{j}} = (\boldsymbol{W}_{j'} : j' \neq \boldsymbol{j}) \sim \nu^{n-1}$。在给定 $\boldsymbol{j} = j$，$\boldsymbol{W}_{-j} = W_{-j}$的条件下，第 i 个参与方随后按如下方式构造其用于 Π 的输入 $\boldsymbol{X}_i$： 设定 $\boldsymbol{X}_i(j) = U_i$，并使用私人随机性采样 $\boldsymbol{X}_i(j')|\boldsymbol{W}_{j'} = W_{j'}$。然后，参与者在其联合输入 $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$ 上运行协议 Π。注意，所用的公共随机性为 $\boldsymbol{R} = (\boldsymbol{j}, \boldsymbol{W}_{-\boldsymbol{j}})$。

为了证明第一个命题，注意到如果 AND 函数的输入 $\boldsymbol{U}$服从 $\boldsymbol{U} \sim \sigma$，那么 $\boldsymbol{X} \sim \mu^0_{\text{prob}}[\nu]$；而如果 $U = 1^k$，那么 $\boldsymbol{X} \sim \mu^1_{\text{prob}}[\nu]$。因此， Λ 相对于输入分布 σ 和 1^k的均匀混合，错误率至多为 2%。因此，相对于

输入 1^k，错误率至多为 4%。由于 $\sigma(0^k) = 1/2$，相对于输入 0^k 的错误率至多为 8%。

对于第二个命题，注意，在给定 $\boldsymbol{R} = R = (j, W_{-j})$ 的条件下，$(\boldsymbol{d}, \boldsymbol{U}, \Lambda(\boldsymbol{U}, R))$ 与 $(\boldsymbol{W}_j, \boldsymbol{X}^j, \pi)$ 的联合分布相同。因此

$$I(\boldsymbol{U}_i : \Lambda(\boldsymbol{U}, R)|\boldsymbol{d}, \boldsymbol{R} = R) = I(\boldsymbol{X}_i(j) : \boldsymbol{\pi}|\boldsymbol{W}_j, \boldsymbol{j} = j, \boldsymbol{W}_{-\boldsymbol{j}} = W_{-j})$$

对 R 取平均可得

$$\begin{aligned} I(\boldsymbol{U}_i : \Lambda(\boldsymbol{U}, R)|\boldsymbol{d}, \boldsymbol{R}) &= \frac{1}{n}\sum_{j\in[n]} I(\boldsymbol{X}_i(j) : \boldsymbol{\pi}|\boldsymbol{W}_j, \boldsymbol{j} = j, \boldsymbol{W}_{-\boldsymbol{j}} = W_{-j}) \\ &= \frac{1}{n}\sum_{j\in[n]} I(\boldsymbol{X}_i(j) : \boldsymbol{\pi}|\boldsymbol{W}). \end{aligned}$$

□

引理的证明2.7。 设 Λ 为命题 2.8给出的协议。那么

$$L = \sum_{i\in[k]} \frac{1}{\nu(i)} I(\boldsymbol{U}_i : \Lambda(\boldsymbol{U}, \boldsymbol{R})|\boldsymbol{d}, \boldsymbol{R}).$$

化简内层各项可得

$$\begin{aligned} \frac{1}{\nu(i)} I(\boldsymbol{U}_i : \Lambda(\boldsymbol{U}, \boldsymbol{R})|\boldsymbol{d}, \boldsymbol{R}) &= \frac{1}{\nu(i)} \sum_{j\in[k]} \nu(j) \cdot I(\boldsymbol{U}_i : \Lambda(\boldsymbol{U}, \boldsymbol{R})|\boldsymbol{d} = j, \boldsymbol{R}) \\ &= I(\boldsymbol{U}_i : \Lambda(\boldsymbol{U}, \boldsymbol{R})|\boldsymbol{d} = i, \boldsymbol{R}) \end{aligned}$$

注意，在给定 $\boldsymbol{d} = i$ 的条件下，$(\boldsymbol{U}_i, \boldsymbol{U})$ 的联合分布与 $(\boldsymbol{b}, e_i[\boldsymbol{b}])$ 相同。因此

$$L = \sum_{i\in[k]} I(\boldsymbol{b} : \Lambda(e_i[\boldsymbol{b}], \boldsymbol{R})|\boldsymbol{R}).$$

□

2.1.3 界定与函数的信息复杂度

我们在本小节证明以下引理，由此结合引理 2.3 、引理 2.5、引理 2.6 和引理2.7，即可证明定理。

引理 2.9。 设 Λ 是一个针对 k 比特 *AND* 函数的（公共随机性）协议，使用公共随机性 $\boldsymbol{R}$，其对输入 0^k 和 1^k的错误率至多为8% 。则

$$\sum_{i\in[k]} I(\boldsymbol{b}, \Lambda(e_i[\boldsymbol{b}], \boldsymbol{R})|\boldsymbol{R}) = \Omega(1).$$

引理2.9 与信息复杂度文献中的先前下界非常相似 [BYJKS04、CKS03、Gro09]。我们需要如下设置。联合采样$\boldsymbol{e} \in [k]$、 $\boldsymbol{V} \in \{0,1\}^k$如下：

1. 均匀采样 $\boldsymbol{e} \in [k]$ 。

2. 给定 $\boldsymbol{e} = e$，均匀采样 $\boldsymbol{V}_e \in \{0,1\}$ ，并对所有 $i \neq e$ 设置 $\boldsymbol{V}_i = 0$ 。

给定一个协议 Λ 使用公共随机性 $\boldsymbol{R}$，其条件信息复杂度为

$$\mathrm{CIC}(\Lambda) = I(\boldsymbol{V} : \Lambda(\boldsymbol{V}, \boldsymbol{R}) | \boldsymbol{e}, \boldsymbol{R}).$$

这个量自然出现在使用信息复杂度研究唯一不相交性的问题中；这一研究始于 [BYJKS04]。Gronemeier [Gro09] 和 Jayram [Jay09] 证明了这个量的紧下界。

定理 2.10。 *[[Gro09, Jay09]]*$CIC(\Lambda) = \Omega(1/k)$。

事实上，证明（尽管未明确如此表述）仅依赖于以下假设： Λ 在全零输入和全一输入上的错误率为 $\leq 30\%$ （完整证明见 Gronemeier 的博士论文 [Gro10]）。因此，该证明适用于我们的协议 Λ。下面的命题将 $\mathrm{CIC}(\Lambda)$与我们要界定的量联系起来，并由此完成引理 2.9的证明，从而也完成定理 2.3的证明。

命题 2.11. $\sum_{i\in[k]} I(\boldsymbol{b} : \Lambda(e_i(\boldsymbol{b}) \ \boldsymbol{R}) | \boldsymbol{R}) = k \cdot CIC(\Lambda)$。

证明。

$$\begin{aligned} k \cdot \mathrm{CIC}(\Lambda) &= k \cdot I(\boldsymbol{V} : \Lambda(\boldsymbol{V}, \boldsymbol{R}) | \boldsymbol{e}, \boldsymbol{R}) \\ &= \sum_{i\in[k]} I(\boldsymbol{V} : \Lambda(\boldsymbol{V}, \boldsymbol{R}) | \boldsymbol{e} = i, \boldsymbol{R}) \\ &= \sum_{i\in[k]} I(\boldsymbol{b} : \Lambda(e_i(\boldsymbol{b}), \boldsymbol{R}) | \boldsymbol{R}). \end{aligned}$$

□

2.2 推广到次概率分布

将定理 2.3 推广到次概率分布会很方便。定义在 ν上的次概率分布 $[k]$ 满足 $\nu(i) \geq 0$ 以及 $\sum \nu(i) \leq 1$。我们将 $\mu^0_{\mathrm{prob}}[\nu]$, $\mu^1_{\mathrm{prob}}[\nu]$ 的定义推广到次概率分布如下。

我们首先描述如何对 $\boldsymbol{X} \sim \mu^0_{\mathrm{prob}}[\nu]$ 进行采样。对于每个 $j \in [n]$，以概率 $\nu(i)$ 对 $\boldsymbol{X}_i(j)$ 进行均匀采样，并对所有 $i \neq i'$ 设置 $\boldsymbol{X}_{i'}(j) = 0$；以概率$1 - \sum \nu(i)$ 对所有 i 设置 $\boldsymbol{X}_i(j) = 0$。对 $\boldsymbol{Y} \sim \mu^1_{\mathrm{prob}}[\nu]$ 进行采样时，我们采用与分布式情形相同的过程：首先采样 $\boldsymbol{X} \sim \mu^0_{\mathrm{prob}}[\nu]$, 然后均匀采样一个 $\boldsymbol{j} \in [n]$，并对所有 $j' \neq j$设置 $\boldsymbol{Y}^{\boldsymbol{j}} = 1^k$ 和 $\boldsymbol{Y}^{j'} = \boldsymbol{X}^{j'}$。我们将 $\mu^0_{\mathrm{prob}}[\nu]$和 $\mu^1_{\mathrm{prob}}[\nu]$ 的等权混合分布记为 $\mu_{\mathrm{prob}}[\nu]$。下面的定理将定理 2.3 推广到次分布。

定理 2.12。 固定 $n, k \geq 1$。令 ν 为 $[k]$ 上的一个次概率分布。令 Π 为一个 $[c_1, \ldots, c_k]$-有界协议，它在输入分布 $\mu_{prob}[\nu]$ 下以错误率2%解决分布式唯一集合不相交问题。那么

$$\sum_{i \in [k]} \frac{c_i}{\nu(i)} = \Omega(n).$$

证明。将 ν 扩展为 $[k+1]$ 上的分布 ν'，令 $\nu'(i) = \nu(i)$（对于 $i \in [k]$），并令 $\nu'(k+1) = 1 - \sum \nu(i)$。将 Π 扩展为一个供 $k+1$ 个参与者使用的协议 Π'，其中参与方 $k+1$ 完全不参与该协议。因此， Π' 是一个 $[c_1, \ldots, c_k, 0]$有界协议。将定理 2.3 应用于 Π' 和 ν'即可得证。 □

2.3 固定集合大小的扩展

现在我们利用已证明的结果推导定理 1.5，即固定集合大小情形下的下界。首先引入一些记号。

Le令 $\mathfrak{s} = [s_1, \ldots, s_k]$ 表示集合大小，其中 $s_i \geq 1$ 和 $\sum s_i \leq n$。定义

$$\mathcal{F}_{\text{size}}[\mathfrak{s}] = \{X \in (\{0,1\}^n)^k : \forall i \in [k], |X_i| = s_i\}.$$

对于 $b \in \{0,1\}$，令 $\mathcal{F}^b_{\text{size}}[\mathfrak{s}] = \mathcal{F}^b \cap \mathcal{F}_{\text{size}}[\mathfrak{s}]$和 $\mu^b_{\text{size}}[\mathfrak{s}]$为 $\mathcal{F}^b_{\text{size}}[\mathfrak{s}]$上的均匀分布。我们的困难分布 $\mu_{\text{size}}[\mathfrak{s}]$ 是 $\mu^0_{\text{size}}[\mathfrak{s}]$ 和$\mu^1_{\text{size}}[\mathfrak{s}]$的等权混合分布。等价地，均匀采样 $\boldsymbol{b} \in \{0,1\}$，并取 $\boldsymbol{X} \sim \mu^{\boldsymbol{b}}_{\text{size}}[\mathfrak{s}]$。我们证明：对于在输入分布 $\mu_{\text{size}}[\mathfrak{s}]$下解决唯一集合不相交问题的协议，存在通信复杂度下界。

定理 2.13。 设 $\mathfrak{s} = [s_1, \ldots, s_k]$ 满足 $\sum s_i \leq n/2$。设 Π 是一个 $[c_1, \ldots, c_k]$-有界的 k-方协议，它在输入分布 $\mu_{size}[\mathfrak{s}]$ 下以1%的错误率解决唯一集合不相交问题。那么

$$\sum_{i \in [k]} \frac{c_i}{s_i} = \Omega(1).$$

显然，定理 2.13蕴含定理 1.5，但事实上二者等价。在证明之前，我们需要以下断言。

断言 2.14。 设 $b \in \{0,1\}, X \in \mathcal{F}^b_{size}[\mathfrak{s}]$。设 $\boldsymbol{\Sigma}$ 是 $[n]$ 的一个随机排列，并令 $\boldsymbol{\Sigma}(X)$表示将 $\boldsymbol{\Sigma}$ 应用于 X所得的结果。那么 $\boldsymbol{\Sigma}(X)$ 在 $\mathcal{F}^b_{size}[\mathfrak{s}]$上均匀分布。

证明。 该断言成立，因为 $[n]$ 上的置换在 $\mathcal{F}^b_{\text{size}}[\mathfrak{s}]$上具有传递性。具体而言，对于任意$X$、$X' \in \mathcal{F}^b_{\text{size}}[\mathfrak{s}]$，存在一个作用于 $[n]$ 上的置换 Σ，使得 $\Sigma(X) = X'$。这意味着，均匀置换会将 X 映射为定义域 $\mathcal{F}^b_{\text{size}}[\mathfrak{s}]$中的一个均匀元素。 □

断言 2.15。 定理 *1.5* 和定理*2.13* 等价。

证明。 我们比较规模为 $\mathfrak{s} =[s_1, \ldots, s_k]$ 的多方唯一集合不相交问题在两种设置下的情形：最坏情况输入和均匀输入。显然，适用于最坏情况输入的协议，也能以相同的通信量

和错误率保证处理均匀输入。反过来，设 $X \in \mathcal{F}^b_{\text{size}}[\mathfrak{s}]$ 是唯一集合不相交问题的任意输入。参与者利用公共随机性，在 $\mathbf{\Sigma}$上抽取一个均匀置换 $[n]$，并各自将其应用于自己的输入。根据命题2.14，我们知道 $\mathbf{\Sigma}(X)$ 服从 $\mu^b_{\text{size}}[\mathfrak{s}]$的分布。现在，他们可以应用一个能在输入$\mu_{\text{size}}[\mathfrak{s}]$下解决唯一集合不相交问题的协议。 □

现在我们转而证明定理 2.13。

定理*2.13*的证明。 首先，注意我们可以假设 $c_i \geq 1$ 对所有 i都成立，因为我们可以将 $c_i = 0$ 的参与者从游戏中移除，因为他们不允许发送任何比特。

设 Π 是定理 2.13中假设的协议。也就是说，该协议是 $[c_1, \ldots, c_k]$-有界的，并且在输入分布 $\mu_{\text{size}}[\mathfrak{s}]$ 下以 1%的错误率解决唯一集合不相交问题，其中$\mathfrak{s} = [s_1, \ldots, s_k]$ 满足 $\sum s_i \leq n/2$。我们将用它设计一个 $[c_1 + 1, \ldots, c_k + 1]$-有界协议 Π'，该协议在一种特定的次概率分布情形下以 2%的错误率解决唯一集合不相交问题，然后应用定理 2.12。

接下来，定义一个次概率分布 ν，其定义在 $[k]$ 上，满足 $\nu(i) = \frac{s_i}{4n}$。我们考虑其对应的分布式输入 $\mu_{\text{prob}}[\nu]$，定义在大小为 $n/2$ 比特的输入上。令 $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k) \sim \mu_{\text{prob}}[\nu]$，其中 $\boldsymbol{X} \in (\{0,1\}^{n/2})^k$。每个 $\boldsymbol{X}_i$ 都服从二项分布 $\text{Bin}(n/2, \nu(i))$，其期望大小为 $\mathbb{E}[|\boldsymbol{X}_i|] = \frac{s_i}{2}$。因此，由霍夫丁界可得，

$$\Pr[|\boldsymbol{X}_i| > s_i] \leq \exp(-s_i/6).$$

令 E 表示存在某个 $i \in [k]$使得 $|\boldsymbol{X}_i| > s_i$ 这一事件。于是

$$\Pr[E] \leq \sum_{i \in [k]} \exp(-s_i/6).$$

我们首先分析 $\Pr[E] \geq 1\%$ 的情形。在这种情况下，由假设可知 $c_i \geq 1$，并且对于所有 $x \geq 1$，都有 $\frac{1}{x} \geq C\exp(-x/6)$，其中 $C > 0$ 为某个绝对常数，因此我们得到

$$\sum_{i \in [k]} \frac{c_i}{s_i} \geq C \sum_{i \in [k]} \exp(-s_i/6) \geq C \Pr[E] = \Omega(1).$$

从现在起，我们假设 $\Pr[E] < 1\%$ 。

现在我们设计协议 Π'。首先，每个参与方检查其输入 X_i 是否满足$|X_i| > s_i$。如果满足，协议就中止。这需要每个参与方发送一个比特；根据假设，协议中止的概率至多为 1%。否则，每个参与方按如下方式将其输入 X_i 扩展为一个大小为 $|Y_i| = s_i$ 的新输入 $Y_i \in \{0,1\}^n$ 。

在协议开始之前，参与者预先约定好两两不相交的子集T_1、……、T_k，其中包含 $|T_i| = s_i$，且支撑集位于最后 $n/2$ 个坐标中（因此它们不会与输入 X_1、……、X_k重叠）。现在，第 i位参与方从 T_i 中任意选取 $s_i - |X_i|$ 个元素，并将其添加到自己的集合 X_i中；我们将新的输入记为 $Y_i \in \{0,1\}^n$。注意，$Y = (Y_1, \ldots, Y_k)$ 满足与 $X = (X_1, \ldots, X_k)$ 相同的承诺；也就是说，它们要么两两不相交，要么拥有一个公共元素，并且除该元素外两两不相交。

我们希望将 Π 应用于 Y。然而目前还不能直接这样做；尽管确实有$Y \in \mathcal{F}^0_{\text{大小}}[\mathfrak{s}]$或 $Y \in \mathcal{F}^1_{\text{size}}[\mathfrak{s}]$，但其在各集合中的分布并不均匀。不过，我们可以在这里应用命题 2.14，使 Y在相应的集合族中均匀分布。参与者利用公共随机性采样一个置换 $\mathbf{\Sigma}$，作用于 $[n]$，并将其应用于 Y。现在我们可以应用 $\Pi(\mathbf{\Sigma}(Y))$，根据假设，其将以 2% 的错误率给出正确结果。证明现在由定理 2.12得出。 □

3 区间系统

回顾一下，我们的计划是利用多方唯一集合不相交问题的下界，来证明针问题的流式算法下界。为了有效地将唯一集合不相交问题的输入嵌入数据流中，我们引入一种称为区间系统的组合构造。

定义 3.1 （区间）。 区间是一个非空集合，其形式为 $I = \{a, a+1, \ldots, b\}$，其中 $a \leq b$。

定义 3.2 （区间系统）。 一个 $[t]$-区间系统是由 $[t]$支撑的 k 个两两不相交区间组成的集合 $F = \{I_1, \ldots, I_k\}$。如果要指定区间的数量，我们称 F 为一个$[t,k]$- 区间系统。

定义 3.3 （随机区间系统）。 随机 $[t]$-区间系统 $\mathcal{F}$是定义在 $[t]$-区间系统 F上的一个分布。同样，随机 $[t,k]$-区间系统 $\mathcal{F}$是定义在 $[t,k]$-区间系统 F上的一个分布。

接下来，我们为区间系统定义一个相应的集合分布 $T \subset [t]$。

定义 3.4 (区间系统的集合分布**)**。 设 F 是一个 $[t]$-区间系统。我们将通过从每个区间 $I \in F$ 中均匀选取一个元素所得到的集合分布 $T \subset [t]$ 记为 $\mathit{Sets}(F)$。

如果 $\mathcal{F}$ 是一个随机区间系统 $[t]$，那么我们按如下方式定义 $\mathit{Sets}(\mathcal{F})$：首先抽样得到 $F \sim \mathcal{F}$，然后从 $\mathit{Sets}(F)$ 中抽样得到 $T \sim$。

注意，如果 $\mathcal{F}$ 是一个随机区间系统 $[t,k]$，那么 Sets($\mathcal{F}$) 是 $[t]$ 中 k-元集合上的一个分布（k-元集合是指大小为 k 的集合）。我们的目标是模拟 $[t]$ 中 k-元集合上的均匀分布。我们称这样的随机区间系统为完美的。

定义 3.5 (完美区间系统**)**。 如果随机区间系统 $\mathcal{F}$ 生成的 $\mathit{Sets}(\mathcal{F})$ 是 $[t]$ 中所有 k-元集合上的均匀分布，则称随机 $[t,k]$-区间系统 $\mathcal{F}$ 为完美的。

构造完美随机化的 $[t,k]$-区间系统有许多方法。例如，一种朴素的方法是采样 k 均匀坐标 $i_1, \ldots, i_k \in [t]$，然后取 $F = \{\{i_1\}, \ldots, \{i_k\}\}$ 上的分布。不过，为了实现高效归约，我们需要区间尽可能长的区间系统。从技术上说，归约的效率将由区间系统的值 这一概念决定。

定义 3.6（区间系统的值）。 设 F 是一个 $[t]$-区间系统。其值为

$$val(F) = \sum_{I \in F} \frac{1}{|I|}.$$

如果 $\mathcal{F}$ 是一个随机化的 $[t]$-区间系统，那么其值为

$$val(\mathcal{F}) = \mathbb{E}_{F \sim \mathcal{F}} [val(F)].$$

为了证明流式算法的强下界，我们需要一个在 $[t,k]$-区间上尽可能低值的完美分布。以下命题给出了相应的下界。

命题 3.7. 设 F 是一个 $[t,k]$-区间系统。那么

$$val(F) \geq \frac{k^2}{t}.$$

证明。 设 $F = \{I_1, \ldots, I_k\}$，其中 $|I_i| = s_i$。我们有 $\sum s_i \leq t$，且 $\mathrm{val}(F) = \sum \frac{1}{s_i}$。当所有 s_i 都相等时，该表达式取得最小值，因此

$$\mathrm{val}(F) \geq k \cdot \frac{k}{\sum s_i} \geq \frac{k^2}{t}.$$

□

本节的主要技术结果是构造一个值接近最优的完美随机化$[t,k]$-区间系统。为此，我们设计了一种随机算法，用于采样 $[t,k]$-区间系统。我们将证明，其输出分布是完美的，且其值接近命题3.7给出的最小值。

下面给出对区间或区间系统进行“平移”的定义，这将十分方便。对于区间 $I = [a,b]$ 和整数 c，定义 $I + c = [a + c, b + c]$。对于区间系统 $F = \{I_1, \ldots, I_k\}$，定义 $F + c = \{I_1 + c, \ldots, I_k + c\}$。

算法 1： SampleIntervalSystem

输入： $t \geq 1$，$k \geq 0$与 $k \leq t$

输出： $[t,k]$-区间系统 F

如果 $k = 0$ **则**
  返回 $F = \{\}$
否则如果 $k = 1$ **则**
  返回 $F = \{[t]\}$
否则
  令 $s = \lceil t/2 \rceil$
  以概率 $\Pr[\boldsymbol{j} = j] = \frac{\binom{s}{j}\binom{t-s}{k-j}}{\binom{t}{k}}$ 采样 $\boldsymbol{j} \in \{0, \ldots, k\}$
  计算 $F_1 = \mathrm{SampleIntervalSystem}(s, \boldsymbol{j})$
  计算 $F_2 = \mathrm{SampleIntervalSystem}(t - s, k - \boldsymbol{j})$
  返回 $F = F_1 \cup (F_2 + s)$
结束

我们将通过运行 SampleIntervalSystem(t,k) 得到的随机 $[t,k]$-区间系统记为 $\mathcal{F}[t,k]$。

命题 3.8。 $\mathcal{F}[t,k]$ 是完美的。

证明。 证明采用关于 k、t 的归纳法。如果 $k=0$ 或 $k=1$，则由算法的基本情形可知结论显然成立。如果 $k \geq 2$，那么我们以与在 $[t]$ 中均匀分布的 k-集合相同的概率，对区间 $[s]$ 中的元素数量 j 进行采样。根据归纳法， F_1 的分布 $\mathcal{F}[s,j]$ 是一个完美的随机化 $[s,j]$ 区间系统；而 F_2 的分布 $\mathcal{F}[t-s,k-j]$ 是一个完美的随机化 $[t-s,k-j]$ 区间系统。命题得证。 □

接下来我们分析 $\mathcal{F}[t,k]$ 的值；为简化分析，我们将范围限定在 t 是 2 的幂的情形。这足以满足我们的应用需求，而且我们预计只需稍作修改，这个界限便可推广到一般的 t。下文假设所有对数均以 2 为底。

引理 3.9。 假设 t 是 2 的幂。那么 $val(\mathcal{F}[t,k]) \leq {}^{k^2\log(2t)}t$。

为了证明引理 3.9，我们需要以下技术性命题，它计算了算法中关于 j 的分布的一阶矩和二阶矩。

命题 3.10。 设 $t,k \geq 1$, t 均为偶数，且 $0 \leq j \leq k$。定义 $p(t,k,j) = \binom{t/2}{j}\binom{t/2}{k-j}\binom{t}{k}$。
那么

$$\sum_{j=0}^{k} p(t,k,j)\cdot j = \frac{k}{2}$$

and

$$\sum_{j=0}^{k} p(t,k,j)\cdot j^2 \leq \frac{k(k+1)}{4}.$$

证明。 令 $s=t/2$。令 T 为 $[t]$ 的大小为 k 的均匀子集。则 $p(t,k,j) = \mathbf{Pr}[|T\cap[s]| = j]$。因此

$$\sum_{j=0}^{k} p(t,k,j)\cdot j = \mathbb{E}_T\left[\sum_{i\in[s]} \mathbf{1}[i\in T]\right] = \sum_{i\in[s]} \Pr[i\in T] = s\cdot\frac{k}{2s} = \frac{k}{2}$$

and

$$\begin{aligned}\sum_{j=0}^{k} p(t,k,j)\cdot j^2 &= \mathbb{E}_T\left[\sum_{i,j\in[s]} \mathbf{1}[i\in T]\cdot\mathbf{1}[j\in T]\right] = \sum_{i,j\in[s]} \Pr[i,j\in T] \\ &= s\cdot\frac{k}{2s} + s(s-1)\frac{k(k-1)}{2s(2s-1)} \leq \frac{k}{2} + \frac{k(k-1)}{4} = \frac{k(k+1)}{4}.\end{aligned}$$

□

引理 3.9 的证明。令 $f(t,k) = t \cdot \mathrm{val}(\mathcal{F}[t,k])$。我们有 $f(t,0) = 0, f(t,1) = 1$ 与 $f(t,k) = 0$，当且仅当 $k > t$。对于 $k \geq 2$， $f(t,k)$ 的定义是递归的。令 $p(t,k,j) = \binom{t/2}{j}\binom{t/2}{k-j}\Big/\binom{t}{k}$。然后

$$\mathrm{val}(\mathcal{F}[t,k]) = \sum_{j=0}^{k} p(t,k,j)\left(\mathrm{val}(\mathcal{F}[t/2,j]) + \mathrm{val}(\mathcal{F}[t/2,k-j])\right).$$

这意味着

$$f(t,k) = 4\sum_{j=0}^{k} p(t,k,j) f(t/2,j).$$

计算 $f(t,2)$ 将很有启发性：

$$f(t,2) = \frac{t}{t-1} + \frac{t-2}{t-1} f(t/2,2) \leq 2 + f(t/2,2) \leq 2\log(t).$$

我们将用归纳法证明

$$f(t,k) \leq k^2 + k(k-1)\log(t).$$

我们已经验证了 $k = 0$、1、2 的情形。对于 $k \geq 3$，根据归纳法：

$$f(t,k) \leq 4\sum_{j=0}^{k} p(t,k,j)\left(j^2 + j(j-1)\log(t/2)\right).$$

应用命题 3.10 可得

$$\begin{aligned} f(t,k) &\leq k(k+1) + k(k-1)\log(t/2) \\ &= 2k + k(k-1)\log(t) \\ &\leq k^2 + k(k-1)\log(t). \end{aligned}$$

最后我们得到

$$\mathrm{val}(\mathcal{F}[t,k]) = \frac{f(t,k)}{t} \leq \frac{k^2 + k(k-1)\log(t)}{t} \leq \frac{k^2\log(2t)}{t}.$$

□

我们将针问题的流式算法应用于一种额外限制：区间系统中区间的总长度必须与 t 保持一定距离。我们将这类区间系统称为有效区间系统。

定义 3.11（有效区间系统）。如果 $\sum_{I\in F}|I| \leq t/2$，则称一个 $[t]$-区间系统 F 为有效的。如果一个随机 $[t]$-区间系统 $\mathcal{F}$ 的支撑集中所有 $[t]$-区间系统 F 都有效，则称该随机区间系统有效。

接下来，我们将展示如何细化一个区间系统，以获得一个有效随机区间系统，同时保留集合的分布，且不会使其值增加太多。

引理 3.12。 假设 $k \le t/6$。设 F 是一个 $[t,k]$-区间系统。则存在一个随机化的$[t,k]$-区间系统 $\mathcal{F}$，满足：

1. 集合$(\mathcal{F})$ $=$集合(F)

2. $val(\mathcal{F}) \le 5 \cdot val($

$F)$*3.* $\mathcal{F}$ 有效

证明。 设 $F = \{I_1, \dots, I_k\}$。给定一个区间 I_i，定义 $\ell_i = \min(3, |I_i|)$。将 I_i 划分为 ℓ_i个区间 $\{I_{i,a} : a \in [\ell_i]\}$ 尽可能等长，并注意到

$$\frac{|I_i|}{5} \le |I_{i,a}| \le \frac{|I_i|}{3} + 1 \quad \forall a \in [\ell_i].$$

设 $p_{i,\,a} = \frac{|I_i}{|I_i|}{}_a|$。我们定义一个随机化的 $[t,k]$-区间系统 $\mathcal{F}$，其中对每个 $i \in [k]$，独立地将 I_i 替换为其子区间之一。具体而言，我们以概率 p_i选择 $a \in [\ell_i],\ a$ 并将 I_i 替换为 $I_{i,\ a}$。现在我们证明这些命题。

1. 注意到，对均匀分布元素 $x \in I_i$ 进行采样，等价于先以概率 $p_{i,a}$采样 $a \in [\ell_i]$，再从中均匀采样一个元素 $x \in I_{i,a}$。这意味着 Sets$(\mathcal{F})$ $=$Sets(F)。

2. 由于对于所有 i, a都有 $|I_{i,a}| \ge |I_i|/5$，因此该命题对 $\mathcal{F}$的支撑集中的任意 F' 都成立，从而对 $\mathcal{F}$也成立。

3. 由于对于所有 i, a都有 $|I_{i,a}| \le (|I_i| + 1)/2$，因此对于 $\mathcal{F}'$ 的支撑集中的任意 $F' = \{I_{1,a_1}, \dots, I_{k,a_k}\}$，我们有

$$\sum_{i \in [k]} |I_{i,a_i}| \le k + \frac{1}{3} \sum_{i \in [k]} |I_i| \le k + \frac{t}{3} \le \frac{t}{2}$$

其中，最后一个不等式由我们假设 $k \le t/6$得出。

□

引理 3.12 也适用于随机化的 $[t,k]$-区间系统，因为可以将其应用于其支撑集中的任意区间系统。下面的引理总结了我们需要的所有事实，只需将其应用于 $\mathcal{F}[t,k]$。

引理 3.13。 设 $k, t \ge 1$。假设 t 是 *2* 的幂且 $k \le t/6$。那么存在一个有效的完美随机化 $[t,k]$-区间系统 $\mathcal{F}$，满足

$$val(\mathcal{F}) \le \frac{10k^2 \log(t)}{t}.$$

4 针问题的下界

本节将结合固定集合大小下唯一集合不相交问题的下界（定理 1.2）与基于区间系统的高效归约（引理 2.13），来证明定理 3.13。

首先，我们回顾一下参数： n 表示定义域的大小， t 表示样本数， p 表示针概率。我们在全文中假设 $n = \Omega(t^2)$ 足够大。我们用 k 表示植入模型中一条数据流里的针数，其中 $k \sim \mathrm{Bin}(t, p)$。我们用 Uniform 表示 $[n]^t$上的均匀分布。

首先，我们说明在 k 固定时如何证明下界。给定一个 $[t, k]$-区间系统 $F = \{I_1, \ldots, I_k\}$，本节中我们假设这些区间按顺序排列，即 I_1 在 I_2之前， I_2又在 I_3之前，依此类推。我们将其对应的大小定义为

$$\mathrm{Sizes}(F) = (|I_1|, \ldots, |I_k|).$$

我们回顾引言中给出的植入流分布的定义，并在此更正式地介绍它。

定义 4.1 (区间系统的植入分布)。 令 F 为一个 $[t]$-区间系统。我们定义其在数据流 $X \in [n]^t$ 上的植入分布 $\mathit{Planted}[F]$ 如下：

1. 采样一根均匀针 $x \in [n]$;
2. 对每个区间 $I \in F$，采样均匀分布的索引 $a_I \in I$ 并设置$X_{a_I} = x$;
3. 对所有 $j \in [n] \setminus \{a_I : I \in F\}$，均匀采样 $X_j \in [n]$ 。

对于 $\mathcal{F}$ 一个随机化的 $[t]$-区间系统，我们通过先采样 $F \sim \mathcal{F}$，然后 $X \sim$ ，来定义其植入分布 $\mathit{Planted}[\mathcal{F}]$ $[F]$。

我们首先对引理 1.7 进行形式化并证明。给定一个流式算法 $\mathcal{ALG}$和两个分布 D_0, D_1 关于数据流，我们称 $\mathcal{ALG}$ 以错误率 D_0， D_1区分 δ，如果在算法运行结束时，最后一位参与方能够猜出输入是从 D_0 还是 D_1 中采样得到的，并且猜对的概率至少为 $1 - \delta$。如果一个流式算法进行 ℓ遍数据流扫描，则称其为 ℓ 遍流式算法。

引理 4.2。 设 F 为一个 $[t, k]$-区间系统，并令 $\mathfrak{s} = \mathit{Sizes}(F)$。设 $\mathcal{ALG}$ 为一个 ℓ-遍流式算法，它以 0.5% 的错误率区分 $\mathit{Planted}[F]$ 与均匀分布，并使用 s 空间。则存在一个通信协议 Π，可在输入分布 $\mu_{\mathit{size}}[\mathfrak{s}]$ 下解决唯一集合不相交问题；其中每位参与方发送 ℓs 比特，错误率为 1%。

证明。 $X = (X_1, \ldots, X_k) \in (\{0,1\}^n)^k$ 是参与方的输入，其中我们假设 $X \sim \mu^b_{\mathrm{size}}[\mathfrak{s}]$ 对某个 $b \in \{0, 1\}$ 成立。参与方的目标是确定 b。

令 $F = \{I_1, \ldots, I_k\}$。令 $J_1, \ldots, J_k$ 为 $[t]$ 的一个划分，其中 $I_i \subset J_i$。首先，每个参与方根据自己的输入 X_i 单独构造一个数据流 $Y_i \in [n]^{J_i}$ 。第 i 个参与方按如下方式生成其数据流：

1. 对每个 $j \in J_i \setminus I_i$，均匀采样 $Y_i(j) \in [n]$。

2. 令 $S_i = \{j \in [n] : X_i(j) = 1\}$，其中 $|S_i| = s_i$ 是一个假设。令 $L_i \in [n]^{s_i}$ 为 S_i的一个随机排列。设 $(Y_i(j) : j \in I_i) = L_i$。

令 $Y = Y_1 \circ \cdots \circ Y_k \in [n]^t$ 为这些数据流的串联。参与者模拟在该数据流上运行 $\mathcal{ALG}$：每位参与者在自己负责的数据流部分上模拟算法，并将流式算法的内部存储器发送给下一位参与者。每一轮结束时，最后一位参与者将内部存储器发回给第一位参与者。因此，每位参与者至多发送 ℓs 比特。最后，我们需要证明这能够区分 $b = 0$ 和 $b = 1$。

最后，我们根据 b的取值计算 Y的分布，并证明当 $b = 0$ 时，Y 的分布接近均匀分布；当 $b = 1$ 时，它接近植入分布 Planted$[F]$。因此，根据假设，该算法能够区分这两种情况，这正是我们的目标。

首先，若 $b = 0$，则 X_1、……、X_k 是 $[n]$ 中大小分别为 s_1、……、s_k 的均匀分布集合，并以两两不相交为条件。因此，Y 的元素在 n 中所有由 t 个互不相同元素组成的选择中均匀分布。由于我们假设 $n = \Omega(t^2)$，Y 与均匀分布之间的统计距离至多为 t^2/n，可以将其调得任意小，例如 0.1%。

类似地，若 $b = 1$，则 X_1、……、X_k 在具有唯一交集的条件下均匀分布。同样，假设 $n = \Omega(t^2)$ 意味着 Y 与 Planted$[F]$ 之间的统计距离可以调得任意小，例如 0.1%。

总之，由于我们假设 $\mathcal{ALG}$ 能以 0.5% 的错误率区分均匀分布与 Planted$[F]$，因此它也能以略高的 1% 错误率区分 $b = 0$ 和 $b = 1$ 对应的 Y 分布。

□

结合引理4.2 和定理 2.13，我们得到以下推论，它形式化了引理1.7。

引理 4.3。设 F 是一个有效的 $[t, k]$-区间系统。设 $\mathcal{ALG}$ 是一个 ℓ-遍流式算法，它以错误率 0.5% 区分 *Planted*$[F]$ 和均匀分布，且使用空间 s。那么

$$\ell s = \Omega\left(\frac{1}{val(F)}\right).$$

证明。设 Π 是由引理 4.2得到的协议，它在输入分布 $\mu_{\text{size}}[\mathfrak{s}]$ 下解决唯一集合不相交问题，适用于 $\mathfrak{s} =$ 大小$(F) = [s_1, \ldots, s_k]$，且每个参与方最多发送 ℓs 比特。由于 F 有效，因此我们有 $\sum s_i \le t/2$。定理 2.13 于是给出

$$\sum_{i \in [k]} \frac{\ell s}{s_i} = \Omega(1).$$

回顾 val$(F) = \sum_{i \in [k]} \frac{1}{s_i}$ 的定义，我们可以将其改写为 $\ell s \cdot$val$(F) = \Omega(1)$。

□

以下引理对引理 1.8进行了形式化，并将引理4.3 推广到随机化区间系统。

引理 4.4。 设 $\mathcal{F}$ 是一个有效的随机化 $[t]$-区间系统。设 $\mathcal{ALG}$ 是一个 ℓ-遍流式算法，以错误率 0.1% 区分 $\mathit{Planted}[\mathcal{F}]$ 与均匀分布，并使用空间 s。那么

$$\ell s = \Omega\left(\frac{1}{val(\mathcal{F})}\right).$$

证明。 样本 $F \sim \mathcal{F}$。由于 $\mathrm{val}(\mathcal{F}) = \mathbb{E}[\mathrm{val}(F)]$,，根据马尔可夫不等式，我们有

$$\Pr_F[\mathrm{val}(F) > 2\mathrm{val}(\mathcal{F})] \le 50\%.$$

接下来，令 $\mathrm{err}(F)$ 表示 $\mathcal{ALG}$区分 $\mathrm{Planted}[F]$与均匀分布的错误率。由于 $\mathrm{Planted}[\mathcal{F}]$是 $\mathrm{Planted}[F]$的混合分布，因此 $\mathrm{err}(F)$ 的平均值就是 $\mathcal{ALG}$区分 $\mathrm{Planted}[\mathcal{F}]$ 与均匀分布的错误率，我们假设该错误率为 0.1%。因此

$$\Pr_F[\mathrm{err}(F) > 0.5\%] \le 20\%.$$

总之，在 F的支撑集 $\mathcal{F}$中，存在某个选择，使得 $\mathrm{val}(F) \le 2\mathrm{val}(\mathcal{F})$ 且 $\mathrm{err}(F) \le 0.5\%$。将引理 4.3 应用于 F，即可得到该引理。 □

现在我们可以最终证明定理 1.2，给出任何能够解决针问题的流式算法的样本空间下界。

定理 1.2的证明。令 $\mathcal{ALG}$为一个 ℓ遍流式算法，它使用 t 个样本，能够以高概率区分均匀分布与植入针分布。由于输入是随机的，我们可以重复运行几次以降低错误率。因此，只需将 t 乘以一个常数因子，就可以假设错误率至多为0.1%，并且 t 是 2 的幂。

对于 $k \le t$ ，令 $\mathcal{F}_k$ 为引理 $[t,k]$ 给出的有效完美随机 3.13 区间系统。我们通过采样 $[t]$ 区间系统 $\mathcal{F}$，并取 $k \sim \mathrm{Bin}(t,p)$来构造一个随机 $\mathcal{F}_k$。注意，$\mathrm{Planted}[\mathcal{F}]$ 与植入针分布相同。如果$\mathcal{ALG}$ 使用了 s比特空间，那么由引理 4.4 可得

$$\ell s = \Omega\left(\frac{1}{\mathrm{val}(\mathcal{F})}\right).$$

为完成证明，我们只需计算 $\mathrm{val}(\mathcal{F})$。对于任意固定的 k ，由引理3.13 可知

$$\mathrm{val}(\mathcal{F}_k) \le \frac{10k^2\log(t)}{t}.$$

由于 $k \sim \mathrm{Bin}(t,p)$，我们有 $\mathbb{E}[k^2] = p(1-p)t + p^2t^2$。由于假设 $p = \Omega(1/t)$，主导项是二次项，因此 $\mathbb{E}[k^2] = \Theta(p^2t^2)$。于是我们得到

$$\mathrm{val}(\mathcal{F}) = O(p^2t\log(t)).$$

重新整理各项即可完成证明，因为这给出了 $\ell p^2 st\log(t) = \Omega(1)$。 □

5 开放问题

我们在定理 1.2 中证明了针问题所需样本与空间复杂度的近乎紧确界，并由此证明了随机流问题中频率估计的类似近乎紧确界。要证明精确界并消除剩余的对数因子，仍然是一个悬而未决的问题。我们提出以下自然猜想。

猜想 5.1. 任何 ℓ 次数据流算法，若能以高概率区分均匀分布模型与植入模型，其中 p 为针概率， t 为样本数， s 为空间， n 为域大小，则满足 $\ell p^2 st = \Omega(1)$。

另一个自然的猜想是去除定理 $\sum s_i \leq n/2$ 中的人为限制 1.5。我们之所以需要这一限制，是因为我们并未直接证明该定理，而是通过归约到非对称乘积分布情形来证明。我们推测，或许存在一种直接证明方法，可以克服这一技术障碍（虽然我们并没有任何需要一般性界的实际应用，但若能得到一个更完整的结果，在美学上会更加令人满意）。

参考文献

[AMOP08] Alexandr Andoni、Andrew McGregor、Krzysztof Onak 和 Rina Panigrahy。随机顺序流中频率矩的更优界。arXiv 预印本 arXiv:0808.2222，2008。2、3、4、6[AMS99] Noga Alon、Yossi Matias 和 Mario Szegedy。频率矩近似的空间复杂度。《计算机与系统科学杂志》，58(1):137–147，1999。2、3、5[BIPW10] Khanh Do Ba、Piotr Indyk、Eric Price 和 David P Woodruff。稀疏恢复的下界。载于第二十一届 ACM-SIAM 离散算法研讨会论文集，第 1190–1197 页。SIAM，2010。7[BR00] Omer Barkol 和 Yuval Rabani。单元探测模型中最近邻搜索及相关问题的更紧界。载于第三十二届 ACM 计算理论研讨会论文集，第 388–396 页，2000。7[BVWY18] Vladimir Braverman、Emanuele Viola、David P Woodruff 和 Lin F Yang。重新审视随机顺序流中的频率矩估计。载于第 45 届国际自动机、语言与编程研讨会（ICALP 2018）。Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik，2018。2、3、4[BYJKS04] Ziv Bar-Yossef、Thathachar S Jayram、Ravi Kumar 和 D Sivakumar。数据流与通信复杂度的信息统计方法。《计算机与系统科学杂志》，68(4):702–732，2004。3、4、5、13

[CCM08] Amit Chakrabarti、Graham Cormode 和 Andrew McGregor。通信与数据流计算的稳健下界。载于第四十届 ACM 计算理论研讨会论文集，第 641–650 页，2008。3[CJP08] Amit Chakrabarti、TS Jayram 和 Mihai Pˇatraşcu。随机排序数据流中选择问题的紧下界。载于第十九届 ACM-SIAM 离散算法研讨会论文集，第 720–729 页，2008。3[CKLM18] Arkadev Chattopadhyay、Michal Kouck`y、Bruno Loff 和 Sagnik Mukhopad- hyay。模拟胜过丰富性：新的数据结构下界。载于第 50 届 ACM SIGACT 计算理论研讨会论文集，第 1013–1020 页，2018。7[CKS03] Amit Chakrabarti、Subhash Khot 和 Xiaodong Sun。集合不相交性多方通信复杂度的近最优下界。载于第 18 届 IEEE 计算复杂性年会论文集，第 107–117 页。IEEE，2003。3、4、5、13[CMVW16] Michael Crouch、Andrew McGregor、Gregory Valiant 和 David P Woodruff。随机流：样本复杂度与空间复杂度。载于第 24 届欧洲算法研讨会（ESA 2016）。Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik，2016。2、3[DDKS16] Itai Dinur、Orr Dunkelman、Nathan Keller 和 Adi Shamir。用于大海捞针问题的内存高效算法。载于国际密码学年会论文集，第 185–206 页。Springer，2016。7[DLOM02] Erik D Demaine、Alejandro López-Ortiz 和 J Ian Munro。有限空间下互联网数据包流的频率估计。载于欧洲算法研讨会论文集，第 348–360 页。Springer，2002。3[GH09] Sudipto Guha 和 Zhiyi Huang。重新审视随机顺序流中的直和定理与空间下界。载于国际自动机、语言与编程研讨会论文集，第 513–524 页。Springer，2009。2、3、4[GM07] Sudipto Guha 和 Andrew McGregor。空间高效的采样。载于人工智能与统计学会议论文集，第 171–178 页。PMLR，2007。3[GM09] Sudipto Guha 和 Andrew McGregor。数据流顺序与次序统计量：随机顺序流中的分位数估计。SIAM计算杂志，38(5):2044–2059，2009。3[Gro09] Andre Gronemeier。AND 函数与不相交性的 NIH 多方信息复杂度渐近最优下界。载于第 26 届计算机科学理论方面国际研讨会。Schloss Dagstuhl-Leibniz-Zentrum für Informatik，2009。5、9、13

[Gro10]André Gronemeier。信息复杂度与基本问题的数据流算法。2010。https://eldorado.tu-dortmund.de/bitstream/2003/27529/1/Gronemeier2010.pdf。

13[HW07] Johan Håstad 和 Avi Wigderson。集合不相交性的随机通信复杂度。《计算理论》，3(1):211–219，2007。5[IW05] Piotr Indyk 和 David Woodruff。数据流频率矩的最优近似。载于第三十七届 ACM 计算理论研讨会论文集，第 202–208 页，2005。3[Jay09] T. S. Jayram。海灵格再度出击：关于与的多方信息复杂度的注记。APPROX '09 / RANDOM '09，第 562–573 页，柏林、海德堡，2009。施普林格出版社。5, 9, 13[JKKR03] TS Jayram、Subhash Khot、Ravi Kumar 和 Yuval Rabani。部分匹配问题的单元探测下界。载于第三十五届 ACM 计算理论研讨会论文集，第 667–672 页，2003。7[KPW21] Akshay Kamath、Eric Price 和 David P Woodruff。一个关于集合不相交性的新证明及其在数据流中的应用。arXiv 预印本 arXiv:2105.11338，2021。3, 4[MNSW95] Peter Bro Miltersen、Noam Nisan、Shmuel Safra 和 Avi Wigderson。数据结构与非对称通信复杂度。载于第二十七届 ACM 计算理论研讨会论文集，第 103–111 页，1995。7[MP80] J Ian Munro 和 Mike S Paterson。有限存储下的选择与排序。《理论计算机科学》，12(3):315–323，1980。3[MPTW12] Andrew McGregor、A Pavan、Srikanta Tirthapura 和 David Woodruff。对子采样数据流上的统计量进行空间高效估计。载于第 31 届 ACM SIGMOD-SIGACT-SIGAI《数据库系统原理》研讨会论文集，第 273–282 页，2012。3[MPTW16] Andrew McGregor、A Pavan、Srikanta Tirthapura 和 David P Woodruff。对子采样数据流上的统计量进行空间高效估计。《算法》，2(74):787–811，2016。3[PT06] Mihai Pătraşcu 和 Mikkel Thorup。前驱搜索的时间-空间权衡。载于第三十八届 ACM 计算理论研讨会论文集，第 232–240 页，2006。7

[YZ22] Guangxu Yang 和 Jiapeng Zhang。提升定理与信息复杂度相遇：集合不相交性的已知与新下界。手稿，2022。5