arXiv:2304.01438v3 [cs.DS] 5 Oct 2024

伪确定性近似计数的紧空间下界

Ofer Grossman MeghalGupta Mark Sellke

摘要

我们研究流式算法中最基本的问题之一：近似计算数据流中的元素数量。众所周知， [Mor78] 提出了一种随机算法，对于长度至多为 N 的流，可在空间 $O(\log\log N)$ 内实现常数因子的近似误差。我们研究该问题的伪确定性复杂度，并证明了一个紧确的Ω(log N)下界，从而解决了[GGMW20]提出的问题。

1 引言

流式算法的研究起源于 Morris 的开创性论文[Mor78], ，该论文提出了一种低内存随机算法，用于近似计数在线到达数据流中的元素数量。粗略地说，其思路是维护一个计数器，使其近似于$\log n$，其中 n 表示截至目前在数据流中已见到的元素数量。算法每遇到一个来自数据流的元素，就以约 $\frac{1}{n}$ 的概率增加计数器。后来，正如[Fla85], 所证明的，Morris 的算法对于长度至多为 N 的流，可在空间 $O(\log\log N)$内实现常数因子的近似误差。

Morris 算法有这样一个性质：对同一数据流多次运行，可能会得到不同的近似结果。也就是说，如果 Alice 和 Bob 对同一数据流运行该算法（但使用不同的随机性），Alice 可能得到某个近似值（例如 2^{30}），而 Bob（运行同一算法，但使用独立的随机性）可能得到不同的近似值（例如2^{29}）。这种现象是不可避免的吗？也就是说，是否存在一种低空间算法，虽然带有随机性，但对于所有数据流，Alice 和 Bob 以高概率都会得到相同的长度近似值？这类算法，即在同一输入上多次运行时以高概率输出相同结果的算法，称为伪确定性算法。

关于数据流近似计数的伪确定性空间复杂度这一问题，最早由 [GGMW20] 提出。我们的主要结果彻底解决了这一问题，给出了紧的 Ω(log N) 下界。事实上，我们的下界适用于一个更简单的阈值版本问题：判断数据流长度至多为 N 还是至少为 M，其中 $N < M$ 为任意整数。此外，该下界仅取决于 N；也就是说，即使 M 远大于 N，该问题仍然很难。

定理 1.1（非正式）。对于任意 $N < M$，要区分长度至多为 N 和至少为 M 的数据流，伪确定性流算法至少需要 Ω(log N) 空间。

与此同时， [BKKS23] 通过不同的技术证明了一个非平凡的 $\Omega(\sqrt{\log n/\operatorname{loglog} n})$ 空间下界。

伪确定性的动机。随机化是算法设计中的一种强大工具。例如，几乎所有经典流算法（例如用于重频项、近似计数、 ℓ_p 近似、在转台算法中查找向量的非零项，以及统计数据流中不同元素数量的算法）若不付出空间复杂度大幅增加的代价，就无法去随机化。然而，与确定性算法相比，随机化算法也有若干缺点：

- 它们存在非零的出错或失败概率。
- 随机　　比特必须经过采样，而这可能不可行或计算开销过大　　e.
- 输出不可预测。也就是说，对同一输入多次运行该算法，可能会得到不同的输出。

上述前两点对随机算法的研究产生了重要影响。例如，第一点促使人们区分蒙特卡洛算法与拉斯维加斯算法，以及 RP 与 BPP，并推动了对朴素放大方法的改进[IZ89，GM20]。第二点推动了对伪随机生成器的研究 [HILL99], 、如何回收随机比特 [IZ89], 以及高效提取器 [Tre01], ，仅举几例。伪确定性的研究伪确定性由 Gat 和 Goldwasser 提出 [GG11] ，旨在解决第三个问题。我们的结果表明，对于任何以较低空间复杂度解决近似计数（或阈值）问题的算法而言，第三个缺点都无法避免。

1.1 相关工作

关于数据流中近似计数的先前工作。流式算法的研究始于 Morris 的工作[Mor78] ， 后来 [Fla85] 给出了严格的分析。如前所述， （随机化的）Morris 计数器所需的状态数比（确定性的）精确计数器少一个对数因子。该近似计数器已成为其他流式算法的有用理论基元 [AJKS02、GS09、KNW10、BDW18、JW19] ， 其性能也已得到广泛评估 [Cve07、Cs¨u10、XKNS21]。

近来，[NY22]确定了近似计数中对误差水平和概率的最优依赖关系。此外， [AAHNY22] 研究了维护多个近似计数器（而非仅一个）的摊还复杂度。

伪确定性研究进展。Gat 和 Goldwasser [GG11] 提出了伪确定性算法，旨在将其应用于密码学。此后，这类算 [序算法17],、平均情形算法 [、并行算法 [、决策树算法[、交互式证明 [、学习算法[、近似算法 [、低空间算法 [和复杂度领域中均对伪确定性算法进行了研究。

理论 [LOS21、DPWV22] 以及其他领域。在其中一些场景中，研究表明，伪确定性算法的表现优于确定性算法，有时甚至能达到随机化算法的最先进水平。

本文[GGMW20] 开启了对伪确定性流式算法的研究。他们考察了若干问题（例如寻找重频项、ℓ_2 近似，以及在转台算法中寻找向量的非零项），并提出近似计数是否存在一种使用 $O(\log\log n)$ 空间位数的伪确定性算法。我们给出了最优的$\Omega(\log\ n)$ 下界。

在本文定稿之际，[BKKS23] 独立证明了一个$\Omega(\sqrt{\log\ n/\log\log\ n})$下界。他们的证明通过归约到单向通信复杂度来完成。相比之下，我们的证明采用了截然不同的方法：通过将流算法建模为马尔可夫链，直接分析其复杂度，并证明在适当意义下它的行为类似于由若干循环部分组成的集合。

1.2 主要结果

我们考虑数据流中的伪确定性近似计数问题。首先回顾伪确定性流式算法的定义：

定义 1. 若存在某个函数 F，使得对于所有有效输入数据流 x，我们称流算法 $\mathcal{A}$ 为伪确定性，

$$\mathbb{P}_r[\mathcal{A}(x,r)=F(x)]\geq 2/3$$

其中 r 是算法采样并使用的随机性 $\mathcal{A}$。

我们的主要结果给出了紧确界 $\Omega(\log\ N)$ 位内存。事实上，我们证明，对于任意函数 $f\colon \mathbb{N}\to\mathbb{N}$，其中 $f(N)>N$，以伪确定性方式区分长度为 $T\leq N$ 和 $T\geq f(N)$ 的数据流需要 $\Omega(\log\ N)$ 位内存。

定义 2（伪确定性近似阈值问题）。固定 N、$M\in\mathbb{N}$。假设$\mathcal{A}$ 是一个伪确定性流式算法，对于数据流长度 T，满足：

1. 对于 $T\leq N$，$F(T)=1$；
2. 对于 $T\geq M$，$F(T)=0$。

那么 w 我们称 $\mathcal{A}$ 解决了 (N,M)-伪确定性近似阈值问题 .

定理 1.2. 对任意 $f\colon \mathbb{N}\to\mathbb{N}$，其中 $f(N)>N$，任何解决 $(N,f(N))$-近似阈值问题的伪确定性算法序列 $\mathcal{A}_N$ 都需要 $\Omega(\log\ N)$ 个空间位数。

由于计数 比阈值判定更难，下面的推论可直接得出 .

推论 1.3. 对于长度至多为 N 的数据流，任何能够以常数乘法因子解决近似计数问题的伪确定性流算法，都需要 $\Omega(\log\ N)$ 位空间。

备注 1.4. 我们的结果还意味着，以伪确定性方式解决重频项问题某个变体的空间下界。考虑这个问题的一个版本：给定一个 $\{0,1\}$值数据流，我们的目标是输出一个至少出现了 10% 时间的比特。由推论 1.3可知，任何解决此问题的伪确定性流算法都需要 $\Omega(\log\ N)$ 位空间。实际上，如果公开信息表明前 $9N$ 位为 0 ，其余位为 1，那么通过计数 $(N, 81N)$ 个 1，我们就将问题转化为近似阈值问题。随机化算法仍然可以用 $O(\log\log\ N)$ 位空间解决该问题，因为只需分别为 0 和 1维护一对 Morris 计数器即可。（注意，尽管重频项通常被认为可以用常数空间确定性地解决，但这类解法一般假设精确计数只需常数空间。）

1.3 马尔可夫链表述

用马尔可夫链的语言重新表述近似阈值问题和伪确定性会很有帮助。注意，任何使用 2^b 个状态的 b 位伪确定性流算法，都可以描述为一条马尔可夫链。在本文余下部分，我们将采用这种表述，并在下文给出其精确定义。

设 $\mathcal{M}_n$ 是状态空间为 $V = \{v_1, \ldots, v_n\}$、初始状态为 $x_0 = v_1$ 的任意马尔可夫链，并令 x_1、x_2、..... 表示之后各个时刻的随机状态。此外，设 $U \subseteq V$ 是 V 的一个指定子集。

定义 3（伪确定性近似阈值判定的马尔可夫链解）。如果 $(\mathcal{M}_n, U)$ 满足以下条件，我们称其为近似阈值问题的 $(N, f(N))$-解：

(i) 对所有 $t \le N$， $\mathbb{P}[x_t \in U] \ge 2/3$；
(ii) 对所有 $t \ge f(N)$， $\mathbb{P}[x_t \in U] \le 1/3$。

我们称 $\mathcal{M}_n$ 是伪确定性的，当且仅当对于所有 $\mathbb{P}[x_t \in U] \in [0, 1/3] \cup [2/3, 1]$ t.

我们将证明以下结论：

定理 1.5. 对于任意伪确定性 $(\mathcal{M}_n, U)$，若其以 $(N, f(N))$-方式求解某个 $f(N) > N$ 的近似阈值问题，则有 $n \ge N^{\Omega(1)}$。

容易看出，定理 1.5 通过将使用b 位内存的流算法视为一个具有 2^b 个状态的马尔可夫链，即可推出定理 1.2 。这里 U 是算法输出1的算法状态集合。

2 技术概述

2.1 示例说明

我们首先介绍两个无法解决伪确定性近似阈值问题的马尔可夫链示例，但它们最终出人意料地说明了一般情形。

示例 1：阈值 Morris 计数器。我们先介绍如何在通常的随机化（非伪确定性）设置下解决近似阈值判定问题：对于 $M = 10N$ 和 $t \in [N, 10N]$，不要求 $\mathbb{P}[x_t \in U] \in [0, 1/3] \cup [2/3, 1]$。我们将使用 Morris 计数器的一个版本，它采用简单的双状态实现（因为我们的目标是判断是否超过阈值，而不是计数）。

我们的马尔可夫链包含状态 (v_1, v_2)，其中 v_2 为终止状态，而从 v_1 到 v_2 的转移以概率 $\frac{1}{3N}$ 发生。形式上，$\mathbb{P}[x_{t+1} = v_2 \mid x_t = v_2] = 1$，且 $\mathbb{P}[x_{t+1} = v_2 \mid x_t = v_1] = \frac{1}{3N}$。令 $U = \{v_1\}$，该链便给出了一个用于 $N < 10N$ 近似阈值判定的随机算法。然而，这个例子并非伪确定性，因为函数 $t \mapsto \mathbb{P}[x_t \in U]$ 会逐渐衰减。特别地，容易看出 $\mathbb{P}[x_{t+1} \in U] \geq 0.99 \cdot \mathbb{P}[x_t \in U]$，因此根据离散时间介值定理，存在 t 使得 $\mathbb{P}[x_t \in U] \notin [0, 1/3] \cup [2/3, 1]$。

示例 2：素数循环。在第二个示例中，我们讨论一种徒劳的尝试：解决一个更简单版本的近似阈值问题，其中已知数据流长度至多为$10M = 100N$。非正式地说，第一步中，马尔可夫链会选取一个素数 p_i。之后的每一步，它只记录当前时间对 p_i取模的结果。我们证明中的一个主要思路（将在下一小节概述）是证明任何马尔可夫链 $\mathcal{M}_n$ 的行为都类似于（一个更复杂版本的）这个示例。

令 $k = [n^{1/4}, 2n^{1/4}]$，并选取 k 个互不相同的素数 $p_1, \ldots\ldots, p_k \in [n^{1/3}, 2n^{1/3}]$。注意到 $\sum_{i=1}^{k} p_i \leq n$，我们可以构造一个马尔可夫链 $\mathcal{M}_n$，其中包含一个初始状态 v_0，以及对每个 $1 \leq i \leq k$都有一个确定性的 p_i-循环。这里，p-循环由顶点$(v_0^{(p)}, \ldots\ldots, v_{p-1}^{(p)})$组成，每一步的状态转移都会确定性地将下标加 1，并对 p 取模。第一个时间步中，$\mathcal{M}_n$ 会移动到v_1，其中 p_i是均匀随机选取的。我们可以尝试通过在 $\mathcal{M}_n$的状态空间中选取一个特殊子集 $U \subseteq \bigcup_{1 \leq i \leq k} \{v_0^{(p_i)}, \ldots\ldots, v_{p_i-1}^{(p_i)}\}$，来解决伪确定性近似阈值问题（数据流长度受限于至多 $100N$）。

我们将在下文论证这是不可能的，这也将为主要证明提供一个有益的热身。注意，根据离散时间介值定理，对于任何这样的解，都必须存在$0 \leq T \leq 9N$，使得 $\mathbb{P}[x_t \in U] \geq 2/3$ 对大约一半的$t \in [T, T+1, \ldots, T+N]$取值成立。要与伪确定性矛盾，只需证明$\mathbb{P}[x_t \in U]$在范围 $t \in [T, T+1, \ldots, T+N]$上的方差上界为 $o(1)$。为此，只需证明存在某个常数 C，

$$\sum_{t \in [T, T+1, \ldots, T+N]} (\mathbb{P}[x_t \in U] - C)^2 \leq o(N). \tag{2.1}$$

我们将令 C 等于预期 $\mathbb{P}[x_t \in U]$ 取到的平均值，即 $C = \mathbb{E}_{i \in [k]} F(p_i)$，其中 $F(p_i)$是属于 U 的 $\{v_0^{(p_i)}, \ldots, v_{p_i-1}^{(p_i)}\}$所占的比例。将左侧的求和项改写后，我们需要给出其上界

$$\sum_{t \in [T, \ldots, T+N]} (\mathbb{P}[x_t \in U] - C)^2 = \sum_{t \in [T, \ldots, T+N]} \left(\frac{1}{k} \sum_{1 \leq i \leq k} \left[\mathbb{I}(v_t^{(p_i)} \in U) - F(p_i) \right] \right)^2.$$

其中，下标 $v^{(p_i)}$ 按模 p_i 计算，而 $\mathbb{I}$ 表示指示函数。展开

这个表达式并交换求和顺序后，得到

$$\frac{1}{k^2}\sum_{1\le i\le k}\left[\sum_{t\in[T,\dots,T+N]}\Big(\mathbb{I}(v_t^{(p_i)}\in U)-F(p_i)\Big)^2\right]$$
$$+\frac{1}{k^2}\sum_{1\le i,j\le k}\left[\sum_{t\in[T,\dots,T+N]}\Big(\mathbb{I}(v_t^{(p_i)}\in U)-F(p_i)\Big)\cdot\Big(\mathbb{I}(v_t^{(p_j)}\in U)-F(p_j)\Big)\right] \tag{2.2}$$

第一项为 $O(N/k) = o(N)$，因此我们只需对第二项进行界定；固定 i, j，并用 $o\ N$ p $\mathbb{I}\ v_t^{(p)}\ \in()$ 对内部 $U\ -F$ 0 表达式作一致估计。对于每个素数，表达式 $()\ (p)$ 在任意长度为 p 的区间上的平均值为 。因此，根据中国剩余定理，(2.2) 中的被加数在每个长度为 p_ip_j 的区间上的平均值为 0。考虑到这一抵消后，剩余项的数量至多为 $p_ip_j \le O(n^{2/3}) \le o(N)$。由于被加数的绝对值至多为 1，综合可得界 (2.1)，从而与伪确定性相矛盾。

2.2 证明概述

为导出矛盾，假设马尔可夫链 $(\mathcal{M}_n, U)$ 是伪确定性的，并且能够解决近似阈值问题。我们证明的主要思路是从 $\mathcal{M}_n$ 中提取一个行为大致类似于上文例 2 的子系统，即由周期长度各不相同的循环组成的集合。我们利用时刻 T 附近的这一行为，此时对于约一半的 $s \in [T, T+$，$\mathbb{P}[x_s \in U] \ge 2/3$ 成立，且 $s \in [T, T+\text{poly}(n)],$；根据离散时间介值定理，这样的时刻存在。最后，我们运用傅里叶分析证明，存在某个 $s \in [T, T+\text{poly}(n)]$，使得 $\mathbb{P}[x_s \in U]$ 的偏差不太大，这与 $(\mathcal{M}_n, U)$ 的伪确定性相矛盾。

中等时间尺度上的循环行为。 第一步是像示例 1 那样，通过考虑一个随机时间，绕开永久性不可逆转移的问题。我们选择一个随机时间 $1 \le t \le \text{poly}(n)$ 并证明，以高概率，该链 在顶点 x_t处表现出循环行为。更准确地说，令 $t+r$ 为下一个满足以下条件的时刻：$x_{t+r} = x_t$ (其中 $r = \infty$ 表示该链永不返回）。我们在引理3.1 中证明，以高概率，

$$\mathbb{P}[r \le 10n] \ge 1/2, \tag{2.3}$$
$$\mathbb{P}[r \le n^{18}] \ge 1 - n^{-16}. \tag{2.4}$$

利用这一保证，我们可以从（随机）状态的视角观察马尔可夫链x_t，方法是考察循环长度序列。界（2.4）使我们能够将这一过程定义至较大的 $\text{poly}(n)$ 个时间步；以高概率，所有循环长度均不超过 n^{18}。这避开了例 1 中的情形：尽管该链最终可能转移到某个终止状态，并且再也不会返回到 x_t，但在某个 $\text{poly}(n)$ 时间窗口内，它通常不会如此。

周期部分分解。接下来，由（2.3）可知，存在一个整数$1 \le m_v \le 10n$，使得在给定 $x_t = v$ 的条件下，每个返回到 v 的循环，其长度恰为 $\Omega(1/n)$ 的概率为 m_v。我们将长度恰为 m_v 的循环称为特殊循环，并进一步以一个经过屏蔽的循环长度过程 ℓ为条件。该过程 ℓ 会隐藏特殊循环，但按顺序揭示其他所有循环长度，直到揭示出数量适中且固定的 $\mathrm{poly}(n)$ 个非特殊循环后停止。[1]

举一个具体例子，假设在顶点 v，经过 3 步后返回v 的概率为 $1/4$，经过 5 步后返回 v 的概率为 $1/4$，经过 7 步后返回 v 的概率为 $1/3$，经过 11 步后返回 v 的概率为 $1/6$，并令 $m_v = 7$。假设在一次从 v出发的马尔可夫链运行中，所经历的循环长度依次为 (3, 5, 7, 11, 7, 5, 3, 7)。那么 $\vec{\ell} =$ (3, 5, 11, 5, 3)。

此后，隐藏的 m_v 循环数近似服从高斯分布，其标准差为 poly(n)。特别地，其概率质量函数在较短尺度上几乎保持不变，这使我们能够近似函数

$$\mathbb{P}[x_s \in U \mid x_t = v]$$

当 s 较大时，用一个 m_v-周期函数来近似。对 x_t 的随机性取平均后，我们将 $\mathbb{P}[x_s \in U]$ 近似为周期为 $m \le 10n$ 的周期函数的平均值。[2] 至此，证明的结构阶段完成。现在我们转而寻找一个时间 s，使得 $\mathbb{P}[x_s \in U] \in [1/3, 2/3]$。

周期分解分析。最后一步是在时间区间 T 上分析这种循环混合行为；在该区间上，约有一半的值满足 $\mathbb{P}[x_s \in U] \ge 2/3$ $s \in [T, T+\text{poly}(n)]$。根据离散时间介值定理，这样的 T必然存在。若不同的周期 m_v 互素，直观而言，在范围 $[T, T + \text{poly}(n)]$ 内，每对循环行为都近似独立，因为 $m_v m_v' \le 100n^2 \ll \text{poly}(n)$。这种两两近似独立性将使我们能够像示例 2 那样，对 $\sum_{s\in[T,T+\mathrm{poly}(n)]}(\mathbb{P}[x_s \in U] - 1/2)^2$给出上界，从而与伪确定性矛盾。

为处理一般的周期长度，关键思路是除去公有素因子，并限制在相应的等差数列上。具体而言，对每个素数 p，令 $e_p \ge 0$ 为满足以下条件的最大整数：m_v 以至少某个常数概率是 p^{e_p}的倍数。随后可以证明，$\beta = \prod_{p\text{素数}} p^{e_p} \le \text{poly}(n)$，因此我们将注意力限制在一个固定的等差数列上，其公差为 β。

将范围限制在这个等差数列上，会使不同的循环表现得几乎像是互素。具体来说，“约化周期长度” $M_v = m_v / \gcd(m_v, \beta)$具有这样的性质：任意固定的素数 p 整除 M_v的概率都很低。这一性质并不意味着不同循环两两独立：对于独立的 M_v、M_v'，$\gcd(M_v, M_v') > 1$ 以高概率成立仍然可能。然而，借助傅里叶分析，我们仍然能够证明，这一性质蕴含着关于$\sum_{s\in[T,T+\mathrm{poly}(n)]}(\mathbb{P}[x_s \in U]-1/2)^2$的一个非平凡上界。如上所述，这样的上界与伪确定性相矛盾，从而完成证明。

[1]严格来说，我们会以某个正概率将长度 m_v 循环判定为非特殊。这是必要的，例如当循环长度 始终 等于 m_v 时。[2]我们使用一种称为 c-覆盖的比较概念来形式化这种近似，见定义 5。

定理 3.1 的证明1.5

假设 $n \geq n_0$ 足够大，并令阈值为 $N > n^{34}$。我们将证明，这样的马尔可夫链无法对任何(N, M)-求解近似计数问题 M。我们将在定理 $[0, 1/3]$ 和 $[2/3, 1]$ 的陈述中分别用 $[0, 10^{-4}]$ 和 $[1 - 10^{-4}, 1]$ 替换区间1.5；根据放大，这两种表述是等价的。

3.1 中等时间尺度上的再现行为

对于任意 $v \in V$，令 μ_v 为从 v 出发的返回时间分布，即从 $x_0 = v$ 开始时，使得 $x_r = v$ 成立的最小 $r \geq 1$ 所对应的分布。（如果此条件之后不再成立，则令 $r = \infty$。）

引理 3.1. 存在 $1 \leq t \leq N/100$ 和一个子集 $S \subseteq V$，使得：

(i) $\mathbb{P}[x_t \in S] \geq 1/2$.

（ii）对于所有 $v \in S$:

$$\mathbb{P}_{r\sim\mu_v}(r \in [0, 10n]) \geq 1/2,$$
$$\mathbb{P}_{r\sim\mu_v}(r \in [0, n^{18}]) \geq 1 - n^{-16}.$$

证明。我们首先证明，$1 \leq t \leq N/100$ 的取值中至少有 0.85 的比例满足：

$$\mathbb{E}_{x_t}\mathbb{P}_{r\sim\mu_{x_t}}(r \in [0, 10n]) \geq 0.85 \tag{3.1}$$

将区间 $[1, N/100]$ 划分为长度为 $10n$ 的区间 $I_j = [1 + (j-1)10n, 10jn]$，其中 $1 \leq j \leq N/1000n$。考虑马尔可夫链 $x_1 \cdots x_{N/100}$ 的一次固定运行。对于每个区间 I_j，至多有 n 个 z_i 的取值在 I_j 中的某个更晚时刻不会返回自身，其中 $i \in I_j$。因此，在马尔可夫链的一次固定运行中，从时刻 t 开始返回 x_t 所需的时间至多为 $10n$ 步，这对至少 0.9 比例的时刻 $1 \leq t \leq N/100$ 成立。对所有运行取平均，可得

$$\mathbb{E}_{t\in[1,N/100]}\mathbb{E}_{x_t}\mathbb{P}_{r\sim\mu_{x_t}}(r \in [1, 10n]) \geq 0.9$$

因此至少有 0。85 比例的 $0 < t < N/100$ 满足式（3.1），根据马尔可夫不等式。用相同的论证，至少有 0。85 比例的 $1 \leq t \leq N/100$,

$$\mathbb{E}_{x_t}\mathbb{P}_{r\sim\mu_{x_t}}(r \in [0, n^{18}]) \geq 1 - 10n^{-17}$$

特别地，存在一个满足这两个条件的值 t，它将作为引理陈述中的 t。

接下来，对3.1)应用马尔可夫不等式可知，存在某个 $S_1 \subseteq V$满足 $\mathbb{P}[x_t \in S_1] \geq 0.6$,

$$\mathbb{P}_{r\sim\mu_v}(r \in [1, 10n]) \geq 1/2.$$

同样，对于某个 $S_2 \subseteq V$，使得 $\mathbb{P}[x_t \in S_2] > 0.9$,

$$\mathbb{P}_{r\sim\mu_v}(r \in [0, n^{18}]) \geq 1 - 100n^{-17} \geq 1 - n^{-16}.$$

取 $S = S_1 \cap S_2$ 即完成证明。 □

3.2 分解为周期部分

从现在起，　我们按照上一小节的方式固定 t, S。对于每个 $v \in S$，定义周期长度

$$m_v \equiv \operatorname*{argmax}_{1 \le m \le 10n} \mu_v(m).$$

从 $x_t = v$ 开始，我们按如下方式定义一个步序列。设 $(C_1, C_2, \dots)$ 是返回位置 x_t 的循环长度序列，即对于 $j \ge 0$，恰好有 $x_{t+C_1+\cdots+C_j}$ 满足 $s \ge t$，且 $x_s = x_t$。当 $C_i = m_v$ 时，我们以概率 C_i 将其指定为特殊循环

$$\mathbb{P}[C_i \text{ special} \mid C_i = m_v] = \frac{n^{-2}\mu_v(\{1, 2, \dots, n^{18}\})}{\mu_v(m_v)}. \tag{3.2}$$

这种概率选择会产生 $\mathbb{P}[C_i$ 个特殊 $\mid C_i \le n^{18}] = n^{-2}$，这将在后文派上用场。在所有其他满足 $C_i \le n^{18}$ 的情况下，我们将 C_i 指定为典型循环。我们将特殊步序列 $(s_1, s_2, \dots) = (m_v, m_v, \dots)$ 定义为特殊循环长度构成的子序列，将典型步序列 $(\ell_1, \ell_2, \dots)$ 定义为典型循环长度构成的序列。注意，长度大于 n^{18} 的循环既不是特殊循环，也不是典型循环。

如果 $x_t = v$，令 τ_v 为 n^{14} 完成典型步骤的首次时刻。（对于所有其他状态 v，我们令 τ_v 未定义或为无穷大。）令 Y 为截至时刻 τ_{x_t} 的特殊循环数量，并令 $s_{\text{tot}} = Y m_v$ 为这些特殊循环的总持续时间。

对于每个 $v \in S$，令 E_v 表示以下事件：$x_t = v$，且截至时刻 τ_v 的所有循环都是特殊循环或典型循环（在该事件中，尤其有 $\tau_v < \infty$）。

引理 3.2。对于任意 $v \in S$，有 $\mathbb{P}[E_v \mid x_t = v] \ge 0.99$。

证明。注意，每个循环长度 C_i 相互独立，且以至少 $1 - n^{-16}$ 的概率不超过 n^{18}（由引理 3.1）。我们发现，以高概率，前 n^{15} 个循环长度都是特殊循环长度或典型循环长度：

$$\mathbb{P}\left[\max_{1 \le i \le n^{15}} C_i \le n^{17}\right] \ge 1 - n^{-1}. \tag{3.3}$$

在条件事件（3.3）下，每个 C_i 独立地以至少 $1/2$ 的概率被标记为典型。因此，前 n^{15} 个循环中的典型步骤数量随机占优于一个二项 $(n^{15}, 1/2)$ 随机变量；尤其是，该数量至少为 n^{14} 的概率不低于 $1 - n^{-1}$。在此事件下，τ_v 会在前 n^{15} 个循环内发生，因此 E_v 成立的总体概率至少为$(1 - n^{-1})(1 - n^{-1}) \ge 0.99$。 □

接下来，我们利用特殊循环和典型循环的概念来描述 τ_v 的分布。回顾一下，τ_v 指的是在以 $x_t = v$ 为条件时，完成 n^{14} 个典型步骤所需的首次时刻。

定义 4. 若概率分布 ν 在 $\mathbb{Z}$ 上是 m-区间周期的，则其支撑集为一个离散区间 $\{I, I+1, \dots, J\}$，并且在此区间内，$\nu(j) = \nu(j+m)$ 仅取决于 j 对 m 取模的结果。ν 的取值范围为 $J - I$。

类似地，若概率分布序列 $\vec{\nu} = (\nu_I, \nu_{I+1}, \dots, \nu_J)$ 定义在 V 上，且 m 为区间周期的，则当且仅当 $\nu_j = \nu_{j+m}$ 仅取决于 j 对 m 取模后的值。

定义 5. 设 μ,ν 是可数集 $\mathcal{I}$ 上的概率分布，且 $0<c<1$ 是一个常数。若对所有 $i\in\mathcal{I}$，都有 $\mu(i)\geq c\nu(i)$，则称 μ 是 ν 的一个 c-覆盖。

令 $\mu_{\mathrm{tot},v,\vec{\ell}}$ 表示在事件 E_v 以及 n^{14} 个典型步骤的序列 $\vec{\ell}$ 条件下，τ_v 的分布。事实上，从 (3.2) 很容易看出，在 (v,E_v) 的条件下，每对相邻典型步骤 (ℓ_{i-1},ℓ_i) 之间的特殊步骤数 z_i 恰好服从几何分布 $\mathrm{Geom}(1-n^{-2})$。因此，对于每个 $j\in\mathbb{Z}_{\geq 0}$，$\mathbb{P}[z_i=j]=(1-n^{-2})n^{-2j}$，且其均值和方差已知为

$$\mathbb{E}[z_i]=\frac{1}{n^2-1},$$
$$\mathrm{Var}[z_i]=\frac{n^2}{(n^2-1)^2}.$$

引理 3.3. 对于每个 $v\in S$ 和 ℓ，存在一个 m_v-区间周期分布 $\nu_{v,\vec{\ell}}$，其取值范围为 $n^6<w_v<n^8$，使得分布 $\mu_{tot,v,\vec{\ell}}$ 是 $\nu_{v,\vec{\ell}}$ 的 0.06-覆盖。

证明. 令 $Z=\sum_{i=1}^{n^{14}}z_i$ 表示截至时间 τ_v 的特殊循环总数。我们首先证明，当 $n\to\infty$ 时，Z 满足中心极限定理。注意，$w_i=z_i-\mathbb{E}[z_i]$ 满足 $\mathbb{E}[w_i]=0,\mathbb{E}[w_i^2]=\Theta(n^{-2})$ 和 $\mathbb{E}[|w_i|^3]=\Theta(n^{-2})$。因此，由 Berry–Esseen 定理可知，令 $\Phi(A)=\mathbb{P}^{y\sim\mathcal{N}(0,1)}[y\leq A]$ 表示标准高斯分布的累积分布函数，并令 μ_Z,σ_Z 分别表示 Z 的均值和标准差：

$$\sup_{A\in\mathbb{R}}\left|\mathbb{P}\left[\frac{Z-\mu_Z}{\sigma_Z}\leq A\right]-\Phi(A)\right|\leq\frac{Cn^{-2}}{n^{-3}\sqrt{n^{14}}}\leq O(1/n^2).$$

特别地，当 n 足够大时，

$$\mathbb{P}[Z\in[\mu_Z-2\sigma_Z,\mu_Z-\sigma_Z]]\geq\Phi(-1)-\Phi(-2)-0.001\geq 0.13,$$
$$\mathbb{P}[Z\in[\mu_Z+\sigma_Z,\mu_Z+2\sigma_Z]]\geq\Phi(2)-\Phi(1)-0.001\geq 0.13.$$

通过取平均值，存在整数 $a_1\in[\mu_Z-2\sigma_Z,\mu_Z-\sigma_Z]$ 和 $a_2\in[\mu_Z+\sigma_Z,\mu_Z+2\sigma_Z]$，使得

$$\mathbb{P}[Z=a_1],\mathbb{P}[Z=a_2]\geq\frac{1}{8\sigma_Z}.$$

接下来，我们声称 Z 的概率质量函数是对数凹的；这意味着

$$\mathbb{P}[Z=a]\geq\frac{1}{8\sigma_Z},\quad\forall a\in[\mu_Z-\sigma_Z,\mu_Z+\sigma_Z].\tag{3.4}$$

由于 $\mu_{\mathrm{tot},\ v,\ \vec{\ell}}$ 恰好是 m_vZ 和 $\sigma_Z=n^7\sqrt{\mathrm{Var}[z_1]}=\sqrt{n^6(1\pm O(1/n))}$ 的概率分布，命题（3.4）立即推出所需结果。只需证明 Z 具有对数凹概率质量函数。

为此，首先注意到 Z 是一个负二项分布，其概率质量函数的精确形式为

$$\mathbb{P}[Z=k]=\binom{k+n^{14}-1}{n^{14}-1}n^{-14k}(1-n^{-2})^{n^{14}}.$$

（这种描述广为人知，可由初等的星与杠论证得出。）因此，只需证明 $k \mapsto \binom{k+n^{14}-1}{n^{14}-1}$ 是对数凹的。为此，我们注意到，对于定义在 $k \in \mathbb{Z}_{\geq 0}$ 上的每个 $c \geq 0$， $k \mapsto k+c$ 都是对数凹的，并且

$$\binom{k+n^{14}-1}{n^{14}-1} = \frac{(k+n^{14}-1)(k+n^{14}-2)\cdots(k+1)}{(n^{14}-1)!}$$

正比于这类序列的乘积。 □

我们指出，对 ℓ 的条件化在上述论证中不起作用，因为它只会平移 $\mu_{tot,\ v,\ \vec{\ell}}$。然而，在下一小节中，在对3.3 进行条件化后再应用引理 ℓ 将十分重要。

对于每个 $v \in S$，典型步序列 $\vec{\ell}$，以及时间 $s > t$，定义概率分布

$$\widetilde{\mu}_s = \widetilde{\mu}_s(v, \vec{\ell})$$

将 $\vec{}$ 在 V 上的分布定义为在 (E_v, ℓ) 条件下 x_s 的条件分布。

引理 3.4. 对于每个 $v \in S$、$\vec{\ell}$ 和 $T' > n^{33}$，存在一个 m_v-区间周期序列 $\vec{\omega} = \vec{\omega}_v = (\omega_{T'}, \ldots, \omega_{T'+n^5})$，其分布被序列 $\widetilde{\mu}_{T'} \ldots \widetilde{\mu}_{T'+n^5}$ 以 0.03 覆盖。

证明。对于任意顶点 u 和时刻 r，定义 $P^r(u)$ 为从顶点 u 出发经过 r 步后马尔可夫链状态的分布。令

$$\operatorname{supp}(\nu_{v,\vec{\ell}}) = \{I, I+1, \ldots, J\} = \{I(\nu_{v,\vec{\ell}}), I(\nu_{v,\vec{\ell}})+1, \ldots, J(\nu_{v,\vec{\ell}})\}$$

为 $\nu_{v,\ \vec{\ell}}$ 的支撑集。此外，令 $\overline{\nu}_{v,\ \vec{\ell}}(0), \ldots, \overline{\nu}_{v,\ \vec{\ell}}(m_v-1)$ 满足 $\nu_{v,\ \vec{\ell}}(j) = \overline{\nu}_{v,\ \vec{\ell}}(j \bmod m)$，其中 $I \leq j \leq J$。

我们显式计算 $T' \leq s \leq T' + n^5$ 的 $\widetilde{\mu}_s$。用正测度不等式表示其差是一个正测度，我们有

$$\begin{aligned}
\widetilde{\mu}_s(v, \vec{\ell}) &\geq \sum_{d=1}^{s} \mathbb{P}[\tau_v = d | (x_d = v, \vec{\ell})] \cdot P^{s-d}(v) \\
&\geq 0.06 \sum_{d=1}^{s} \nu_{v,\vec{\ell}}(d) \cdot P^{s-d}(v) \\
&= 0.06 \sum_{u=I-T'}^{J-T'-n^5} \nu_{v,\vec{\ell}}(s-u) \cdot P^u(v) \\
&\geq 0.06 \left(\frac{J-I-n^5-2m}{m}\right) \cdot \sum_{i=1}^{m} \overline{\nu}_{v,\vec{\ell}}(s-i \mod m) \sum_{\substack{I-T' \leq u \leq J-T'-n^5 \\ u \equiv i \bmod m}} P^u(v).
\end{aligned}$$

注意，这个下界在给定范围内恰好以 s 为周期，周期为 m。下面证明它具有显著的总质量。注意

$$\begin{aligned}
&\left(\frac{J-I+n^5+2m}{m}\right) \cdot \sum_{i=1}^{m} \overline{\nu}_{v,\vec{\ell}}(s-i \mod m) \sum_{\substack{I-T' \leq u \leq J-T'-n^5 \\ u \equiv i \bmod m}} P^u(v) \\
&\geq \sum_{t=1}^{s} \nu_{v,\vec{\ell}}(t) \cdot P^{s-t}(v)
\end{aligned}$$

的总质量至少为 0.99，因为显然 $\mathbb{P}[\tau_v < s \mid E_v] \geq 0.99$（由 s_{tot} 上的马尔可夫不等式可得）。由于 $J - I \geq n^6$，我们有

$$\frac{J - I - n^5 - 2m}{J - I + n^5 + 2m} \geq 0.99$$

因此，我们得出上述下界的总质量至少为 0.03。 □

3.3 周期分解分析

我们固定引理 m_v 中的取值$^{\nu_v \text{和分布} \vec{\ell}}$，并取自3.3。接下来，我们按如下方式构造一个“全局周期长度” β。对于每个素数 p，令 e_p 为满足以下条件的最大值

$$\mathbb{P}[p^{e_p} \text{ divides } m_v \mid v \in S] \geq 0.55$$

并定义 $\beta = \prod_p p^{e_p}$。

引理 3.5。成立：$\beta \leq n^2$。

证明。只需注意到

$$\log(10n) \geq \mathbb{E}[\log(m_v) \mid v \in S] \geq \log(\beta^{0.55}).$$

前者成立，因为 $m_v \leq 10n$；后者则由马尔可夫不等式得出。因此 $\beta \leq (10n)^{20/11} \leq n^2$。 □

令 $M_v = m_v/$ 为 m_v和 β的最大公约数。根据定义，对于任意素数 p，有

$$\mathbb{P}[p \text{ divides } M_{x_t} \mid x_t \in S] \leq 0.55. \tag{3.5}$$

回顾第 1.3 小节，对于每个时刻 s令 $F(s) = 0$当 $\mathbb{P}[x_s \in U] \in [0, 1/3]$且 $F(s) = 1$，否则。此外，定义

$$\overline{F}_a = \frac{1}{\lfloor n^5/\beta \rfloor} \sum_{i=1}^{\lfloor n^5/\beta \rfloor} F((a+i)\beta).$$

引理 3.6。存在 $a \in \mathbb{N}$，使得 $|\overline{F}_a - 1/2| \leq 0.01$。

证明。由离散时间介值定理可得。事实上，由于 $N \geq n^5$，我们有$\overline{F}_0 = 1$；而对于 $a \geq f(N)$，$\overline{F}_a = 0$。显然，对于所有 a，都有 $|\overline{F}_{a+1} - \overline{F}_a| \leq 0.01$。 □

在证明余下部分中，令 $T = a\beta$对于 a 如引理 3.6 中所述。注意到 $T > n^{33}$，因此满足引理 T' 中 3.4 对 $T \leq s \leq T + n^5$的要求，令

$$\overline{\omega}_s = \mathbb{E}^{x_t, \vec{\ell}}[\mathbb{P}^{v \sim \omega_{x_t, \ell, s}}[v \in U] \mid x_t \in S \text{ and } E_{x_t}].$$

（注意，$\{x_t \in S$和 $E_{x_t}\}$只是一个事件。）

引理 3.7. $\overline{\omega}_s \geq 0.99$若 $F(s) = 1$且 $\overline{\omega}_s \leq 0.01$若 $F(s) = 0$。

证明. 我们重点讨论后一个结论，因为两者是对称的。

给定 x_t，生成 $w \sim P^{s-t}(x_t)$，使其为从原始轨迹分布中抽取的马尔可夫链在时刻 s 的状态。根据期望的贝叶斯公式，由于事件 $\{x_t \in S \text{ 和 } E_{x_t}\}$ 的概率至少为 $1/3$，对其进行条件化后，任意非负随机变量的期望最多增加至原来的 3 倍。特别地：

$$\begin{aligned}\mathbb{E}^{x_t,\vec{\ell}}[\mathbb{P}[w \in U\,] \mid x_t \in S \text{ and } E_{x_t}] &\le \frac{\mathbb{E}^{x_t,\vec{\ell}}[\mathbb{P}[w \in U\,]]}{\mathbb{P}[x_t \in S \text{ and } E_{x_t}]} \\ &= \frac{\mathbb{P}[x_s \in U]}{\mathbb{P}[x_t \in S \text{ and } E_{x_t}]} \\ &\le 3 \cdot 10^{-4}.\end{aligned}$$

给定 x_t 和 ℓ，我们还令 $\widetilde{w} \sim \omega_{x_t,\,\vec{\ell},\,s}$ 对于 $\omega_{x_t,\,\vec{\ell},\,s}$ 如引理 3.4所述。于是，引理 0.03-覆盖保证3.4 直接推出

$$\begin{aligned}&\mathbb{E}^{x_t,\vec{\ell}}[\mathbb{P}[w \in U\,] \mid x_t \in S \text{ and } E_{x_t}] \\ &\ge 0.03 \cdot \mathbb{E}^{x_t,\vec{\ell}}[\mathbb{P}[\widetilde{w} \in U\,] \mid x_t \in S \text{ and } E_{x_t}] \\ &= 0.03 \cdot \overline{\omega}_s.\end{aligned}$$

合并上述结果即可完成证明。 □

引理 3.8。结论成立：

$$\left| \frac{1}{\lfloor n^5/\beta \rfloor} \sum_{i=1}^{\lfloor n^5/\beta \rfloor} \overline{\omega}_{(a+i)\beta} - \frac{1}{2} \right| \le 0.02.$$

证明。 根据引理 3.7， $|\overline{\omega}_s - F(s)| \le 0.01$ 对于每个 s。因此

$$\left| \frac{1}{\lfloor n^5/\beta \rfloor} \sum_{i=1}^{\lfloor n^5/\beta \rfloor} \overline{\omega}_{(a+i)\beta} - \frac{1}{2} \right| \le |\overline{F}_a - 1/2| + 0.01 \le 0.02.$$

□

我们将使用以下引理，其证明见第 4节。

引理 3.9. 以下结论对 $L = n^3$ 成立。固定 n，并令 $S = (S_1, S_2, \ldots, S_L) \in [-1, 1]^L$ 为一个随机序列，该序列以某个随机 $m = m(S) \le 10n$ 为周期，周期为 m。假设

$$-L/10 \le \mathbb{E} \sum_{d=1}^{L} S_d \le L/10 \tag{3.6}$$

并且对每个素数 p,

$$\mathbb{P}[p \textit{ divides } m] \le 0.55.$$

那么存在一个指标 $1 \le d \le L$ 使得

$$|\mathbb{E}[S_d]| \le 0.9.$$

综合以上结果，我们最终可以证明定理1.5。

定理*1.5*的证明：给定一个假定的$(\mathcal{M}_n, U)$，我们应用引理3.9 wi th

$$S = S_{x_d,\vec{\ell}} = \mathbb{P}^{v\sim\omega_{x_d,\ell,s}}[v \in U].$$

其中我们通过 $[0,1] \to [-1,1]$ 映射 $a \mapsto 2a-1$。根据引理3.8，由于$\mathbb{E}[S_d] = \overline{\omega}_{(a+d)\beta}$，第一个条件成立。由（3.5），第二个条件成立。因此，引理3.9 意味着存在 s 使得 $\overline{\omega}_s \in [0.02, 0.98]$。这与上述引理3.7 矛盾，从而导出矛盾并完成证明。 □

4 引理3.9的证明

这里我们使用3.9 傅里叶分析来证明引理。对于每个 $m \geq 1$，令

$$Z_m = \{0, 1/m, \ldots, (m-1)/m\}$$

^并令 $Z_m^* = Z_m \backslash \{0\}$。给定一个 m-周期序列 $S = (S_1, S_2, \ldots)$，令 $S: [0,1) \to \mathbb{C}$ 为其傅里叶变换

$$\widehat{S}(\alpha) = \begin{cases} \frac{1}{m}\sum_{j=1}^{m} e^{2\pi i j\alpha} S_j, & \alpha \in Z_m, \\ 0, & \alpha \notin Z_m\,. \end{cases} \tag{4.1}$$

定义 (4.1) 与将 S 视为定义在 $\mathbb{Z}$ 上的周期函数时的傅里叶变换相符，因此与周期 m无关——我们可以将 S 视为 km 对任意 $k \geq 1$ 都具有周期性，且 $\sim S$ 仍然一致。这使我们能够对周期长度不同的序列对 S, S' 使用傅里叶分析 m, m'。事实上，给定 $M = mm'$ 我们可以定义：

$$\langle S, S' \rangle = \frac{1}{M}\sum_{i=1}^{M} S_i S_i' \tag{4.2}$$

$$\|S\|_{L^2}^2 = \frac{1}{M}\sum_{i=1}^{M} S_i^2 \tag{4.3}$$

$$\langle \widehat{S}, \widehat{S'} \rangle = \sum_{\alpha\in Z_M} \widehat{S}(\alpha)\widehat{S'}(\alpha) \tag{4.4}$$

$$\|\widehat{S}\|_{\ell^2}^2 = \sum_{\alpha\in Z_m} |\widehat{S}(\alpha)|^2. \tag{4.5}$$

^ ^ ^帕塞瓦尔恒等式对模 M的结论意味着 $\langle S, S' \rangle = \langle S, S' \rangle$ 和 $\|S\|_{L^2} = \|S\|_{L^2}$。

下一个命题通过傅里叶平均来近似在 1, 2, . . ., L 上的平均值。以下我们令 μ 为随机序列 S的概率分布，并回顾：几乎必然地，$S \sim \mu$ 是 $m(S)$ 周期的。

命题4.1。则

$$\left|\frac{1}{L}\sum_{d=1}^{L}\mathbb{P}^{S\sim\mu}[S_d=1]^2-\mathbb{E}^{S,S'\sim\mu}\langle\widehat{S},\widehat{S'}\rangle\right|\le O(n^2/L), \tag{4.6}$$

$$\left|\left(\frac{1}{L}\mathbb{E}^{S\sim\mu}\sum_{d=1}^{L}S_d\right)^2-\mathbb{E}^{S,S'\sim\mu}[\widehat{S}(0)\widehat{S'}(0)]\right|\le O(n^2/L). \tag{4.7}$$

证明。注意，根据定义，

$$\frac{1}{L}\sum_{d=1}^{L}\mathbb{P}^{S\sim\mu}[S_d=1]^2=\frac{1}{L}\mathbb{E}^{S,S'\sim\mu}\big\langle(S_1\ldots,S_L),(S'_1,\ldots,S'_T)\big\rangle$$

其中 $\langle(S_1\ldots,S_L),(S'_1,\ldots,S'_T)\rangle=\frac{1}{L}\sum_{d=1}^{L}S_dS'_d$。由于 $M=mm'\le 100n^2$，

$$\left|\frac{1}{L}\mathbb{E}^{S,S'\sim\mu}\big\langle(S_1\ldots,S_L),(S'_1,\ldots,S'_T)\big\rangle-\langle S,S'\rangle\right|\le O(n^2/L).$$

$\widehat{\ }$ $\widehat{\ }$结合以上结果并利用 $\langle S,\ S'\rangle=\langle S,\ S'\rangle$ 根据帕塞瓦尔恒等式，式（4.6）的证明即告完成。式（4.7）的证明类似，因为

$$\left|\frac{1}{L}\mathbb{E}^{S\sim\mu}\sum_{d=1}^{L}S_d-\mathbb{E}^{S\sim\mu}[\hat{S}(0)]\right|=\left|\frac{1}{L}\mathbb{E}^{S\sim\mu}\sum_{d=1}^{L}S_d-\frac{1}{m(S)}\mathbb{E}^{S\sim\mu}\sum_{i=1}^{m(S)}S_i\right|\le O(n/L).$$

事实上 $(\mathbb{E}^{S\sim\mu}[\widehat{S}(0)])^2=\mathbb{E}^{S,\ S'\sim\mu}[\widehat{S}(0)\ \widehat{S'}(0)],$，且函数 $x\mapsto x^2$ 在 $x\in[-1,1]$ 上是利普希茨连续的。□

引理 3.9 的证明思路是 S，S' 具有相似周期的可能性很小。在展开 $\widehat{\ }$这一论证时，我们必须将偏差 $S(0)$ 与 $\alpha\in(0,\ 2\pi)$ 的贡献分别处理。下一个命题给出了处理后一类贡献的估计；请注意，下面的最后一项混合了 S，S'。

命题 4.2。对于 m 和 m'-周期函数 $S,S':\mathbb{Z}\to[-1,1]$，

$$\langle\widehat{S},\widehat{S'}\rangle\le\widehat{S}(0)\widehat{S'}(0)+\|\widehat{S}|_{Z^*_{m'}}\|_{\ell^2}$$

$\widehat{\ }$ $\widehat{\ }$其中 $S|_{Z^*_{m'}}$ 表示 S 对 $Z^*_{m'}$的限制。

$\widehat{\ }$ $\widehat{\ }$证明。被加数 $S(\alpha)S'(\alpha)$ 在 (4.4) 中仅当 $\alpha\in Z_g$ 对于 $g=$最大公约数(m,m') 时才非零。这给出了如下上界：

$$\widehat{S}(0)\widehat{S'}(0)+\|\widehat{S}|_{Z^*_g}\|_{\ell^2}\cdot\|\widehat{S'}|_{Z^*_g}\|_{\ell^2}.$$

这立即推出该结论，因为

$$\|\widehat{S}|_{Z^*_g}\|_{\ell^2}\le\|\widehat{S}|_{Z^*_{m'}}\|_{\ell^2}$$

（事实上，等号成立），而

$$\|\widehat{S'}|_{Z_g^*}\|_{\ell^2} \le \|\widehat{S'}|_{Z_M}\|_{\ell^2} = \|S'\|_{L^2} \le 1.$$

□

引理 4.3. 假设 S, S' 独立同分布。$\sim \mu$ 并且对于任意素数 p，$\mathbb{P}^{S\sim\mu}[p \text{ 整除 } m(S)] \le \varepsilon$。那么

$$\mathbb{E}^{S\sim\mu}\left[\|\widehat{S}|_{Z_{m'}^*}\|_{\ell^2}\right] \le \sqrt{\varepsilon}$$

证明。由假设直接可得，对于任意固定的 $\alpha \neq 0$，有 $\mathbb{P}[\alpha \in Z_{m'}^*] \le \varepsilon$。因此，利用期望的线性性可得

$$\mathbb{E}^{S\sim\mu}\left[\|\widehat{S}|_{Z_{m'}^*}\|_{\ell^2}^2\right] \le \varepsilon.$$

应用詹森不等式即可完成证明。 □

引理 ***3.9***的证明：对 (4.6) 应用马尔可夫不等式，并注意到 $n^2/L \le 1/n$ 只需证明

$$\mathbb{E}^{S,S'\sim\mu}\langle\widehat{S}, \widehat{S'}\rangle \overset{?}{\le} 0.8 < 0.9^2. \tag{4.8}$$

我们先处理左侧的偏差项。结合 (4.7) 和 (3.6) 可得

$$\mathbb{E}^{S,S'\sim\mu}[\widehat{S}(0)\widehat{S'}(0)] \le 0.02.$$

对于其余各项 $\alpha \neq 0$，结合命题4.2 与引理 4.3 （根据假设，该引理在$\varepsilon = 0.55$ 条件下成立）可得

$$\mathbb{E}^{S,S'\sim\mu} \sum_{\alpha\in Z_M^*} \widehat{S}(\alpha)\widehat{S'}(\alpha) \le \sqrt{0.55} \le 0.75.$$

将各项贡献相加，得到（4.8），从而完成证明。 □

致谢

感谢 Yang Liu 的有益讨论，并感谢 Rachel Zhang、Naren Manoj 和 Themistoklis Haris 对稿件的评论。Ofer Grossman 在麻省理工学院攻读研究生期间获得了美国国家科学基金会研究生研究奖学金计划资助；本研究的部分工作完成期间，他曾访问西蒙斯理论计算研究所。Meghal Gupta 在加州大学伯克利分校攻读研究生期间获得了校长奖学金资助。

参考文献

[AAHNY22] Ishaq Aden-Ali、Yanjun Han、Jelani Nelson 和 Huacheng Yu。《近似计数的摊还复杂度》。arXiv 预印本 arXiv:2211.03917，2022。

[AJKS02] Mikl´os Ajtai、TS Jayram、Ravi Kumar 和 D Sivakumar。《数据流中的逆序数近似计数》。载于第三十四届 ACM 计算理论研讨会论文集，第 370–379 页，2002 年。[BDW18] Arnab Bhattacharyya、Palash Dey 和 David P Woodruff。《插入流中 ℓ^1-重频项及相关问题的最优算法》。ACM 算法汇刊（TALG），15(1):1–27，2018 年。[BKKS23] Vladimir Braverman、Robert Krauthgamer、Aditya Krishnan 和 Shay Sapir。《数据流中伪确定性计数的下界》。载于第 50 届自动机、语言与编程国际研讨会（ICALP 2023），莱布尼茨国际信息学会议论文集第 261 卷，第 30:1–30:14 页，德国达格施图尔，2023 年。施洛斯达格施图尔－莱布尼茨信息学中心。¨ur Informatik.[Cs¨u10] Mikl´os Cs¨ur¨os。《使用浮点计数器进行近似计数》。载于 COCOON，第 6196 卷，第 358–367 页。施普林格，2010 年。[Cve07] Andrej Cvetkovski。《一种使用有限内存资源进行近似计数的算法》。ACM SIGMETRICS 性能评估评论，35(1):181–190，2007 年。[DPV18] Peter Dixon、A Pavan 和 NV Vinodchandran。《伪确定性近似算法》。载于第 43 届数学基础计算机科学国际研讨会（MFCS 2018）。施洛斯达格施图尔－莱布尼茨信息学中心，2018 年。[DPV21] Peter Dixon、A Pavan 和 NV Vinodchandran。《多重伪确定性计算的完全问题》。理论计算机科学创新，2021 年。[DPWV22] Peter Dixon、A Pavan、Jason Vander Woude 和 NV Vinodchandran。《伪确定性：承诺与下界》。载于第 54 届 ACM SIGACT 计算理论研讨会论文集，第 1552–1565 页，2022 年。[Fla85] Philippe Flajolet。《近似计数：详细分析》。BIT 数值数学，25(1):113–134，1985 年。[GG11] Eran Gat 和 Shafi Goldwasser。《具有唯一答案的概率搜索算法及其密码学应用》。载于计算复杂性电子研讨会（ECCC），第 18 卷，第 136 页，2011 年。[GG17] Shafi Goldwasser 和 Ofer Grossman。《伪确定性 NC 中的二分图完美匹配》。载于第 44 届自动机、语言与编程国际研讨会（ICALP 2017）。施洛斯达格施图尔－莱布尼茨信息学中心，2017 年。[GG21] Sumanta Ghosh 和 Rohit Gurjar。《拟阵交：从搜索到加权判定的伪确定性并行归约》。载于近似、随机化与组合优化：算法与技术（APPROX/RANDOM 2021）。施洛斯达格施图尔－莱布尼茨信息学中心，2021 年。¨GGH[Shafi Goldwasser、Ofer Grossman 和 Dhiraj Holden。《伪确定性证明》。载于第 9 届理论计算机科学创新会议（ITCS 2018）。施洛斯达格施图尔－莱布尼茨信息学中心，2018 年。

[GGH19] Michel Goemans、Shafi Goldwasser 和 Dhiraj Holden。双重高效的伪确定性证明。arXiv 预印本 arXiv:1910.00994，2019年。[GGMW20] Shafi Goldwasser、Ofer Grossman、Sidhanth Mohanty 和 David P Woodruff。伪确定性数据流处理。载于第11届理论计算机科学创新会议（ITCS 2020）。施洛斯达格施图尔－莱布尼茨信息学中心，2020年。[GGR13] Oded Goldreich、Shafi Goldwasser 和 Dana Ron。论伪确定性算法的可能性与局限。载于第4届理论计算机科学创新会议论文集，第127–138页。ACM，2013年。[GIPS21] Shafi Goldwasser、Russell Impagliazzo、Toniann Pitassi 和 Rahul Santhanam。论NP搜索问题的伪确定性查询复杂度。载于第36届计算复杂性会议论文集，第1页，2021年。[GL19] Ofer Grossman 和 Yang P Liu。对数空间中的可复现性与伪确定性。载于第30届ACM-SIAM离散算法年会论文集，第606–620页。SIAM，2019年。[GM20] Ofer Grossman 和 Dana Moshkovitz。无缓慢降速的放大与去随机化。SIAM计算杂志，49(5):959–998，2020年。[Gol19] Oded Goldreich。多重伪确定性算法。载于计算复杂性电子研讨会（ECCC），2019年。[Gro15] Ofer Grossman。以伪确定性方式寻找本原根。载于计算复杂性电子研讨会（ECCC），第22卷，第207页，2015年。[GS09] Andr´e Gronemeier 和 Martin Sauerhoff。应用近似计数计算长数据流的频率矩。《计算系统理论》，44:332–348，2009年。[HILL99] Johan H˚astad、Russell Impagliazzo、Leonid A Levin 和 Michael Luby。由任意单向函数构造伪随机生成器。SIAM计算杂志，28(4):1364–1396，1999年。[Hol17] Dhiraj Holden。关于BPP搜索问题的无条件次指数时间伪确定性算法的注记。arXiv 预印本 arXiv:1707.05808，2017年。[IZ89] Russell Impagliazzo 和 David Zuckerman。如何回收随机比特。载于FOCS，第30卷，第248–253页，1989年。[JW19] Rajesh Jayaram 和 David P Woodruff。迈向流式与分布式模型中矩估计的最优界。载于《近似、随机化与组合优化：算法与技术》（APPROX/RANDOM 2019）。施洛斯达格施图尔－莱布尼茨信息学中心，2019年。[KNW10] Daniel M Kane、Jelani Nelson 和 David P Woodruff。论草图算法与流式处理中小范数的精确空间复杂度。载于第21届ACM-SIAM离散算法年会论文集，第1161–1178页。SIAM，2010年。[LOS21] Zhenjian Lu、Igor C Oliveira 和 Rahul Santhanam。伪确定性算法与概率时间的结构。载于第53届ACM SIGACT理论计算年会论文集，第303–316页，2021年。

[Mor78] Robert Morris。在小型寄存器中计数大量事件。《ACM 通讯》，21(10):840–842，1978年。[NY22] Jelani Nelson 和 Huacheng Yu。近似计数的最优界。载于第41届 ACM SIGMOD-SIGACT-SIGAI 数据库系统原理研讨会论文集，2022。[OS17] Igor C Oliveira 和 Rahul Santhanam。亚指数时间中的伪确定性构造。载于第49届 ACM SIGACT 年度计算理论研讨会论文集，第665–677页，2017年。[OS18] Igor C Oliveira 和 Rahul Santhanam。伪去随机化学习与近似。载于LIPIcs－莱布尼茨国际信息学论文集，第116卷。施洛斯达格施图尔－莱布尼茨信息学中心，2018年。[Tre01] Luca Trevisan。提取器与伪随机生成器。《ACM 汇刊》，48(4):860–879，2001年。[XKNS21] Jingyi Xu、Sehoon Kim、Borivoje Nikolic 和 Yakun Sophia Shao。采用近似计数算法的高效内存硬件性能计数器。载于2021年IEEE系统与软件性能分析国际研讨会（ISPASS），第226–228页。IEEE，2021年。