arXiv:2301.05605v1 [cs.DS] 13 Jan 2023

流式频率矩估计的差分隐私持续发布

Alessandro Epasto <aepasto@google.com>, Jieming Mao <maojm@google.com>, Andres Munoz Medina <ammedina@google.com>, VahabMirrokni <mirrokni@google.com>, Sergei Vassilvitskii <sergeiv@google.com>Peilin Zhong <peilinz@google.com>

2022年4月

摘要

流式模型是一种处理大规模数据的常用计算方法。在这种设定中，数据以数据流的形式到达，目标是在仅遍历数据流一遍且尽可能少占用空间的情况下，计算所需的量（通常为数据统计量）。

受数据隐私重要性的启发，我们在持续发布设置下开发了差分隐私流式算法；在该设置中，算法在每个时间戳的输出并集都必须满足差分隐私。具体而言，我们研究该设置下的基本 ℓ_p($p \in [0, +\infty)$) 频率矩估计问题，并给出一种 ε-差分隐私算法，该算法可实现 $(1+\eta)$-相对近似 ($\forall \eta \in (0, 1)$)，加性误差为多对数(Tn)，且使用多对数(Tn) $\max(1, n^{1-2/p})$ 空间，其中 T 是数据流的长度， n 是元素全集的大小。即使在非隐私设定下，我们的空间复杂度也在多对数因子范围内近乎最优。

为获得这些结果，我们首先将差分隐私持续发布模型下的若干基本操作（例如不同元素计数、重频项和低频元素计数）归约为相同设置下更简单的计数/求和问题。基于这些基本操作，我们开发了一种差分隐私持续发布层集估计方法，以解决 ℓ_p 频率矩估计问题。

我们还简单地将结果扩展到了更难的滑动窗口模型中，在该模型下，必须对过去的 W 个数据项持续维护统计量。

1 引言

数据隐私是部署现实世界计算系统时的核心关切。在关于计算隐私的浩瀚文献中，Hsu 等人 [2021], 提出的差分隐私（DP）概念 Dwork[2008], Dwork 等人 [2014] 十多年来一直是事实上的标准。差分隐私的经典定义假设数据是静态的 Dwork [2008], ，且数据管理者有兴趣在数据集上获取预先确定的数量的查询答案。

然而，现实世界的应用通常需要分析自然且快速增长的用户数据集。流行的流式模型计算 Sr. [1978], Alon 等人 [1996], 便展示了这一点：数据会随着时间推移不断到达，算法在每次更新时都会输出一个新解。在这类流式应用中，差分隐私的持续发布模型 Dwork 等人 [2010a], Hubert Chan 等人 [2010] 提供了严格的隐私保证：对于算法的所有输出，观察者在信息论意义上能够获知数据流中某个数据点是否存在的能力受到限制。

本文聚焦于隐私流式算法领域的两个基本挑战：仅插入模型（又称流式模型）和滑动窗口模型。在前一种模型中，数据管理者接收数据流 $a_1, a_2, \ldots$，并在每个时刻 t 发布一个依赖于截至该时刻所接收全部数据的统计查询。在后一种模型中，计算仅依赖于数据管理者最近观察到的 W 个数据项。

与流式模型相比，滑动窗口模型通常在实际应用中更具相关性，因为它能够考虑信息时效性；在某些情形下，它也可能是法律或隐私方面的要求。例如，在某些情况下，通用数据保护条例（GDPR）等数据隐私法[1] 不允许无限期保留用户数据。

我们的主要贡献是针对流式模型和滑动窗口模型下的一系列基础流式问题，提出隐私算法。

示例：Privacy Sandbox。我们展示了研究结果的一项具体实际应用。作为$Privacy\ Sandbox$ 计划的一部分，Chrome 开发了一系列 API，旨在减少跨站跟踪，同时支持数字广告生态系统。其中一项提案的关键组成部分是*k*-匿名服务器。[2] 该服务器确保每个报告给广告商的广告创意，在特定时间窗口内都至少赢得相应的拍卖 k 次。抽象掉具体细节后，这个问题要求计算滑动窗口中的不同元素数量。此外，为了进一步加强隐私保护，计算本身应具备差分隐私，这正是本文研究的情形。

前面的例子阐明了在持续发布设置下，研究具有差分隐私的不同元素计数问题的滑动窗口算法的具体动机。本文其余部分将对这一模型以及我们在该领域取得的结果进行形式化说明。

具体而言，我们研究一类比输入数据的不同元素计数问题更一般的统计量： ℓ_p 频率矩问题。ℓ_p 频率矩是各元素频率的 p次方之和。对于 $p = 0$，不同元素的数量是一个特例。 ℓ_p 频率矩问题是流式处理文献中最基础的问题之一。第一个非平凡的 $p = 1$ 算法是 Morris [1978]。此后，Flajolet 和 Martin [1985] 首次研究了 $p = 0$的情形。Alon 等人 [1996] 开创了对 $p = 2$ 及其他 $p \in [0, \infty$的研究）。经过数十年的发展，对于所有 $p \in [0, \infty$，当前最佳的 ℓ_p 频率矩估计算法所需空间已接近最优，也就是说，它们几乎达到了已证明的空间下界（例如，Ganguly [2011], Kane 等人 [2011,2010], Flajolet 等人 [2007], Li 和 Woodruff [2013]）。

然而，即使在非持续发布设定下，差分隐私流式频率矩的研究现状也不明朗。现有大多数工作只研究了 $p = 0$、1、2（更多讨论见第 1.4 节）。Wang 等人 [2022] 研究了一般的 $p \in (0, 1]$）。近期一项独立工作中，Blocki 等人 [2022] 研究了所有 $p \in [0, \infty$。但是，Blocki 等人 [2022], 和 Wang 等人 [2022]

[1] https://gdpr-info.eu/art-17-gdpr/ [2] https://github.com/WICG/turtledove/blob/main/FLEDGE_k_匿名性_server.md

研究了持续发布设置。在差分隐私持续发布设定下，Bolot 等人 [2013] 研究了不同元素计数，但并未在低空间流式处理场景下进行研究。在差分隐私流式持续发布设定下，现有工作 Dwork [2008], 和 Hubert Chan 等人 [2010] 仅研究了 $p = 1$ 这一情形。在差分隐私低空间流式持续发布设定下，尚无已知的 $p \neq 1$ 算法。

对于滑动窗口模型中的 ℓ_p 频率矩，已有一些已知技术 Datar 等人[2002], Braverman 和 Ostrovsky [2007] ，可以通过使用少量额外空间，将流式算法转换为滑动窗口算法。我们将展示如何扩展这些技术，从而将我们的差分隐私流式持续发布算法转换为差分隐私滑动窗口持续发布算法。

1.1 计算模型

本文考虑一种带有 T 时间戳的流式设定。在每个时间戳 $t \in [T]$，我们都会收到一个输入 $a_t \in \mathcal{U} \cup \{\perp\}$，其中 $\mathcal{U}$ 表示所有可能输入元素的全集，而 $\perp$表示空集。我们有时也会考虑整数输入流，即在每个时间戳$t \in [T]$，我们都会收到一个输入 $a_t \in \mathbb{Z}$。目标是根据这些输入计算某个函数 $g(\cdot)$。本文考虑持续发布模型，而非仅在数据流结束时计算 g ：

- **持续发布模型**：在每个时间戳 $t \in [T]$，我们希望根据数据 $a_1, a_2, \cdots, a_t$ 输出 $g(\cdot)$。

根据 g 所依据的输入范围，我们考虑两种不同的流式模型。

- **（仅插入）流式模型**：在此模型中， g 取决于所有过去的输入，也就是说，在时间戳 $t \in [T]$，我们希望计算 $g(a_1, ..., a_t)$。除非另有说明，本文中提到的流式模型均指仅插入流式模型。

- **滑动窗口模型**：在此模型中，我们有一个表示窗口大小的参数 $W \in \mathbb{Z}_{\geq 1}$ 。 g 取决于最近的 W 个输入，也就是说，在时间戳 $t \in [T]$，我们希望计算$g(a_{\max(t-W+1,1)} \cdots, a_t)$。

本文特别关注空间复杂度相对于 T 为次线性的算法。对于一个（子）流 $\mathcal{S}$，我们用 $\|\mathcal{S}\|_p^p$ 表示 $\mathcal{S}$ 的 ℓ_p 阶频率矩。特别地，对于 $p = 0$, $\|\mathcal{S}\|_0$ 表示不同元素的数量。对于 $p = 1$, $\|\mathcal{S}\|_1$ 表示非空元素的数量。详细的符号约定和定义请参阅预备知识部分（第 2 节）。

1.2 我们的结果与先前工作的比较

本节简要概述我们的结果，并与先前工作进行比较。我们用n 表示全集 $\mathcal{U}$ 的大小。我们用 (α, γ)-近似来指定近似保证，其中乘法因子为 α ，加性误差为 γ。在所有结果中，我们都使用 $\varepsilon \geq 0$ 作为差分隐私参数，并将相对误差设为 $\eta \in (0,0.5)$。

1.2.1 差分隐私流式持续发布算法

我们开发了一系列差分隐私算法，用于在流式持续发布模型中解决频率矩估计及其相关问题。

ℓ_p 频率矩估计($p \in [0,\infty)$)。我们的主要结果是一种通用的 ℓ_p 频率矩估计算法，适用于所有 $p \in [0,\infty)$。

定理 1.1 (ℓ_p 频率矩，定理 5.31 的非正式版本。存在一种 ε-差分隐私算法，在流式持续发布模型中，以至少0.9的概率，始终为每个时间戳 $\left(1+\eta,\ \left(\frac{\log(Tn}{\eta\varepsilon}\right)^{O(\max(1,p))}\right)$ 近似 $\|\mathcal{S}\|_p^p$，其中 t，$\mathcal{S}$ 表示截至时间戳 t的数据流。该算法使用的空间为$\max(1, n^{1-2/p}) \cdot \left(\frac{\log(Tn}{\eta\varepsilon}\right)^{O(\max(1,p))}$.)

据我们所知，我们是首个在差分隐私流式持续发布设置下研究一般 ℓ_p 频率矩估计问题的工作。Dwork 等人 [2010a]和 Hubert Chan 等人 [2010] 研究了同一设置下的求和问题，其中求和问题可以看作 $p = 1$的特例。Wang 等人 [2022] 研究了 $p \in$ 的流式ℓ_p 频率矩估计（$0, 1]$ 基于 p-稳定分布），与此同时，独立开展的 Blocki 等人 [2022] 研究了 $p \in$ 的情形（$0, 1],$，但尚不清楚如何将其技术推广到持续发布模型，即他们的方法只能为数据流结束时的输出提供差分隐私保证。此外，Wang 等人 [2022] 的方法无法在任意小的 $\varepsilon > 0$下实现 ε-DP，他们还提到，其技术可能不易扩展到 $p > 1$的情形。

即使与非隐私流式频率矩估计算法相比，我们的空间使用量在相差多对数因子的范围内仍接近最优：对于 $p \le 2$，我们的算法和先前的非隐私算法（例如 Kane 等人 [2010]）所需的空间都是多对数级的；对于 $p > 2$，我们的算法和先前的非隐私算法 Indyk 和 Woodruff [2005] 所需的空间都是 $\tilde{O}(n^{1-2/p})$[3]。请注意，即使在非隐私情形下，$\Omega(n^{1-2/p})$ 空间也是 $p > 2$ 的一个已证下界 Saks 和 Sun [2002], Bar-Yossef 等人 [2004]。

求和（ℓ_1 频率矩估计）。与 ℓ_p频率矩估计问题相关的最简单问题是求和问题：目标是计算输入数值的总和。请注意，ℓ_1 频率矩估计是对二进制数据流求和的一个特例，即我们将 $\perp$ 视为 0，并将所有其他元素视为 1。

定理 1.2 （非负流求和，定理 3.3 的非正式版本）。 在流式持续发布模型中，存在一种用于求和问题的 ε-差分隐私算法。若保证输入数值均为非负数，则以至少 0.9 的概率，对于任意时间戳 $t \in [T]$，输出始终是所有输入数值之和的（$1+\eta$，$O_{\varepsilon,\ \eta}$（对数 T））-近似。该算法使用 O（1）空间。

求和问题由 Dwork 等人 [2010a], Hubert Chan 等人 [2010] 在差分隐私流式持续发布模型中进行了研究。他们的近似具有 $O(\log^{2.5} T)$ 加性误差。在我们的工作中，我们证明，如果允许 ($1+\eta$) 相对误差，并且只处理包含非负数的数据流，就可以将加性误差降低到 $O(\log\ T)$。当我们

[3]我们用 $\tilde{O}(g)$ 表示 $g \cdot \text{poly}(\log(g))$

在流式模型中无法避免某些问题（例如不同元素数量）产生相对误差，但仍需要将求和作为子程序时，这一结果便很有用。

不同元素计数（ℓ_0 频率矩估计）。不同元素计数是流式处理文献中的一个基础问题。目标是估计数据流中出现的不同元素数量。我们提供了一种用于不同元素计数的差分隐私流式持续发布算法。

定理 1.3（不同元素数量，推论 4.11 的非正式版本）。存在一种针对流式持续发布模型中不同元素数量的 ε-差分隐私算法。以至少 0.9 的概率，对于每个时间戳$t \in [T]$，其输出始终是一个 $\left(1+\eta, O_{\varepsilon,\eta}\left(\log^2(T)\right)\right)$-近似值。该算法使用 $\mathrm{poly}\left(\frac{\log(T)}{\eta \min(\varepsilon,1)}\right)$ 空间。

在非隐私流式设定中，可以通过 Flajolet 和 Martin 的草图算法 [1985] 及其变体（例如 Flajolet 等人 [2007]）解决不同元素计数问题。一些近期工作 Choi 等人 [2020], Smith 等人 [2020] 将这些草图技术扩展到差分隐私流式处理设置，以实现不同元素计数。然而，目前尚不清楚如何将这些技术扩展到持续发布设置。Bolot 等人[2013]研究了不同元素计数的持续发布。不过，Bolot 等人 [2013] 的工作并非针对低空间流式处理场景。

频率和 ℓ_2 频率矩的估计。ℓ_2 频率矩估计的目标是估计各元素频率的平方和。我们提出一种差分隐私流式持续发布 CountSketch Charikar 等人的算法，并用它来估计 ℓ_2 频率矩以及每个元素的频率。

定理 1.4 (频率和 ℓ_2 频率矩，定理 5.4 的非正式版本**)**。存在一个 ε-差分隐私算法，在流式持续发布模型中，以至少0.9的概率，对每个时间戳 $t \in [T]$ 始终输出：

1. $\hat{f}_a$ 对于所有满足 $a \in \mathcal{U}$ 的 $|f_a - \hat{f}_a| \leq \eta \|\mathcal{S}\|_2 + \tilde{O}_{\varepsilon,\eta}\left(\log^{3.5}(Tn)\right)$，其中 $\mathcal{S}$ 表示截至时间戳 t 的数据流，而 f_a 表示 $\mathcal{S}$ 中 a 的频率，
2. $\hat{F}_2$ *such that* $|\hat{F}_2 - \|\mathcal{S}\|_2^2| \leq \eta \|\mathcal{S}\|_2^2 + \tilde{O}_{\varepsilon,\eta}\left(\log^7(Tn)\right)$

该算法使用 $O\left(\frac{\log(Tn)}{\eta^2} \cdot \log(T)\right)$ 空间。

尽管已有一系列工作研究了差分隐私 ℓ_2 频率矩（例如，Blocki 等人 [2012],Sheffet [2017], Bu 等人 [2021]），但这些工作都没有考虑流式持续发布设置，而且目前尚不清楚如何将以往的技术扩展到持续发布设置。

ℓ_p 重频项。在 ℓ_p 重频项问题中，给定参数 k，目标是找出其频率的 p 次方至少达到 ℓ_p 频率矩的 $1/k$ 比例的元素。通过扩展我们的差分隐私流式持续发布 CountSketch 算法，我们得到了一种差分隐私流式持续发布 ℓ_p 重频项算法。

定理 1.5（ℓ_p 适用于所有 的重频项 $p \in [0, \infty)$，定理 5.10 的非正式表述）。在流式持续发布模型中，存在一个ε-差分隐私算法，使得其以至少 0.9 的概率ˆ始终输出一个集合 $H \subseteq \mathcal{U}$ 和一个函数 f： $H \to \mathbb{R}$，对每个时间戳 $t \in [T]$ 均满足

1. $\forall a \in$ ˆH, $f(a) \in (1 \pm \eta) \cdot f_a$ 其中， f_a 是数据流 $\mathcal{S}$ 中截至时间戳的 a 的频率 t,

2. $\forall a \in \mathcal{U}$，如果 $f_a \geq \frac{1}{\varepsilon\eta} \cdot$ 多项式 $\left(\text{对数}\left(\frac{T \cdot k \cdot n}{\eta}\right)\right)$ 且 $f_a^p \geq \|\mathcal{S}\|_p^p / k$ 那么 $a \in H$，

3. H 的大小至多为 $O\left(\log(Tn) \cdot 2^p \cdot k\right)$。

该算法使用 $\max(1, n^{1-2/p}) \cdot \frac{k^3}{\eta^2} \cdot \text{poly}\left(\log\left(T \cdot k \cdot n\right)\right)$ 空间。

据我们所知，尽管 Chan 等人 [2012], ℓ_p 已研究过差分隐私流式持续发布 ℓ_1 重频项问题（针对 $p \neq 1$），但此前尚未在差分隐私流式持续发布设定下研究过 $p > 2$ 重频项问题。请注意，即使在非隐私设定下，Saks 和 Sun [2002], 、Bar-Yossef 等人[2004]给出的 $\Omega(n^{1-2/p})$ 也是 ℓ_p 重频项所需空间的下界。

1.2.2 差分隐私滑动窗口持续发布算法 s

Braverman 和 Ostrovsky 的平滑直方图 [2007] 是一种通用算法框架；如果我们要计算的目标函数具有某些良好性质，它便可以将相对近似流式算法转换为相对近似滑动窗口算法。

我们推广了平滑直方图，使其在目标函数具有良好性质时，能够将同时具有相对误差和加性误差的近似流式算法转换为同时具有相对误差和加性误差的近似滑动窗口算法。此外，我们证明，如果流式算法在持续发布设置下满足差分隐私，那么转换后的滑动窗口算法在持续发布设置下也满足差分隐私。

通过应用我们推广后的平滑直方图方法，并在加性误差和空间使用量方面付出比我们的差分隐私流式持续发布算法多 poly $\left(\frac{\log T}{\eta}\right)$ 倍的代价，我们给出了以下差分隐私滑动窗口持续发布算法：

1. ℓ_p 矩估计（见推论 6.10），

2. 求和（见推论 6.7），

3. 不同元素计数（见推论 6.8），4. ℓ_2 矩估计
（见推论 6.9）。

1.3 我们的技术

本节简要介绍我们算法的核心思路。我们提出了一系列技术，将我们在差分隐私流式持续发布设定中考虑的几乎所有问题，都归约为该设定下的求和问题。

通过分组实现更优的加性误差。为了说明在差分隐私流式持续发布算法中使用分组的直觉，我们从以下简单问题开始：给定一个数字数据流 $c_1, c_2, \cdots, c_T$，其中每个 c_i 至少为 $10 \cdot \ln(10 \cdot T)/\varepsilon$，目标是以至少 0.9 的概率，对每个前缀 t 输出 $\sum_{j=1}^{t} c_j$ 的一个 (1 ± 0.1) 近似，并且我们希望所有输出构成的集合满足 ε-DP，即持续发布结果满足 ε-DP。解决上述问题的一种简单方法是发布一组带噪声的数字流 $\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_T$，其中 $\forall t \in [T], \hat{c}_t = c_t + \text{Lap}(1/\varepsilon)$，并针对每个前缀 $t \in [T]$ 报告 $\sum_{j=1}^{t} \hat{c}_j$。不难看出，$(\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_T)$ 满足 ε-DP。由于报告的近似前缀和仅取决于 $(\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_T)$，因此持续发布结果满足 ε-DP。此外，以至少

0.9 的概率， $\forall t \in [T], |c_t - \hat{c}_t| \leq \ln(10 \cdot T)/\varepsilon$。由于 c_t 至少为 $10 \ln(10 \cdot N)/\varepsilon$，因此有 $0.9c_t \leq \hat{c}_t \leq 1.1c_t$，这意味着每个报告的近似前缀和都是一个 (1 ± 0.1) 近似。

为推广上述思路，我们提出一种分组方法，以差分隐私的方式将数据流中的连续数字分组，使每组的总计数足够大。为实现分组，我们需要迭代应用稀疏向量技术（例如，参见 Dwork 等人 [2014]）。类似的思路也出现在 Dwork 等人 [2015] 中，该方法在数据流稀疏时，对求和问题给出了比 Dwork 等人 [2010a] 更好的加性误差保证。相比之下，我们的加性误差保证始终优于 Dwork 等人 [2010a] 和 Dwork 等人 [2015]，同时允许额外的 $(1 + \varepsilon)$ 相对近似。

不同元素计数。我们将说明如何把不同元素计数问题归约为求和问题。假设元素全集较小，我们就能够跟踪数据流中已经出现过的元素集合。然后，我们可以创建一个由 $\{0, 1\}$构成的二进制数据流，其中 1 表示出现了一个新元素，0 表示输入元素已经出现过或为空。因此，在时间戳 t 处，二进制数据流的总和恰好等于不同元素的数量。此外，如果将输入流中的一个元素从 a 改为 b，二进制数据流中只有常数个位置会翻转：考虑这种改变 $a \to \perp \to b$。如果 a 不是它在输入流中首次出现，那么将 a 改为 $\perp$ 不会导致二进制数据流发生任何变化。如果 a 是它在输入流中首次出现，那么将 a 改为 $\perp$ 会使二进制数据流中相应的 1 变为 0，并使输入流中 a 原本第二次出现的位置所对应的 0 变为 1。因此，最多会影响二进制数据流中的 2 个条目。类似地，将 $\perp$ 改为 b 最多会导致二进制数据流中的 2 个条目发生变化。因此，二进制数据流具有低敏感度，这意味着差分隐私流式持续发布求和算法能够以较小的加性误差，对不同元素的数量给出良好近似。接下来，我们讨论如何处理较大的全集。对于较大的全集，我们可以尝试不同的采样率 $1/2$、$1/4$、$1/8 \cdots$、$1/T$。应当存在某个采样率，使得：(1) 如果将采样元素哈希到哈希桶中，以较高概率不会发生碰撞；(2) 采样数量远大于求和子程序导致的加性误差，因此我们能够对不同采样元素的数量作出良好的相对近似。随后，我们可以用不同采样元素的数量来估计输入流中不同元素的数量。

CountSketch 和 ℓ_p 重频项。令 $h\colon \mathcal{U} \to [k]$ 为一个哈希函数，它将元素均匀地哈希到 k 个哈希桶中。令 $g\colon \mathcal{U} \to \{-1, 1\}$ 以相同概率将每个元素随机映射为 -1 或 1。CountSketch 是由 k 个数构成的元组（z_1、z_2、$\cdots$、z_k），其中 z_i 是哈希到桶 i中的元素的加权频率之和，而元素a 的频率权重为 $g(a)$。在输入数据流中将 a 改为 b，最多会改变 2 个桶：桶 i 包含a，桶 j 包含 b。由于 $|g(a)| \leq 1$，z_i 和 z_j 至多会改变 1。我们可以使用 Dwork 等人的差分隐私流式持续发布求和算法 [2010以及 Hubert Chan 等人的], [2010]，来估计（z_1、z_2、$\cdots$、z_k），使得每个估计值 $\hat{}z_i$的 z_i仅有 poly(log T) 的加性误差，并且（$\hat{}z_1$、$\hat{}z_2$、$\cdots$、$\hat{}z_k$）在流式持续发布模型下满足差分隐私。假设 ℓ_2频率矩远大于 poly(log T)，那么加性误差就会转化为相对误差，我们可以使用 $\hat{}z_1$、$\hat{}z_2$、$\cdots$、$\hat{}z_k$来获得 ℓ_2 频率矩的良好相对近似。同样地，如果某个元素的频率远大于 poly(log(T))，那么 poly(log(T))相对于该频率就会成为较小的误差；根据 CountSketch 的标准分析，我们能够判断它是否为 ℓ_2 重频项。因此，我们可以使用这种差分隐私流式持续发布 CountSketch，以 $(1 + \eta)$ 相对误差和 poly(log T) 加性误差估计 ℓ_2 频率矩，还可以用这种 CountSketch 找出所有至少达到 poly(log T)

且属于 ℓ_2 重频项的元素。

注意，对于 $p \leq 2$，如果 a 的频率最大，且它是一个 $(1/k)$-ℓ_p 重频项，那么 a必定是一个 $(1/k)$-ℓ_2 重频项。对于 $p > 2$，如果 a 是一个 $1/k$-ℓ_p 重频项，那么 a 必定是一个$1/(kn^{1-2/p})$-ℓ_2 重频项。因此，通过某种哈希技术，我们可以使用 ℓ_2 重频项算法来构造 ℓ_p 重频项算法。但是，由于 ℓ_2 重频项算法只能报告频率大于 poly(log T) 的元素，因此得到的 ℓ_p重频项算法也只能报告频率大于 poly(log T) 的元素。

ℓ_p 频率矩估计。从较高层面来说，我们希望在差分隐私流式持续发布设定中模拟 Indyk 和 Woodruff [2005] 的层集估计思路。具体而言，令$\alpha = 1 + \eta$，令 f_a 表示 a 的频率，并令 $G_i = \{a \mid f_a \in (\alpha^i, \alpha^{i+1}]\}$。那么 $\sum_i |G_i| \cdot (\alpha^i)^p$是频率矩的良好近似。若 G_i 满足 $|G_i| \cdot (\alpha^i)^p$ 至少占 ℓ_p 矩的 $\Omega_\alpha(1/\log(T))$ 比例，我们就称其为贡献的。由于非贡献元素对 ℓ_p 频率矩的总贡献很小，因此不难看出，$\sum$贡献的 $_{G_i} |G_i| \cdot (\alpha^i)^p$仍然是 ℓ_p 频率矩的良好近似。由此，我们只需估计每个贡献的 G_i的大小。根据贡献的定义，不难看出，如果 G_i 是贡献的，那么 α^i 较大或 $|G_i|$ 较大。事实上，正如 Indyk 和 Woodruff [2005],所指出的，对于每个贡献的层集 G_i，都存在合适的采样概率，使得采样后从 G_i 中采样到至少 poly(log T) 个元素，并且从 G_i 采样到的所有元素，在从全集 $\mathcal{U}$采样到的所有元素中，都是至少 $1/$poly(log T)-ℓ_p 重频项。理想情况下，我们可以尝试不同的采样率 $1, 1/2, 1/4, 1/8, \cdots, 1/T$，并使用我们的 ℓ_p重频项算法报告重频项，再估计每个 i对应的 $|G_i|$。然而，对于满足$\alpha^i \ll$ poly(log T) 的 i，我们的差分隐私流式持续发布 ℓ_p 重频项算法不会报告来自 G_i的任何元素。因此，我们必须另寻方法来估计 $|G_i|$，而不能使用重频项算法。

与不同元素计数类似，我们先考虑全集大小较小的情形，这样我们就可以跟踪集合 $A_1, A_2, \cdots, A_k$，其中 $k =$ 为 poly(log T)，而 A_i 是所有频率恰好为 i 的元素组成的集合。我们可以构造数据流 $\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_k$，其数值取自 $\{-1, 0, 1\}$。在数据流处理过程中，当我们看到一个输入元素 a，且 a 属于集合 A_i 时，由于 a 的频率增加，我们就将 a 移到集合 A_{i+1}。与此同时，我们将 -1 追加到数据流 $\mathcal{S}_i$，将 1 追加到数据流 $\mathcal{S}_{i+1}$，并将 0 追加到 $\mathcal{S}_j$，其中 $j \neq i, i+1$。不难验证，$\mathcal{S}_l$ 的总和始终等于 $|A_l|$，也就是频率恰好为 l 的元素数量。此外，与不同元素计数的分析类似，如果改变输入流中的一个元素，则每个 $\mathcal{S}_l$ 至多会受到 4 个条目的影响。因此，$(\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_k)$ 的总敏感度至多为 $O(k)$。所以，我们可以使用差分隐私持续发布求和，以加性误差 poly(log T) 估计每个 $\mathcal{S}_l$ 的总和。由此，我们可以对每个 i 以 $\alpha^i \ll$ poly(log T) 估计 $|G_i|$，加性误差为 poly(log T)，且在近似 ℓ_p 频率矩时只会引入至多 poly(log T) 的加性误差。

现在回到全集大小较大的情形。在这种情形下，我们可以使用前文讨论过的、用于不同元素计数的类似哈希和子采样技术，来估计每个 $l \in [k]$ 的 $|A_l|$。

1.4 相关工作

Dwork 等人 [2010a] 和 Hubert Chan 等人 [2010] 开创了持续发布模型中差分隐私的研究，并提出了用于计算求和的二叉树机制。Bolot 等人 [2013] 和 Perrier 等人 [2019] 将他们的结果推广到衰减求和、无空间限制的不同元素计数，以及实值数据的求和。Song 等人

[2018], Fichtenberger 等人 [2021] 研究了差分隐私持续发布模型下的图问题。Jain 等人 [2012], Smith 和 Thakurta [2013], Agarwal 和 Singh [2017] 研究了差分隐私在线学习。Jain 等人 [2021] 首次证明了在差分隐私下，持续发布模型与批处理模型之间存在误差方面的多项式分离。Upadhyay [2019] 研究了差分隐私滑动窗口模型中的重频项。

针对 $p = 0$（不包括持续发布）的差分隐私频率矩估计已有大量研究：Mir 等人 [2011], Dwork 等人 [2010b], Blocki 等人 [2012], Sheffet [2017],Choi 等人 [2020], Smith 等人 [2020], Bu 等人 [2021]。Wang 等人 [2022] 研究了具有低空间复杂度的 $p \in (0, 1]$（不包括持续发布）的频率矩估计。近期一项并行开展的独立工作 Blocki 等人 [2022] 也研究了具有低空间复杂度的 $p \in [0, \infty$，但同样未涉及持续发布设置。Mir 等人 [2011], Dwork 等人 [2010b] 在低空间流式处理场景下研究了差分隐私 ℓ_1 重频项问题，但未涉及持续发布设置。Chan 等人 [2012] 在低空间持续发布流式处理场景下研究了差分隐私 ℓ_1 重频项问题。然而，目前尚不清楚如何将他们的技术扩展到 $p \neq 1$的 l_p情形。

ℓ_p 频率矩估计和 ℓ_p 重频项在非隐私流式处理文献中得到了广泛研究。对于 ℓ_p 频率矩估计，例如，Flajolet 和 Martin[1985], Flajolet 等人 [2007], Durand 和 Flajolet [2003] 可解决 $p = 0$，Alon 等人 [1996], Charikar 等人 [2002], Thorup 和 Zhang [2004] 可解决 $p = 2$，Kane 等人 [2010], Indyk [2006], Li [2008], Kane 等人[2011] 可解决 $p \in$ (0, 2)，而 Indyk 和 Woodruff [2005], Andoni 等人 [2011], Andoni [2017] 可解决 $p > 2$。对于 ℓ_p 重频项，例如，Cormode 和 Muthukrishnan [2005],Misra 和 Gries [1982] 可解决 $p = 1$，Charikar 等人 [2002] 可解决 $p = 2$，Jowhari 等人 [2011] 可解决$p \in$ (0, 2)，而 Indyk 和 Woodruff [2005], Andoni 等人 [2011] 可解决 $p > 2$。

2 预备知识

2.1 符号约定

本文中，对于 $n \geq 1$，我们用 $[n]$ 表示集合 $\{1, 2, \cdots, n\}$。如果没有歧义，对于$i \leq j \in \mathbb{Z}$，我们有时用 $[i, j]$ 表示整数集 $\{i, i + 1, \cdots, j\}$，而不是实数集 $\{a \in \mathbb{R} \mid i \leq a \leq j\}$。我们用 $f_a(a_1, a_2, \cdots, a_k)$ 表示a 在序列 $(a_1, a_2, \cdots, a_k)$ 中的频率，即 $f_a(a_1, a_2, \cdots, a_k) = |\{i \in [k] \mid a_i = a\}|$。如果上下文中序列 $(a_1, a_2, \cdots, a_k)$ 已明确，我们简写为 f_a 表示 a 的频率。我们用 $\mathbf{1}_{\mathcal{E}}$表示一个指示函数，即当条件 $\mathcal{E}$ 成立时为 $\mathbf{1}_{\mathcal{E}} = 1$，当条件 $\mathcal{E}$ 不成立时为 $\mathbf{1}_{\mathcal{E}} = 0$。

对于 $\alpha \geq 1$、$\gamma \geq 0$，若 $\frac{1}{\alpha} \cdot x - \gamma \leq y \leq \alpha \cdot x + \gamma$，则 y 是 x 的 (α, γ)-近似。若 y 是 x 的 $(1, \gamma)$-近似，则称 y 是 x 的加性误差为 γ 的近似。若 y 是 x 的 $(\alpha, 0)$-近似，则称 y 是 x 的 α- 近似。我们用 $a \pm b$ 表示实数区间 $[a - |b|, a + |b|]$。对于实数集合 S，我们用 $S \pm b$ 表示集合 $\bigcup_{a \in S} a \pm b$，并用 $S \cdot c$ 表示集合 $\bigcup_{a \in S} a \cdot c$。我们用 Lap($b$) 表示尺度参数为 b 的拉普拉斯分布，即 Lap(b) 的概率密度函数为 $\frac{1}{2b} \exp(|x|/b)$

2.2 待计算函数

我们研究流式处理文献中的几种基本函数。当输入为整数时，我们考虑在（子）流（$a_i, \cdots, a_j$）上求解求和问题:

- 数值求和：**求和**$(a_i, \cdots, a_j)$：$= \sum_{k=i}^{j} a_k$

当输入来自 $\mathcal{U} \cup \{\perp\}$ 时，我们考虑基于（子）流 $(a_i, \cdots, a_j)$ 中元素频率的函数 $g(a_i, \cdots, a_j)$。

- 非空元素数量： $\|(a_i, ..., a_j)\|_1 := \sum_{a \in \mathcal{U}} f_a(a_i, ..., a_j)$.
- 不同元素数量： $\|(a_i, ..., a_j)\|_0 := \sum_{a \in \mathcal{U}} \mathbf{1}_{f_a(a_i, ..., a_j) > 0}$.
- ℓ_p-频率矩： $\|(a_i, ..., a_j)\|_p^p := \sum_{a \in \mathcal{U}} f_a(a_i, ..., a_j)^p$.
- ℓ_p-重频项： $(1/k)$-ℓ_p-$\mathbf{HH}(a_i, ..., a_j) := \{a \in \mathcal{U} \mid f_a(a_i, \cdots, a_j)^p \geq \|(a_i, \cdots, a_j)\|_p^p / k\}$.

请注意， $\|(a_i, ..., a_j)\|_1$ 是带有二进制输入的**求和**$(a_i, \cdots, a_j)$的一个特殊情形。

2.3 差分隐私

相邻数据流： 考虑两个数据流 $\mathcal{S} = (a_1, a_2, \cdots, a_T)$ 和 $\mathcal{S}' = (a'_1, a'_2, \cdots, a'_T)$。如果至多存在一个时间戳 $t \in [T]$ 使得 (1). $|a_t - a'_t| \leq 1$ （仅当输入被视为整数时才需要）(2). $\forall i \neq t, a_i = a'_i$，那么我们称 $\mathcal{S}$ 和 $\mathcal{S}'$ 为相邻数据流。

定义 2.1 (差分隐私)。 我们称算法 A 满足 ε-差分隐私（*DP*），如果对于任意两个相邻数据流 $\mathcal{S}, \mathcal{S}'$ 以及任意输出集合 $\mathcal{O}$,

$$\Pr[A(\mathcal{S}) \in \mathcal{O}] \leq e^{\varepsilon} \cdot \Pr[A(\mathcal{S}') \in \mathcal{O}].$$

请注意，在持续发布模型中，定义 2.1 提到的输出 $A(\mathcal{S})$ 是算法 A 在数据流 $\mathcal{S}$ 的每个时间戳上的完整输出历史。

定义 2.2 (数据流之间的距离)。 考虑两个数据流 $\mathcal{S}$ 和 $\mathcal{S}'$。如果 d 是满足以下条件的最小整数：存在数据流序列 $\mathcal{S}_0, \mathcal{S}_1, \cdots, \mathcal{S}_d$，其中 $\mathcal{S}_0 = \mathcal{S}, \mathcal{S}_d = \mathcal{S}'$ 与 $\forall i \in [d]$, $\mathcal{S}_i$ 和 $\mathcal{S}_{i-1}$ 是相邻数据流，那么 $\mathcal{S}$ 和 $\mathcal{S}'$ 之间的距离为 $\mathrm{dist}(\mathcal{S}, \mathcal{S}') = d$。

定义 2.3 (数据流映射的敏感度)。 设 $\mathcal{F}$ 是一个映射，它将给定的输入流 $\mathcal{S}$映射为一个数据流元组 $(\mathcal{F}_1(\mathcal{S})\ \mathcal{F}_2(\mathcal{S})\ \cdots,\ \mathcal{F}_k(\mathcal{S}))$。 $\mathcal{F}$的敏感度是满足以下条件的最小值 s：对于任意两个相邻数据流 $\mathcal{S}$和 $\mathcal{S}'$, $\sum_{i \in [k]} \mathrm{dist}(\mathcal{F}_i(S)\ \mathcal{F}_i(S')) \leq s$。

定理 *2.4*（组合，*Dwork* 等人 [2014]）。设 $\mathcal{F}$ 是一个映射，它将给定的输入流 $\mathcal{S}$映射为一个数据流元组（$\mathcal{F}_1(\mathcal{S})$ $\mathcal{F}_2(\mathcal{S})$ $\cdots$，$\mathcal{F}_k(\mathcal{S})$）。设 A_1、A_2、$\cdots$、A_k是 k ε-差分隐私算法。设 A是一个算法，使得对于某个函数 $M(\cdot)$，$A(\mathcal{S}) = M(A_1(\mathcal{F}_1(\mathcal{S}))\ A_2(\mathcal{F}_2(\mathcal{S}))\ \cdots,\ A_k(\mathcal{F}_k(\mathcal{S})))$。那么算法 A 是 $(s\varepsilon)$-差分隐私。

定理 2.4 的示例用法： 我们论文中定理 2.4 的一些示例用法如下所示:

- **输入流上多个算法的组合：** 假设以下每个算法 A_1、A_2、$\cdots$、A_k均为 ε-差分隐私，则对于任意函数 $M(\cdot)$，$M(A_1(\mathcal{S})\ A_2(\mathcal{S})\ \cdots,\ A_k(\mathcal{S}))$ 是 $(k\varepsilon)$-差分隐私。
- 不相交子流上的算法组合：对于输入流 $\mathcal{S} = (a_1、a_2、\cdots、a_T)$，我们将 $\mathcal{S}$划分为 $\mathcal{S}_1 = (a_1、1、a_1、2、\cdots、a_1、_T)$ $\mathcal{S}_2 = (a_2、1、a_2、2、\cdots、a_2、_T)$ $\cdots$、$\mathcal{S}_k = (a_k、1、a_k、2、\cdots、a_k、_T)$，也就是说，$\forall t \in [T]$, 只有一个 $i \in [k]$ 使得 $a_{i,\ t} = a_t$，且 $\forall i' \neq i, a_{i,\ t} = \perp$（或 0）。显然，这种划分的敏感度为 1。因此，如果每个 $A_1, A_2, \cdots, A_k$ 都是一个 ε-差分隐私算法，那么对于任意函数 $M(\cdot)$, $M(A_1(\mathcal{S}_1)\ A_2(\mathcal{S}_2)\ \cdots, A_k(\mathcal{S}_k))$ 都是 ε-差分隐私。

2.4 流式持续发布求和与计数

对于流式持续发布模型中的求和问题，Dwork 等人 [2010a], Hubert Chan 等人 [2010]提出了二叉树机制。该机制实现了多对数级加性误差，且只需对数级空间。此外，它还可以处理数据流中的负数。

定理 2.5 (Dwork 等人 [2010a], Hubert Chan 等人 [2010])。令 $\varepsilon \geq 0, \xi \in (0, 0.5)$，存在一个$\varepsilon$-差分隐私算法，可在流式持续发布模型中进行求和。以 $1-\xi$ 的概率，对于每个时间戳 $t \in [T]$，输出的加性误差始终至多为 $O\left(\frac{1}{\varepsilon}\log^{2.5}(T)\log\left(\frac{1}{\xi}\right)\right)$。该算法使用 $O(\log(T))$ 空间。

2.5 概率工具

引理 2.6 (Bellare 和 Rompel [1994])。设 $\lambda \geq 4$ 为偶数。设 X 是由 $n\lambda$-路独立随机变量的和，这些随机变量的取值属于 $[0,1]$。设 $\mu = E[X]$ 和 $A > 0$。则有

$$\Pr\left[|X-\mu| > A\right] \leq 8 \cdot \left(\frac{\lambda\mu + \lambda^2}{A^2}\right)^{\lambda/2}$$

引理 *2.7*（通过中位数技巧提升成功概率）。假设 X 是一个值为v 的随机估计量，且 X 以至少 $2/3$ 的概率是 v的（α，γ）-近似。那么，对于 $\xi \in \in (0, 0.5)$，如果我们抽取 $k = \lceil 50\log(1/\xi)\rceil$ 个 X的独立副本 X_1、X_2、$\cdots$、X_k，则其中位数以至少 $1-\xi$的概率是 v的（α，γ）-近似。

证明。我们称 X_i 是良好的，当且仅当 X_i 是（α，γ）-近似于 v。如果中位数不是良好的，那么 $\sum_{i\in[k]} \mathbf{1}_{X_i \text{是良好的}} < \frac{k}{2}$。根据 Chernoff 界，$\Pr[\text{中位数不是良好的}] \leq \Pr[\sum_{i\in[k]} \mathbf{1}_{X_i \text{是良好的}} < \frac{k}{2}] \leq \Pr[E[\sum_{i\in[k]} \mathbf{1}_{X_i \text{是良好的}}] - \frac{k}{2} > \frac{k}{6}] \leq e^{-k/48} \leq \xi.$ □

3 Co误差更优的持续发布求和

本节将证明，如果允许相对近似，且输入流仅包含非负数，那么我们可以得到一种具有加性误差的持续发布求和算法

优于定理 2.5。

算法 1： 计数流分组。

输入：非负数流 $c_1, c_2, \cdots, c_T$、差分隐私参数 $\varepsilon > 0$、近似参数 $\eta \in (0, 0.5)$ 和失败概率 $\xi \in (0, 1)$。输出：由分组噪声计数组成的组流 $(\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_T)$ 令 $\varepsilon_0 \leftarrow \varepsilon/2$。初始化组索引 $i \leftarrow 1$、当前组 $G_1 \leftarrow \emptyset$ 和阈值$\tau_1 \leftarrow \left(\frac{1}{\eta} + 1\right) \cdot \frac{7}{\varepsilon_0} \cdot \ln(3 \cdot T/\gamma) + \text{Lap}(2/\varepsilon_0)$。//$G_i$ 仅用于分析。对于 $t = 1$ 到 T，执行$G_i \leftarrow G_i \cup \{t\}$。令 $\nu_t \leftarrow \text{Lap}(4/\varepsilon_0)$。如果 $\nu_t + \sum_{j \in G_i} c_j \geq \tau_i$，则$\hat{c}_t \leftarrow \text{Lap}(1/\varepsilon_0) + \sum_{j \in G_i} c_j$.$i \leftarrow i + 1$.$\tau_i \leftarrow \left(\frac{1}{\eta} + 1\right) \cdot \frac{7}{\varepsilon_0} \cdot \ln(3 \cdot T/\gamma) + \text{Lap}(2/\varepsilon_0)$。$G_i \leftarrow \emptyset$。否则$\hat{c}_t \leftarrow 0$。结束

引理 3.1。 算法 *1* 的输出流 $\hat{c}_1$，$\hat{c}_2$，$\cdots$，$\hat{c}_T$满足 ε-差分隐私。

证明思路是迭代应用稀疏向量技术。证明见附录 A.1。

引理 3.2。 设 $\hat{c}_1$，$\hat{c}_2$，$\cdots$，$\hat{c}_T$为算法 *1* 的输出流。那么，以至少$1 - \xi$的概率，$\forall l$，r 满足 $1 \leq l \leq r \leq T$,

$$(1-\eta)\sum_{j=l}^{r} c_j - \left(\frac{1}{\eta} + 4\right) \cdot \frac{7}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi) \leq \sum_{j=l}^{r} \hat{c}_j \leq (1+\eta)\sum_{j=l}^{r} c_j + \left(\frac{1}{\eta} + 4\right) \cdot \frac{7}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi).$$

引理 3.2 的证明见附录 A.2

定理 3.3 (非负流求和)。设 $\varepsilon \geq 0$，$\xi \in (0, 0.5)$，则存在一种用于流式持续发布模型中求和的 ε-差分隐私算法。若保证输入数值均为非负数，则以至少 $1 - \xi$ 的概率，对于任意时间戳 $t \in [T]$，输出始终是求和问题的 $\left(1 + \eta, O\left(\frac{\log(T/\xi)}{\varepsilon\eta}\right)\right)$-近似。该算法使用 $O(1)$ 空间。

证明。根据引理 3.1，算法 1 的输出流满足 ε-差分隐私，因此，我们只需对 $(\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_T)$ 求解非隐私求和问题。近似保证由引理 3.2 给出。注意，我们不需要存储 G_i，只需在任意时间戳维护 G_i 中数值的总和。因此，所需总空间为 $O(1)$。 □

4 持续发布的不同元素数量

在本节中，我们将展示如何利用 ε-差分隐私流式持续发布求和来解决 ε-差分隐私流式持续发布不同元素数量问题。在第 4.1 节中，我们将介绍如何估计

全集较小时的不同元素数量。在第 4.2 节中，我们通过子采样，将大全集中的不同元素数量问题归约为小全集中的不同元素数量问题。

4.1 小全集的不同元素数量

算法 2: 小全集的不同元素数量

输入： 一个元素数据流 $\mathcal{S}$，元素为 a_1、a_2、$\cdots$、$a_T \in \mathcal{U} \cup \{\perp\}$，并保证 $|\mathcal{U}| \leq m$。
参数： 相对近似因子 $\alpha \geq 1$ 和加性近似因子 $\gamma \geq 0$
取决于流式持续发布求和算法。
//参见定理 2.5。
输出： 在每个时间戳 t 估计不同元素数量。
初始化一个空数据流 $\mathcal{C}$。
设 $S \leftarrow \emptyset$。
对于数据流 $\mathcal{S}$ 中的每个 a_t **执行**
  如果 $a_t \notin S$ 且 $a_t \neq \perp$，**则**
    $S \leftarrow S \cup \{a_t\}$。
    在数据流 $\mathcal{C}$ 的末尾追加 1。
  end
  else
    在数据流 $\mathcal{C}$ 的末尾追加 0。
  end
  输出 $\mathcal{C}$ 总计数的 (α, γ)-近似。
end

引理 4.1. 在任意时刻 $t \in [T]$ 结束时，算法 *2* 的输出是不同元素数量的 (α, γ)-近似。

证明。 当且仅当我们看到一个新的非空元素时，才会在数据流 $\mathcal{C}$ 中追加 1，因此， $\mathcal{C}$ 中的总计数在任意时刻 $t \in [T]$ 结束时始终等于不同元素的数量。因此， $\mathcal{C}$ 总计数的 (α, γ)-近似就是不同元素数量的 (α, γ)-近似。

□

引理 4.2. 如果算法 *2* 中用于持续发布 $\mathcal{C}$ 近似总计数的算法满足 ε-差分隐私，那么算法 *2* 在持续发布模型中满足 5ε-差分隐私。

证明。 考虑两个相邻数据流 $\mathcal{S} = (a_1, a_2, \cdots, a_T)$ 和 $\mathcal{S}' = (a'_1, a'_2, \cdots, a'_T)$，它们是 $\mathcal{U}$ 中的元素流，且仅在时间戳 t 处不同，即 $a_t \neq a'_t$。考虑生成的计数流 $\mathcal{C} = (c_1, c_2, \cdots, c_T)$ 与 $\mathcal{C}' = (c'_1, c'_2, \cdots, c'_T)$ 之间的差异。

考虑任意时间戳 $i \in [T]$，如果 $a_i \neq a_t$ 且 $a_i \neq a'_t$，很容易验证 $c_i = c'_i$。假设$i \neq t$。如果 $a_i = a_t = u \in \mathcal{U}$，但 a_i 是 u 在 $\mathcal{S}$中第三次出现或更晚，那么 $a'_i = a_i = u$ 就是 u 在 $\mathcal{S}'$ 中第二次出现或更晚，这意味着 $c_i = c'_i = 0$。类似地，如果 $a_i = a'_t = u \in \mathcal{U}$是 u 在 $\mathcal{S}$中第三次或更晚出现，我们也可以证明 $c_i = c'_i = 0$。因此，数据流 $\mathcal{C}$的敏感度至多为 5：只有当 $i = t$或 a_i是 a_t的第一次或第二次出现时， $a_{t'}$、 c_i才可能与 c'_i不同。因此，如果我们使用一个 ε-差分隐私算法来持续发布$\mathcal{C}$的总计数，算法 2 持续发布的输出满足 $(5 \cdot \varepsilon)$-差分隐私。 □

定理 4.3。 令 $\varepsilon \geq 0, \xi \in (0, 0.5)$，假设存在一种 ε-差分隐私流式持续发布求和算法（用于非负数流），其空间复杂度为 J，并且以至少$1-\xi$的概率始终为每个时间戳输出一个(α, γ)-近似。对于全集大小至多为 m 的数据流，在流式持续发布模型下，存在一种计算不同元素数量的(5ε)-差分隐私算法。该算法以至少 $1-\xi$的概率始终为每个时间戳 $t \in [T]$输出一个(α, γ)-近似。该算法使用 $O(m + J)$ 空间。

证明。 考虑算法 2。近似保证由引理 4.1 证明。差分隐私保证由引理 4.2 证明。在证明的剩余部分，我们只需证明空间使用量。由于 $|\mathcal{U}| \leq m$，维护集合 S 所需的空间至多为 m。对 $\mathcal{C}$ 上的求和问题持续发布 (α, γ)-近似所需的空间至多为 $\mathcal{J}$。因此，所需的总空间至多为 $O(m + J)$。

□

将上述定理与定理 2.5 结合，可得到以下推论 y.

推论 4.4（小全集上的流式持续发布不同元素）。 对于全集大小至多为 m 的数据流，在流式持续发布模型中，存在一个 ε-差分隐私算法，用于计算不同元素数量。以至少 $1-\xi$ 的概率，对于每个时间戳 $t \in [T]$，输出的加性误差始终至多为 $O\left(\frac{1}{\varepsilon}\log^{2.5}(T)\log\left(\frac{1}{\xi}\right)\right)$。该算法使用 $O($ $m + \log(T)$ 空间。

结合定理 4.3 与定理 3.3，我们得到以下推论:

推论 4.5（小全集下流式持续发布不同元素，具有更优的加性误差）。 在持续发布模型中，对于全集大小至多为 m 的数据流，存在一个用于计算不同元素数量的 ε-差分隐私算法。以至少 $1-\xi$ 的概率，对于每个时间戳 $t \in [T]$，输出的加性误差始终是不同元素数量的一个 $\left(1+\eta, O\left(\frac{\log(T/\xi)}{\varepsilon\eta}\right)\right)$-近似。该算法使用 $O(m)$ 空间。

4.2 一般全集的不同元素数量

算法 3： 通过子采样计算不同元素数量

输入： 元素数据流 $\mathcal{S}$ a_1、a_2、$\cdots$、$a_T \in \mathcal{U} \cup \{\perp\}$，以及误差参数 $\eta \in (0, 0.5)$。

参数： 相对近似因子 $\alpha \geq 1$ 和加性近似因子 $\gamma \geq 0$
取决于用于计算不同元素数量的流式持续发布算法
适用于元素全集较小的数据流。 //参见定理 4.3。

输出： 不同元素数量的估计 $\|\mathcal{S}\|_0$。

$L \leftarrow \lceil \log \min(|\mathcal{U}|, T) \rceil, \lambda \leftarrow 2\log(1000L)$ $m \leftarrow 100L \cdot \left(16\alpha \text{最大值 } \left(\gamma/\eta, 32\alpha\lambda/\eta^2\right)\right)^2$.

令 $h : \mathcal{U} \to [m]$ 为一个两两独立哈希函数。
//这里我们将 $[m]$ 视为大小为 m 的元素全集，而不是整数集。

令 $g\colon \mathcal{U} \to [L] \cup \{\perp\}$ 为一个 λ-wise 独立哈希函数，并且
$\forall a \in \mathcal{U}, i \in [L]$，$\Pr[g(a) = i] = 2^{-i}$，$\Pr[g(a) = \perp] = 2^{-L}$。

初始化空数据流 $\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_L$。

对于数据流 $\mathcal{S}$ 中的每个 a_t **执行**

  对于 $i \in [L]$ **执行**

    如果 $a_t \neq \perp$ 且 $g(a_t) = i$ **则**
      将 $h(a_t)$追加到数据流 $\mathcal{S}_i$ 的末尾。
    end
    else
      将 $\perp$ 追加到数据流 $\mathcal{S}_i$的末尾。
    end

  end

  $\forall i \in [L]$，计算$\hat{s}_i$，使其为 $\|\mathcal{S}_i\|_0$的(α, γ)-近似。

  找出满足$\hat{s}_i \geq \max\left(\gamma/\eta, 32\alpha\lambda/\eta^2\right)$的最大 $i \in [L]$，并输出$\hat{s}_i \cdot 2^i$。

  如果不存在这样的 i，则输出 0。

end

引理 4.6。 考虑任意时间戳 $t \in [T]$。令 v 为算法 *3* 的输出。以至少 0.9 的概率，v 是 $\|(a_1, a_2, \cdots, a_t)\|_0$ 的 $((1 + O(\eta))\alpha, O(\alpha^2 \max(\gamma/\eta, \alpha \log(L)/\eta^2)))$-近似。

为证明引理 4.6，我们需要以下中间结论。考虑一个时间戳$t \in [T]$。令 $\mathcal{S}$表示时间戳 t处的输入流，即 $\mathcal{S} = (a_1, a_2, \cdots, a_t)$。令 $G_i = \{a_j \mid g(a_j) = i, j \leq t\}$，其中 $i \in [L]$。

断言 4.7。 $\forall i \in [L]$，如果 $\|\mathcal{S}\|_0 \geq 2^i \cdot 4\lambda/\eta^2$，$\Pr[|G_i| \in (1 \pm \eta) \cdot \|\mathcal{S}\|_0/2^i] \geq 1 - 0.01/L$。否则，$\Pr[||G_i| - \|\mathcal{S}\|_0/2^i| \leq 4\lambda/\eta] \geq 1 - 0.01/L$。

证明。 假设 $\|\mathcal{S}\|_0 \geq 2^i \cdot 4\lambda/\eta^2$。根据引理 2.6，我们有：

$$\begin{aligned}
&\Pr\left[\left||G_i| - \|\mathcal{S}\|_0/2^i\right| > \eta \cdot \|\mathcal{S}\|_0/2^i\right] \\
\leq &8 \cdot \left(\frac{\lambda \cdot \|\mathcal{S}\|_0/2^i + \lambda^2}{\left(\eta \cdot \|\mathcal{S}\|_0/2^i\right)^2}\right)^{\lambda/2} \\
\leq &0.01/L,
\end{aligned}$$

其中，最后一个不等式由 $\lambda = 2\log(1000L)$ 和 $\|\mathcal{S}\|_0 \geq 2^i \cdot 4\lambda/\eta^2$ 得出。

假设 $\|\mathcal{S}\|_0 \leq 2^i \cdot 4\lambda/\eta^2$，再次应用引理 2.6，我们得到：

$$\Pr\left[\left||G_i| - \|\mathcal{S}\|_0/2^i\right| > 4\lambda/\eta\right]$$

$$
\begin{aligned}
&\leq 8 \cdot \left(\frac{\lambda \cdot \|\mathcal{S}\|_0 / 2^i + \lambda^2}{(4\lambda/\eta)^2} \right)^{\lambda/2} \\
&\leq 0.01/L,
\end{aligned}
$$

其中，最后一个不等式由 $\|\mathcal{S}\|_0/2^i \leq 4\lambda/\eta^2$ 和 $\lambda = 2 \cdot \log(1000L)$ 得出。 □

断言 4.8. 对于 $i \in [L]$，在给定 $|G_i| \leq 16\alpha$最大 $\left(\gamma/\eta, 32\alpha\lambda/\eta^2\right)$ 的条件下，$|G_i| = \|\mathcal{S}_i\|_0$ 的概率至少为 $1 - 0.01/L$。

证明。 由于 h 两两独立， $\forall a, b \in G_i, \Pr[h(a) = h(b)] = 1/m$。由于 $m = 100L$ {16α最大 $\left(\gamma/\eta, 32\alpha\lambda/\eta^2\right)\big)^2$，根据并集界，$\Pr[\exists a \neq b \in G_i, h(a) = h(b)] \leq |G_i|^2/m \leq 0.01/L$。

□

令 $\mathcal{E}$ 表示以下两项均成立的事件：

1. $\forall i \in [L]$，其中 $2^i \cdot 4\lambda/\eta^2 \leq \|\mathcal{S}\|_0, |G_i| \in (1 \pm \eta) \cdot \|\mathcal{S}\|_0/2^i$。
2. $\forall i \in [L]$，其中 $2^i \cdot 4\lambda/\eta^2 > \|\mathcal{S}\|_0, |G_i| \in \|\mathcal{S}\|_0/2^i \pm 4\lambda/\eta$。

根据断言 4.7， $\mathcal{E}$ 发生的概率至少为 0.99。令 $\mathcal{E}'$ 为以下事件： $\forall i \in [L]$，且 $|G_i| \leq 16\alpha$max($\gamma/\eta, 32\alpha\lambda/\eta^2$)， $|G_i| = \|\mathcal{S}_i\|_0$。根据断言 4.8， $\mathcal{E}'$ 发生的概率至少为 0.99。

接下来，我们将证明引理 4.6。

引理 *4.6* 的证明。 在此证明中，我们以事件 $\mathcal{E}$ 和 $\mathcal{E}'$为条件。注意， $\mathcal{E}$ 和 $\mathcal{E}'$ 同时发生的概率至少为 0.98。

考虑 $\|\mathcal{S}\|_0 \geq 8\alpha \cdot$max $\left(\gamma/\eta, 32\alpha\lambda/\eta^2\right)$ 的情形。令 $i^* \in [L]$ 为满足 $\|\mathcal{S}\|_0/2^{i^*} \geq 4\alpha \cdot$max($\gamma/\eta$, $32\alpha\lambda/\eta^2$) 的最大值。根据事件 $\mathcal{E}$，我们有 $|G_{i^*}| \in (1 \pm \eta) \cdot \|\mathcal{S}\|_0/2^{i^*}$。由于我们对 i^*的选择，我们有 $\|\mathcal{S}\|_0/2^{i^*} \leq 8\alpha \cdot$max $\left(\gamma/\eta, 32\alpha\lambda/\eta^2\right)$。因此，$|G_{i^*}| \leq 16\alpha$max $\left(\gamma/\eta, 32\alpha\lambda/\eta^2\right)$。根据事件 $\mathcal{E}'$，我们有 $\|\mathcal{S}_{i^*}\|_0 = |G_{i^*}|$。因此，我们有$\hat{s}_{i^*} \geq \|\mathcal{S}_{i^*}\|_0/\alpha - \gamma \geq |G_{i^*}|/\alpha - \gamma$.。由于 $|G_{i^*}| \geq 2\alpha \cdot$max $\left(\gamma/\eta, 32\alpha\lambda/\eta^2\right)$，我们有$\hat{s}_{i^*} \geq$max $\left(\gamma/\eta, 32\alpha\lambda/\eta^2\right)$。因此，算法 3 将对某个 $i' \geq i^*$输出 $\hat{s}_{i'} \cdot 2^{i'}$。根据事件 $\mathcal{E}$，我们知道 $|G_{i'}| \leq$max($2 \cdot \|\mathcal{S}\|_0/2^{i^*}, \|\mathcal{S}\|_0/2^{i^*} + 4\lambda/\eta$) $\leq 16\alpha$max($\gamma/\eta, 32\alpha\lambda/\eta^2$)。根据事件 $\mathcal{E}'$，我们有 $\|\mathcal{S}_{i'}\|_0 = |G_{i'}|$。由于算法输出 $\hat{s}_{i'} \cdot 2^{i'}$，我们知道$\hat{s}_{i'} \geq$max($\gamma/\eta, 32\alpha\lambda/\eta^2$)，这意味着

$$
\begin{aligned}
|G_{i'}| &= \|\mathcal{S}_{i'}\|_0 \\
&\geq (\hat{s}_{i'} - \gamma)/\alpha \\
&\geq (1 - \eta)\hat{s}_{i'}/\alpha \\
&\geq 16\lambda/\eta^2.
\end{aligned}
$$

根据事件 $\mathcal{E}$，我们有 $|G_{i'}| \in (1 \pm \eta) \cdot \|\mathcal{S}\|_0/2^{i'}$。因此，我们有

$$
\begin{aligned}
\hat{s}_{i'} &\leq \alpha \|\mathcal{S}_{i'}\|_0 + \gamma \\
&= \alpha |G_{i'}| + \gamma \\
&\leq \frac{\alpha}{1 - \eta} \cdot |G_{i'}|
\end{aligned}
$$

$$\leq \frac{1+\eta}{1-\eta} \cdot \alpha \cdot \|\mathcal{S}\|_0/2^{i'}$$
$$\leq (1+4\eta)\alpha\|\mathcal{S}\|_0/2^{i'},$$

其中，第二个不等式由 $\hat{s}_{i'} \geq \gamma/\eta$ 得出，最后一个不等式由 $\eta \leq 0.5$ 得出。同样，我们有：

$$\begin{aligned}\hat{s}_{i'} &\geq \|\mathcal{S}_{i'}\|_0/\alpha - \gamma \\ &= |G_{i'}|/\alpha - \gamma \\ &\geq |G_{i'}|/((1+\eta)\alpha) \\ &\geq \frac{1-\eta}{1+\eta} \cdot \frac{1}{\alpha} \cdot \|\mathcal{S}\|_0/2^{i'} \\ &\geq (1-4\eta)/\alpha \cdot \|\mathcal{S}\|_0/2^{i'},\end{aligned}$$

其中，第二个不等式由 $\hat{s}_{i'} \geq \gamma/\eta$ 得出。

接下来，考虑 $\|\mathcal{S}\|_0 < 8\alpha \cdot \max(\gamma/\eta, 32\alpha\lambda/\eta^2)$ 的情形。算法 3 的输出要么是 0，要么是对某个 $i' \in [L]$ 输出 $\hat{s}_{i'} \cdot 2^{i'}$。假设其输出为 $\hat{s}_{i'} \cdot 2^{i'}$。我们有 $\hat{s}_{i'} \geq \max\left(\gamma/\eta, 32\alpha\lambda/\eta^2\right)$，这意味着

$$|G_{i'}| \geq \|\mathcal{S}_{i'}\|_0 \geq (\hat{s}_{i'} - \gamma)/\alpha \geq (1-\eta)\hat{s}_{i'}/\alpha \geq 16\lambda/\eta^2.$$

根据事件 $\mathcal{E}$，我们有 $|G_{i'}| \leq 2 \cdot \|\mathcal{S}\|_0/2^{i'}$。因此

$$\begin{aligned}\hat{s}_{i'} \cdot 2^{i'} &\leq (\alpha\|\mathcal{S}_{i'}\|_0 + \gamma) \cdot 2^{i'} \\ &\leq \frac{\alpha}{1-\eta} \cdot \|\mathcal{S}_{i'}\|_0 \cdot 2^{i'} \\ &\leq \frac{\alpha}{1-\eta} \cdot |G_{i'}| \cdot 2^{i'} \\ &\leq 4\alpha\|\mathcal{S}\|_0 \\ &\leq 32\alpha^2 \max(\gamma/\eta, 32\alpha\lambda/\eta^2).\end{aligned}$$

□

定理 4.9。 设 $\varepsilon \geq 0, \xi, \xi' \in (0, 0.5)$ $\eta \in (0, 0.5)$，假设存在一个 ε-差分隐私算法，用于在流式持续发布模型中处理全集大小至多为 $100\log(\min(|\mathcal{U}|, T)) \cdot (16\alpha \max(\gamma/\eta, 32\alpha \cdot 2\log(1000\log(\min(|\mathcal{U}|, T)))/\eta^2))^2$ 的数据流中的不同元素数量，该算法使用 J 空间，并且以至少 $1-\xi$ 的概率，对每个时间戳始终输出一个 (α, γ)-近似。则存在一个 $(\varepsilon' = \lceil 50\log(T/\xi')\rceil\varepsilon)$-差分隐私算法，用于在流式持续发布模型中处理全集为 $\mathcal{U}$ 的数据流中的不同元素数量。以至少 $1 - \xi' - \log(\min(|\mathcal{U}|, T)) \cdot \lceil 50\log(T/\xi')\rceil \cdot \xi$ 的概率，该算法对每个时间戳 $t \in [T]$ 始终输出一个 $((1+O(\eta))\alpha, O(\alpha^2 \max(\gamma/\eta, \alpha\log\log(\min(|\mathcal{U}|, T))/\eta^2)))$-近似。该算法使用 $O(J \cdot \log(\min(|\mathcal{U}|, T)) \cdot \log(T/\xi'))$ 空间。

证明。 根据我们对 $\mathcal{S}_1$、$\mathcal{S}_2$、$\cdots$、$\mathcal{S}_L$ 的构造以及敏感度的定义（定义 2.3），$(\mathcal{S}_1$、$\mathcal{S}_2$、$\cdots$、$\mathcal{S}_L)$ 的敏感度为 1。由于对每个 $i \in [L]$ 报告 $\hat{s}_i$ 的算法是

ε-差分隐私，算法 3 满足 ε-差分隐私。根据引理 4.6，对于任意 $t \in [T]$，在$\hat{s}_i$ 是 $\|\mathcal{S}\|_0$的一个 (α、 γ)-近似这一条件下，以至少 0.9 的概率，输出是一个 ((1+$O(\eta)$)α、 $O(\alpha^2 \max(\gamma/\eta$、 $\alpha \log\log(\min(|\mathcal{U}|$、 $T))/\eta$ ^{2})近似。为了将概率提高到 $1-\xi'$，使得 $\forall t \in [T]$ 近似保证始终成立，我们需要运行 $\lceil 50\log(T/\xi')\rceil$ 个算法 3 的独立副本，并在每个时间戳取输出的中位数。因此，整体算法满足 ($\lceil 50\log(T/\xi')\rceil \cdot \varepsilon$)-差分隐私。

接下来，考虑每个$\hat{s}_i$ 都是 $\|\mathcal{S}_i\|_0$的 (α, β)-近似的成功概率。对所有 i 以及算法 3 的所有独立副本应用并集界，成功概率至少为 $1 - \log(\min(|\mathcal{U}|, T)) \cdot \lceil 50\log(T/\xi')\rceil \cdot \xi$。

最后，考虑空间使用量。考察算法 3 的每个运行副本。哈希函数$h(\cdot)$占用 $O(1)$ 空间。哈希函数 $g(\cdot)$占用 $O(\lambda) = O(\log\log(\min(|\mathcal{U}|, T)))$ 空间。为了持续发布所有 i 的$\hat{s}_i$，我们需要使用 $O(J \cdot \log(\min(|\mathcal{U}|, T)))$ 空间。因此，所有副本所需的总空间至多为 $O(J \cdot \log(\min(|\mathcal{U}|, T)) \cdot \log(T/\xi'))$。

将推论 4.4 代入上述定理，并令 $\xi = \xi'/2\log(\min(|\mathcal{U}|,T))\cdot\lceil 50\log(2T/\xi')\rceil$、$\varepsilon = \lceil 50\varepsilon'\log(2T/\xi')\rceil$、$\alpha = 1$ $\gamma = O(\frac{1}{\varepsilon}$ T $1/\xi$ $J = O$ T $+100$ $|\mathcal{U}|$ T $\cdot$ 16α γ/η $32\alpha\cdot$ 、 $(\log^{2.5}(\,)\log(\,)$ 2 1000 T ^{2} og( )）以及（log( ) log(min(,)) (max(, log(log(min(|$\mathcal{U}$|,))))/η^2))），可得以下推论。

推论 4.10（流式持续发布不同元素）。对于 $\eta \in (0,0.5)$，存在一个ε-差分隐私算法，用于在元素全集为 $\mathcal{U}$ 的流式持续发布模型中计算数据流的不同元素数量。以至少 $1-\xi$ $1+$ 的概率，对于每 η T/ξ $^{2.5}$ T $1/\xi$ T/ξ $t \in [T]$ 个时间戳，输出始终是一个（, $O\left(\max\left(\frac{\log()\log()\log()\log\log()}{\eta\varepsilon}, \log\log_{\eta^2} T\right)\right)$）-近似。该算法使用$\mathrm{poly}\left(\log(T/\xi)\,\eta\,\min(\varepsilon,1)\right)$ 空间。

将推论 4.5 代入定理 4.9，其中 $\xi = \xi'/2\log(\min(|\mathcal{U}|,T))\cdot\lceil 50\log(2T/\xi')\rceil$, $\varepsilon = \lceil 50\varepsilon'\log($ $2T/\xi')\rceil$,$\alpha = 1+\eta, \gamma = O\left(\frac{\log(T/\xi)}{\varepsilon\eta}\right)$ 且

$$J = O(100\log(\min(|\mathcal{U}|, T)) \cdot (16\alpha\max(\gamma/\eta, 32\alpha \cdot 2\log(1000\log(\min(|\mathcal{U}|, T)))/\eta^2))^2),$$

我们得到以下推论。

推论 4.11 (流式持续发布不同元素，在 $\log(T)$ 上有更优的依赖关系)。对于$\eta \in (0,0.5)$，存在一种 ε-差分隐私算法，在流式持续发布模型中，对元素全集为 $\mathcal{U}$ 的数据流计算不同元素数量。以至少 $1-\xi$的概率，对于每个时间戳 $t \in [T]$，输出始终是一个$(1+O(\eta)$ $O\left(\log^2\left(\frac{T/\xi}{\eta^2\varepsilon}\right)\right)$)-近似。该算法使用$\mathrm{poly}\left(\log(T/\xi)\,\eta\,\min(\varepsilon,1)\right)$ 空间。

5 持续发布的重频项与频率矩估计

本节介绍用于 ℓ_p 重频项和频率矩估计的 ε-差分隐私流式持续发布算法。在第 5.1 节中，我们介绍持续发布模型下的 ε-差分隐私 CountSketch Charikar 等人[2002] 算法。CountSketch 用于 ℓ_2 重频项和ℓ_2 矩估计。在第 5.2 节中，我们说明如何利用 ℓ_2 重频项解决 ℓ_p 重频项问题。在第 5.3 节中，我们说明如何估计低频率元素的数量。在

第 5.4 节中，我们说明如何利用 ℓ_p 重频项和低频元素估计器来估计 ℓ_p 频率矩。

5.1 持续发布 CountSketch

算法 4： 持续发布 CountSketch

输入： 一个由元素 a_1、a_2、$\cdots$、$a_T \in \mathcal{U} \cup \{\perp\}$ 构成的数据流 $\mathcal{S}$，以及一个参数 $k \in \mathbb{Z}_{\geq 1}$。

参数： 相对近似因子 $\alpha \geq 1$ 和加性近似因子 $\gamma \geq 0$
取决于流式持续发布求和算法。
//参见定理 2.5。

输出： 在每个时间戳 t 处的一个元组 $(z_1, z_2, \cdots, z_k)$。

令 $h\colon \mathcal{U} \to [k]$ 为一个四重独立哈希函数，使得 $\forall a \in \mathcal{U}, i \in [k], \Pr[h(a) = i] = \frac{1}{k}$.

令 $g\colon \mathcal{U} \to \{-1, 1\}$ 为一个四重独立哈希函数，使得 $\forall a \in \mathcal{U}, \Pr[g(a) = 1] = \frac{1}{2}$。

初始化空数据流 $\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_k$。

对于数据流 $\mathcal{S}$ 中的每个 a_t **执行**

  如果 $a_t = \perp$，**则**

    在每个数据流 $\mathcal{S}_1$、$\mathcal{S}_2$、$\cdots$、$\mathcal{S}_k$ 的末尾追加 0。

  end

  else

    在数据流 $\mathcal{S}_{h(a_t)}$ 的末尾追加 $g(a_t)$，并在每个数据流 $\mathcal{S}_i$ 的末尾追加 0
    对于 $i \neq h(a_t)$。

  end

  输出一个元组（z_1、z_2、$\cdots$、z_k），其中 z_i 是 $\mathcal{S}_i$ 的总计数估计值，具有加性误差至多为 γ。

end

引理 5.1（差分隐私保证）。如果算法 *4* 中为每个 $i \in [k]$ 持续发布 $\mathcal{S}_i$ 近似总计数的子程序满足 ε-差分隐私，则算法 *4* 满足 2ε-差分隐私。

证明。考虑两个相邻数据流 $\mathcal{S}$ 和 $\mathcal{S}'$，它们的唯一区别是第 t 个元素，即 $a_t \neq a'_t$。考虑算法 4 中与之对应的数据流 $\mathcal{S}_1$、$\mathcal{S}_2$、$\cdots$、$\mathcal{S}_k$ 和 $\mathcal{S}'_1$、$\mathcal{S}'_2$、$\cdots$、$\mathcal{S}'_k$。对于 $i \in [k]$ 和 $j \neq t$，$\mathcal{S}_i$ 中的第 j 个数应与 $\mathcal{S}'_i$中的第 j 个数相同。因此，我们只需考虑每个 $i \in [k]$下 $\mathcal{S}_i$ 和 $\mathcal{S}'_i$ 中的第 t 个数。由于 a_t 只能使 $\mathcal{S}_{h(a_t)}$ 中的第 t 个数非零，a'_t 只能使 $\mathcal{S}'_{h(a'_t)}$ 中的第 t 个数非零，并且非零数只能是 ± 1，所以敏感度至多为 2。因此，如果我们使用 ε- 差分隐私算法持续发布 $\mathcal{S}_1$、$\mathcal{S}_2$、$\cdots$、$\mathcal{S}_k$ 的总计数，那么算法 4 持续发布的输出满足 2ε-差分隐私。

□

引理 5.2 (频繁元素的良好近似)。考虑任意 $a \in \mathcal{U}$ 和任意时间戳$t \in [T]$。令 f_a 表示 a 在 a_1，a_2，$\cdots$，a_t中的频率。令 $(z_1$，z_2，$\cdots$，$z_k)$ 表示算法 *4* 在时间戳 t处的输出。那么 $\forall \eta \in (0,\ 0。5)$，其概率至少为 $1 - 1/(k\eta^2)$，$|f_a - g(a) \cdot z_{h(a)}| \leq \eta \cdot \sqrt{\sum_{b \in \mathcal{U}} f_b^2} + \gamma$.

证明。令 $\hat{z}_{h(a)}$ 表示在时间戳 t 时数据流 $\mathcal{S}_{h(a)}$ 的真实总计数。根据 Charikar 等人 [2002], 提出的原始 CountSketch，我们有 $\Pr\left[|f_a - g(a) \cdot \hat{z}_{h(a)}| \leq \eta\sqrt{\sum_{b \in \mathcal{U}} f_b^2}\right] \geq 1 - 1/(k\eta^2)$。由于

$$|f_a - g(a) \cdot z_{h(a)}| \leq |f_a - g(a) \cdot \hat{z}_{h(a)}| + |g(a)| \cdot |z_{h(a)} - \hat{z}_{h(a)}| \leq |f_a - g(a) \cdot \hat{z}_{h(a)}| + \gamma,$$

我们以至少 $1-1/(k\eta^2)$ 的概率有 $|f_a - g(a)\cdot z_{h(a)}| \le \eta\cdot\sqrt{\sum_{b\in\mathcal{U}} f_b^2} + \gamma$。 □

引理 5.3 (ℓ_2 频率矩估计)。考虑任意时间戳 $t\in[T]$。对于 $a\in\mathcal{U}$，令 f_a 为 $a_1,a_2,\cdots,a_t$中 a 的频率。令 $(z_1,z_2,\cdots,z_k)$ 为算法 *4* 在时间戳 t $\forall\eta\in$ 0 0.5 $1-100/\ k\eta^2$ $|\sum_{i=1}^k z_i^2 - \sum_{a\in\mathcal{U}} f_a^2| \le$ 处的 $\eta\sum_{a\in\mathcal{U}} f_a^2 + 4k\gamma^2/\eta$ 输出。那么(,)，其成立的概率至少为(),

证明。令 $F_2 = \sum_{a\in\mathcal{U}} f_a^2$。令 $Z = \sum_{i\in[k]} z_i^2$。对于 $i\in[k]$，令 $\hat{}z_i$为数据流 $\mathcal{S}_i$在时间戳 t处的真实总计数。令 $\hat{}\ Z = \sum_{i\in[k]} \hat{}z_i^2$。根据 CountSketch 的分析，Charikar 等人 [2002],Thorup 和 Zhang [2004], 我们有 $\Pr\left[|F_2 - \hat{}Z| \le \eta/4\cdot F_2\right] \ge 1 - 100/(k\eta^2)$。接下来，我们以 $|F_2 - \hat{}\ Z| \le \eta/4\cdot F_2$为条件。

我们有

$$\begin{aligned}&|F_2 - Z|\\ \le&|F_2 - \hat{Z}| + |\hat{Z} - Z|\\ \le&\eta/4\cdot F_2 + \sum_{i=1}^k |z_i^2 - \hat{z}_i^2|.\end{aligned}$$

对 $i\in[k]$，记 $z_i = \hat{}z_i + v_i$。我们有 $|v_i| \le \gamma$。根据凸性，我们有:

$$(1-\eta/4)\hat{z}_i^2 - 4v_i^2/\eta \le (\hat{z}_i + v_i)^2 \le \hat{z}_i^2/(1-\eta/4) + 4v_i^2/\eta.$$

由于 $\eta\in(0, 0.5)$，$1/(1-\eta/4) \le (1+\eta/2)$。因此，$|z_i^2 - \hat{}z_i^2| \le \eta/2\cdot\hat{}z_i^2 + 4v_i^2/\eta$。我们有:

$$|F_2 - Z| \le \eta/4\cdot F_2 + \eta/2\cdot\hat{Z} + 4k\gamma^2/\eta \le \eta/4\cdot F_2 + \eta/2\cdot 3/2\cdot F_2 + 4k\gamma^2/\eta \le \eta F_2 + 4k\gamma^2/\eta.$$

□

定理 5.4 (流式持续发布 ℓ_2 频率估计器)。设 $\varepsilon>0$，$\eta\in(0,\ 0。5)$ $\xi\in(0,\ 0。5)$。存在一个在流式持续发布模型中的 ε-差分隐私算法，使得以至少 $1-\xi$ 的概率，对于每个时间戳 $t\in[T]$，它始终输出:

1. $\hat{}\ f_a$对每个满足 $|f_a - \hat{}f_a| \le \eta\|\mathcal{S}\|_2 + O\left(\frac{\log(T/\xi)+\log(|\mathcal{U}|)\varepsilon}{\cdot\log^{2.5}(T)\cdot\log\left(\frac{\log(T/\xi)+\log(|\mathcal{U}|)}{\xi\eta}\right)}\right)$的 $a\in\mathcal{U}$，其中 $\mathcal{S}$ 表示数据流 $(a_1,\ a_2,\ \cdots,\ a_t)$，$f_a$ 表示 $\mathcal{S}$ 中 a 的频率，

2. $\hat{F}_2$ *such that* $|\hat{F}_2 - \|\mathcal{S}\|_2^2| \le \eta\|\mathcal{S}\|_2^2 + O\left(\frac{(\log(T/\xi)+\log(|\mathcal{U}|))^2}{\varepsilon^2\eta^3}\cdot\log^5(T)\cdot\log^2\left(\frac{\log(T/\xi)+\log(|\mathcal{U}|)}{\xi\eta}\right)\right)$ 该算法使用 $O\left(\frac{\log(T/\xi)+\log(|\mathcal{U}|)}{\eta^2}\cdot\log(T)\right)$ 空间。

证明。假设我们设置 $k = 400/\eta^2$。根据引理 5.2 和引理 5.3，对于每个特定的时间戳 $t\in[T]$ 和 $a\in\mathcal{U}$，近似保证以至少 $2/3$ 的概率成立。为了将近似保证的成功概率提升至 $1-\xi/2$，并使其同时对所有 $t\in[T]$ 和所有 $a\in|\mathcal{U}|$ 成立，根据引理 2.7，我们运行 $\lceil 50(\log(2T/\xi) + \log(|\mathcal{U}|))\rceil$ 个算法 4 的副本，并取每个估计器的中位数。

对于每个 $i\in[k]$的 $\mathcal{S}_i$ 求和问题，我们应用定理 2.5。由于我们运行了 $\lceil 50(\log(2T/\xi)+\log(|\mathcal{U}|))\rceil$ 个算法 4 的副本，如果最终需要一个 ε-差分隐私算法，则每个求和

子程序都需要满足 $\varepsilon/(2\cdot\lceil 50(\log(2T/\xi)+\log(|\mathcal{U}|))\rceil)$-差分隐私，这依据引理 5.1。为了同时确保每次运行求和子程序的调用都以至少 $1-\xi/2$ 的概率成功，我们需要对所有求和调用应用并集界，因此每次运行求和子程序的成功概率应至少为 $1-\xi/(2\cdot\lceil 50(\log(2T/\xi)+\log(|\mathcal{U}|))\rceil\cdot k) = 1-\xi/(2\cdot\lceil 50(\log(2T/\xi)+\log(|\mathcal{U}|))\rceil\cdot 400/\eta^2)$。因此，根据定理 2.5，我们有 $\alpha=1$ 和 $\gamma = O\left(\frac{\log(T/\xi)+\log(|\mathcal{U}|)}{\varepsilon}\cdot\log^{2.5}(T)\cdot\log\left(\frac{\log(T/\xi)+|\mathcal{U}|}{\xi\eta^2}\right)\right)^{\log()}$。

最后，我们来考察总空间使用量。由于我们调用了 $(\lceil 50(\log(2T/\xi)+\log(|\mathcal{U}|))\rceil\cdot 400/\eta^2)$ 次求和子程序，执行这些子程序所需的空间为 $O\left(\frac{\log(T/\xi)+\log(|\mathcal{U}|)\cdot\eta^2}{}\cdot\log(T)\right)$。还需要 $O(1)$ 的额外空间。因此，所需的总空间为 $O\left(\frac{\log(T/\xi)+\log(|\mathcal{U}|)}{\eta^2}\cdot\log(T)\right)$ □

5.2 持续发布的 ℓ_p 重频项

通过应用 CountSketch，我们能够构造 ℓ_p 重频项。

算法 5: 持续发布的 ℓ_p 重频项 ($p\in[0,\infty)$)

输入： 一个由元素 a_1、a_2、$\cdots$、$a_T\in\mathcal{U}\cup\{\perp\}$组成的数据流 $\mathcal{S}$，一个参数 $k\in\mathbb{Z}_{\geq 1}$，一个误差参数 $\eta\in(0,0.5)$。

参数： 取决于流式持续发布 CountSketch 算法。 //参见定理 5.4。

输出： 一组元素及其估计频率 $H\subseteq\mathcal{U}$ 时间戳 t。$\hat{f}\colon H\to\mathbb{R}_{\geq 0}$ 在每个

设 $\phi\geq\max\left(|\mathcal{U}|^{1-2/p},1\right)$。设 $m=10k^2$。

设 $h:\mathcal{U}\to[m]$ 为一个两两独立哈希函数，其中 $\forall a\in\mathcal{U},i\in[m],\Pr[h(a)=i]=1/m$。

初始化空数据流 $\mathcal{S}_1,\mathcal{S}_2,\cdots,\mathcal{S}_m$。**对于** 每个 a_t 在数据流 $\mathcal{S}$ **执行**

  如果 $a_t=\perp$ **那么**

    将 $\perp$ 追加到每个数据流 $\mathcal{S}_1,\mathcal{S}_2,\cdots,\mathcal{S}_m$的末尾。

  end

  else

    将 a_t 追加到数据流 $\mathcal{S}_{h(a_t)}$ 的末尾，并将 $\perp$ 追加到每个数据流 $\mathcal{S}_i$ 的末尾，以便 $i\neq h(a_t)$。

  end

  对于 $i\in[m]$，计算 $\hat{}F_{2,i}$，它是 $\|\mathcal{S}_i\|_2^2$的 $(1.1,\gamma_1)$-近似。

  对于 $a\in\mathcal{U}$，计算 $\hat{}f_a$，它是 f_a的 $(1,(\eta/16)/(10\sqrt{\phi k})\cdot\|\mathcal{S}_{h(a)}\|_2+\gamma_2)$-近似，$a$ 在 $\mathcal{S}$ 中的频率（或等价地，在 $\mathcal{S}_{h(a)}$ 中）。

  对于 $a\in\mathcal{U}$，如果 $\hat{}f_a^2\geq\frac{\hat{F}_{2,h(a)}+\gamma_1}{25\phi k}+\frac{512\gamma_2^2}{\eta^2}$，将 a 添加到 $\hat{}H$中。

  设 $H\subseteq\hat{}H$ 仅保留满足 $\hat{}f_a$ 是其中前-$\left(\left(\frac{1+\eta}{1-\eta}\right)^p\cdot k\right)$ 个值之一的元素 a

  $\{\hat{}f_b\mid b\in\hat{}H\}$。对于每个 $a\in H$，报告 $\hat{}f(a)\leftarrow\hat{}f_a$。

end

引理 5.5 (差分隐私保证)。 如果 $\forall i\in[m]$ 算法 *5* 中用于持续发布$\hat{}$ $\hat{}F_{2,i}$ 且对于所有满足 $h(a)=i$ 的 a，该子程序都是 ε-差分隐私，那么算法 *5* 就是 2ε-差分隐私。

证明。 考虑两个相邻数据流 $\mathcal{S}$ 和 $\mathcal{S}'$，它们唯一的差异是第 t 个元素，即 $a_t\neq a_t'$。如果 $a_t\neq\perp$，则 $\mathcal{S}_{h(a_t)}$ 与$\mathcal{S}'_{h(a_t)}$之间至多只有一个元素不同。类似地，如果 $a_t'\neq\perp$，则 $\mathcal{S}_{h(a_t')}$ 与

$\hat{\mathcal{S}}'_{h(a'_t)}$之间至多只有一个元素不同。因此，如果对于每个 $i \in [m]$，对每个满足 $h(a) = i$ 的 $a \in \mathcal{U}$ 释放 $F_{2,i}$ 和 f_a 的持续发布算法都是 ε-差分隐私，则整体算法是 2ε-差分隐私。 □

引理 5.6. 在任意时间戳 $t \in [T]$, $\forall a \in \mathcal{U}$处，若 $a \in \hat{H}$，$(1-\eta)f_a^2 \le \hat{f}^2(a) \le (1+\eta)f_a^2$其中 f_a是 a在 $a_1, a_2, \cdots, a_t$中的频率。

证明。由于 $a \in H$，我们有：

$$\hat{f}_a^2 \ge \frac{\hat{F}_{2,h(a)} + \gamma_1}{25\phi k} + \frac{512\gamma_2^2}{\eta^2}.$$

因此：

$$\begin{aligned}
\frac{\eta}{16} \cdot \hat{f}_a^2 &\ge \frac{16}{\eta} \cdot \left(\frac{2 \cdot (\eta/16)^2}{100\phi k} \cdot \left(2\hat{F}_{2,h(a)} + 2\gamma_1 \right) + 2\gamma_2^2 \right) \\
&\ge \frac{16}{\eta} \cdot \left(\frac{2 \cdot (\eta/16)^2}{100\phi k} \cdot \|\mathcal{S}_{h(a)}\|_2^2 + 2\gamma_2^2 \right) \\
&= \frac{16}{\eta} \cdot \left(2 \cdot \left(\frac{\eta/16}{10\sqrt{\phi k}} \cdot \|\mathcal{S}_{h(a)}\|_2 \right)^2 + 2\gamma_2^2 \right) \\
&\ge \frac{16}{\eta} \cdot \left(\frac{\eta/16}{10\sqrt{\phi k}} \cdot \|\mathcal{S}_{h(a)}\|_2 + \gamma_2 \right)^2
\end{aligned}$$

由凸性可得：

$$\hat{f}_a^2 \ge (1-\eta/16) \cdot f_a^2 - \frac{16}{\eta} \cdot \left(\frac{\eta/16}{10\sqrt{\phi k}} \cdot \|\mathcal{S}_{h(a)}\|_2 + \gamma_2 \right)^2 \ge (1-\eta/16) \cdot f_a^2 - \eta/16 \cdot \hat{f}_a^2$$

and

$$\hat{f}_a^2 \le 1/(1-\eta/16) \cdot f_a^2 + \frac{16}{\eta} \cdot \left(\frac{\eta/16}{10\sqrt{\phi k}} \cdot \|\mathcal{S}_{h(a)}\|_2 + \gamma_2 \right)^2 \le 1/(1-\eta/16) \cdot f_a^2 + \eta/16 \cdot \hat{f}_a^2.$$

由于 $\eta \in (0, 0.5)$，我们有 $(1-\eta)f_a^2 \le \hat{f}_a^2 \le (1+\eta)f_a^2$。 □

引理 5.7. 在任意时间戳 t处，算法 $\mathcal{S}$ 的输出 H大小至多为 $\left(\frac{1+\eta}{1-\eta}\right)^p \cdot k$。

证明。请注意，H只保留来自 $\hat{H}$ 的前 $\left(\left(\frac{1+\eta}{1-\eta}\right)^p \cdot k\right)$ 个值。 □

引理 5.8。 在任意时间戳 t，考虑任意 $a \in \mathcal{U}$。令 f_a 为 $a_1, a_2, \cdots, a_t$ 中 a 的频率。如果 $f_a \ge 4\sqrt{\gamma_1/(\phi k)} + 512\gamma_2^2/\eta^2$ 且 $f_a^p \ge \|\mathcal{S}\|_p^p/k$，则其概率至少为 0.9，$a \in \hat{H}$。

证明。在此证明中，我们考虑时间戳 t 处的所有数据流和变量。假设 $f_a^p \ge \|\mathcal{S}\|_p^p/k$。令 $B = \{b \in \mathcal{U} \mid f_b^p \ge \|\mathcal{S}\|_p^p/k\}$。则以至少 0.9 的概率，$\forall b \in B \setminus \{a\}, h(a) \ne h(b)$。在证明的其余部分，我们以 $\forall b \in B \setminus \{a\}, h(a) \ne h(b)$ 为条件。

情况 1, $p \le 2$**:** 在这种情形下，我们有 $\phi = 1$。$\forall x \in \mathcal{U}$ 结合 $h(x) = h(a)$，我们有 $f_x \le f_a$，这意味着 $f_x^{p-2} \ge f_a^{p-2}$，因为 $p \le 2$。由于 $\forall x \in \mathcal{U}$，$\frac{f_x^p}{f_x^2} \le 1$，我们有

$$\frac{\|\mathcal{S}_{h(a)}\|_p^p}{\|\mathcal{S}_{h(a)}\|_2^2} = \frac{\sum_{x \in \mathcal{U}: h(x) = h(a)} f_x^p}{\sum_{x \in \mathcal{U}: h(x) = h(a)} f_x^2} \ge \min_{x \in \mathcal{U}: h(x) = h(a)} \frac{f_x^p}{f_x^2} = \frac{f_a^p}{f_a^2},$$

其中 i 这意味着 $f_a^2/\|\mathcal{S}_{h(a)}\|_2^2 \geq f_a^p/\|\mathcal{S}_{h(a)}\|_p^p \geq f_a^p/\|\mathcal{S}\|_p^p \geq 1/k$，因此 $f_a^2 \geq \|\mathcal{S}_{h(a)}\|_2^2/(\phi k)$.

情况 2， $p > 2$**：** 在这种情形下，我们有 $\phi = |\mathcal{U}|^{1-2/p}$。由于 $f_a^p \geq \|\mathcal{S}\|_p^p/k$，我们有：

$$f_a \geq \|\mathcal{S}\|_p/k^{1/p} \geq \|\mathcal{S}\|_p/k^{1/2} \geq \|\mathcal{S}\|_2/(k^{1/2} \cdot |\mathcal{U}|^{1/2-1/p}),$$

其中，第二个不等式由 $k^{1/2} \geq k^{1/p}$ for $p > 2$ 得出，第三个不等式由霍尔德不等式 $\|\mathcal{S}\|_2 \leq |\mathcal{U}|^{1/2-1/p} \cdot \|\mathcal{S}\|_p$ 得出。因此， $f_a^2 \geq \|\mathcal{S}\|_2^2/(\phi k)$。

因此，在上述两种情形下，我们始终有 $f_a^2 \geq \|\mathcal{S}\|_2^2/(\phi k)$。

根据凸性，我们有

$$\begin{aligned}\hat{f}_a^2 &\geq (1-\eta/16) \cdot f_a^2 - \frac{16}{\eta} \cdot \left(\frac{\eta/16}{10\sqrt{\phi k}} \cdot \|\mathcal{S}_{h(a)}\|_2 + \gamma_2\right)^2 \\ &\geq (1-\eta/16) \cdot f_a^2 - \frac{16}{\eta} \cdot \left(\frac{2(\eta/16)^2}{100\phi k} \cdot \|\mathcal{S}_{h(a)}\|_2^2 + 2\gamma_2^2\right)\end{aligned}$$

因此我们有：

$$\begin{aligned}\hat{f}_a^2 &\geq \frac{1}{2} \cdot \|\mathcal{S}_{h(a)}\|_2^2/(\phi k) - \frac{16}{\eta} \cdot \left(\frac{2(\eta/16)^2}{100\phi k} \cdot \|\mathcal{S}_{h(a)}\|_2^2 + 2\gamma_2^2\right) \\ &\geq \frac{1}{2} \cdot \frac{1}{2} \cdot (\hat{F}_{2,h(a)} - \gamma_1)/(\phi k) - \frac{16}{\eta} \cdot \left(\frac{2(\eta/16)^2}{100\phi k} \cdot 2 \cdot (\hat{F}_{2,h(a)} + \gamma_1) + 2\gamma_2^2\right) \\ &= \left(\frac{1}{4} - \frac{\eta/16}{25}\right) \cdot \frac{\hat{F}_{2,h(a)}}{\phi k} - \left(\frac{1}{4} + \frac{\eta/16}{25}\right) \cdot \frac{\gamma_1}{\phi k} - \frac{32}{\eta} \cdot \gamma_2^2 \\ &\geq \frac{1}{5} \cdot \frac{\hat{F}_{2,h(a)}}{\phi k} - \frac{1}{3} \cdot \frac{\gamma_1}{\phi k} - \frac{32}{\eta^2} \cdot \gamma_2^2 \\ &\geq 2 \cdot \left(\frac{\hat{F}_{2,h(a)} + \gamma_1}{25\phi k} + \frac{512\gamma_2^2}{\eta^2}\right) - \left(\frac{2048\gamma_2^2}{\eta^2} + \frac{2}{3} \cdot \frac{\gamma_1}{\phi k} - \frac{3}{25\phi k} \cdot \hat{F}_{2,h(a)}\right) \qquad (1)\end{aligned}$$

另一方面，根据凸性，我们有：

$$\begin{aligned}\hat{f}_a^2 &\geq (1-\eta/16) f_a^2 - \frac{16}{\eta} \cdot \left(\frac{\eta/16}{10\sqrt{\phi k}} \|\mathcal{S}_{h(a)}\|_2 + \gamma_2\right)^2 \\ &\geq (1-\eta/16) f_a^2 - \frac{16}{\eta} \cdot \left(\frac{2(\eta/16)^2}{100\phi k} \|\mathcal{S}_{h(a)}\|_2^2 + 2\gamma_2^2\right)\end{aligned}$$

因此

$$\begin{aligned}\hat{f}_a^2 &\geq \frac{1}{2} \cdot 16 \cdot (\gamma_1/(\phi k) + 512\gamma_2^2/\eta^2) - \frac{16}{\eta} \cdot \left(\frac{2(\eta/16)^2}{100\phi k} \cdot 2 \cdot (\hat{F}_{2,h(a)} + \gamma_1) + 2\gamma_2^2\right) \\ &= \left(8 - \frac{\eta/16}{25k}\right) \cdot \frac{\gamma_1}{\phi k} + \left(\frac{4096}{\eta^2} - \frac{32}{\eta}\right) \gamma_2^2 - \frac{\eta/16}{25\phi k} \cdot \hat{F}_{2,h(a)} \\ &\geq \frac{2}{3} \cdot \frac{\gamma_1}{\phi k} + \frac{2048}{\eta^2} \cdot \gamma_2^2 - \frac{3}{25\phi k} \cdot \hat{F}_{2,h(a)} \qquad (2)\end{aligned}$$

根据l o 结合方程（1） + 方程（2），可得 ˆ $f_a^2 \geq \frac{\hat{F}_{2,h(a)}+\gamma_1}{25\phi k} + \frac{512\gamma_2^2}{\eta^2}$ 因此， $a \in \hat{H}$. □

引理 5.9。对于任意时间戳 t，如果 $f_a^p \geq \|\mathcal{S}\|_p^p/k$ 并且 $a \in \hat{} H$，那么 $a \in H$。

证明。我们通过反证法证明该结论。假设 $a \notin H$，存在一个子集 $Q \subseteq H$ with$|Q| \geq \left(\frac{1+\eta}{1-\eta}\right)^p \cdot k$，使得 $\forall b \in Q, \hat{} f_b \geq \hat{} f_a$。根据引理 5.6，我们有 $\forall b \in Q, f_b^p \geq \left(\frac{1-\eta}{1+\eta}\right)^p \cdot f_a^p \geq \left(\frac{1-\eta}{1+\eta}\right)^p \cdot \|\mathcal{S}\|_p^p/k$。于是得到 $\sum_{b\in Q} f_b^p \geq \|\mathcal{S}\|_p^p$，这与假设矛盾。 □

定理 5.10 (ℓ_p所有 $p \in [0, \infty$ 重频项)。设 $\varepsilon > 0$，$\eta \in (0,\ 0。5)$ $k \geq 1$，$\xi \in (0,\ 0。5)$。设 $\phi = \max(1,\ |\mathcal{U}|^{1-2/p})$。在流式持续发布模型中，存在一个 ε-差分隐私算法 $\hat{}$，使得以至少 $1-\xi$ 的概率，它始终输出一个集合 $H \subseteq \mathcal{U}$ 和一个函数 $f:\ H \to \mathbb{R}$，对每个时间戳 $t \in [T]$: 满足

$\hat{}$1. $\forall a \in H$, $f(a) \in (1 \pm \eta) \cdot f_a$ 其中 f_a 是数据流 $\mathcal{S} = (a_1, a_2, \cdots, a_t)$ 中 a 的频率，*2.* $\forall a \in \mathcal{U}$，如果 $f_a \geq \frac{1}{\varepsilon\eta} \cdot log^C\left(\frac{T\cdot k\cdot|\mathcal{U}|}{\xi\eta}\right)$，其中某个常数 $C > 0$ 足够大，且 $f_a^p \geq \|\mathcal{S}\|_p^p/k$，那么 $a \in H$，*3.* H 的大小至多为 $O\left((log(T/\xi) + log(|\mathcal{U}|)) \cdot \left(\frac{1+\eta}{1-\eta}\right)^p \cdot k\right)$。

该算法使用 $\frac{\phi k^3}{\eta^2}$ · 多项式 $\left(\text{对数}\ \left(\frac{T\cdot k\cdot|\mathcal{U}|}{\xi}\right)\right)$ 空间。

证明。第一条性质由引理 5.6 得出。

根据引理 5.8 和引理 5.9，对于任意特定的 $t \in [T]$ 和 $a \in \mathcal{U}$，该保证以 0.9 的概率成立。为了提高概率，使该保证以至少 $1-\xi/2$ 的概率同时对所有 $t \in [T]$ 和 $a \in \mathcal{U}$ 成立，我们需要重复运行算法 5$\lceil 50(\log(2T/\xi) + \log(|\mathcal{U}|))\rceil$ 次，并令时间戳 t 处的最终 H 为时间戳 $\hat{}\ \hat{}$ t处所有输出 H 的并集；令时间戳 t处的最终 $f(a)$为时间戳 t处的任意输出 $f(a)$。根据引理 5.7，单次运行算法 5 所得到的输出 H至多为 $\left(\frac{1+\eta}{1-\eta}\right)^p \cdot k$。因此，最终输出 H的大小至多为 $O\left((\log(T/\xi) + \log(|\mathcal{U}|)) \cdot \left(\frac{1+\eta}{1-\eta}\right)^p \cdot k\right)$，这就证明了第二条性质。

我们对每个$i \in [m]$的频率和 ℓ_2 频率矩估计器应用定理 5.4。由于我们运行了 $\lceil 50(\log(2T/\xi) + \log(|\mathcal{U}|))\rceil$ 个算法 5 的副本，如果我们最终希望得到一个 ε-差分隐私算法，根据引理 5.5，每个频率和 ℓ_2 频率矩子程序都需要满足 $\varepsilon/(4\lceil 50(\log(2T/\xi) + \log(|\mathcal{U}|))\rceil)$-差分隐私。为使频率和 ℓ_2 频率矩子程序在每次运行时均以至少$1-\xi/2$的概率成功，我们需要对所有子程序调用应用并集界，因此子程序的每次运行都应以至少 $1-\xi/(4\cdot\lceil 50(\log(T/\xi) + \log(|\mathcal{U}|))\rceil \cdot m) = 1-\xi/(4\cdot\lceil 50(\log(T/\xi) + \log(|\mathcal{U}|))\rceil \cdot 10k^2)$的概率成功。请注意，我们还需要将定理 5.4 中的 η 重新缩放为算法 5 中用于估计每个元素频率的$(\eta/16)/(10\sqrt{\phi k})$，并将定理 5.4 中的 $\eta = 0.01$设置为用于估计 ℓ_2 频率矩的值。因此，根据定理 5.4，我们有

$$\gamma_1 = \frac{1}{\varepsilon^2} \cdot \text{poly}\left(\log\left(\frac{T\cdot k\cdot|\mathcal{U}|}{\xi}\right)\right)$$

and

$$\gamma_2 = \frac{1}{\varepsilon} \cdot \mathrm{poly}\left(\log\left(\frac{T \cdot k \cdot |\mathcal{U}|}{\xi\eta}\right)\right)$$

因此，第二个性质可由引理 5.8、引理 5.9 以及我们的概率放大论证推出。

最后，我们来考察空间使用量。由于我们运行了 $O(\log(T/\xi) + \log(|\mathcal{U}|))$ 个算法 5 的副本，且每个副本都调用 $O(k^2)$ 个频率估计器和 ℓ_2 个频率矩估计器，并使用上述讨论的参数。根据定理 5.4，总空间使用量至多为 $\frac{\phi k^3}{\eta^2} \cdot \mathrm{poly}\left(\log\left(\frac{T \cdot k \cdot |\mathcal{U}|}{\xi}\right)\right)$。 □

5.3 差分隐私持续发布低频元素计数

在本节中，我们给出一种差分隐私持续发布算法，用于统计频率为某个（较低）值的元素数量。与我们的不同元素计数算法类似，我们首先考虑元素全集较小的情形。

5.3.1 小全集中的低频元素数量

算法 6： 小全集中的低频元素数量

输入： 一个由元素 $a_1, a_2, \cdots, a_T \in \mathcal{U} \cup \{\perp\}$ 构成的数据流 $\mathcal{S}$，保证 $|\mathcal{U}| \le m$，以及一个目标频率 k。

参数： 相对近似因子 $\alpha \ge 1$ 和加性近似因子 $\gamma \ge 0$ 取决于流式持续发布求和算法。 //参见定理 2.5。

输出： 估计每个 $i \in [k]$ 在每个时间戳上频率恰为 i 的元素数量 时间戳 t。

初始化空数据流 $\mathcal{C}_1$、$\mathcal{C}_2$、$\cdots$、$\mathcal{C}_k$。

对于每个 $a \in \mathcal{U}$，初始化频率 $f(a) \leftarrow 0$

对于每个 a_t 在数据流 $\mathcal{S}$ 中**执行**

  如果 $a_t \neq \perp$, $f(a_t) \leftarrow f(a_t) + 1$。

  对于 每个 $i \in [k]$ **执行**

    如果 $a_t \neq \perp$ 且 $f(a_t) = i + 1$ **则**

      将 -1 追加到 $\mathcal{C}_i$末尾。

    end

    否则，如果 $a_t \neq \perp$ 且 $f(a_t) = i$ **则**

      在 $\mathcal{C}_i$ 末尾追加 1。

    end

    else

      在 $\mathcal{C}_i$ 末尾追加 0。

    end

  end

  对于每个 $i \in [k]$，输出 $\hat{s}_i$，它是 $\mathcal{C}_i$ 的总计数的 (α, γ)-近似。

end

引理 5.11. 在任意时刻 $t \in [T]$, $\forall i \in [k]$ 结束时，算法 6 的输出 $\hat{s}_i$ 是 $|\{a \in \mathcal{U} \mid f_a = i\}|$，即频率准确的元素集合的大小的 (α, γ)-近似i。

证明。容易看出，算法 6 维护 $f(a)$，使得对于任意 $a \in \mathcal{U}$，$f(a) = f_a$。对于 $i \in [k]$，如果添加 a_t 之前，a_t 的频率为 i，我们就将 -1 追加到 $\mathcal{C}_i$。如果添加 a_t 之后，a_t 的频率变为 i，我们就将 1 追加到 $\mathcal{C}_i$。注意，如果 a_t 的频率取值不属于上述两种情形，就不会影响频率为 i的元素数量。因此，$\mathcal{C}_i$ 的总计数始终等于频率为 i的元素数量。由此，$\hat{s}_i$ 是对频率为 i的元素数量的一个 (α, γ)-近似。 □

引理 5.12。如果持续发布 $\mathcal{C}_i$ 的近似总计数的算法对每个 $i \in [k]$ 都满足 ε-差分隐私，那么算法 6 在持续发布模型中满足 $(8k\varepsilon)$-差分隐私。

证明。考虑两个相邻数据流 $\mathcal{S} = (a_1, a_2, \cdots, a_T)$ 和 $\mathcal{S}' = (a'_1, a'_2, \cdots, a'_T)$，它们由 $\mathcal{U}$ 中的元素构成，且仅在时间戳 t 处不同，即 $a_t \neq a'_t$。固定任意 $i \in [k]$，考察相应计数流 $\mathcal{C}_i = (c_1, c_2, \cdots, c_T)$ 和 $\mathcal{C}'_i = (c'_1, c'_2, \cdots, c'_T)$ 之间的差异。

考虑一个中间相邻数据流 $\mathcal{S}'' = (a_1, a_2, \cdots, a_{t-1}, \perp, a_{t+1}, \cdots, a_T)$。令 $\mathcal{C}''_i = (c''_1, c''_2, \cdots, c''_T)$ 为与 $\mathcal{S}''$ 对应的计数流。$\mathcal{C}_i$ 与 $\mathcal{C}'_i$ 之间的总差异不超过 $\mathcal{C}_i$ 与 $\mathcal{C}''$ 之间的总差异和 $\mathcal{C}'_i$ 与 $\mathcal{C}''_i$ 之间的总差异之和。

考虑 $\mathcal{C}_i$ 与 $\mathcal{C}''_i$之间的差异。若 $a_t = \perp$, $\mathcal{C}''_i$ 与 $\mathcal{C}_i$完全相同。假设$a_t = a \in \mathcal{U}$ 是 a 在 $\mathcal{S}$中第 j 次出现。若 $j > i$，将 a_t 改为 $\perp$ 不会影响 $\mathcal{C}_i$。假设$j \leq i$。令 $a_{t_1}, a_{t_2}, a_{t_3}$ 分别为 a 在 $\mathcal{S}$中第 i 次、第 $(i+1)$ 次和第 $(i+2)$ 次出现。然后很容易验证 $c_{t_1} = 1, c''_{t_1} = 0, c_{t_2} = -1, c''_{t_2} = 1, c_{t_3} = 0, c''_{t_3} = -1$，并且对于任何其他$p \neq t_1, t_2, t_3$，都有 $c_p = c''_p$。因此，$\mathcal{C}_i$与 $\mathcal{C}''_i$之间的总差异至多为 4。

同理，$\mathcal{C}'_i$ 与 $\mathcal{C}''_i$之间的总差异至多为 4。因此，$\mathcal{C}_i$与 $\mathcal{C}'_i$之间的总差异至多为 8。因此，$\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_k$的总敏感度至多为 $8k$。如果我们使用一个 ε-差分隐私算法，持续发布每个 $i \in [k]$对应的 $\mathcal{C}_i$ 总计数，那么算法 6 持续发布的输出满足 $(8k\varepsilon)$-差分隐私。 □

定理 *5.13*（小全集中低频元素的流式持续发布计数）。令 $k \geq 1, \varepsilon \geq 0, \xi \in (0, 0.5)$。假设全集 $\mathcal{U}$ 的大小至多为 m。在流式持续发布模型中，存在一个 ε- 差分隐私算法，使得以至少 $1 - \xi$ 的概率，对于每个时间戳 t，它始终输出 k 个数$\hat{s}_1$、$\hat{s}_2$、$\cdots$、$\hat{s}_k$，满足 $\forall i \in [k]$: $\hat{s}_i$ 是$|\{a \in \mathcal{U} \mid f_a = i\}|$ 的近似，其加性误差为 $O\left(\frac{k}{\varepsilon} \cdot \log^{2.5}(T) \log\left(\frac{k}{\xi}\right)\right)$。该算法使用 $O(m + k \log(T))$ 的空间。

证明。考虑算法 6，我们使用定理 2.5 作为求和子程序。根据引理 5.12，如果希望最终算法满足 ε-差分隐私，则每个子程序必须满足 $(\varepsilon/(8k))$-差分隐私。此外，如果希望总体成功概率为 $1 - \xi$，则每个求和子程序的成功概率应至少为 $1 - \xi/k$。应用定理 2.5，可得 $\alpha = 1$ 和 $\gamma = O\left(\frac{k}{\varepsilon} \cdot \log^{2.5}(T) \log\left(\frac{k}{\xi}\right)\right)$。

我们需要 $O(m)$ 的空间来维护元素的频率。运行 k 求和子程序需要 $O(k \log(T))$ 的空间。因此，总空间使用量为 $O(m + k \log(T))$。 □

5.3.2 一般全集的低频元素数量

算法 7： 通过子采样计算低频元素数量

输入： 一个包含元素 $\mathcal{S}$ a_1、a_2、$\cdots$、$a_T \in \mathcal{U} \cup \{\perp\}$的数据流，一个误差参数 $\eta \in (0, 0.5)$，以及一个参数 $k \geq 1$。

参数： 加性近似因子 γ_1，取决于不同元素数量的流式持续发布，不同元素数量、相对近似因子 $\alpha_2 \geq 1$ 和加性近似因子 $\gamma_2 \geq 0$ 取决于以下流式持续发布计数：小全集中的低频元素。

//参见推论 4.10 和定理 5.13。

输出： 估计频率恰好为 i 的元素数量，并在每个 $i \in [k]$ 的每个时间戳 t。

$L \leftarrow \lceil$对数最小值$(|\mathcal{U}|, T)\rceil$，$\lambda \leftarrow 2$对数$(1000k)$ $m \leftarrow 100 \cdot (25600\lambda/\eta^2)^2$。

令 $h : \mathcal{U} \to [m]$ 为一个两两独立哈希函数。

令 $g\colon \mathcal{U} \to [L] \cup \{\perp\}$ 为一个 λ-路独立哈希函数，并且 $\forall a \in \mathcal{U}, i \in [L], \Pr[g(a) = i] = 2^{-i}, \Pr[g(a) = \perp] = 2^{-L}$.

初始化空数据流 $\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_L$。

对于数据流 $\mathcal{S}$ 中的每个 a_t **执行**

  对于 $i \in L$ **执行**

    如果 $a_t \neq \perp$ 且 $g(a_t) = i$ **则**

      将 $h(a_t)$追加到数据流 $\mathcal{S}_i$的末尾。

    end

    else

      将 $\perp$ 追加到数据流 $\mathcal{S}_i$的末尾。

    end

  end

  计算 $\hat{}d$，它是对 $\mathcal{S}$中不同元素数量的 $(1.1, \gamma_1)$-近似。

  对于 $i \in [L]$，计算 $\hat{}s_{i,\,1}$、$\hat{}s_{i,\,2}$、$\cdots$、$\hat{}s_{i,\,k}$，其中 $\hat{}s_{i,\,j}$是对不同元素数量的 (α_2, γ_2)-近似数据流 $\mathcal{S}_i$ 中的元素，其中每个元素的频率恰好为 j。

  如果 $\hat{}d \leq \max(3\gamma_1, 64\lambda/\eta^2)$，则设置$\hat{s}_1 = \hat{s}_2 = \cdots = \hat{s}_k = 0$。

  否则，找到满足 $2^{i^*} \cdot (64\lambda/\eta^2) \leq \hat{}d$ 的最大 $i^* \in [L]$，并且 $\forall j \in [k]$，设置$\hat{s}_j = \hat{}s_{i^*,\,j} \cdot 2^i$。

  输出$\hat{s}_1$、$\hat{s}_2$、$\cdots$、$\hat{s}_k$。

end

$\hat{}$引理 5.14。 如果算法 7 中持续发布 d 的子程序是 ε-差分隐私，并且持续发布 $\{\hat{}s_{i,\,1}, \hat{}s_{i,\,2}, \cdots, \hat{}s_{i,\,k}\}$ 的子程序对于每个 $i \in [k]$ 也都是 ε-差分隐私，则算法 7 是 3ε-差分隐私。

证明。考虑两个相邻数据流 $\mathcal{S}$ 和 $\mathcal{S}'$，它们唯一的区别在于第 t 个元素，即 $a_t \neq a'_t$。考虑它们在算法 7 中对应的数据流 $\mathcal{S}_1$、$\mathcal{S}_2$、$\cdots$、$\mathcal{S}_L$ 和 $\mathcal{S}'_1$、$\mathcal{S}'_2$、$\cdots$、$\mathcal{S}'_L$。对于 $i \in [L]$ 和 $j \neq t$，$\mathcal{S}_i$ 中第 j 个元素应与 $\mathcal{S}'_i$ 中第 j 个元素相同。由于 a_t 只能使 $\mathcal{S}_{h(a_t)}$ 中的第 t 个元素与 $\mathcal{S}'_{h(a_t)}$ 中的第 t 个元素不同，而 a'_t 只能使 $\mathcal{S}_{h(a'_t)}$ 中的第 t 个元素与 $\mathcal{S}'_{h(a'_t)}$ 中的第 t 个元素不同，因此 $\{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_L\}$ 的总敏感度至多为 2。因此，如果 d 的持续发布满足 ε-差分隐私，并且对于每个 $i \in [L]$，$\hat{}s_{i,1}$、$\hat{}s_{i,2}$、$\cdots$、$\hat{}s_{i,\,k}$ 的持续发布也满足 ε-差分隐私，那么$\hat{s}_1$、$\hat{s}_2$、$\cdots$、$\hat{s}_k$ 的持续发布满足 (3ε)-差分隐私。 □

引理 5.15。 考虑任意时间戳 $t \in [T]$。 $\forall a \in \mathcal{U}$，令 f_a 表示 a 在 a_1，a_2，$\cdots$，a_t中的频率。以至少 0.9 的概率，$\forall j \in [k]$，算法 7 的输出 $\hat{s}_j$ 是一个

$\left(\alpha_2, \left(\alpha_2\eta + \frac{\eta^2\gamma_2}{32\lambda}\right) \cdot \|\mathcal{S}\|_0 + 6\gamma_1 + 128\lambda/\eta^2\right)$-近似于 $|\{a \in \mathcal{U} \mid f_a = j\}|$，其中 $\|\mathcal{S}\|_0$ 是 a_1，a_2，$\cdots$，a_t中的不同元素数量。

为了证明引理 5.15，我们需要以下中间结论。考虑时间戳 $t \in [T]$.$\forall a \in \mathcal{U}$，令 f_a 表示 $a_1, a_2, \cdots, a_t$中 a 的频率。若算法 7 中存在 i^*，我们定义$G = \{a_j \mid g(a_j) = i^*, j \in [t]\}$。令 $\|\mathcal{S}\|_0$ 表示 $a_1, a_2, \cdots, a_t$中的不同元素数量。对于 $l \in [k]$，定义 $G_l = \{a \in G \mid f_a = l\}$。

断言 5.16. 概率至少为0.98时，$\forall a \neq a' \in G, h(a) \neq h(a')$。

证明。由于 $E[|G|] = \|\mathcal{S}\|_0/2^{i^*}$，根据马尔可夫不等式，$|G| \leq 100 \cdot \|\mathcal{S}\|_0/2^{i^*}$ 成立的概率至少为 0.99。由于算法 7 可以找到 i^*，我们有 $\|\mathcal{S}\|_0 \leq 1.1 \cdot (d + \gamma_1) \leq 2d$。根据 i^* 的选取，我们有 $2^{i^*} \geq \hat{}d/(128\lambda/\eta^2)$。因此，$\|\mathcal{S}\|_0/2^{i^*} \leq 256\lambda/\eta^2$。以至少 0.99 的概率，$|G| \leq 25600\lambda/\eta^2$。根据并集界和马尔可夫不等式，$\Pr[\forall a \neq a' \in G, h(a) \neq h(a') \mid |G| \leq 25600\lambda/\eta^2] \geq 1 - |G|^2/m \geq 0.99$。□

断言 5.17. $\forall l \in [k]$，$\Pr[2^{i^*} \cdot |G_l| \in |\{a \in U \mid f_a = l\}| \pm \eta\|\mathcal{S}\|_0] \geq 1 - 0.01/k$。

证明。设 $s_l = |\{a \in U \mid f_a = l\}|$。由于可以找到 i^*，我们有 $d \geq 3\gamma_1$。注意，$s_l \leq \|\mathcal{S}\|_0 \leq 1.1(\hat{}d + \gamma_1) \leq 2\hat{}d \leq 2^{i^*} \cdot 4 \cdot 64\lambda/\eta^2$。应用引理 2.6，$\forall l \in [k]$ 我们有：

$$\begin{aligned}
&\Pr\left[\left||G_l| - s_l/2^{i^*}\right| > 32\lambda/\eta\right] \\
\leq &8 \cdot \left(\frac{\lambda \cdot s_l/2^{i^*} + \lambda^2}{(32\lambda/\eta)^2}\right)^{\lambda/2} \\
\leq &0.01/k,
\end{aligned}$$

其中，最后一个不等式由 $s_l/2^{i^*} \leq 4 \cdot 64\lambda/\eta^2$ 和 $\lambda = 2 \cdot \log(1000k)$ 得出。由于 $\|\mathcal{S}\|_0 \geq (\hat{} d - \gamma_1)/1.1 \geq \hat{}d/2 \geq 2^{i^*} \cdot 32\lambda/\eta^2$，我们有

$$\Pr\left[\left||G_l| - s_l/2^{i^*}\right| > \eta\|\mathcal{S}\|_0/2^{i^*}\right] \leq 0.01/k.$$

□

接下来我们给出引理 5.15 的证明

引理 *5.15* 的证明。令 $\mathcal{E}$ 表示事件：$\forall a \neq a' \in G, h(a) \neq h(a')$。令 $\forall l \in [k], s_l = |\{a \in U \mid f_a = l\}|$。令 $\mathcal{E}'$ 表示事件：$\forall l \in [k], 2^{i^*} \cdot |G_l| \in s_l \pm \eta\|\mathcal{S}\|_0$。根据断言 5.16 和断言 5.17，$\mathcal{E}$ 和 $\mathcal{E}'$ 同时发生的概率至少为 0.97。在证明的其余部分中，我们以事件 $\mathcal{E}$ 和 $\mathcal{E}'$同时发生为条件。

首先，考虑 $\hat{} d < \max(3\gamma_1, 64\lambda/\eta^2)$ 这一情形。由于 $\hat{}d$ 是 $\|\mathcal{S}\|_0$ 的 $(1.1, \gamma_1)$-近似，因此我们有 $\|\mathcal{S}\|_0 \leq 6\gamma_1 + 128\lambda/\eta^2$。在这种情形下，$\forall l \in [k]$，$\hat{s}_l = 0$ 和 $|\{a \in \mathcal{U} \mid f_a = l\}| \leq \|\mathcal{S}\|_0 \leq 6\gamma_1 + 128\lambda/\eta^2$。

接下来，考虑 $\hat{} d \geq \max(3\gamma_1, 64\lambda/\eta^2)$ 的情形。对于每个 $j \in [k]$，令 $s_{i^*,l}$ 表示 $\mathcal{S}_l$ 中频率恰好为 j 的元素数量。根据事件 $\mathcal{E}$，我们有$\forall l \in [k], s_{i^*,l} = |G_l|$。根据事件 $\mathcal{E}'$, $\forall l \in [k]$，我们有：

$$\hat{s}_l \geq 2^{i^*} \cdot \left(\frac{1}{\alpha_2} \cdot |G_l| - \gamma_2\right) \geq \frac{1}{\alpha_2} \cdot (s_l - \eta\|\mathcal{S}\|_0) - 2^{i^*} \cdot \gamma_2$$

$$\geq \frac{1}{\alpha_2} \cdot s_l - \left(\eta + \frac{\eta^2\gamma_2}{32\lambda}\right) \cdot \|\mathcal{S}\|_0,$$

其中，第一个不等式由 $\hat{s}_{i^*,l}$ 是 $s_{i^*,l}$ 的 (α_2, γ_2)-近似以及 $s_{i^*,l} = |G_l|$ 得出，第二个不等式由事件 $\mathcal{E}'$ 得出，第三个不等式由 $\alpha_2 \geq 1$ 和 $\|\mathcal{S}\|_0 \geq (\hat{d} - \gamma_1)/1.1 \geq \hat{d}/2 \geq 2^{i^*} \cdot 32\lambda/\eta^2$ 得出。同理，$\forall l \in [k]$，我们有：

$$\begin{aligned}\hat{s}_l &\leq 2^{i^*} \cdot (\alpha_2 \cdot |G_l| + \gamma_2) \leq \alpha_2 \cdot (s_l + \eta\|\mathcal{S}\|_0) + 2^{i^*} \cdot \gamma_2 \\ &\leq \alpha_2 \cdot s_l + \left(\alpha_2\eta + \frac{\eta^2\gamma_2}{32\lambda}\right) \cdot \|\mathcal{S}\|_0.\end{aligned}$$

□

定理 5.18 (低频元素计数的流式持续发布)。设 $k \geq 1, \varepsilon \geq 0, \xi \in (0, 0。5)$ $\eta \in (0, 0。5)$。存在一个 ε-差分隐私算法，在流式持续发布模型中，以至少 $1 - \xi$ 的概率始终输出数字 $k\,\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_k$，对每个时间戳 t，只要满足 $\forall i \in [k]$，$\hat{s}_i$ 就是 $|\{a \in \mathcal{U} \mid f_a = i\}|$ 的一个近似，其加性误差为：

$$\left(\eta + \eta^2 \cdot \frac{k}{\varepsilon} \cdot \text{poly}\left(\log\left(\frac{Tk}{\xi}\right)\right)\right) \cdot \|\mathcal{S}\|_0 + \frac{1}{\varepsilon} \cdot \text{poly}\left(\log\left(\frac{T}{\xi}\right)\right) + O\left(\frac{\log k}{\eta^2}\right)$$

该算法使用 $\frac{1}{\eta^4} \cdot$ 多项式 $\left(\text{对数}(T \cdot k/\xi)\min(\varepsilon, 1)\right)$ 空间。

证明。根据引理 5.15，近似保证以至少 0.9 的概率成立。若要将所有 $t \in [T]$ 的成功概率提升至 $1 - \xi/3$，我们需要运行 $\lceil 50\log(3T/\xi) \rceil$ 个算法 7 的独立副本。由于我们运行了 $\lceil 50\log(3T/\xi) \rceil$ 个算法 7 的副本，并且根据引理 5.14，若要使最终算法满足 ε-差分隐私，则对于每个 $i \in [k]$，流式持续发布 $\{\hat{s}_{i,1}, \hat{s}_{i,2}, \cdots, \hat{s}_{i,k}\}$ 的每个子程序都需要满足 $(\varepsilon/(3 \cdot \lceil 50\log(2T/\xi) \rceil)$-差分隐私，并且流式持续发布 d 的子程序也需要满足 $(\varepsilon/(3 \cdot \lceil 50\log(3T/\xi) \rceil)$-差分隐私。为了同时确保算法 7 的所有独立副本中，流式持续发布 $\{\hat{s}_{i,1}, \hat{s}_{i,2}, \cdots, \hat{s}_{i,k}\}$ 的每次调用都以至少 $1 - \xi/3$ 的概率满足所需的 (α_2, γ_2)-近似，我们需要使每个特定 $i \in [L]$ 和算法 7 的特定副本所对应的 $\{\hat{s}_{i,1}, \hat{s}_{i,2}, \cdots, \hat{s}_{i,k}\}$，以至少 $1 - \xi/(3 \cdot \lceil 50\log(3T/\xi) \rceil \cdot L)$ 的概率满足近似保证。为了同时确保算法 7 的所有独立副本中，流式持续发布 d 的每次调用都以至少 $1 - \xi/3$ 的概率满足所需的 $(1.1, \gamma_1)$-近似，我们需要使算法 7 的每个特定副本所对应的 d，以至少 $1 - \xi/(3 \cdot \lceil 50\log(3T/\xi) \rceil)$ 的概率满足近似保证。然后，根据推论 4.10，我们有 $\gamma_1 = \frac{1}{\varepsilon} \cdot \text{poly}\left(\log\left(\frac{T}{\xi}\right)\right)$。根据定理 5.13，我们有 $\alpha_2 = 1$ 和 $\gamma_2 = \frac{k}{\varepsilon} \cdot \text{poly}\left(\log\left(\frac{Tk}{\xi}\right)\right)$。将上述参数代入引理 5.15，可得 $\forall i \in [k]$，$\hat{s}_i$ 是 $|\{a \in \mathcal{U} \mid f_a = i\}|$ 的一个近似，其加性误差为：

$$\left(\eta + \eta^2 \cdot \frac{k}{\varepsilon} \cdot \text{poly}\left(\log\left(\frac{Tk}{\xi}\right)\right)\right) \cdot \|\mathcal{S}\|_0 + \frac{1}{\varepsilon} \cdot \text{poly}\left(\log\left(\frac{T}{\xi}\right)\right) + O\left(\frac{\log k}{\eta^2}\right).$$

接下来，考虑空间使用量。根据推论 4.10，发布算法 7 的所有副本的 $\hat{d}$ 所需的总空间为 $\text{poly}\left(\frac{\log(T/\xi)}{\varepsilon 1}\min(,)\right)$。根据定理 5.13，发布所有算法 7 副本的 $\{\hat{s}_{i,1}、\hat{s}_{i,2}、\cdots、\hat{s}_{i,k}\}$ 所需的总空间为 $O(\log(T/\xi) \cdot L \cdot (m + \log(T))) = \frac{1}{\eta^4} \cdot \text{poly}\left(\log\left(\frac{T \cdot k}{\xi}\right)\right)$。因此，总空间至多为 $\frac{1}{\eta^4} \cdot \text{poly}\left(\log(T \cdot k/\xi)\min(\varepsilon, 1)\right)$。

□

5.4 ℓ_p 矩估计

本节将说明如何利用我们的 ℓ_p 重频项以及低频元素数量估计器来解决 ℓ_p 矩估计问题。

算法 8： ℓ_p 频率矩估计

输入：</style}一个由元素 $\mathcal{S}$组成的数据流 a_1、a_2、$\cdots$、$a_T \in \mathcal{U} \cup \{\perp\}$，以及误差参数 $\eta \in (0, 0.5)$。

参数：</style}一个 τ 阈值参数，其取决于重频项算法；一个相对近似因子 α 和一个加性近似因子 γ，其取决于低频元素计数估计算法。//参见定理 5.10 和定理 5.18。

输出： 在每个时间戳 t处对 ℓ_p 频率矩进行估计。

令 β' 从 $[1/2, 1]$ 中均匀随机抽取，并令 $\beta \in \beta' \pm (\eta/T)^C$，其中某个常数足够大常数 $C > 0$。//因此，β 可用 $\Theta(\log(T/\eta))$ 位表示。

令 q_1^* 为满足 $\beta(1+\eta)^q$ 的最小整数 $\hat{}_1^* > \tau$ 且令 q_2^* 为满足条件的最小整数 $\beta(1+\eta)^{q_2^*+1} \geq T$，并且对于任意 $q \in [q_1^*, q_2^*]$，定义区间 $I_q = (\beta(1+\eta)^q, \beta(1+\eta)^{q+1}]$。

$[L \leftarrow \log(|\mathcal{U}|), \lambda \leftarrow 2 \cdot \log(1000(L+1)\log(4T)/\eta)$.

令 k 为满足 $k \leq \beta(1+\eta)^{q_1^*}$ 的最大整数。令 $B \leftarrow \left(\frac{\log(4T)}{\eta}\right) \cdot 100(L+1) \cdot \frac{32\lambda}{\eta^3} \cdot (1+\eta)^p$。

令 $g : \mathcal{U} \to [L] \cup \{\perp\}$ 为一个 λ-wise 独立的哈希函数，并且 $\forall a \in \mathcal{U}, i \in [L], \Pr[g(a) = i] = 2^{-i}, \Pr[g(a) = \perp] = 2^{-L}$.

初始化空数据流 $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_L$。

对数据流 $\mathcal{S}$ 中的每个 a_t **执行**

  如果 $a_t \neq \perp$，**则**

    将 a_t 追加到 $\mathcal{S}_0$末尾。

    对于 $i \in [L]$，如果 $g(a_t) = i$，则将 a_t 追加到 $\mathcal{S}_i$；否则将 $\perp$ 追加到 $\mathcal{S}_i$。

  end

  else

    对于 $i \in [L] \cup \{0\}$，将 $\perp$ 追加到 $\mathcal{S}_i$末尾。

  end

  对于每个 $i \in [L] \cup \{0\}$，计算一个集合 $H_i \subseteq \mathcal{U}$ 以及一个函数 $\hat{f}_i$：满足以下条件的 $H_i \to \mathbb{R}_{\geq 0}$：

  1. $\forall a \in H_i$，$(1-\eta') \cdot f_a \leq \hat{} f_i(a) \leq (1+\eta') \cdot f_a$，其中 f_a 是 a 在 a_1、a_2、$\cdots$、a_t 中的频率，且 η' 满足 $\eta' \leq \frac{\eta}{10000(L+1)|H_i|}$.
  2. 如果 $f_a \geq \tau$ 且 $f_a^p \geq \|\mathcal{S}_i\|_p^p / B$，则出现在 $\mathcal{S}_i$中的 $\forall a \in \mathcal{U}$。

  计算 $\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_k$，其中 $\forall l \in [k], \hat{s}_l$ 是 $|\{a \in \mathcal{U} \mid f_a = l\}|$ 的 (α, γ)-近似。

  对于 $q \in [q_1^*, q_2^*]$，**执行**

    初始化 $\hat{z}_q = 0$.

    对于 $i \in [L] \cup \{0\}$，**执行**

      如果 $|\{a \in H_i \mid \hat{f}_i(a) \in I_q\}| \geq 8\lambda/\eta^2$ 或 $i = 0$ **则**

        $\hat{z}_q \leftarrow \max(\hat{z}_q, |\{a \in H_i \mid \hat{} f_i(a) \in I_q\}| \cdot 2^i)$.

      end

    end

  end

  ^输出 $F_p = \sum_{l \in [k]} \hat{s}_l \cdot l^p + \sum_{q \in [\hat{q}_1^*, q_2^*]} \hat{z}_q \cdot (\beta(1+\eta)^q)^p$

end

引理 5.19。考虑算法 *8* 中的子程序。如果持续发布$\hat{} s_1$，$\hat{s}_2$，$\cdots$，$\hat{s}_k$ 的算法是 ε-差分隐私，并且对于每个 $i \in [L] \cup \{0\}$，持续发布(H_i, f_i) 的算法是ε-差分隐私。算法 *8* 在持续发布模型中是 4ε-差分隐私。

证明。考虑两个相邻数据流 $\mathcal{S}=(a_1,a_2,\cdots,a_T)$ 和 $\mathcal{S}'=(a_1',a_2',\cdots,a_T')$，它们仅在时间戳 t 处不同，即 $a_t\neq a_t'$。考虑对应的数据流 $\mathcal{S}_0,\mathcal{S}_1,\cdots,\mathcal{S}_L$和 $\mathcal{S}_0',\mathcal{S}_1',\cdots,\mathcal{S}_L'$。 $\forall i\in[L]\cup\{0\},j\neq t$， $\mathcal{S}_i$ 的第 j 个元素应与 $\mathcal{S}_i'$ 的第 j 个元素相同。此外， $\forall i\in[L]$ 在 $i\neq g(a_t)$ $i\neq g(a_t')$ 的情况下， $\mathcal{S}_i$ 的第 t 个元素也与 $\mathcal{S}_i'$ 的第 t 个元素相同。因此，最多有 3 个数据流 $\mathcal{S}_0$,$\mathcal{S}_{g(a_t)}$ 和 $\mathcal{S}_{g(a_t')}$ 可能分别与 $\mathcal{S}_0'$,$\mathcal{S}_{g(a_t)}'$ 和 $\mathcal{S}_{g(a_t')}'$ 不同。因此，输出 $\{(H_0,f_0),(H_1,f_1)\cdots,(H_L,f_L)\}$ 满足 3ε-差分隐私。由于 $(\hat{s}_1,\hat{s}_2,\cdots,\hat{s}_k)$ 满足 ε-差分隐私，且最终输出仅取决于 $(\hat{s}_1,\hat{s}_2,\cdots,\hat{s}_k)$ 和 $\hat{}\hat{}\hat{}\{(H_0,f_0),(H_1,f_1)\cdots,(H_L,f_L)\}$，因此最终输出满足 4ε- 差分隐私。 □

在本节余下部分，我们来分析算法 8 的近似保证。考虑时间戳 $t\in[T]$ 和 $\forall a\in\mathcal{U}$，令 f_a 表示 a 在 $a_1,a_2,\cdots,a_t$中的频率。令$\mathcal{S},\mathcal{S}_0,\mathcal{S}_1,\cdots,\mathcal{S}_L$ 表示截至时间戳 t的数据流。令 $\mathcal{I}=\{[1,1],[2,2],\cdots,[k,k],I_{q_1^*},I_{q_1^*+1},\cdots,I_{q_2^*}\}$。根据我们对 k,q_1^*,q_2^*的选择，有如下观察:

观察 5.20。 $\sum_{I\in\mathcal{I}}\sum_{a\in\mathcal{U}:\ f_a\in I}f_a^p=\|\mathcal{S}\|_p^p$。

定义 5.21。 对于任意 $I\in\mathcal{I}$，如果 $\sum_{a\in\mathcal{U}:\ f_a\in I}f_a^p\geq\eta\|\mathcal{S}\|_p^p/(q_2^*-q_1^*+1)$ 或者 I 对于某个 $i\in[k]$是 $\{i\}$，那么区间 I 就是贡献的。

5.4.1 高频元素分析

在本节中，我们证明 $\sum_{q\in[q_1^*,q_2^*]}\hat{}z_q\cdot(\beta(1+\eta)^q)^p$是 $\sum_{q\in[q_1^*,q_2^*]}\sum_{a\in\mathcal{U}:\ f_a\in I_q}f_a^p$ 和 $\sum_{q\in[q_1^*,q_2^*]:\ I_q\text{是贡献的}}\sum_{a\in\mathcal{U}:\ f_a\in I_q}f_a^p$ 的良好近似。

下面的引理说明，以较高概率，任意特定采样元素的频率都应充分远离区间 $I_{q_1^*}$、 $I_{q_1^*+1}$、 $\cdots$、 $I_{q_2^*}$ 的边界。

引理 5.22 (Indyk 和 Woodruff [2005])。 考虑算法 *8* 中的 β。任取一个 $f\in[T]$ 和任意一个 $r\geq2(\eta/T)^{C-1}$。

$$\Pr_{\beta'}\left[\min_{q\in\{q_1^*,q_1^*+1,\cdots,q_2^*,q_2^*+1\}}|f-\beta(1+\eta)^q|<r\right]\leq\frac{100r}{\eta\cdot f}$$

$\hat{}$ 通过应用上述引理，我们证明以高概率，可以使用 $f_i(a)$ 将 a正确分类到 $I_{q_1^*},I_{q_1^*+1},I_{q_1^*+2},\cdots,I_{q_2^*}$ 中的正确区间。

$\hat{}$**引理 5.23**。 概率至少为0.99时， $\forall i\in[L]\cup\{0\},\forall a\in H_i$ 和 $\forall q\in[q_1^*,q_2^*]$, $f_i(a)\in I_q$当且仅当 $f_a\in I_q$。

证明。 根据引理 5.22，对于任意 $i\in[L]\cup\{0\}$ 和任意 $a\in H_i$，我们有:

$$\Pr\left[\min_{q\in\left\{q_1^*,q_1^*+1,\cdots,q_2^*+1\right\}}|f_a-\beta(1+\eta)^q|<\eta' f_a\right]\leq\frac{1}{100\cdot(L+1)\cdot|H_i|},$$

其中，不等式由 $\eta'\leq\eta/(10000(L+1)|H_i|)$ ~~得出。对所有~~ $i\in[L]\cup\{0\}$ 和所有 $a\in H_i$ 应用并集界后，以至少 0.99 的概率，以下事件成立： $\forall i\in[L]\cup\{0\},\forall a\in H_i,\forall q\in\{q_1^*,q_1^*+1,\cdots,q_2^*+1\}$,

1. 如果 $f_a\leq\beta(1+\eta)^q$，那么 $\hat{}$ $f_i(a)\leq f_a+\eta' f_a\leq\beta(1+\eta)^q$,

ˆ 2. 如果 $f_a > \beta(1+\eta)^q$，那么 $f_i(a) \geq f_a - \eta' f_a > \beta(1+\eta)^q$。

ˆ 因此，以至少 0.99 的概率，$\forall i \in [L] \cup \{0\}$、$\forall a \in H_i$ 和 $\forall q \in [q_1^*, q_2^*]$、$f_i(a) \in I_q$ 当且仅当 $f_a \in I_q$。 □

令 G_i 表示出现在数据流 $\mathcal{S}_i$ 中的元素集合，并令 $G_{i,q}$ 表示出现在数据流 $\mathcal{S}_i$ 中且其频率位于区间 I_q 内的元素集合。形式化地，$\forall i \in [L]$，令 $G_i = \{a_j \mid g(a_j) = i、\ j \leq t\}$、$G_0 = \{a_j \mid j \leq t\}$，并令 $\forall q \in \{q_1^*, q_1^*+1, \cdots, q_2^*\}$、$\forall i \in [L] \cup \{0\}$，令 $G_{i,q} = \{a \in G_i \mid f_a \in I_q\}$。对于 $q \in [q_1^*, q_2^*]$，令 $z_q = |G_{0,q}|$，即位于数据流 $\mathcal{S}$ 中且频率处于范围 I_q 内的元素数量。

引理 5.24。 $\forall i \in [L] \cup \{0\}, q \in [q_1^*, q_2^*]$，若 $i = 0$ 或 $z_q \geq 2^i \cdot 4\lambda/\eta^2$，$\Pr[|G_{i,q}| \in (1 \pm \eta) \cdot z_q/2^i] \geq 1 - 0.01/((L+1) \cdot \log(4T)/\eta)$。否则，$\Pr[||G_{i,q}| - z_q/2^i| \leq 4\lambda/\eta] \geq 1 - 0.01/((L+1) \cdot \log(4T)/\eta)$。

证明。 考虑任意 $i \in [L] \cup \{0\}$ 和 $q \in [q_1^*, q_2^*]$。若 $i = 0$，根据定义 $z_q = |G_{0,q}|$。假设 $z_q \geq 2^i \cdot 4\lambda/\eta^2$。由引理 2.6 可得：

$$\begin{aligned}
&\Pr\left[\left||G_{i,q}| - z_q/2^i\right| > \eta \cdot z_q/2^i\right] \\
\leq &8 \cdot \left(\frac{\lambda \cdot z_q/2^i + \lambda^2}{(\eta \cdot z_q/2^i)^2}\right)^{\lambda/2} \\
\leq &0.01/\left((L+1) \cdot \log(4T)/\eta\right),
\end{aligned}$$

其中，最后一个不等式由 $\lambda = 2 \cdot \log(1000(L+1)\log(4T)/\eta)$ 以及 $z_q \geq 2^i \cdot 4\lambda/\eta^2$ 推出。假设 $z_q \leq 2^i \cdot 4\lambda/\eta^2$，再次应用引理 2.6，可得：

$$\begin{aligned}
&\Pr\left[\left||G_{i,q}| - z_q/2^i\right| > 4\lambda/\eta\right] \\
\leq &8 \cdot \left(\frac{\lambda \cdot z_q/2^i + \lambda^2}{(4\lambda/\eta)^2}\right)^{\lambda/2} \\
\leq &0.01/\left((L+1) \cdot \log(4T)/\eta\right),
\end{aligned}$$

wher最后一个不等式由 $z_q/2^i \leq 4\lambda/\eta^2$ 和 $\lambda = 2 \cdot \log(1000(L+1)\log(4T)/\eta$ 推出). □

我们将事件 $\mathcal{E}_1$ 和 $\mathcal{E}_2$ 定义如下：$\mathcal{E}_1$ 表示事件 $\forall i \in [L] \cup \{0\}, \forall a \in H_i, \forall q \in$ˆ$[q_1^*, q_2^*], f_i(a) \in I_q$ 当且仅当 $f_a \in I_q$。$\mathcal{E}_2$ 表示事件：

1. $\forall i \in [L] \cup \{0\}$、$\forall q \in [q_1^*, q_2^*]$，如果 $z_q \geq 2^i \cdot 4\lambda/\eta^2$ 或 $i = 0$、$|G_{i,q}| \cdot 2^i \in (1 \pm \eta) z_q$。
2. $\forall i \in [L], \forall q \in [q_1^*, q_2^*]$, if $z_q < 2^i \cdot 4\lambda/\eta^2$, $|G_{i,q}| \in z_q/2^i \pm 4\lambda/\eta$.

根据引理 5.22，$\mathcal{E}_1$ 发生的概率至少为 0.99。根据引理 5.24，由于 $\forall q \in [q_1^*, q_2^*]$、$z_q = |G_{0,q}|$ 和 $q_2^* - q_1^* + 1 \leq \log(4T)/\eta$，对所有$i \in [L] \cup \{0\}$ 和所有 $q \in [q_1^*, q_2^*]$应用并集界，$\mathcal{E}_2$ 发生的概率至少为 0.99。

引理 5.25（高频矩估计的上界）。在 $\mathcal{E}_1$ 和 $\mathcal{E}_2$的条件下，

$$\sum_{q \in [q_1^*, q_2^*]} \hat{z}_q \cdot (\beta(1+\eta)^q)^p \leq (1+\eta) \cdot \sum_{q \in [q_1^*, q_2^*]} \sum_{a \in \mathcal{U}, f_a \in I_q} f_a^p$$

ˆ证明。由于事件 $\mathcal{E}_1$，我们有 $\forall i \in [L]\cup\{0\}, \forall q \in [q_1^*, q_2^*], \{a \in H_i \mid f_i(a) \in I_q\} \subseteq G_{i,\ q}$。注意，对于 $q \in [q_1^*, q_2^*]$，ˆz_q 要么为 0，要么是某个满足 $i' = 0$ 的 i' 所对应的 ˆ$z_q = 2^{i'} \cdot |\{a \in H_{i'} \mid \hat{f}_{i'}(a) \in I_q\}|$，或者是 ˆ$f_{i'}(a) \in I_q\}| \geq 8\lambda/\eta^2$。由于 $\forall q \in [q_1^*, q_2^*], \forall i \in [L]\cup\{0\}, |\{a \in H_i \mid \hat{f}_i(a) \in I_q\}| \leq |G_{i,\ q}|$，$\forall q \in [q_1^*, q_2^*]$，若 ˆ$z_q \neq 0$，则存在某个 $i' \in [L]\cup\{0\}$ 使得：

$$\begin{aligned}\hat{z}_q &= 2^{i'} \cdot |\{a \in H_{i'} \mid \hat{f}_{i'}(a) \in I_q\}| \\ &\leq 2^{i'} \cdot |G_{i',q}| \\ &\leq (1+\eta) \cdot z_q,\end{aligned}$$

其中，第二个不等式由 $|G_{i',q}| \geq 8\lambda/\eta^2$ 推出，这意味着根据事件 $\mathcal{E}_2$，$|G_{i',q}| \cdot 2^{i'} \in (1\pm\eta) z_q$。

因此，

$$\begin{aligned}&\sum_{q\in[q_1^*,q_2^*]} \hat{z}_q \cdot (\beta(1+\eta)^q)^p \\ \leq &(1+\eta) \sum_{q\in[q_1^*,q_2^*]} z_q \cdot (\beta(1+\eta)^q)^p \\ \leq &(1+\eta) \sum_{q\in[q_1^*,q_2^*]} \sum_{a\in\mathcal{U}: f_a\in I_q} (\beta(1+\eta)^q)^p \\ \leq &(1+\eta) \sum_{q\in[q_1^*,q_2^*]} \sum_{a\in\mathcal{U}: f_a\in I_q} f_a^p,\end{aligned}$$

其中，第一个不等式由 ˆ$z_q \leq (1+\eta) z_q$ 得出，第二个不等式由 z_q 的定义（即 $z_q = |\{a \in \mathcal{U} \mid f_a \in I_q\}|$）得出，最后一个不等式由 $f_a \geq \beta(1+\eta)^q$ if $f_a \in I_q$ 得出。 □

将事件 $\mathcal{E}_3$ 定义为：$\forall i \in [L]\cap\{0\}, \|\mathcal{S}_i\|_p^p \leq 100(L+1) \cdot \|\mathcal{S}\|_p^p / 2^i$。

引理 5.26。 $\mathcal{E}_3$ 发生的概率至少为 0.99。

证明。考虑 $i \in [L]\cup\{0\}$。我们有 $E\left[\|\mathcal{S}_i\|_p^p\right] = \sum_{a\in\mathcal{U}} \mathbf{Pr}[g(a) = i] \cdot f_a^p = \|\mathcal{S}\|_p^p / 2^i$。根据马尔可夫不等式，以至少 $1 - 1/(100(L+1))$ 的概率，$\|\mathcal{S}_i\|_p^p \leq 100(L+1) \cdot \|\mathcal{S}\|_p^p / 2^i$。对所有 $i \in [L]\cup\{0\}$ 使用并集界后，$\mathcal{E}_3$ 发生的概率至少为 0.99。 □

在剩余的分析中，我们也以 $\mathcal{E}_3$ 为条件。

引理 5.27（贡献的高频矩估计的下界）。以 $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ 为条件。

$$\sum_{q\in[q_1^*,q_2^*]} \hat{z}_q \cdot (\beta(1+\eta)^q)^p \geq (1-\eta)^{p+1} \cdot \sum_{q\in[q_1^*,q_2^*]: I_q \text{ is contributing}} \sum_{a\in\mathcal{U}, f_a\in I_q} f_a^p$$

证明。考虑任意一个贡献的 $q \in [q_1^*, q_2^*]$。

情况 1： $z_q \leq 16\lambda/\eta^2$。根据定义 5.21，我们有 $\sum_{a\in\mathcal{U}:\ f_a\in I_q} f_a^p \geq \eta\|\mathcal{S}\|_p^p / (q_2^* - q_1^* + 1)$，这意味着 $\eta\|\mathcal{S}_0\|_p^p / (q_2^* - q_1^* + 1) \leq z_q \cdot (\beta(1+\eta)^{q+1})^p \leq 16\lambda/\eta^2 \cdot (1+\eta)^p \cdot (\beta(1+\eta)^q)^p$。

由于 $B \geq (q_2^* - q_1^* + 1) \cdot 16\lambda \cdot (1+\eta)^p/\eta^3, \forall a \in \mathcal{U}$ 在 $f_a \in I_q$ 的情况下，$f_a^p \geq \|\mathcal{S}_0\|_p^p/B$。因此，$f_a \in I_q$ 时的 $\forall a \in \mathcal{U}$，我们有 $a \in H_0$。根据事件 $\mathcal{E}_1$，我们有 $|\{a \in H_0 \mid f_0(a) \in I_q\}| = |G_{0,q}| = z_q$。因此，我们有 $\hat{z}_q \geq |\{a \in H_0 \mid f_0(a) \in I_q\}| \geq z_q$。

情况 2： $z_q > 16\lambda/\eta^2$。令 $i^* \in \{0\} \cup [L]$ 为满足 $z_q/2^{i^*} \geq 16\lambda/\eta^2$ 的最大值。根据事件 $\mathcal{E}_2$，我们有 $|G_{i^*,q}| \geq 8\lambda/\eta^2$。

由于 q 是贡献的，我们有：

$$\begin{aligned}
&z_q \cdot (\beta(1+\eta)^{q+1})^p \\
\geq &\sum_{a \in \mathcal{U}: f_a \in I_q} f_a^p \\
\geq &\eta \|\mathcal{S}\|_p^p/(q_2^* - q_1^* + 1) \\
\geq &2^{i^*} \cdot \eta \|\mathcal{S}_{i^*}\|_p^p/((q_2^* - q_1^* + 1) \cdot 100(L+1)),
\end{aligned}$$

其中，最后一个不等式由事件 $\mathcal{E}_3$ 得出。因此，我们有

$$\begin{aligned}
&(\beta(1+\eta)^q)^p \\
\geq &\eta \|\mathcal{S}_{i^*}\|_p^p/((q_2^* - q_1^* + 1) \cdot 100(L+1) \cdot (z_q/2^{i^*}) \cdot (1+\eta)^p) \\
\geq &\eta \|\mathcal{S}_{i^*}\|_p^p/((q_2^* - q_1^* + 1) \cdot 100(L+1) \cdot (32\lambda/\eta^2) \cdot (1+\eta)^p),
\end{aligned}$$

其中，最后一个不等式由 $z_q/2^{i^*} \leq 32\lambda/\eta^2$ 得出。由于 $B \geq ((q_2^* - q_1^* + 1) \cdot 100(L+1) \cdot (32\lambda/\eta^2) \cdot (1+\eta)^p)/\eta$，我们有 $\forall a \in \mathcal{U}$，其中 $a \in G_{i^*,q}$，$f_a^p \geq \|\mathcal{S}_{i^*}\|_p^p/B$ 且 $f_a > \beta(1+\eta)^q > \tau$，这意味着 $a \in H_{i^*}$。根据事件 $\mathcal{E}_1$，我们有 $\{a \in H_{i^*} \mid f_{i^*}(a) \in I_q\} = G_{i^*,q}$。因此，$|\{a \in H_{i^*} \mid \hat{f}_{i^*}(a) \in I_q\}| \geq 8\lambda/\eta^2$。最后，此外，根据事件 $\mathcal{E}_2$，我们有 $\hat{z}_q \geq |\{a \in H_{i^*} \mid \hat{f}_{i^*}(a) \in I_q\}| \cdot 2^{i^*} \geq (1-\eta)z_q$。

因此，在任何情形下，我们都有 $\hat{z}_q \geq (1-\eta)z_q$，并且：

$$\begin{aligned}
&\sum_{q \in [q_1^*, q_2^*]} \hat{z}_q \cdot (\beta(1+\eta)^q)^p \\
\geq &\sum_{q \in [q_1^*, q_2^*]: I_q \text{ is contributing}} \hat{z}_q \cdot (\beta(1+\eta)^q)^p \\
\geq &\sum_{q \in [q_1^*, q_2^*]: I_q \text{ is contributing}} (1-\eta) z_q \cdot (\beta(1+\eta)^q)^p \\
= &\frac{1-\eta}{(1+\eta)^p} \sum_{q \in [q_1^*, q_2^*]: I_q \text{ is contributing}} \sum_{a \in \mathcal{U}: f_a \in I_q} (\beta(1+\eta)^{q+1})^p \\
\geq &\frac{1-\eta}{(1+\eta)^p} \sum_{q \in [q_1^*, q_2^*]: I_q \text{ is contributing}} \sum_{a \in \mathcal{U}: f_a \in I_q} f_a^p \\
\geq &(1-\eta)^{p+1} \sum_{q \in [q_1^*, q_2^*]: I_q \text{ is contributing}} \sum_{a \in \mathcal{U}: f_a \in I_q} f_a^p,
\end{aligned}$$

其中，等式由定义 $z_q = |\{a \in \mathcal{U} \mid f_a \in I_q\}|$ 得出。 □

5.4.2 低频元素分析

In t本节中，我们证明 $\sum_{l\in[k]}\hat{s}_l\cdot l^p$ 是 $\sum_{l\in[k]}\sum_{a\in\mathcal{U}:\ f_a=l}f_a^p$ 的良好近似。

引理 5.28 (低频矩的近似)。 $\sum_{l\in[k]}\hat{s}_l\cdot l^p$ 是 $\sum_{l\in[k]}\sum_{a\in\mathcal{U}:f_a=l}f_a^p$ 的（α，$\gamma\cdot(2\tau)^{p+1}$）-近似。

证明。对于 $l\in[k]$，令 $s_l=|\{a\in\mathcal{U}\mid f_a=l\}|$。我们有：

$$\begin{aligned}&\sum_{l\in[k]}\hat{s}_l\cdot l^p\\ \geq&\sum_{l\in[k]}(\frac{1}{\alpha}\cdot s_l-\gamma)\cdot l^p\\ \geq&\frac{1}{\alpha}\cdot\sum_{l\in[k]}\sum_{a\in\mathcal{U}:f_a=l}f_a^p-k\cdot\gamma\cdot k^p\\ \geq&\frac{1}{\alpha}\cdot\sum_{l\in[k]}\sum_{a\in\mathcal{U}:f_a=l}f_a^p-\gamma\cdot(2\tau)^{p+1},\end{aligned}$$

其中，最后一个不等式由我们的 k 选择蕴含 $k\leq 2\tau$ 得出。类似地，我们有：

$$\begin{aligned}&\sum_{l\in[k]}\hat{s}_l\cdot l^p\\ \leq&\sum_{l\in[k]}(\alpha\cdot s_l+\gamma)\cdot l^p\\ \leq&\alpha\cdot\sum_{l\in[k]}\sum_{a\in\mathcal{U}:f_a=l}f_a^p-\gamma\cdot(2\tau)^{p+1}.\end{aligned}$$

□

5.4.3 将高频矩与低频矩结合起来

引理 5.29。 $\sum$贡献的 $\sum_{I\in\mathcal{I}}\sum_{a\in\mathcal{U}:\ f_a\in I}f_a^p\geq\ (1-\eta)\ \cdot\|\mathcal{S}\|_p^p$。

证明。

$$\begin{aligned}&\sum_{\text{contributing }I\in\mathcal{I}}\sum_{a\in\mathcal{U}:f_a\in I}f_a^p\\ =&\sum_{I\in\mathcal{I}}\sum_{a\in\mathcal{U}:f_a\in I}f_a^p-\sum_{\text{non-contributing }I\in\mathcal{I}}\sum_{a\in\mathcal{U}:f_a\in I}f_a^p\\ \geq&\|\mathcal{S}\|_p^p-(q_2^*-q_1^*+1)\cdot\eta\cdot\|\mathcal{S}\|_p^p/(q_2^*-q_1^*+1)\\ \geq&(1-\eta)\cdot\|\mathcal{S}\|_p^p.\end{aligned}$$

□

引理 5.30。 考虑任意时间戳 $t \in [T]$。 $\forall a \in \mathcal{U}$，令 f_a 表示 a 在$a_1, a_2, \cdots, a_t$中的频率。以至少0.9的概率，算法 *8* 的输出$\hat{}$ F_p是对 $\|\mathcal{S}\|_p^p$的 $\left($最大值 $\left(\frac{\alpha}{1-\eta}, (1+2\eta)^{p+2}\right), \gamma \cdot (2\tau)^{p+1}\right)$-近似。

证明。 根据引理 5.25 和引理 5.27，我们有：

$$(1-\eta)^{p+1} \sum_{q \in [q_1^*, q_2^*]: I_q \text{ is contributing}} \sum_{a \in \mathcal{U}, f_a \in I_q} f_a^p \leq \sum_{q \in [q_1^*, q_2^*]} \hat{z}_q \cdot (\beta(1+\eta)^q)^p \leq (1+\eta) \sum_{q \in [q_1^*, q_2^*]} \sum_{a \in \mathcal{U}, f_a \in I_q} f_a^p.$$

根据引理 5.28，我们有：

$$\frac{1}{\alpha} \cdot \sum_{l \in [k]} \sum_{a \in \mathcal{U}: f_a = l} f_a^p - \gamma \cdot (2\tau)^{p+1} \leq \sum_{l \in [k]} \hat{s}_l \cdot l^p \leq \alpha \cdot \sum_{l \in [k]} \sum_{a \in \mathcal{U}: f_a = l} f_a^p + \gamma \cdot (2\tau)^{p+1}$$

因此，我们有：

$$\begin{aligned}
\hat{F}_p &\geq \min\left(\frac{1}{\alpha}, (1-\eta)^{p+1}\right) \sum_{I \in \mathcal{I}: I \text{ is contributing}} \sum_{a \in \mathcal{U}: f_a \in I} f_a^p - \gamma \cdot (2\tau)^{p+1} \\
&\geq \min\left(\frac{1-\eta}{\alpha}, (1-\eta)^{p+2}\right) \|\mathcal{S}\|_p^p - \gamma \cdot (2\tau)^{p+1} \\
&\geq \min\left(\frac{1-\eta}{\alpha}, \frac{1}{(1+2\eta)^{p+2}}\right) \|\mathcal{S}\|_p^p - \gamma \cdot (2\tau)^{p+1}
\end{aligned}$$

其中第二步由引理 5.29 得出。同理，我们有：

$$\begin{aligned}
\hat{F}_p &\leq \max(\alpha, 1+\eta) \sum_{I \in \mathcal{I}} \sum_{a \in \mathcal{U}: f_a \in I} f_a^p + \gamma \cdot (2\tau)^{p+1} \\
&= \max(\alpha, 1+\eta) \cdot \|\mathcal{S}\|_p^p + \gamma \cdot (2\tau)^{p+1}
\end{aligned}$$

□

定理 5.31 (流式持续发布 ℓ_p 频率矩估计**)**。 设 $p > 0, \varepsilon \geq 0, \xi \in (0, 0.5)$ $\eta \in (0, 0.5)$。在流式持续发布模型中，存在一个 ε-差分隐私算法，使得其始终以至少 $1-\xi$ 的概率输出对 $\|\mathcal{S}\|_p^p$的一个 $\left(1+\eta, \left(\frac{\log(T|\mathcal{U}|/\xi}{\eta\varepsilon}\right)^{O(\max(1,p))}\right)$-近似。该算法使用的空间至多为

$$\phi \cdot \left(\frac{\log(T|\mathcal{U}|/\xi)}{\eta\varepsilon}\right)^{O(\max(1,p))},$$

其中 $\phi = \max(1, |\mathcal{U}|^{1-2/p})$。

证明。 为了将引理 5.30 中近似保证成立的概率提升至 $1 - \xi/3$，并同时对所有时间戳 $t \in T$成立，我们运行 $\lceil 50 \log(3T/\xi) \rceil$ 个算法 8 的独立副本，并取输出的中位数。由于我们运行 $\lceil 50 \log(3T/\xi) \rceil$ 个算法 8 的独立副本，根据引理 5.19，若要使最终算法满足 ε-差分隐私，则流式持续发布 (H_i, f_i) 的每个子程序$\hat{}$对于 $i \in [L] \cup \{0\}$ 都需要满足 $(\varepsilon/(4 \cdot \lceil 50 \log(3T/\xi) \rceil))$-差分隐私

；流式持续发布 $\{\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_k\}$ 的子程序也需要满足 $(\varepsilon/(4\lceil 50\log(3T/\xi)\rceil)$-差分隐私。为了以至少 $1-\xi/3$ 的概率，使算法 8 的所有独立副本中，流式持续发布 (H_i, f_i) 的每次子程序调用都满足算法 8 中所述的期望性质，我们需要使 (H_i, f_i) 对每个特定的 $i \in [L] \cup \{0\}$ 以及算法 8 的某个特定副本，以至少 $1-\xi/(4\cdot\lceil 50\log(3T/\xi)\rceil\cdot(L+1))$ 的概率满足该性质。为了以至少 $1-\xi/3$ 的概率，使算法 8 的所有独立副本中，流式持续发布 $\{\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_k\}$ 的每次子程序调用都满足算法 8 中所述的期望性质，我们需要使 $\{\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_k\}$ 对每个特定的 $i \in [L] \cup \{0\}$ 以及算法 8 的某个特定副本，以至少 $1-\xi/(3\cdot\lceil 50\log(3T/\xi)\rceil)$ 的概率满足期望性质。根据算法 8，我们有

$$B = \Theta\left(\frac{\log(T)\log(|\mathcal{U}|)\log(\log(T|\mathcal{U}|)/\eta)\cdot(1+\eta)^p}{\eta^4}\right).$$

因此，根据定理 5.10，算法 8 中 $|H_i|$ 的大小至多为 $\mathrm{poly}\left(\log(T|\mathcal{U}|/\xi)_\eta\right)\cdot 2^{O(p)}$。因此，我们选择 $\eta' = 1/\left(\mathrm{poly}\left(\log(T|\mathcal{U}|/\xi)_\eta\right)\cdot 2^{O(p)}\right)$。接着，根据定理 5.10，我们有

$$\tau = \frac{1}{\varepsilon}\cdot\mathrm{poly}\left(\frac{\log(T\cdot|\mathcal{U}|/\xi)}{\eta}\right)\cdot 2^{O(p)}.$$

根据定理 5.18，我们有 $\alpha = 1$，并且我们选择 γ 为

$$\left(\eta'' + \eta''^2\cdot\frac{\tau}{\varepsilon}\cdot\mathrm{poly}\left(\log\left(\frac{T\tau}{\xi}\right)\right)\right)\cdot\|\mathcal{S}\|_0 + \frac{1}{\varepsilon}\cdot\mathrm{poly}\left(\log\left(\frac{T}{\xi}\right)\right) + O\left(\frac{\log\tau}{\eta''^2}\right),$$

其中我们选择 η'' 为

$$\frac{\varepsilon\eta}{\tau^{O(\max(1,p))}\mathrm{poly}\left(\frac{\log(T|\mathcal{U}|/\xi)}{\eta}\right)} = \frac{1}{\left(\frac{\log(T|\mathcal{U}|/\xi)}{\eta\varepsilon}\right)^{O(\max(1,p))}}$$

使得

$$\begin{aligned}\gamma\cdot(2\tau)^{p+1} &\le \eta\|\mathcal{S}\|_0 + \left(\frac{\log(T|\mathcal{U}|/\xi)}{\eta\varepsilon}\right)^{O(\max(1,p))}\\ &\le \eta\|\mathcal{S}\|_p^p + \left(\frac{\log(T|\mathcal{U}|/\xi)}{\eta\varepsilon}\right)^{O(\max(1,p))}.\end{aligned}$$

注意，$\eta\|\mathcal{S}\|_p^p$ 变成了相对误差。因此，根据引理 5.30，输出 $\hat{} F_p$ 是一个$\left((1+\eta)^{O(\max(1,p))}, \left(\log(T|\mathcal{U}|/\xi\right)_\eta^{O(\max(1,p))}\right)$-近似。

接下来，考虑空间使用量。根据定理 5.10，运行所有重频项子程序所需的总空间至多为

$$\phi\cdot 2^{O(\max(1,p))}\cdot\mathrm{poly}\left(\frac{\log(T|\mathcal{U}|/\xi)}{\eta}\right).$$

根据定理 5.18，对于算法 8 的所有运行副本，计算 $\{\hat{s}_1、\hat{s}_2、\cdots、\hat{s}_k\}$ 所需的总空间至多为

$$\left(\frac{\log(T|\mathcal{U}|/\xi)}{\eta\varepsilon}\right)^{O(\max(1,p))}.$$

因此，所需的总空间至多为

$$\phi\cdot\left(\frac{\log(T|\mathcal{U}|/\xi)}{\eta\varepsilon}\right)^{O(\max(1,p))}.$$

□

6 滑动窗口持续发布算法的扩展

本节简要回顾 Braverman 和 Ostrovsky [2007] 提出的平滑直方图技术，该技术可将任意（非私有）流式算法转换为（非私有）滑动窗口算法。原始框架仅支持具有相对误差且没有加性误差的近似算法。本节将介绍如何扩展该框架，使其也支持加性误差。

假设有两个数据流 $\mathcal{A}=(a_1,a_2,\cdots,a_{t_1})$ 和 $\mathcal{B}=(b_1,b_2,\cdots,b_{t_2})$。我们用 $\mathcal{A}\cup\mathcal{B}$表示两个数据流的拼接，即 $\mathcal{A}\cup\mathcal{B}=(a_1,a_2,\cdots,a_{t_1},b_1,b_2,\cdots,b_{t_2})$。如果 $\mathcal{B}$是 $\mathcal{A}$的后缀，即 $\exists i\in[t_1]$ 使得 $a_i=b_1,a_{i+1}=b_2,\cdots,a_{t_1}=b_{t_2}$，那么我们将其记为 $\mathcal{B}\subseteq_r\mathcal{A}$。

定义 6.1 (Braverman 和 Ostrovsky 的平滑函数 [2007])。 设 $g(\cdot)$ 是定义在数据流上的函数。函数 $g(\cdot)$ 是(ζ,β)-平滑的，当且仅当：

1. $\forall\mathcal{A},0\le g(\mathcal{A})\le$多项式$(T)$。
2. $\forall\mathcal{A},\ \mathcal{B}$ 其中 $\mathcal{B}\subseteq_r\mathcal{A}$, $g(\mathcal{A})\ge g(\mathcal{B})$
3. 对任意 $\eta\in(0,1)$，存在 $\zeta(\eta,g)$ 和 $\beta(\eta,g)$ 使得
   (a) $0<\beta\le\zeta<1$.
   (b) *If* $\mathcal{B}\subseteq_r\mathcal{A}$ *and* $(1-\beta)g(\mathcal{A})\le g(\mathcal{B})$ *then* $(1-\zeta)g(\mathcal{A}\cup\mathcal{C})\le g(\mathcal{B}\cup\mathcal{C})$ *for any stream* $\mathcal{C}$.

引理 6.2 (频率矩的平滑性，Braverman 和 Ostrovsky [2007])。 对于 $p>1$,$\|\mathcal{S}\|_p^p$是 $\left(\eta,\left(\frac{\eta}{p}\right)^p\right)$-平滑的。对于 $0<p\le1$, $\|\mathcal{S}\|_p^p$是 (η,η)-平滑的。 $\|\mathcal{S}\|_0$是 (η,η)-平滑的。

引理 6.3。 令 $0\le Z\le$多项式(T)。若 $g(\mathcal{S}):=\|\mathcal{S}\|_p^p+Z$，则 $g(\cdot)$是 $\left(\eta,\left(\frac{\eta}{p}\right)^p\right)$-平滑的（若 $p>1$），而 $g(\cdot)$是$(\eta,\ \eta)$-平滑的（若 $0<p\le1$）。若 $g(\mathcal{S})=\|\mathcal{S}\|_0+Z$，则 $g(\cdot)$是$(\eta,\ \eta)$-平滑的。

证明。 对于引理陈述中的 $g(\cdot)$，显然满足定义 6.1 中的前两项要求。因此，我们只需证明定义 6.1 中的第三项要求。

令 $\mathcal{A}$和 $\mathcal{B}$满足 $\mathcal{B}\subseteq_r\mathcal{A}$且 $(1-\beta)g(\mathcal{A})\le g(\mathcal{B})$。考虑任意数据流 $\mathcal{C}$。我们将证明 $(1-\zeta)g(\mathcal{A}\cup\mathcal{C})\le g(\mathcal{B}\cup\mathcal{C})$。考虑情形 $g(\mathcal{S})=\|\mathcal{S}\|_p^p+Z$。我们构造一个辅助数据流 $\mathcal{X}$，使得 $\mathcal{X}$中的元素不会出现在 $\mathcal{A}$、$\mathcal{B}$、$\mathcal{C}$和 $\|\mathcal{X}\|_p^p=Z$中的任何一个中。于是，我们有 $(1-\beta)\|\mathcal{A}\cup\mathcal{X}\|_p^p=(1-\beta)g(\mathcal{A})\le g(\mathcal{B})=\|\mathcal{B}\cup\mathcal{X}\|_p^p$。根据 $\|\mathcal{S}\|_p^p$的平滑性

（引理 6.2），我们有 $(1-\zeta)g(\mathcal{A}\cup\mathcal{C}) = (1-\zeta)\|\mathcal{A}\cup\mathcal{X}\cup\mathcal{C}\|_p^p \le \|\mathcal{B}\cup\mathcal{X}\cup\mathcal{C}\|_p^p = g(\mathcal{B}\cup\mathcal{C})$。根据引理 6.2，$g(\cdot)$在 $p>1$时是 $\left(\eta, \left(\frac{\eta}{p}\right)^p\right)$-平滑的，而 $g(\cdot)$在 $0<p\le 1$时是 (η,η)-平滑的。

类似地，通过构造 $\|\mathcal{X}\|_0 = Z$ 并采用类似的论证，我们可以证明 $g(\mathcal{S}) = \|\mathcal{S}\|_0 + Z$ 是 (η,η)-平滑的。

□

引理 6.4. 设 $\alpha \ge 1, \gamma \ge 0$。若 g' 是 g 的(α,γ)-近似，则当 $Z \ge \frac{\alpha}{\alpha-1}\cdot\gamma$ 时，$g'+Z$ 是 $g+Z$ 的 α- 近似。

证明。我们有：

$$\begin{aligned} g'+Z &\ge \frac{1}{\alpha}\cdot g - \gamma + Z \\ &\ge \frac{1}{\alpha}\cdot g - \frac{\alpha-1}{\alpha}\cdot Z + Z \\ &\ge \frac{1}{\alpha}\cdot (g+Z). \end{aligned}$$

另一方面，

$$\begin{aligned} g'+Z &\le \alpha\cdot g + \gamma + Z \\ &\le \alpha\cdot g + (\alpha-1)Z + Z \\ &\le \alpha\cdot (g+Z). \end{aligned}$$

□

定理 *6.5*（平滑直方图算法框架 *Braverman* 和 *Ostrovsky* [2007]）。设 $\eta \in (0, 0.5)$。设 $g(\cdot)$ 是一个 (ζ,β)-平滑函数。若存在一个流式算法 Λ，能够以至少 $1-\xi$ 的概率，同时对所有时间戳 $t\in[T]$ 维护 $g(\cdot)$ 的一个$(\frac{1}{1-\eta})$-近似，且使用空间 $h(\eta,\xi)$，则存在一个滑动窗口算法 Λ'，能够以至少 $1-\xi$ 的概率，同时对所有时间戳 $t\in[T]$，在滑动窗口上维护 $g(\cdot)$ 的一个$\left(\frac{1}{1-\eta-\zeta}\right)$-近似，且使用空间 $O\left(\frac{\log T}{\beta}\cdot h(\eta,\xi\beta/\log(T))\right)$。

此外，在任意时间戳 $t\in[T]$，Λ' 都会启动一个新的 Λ 实例，该实例将数据流中的第 t 个元素视为数据流的起点，而 Λ' 最多只保留 $O(\log(T)/\beta)$ 个过去的 Λ 实例（这些实例从不同的时间戳开始）。在任意时间戳 t，Λ' 的输出仅取决于其所维护实例 Λ的输出；是否将某个实例 Λ 保留到时间戳 $t+1$，也仅取决于其所维护实例 Λ 在时间戳 t 的输出。

我们能够将上述平滑直方图框架扩展到差分隐私持续发布的场景。

定理 *6.6*（差分隐私持续发布模型的平滑直方图）。设 $g(\cdot)$ 是一个 (ζ,β)-平滑函数。若存在一个 ε'-差分隐私流式持续发布算法 Λ，能够以至少 $1-\xi$ 的概率，同时对所有时间戳 $t\in[T]$ 维护 $g(\cdot)$ 的 $\left(\frac{1}{1-\eta}\right)$-近似，且使用空间 $h(\eta,\xi)$，则存在一个 ε-差分隐私滑动窗口持续发布算法Λ'，其 $\varepsilon = O(\varepsilon'\beta/\log(T))$，能够以至少 $1-\xi$ 的概率，对所有时间戳 $t\in[T]$ 的滑动窗口维护 $g(\cdot)$ 的 $\left(\frac{1}{1-\eta-\zeta}\right)$-近似，并使用空间 $O\left(\frac{\log T}{\beta}\cdot h(\eta,\xi\beta/\log(T))\right)$。

证明。 近似保证、空间保证和成功概率可直接由定理 6.5 得出。在证明的剩余部分，我们证明差分隐私保证。

设 $\Lambda_1, \Lambda_2, \cdots, \Lambda_T$ 是 Λ 的实例，其中 Λ_t 于时间戳 t 启动。设 $o_{t,1}, o_{t,2}, \cdots, o_{t,T}$ 是 Λ_t 的输出；如果在时间戳 j， Λ_t 尚未启动，或已被 Λ' 踢出，那么 $o_{t,j} = \perp$。根据定理 6.5， Λ' 在所有时间戳 $t \in [T]$ 上的输出由 $\{o_{i,j} \mid i, j \in [T]\}$ 决定。因此，我们只需证明 $\{o_{i,j} \mid i, j \in [T]\}$ 满足 ε-差分隐私。考虑两个相邻数据流 $\mathcal{S}$ 和 $\mathcal{S}'$，它们只有第 r 个元素不同。固定一个可能的配置 $\{o_{i,j} \mid i, j \in [T]\}$。根据定理 6.5，至多存在 $O(\log(T)/\beta)$ 个不同的 $i \in [T]$，使得 $o_{i,r} \neq \perp$。设这样的 i 集合为 I。由于对于 $i \in I$，每个 Λ_i 在流式持续发布设置下都满足 ε'-差分隐私，因此我们有：

$$\begin{aligned}\Pr\left[\forall i \in I, j \in [T], o_{i,j}(\mathcal{S}) = o_{i,j}\right] &\leq \exp(\varepsilon' \cdot |I|) \cdot \Pr\left[\forall i \in I, j \in [T], o_{i,j}(\mathcal{S}') = o_{i,j}\right] \\ &\leq \exp(\varepsilon) \cdot \Pr\left[\forall i \in I, j \in [T], o_{i,j}(\mathcal{S}') = o_{i,j}\right].\end{aligned}$$

另一方面，我们有：

$$\begin{aligned}&\Pr\left[\forall i \notin I, j \in [T], o_{i,j}(\mathcal{S}) = o_{i,j} \mid o_{i,j}(\mathcal{S}) = o_{i,j} \forall i \in I, j \in [T]\right] \\ =&\Pr\left[\forall i \notin I, j \in [T], o_{i,j}(\mathcal{S}') = o_{i,j} \mid o_{i,j}(\mathcal{S}') = o_{i,j} \forall i \in I, j \in [T]\right]\end{aligned}$$

因此，该算法满足 ε-差分隐私。 □

结合定理 3.3、引理 6.3、引理 6.4 和定理 6.6，我们可以得到以下用于非负数求和的滑动窗口持续发布算法。

推论 6.7 (非负数的滑动窗口求和)。 令 $\eta \in (0, 0.5)$ $\varepsilon \geq 0$， $\xi \in (0$， *...?*

通过结合推论 4.11、引理 6.3、引理 6.4 和定理 6.6，我们能够得到以下用于不同元素数量的滑动窗口持续发布算法。

推论 *6.8*（滑动窗口持续发布不同元素）。对于 $\eta \in (0, 0.5)$ $\varepsilon \geq 0, \xi \in (0, 0.5)$，存在一个 ε-差分隐私算法，用于在滑动窗口持续发布模型中计算元素全集为 $\mathcal{U}$ 的数据流中的不同元素数量。以至少 $1-\xi$ 的概率，对于每个时间戳 $t \in [T]$，输出始终是一个 $(1+\eta, O\left(\log^2(T/(\eta\xi \quad T \quad \eta^4\varepsilon\right))) \log(\,)$-近似。该算法使用 $poly\left(\log(T/\xi)\,\eta\, min(\varepsilon, 1)\right)$ 空间。

结合定理 5.4、引理 6.3、引理 6.4 和定理 6.6，我们可以得到以下用于 ℓ_2 频率矩的滑动窗口持续发布算法。

推论 *6.9*（滑动窗口持续发布 ℓ_2 频率矩）。设 $\varepsilon > 0, \eta \in (0, 0.5)$ $\xi \in (0, 0.5)$。在滑动窗口持续发布模型中，存在一个 ε-差分隐私算法，使得以至少 $1-\xi$ 的概率，对每个满足 $|\hat{}F_2 - \|\mathcal{S}\|_2^2| \leq \eta\|\mathcal{S}\|_2^2 + O\left((\log(T/(\xi\eta)) + \log(|\mathcal{U}|))^2 \log^2(T)\, \varepsilon^2\eta^8 \quad \cdot \log^5(T) \cdot \log^2\left(\log(T/\xi) + \quad |\mathcal{U}| \quad \xi\eta\right)\right) \log(\,)$ 的时间戳 $t \in [T]$，它始终输出 $\hat{}F_2$。其中，$\mathcal{S}$ 表示时间戳 t 时对应于最新 W 个元素的子数据流。该算法使用 $O\left(\frac{\log(T/(\xi\eta)) + \log(|\mathcal{U}|)}{\eta^4} \cdot \log^2(T)\right)$ 空间。

通过结合定理 5.31、引理 6.3、引理 6.4 和定理 6.6，我们可以得到以下用于 ℓ_p 频率矩的滑动窗口持续发布算法。

推论 6.10（滑动窗口持续发布 ℓ_p 频率矩）。令 $p > 0, \varepsilon \geq 0, \xi \in (0, 0。5)$ $\eta \in (0, 0。5)$。存在一个 ε-差分隐私算法，在滑动窗口持续发布模型中，以至少 $1-\xi$ 的概率始终输出一个对 $\|\mathcal{S}\|_p^p$ 的 $\left(1+\eta, \left(\frac{\log(T|\mathcal{U}|/\xi}{\eta\varepsilon}\right)^{O(p)}\right)$-近似，其中 $\mathcal{S}$ 表示时间戳为 t 时对~~应于最新~~ W 个元素的子流。该算法使用的空间至多

$$\phi \cdot \left(\frac{\log(T|\mathcal{U}|/\xi)}{\eta\varepsilon}\right)^{O(p)},$$

其中 $\phi = \max(1, |\mathcal{U}|^{1-2/p})$。

参考文献

Naman Agarwal 和 Karan Singh。在线学习中差分隐私的代价。载于 Doina Precup 和 Yee Whye Teh 编，《第*34*届国际机器学习会议论文集》，第 70 卷，收录于《机器学习研究论文集》，第 32–40 页。PMLR，2017年8月6–11日。网址：https://proceedings.mlr.press/v70/agarwal17a.html。

Noga Alon、Yossi Matias 和 Mario Szegedy。频率矩近似的空间复杂度。载于第二十八届 *ACM*计算理论研讨会论文集，第 20–29 页，1996 年。

Alexandr Andoni。通过最大稳定性计算高阶频率矩。载于*2017* 年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 6364–6368 页。IEEE，2017 年。

Alexandr Andoni、Robert Krauthgamer 和 Krzysztof Onak。通过精确采样实现流算法。载于 *2011*年 *IEEE* 第 *52* 届计算机科学基础年会，第 363–372 页。IEEE，2011年。

Ziv Bar-Yossef、Thathachar S Jayram、Ravi Kumar 和 D Sivakumar。针对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 年。

Mihir Bellare 和 John Rompel。随机性高效的无偏采样。载于第*35*届计算机科学基础年会论文集，第276–287页。IEEE，1994年。

Jeremiah Blocki、Avrim Blum、Anupam Datta 和 Or Sheffet。Johnson–Lindenstrauss变换本身保留差分隐私。载于第*53*届*IEEE*计算机科学基础年会，*FOCS 2012*，美国新泽西州新不伦瑞克，*2012*年*10*月*20*日至*23*日，第410–419页。IEEE计算机学会，2012年。doi：10.1109/FOCS.2012.67。URL：https://doi.org/10.1109/FOCS.2012.67。

Jeremiah Blocki、Elena Grigorescu、Tamalika Mukherjee 和 Samson Zhou。如何使你的近似算法具有隐私性：一种针对低敏感度函数的可调近似算法的黑盒差分隐私变换。*arXiv* 预印本 *arXiv:2210.03831*，2022。

Jean Bolot、Nadia Fawaz、S. Muthukrishnan、Aleksandar Nikolov 和 Nina Taft。数据流上的私有衰减谓词和。载于第*16*届国际数据库理论会议论文集，ICDT '13，第284–295页，纽约州纽约，美国，2013年。计算机协会。ISBN 9781450315982。doi: 10.1145/2448496.2448530。URL https://doi.org/10.1145/2448496.2448530。

Vladimir Braverman和Rafail Ostrovsky。滑动窗口的平滑直方图。发表于第*48*届*IEEE*计算机科学基础年度研讨会（*FOCS'07*），第283–293页，2007年。doi：10.1109/FOCS.2007.55。

Zhiqi Bu、Sivakanth Gopi、Janardhan Kulkarni、Yin Tat Lee、Judy Hanwen Shen和Uthaipon Tantipongpipat。通过JL投影实现快速且节省内存的差分隐私随机梯度下降。发表于Marc'Aurelio Ranzato、Alina Beygelzimer、Yann N. Dauphin、Percy Liang和Jennifer Wortman Vaughan编的神经信息处理系统进展第*34*卷：*2021*年神经信息处理系统年会（*NeurIPS 2021*），*2021*年*12*月*6*日至*14*日，线上举行，第19680–19691页，2021年。URL https://proceedings.neurips.cc/paper/2021/hash/a3842ed7b3d0fe3ac263bcabd2999790-摘要。HTML。

T-H Hubert Chan、Mingfei Li、Elaine Shi 和 Wenchang Xu。《分布式数据流中重频项的差分隐私持续监测》。载于隐私增强技术研讨会，第 140–159 页。施普林格，2012。Moses Charikar、Kevin Chen 和 Martin Farach-Colton。《在数据流中查找频繁项》。载于自动机、语言与编程国际学术讨论会，第 693–703 页。施普林格，2002。Seung Geol Choi、Dana Dachman-Soled、Mukul Kulkarni 和 Arkady Yerukhimovich。《面向大规模统计量的差分隐私多方草图》。密码学电子印本档案，论文 2020/029，2020。网址：https://eprint.iacr.org/2020/029。https://eprint.iacr.org/2020/029。Graham Cormode 和 Shan Muthukrishnan。《改进的数据流摘要：计数最小草图及其应用》。算法杂志，55(1):58–75，2005。Mayur Datar、Aristides Gionis、Piotr Indyk 和 Rajeev Motwani。《在滑动窗口上维护数据流统计量》。*SIAM*计算杂志，31(6):1794–1813，2002。Marianne Durand 和 Philippe Flajolet。《大基数的 LogLog 计数》。载于欧洲算-法研讨会，第 605–617 页。施普林格，2003。Cynthia Dwork。《差分隐私：研究成果综述》。载于计算模型理论与应用国际会议，第 1–19 页。施普林格，2008。Cynthia Dwork、Moni Naor、Toniann Pitassi 和 Guy N. Rothblum。《持续观察下的差分隐私》。载于 Leonard J. Schulman（编），第*42*届*ACM*计算理论研讨会论文集，*STOC*，第 715–724 页。ACM，2010a。Cynthia Dwork、Moni Naor、Toniann Pitassi、Guy N Rothblum 和 Sergey Yekhanin。《全程隐私数据流算法》。载于计算机科学，第 66–80 页，2010b。

Cynthia Dwork、Aaron Roth 等。《差分隐私的算法基础》。《理论计算机科学基础与趋势》，9(3–4):211–407，2014。Cynthia Dwork、Moni Naor、Omer Reingold 和 Guy N. Rothblum。《通过私有划分实现矩形查询的纯差分隐私》。载于第*21*届密码学进展国际会议论文集第二部分——*ASIACRYPT 2015*，第*9453*卷，第 735–751 页，柏林，海德堡，2015。施普林格出版社。ISBN 9783662487990。doi: 10.1007/978-3-662-48800-3_30。网址：https://doi.org/10.1007/978-3-662-48800-3_30。Hendrik Fichtenberger、Monika Henzinger 和 Wolfgang Ost。《持续观察下图的差分隐私算法》。载于 Petra Mutzel、Rasmus Pagh 和 Grzegorz Her-man（编），第*29*届欧洲算法年会（*ESA 2021*），*2021*年*9*月*6*日至*8*日，葡萄牙里斯本（线上会议），第204卷*LIPIcs*，第 42:1–42:16 页。达格施图尔城堡—莱布尼茨信息学中心，2021。doi: 10.4230/LIPIcs.ESA.2021.42。网址：https://doi.org/10.4230/LIPIcs.ESA.2021.42。Philippe Flajolet 和 G Nigel Martin。《数据库应用的概率计数算法》。《计算机与系统科学杂志》，31(2):182–209，1985。Philippe Flajolet、Éric Fusy、Olivier Gandouet 和 Frédéric Meunier。《HyperLogLog：近乎最优基数估计算法的分析》。载于《离散数学与理论计算机科学》，第 137–156 页。《离散数学与理论计算机科学》，2007。Sumit Ganguly。《高阶频率矩的多项式估计量》。*arXiv*预印本 *arXiv:1104.4552*，2011。Hsiang Hsu、Natalia Martinez、Martin Bertran、Guillermo Sapiro 和 Flavio P. Calmon。《从统计、信息与估计理论角度看隐私的综述》。*IEEE BITS*信息论杂志，1(1):45–56，2021。T-H. Hubert Chan、Elaine Shi 和 Dawn Song。《统计数据的隐私与持续发布》。载于 Samson Abramsky、Cyril Gavoille、Claude Kirchner、Friedhelm Meyer auf der Heide 和 Paul G. Spirakis（编），自动机、语言与编程，第 405–417 页，柏林，海德堡，2010。施普林格柏林海德堡。ISBN 978-3-642-14162-1。Piotr Indyk。《稳定分布、伪随机生成器、嵌入与数据流计算》。《*ACM*学报》（*JACM*），53(3):307–323，2006。Piotr Indyk 和 David Woodruff。《数据流频率矩的最优近似》。载于第*37*届*ACM*计算理论年会论文集，第 202–208 页，2005。Palak Jain、Sofya Raskhodnikova、Satchit Sivakumar 和 Adam D. Smith。《持续观察下差分隐私的代价》。*CoRR*，abs/2112.00828，2021。网址：https://arxiv.org/abs/2112.00828。Prateek Jain、Pravesh Kothari 和 Abhradeep Thakurta。《差分隐私在线学习》。载于 Shie Mannor、Nathan Srebro 和 Robert C. Williamson（编），第*25*届学习理论年会论文集，第 23 卷《机器学习研究论文集》，第 24.1–24.34 页，爱丁堡，苏格兰，2012年6月25–27日。PMLR。网址：https://proceedings.mlr.press/v23/jain12.html。

Hossein Jowhari、Mert Sağlam 和 Gábor Tardos。lp采样器的紧确界、在数据流中查找重复项及相关问题。载于第三十届 *ACM SIGMOD-SIGACT-SIGART* 数据库系统原理研讨会论文集，第49–58页，2011年。Daniel M Kane、Jelani Nelson 和 David P Woodruff。关于草图化和流式处理小范数的精确空间复杂度。载于第二十一届 *ACM-SIAM*离散算法研讨会论文集，第1161–1178页。SIAM，2010。Daniel M Kane、Jelani Nelson、Ely Porat 和 David P Woodruff。以最优空间快速估计数据流中的矩。载于第四十三届 *ACM*计算理论研讨会论文集，第745–754页，2011年。Ping Li。使用稳定随机投影估计 l α ($0< \alpha \leq 2$) 中降维的估计器与尾界。载于第十九届 *ACM-SIAM*离散算法研讨会论文集，第10–19页，2008。Yi Li 和 David P Woodruff。小误差高频矩估计的紧确下界。载于近似、随机化与组合优化：算法与技术，第623–638页。施普林格，2013。Darakhshan Mir、Shan Muthukrishnan、Aleksandar Nikolov 和 Rebecca N Wright。通过草图统计实现全知者隐私算法。载于第三十届 *ACM SIGMOD-SIGACT-SIGART* 数据库系统原理研讨会论文集，第37–48页，2011年。Jayadev Misra 和 David Gries。查找重复元素。计算机程序科学，2（2）：143–152，1982年。Robert Morris。在小型寄存器中计数大量事件。《*ACM*通讯》，21(10)：840–842，1978年。Victor Perrier、Hassan Jameel Asghar 和 Dali Kaafar。实值数据流的隐私持续发布。载于第*26*届年度网络与分布式系统安全研讨会（*NDSS 2019*），圣迭戈，加利福尼亚州，美国，*2019*年*2*月*24*日至*27*日。互联网协会，2019。网址：https://www.ndss-symposium.org/ndss-paper/private-continual-release-of-real-val

Michael Saks 和 Xiaodong Sun。数据流模型中距离近似的空间下界。载于第*26*届年度 *ACM* 计算理论研讨会论文集，第 360–369 页，2002 年。Or Sheffet。差分隐私普通最小二乘法。载于 Doina Precup 和 Yee Whye Teh 编，第*34*届国际机器学习会议，*ICML 2017*，悉尼，新南威尔士州，澳大利亚，*2017*年*8*月*6*日至*11*日论文集，《机器学习研究论文集》第 70 卷，第 3105–3114 页。PMLR，2017 年。网址：http://proceedings.mlr.press/v70/sheffet17a.html。Adam Smith 和 Abhradeep Thakurta。面向全信息和赌博机设定下隐私在线学习的（近乎）最优算法。载于神经信息处理系统国际会议第*26*届论文集——第*2*卷，NIPS'13，第 2733–2741 页，美国纽约州雷德胡克，2013 年。Curran Associates Inc。

Adam D. Smith、Shuang Song 和 Abhradeep Thakurta。Flajolet-Martin 草图本身即可保持差分隐私：以最小空间进行隐私计数。载于 Hugo Larochelle、Marc'Aurelio Ranzato、Raia Hadsell、Maria-Florina Balcan 和 Hsuan-Tien Lin 编，神经信息处理系统进展 *33*：神经信息处理系统年度会议 *2020，NeurIPS 2020，2020*年*12*月*6*日至*12*日，线上，2020 年。网址：https://proceedings.neurips.cc/paper/2020/hash/e3019767b1b23f82883c9850356b71d6-Abstract.html。

Shuang Song、Susan Little、Sanjay Mehta、Staal Vinterbo 和 Kamalika Chaudhuri。差分隐私图统计量的持续发布，2018年。网址：https://arxiv.org/abs/1809.02575。

Robert H. Morris Sr. 使用小型寄存器计数大量事件。*Commun. ACM*，21(10)：840–842，1978年。

Mikkel Thorup 和 Yin Zhang。基于制表的 4-通用哈希及其在二阶矩估计中的应用。载于 *SODA*，第 4 卷，第 615–624 页，2004。

Jalaj Upadhyay。滑动窗口模型下的亚线性空间隐私算法。载于 Kamalika Chaudhuri 和 Ruslan Salakhutdinov 编，第 *36* 届国际机器学习会议论文集，第 97 卷《机器学习研究论文集》，第 6363–6372 页。PMLR，2019 年 6 月 9–15 日。网址：https://proceedings.mlr.press/v97/upadhyay19a.html。

Lun Wang、Iosif Pinelis 和 Dawn Song。具有多对数空间的差分隐私分数阶频率矩估计。载于第十届国际学习表征会议，*ICLR 2022*，线上会议，*2022*年*4*月*25*日至*29*日。OpenReview.net，2022年。网址：https://openreview.net/forum?id=7I8LPkcx8V。

第3节中遗漏的细节

A.1 引理 3.1 的证明

证明。我们首先证明，输出分组 G_1、G_2、$\cdots$、G_m 是 ε_0-差分隐私的。令 O_1、O_2、$\cdots$、O_m 为任意固定的分组。令 c'_1、c'_2、$\cdots$、c'_T 为任意相邻数据流，即 $\exists q \in [T]$，满足 $|c_q - c'_q| \leq 1$ 且 $\forall j \neq q$、$c_j = c'_j$。令 G'_1、G'_2、$\cdots$、$G'_{m'}$ 为该相邻数据流的输出分组。假设对于某个 $r \in [m]$有 $q \in O_r$。考虑 $\Pr[(G_1, G_2, \cdots, G_m) = (O_1, O_2, \cdots, O_m)]$ 和 $\Pr[(G'_1, G'_2, \cdots, G'_{m'}) = (O_1, O_2, \cdots, O_m)]$（在此情形下 $m' = m$）。我们有：

$$
\begin{aligned}
&\frac{\Pr[(G_1, G_2, \cdots, G_m) = (O_1, O_2, \cdots, O_m)]}{\Pr[(G'_1, G'_2, \cdots, G'_m) = (O_1, O_2, \cdots, O_m)]} \\
=&\frac{\Pr[(G_1, G_2, \cdots, G_{r-1}) = (O_1, O_2, \cdots, O_{r-1})]}{\Pr\left[(G'_1, G'_2, \cdots, G'_{r-1}) = (O_1, O_2, \cdots, O_{r-1})\right]} \cdot \frac{\Pr[G_r = O_r \mid (G_1, G_2, \cdots, G_{r-1}) = (O_1, O_2, \cdots, O_{r-1})]}{\Pr\left[G'_r = O_r \mid (G'_1, G'_2, \cdots, G'_{r-1}) = (O_1, O_2, \cdots, O_{r-1})\right]} \\
&\cdot \frac{\Pr[(G_{r+1}, G_{r+2}, \cdots, G_m) = (O_{r+1}, O_{r+2}, \cdots, O_m) \mid (G_1, G_2, \cdots, G_r) = (O_1, O_2, \cdots, O_r)]}{\Pr\left[(G'_{r+1}, G'_{r+2}, \cdots, G'_m) = (O_{r+1}, O_{r+2}, \cdots, O_m) \mid (G'_1, G'_2, \cdots, G'_r) = (O_1, O_2, \cdots, O_r)\right]}
\end{aligned}
$$

由于 $\forall j \in O_1 \cup O_2 \cup \cdots \cup O_{r-1}, c_j = c'_j$，在前缀 $O_1 \cup O_2 \cup$ 上运行算法 1 的行为 $\cdots \cup O_{r-1}$对 c_1、c_2、$\cdots$、c_T 的运行行为，与对其前缀 $O_1 \cup O_2 \cup \cdots \cup O_{r-1}$（即 c'_1、c'_2、$\cdots$、c'_T）运行时的行为相同。因此，我们有 $\Pr[(G_1,G_2,\cdots,G_{r-1})=(O_1,O_2,\cdots,O_{r-1})] \Pr[(G'_1,G'_2,\cdots,G'_{r-1})=(O_1,O_2,\cdots,O_{r-1})] = 1$。类似地，由于 $\forall j \in$

$O_{r+1} \cup O_{r+2} \cup \cdots \cup O_m$、$c_j = c'_j$，当 $G_r = G'_r = O_r$ 时，对 c_1 的后缀 $O_{r+1} \cup O_{r+2} \cup \cdots \cup O_m$（即 c_2、$\cdots$、c_T）运行算法 1 的行为，与对 c'_1、c'_2、$\cdots$、c'_T 的相同后缀运行时的行为相同。因此，我们有 $\Pr[(G_{r+1}, G_{r+2}, \cdots, G_m) = (O_{r+1}, O_{r+2}, \cdots, O_m) \mid (G_1, G_2, \cdots, G_r) = (O_1, O_2, \cdots, O_r)] \Pr[(G'_{r+1}, G'_{r+2}, \cdots, G'_m) = (O_{r+1}, O_{r+2}, \cdots, O_m) \mid (G'_1, G'_2, \cdots, G'_r) = (O_1, O_2, \cdots, O_r)]$。因此，我们有：

$$\begin{aligned}
&\frac{\Pr\left[(G_1, G_2, \cdots, G_m) = (O_1, O_2, \cdots, O_m)\right]}{\Pr\left[(G'_1, G'_2, \cdots, G'_m) = (O_1, O_2, \cdots, O_m)\right]} \\
=&\frac{\Pr\left[G_r = O_r \mid (G_1, G_2, \cdots, G_{r-1}) = (O_1, O_2, \cdots, O_{r-1})\right]}{\Pr\left[G'_r = O_r \mid (G'_1, G'_2, \cdots, G'_{r-1}) = (O_1, O_2, \cdots, O_{r-1})\right]}.
\end{aligned}$$

假设 $O_r = \{x+1, \cdots, x+k\}$。

考虑 $r \neq m$ 的第一种情形。在这种情形下，我们有

$$\begin{aligned}
&\Pr\left[G_r = O_r \mid (G_1, G_2, \cdots, G_{r-1}) = (O_1, O_2, \cdots, O_{r-1})\right] \\
=&\Pr\left[G_r = O_r \mid G_{r-1} = O_{r-1}\right] \\
=&\Pr\left[\left(\forall j \in [k-1], \nu_{x+j} + \sum_{b=1}^{j} c_{x+b} < \tau_r\right) \bigwedge \left(\nu_{x+k} + \sum_{b=1}^{k} c_{x+b} \geq \tau_r\right)\right].
\end{aligned}$$

现在，令 ν_{x+1}、ν_{x+2}、$\cdots$、ν_{x+k-1} 固定，并令 $g = \max_{j \in [k-1]} \nu_{x+j} + \sum_{b=1}^{j} c_{x+b}$。于是，

$$\begin{aligned}
&\Pr_{\nu_{x+k}, \tau_r}\left[G_r = O_r \mid G_{r-1} = O_{r-1}\right] \\
=&\Pr_{\nu_{x+k}, \tau_r}\left[\tau_r \in (g, \nu_{x+k} + \sum_{b=1}^{k} c_{x+b}]\right] \\
=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\nu_{x+k}}(v) \cdot p_{\tau_r}(\tau) \cdot \mathbf{1}\left(\tau \in (g, v + \sum_{b=1}^{k} c_{x+b})\right) \mathrm{d}v\mathrm{d}\tau
\end{aligned} \tag{3}$$

其中，$p_{\nu_{x+k}}(\cdot)$ 和 $p_{\tau_r}(\cdot)$ 分别是 ν_{x+k}和 τ_r的概率密度函数。令 $g' = \max_{j \in [k-1]} \nu_{x+j} + \sum_{b=1}^{j} c'_{x+b}$。令 $v' = v + g - g' + \sum_{b=1}^{k} c'_{x+b} - \sum_{b=1}^{k} c_{x+b}$。令 $\tau' = \tau + g - g'$。由于 $|c_q - c'_q| \leq 1$，容易看出 $|v' - v| \leq 2$和 $|\tau - \tau'| \leq 1$。注意，$\mathrm{d}v' = \mathrm{d}v$且 $\mathrm{d}\tau' = \mathrm{d}\tau$。因此，公式（3）等于以下内容：

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\nu_{x+k}}(v') \cdot p_{\tau_r}(\tau') \cdot \mathbf{1}\left(\tau + g - g' \in \left(g, v + g - g' + \sum_{b=1}^{k} c'_{x+b}\right)\right) \mathrm{d}v\mathrm{d}\tau \\
=&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\nu_{x+k}}(v') \cdot p_{\tau_r}(\tau') \cdot \mathbf{1}\left(\tau \in \left(g', v + \sum_{b=1}^{k} c'_{x+b}\right)\right) \mathrm{d}v\mathrm{d}\tau \\
\leq&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(\varepsilon_0/2) \cdot p_{\nu_{x+k}}(v) \cdot \exp(\varepsilon_0/2) \cdot p_{\tau_r}(\tau) \cdot \mathbf{1}\left(\tau \in \left(g', v + \sum_{b=1}^{k} c'_{x+b}\right)\right) \mathrm{d}v\mathrm{d}\tau \\
=&\exp(\varepsilon_0) \cdot \Pr_{\nu_{x+k}, \tau_r}\left[\tau_r \in (g', \nu_{x+k} + \sum_{b=1}^{k} c'_{x+b}]\right] \\
=&\exp(\varepsilon_0) \cdot \Pr_{\nu_{x+k}, \tau_r}\left[G'_r = O_r \mid G'_{r-1} = O_{r-1}\right]
\end{aligned}$$

$$= \exp(\varepsilon_0) \cdot \Pr_{\nu_{x+k}, \tau_r} \left[G'_r = O_r \mid (G'_1, G'_2, \cdots, G'_{r-1}) = (O_1, O_2, \cdots, O_{r-1}) \right].$$

接下来，考虑第二种情形，其中 $r = m$。在这种情形下，我们有

$$\begin{aligned} & \Pr\left[G_r = O_r \mid (G_1, G_2, \cdots, G_{r-1}) = (O_1, O_2, \cdots, O_{r-1}) \right] \\ = & \Pr\left[G_r = O_r \mid G_{r-1} = O_{r-1} \right] \\ = & \Pr\left[\forall j \in [k], \nu_{x+j} + \sum_{b=1}^{j} c_{x+b} < \tau_r \right]. \end{aligned}$$

现在，固定 $\nu_{x+1}, \nu_{x+2}, \cdots \nu_{x+k}$，并令 $g = \max_{j \in [k]} \nu_{x+j} + \sum_{b=1}^{j} c_{x+b}$。令 $g' = \max_{j \in [k]} \nu_{x+j} + \sum_{b=1}^{j} c'_{x+b}$。由于 $|c_q - c'_q| \leq 1$，我们有 $|g - g'| \leq 1$。于是，

$$\Pr_{\tau_r}[\tau_r > g] \leq \exp(\varepsilon_0) \cdot \Pr_{\tau_r}[\tau_r > g'].$$

因此，我们有

$$\begin{aligned} & \Pr\left[G_r = O_r \mid (G_1, G_2, \cdots, G_{r-1}) = (O_1, O_2, \cdots, O_{r-1}) \right] \\ \leq & \exp(\varepsilon_0) \cdot \Pr\left[G'_r = O_r \mid (G'_1, G'_2, \cdots, G'_{r-1}) = (O_1, O_2, \cdots, O_{r-1}) \right]. \end{aligned}$$

因此，我们可以得出结论：$(G_1, G_2, \cdots, G_m)$ 始终满足 ε_0 差分隐私。

注意，对于任意固定的 $(O_1, O_2, \cdots, O_m)$，并在满足 $(G_1, G_2, \cdots, G_m) = (O_1, O_2, \cdots, O_m)$ 的条件下，$\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_T$ 根据拉普拉斯机制满足 ε_0-差分隐私。因此，根据组合定理，输出流 $\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_T$ 满足 ε-差分隐私。□

A.2 引理 3.2 的证明

令 G_1、G_2、$\cdots$、G_m 为算法 1 生成的组。令 $\tilde{c}_1$、$\tilde{c}_2$、$\cdots$、$\tilde{c}_m$ 为 $\forall i \in [m]$、$\tilde{c}_i = \hat{c}_{\max(G_i)}$，即 $\tilde{c}_i$ 是组 G_i 的带噪计数。

引理 A.1. 以至少 $1-\xi$ 的概率，算法 *1* 的输出流满足以下性质：

1. $\forall i \in [m-1], \sum_{j \in G_i \setminus \{\max_{j' \in G_i} j'\}} c_j \leq \frac{7}{\eta \varepsilon_0} \cdot \ln(3 \cdot T/\xi) + \frac{13}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi)$.
2. $\sum_{j \in G_m} c_j \leq \frac{7}{\eta \varepsilon_0} \cdot \ln(3 \cdot T/\xi) + \frac{13}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi)$.
3. $\forall i \in [m-1], (1-\eta) \sum_{j \in G_i} c_j \leq \tilde{c}_i \leq (1+\eta) \sum_{j \in G_i} c_j$.

证明。令 $\mathcal{E}$ 表示事件

1. $\forall i \in [m], \left| \tau_i - \left(\frac{1}{\eta} + 1 \right) \cdot \frac{7}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi) \right| \leq \frac{2}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi)$.
2. $\forall t \in [T], |\nu_t| \leq \frac{4}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi)$. 3. $\forall i \in [m-1], |\tilde{c}_i - \sum_{j \in G_i} c_j| \leq \frac{1}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi)$

根据拉普拉斯噪声的累积分布函数，很容易证明，通过对所有 $i \in [m], t \in [T]$ 使用并集界， $\mathcal{E}$ 以至少 $1-\xi$ 的概率发生。在证明的剩余部分中，我们以事件 $\mathcal{E}$ 为条件。

考虑性质1。任取 $i \in [m-1]$。令 $t = \max_{j \in G_i} j - 1$。则有$\nu_t + \sum_{j \in G_i,\ j \le t} c_j \le \tau_i$。由于$|\nu_t| \le \frac{4}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi)$ 且 $\tau_i \le \frac{7}{\eta \cdot \varepsilon_0} \cdot \ln(3 \cdot T/\xi) + \frac{9}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi)$，因此有 $\sum_{j \in G_i,\ j \le t} c_j \le \frac{7}{\eta \varepsilon_0} \cdot \ln(3 \cdot T/\xi) + \frac{13}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi)$。

性质2的证明与性质1的证明相同。

考虑性质3。任取 $i \in [m-1]$。令 $t = \max_{j \in G_i} j$。则有 $\nu_t + \sum_{j \in G_i} c_j \ge \tau_i$。由于 $|\nu_t| \le \frac{4}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi)$ 且 $\tau_i \ge \frac{7}{\eta \cdot \varepsilon_0} \cdot \ln(3 \cdot T/\xi) + \frac{5}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi)$，因此有 $\sum_{j \in G_i} c_j \ge \frac{7}{\eta \varepsilon_0} \cdot \ln(3 \cdot T/\xi)$。注意，$|\tilde{c}_i - \sum_{j \in G_i} c_j| \le \frac{1}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi)$。因此，我们知道$\sum_{j \in G_i} c_j \ge \frac{7}{\eta \varepsilon_0} \cdot \ln(3 \cdot T/\xi)$。因此，$\forall i \in [m-1]$ 有 $(1-\eta) \sum_{j \in G_i} c_j \le \tilde{c}_i \le (1+\eta) \sum_{j \in G_i} c_j$。 □

现在，我们可以证明引理 3.2。

证明　引理 *3.2* 的证明。以至少 $1-\xi$ 的概率，引理 A.1 中列出的性质成立 .

令 $l' \in [m]$ 为满足 $\max_{j \in G_{l'}} j \ge l$ 的最小索引，并令 $r' \in [m]$ 为满足 $\max_{j \in G_{r'}} j \le r$ 的最大索引。

我们有:

$$\begin{aligned}
\sum_{j=l}^{r} \hat{c}_j = & \sum_{j'=l'}^{r'} \tilde{c}_{j'} \\
\ge & (1-\eta) \sum_{j'=l'}^{r'} \sum_{j \in G_{j'}} c_j - \left(\frac{7}{\eta \varepsilon_0} \cdot \ln(3 \cdot T/\xi) + \frac{13}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi) \right) \\
\ge & (1-\eta) \sum_{j=l}^{\max_{b \in G_{r'}} b} c_j - \left(\frac{7}{\eta \varepsilon_0} \cdot \ln(3 \cdot T/\xi) + \frac{13}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi) \right) \\
\ge & (1-\eta) \sum_{j=l}^{r} c_j - \frac{7}{\eta \varepsilon_0} \cdot \ln(3 \cdot T/\xi) - \frac{26}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi),
\end{aligned}$$

其中，第一个不等式由引理 A.1 的性质3和性质2得出，第二个不等式由 l' 的选取方式得出，第三个不等式由引理 A.1 的性质1得出。

同理，我们可以证明

$$\begin{aligned}
\sum_{j=l}^{r} \hat{c}_j = & \sum_{j'=l'}^{r'} \tilde{c}_{j'} \\
\le & (1+\eta) \sum_{j'=l'}^{r'} \sum_{j \in G_{j'}} c_j + \left(\frac{7}{\eta \varepsilon_0} \cdot \ln(3 \cdot T/\xi) + \frac{13}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi) \right) \\
\le & (1+\eta) \sum_{j=\min_{b \in G_{l'}} b}^{r} c_j + \left(\frac{7}{\eta \varepsilon_0} \cdot \ln(3 \cdot T/\xi) + \frac{13}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi) \right) \\
\le & (1+\eta) \sum_{j=l}^{r} c_j + \frac{7}{\eta \varepsilon_0} \cdot \ln(3 \cdot T/\xi) + \frac{26}{\varepsilon_0} \cdot \ln(3 \cdot T/\xi),
\end{aligned}$$

其中，第一个不等式由引理 A.1 的性质3和性质2推出，第二个不等式由 r' 的选取推出，第三个不等式则由引理 A.1 的性质1推出。 □