arXiv:2302.05707v1 [cs.CR] 11 Feb 2023

关于有界空间下的差分隐私与自适应数据分析

Itai Dinur* Uri Stemmer† David P. Woodruff‡ Samson Zhou§

2023年2月11日

摘要

我们研究差分隐私与自适应数据分析这两个相关领域的空间复杂度。具体而言，

1. 在标准密码学假设下，我们证明存在一个问题 P：与不考虑隐私时所需的空间相比，采用差分隐私高效解决该问题需要多出指数级的空间。据我们所知，这是隐私算法与非隐私算法之间关于空间复杂度的首次分离。

2. 自适应数据分析这一研究方向致力于理解：回答一系列自适应查询所需的样本数量。我们从基础层面重新审视以往的下界，并表明这些下界源于空间瓶颈，而非采样瓶颈。

为获得我们的结果，我们定义并构造了一种具有多个密钥的加密方案，该方案旨在以一种非常特殊的方式抵御有限程度的密钥泄漏。

1 引言

查询到通信提升定理使我们能够将给定函数的查询复杂度下界 f 转化为相关函数$\hat{}f$的通信复杂度下界。自 Raz 和 McKenzie 的开创性工作以来 [74], ，人们提出并应用了若干此类提升定理，在各种设定下获得了新的通信复杂度下界。

在密码学领域，人们也取得了相关成果，其出发点是在理想化模型（例如随机预言机模型）[8]中，对手解决密码分析问题所需查询复杂度的下界。随后，该查询复杂度下界被

*本-古里安大学。dinuri@bgu.ac.il。部分研究经费由以色列科学基金会（项目编号 1903/20）和欧洲研究理事会 ERC 起步基金协议（编号 757731，LightCrypt）资助。

†特拉维夫大学和 Google Research。u@uri.co.il。部分经费由以色列科学基金会（项目编号 1871/19）以及 Len Blavatnik 和 Blavatnik 家族基金会提供。

‡卡内基梅隆大学。dwoodruf@andrew.cmu.edu。部分工作是在访问 Google Research 期间完成的。另获西蒙斯研究员奖以及美国国家科学基金会 CCF-1815840 号项目资助。§加州大学伯克利分校和莱斯大学。samsonzhou@gmail.com。部分工作是在卡内基梅隆大学任职期间完成的。部分经费由西蒙斯研究员奖以及美国国家科学基金会 CCF-1815840 号项目提供。

提升为查询-空间下界，用于刻画解决同一问题的非均匀（预处理）对手 [84, 25, 24]。

基于这些研究方向中发展出的思想，我们提出了一种新技术，可将差分隐私和自适应数据分析背景下问题的采样下界转化为空间下界。在介绍我们的结果之前，我们先说明所研究问题设定的动机。

1.1 差分隐私

差分隐私 [35] 是一种隐私的数学定义，旨在支持对数据集进行统计分析，同时强有力地保证个体级信息不会泄露。通俗地说，如果一个分析数据的算法足够稳健，即其结果分布不会过度依赖任何单个数据点，那么该算法就满足差分隐私。形式上，

定义 1.1（[35]）。令 $\mathcal{A} : X^* \to Y$ 为一个随机化算法，其输入为数据集$D \in X^*$。若对于任意两个仅相差一个数据点的数据集 D, D'（此类数据集称为相邻数据集），以及任意结果集合 $F \subseteq Y$，均有 $\Pr[\mathcal{A}(D) \in F] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(D') \in F] + \delta$.，则算法 $\mathcal{A}$ 满足（ε, δ）-差分隐私（DP）。

为了解释这个定义，设 D 是一个包含 n 个数据点的数据集，每个数据点代表一个人的信息。假设 Alice 知道除其中一个数据点（比如 Bob 的数据点）之外的所有数据点。现在，假设我们计算 $z \leftarrow \mathcal{A}(D)$，并将 z 提供给 Alice。如果 $\mathcal{A}$ 是差分隐私的，那么 Alice 几乎无法了解到 Bob 的数据点，因为无论 Bob 的数据点是什么，z 的分布都大致相同。

过去几年中，我们见证了差分隐私在各种场景下的研究呈爆发式增长。特别是，一系列富有成果的研究专注于设计空间复杂度较低的差分隐私算法，主要应用于流式处理场景。这一方向的研究包括 [36,64, 9, 5, 14, 1, 59, 78, 69, 70, 13, 87, 10, 18, 85, 86, 90]。这些研究都取得了积极成果，并针对各种问题提出了空间复杂度较低的差分隐私算法。事实上，其中一些研究表明，经典流式算法基本上无需修改，本身就具有差分隐私性。例如，Blocki 等人 [9] 表明，Johnson-Lindenstrauss 变换本身就能保护差分隐私；Smith 等人 [78] 则证明，经典的 Flajolet-Martin 草图也具有这一性质。

鉴于这些积极结果，人们可能会认为，小空间算法特别适合差分隐私，因为这些算法本来就不会保留太多关于输入的信息。

问题 1.2。差分隐私是否需要更多空间?

1.1.1 我们关于有界空间下差分隐私的结果

我们对问题1.2 给出肯定回答，即我们证明，差分隐私可能需要更多空间。为此，我们提出一个问题：在非隐私情况下，只需少量空间即可解决，但在隐私情况下则需要大量空间。首先，我们来考察下面这个玩具问题，它为上述问题提供了一些答案。

一个无趣的玩具问题。回忆一下， F_2 （数据流的第二频率矩）估计的乘法近似误差为 $1+\alpha$ ，其空间下界为 $\Omega(1/\alpha^2)$ [88]。这立即表明，对于“输出数据流的最后一个元素，或数据流$1+\alpha$值的一个 F_2 -近似”这一问题，存在分离。在非隐私设置中，输出最后一个元素所需的空间与 α无关；但在隐私设置中，算法必须（以隐私方式）估计 F_2 ，因此至少需要 $1/\alpha^2$ 空间。当然，我们也可以用F_2 替换为标准非隐私模型中具有较大空间下界的其他任务。

我们认为这个玩具问题并不有趣，因为从宏观来看，我们的目标是说明：在某些情况下，私密地计算某个结果所需的空间，远远大于非私密地计算“同一结果”所需的空间。然而，在这个玩具问题中，私密算法和非私密算法可以说并非在计算“同一结果”。为解决这个问题，我们将关注由一个函数 （其取值范围为某个度量空间）定义的问题，而所需完成的任务是近似该函数的值。请注意，这种表述排除了玩具问题作为问题1.2的有效答案；并且在这种表述下，从形式上说，每个用于解决该任务的算法都必须计算（或近似）“同一结果”。

让我们更精确地描述所考虑的设定。为了简化表述，我们不研究流式模型，而是关注以下计算模型。[1] 考虑一个在数据集 D 上实例化的算法，之后该算法旨在回答关于 D的查询。如果在获得查询之前，该算法将 D 缩减为一个至多包含 s 比特的摘要 z ，我们就称该算法使用的空间为 s 。随后，在收到查询时，算法仅使用该摘要z （无法额外访问原始数据集 D）来回答查询。形式上，我们考虑由函数 $P\colon X^* \times Q \to M$定义的问题，其中 X 是数据域， Q 是可能查询的族，而 M 是度量空间。

定义 1.3. 我们称 $\mathcal{A}=(\mathcal{A}_1,\mathcal{A}_2)$ 以空间复杂度 s、样本复杂度 n、误差 α 和置信度 β 求解问题 $P\colon X^*\times Q\to M$，如果

1. $\mathcal{A}_1\colon X^*\to\{0,1\}^s$ 是一个预处理过程，它接收一个数据集 D ，并输出一个 s 比特的字符串。
2. 对于每个输入数据集 $D\in X^n$ 和每个查询 $q\in Q$ ，都有

$$\Pr_{\substack{z\leftarrow\mathcal{A}_1(D)\\ a\leftarrow\mathcal{A}_2(z,q)}}[|a-P(D,q)|\le\alpha]\ge 1-\beta.$$

我们给出以下定理。

定理 1.4（非正式表述）。令 $d\in\mathbb{N}$ 为控制问题规模的参数（具体而言，来自 X 的数据点可用 $\mathrm{polylog}(d)$ 比特表示，而来自 Q 的查询可用 $\mathrm{poly}(d)$ 比特表示）。存在一个问题 $P\colon X^*\times Q\to M$，使得以下结论成立。

1. P 可以通过非隐私方式，使用$\mathrm{polylog}(d)$ 比特的空间求解。% 形式化表述见引理 *3.2* 。*2.* P 可以通过隐私方式求解，所需样本复杂度和空间复杂度为 $\tilde{O}(\sqrt{d})$。% 形式化表述见引理 *3.4* 。

[1]不过，我们指出，我们的所有结果都可以推广到流式处理场景。参见备注 3.3。

3. 假设存在一种次指数安全的对称密钥加密方案，那么对于求解 $\mathcal{A}$ 问题的每个计算高效的差分隐私算法 P，其空间复杂度都必须为 $\tilde{\Omega}(\sqrt{d})$，即使其样本复杂度是 d 的一个较大多项式。此外，即使只要求 $\mathcal{A}$ 满足计算意义下的差分隐私（也就是说，我们构造的针对 $\mathcal{A}$ 的对手在计算上是高效的），这一结论仍然成立。% 形式化表述见推论 *3.13* 。

请注意，这在 d 中体现为非隐私空间复杂度与隐私空间复杂度之间的指数分离。我们强调，私密地解决 P 的难点并非源于样本不足。事实上，根据条目 2，$\tilde{O}(\sqrt{d})$ 个样本足以私密地解决这个问题。然而，条目 3 指出，除非算法使用大空间，否则即使样本数量远超所需，也无法私密地解决这个问题。

据我们所知，这是首个区分隐私算法与非隐私算法空间复杂度的结果。诚然，我们为证明上述定理而定义的问题 P 有些不自然。相比之下，我们关于自适应数据分析的否定性结果（接下来将对此进行综述）针对的是文献中研究的典型问题。

1.2 自适应数据分析

设想一位数据分析师想要检验某个特定的研究假设。分析师获取相关数据，对假设进行评估，并（例如）得出该假设不成立的结论。根据这些发现，分析师随后决定检验第二个假设，并使用同一批数据对其进行评估（获取新数据可能成本过高，甚至根本不可能）。也就是说，分析师会自适应地选择假设，而这种选择取决于此前与数据的交互。因此，这些发现不再受到经典统计理论的支持；经典统计理论假设，在收集数据之前就已确定要检验的假设，而分析师也面临着对数据过拟合的风险。

从 [34], 自适应数据分析（ADA）这一研究方向开始，旨在设计能够在此类情境下可证明地保证统计有效性的方法。具体而言，目标是设计一种机制 $\mathcal{A}$，它首先获取一个数据集 D，其中包含 t 来自某个未知分布 $\mathcal{P}$的独立同分布样本，然后回答 k 自适应选择的查询，针对 $\mathcal{P}$。重要的是，$\mathcal{A}$ 的答案必须针对底层分布 $\mathcal{P}$准确，而不只是针对经验数据集 D。这里的核心问题是：

问题 1.5。 $\mathcal{A}$ 需要多少个样本（也就是说，t 应该是多少），才能支撑 k 这样的自适应查询?

为了应对最坏情况分析者，我们假设分析者具有对抗性，试图使机制失效。如果一种机制能够在面对这样的对抗性分析者时维持效用，那么它就能在面对任何分析者时维持效用。形式上，ADA 研究方向所探讨的典型问题被定义为机制 $\mathcal{A}$ 与对手$\mathcal{B}$之间的双人博弈。参见图1。

定义 1.6 ([34])。 机制 $\mathcal{A}$ 对于 (α, β)-准确地处理 k 定义域 X，使用样本量 t，如果对于每个对手 $\mathcal{B}$ (与 $\mathcal{A}$ 在图 *1*所指定的游戏中交互) 均满足

$$\Pr\left[\exists j \in [k] \ s.t. \ |z_j - h_j(\mathcal{P})| > \alpha\right] \leq \beta,$$

1. 对手 $\mathcal{B}$ 选择数据域 X上的分布 $\mathcal{P}$。 2. 机制 $\mathcal{A}$ 获取一个样本 $S \sim \mathcal{P}^t$，其中包含来自 $\mathcal{P}$ 的 t 个独立同分布样本。 3. 对于 k 轮次 $j = 1,2,\ldots,k$:

- 对手选择一个函数 $h_j : X \to \{-1, 0, 1\}$，该函数可以根据机制此前给出的所有答案来确定。
- 机制获取 h_j 并给出答案 z_j，然后将其提供给 $\mathcal{B}$。

图 1：机制 $\mathcal{A}$ 与对手 $\mathcal{B}$之间的双人博弈。

其中 $h_j(\mathcal{P}) = \mathbb{E}_{x\sim\mathcal{P}}[h_j(x)]$是函数 h_j在底层分布 $\mathcal{P}$上的“真实”值。

继 Dwork 等人 [34], 的工作之后，这个问题吸引了大量研究 [53, 81, 4, 68, 61, 57, 77, 76, 44, 45, 33, 75, 26, 43, 42, 3], 其中大多数研究都着重于了解自适应数据分析需要多少样本 （即聚焦于问题 1.5）。特别是，目前已知以下几乎匹配的界。

定理 1.7 ([34, 4])。 存在一种计算高效机制，对于 (0。1，0。1)- 准确的 k 个查询，所需样本量为 $t = \tilde{O}\left(\sqrt{k}\right)$。

定理 1.8（[53, 81], 非正式表述）。假设单向函数存在，则对于 k 个查询，任何具有计算效率且达到 (0.1, 0.1) 准确度的机制，其样本量至少为$t = \Omega\left(\sqrt{k}\right)$。

1.2.1 我们关于有界空间下自适应数据分析的结果

此前关于 ADA 问题的所有工作都将其视为抽样问题，传达出这样一个信息：“自适应数据分析比非自适应数据分析需要更多样本”。在本文中，我们从基础层面重新审视 ADA 问题，并提出：

问题 1.9。对于 ADA 问题而言，是否存在比样本数量更根本的瓶颈?

考虑一种机制 $\mathcal{A}$，它最初会获得底层分布 $\mathcal{P}$ 的完整描述，但必须将该描述压缩成摘要 z，其描述长度与 $\mathcal{P}$ 中 t 个样本的描述长度相同。（我们将 $\mathcal{A}$ 的空间复杂度视为 z 的比特数。）之后，$\mathcal{A}$ 需要使用 z 回答 k 个自适应查询，而不能再访问 $\mathcal{P}$。与只能从 $\mathcal{P}$ 获得 t 个样本的机制相比，这是否赋予 $\mathcal{A}$ 更强的能力?

我们证明，一般来说，答案是否定的。具体而言，我们证明如下定理。

定理 1.10 （定理 4.1 的非正式版本）。 假设单向函数存在，那么每个对 (0.1,0.1)-准确的机制（计算高效） k 查询的计算高效机制，其空间复杂度 至少为 $\Omega\left(\sqrt{k}\right)$。

事实上，在该定理的正式版本中（见定理 4.1），我们证明空间复杂度至少为 $\Omega\left(\sqrt{k}\right)$ 倍于域元素的表示长度。我们认为，这是对先前关于 ADA 问题的下界的重大加强：并不是机制没有获得足够多的关于 $\mathcal{P}$的信息，而是它无法以能够高效回答

自适应查询的方式压缩这些信息。这推广了 k 的否定性结果，因为抽样 [53, 81], 个 $t = \sqrt{k}$ 点来存储关于 $\mathcal{P}$ 的信息，只是其中一种特定方式 $\mathcal{P}$。

1.3 我们的方法

我们通过结合密码学、隐私、学习理论、通信复杂度和信息论等多个研究领域的技术，得出了我们的结果。

1.3.1 我们的方法：多实例抗泄漏方案

我们定义并构造的主要密码学工具，是一种适用的多密钥加密方案，其设计旨在以一种非常特定的方式抵御一定程度的密钥泄漏。具体而言，该方案包含 n 个底层加密方案的实例，各自使用独立密钥（每个密钥长度为 λ 比特）。这些密钥最初交给对手，对手将其压缩为一个包含 $s \ll n \cdot \lambda$ 比特的摘要 z。在此阶段之后，每个实例都会独立设置一个额外参数，该参数是公开的，但在初始阶段对对手而言未知。随后，对手获得了使用这些密钥加密的明文的加密结果。我们要求，在给定 z 和公开参数的情况下，除少数密钥对应的明文外（密钥数量取决于 s，但不取决于 n），使用每个密钥加密的明文在计算上仍然是隐藏的。

我们称该方案为多实例抗泄漏方案（或 MILR 方案），以强调这样一个事实：尽管对手的泄漏是关于所有 n 密钥的任意函数，但方案本身由 n 个完全独立的实例组成。

MILR 方案的效率由两个参数衡量：（1）在其下加密（可能）不安全的密钥数量；（2）安全参数 λ 的损失。我们构造的方案在这两个参数上都达到了最优，仅相差一个乘法常数因子。首先，除 $O\left(\frac{s}{\lambda}\right)$ 个密钥之外，所有密钥下的加密都仍然是隐藏的。这基本上是最优的，因为对手可以定义 z 来存储前 $\frac{s}{\lambda}$ 个密钥。其次，我们在安全参数 λ 上损失了一个常数因子。我们构造的另一个优势是，其内部参数不依赖于 s。如果允许这种依赖，那么在某些情形下（特别是当 $s \leq o(n \cdot \lambda)$ 时），就可以对方案进行微调，使效率参数的损失降为乘法 $1 + o(1)$；不过，这对我们的应用影响不大。

我们的构造可以说是最自然的方案。初始阶段之后，要使用 λ 位密钥加密明文，我们首先应用一个提取器（以公开参数作为种子）将其哈希为更短的密钥，再用这个密钥通过底层加密方案加密明文。

MILR 方案与抗泄漏密码学领域中发展起来的方案有关（参见 [17, 39, 40, 71, 79, 56, 58]），该领域通常使用随机性提取这一基本技术。然而，抗泄漏密码学主要研究密码系统抵御侧信道攻击的能力，而我们的模型并非旨在形式化对此类攻击的安全性，并且具有若干在该领域并不常见的性质（例如，以一种本质上会使其中某些实例不安全的方式保护加密方案的多个独立实例）。因此，抗泄漏密码学领域发展出的高级密码系统要么不适用于我们的目的，要么过于复杂或效率低下。

尽管我们的构造很简单，但要证明它具备所声称的抗泄漏安全性，过程却有些技术性。我们强调，我们不依赖针对特定问题的困难性假设，也不假设底层加密方案具有诸如抵御相关密钥攻击的特殊性质。相反，我们的证明基于 Unruh 提出的预采样技术 [84]，该技术用于证明密码系统在随机预言机模型下能够抵御非一致对手。最近，[25, 24, 32] 基于通信复杂度领域发展出的工具，对这一技术进行了改进和优化 [49, 62]。我们将该技术用于证明计算（而非信息论）情形下的安全性，这似乎要求假设底层加密方案能够抵御非一致对手（尽管这被视为一项标准假设）。

1.3.2 我们的方法：隐私需要更多空间

我们设计了一个问题，在非隐私条件下只需非常小的空间复杂度即可求解，但在隐私条件下却需要较大的空间复杂度。为此，我们将一个关于以隐私方式求解特定问题所需样本复杂度的已知否定性结果，提升为相关问题的空间下界。我们所研究的问题是所谓的、参数为 d 的单向边际问题；该问题存在采样下界。在这个问题中，我们的输入是一个数据集$D \in (\{0,1\}^d)^*$，其中包含一组二进制向量，每个向量的长度为 d。我们的目标是输出一个向量 $\vec{a} \in [0,1]^d$，使其在较小的 L_∞ 误差范围内近似输入向量的平均值，例如误差不超过 $1/10$。也就是说，我们希望向量 $\vec{a} \in [0,1]^d$ 满足：$\left\| \vec{a} - \frac{1}{|D|} \sum_{\vec{x} \in D} \vec{x} \right\|_\infty \leq \frac{1}{10}$.

我们称该问题的算法具有样本复杂度 n，如果对于每个大小为 n 的输入数据集，它都能以至少 0.9 的概率输出一个良好的解。差分隐私领域中最基础的结果之一表明，该问题需要一个很大的数据集：

定理 1.11（[16], 非正式表述）。任何用于求解参数为 d 的单向边缘问题的差分隐私算法，其样本复杂度都必须为 $n = \Omega(\sqrt{d})$。

为了将这一采样下界提升为空间下界，我们考虑这样一个问题：输入数据集包含 n 个密钥 $\vec{x} = (x_1, \ldots, x_n)$（从我们的 MILR 方案中抽样得到）。然后，算法必须将该数据集压缩成一个包含 s 比特的摘要 z。之后，算法会收到一个由一组 n 个密文指定的“查询”，其中每个密文都是一个 d 比特向量的加密结果。目标任务是近似明文输入向量的平均值。直观来说，如果算法的空间为 $s \ll \sqrt{d}$，那么根据我们的 MILR 方案的性质，它最多只能解密这些 d 比特向量中的 $\approx s \ll \sqrt{d}$ 个，因此它试图用少于 $\sqrt{d}$ 个样本来求解单向边缘问题。我们证明，可以将这一论证形式化，从而导出矛盾。

1.3.3 我们的方法：自适应数据分析关乎空间

正如我们所提到的，Hardt 和 Ullman [53] 以及 Steinke 和 Ullman [81] 证明了 ADA 问题需要很大的样本复杂度（见定理1.8）。具体而言，他们证明存在一个计算高效的对手，能够使每个高效机制在回答自适应查询时失败。回想一下，ADA 博弈（见图1）从对手选择底层分布开始。

我们将 [53, 81] 的否定性结果提升为 空间 下界。为此，我们设计了另一种对手：它首先为我们的 MILR 方案采样一大批 密钥，然后

将目标分布定义为在这些采样密钥上均匀分布。回想一下，在我们的设定中，机制会获得该目标分布的精确描述，随后必须将此描述压缩至 s 比特。然而，根据我们的 MILR 方案的性质，这意味着该机制只能解密至多对应于 $\approx s/\lambda$ 个密钥的密文。接着我们证明，在这些条件下，可以模拟 [53, 81] 的对手；在他们的设定中，“输入样本”对应于这样一组密钥索引：对于这些索引，机制能够解密相应的密文。

1.4 其他相关工作

查询到通信提升定理已被通信复杂度领域的许多研究采用，例如
[51, 49, 29, 22, 89, 52, 46, 21, 63, 55, 50, 20, 19]。如前所述，为了得到关于差分隐私算法的否定性结果，我们将 Bun 等人 [16] 针对单向边缘问题的采样下界提升为空间下界。差分隐私文献中还存在许多其他采样下界；这里我们只综述其中几项。

Dwork 等人 [37] 使用叛徒追踪方案证明了差分隐私的计算采样下界。Ullman [83]，使用指纹编码构造了一种新颖的叛徒追踪方案，并得到了更强的差分隐私计算采样下界，从而扩展了他们的结果。Ullman 的构造可视为单向边缘问题的一种加密变体。Bun 等人 [15] 和 Alon 等人 [2] 表明，存在一些简单的学习任务，解决这些任务所需的样本数在采用差分隐私时渐近地更多（与非隐私样本复杂度相比）。Feldman [41] 和 Brown 等人 [12] 表明，存在一些学习问题，其中每个近乎最优的学习算法（相对于其所需输入样本数达到近乎最优的误差）都必须记忆其输入数据的很大一部分。我们强调，所有这些结果与我们的结果有本质区别，因为它们讨论的是抽样，而不是空间。

Steinhardt 等人 [80] 询问，在附加空间约束的条件下学习某个概念类，是否存在样本复杂度上的分离，并猜想奇偶学习问题能够体现这种分离。Raz [73] 证明了该猜想：他表明，虽然奇偶学习可以使用 $O(n)$ 个样本和 $O(n^2)$ 空间解决，但任何使用 $o(n^2)$ 个样本的奇偶学习算法都需要指数数量的样本。随后， [60] 表明，在某些超线性空间约束下，学习线性规模的 DNF 公式、线性规模决策树和对数规模合取函数是不可行的。另有一系列活跃的研究进一步刻画了空间约束下学习概念类 [66, 72, 67, 6, 47, 27]和分布 [23, 30,31]的局限性。我们指出，尽管空间约束导致样本复杂度提高这一结论，在概念上与我们的自适应数据分析结果相似，但学习概念类这一问题并非准确响应自适应查询所固有的问题。

1.5 通信复杂度的应用

最后，我们的论证还给出了分布式单向通信复杂度下界；当目标是以极低的成功概率计算一个关系时，这些下界很有用。据我们所知，现有的查询到通信提升定理（例如 [74, 51, 49]）并未考虑我们在此处研究的问题和输入分布。粗略地说，我们证明：如果任何用于计算函数 f 的基于抽样的协议都需要 k 个样本 $(a_1, b_1) \ldots, (a_k, b_k)$，其中 $a_i \in \{0,1\}^t$ 对每个 $i \in [k]$ 都成立，那么任何计算

$f(A,B)$ 的单向协议在此设定下都必须使用 $\Omega(kt)$ 通信。

更准确地说，在双人单向通信博弈中，Alice 和 Bob 分别获得输入 A 和 B，目标是让 Alice 向 Bob 发送尽可能少的信息，使 Bob 能够计算某个预先指定的函数 f 所对应的 $f(A,B)$。协议 Π 的通信代价，是在所有可能的输入 A下，Alice 发送的最大消息长度（以比特计）；而（随机化）通信复杂度，则是成功概率至少为2 3的协议所需的最小通信代价。在分布式设定中，A 和 B 还会从一个已知的底层分布中抽取。

在我们的分布式设定中，假设 Alice 持有 m 个相互独立且均匀分布的数 $a_1,\ldots,a_m$，满足以下两种情况之一：对所有 $i\in[m]$，$a_i\in GF(p)$；或者对所有 $i\in[m]$，当 t 足够大时，$a_i\in GF(2^t)$。再假设 Bob 持有 m 个来自同一域、相互独立且均匀分布的数 $b_1,\ldots,b_m$，满足以下两种情况之一：$GF(p)$ 或 $GF(2^t)$。那么，对于任意函数 $f(\langle a_1,b_1\rangle,\ldots,\langle a_m,b_m\rangle)$，其中点积是在$GF(2)$上计算的；或者对于 $f(a_1\cdot b_1,\ldots,a_m\cdot b_m)$，其中乘积是在 $GF(p)$上计算的，计算 f 且成功概率为 σ 时的随机化单向通信复杂度，等于 Alice 需要向 Bob 发送的 $a_1,\ldots,a_m$ 中的样本数量，才能使 Bob 以概率$\sigma-\varepsilon$计算出 f。许多这类问题的采样下界都很容易证明，例如 $\sum_i a_i\cdot b_i \pmod p$ 或 MAJ$(\langle a_1,b_1\rangle,\ldots,\langle a_m,b_m\rangle)$，并且这些下界可立即转化为通信复杂度下界。这些下界的主要直觉是，数 $b_1,\ldots,b_m$ 可以看作 Bob 持有的哈希函数，因此，如果 Bob 仅持有这些数中的一个小子集，我们就可以应用剩余哈希引理的一个变体。完整细节见附录 E。

1.6 论文结构

本文其余部分的结构如下。MILR 方案在第 2 节中定义。关于差分隐私和自适应数据分析的结果分别在第 3 节和第 4 节中介绍。我们在第 5 节中构造 MILR 方案，并在第 6 节中证明其安全性。为便于阅读，这些章节中的部分技术细节放在附录 B、C 和 D 中。附录A 介绍密码学中的标准预备知识。最后，附录E 详细讨论单向协议中样本复杂度与通信复杂度之间的关系。

2 多实例抗泄漏方案

我们将多实例抗泄漏方案（或称 MILR 方案）定义为由以下高效算法组成的元组 (Gen, Param,Enc,Dec)，其语法如下：

- Gen 是一种随机化算法，输入安全参数 λ，输出一个 λ 位密钥。形式化地，$x\leftarrow$ Gen(1^λ)。
- Param 是一种随机化算法，输入安全参数 λ，输出一个 poly(λ) 位公开参数。形式化地，$p\leftarrow$ Param(1^λ)。
- Enc 是一种随机化算法，其输入为密钥 x、公开参数 p 和消息 $m\in\{0,1\}$，输出为密文 $c\in\{0,1\}^{\text{poly}(\lambda)}$。形式化地，$c\leftarrow$ Enc(x,p,m)。
- Dec 是一种确定性算法，其输入为密钥 x、公开参数 p 和密文 c，输出为解密后的消息 m'。如果密文 c 是.....的加密

m 使用密钥 x 和参数 p 加密得到，则 $m' = m$。形式化地说，如果 $c \leftarrow \mathsf{Enc}(x, p, m)$，那么 $\mathsf{Dec}(x, p, c) = m$ 的概率为 1。

为了定义 MILR 方案的安全性，设 $n \in \mathbb{N}$，令 $\vec{x} = (x_1, \ldots, x_n)$ 为密钥向量，并令 $\vec{p} = (p_1, \ldots, p_n)$ 为公开参数向量（通过调用 Param 为每个方案设置一次）。令 $J \subseteq [n]$ 为一个子集，称为 “ 隐藏坐标”。现在考虑一对预言机 $\mathcal{E}_1(\vec{x}, \vec{p}, J, \cdot, \cdot)$ 和 $\mathcal{E}_0(\vec{x}, \vec{p}, J, \cdot, \cdot)$，它们具有以下性质。

1. $\mathcal{E}_1(\vec{x}, \vec{p}, J, \cdot, \cdot)$ 接受一个密钥 $j \in [n]$ 的索引和一条消息 m 作为输入，并返回 $\mathsf{Enc}(x_j, p_j, m)$。2. $\mathcal{E}_0(\vec{x}, \vec{p}, J, \cdot, \cdot)$ 接受相同的输入。如果 $j \in J$，则结果为 $\mathsf{Enc}(x_j, p_j, 0)$，否则结果为 $\mathsf{Enc}(x_j, p_j, m)$。

定义 2.1. 设 λ 为安全参数。设 $\Gamma : \mathbb{R} \to \mathbb{R}$ 和 $\overline{\tau} : \mathbb{R}^2 \to \mathbb{R}$ 为实值函数。若满足以下条件，则 MILR 方案 (Gen, Param, Enc, Dec) 对空间受限预处理对手是 $(\Gamma, \overline{\tau})$-安全的。

(1) 多重语义安全：对于每个 $n = \mathrm{poly}(\Gamma(\lambda))$ 以及每个 $\mathrm{poly}(\Gamma(\lambda))$ 时间对手 $\mathcal{B}$，都存在一个可忽略函数 negl，使得

$$\left| \Pr_{\vec{x},\vec{p},\mathcal{B},\mathrm{Enc}} \left[\mathcal{B}^{\mathcal{E}_0(\vec{x},\vec{p},[n],\cdot,\cdot)}(\vec{p}) = 1 \right] - \Pr_{\vec{x},\vec{p},\mathcal{B},\mathrm{Enc}} \left[\mathcal{B}^{\mathcal{E}_1(\vec{x},\vec{p},[n],\cdot,\cdot)}(\vec{p}) = 1 \right] \right| \leq \mathrm{negl}(\Gamma(\lambda)).$$

也就是说，获得公开参数但无法获得密钥的计算受限对手，无法分辨自己是在与 $\mathcal{E}_0$ 交互，还是在与 $\mathcal{E}_1$ 交互。

(2) 针对有界预处理对手的多重安全性：对于每个 $n = \mathrm{poly}(\Gamma(\lambda))$、每个 $s \leq n \cdot \lambda$ 以及每个预处理过程 $F : \left(\{0,1\}^\lambda\right)^n \to \{0,1\}^s$（可能是随机化的），都存在一个随机函数 $J = J(F, \vec{x}, z, \vec{p}) \subseteq [n]$，该函数接收一组密钥 $\vec{x}$、公开参数 $\vec{p}$ 和一个元素 $z \leftarrow F(\vec{x})$，并返回一个大小为 $|J| \geq \tau := n - \overline{\tau}(\lambda, s)$ 的子集，使得对于每个 $\mathrm{poly}(\Gamma(\lambda))$ 时间算法 $\mathcal{B}$，都存在一个可忽略函数 negl 满足

$$\left| \Pr_{\substack{\vec{x},\vec{p},\mathcal{B},\mathrm{Enc} \\ z \leftarrow F(\vec{x}) \\ J \leftarrow J(F,\vec{x},z,\vec{p})}} \left[\mathcal{B}^{\mathcal{E}_0(\vec{x},\vec{p},J,\cdot,\cdot)}(z, \vec{p}) = 1 \right] - \Pr_{\substack{\vec{x},\vec{p},\mathcal{B},\mathrm{Enc} \\ z \leftarrow F(\vec{x}) \\ J \leftarrow J(F,\vec{x},z,\vec{p})}} \left[\mathcal{B}^{\mathcal{E}_1(\vec{x},\vec{p},J,\cdot,\cdot)}(z, \vec{p}) = 1 \right] \right| \leq \mathrm{negl}(\Gamma(\lambda)).$$

也就是说，即使我们的密钥有 s 比特泄露（由作用于密钥的预处理函数 F 计算得出），关于 J 的密钥进行的加密仍然是计算不可区分的。

注 2.2。当 Γ 为恒等函数时，我们简称该方案为 $\overline{\tau}$-安全。请注意，在这种情况下，该安全性可抵御所有运行时间关于安全参数 λ 为多项式的对手。我们还将假设存在一种次指数安全加密方案。具体而言，存在一个常数 $\nu > 0$，使得该方案对于 $\Gamma(\lambda) = 2^{\lambda^\nu}$ 是 $(\Gamma, \overline{\tau})$-安全的。也就是说，我们假设存在这样一种方案，其安全性可抵御所有运行时间关于 2^{λ^ν} 为多项式的对手。

在第 5 和 6 节中，我们证明以下定理。

定理 2.3. 设 $\Omega(\lambda) \leq \Gamma(\lambda) \leq 2^{o(\lambda)}$。如果存在一个针对非一致对手达到 $\Gamma(\lambda)$ 安全性的加密方案，那么就存在一个 MILR 方案，该方案能够抵御空间有界的非均匀预处理对手，并达到 $(\Gamma(\lambda), \overline{\tau})$ 安全性，其中 $\overline{\tau}(\lambda, s) = \frac{2s}{\lambda} + 4$。

3 差分隐私的空间困难性

考虑一个算法，它在数据集 D 上实例化，随后旨在回答关于 D 的查询。我们称此类算法的空间为 s，如果它在获取查询之前，将 D 压缩为摘要 z，其中最多包含 s 比特。随后，在获得查询时，算法仅使用摘要 z 来回答查询（不再访问原始数据集 D）。

设 (Gen, Param, Enc, Dec) 是一个安全参数为 λ 的 MILR 方案。在样本复杂度为 n 的 (λ, d)-解密平均值（DA）问题中，输入数据集包含 n 个密钥，即 $D = (x_1, \ldots, x_n) \in \left(\{0,1\}^\lambda\right)^n$。查询 $q = ((p_1, c_1) \ldots, (p_n, c_n))$ 由 n 对 (p_i, c_i) 指定，其中 p_i 是公开参数，c_i 是密文，它是一个长度为 d 的二进制向量的加密结果。目标是发布一个向量 $\vec{a} = (a_1, \ldots, a_d) \in [0,1]^d$，使其近似于“解密平均向量（dav）”，其定义为 $\mathrm{dav}_q(D) = \frac{1}{n}\sum_{i=1}^n \mathrm{Dec}(x_i, p_i, c_i)$。

定义 3.1. 设 $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ 是一个算法，其中 $\mathcal{A}_1 : \left(\{0,1\}^\lambda\right)^n \to \{0,1\}^s$ 是预处理过程，它将包含 n 个密钥的数据集概括为 s 比特；$\mathcal{A}_2$ 是“响应算法”，它接收 $\mathcal{A}_1(D)$ 的结果以及查询 q。如果输出是满足 $\|\vec{a} - \mathrm{dav}_q(D)\|_\infty \leq \frac{1}{10}$. 的向量 $\vec{a}$ 的概率至少为 $9/10$，则称 $\mathcal{A}$ 解决了 DA 问题。

不考虑隐私时，DA 问题几乎是平凡的。具体来说，

引理 3 .2. (λ, d)-DA 问题可以使用空间 $s = O(\lambda \log(d))$ 高效求解 .

证明。预处理算法 $\mathcal{A}_1$ 对输入密钥进行 $O(\log d)$ 次采样。随后，算法 $\mathcal{A}_2$ 接收查询 q，并使用采样密钥估计 dav 向量。根据 Chernoff 界即可得出该引理。

□

备注 3.3. 如前所述，为简化表述，在我们的计算模型中，我们将算法 $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ 的空间复杂度视为算法 $\mathcal{A}_1$ 输出的大小。不过，我们指出，我们的分离结果也适用于一种流式模型，在该模型中，$\mathcal{A}_1$ 和 $\mathcal{A}_2$ 都必须使用小空间。为此，请注意，在上述证明中，算法 $\mathcal{A}_1$ 本身已经使用小空间（而不只是输出长度小），因为它只从输入数据集中采样 $O(\log d)$ 个密钥。现在，我们分析算法 $\mathcal{A}_2$以流式方式读取查询 q 时的空间复杂度。回想一下，查询 q 包含 n 个公开参数 $p_1, \ldots, p_n$ 和 n 个密文 $c_1, \ldots, c_n$，其中每个c_i 都是一个 d 位向量的加密，记为 $y_i \in \{0,1\}^d$。为了使 $\mathcal{A}_2$ 能够以小空间读取 q，我们按如下顺序排列它：$q = (p_1, \ldots, p_n), (c_1,1, \ldots, c_n,1) \ldots, (c_1,d, \ldots, c_n,d)$ $q_0 \circ q_1 \circ \cdots \circ q_d$. 这里，$c_{i,j} = \mathrm{Enc}(x_i, p_i, y_i[j])$ 是使用密钥 x_i 和公开参数 p_i 对 j 的第 j 位进行加密。请注意，数据流的第一部分 q_0包含公开参数，之后每一部分 q_j 都包含所有 n个输入向量的第 j位的加密。按此顺序排列查询后，算法$\mathcal{A}_2$ 首先读取 q_0，并存储与 $\mathcal{A}_1$所存储的密钥对应的 $O(\log d)$ 个公开参数。然后，对于每个 $j \in [d]$，在读取 q_j时，算法 $\mathcal{A}_2$ 使用采样密钥估计第 j 个坐标的平均值。接着，算法 $\mathcal{A}_2$ 输出这个估计值，并继续处理下一个坐标。该过程可以用 $\mathrm{poly}(\lambda \log(d))$ 的空间复杂度实现。

因此，在没有隐私约束的情况下，DA 问题可以用小空间解决。现在我们证明，假设输入数据集足够大，DA 问题可以很容易地在大空间下以差分隐私方式解决。具体而言，

引理 3.4. 存在一种计算高效的 (ε, δ)-差分隐私算法，使用空间 $s = O\left(\frac{1}{\varepsilon} \cdot \sqrt{d \cdot \log(\frac{1}{\delta})} \cdot \lambda \cdot \log\ d\right)$ 求解 (λ, d)-DA 问题，前提是输入数据集的大小满足 $n = \Omega(s)$（足够大）。

证明。 预处理算法 $\mathcal{A}_1$ 抽取 $\approx \sqrt{d}$ 个密钥。根据差分隐私的标准组合定理 [38], 这足以使响应算法 $\mathcal{A}_2$ 能够以隐私方式近似目标向量的每个 d 坐标。

□

因此，DA 问题可以使用小空间以非隐私方式求解，也可以使用大空间以隐私方式求解。我们的下一个目标是证明，私密地求解这一问题确实需要大空间。在证明这一点之前，我们先介绍计算差分隐私和指纹编码方面的一些其他预备知识。

3.1 计算差分隐私和指纹编码的预备知识

Beimel 等人 [7] 和 Mironov 等人 [65]定义了计算差分隐私。设 $\mathcal{A}$ 是一个作用于数据集的随机化算法（机制）。计算差分隐私通过挑战者与对手之间的双人博弈来定义，双方运行一对算法 $(Q,\ T)$。博弈开始时，对手 Q 选择一对大小均为 n 的相邻数据集 $(D_0,\ D_1)$，以及一个任意字符串 r （我们认为它代表对手的内部状态）。随后，挑战者均匀抽取一位 b，并应用 $\mathcal{A}$ (D_b) 以获得一个结果 a。接着，在 a 上应用 $T(r,\ \cdot)$，并尝试猜出 b。形式化地，

定义 3.5。设 λ 为安全参数，ε 为常数，且 $\delta\colon\ \mathbb{R} \to \mathbb{R}$ 为一个函数。若对于每个 $n = \text{poly}$ (λ) 以及每个非均匀 poly(λ) 时间对手 $(Q,\ T)$，都存在一个可忽略函数 negl，使得随机化算法 $\mathcal{A}\colon\ X^* \to Y$ 是 $(\varepsilon,\ \delta)$-计算差分隐私的（CDP）：

$$\Pr_{\substack{(r,D_0,D_1)\leftarrow Q\\ \mathcal{A},T}}[T(r,\mathcal{A}(D_0))=1] \le e^{\varepsilon} \cdot \Pr_{\substack{(r,D_0,D_1)\leftarrow Q\\ \mathcal{A},T}}[T(r,\mathcal{A}(D_1))=1] + \delta(n) + \text{negl}(\lambda).$$

定义 3.6。设 ε 为常数，$\delta = \delta(\lambda)$ 为一个函数。给定两个概率分布族 $\mathcal{X} = \{X_\lambda\}_{\lambda\in\mathbb{N}}$ 和 $\mathcal{Y} = \{Y_\lambda\}_{\lambda\in\mathbb{N}}$，如果对于每个非一致概率多项式时间区分器 D，都存在一个可忽略函数 negl，使得 $\Pr_{x\leftarrow X_\lambda}[D(x) = 1] \le e^{\varepsilon} \cdot \Pr_{y\leftarrow Y_\lambda}[D(y) = 1] + \delta(\lambda) + \text{negl}(\lambda)$，反之亦然，则记作 $\mathcal{X} \approx_{\varepsilon,\delta} \mathcal{Y}$。

我们回顾指纹码的概念。Boneh 和 Shaw [11] 将其作为一种用于对数字内容进行水印的密码学工具引入。自 Bun、Ullman 和 Vadhan [16]，的工作以来，指纹编码在各种情形下证明差分隐私的下界方面发挥了关键作用。

（抗合谋）指纹码是一种方案，用于向 n 个用户分发码字 w_1、$\cdots$、w_n，并能唯一追溯到每个用户。此外，如果一组（至多 k 个）用户

将其码字合成为一个海盗码字 $\hat{w}$，那么仍然可以将该海盗码字追溯到参与合成它的某个用户。当然，如果不对海盗如何生成合成码字作任何假设，就不可能实现安全追踪。因此，根据标记假设对海盗施加限制：合成码字在每个位置上都必须与至少一个“真实”码字一致。也就是说，在索引 j 处，$w_i[j] = b$ 对每个 $i \in [n]$ 都成立时，海盗必须输出 $\hat{w}$，且 $\hat{w}[j] = b$ 也成立。[2]

定义 3.7 ([11, 82]). 一个 k-抗合谋指纹码，长度为 d，适用于 n 个用户，失败概率为 γ，或简称 (n, d, k, γ)-指纹码，是一对随机变量 $C \in \{0,1\}^{n\times d}$ 和Trace： $\{0,1\}^d \to 2^{[n]}$，满足以下条件。对于所有对手 P: $\{0,1\}^{k\times d} \to \{0,1\}^d$ 和，满足 $|S| = k$,

$$\Pr_{C,\mathrm{Trace},P}[(\forall 1 \le j \le d \ \ \exists i \in [n] \ s.t. \ P(C_S)[j] = c_i[j]) \wedge (\mathrm{Trace}(P(C_S)) = \emptyset)] \le \gamma,$$

and

$$\Pr_{C,\mathrm{Trace},P}[\mathrm{Trace}(P(C_S)) \cap ([n] \setminus S) \neq \emptyset] \le \gamma,$$

其中 C_S 包含由 S给出的 C 的各行。

注 3.8. 如前所述，条件 $\{\forall 1 \le j \le d \ \ \exists i \in [n]$ s.t. $P(C_S)[j] = c_i[j]\}$ 称为“标记假设”。第二个条件称为“误指控小概率”。因此，如果对手 P 保证其输出满足标记假设，那么以至少 $1 - 2\gamma$ 的概率，算法 Trace 会输出一个索引 $i \in S$。

定理 3.9（[11, 82, 81]）。对于每个 $1 \le k \le n$，都存在一种面向 n 个用户、长度为 $d = O(k^2 \cdot \log n)$ 的 k-抗合谋指纹码，其失败概率为 $\gamma = \frac{1}{n^2}$，且具有可高效计算的 Trace 函数。

我们指出，存在满足定理 3.9所述保证的自适应和非自适应指纹编码构造；我们采用非自适应版本。

3.2 DA 问题的一个否定性结果

我们关于空间受限差分隐私算法的主要否定性结果如下。

定理 3.10。设 Π = (Gen, Param, Enc, Dec) 是一个安全参数为 λ、能够抵抗空间受限预处理对手的 $(\Gamma, \overline{\tau})$-安全 MILR 方案。设 $d \le \mathrm{poly}(\Gamma(\lambda))$ 和 $n \le \mathrm{poly}(\Gamma(\lambda))$ 是 λ 的函数。设 ε 为常数，且 $\delta \le \frac{1}{4n(e^\varepsilon+1)} = \Theta(\frac{1}{n})$。对于任何解决 (λ, d)-DA 问题、运行时间为 $\mathrm{poly}(\Gamma(\lambda))$ 的 (ε, δ)-计算差分隐私算法，其样本复杂度为 n、空间复杂度为 s，均有 $\overline{\tau}(\lambda, s) = \Omega\left(\sqrt{\frac{d}{\log n}}\right)$.

证明。设 $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ 是一个多项式$(\Gamma(\lambda))$ 时间的计算差分隐私算法，用于解决 (λ, d)-DA 问题，其样本复杂度为 $n = \mathrm{poly}(\Gamma(\lambda))$，空间复杂度为 s。记 $\overline{\tau} = \overline{\tau}(\lambda, s)$，并反设 $\overline{\tau} = O\left(\sqrt{\frac{d}{\log n}}\right)$（足够小）。我们构造以下对手 $\mathcal{B}$，用于对抗一个$(n+1, d, \overline{\tau}, \frac{1}{n^2})$-指纹码（根据定理 3.9 及反设，这种码必定存在）。

[2]我们采用 [81], 给出的标记假设表述，这与通常考虑的表述略有不同。

1. 输入为 n 个码字 $w_1, \ldots, w_n \in \{0,1\}^d$。2. 从 Gen($1^\lambda$) 中抽取 n 个密钥 $x_1, \ldots, x_n \sim$。3. 令 $z \leftarrow \mathcal{A}_1(x_1, \ldots, x_n)$。4. 从 Param($1^\lambda$) 中抽取 n 个公开参数 $p_1, \ldots, p_n \sim$。5. 对于 $i \in [n]$，令 $c_i \leftarrow \text{Enc}(x_i, p_i, w_i)$。6. 令 $\vec{a} \leftarrow \mathcal{A}_2(z, (p_1, c_1) \ldots, (p_n, c_n))$。7. 将其坐标舍入至 $\{0,1\}$ 后，输出 $\vec{a}$。

我们将 $\mathcal{B}$ 视为指纹码的对手。实际上，它的输入是一组码字，输出是一个长度为 d 的二进制向量。注意，如果 $\mathcal{A}$ 解决了 DA 问题（即近似解密后的平均向量），那么对于每个坐标，$\mathcal{B}$ 的结果都必须与至少一个输入码字一致；换言之，它满足标记假设（参见注释 3.8）。

注释 3.11。在继续给出正式证明之前，我们先概述其结构（本注释可以忽略，正式证明并不需要它）。简单来说，我们将证明：

(1) 算法 $\mathcal{B}$ 相对于 码字 集合而言是计算差分隐私的（尽管我们对 $\mathcal{A}$ 的假设是，它相对于 密钥 而言具有隐私性）。(2) 利用 MILR 方案的性质，我们将证明 $\mathcal{B}$ 实际上必须忽略其绝大多数输入，至多使用 $\overline{\tau}$ 个码字。这意味着 $\mathcal{B}$ 实际上是一个只处理 $\overline{\tau}$ 个码字（而非它收到的 n 个码字作为输入）的指纹码对手。(3) 差分隐私领域中的一个已知结论（始于 *[16]*）指出，成功的指纹码对手不可能满足差分隐私，因为这会与追踪算法能够恢复其某个输入点这一事实相矛盾。我们在此处的关键在于，$\mathcal{B}$ （实际上）只使用 $\overline{\tau}$ 个码字，因此，为了导出矛盾，只需使用码字长度短得多的指纹码（长度仅为 $\approx \overline{\tau}^2$，而非 $\approx n^2$）。这意味着 DA 问题的难度取决于 $\mathcal{A}$ 的空间（它控制着$\overline{\tau}$），而非输入大小（即 n）。

回顾一下，$\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ 关于密钥满足计算差分隐私。我们首先证明，$\mathcal{B}$ 关于码字满足计算差分隐私。为此，设 Q 是一个对手（如定义 3.5 所述），它输出一对相邻数据集 $(\vec{w}, \vec{w'})$，每个数据集都包含 n 个码字，并附带一个状态 r。给定 $(\vec{w}, \vec{w'})$，我们用 $\ell = \ell(\vec{w}, \vec{w'}) \subseteq [n]$ 表示 $\vec{}$ 这个索引，在该索引处 $\vec{w}, w'$ 不同。我们还用 x_0 表示另一个密钥，它独立于算法 $\mathcal{B}$ 抽样得到的密钥 $x_1, \ldots, x_n$。根据算法 $\mathcal{A}$ 的隐私保证以及加密方案的语义安全性

（参见定义 2.1）可知

$$
\begin{aligned}
&\langle r, \mathcal{B}(\vec{w})\rangle \equiv \\
&\quad \equiv \langle r, \mathcal{A}_2\left(\mathcal{A}_1(x_1, ..., x_\ell, ..., x_n), \vec{p}, \mathrm{Enc}(x_1, p_1, w_1), ..., \mathrm{Enc}(x_\ell, p_\ell, w_\ell), ..., \mathrm{Enc}(x_n, p_n, w_n)\right)\rangle \\
&\approx_{(\varepsilon,\delta)} \langle r, \mathcal{A}_2\left(\mathcal{A}_1(x_1, ..., x_0, ..., x_n), \vec{p}, \mathrm{Enc}(x_1, p_1, w_1), ..., \mathrm{Enc}(x_\ell, p_\ell, w_\ell), ..., \mathrm{Enc}(x_n, p_n, w_n)\right)\rangle \\
&\quad \equiv_c \langle r, \mathcal{A}_2\left(\mathcal{A}_1(x_1, ..., x_0, ..., x_n), \vec{p}, \mathrm{Enc}(x_1, p_1, w_1), ..., \mathrm{Enc}(x_\ell, p_\ell, w'_\ell), ..., \mathrm{Enc}(x_n, p_n, w_n)\right)\rangle \\
&\approx_{(\varepsilon,\delta)} \langle r, \mathcal{A}_2\left(\mathcal{A}_1(x_1, ..., x_\ell, ..., x_n), \vec{p}, \mathrm{Enc}(x_1, p_1, w_1), ..., \mathrm{Enc}(x_\ell, p_\ell, w'_\ell), ..., \mathrm{Enc}(x_n, p_n, w_n)\right)\rangle \\
&\quad \equiv \langle r, \mathcal{B}(\vec{w'})\rangle.
\end{aligned}
$$

因此，算法 $\mathcal{B}$ 是 $(2\varepsilon, (e^\varepsilon + 1)\delta)$-计算差分隐私的。现在考虑算法 $\mathcal{B}$ 的以下变体，记为 $\mathcal{B}$。相对于 $\mathcal{B}$ 的修改以红色标出。

1. 输入为 n 个码字 $w_1, \ldots, w_n \in \{0,1\}^d$。2. 采样 n 个密钥 $x_1, \ldots, x_n \sim \mathsf{Gen}(1^\lambda)$。3. 令 $z \leftarrow \mathcal{A}_1(x_1, \ldots, x_n)$。4. 采样 n 个公开参数 $p_1, \ldots, p_n \sim \mathsf{Param}(1^\lambda)$。5. 令 $J \leftarrow J(\mathcal{A}_1, \vec{x}, z, \vec{p})$ 为定义 2.1 保证存在的坐标子集，其大小为 $|J| = n - \overline{\tau}$。6. 对于 $i \in J$，令 $c_i \leftarrow \mathsf{Enc}(x_i, p_i, 0)$。7. 对于 $i \in [n] \setminus J$，令 $c_i \leftarrow \mathsf{Enc}(x_i, p_i, w_i)$。8. 令 $\vec{a} \leftarrow \mathcal{A}_2(z, (p_1, c_1) \ldots, (p_n, c_n))$。9. 输出 $\vec{a}$，将其坐标四舍五入到 $\{0,1\}$。

ˆ 注 3.12. 注意，算法 $\mathcal{B}$ 不一定是计算高效的，因为计算 $J(\mathcal{A}_1, \vec{x}, z, \vec{p})$ 可能并不高效。尽管如此，正如我们接下来要展示的，这仍足以ˆ 得出矛盾并完成下界的证明。具体而言，我们将证明 $\mathcal{B}$是计算差分隐私的（相对于码字而言），并且它是针对指纹码的成功对手。即使 $\mathcal{B}$ 本身是一个非高效机制，这也会导致矛盾。

现在我们证明，根据 MILR 方案的多重安全性（见定义 2.1），$\mathcal{B}$ 和 $\mathcal{B}$ 的结果分布在计算上不可区分。具体而言，我们要证明，对于定义 3.5中的每个高效对手 (Q, T)，都有

$$
|\Pr[T(r, \mathcal{B}(\vec{w})) = 1] - \Pr[T(r, \hat{\mathcal{B}}(\vec{w})) = 1]| \leq \mathrm{negl}(\Gamma(\lambda)).
$$

注意，这里两个表达式使用的是同一个数据集 $\vec{w}$（不含相邻数据集 $\vec{w'}$）。为证明这一点，考虑以下算法，记作 $\mathcal{W}$，我们将其视为 MILR 方案的对手。该算法只能通过预言机 $\mathcal{E}$访问加密预言机。

1. $\mathcal{W}$的输入为 n 个码字 $w_1, \ldots, w_n \in \{0,1\}^d$、一个元素 z（据称由$\mathcal{A}_1$计算得出），以及一组 n 公共参数 $p_1, \ldots, p_n$。
2. 对于 $i \in [n]$ 令 $c_i \leftarrow \mathcal{E}(i, w_i)$。

3. 令 $\vec{a} \leftarrow \mathcal{A}_2(z, (p_1, c_1) \ldots, (p_n, c_n))$.

4. 输出 $\vec{a}$，将其坐标舍入到 $\{0, 1\}$。

现在，根据 MILR 方案的多重安全性，我们有

$$\left| \Pr_{Q,T,\mathcal{B}}[T(r, \mathcal{B}(\vec{w})) = 1] - \Pr_{Q,T,\mathcal{B}}[T(r, \hat{\mathcal{B}}(\vec{w})) = 1] \right|$$

$$= \left| \Pr_{\substack{\vec{x},\vec{p},\mathcal{W},Q,T,\mathrm{Enc} \\ z \leftarrow \mathcal{A}_1(\vec{x}) \\ J \leftarrow J(\mathcal{A}_1,\vec{x},z,\vec{p})}} [Q(r, \mathcal{W}^{\mathcal{E}_1(\vec{x},\vec{p},J,\cdot,\cdot)}(\vec{w}, z, \vec{p})) = 1] - \Pr_{\substack{\vec{x},\vec{p},\mathcal{W},Q,T,\mathrm{Enc} \\ z \leftarrow \mathcal{A}_1(\vec{x}) \\ J \leftarrow J(\mathcal{A}_1,\vec{x},z,\vec{p})}} [Q(r, \mathcal{W}^{\mathcal{E}_0(\vec{x},\vec{p},J,\cdot,\cdot)}(\vec{w}, z, \vec{p})) = 1] \right|$$

$$\leq \mathrm{negl}(\Gamma(\lambda)).$$

ˆ 因此，我们有 $\mathcal{B} \equiv_c \mathcal{B}$，并且 $\mathcal{B}$ 是 $(2\varepsilon, (e^{\varepsilon} + 1)\delta)$-计算差分隐私。ˆ 因此，算法 $\mathcal{B}$ 也是 $(2\varepsilon, (e^{\varepsilon} + 1)\delta)$-计算差分隐私（相对于输入码字）。注意，算法 $\mathcal{B}$ 忽略了除 $N - |J| = \overline{\tau}$ 之外的所有码字；此外，选择忽略哪些码字与码字本身无关。现在考虑以下思想实验。

1. 为指纹码采样一个码本 $w_0, w_1, \ldots, w_n$。ˆ2. 在 $(w_1, \ldots, w_n)$ 上运行 $\mathcal{B}$。ˆ 3. 对 $\mathcal{B}$ 的结果运行 Trace，并返回其输出。

由于 ˆ$\mathcal{B}$ 只考虑 $\overline{\tau}$ 个码字，根据指纹码的性质，Trace 的结果以至少 $1 - \frac{1}{n^2} \geq \frac{1}{2}$ ˆ 的概率是一个未被 $\mathcal{B}$ 忽略的码字的坐标，特别地，它是介于 1 和 n 之间的一个坐标。因此，必然存在一个坐标 $i^* \neq 0$，使得该思想实验以至少 $\frac{1}{2n}$ 的概率输出它。现在考虑以下经过修改的思想实验。

1. 为指纹码采样一个码本 w_0，w_1，. . .，w_n。ˆ2. 在 ($\mathcal{B}$，w_1，. . .，w_{i^*-1}，w_0，w_{i^*+1}，. . .，w_n) 上运行 ˆ，并返回其输出。

ˆ由于 $\mathcal{B}$ 是计算差分隐私的，且 Trace 是高效算法，因此在第二个思想实验中输出 i^* 的概率与前一个思想实验中的大致相同，具体而言，至少为

$$e^{-2\varepsilon} \left(\frac{1}{2n} - (e^{\varepsilon} + 1)\delta - \mathrm{negl}(\Gamma(\lambda)) \right) \geq e^{-2\varepsilon} \left(\frac{1}{4n} - \mathrm{negl}(\Gamma(\lambda)) \right) = \Omega\left(\frac{1}{n} \right).$$

然而，根据 FPC（误指控小概率）的保证，在第二个实验中输出 i^* 的概率至多为 $\frac{1}{n^2}$。这与算法 $\mathcal{A}$ 的存在相矛盾。

□

以下推论可通过将定理 3.10 应用于我们的 MILR 方案而得到，该方案如定理 2.3所述。

推论 3.13。设 $\Omega(\lambda) \leq \Gamma(\lambda) \leq 2^{o(\lambda)}$，且 $d \leq \text{poly}(\Gamma(\lambda))$。若存在一个可抵御非一致对手的 $\Gamma(\lambda)$-安全加密方案，则存在一个 MILR 方案，使得相应的 (λ, d)-DA 问题需要大空间才能以保护隐私的方式求解。具体而言，设$n \leq \text{poly}(\Gamma(\lambda))$，设 ε 为常数，且 $\delta \leq \frac{1}{4n(e^{\varepsilon}+1)} = \Theta(\frac{1}{n})$。对于 (λ, d)-DA 问题，任何样本复杂度为 n 的 $\text{poly}(\Gamma(\lambda))$ 时间 (ε, δ)-计算差分隐私算法，其空间复杂度都必须为$s = \Omega\left(\lambda \cdot \sqrt{d \log n}\right)$.

4 自适应数据分析（ADA）的空间复杂度困难性

考虑这样一种机制：它首先接收一个样本作为输入，其中包含 t 个来自某个底层（未知）分布的独立同分布样本 $\mathcal{D}$，然后回答 k 个自适应选择的统计查询，其对象为 $\mathcal{D}$。重要的是，答案必须相对于底层分布准确，而不能仅仅相对于经验样本准确。这里的难点在于，随着查询被自适应地选择，这种交互可能很快导致过拟合，也就是说，得到的答案仅相对于经验样本准确，而相对于底层分布并不准确。Dwork 等人提出了这一基本问题，我们称之为自适应数据分析（ADA）问题。[34] 他们将该问题与差分隐私联系起来，并指出差分隐私可用作防止过拟合的手段。直观地说，当答案揭示的是输入样本特有的属性，而非总体的属性时，就会发生过拟合；而这正是差分隐私旨在防止泄露的内容。

Hardt、Steinke 和 Ullman [53, 81] 针对自适应数据分析问题给出了否定性结果。具体而言，他们证明，给定 t 个样本后，在计算上难以回答超过 $k = O(t^2)$ 个自适应查询。我们证明，ADA 问题的困难性实际上更加根本；事实上，这源于空间瓶颈，而非采样瓶颈。非正式地说，我们证明，即使在预处理阶段，机制获得了底层分布 $\mathcal{D}$的完整描述，随后只被要求存储关于该分布的有限信息（其数量等于 t 个样本的表示长度，并且来自$\mathcal{D}$），同样的困难性结果仍然成立。因此，并不是机制没有获得关于 $\mathcal{D}$的足够信息；而是它无法以一种支持 t^2 自适应查询的方式压缩这些信息。这推广了 [53, 81], 的否定性结果，因为从 $\mathcal{D}$ 中抽样 t 个点只是试图存储关于 $\mathcal{D}$的信息的一种特定方式。

考虑 AdaptiveGameSpace：机制最初会获得底层分布的完整描述，但必须将其压缩为一个 s 位的摘要 z。为强调机制无法额外访问底层分布，我们将其视为两个机制 $\mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$：其中 $\mathcal{A}_1$ 计算摘要 z，而 $\mathcal{A}_2$ 根据 z回答查询。我们将 $s = |z|$ 视为此类机制 $\mathcal{A}$的空间复杂度。

关于 ADA 问题，我们的主要定理如下。

定理 4.1。设 $\Omega(\lambda) \leq \Gamma(\lambda) \leq 2^{o(\lambda)}$，且 $k \leq \text{poly}(\Gamma(\lambda))$。若存在一个能够抵御非一致对手的 $\Gamma(\lambda)$-安全加密方案，则存在一个 $\text{poly}(\Gamma(\lambda))$-时间对手$\mathcal{B}$，使得以下结论成立。设 $\mathcal{A}$=$(\mathcal{A}_1, \mathcal{A}_2)$ 是一个 $\text{poly}(\Gamma(\lambda))$-时间机制，其空间为 com-

算法 1 AdaptiveGameSpace($\mathcal{A}$=($\mathcal{A}_1, \mathcal{A}_2$) $\mathcal{B}, s, k$)

1. 对手 $\mathcal{B}$ 选择定义域 $\mathcal{X}$ 上的一个分布 $\mathcal{D}$。2. 机制 $\mathcal{A}_1$ 获取 $\mathcal{D}$，并将其概括为 s 比特，记作 z。3. 机制 $\mathcal{A}_2$ 以 z 实例化。4. 对于轮次 $i = 1,2,\ldots,k$: (a) 对手 $\mathcal{B}$ 指定一个查询 $q_i : \mathcal{X} \to \{-1,0,1\}$(b) 机制 $\mathcal{A}_2$ 获取 q_i，并返回一个答案 $a_i \in [-1,1]$(c) 将 a_i 提供给 $\mathcal{A}$。5. 若 $\exists i$ s.t. $|a_i - \mathbb{E}_{y\sim\mathcal{D}}[q_i(y)]| > 1/10$，则游戏结果为 1；否则为 0。

复杂度 $s \le O\left(\lambda \cdot \sqrt{k}\right)$（足够小）。那么，$\Pr[\mathit{AdaptiveGameSpace}(\mathcal{A}, \mathcal{B}, s, k) = 1] > \frac{2}{3}$.

此外，对手定义的底层分布 $\mathcal{B}$ 可以用$O(\sqrt{k}\cdot\lambda)$比特完整描述，可以在 poly(Γ(λ)) 时间内进行抽样，并且从该分布中抽取的元素可以用 $O(\lambda + \log(k))$ 比特表示。

从某种意义上说，定理中“此外”这一部分表明，我们的对手所选择的分布并不复杂。具体而言，即使机制的空间与底层分布的完整描述长度呈线性关系（且允许高效地对其进行抽样），我们的否定性结果依然成立。如果机制的空间再大一些（仅大一个常数倍），它就能存储底层分布的完整描述，并回答无界数量的自适应查询。定理 4.1 的形式化证明推迟至附录B。这里我们只给出一个非正式（且过度简化的）证明概要。

4.1 非正式证明概要

令 k 表示对手发出的查询数量。我们的任务是证明，存在一个对手能够击败每个在 $\mathcal{A}_{\text{space}}$ 中进行博弈的高效机制，AdaptiveGameSpace前提是该机制使用的空间为 $s \ll \sqrt{k}$。根据 [53, 81]，我们知道存在一个对手 $\mathcal{B}_{\text{sample}}$，只要其样本复杂度为 $t \ll \sqrt{k}$，就能击败每个在标准 ADA 博弈（图1所定义的博弈）中进行博弈的高效机制。我们设计一个对手 $\mathcal{B}_{\text{space}}$，使其在 AdaptiveGameSpace中进行博弈的方式能够模拟 $\mathcal{B}_{\text{sample}}$。现在，我们将详细说明 $\mathcal{B}_{\text{space}}$构造中的关键要点，以及它们与 $\mathcal{B}_{\text{sample}}$的联系。

回想一下，两种博弈都以对手指定底层分布开始。关于对手 $\mathcal{B}_{\text{sample}}$（来自 [53, 81]）的一个有用事实是，它指定的分布在一个大小为 $n = \Theta(t)$ 的小点集上均匀分布（这些点 n 对与之对抗的机制$\mathcal{B}_{\text{sample}}$而言是未知的）。我们的对手 $\mathcal{B}_{\text{space}}$首先从我们的 MILR 方案中抽取 n 个独立密钥（x_1、……、x_n），然后将目标分布 $\mathcal{D}_{\text{space}}$ 定义为在集合$\{(j, x_j)\}_{j\in[n]}$上均匀分布。回想一下，在 AdaptiveGameSpace 中，该目标分布会提供给机制$\mathcal{A}_{\text{space}}$，由机制将其压缩为一个包含 s 比特的摘要 z。在此阶段之后，根据我们的 MILR 方案的安全性，应当存在一个大小为 $|J| = n - \overline{\tau}$ 的大集合 $J \subseteq [n]$，其中对应的是尚未被 $\mathcal{A}_{\text{space}}$攻破的密钥。记作 $I = [n] \setminus J$。

我们的对手 $\mathcal{B}_{\text{space}}$ 现在按如下方式模拟 $\mathcal{B}_{\text{sample}}$。首先，令 $\mathcal{D}_{\text{sample}}$ 表示由 $\mathcal{B}_{\text{sample}}$选择的目标分布，并令 $m_1,\ldots,m_n$表示其支撑集。随后，我们的对手对 n个公开参数 $p_1,\ldots,p_n$进行采样，并使用相应的密钥加密支撑集 m_j 中的每个点

和公开参数。具体而言， $c_j \leftarrow \mathrm{Enc}(x_j, p_j, m_j)$。% 这是一种过度简化。由于技术原因，实际的构造有所不同。

现在，对于 $\mathcal{B}_{\mathrm{sample}}$ 指定的每个查询 q，我们的对手都会输出由 $f_q(j,x) = q(\mathrm{Dec}(x, p_j, c_j))$ 定义的查询 f_q。随后，我们的对手从机制 $\mathcal{A}_{\mathrm{space}}$ 获取答案 a，并将 a 输入给 $\mathcal{B}_{\mathrm{sample}}$。注意， f_q 相对于 $\mathcal{D}_{\mathrm{space}}$ 的“真实”值，与 q 相对于 $\mathcal{D}_{\mathrm{sample}}$ 的“真实”值相同。因此，如果 $\mathcal{A}_{\mathrm{space}}$ 在这场与我们的对手 $\mathcal{B}_{\mathrm{space}}$ 对抗的博弈中保持准确性，那么在 $\mathcal{B}_{\mathrm{space}}$ 内部运行的模拟中， $\mathcal{B}_{\mathrm{space}}$ 就能对抗 $\mathcal{B}_{\mathrm{sample}}$ 保持效用。直观来说，我们希望据此得出矛盾，因为 $\mathcal{B}_{\mathrm{sample}}$ 会击败它所面对的每个高效机制。但这种说法并不准确，因为 $\mathcal{B}_{\mathrm{space}}$ 看到了目标分布 $\mathcal{D}_{\mathrm{sample}}$ 的完整描述，而 $\mathcal{B}_{\mathrm{sample}}$ 只能愚弄那些最多只能看到来自该目标分布的 t 个样本的机制。

$\widehat{}$为了解决这一问题，我们考虑对手的以下修改变体，称为 $\mathcal{B}_{\mathrm{space}}$。$\widehat{}$ 的修改之处在于，$\mathcal{B}_{\mathrm{space}}$ 无法看到 $\mathcal{D}_{\mathrm{sample}}$ 的完整描述，而只能看到 $\mathcal{D}_{\mathrm{sample}}$ 支撑集上与集合 $I = [n] \setminus J$ 中索引对应的点。随后，在$\widehat{}$生成密文 c_j 时，修改后的对手 $\mathcal{B}_{\mathrm{space}}$ 会加密零，而不是它缺失的点 m_j。根据我们的 MILR 方案的安全性，机制 $\mathcal{A}_{\mathrm{space}}$ 无法察觉这一修改，因此，假设它能在面对原始对手 $\mathcal{B}_{\mathrm{space}}\widehat{}$ 时保持准确性，那么它面对修改后的对手 $\mathcal{B}_{\mathrm{space}}$ 时也能保持准确性。与之前一样，这意味着$\sim\mathcal{B}_{\mathrm{space}}$ 能对抗模拟出的 $\mathcal{B}_{\mathrm{sample}}$ 保持准确性。直观来说，这就导致了矛盾，$\widehat{}$因为 $\mathcal{B}_{\mathrm{space}}$ 只使用了目标分布 $\mathcal{D}_{\mathrm{sample}}$ 中的 $\overline{\tau} \leq t$ 个点。

我们强调，这个证明概要过于简化且不准确。具体而言，以下两个$\widehat{}$技术问题需要解决：(1) 的确，$\mathcal{B}_{\mathrm{space}}$ 仅使用了 $\overline{\tau}$ 目标分布支撑集中的点， $\mathcal{D}_{\mathrm{sample}}$但这些点不一定是从采样得到的 $\mathcal{D}_{\mathrm{sample}}$；以及$\widehat{}$(2) 修改后的对手 $\mathcal{B}_{\mathrm{space}}$ 在计算上并非高效，因为计算集合 J 并不高效。我们在附录B中解决这些问题以及本文中的其他非正式之处。

5 从语义安全加密方案构造 MILR 方案

构造。设 $\lambda' \leq \lambda$ 满足 $\lambda = \mathrm{poly}(\lambda')$。给定一个加密方案 $\Pi' = (\mathrm{Gen}', \mathrm{Enc}', \mathrm{Dec}')$，使得 $\mathrm{Gen}'(1^{\lambda'})$ 输出一个在 $\{0,1\}^{\lambda'}$ 上均匀分布的密钥（即 $x' \leftarrow_R \{0,1\}^{\lambda'}$），我们按如下方式构造一个 MILR 方案 $\Pi = (\mathrm{Gen}, \mathrm{Param}, \mathrm{Enc}, \mathrm{Dec})$：

- Gen：输入 1^{λ}，返回 $x \leftarrow_R \{0,1\}^{\lambda}$。
- Param：输入 1^{λ}，令 $\mathcal{G}$ 为定义域为 $\{0,1\}^{\lambda}$、值域为 $\{0,1\}^{\lambda'}$的通用哈希函数族。返回 $g \leftarrow_R \mathcal{G}$（的描述）。
- Enc：输入 (x, p, m)，解析 $g := p$（作为哈希函数的描述），令 $x' = g(x)$，并返回 $\mathrm{Enc}'(x', m)$。
- Dec：输入 (x, p, c)，解析 $g := p$，令 $x' = g(x)$，并返回 $\mathrm{Dec}'(x', c)$。

使用通用哈希函数族的标准构造，所有算法的运行时间都是关于 λ 的多项式。此外，如果 $c \leftarrow \mathrm{Enc}(x, p, m)$，那么 $\mathrm{Dec}(x, p, c) = m$ 以概率 1 成立（因为这

对 Enc′ 和 Dec′ 都成立）。

以下两个定理（分别对应定义 2.1 中的两种安全性）确立了 Π 的安全性，并证明定理 2.3。

定理 5.1（多重语义安全）。设 $\Omega(\lambda) \leq \Gamma(\lambda) \leq 2^{o(\lambda)}$ 和 $\lambda' = 0.1\lambda$。若 Π′ 是针对均匀对手（分别为非一致对手）的 $\Gamma(\lambda')$-安全的，则 Π 是针对均匀对手（分别为非一致对手）的 $\Gamma(\lambda)$-安全的。

定理 5.2（针对有界预处理对手的多重安全性）。设 $\Omega(\lambda) \leq \Gamma(\lambda) \leq 2^{o(\lambda)}$, $\lambda' = 0.1\lambda$ （如定理 5.1 中所述）。若 Π′ 针对非一致对手是 $\Gamma(\lambda')$-安全的，则 Π针对空间有界的非均匀预处理对手是 $(\Gamma(\lambda)\ \tau)$- 安全的，其中 $\tau(\lambda, s) = \frac{2s}{\lambda} + 4$。

注释 5.3。由于 $\Gamma(\lambda) \leq 2^{o(\lambda)}$ and $\lambda' = 0.1\lambda$，因此 $\text{poly}(\Gamma(\lambda')) = \text{poly}(\Gamma(\lambda))$。因此，为简单起见，我们将所有对手（包括针对 Π′运行的对手）的运行时间和优势都作为 λ 的函数进行分析。

我们在附录 5.1 中证明定理C 和定理5.2 在第 6节中。

6 多重 -安全性：针对有界预处理对手 y

本节将证明定理 5.2。证明需要特定的定义和符号，具体如下。

6.1 预备知识

记号。给定由元素组成的序列 $X = (X_1, \ldots, X_n)$ 和子集 $I \subseteq [n]$，我们用 X_I 表示由坐标属于 I 的元素组成的序列。

对于随机变量 X，用 $H_\infty(X)$ 表示其最小熵。对于值域相同的随机变量 X、Y，用 $\Delta(X, Y)$ 表示它们的分布之间的统计距离。若 $\Delta(X, Y) \leq \gamma$，则称 X 和 Y 是 γ-接近的。我们用记号 $X \leftarrow_R \mathcal{X}$ 表示随机变量 X 是从集合 $\mathcal{X}$ 中均匀随机选取的。

稠密且比特 -固定源。 我们将使用以下定义（参见 [25, 定义 1]）。

定义 6.1。$(n, 2^\lambda)$-源是一个值域为 $(\{0,1\}^\lambda)^n$ 的随机变量 X。若一个源

- $(1-\delta)$-稠密的，若对于每个子集 $I \subseteq [n]$， $H_\infty(X_I) \geq (1-\delta) \cdot |I| \cdot \lambda$，

  如果其至多在 k 个坐标上取固定值，并且在其余坐标上是（$1-\delta$）-稠密的，则称其为 •（k， $1-\delta$）-稠密的，

  如果其至多在 k 个坐标上取固定值，并且在其余坐标上均匀分布，则称其为 • k-位固定的。

也就是说，$(1-\delta)$-稠密源中任意条目子集的最小熵，至多比均匀随机源的情况低 δ 的一个比例。

6.2 密钥泄漏引理

令 $X=(X_1,\ldots,X_n)\in(\{0,1\}^\lambda)^n$ 为随机变量，表示从 Π 中独立且均匀随机选取的 n 个密钥。令 $Z:=F(X)$ 为表示对手泄漏的随机变量。对于 $z\in\{0,1\}^s$，令 X_z 为从 X 的分布中、在给定 $F(X)=z$的条件下选取的随机变量。

我们将 $G:=(G_1,\ldots,G_n)$ 和 $G(X):=(G_1(X_1)\ldots,G_n(X_n))$ 记为 Π 的哈希函数（公开参数）对应的随机变量。对于长度不同的序列，我们也会使用类似的记法（具体含义可从上下文中看出）。

定理 5.2 的证明基于下面的引理（在第 6.4 节中证明），该引理分析了（G、Z、$G(X)$ 的联合分布。

引理 6.2. 设 $F\colon(\{0,1\}^\lambda)^n\to\{0,1\}^s$ 是任意函数，$X=(X_1,\ldots,X_n)\leftarrow_R(\{0,1\}^\lambda)^n$，并记 $Z:=F(X)$。设 $\mathcal{G}$ 是一个定义域为 $\{0,1\}^\lambda$、值域为 $\{0,1\}^{\lambda'}$ 的通用哈希函数族，且 $G\leftarrow_R(\mathcal{G})^n$。设 $\delta>0$、$\gamma>0$、$s'>s$ 为满足 $(1-\delta)\lambda>\lambda'+\log n+1$ 的参数。

那么，存在一族由凸组合构成的$V_{G,\ Z}=\{V_{\vec{g},z}\}_{\vec{g}\in(\mathcal{G})^n,\ z\in\{0,1\}^s}$，这些凸组合是 k-位固定（$n,2^{\lambda'}$）-源，且满足 $k=s'+\log_{\delta\cdot\lambda}\frac{1/\gamma}{}$，使得

$$\Delta[(G,Z,G(X)),(G,Z,V_{G,Z})]\le\sqrt{2^{-(1-\delta)\lambda+\lambda'+\log n}}+\gamma+2^{s-s'}.$$

我们得到以下推论（它蕴含了定理 5.2 的参数）。

推论 6.3。 在引理 *6.2* 的设定下，假设 $n<2^{0.15\lambda}$ 且 λ充分大，则参数 $s'=s+\lambda$、$\lambda'=0.1\lambda$、$\delta=0.5$、$\gamma=2^{-\lambda}$ 满足

$$\Delta[(G,Z,G(X)),(G,Z,V_{G,Z})]\le 2^{-0.1\lambda},\qquad and\qquad k=\frac{2s}{\lambda}+4.$$

证明。令 $s'=s+\lambda$, $\lambda'=0.1\lambda$, $\delta=0.5,\gamma=2^{-\lambda}$。于是，当 λ 充分大时，$(1-\delta)\lambda=0.5\lambda>0.1\lambda+0.15\lambda+1>\lambda'+\log n+1$（并且引理 6.2 的条件成立）。因此（当 λ 充分大时）：$\Delta[(G,Z,G(X))$，$(G,Z,V_{G,\ Z})]\le\sqrt{2^{-(1-\delta)\lambda+\lambda'+\log n}}+\gamma+2^{s-s'}=\sqrt{2^{-0.4\lambda+\log n}}+2^{-\lambda}+2^{-\lambda}\le 2^{-0.1\lambda}$，且 $k=s'+\frac{1/\gamma}{\log_{\delta\cdot\lambda}}=\frac{s+\lambda+\lambda}{0.5\cdot\lambda}=\frac{2s}{\lambda}+4$。

□

6.3 定理5.2的证明

利用引理**6.2** 证明定理**5.2**。 在证明定理5.2之前，我们解释一下为什么需要引理6.2 以及它在证明中如何使用。

证明比引理 6.2更弱的一些结论很容易，但这些结论似乎不足以构造 MILR 方案（即证明定理 5.2）。例如，很容易证明，以高概率而言，给定泄漏 z 和哈希函数 $\vec{g}$，存在一个由（经过哈希处理的）密钥构成的大子集，其中每个密钥都近似服从均匀分布。然而，对手可能掌握该子集中各密钥之间的各种关系，而目前尚不清楚如何

在不对加密方案抵抗相关密钥攻击的能力作出假设的情况下证明其安全性。

此外，考虑一个更强的陈述：它断言，给定泄漏z和哈希函数$\vec{g}$后，以高概率存在一个较大的（哈希后）密钥子集，这些密钥联合均匀分布。我们认为，即使这个更强的陈述也不足以证明安全性，因为它没有考虑子集之外的其余密钥。特别地，考虑这样一种情形：对手能够恢复子集之外的一些弱密钥。结合这项额外知识和泄漏 z，原有的密钥子集可能不再均匀分布（并且可能遭受显著的熵损失）。

引理 6.2 基本断言，即使将 z、$\vec{g}$ 以及所有其余密钥都提供给对手，仍然存在一个密钥子集，使其近似联合均匀分布。更具体地说，给定哈希函数 $\vec{g}$ 和泄漏 z，根据该引理，哈希密钥的分布$\vec{g}(X)$是$V_{\vec{g},z}$位固定源的k凸组合（或与之接近）。在定理 5.2 的证明中，我们将通过向对手提供 k 哈希密钥来固定这样一个位固定源（我们会谨慎处理，确保对手的优势不会发生显著变化）。由于从对手的视角看，其余哈希密钥均匀分布，因此这些密钥的安全性可由底层加密方案的语义安全性推出。

定理 5.2 的证明。固定一个预处理过程 F，并令 λ 足够大，$n < 2^{0.15\lambda}$。根据引理 6.2（参数取为推论 6.3 中的设置），存在一个由 $V_{G,\ Z}$ 个凸组合 $V_{\vec{g},z}$ k n $2^{\lambda'}$ $k = \frac{2s}{\lambda} + 4$ $\Delta[\ G\ Z\ G\ X\ ,\ G\ Z\ V_{G\ Z}\ ] \leq$，构成的族，其中每个凸组合都是 -位固定 (,)-源，使得 (, , ( )) (, ,,)
$2^{-0.1\lambda}$.

对索引集 J 进行抽样并简化分布。我们首先定义定义 2.1 中的预言机 $\mathcal{E}_0$ 和 $\mathcal{E}_1$ 如何对 J 进行抽样。在对随机变量 $(G, Z, V_{G,Z})$ 进行抽样时，随机变量 J 的定义很自然：给定 $\vec{g} \in (\mathcal{G})^n$，$z \in \{0,1\}^s$，从凸组合 $V_{\vec{g},z}$ 中抽取一个 k-位固定源（根据其权重），并令 J 为至少包含 $n-k$ 个未固定索引的集合。这定义了随机变量 $(G, Z, V_{G,Z}, J)$ 上的联合分布。另一种从该分布中抽样的方法是先对变量 $(G, Z, V_{G,Z})$ 进行抽样，然后根据其边缘分布对 J 进行抽样。这定义了一个用于对 J 进行抽样的随机化过程。尽管这些预言机不会对 $(G, Z, V_{G,Z})$ 进行抽样，我们仍使用相同的抽样过程，在给定样本 $(\vec{g}, z, \vec{g}(\vec{x}))$ 的情况下对 J 进行抽样（若样本 $(\vec{g}, z, \vec{g}(\vec{x}))$ 不在 $(G, Z, V_{G,Z})$ 的分布支撑集中，则定义 $J = [n]$）。

考虑一个运行时间为 poly(Γ(λ)) 的算法 $\mathcal{B}$。由于 $\mathcal{B}$ 向 $\mathcal{E}_0$ 和 $\mathcal{E}_1$ 发出的加密查询由哈希密钥 $\vec{g}(\vec{x})$ 回答，因此给定 $\vec{g}(\vec{x})$ 后，$\mathcal{B}$ 与 $\mathcal{E}_0$ 和 $\mathcal{E}_1$ 之间的交互不再依赖于 $\vec{x}$。因此，对于 $t = 0$，我们定义 $\mathcal{E}_t^{(1)}(\vec{g}(\vec{x})\ \vec{g}, J, \cdot, \cdot)$，用以模拟 $\mathcal{E}_t(\vec{x}, \vec{g}, J, \cdot, \cdot)$ 与 $\mathcal{B}$ 的交互。预言机不再抽样 $\vec{x}$，而是直接根据交互开始前的联合分布抽样 $(\vec{g}, z, \vec{g}(\vec{x})\ J)$。为简化记号，我们用 $\mathcal{D}_1$ 表示该联合分布。记

$$\mathrm{Adv}_{\mathcal{B}}(\lambda) = \left| \Pr_{\substack{\mathcal{B},\mathrm{Enc} \\ (\vec{g},z,\vec{g}(\vec{x}),J)\leftarrow\mathcal{D}_1}} \left[\mathcal{B}^{\mathcal{E}_0^{(1)}(\vec{g}(\vec{x}),\vec{g},J,\cdot,\cdot)}(z,\vec{g}) = 1\right] - \Pr_{\substack{\mathcal{B},\mathrm{Enc} \\ (\vec{g},z,\vec{g}(\vec{x}),J)\leftarrow\mathcal{D}_1}} \left[\mathcal{B}^{\mathcal{E}_1^{(1)}(\vec{g}(\vec{x}),\vec{g},J,\cdot,\cdot)}(z,\vec{g}) = 1\right] \right|.$$

接下来要证明 $\mathrm{Adv}_{\mathcal{B}}(\lambda) \leq \mathrm{negl}(\Gamma(\lambda))$）。

使用引理 6.2 将其转换为位固定源的凸组合族。我们有

$$\Delta[(G, Z, G(X), J), (G, Z, V_{G,Z}, J)] \leq \Delta[(G, Z, G(X)), (G, Z, V_{G,Z})] \leq 2^{-0.1\lambda},$$

其中，第一个不等式由数据处理不等式得出，因为对两个分布中的三个变量应用相同的函数即可计算出 J；第二个不等式则由推论 6.3 得出。因此，对于 $t = 0$，我们将从 $\mathcal{D}_1$ 中采样的 $\mathcal{E}_t^{(1)}$ 替换为从 $(G, Z, V_{G,\ Z}, J)$ 的联合分布中采样的 $\mathcal{E}_t^{(2)}$，并将该联合分布记为 $\mathcal{D}_2$。由于 $\mathcal{B}$ 和 Enc 使用独立的随机性，根据三角不等式，总惩罚至多为 $2 \cdot 2^{-0.1\lambda}$，即

$$\left| \Pr_{\substack{\mathcal{B},\mathrm{Enc} \\ (\vec{g},z,\vec{y},J)\leftarrow\mathcal{D}_2}} \left[\mathcal{B}^{\mathcal{E}_0^{(2)}(\vec{y},\vec{g},J,\cdot,\cdot)}(z,\vec{g}) = 1\right] - \Pr_{\substack{\mathcal{B},\mathrm{Enc} \\ (\vec{g},z,\vec{y},J)\leftarrow\mathcal{D}_2}} \left[\mathcal{B}^{\mathcal{E}_1^{(2)}(\vec{y},\vec{g},J,\cdot,\cdot)}(z,\vec{g}) = 1\right] \right| \geq \mathrm{Adv}_{\mathcal{B}}(\lambda) - 2^{-0.1\lambda+1},$$

其中，我们将从 $\mathcal{D}_2$ 中采样的结果记为 $(\vec{g}, z, \vec{y}, J)$。

向对手提供额外输入。考虑一个针对 Π 的（可能）更强大的 $\mathrm{poly}(\Gamma(\lambda))$ 时间算法 $\mathcal{B}_1$，其输入由 $(z, \vec{g}, J, \vec{y}_{\overline{J}})$ 构成，其中 $\overline{J} = [n] \backslash J$。具体而言，除了 $(z, \vec{g})$ 之外，输入还包括 J 以及哈希密钥 $\vec{y}_{\overline{J}} \in \left(\{0,1\}^{\lambda'}\right)^{n-|J|}$（注意，这些参数定义了一个 $|\overline{J}|$-位固定源）。我们将 $in = (z, \vec{g}, J, \vec{y}_{\overline{J}})$ 记为，并且

$$\mathrm{Adv}_{\mathcal{B}_1}(\lambda) = \left| \Pr_{\substack{\mathcal{B}_1,\mathrm{Enc} \\ (\vec{g},z,\vec{y},J)\leftarrow\mathcal{D}_2}} \left[\mathcal{B}_1^{\mathcal{E}_0^{(2)}(\vec{y},\vec{g},J,\cdot,\cdot)}(in) = 1\right] - \Pr_{\substack{\mathcal{B}_1,\mathrm{Enc} \\ (\vec{g},z,\vec{y},J)\leftarrow\mathcal{D}_2}} \left[\mathcal{B}_1^{\mathcal{E}_1^{(2)}(\vec{y},\vec{g},J,\cdot,\cdot)}(in) = 1\right] \right|.$$

接下来，我们证明，对于任意这样的 $\mathrm{poly}(\Gamma(\lambda))$ 时间算法 $\mathcal{B}_1$，$\mathrm{Adv}^{\mathcal{B}_1}(\lambda) \leq \mathrm{negl}(\Gamma(\lambda))$。由于任何输入为 $(z, \vec{g})$ 的算法 $\mathcal{B}$ 都可以由一个输入为 in 且运行时间相近的算法 $\mathcal{B}_1$ 模拟，因此这意味着 $\mathrm{Adv}_{\mathcal{B}}(\lambda) - 2^{-0.1\lambda+1} \leq \mathrm{negl}(\Gamma(\lambda))$，进而 $\mathrm{Adv}_{\mathcal{B}}(\lambda) \leq \mathrm{negl}(\Gamma(\lambda))$，证明至此完成。

固定对手的输入。由于 $\mathcal{E}_0^{(2)}$ 和 $\mathcal{E}_1^{(2)}$ 都从同一分布中抽取 $\mathcal{B}_1$ 的输入，根据平均论证，存在一个输入 $in^* = in_\lambda^* = (z^*, \vec{g}^*, J^*, \vec{y}^*_{\overline{J^*}})$，使得将输入固定为 in^*，并从以 in^* 为条件的 $\mathcal{D}_2$ 中抽样时，$\mathcal{B}_1$ 的优势仍至少与原来一样大。注意，给定从 $\mathcal{D}_2$ 中抽取的 in^*，这等价于从由 $(\overline{J^*}, \vec{y}^*_{\overline{J^*}})$ 定义的 $|\overline{J^*}|$-位固定源中抽样，即选择 $\vec{w} \leftarrow_R (\{0,1\}^{\lambda'})^{|J^*|}$。因此，

$$\left| \Pr_{\substack{\mathcal{B}_1,\mathrm{Enc} \\ \vec{w}\leftarrow_R(\{0,1\}^{\lambda'})^{|J^*|}}} \left[\mathcal{B}_1^{\mathcal{E}_0^{(2)}(in^*,\vec{w},\cdot,\cdot)}(in^*) = 1\right] - \Pr_{\substack{\mathcal{B}_1,\mathrm{Enc} \\ \vec{w}\leftarrow_R(\{0,1\}^{\lambda'})^{|J^*|}}} \left[\mathcal{B}_1^{\mathcal{E}_1^{(2)}(in^*,\vec{w},\cdot,\cdot)}(in^*) = 1\right] \right| \geq \mathrm{Adv}_{\mathcal{B}_1}(\lambda).$$

利用预处理，将 Π的安全性归约到（多实例）安全性Π′。 现在，我们使用 $\mathcal{B}_1$ 来定义一个非均匀 poly(Γ(λ)-时间对手 $\mathcal{B}_2$ （不进行预处理），使其针对 $|J^*|$ 个 Π′ 实例运行，且其优势至少为 Adv$_{\mathcal{B}_1}(\lambda)$。根据 Π′ 的语义安全性以及混合论证（类似于定理 5.1的证明），可得 Adv$_{\mathcal{B}_1}(\lambda) \leq$ negl(Γ(λ)，证明得证。

算法 2 $\mathcal{B}^{\mathcal{E}^{(3)}_{\cdot\,()(\vec{x}',[|J^*|],\cdot,\cdot)\ 2}()}$

设定：$\mathcal{B}_2$ 是一个非均匀对手，它针对 $|J^*_{\vec{\cdot}}|$ 个由 Π′（由 $\mathcal{E}$ 定义）实例运行 $\mathcal{E}^{(2)}_{\cdot\,()(\cdot))}$。它获得 $in^* = in^*_\lambda = (z^*, \vec{g}^*, J^*, y^*_{\overline{J^*}})$ 作为辅助信息。$\mathcal{B}_2$可以访问 $\mathcal{B}^{(3)}_{\cdot\,()(in^*,\vec{w},\cdot,\cdot)1}(in^*)$，该算法针对Π运行。

1. $\mathcal{B}_2$ 将 in^* 作为输入提供给 $\mathcal{B}_1$。2. $\mathcal{B}_2$ 按如下方式回答 $\mathcal{B}_1$ 的每个查询 (j,m)：
   - 如果 $j \in \overline{J^*}$，$\mathcal{B}_2$使用辅助信息字符串 in^*（其中包含 $\vec{y}^*_j$）计算答案 Enc′$(\vec{y}^*_j, m)$，并将其交给 $\mathcal{B}_1$。
   - 若 $j \in J^*$，$\mathcal{B}_2$会将查询 (j,m) 转换为 (j',m)，其中 $j' \in [|J^*|]$是通过将 j映射到 J^*（忽略 $\overline{J^*}$中的索引）得到的。$\mathcal{B}_2$随后会用 (j',m) 查询其预言机，并将答案转发给 $\mathcal{B}_1$。
3. $\mathcal{B}_2$ 输出与 $\mathcal{B}_1$相同的结果。

对手 $\mathcal{B}_2$ 如算法 2所示。注意，$\mathcal{B}_2$ 在给定输入 $\mathcal{B}_1$ 的情况下，能够完美模拟 in^*的预言机，因此其优势至少为 Adv$_{\mathcal{B}_1}(\lambda)$，如所述。最后，其运行时间为多项式时间 poly(Γ(λ)。

□

6.4 引理 6.2 的证明

证明概述。我们首先在引理 6.4 中证明，$G = (G_1, \ldots, G_t)$（对于某个 $t \in [n]$）是一个良好提取器，前提是其输入 $Y = (Y_1, \ldots, Y_t)$ 是 $(1-\delta)$-稠密的，也就是说，对于每个坐标子集，它都具有足够的最小熵（精确表述见引理 6.4）。具体而言，我们证明 $(G, G(Y))$ 与 (G, U) 统计接近，其中 U在 $(\{0,1\}^{\lambda'})^t$ 上均匀分布。证明采用剩余哈希引理 [54] 的一种变体：将一系列哈希函数 $(G_1, \ldots, G_t)$ 分别应用于输入的每个块（而不是对整个输入应用单个哈希函数）。我们注意到，在不同的通信复杂度设置下，[49, 中证明了一个相关引理（引理 13] ）。我们的变体适用于不同（且大多更宽）的参数范围，例如 δ 的不同取值以及所提取的比特数 $t \cdot \lambda'$，这些都与我们的设置相关。[32, 28] 中还给出了一些其他（相关性稍弱的）结果。

证明的其余部分见附录 D，与 [25, 引理 1] 有些相似。我们固定$z \in \{0,1\}^s$，使得 $H_\infty(X_z)$ 足够高。引理 D.1（在 [25, 49, 62] 中证明）指出，X_z（在应用 G 之前）接近于 $(k, 1-\delta)$-稠密源的一个凸组合（其中 k 的值足够小）。因此，我们对每个这样的 $(1-\delta)$-稠密源中最小熵较高的坐标应用引理 6.4。由此可得引理 D.2，该引理指出 $(G, G(X_z))$ 与 $(G, V^*_{G,z})$ 统计接近，其中 $V^*_{G,z}$ 是由 k 位固定的 $(n, 2^{\lambda'})$-源构成的凸组合族，以 $\vec{g} \in \mathcal{G}^n$ 为索引。

命题 1 （这是一个相当标准的结论）表明，以极高的概率（关于 X 和 F，

设定 $z \leftarrow F(X)$)， $H_\infty(X_z)$ 确实足够高，可以使用参数良好的引理 D.2。对于每个这样的 z，我们定义 $V_{G,z}$ 为上文所述的 $V^*_{G,z}$，这实质上得到族 $V_{G,Z}$ （ z 中其余 $H_\infty(X_z)$ 较低的取值很容易处理，且统计距离的损失很小）。因此，结合命题 1 和引理 D.2，可以界定 $(G, Z, G(X))$ 与 (G, Z, V_G, z) 之间的统计距离，从而证明引理 6.2。

从稠密源中进行分块提取。

引理 6.4. 设 $Y = (Y_1, \ldots, Y_t) \in (\{0,1\}^\lambda)^t$ 是一个 $(t, 2^\lambda)$-源，并且对于 $0 < \delta < 1$，它是 $(1-\delta)$-稠密的。设 $\mathcal{G}$ 是一个定义域为 $\{0,1\}^\lambda$、值域为 $\{0,1\}^{\lambda'}$ 的通用哈希函数族。那么，对于$G \leftarrow_R (\mathcal{G})^t$ 和 $U \leftarrow_R (\{0,1\}^{\lambda'})^t$,

$$\Delta[(G, G(Y)), (G, U)] \leq \sqrt{2^{-(1-\delta)\lambda + \lambda' + \log t}},$$

假设 $(1-\delta)\lambda > \lambda' + \log t + 1$。

证明. 设 $d := \log |\mathcal{G}|$。对于随机变量 Q，以及 Q 的一个独立副本 Q'，我们用$Col[Q] = \Pr[Q = Q']$ 表示 Q 的碰撞概率。我们有

$$\begin{aligned} &Col[(G, G(Y))] = \Pr_{G,Y,G',Y'}[(G, G(Y')) = (G', G'(Y'))] = \\ &\Pr_{G,G'}[G = G'] \cdot \Pr_{G,Y,Y'}[G(Y) = G(Y')] = 2^{-t \cdot d} \cdot \Pr_{G,Y,Y'}[G(Y) = G(Y')]. \end{aligned} \tag{1}$$

对于序列 $Y_1, \ldots, Y_t, Y'_1, \ldots, Y'_t$，定义 $C = |\{i \mid Y_i = Y'_i\}|$。现在我们对表达式 $\Pr[C = c]$ 进行上界估计。

回顾一下， Y是一个（$1-\delta$）-稠密源，也就是说，对于每个子集 $I \subseteq [t]$， $H_\infty(Y_I) \geq (1-\delta) \cdot |I| \cdot \lambda$。取一个满足 $|I| = c$ 的子集 $I \subseteq [t]$。于是,

$$\begin{aligned} &\Pr[Y_I = Y'_I] = \sum_{y_I \in (\{0,1\}^\lambda)^c} (\Pr[Y_I = y_I])^2 \leq \\ &\max_{y_I}\{\Pr[Y_I = y_I]\} \cdot \sum_{y_I \in (\{0,1\}^\lambda)^c} \Pr[Y_I = y_I] \leq 2^{-(1-\delta)\cdot c \cdot \lambda}. \end{aligned}$$

因此,

$$\begin{aligned} &\Pr[C = c] \leq \sum_{\{I \subseteq [t] \mid |I| = c\}} \Pr[Y_I = Y'_I] \leq \binom{t}{c} \cdot 2^{-(1-\delta)\cdot c \cdot \lambda} \leq \\ &t^c \cdot 2^{-(1-\delta)\cdot c \cdot \lambda} = 2^{c \cdot (-(1-\delta)\lambda + \log t)}. \end{aligned} \tag{2}$$

我们有

$$\Pr_{G,Y,Y'}[G(Y) = G(Y')] = \sum_{c=0}^{t} \Pr_{Y,Y'}[C = c] \cdot \Pr_{G,Y,Y'}[G(Y) = G(Y') \mid C = c].$$

对于每个满足 $Y_i \neq Y'_i$ 的坐标 i，当 G_i 从通用哈希函数族中均匀选取时，$\Pr_{G_i}[G_i(Y_i) = G_i(Y'_i)] = 2^{-\lambda'}$。由于 $G = (G_1, \ldots, G_t)$ 包含 t 个从 $\mathcal{G}$ 中均匀选取的独立副本，

$$\Pr_G[G(Y) = G(Y') \mid C = c] = 2^{-\lambda' \cdot (t-c)}.$$

因此，利用（2），我们得到

$$\Pr_{G,Y,Y'}[G(Y) = G(Y')] = \sum_{c=0}^{t} \Pr[C = c] \cdot 2^{-\lambda' \cdot (t-c)} \leq$$
$$\sum_{c=0}^{t} 2^{c \cdot (-(1-\delta)\lambda + \log t)} \cdot 2^{-\lambda' \cdot (t-c)} = 2^{-\lambda' \cdot t} \cdot \sum_{c=0}^{t} 2^{-c \cdot ((1-\delta)\lambda - \lambda' - \log t)} =$$
$$2^{-\lambda' \cdot t} \cdot (1 + \sum_{c=1}^{t} 2^{-c \cdot ((1-\delta)\lambda - \lambda' - \log t)}) \leq 2^{-\lambda' \cdot t} \cdot (1 + 2^{-(1-\delta)\lambda + \lambda' + \log t + 1}),$$

其中，最后一个不等式利用了假设 $(1-\delta)\lambda > \lambda' + \log t + 1$。

将分布视为 $\{0,1\}^{t \cdot d + t \cdot \lambda'}$ 上的向量（并滥用记号），将上述表达式代入（1），可推出

$$\|(G, G(Y)) - (G, U)\|_2^2 = Col[(G, G(Y))] - 2^{-t \cdot d - t \cdot \lambda'} \leq$$
$$2^{-t \cdot d - t \cdot \lambda'} \cdot (1 + 2^{-(1-\delta)\lambda + \lambda' + \log t + 1}) - 2^{-t \cdot d - t \cdot \lambda'} = 2^{-t \cdot d - t \cdot n' - (1-\delta)\lambda + \lambda' + \log t + 1}.$$

最后，利用柯西–施瓦茨不等式，我们得出

$$\Delta[(G, G(Y)), (G, U)] \leq 1/2 \cdot \sqrt{2^{t \cdot d + t \cdot \lambda'}} \cdot \|(G, G(Y)) - (G, U)\|_2 \leq$$
$$1/2 \cdot \sqrt{2^{t \cdot d + t \cdot \lambda'}} \cdot \sqrt{2^{-t \cdot d - \lambda' \cdot t - (1-\delta)\lambda + \lambda' + \log t + 1}} < \sqrt{2^{-(1-\delta)\lambda + \lambda' + \log t}}.$$

□

A 密码学的其他预备知识

概率系综是一个无限的概率分布序列。计算不可区分性定义如下。

定义 A.1。若对于任意非一致概率多项式时间区分器 D，都存在一个可忽略函数 negl，使得：则称两个概率分布族 $\mathcal{X} = \{X_\lambda\}_{\lambda \in \mathbb{N}}$ 和 $\mathcal{Y} = \{Y_\lambda\}_{\lambda \in \mathbb{N}}$ 在计算上不可区分，记作 $\mathcal{X} \equiv_c \mathcal{Y}$。

$$\left| \Pr_{x \leftarrow X_\lambda}[D(x) = 1] - \Pr_{y \leftarrow Y_\lambda}[D(y) = 1] \right| \leq \text{negl}(\lambda).$$

定义 A.2 ([48])。 一个 非一致性多项式时间算法 是一个二元组 $(M, \overline{a})$，其中 M 是一个双输入多项式时间图灵机，而 $\overline{a} = a_1, a_2, \ldots$ 是一个无限字符串序列

使得 $|a_\lambda| = \text{poly}(\lambda)$。对于每个 x，我们考虑 M 在输入对 $(x, a_{|x|})$ 上的计算。直观地说，a_λ 可以看作是从“外部”提供的额外“辅助信息”（与输入 $x \in \{0,1\}^\lambda$ 一起）。我们强调，对于所有长度相同的输入（即 λ），M 获得的辅助信息都相同（即 a_λ）。

等价地，非一致性多项式时间算法是一个无限的图灵机序列，$M_1, M_2, \ldots$，满足 M_λ 的描述长度及其在长度为 λ 的输入上的运行时间，均由关于 λ 的多项式所界定（该多项式对整个序列固定）。机器 M_λ 仅用于长度为 λ的输入。

加密方案。加密方案是由以下高效算法组成的三元组 (Gen, Enc,Dec)，其语法如下：

- Gen 是一个随机化算法，输入安全参数 λ，输出一个 λ位的密钥。形式化地，$x \leftarrow \text{Gen}(1^\lambda)$。
- Enc 是一个随机化算法，输入密钥 x、公开参数 p和消息 $m \in \{0,1\}$，输出一个密文 $c \in \{0,1\}^{\text{poly}(\lambda)}$。形式化地，$c \leftarrow \text{Enc}(x, m)$。
- Dec 是一个确定性算法，输入密钥 x、公开参数 p和密文 c，输出解密后的消息 m'。如果密文 c 是使用密钥 x 和参数 p对 m 进行加密得到的，那么 $m' = m$。形式化地，如果 $c \leftarrow \text{Enc}(x, m)$，则 $\text{Dec}(x, c) = m$，且其概率为 1。

让我们回顾一下加密方案安全性的形式化定义。考虑一对预言机$\mathcal{E}_0$和 $\mathcal{E}_1$，其中 $\mathcal{E}_1(x, \cdot)$接收消息 m 作为输入，并返回 $\text{Enc}(m, x)$；而 $\mathcal{E}_0(x, \cdot)$接收相同的输入，但返回 $\text{Enc}(0, x)$。如果不存在计算高效的对手能够判断其交互的是 $\mathcal{E}_0$ 还是 $\mathcal{E}_1$，则加密方案（Gen、Enc、Dec）是安全的。形式化地说，

定义 A.3. 设 λ 为安全参数，$\Gamma = \Gamma(\lambda)$为 λ的函数。若对于每个运行时间为 $\text{poly}(\Gamma(\lambda))$ 的对手 $\mathcal{B}$，都存在一个可忽略函数 negl，使得：

$$\left| \Pr_{\substack{x \leftarrow \text{Gen}(1^\lambda) \\ \mathcal{B}, \text{Enc}}} \left[\mathcal{B}^{\mathcal{E}_0(x,\cdot)} = 1 \right] - \Pr_{\substack{x \leftarrow \text{Gen}(1^\lambda) \\ \mathcal{B}, \text{Enc}}} \left[\mathcal{B}^{\mathcal{E}_1(x,\cdot)} = 1 \right] \right| \leq \text{negl}(\Gamma(\lambda)),$$

其中概率　　是对抽样 $x \leftarrow \text{Gen}(1^\lambda)$以及 $\mathcal{B}$ 和 Enc 的随机性取的　　.

B 定理 4.1 的证明

本节证明定理 4.1。我们首先回顾 [53, 81] 的其他预备知识。作为证明其关于 ADA 问题的否定性结果的铺垫，[53, 81] 针对一族受限的机制（称为自然机制）给出了一个否定性结果。这些算法只能在给定的样本点上评估查询。形式化地说，

定义 B.1（[53]）。若机制 $\mathcal{A}$ 在收到样本 $Y = (y_1, \ldots, y_t) \in X^t$ 和查询 $q: X \to \{-1, 0, 1\}$后，算法返回的答案仅是 $(q(y_1) \ldots, q(y_t))$ 的函数，则称该机制是自然的。特别地，它不能自行对其他数据点计算 q。

考虑下述自适应博弈变体，该博弈曾出现在图 1 （引言中）中，其中底层分布固定为均匀分布。参见算法 AdaptiveGameOne。

算法 3 AdaptiveGameOne($\mathcal{A}, \mathcal{B}, t, k$)

记号： t 是样本量， $n = 2000t$ 是定义域大小

。1. 机制 $\mathcal{A}$ 获取样本 $Y = (y_1, \ldots, y_t)$，其中包含来自 $[n]$ 上均匀分布的 t 个独立同分布样本。该样本不会提供给 $\mathcal{B}$。2. 对于轮次 $i = 1, 2, \ldots, k$：(a) 对手 $\mathcal{B}$ 指定查询 q_i : $[n] \to \{-1, 0, 1\}$(b) 机制 $\mathcal{A}$ 获取 q_i ，并返回答案 $a_i \in [-1, 1]$% 如果 $\mathcal{A}$ 是 自然数 ，则它只能获得 $(q_i(y_1) \ldots, q_i(y_t))$，而不是 q_i的完整描述。(c) 将 a_i 提供给 $\mathcal{B}$。3. 如果存在 $\exists i$ 使得 $\left| a_i - \frac{1}{n} \sum_{j \in [n]} q_i(j) \right| > 1/10$，则该游戏的结果为 1；否则为 0。

定理 B.2 ([53, 81])。 存在一个计算高效的对手 $\mathcal{B}$ 使得对于任意自然 机制 $\mathcal{A}$ 以及任意足够大的 t 和 $k = \Theta(t^2)$ 都有

$$\Pr[\mathit{AdaptiveGameOne}(\mathcal{A}, \mathcal{B}, t, k) = 1] > \frac{3}{4}.$$

我们指出，即使 M 在计算上不受限制，该定理仍然成立。也就是说，存在一个计算高效的对手 $\mathcal{A}$ ，它能在 ADA 问题中击败每一种自然机制 M ，即使底层分布是均匀分布也是如此。

我们的第一个观察是，即使在博弈的第 1 步中，机制选择其 t 个输入点，而非对它们进行均匀抽样，[53, 81] 的证明仍然成立。我们将这种机制称为选择机制。参见算法 AdaptiveGameTwo。

算法 4 AdaptiveGameTwo($\mathcal{A}, \mathcal{B}, t, k$)

记号： t 表示样本量， $n = 2000t$ 表示定义域大小。

1. 机制 $\mathcal{A}$ 选择一个大小为 t的子集 $Y = (y_1, \ldots, y_t) \subseteq [n]$ 。该子集不会告知 $\mathcal{B}$。 2. 对于轮次 $i = 1, 2, \ldots, k$: (a) 对手 $\mathcal{B}$ 指定一个查询 $q_i : [n] \to \{-1, 0, 1\}$(b) 机制 $\mathcal{A}$ 获取 q_i 并给出答案 $a_i \in [-1, 1]$% 如果 $\mathcal{A}$ 是自然机制，则它只会获得 $(q_i(y_1) \ldots, q_i(y_t))$，而不是 q_i的完整描述。(c) 将 a_i 提供给 $\mathcal{B}$3. 如果存在 $\exists i$ 使得 $\left| a_i - \frac{1}{n} \sum_{j \in [n]} q_i(j) \right| > 1/10$，则博弈的结果为 1；否则为 0。

定理 B.3 （ [53, 81]的扩展）。 存在一个计算高效的对手 $\mathcal{B}$ ，使得对于每个自然数 机制 $\mathcal{A}$ 以及所有足够大的 t 和 $k = \Theta(t^2)$ ，都有

$$\Pr[\mathit{AdaptiveGameTwo}(\mathcal{A}, \mathcal{B}, t, k) = 1] > \frac{3}{4}.$$

注 B.4。这个扩展对我们的目的而言并非必需，但它简化了我们的分析。回想一下，我们最终想要证明，对于能够

对底层分布进行预处理，并存储有限量与之相关的信息的高效机制，存在一个否定性结果。直观地说，利用我们对 MILR 的定义，我们将迫使此类机制只存储关于底层分布中有限多个点的有用信息。但这些点不一定是从底层分布中抽取的，因此，对我们而言，依赖于关于选择自然机制的困难性结果会更方便。

B.1 从选择性和自然机制到空间受限性

接下来，我们将选择性和自然机制的否定性结果（定理 B.3）“提升”，从而得到空间下界。为此，回顾第 AdaptiveGameSpace 节介绍的 4：机制最初会获得底层分布的完整描述，但必须将其压缩为 s 位摘要 z。

我们证明以下定理。

定理 B.5。假设存在一个 MILR 方案 $\Pi = (\text{Gen},\text{Param},\text{Enc},\text{Dec})$，其安全参数为 λ，并且能够抵抗空间受限预处理对手，达到 $(\Gamma,\overline{\tau})$-安全。设 $k \leq \text{poly}(\Gamma(\lambda))$ 是关于 λ 的函数。则存在一个运行时间为 $\text{poly}(\Gamma(\lambda))$ 的对手 $\mathcal{B}$，使得以下结论成立。设 $\mathcal{A}=(\mathcal{A}_1,\mathcal{A}_2)$ 是一个运行时间为 $\text{poly}(\Gamma(\lambda))$、空间复杂度为 s 的机制，且 $\overline{\tau}(\lambda,s) \leq O\left(\sqrt{k}\right)$（足够小）。那么，

$$\Pr[\mathit{AdaptiveGameSpace}(\mathcal{A},\mathcal{B},s,k)=1] > \frac{2}{3}.$$

此外，对手 $\mathcal{B}$ 定义的底层分布可以用$O(\sqrt{k}\cdot\lambda)$ 比特完整描述，可以在多项式$(\Gamma(\lambda))$ 时间内采样，并且从该分布中采样的元素可以用 $O(\lambda+\log(k))$ 比特表示。

证明。我们的任务是证明，存在一个对手，能够使在 AdaptiveGameSpace 中运行的每个（高效）机制都失败。我们知道，存在一个对手，能够使在 AdaptiveGameTwo 中运行的每个自然机制都失败。为了弥合这一差距，我们考虑一个名为 WrapA 的包装算法，它可以作为包装器用于为 AdaptiveGameSpace 设计的机制 $\mathcal{A}$，将其转化为在 AdaptiveGameTwo 中运行的自然机制。考虑算法 WrapA。

我们强调，算法 WrapA 不一定是计算高效的，因为计算 $J(\mathcal{A}_1,\ \vec{x},\ z,\ \vec{p})$ 可能并不高效。这并不影响我们的分析，因为定理 B.3 即使对于计算能力不受限的机制（只要它们是自然机制）也成立。实际上，对于每个机制 $\mathcal{A}$ (为 AdaptiveGameSpace设计）我们都有 WrapA$\circ\mathcal{A}$ 是一个 自然的 机制，可用于在 AdaptiveGameTwo中运行。因此，定理 $\mathcal{B}$ B.3 保证的对手 B.3 会使 WrapA$\circ\mathcal{A}$在 AdaptiveGameTwo中失败，前提是

$$\overline{\tau}(\lambda,s) = |I| \leq t = \Theta(\sqrt{k}), \tag{3}$$

其中，I 是算法5 的第 WrapA 步中的集合，t 是AdaptiveGameTwo中的样本量。我们继续分析，并假设不等式（3）成立。

回想一下，我们的任务是设计一个对手，使 $\mathcal{A}$ 在 AdaptiveGameSpace 中失效。为此，我们考虑算法 WrapA 的一个修改版本，称为 WrapB。

算法 5 WrapA

设置：算法 WrapA 可以访问一个算法 $\mathcal{A}=(\mathcal{A}_1,\mathcal{A}_2)$，其空间复杂度为 s（专为 AdaptiveGameSpace 设计），并与一个对手 $\mathcal{B}$（专为 AdaptiveGameTwo 设计）交互。1. 令 (Gen, Param, Enc, Dec) 为一个具有安全参数 λ 的 $(\Gamma,\bar{\tau})$-安全 MILR 方案。2. 从 Gen 中采样 n 个密钥 $x_1,\ldots,x_n\in\{0,1\}^\lambda$，并定义 $\mathcal{D}$ 为集合 $\{(j,x_j)\}_{j=1}^n$ 上的均匀分布。3. 令 $z\leftarrow\mathcal{A}_1(\mathcal{D})$。4. 从 Param 中采样 n 个公开参数 $p_1,\ldots,p_n\in\{0,1\}^{\mathrm{poly}(\lambda)}$。5. 令 $J\leftarrow J(\mathcal{A}_1,\vec{x},z,\vec{p})\subseteq[n]$ 为“隐藏坐标”集合，如定义 2.1 所示。记作 $I=[n]\setminus J$。集合 I 对应于 AdaptiveGameTwo 第 1 步中的子集 Y。6. 使用 z 实例化 $\mathcal{A}_2$。7. 对于轮次 $i=1,2,\ldots,k$：(a) 获得一个查询 q_i，其范围限于坐标 I。也就是说，对于每个 $j\in I$，我们获得 $q_i(j)$。(b) 对于 $j\in I$，令 $c_{i,j}=\mathrm{Enc}(x_i,p_i,q_i(j))$。对于 $j\notin I$，令 $c_{i,j}=\mathrm{Enc}(x_i,p_i,0)$。(c) 定义查询 f_i：$[n]\times\{0,1\}^\lambda\to\{-1,0,1\}$，其中 $f_i(j,x)=\mathrm{Dec}(x,p_j,c_{i,j})$。$f_i$ 的描述包含 $\{c_{i,j}\}_{j\in[n]}$ 和 $\{p_j\}_{j\in[n]}$。(d) 将 f_i 提供给 $\mathcal{A}_2$，并获得一个答案 $a_i\in[-1,1]$。(e) 输出 a_i。

算法 6 WrapB

设置：算法 WrapB 可以访问算法 $\mathcal{A}=(\mathcal{A}_1,\mathcal{A}_2)$，其空间复杂度为 s（为 AdaptiveGameSpace 设计），并与对手 $\mathcal{B}$（为 AdaptiveGameTwo 设计）交互。1. 设 (Gen, Param, Enc, Dec) 是一个 $(\Gamma,\bar{\tau})$-安全的 MILR 方案，其安全参数为 λ。2. 从 Gen 中采样 n 个密钥 $x_1,\ldots,x_n\in\{0,1\}^\lambda$，并定义 $\mathcal{D}$ 为集合 $\{(j,x_j)\}_{j=1}^n$ 上的均匀分布。3. 令 $z\leftarrow\mathcal{A}_1(\mathcal{D})$。4. 从 Param 中采样 n 个公开参数 $p_1,\ldots,p_n\in\{0,1\}^{\mathrm{poly}(\lambda)}$。5. 令 $J\leftarrow J(\mathcal{A}_1,\vec{k},z,\vec{p})\subseteq[n]$ 为定义 2.1 中所述的“隐藏坐标”集合。记为 $I=[n]\setminus J$。集合 I 对应于 AdaptiveGameTwo 第 1 步中的子集 Y。6. 使用 z 实例化 $\mathcal{A}_2$。7. 对于轮次 $i=1,2,\ldots,k$：(a) 获取一个查询 q_i（不限于坐标 I）。(b) 对于 $j\in[n]$，令 $c_{i,j}=\mathrm{Enc}(x_i,p_i,q_i(j))$。(c) 定义查询 f_i：$[n]\times\{0,1\}^\lambda\to\{-1,0,1\}$，其中 $f_i(j,x)=\mathrm{Dec}(x,p_j,c_{i,j})$。$f_i$ 的描述包含 $\{c_{i,j}\}_{j\in[n]}$ 和 $\{p_j\}_{j\in[n]}$。(d) 将 f_i 提供给 $\mathcal{A}_2$，并获得答案 $a_i\in[-1,1]$。(e) 输出 a_i。

与 WrapA 不同，经过修改的机制 WrapB 并不保证所得的 WrapB$\circ\mathcal{A}$ 是自然的。不过，正如我们接下来要说明的，对于每个计算高效机制 $\mathcal{A}$，WrapB$\circ\mathcal{A}$ 在 AdaptiveGameTwo 中仍然会失败。为此，考虑算法 WrapC，它是 WrapA 和 WrapB 的另一种变体，只能通过预言机 $\mathcal{E}$访问加密。我们将 WrapC 视为针对 MILR 方案的对手。

算法 7 WrapC

设定：算法 WrapC 可以访问一个空间复杂度为 s 的算法 $\mathcal{A}$=($\mathcal{A}_1, \mathcal{A}_2$)（为 AdaptiveGameSpace 设计），并与一个对手 $\mathcal{B}$ （为 AdaptiveGameTwo 设计）交互。它还可以访问加密预言机 $\mathcal{E}$。

输入：一个元素 z （据称由 $\mathcal{A}_1$ 计算得到），以及一组 n 公共参数p_1，.....，$p_n \in \{0,1\}^{\mathrm{poly}(\lambda)}$。

1. 用 z 实例化 $\mathcal{A}_2$。 2. 对于轮次 $i = 1, 2, \ldots, k$： (a) 获取一个查询 q_i （不受坐标集合的限制） (b) 对于 $j \in [n]$，令 $c_{i,j} = \mathcal{E}(i, q_i(j))$。 (c) 定义查询 f_i： $[n] \times \{0,1\}^\lambda \to \{-1, 0, 1\}$，其中 $f_i(j, x)$ $=\mathrm{Dec}(x, p_j, c_{i,j})$。 f_i 的描述包含 $\{c_{i,j}\}_{j\in[n]}$ 和 $\{p_j\}_{j\in[n]}$。 (d) 将 f_i 提供给 $\mathcal{A}_2$，并获取答案 $a_i \in [-1, 1]$。 (e) 输出 a_i。

注意，如果 WrapC 可以访问预言机 $\mathcal{E}_1$，那么它的行为与 WrapA 完全相同；如果它可以访问预言机 $\mathcal{E}_0$，那么它的行为与 WrapB 完全相同。形式化地说，

$$
\begin{aligned}
&\Pr\left[\mathtt{AdaptiveGameTwo}\left((\mathtt{WrapB}\circ\mathcal{A}), \mathcal{B}, t, k\right) = 1\right] \\
&= \Pr_{\substack{\vec{x},\vec{p},\mathcal{B},\mathcal{A}_2,\mathrm{Enc} \\ z\leftarrow\mathcal{A}_1(\vec{x}) \\ J\leftarrow J(\mathcal{A}_1,\vec{x},z,\vec{p})}}\left[\mathtt{AdaptiveGameTwo}\left((\mathtt{WrapC}^{\mathcal{E}_0(\vec{x},\vec{p},J,\cdot,\cdot)}(z,\vec{p})\circ\mathcal{A}), \mathcal{B}, t, k\right) = 1\right] \\
&\geq \Pr_{\substack{\vec{x},\vec{p},\mathcal{B},\mathcal{A}_2,\mathrm{Enc} \\ z\leftarrow\mathcal{A}_1(\vec{x}) \\ J\leftarrow J(\mathcal{A}_1,\vec{x},z,\vec{p})}}\left[\mathtt{AdaptiveGameTwo}\left((\mathtt{WrapC}^{\mathcal{E}_1(\vec{x},\vec{p},J,\cdot,\cdot)}(z,\vec{p})\circ\mathcal{A}), \mathcal{B}, t, k\right) = 1\right] - \mathrm{negl}(\Gamma(\lambda)) \\
&= \Pr\left[\mathtt{AdaptiveGameTwo}\left((\mathtt{WrapA}\circ\mathcal{A}), \mathcal{B}, t, k\right) = 1\right] - \mathrm{negl}(\Gamma(\lambda)) \\
&\geq \frac{3}{4} - \mathrm{negl}(\Gamma(\lambda)) \geq \frac{2}{3}.
\end{aligned}
$$

最后，注意到我们不必将 WrapB 与 $\mathcal{A}$ 组合（并将它们视为单个机制），而是可以将其与对手 $\mathcal{B}$ 组合，从而得到一个在 AdaptiveGameSpace 中进行博弈的组合对手。也就是说，

$$
\mathtt{AdaptiveGameTwo}\left((\mathtt{WrapB}\circ\mathcal{A}), \mathcal{B}, t, k\right) \equiv \mathtt{AdaptiveGameSpace}\left(\mathcal{A}, (\mathtt{WrapB}\circ\mathcal{B}), s, k\right),
$$

因此，

$$\begin{aligned}&\Pr\left[\texttt{AdaptiveGameSpace}\left(\mathcal{A},(\texttt{WrapB}\circ\mathcal{B}),s,k\right)=1\right]\\&\quad=\Pr\left[\texttt{AdaptiveGameTwo}\left((\texttt{WrapB}\circ\mathcal{A}),\mathcal{B},t,k\right)=1\right]\\&\quad\geq\frac{2}{3}.\end{aligned}$$

□

定理 4.1 现在可通过将定理 B.5 应用于我们的 MILR 方案（如定理 2.3所述）得到。

C 第 5 节中缺失的证明

定理 *5.1* 的证明。 我们证明均匀对手的情形。非均匀对手的证明类似。设 $\mathcal{B}$ 是一个关于 $\Gamma(\lambda)$ 的多项式时间对手，用于 Π。记

$$\mathrm{Adv}_{\mathcal{B}}(\lambda)=\left|\Pr_{\vec{x},\vec{g},\mathcal{B},\mathrm{Enc}}\left[\mathcal{B}^{\mathcal{E}_0(\vec{x},\vec{g},[n],\cdot,\cdot)}(\vec{g})=1\right]-\Pr_{\vec{x},\vec{g},\mathcal{B},\mathrm{Enc}}\left[\mathcal{B}^{\mathcal{E}_1(\vec{x},\vec{g},[n],\cdot,\cdot)}(\vec{g})=1\right]\right|.$$

我们的目标是证明 $\mathrm{Adv}_{\mathcal{B}}(\lambda)\leq\mathrm{negl}(\Gamma(\lambda))$.

从 Π' 的（多实例）安全性归约 Π 的安全性。我们将使用 $\mathcal{B}$ 构造一个对手 $\mathcal{B}_1$，针对 Π' 的 n 个实例，且这些实例使用独立密钥。我们定义与 Π' 对应的预言机。

1. $\mathcal{E}_1^{(1)}(\vec{x'},[n],\cdot,\cdot)$ 接收一个密钥 $j\in[n]$ 的索引和一条消息 m 作为输入，并返回 $\mathrm{Enc}'(x'_j,m)$。
2. $\mathcal{E}_0^{(1)}(\vec{x'},[n],\cdot,\cdot)$ 接收相同的输入，并返回 $\mathrm{Enc}'(x'_j,0)$。

$\mathcal{B}_1$ 的优势定义为

$$\mathrm{Adv}_{\mathcal{B}_1}(\lambda)=\left|\Pr_{\vec{x'},\mathcal{B}_1,\mathrm{Enc}'}\left[\mathcal{B}_1^{\mathcal{E}_0^{(1)}(\vec{x'},[n],\cdot,\cdot)}()=1\right]-\Pr_{\vec{x'},\mathcal{B}_1,\mathrm{Enc}'}\left[\mathcal{B}_1^{\mathcal{E}_1^{(1)}(\vec{x'},[n],\cdot,\cdot)}()=1\right]\right|.$$

算法 8 $\mathcal{B}_1^{\mathcal{E}^{(1)}\,\vec{x'}\,(\cdot)}()$

设置： $\mathcal{B}_1$ 针对 n 个 Π' 实例运行，并可访问 $\mathcal{B}^{\mathcal{E}_{(\cdot)}(\vec{x},\ \vec{g},\ [n],\ \cdot,\ \cdot)}(\vec{g})$，后者针对$\Pi$运行。

1. $\mathcal{B}_1$ 选择 $\vec{g}\leftarrow_R\mathcal{G}^n$ 并将其作为输入提供给 $\mathcal{B}$ 。2. $\mathcal{B}_1$ 将 $\mathcal{B}$ 的每个查询（j， m）转发给其预言机，并将答案转发回 $\mathcal{B}$。3. $\mathcal{B}_1$ 输出与 $\mathcal{B}$相同的结果。

对手 $\mathcal{B}_1$ 在算法8 中给出，我们将分析其运行时间和优势。

为此，我们首先考虑给定 $\vec{g}$时， Π的哈希密钥的分布。每个密钥$x\in\{0,1\}^\lambda$ 均独立且均匀地选取，并使用一个

从通用哈希函数族中均匀选取的哈希函数，将其哈希为比特。剩余哈希引理的一个标准应用 [54] 表明，每个密钥的分布与 $2^{-\Omega(\lambda)}$均匀分布的统计距离至多为。由于密钥是独立选取的，且 $n < 2^{-o(\lambda)}$，应用三角不等式（对所有 n 个密钥的统计距离求和）可知，它们的联合分布与 $2^{-\Omega(\lambda)}$ 均匀分布的统计距离至多为（给定 $\vec{g}$）。

由于 $\mathcal{B}_1$ 的密钥 $\vec{x'} \in (\{0,1\}^{\lambda'})^n$ 独立于 $\vec{g}$ 均匀分布，因此对 $\mathcal{B}$ 的模拟会产生 $2^{-\Omega(\lambda)}$ 的加性损失，即

$$\mathrm{Adv}_{\mathcal{B}_1}(\lambda) \geq \mathrm{Adv}_{\mathcal{B}}(\lambda) - 2^{-\Omega(\lambda)}.$$

$\mathcal{B}_1$ 的运行时间被 poly(Γ(λ)) 所界定（与 $\mathcal{B}$ 的运行时间相同）。

将 Π' 的（多实例）安全性归约为单实例安全性。接下来，我们利用多实例对手 $\mathcal{B}_1$，构造一个针对单个 Π' 实例（具有 $n = 1$）的对手 $\mathcal{B}_2$。$\mathcal{B}_2$ 的优势定义为

$$\mathrm{Adv}_{\mathcal{B}_2}(\lambda) = \left| \Pr_{\vec{x'},\mathcal{B}_2,\mathrm{Enc}'} \left[\mathcal{B}_2^{\mathcal{E}_0^{(1)}(\vec{x'},[1],\cdot,\cdot)}() = 1 \right] - \Pr_{\vec{x'},\mathcal{B}_2,\mathrm{Enc}'} \left[\mathcal{B}_2^{\mathcal{E}_1^{(1)}(\vec{x'},[1],\cdot,\cdot)}() = 1 \right] \right|.$$

算法 9 $\mathcal{B}^{\mathcal{E}^{(1)}_{\cdot}()(\vec{x'},\ [1],\ \cdot,\ \cdot)}$ 2 ()

设定：$\mathcal{B}_2$ 针对 Π' 的单个实例运行，并可访问 $\mathcal{B}^{\mathcal{E}^{(1)}_{\cdot}()(\vec{x'},\ [n],\ \cdot,\ \cdot)}$ 1 ()，其针对 Π' 的 n 个实例运行。

1. $\mathcal{B}_2$ 均匀随机选择一个索引 $i \in [n]$，并选择 $n-1$ 个密钥x'_1、.....、x'_{i-1}、x'_{i+1}、.....、x'_n，每个密钥都从 $\{0,1\}^{\lambda'}$中均匀随机选取。2. $\mathcal{B}_2$ 按如下方式回答 $\mathcal{B}_1$ 的每个查询 (j,m):

   • 如果 $i = j$，$\mathcal{B}_2$ 就将查询 $(1,m)$ 转发给其预言机，并将答案返回给 $\mathcal{B}_1$。
   • 如果 $j < i$，$\mathcal{B}_2$ 就将 $\mathrm{Enc}'(x'_j, m)$ 返回给 $\mathcal{B}_1$。• 如果 $j > i$，$\mathcal{B}_2$ 就将 $\mathrm{Enc}'(x'_j, 0)$ 返回给 $\mathcal{B}_1$。

3. $\mathcal{B}_2$ 输出与 $\mathcal{B}_1$相同的结果。

对手 $\mathcal{B}_2$ 在算法9中给出。为了关联 $\mathrm{Adv}^{\mathcal{B}_2}(\lambda)$ 与 $\mathrm{Adv}^{\mathcal{B}_1}(\lambda)$，我们使用混合论证。具体而言，对于 $i \in \{0,\ldots,n\}$ 定义 $\mathcal{E}^{(i)}$ 如下。

- $\mathcal{E}_{(i)}(\vec{x'},[n],\cdot,\cdot)$ 将密钥 $j \in [n]$ 的索引和消息 m作为输入。若 $j \leq i$，则返回 $\mathrm{Enc}'(x'_j, m)$。否则，$j > i$并返回 $\mathrm{Enc}'(x'_j, 0)$。

注意，$\mathcal{E}^{(0)} = \mathcal{E}_0^{(1)}$ 和 $\mathcal{E}^{(n)} = \mathcal{E}_1^{(1)}$。对于 $i \in [n]$，定义

$$\mathrm{Adv}^{(i)}_{\mathcal{B}_1}(\lambda) = \left| \Pr_{\vec{x'},\mathcal{B}_1,\mathrm{Enc}'} \left[\mathcal{B}_1^{\mathcal{E}^{(i-1)}(\vec{x'},[n],\cdot,\cdot)}() = 1 \right] - \Pr_{\vec{x'},\mathcal{B}_1,\mathrm{Enc}'} \left[\mathcal{B}_1^{\mathcal{E}^{(i)}(\vec{x'},[n],\cdot,\cdot)}() = 1 \right] \right|.$$

我们有

$$\mathrm{Adv}_{\mathcal{B}_1}(\lambda) = \sum_{i=1}^{n} \left(\Pr_{\vec{x'},\mathcal{B}_1,\mathrm{Enc}'} \left[\mathcal{B}_1^{\mathcal{E}^{(i-1)}(\vec{x'},[n],\cdot,\cdot)}() = 1 \right] - \Pr_{\vec{x'},\mathcal{B}_1,\mathrm{Enc}'} \left[\mathcal{B}_1^{\mathcal{E}^{(i)}(\vec{x'},[n],\cdot,\cdot)}() = 1 \right] \right) \leq$$

$$\sum_{i=1}^{n} \left| \Pr_{\vec{x'},\mathcal{B}_1,\mathrm{Enc}'} \left[\mathcal{B}_1^{\mathcal{E}^{(i-1)}(\vec{x'},[n],\cdot,\cdot)}() = 1 \right] - \Pr_{\vec{x'},\mathcal{B}_1,\mathrm{Enc}'} \left[\mathcal{B}_1^{\mathcal{E}^{(i)}(\vec{x'},[n],\cdot,\cdot)}() = 1 \right] \right| = \sum_{i=1}^{n} \mathrm{Adv}_{\mathcal{B}_1}^{(i)}(\lambda).$$

根据 $\mathcal{B}_2$ 的预言机，$\mathcal{B}_1$ 会与 $\mathcal{E}^{(i-1)}$ 或 $\mathcal{E}^{(i)}$ 交互，因此在选定 i 的条件下，$\mathcal{B}_2$ 的优势为 $\mathrm{Adv}_{\mathcal{B}_1}^{(i)}(\lambda)$。由于 $i \in [n]$ 服从均匀分布，

$$\mathrm{Adv}_{\mathcal{B}_2}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Adv}_{\mathcal{B}_1}^{(i)}(\lambda) \geq \frac{1}{n} \mathrm{Adv}_{\mathcal{B}_1}(\lambda).$$

$\mathcal{B}_2$ 的运行时间上界为 $n \cdot \mathrm{poly}(\Gamma(\lambda)) = \mathrm{poly}(\Gamma(\lambda))$。因此，关于 Π' 的安全性假设意味着 $\mathrm{Adv}_{\mathcal{B}_2}(\lambda) \leq \mathrm{negl}(\Gamma(\lambda))$。由此我们得出，$\mathrm{Adv}_{\mathcal{B}_1}(\lambda)$，进而 $\mathrm{Adv}_{\mathcal{B}}(\lambda)$，在 $\Gamma(\lambda)$ 中是可忽略的。

□

第 D 节中缺失的证明6

以下结果取自 [25, 命题 2]（其原始版本归功于 [49, 62]）。

引理 D.1。固定 $z \in \{0,1\}^s$，并令 $s_z = \lambda \cdot n - H_\infty(X_z)$ 为 X_z 的最小熵亏损。令 $\gamma > 0$ 为任意值。那么，对于每个 $\delta > 0$，X_z 与有限个 $(k, 1-\delta)$-稠密源的凸组合之间的距离至多为 γ

$$k = \frac{s_z + \log 1/\gamma}{\delta \cdot \lambda}.$$

现在证明以下引理。

引理 D.2. 固定 $z \in \{0,1\}^s$，并令 $s_z = \lambda \cdot n - H_\infty(X_z)$。

令 $\mathcal{G}$ 为定义域为 $\{0,1\}^\lambda$、值域为 $\{0,1\}^{\lambda'}$ 的通用哈希函数族，并令 $G \leftarrow_R \mathcal{G}^n$。令 $\delta > 0$、$\gamma > 0$ 为参数，并假设 $(1-\delta)\lambda > \lambda' + \log\, n + 1$。

那么，存在一个凸组合族 $V^*_{G,\ z} = \{V^*_{\vec{g},z}\}_{\vec{g} \in (\mathcal{G})^n}$，其成员为 k 位固定 $(n, 2^{\lambda'})$-源，其中

$$k = \frac{s_z + \log 1/\gamma}{\delta \cdot \lambda}$$

使得

$$\Delta[(G, G(X_z)), (G, V^*_{G,z})] \leq \sqrt{2^{-(1-\delta)\lambda + \lambda' + \log n}} + \gamma.$$

引理D.2 基于引理 6.4 和 D.1。

证明。固定 $z \in \{0,1\}^s$。令 s_z、$\gamma > 0$、$\delta > 0$ 如引理所述。

然后，根据引理D.1（应用参数 s_z、γ、δ），$\Delta[X_z, X^*_z] \leq \gamma$，其中 X^*_z是有限多个 $(k, 1-\delta)$-稠密 $(n, 2^\lambda)$-源的凸组合，且 $k = {}^{s_z + \log_{\delta \cdot \lambda}}{}^{1/\gamma}$。

设 X^* 是凸组合 X_z^* 中的一个 $(k, 1-\delta)$-稠密源。设 V^* 是对应的 k-位固定源（X^* 和 V^* 在相同的索引上固定为相同的值）。

令 $J \subseteq [n]$ 表示 X^* 未固定的索引集合，并进一步记 $t := |J| \geq n - k$。那么，X_J^* 是一个 $(t, 2^\lambda)$-源，且是 $(1-\delta)$-稠密的。

由于我们假设 $(1-\delta)\lambda > \lambda' + \log n + 1 > \lambda' + \log t + 1$，因此可以将引理 6.4 应用于 X_J^*（其中 δ 如上所述）。我们得出

$$\Delta[(G, G(X_J^*)), (G, U)] \leq \sqrt{2^{-(1-\delta)\lambda+\lambda'+\log t}} < \sqrt{2^{-(1-\delta)\lambda+\lambda'+\log n}},$$

其中 $U \leftarrow_R (\{0,1\}^{\lambda'})^t$。

由于 $\overline{J}$ 在 X^* 和 V^* 中其余的索引都固定为相同的值，

$$\Delta[(G, G(X^*)), (G, V_G^*)] \leq \sqrt{2^{-(1-\delta)\lambda+\lambda'+\log n}},$$

其中 $V_G^* \in (\{0,1\}^{\lambda'})^n$ 是一个由 $|\overline{J}|$ 位固定源组成的族，它由每个 $\vec{g} \in \mathcal{G}^n$ 和 V^* 得到：将 $\overline{J}$ 的索引设为它们在 $\vec{g}(V^*)$ 中对应的值，同时 J 的索引在 $(\{0,1\}^{\lambda'})^t$ 中服从均匀分布。

令 $(G, V_{G,z}^*)$ 由将 $(G, G(X_z^*))$ 中的每个 $(G, G(X^*))$ 替换为相应的 (G, V_G^*) 得到。那么，上述结论意味着

$$\Delta[(G, G(X_z^*)), (G, V_{G,z}^*)] \leq \sqrt{2^{-(1-\delta)\lambda+\lambda'+\log n}}.$$

因此，

$$\begin{aligned}&\Delta[(G, G(X_z)), (G, V_{G,z}^*)] \leq \\ &\Delta[(G, G(X_z)), (G, G(X_z^*))] + \Delta[(G, G(X_z^*)), (G, V_{G,z}^*)] \leq \gamma + \sqrt{2^{-(1-\delta)\lambda+\lambda'+\log n}},\end{aligned}$$

这里用到了以下事实：

$$\Delta[(G, G(X_z)), (G, G(X_z^*))] \leq \Delta[X_z, X_z^*] \leq \gamma,$$

因为 G 独立于 X_z 和 X_z^*。

□

引理 **6.2** 的证明。引理6.2 基于引理 D.2 以及下面的命题；该命题给出了 X 和 F 使得 $H_\infty(X_z)$ 的概率上界，并表明其概率很低。

命题 1。设 $s' > s$。那么，$\Pr_{z \leftarrow F(X)}[H_\infty(X_z) < n \cdot \lambda - s'] < 2^{s-s'}$。

证明。首先，由于随机函数是确定性函数的凸组合，我们不失一般性地假设 F 是确定性的。固定满足 $H_\infty(X_z) < n \cdot \lambda - s'$ 的 $z \in \{0,1\}^s$。由于 $X \in (\{0,1\}^\lambda)^n$ 服从均匀分布，因此 $\Pr_X[F(X) = z] < 2^{-s'}$。

设 $K = \{z \in \{0,1\}^s \mid H_\infty(X_z) < n \cdot \lambda - s'\}$。那么，

$$\Pr_{z \leftarrow F(X)}[H_\infty(X_z) < n \cdot \lambda - s'] = \sum_{z \in K} \Pr_X[F(X) = z] < \sum_{z \in K} 2^{-s'} \leq 2^s \cdot 2^{-s'} = 2^{s-s'}.$$

□

引理的证明 6.2。 设 $\delta > 0$， $\gamma > 0$， $s' > s$ 为参数。

考虑 $z \in \{0,1\}^s$ 具有 $s_z \leq s'$（即 $H_\infty(X_z) \geq n \cdot \lambda - s'$）。由于 $(1-\delta)\lambda > \lambda' + \log\, n + 1$，我们调用引理 D.2 并推得存在一个由有限多个$k$-位固定分布构成的凸组合族 $V^*_{G,z}$，使得

$$\Delta[(G, G(X_z)), (G, G(V^*_{G,z}))] \leq \sqrt{2^{-(1-\delta)\lambda+\lambda'+\log n}} + \gamma, \tag{4}$$

and

$$k = \frac{s' + \log 1/\gamma}{\delta \cdot \lambda}.$$

定义该族

$$V_{G,Z} = \{V_{G,z}\}_{z\in\{0,1\}^s} = \{V_{\vec{g},z}\}_{\vec{g}\in(\mathcal{G})^n, z\in\{0,1\}^s}$$

如下：对于每个满足 $s_z \leq s'$ 的 z，令 $V_{G,z} := V^*_{G,z}$；而对于每个满足 $s_z > s'$ 的 z，令$V_{G,z}$为 $(\{0,1\}^{\lambda'})^n$ 上的均匀分布。于是，$V_{G,Z}$ 是由 k-位固定的 $(n, 2^{\lambda'})$-源构成的凸组合族$V_{\vec{g},z}$，其中 $k = {}^{s'+}{}_{\log_{\delta\cdot\lambda}}{}^{1/\gamma}$。为完成证明，剩下需要界~~定$\Delta[(G,$~~$Z, G(X)), (G, Z, V_{G,Z})]$。

将 $\mathcal{E}$ 记为事件 $s_z < s'$。那么，

$$\begin{aligned}
&\Delta[(G, Z, G(X)), (G, Z, V_{G,Z})] = \mathop{\mathbb{E}}_{z\leftarrow F(X)} \Delta[(G, G(X_z)), (G, V_{G,z})] \leq \\
&\Pr[\mathcal{E}] \cdot 1 + \mathop{\mathbb{E}}_{z\leftarrow F(X)|\neg\mathcal{E}} \Delta[(G, G(X_z)), (G, V_{G,z})] \leq \\
&2^{s-s'} + \sqrt{2^{-(1-\delta)\lambda+\lambda'+\log n}} + \gamma,
\end{aligned}$$

其中，最后一个不等式依据命题 1 和（4）。

□

E 通信复杂度的应用

在双人单向通信博弈中，输入 A 和 B 分别交给 Alice 和 Bob。目标是让 Alice 向 Bob 发送尽可能少的信息，使 Bob 能够针对某个预先确定的函数 B 计算 $f(A, B)$。协议 Π 的通信代价是 Alice 在所有可能的输入 X 下发送的最大消息的比特数；而（随机化）通信复杂度则是在成功概率至少为 2 3 的协议中，最小的通信代价。在分布式设定中， X 和 Y 还会从某个已知的底层分布中抽取。

在我们的设定中，假设 Alice 有 n 个独立且均匀分布的数 $a_1, \ldots, a_n$，满足以下两种情况之一：对于所有 $i \in [n]$，$a_i \in GF(p)$；或者对于所有 $i \in [n]$，当 t 足够大时， $a_i \in GF(2^t)$。再假设 Bob 有 n 个来自同一域的独立且均匀分布的数 $b_1, \ldots, b_n$，满足 $GF(p)$ 或 $GF(2^t)$。那么，对于任意函数 $f(\langle a_1, b_1\rangle, \ldots, \langle a_n, b_n\rangle)$，其中点积在 $GF(2)$ 上计算；或者 $f(a_1 \cdot b_1, \ldots, a_n \cdot b_m n)$，其中乘积在 $GF(p)$ 上计算，该函数都具有如下性质：以概率 σ 计算 f 的随机化单向通信复杂度，等于 Alice 需要发送给 Bob 的、来自 a_1,... , a_n 的样本数量，以便以概率

$\sigma-\varepsilon$。很容易为许多这类问题证明采样下界，例如 $\sum_i a_i \cdot b_i$ (模 p) 或 MAJ($\langle a_1, b_1\rangle, \ldots, \langle a_n, b_n\rangle$)；根据上述结果，这会立即转化为通信复杂度下界。这些下界背后的主要直觉是，数值 $b_1, \ldots, b_n$ 可以看作 Bob 拥有的哈希函数。

在推论 6.3的背景下，Alice 的输入 A 是哈希函数的输入 $X = (X_1, \ldots..., X_n)$，而 Bob 的输入 B 是哈希函数集合 $G = (G_1, \ldots..., G_n)$。特别地，我们有 $X_i \in \{0,1\}^\lambda$ 以及 $G_i \in \{0,1\}^t$ 对于每个 $i \in [n]$。非正式地说，推论 6.3 表明，如果一个抽样协议需要 $\Omega(\eta^2 n)$ 个样本才能达到 $\frac{1}{2} + \frac{\eta}{2}$ 的成功概率，即 $\frac{\eta}{2}$ 优势，那么任何达到 $\frac{1}{2} + \eta$ 的成功概率，即 η 优势的算法，都必须使用 $\Omega(\eta^2 n\lambda)$ 的通信量。

定理 E.1. 设将 $X = (X_1, \ldots..., X_n)$（对于所有 $i \in [n]$， $X_i \in \{0,1\}^\lambda$）给 Alice，并将 $G = (G_1, \ldots..., G_n)$（对于所有 $i \in [n]$， $G_i \in \{0,1\}^t$）给 Bob。假设对于函数 f： $(\{0,1\}^\lambda)^n \times (\{0,1\}^\lambda)^n \to \{0,1\}$，任何计算 $f(X,G)$ 且成功概率至少为 $\frac{1}{2} + \frac{\eta}{2}$ 的协议 Π，都需要 k 个样本 (X_i, G_i)，其中 k 是依赖于 η 的某个量。那么，任何计算 $f(X,G)$ 且成功概率至少为 $\frac{1}{2} + \eta$ 的单向协议，都必须使用 $\Omega(s)$ 的通信量，其中 $s = \frac{k\lambda}{2} - 2\log_2 \frac{6}{\eta}$。

证明。我们将 h 定义为满足以下条件的函数：

$$f(X,G) = h(G_1(X_1), G_2(X_2), \ldots, G_n(X_n)).$$

在推论 $t = \lambda' = 1$、 $\delta = 1/2$、 $n < 2^{\lambda/3}$、 $\gamma = \eta/6$、 $s = \Theta(\eta^2 n\lambda)$ 和 $s' = s + \log_2(6/\eta)$ 的表述中，取 6.3，那么当 λ 足够大，使得 $2^{-c\lambda} < \frac{\eta}{6}$ 对某个固定常数$c > 0$ 成立时，我们有

$$\Delta[P_{(G,Z,G(X))}, P_{(G,Z,V)}] \le \sqrt{2^{-(1-\delta)\lambda+\lambda'+\log n}} + \gamma + 2^{s-s'} \le \sqrt{2^{-\lambda/2+1+\lambda/3}} + \eta/6 + \eta/6 \le \eta/2,$$

其中 V是一个 $((G,Z)\ k)$-凸位固定 $(n, 2^{\lambda'})$-源，并且 $k = {}^{2}(s+2\log_2(6/\eta))\lambda$。

假设存在一个协议 Π，它以大于 $\frac{1}{2} + \eta$ 的概率计算 $f(G,X)$，并使用 $o(s)$ 空间。根据推论 6.3，成功概率是所有满足 $k' < k$ 的 $((G,Z)\ k')$-凸位固定 $(n, 2^{\lambda'})$-源对应成功概率的凸组合。因此，存在一个协议 Π'，由 Alice 生成消息 z，使得 Bob 在 g 上的成功概率至少为 $\frac{1}{2} + \eta$。根据全概率公式，Bob 的成功概率等于在凸组合中以每个 k'-位固定源为条件时 Bob 的成功概率的加权平均，因此存在一个成功概率至少为 $\frac{1}{2} + \frac{\eta}{2}$ 的 k'-位固定源。于是，Bob 在 g_1、……、g_n 上的后验分布，在至多包含 k' 个索引的固定集合 S 之外的所有坐标上联合均匀分布，且统计距离可忽略。由于每个 g_i 都是通用的，Bob 的后验分布与 Alice 对 S 中的坐标进行采样并将这些坐标发送给 Bob 时的分布相同。然而，由于 $|S| \le k' < k$，这与以下假设矛盾：任何以至少 $\frac{1}{2} + \frac{\eta}{2}$ 的概率计算 $f(X,G)$ 的协议 Π，对于某个依赖于 η 的量 k，都需要 k 个样本 (X_i, G_i)。

□

我们可以在此情形下将推论 6.3 应用于任何具有如下形式的关系 $f(X,\ G)$

$$h(G_1(X_1), G_2(X_2), \ldots, G_n(X_n))$$

对于某个关系 h，前提是 G_i 相互独立，且取自值域为 $\{0,1\}^{\lambda'}$ 的通用哈希函数族。事实上，将该定理中的 Z 视为从 Alice 发送给 Bob 的消息，推论 6.3 表明，分布 $P_{(G,Z,G(X))}$ 和 $P_{(G,Z,V)}$ 的变差距离很小，其中 V 是一个 $((G,Z)\ k)$-凸比特固定源。因此，Bob 在获知 $G=g$ 和 $Z=z$ 以及某些特定取值 g 和 z 后，无法区分 $g(X)$ 与 V；其中 V 是一个 $((g,z)\ k)$-凸比特固定源，这意味着在给定 $G=g$ 和 $Z=z$ 的条件下，V 是 k-位固定源的分布的凸组合，也就是说，是定义在 $(\{0,1\}^{\lambda'})^n$ 上的分布，在至多 k 个坐标上固定，而其余坐标服从均匀分布。

由于对 g 的平均成功概率至少为 $1/2+\eta$，这意味着存在一个 g，使其成功概率至少为 $1/2+\eta$；我们可以使用这个特定的 $g=(g_1,\ldots,g_n)$ 推导下界。类似地，我们可以固定 Alice 的一条消息 z，使得 Bob 在 g 上的成功概率至少为 $1/2+\eta$。现在，根据全概率公式，Bob 的成功概率是 Bob 在构成 V 的凸组合中的每个 k-位固定源条件下的成功概率的加权平均；因此，如果 Bob 在 (g,z) 上的成功概率至少为 $1/2+\eta$，那么就存在一个 k- 位固定源，其成功概率至少为 $1/2+\eta$。这又意味着，Bob 关于 $g_i(X_i)$ 的后验分布，在除固定索引集合 S 之外的所有坐标上联合均匀分布，且统计距离可忽略不计。

上述结果特别意味着，在 S 之外，g_i 不可能是零函数，否则 $g_i(X_i)$ 就不可能服从均匀分布。但对于许多通用族，如果固定 $g_i\neq 0$，那么 $g_i(X_i)$ 服从均匀分布就意味着 X_i 服从均匀分布。特别是，当 $g_i(X_i)=\langle g_i,X_i\rangle$ 模 2 时，这一结论成立。在这种情况下，Bob 的后验分布与 Alice 仅对坐标 $i\in S$ 进行抽样并将对应的 X_i 个值发送给 Bob 后得到的分布相同。因此，如果我们证明成功概率为 $1/2+\eta/2$ 的采样下界，就能得到成功概率为 $1/2+\eta$ 的任意单轮通信下界，前提是参数设置得当，使统计距离至多为例如 $\eta/2$。

例如，假设 h 是多数函数，并且对每个 i，对于比特串 g_i 和 X_i，$g_i(X_i)=\langle g_i,X_i\rangle$ 模 2。众所周知，所有此类 g_i 构成一个通用族。要在 (G,X) 上的均匀分布下达到 $1/2+\eta$ 的成功概率，假设 $\lambda=\Omega(\log 1/\eta)$，使得 $g_i=0^\lambda$ 不太可能发生（并且对于任意固定的 $g_i\neq 0^\lambda$，当 X_i 服从均匀分布时，$\langle g_i,X_i\rangle$ 服从均匀分布），标准结论是：采样协议必须对 $\Theta(\eta^2 n)$ 个坐标 i 进行抽样并发送给 Bob，然后由 Bob 输出经验多数。在推论 6.3 中，我们可以令 $\lambda'=1$, $\delta=1/2$, $n<2^{\lambda/3}$, $\gamma=\eta/6$, $s=\Theta(\eta^2 n\lambda)$，以及 $s'=s+\log_2(6/\eta)$，从而得出对于 $\eta/6>2^{-c\lambda}$ 和一个较小的常数 $c>0$，

$$\Delta[P_{(G,Z,G(X))},P_{(G,Z,V)}]\le\sqrt{2^{-(1-\delta)\lambda+\lambda'+\log n}}+\gamma+2^{s-s'}\le\sqrt{2^{-\lambda/2+1+\lambda/3}}+\eta/6+\eta/6\le\eta/2,$$

其中 V 是一个 $((G,Z)\ k)$-凸比特固定 $(n,2^{\lambda'})$-源，且 $k=$ $^{2}(s+2$ 对数$_2(6/\eta))\lambda$。根据采样下界，要达到 $1/2+\eta/2$ 成功概率，需要 $k=\Omega(\eta^2 n\lambda)$，这意味着作为单向任意通信下界，需要$s=\Omega(\eta^2 n\lambda-2$对数$_2(6/\eta))$，因为单向协议的成功概率至多为 $1/2+\eta/2+\eta/2=1/2+\eta$。

参考文献

[1] Alabi, D., Ben-Eliezer, O., Chaturvedi, A.：有界空间差分隐私分位数。CoRR abs/2201.03380（2022）2[2] Alon, N., Livni, R., Malliaris, M., Moran, S.：私有PAC学习蕴含有限Littlestone维度。载于：STOC。第852–860页。ACM（2019）8[3] Bassily, R., Freund, Y.：典型稳定性。arXiv预印本 arXiv:1604.03336（2016）5[4] Bassily, R., Nissim, K., Smith, A. D., Steinke, T., Stemmer, U., Ullman, J.R.：自适应数据分析的算法稳定性。SIAM计算杂志 50(3)（2021）5[5] Bassily, R., Nissim, K., Stemmer, U., Thakurta, A.：实用的局部隐私重频项。机器学习研究杂志 21，16:1–16:42（2020）2[6] Beame, P., Gharan, S.O., Yang, X.：从随机求值中学习有限函数的时间—空间权衡及其在多项式中的应用。载于：学习理论会议，COLT。第843–856页（2018）8[7] Beimel, A., Nissim, K., Omri, E.：分布式私有数据分析：同时解决“如何”和“是什么”。载于：CRYPTO。计算机科学讲义，第5157卷，第451–468页。施普普林格（2008）12[8] Bellare, M., Rogaway, P.：随机预言机具有实用性：设计高效协议的范式。载于：Denning, D.E., Pyle, R., Ganesan, R., Sandhu, R.S., Ashby, V.（编）：CCS 1993。第62–73页。ACM（1993）1[9] Blocki, J., Blum, A., Datta, A., Sheffet, O.：Johnson-Lindenstrauss 变换本身保留差分隐私。载于：FOCS。第410–419页。IEEE计算机学会（2012）2[10] Bolot, J., Fawaz, N., Muthukrishnan, S., Nikolov, A., Taft, N.：数据流上的私有衰减谓词和。载于：ICDT。第284–295页。ACM（2013）2[11] Boneh, D., Shaw, J.：数字数据的抗串谋指纹编码。IEEE信息论汇刊 44(5)，1897–1905（1998）12, 13[12] Brown, G., Bun, M., Feldman, V., Smith, A., Talwar, K.：何时必须记忆无关训练数据才能实现高精度学习？载于：第53届ACM SIGACT计算理论研讨会论文集。第123–132页（2021）8[13] Bu, Z., Gopi, S., Kulkarni, J., Lee, Y.T., Shen, H., Tantipongpipat, U.：通过JL投影实现快速且高效利用内存的差分隐私随机梯度下降。神经信息处理系统进展 34，19680–19691（2021）2[14] Bun, M., Nelson, J., Stemmer, U.：重击者与局部隐私结构。ACM算法汇刊 15(4)，51:1–51:40（2019）2[15] Bun, M., Nissim, K., Stemmer, U., Vadhan, S.P.：阈值函数的差分隐私发布与学习。载于：FOCS。第634–649页。IEEE计算机学会（2015）8[16] Bun, M., Ullman, J.R., Vadhan, S.P.：指纹编码与近似差分隐私的代价。载于：STOC。第1–10页。ACM（2014）7, 8, 12, 14

[17] Canetti, R., Dodis, Y., Halevi, S., Kushilevitz, E., Sahai, A.：抗暴露函数与全有或全无变换。载于：Preneel, B.（编）《EUROCRYPT 2000》。计算机科学讲义，第1807卷，第453–469页。施普林格（2000年）6[18] Chan, T.H.H., Li, M., Shi, E., Xu, W.：基于分布式数据流的重击者差分隐私持续监测。载于：隐私增强技术国际研讨会。第140–159页。施普林格（2012年）2[19] Chattopadhyay, A., Filmus, Y., Koroth, S., Meir, O., Pitassi, T.：利用内积实现BPP的查询到通信提升。载于：第46届自动机、语言与编程国际学术讨论会（ICALP 2019）。Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik（2019年）8[20] Chattopadhyay, A., Filmus, Y., Koroth, S., Meir, O., Pitassi, T.：利用低差异度小工具实现查询到通信提升。SIAM计算杂志 50(1)，171–210（2021年）8[21] Chattopadhyay, A., Kouck`y, M., Loff, B., Mukhopadhyay, S.：模拟胜过丰富性：新的数据结构下界。载于：第50届ACM SIGACT计算理论年会论文集。第1013–1020页（2018年）8[22] Chattopadhyay, A., Kouck`y, M., Loff, B., Mukhopadhyay, S.：通过伪随机性质得到的模拟定理。计算复杂性28(4)，617–659（2019年）8[23] Chien, S., Ligett, K., McGregor, A.：稳健统计与分布检验的空间高效估计。载于：计算机科学创新——ICS论文集。第251–265页（2010年）8[24] Coretti, S., Dodis, Y., Guo, S.：随机置换、理想密码和通用群模型中的非一致性界。载于：Shacham, H., Boldyreva, A.（编）《CRYPTO 2018》。计算机科学讲义，第10991卷，第693–721页。施普林格（2018年）2, 7[25] Coretti, S., Dodis, Y., Guo, S., Steinberger, J.P.：随机预言机与非一致性。载于：Nielsen, J.B., Rijmen, V.（编）《EUROCRYPT 2018》。计算机科学讲义，第10820卷，第227–258页。施普林格（2018年）2, 7, 20, 24, 34[26] Cummings, R., Ligett, K., Nissim, K., Roth, A., Wu, Z.S.：具有稳健泛化保证的自适应学习。载于：学习理论会议。第772–814页。PMLR（2016年）5[27] Dagan, Y., Shamir, O.：用少量内存和通信检测相关性。载于：学习理论会议，COLT。第1145–1198页（2018年）8[28] Dai, W., Tessaro, S., Zhang, X.：对称加密中的超线性时间-内存权衡。载于：TCC 2020。第335–365页（2020年）24[29] De Rezende, S.F., Nordstr¨om, J., Vinyals, M.：有限交互如何阻碍真实通信（以及这对证明和电路复杂性意味着什么）。载于：2016年IEEE第57届计算机科学基础年度研讨会（FOCS）。第295–304页。IEEE（2016年）8[30] Diakonikolas, I., Gouleakis, T., Kane, D.M., Rao, S.：离散分布检验中的通信与内存高效方法。载于：Beygelzimer, A., Hsu, D.（编）学习理论会议，COLT。第1070–1106页（2019年）8

[31] Diakonikolas, I., Kane, D.M., Pensia, A., Pittas, T.：高维稳健统计的流式算法。载于：国际机器学习会议，ICML。第5061–5117页（2022年）8[32] Dodis, Y., Farshim, P., Mazaheri, S., Tessaro, S.：迈向击败植入后门的随机预言机：有界自适应性下的不可区分性。载于：Pass, R., Pietrzak, K.（编）TCC 2020。计算机科学讲义，第12552卷，第241–273页。施普林格（2020年）7, 24[33] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A.：自适应数据分析中的泛化与留出集复用。载于：NIPS。第2350–2358页（2015年）5[34] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A.L.：在自适应数据分析中保持统计有效性。载于：第四十七届ACM计算理论年会论文集。第117–126页（2015年）4, 5, 17[35] Dwork, C., McSherry, F., Nissim, K., Smith, A.：在隐私数据分析中根据敏感度校准噪声。隐私与保密期刊7(3)，17–51（2016年）2[36] Dwork, C., Naor, M., Pitassi, T., Rothblum, G.N., Yekhanin, S.：全隐私流式算法。载于：ICS。第66–80页。清华大学出版社（2010年）2[37] Dwork, C., Naor, M., Reingold, O., Rothblum, G.N., Vadhan, S.：差分隐私数据发布的复杂性：高效算法与困难性结果。载于：第四十一届ACM计算理论年会论文集。第381–390页（2009年）8[38] Dwork, C., Rothblum, G.N., Vadhan, S.P.：增强与差分隐私。载于：FOCS。第51–60页。IEEE计算机学会（2010年）12[39] Dziembowski, S.：通过有界存储模型实现入侵韧性。载于：Halevi, S., Rabin, T.（编）TCC 2006。第3876卷，第207–224页。施普林格（2006年）6[40] Dziembowski, S., Pietrzak, K.：抗泄漏密码学。载于：FOCS 2008。第293–302页。IEEE计算机学会（2008年）6[41] Feldman, V.：学习是否需要记忆？一个关于长尾的简短故事。载于：第52届ACM SIGACT计算理论年会论文集。第954–959页（2020年）8[42] Feldman, V., Steinke, T.：通过稳定中位数实现自适应选择的估计量的泛化。载于：学习理论会议。第728–757页。PMLR（2017年）5[43] Feldman, V., Steinke, T.：在自适应数据分析中根据方差校准噪声。载于：学习理论会议。第535–544页。PMLR（2018年）5[44] Fish, B., Reyzin, L., Rubinstein, B.I.P.：亚线性时间自适应数据分析。载于：ISAIM（2018年）5[45] Fish, B., Reyzin, L., Rubinstein, B.I.P.：在自适应数据分析中抽样而不损害准确性。载于：ALT。机器学习研究论文集，第117卷，第297–318页。PMLR（2020年）5

[46] Garg, A., G¨o¨os, M., Kamath, P., Sokolov, D.：基于分辨率的单调电路下界。载于：第50届ACM SIGACT计算理论年会论文集。第902–911页（2018年）8[47] Garg, S., Raz, R., Tal, A.：用于学习的基于提取器的时间-空间下界。载于：第50届ACM SIGACT计算理论年会论文集，STOC。第990–1002页（2018年）8[48] Goldreich, O.：《密码学基础——第1卷：基本技术》。剑桥大学出版社（2001年）26[49] G¨o¨os, M., Lovett, S., Meka, R., Watson, T., Zuckerman, D.：矩形是非负联合函数。载于：Servedio, R.A., Rubinfeld, R.（编）STOC 2015。第257–266页。ACM（2015年）7, 8, 24, 34[50] G¨o¨os, M., Pitassi, T.：通过临界块敏感度得到通信下界。载于：第46届ACM计算理论年会论文集。第847–856页（2014年）8[51] G¨o¨os, M., Pitassi, T., Watson, T.：确定性通信与划分数。载于：2015年IEEE第56届计算机科学基础年会。第1077–1088页。IEEE（2015年）8[52] G¨o¨os, M., Pitassi, T., Watson, T.：针对有界错误概率多项式时间的查询到通信提升。《SIAM计算杂志》49(4)，FOCS17–441（2020年）8[53] Hardt, M., Ullman, J.R.：在交互式数据分析中防止假发现很困难。载于：FOCS。第454–463页。IEEE计算机学会（2014年）5, 6, 7, 8, 17, 18, 27, 28[54] H˚astad, J., Impagliazzo, R., Levin, L.A., Luby, M.：从任意单向函数构造伪随机生成器。《SIAM计算杂志》28(4)，1364–1396（1999年）24, 33[55] Hatami, H., Hosseini, K., Lovett, S.：异或函数协议的结构。《SIAM计算杂志》47(1)，208–217（2018年）8[56] Hazay, C., L´opez-Alt, A., Wee, H., Wichs, D.：基于最小假设的抗泄漏密码学。《密码学杂志》29(3)，514–551（2016年）6[57] Jung, C., Ligett, K., Neel, S., Roth, A., Sharifi-Malvajerdi, S., Shenfeld, M.：差分隐私泛化保证的新分析（特邀论文）。载于：STOC。第9页。ACM（2021年）5[58] Kalai, Y.T., Reyzin, L.：抗泄漏密码学综述。载于：Goldreich, O.（编）《为密码学奠定坚实基础：论Shafi Goldwasser和Silvio Micali的工作》，第727–794页。ACM（2019年）6[59]Kaplan, H., Stemmer, U.：关于使用差分隐私清理数据流的简述。CoRR abs/2111.13762（2021年）2[60] Kol, G., Raz, R., Tal, A.：稀疏奇偶校验学习的时间-空间困难性。载于：第49届ACM SIGACT计算理论年会论文集，STOC。第1067–1080页（2017年）8

[61] Kontorovich, A., Sadigurschi, M., Stemmer, U.：基于相关观测的自适应数据分析。载于：ICML。《机器学习研究论文集》，第162卷，第11483–11498页。PMLR（2022年）5[62] Kothari, P.K., Meka, R., Raghavendra, P.：用合取析取函数逼近矩形，以及线性规划松弛的弱指数下界。载于：Hatami, H., McKenzie, P., King, V.（编）STOC 2017。第590–603页。ACM（2017年）7, 24, 34[63] Loff, B., Mukhopadhyay, S.：相等的提升定理。载于：第36届国际计算机科学理论方面研讨会（STACS 2019）。Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik（2019年）8[64] Mir, D., Muthukrishnan, S., Nikolov, A., Wright, R.N.：通过草图统计实现全程隐私算法。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）2[65] Mironov, I., Pandey, O., Reingold, O., Vadhan, S.P.：计算差分隐私。载于：CRYPTO。《计算机科学讲义》，第5677卷，第126–142页。施普林格（2009年）12[66] Moshkovitz, D., Moshkovitz, M.：混合蕴含空间受限学习的下界。载于：第30届学习理论会议，COLT 2017论文集。第1516–1566页（2017年）8[67] Moshkovitz, D., Moshkovitz, M.：熵采样器与空间受限学习的强一般性下界。载于：第9届理论计算机科学创新会议，ITCS 2018。第28:1–28:20页（2018年）8[68] Nissim, K., Smith, A., Steinke, T., Stemmer, U., Ullman, J.：后选择泛化的局限。《神经信息处理系统进展》第31卷（2018年）5[69] Pagh, R., Stausholm, N.M.：高效的差分隐私 f_0 线性草图。载于：ICDT。《莱布尼茨国际信息学论文集》，第186卷，第18:1–18:19页。Schloss Dagstuhl - Leibniz-Zentrum f¨ur Informatik（2021年）2[70] Pagh, R., Thorup, M.：私有 CountSketch 的改进效用分析。CoRR abs/2205.08397（2022年）2[71] Pietrzak, K.：一种抗泄漏操作模式。载于：Joux, A.（编）EUROCRYPT 2009。《计算机科学讲义》，第5479卷，第462–482页。施普林格（2009年）6[72] Raz, R.：一大类学习问题的时间—空间下界。载于：第58届IEEE计算机科学基础年会，FOCS。第732–742页（2017年）8[73] Raz, R.：快速学习需要良好的内存：奇偶学习的时间—空间下界。《美国计算机协会学报》66(1)，3:1–3:18（2019年）8[74] Raz, R., McKenzie, P.：单调NC层次的分离。《组合》19(3)，403–435（1999年）1, 8[75] Rogers, R., Roth, A., Smith, A., Thakkar, O.：最大信息、差分隐私与后选择假设检验。载于：2016年IEEE第57届计算机科学基础年度研讨会（FOCS）。第487–494页。IEEE（2016年）5[76] Shenfeld, M., Ligett, K.：自适应泛化的必要且充分稳定性概念。载于：NeurIPS。第11481–11490页（2019年）5

[77] Shenfeld, M., Ligett, K.：自适应性下的泛化：贝叶斯视角。CoRR abs/2106.10761（2021）5[78] Smith, A.D., Song, S., Thakurta, A.：Flajolet-Martin 草图本身即可保持差分隐私：以最小空间实现私密计数。载于：NeurIPS（2020）2[79] Standaert, F., Pereira, O., Yu, Y., Quisquater, J., Yung, M., Oswald, E.：实践中的抗泄漏密码学。载于：Sadeghi, A., Naccache, D.（编）《迈向硬件内生安全——基础与实践》，第99–134页。信息安全与密码学，施普林格（2010）6[80] Steinhardt, J., Valiant, G., Wager, S.：内存、通信与统计查询。载于：《第29届学习理论会议论文集》（COLT），第1490–1516页（2016）8[81] Steinke, T., Ullman, J.R.：交互式指纹码与阻止假发现的难题。载于：COLT。《JMLR研讨会与会议论文集》，第40卷，第1588–1628页。JMLR.org（2015）5, 6, 7, 8, 13, 17, 18, 27, 28[82] Tardos, G.：最优概率指纹码。《美国计算机协会会刊》（JACM）55(2)，1–24（2008）13[83] Ullman, J.：回答 $n^{2+o(1)}$ 差分隐私计数查询很困难。载于：《第四十五届ACM计算理论年会论文集》，第361–370页（2013）8[84] Unruh, D.：随机预言机与辅助输入。载于：Menezes, A.（编）《CRYPTO 2007》。计算机科学讲义，第4622卷，第205–223页。施普林格（2007）2, 7[85] Upadhyay, J.：滑动窗口模型下的亚线性空间私密算法。载于：ICML。《机器学习研究论文集》，第97卷，第6363–6372页。PMLR（2019）2[86] Upadhyay, J., Upadhyay, S.：滑动窗口模型下私密矩阵分析的框架。载于：ICML。《机器学习研究论文集》，第139卷，第10465–10475页。PMLR（2021）2[87] Wang, L., Pinelis, I., Song, D.：采用多对数空间的差分隐私分数阶频率矩估计。arXiv预印本 arXiv:2105.12363（2021）2[88] Woodruff, D.P.：所有频率矩的最优空间下界。载于：Munro, J.I.（编）《第十五届ACM-SIAM离散算法年会论文集》（SODA 2004），美国路易斯安那州新奥尔良，2004年1月11–14日，第167–175页。工业与应用数学学会（2004）3[89] Wu, X., Yao, P., Yuen, H.S.：利用内积组件进行Raz–McKenzie模拟。载于：电子计算复杂性讨论会，第24卷，第10页（2017）8[90] Zhao, F., Qiao, D., Redberg, R., Agrawal, D., Abbadi, A.E., Wang, Y.X.：差分隐私线性草图：高效实现与应用。arXiv预印本 arXiv:2205.09873（2022）2