arXiv:2603.04635v1 [stat.ML] 4 Mar 2026

用于分布独立性检验的最优预测增强算法

Maryam Aliakbarpour* Alireza Azizi Ria Stevens 莱斯大学 莱斯大学 莱斯大学maryama@rice.edu alireza.azizi@rice.edu ria.stevens@rice.edu

2026年3月6日

摘要

独立性检验是统计推断中的一项基础问题：给定来自多个随机变量的联合分布 p 的样本，目标是确定该分布是否为乘积分布，或者在全变差距离下与所有乘积分布都相差至少 ϵ。在非参数有限样本情形中，这项任务以代价高昂著称，因为极小极大样本复杂度随支撑集大小呈多项式增长。在这项工作中，我们利用 *augmented distribution testing* 框架，突破了这些最坏情况限制。我们设计了结合辅助预测信息的独立性检验器，但这种信息可能并不可信。我们的框架确保检验器保持稳健：无论预测质量如何，都能保证最坏情况下的有效性；而预测准确时，样本效率则会显著提升。我们的主要贡献包括：(i) 一种用于离散分布的双变量独立性检验器，可根据预测误差自适应地降低样本复杂度；(ii) 将其推广到高维多变量情形，以检验 d 个随机变量的独立性；以及 (iii) 给出匹配的极小极大下界，证明我们的检验器达到了最优样本复杂度。

*Dep计算机科学系与 Ken Kennedy 研究所 研究所 ute

1 引言

独立性检验是统计推断中的核心问题，应用涵盖医学领域 [Cor51，Sas97，HYF+22]，也涉及机器学习中的因果发现和特征选择等问题 [BJ02，SRB07，SSG+07，GFT+07，ZPJS11，BPZL12，CLZ22]。形式上，给定来自多个随机变量的联合分布的样本，目标是确定这些变量在统计上是否相互独立，或是否存在某种依赖性。

尽管诸如 χ^2 Pea [rson00] 检验被广泛使用，但它们通常无法为非参数有限样本情形提供保证（例如，控制较小的第一类和第二类错误）。相反，其有效性要么源于严格的参数假设，要么依赖于渐近性质，而这些性质只有在样本数量趋于无穷时才成立。若缺乏此类假设，独立性检验的成本往往极高：其极小极大样本复杂度随支撑集大小 [BFF+01 呈多项式增长，DK16]。这一高昂的下界给样本高效推断带来了严峻挑战。

为突破这些最坏情形下的局限，近期一系列研究指出，分析者不必在真空中开展工作。在许多现代数据科学场景中，人们通常拥有辅助信息：来自不同机构的公共数据集，或源自历史数据、生成模型或领域特定启发式方法的分布。尽管此类信息可能十分丰富，却往往缺乏正式的质量保证。一个自然的问题是，能否以严谨的方式利用这类不可信的 预测信息——设计出在预测较差时仍保持数学可靠性（且具备鲁棒性）、而在预测准确时又能显著提高样本效率的检验方法。

近年来，增强算法框架应运而生，旨在弥合这一差距，并将辅助信息引入经典统计问题 [CR14、CRS15、OS18、EIN+21、AIRS24]等工作及后续研究 [算法26]。这一范式在多种分布检验任务中展现出巨大潜力，例如均匀性检验、恒等性检验（拟合优度检验）和接近性检验（等价性检验） [AIRS24、ABCR25]。在增强检验框架中，检验器既能获得来自 p真实数据分布的样本，也能获得一个预测分布 $\hat{}p$。检验器必须具备鲁棒性，无论 $\hat{}p$为何，都能维持最坏情形下的有效性，同时利用准确的预测超越经典的极小极大下界。

在本文中，我们将增强分布检验框架应用于离散支撑集上的独立性检验问题。我们的主要贡献如下： *i*) 我们提出了一种针对二元分布的独立性检验器，其样本复杂度有限，并具有最坏情形错误保证。该检验器利用预测分布，并根据预测质量自适应地降低样本复杂度，同时不损害检验的准确度。 *ii*) 我们将这一结果推广到高维情形，用于检验多个随机变量的独立性。 *iii*) 我们给出匹配的下界，表明我们的检验器达到了最优样本复杂度。

1.1 问题陈述

我们采用分布性质测试的框架 [Gol17]。在此设定下，给定对分布 p的样本访问权限，检验器必须以高概率输出接受 ACCEPT，如果 p 属于某个性质 $\mathcal{P}$（视为一个分布集合），并且 REJECT 输出 p，如果 ϵ与集合$\mathcal{P}$中的每个分布的全变差距离至少为 $\mathcal{P}$。这里，我们将独立性视为该性质——具体而言，它是定义域某一给定分解上的所有乘积分布所构成的集合。

在增强设定下， [AIRS24], 除了可以对分布进行样本访问 p之外，我们还会获得一个预测分布 $\hat{}p$ 以及一个建议误差界 α，用于表示对 $\hat{}p$准确度的声称。我们假设完全了解 $\hat{}p$ （即已知其概率质量函数）。在此框架下，增强独立性检验器除了输出 ACCEPT 或 REJECT之外，还有第三种选择：如果预测的实际质量低于建议误差 INACCURATE INFORMATION ，它可以选择不回答检验，而输出 α。然而，只要预测足够准确（即建议误差 α 正确），检验器就必须给出答案。至关重要的是，检验器不能输出错误答案；如果检验器决定回答检验，那么无论预测质量如何，其输出都必须以高概率正确。形式化地，我们将所需的增强检验器定义如下：

定义 1.1（增强独立性检验器）。设 $\alpha \in [0,1], \epsilon \in$ (0,1) $\delta \in$ (0, 1) $\delta \in$。对于所有 $i \in [d]$，令 $n_i \geq 2$ 为表示定义域大小的整数。设 p 和 $\hat{}p$ 为定义域 $\prod_{i=1}^{d}[n_i]$ 上的分布，并令 $p_1, \ldots, p_d$ 表示 p 的边缘分布。若算法 $\mathcal{A}$ 可对 p 进行样本访问，并可显式访问 $\hat{}p$，且输出 ACCEPT、 REJECT 或 INACCURATE INFORMATION，则称其为一个（α, ϵ, δ）增强独立性检验器，若对于所有 p，它都满足：

- 如果 $p = p_1 \times p_2 \times \cdots \times p_d$， $\mathcal{A}$ 以至多 $\delta/2$ 的概率输出 REJECT。
- 如果 p 与所有乘积分布都相差至少 ϵ， $\mathcal{A}$ 以至多 $\delta/2$ 的概率输出 ACCEPT。
- 如果 $\mathrm{d_{TV}}(p, \hat{}p) \leq \alpha$， $\mathcal{A}$ 以至多 $\delta/2$ 的概率输出 INACCURATE INFORMATION。

注 1（寻找预测误差。）。请注意，在许多实际场景中，预测误差 α 可能无法预先得知。不过，我们的增强检验器具有鲁棒性，因此非常适合自适应地确定所需的误差水平。具体而言，通过采用 SEARCH*AIRS* 的 [元算法*24*]，，我们可以从最小值开始， $\alpha^\star$ 寻找真实误差 α，并不断增大 α 的取值，反复调用增强检验器，直到算法给出明确答案（即不再输出 INACCURATE INFORMATION）。该过程的期望样本复杂度为 $O(f(\alpha^\star))$，其中 $f(\alpha^\star)$ 表示提供预测的真实误差 $\alpha^\star$ 时检验器的样本复杂度。本文余下部分均假设给定一个固定的 α，并基于该误差界推导样本复杂度。

1.2 我们的主要结果

在本文中，我们为二维和 d 维分布的增强独立性检验给出了匹配的上界和下界。我们的贡献总结在以下定理中。

定理 *2*（定理 *4*、*6*、*7* 和 *8* 的非正式版本）。设 $d \geq 2$。设 p 为 $n_1 \times \ldots \times n_d$ 上的未知分布，并设 $\hat{}p$ 为同一定义域上对 p 的已知预测。设 $N := \prod_{i=1}^{d} n_i$ 表示总域大小。对于每个 $\epsilon \in$ *(0, 1)* 和 $\alpha \in [0,1]$，在给定预测 $\hat{}p$ 的情况下，对 p 进行 $(\alpha, \epsilon, \delta = 0.1)$-增强独立性检验的样本复杂度为

$$\Theta\left(\max_{j\in[d]}\left(\frac{\sqrt{N}}{\epsilon^2}, \frac{n_j^{1/3}N^{1/3}\alpha^{1/3}}{\epsilon^{4/3}}\right)\right). \tag{1}$$

二维增强独立性检验器及样本复杂度上界在第2.2节中讨论，而相应的下界在第4节中建立。 d维检验器及上界在第3节中给出，匹配的下界在第5节中证明。

注 3 (成功放大)。在本文中，我们考虑一个常数失败概率（$\delta = 0.1$）。利用标准放大技术，我们可以将 δ 设得任意小，代价是样本复杂度增加一个乘法因子 $\mathcal{O}(\log(1/\delta))$。

1.3 技术概述

1.3.1 背景：扁平化

在构造上界时，我们采用了标准分布检验文献中的一种技术，称为扁平化 [DK16]。该技术通过重新分配原始分布中高概率元素的概率质量，将一个分布检验问题转化为定义域更大的等价问题。扁平化大幅降低了该分布的 ℓ_2-范数，从而能够使用样本高效的检验算法。

标准展平。扁平化将分布中每个定义域元素的概率质量划分为大小相等的元素级桶，从而定义一个定义域更大的新分布。形式上，给定定义域为 $[n]$ 的分布 p 以及一组桶大小 b_i （对所有 $i \in [n]$），扁平化将每个元素 i 映射为一个元组（$i,\ \ell$），其中 ℓ 从 $[b_i]$中均匀采样。该操作定义了一个新分布 $p^{(F)}$，使得：

$$p^{(F)}(i,\ell) = \frac{p(i)}{b_i} \tag{2}$$

扁平化　　满足以下两个关键性质，这些性质由 [DK16], 证明，并使其能够用于分布检验：　　ng:

1. 如果使用相同的桶大小对两个分布 p 和 q 进行扁平化，则它们的 ℓ_1距离保持不变。因此，与其将 p 和 q 提供给检验器，不如提供 $p^{(F)}$ 和 $q^{(F)}$，并期待得到相同的结果。

2. 可以利用来自 p 的一组初始样本来识别哪些元素具有较大的概率质量。通过将桶大小 b_i 设置为与这些经验频率成比例，可以有效地将高概率元素"分散"到定义域中的多个点上。这可以确保所得的展平分布 $p^{(F)}$ 在期望意义下具有足够小的 ℓ_2范数，从而允许进行节省样本的检验。

增强展平。[AIRS24] 将 [DK16] 的扁平化技术扩展到增强设定中，使预测值 $\hat{}p$能够指导为每个元素创建的桶数量。除了利用从对 p进行抽样所获得的信息外，该方法还会对在预测下具有较大概率质量的元素进行扁平化。当预测准确时，这可以进一步降低 ℓ_2范数。

增强展平技术 [AIRS24] 首先抽取一个大小为 $\mathcal{S}$ 泊松分布(s) 的样本集，该样本集来自 p。然后，对于所有 $i \in [n]$，令 N_i 表示 $i \in \mathcal{S}$ 出现的次数，并将所有 i 的桶数量选为：

$$b_i = \lfloor n \cdot \hat{p}(i) \rfloor + N_i + 1. \tag{3}$$

该表达式的第一项体现了预测的影响：为预测概率质量较高的元素分配更多桶。第二项体现了样本集提供的信息：为在 $\mathcal{S}$中更频繁出现的元素分配更多桶。 [AIRS24] 证明，当 $\hat{}p$ 与 α的接近程度为 p时，由该增强展平定义的分布 $p^{(F)}$ 具有较小的 ℓ_2^2范数：

$$\mathbf{E}\left[\left\|p^{(F)}\right\|_2^2\right] \leq \frac{2\alpha}{s} + \frac{4}{n}. \tag{4}$$

详见事实A.4 。

通过展平进行接近性测试。 对两个分布在 $[n]$上的接近性进行测试，需要判别 $p = q$ 还是 $\|p - q\|_1 \geq \epsilon$，且成功概率至少为 0。9。该任务的极小极大样本复杂度为 $\Theta(n^{2/3}/\epsilon^{4/3} + \sqrt{n}/\epsilon^2)$ [DK16]。该领域的一项关键进展是开发出接近性测试器，其样本复杂度取决于底层分布的 ℓ_2范数，而不是原始支撑集大小 [CDVV14]。具体而言，如果这些分布的 ℓ_2范数至多为 b，则可以使用 $O(n\,b/\epsilon^2)$ 个样本进行检验；在某些情况下，这比最坏情况的界所暗示的复杂度高效得多。 ℓ_2范数与检验复杂度之间的这种关系，是展平技术的核心动机：通过降低分布的 ℓ_2范数，可以利用这些改进后的检验器来实现最优样本复杂度。事实上，这一方法对于确立标准独立性检验的最优界至关重要 [DK16]。

通过扁平化进行独立性检验。 独立性检验可以被视为判别联合分布 p 是否等于其边缘分布的乘积， $p_1 \times p_2$ 还是 3ϵ与其相距 [BFF+01]。从这一角度来看，可以使用来自 p的样本来模拟来自 p和 $p_1 \times p_2$的样本，随后使用接近性检验器来确定独立性。

虽然对联合定义域进行直接扁平化 $[n]\times[m]$ 其中 $n \geq m$ 会导致次优的样本复杂度，[DK16]达到了最优速率

$$O\left(\frac{n^{2/3}m^{1/3}}{\epsilon^{4/3}} + \frac{\sqrt{nm}}{\epsilon^2}\right).$$

粗略地说，他们的方法依赖于构造展平的乘积：根据各自的边缘分布独立地展平每个维度，从而确保独立分布 p被转换为一个联合展平分布，且其 ℓ_2范数较低。这使得该分布适用于 [CDVV14], 提出的改进的接近性检验器；对于 ℓ_2范数较小的分布，该检验器的效率非常高。

1.3.2 我们的结果概述

二维增强独立性检验的上界。我们的方法将增强展平与乘积展平技术相结合，利用预测$\hat{}p$ 更高效地降低边缘分布的 ℓ_2范数。该方法的正确性取决于一个关键的验证步骤：如果$\hat{}p$ 准确，则展平后的边缘分布可以保证具有较低的 ℓ_2范数。反之，若违反这一界限，就能作为预测不准确的证据。此外，如果 p相互独立，展平后的联合分布也必然具有较低的 ℓ_2范数。我们的算法会明确估计这个联合范数；若其超过预期阈值，即可作为非独立性的证据，使我们能够REJECT 立即采取行动。这样，我们要么通过范数增大及早发现依赖性，要么继续处理表现“良好”的分布，从而高效检验其与边缘分布的接近性。

二维增强独立性检验的下界。 二维增强独立性检验的下界分为两种情形，具体取决于 n、m、α 和 ϵ之间的关系。第一种情形给出 $\Omega(\sqrt{nm}/\epsilon^2)$ 的下界，该下界通过从标准独立性检验的困难实例 [ADK15、DK16]归约而得。第二种情形给出 $\Omega\left(n^{2/3}m^{1/3}\alpha^{1/3}\epsilon^{-4/3}\right)$ 的下界。为证明此下界，我们构造两个在信息论上不可区分的分布族，但增强检验器必须能够将它们判别开来。这一构造建立在先前工作 [Val11、BFR+13、DK16] 所确立的技术之上：分布与某性质的距离隐藏在概率质量较低的“轻”元素中，而大部分概率质量集中在两个分布族中完全相同的“重”行上。这里，行指固定第一坐标后得到的第二坐标上的概率分布。

在标准设定中，重行服从均匀分布，而轻行在一个分布族中服从均匀分布，在另一个分布族中则与均匀分布相距 ϵ。这种构造迫使检验器抽取足够多的样本，才能有效地找出重行。然而，在我们的增强设定中，预测可能会向检验器揭示重行的位置。我们将这一技巧扩展到增强设定，方法是设计分布族，使得只有行的 $\mathcal{O}(\alpha)$ 因子是重行。通过选择均匀分布作为预测，我们便困住了检验器：该预测既不提供任何关于哪些行是重行或轻行的信息，又与真实分布相距 α。这迫使检验器输出 ACCEPT或 REJECT，而不是 INACCURATE INFORMATION——本质上要求它区分这两个分布族。

接下来，我们给出一个信息论论证，说明如果样本不足，这两个分布族便无法区分。遵循 [DK16], 的框架，我们界定一个指示变量（表示某个分布所属的分布族）与从该分布抽取的样本集之间的互信息。为此，我们证明，只有当检验器在轻行中观察到碰撞（即从同一行抽取多个样本）时，互信息才会显著。重行不会贡献互信息，因为它们在两个分布族中完全相同；而如果轻行中没有发生碰撞，检验器便无法得知该行服从均匀分布，还是与均匀分布相距 ϵ。在我们的困难实例中，除非样本复杂度很高，否则检验器在轻行中观察到足够多次碰撞的概率很低。

上界： d**维增强独立性检验**。将增强独立性检验扩展到 d 维分布的一种朴素方法，是分别对所有边缘分布应用增强展平

，然后在展平后的联合分布与展平边缘分布的乘积之间进行接近性检验。然而，增强展平会使每个坐标的定义域大小至少增加一个常数倍（大于 2）。对于一般的 d，这会使样本复杂度额外增加 $2^{\mathcal{O}(d)}$ 倍；当 d 较大时，这将变得难以承受。相比之下，在低维情况下，这一开销仍然无碍。事实上，我们利用这一方法构造了一个三维检验器，并将其作为高维构造的构建模块。为获得一般检验器，我们将 d 个坐标划分为至多三个组，每组的总域大小至多为 $\sqrt{N}$，其中 $N := \prod_{i=1}^{d} n_i$。如果某个坐标的定义域大小大于 $\sqrt{N}$，我们就将其单独划为一个单元素集合，并把其余坐标合并为第二组。完成划分后，我们首先调用二维或三维增强检验器，检查所得各组之间的独立性。如果该检验输出 ACCEPT，则还需验证每组内部的独立性。根据构造，每组的总域大小至多为 $\sqrt{N}$，因此只需使用 $\mathcal{O}(\sqrt{N}/\epsilon^2)$ 个样本，就可以将每组上的经验分布学习到 ℓ_1 误差（$\Theta(\epsilon)$）以内。

通过明确界定学习边缘分布所引入的误差，我们证明，对于乘积分布，学习得到的经验分布也接近其边缘分布的乘积。因此，可以对经验分布进行独立性检验。最后这一步完全是计算性的，不需要额外的样本。

下界： **d维增强独立性检验**。我们对 d 维增强独立性检验给出的下界，可由二维增强独立性检验问题的一个归约得到。尽管这一归约并不适用于所有分布，我们证明，d 维检验器仍然能够判别用于证明二维情形下界的特定困难实例。具体而言，我们证明，可以将一个二维分布重塑为 d维分布，同时保持其与乘积分布以及预测之间的距离。该论证依赖于二维情形下构造的困难实例的一项结构性质，即相应的乘积分布是均匀分布。因此，我们可以将二维样本重塑为d维样本，应用 d 维检验器，并返回其输出，从而解决二维增强独立性检验问题。

1.4 相关工作

分布检验。自 [RS96, GGR98] 和 [Bat01]，首次提出以来，许多研究一直在探讨用于检验分布不同性质的算法。这一领域的核心问题包括恒等性检验 [VV17、DKN15], 均匀性检验 [Pan08]，以及接近性检验 [BFR+00, ADJ+12, BFR+13, CDVV14]。例如， [BKR04、Val11、GR11、ILR12、LRR13、DDS+13、ADK15、DK16、ABR16、CDGR18、AGP+19、AKR19、AS20] 等研究也探讨了这些性质及其他性质。关于该领域的综述，请参阅 [Rub12、Gol17、Can20、Can22]。

独立性检验。自 [Pea00], 开创性工作提出著名的 χ^2-检验以来，检验两个分布是否独立的问题便一直受到研究。现代研究着重于开发样本高效算法。特别是， [Bat01] 和 [LRR13] 提出了首批样本复杂度次线性的算法。 [ADK15] 和 [DK16] 改进了这些结果，后者为标准设定下的二维和 d维独立性检验给出了紧确的样本复杂度界。相关问题包括 k-路独立性检验 [AAK+07，RX10] 以及条件独立性检验[CDKS18]。

分布检验增强算法。为了绕过分布检验的标准下界，近期研究假设了比随机抽样更强的访问模型。这些模型包括概率质量函数的精确或近似预言机 [CR14，OS18，EIN+21]，以及能够从定义域的某个子集进行抽样的条件预言机 [CRS15]。与我们的工作最相关的是， [AIRS24] 开创了增强分布检验的研究，其中检验器会获得对真实分布的预测

以及对该预测准确度的估计。最近， [ABCR25] 将这一框架扩展到了隐私设置。在这项工作中，我们基于 [AIRS24], 的模型，研究用于独立性检验的增强算法。

2 二元增强独立性检验的上界

2.1 用于独立性检验的增强展平

文献 [AIRS24] 中介绍的增强展平技术1.3.1 适用于一维定义域上的分布 $[n]$。本节将这一方法与文献 [DK16] 中的扁平化乘积技术相结合，针对二维定义域上的分布开发一种增强展平技术 $[n] \times [m]$。

假设给定一个定义在 $[n] \times [m]$ 上的分布 p，其边缘分布为 p_1 和 p_2。为了对 p进行扁平化，我们从 p_1 中抽取 Poi(s_1) 个样本，组成样本集 $\mathcal{S}_1$ ；并独立地从 p_2 中抽取 Poi(s_2) 个样本，组成样本集 $\mathcal{S}_2$。我们构造两个独立的逐坐标增强扁平化 F_1 和 F_2，其中桶的数量 $b_i^{(1)}$ 和 $b_j^{(2)}$ 根据（3）选取；其中 F_1 使用预测值$\hat{}p_1$ 和样本集 $\mathcal{S}_1$，而 F_2 使用预测值$\hat{}p_2$ 和样本集 $\mathcal{S}_2$。接着，我们在 $[n]\times[m]$ 上定义全局扁平化 F，为每一对(i, j)分配$b_i^{(1)} b_j^{(2)}$ 个桶。这等价于将每个(i, j)映射到一个元组(i, j, k, ℓ)所得到的扁平化，其中 k和 ℓ 分别从 $[b_i^{(1)}]$ 和 $[b_j^{(2)}]$中均匀选取。因此，当 p是一个乘积分布时， $p^{(F)} = p^{(F_1)}_1 \times p_2^{(F_2)}$。

2.2 二元增强独立性检验器

本节给出一种针对二元分布的增强独立性检验器，其样本复杂度达到最优。给定对分布的样本访问 p、对预测 $\hat{}p$ 的完整知识以及对预测准确度的估计值 α，该算法按照第 $[n] \times [m]$2.1 节所述的过程，对定义域进行增强展平。

然后，它通过比较展平后的边缘分布的范数与 d α $(_{\mathrm{TV}}, p\hat{} \leq \alpha)$ 情况下的期望值，检验该估计值 INACCURATEINFORMATION是否可靠。如果发现估计值不可靠，则输出 REJECT。否则，执行第二个验证步骤，确保展平后的联合分布与乘积分布的范数足够接近。如果二者不够接近，则联合分布必然与乘积分布相距甚远，检验器输出 [。如果二者足够接近，算法便使用 CDVV的接近性检验器比较这两个展平分布，并输出该检验的结果。

检验器的伪代码见算法 1。在定理4中，我们证明其正确性及其样本复杂度。

定理 *4*。设 n、 $m \in N$ 满足 $n \geq m \geq 2$。对所有 ϵ、 $\delta \in (0, 1)$ $\alpha \in [0, 1]$, $[n] \times [m]$ 上的未知分布 p 以及 $[n] \times [m]$ 上的已知分布 $\hat{}p$，算法 *1* 是一个 $(\alpha, \epsilon, \delta = 0.1)$-增强独立性检验器，样本复杂度为 $O\left(max\left\{\frac{\sqrt{nm}}{\epsilon^2}, \frac{n^{2/3}m^{1/3}\alpha^{1/3}}{\epsilon^{4/3}}\right\}\right)$。

证明。

正确性 我们证明，算法 1 满足增强独立性检验器的三项要求，如定义 1.1所述。在整个证明中，我们将常数设为 $c = 120$ 和 $c' = 160$。

算法1 的正确性证明首先表明，以高概率，每个展平分布的ℓ_2^2 范数估计值都是准确的。在这些事件成立的条件下，该算法满足增强独立性检验器的三项要求：(i) 若预测 α-接近于 p，则该算法以高概率不输出 INACCURATE INFORMATION；(ii) 若 p远离乘积分布，则该算法不输出 ACCEPT；(iii) 若 p 是乘积分布，则该算法不输出 REJECT。

算法 1： 二维分布增强独立性检验器

输入： 对定义在 $[n]\times[m]$ 上且边缘分布为 p_1 和 p_2 的分布 p 的样本访问；预测 $\hat{}p$；预测误差的估计值 α；接近参数 ϵ；置信度 δ

Augmented-Independence-Tester-2D $(n, m, p, \hat{}p, \alpha, \epsilon, \delta = 0.1)$**:**

$c, c' \leftarrow$ 常数

$s_1 \leftarrow$ 最小值 $\left(n^{2/3}m^{1/3}\alpha^{1/3}\epsilon^{-4/3}, n\alpha\right)$， $s_2 \leftarrow m\alpha$

$\hat{s}_1 \leftarrow$泊松分布(s_1)， $\hat{s}_2 \leftarrow$泊松分布(s_2)

如果 $\hat{s}_1 > c's_1$ 或 $\hat{s}_2 > c's_2$ **则**

  返回 REJECT

end

$\mathcal{S}_1 \leftarrow \hat{s}_1$ 个来自 p_1 的样本， $\mathcal{S}_2 \leftarrow \hat{s}_2$ 个来自 p_2 的样本

$\tau_1 \leftarrow \frac{2\alpha}{s_1} + \frac{4}{n}$， $\tau_2 \leftarrow \frac{2\alpha}{s_2} + \frac{4}{m}$

对于 $i \in [n]$ 和 $j \in [m]$，**执行**

  构造 $b_i^{(1)}, b_j^{(2)}$ 如（3）中所示，使用样本集 $\mathcal{S}_1$ 和 $\mathcal{S}_2$

end

F_1， F_2， $F \leftarrow [n]$ 的展平，， $[m]$， $[n]\times[m]$ 给定桶 $b_i^{(1)}$，对于所有 i，以及对于所有 j 的 $b_j^{(2)}$；// 参见第 2.1

$L_1 \leftarrow$ 估计值-ℓ_2^2 $\left(p_1^{(F_1)}, \sum_{i=1}^n b_i^{(1)}, 1/120\right)$ // 参见事实 A.2

$L_2 \leftarrow$ 估计值-ℓ_2^2 $\left(p_2^{(F_2)}, \sum_{j=1}^m b_j^{(2)}, 1/120\right)$

如果 $L_1 > c\tau_1$ 或 $L_2 > c\tau_2$ **则**

  返回 INACCURATE INFORMATION

end

$L \leftarrow$ 估计值-ℓ_2^2 $\left(p^{(F)}, \left(\sum_{i=1}^n b_i^{(1)}\right)\left(\sum_{j=1}^m b_j^{(2)}\right), 1/120\right)$

如果 $L > 10c^2\tau_1\tau_2$ **则**

  返回 REJECT

end

返回 接近度测试器 $\left(p^{(F)}, p_1^{(F_1)} \times p_2^{(F_2)}, 4c^2\tau_1\tau_2, \epsilon, 1/80\right)$ // 参见事实 A.3

结束

根据事实A.2，估计值 L_1、 L_2和 L 分别对应于扁平化分布 $p_1^{(F_1)}$、 $p_2^{(F_2)}$和 $p^{(F)}$的范数，并且以高概率处于真实范数的常数因子范围内。具体而言，根据事实A.2以及并集界，以下各项以至少 $1 - 1/40$的概率成立：

$$\frac{\left\|p_1^{(F_1)}\right\|_2^2}{2} \le L_1 \le \frac{3\left\|p_1^{(F_1)}\right\|_2^2}{2}, \quad \frac{\left\|p_2^{(F_2)}\right\|_2^2}{2} \le L_2 \le \frac{3\left\|p_2^{(F_2)}\right\|_2^2}{2} \quad \text{and} \quad \frac{\left\|p^{(F)}\right\|_2^2}{2} \le L \le \frac{3\left\|p^{(F)}\right\|_2^2}{2}. \tag{5}$$

在证明剩余部分中，我们将始终以该事件为条件。

我们现在证明，当预测与该分布 α接近时，输出INACCURATE INFORMATION 在条件（5）下的概率至多为0.025。假设$\mathsf{d}_{\mathrm{TV}}(p, \hat{}p) \le \alpha$。注意该算法仅可能输出 INACCURATE INFORMATION 于第17行，当 $L_1 > c\tau_1$ 或 $L_2 > c\tau_2$时。应用并集界以及如下事实：通过对（5）进行条件化，我们可以用常数倍上界 L_1 与 L_2 的常数倍

的 $\left\|p^{(F_1)}{}_1\right\|_2^2$ 与 $\left\|p^{(F_2)}{}_2\right\|_2^2$，输出 INACCURATE INFORMATION 的概率至多为：

$$\begin{aligned}
&\mathbf{Pr}[\text{INACCURATE INFORMATION} \mid \mathrm{d_{TV}}(p,\hat{p}) \le \alpha, (5)] && (6)\\
&= \mathbf{Pr}[L_1 > c\tau_1 \text{ or } L_2 > c\tau_2 \mid \mathrm{d_{TV}}(p,\hat{p}) \le \alpha, (5)] && (7)\\
&\le \mathbf{Pr}\left[\frac{3}{2}\left\|p_1^{(F_1)}\right\|_2^2 > c\tau_1 \mid \mathrm{d_{TV}}(p,\hat{p}) \le \alpha\right] \\
&\qquad + \mathbf{Pr}\left[\frac{3}{2}\left\|p_2^{(F_2)}\right\|_2^2 > c\tau_2 \mid \mathrm{d_{TV}}(p,\hat{p}) \le \alpha\right] && (8)
\end{aligned}$$

根据事实 A.4，当 $\mathrm{d_{TV}}(p, \hat{p}) \le \alpha$时，$\ell_2^2$-范数的期望 $p_1^{(F_1)}$ 和 $p_2^{(F_2)}$ 分别以上界 τ_1 和 τ_2。结合马尔可夫不等式，该事实推出式 (8) 的上界为：

$$\begin{aligned}
&\mathbf{Pr}[\text{INACCURATE INFORMATION} \mid \mathrm{d_{TV}}(p,\hat{p}) \le \alpha] && (9)\\
&\le \frac{\mathbf{E}\left[\frac{3}{2}\left\|p_1^{(F_1)}\right\|_2^2 \mid \mathrm{d_{TV}}(p,\hat{p}) \le \alpha\right]}{c\tau_1} + \frac{\mathbf{E}\left[\frac{3}{2}\left\|p_2^{(F_2)}\right\|_2^2 \mid \mathrm{d_{TV}}(p,\hat{p}) \le \alpha\right]}{c\tau_2} && (10)\\
&\le \frac{3\tau_1}{2c\tau_1} + \frac{3\tau_1}{2c\tau_2} && (11)\\
&= 1/40, && (12)
\end{aligned}$$

其中，最后一个等式是通过代入 $c = 120$ 得到的。

接下来，我们证明在给定条件 (5) 下，接近性检验器的失败概率至多为 1/80。根据事实 A.3，只要参数 $b = 4c^2\tau_1\tau_2$ 以上界两个输入分布的 ℓ_2^2- 范数的较小值即可。我们的算法通过在第 16 行和第 20 行验证边缘分布及分布本身的范数，确保满足这一条件。具体而言，只有当第23行调用第16 行和第20 行的结果均为 false 时，才会调用接近性检验器。这意味着调用接近性检验器时，我们有以下界：

$$L_1 \le c\tau_1, \quad L_2 \le c\tau_2, \quad \text{and} \quad L \le 10c^2\tau_1\tau_2. \tag{13}$$

结合（5），这些界意味着展平分布的平方范数满足以下界：

$$\left\|p_1^{(F_1)}\right\|_2^2 \le 2c\tau_1, \quad \left\|p_2^{(F_2)}\right\|_2^2 \le 2c\tau_2, \quad \text{and} \quad \left\|p^{(F)}\right\|_2^2 \le 20c^2\tau_1\tau_2. \tag{14}$$

此外，由于 ℓ_2^2-范数 $p^{(F_1)}{}_1 \times p_2^{(F_2)}$ 等于其边缘分布平方范数的乘积，因此其上界为 $4c^2\tau_1\tau_2$。因此，提供给标准接近性检验器的分布的平方范数的最小值至多为 $4c^2\tau_1\tau_2$，这与我们对参数 b的设定相符。因此，我们可以应用事实 A.3 ，并保证接近性检验器的失败概率至多为 1/80。

现在证明，如果 p与任一乘积分布的距离至少为 ϵ，那么算法 1 输出 ACCEPT 的概率至多为 1/80。注意，算法仅在第 23 行输出 ACCEPT ，此时针对 $p^{(F)}$ 和 $p^{(F_1)}{}_1 \times p_2^{(F_2)}$ 的标准接近性检验器输出 ACCEPT。然而，当 p 与所有乘积分布的距离至少为 ϵ时，由于以下观察，$p^{(F)}$ 必定与 $p^{(F_1)}{}_1 \times p^{(F_2)}{}_2$ 的距离至少为 ϵ：(i) 如果 p 与任一乘积分布的距离至少为 ϵ，那么它与其边缘分布的乘积 $p_1 \times p_2$的距离必定至少为 ϵ；(ii) 根据 F的构造，$(p_1 \times p_2)^{(F)} = p^{(F_1)}{}_1 \times p_2^{(F_2)}$；(iii) 扁平化保持全变差距离 [DK16]。因此，接近性检验器的输入相距至少为 ϵ，且接近性检验器输出 ACCEPT的概率至多为 1/80。

最后，我们证明，若 p是乘积分布，则在给定（ REJECT ）的条件下，算法输出 1/40的概率至多为5。注意，若发生以下任一事件，算法将输出 REJECT：（i）$\hat{s}_1 > c's_1$ 或$\hat{s}_2 > c's_2$（第6行）；（ii）接近性检验器输出 REJECT （第23行）；（iii）$L > 10c^2\tau_1\tau_2$（第21行）。我们证明，在给定（5）且 p 是乘积分布的条件下，（i）和（ii）发生的概率很低，

而（iii）不会发生。

首先，回顾一下，$\hat{s}_1$ 和 $\hat{s}_2$ 是均值分别为 s_1 和 s_2 的泊松随机变量。因此，根据马尔可夫不等式和并集界，我们可以将 $\hat{s}_1 > c's_1$ 或 $\hat{s}_2 > c's_2$ 的概率界定为 $1/80$。此外，如果 $p = p_1 \times p_2$，那么 $p^{(F)} = (p_1 \times p_2)^{(F)}$，且接近性检验器输出 REJECT 的概率至多为 $1/80$。最后，为了证明 (iii) 不会发生，注意若 p 是乘积分布，则展平分布 $p^{(F)}$ 等于其展平边缘分布的乘积，$p_1^{(F_1)} \times p_2^{(F_2)}$。特别地，这两个分布的平方范数必定相同。然而，正如我们在证明接近性检验器成功时所论证的，我们可以将展平边缘分布的乘积的平方范数界定为

$$\left\| p_1^{(F_1)} \times p_2^{(F_2)} \right\|_2^2 \le 4c^2\tau_1^2\tau_2^2. \tag{15}$$

在条件（5）下，估计值 L 被 ℓ_2^2-范数（属于 $p^{(F)}$）的常数倍上界。将这一事实与（15）结合，我们有

$$L \le \frac{3}{2}\left\| p^{(F)} \right\|_2^2 = \frac{3}{2}\left\| p_1^{(F_1)} \times p_2^{(F_2)} \right\|_2^2 \le 6c^2\tau_1^2\tau_2^2. \tag{16}$$

因此，如果 p 是乘积分布，$L > 10c^2\tau_1^2\tau_2^2$ 将永远不会求值为真，且该算法不会输出 REJECT 于第 21 行。最终，通过对事件 (i) 和 (ii) 取并集界，该算法在乘积分布上输出 REJECT 的概率至多为 $1/80$，条件为（5）。

我们至此已经证明，范数的估计值以至少 $1-1/40$ 的概率满足 (5)。在该事件的条件下，我们已证明：(i) 当 $\mathsf{d}_{\mathrm{TV}}(p, \hat{p}) \le \alpha$ 时，检验器输出信息不准确的概率至多为 $1/40$；(ii) 当 p 与乘积的距离至少为 ϵ 时，检验器输出 ACCEPT 的概率至多为 $1/80$；以及 (iii) 当 p 为乘积分布时，检验器输出 REJECT 的概率至多为 $1/40$。通过对范数估计值的失败以及检验器其余部分的条件失败取并集界，每种情形下的总失败概率至多为 $1/20$。因此，算法 1 是一个 $(\alpha, \epsilon, 0.1)$-增强独立性检验器。

样本复杂度算法1 必须从以下三个环节中抽取样本：p 对分布进行扁平化、估计扁平化分布的平方范数，以及在扁平化分布上运行标准接近性检验器。

为构造扁平化结果，我们最多抽取 $c's_1 + c's_2 = \mathcal{O}(s_1 + s_2)$ 个样本。根据事实A.4中描述的扁平化构造方法 F_1，$p_1^{(F_1)}$ 的定义域大小为：

$$\sum_{i=1}^{n} (\lfloor n\hat{p}_1(i) \rfloor + N_i + 1) \le \sum_{i=1}^{n} (n\hat{p}_1(i) + 1) + \sum_{i=1}^{n} N_i + \sum_{i=1}^{n} 1 = 3n + s_1 = \mathcal{O}(n). \tag{17}$$

同理，$p_2^{(F_2)}$ 的定义域大小至多为 $3m + s_2 = \mathcal{O}(m)$。因此，$p^{(F)}$ 的定义域大小为 $\mathcal{O}(nm)$。

根据事实 A.2，估计定义域大小为 n 的分布的 ℓ_2-范数需要 $\mathcal{O}(\sqrt{n})$ 个样本。因此，要估计我们的扁平化分布的范数 $p_1^{(F_1)}, p_2^{(F_2)}$ 和 $p^{(F)}$，我们必须分别抽取 $\mathcal{O}(\sqrt{n})$、$\mathcal{O}(\sqrt{m})$ 和 $\mathcal{O}(\sqrt{nm})$ 个样本。

应用事实 A.3，标准接近性检验器分别从 $p^{(F)}$ 和 $p_1^{(F_1)} \times p_2^{(F_2)}$ 中抽取 $\mathcal{O}(nm\sqrt{\tau_1\tau_2}/\epsilon^2)$ 个样本。注意，我们可以通过从 $p^{(F)}$ 中抽取两个独立样本，并返回其中一个点的第一个坐标和另一个点的第二个坐标，从而从 $p_1^{(F_1)} \times p_2^{(F_2)}$ 中抽取一个样本。因此，接近性检验器的总体样本复杂度为 $\mathcal{O}(nm\sqrt{\tau_1\tau_2}/\epsilon^2)$。

最终，替 代入 τ_1 并 τ_2 及其实现（第 9 行），并包含每个分量，th e

算法 1 的样本复杂度为:

$$\mathcal{O}\left(s_1+s_2+\sqrt{n}+\sqrt{m}+\sqrt{nm}+\frac{nm}{\epsilon^2}\sqrt{\left(\frac{2\alpha}{s_1}+\frac{4}{n}\right)\left(\frac{2\alpha}{s_2}+\frac{4}{m}\right)}\right) \tag{18}$$

$$=\mathcal{O}\left(s_1+m\alpha+\sqrt{nm}+\frac{nm}{\epsilon^2}\sqrt{\left(\frac{\alpha}{s_1}+\frac{1}{n}\right)\left(\frac{\alpha}{m\alpha}+\frac{1}{m}\right)}\right) \tag{19}$$

$$=\mathcal{O}\left(s_1+m\alpha+\sqrt{nm}+\frac{n\sqrt{m}}{\epsilon^2}\sqrt{\left(\frac{\alpha}{s_1}+\frac{1}{n}\right)}\right), \tag{20}$$

其中第二行是通过代入 $s_2=m\alpha$ 得出的。回想一下， $s_1=$最小值 $\left(\frac{n^{2/3}m^{1/3}\alpha^{1/3}}{\epsilon^{4/3}},n\alpha\right)$。分别代入这些可能性后，算法1的总体样本复杂度为:

$$\mathcal{O}\left(\max\left(\frac{n^{2/3}m^{1/3}\alpha^{1/3}}{\epsilon^{4/3}},\frac{\sqrt{nm}}{\epsilon^2}\right)\right). \tag{21}$$

□

3 多元增强独立性检验的上界

首先注意到，二维增强独立性检验器可以直接推广到三维情形；该三维检验器的描述和样本复杂度分析见附录B。我们高维构造的核心思想是以近乎均衡的方式将坐标划分为至多三组，使得每个包含多个坐标的组的总域大小至多为 $\sqrt{N}$，其中 N表示完整定义域的大小。由于组数至多为三，我们可以使用二维或三维增强独立性检验器来验证各组之间的独立性。此外，每组定义域大小的上界确保我们可以仅使用 $\tilde{\epsilon}$接近真实分布的经验分布，并且只需 $\mathcal{O}(\sqrt{N}/\epsilon^2)$个样本，这由事实A.6得出。随后，使用算法3检验每组内部的独立性，其正确性和样本复杂度将在第3.1节中分析。现在，我们将介绍划分策略，以完成算法2 的构造，使其成为用于 d维分布的增强独立性检验器。具体而言，我们将定义域大小最大的坐标单独分出（假设其为 n_1），并根据 n_1 大于还是小于$\sqrt{N}$，将其余坐标划分为另外一组或两组:

—

如果 $n_1\geq\sqrt{N}$，则将其余坐标归入一个单独的块 $B=\{2,3,\ldots,d\}$，其诱导定义域大小满足 $N_B\leq\sqrt{N}$。在此情况下，将分布 p 视为定义在 $[n_1]\times[N_B]$上的二维分布，并应用二维增强独立性检验器来检验第一个坐标与该块之间的独立性 B。在接受的条件下，算法随后调用算法3 来验证 B中的坐标之间是否相互独立。

否则，如果 $n_1<\sqrt{N}$，则将其余坐标划分为两个不相交的块 B和 C。令 t为满足 $\prod_{i=1}^{t}n_i>\sqrt{N}$ 的最小索引，其中 $\prod_{i=1}^{t-1}n_i\leq\sqrt{N}$。定义 $B=\{2,3,\ldots,t\}$和 $C=\{t+1,\ldots,d\}$。第一个不等式确保 $N_C\leq\sqrt{N}$，而将第二个不等式乘以 n_t/n_1 （给定 $n_t\leq n_1$）可确保 $N_B\leq\sqrt{N}$。随后将该分布视为定义在 $[n_1]\times[N_B]\times[N_C]$上的三维分布，并应用三维增强独立性检验器。如果该检验输出 ACCEPT，则随后使用算法 3 分别检验块 B 和 C内部的独立性，并且仅当两项检验均成功时，算法才输出 ACCEPT。

算法 2： 针对 d维分布的增强独立性检验器

输入： 对 p 上的 $\prod_{i=1}^{d}[n_i]$ 进行样本访问（假设 $n_1 \geq n_2 \geq \cdots \geq n_d \geq 2$）；预测 $\hat{p}$；预测误差的估计值 预测误差 $\alpha \in [0,1]$；接近参数 $\epsilon \in$ （0，1）；失败概率 $\delta = 0.1$

增强独立性检验器-d维 $(p,\ \hat{p},\ \alpha,\ \epsilon,\ \delta)$ **：**

- $\epsilon' \leftarrow \epsilon/12$
- $\delta' \leftarrow \delta/5$
- $N \leftarrow \prod_{i=1}^{d} n_i$
- **如果** $n_1 \geq \sqrt{N}$，**则**
  - $B \leftarrow \{2,3,\ldots,d\}$ $N_B \leftarrow \prod_{i\in B} n_i$
  - 将 p 视为定义在 $[n_1] \times \prod_{i\in B}[n_i]$ 上的二维分布
  - $out \leftarrow$ Augmented-Independence-Tester-2D $(n_1,\ N_B,\ p,\ \hat{p},\ \alpha,\ \epsilon',\ \delta')$ // 检验 $p = p_1 \times p_B$
  - **如果** $out \in \{\text{Inaccurate Information}, \text{Reject}\}$ **则**
    - **返回** out
  - **end**
  - $p_B \leftarrow \pi_B(p)$ // 将 p 的样本投影到 B 中的坐标
  - **返回** TestIndependenceByLearning $(B,\ p_B,\ \epsilon',\ \delta')$ // 检验 $p_B = \prod_{i\in B} p_i$
- **else**
  - 选择满足 $\prod_{i=1}^{t}$ 的最小 t $\prod_{i=1}^{t} n_i \geq \sqrt{N}$
  - $B \leftarrow \{2,\ldots,t\}$，$C \leftarrow \{t{+}1,\ldots,d\}$
  - $N_B \leftarrow \prod_{i\in B} n_i$，$N_C \leftarrow \prod_{i\in C} n_i$
  - 将 p 视为定义在 $[n_1] \times \prod_{i\in B}[n_i] \times \prod_{i\in C}[n_i]$ 上的三维分布
  - $out \leftarrow$ 增强型三维独立性测试器 $(n_1,\ N_B,\ N_C,\ p,\ \hat{p},\ \alpha,\ \epsilon',\ \delta')$ // 测试是否 $p = p_1 \times p_B \times p_C$
  - **如果** $out \in \{\text{Inaccurate Information}, \text{Reject}\}$，**则**
    - **返回** out
  - **end**
  - $p_B \leftarrow \pi_B(p)$ // 将 p 的样本投影到 B 中的坐标
  - $p_C \leftarrow \pi_C(p)$ // 将 p 的样本投影到 C 中的坐标
  - $out_B \leftarrow$ TestIndependenceByLearning $(B,\ p_B,\ \epsilon',\ \delta')$ // 检验 $p_B = \prod_{i\in B} p_i$
  - $out_C \leftarrow$ TestIndependenceByLearning $(C,\ p_C,\ \epsilon',\ \delta')$ // 检验 $p_C = \prod_{i\in C} p_i$
  - **如果** $out_B = \text{Reject}$ 或 $out_C = \text{Reject}$ **则**
    - **返回** Reject
  - **else**
    - **返回** Accept
  - **end**
- **end**

结束

3.1 通过学习检验独立性

在本节中，我们描述一种基于学习的过程，用于在坐标集内检验独立性S ，即通过构造 S上诱导分布的经验直方图，计算其单变量边缘分布，并验证学习得到的分布是否分解为其边缘分布的乘积。我们算法的伪代码见算法3，其正确性与样本复杂度在定理5中加以分析。

算法3： TestIndependenceByLearning

输入： 对 p_S 在 $\prod$上的样本访问 $_{i\in S}[n_i]$；指标集 $S \subseteq [d]$，其中 $N_S := \prod_{i\in S} n_i$；接近度参数 $\epsilon \in (0,\ 1)$；失败概率 δ

TestIndependenceByLearning $(S, p_S, \epsilon, \delta)$**:**

$\delta' \leftarrow \delta/(|S|+1)$

$\eta \leftarrow \epsilon/7$ // 预算：学习精度 η 和乘积检验松弛量

$t \leftarrow \Theta\left(\frac{N_S+\log(\ \ 1/\delta'}{\eta^2}\right)$ // 在大小为 N_S 的定义域上进行直方图学习

抽取 t独立同分布样本 $X_S^{(1)}, \ldots\ldots, X_S^{(t)} \sim p_S$

构造关于 $\prod$的经验分布 $\tilde{p}_S$ $_{i\in S}$来自 $\{X_S^{(j)}\}_{j=1}^{t}[n_i]$

计算单变量边缘分布 $\{\tilde{p}_i\}_{i\in S}$ 和乘积分布 $\tilde{p}_\Pi \leftarrow \prod$ $_{i\in S}\tilde{p}_i$

$\Delta \leftarrow \|\tilde{p}_S - \tilde{p}_\Pi\|_{\mathsf{TV}}$ // 由于 $\tilde{p}_S$ 是显式的，因此进行精确计算

如果 $\Delta \le 6\eta$ **那么**

  返回 Accept

else

  返回 Reject

end

结束

定理 5。设 S 是一个坐标集，p_S 是定义在 $\prod_{i\in S}[n_i]$ 上的一个分布，其总域大小为$N_S := \prod_{i\in S} n_i$。对于任意接近参数 $\epsilon \in (0,\ 1)$ 以及常数失败概率 δ，存在一种算法，给定来自 p_S的 $t = \Theta\left(\frac{N_S}{\epsilon^2}\right)$ 个独立同分布样本后，能够以至少 $1-\delta$ 的概率判别以下两种情况：p_S 是乘积分布，或者 p_S 与任意乘积分布的全变差距离至少为 ϵ。

证明。我们通过证明算法 3 能够正确判别乘积分布与 ϵ-远离独立性的分布，且使用所述数量的样本，从而证明该定理。为此，我们首先在引理 3.1中证明经验分布及其边缘分布与真实分布及其相应边缘分布接近。随后利用这一事实证明算法 3的正确性。

引理3.1。设 $S \subset [d]$ 为任意坐标集，并令 p_S 表示定义在 $\prod_{i\in S}[n_i]$上的诱导分布，其定义域大小为 $N_S := \prod_{i\in S} n_i$。对于任意接近参数 $\eta \in (0,\ 1)$ 及常数失败参数$\delta \ge 10^{-4}$，使用 $t = {}^{N_S+}{}_{\log((|S|+1)/\delta)\ \eta^2}$ 个独立同分布样本，我们可以构造经验分布 $\tilde{p}_S$，使得以至少 $1-\delta$ 的概率，以下两个陈述同时成立：

(i) $\|\tilde{p}_S - p_S\|_{\mathsf{TV}} \le \eta$.

(ii) $\sum_{i\in S} \|\tilde{p}_i - p_i\|_{\mathsf{TV}} \le 5\eta$,

这里，$\tilde{p}_i$ 和 p_i 分别表示 $\tilde{p}_S$ 和 p_S 在坐标 i上的单变量边缘分布。

证明。令 $\delta' := \delta/(|S|+1)$。我们将证明，两个不等式 (i)–(ii) 以至少 $1-\delta$ 的概率同时成立，方法是建立 $|S|+1$ 高概率事件并应用并集界。

令 $\mathcal{E}_0 = \{\|\tilde{p}_S - p_S\|_{\mathsf{TV}} \le \eta\}$ 表示 (i)中的事件。将事实 A.6 应用于诱导分布 p_S，其定义域大小为 N_S，失败概率为 δ'。根据给定的 t选择，可得

$$\mathbf{Pr}[\mathcal{E}_0] = \mathbf{Pr}[\|\tilde{p}_S - p_S\|_{\mathsf{TV}} \le \eta] \ge 1 - \delta'.$$

所有边缘分布之间的全变差距离的一个平凡界为 $\sum_{i\in S} \|\tilde{p}_i - p_i\|_{\mathsf{TV}} \le |S|\cdot\eta$，但通常对我们的目的而言过于宽松。为了证明 (ii) 中所述的更紧上界，我们利用这样一个事实：对每个 $i \in S$，经验直方图 $\tilde{p}_S$ 的单变量边缘分布 $\tilde{p}_i$，与将相同的 t 样本投影到坐标 i上得到的经验分布一致。因此，我们可以将推论 A.7 应用于单变量边缘分布 p_i（其定义域大小为 n_i），并取失败概率 δ'，从而对每个 $i \in S$，

$$\mathbf{Pr}\left[\|\tilde{p}_i - p_i\|_{\mathsf{TV}} \le \sqrt{\frac{n_i + \log\left(\frac{|S|+1}{\delta}\right)}{t}}\right] \ge 1 - \delta'.$$

令 $\mathcal{E}_i$ 表示上述事件。对 $\mathcal{E}_0$ 和 $\{\mathcal{E}_i\}_{i\in S}$ 这 $|S|+1$ 个事件应用并集界，

$$\mathbf{Pr}\left[\mathcal{E}_0 \cap \bigcap_{i\in S} \mathcal{E}_i\right] \ge 1-(|S|+1)\delta' = 1-\delta.$$

以此交集事件为条件。将边缘分布的界对 $i \in S$ 求和，得到 s

$$\sum_{i\in S} \|\tilde{p}_i - p_i\|_{\mathsf{TV}} \le \sum_{i\in S} \sqrt{\frac{n_i + \log\left(\frac{|S|+1}{\delta}\right)}{t}}.$$

由于 $t \ge N_S/\eta^2$，我们有 $1/\sqrt{t} \le \eta/\sqrt{N_S}$。因此，

$$\begin{aligned}
\sum_{i\in S} \sqrt{\frac{n_i + \log\left(\frac{|S|+1}{\delta}\right)}{t}} &\le \eta \sum_{i\in S} \sqrt{\frac{n_i + \log\left(\frac{|S|+1}{\delta}\right)}{N_S}} \\
&\le \eta \sum_{i\in S} \left(\sqrt{\frac{n_i}{N_S}} + \sqrt{\frac{\log\left(\frac{|S|+1}{\delta}\right)}{N_S}}\right) && ((\sqrt{a+b} \le \sqrt{a} + \sqrt{b})) \\
&= \eta \sum_{i\in S} \left(\sqrt{\frac{1}{\prod_{j\in S,\, j\ne i} n_j}} + \sqrt{\frac{\log\left(\frac{|S|+1}{\delta}\right)}{\prod_{j\in S} n_j}}\right) && (\text{definition of } N_S) \\
&\le \eta \left(\frac{|S|}{\sqrt{2}^{|S|-1}} + \frac{|S|\sqrt{\log\left(\frac{|S|+1}{\delta}\right)}}{\sqrt{2}^{|S|}}\right) && (\forall j,\ n_j \ge 2) \\
&\le 5\eta.
\end{aligned}$$

由于括号中的表达式随 δ 减小而增大，因此只需在 $\delta = 10^{-4}$处对该表达式进行上界估计。为控制其对 $|S|$ 的依赖性，考虑数列 $a_d := \frac{d}{\sqrt{2}^{d-1}}$ 和 $b_d := \frac{\sqrt[d]{\log((d+1)/\delta)}}{\sqrt{2}^{d}}$。直接应用比值判别法可知，对所有 $d \ge 3$，均有 $a_{d+1}/a_d < 1$ 和 $b_{d+1}/b_d < 1$，因此当 $|S| \ge 3$时，括号中的表达式关于 $|S|$ 递减。因此，其在所有可行的 $|S|$ 中取得的最大值出现在$|S| = 3$处。代入 $|S| = 3$ 和 $\delta = 10^{-4}$，得到数值上界 5。

总而言之，在交集事件 $\mathcal{E}_0 \cap \bigcap_{i\in S} \mathcal{E}_i$（其发生概率至少为 $1-\delta$）上，（i）和（ii）同时成立，证明得证。

□

算法 3 的正确性证明：首先，假设 p_S 是乘积分布，即 $p_S = \prod_{i\in S} p_i$。我们将证明，其经验直方图 $\tilde{p}_S$ 与 6η其边缘分布的乘积相差 $\tilde{p}_\Pi := \prod_{i\in S} \tilde{p}_i$，其概率至少为 $1-\delta$，利用引理 3.1。根据三角不等式以及乘积测度的标准混合（望远镜）论证，我们有：

$$\begin{aligned}
\|\tilde{p}_S - \tilde{p}_\Pi\|_{\mathsf{TV}} &\le \|\tilde{p}_S - p_S\|_{\mathsf{TV}} + \|p_S - \tilde{p}_\Pi\|_{\mathsf{TV}} \\
&= \|\tilde{p}_S - p_S\|_{\mathsf{TV}} + \left\| \prod_{i\in S} p_i - \prod_{i\in S} \tilde{p}_i \right\|_{\mathsf{TV}} \\
&\le \|\tilde{p}_S - p_S\|_{\mathsf{TV}} + \sum_{i\in S} \|p_i - \tilde{p}_i\|_{\mathsf{TV}} \\
&\le \eta + 5\eta = 6\eta
\end{aligned} \tag{22}$$

接下来，假设 p_S 与任何乘积分布的距离至少为 ϵ，具体而言为 $\tilde{p}_\Pi$。我们可以证明，$\tilde{p}_S$ 以至少 $1-\delta$ 的概率，与其边缘分布的乘积相差 $(\epsilon-\eta)$：

$$\begin{aligned}
\|\tilde{p}_S - \tilde{p}_\Pi\|_{\mathsf{TV}} &= \|(p_S - \tilde{p}_\Pi) - (p_S - \tilde{p}_S)\|_{\mathsf{TV}} \\
&\ge \|p_S - \tilde{p}_\Pi\|_{\mathsf{TV}} - \|p_S - \tilde{p}_S\|_{\mathsf{TV}} \\
&> \epsilon - \eta
\end{aligned} \tag{23}$$

结合（22）和（23），如果我们选择 $\eta = \epsilon/7$，那么完备性区域和可靠性区域不会重叠。因此，以至少 $1-\delta$的概率，算法 3 能够正确检验坐标 S上的独立性。

样本复杂度分析：首先注意，一旦估计出经验分布 $\tilde{p}_S$，算法后续的所有步骤都只是计算操作，无需再从 p_S 获取额外样本。学习 $\tilde{p}_S$ 所需的样本数为：

$$\mathcal{O}\left(\frac{N_S + \log((|S|+1)/\delta)}{\epsilon^2}\right) = \mathcal{O}\left(\frac{N_S + \log(|S|+1) + \log(1/\delta)}{\epsilon^2}\right) = \mathcal{O}\left(\frac{N_S}{\epsilon^2}\right),$$

其中，最后一个等式成立是因为 $N_S = \prod_{i\in S} n_i \ge 2^{|S|}$，因此 $\log(|S|+1) = O(\log N_S) = o(N_S)$，且 δ 是常数。

□

3.2 多变量增强独立性检验器的分析

定理 6。设 p 是定义域 $\prod_{i=1}^d [n_i]$ 上的一个分布，且 $\hat{p}$ 是对同一定义域上 p 的一个预测。不失一般性，假设 $n_1 \ge n_2 \ge \cdots \ge n_d \ge 2$。令 $N := \prod_{i=1}^d n_i$ 表示总域大小。对于所有 $\epsilon \in (0,\ 1)$ 以及 $\alpha \in [0,1]$，都存在一种算法，它能够对 p 进行样本访问，并对 $\hat{p}$ 进行显式访问；该算法是一个 $(\alpha,\ \epsilon,\ \delta = 0.1)$ 增强独立性检验器，其样本复杂度为 $O\left(\max\left\{\frac{\sqrt{N}}{\epsilon^2},\ \frac{n_1^{1/3} N^{1/3} \alpha^{1/3}}{\epsilon^{4/3}}\right\}\right)$。

证明。 我们首先证明，算法 2 以高概率成功完成增强独立性检验，然后给出其样本复杂度的上界。

正确性证明： 我们验证算法2 满足定义 1.1 中的三个条件，从而证明其正确性：

(i) 如果 p 是乘积分布，算法可能在两个步骤中错误地拒绝。首先，算法调用二维或三维增强独立性检验器。根据这些检验器的正确性，此步骤中拒绝的概率至多为 $\delta'/2 = 0.01$。其次，算法可能在基于学习的

独立性检验所针对的集合 B 或 C上。由于 p 是乘积分布，其边缘分布 p_B 和 p_C 也都是乘积分布。根据定理 5，每次调用算法 3 拒绝的概率至多为 $\delta' = 0.02$。对所有可能的拒绝事件应用并集界，算法 2 输出 REJECT的总概率至多为 $(0.01 + 0.02 + 0.02) \leq \frac{\delta}{2}$

(ii) 如果 p 与任意乘积分布都相距至少 ϵ，我们证明，其中一个顺序检验器仅以较低概率检测出这种差距并输出 ACCEPT 。由于算法的全局接受遵循合取规则，这个单项低概率构成算法输出 ACCEPT 总概率的上界。我们给出三维情形的论证；二维情形的证明类似。考虑定义域分别为 B 和 C 的两个任意乘积分布，记为 $r_B = \prod_{i \in B} r_i$ 和 $r_C = \prod_{i \in C} r_i$。构造一个在相同定义域 p 上的乘积分布，记为 $r = p_1 \times r_2 \times \ldots \times r_d$。于是有：

$$\begin{aligned} \epsilon \leq \|p - r\| &\leq \|p - p_1 \times p_B \times p_C\| + \|p_1 \times p_B \times p_C - p_1 \times r_B \times r_C\| \\ &\leq \|p - p_1 \times p_B \times p_C\| + \|p_B - r_B\| + \|p_C - r_C\| \end{aligned} \tag{24}$$

因此，右侧的三个项中至少有一项必须大于 24。 $\epsilon/3$我们逐一分析：

- 如果 $\|p - p_1 \times p_B \times p_C\| \geq \epsilon/3$，那么根据引理 A.1 的逆否命题， p 与定义域 $\frac{\epsilon}{\mathrm{T2}}$ 上的任意乘积分布都相距至少 $[n_1] \times [N_B] \times [N_C]$。由于调用三维增强检验器时使用的参数为$\epsilon' = \frac{\epsilon}{12}$ 和 $\delta' = \frac{\delta}{5}$，它将以至多 ACCEPT 的概率 $\frac{\delta'}{2} = 0.01$
- 如果 $\|p_B - r_B\| \geq \epsilon/3 > \epsilon' = \epsilon/12$，则 p_B 与任意乘积分布相距 ϵ'-远。根据定理5，将算法3 应用于集合 B 时，输出 ACCEPT 的概率至多为 $\delta' = 0.02$。
- 如果 $\|p_C - r_C\| \geq \epsilon/3$，则同样的论证适用于集合 C，且算法3 输出 ACCEPT 的概率至多为 $\delta' = 0.02$。

（iii）算法2 仅能在初次调用二维或三维增强独立性检验器时输出 INACCURATE INFORMATION 。由于全变差距离在边缘化下会收缩，任何全局全变差保证都会立即蕴含二维/三维调用所使用的边缘分布具有相同保证。根据这些检验器的正确性保证，这种情况发生的概率至多为 $\delta'/2 = 0.01$。至此完成正确性证明。

样本复杂度分析：在第一种情形中，我们知道 $n_1 \geq \sqrt{N} \geq N_B$。因此，定理4可用于计算以接近参数 $\epsilon' = \epsilon/12$ 执行二维增强检验的样本复杂度，如下所示：

$$\mathcal{O}\left(\max\left\{\frac{\sqrt{n_1 N_B}}{(\epsilon')^2}, \frac{n_1^{2/3} N_B^{1/3} \alpha^{1/3}}{(\epsilon')^{4/3}}\right\}\right) = \mathcal{O}\left(\max\left\{\frac{\sqrt{N}}{\epsilon^2}, \frac{n_1^{1/3} N^{1/3} \alpha^{1/3}}{\epsilon^{4/3}}\right\}\right) \tag{25}$$

在第二种情况下，根据构造，我们知道 n_1， N_B， $N_C \leq \sqrt{N}$。令 $n_2 = N_B$， $n_3 = N_C$表示算法 4在样本复杂度中涉及的定义域大小。使用接近参数 $\epsilon' = \epsilon/12$ 执行三维增强检验的样本复杂度如下：

$$\mathcal{O}\left(\max_l\left\{\frac{\sqrt{N}}{(\epsilon')^2}, \frac{n_l^{1/3} N^{1/3} \alpha^{1/3}}{(\epsilon')^{4/3}}\right\}\right) \leq \mathcal{O}\left(\max\left\{\frac{\sqrt{N}}{\epsilon^2}, \frac{N^{1/6} N^{1/3} \alpha^{1/3}}{\epsilon^{4/3}}\right\}\right) = \mathcal{O}\left(\frac{\sqrt{N}}{\epsilon^2}\right) \tag{26}$$

对于这两种情况，根据定理 5，以准确度 $\eta = \epsilon'/7 = \epsilon/84$ 学习 p_B 或p_C 的经验分布所需的样本数也有如下界：

$$\mathcal{O}\left(\frac{N_B}{\eta^2}\right) + \mathcal{O}\left(\frac{N_C}{\eta^2}\right) \leq \mathcal{O}\left(\frac{\sqrt{N}}{\epsilon^2}\right), \tag{27}$$

其中，最后一个不等式由 $N_B,\ N_C \le \sqrt{N}$ 得出。因此，定理 6 的样本复杂度可通过结合（25）、（26）和（27）得到。 □

4 二元增强独立性检验的下界

本节中，我们为二维分布的增强独立性检验建立一个下界，该下界与第 2.2 节给出的上界相匹配。

定理 7. 设 p 是定义在 $[n]\times[m]$ 上的一个分布，满足 $n \ge m \ge 2$ 且 $\log n \le m$。假设给定 $\alpha \in [0,1]$ 和 $\epsilon \in (0,\ 1)$。任何 $(\alpha, \epsilon, \delta := 0.1)$增强独立性检验器若要成功，必须使用 $\Omega\left(\max\left\{\frac{\sqrt{nm}}{\epsilon^2},\ \frac{n^{2/3}m^{1/3}\alpha^{1/3}}{\epsilon^{4/3}}\right\}\right)$个样本。

证明。 根据 n, m, α 与 ϵ之间的关系，我们分两种情况讨论。每种情况分别给出所需下界中的一项。为证明第一项，我们从标准独立性检验进行归约。为证明下界的第二项，我们采用信息论方法，证明要判别两个分布族——其中一个由乘积分布组成，另一个由与任何乘积分布都相距甚远的分布组成——至少需要所述的样本复杂度。

4.1 情形 1： $\alpha \le \frac{\sqrt{m}}{\sqrt{n}\epsilon^2}$

在这种情形下，下界的主导项为 $\Omega\left(\sqrt{nm}/\epsilon^2\right)$。这个界可通过使用标准独立性检验的困难实例进行归约得到。具体而言，设 $\mathcal{A}_{aug}$ 是一个增强型独立性检验器。我们构造一个标准独立性检验器 $\mathcal{A}_{std}$，从分布 p 中抽取样本，并将这些样本传递给 $\mathcal{A}_{aug}$。$\mathcal{A}_{std}$ 将 $\mathcal{A}_{aug}$ 的预测设为均匀分布，将准确度估计设为 $\alpha \le \sqrt{m}/(m\epsilon^2)$，将接近参数设为 ϵ，并将失败概率设为 δ。如果 $\mathcal{A}_{aug}$ 输出 REJECT 或 INACCURATE INFORMATION，则 $\mathcal{A}_{std}$ 输出 REJECT。否则，如果 $\mathcal{A}_{aug}$ 输出 ACCEPT，则 $\mathcal{A}_{std}$ 输出 ACCEPT。

标准独立性检验的已有下界（例如 [ADK15、DK16]）构造了一个困难实例，其中包含均匀分布和一个与任何乘积分布都相距 ϵ 的分布。我们考虑同一个困难实例。

在困难实例的 ACCEPT 情形中，当分布均匀时，分布与预测之间的距离为 0，因此满足 $\mathrm{d}_{\mathrm{TV}}(p,\ \hat{p}) \le \alpha$。因此，$\mathcal{A}_{aug}$ 以至多 $\delta/2$ 的概率输出 INACCURATE INFORMATION。它输出 REJECT 的概率也至多为 $\delta/2$，这意味着 $\mathcal{A}_{std}$ 输出 REJECT 的概率至多为 δ。在困难实例的 REJECT 情形中，当分布与乘积分布相距 ϵ 时，$\mathcal{A}_{aug}$——因此 $\mathcal{A}_{std}$——输出 ACCEPT 的概率至多为 $\delta/2$。因此，$\mathcal{A}_{std}$ 成功检验了该困难实例的独立性，而标准独立性检验的下界 $\Omega\left(\sqrt{nm}/\epsilon^2\right)$ 也适用于增强独立性检验。

4.2 情形 2： $\alpha > \frac{\sqrt{m}}{\sqrt{n}\epsilon^2}$

在这种情形下，下界的主导项为

$$\Omega\left(\frac{n^{2/3}m^{1/3}\alpha^{1/3}}{\epsilon^{4/3}}\right). \tag{28}$$

我们通过构造两个增强独立性检验器必须能够区分的分布族来证明这一界限，然后证明除非样本数量足够大，否则从信息论角度无法区分这两个分布族。

设 k 为一个整数，使得存在一个足够小的常数 $c>0$，满足

$$k \le c \cdot \left(\frac{n^{2/3} m^{1/3} \alpha^{1/3}}{\epsilon^{4/3}} \right). \tag{29}$$

给定 k，我们可以按如下方式构造增强独立性检验的困难实例。令 $\hat{}p$ 为 $[n] \times [m]$ 上的均匀分布，并令 $\epsilon' \in (0,1)$ 和 $\alpha' \in [0,1]$，其定义将在后文给出。然后，按照以下过程构造 p:

1. 令 X 以相等的概率取 0 或 1。

2. 按如下方式抽取 c_i：

$$c_i = \begin{cases} \frac{1}{k} & \text{w.p. } \frac{\alpha' k}{n} \\ \frac{1}{n} & \text{w.p. } 1 - \frac{\alpha' k}{n} \end{cases}. \tag{30}$$

3. 如果 $X=0$，则对所有 $i \in [n], j \in [m]$，令 $\mu(i,j) = \frac{1}{m}$。

4. 如果 $X=1$，则对所有 $i \in [n], j \in [m]$，按如下方式抽取 $\mu(i,j)$:

$$\mu(i,j) = \begin{cases} \frac{1}{m} & \text{if } c_i = \frac{1}{k} \\ \frac{1+\epsilon'_{ij}}{m} & \text{if } c_i = \frac{1}{n}, \text{ where } \epsilon'_{ij} = \begin{cases} \epsilon' & \text{w.p. } 1/2 \\ -\epsilon' & \text{w.p. } 1/2 \end{cases} \end{cases}. \tag{31}$$

5. 对所有 $i \in [n], j \in [m]$，将 $q(i,j)$ 定义为:

$$q(i,j) = c_i \cdot \mu(i,j) \tag{32}$$

6. 构造一个分布 p:

$$p(i,j) = \frac{q(i,j)}{s_i \cdot C}, \tag{33}$$

其中

$$s_i = \sum_{j=1}^{m} \mu(i,j) \quad \text{and} \quad C = \sum_{i=1}^{n} c_i. \tag{34}$$

假设存在一个算法 $\mathcal{A}$，用于增强独立性检验，且其成功概率至少为 0.9。给定分布 p 在 $[n] \times [m]$ 上的一组 $\mathcal{S}$ 个样本、预测值 $\hat{}p$ 以及预测误差的估计值 α，$\mathcal{A}$ 返回 ACCEPT、REJECT 或 INACCURATE INFORMATION。

我们将证明，收到来自 p 的样本后，该检验器必须以高概率判别 $X=0$ 和 $X=1$。然而，我们不直接从 p 抽取样本集 $\mathcal{S}$，而是对所有满足 $\in [n] \times [m]$ 的 (i,j)，抽取 $a_{i,j} \sim \text{Poi}(k \cdot q(i,j))$，并令 $a_{i,j}$ 表示 $\mathcal{S}$ 中 (i,j) 的副本数。只要 $|\mathcal{S}| \le k/100$，这一过程就等价于从 p抽取 $\mathcal{O}(k)$ 个样本。在引理 4.1 中，我们证明，以至少 0.99 的概率：(i) $|\mathcal{S}| \ge k/100$；(ii) 如果 $X=0$，则 p 是一个乘积分布，且与预测值 $\hat{}p$ 的距离至多为 α；(iii) 如果$X=1$，那么 p 与乘积分布相距甚远。

引理 *4.1.* 设 $n \ge m$ 为足够大的整数，使得 $m \ge \log(n)$。设 k 为整数，且令$\epsilon \in (0, 1/192)$ $\alpha \in (0,1)$。假设 $\epsilon' = 192\epsilon$与 $\alpha' = 2\alpha/3$。给定 ϵ'、α'，令 X、q和 p按方程 *(30)* 至 *(34)* 抽取。令 $\hat{}p$ 为 $[n] \times [m]$上的均匀分布。令 $\mathcal{S}$ 为一个多重集，使得对所有 $(i,j) \in [n] \times [m]$，$(i,j) \in \mathcal{S}$的副本数为泊松分布$(k \cdot q(i,j))$。则以下事件以至少 *0.95* 的概率同时成立：

1. $|\mathcal{S}| \ge k/100$。
2. 如果 $X=0$, p 是乘积分布，且与 $\hat{}p$ 的距离至多为 α。

3. 若 $X=1$, p 与所有乘积分布的距离至少为 ϵ。

引理 4.1 的证明推迟至附录 C.1。如果 p 和 $\mathcal{S}$ 满足引理中描述的全部三个标准，我们就称 $(p,\mathcal{S})$ 有效。当 $(p,\mathcal{S})$ 有效时，我们观察到 $\mathcal{A}$ 能以至少 0.925 的概率判别$X=0$ 和 $X=1$。具体而言，假设存在一个算法 $\mathcal{A}'$，它接收来自有效构造 $(p,\mathcal{S})$ 的 $\mathcal{S}$，以及预测 $\hat{}p$ 和参数 ϵ' 和 α'，并输出

$$\mathcal{A}'(\mathcal{S},\hat{p},\alpha,\epsilon)=\begin{cases}0 & \text{if } \mathcal{A}(\mathcal{S},\hat{p},\alpha',\epsilon')=\text{ACCEPT}\\ 1 & \text{if } \mathcal{A}(\mathcal{S},\hat{p},\alpha',\epsilon')=\text{REJECT or INACCURATE INFORMATION.}\end{cases} \tag{35}$$

若 $X=0$，那么，由于 $(p,\ \mathcal{S})$ 是有效的，p 是乘积分布，且 $\hat{}p$ 与 p 的距离为 α'。由于 $\mathcal{A}$ 是一个 $(\epsilon',\ 0。1)$-增强独立性检验器，这意味着 $\mathcal{A}$ 输出 REJECT 或 INACCURATEINFORMATION 的概率至多为 0。05。应用并集界可得，$\mathcal{A}$ 既不输出 REJECT 也不输出 INACCURATE INFORMATION 的概率至少为 0.9。类似地，若 $X=1$, p 与某个乘积分布的距离为 ϵ'，则 $\mathcal{A}$ 输出 ACCEPT 的概率至多为 0.05。根据全概率公式，我们最终得到：

$$\mathbf{Pr}[\mathcal{A}'(\mathcal{S},\hat{p},\alpha',\epsilon')=X\mid(p,\mathcal{S})\text{ valid}]\geq 0.925. \tag{36}$$

然而，通过信息论论证，我们可以证明没有算法能够在输入 X 时 $\mathcal{S}$以大于 0。51 的概率恢复。具体而言，我们将证明，给定 k （按（29）中的方式选取），样本集与 X 之间的互信息为 $o(1)$。信息论中的一个标准结果随即表明，没有算法能够以高于随机猜测的概率恢复 X。

将 $\mathcal{S}$表示为一组 n随机计数向量 $A_i=(a_{i,\ 1},\ a_{i,\ 2},\ \cdots,\ a_{i,\ m})$。接下来，我们将证明 X 与这组向量之间的互信息很小。首先，我们应用互信息链式法则，并利用给定 X 时 A_i 各项之间的条件独立性：

$$I(X;A_1,A_2,\cdots,A_n)=\sum_{i=1}^{n}I(X;A_i\mid A_{1:i-1})\leq\sum_{i=1}^{n}I(X;A_i)=nI(X;A), \tag{37}$$

由于每个 A_i 都服从相同的分布，令 A 表示任意一个 A_i，并将（37）改写为：

$$I(X;A_1,\cdots,A_n)\leq n\cdot I(X;A). \tag{38}$$

在引理 4.2中，我们对 X 与 A之间的互信息进行界定：

引理 4.2。设 $\alpha\in[0,1]$,$\epsilon\in(0,\ 1)$ $k\in\mathbb{N}$。设 X以相同的概率取值为0或1。设 q是一个定义在 $[n]\times[m]$ 上的测度，按照 (32) 中的方法，利用 X构造而成。对于任意的 $i\in[n]$，设 $A:=A_i$ 是一个计数向量，满足 $a_{i,\ j}\sim$ 泊松分布$(kq(i,j))$ 对所有 $j\in[m]$均成立。则 X 与 A 之间的互信息满足如下界：

$$I(X;A)\leq\mathcal{O}\left(\frac{\epsilon^4k^3}{n^3m\alpha}\right). \tag{39}$$

引理的证明4.2 推迟至附录C.2。结合引理4.2 与（38），可得关于 $I(X;\ A_1,\ \cdots,\ A_n)$ 的如下界：

$$I(X;A_1,\cdots,A_n)\leq n\cdot\mathcal{O}\left(\frac{\epsilon^4k^3}{n^3m\alpha}\right)\leq\mathcal{O}\left(\frac{\epsilon^4k^3}{n^2m\alpha}\right). \tag{40}$$

当 k为 $o\left(\frac{n^{2/3}m^{1/3}\alpha^{1/3}}{\epsilon^{4/3}}\right)$时，如（29）中所确立的，该界表明 $I(X;\ A_1,\cdots,A_n)=o(1)$。因此我们可以应用事实 A.5 并得出结论：没有任何算法推断 X 的值的概率能够超过 0.51。具体而言，$\mathcal{A}'$ 能够恢复 X 在任意样本集 $\mathcal{S}$上（无论有效与否）的概率小于 0.51。将该界与如下事实相结合：$(p,\mathcal{S})$ 以至少 0.99 的概率有效（由引理 4.1），我们有

$\mathcal{A}$ 在有效输入上恢复 X 的概率的如下界：

$$\mathbf{Pr}[\mathcal{A}'(\mathcal{S},\hat{p},\alpha',\epsilon')=X \mid (p,\mathcal{S}) \text{ valid}] = \frac{\mathbf{Pr}[\mathcal{A}'(\mathcal{S},\hat{p},\alpha',\epsilon')=X,\ (p,\mathcal{S}) \text{ valid}]}{\mathbf{Pr}[(p,\mathcal{S}) \text{ valid}]} \tag{41}$$

$$\leq \frac{\mathbf{Pr}[\mathcal{A}'(\mathcal{S},\hat{p},\alpha',\epsilon')=X]}{\mathbf{Pr}[(p,\mathcal{S}) \text{ valid}]} \tag{42}$$

$$< \frac{0.51}{0.95} \tag{43}$$

这与 (36) 相矛盾；该结论在假设增强独立性检验器在 $\mathcal{S}$ 上成功的前提下，证明了 $\mathcal{A}'$ 能够在有效输入上以至少 X 0.925 的概率恢复。因此，不存在一个接收 $k = o\left(\frac{n^{2/3}m^{1/3}\alpha^{1/3}}{\epsilon^{4/3}}\right)$ 个样本并在 p 上成功的增强独立性检验器。 □

5 多变量增强独立性检验的下界

本节通过证明 d 维独立性检验至少与二维独立性检验一样困难，给出高维增强独立性检验的下界。这些下界是紧的，与定理 6 中给出的上界相匹配。

定理 8。 令 p为乘积域 $\prod_{i=1}^{d}[n_i]$上的一个分布，且$\hat{}p$ 为 p 在同一乘积域上的一个预测。不失一般性，假设 $n_1 \geq n_2 \geq \ldots \geq n_d \geq 2$。令 $N := \prod_{i=1}^{d} n_i$ 表示总域大小。固定参数 $\alpha \in [0,1]$ 和 $\epsilon \in (0,\ 1/3)$。任何 (α、 ϵ、 δ)增强独立性检验器都必须使用$\Omega\left(\max_j\left\{\frac{\sqrt{N}}{\epsilon^2}、\ \frac{n_j^{1/3}N^{1/3}\alpha^{1/3}}{\epsilon^{4/3}}\right\}\right)$个样本才能成功。

证明。 我们将二维独立性检验的困难实例（如第 4 节所述）归约为d维独立性检验。假设存在一个用于 d维增强独立性检验的算法 $\mathcal{A}_d$，用于检验定义在 $[n_1]\times\ldots[n_d]$上的分布。考虑以下用于二维增强独立性检验的算法：该算法可访问分布 $\tilde{p}$ 在 $[n_1]\times\left[\prod_{i=2}^{d} n_i\right]$ 上的样本，并以ˆ为输入，给定预测 $\tilde{p}$、接近参数 ϵ 以及预测误差的估计值 α：

1. 从 $\tilde{p}$ 中抽取 k 个样本。

2. 将每个样本重塑为 $[n_1]\times[n_2]\times\ldots\times[n_d]$ 上的样本：保持第一个坐标不变，并将第二个坐标映射为 $[n_2]\times\ldots\times[n_d]$ 中的一个元组。此映射将第二维展开为第 2 到 d 维，其中最后一维变化最快。

ˆ3. 通过相同的映射将 $\tilde{p}$ 重塑为 $\hat{}p$ 。4. 返回在重塑后的样本、重塑后的预测、 ϵ 和 α上运行 $\mathcal{A}_d$ 的结果。

我们通过证明该算法在二维增强独立性检验的困难实例上能够成功，来完成归约。设 $\tilde{p}$ 是定义在 $[n_1]\times\left[\prod_{i=2}^{d} n_i\right]$ 上、接近参数为 ϵ的二维独立性检验困难实例。若 $\alpha \leq \frac{\sqrt{\prod_{i=2}^{d} n_i}}{\sqrt{n_1}\epsilon^2}$，则将 $\tilde{p}$ 视为二维下界第一种情形的困难实例，具体见第4.1节。否则，将 $\tilde{p}$ 视为二维下界第二种情形的困难实例，具体见第4.2节。设 p 为重塑后样本的分布。我们证明，若 d维检验器在 p上成功，则二维检验器在 $\tilde{p}$上也能成功。

首先请注意，所描述的重塑是分布的定义域元素之间的一种映射，不会影响这些元素的概率。由于全变差距离在

这种重塑下保持不变，因此重塑后的预测误差（即其与重塑后分布之间的距离）与原始预测的误差相同。

接下来，考虑 ACCEPT 二维下界每种情形的实例。对于该下界的第一种情形（第4.1节），回顾一下，ACCEPT 实例是均匀分布。经过重塑后，p 仍然是均匀分布，因此也是独立的。对于第二种情形（第4.2节），ACCEPT 实例是两个独立分布的乘积，其中第一个边缘分布是随机的，第二个边缘分布是均匀分布。重塑这个均匀的第二个边缘分布后，会得到定义在$[n_2] \times \ldots \times [n_d]$上的均匀分布，因此也是乘积分布。由于第一个边缘分布与原来的第二个边缘分布相互独立，因此它必定也与重塑后的边缘分布相互独立。因此，重塑后的分布是乘积分布。

对于 REJECT 实例，在下界的两种情形中，我们知道 $\tilde{p}$ 与任意乘积分布的距离至少为 ϵ，并希望证明 p 与任意乘积分布的距离也至少为 ϵ。我们将证明这一蕴含关系的逆否命题：如果存在定义在 $[n_1] \times \ldots \times [n_d]$ 上的乘积分布 q，使得 $\mathrm{d}_{\mathrm{TV}}(p, q) \leq \epsilon$，那么必定存在定义在 $[n_1] \times \left[\prod_{i=2}^{d} n_i\right]$ 上的乘积分布 $\tilde{q}$，使得 $\mathrm{d}_{\mathrm{TV}}(\tilde{p}, \tilde{q}) \leq \epsilon$。假设这样的分布 q 存在。按照上文所述的映射，将 q 重塑为定义在 $[n_1] \times \left[\prod_{i=2}^{d} n_i\right]$ 上的分布 $\tilde{q}$。由于 q 是乘积分布，其各个坐标彼此独立。特别地，第一个坐标与其余坐标构成的元组相互独立，而 $\tilde{q}$的边缘分布分别对应于第一个坐标和其余坐标构成的元组的分布，因此它是乘积分布。此外，由于重塑保持全变差距离不变，$\mathrm{d}_{\mathrm{TV}}(\tilde{p}, \tilde{q}) = \mathrm{d}_{\mathrm{TV}}(p, q) \leq \epsilon$。因此，$\tilde{p}$ 与某个乘积分布的距离至多为 ϵ，逆否命题得证。

最终，我们得到：（i）重塑后，分布与预测之间的距离保持不变；（ii）当 $\tilde{p}$是乘积分布时，p 也是乘积分布；（iii）当 $\tilde{p}$ 与其支撑集上的所有乘积分布都相距至少ϵ时，p 也与其支撑集上的所有乘积分布相距甚远。由于 $\mathcal{A}_d$ 是增强型 d维独立性检验器，$\mathcal{A}_2$ 可以判别 p 是乘积分布，还是与某个乘积分布相距至少 ϵ。换言之，$\mathcal{A}_d$ 对重塑后的分布 p 作出正确判定，就能使 $\mathcal{A}_2$ 对困难实例 $\tilde{p}$作出正确判定。因此，$\mathcal{A}_2$ 是一个能在 $\tilde{p}$上成功的增强型二维独立性测试器。由此，我们可以用增强型 $\tilde{p}$独立性检验的样本复杂度，给增强型 d维分布独立性检验的样本复杂度建立下界。代入第 4 节中得到的界可知，对于 $[n_1] \times \ldots [n_d]$ 上的增强独立性检验，若邻近参数为 ϵ 且预测误差为 α，则任何算法都至少需要 $\Omega\left(\max_j \left\{\frac{\sqrt{N}}{\epsilon^2}, \frac{n_j^{1/3} N^{1/3} \alpha^{1/3}}{\epsilon^{4/3}}\right\}\right)$ 个样本。

□

6 致谢

R.S. 感谢 Ken Kennedy 研究所研究集群基金、由能源 HPC 会议和莱斯大学计算机科学系资助的 Ken Kennedy 研究所计算科学与工程招聘奖学金，以及 Ken Kennedy 研究所 2025/26 年 Andrew Ladd 纪念计算机科学卓越研究生奖学金提供的部分支持。

参考文献

[AAK+07] Noga Alon、Alexandr Andoni、Tali Kaufman、Kevin Matulef、Ronitt Rubinfeld 和 Ning Xie。检验 k-路独立性和近似 k-路独立性。载于第三十九届 *ACM*计算理论研讨会论文集，第 496–505 页，2007。

[ABCR25] Maryam Aliakbarpour、Arnav Burudgunte、Cl´ement Cannone 和 Ronitt Rubinfeld。通过利用未经验证的辅助数据改进隐私分布检验，2025年。

[ABR16] Maryam A liakbarpour、Eric Blais 和 Ronitt Rubinfeld。junta 分布的学习与测试 s.

华录用会议论文集，2016年6月23日至26日，第29届学习理论会议，页码19–46，2016年。[ADJ+12] Jayadev Acharya、Hirakendu Das、Ashkan Jafarpour、Alon Orlitsky、Shengjun Pan 和 Ananda Theertha Suresh。竞争分类与接近性检验。载于 Shie Mannor、Nathan Srebro 和 Robert C. Williamson 编，《COLT 2012——第25届学习理论年度会议》，2012年6月25日至27日，苏格兰爱丁堡，第23卷《JMLR 论文集》，页码22.1–22.18。JMLR.org，2012年。[ADK15] Jayadev Acharya、Constantinos Daskalakis 和 Gautam Kamath。分布性质检验的最优方法。载于 C. Cortes、N. Lawrence、D. Lee、M. Sugiyama 和 R. Garnett 编，《神经信息处理系统进展》，第28卷。Curran Associates, Inc.，2015年。[ADKR19] Maryam Aliakbarpour、Ilias Diakonikolas、Daniel Kane 和 Ronitt Rubinfeld。通过样本置换对分布进行隐私检验。载于《神经信息处理系统进展32：2019年神经信息处理系统年度会议》，NeurIPS 2019，2019年12月8日至14日，加拿大不列颠哥伦比亚省温哥华，页码10877–10888，2019年。[AGP+19] Maryam Aliakbarpour、Themis Gouleakis、John Peebles、Ronitt Rubinfeld 和 Anak Yodpinya-nee。迈向一般偏序集上分布单调性的检验。载于《第32届学习理论会议论文集》，COLT，页码34–82，2019年。[AIRS24] Maryam Aliakbarpour、Piotr Indyk、Ronitt Rubinfeld 和 Sandeep Silwal。离散分布增强检验的最优算法。《神经信息处理系统进展》，37:11186–11227，2024年。[AKR19] Maryam Aliakbarpour、Ravi Kumar 和 Ronitt Rubinfeld。离散分布混合的检验。载于 Alina Beygelzimer 和 Daniel Hsu 编，《第32届学习理论会议论文集》，《机器学习研究论文集》第99卷，页码83–114，美国菲尼克斯，2019年6月25日至28日。PMLR。[alg26] 带预测的算法。https://algorithms-with-predictions.github.io/，2026年。一个全面、由社区推动的学习增强算法研究资源。[AS20] Maryam Aliakbarpour 和 Sandeep Silwal。用少量样本检验多个分布的性质。载于 Thomas Vidick 编，《第11届理论计算机科学创新会议（ITCS 2020）》，《莱布尼茨国际信息学论文集》（LIPIcs）第151卷，页码69:1–69:41，德国达格施图尔，2020年。Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik。[Bat01] Tugkan Batu。《分布性质检验》。博士论文，康奈尔大学，2001年。[BFF+01] Tugkan Batu、Lance Fortnow、Eldar Fischer、Ravi Kumar、Ronitt Rubinfeld 和 Patrick White。随机变量独立性与同一性的检验。载于《第42届计算机科学基础年度研讨会》，FOCS 2001，2001年10月14日至17日，美国内华达州拉斯维加斯，页码442–451，2001年。[BFR+00] Tugkan Batu、Lance Fortnow、Ronitt Rubinfeld、Warren D. Smith 和 Patrick White。分布接近性检验。载于《第41届计算机科学基础年度研讨会》，FOCS 2000，2000年11月12日至14日，美国加利福尼亚州雷东多海滩，页码259–269，2000年。[BFR+13] Tugkan Batu、Lance Fortnow、Ronitt Rubinfeld、Warren D. Smith 和 Patrick White。离散分布接近性检验。《ACM学报》，60(1):4:1–4:25，2013年。[BJ02] Francis R. Bach 和 Michael I. Jordan。树依赖分量分析。载于 Adnan Darwiche 和 Nir Friedman 编，《UAI ’02：第18届不确定性人工智能会议论文集》，加拿大阿尔伯塔省埃德蒙顿阿尔伯塔大学，2002年8月1日至4日，页码36–44。Morgan Kaufmann，2002年。

[BKR04] Tugkan Batu、Ravi Kumar 和 Ronitt Rubinfeld。用于检验单调分布和单峰分布的次线性算法。载于《第36届ACM计算理论研讨会年度会议论文集》，美国伊利诺伊州芝加哥，2004年6月13日至16日，页码381–390，2004年。[BPZL12] Gavin Brown、Adam Craig Pocock、Ming-Jie Zhao 和 Mikel Luj´an。条件似然最大化：信息论特征选择的统一框架。《机器学习研究期刊》，13:27–66，2012年。[Can20] Cl´ement L. Canonne。《分布检验综述：你的数据很大，但它是蓝色的吗？》研究生调查系列第9号。计算理论文库，2020年。[Can22] Cl´ement L. Canonne。分布检验的主题与技术：一个有偏但具代表性的样本。《计算机、通信与信息理论基金会趋势》，19(6):1032–1198，2022年。[CDGR18] Cl´ement L. Canonne、Ilias Diakonikolas、Themis Gouleakis 和 Ronitt Rubinfeld。离散分布形状限制的检验。《计算机理论系统》，62(1):4–62，2018年。[CDKS18] Cl´ement L Canonne、Ilias Diakonikolas、Daniel M Kane 和 Alistair Stewart。离散分布条件独立性的检验。载于《2018年信息论与应用研讨会（ITA）》，页码1–57。IEEE，2018年。[CDVV14] Siu-on Chan、Ilias Diakonikolas、Paul Valiant 和 Gregory Valiant。离散分布接近性检验的最优算法。载于 SODA，页码1193–1203，2014年。[CLZ22] Zhanrui Cai、Runze Li 和 Yaowu Zhang。一种用于因果发现的无分布假设条件独立性检验。《机器学习研究期刊》，23:85:1–85:41，2022年。[Cor51] Jerome Cornfield。基于临床数据估计比较率的方法：在肺癌、乳腺癌和宫颈癌中的应用。《美国国家癌症研究所杂志》，11(6):1269–1275，1951年6月。[CR14] Cl´ement L. Canonne 和 Ronitt Rubinfeld。聚合数据背后概率分布的检验。载于《自动机、语言与程序设计——第41届国际研讨会（ICALP）》，页码283–295，2014年。[CRS15] Cl´ement L Canonne、Dana Ron 和 Rocco A Servedio。使用条件样本检验概率分布。《SIAM计算期刊》，44(3):540–616，2015年。[DDS+13]Constantinos Daskalakis、Ilias Diakonikolas、Rocco A. Servedio、Gregory Valiant 和 Paul Valiant。k-众数分布检验：通过归约实现的最优算法。载于《第24届ACM-SIAM离散算法年度研讨会论文集》，SODA '13，页码1833–1852，美国宾夕法尼亚州费城，2013年。工业与应用数学学会。[DK16] Ilias Diakonikolas 和 Daniel M. Kane。离散分布性质检验的新方法。载于《IEEE第57届计算机科学基础年度研讨会》，FOCS 2016，2016年10月9日至11日，美国新泽西州新不伦瑞克凯悦酒店，页码685–694，2016年。[DKN15] Ilias Diakonikolas、Daniel M. Kane 和 Vladimir Nikishkin。结构化分布同一性的检验。载于 SODA，页码1841–1854，2015年。[EIN+21] Talya Eden、Piotr Indyk、Shyam Narayanan、Ronitt Rubinfeld、Sandeep Silwal 和 Tal Wagner。次线性时间内基于学习的支撑集估计。载于《第9届国际学习表征会议》，ICLR 2021，线上会议，奥地利，2021年5月3日至7日。OpenReview.net，2021年。[GFT+07] Arthur Gretton、Kenji Fukumizu、Choon Hui Teo、Le Song、Bernhard Sch¨olkopf 和 Alexander J. Smola。独立性的核统计检验。载于《神经信息处理系统进展》（NIPS），页码585–592。Curran Associates, Inc.，2007年。[GGR98] Oded Goldreich、Shafi Goldwasser 和 Dana Ron。性质检测及其与学习和近似的联系。《美国计算机学会会刊》，45(4):653–750，1998年。

[Gol17] Oded Goldreich。《性质检测导论》。剑桥大学出版社，2017 年。[GR11] Oded Goldreich 和 Dana Ron。《有界度图扩张性的检测》，第 68–75 页。施普林格出版社，柏林、海德堡，2011 年。

[HYF+22] 侯杰、叶秀芬、冯伟兴、张巧生、韩亚桐、刘宇松、李宇和魏玉芬。距离相关在基因共表达网络分析中的应用。BMC Bioinform., 23(1):81, 2022.[ILR12] Piotr Indyk、Reut Levi 和 Ronitt Rubinfeld。以次线性时间近似并检验 k-直方图分布。载于 PODS，第 15–22 页，2012.[LRR13] Reut Levi、Dana Ron 和 Ronitt Rubinfeld。分布集合性质的检验。Theory of Computing, 9(8):295–347, 2013.[MU05] Michael Mitzenmacher 和 Eli Upfal。《概率与计算：随机算法与概率分析》。剑桥大学出版社，纽约，美国，2005.[OS18] Krzysztof Onak 和 Xiaorui Sun。揭示概率的样本。载于人工智能与统计国际会议，机器学习研究论文集第 84 卷，第 2018–2026 页。PMLR，2018.[Pan08] Liam Paninski。针对极其稀疏抽样的离散数据进行均匀性检验的基于巧合的检验。IEEE Trans. Inf. Theory, 54(10):4750–4755, 2008.[Pea00] 卡尔·皮尔逊。X. 关于如下判据：在相关变量系统的情形下，给定系统相对于其可能值的偏差，是否可以合理地认为它是由随机抽样产生的。《伦敦、爱丁堡和都柏林哲学杂志与科学杂志》，50(302):157–175, 1900.[RS96] Ronitt Rubinfeld 和 Madhu Sudan。多项式的稳健刻画及其在程序检测中的应用。SIAM J. Comput., 25(2):252–271, 1996.[Rub12] Ronitt Rubinfeld。驯服大型概率分布。XRDS, 19(1):24–28, 2012.[RX10] Ronitt Rubinfeld 和 Ning Xie。乘积空间上非均匀 k-元独立分布的检验。载于自动机、语言与编程国际学术研讨会，第 565–581 页。施普林格，2010.[Sas97] Peter D Sasieni。从基因型到基因：将样本量扩大一倍。生物统计学，第 1253–1261 页，1997.[SRB07] G´abor J. Sz´ekely、Maria L. Rizzo 和 Nail K. Bakirov。通过距离相关测量和检验依赖性。《统计学年刊》，35(6):2769–2794, 2007.[SSG+07] Le Song、Alex Smola、Arthur Gretton、Karsten M Borgwardt 和 Justin Bedo。通过依赖性估计进行监督特征选择。载于第 24 届机器学习国际会议论文集，第 823–830 页，2007.[Val11] Paul Valiant。分布对称性质的检验。SIAM J. Comput., 40(6):1927–1968, 2011.[VV17] Gregory Valiant 和 Paul Valiant。自动不等式证明器与实例最优恒等性检验。SICOMP, 46(1):429–455, 2017.[ZPJS11] Kun Zhang、Jonas Peters、Dominik Janzing 和 Bernhard Sch¨olkopf。基于核的条件独立性检验及其在因果发现中的应用。载于 F´abio Gagliardi Cozman 和 Avi Pfeffer 编，UAI 2011：第二十七届人工智能不确定性会议论文集，西班牙巴塞罗那，2011 年 7 月 14–17 日，第 804–813 页。AUAI出版社，2011。

附录

A 背景结果

下面列出我们在证明过程中使用的一些有用事实、定义、引理和定理。

A.1 分布检验结果

引理 *A.1* （[*BFF*+01，*LRR*13]）。设 p 和 q 是定义在 $[n]\times[m]$ 上的分布，且 $\epsilon\in[0,1]$。设 p_1 和 p_2 表示 p 的边缘分布。若 q 是乘积分布，且 $d_{\mathrm{TV}}(p,q)\leq\epsilon/3$，则 $d_{\mathrm{TV}}(p,p_1\times p_2)\leq\epsilon$。

事实 A.2 ([GR11，AIRS24])。存在一种算法，它在获得未知分布的样本访问权限后 p 在 $[n]$上，抽取 $\mathcal{O}(\sqrt{n}\log(1/\delta))$ 个样本，来自 p 并输出一个估计值 L，用于估计 ℓ_2^2范数，满足以下条件的概率至少为 $1-\delta$，

$$\frac{\|p\|_2^2}{2}\leq L\leq\frac{3\|p\|_2^2}{2}. \tag{44}$$

事实 A.3 ([CDVV14，ADKR19])。存在一种算法，

$$\textsc{ClosenessTest}(p,q,b,\epsilon,\delta), \tag{45}$$

该算法可访问分布 p 和 q 在 $[n]$上的样本，并获得参数 b、ϵ和 δ。如果 $b\geq\min(\|p\|_2^2,\|q\|_2^2)$，该算法可以以至少 $1-\delta$的概率判别 $p=q$或 $\mathrm{d_{TV}}(p,q)>\epsilon$，使用 $\mathcal{O}(n\sqrt{b}/\epsilon^2\log(1/\delta))$ 个样本。

事实A.4 ([AIRS24])。设 $\alpha\in[0,1]$ 和 $\nu\leq1$ 为两个给定参数。设 p 和 $\hat{}p$ 为定义在$[n]$ 上的分布，满足 $\mathrm{d_{TV}}(p,\hat{}p)\leq\alpha$。假设我们从 泊松分布$(s)$ 中抽取一组 p 样本，并构造一个展平分布， $p^{(F)}$，其桶按如下方式选取：

$$b_i:=\left\lfloor\frac{\hat{p}(i)}{\nu}\right\rfloor+N_i+1. \tag{46}$$

其中， N_i 是样本集中元素 i 的频数。取期望时，这种扁平化会使该分布的定义域大小增加$\frac{1}{\nu}+s$。此外，$p^{(F)}$ 的期望 ℓ_2^2-范数有如下界：

$$\mathbf{E}\left[\left\|p^{(F)}\right\|_2^2\right]\leq\frac{2\alpha}{s}+4\nu. \tag{47}$$

A.2 信息论结果

事实 A.5 ([DK16]引理 3.2 的逆否命题。设 X 是一个均匀随机比特，且设 A 是一个随机变量。如果 $I(X;A)<2\cdot10^{-4}$，那么对所有函数 f，$\mathbf{Pr}[f(A)=X]<0.51$。

A.3 抽样结果

事实 A.6。设 q 是 $[m]$ 上的任意分布。对于任意 $\epsilon,\delta\in(0,1)$，给定 $t\geq\frac{m+\log(1/\delta)}{\epsilon^2}$个独立样本$X_1$、……、$X_t\sim q$，经验分布 $\tilde{q}_i:=\frac{1}{t}\sum_{j=1}^{t}\mathbb{I}\{X_j=i\}$ 以至少 $1-\delta$ 的概率满足 $\|q-\tilde{q}\|_{\mathsf{TV}}\leq\epsilon$。

证明。每个元素 $t\cdot\tilde{q}_i$ 都服从参数为 t 和 q_i的二项分布。因此，向量 $(t\tilde{q}_1,\cdots,t\tilde{q}_m)$ 服从参数为 t和 $(q_1,\cdots,q_m)$ 的多项分布。Bretagnolle-Huber-Carol

不等式可用于对全变差距离进行如下界定：

$$\begin{aligned}
&\mathbf{Pr}\left[\sum_{i=1}^{m}|t\cdot\tilde{q}_i - t\cdot q_i| \geq 2t\epsilon\right] \leq 2^m e^{-2t\epsilon^2} \\
&\mathbf{Pr}\left[\sum_{i=1}^{m}|\tilde{q}_i - q_i| \geq 2\epsilon\right] \leq \delta \qquad \text{if } t \geq \frac{m+\log 1/\delta}{\epsilon^2} \\
\Rightarrow \quad &\mathbf{Pr}[\|\tilde{q}-q\|_{\mathsf{TV}} \leq \epsilon] \geq 1-\delta \qquad \text{if } t \geq \frac{m+\log 1/\delta}{\epsilon^2}
\end{aligned} \tag{48}$$

□

推论 A.7. 令 q 为 $[m]$ 上的任意分布。对于任意 $\delta \in (0,1)$ 以及任意数量的样本 t，经验分布 $\tilde{q}$ 满足

$$\mathbf{Pr}\left[\|\tilde{q}-q\|_{\mathsf{TV}} \leq \sqrt{\frac{m+\log 1/\delta}{t}}\right] \geq 1-\delta \tag{49}$$

B 三维增强独立性检验

本节研究三维离散分布 p 的独立性检验问题，该分布支撑于 $[n_1]\times[n_2]\times[n_3]$，并给定同一定义域上的预测分布 $\hat{}p$。令 p_1、p_2、和 p_3 分别表示 p 的边缘分布。不失一般性，假设 $n_1 \geq n_2 \geq n_3$。

算法 4 的正确性证明可由二维情形的类似论证得出。接下来，我们分析该算法的样本复杂度。它必须从 p 中抽取样本，以扁平化这些分布、估计扁平化分布的平方范数，并在扁平化分布上运行标准接近性检验器。

为构造扁平化，抽取至多 $240(s_1+s_2+s_3) = \mathcal{O}(s_1+s_2+s_3)$ 个样本。

根据 事实 A.4 中描述的扁平化 F_ℓ 构造，定义域大小 $p^{(F_\ell)}$ ℓ 为 ：

$$\sum_{i_\ell=1}^{n_\ell}(\lfloor n_\ell\cdot\hat{p}_\ell(i_\ell)\rfloor + N_{i_\ell}+1) \leq \sum_{i_\ell=1}^{n_\ell}(n_\ell\cdot\hat{p}_\ell(i_\ell)+1) + \sum_{i_\ell=1}^{n_\ell}N_i + \sum_{i_\ell=1}^{n_\ell}1 = 3n_\ell + s_\ell = \mathcal{O}(n_\ell). \tag{50}$$

因此， $p^{(F)}$ 的定义域大小为 $\mathcal{O}(n_1n_2n_3)$。根据事实 A.2，估计定义域大小为 n 的分布的 ℓ_2-范数需要 $\mathcal{O}(\sqrt{n})$ 个样本。因此，要估计扁平化后分布 $p_1^{(F_1)}, p_2^{(F_2)}, p_3^{(F_3)}$ 和 $p^{(F)}$ 的范数，我们必须分别抽取 $\mathcal{O}\left(\sqrt{n_1}\right)$，$\mathcal{O}\left(\sqrt{n_2}\right), \mathcal{O}\left(\sqrt{n_3}\right)$ 和 $\mathcal{O}\left(\sqrt{n_1n_2n_3}\right)$ 个样本。

应用事实 A.3，标准接近性检验器从 $p^{(F)}$和 $p^{(F_1)}{}_1 \times p^{(F_2)}{}_2 \times p_3^{(F_3)}$ 各抽取 $\mathcal{O}(n_1n_2n_3\sqrt{\tau_1\tau_2\tau_3}/\epsilon^2)$ 个样本。注意，我们可以通过从 $p^{(F)}$ 抽取三个独立样本，并分别返回一个点的第一个坐标、另一个点的第二个坐标以及最后一个点的第三个坐标，来从 $p^{(F_1)}{}_1 \times p^{(F_2)}{}_2 \times p_3^{(F_3)}$ 抽取一个样本。因此，接近性检验器的总体样本复杂度为 $\mathcal{O}(n_1n_2n_3\sqrt{\tau_1\tau_2\tau_3}/\epsilon^2)$。

最终，将 τ_1、τ_2和 τ_3 替换为它们的实现（在第 15 行），考虑 $s_2 = n_2\alpha$ 和

算法 4： 三维分布的增强独立性检验器

输入： 对分布 p 的样本访问，该分布定义在 $[n_1]\times[n_2]\times[n_3]$ 上，边缘分布为 p_1，p_2，p_3；预测 $\hat{p}$ 定义在相同定义域上；预测误差的估计值 α；接近参数 ϵ；置信度 δ

增强型三维独立性测试器$(n_1, n_2, n_3, p_1, p_2, p_3, \hat{p}, \alpha, \epsilon)$**：**

$s_1 \leftarrow$ 最小值$\left(n_1\alpha, \frac{n_1^{2/3}n_2^{1/3}n_3^{1/3}\alpha^{1/3}}{\epsilon^{4/3}}\right)$，$s_2 \leftarrow n_2\alpha$，$s_3 \leftarrow n_3\alpha$

$\hat{s}_1 \leftarrow$泊松分布$(s_1), \hat{s}_2 \leftarrow$泊松分布$(s_2), \hat{s}_3 \leftarrow$泊松分布$(s_3)$

如果 $\hat{s}_1 > 240s_1$ 或 $\hat{s}_2 > 240s_2$ 或 $\hat{s}_3 > 240s_3$ **则**

  返回 REJECT

end

对于 $\ell \in \{1,2,3\}$，**执行**

  $S_\ell \leftarrow \hat{s}_\ell$ 个来自 p_ℓ 的样本

  对于 $i_\ell \in [n_\ell]$，**执行**

    $N_{i_\ell} \leftarrow$在 S_ℓ中的 i_ℓ 频数

    $b_{i_\ell}^{(\ell)} \leftarrow \lfloor n_\ell \cdot \hat{p}_\ell(i_\ell)\rfloor + N_{i_\ell} + 1$ // 参见公式 3

  end

  $F_\ell \leftarrow$ 给定桶的扁平化 $[n_\ell]$ 对于所有 $b_{i_\ell}^{(\ell)}$ // 参见第 2.1

  $L_\ell \leftarrow$ 估计值-ℓ_2^2 $\left(p_\ell^{(F_\ell)}, \sum_{i_\ell=1}^{n_\ell} b_{i_\ell}^{(\ell)}, 1/180\right)$ // 事实 A.2 的算法

  $\tau_\ell \leftarrow \frac{2\alpha}{s_\ell} + \frac{4}{n_\ell}$

end

如果 $\exists \ell$ 使得 $L_\ell > 180\tau_\ell$ **则**

  返回 INACCURATE INFORMATION

end

$F \leftarrow [n_1]\times[n_2]\times[n_3]$ 给定桶的扁平化 $b_{i_\ell}^{(\ell)}$ 对于所有 i_ℓ 以及对于 $\ell \in \{1,2,3\}$

$L \leftarrow$ 估计值-ℓ_2^2 $\left(p^{(F)}, \prod_{\ell=1}^{3}\left(\sum_{i_\ell=1}^{n_\ell} b_{i_\ell}^{(\ell)}\right), 1/180\right)$

如果 $L > (6\cdot 180^3)\tau_1\tau_2\tau_3$ **则**

  返回 REJECT

end

返回 接近度测试器 $\left(p^{(F)}, p^{(F_1)}{}_1 \times p^{(F_2)}{}_2 \times p^{(F_3)\,3}, (6\cdot 180^3)\cdot\tau_1\tau_2\tau_3, \epsilon, 1/120\right)$

结束

$s_3 = n_3\alpha$，并计入每个分量，算法 4 的样本复杂度为：

$$\begin{aligned}&\mathcal{O}\left(s_1+s_2+s_3+\sqrt{n_1}+\sqrt{n_2}+\sqrt{n_3}+\sqrt{n_1n_2n_3}+\frac{n_1n_2n_3}{\epsilon^2}\sqrt{\left(\frac{2\alpha}{s_1}+\frac{4}{n_1}\right)\left(\frac{6}{n_2}\right)\left(\frac{6}{n_3}\right)}\right)\\ =&\mathcal{O}\left(s_1+(n_2+n_3)\alpha+\sqrt{n_1n_2n_3}+\frac{n_1\sqrt{n_2n_3}}{\epsilon^2}\sqrt{\left(\frac{\alpha}{s_1}+\frac{1}{n_1}\right)}\right).\end{aligned} \tag{51}$$

回顾 $s_1 = \min\left(n_1\alpha, \frac{n_1^{2/3}n_2^{1/3}n_3^{1/3}\alpha^{1/3}}{\epsilon^{4/3}}\right)$。将上述每种可能情形代入后，算法 4的总体样本复杂度为：

$$\mathcal{O}\left(\max\left\{\frac{\sqrt{n_1n_2n_3}}{\epsilon^2}, \frac{n_1^{2/3}n_2^{1/3}n_3^{1/3}\alpha^{1/3}}{\epsilon^{4/3}}\right\}\right) \tag{52}$$

C 第4节省略的证明

C.1 引理4.1的证明

引理 *4.1*。设 $n \geq m$ 为充分大的整数，使得 $m \geq$*log(n)*。设 k 为一个整数，并令$\epsilon \in$*(0,* $1/192$*)* $\alpha \in$*(0, 1)*。假设 $\epsilon' = 192\epsilon$与 $\alpha' = 2\alpha/3$。给定 ϵ'、α'，令 X、 q与 p按式 *(30)* 至 *(34)* 抽取。令 $\hat{}p$ 为 $[n] \times [m]$上的均匀分布。令 $\mathcal{S}$ 为一个多重集，使得对所有 *(i, j)* $\in [n] \times [m]$，*(i, j)* $\in \mathcal{S}$的副本数为泊松分布 *(*$k \cdot q$*(i, j))*。则以至少 *0.95* 的概率，以下事件全部成立：

1. $|\mathcal{S}| \geq k/100$。*2.* 如果 $X = 0$, p 是一个乘积分布，且与 $\hat{}p$ 的距离至多为 α。*3.* 如果 $X = 1$, p 与所有乘积分布的距离至少为 ϵ。

证明。令 $\mathcal{H}$ 为"重"行的集合——即 $[n]$ 的子集，满足 $c_i = \frac{1}{k}$。为了证明 $\mathcal{S}$ 以高概率成立，我们将证明以下事件以高概率发生，并说明在这些事件成立的条件下， $\mathcal{S}$ 有效。这些事件是：

1. For all $i \in \mathcal{H}^{\mathsf{c}}$, $\left|\sum_{j=1}^{m} \frac{\epsilon'_{ij}}{m}\right| \leq \frac{\epsilon}{2}$.
2. For all $j \in [m]$, $\left|\sum_{i \in \mathcal{H}^{\mathsf{c}}} \frac{\epsilon'_{ij}}{n}\right| \leq \frac{\epsilon}{2}$.
3. $|\mathcal{H}| \leq \frac{3}{2}\alpha' k$.

利用标准集中不等式可以证明，这些事件中的每一个都以高概率发生。首先，对所有 i应用霍夫丁界和并集界，我们发现事件 1 发生的概率至少为 0.99：

$$\mathbf{Pr}\left[\exists i \in \mathcal{H}^{\mathsf{c}} \ : \left|\sum_{j=1}^{m} \frac{\epsilon'_{ij}}{m}\right| \geq \epsilon' \sqrt{\frac{2}{m}\log(50n)}\right] \leq 0.01\,. \tag{53}$$

类似地，对所有 j应用霍夫丁界和并集界，我们发现事件 2 发生的概率至少为 0.99：

$$\mathbf{Pr}\left[\exists j \in [m] \ : \left|\sum_{i \in \mathcal{H}^{\mathsf{c}}} \frac{\epsilon'_{ij}}{n}\right| \geq \epsilon' \sqrt{\frac{2}{n}\log(50m)}\right] \leq 0.01\,. \tag{54}$$

最后，利用切尔诺夫界以及 $|\mathcal{H}|$的期望为 $\alpha' k$这一事实，我们可以将 $|\mathcal{H}|$ 较大的概率界定为：

$$\mathbf{Pr}\left[|\mathcal{H}| \geq \left(1 + \sqrt{\frac{3}{\alpha' k}\log(100)}\right)\alpha' k\right] \leq 0.01\,. \tag{55}$$

注意，（55）直接表明事件3以高概率发生。特别地，由于我们正在建立下界的第二项，我们有 $\alpha \geq \frac{\sqrt{m}}{\sqrt{n\epsilon^2}}$。由于在这种情况下，下界的第一项必然适用于样本数量，我们还有~~$k \geq$~~ $\frac{\sqrt{nm}}{\epsilon^2}$。将这两个不等式结合起来可得$\alpha k > \frac{m}{\epsilon^4}$。对于足够大的 m和足够小的 ϵ，该界意味着 $\frac{\alpha' k}{2} > 3\log(100)$。因此，~~(55)~~意味着 $|\mathcal{H}| \geq \frac{3}{2}\alpha' k$ 以至少0.01的概率成立。

X = 0 意味着 p 是乘积，且 α与 ˆp 相差不超过。首先，注意如果 $X = 0$，那么 p 按构造即为乘积分布。此外，在事件 3条件下，当 $X = 0$ 且 $\alpha' k \ll 1$时， $p = \hat{}p$。当 $\alpha' k > 1$时，

我们将证明，以高概率有 $\mathrm{d_{TV}}(p、\hat{}p) \leq \alpha$。考虑 $\mathrm{d_{TV}}(p、\hat{}p)$:

$$\mathrm{d_{TV}}\left(p,\hat{p}\right) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left|\frac{\mu(i,j)c_i}{s_i\cdot C} - \frac{1}{nm}\right| \tag{56}$$

$$= \frac{1}{2}\left(\sum_{i\in\mathcal{H}}\sum_{j=1}^{m}\left|\frac{1}{kmC} - \frac{1}{nm}\right| + \sum_{i\in\mathcal{H}^c}\sum_{j=1}^{m}\left|\frac{1}{nmC} - \frac{1}{nm}\right|\right). \tag{57}$$

$$= \frac{1}{2}\left(|\mathcal{H}|\left|\frac{1}{kC} - \frac{1}{n}\right| + |\mathcal{H}^c|\left|\frac{1}{nC} - \frac{1}{n}\right|\right) \tag{58}$$

$$= \frac{1}{2}\left(|\mathcal{H}|\left|\frac{1}{kC} - \frac{1}{nC} + \frac{1}{nC} - \frac{1}{n}\right| + (n-|\mathcal{H}|)\left|\frac{1}{nC} - \frac{1}{n}\right|\right). \tag{59}$$

应用三角不等式，我们可以将（59）上界估计为

$$\mathrm{d_{TV}}\left(p,\hat{p}\right) \leq \frac{1}{2}\left(|\mathcal{H}|\left|\frac{1}{kC} - \frac{1}{nC}\right| + |\mathcal{H}|\left|\frac{1}{nC} - \frac{1}{n}\right| + (n-|\mathcal{H}|)\left|\frac{1}{nC} - \frac{1}{n}\right|\right) \tag{60}$$

$$= \frac{1}{2}\left(|\mathcal{H}|\left|\frac{1}{kC} - \frac{1}{nC}\right| + n\cdot\left|\frac{1}{nC} - \frac{1}{n}\right|\right) \tag{61}$$

$$= \frac{1}{2}\left(\frac{|\mathcal{H}|}{C}\left|\frac{1}{k} - \frac{1}{n}\right| + \left|\frac{1}{C} - 1\right|\right). \tag{62}$$

由于 $k \leq n/2$，我们知道 $|1/k - 1/n| = 1/k - 1/n \leq 1/k$。因此，我们有

$$\mathrm{d_{TV}}\left(p,\hat{p}\right) \leq \frac{1}{2}\left(\frac{|\mathcal{H}|}{Ck} + \left|\frac{1}{C} - 1\right|\right). \tag{63}$$

为证明（63）的上界为 α，我们现在证明 $\frac{|\mathcal{H}|}{Ck} \leq \mathcal{O}(\alpha')$ 和 $\left|\frac{1}{C} - 1\right| \leq \mathcal{O}(\alpha')$，然后相应地选择$\alpha'$。首先，回顾事件 3，我们可以将 $|\mathcal{H}|$ 上界估计为 $3\alpha' k/2$。直接计算可得 $C \geq 1$:

$$C = \frac{|\mathcal{H}|}{k} + \frac{n-|\mathcal{H}|}{n} = 1 + \frac{|\mathcal{H}|}{k}\left(1 - \frac{k}{n}\right) \geq 1 + \frac{|\mathcal{H}|}{2k} \geq 1, \tag{64}$$

应用（64）以及我们对 $|\mathcal{H}|$的上界估计，可得

$$\frac{|\mathcal{H}|}{Ck} \leq \frac{3\alpha' k}{2k} = \frac{3}{2}\alpha'. \tag{65}$$

此外，我们可以应用事件 3 将 C 的上界估计为

$$C = \frac{|\mathcal{H}|}{k} + \frac{n-|\mathcal{H}|}{n} = 1 + \frac{|\mathcal{H}|}{k}\left(1 - \frac{k}{n}\right) \leq 1 + \frac{|\mathcal{H}|}{k} \leq 1 + \frac{3}{2}\alpha'. \tag{66}$$

我们可以利用界 $1 \leq C \leq 1 + \frac{3}{2}\alpha'$将 $|1 - 1/C|$ 上界估计为:

$$0 \leq 1 - \frac{1}{C} \leq 1 - \frac{1}{1+\frac{3}{2}\alpha'} = \frac{\frac{3}{2}\alpha'}{1+\frac{3}{2}\alpha'} \leq \frac{3}{2}\alpha' \tag{67}$$

由（65）和（67），我们可以将（63）的上界写为

$$\mathrm{d_{TV}}\left(p,\hat{p}\right) \leq \frac{1}{2}\left(\frac{3}{2}\alpha' + \frac{3}{2}\alpha'\right) = \frac{3}{2}\alpha'. \tag{68}$$

选择 $\alpha' = \frac{2}{3}\alpha$ 可确保 $\mathrm{d_{TV}}(p,\ \hat{}p) \leq \alpha$。

X = 1 意味着 p 与乘积分布相距甚远。 $p(i,\ j) = \frac{\mu(i,\ j)\cdot c_i}{s_i \cdot C}$ 的边缘分布为:

$$p_1(i) = \frac{c_i}{C} \qquad p_2(j) = \sum_{i=1}^{n} \frac{\mu(i,j)\cdot c_i}{s_i C} \tag{69}$$

我们想要证明 $\mathrm{d_{TV}}(p,\ p_1 \times p_2) > \epsilon$。同样，设 $\mathcal{H}$ 为"重"行的集合，并在事件1、2和 3的条件下考虑。考察 p 与 $p_1 \times p_2$ 之间的距离。我们有:

$$\mathrm{d_{TV}}(p, p_1 \times p_2) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left|\frac{\mu(i,j)}{s_i}\cdot\frac{c_i}{C} - \frac{c_i}{C}\cdot\sum_{i'=1}^{n}\frac{\mu(i',j)}{s_{i'}}\cdot\frac{c_{i'}}{C}\right| \tag{70}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\frac{c_i}{C}\sum_{j=1}^{m}\left|\frac{\mu(i,j)}{s_i} - \sum_{i'=1}^{n}\frac{\mu(i',j)}{s_{i'}}\cdot\frac{c_{i'}}{C}\right| \tag{71}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\frac{c_i}{C}\sum_{j=1}^{m}\left|\frac{\mu(i,j)}{s_i} - \mu(i,j) + \mu(i,j) - \frac{1}{m} + \frac{1}{m} - \sum_{i'=1}^{n}\frac{\mu(i',j)}{s_{i'}}\cdot\frac{c_{i'}}{C}\right|. \tag{72}$$

应用反三角不等式，可得该量的下界:

$$\mathrm{d_{TV}}(p, p_1 \times p_2) \geq \frac{1}{2}\sum_{i=1}^{n}\frac{c_i}{C}\sum_{j=1}^{m}\left|\mu(i,j) - \frac{1}{m}\right| - \left|\frac{\mu(i,j)}{s_i} - \mu(i,j)\right| - \left|\frac{1}{m} - \sum_{i'=1}^{n}\frac{\mu(i',j)}{s_{i'}}\cdot\frac{c_{i'}}{C}\right|. \tag{73}$$

为了证明 $\mathrm{d_{TV}}(p,\ p_1 \times p_2) > \epsilon$，对于固定的 i 和 j，我们分别对（73）中内层求和的每一项进行界定，并应用事件 1、2 和 3的条件。注意，对第二项和第三项应用上界时，由于其系数为负，会得到 TV 距离的下界。第一项为

$$\left|\mu(i,j) - \frac{1}{m}\right| = \begin{cases}\left|\frac{1}{m} - \frac{1}{m}\right| & i \in \mathcal{H} \\ \left|\frac{1+\epsilon'_{ij}}{m} - \frac{1}{m}\right| & i \in \mathcal{H}^{\mathsf{c}}\end{cases} = \begin{cases}0 & i \in \mathcal{H} \\ \frac{\epsilon'}{m} & i \in \mathcal{H}^{\mathsf{c}}\end{cases}. \tag{74}$$

第二项为

$$\left|\frac{\mu(i,j)}{s_i} - \mu(i,j)\right| = \mu(i,j)\left|\frac{1}{s_i} - 1\right|. \tag{75}$$

当 $i \in \mathcal{H}$, $s_i = \sum_{j=1}^{m}\frac{1}{m} = 1$ 时。因此，我们可以根据 $i \in \mathcal{H}$ 和 $i \notin \mathcal{H}$ 的情形将这一项拆分为

$$\left|\frac{\mu(i,j)}{s_i} - \mu(i,j)\right| = \begin{cases}0 & i \in \mathcal{H} \\ \frac{1+\epsilon'_{ij}}{m}\left|\frac{1}{s_i} - 1\right| & i \in \mathcal{H}^{\mathsf{c}}\end{cases}. \tag{76}$$

利用事件 1 可知，对所有 $i \in \mathcal{H}^{\mathsf{c}}$，$\left|\sum_{j=1}^{m}\frac{\epsilon'_{ij}}{m}\right| \leq \epsilon/2$ 根据定义 $s_i = \sum_{j=1}^{m}\mu(i,\ j) = 1 + \sum_{j=1}^{m}\frac{\epsilon'_{ij}}{m}$，有 $s_i \in (1-\epsilon/2,\ 1+\epsilon/2)$，这意味着 $|1/s_i - 1| \leq \frac{\epsilon/2}{1-\epsilon/2} \leq \epsilon$ 并得到简化后的界:

$$\left|\frac{\mu(i,j)}{s_i} - \mu(i,j)\right| \leq \begin{cases}0 & i \in \mathcal{H} \\ \frac{1+\epsilon'_{ij}}{m}\cdot\epsilon & i \in \mathcal{H}^{\mathsf{c}}\end{cases}. \tag{77}$$

第三项为

$$\left|\frac{1}{m}-\sum_{i'=1}^{n}\frac{\mu(i',j)}{s_{i'}}\cdot\frac{c_{i'}}{C}\right| = \frac{1}{m}\left|1-\frac{1}{C}\left(\sum_{i'\in\mathcal{H}}\frac{1}{k}+\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{s_{i'}}\cdot\frac{1+\epsilon'_{i'j}}{n}\right)\right| \tag{78}$$

$$= \frac{1}{m}\left|1-\frac{1}{C}\left(\sum_{i'\in\mathcal{H}}\frac{1}{k}+\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{n}-\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{n}+\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{s_{i'}}\cdot\frac{1+\epsilon'_{i'j}}{n}\right)\right| \tag{79}$$

$$= \frac{1}{m}\left|1-\frac{1}{C}\left(C-\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{n}+\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{s_{i'}}\cdot\frac{1+\epsilon'_{i'j}}{n}\right)\right| \tag{80}$$

$$= \frac{1}{mC}\left|\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{n}-\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{s_{i'}}\cdot\frac{1+\epsilon'_{i'j}}{n}\right| \tag{81}$$

$$= \frac{1}{mC}\left|\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{n}\left(1-\frac{1}{s_{i'}}\right)-\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{\epsilon'_{i'j}}{ns_{i'}}\right| \tag{82}$$

$$= \frac{1}{mC}\left|\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{n}\left(1-\frac{1}{s_{i'}}\right)+\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\left(1-\frac{1}{s_{i'}}\right)\frac{\epsilon'_{i'j}}{n}-\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{\epsilon'_{i'j}}{n}\right| \tag{83}$$

$$\leq \frac{1}{mC}\left(\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{n}\left|1-\frac{1}{s_{i'}}\right|+\left|\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\left(1-\frac{1}{s_{i'}}\right)\frac{\epsilon'_{i'j}}{n}\right|+\left|\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{\epsilon'_{i'j}}{n}\right|\right), \tag{84}$$

其中，最后一步应用了三角不等式。同样，对于所有 $i'\in\mathcal{H}^{\mathsf{c}}$，$|1-1/s_{i'}|\leq\frac{\epsilon/2}{1-\epsilon/2}\leq\epsilon$，我们可以应用柯西-施瓦茨不等式来界定式（84）中的中间项：

$$\left|\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\left(1-\frac{1}{s_{i'}}\right)\frac{\epsilon'_{i'j}}{n}\right| \leq \sqrt{\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\left(1-\frac{1}{s_{i'}}\right)^2}\cdot\sqrt{\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\left(\frac{\epsilon'_{i'j}}{n}\right)^2} \tag{85}$$

$$\leq \sqrt{|\mathcal{H}^{\mathsf{c}}|\,\epsilon^2}\cdot\sqrt{|\mathcal{H}^{\mathsf{c}}|\left(\frac{\epsilon'}{n}\right)^2} \tag{86}$$

$$= \frac{|\mathcal{H}^{\mathsf{c}}|}{n}\epsilon\epsilon'. \tag{87}$$

应用对所有 $j\in[m]$，$\left|\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{\epsilon'_{i'j}}{n}\right|\leq\epsilon/2$（即事件 2）的条件，并结合（87）和$|1-1/s_{i'}|\leq\epsilon$，我们可以对第三项作如下界定

$$\left|\frac{1}{m}-\sum_{i'=1}^{n}\frac{\mu(i',j)}{s_{i'}}\cdot\frac{c_{i'}}{C}\right| \leq \frac{1}{mC}\left(\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{1}{n}\left|1-\frac{1}{s_{i'}}\right|+\left|\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\left(1-\frac{1}{s_{i'}}\right)\frac{\epsilon'_{i'j}}{n}\right|+\left|\sum_{i'\in\mathcal{H}^{\mathsf{c}}}\frac{\epsilon'_{i'j}}{n}\right|\right) \tag{88}$$

$$\leq \frac{1}{mC}\cdot\left(\frac{|\mathcal{H}^{\mathsf{c}}|}{n}\epsilon+\frac{|\mathcal{H}^{\mathsf{c}}|}{n}\epsilon\epsilon'+\frac{\epsilon}{2}\right) \tag{89}$$

$$= \frac{|\mathcal{H}^{\mathsf{c}}|}{nmC}(\epsilon+\epsilon\epsilon')+\frac{\epsilon}{2mC}. \tag{90}$$

将（74）、（77）和（90）代回（73），得到以下界：

$$\mathrm{d_{TV}}\left(p, p_1 \times p_2\right) \geq \frac{1}{2}\sum_{i=1}^{n}\frac{c_i}{C}\sum_{j=1}^{m}\left|\mu(i,j)-\frac{1}{m}\right| - \left|\frac{\mu(i,j)}{s_i}-\mu(i,j)\right| - \left|\frac{1}{m}-\sum_{i'=1}^{n}\frac{\mu(i',j)}{s_{i'}}\cdot\frac{c_{i'}}{C}\right| \tag{91}$$

$$\geq \frac{1}{2}\left(\sum_{i\in\mathcal{H}^{\mathsf{c}}}\frac{c_i}{C}\sum_{j=1}^{m}\frac{\epsilon'}{m} - \sum_{i\in\mathcal{H}^{\mathsf{c}}}\frac{c_i}{C}\sum_{j=1}^{m}\frac{1+\epsilon'_{ij}}{m}\cdot\epsilon - \sum_{i=1}^{n}\frac{c_i}{C}\sum_{j=1}^{m}\frac{|\mathcal{H}^{\mathsf{c}}|}{nmC}\left(\epsilon+\epsilon\epsilon'\right) + \frac{\epsilon}{2mC}\right) \tag{92}$$

$$= \frac{1}{2}\left(\frac{|\mathcal{H}^{\mathsf{c}}|\,\epsilon'}{nC} - \frac{|\mathcal{H}^{\mathsf{c}}|}{nC}\epsilon - \left(\sum_{i\in\mathcal{H}^{\mathsf{c}}}\frac{1}{nC}\sum_{j=1}^{m}\frac{\epsilon'_{ij}}{m}\right)\epsilon - \frac{|\mathcal{H}^{\mathsf{c}}|}{nC}\left(\epsilon+\epsilon\epsilon'\right) - \frac{\epsilon}{2C}\right), \tag{93}$$

再次，通过对事件 1进行条件化，可得 $\left|\sum_{j=1}^{m}\frac{\epsilon'_{ij}}{m}\right| \leq \epsilon/2$ 对所有 $i \in \mathcal{H}^{\mathsf{c}}$均成立，从而

$$\mathrm{d_{TV}}\left(p, p_1 \times p_2\right) \geq \frac{1}{2C}\left(\epsilon'\left(\frac{|\mathcal{H}^{\mathsf{c}}|}{n}\left(1-\epsilon\right)\right) - \frac{|\mathcal{H}^{\mathsf{c}}|}{n}\left(\epsilon+\frac{\epsilon}{2}+\epsilon\right) - \frac{\epsilon}{2}\right), \tag{94}$$

我们还可以将 $\frac{|\mathcal{H}^{\mathsf{c}}|}{n}$ 的上界和下界表示为

$$\frac{1}{4} \leq 1-\frac{3\alpha'}{4} \leq \frac{n-3\alpha' k/2}{n} \leq \frac{|\mathcal{H}^{\mathsf{c}}|}{n} \leq \frac{n}{n} = 1, \tag{95}$$

这意味着（94）可界定为

$$\mathrm{d_{TV}}\left(p, p_1 \times p_2\right) \geq \frac{1}{2C}\left(\frac{\epsilon'}{4}(1-\epsilon) - 3\epsilon\right). \tag{96}$$

鉴于（ $C \leq 1+3\alpha'/2 \leq 5/2$ ）中已部分建立了上界 66，最终可得

$$\mathrm{d_{TV}}\left(p, p_1 \times p_2\right) \geq \frac{1}{5}\left(\frac{\epsilon'}{4}(1-\epsilon) - 3\epsilon\right). \tag{97}$$

任何满足 ϵ' 的 $\epsilon' \geq \frac{72\epsilon}{1-\epsilon}$都能保证 $\mathrm{d_{TV}}\left(p, p_1 \times p_2\right) \geq 3\epsilon$。例如，令 $\epsilon' = 192\epsilon$，即可确保 $\mathrm{d_{TV}}\left(p, p_1 \times p_2\right) \geq 3\epsilon$ 对所有 $\epsilon \leq 1/6$均成立。应用引理 A.1，这意味着对于所有乘积分布 q，$\mathrm{d_{TV}}\left(p, q\right) \geq \epsilon$。

$|\mathcal{S}|$ 有界 现在我们证明，第一个事件—— $\mathcal{S}$ 的大小以下界为 $k/100$——以高概率成立。令$a_{i,\ j}$表示 $\mathcal{S}$ 中（$i,\ j$）的副本数。根据构造，$a_{i,\ j}$服从 $\mathsf{Poi}(k\cdot q(i,\ j))$，而 $\mathcal{S}$ 的大小是所有$a_{i,\ j}$的求和。由于泊松随机变量的求和本身仍是泊松随机变量，

$$|\mathcal{S}| = \sum_{(i,j)} a_{i,j} \sim \mathsf{Poi}\left(\sum_{(i,j)} k\cdot q(i,j)\right) = \mathsf{Poi}\left(k\cdot\|q\|_1\right) \tag{98}$$

为了证明 $|\mathcal{S}|$ 以高概率至少为 $k/100$，我们通过霍夫丁界证明 $\|q\|_1$ 为 $\Theta(1)$，并对泊松随机变量应用切尔诺夫界。考虑 q的范数：

$$\|q\|_1 = \sum_{i\in[n]}\sum_{j\in m}\left|c_i\cdot\mu(i,j)\right| = \sum_{i\in\mathcal{H}} c_i \sum_{j\in m}\left|\mu(i,j)\right| + \sum_{i\in\mathcal{H}^{\mathsf{c}}} c_i \sum_{j\in m}\left|\mu(i,j)\right|. \tag{99}$$

当 $X=0$，μ（$i,\ j$）始终为 $1/m$。因此，（99）为：

$$\|q\|_1 = \sum_{i\in\mathcal{H}} c_i \sum_{j\in[m]}\frac{1}{m} + \sum_{i\in\mathcal{H}^{\mathsf{c}}} c_i \sum_{j\in[m]}\frac{1}{m} = \sum_{i\in\mathcal{H}} c_i + \sum_{i\in\mathcal{H}^{\mathsf{c}}} c_i = C. \tag{100}$$

另一方面，当 $X=1$，$\mu(i,\ j)$ 取决于 i。尽管仍然成立的是 $\mu(i,\ j)=1/m$ 当$i\in\mathcal{H}$，当 $i\in\mathcal{H}^{\mathsf{c}}$，$\mu(i,\ j)$ 为 $(1+\epsilon'_{ij})/m$。在这种情况下，(99) 为：

$$\|q\|_1=\sum_{i\in\mathcal{H}}c_i\sum_{j\in[m]}\frac{1}{m}+\sum_{i\in\mathcal{H}^{\mathsf{c}}}c_i\sum_{j\in[m]}\left|\frac{1+\epsilon'_{ij}}{m}\right| \tag{101}$$

$$=\sum_{i\in\mathcal{H}}c_i\sum_{j\in[m]}\frac{1}{m}+\sum_{i\in\mathcal{H}^{\mathsf{c}}}c_i\sum_{j\in[m]}\frac{1+\epsilon'_{ij}}{m} \tag{102}$$

$$=\sum_{i\in\mathcal{H}}c_i+\sum_{i\in\mathcal{H}^{\mathsf{c}}}c_i+\sum_{i\in\mathcal{H}^{\mathsf{c}}}c_i\sum_{j\in[m]}\frac{\epsilon'_{ij}}{m} \tag{103}$$

$$=C+\sum_{i\in\mathcal{H}^{\mathsf{c}}}\sum_{j\in[m]}\frac{\epsilon'_{ij}}{nm}. \tag{104}$$

由于每个 $i,\ j$ 在 (104) 中的项都$\pm\frac{\epsilon'}{nm}$，我们可以对它们的求和应用霍夫丁界：

$$\mathbf{Pr}\left[\left|\sum_{i\in\mathcal{H}^{\mathsf{c}}}\sum_{j\in[m]}\frac{\epsilon'_{ij}}{nm}\right|\geq 0.1\right]\leq 2\exp\left(-\frac{0.005\cdot n^2m}{|\mathcal{H}^{\mathsf{c}}|\,\epsilon'^2}\right). \tag{105}$$

根据 (95)，我们知道 $|\mathcal{H}^{\mathsf{c}}|\leq n$，或者等价地，$-\frac{1}{|\mathcal{H}^{\mathsf{c}}|}\leq-\frac{1}{n}$。代入这个不等式后，霍夫丁界可简化为：

$$\mathbf{Pr}\left[\left|\sum_{i\in\mathcal{H}^{\mathsf{c}}}\sum_{j\in[m]}\frac{\epsilon'_{ij}}{nm}\right|\geq 0.1\right]\leq 2\exp\left(-\frac{0.005\cdot nm}{\epsilon'^2}\right)\leq 0.01, \tag{106}$$

其中，只要 nm/ϵ'^2 足够大，最后一个不等式便成立。结合 (中确立的 C 界 64) 和 (66) 中的界，当 $X=0$ 时，

$$1\leq\|q\|_1=C\leq 2.5, \tag{107}$$

并且，当 $X=1$ 时，以至少 0.99 的概率，

$$0.9\leq C-0.1\leq\|q\|_1\leq C+0.1\leq 2.6\,. \tag{108}$$

由于 $\|q\|=\Theta(1)$ 以高概率成立，无论 X 如何，$|\mathcal{S}|\sim$ 泊松分布$(k\|q\|_1)$ 都应以高概率为 $\Theta(k)$。现在我们将证明 $|\mathcal{S}|$ 上的高概率下界 $k/100$（对于这个引理，上界并非必要）。首先注意，给定 $0.9\leq\|q\|_1$ 后，泊松分布$(k\cdot 0.9)$ 在随机意义上小于泊松分布$(k\|q\|_1)$，这意味着

$$\mathbf{Pr}\left[|\mathcal{S}|\leq\frac{k}{100}\mid 0.9\leq\|q\|_1\right]=\mathbf{Pr}_{S\sim\mathtt{Poi}(k\|q\|_1)}\left[S\leq\frac{k}{100}\mid 0.9\leq\|q\|_1\right]\leq\mathbf{Pr}_{S\sim\mathtt{Poi}(0.9k)}\left[S\leq\frac{k}{100}\right]. \tag{109}$$

应用泊松分布特有的切尔诺夫界（例如，参见 [MU05]中的定理5.4），我们可以将（109）界定为

$$\mathbf{Pr}\left[|\mathcal{S}|\leq\frac{k}{100}\mid 0.9\leq\|q\|_1\right]\leq\frac{e^{-0.9k}\cdot(0.9k\cdot e)^{k/100}}{(k/100)^{k/100}}\leq 0.01\,, \tag{110}$$

其中，最后一个不等式对任意 $k \geq 6$都成立。结合（106）和（110），并利用全概率公式，最终可得，在事件1、2 和3的条件下，

$$\mathbf{Pr}\left[|\mathcal{S}| \geq \frac{k}{100}\right] \leq 0.02\,. \tag{111}$$

综上所述。 到目前为止，我们已经证明事件 1、2 和 3 各自发生的概率至少为 0。99。根据并集界，这三个事件同时发生的概率至少为 0。97。在这些事件发生的条件下，我们已经证明：(i) 若 $X = 0$，则 p 是一个乘积分布，且 $\mathsf{d}_{\mathrm{TV}}(p, \hat{p}) \leq \alpha$；(ii) 若 $X = 1$，则 p与所有乘积分布都相差至少 ϵ；并且 (iii) $|\mathcal{S}| \leq k/100$ 以至少 0。98 的概率成立。最终，根据链式法则，三个目标同时成立的概率至少为 $0.97 \cdot 0.98 = 0.9506 > 0.95$。 □

C.2 引理 4.2 的证明

引理 4.2。 令 $\alpha \in [0,1], \epsilon \in (0,1)$ $k \in \mathbb{N}$。令 X 以相同概率取值为 0 或 1 。令 q 为按照 (32) 使用 X构造的 $[n] \times [m]$ 上的测度。对于任意 $i \in [n]$，令 $A := A_i$ 为一个计数向量 m，满足 $a_{i,\,j} \sim$ 泊松分布$(kq(i,j))$ 对所有 $j \in [m]$ 均成立。那么， X 与 A 之间的互信息满足如下界：

$$I(X;A) \leq \mathcal{O}\left(\frac{\epsilon^4 k^3}{n^3 m \alpha}\right). \tag{39}$$

证明。 对于所有 $v \in \mathbb{N}^m$，定义 $x_{\min}(v)$ 和 $x_{\max}(v)$ 如下：

$$x_{\min}(v) = \arg\min_{x \in \{0,1\}} \left(\mathbf{Pr}[A = v \mid X = x]\right) \tag{112}$$

$$x_{\max}(v) = \arg\max_{x \in \{0,1\}} \left(\mathbf{Pr}[A = v \mid X = x]\right) \tag{113}$$

注意，在证明余下部分中，我们将省略 v，写作 $x_{\min}$ 和 $x_{\max}$。 v 的含义将从上下文中明确。引理 C.1 使我们能够将 X 与 A 之间的互信息公式改写为便于分析的形式。

引理 C.1。 设 X 为 0 或 1，且二者出现的概率相等。设 A 为一个取值于 $\mathbb{N}^m$的随机向量。对于所有 $v \in \mathbb{N}^m$，令 $x_{\min}(v)$ 和 $x_{\max}(v)$ 如 (113)所定义。那么， X与 A 之间的互信息满足：

$$I(X;A) = \sum_{v \in \mathbb{N}^m} \mathcal{O}\left(\mathbf{Pr}[A = v \mid X = x_{\min}(v)]\left(1 - \frac{\mathbf{Pr}[A = v \mid X = x_{\max}(v)]}{\mathbf{Pr}[A = v \mid X = x_{\min}(v)]}\right)^2\right). \tag{114}$$

引理 C.3 的证明见附录C.1。

为继续证明我们的界，我们将引理 C.1 中给出的求和拆分为三个独立的求和：分别对所有满足 v $\|v\|_1 = 0$、 $\|v\|_1 = 1$以及 $\|v\|_1 \geq 2$的情况求和。直观地说，互信息的主要贡献来自第三种情况，即至少两个样本落入相应的行。在轻行中，算法可以通过比较不同分箱中的命中次数，获得一些关于其是否均匀，或是否与均匀性相差 ϵ的距离的信息。

回想一下， A 对应于定义域中的某一行 i，该行由随机变量 c_i刻画。当满足 $c_i = 1/n$时，这一行是“轻”的；当满足 $c_i = 1/k$时，则是“重”的。根据构造，当 $c_i = 1/k$时，每个$a_{ij} \sim$ 都服从 Poi（$1/m$），且与 X无关。因此，对于所有 v，

$$\mathbf{Pr}[A = v, c_i = 1/k \mid X = x_{\min}(v)] = \mathbf{Pr}[A = v, c_i = 1/k \mid X = x_{\max}(v)]. \tag{115}$$

由此可知 公式（115）、全概率公式以及我们对 $x_{\min}$ 和 $x_{\max}$ 的定义， we

还可以得出，对于所有 v,

$$\mathbf{Pr}[A = v, c_i = 1/n \mid X = x_{\min}(v)] \leq \mathbf{Pr}[A = v, c_i = 1/n \mid X = x_{\max}(v)]. \tag{116}$$

利用加性平移下比值的单调性，我们可以证明，方程（115）和（116）结合全概率公式，可推出对所有 v,

$$\mathcal{O}\left(\left(1-\frac{\mathbf{Pr}[A = v \mid X = x_{\max}(v)]}{\mathbf{Pr}[A = v \mid X = x_{\min}(v)]}\right)^2\right) \leq \mathcal{O}\left(\left(1-\frac{\mathbf{Pr}[A = v \mid X = x_{\max}(v), c_i = 1/n]}{\mathbf{Pr}[A = v \mid X = x_{\min}(v), c_i = 1/n]}\right)^2\right). \tag{117}$$

我们将此不等式应用于 $\|v\|_1 = 0$ 和 $\|v\|_1 = 1$ 两种情况，得到互信息的以下界：

$$\begin{aligned} I(X;A) = &\sum_{v:\|v\|_1=0} \mathcal{O}\left(\mathbf{Pr}[A = v \mid X = x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A = v \mid X = x_{\max}(v), c_i = 1/n]}{\mathbf{Pr}[A = v \mid X = x_{\min}(v), c_i = 1/n]}\right)^2\right) \\ &+ \sum_{v:\|v\|_1=1} \mathcal{O}\left(\mathbf{Pr}[A = v \mid X = x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A = v \mid X = x_{\max}(v), c_i = 1/n]}{\mathbf{Pr}[A = v \mid X = x_{\min}(v), c_i = 1/n]}\right)^2\right) \\ &+ \sum_{v:\|v\|_1\geq 2} \mathcal{O}\left(\mathbf{Pr}[A = v \mid X = x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A = v \mid X = x_{\max}(v)]}{\mathbf{Pr}[A = v \mid X = x_{\min}(v)]}\right)^2\right). \end{aligned} \tag{118}$$

在引理 C.2、C.3和C.4中，我们分别对每一项进行界定。合并这些界后，可得

$$I(X;A) \leq \mathcal{O}\left(\frac{\epsilon^4 k^4}{n^4 m^2}\right) + \mathcal{O}\left(\frac{\epsilon^4 k^3}{n^3 m^2}\right) + \mathcal{O}\left(\frac{\epsilon^4 k^3}{n^3 m\alpha}\right) = \mathcal{O}\left(\frac{\epsilon^4 k^3}{n^3 m\alpha}\right). \tag{119}$$

□

C.3 引理的证明C.1

引理 C.1。设 X 以相等的概率取0 或1 。设 A 是一个取值于$\mathbb{N}^m$的随机向量。对于所有 $v \in \mathbb{N}^m$，令 $x_{\min}(v)$ 和 $x_{\max}(v)$ 如 (113) 中所定义。则 X 与 A 之间的互信息满足：

$$I(X;A) = \sum_{v\in\mathbb{N}^m} \mathcal{O}\left(\mathbf{Pr}[A = v \mid X = x_{\min}(v)]\left(1 - \frac{\mathbf{Pr}[A = v \mid X = x_{\max}(v)]}{\mathbf{Pr}[A = v \mid X = x_{\min}(v)]}\right)^2\right). \tag{114}$$

证明。互信息的标准定义如下：

$$\begin{aligned} I(X;A) &= \sum_v \sum_{x\in\{0,1\}} \mathbf{Pr}[A = v, X = x] \log\left(\frac{\mathbf{Pr}[A = v, X = x]}{\mathbf{Pr}[X = x]\mathbf{Pr}[A = v]}\right) \\ &= \sum_v \sum_{x\in\{0,1\}} \mathbf{Pr}[X = x]\mathbf{Pr}[A = v \mid X = x] \log\left(\frac{\mathbf{Pr}[A = v \mid X = x]}{\mathbf{Pr}[A = v]}\right) \end{aligned}$$

为简洁起见，我们引入以下记号：

- $\beta = \mathbf{Pr}[A = v \mid X = x_{\max}(v)]$
- $\gamma = \mathbf{Pr}[A = v \mid X = x_{\min}(v)]$
- $p = \mathbf{Pr}[A = v] = \frac{1}{2}\mathbf{Pr}[A = v \mid X = x_{\max}(v)] + \frac{1}{2}\mathbf{Pr}[A = v \mid X = x_{\min}(v)] = \frac{\beta+\gamma}{2}$

将这些代入互信息的定义，得到：

$$\begin{aligned}
I(X;A) &= \sum_v \sum_{x\in\{0,1\}} \mathbf{Pr}[X=x]\mathbf{Pr}[A=v \mid X=x] \log\left(\frac{\mathbf{Pr}[A=v \mid X=x]}{\mathbf{Pr}[A=v]}\right) \\
&= \sum_v \frac{1}{2}\beta \log\left(\frac{\beta}{p}\right) + \frac{1}{2}\gamma \log\left(\frac{\gamma}{p}\right) \\
&= \sum_v \frac{1}{2}\beta\log\beta + \frac{1}{2}\gamma\log\gamma - \frac{\beta+\gamma}{2}\log p \\
&= \sum_v \frac{\beta-p}{2}\log\beta + \frac{\gamma-p}{2}\log\gamma + \frac{p}{2}(\log\beta+\log\gamma) - p\log p \\
&= \sum_v \frac{\beta-\gamma}{4}\log\beta + \frac{\gamma-\beta}{4}\log\gamma + p\log\sqrt{\beta\gamma} - p\log p \\
&= \sum_v \frac{\beta-\gamma}{4}\log\left(\frac{\beta}{\gamma}\right) + p\log\left(\frac{\sqrt{\beta\gamma}}{p}\right) \\
&\le \sum_v \frac{\beta-\gamma}{4\ln(2)}\left(\frac{\beta}{\gamma}-1\right) \\
&= \sum_v \frac{1}{4\ln(2)}\frac{(\beta-\gamma)^2}{\gamma} \\
&= \sum_v \frac{\gamma}{4\ln(2)}\left(\frac{\beta}{\gamma}-1\right)^2,
\end{aligned}$$

其中，最后一个不等式成立，是因为 $\ln(x) \le x-1$ 且几何平均数始终小于或等于算术平均数，即 $\sqrt{\beta\gamma} \le p = \frac{\beta+\gamma}{2}$。将 β 和 γ 的表达式代入，即可得到互信息的所需上界：

$$I(X;A) = \sum_v \mathcal{O}\left(\mathbf{Pr}[A=v \mid X=x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A=v \mid X=x_{\max}(v)]}{\mathbf{Pr}[A=v \mid X=x_{\min}(v)]}\right)^2\right).$$

□

C.4 辅助引理

引理 *C.2.* 对任意 $i \in [n]$，令 $A := A_i$ 为一个包含 m 个计数的向量，使得对于所有 $j \in [m]$，$a_{i,j} \sim$ 服从 *Poi(kq(i, j))*，其中 q 按式 *(31)* 构造。对于所有 $v \in \mathbb{N}^m$，按式 *(113)* 定义 $x_{\min}(v)$ 和 $x_{\max}(v)$。则下式成立。

$$\begin{aligned}
&\sum_{v:\|v\|_1=0} \mathcal{O}\left(\mathbf{Pr}[A=v \mid X=x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A=v \mid X=x_{\max}(v), c_i=1/n]}{\mathbf{Pr}[A=v \mid X=x_{\min}(v), c_i=1/n]}\right)^2\right) \\
&\qquad \le \mathcal{O}\left(\frac{\epsilon^4 k^4}{n^4 m^2}\right).
\end{aligned}$$

证明。首先请注意，只有一个 v 满足 $\|v\|_1 = 0$，因此可以移除求和，并将 v 替换为 $\mathbf{0}$。我们还可以通过将 $\mathbf{Pr}[A=v \mid X=x_{\min}(v)]$ 上界为

1。这样我们得到：

$$\sum_{v:\|v\|_1=0} \mathcal{O}\left(\mathbf{Pr}[A=v \mid X=x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A=v \mid X=x_{\max}(v), c_i=1/n]}{\mathbf{Pr}[A=v \mid X=x_{\min}(v), c_i=1/n]}\right)^2\right)$$
$$\leq \mathcal{O}\left(\left(1-\frac{\mathbf{Pr}[|A|=0 \mid X=x_{\max}(\mathbf{0}), c_i=1/n]}{\mathbf{Pr}[|A|=0 \mid X=x_{\min}(\mathbf{0}), c_i=1/n]}\right)^2\right). \tag{120}$$

为界定此表达式，我们计算 **概率**$[|A|=0 \mid X=0, c_i=1/n]$和 **概率**$[|A|=0 \mid X=1, c_i=1/n]$。由于在给定 $X=x$ 的条件下，每个 a_{ij} 都是独立同分布的， $|A|=0$ 的概率等于$a_{ij}=0$的概率的乘积。因此，我们有：

$$\mathbf{Pr}[|A|=0 \mid X=0, c_i=1/n]=\prod_{j=1}^{m}\mathbf{Pr}[a_{i,j}=0]=\prod_{j=1}^{m}e^{-\frac{k}{nm}}=e^{-\frac{k}{n}},$$

and

$$\begin{aligned}\mathbf{Pr}[|A|=0 \mid X=1, c_i=1/n] &= \prod_{j=1}^{m}\mathbf{Pr}[a_{i,j}=0]\\ &= \prod_{j=1}^{m}\frac{1}{2}\left(e^{-\frac{k(1+\epsilon)}{nm}}+e^{-\frac{k(1-\epsilon)}{nm}}\right)\\ &= \prod_{j=1}^{m}e^{-\frac{k}{nm}}\left(\frac{e^{-\frac{k\epsilon}{nm}}+e^{\frac{k\epsilon}{nm}}}{2}\right)\\ &= e^{-\frac{k}{n}}\cdot\cosh^m\left(\frac{k\epsilon}{nm}\right).\end{aligned}$$

由于双曲余弦$(x)\geq 1$，我们知道 $x_{\max}=1$ 和 $x_{\min}=0$。因此，所讨论的概率之比有如下上界：

$$\cosh^m\left(\frac{k\epsilon}{nm}\right)\leq e^{\frac{m}{2}\left(\frac{k\epsilon}{nm}\right)^2}\leq 1+\frac{k^2\epsilon^2}{n^2m},$$

其中我们使用了双曲余弦$(x)\leq e^{x^2/2}$，且 $e^x\leq 1+2x$ 对所有 $x\in[0,1]$均成立。这意味着 (120) 满足：

$$\mathcal{O}\left(\left(1-\frac{\mathbf{Pr}[|A|=0 \mid X=x_{\max}(\mathbf{0}), c_i=1/n]}{\mathbf{Pr}[|A|=0 \mid X=x_{\min}(\mathbf{0}), c_i=1/n]}\right)^2\right)\leq\mathcal{O}\left(\frac{k^4\epsilon^4}{n^4m^2}\right).$$

□

引理 *C.3.* 对于任意 $i\in[n]$，令 $A\coloneqq A_i$ 为一个由 m 个计数组成的向量，使得对所有 $j\in[m]$，$a_{i,j}\sim$ 服从泊松分布 $(kq(i,j))$，其中 q 按照 *(31)* 中的方式构造。对于所有 $v\in\mathbb{N}^m$，定义 $x_{\min}(v)$ 和 $x_{\max}(v)$，如 *(113)* 所示。则有以下结论。

$$\sum_{v:\|v\|_1=1} \mathcal{O}\left(\mathbf{Pr}[A=v \mid X=x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A=v \mid X=x_{\max}(v), c_i=1/n]}{\mathbf{Pr}[A=v \mid X=x_{\min}(v), c_i=1/n]}\right)^2\right)$$
$$\leq \mathcal{O}\left(\frac{\epsilon^4k^3}{n^3m^2}\right).$$

证明。 存在 $\binom{m}{1}$ 个标准基向量，使得 $|v|_1=1$，每个向量在坐标 j' 处为 1，其余位置均为 0。考虑以下条件概率，可以通过计算 $A=v$ 在给定 $X=x$ 和 $c_i=1/n$ 的条件下的概率得到：

给定 $X = x$ 时，每个 $a_{i,\ j}$ 的独立性：

$$\begin{aligned}\mathbf{Pr}[A = v \mid X = 0, c_i = 1/n] &= \mathbf{Pr}[a_{i,j'} = 1] \prod_{j \neq j'} \mathbf{Pr}[a_{i,j} = 0] \\ &= \frac{k}{nm} e^{-\frac{k}{nm}} \left(e^{-\frac{k}{nm}}\right)^{m-1} \\ &= \frac{k}{nm} e^{-\frac{k}{n}}, \end{aligned} \tag{121}$$

$$\begin{aligned}&\mathbf{Pr}[A = v \mid X = 1, c_i = 1/n] = \mathbf{Pr}[a_{i,j'} = 1] \prod_{j \neq j'} \mathbf{Pr}[a_{i,j} = 0] \\ &= \frac{1}{2}\left(\frac{k(1+\epsilon)}{nm} e^{-\frac{k(1+\epsilon)}{nm}} + \frac{k(1-\epsilon)}{nm} e^{-\frac{k(1-\epsilon)}{nm}}\right) \prod_{j \neq j'} \frac{1}{2}\left(e^{-\frac{k(1+\epsilon)}{nm}} + e^{-\frac{k(1-\epsilon)}{nm}}\right) \\ &= \frac{k}{nm} e^{-\frac{k}{nm}} \left(\cosh\left(\frac{k\epsilon}{nm}\right) - \epsilon \sinh\left(\frac{k\epsilon}{nm}\right)\right) \prod_{j \neq j'} e^{-\frac{k}{nm}} \cosh\left(\frac{k\epsilon}{nm}\right) \\ &= \frac{k}{nm} e^{-\frac{k}{n}} \cosh^m\left(\frac{k\epsilon}{nm}\right) \left(1 - \epsilon \tanh\left(\frac{k\epsilon}{nm}\right)\right), \end{aligned} \tag{122}$$

现在通过设定 $u := \frac{k\epsilon}{nm}$ 并应用不等式 $1 \leq$ 双曲余弦$^m(u) \leq e^{mu^2/2}$，以及 $u - u^3/3 \leq$ 双曲正切(u) $\leq u$，对所有 $u \geq 0$，来界定（121）与（122）中的概率之比。注意，由于 $mu \leq \frac{\epsilon}{2}$，$\frac{k}{n} \leq 0.5$。 – –

$$\begin{aligned} &1 - \epsilon u \leq \cosh^m(u)\,(1 - \epsilon \tanh(u)) \leq e^{\frac{mu^2}{2}} \left(1 - \epsilon u + \epsilon \frac{u^3}{3}\right) \\ &1 - \epsilon u \leq \cosh^m(u)\,(1 - \epsilon \tanh(u)) \leq e^{\frac{\epsilon u}{4}} \left(1 - \frac{11}{12}\epsilon u\right) \leq 1 \\ \Rightarrow \quad &\frac{1}{1 - \epsilon u} \geq \frac{1}{\cosh^m(u)\,(1 - \epsilon \tanh(u))} \geq 1 \end{aligned} \tag{123}$$

由于 $\frac{k}{n} \leq 0$。5，因此有 $0 \leq \epsilon u \leq 0.5$。所以，我们可以使用不等式 $\frac{1}{1-\epsilon u} \leq 1 + 2\epsilon u$，从而 s:

$$1 \leq \frac{1}{\cosh^m\left(\frac{k\epsilon}{nm}\right)\left(1 - \epsilon \tanh(\frac{k\epsilon}{nm})\right)} \leq 1 + \frac{2k\epsilon^2}{nm} \tag{124}$$

下一步是证明 $\sum_{v:\ \|v\|_1=1} \mathbf{Pr}[A = v \mid X = x_{\min}(v)] = \mathcal{O}(k/n)$。首先，根据 $x_{\min}(v)$的定义，我们有

$$\sum_{v:\|v\|_1=1} \mathbf{Pr}[A = v \mid X = x_{\min}(v)] \leq \sum_{v:\|v\|_1=1} \mathbf{Pr}[A = v \mid X = 0]. \tag{125}$$

接下来我们可以界定 通过应用全概率公式对 c 的不同情况进行考量，得到这个求和 $_i$

对于每一项：

$$\begin{aligned}\sum_{v:\|v\|_1=1}\mathbf{Pr}[A=v\mid X=0] &= \sum_{v:\|v\|_1=1}\left[\frac{\alpha k}{n}\mathbf{Pr}[A=v\mid X=0,c_i=1/k]\right.\\ &\qquad\left.+\left(1-\frac{\alpha k}{n}\right)\mathbf{Pr}[A=v\mid X=0,c_i=1/n]\right]\\ &= \sum_{v:\|v\|_1=1}\frac{\alpha k}{n}\left(\frac{1}{m}e^{-\frac{1}{m}}\left(e^{-\frac{1}{m}}\right)^{m-1}\right)+\left(1-\frac{\alpha k}{n}\right)\frac{k}{nm}e^{-\frac{k}{n}}\\ &= \frac{k}{n}\left(\alpha e^{-1}+\left(1-\frac{\alpha k}{n}\right)e^{-\frac{k}{n}}\right)\\ &\le \mathcal{O}\left(\frac{k}{n}\right)\end{aligned}\tag{126}$$

最后一个不等式成立是因为 $0\le\frac{k}{n}\le$–0.5；因此，括号内的项在$[e^{-\frac{1}{2}},1+e^{-1}]$内保持为常数。我们可以结合式（124）和（126）中的界，得到所需的上界 $\mathcal{O}\left(\frac{k}{n}\left(\frac{k\epsilon^2}{nm}\right)^2\right)=\mathcal{O}\left(\frac{k^3\epsilon^4}{n^3m^2}\right)$，从而完成证明。

□

引理 *C.4.* 对任意 $i\in[n]$，令 $A:=A_i$ 为一个由 m 个计数组成的向量，使得对所有 $j\in[m]$，$a_{i,j}\sim$ 服从泊松分布 *Poi(kq(i, j))*，其中 q 按式（*31*）构造。对所有 $v\in N^m$，定义 $x_{\min}(v)$ 和 $x_{\max}(v)$，如式（*113*）所示。则有以下结论。

$$\sum_{v:\|v\|_1\ge 2}\mathcal{O}\left(\mathbf{Pr}[A=v\mid X=x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A=v\mid X=x_{\max}(v)]}{\mathbf{Pr}[A=v\mid X=x_{\min}(v)]}\right)^2\right)\le\mathcal{O}\left(\frac{\epsilon^4k^3}{n^3m\alpha}\right)$$

为证明引理 C.4，我们使用引理 C.5 证明可以将给定表达式界定为 $\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot I(X;\ A\mid c_i=1/n)$. 然后，利用链式法则，我们可以将 $I(X;\ A\mid c_i=1/n)$界定为对所有 j的$I(X;\ a_{i,\ j}\mid c_i=1/n)$求和。接着，我们使用引理 C.6 的结果来界定该求和中的每一项。

证明。 令 v 为 A 的一个可能取值，且满足 $\|v\|_1\ge 2$。考虑互信息界中涉及 v的项：

$$\mathbf{Pr}[A=v\mid X=x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A=v\mid X=x_{\max}(v)]}{\mathbf{Pr}[A=v\mid X=x_{\min}(v)]}\right)^2.\tag{127}$$

应用全概率公式，等式（127）等于：

$$\begin{aligned}&(\mathbf{Pr}[A=v,c_i=1/k\mid X=x_{\min}(v)]+\mathbf{Pr}[A=v,c_i=1/n\mid X=x_{\min}(v)])\\ &\quad\cdot\left(1-\frac{\mathbf{Pr}[A=v,c_i=1/k\mid X=x_{\max}(v)]+\mathbf{Pr}[A=v,c_i=1/n\mid X=x_{\max}(v)]}{\mathbf{Pr}[A=v,c_i=1/k\mid X=x_{\min}(v)]+\mathbf{Pr}[A=v,c_i=1/n\mid X=x_{\min}(v)]}\right)^2.\end{aligned}\tag{128}$$

注意，如果 $c_i=1/k$, A 不依赖于 X，那么无论 X 是 0 还是 1， $a_{ij}\sim$都服从泊松分布$(1/m)$。在接下来的几步证明中，为简洁起见，我们采用以下记号：

- $\beta=\mathbf{Pr}[A=v,c_i=1/n\mid X=x_{\max}(v)]$,
- $\gamma=\mathbf{Pr}[A=v,c_i=1/n\mid X=x_{\min}(v)]$,
- $\kappa=\mathbf{Pr}[A=v,c_i=1/k\mid X=x_{\max}(v)]=\mathbf{Pr}[A=v,c_i=1/k\mid X=x_{\min}(v)]$.

使用这一记号，方程（128）可改写为：

$$(\kappa+\gamma)\left(1-\frac{\kappa+\beta}{\kappa+\gamma}\right)^2=(\kappa+\gamma)\left(\frac{\gamma-\beta}{\kappa+\gamma}\right)^2=\frac{1}{\kappa+\gamma}\cdot(\gamma-\beta)^2.$$

根据引理 C.5，对于所有满足 v 的 $\|v\|_1\geq 2$，以及所有 x，我们有：

$$\mathbf{Pr}[A=v,c_i=1/n\mid X=x]=\mathcal{O}\left(\frac{k}{n\alpha}\right)\mathbf{Pr}[A=v,c_i=1/k\mid X=x].$$

应用该引理，令 $x=x_{\min}(v)$ 并按我们的记号改写界，可得：

$$\begin{aligned}
&\gamma\leq\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\kappa\\
&\implies\gamma+\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\gamma\leq\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\kappa+\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\gamma\\
&\implies\gamma+\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\gamma\leq\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot(\kappa+\gamma)\\
&\implies\frac{1}{\kappa+\gamma}\leq\frac{\mathcal{O}\left(\frac{k}{n\alpha}\right)}{\gamma\left(1+\mathcal{O}\left(\frac{k}{n\alpha}\right)\right)}\\
&\implies\frac{1}{\kappa+\gamma}\leq\frac{\mathcal{O}\left(\frac{k}{n\alpha}\right)}{\gamma}.
\end{aligned}$$

因此，我们可以对等式（129）给出下界：

$$\frac{1}{\kappa+\gamma}\cdot(\gamma-\beta)^2\leq\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\frac{(\gamma-\beta)^2}{\gamma}=\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\gamma\cdot\left(\frac{\gamma-\beta}{\gamma}\right)^2=\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\gamma\cdot\left(1-\frac{\beta}{\gamma}\right)^2.$$

代入 β 和 γ 的值后，此界为：

$$\begin{aligned}
&\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\mathbf{Pr}[A=v,c_i=1/n\mid X=x_{\min}(v)]\cdot\left(1-\frac{\mathbf{Pr}[A=v,c_i=1/n\mid X=x_{\max}(v)]}{\mathbf{Pr}[A=v,c_i=1/n\mid X=x_{\min}(v)]}\right)^2\\
&\quad=\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\mathbf{Pr}[A=v\mid X=x_{\min}(v),c_i=1/n]\,\mathbf{Pr}[c_i=1/n]\\
&\qquad\cdot\left(1-\frac{\mathbf{Pr}[A=v\mid X=x_{\max}(v),c_i=1/n]\,\mathbf{Pr}[c_i=1/n]}{\mathbf{Pr}[A=v\mid X=x_{\min}(v),c_i=1/n]\,\mathbf{Pr}[c_i=1/n]}\right)^2\\
&\quad=\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\left(1-\frac{\alpha k}{n}\right)\cdot\mathbf{Pr}[A=v\mid X=x_{\min}(v),c_i=1/n]\\
&\qquad\cdot\left(1-\frac{\mathbf{Pr}[A=v\mid X=x_{\max}(v),c_i=1/n]}{\mathbf{Pr}[A=v\mid X=x_{\min}(v),c_i=1/n]}\right)^2\\
&\quad\leq\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\mathbf{Pr}[A=v\mid X=x_{\min}(v),c_i=1/n]\cdot\left(1-\frac{\mathbf{Pr}[A=v\mid X=x_{\max}(v),c_i=1/n]}{\mathbf{Pr}[A=v\mid X=x_{\min}(v),c_i=1/n]}\right)^2.
\end{aligned}\tag{129}$$

到目前为止，我们已经证明，对于满足 $\|v\|_1\geq 2$ 的固定 v：

$$\begin{aligned}
&\mathbf{Pr}[A=v\mid X=x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A=v\mid X=x_{\max}(v)]}{\mathbf{Pr}[A=v\mid X=x_{\min}(v)]}\right)^2\\
&\quad\leq\mathcal{O}\left(\frac{k}{n\alpha}\right)\cdot\mathbf{Pr}[A=v\mid X=x_{\min}(v),c_i=1/n]\cdot\left(1-\frac{\mathbf{Pr}[A=v\mid X=x_{\max}(v),c_i=1/n]}{\mathbf{Pr}[A=v\mid X=x_{\min}(v),c_i=1/n]}\right)^2
\end{aligned}\tag{130}$$

我们可以应用这个界来界定所有满足 $\|v\|_1 \geq 2$ 的 v 上的求和：

$$\begin{aligned}
&\sum_{v:\|v\|_1\geq 2} \mathcal{O}\left(\mathbf{Pr}[A=v \mid X=x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A=v \mid X=x_{\max}(v)]}{\mathbf{Pr}[A=v \mid X=x_{\min}(v)]}\right)^2\right) \\
&\quad\leq \sum_{v:\|v\|_1\geq 2} \mathcal{O}\left(\frac{k}{n\alpha}\cdot \mathbf{Pr}[A=v \mid X=x_{\min}(v), c_i=1/n]\cdot\left(1-\frac{\mathbf{Pr}[A=v \mid X=x_{\max}(v), c_i=1/n]}{\mathbf{Pr}[A=v \mid X=x_{\min}(v), c_i=1/n]}\right)^2\right) \\
&\quad\leq \mathcal{O}\left(\frac{k}{n\alpha}\right)\sum_{v} \mathcal{O}\left(\mathbf{Pr}[A=v \mid X=x_{\min}(v), c_i=1/n]\cdot\left(1-\frac{\mathbf{Pr}[A=v \mid X=x_{\max}(v), c_i=1/n]}{\mathbf{Pr}[A=v \mid X=x_{\min}(v), c_i=1/n]}\right)^2\right) \\
&\quad= \mathcal{O}\left(\frac{k}{n\alpha}\right)\Theta\left(I\left(X; A \mid c_i=\frac{1}{n}\right)\right). \qquad (131)
\end{aligned}$$

互信息链式法则允许我们对 $I(X; A \mid c_i = 1/n)$ 进行如下界定：

$$I\left(X; A \mid c_i=\frac{1}{n}\right) \leq \sum_{j=1}^{m} I\left(X; a_{i,j} \mid c_i=\frac{1}{n}\right)$$

当 $c_i = 1/n$， $a_{i,\ j}$ 服从引理C.6中所考虑的随机变量的分布。该引理允许我们对 X 与样本集中第 i 行的每个元素之间的互信息进行界定。
应用引理给出的界，我们得到：

$$\sum_{j=1}^{m} I\left(X; a_{i,j} \mid c_i=\frac{1}{n}\right) \leq \sum_{j=1}^{m} \mathcal{O}\left(\frac{\epsilon^4 k^2}{n^2 m^2}\right) = \mathcal{O}\left(\frac{\epsilon^4 k^2}{n^2 m}\right) \qquad (132)$$

最终，结合方程（131）和方程（132），我们得到所需的界 ：

$$\sum_{v:\|v\|_1\geq 2} \mathcal{O}\left(\mathbf{Pr}[A=v \mid X=x_{\min}(v)]\left(1-\frac{\mathbf{Pr}[A=v \mid X=x_{\max}(v)]}{\mathbf{Pr}[A=v \mid X=x_{\min}(v)]}\right)^2\right) \leq \mathcal{O}\left(\frac{\epsilon^4 k^3}{n^3 m\alpha}\right)$$

□

引理 C.5. 令 $x \in \{0,1\}, v \in \mathbb{N}^m$ 满足 $\|v\|_1 \geq 2\}$。令 c_i 为一个随机变量， q 为一个分布，根据(31) 定义，其中 $X = x$。对于任意的 $i \in [n]$，令 $A := A_i$ 为一个由 m 个计数组成的向量，使得 $a_{i,\ j} \sim$ 泊松分布$(kq(i, j))$ 对所有 $j \in [m]$ 均成立。则有以下结论。

$$\mathbf{Pr}[A=v, c_i=1/n \mid X=x] = \mathcal{O}\left(\frac{k}{n\alpha}\right)\mathbf{Pr}[A=v, c_i=1/k \mid X=x].$$

证明。 方程右侧与 X的具体取值无关，因为它对应于满足 $c_i = 1/k$ 的行。对于这些行，每个元素都服从泊松分布 $a_{i,\ j} \sim$ 泊松分布$(1/m)$。

$$\begin{aligned}
\mathbf{Pr}[A=v, c_i=1/k \mid X=x] &= \mathbf{Pr}[c_i=1/k]\mathbf{Pr}[A=v \mid X=x, c_i=1/k] \\
&= \frac{\alpha k}{n}\prod_{j=1}^{m}\left(\frac{1}{m}\right)^{v_j} e^{-\frac{1}{m}}\frac{1}{v_j!} = \frac{\alpha k e^{-1}}{n m^{\|v\|_1}}\prod_{j=1}^{m}\frac{1}{v_j!} \qquad (133)
\end{aligned}$$

为证明该引理，我们分别对两种情况 $X = 0$ 和 $X = 1$下的左侧取上界，然后证明

这两个界中较大的一个满足所声称的渐近条件。

$$\begin{aligned}\mathbf{Pr}[A=v, c_i=1/n \mid X=0] &= \mathbf{Pr}[c_i=1/n]\mathbf{Pr}[A=v \mid X=0, c_i=1/n] \\ &= \left(1-\frac{\alpha k}{n}\right)\prod_{j=1}^{m}\left(\frac{k}{nm}\right)^{v_j} e^{-\frac{k}{nm}}\frac{1}{v_j!} \\ &\leq \left(\frac{k}{nm}\right)^{\|v\|_1} e^{-\frac{k}{n}}\prod_{j=1}^{m}\frac{1}{v_j!} \end{aligned} \tag{134}$$

$$\begin{aligned}&\mathbf{Pr}[A=v, c_i=1/n \mid X=1] \\ &= \mathbf{Pr}[c_i=1/n]\mathbf{Pr}[A=v \mid X=1, c_i=1/n] \\ &= \left(1-\frac{\alpha k}{n}\right)\prod_{j=1}^{m}\frac{1}{2}\left(\left(\frac{k\,(1+\epsilon)}{nm}\right)^{v_j} e^{-\frac{k(1+\epsilon)}{nm}} + \left(\frac{k\,(1-\epsilon)}{nm}\right)^{v_j} e^{-\frac{k(1-\epsilon)}{nm}}\right)\frac{1}{v_j!} \\ &\leq \left(\frac{k}{nm}\right)^{\|v\|_1} e^{-\frac{k}{n}}\prod_{j=1}^{m}\frac{(1+\epsilon)^{v_j}}{2}\left(e^{-\frac{k\epsilon}{nm}} + \left(\frac{1-\epsilon}{1+\epsilon}\right)^{v_j} e^{\frac{k\epsilon}{nm}}\right)\frac{1}{v_j!} \\ &\leq \left(\frac{k(1+\epsilon)}{nm}\right)^{\|v\|_1} e^{-\frac{k(1-\epsilon)}{n}}\prod_{j=1}^{m}\frac{1}{v_j!} \end{aligned} \tag{135}$$

显然，式（135）中的上界大于式（134）中的上界。因此，只需证明，对于任意满足 $O(\frac{k}{n\alpha}$ 的 v，式（$\|v\|_1 \geq 2$）中概率与式（135）和（133）中概率的比值为 $O(\frac{k}{n\alpha})$ 量级：—

$$\frac{\left(\frac{k(1+\epsilon)}{nm}\right)^{\|v\|_1} e^{-\frac{k(1-\epsilon)}{n}}\prod_{j=1}^{m}\frac{1}{v_j!}}{\frac{\alpha k e^{-1}}{nm^{\|v\|_1}}\prod_{j=1}^{m}\frac{1}{v_j!}} = \frac{k}{n\alpha}\cdot(1+\epsilon)^2\left(\frac{k(1+\epsilon)}{n}\right)^{\|v\|_1-2} e^{1-\frac{k(1-\epsilon)}{n}} \leq 4e\frac{k}{n\alpha} = \mathcal{O}\left(\frac{k}{n\alpha}\right) \tag{136}$$

□

引理 C.6。 设 $\alpha \in [0,1], \epsilon \in (0,1)$ $k \in \mathbb{N}$。设 X 以相同概率取值为 0 或 1。设 q 是定义在 $[n]\times[m]$ 上、按式 (31) 使用 X 构造的分布。对于任意对应于满足 $c_i = \frac{1}{n}$ 的行的 $i \in [n]$，设 $A := A_i$ 是由 m 个计数组成的向量，满足 $a_{i,j} \sim$ 泊松分布$(kq(i,j))$ 对所有 $j \in [m]$ 均成立。那么， X 与 a_{ij} 之间的互信息有如下界：

$$I\left(X; a_{ij} \mid c_i = \frac{1}{n}\right) = \mathcal{O}\left(\frac{\epsilon^4 k^2}{n^2 m^2}\right)$$

证明。 对于所有 l，定义 $x_{\min}(l)$ 和 $x_{\max}(l)$ 如下：

$$x_{\min}(l) = \arg\min_{x\in\{0,1\}}\left(\mathbf{Pr}[a=l \mid X=x]\right), \quad x_{\max}(l) = \arg\max_{x\in\{0,1\}}\left(\mathbf{Pr}[a=l \mid X=x]\right). \tag{137}$$

采用与附录 C.3中类似的论证，我们可以证明互信息由下式给出：

$$I(X;a) = \sum_{l=0}^{\infty}\mathcal{O}\left(\mathbf{Pr}[a=l \mid X=x_{\min}(l)]\left(1-\frac{\mathbf{Pr}[a=l \mid X=x_{\max}(l)]}{\mathbf{Pr}[a=l \mid X=x_{\min}(l)]}\right)^2\right) \tag{138}$$

对于任意 l，我们有：

$$\mathbf{Pr}[a=l \mid X=0] = \frac{e^{-\frac{k}{nm}}\left(\frac{k}{nm}\right)^l}{l!} \tag{139}$$

$$\begin{aligned}\mathbf{Pr}[a=l \mid X=1] &= \frac{1}{2}\left(\frac{e^{-\frac{k(1+\epsilon)}{nm}}\left(\frac{k(1+\epsilon)}{nm}\right)^l}{l!}\right) + \frac{1}{2}\left(\frac{e^{-\frac{k(1-\epsilon)}{nm}}\left(\frac{k(1-\epsilon)}{nm}\right)^l}{l!}\right) \\ &= \frac{e^{-\frac{k}{nm}}\left(\frac{k}{nm}\right)^l}{l!}\left(\frac{e^{-\frac{k\epsilon}{nm}}(1+\epsilon)^l + e^{\frac{k\epsilon}{nm}}(1-\epsilon)^l}{2}\right) \end{aligned} \tag{140}$$

$$\begin{aligned}\frac{\mathbf{Pr}[a=l \mid X=1]}{\mathbf{Pr}[a=l \mid X=0]} &= \frac{e^{-\frac{k\epsilon}{nm}}(1+\epsilon)^l + e^{\frac{k\epsilon}{nm}}(1-\epsilon)^l}{2} \\ &= \cosh\left(\frac{k\epsilon}{nm}\right)\left(\sum_{i=0}^{\lfloor l/2\rfloor}\binom{l}{2i}\epsilon^{2i}\right) - \sinh\left(\frac{k\epsilon}{nm}\right)\left(\sum_{i=0}^{\lfloor \frac{l-1}{2}\rfloor}\binom{l}{2i+1}\epsilon^{2i+1}\right) \end{aligned} \tag{141}$$

我们将把 (138) 中的求和按 $l=0$、$l=1$、$2 \le l < 2/\epsilon$ 和 $l \ge 2/\epsilon$ 拆分。对于 $l=0$，(141) 中的比值等于双曲余弦 $\left(\frac{k\epsilon}{nm}\right)$，它始终大于 1 且小于 $1+\left(\frac{k\epsilon}{nm}\right)^2$。因此，互信息的贡献可界定为

$$\mathbf{Pr}[a=0 \mid X=x_{\min}(l)]\left(1-\frac{\mathbf{Pr}[a=0 \mid X=x_{\max}(l)]}{\mathbf{Pr}[a=0 \mid X=x_{\min}(l)]}\right)^2 \le \left(\frac{k\epsilon}{nm}\right)^4. \tag{142}$$

对于 $l=1$，采用与推导 (123) 类似的推理，我们可以验证 $x_{\max}(l)=0$ 以及 $x_{\min}(l)=1$，并可将 (141) 中比值的倒数界定如下：

$$\begin{aligned}\frac{\mathbf{Pr}[a=1 \mid X=1]}{\mathbf{Pr}[a=1 \mid X=0]} &= \cosh\left(\frac{k\epsilon}{nm}\right) - \epsilon\sinh\left(\frac{k\epsilon}{nm}\right) \\ &= \cosh\left(\frac{k\epsilon}{nm}\right)\left(1-\epsilon\tanh\left(\frac{k\epsilon}{nm}\right)\right) \\ &\ge 1\cdot\left(1-\frac{k\epsilon^2}{nm}\right) \\ \implies \frac{\mathbf{Pr}[a=1 \mid X=x_{\max}(l)]}{\mathbf{Pr}[a=1 \mid X=x_{\min}(l)]} &= \frac{1}{\cosh\left(\frac{k\epsilon}{nm}\right) - \epsilon\sinh\left(\frac{k\epsilon}{nm}\right)} \le \frac{1}{1-\frac{k\epsilon^2}{nm}} \le 1+\frac{2k\epsilon^2}{nm}. \end{aligned}$$

因此，互信息的贡献可界定为

$$\mathbf{Pr}[a=1 \mid X=x_{\min}(l)]\left(1-\frac{\mathbf{Pr}[a=1 \mid X=x_{\max}(l)]}{\mathbf{Pr}[a=1 \mid X=x_{\min}(l)]}\right)^2 \le \mathcal{O}\left(\frac{k^2\epsilon^4}{n^2m^2}\right). \tag{143}$$

对于 $2 \le l < 2/\epsilon$，利用不等式 双曲余弦$(u) \le 1+u^2$ 对所有 $u \in [0,1]$，我们可以将 (141) 中的比值上界如下：

$$\frac{\mathbf{Pr}[a=l \mid X=1]}{\mathbf{Pr}[a=l \mid X=0]} \le \left(1+\left(\frac{k\epsilon}{nm}\right)^2\right)\left(\sum_{i=0}^{\lfloor l/2\rfloor}\binom{l}{2i}\epsilon^{2i}\right) \tag{144}$$

$$= 1+\epsilon^2\sum_{i=1}^{\lfloor l/2\rfloor}\binom{l}{2i}\epsilon^{2i-2} + \left(\frac{k\epsilon}{nm}\right)^2\sum_{i=0}^{\lfloor l/2\rfloor}\binom{l}{2i}\epsilon^{2i} \tag{145}$$

现在，我们将 (145) 右侧的每一项分别上界如下：

$$\sum_{i=1}^{\lfloor l/2\rfloor}\binom{l}{2i}\epsilon^{2i-2}\le\sum_{i=1}^{\lfloor l/2\rfloor}\left(\frac{el}{2i}\right)^{2i}\epsilon^{2i-2}=l^2\sum_{i=1}^{\lfloor l/2\rfloor}\left(\frac{el\epsilon}{2i}\right)^{2i-2}(\frac{e}{2i})^2\le l^2\sum_{i=1}^{\infty}\left(\frac{e}{i}\right)^{2i}\le 12l^2. \tag{146}$$

第一个不等式由标准的斯特林型界 $\binom{n}{m}\le\binom{en}{m}^m$ 得出。第二个不等式利用了我们处于 $l<2/\epsilon$ 这一情形的假设，由此可知 $\frac{el\epsilon}{2i}\le\frac{e}{i}$。最后一个不等式成立，是因为该级数收敛到一个有限常数（事实上小于 12）；这可直接由根值判别法得出。

$$\sum_{i=0}^{\lfloor l/2\rfloor}\binom{l}{2i}\epsilon^{2i}=\frac{(1+\epsilon)^l+(1-\epsilon)^l}{2}\le(1+\epsilon)^l\le e^{l\epsilon}\le e^2 \tag{147}$$

等式由 $(1+\epsilon)^l$ 和 $(1-\epsilon)^l$ 的二项式展开得到，其中奇数次项因符号相反而相互抵消，只剩下偶数次项。第一个不等式成立，因为 $(1-\epsilon)^l\le(1+\epsilon)^l$。第二个不等式利用了标准界 $1+u\le e^u$。最后一个不等式由以下事实得出：$l<2/\epsilon$。

将上述界代入（145）后，我们得到：

$$\begin{aligned}\frac{\mathbf{Pr}[a=l\mid X=1]}{\mathbf{Pr}[a=l\mid X=0]}&\le 1+\epsilon^2\sum_{i=1}^{\lfloor l/2\rfloor}\binom{l}{2i}\epsilon^{2i-2}+\left(\frac{k\epsilon}{nm}\right)^2\sum_{i=0}^{\lfloor l/2\rfloor}\binom{l}{2i}\epsilon^{2i}\\&\le 1+\epsilon^2\left(12l^2+e^2\left(\frac{k}{nm}\right)^2\right)\\&=1+\mathcal{O}(\epsilon^2l^2)\end{aligned} \tag{148}$$

接下来，我们通过计算均值为 $\lambda=\frac{k}{nm}$ 的泊松随机变量的四阶矩，来对这部分对互信息的贡献进行上界估计。该计算利用了以下事实：对于 $l\ge 2$, l^4 $l\ l-1\ \ l-2\ \ l-3\ +l\ l-1$ — $1\le\frac{l^4}{l\ l-1\ \ l-2\ \ l-3\ +l\ l-1}\le$ 与 ()()()() 的比值有一个常数上界，具体为 ()()()()：

$$\begin{aligned}&\sum_{l=2}^{2/\epsilon}\mathbf{Pr}[a=l\mid X=x_{\min}(l)]\left(1-\frac{\mathbf{Pr}[a=l\mid X=x_{\max}(l)]}{\mathbf{Pr}[a=l\mid X=x_{\min}(l)]}\right)^2\\&\le\sum_{l=2}^{2/\epsilon}\frac{e^{-\lambda}\lambda^l}{l!}\epsilon^4l^4\\&\le 15\epsilon^4\sum_{l=2}^{2/\epsilon}\frac{e^{-\lambda}\lambda^l}{l!}\left(l(l-1)(l-2)(l-3)+l(l-1)\right)\\&\le 15\epsilon^4\left(\lambda^4+\lambda^2\right)\\&=\mathcal{O}\left(\left(\frac{k}{nm}\right)^2\epsilon^4\right).\end{aligned} \tag{149}$$

对于 $l\ge 2/\epsilon$，不难证明，式（141）中的比值有如下上界：$(1+\epsilon)^l e^{-\lambda\epsilon}$ 其中$\lambda=\frac{k}{nm}$。接着，对泊松分布的尾部应用切尔诺夫界，并利用以下事实，即 $x^2e^{1-x}\le x$ 对所有

$x \in [0,1]$:

$$\sum_{l \geq 2/\epsilon} \mathbf{Pr}[a = l \mid X = x_{\min}(l)] \left(1 - \frac{\mathbf{Pr}[a = l \mid X = x_{\max}(l)]}{\mathbf{Pr}[a = l \mid X = x_{\min}(l)]}\right)^2 \tag{150}$$

$$\leq \sum_{l \geq 2/\epsilon} \frac{e^{-\lambda}\lambda^l}{l!} \left(1 - (1+\epsilon)^l e^{-\lambda\epsilon}\right)^2$$

$$\leq \sum_{l \geq 2/\epsilon} \frac{e^{-\lambda}\lambda^l}{l!} \left(1 + (1+\epsilon)^{2l} e^{-2\lambda\epsilon}\right) \tag{151}$$

$$= \mathbf{Pr}[\mathrm{Poi}(\lambda) \geq 2/\epsilon] + e^{\lambda\epsilon^2} \mathbf{Pr}[\mathrm{Poi}\left(\lambda(1+\epsilon)^2\right) \geq 2/\epsilon]$$

$$\leq \left(\frac{\lambda}{2/\epsilon}\right)^{2/\epsilon} e^{2/\epsilon - \lambda} + e^{\lambda\epsilon^2} \left(\frac{\lambda(1+\epsilon)^2}{2/\epsilon}\right)^{2/\epsilon} e^{2/\epsilon - \lambda(1+\epsilon)^2}$$

$$= \left(\frac{e}{4}(\lambda\epsilon)^2 e^{1-\lambda\epsilon}\right)^{1/\epsilon} + \left(\frac{e^{1-2\lambda\epsilon^2}(1+\epsilon)^4}{4}(\lambda\epsilon)^2 e^{1-\lambda\epsilon}\right)^{1/\epsilon}$$

$$\leq (\lambda\epsilon)^{1/\epsilon} + \left((1+\epsilon)^4 \lambda\epsilon\right)^{1/\epsilon} \tag{152}$$

当 ϵ 足够小时，即 $\epsilon \leq \frac{1}{4}$ 时，式（150）中的上界为 $O(\lambda^4\epsilon^4)$ 的量级。结合式（142）、（143）和（149）中的界，即可得到所声称的互信息上界，并完成证明。

□