arXiv:2304.01374v1 [cs.DS] 3 Apr 2023

奇偶迹下的分布检验

Renato Ferreira Pinto Jr.*滑铁卢大学
r4ferrei@uwaterloo.ca

Nathaniel Harms†洛桑联邦理工学院
nathaniel.harms@epfl.ch

摘要

分布测试是一项基础统计任务，有着诸多应用，但我们关注的是一类问题：样本中的系统性错误标记使我们无法应用现有理论。为将分布测试应用于这些问题，我们引入奇偶迹下的分布检验，在这种设定下，算法接收一个有序样本 S，其中每个元素仅有最低有效位可供观察。这一抽象揭示了以下三个重要问题之间的联系，并由此得到新的上界和下界：

1. 在混淆收集器下的分布测试中，样本收集器可能无法区分定义域中彼此接近的元素（例如，机器学习分类器）。我们证明了定义域具有环或路径结构时，混淆收集器下分布测试的界。
2. 关于基础问题检验与学习的近期研究通过从分布测试进行归约，建立了分布无关的、基于样本的性质测试的紧下界，但其紧致性仅适用于对称性质。奇偶迹使更广泛的一类等价关系能够应用于非对称性质，同时以不同的技术重现并加强了许多先前结果。3. 我们首次给出了在广受研究的迹重构模型下进行性质测试的结果。在该模型中，目标是仅根据字符串 x 的独立随机迹，检验未知字符串是否满足某种性质，或与满足该性质相距甚远。

我们的主要技术结果是：在奇偶迹下，对自然数集上的分布进行均匀性检验，其紧确界为 $\widetilde{\Theta}\left((n/\epsilon)^{4/5}+\sqrt{n/\epsilon^2}\right)$，并由此得到上述问题的相关结果。

*部分经费由NSERC加拿大研究生奖学金博士奖资助。
†部分经费由NSERC博士后奖学金资助。作者在滑铁卢大学任职期间完成了这项工作的很大一部分。

目录

1 引言 **1**

1.1 奇偶迹下的分布检验.2 1.2 混淆收集器下的分布检验.3 1.3 分布无关的基于样本的性质测试.5 1.4 轨迹重建的属性测试.7 1.5 证明概述.9 1.6 讨论与开放问题.10

2 预备知识与上界的统一框架 11

2.1 记号.11 2.2 混淆收集器：定义与术语.11 2.3 奇偶迹：定义与术语.12 2.4 路径结构和环结构的泊松随机变量.13 2.5 共享分析.15

3 混淆收集器模型中的均匀性测试 21

3.1 简单情形：高度集中的分布.22 3.2 检验统计量的期望值.24 3.3 检验统计量的集中性.31 3.4 测试器的正确性.34

4 性质测试奇偶迹模型中的均匀性测试 el 36

4.1 小 ϵ 情形.37 4.2 简单情形：不平衡且高度集中的分布.37 4.3 检验统计量的期望值.39 4.4 检验统计量的集中性.51 4.5 针对大规模 ϵ的测试器正确性.57

5 奇偶迹模型中均匀性检验的下界 59

5.1 论证概要.60 5.2 YES 和 NO 分布的构造.62 5.3 部分指纹及其概率.63 5.4 YES 和 NO 多米诺骨牌的表现相似.64 5.5 信息界.67

6 分布无 基于样本的性质测试74

6.1 带标签分布.75 6.2 编辑距离.76 6.3 带标签分布测试.77 6.4 通过学习进行测试.79 6.5 密度性质与奇偶迹下的分布检验.81 6.6 均匀测试 k-交替函数.86 6.7 无承诺测试 k-交替与均匀 k-交替函数.87 6.8 分布无关的基于样本的性质测试.88

7 迹重构模型中的性质测试 91

7.1 单迹上界.93 7.2 多迹上界.97 7.3 下界.99

参考资料 103

泊松化与增强108 第3节缺失的证明109 第4节缺失的证明110 编辑距离证明113 带标签分布测试与奇偶迹的比较128

1 引言

仅使用随机样本来对未知概率分布 $\mathcal{D}$ 作出判断，是一种基本的统计任务。判断 $\mathcal{D}$ 是否满足某种性质，或者是否与所有满足该性质的分布都相距很远（依据某种距离度量），正是分布测试算法的目的。分布测试本身已有深入研究，也很有趣，并且还有许多实用应用。但我们关注的是一些问题：数据的系统性错误标记使我们无法应用现有理论。因此，我们定义奇偶迹下的分布测试，以帮助理解这些问题。在定义这一抽象概念之前，我们先来解释这些问题。

1. 混淆收集器下的分布检验。我们希望仅使用来自 $\mathcal{D}$ 的随机样本 S，对定义域为 $\mathcal{X}$ 的未知分布 $\mathcal{D}$ 作出判断。不幸的是，S 已由一个无法区分 $\mathcal{X}$ 中某些元素的实体收集或标记。也许我们的林地植物样本由一名无法区分黑云杉和白云杉，或红枫和糖枫的研究助理制成表格，并错误地将所有云杉和所有枫树分别合并计数[1]。或者，样本由机器学习分类器进行标记；对于每一对元素 x、$y \in \mathcal{X}$，它都有一定概率未学会将 x 与 y 区分开，而是将 x 和 y 的所有样本混在一起。或者，样本标签经过某个函数哈希处理，而该函数在 $\mathcal{X}$ 中相近的元素之间引入了碰撞。又或者，我们希望了解化石按年份的分布，但无法区分来自年份 x 的化石和来自年份 y 的化石，除非某个随机地质事件在 x 年与 y 年之间的岩石中留下痕迹。近期学习理论研究指出，这类问题在应用文献中很常见，但从理论上对此知之甚少 [FKKT21]。我们为这类问题引入一个模型，称为混淆收集器下的分布检验。

2. 分布无关的基于样本的性质测试。[GGR98] 的检验与学习问题是性质测试中的基本问题之一。它探讨哪些性质能够以比学习更高的效率进行测试。分布无关的基于样本的性质测试是与标准PAC学习相对应的性质测试模型，因此，理解该模型中的检验与学习问题，对于性质测试的许多标准动机至关重要[GGR98]。近期有关检验与学习的进展利用与分布测试的联系，获得了新的上界 [GR16] 和下界 [ES20, BFH21, CP22]，推进了性质测试的发展；特别是，这些进展揭示了函数的对称性质与分布测试之间的等价关系 $[n] \to \{0,1\}$（即在 $[n]$上的排列下封闭的性质） [GR16, BFH21]。然而，这些技术尚无法回答诸如 k-交替函数和半空间等重要性质的检验与学习问题，因为这些性质并非对称性质，无法建立相同的等价关系。因此， [BFH21]的下界与PAC学习所给出的上界之间仍存在差距，需要新的技术来弥合。

奇偶迹下的分布测试克服了 [BFH21]的一些局限性，首次揭示了分布测试与对非对称性质进行测试之间的等价关系。我们用一种不同的技术恢复了 [BFH21] 中的许多下界；与检验和学习相比，这种技术能带来更强的结论。此外，我们还为联合函数-分布性质的分布无关、基于样本的检验获得了新的紧致正面结果，为分布无关的基于样本的检验增添了少数几个正面结果： [GR16、RR20、RR21、BFH21]。

[1]感谢生态学家Julie Messier教授提供这些例子，这些物种很容易被学生混淆。

3. 轨迹重建的性质测试。轨迹重建是一个优美的问题，最早由 [Lev01, BKKM04] 提出。与我们相关的是删除信道下的轨迹重建，近期这一问题受到了广泛关注（例如 [HMPW08, DOS17, NP17, HL20, KMMP21, GSZ22, CDL+22, Rub22, CDL+23]）。问题如下：给定一个字符串 $x \in \{0,1\}^N$ 和一个删除率 $\delta \in (0,1)$。对 x 中的每个字符，都以概率 δ 相互独立地将其删除，从而得到一个子字符串，这样便可生成一条随机轨迹。算法获得一系列相互独立的轨迹，并须尽可能少地使用这些轨迹来重构字符串 x。这一问题通常以计算生物学为背景，其中删除信道是生物学家获取DNA字符串的典型方式的一种简化模型：他们无法获得“真实”的DNA，只能得到若干份受损副本。关于生物学应用的综述，参见 [BPRS20]。

迹重构的分析一直非常困难，已知的最佳下界 $\widetilde{\Omega}(N^{3/2})$ [Cha21a] 与上界 $\exp(\widetilde{O}(N^{1/5}))$ [Cha21b] 之间存在巨大差距（其中隐藏常数取决于 δ）。然而，如果目标只是对未知字符串 x作出判断，那么完全进行 重构 可能并无必要。我们提出在迹重构模型中进行 性质测试；据我们所知，该模型尚未得到研究。目标很简单：仅根据 x 的迹作出判断，而不将 x 完全重构。就标准生物学应用而言，我们希望根据若干个受损副本对一条 DNA 字符串作出判断，而不对其进行重构。我们证明了迹重构中首批非平凡的性质测试结果，这些结果源于该问题与奇偶迹下的分布测试之间的等价性。

本文。 标准分布测试算法根据直方图作出判断，直方图统计样本中定义域的每个元素出现的次数。上述每个问题的共同挑战在于，要应用分布测试，测试器需要处理样本的一种结构化错标记，而这种错标记会阻止测试器构造直方图。奇偶迹下的分布测试是对这一挑战的抽象。本文旨在将该模型与上述问题联系起来，并通过证明最基本的分布测试任务一致性测试的紧界，开始理解这一模型。与标准模型相比，这一问题的分析难度显著更高；我们认为，在推进到后文将讨论的某些更困难问题之前，有必要先完成这一分析。

1.1 奇偶迹下的分布检验

现在让我们定义奇偶迹下的分布检验。设 Π 是 $\mathbb{N}$ 上概率分布的一个性质（即集合）。与标准分布检验一样，对于 $\mathbb{N}$ 上的分布 π，奇偶迹下的分布测试器必须接受任意输入 $\pi \in \Pi$（概率为 $2/3$），并拒绝任何与 Π 相距至少 ϵ 的输入 π（概率为 $2/3$），也就是说，该输入与任意 $\pi' \in \Pi$ 的距离至少为 ϵ。（标准分布检验通常使用全变差距离，但我们将看到，在这种情况下它并非自然的选择。）测试器接收的不是来自分布 π 的样本 S，而是 S 的 奇偶迹，记为 $\mathsf{trace}(S)$，定义如下。对于任意大小为 m 的多重集 $S \subset \mathbb{N}$，将 $S = \{x_1, x_2, \ldots, x_m\}$ 按 $x_1 \le x_2 \le \cdots \le x_m$ 排序，并将 $\mathsf{par}(x)$ 定义为 $= (x \mod 2)$，即 x 的奇偶性。然后

$$\mathsf{trace}(S) := (\mathsf{par}(x_1), \mathsf{par}(x_2), \ldots, \mathsf{par}(x_m)) .$$

例如，对于样本 $S = \{5,1,6,2,4,2\}$，算法收到迹$(S) = 100010$，它是 (1，2，2，4，5，6) 的奇偶性字符串。请注意，例如，当 n 为偶数时，在奇偶迹下，$\{1,\ldots,n\}$ 上的均匀分布与 $\{n+1,\ldots,2n\}$ 上的均匀分布不可区分，尽管它们的全变差距离为 1，因此，哪些分布检验任务在奇偶迹下可行并非显而易见。

均匀性检验。为了开始理解奇偶迹，我们来考虑均匀性检验问题（例如，[GR00、Pan08、ADK15、DKN15b], 参见 [Can22]）。目标是接受 $[2n]$ 上的均匀分布，并拒绝那些与均匀分布相距 ϵ 的 $[2n]$ 上的分布。可以看出，在奇偶迹下进行均匀性检验确实可行，即使以全变差距离衡量也是如此，这可以通过优惠券收集器论证得到。收到大小为 $\Theta(n\log n)$ 的迹后，以高概率，要么迹包含定义域 $[2n]$中的每个元素，要么可以安全地拒绝。如果迹包含定义域中的每个元素，算法就能推断出每个样本点的确切身份，并模拟标准分布检验器，从而得到 $O(n\log n+\sqrt{n}/\epsilon^2)$ 的界；这由标准模型下紧确的 $\Theta(\sqrt{n}/\epsilon^2)$ 界推出 [Pan08、VV17a、DGPP18、DGPP19]。

样本量为 $o(n)$ 是否足够，并非一目了然。本文的主要技术贡献是为这个问题建立紧界：确实可以实现次线性样本量，但标准模型与奇偶迹模型之间存在差距。我们将在 第 1.2 节 和第 1.5 节 讨论证明，因为混淆收集器模型将作为预热。

定理 1.1 (非正式表述；参见定理 4.1)。 在奇偶迹下，对定义域 $[2n]$上的分布进行均匀性检验，若以全变差距离衡量，则需要的样本量为 $\widetilde{\Theta}\left(\left(\frac{n}{\epsilon}\right)^{4/5}+\frac{\sqrt{n}}{\epsilon^2}\right)$。

编辑距离。 定理 1.1 使用全变差距离，但这并非总是可行。两个分布的全变差距离可能为 1，却在奇偶迹下不可区分，因此全变差距离并非最自然的度量。为了将奇偶迹模型与本文讨论的其他问题联系起来，我们需要一种新的度量。我们定义分布的编辑距离伪度量；它具有这样一种理想性质：当且仅当编辑距离为 0 时，两个分布 π 和 π' 在奇偶迹下不可区分。我们将分布 π 在 $\mathbb{N}$ 上视为交替的“分数字符串”，

$$1^{\pi(1)}0^{\pi(2)}1^{\pi(3)}\cdots 1^{\pi(2i-1)}0^{\pi(2i)}\cdots$$

其中 b^p 表示 b 重复 p 次（该次数可以是分数）。然后，通过从 m中 π 采样 m 个相互独立的随机字符，其概率与它们的分数重复次数成正比 p，并按顺序将它们连接起来，即可得到大小为 m的 π 奇偶迹。奇偶迹的分布在某些“免费”编辑操作下保持不变，例如 $b^p\to b^{p/2}b^{p/2}$、$a^pc^q\to a^pb^0c^q$或 $a^pb^0c^q\to a^pc^q$；而其他“昂贵”的编辑操作，例如 $b^pc^q\to b^{p-\delta}c^{q+\delta}$，可能会改变迹的分布。编辑距离是将一个分布转换为另一个分布所需的代价；参见 定义 6.10。

1.2 使用混淆收集器进行分布性质测试

我们引入混淆收集器 来建模分布测试问题：算法收到的随机样本 S 被系统性地错误标记；回想一下第一页中的例子。为了形式化这个问题，设想定义域中的任意两个元素 x 和 y，在样本 S 中，x 和 y 的所有出现项都可能以某种概率被合并并计数在一起。这些合并必须具有传递性，因此，元素对 (x,y)、(y,z) 或 (x,z) 被合并的概率并非相互独立。这意味着必须有某种结构化随机过程来合并定义域元素，我们选择将其建模如下。

设 p 是定义在（有限）结构化定义域 V上的一个分布，其结构由“基图” $G=(V,E)$给出。例如，G 可以是一棵表示某个动物群分类体系的树。分布测试算法有一个参数 $\eta\in(0,1]$ 称为分辨率（表示分类器的准确度），并接收一个大小为 m 的随机样本 S，其生成方式如下。首先，以概率

$1-\eta$，并令 C_1、……、C_t 为其连通分量。对每个 C_i，令 $c_i \in C_i$ 为该连通分量的任意代表元。然后从 p 中采样一个包含 m 个独立点的集合 S'，并用其所属连通分量的代表元 c_i 标记每个 $s \in S'$。所得样本 S 将提供给算法。注意，在分辨率为 $\eta = 1$时，图 H 是一个独立集，且每个元素都带有正确的标签，因此该模型就成为标准的分布测试模型。

给定一个关于分布的性质 Π，其定义域为 V，参数为 ϵ，对于 Π 的测试器，其分辨率为 η，样本复杂度为 m，必须接受（概率为[2] 2/3）任意属于 Π 的分布，并拒绝（概率为 2/3）所有与 ϵ 中任意分布的全变差距离至少为 Π 的分布。

结果。当基图 G 是环或路径时，我们得到了相应结果。这涵盖了定义域为 $[n]$ 、且只有当随机“分隔符”出现在两个元素之间时，这些定义域元素才能被区分的情形（例如首页给出的化石示例，或样本标签经过单调哈希函数[3]随机哈希的情形）。

定理 1.2（非正式表述；参见 定理 3.23。）。设 $G = (V, E)$ 是一个有 n 个顶点的路径或环，设$\epsilon \geq \tilde{\Omega}(n^{-1/4})$ 且 $\eta \geq \tilde{\Omega}(n^{-1/5}\epsilon^{-4/5})$。那么，均匀性检验需要 $\widetilde{O}\left(\frac{\sqrt{n}}{\epsilon^2\eta^{3/2}}\right)$ 个样本。

这在具有分辨率 $\eta = 1$ 的均匀性检验的最优$\tilde{\Theta}(\sqrt{n}/\epsilon^2)$ 界与当 η 小到定理所允许的程度时的 $\widetilde{O}\left((n/\epsilon)^{4/5}\right)$ [4] 之间进行插值。

技巧。我们将在定理 1.1和1.2 中更详细地介绍我们的方法，见第 1.5 节，但这里先简要说明主要思路。定理 1.2可视为定理 1.1的预热版本；证明后者时，我们考虑了一个类似的加权环上的问题，尽管混淆收集器在处理分辨率参数时也带来了独特的挑战 η。

令 $\boldsymbol{T}_i$ 表示元素 i 在样本中的重数。一个标准的简化是假设 $\boldsymbol{T}_i$ 服从相互独立的泊松分布Poi($m \cdot p(i)$)。随机图 $\boldsymbol{H}$ 使观测变量之间产生依赖关系；我们用 $\boldsymbol{\Phi}$ 表示描述这些依赖关系的随机布尔矩阵，其中$\boldsymbol{\Phi}_{i,\ j} = 1$ 当且仅当顶点 i，j 属于同一个连通分量。我们的证明归结为对随机二次型 $\boldsymbol{T}^\top \boldsymbol{\Phi} \boldsymbol{T}$的分析。关于二次型的集中界 $\boldsymbol{X}^\top A \boldsymbol{X}$ 已有研究（包括 Hanson-Wright 型不等式，参见例如[GSS21]），但我们尚未发现适用于矩阵 A 本身也是随机变量时的界；我们所需类型的不等式或许也具有独立研究价值。更多详情见第 1.5 节。

相关工作。近期机器学习和分布测试文献中，研究者出于类似动机相互独立地提出了相似的模型。关于应用领域的讨论，参见 [FKKT21]及其中的参考文献。 [FKKT21]提出了一个与我们不同的模型：定义域的分割更为一般，但对每个样本点都会重新独立采样[5]，，并研究学习问题。在分布测试领域， [GR22、CFG+22]研究了一种样本包含“巨型对象”的模型，这些对象本身支持查询，因此算法同样无法获得直方图，而必须对样本进行查询。 [CW21]研究了可保证为排列的系统性错误标记。在 [CW20], 目标是检验是否存在

[2]与标准测试器不同，我们不能仅仅通过重复运行测试器来提高成功概率，因为成功概率部分取决于分辨率 η。[3]单调哈希函数是指保持键的顺序不变的哈希函数，参见例如

[4] [23] AFCK请注意，定理中对 η 的下界是必要的。例如，作者标注的林地动物群样本，其分辨率为 $\eta = 0$，无论样本量多大，都无法据此做出任何判断。[5]简而言之，造成差异的原因是 [FKKT21] 假设样本点可能由具有不同分类的不同实体标注，而我们假设样本点由一个分类不完善的实体标注。

一个由区间构成的分割，使输入分布 p等于参考分布q。其他样本信息不完美的模型还包括局部隐私检验[GR18、She18、ACFT19] 以及信息约束下的推断 [ACT19、ACT20、ACF+21]。

1.3 分布无关的基于样本的性质测试

我们关注一个基础的检验与学习问题 [GGR98], ，尤其是与标准 PAC 学习相对应的分布无关的基于样本的性质测试模型。这对于性质测试的一些拟议应用至关重要，例如模型选择（即为学习选择一个合适的假设类 $\mathcal{H}$ ） [GGR98]。我们认为，性质测试与分布测试之间的正式联系对于理解检验与学习问题至关重要；这些联系最早由 [GR16], 阐述，但他们的结果仅适用于对称函数性质（即在定义域上的排列作用下封闭的性质）。

正如 [GGR98、BFH21], 所指出的，检验与学习本质上就是检验与 VC 维数的问题，因为对一个假设类 $\mathcal{H}$ 进行 PAC 学习所需的样本量（忽略误差 ϵ）为 $\Theta(\mathrm{VC})$，其中 VC 是 $\mathcal{H}$的 VC 维数。因此，目标是确定哪些类 $\mathcal{H}$ 可以使用 $o(\mathrm{VC})$ 个样本进行测试。对于许多重要的假设类，包括 $\mathbb{R}^n$上的半空间以及 $\mathbb{R}$上的 k-交替函数， [BFH21] 通过定义“下 VC 维数”，并利用它构造一个从支持集大小区分问题的归约（见 [RRSS09、VV11、WY19]），证明了下界 $\Omega\left(\mathrm{VC}\log\mathrm{VC}\right)$。该问题要求判断 $[n]$ 上的一个分布的支持集大小至多为 αn还是至少为 βn。在某些情况下，这个界是紧的，这是因为 [GR16] 针对某些对称性质给出了一个 $O\left(\mathrm{VC}\log\mathrm{VC}\right)$ 界，并将问题反向归约为测试支持集大小。

这使得半空间等许多最重要的假设类在测试与学习所需样本量之间存在差距。与 [GR16], 一样，我们认为 $\Omega\left(\mathrm{VC}\log\mathrm{VC}\right)$与 $O(\mathrm{VC})$ 之间的差距意义重大；首先，因为测试能否以低于 PAC 学习所需样本量的次线性 样本量完成，仍是一个悬而未决的问题；其次，因为这与分布测试，尤其是支持集大小估计，密切相关。在这些问题中，对数因子令人意外且十分重要 [RRSS09、VV11、WY19]。遗憾的是， [BFH21] 的方法无法弥合这一差距，因为粗略地说，分布测试与性质测试之间关系的紧密程度，在对称性质处已达到极限。

我们的目标是建立一种更强的分布测试与性质测试之间的关系，从而突破这一局限。在奇偶迹下进行分布测试是朝这一目标迈出的一步。考虑 k-交替函数这一（非对称）性质；这类函数 $\mathbb{R}\to\{0,1\}$ 在 0 和 1 之间交替的次数至多为 k （等价地，这是由k 个区间构成的并集类），该性质已在 [KR00、Nee14、BBBY12、BH18、CGG+19、BFH21], 中得到研究，但检验与学习之间的关系仍是开放问题。下面给出我们的方法的第一个例子：

定理 *1.3*（非正式表述；参见定理 *6.38*）。设 $m_1(k,\epsilon)$ 是在奇偶迹下，检验某个分布的支撑集大小是否为 k，或其与支撑集大小为 k 的分布之间的编辑距离是否至少为 ϵ 所需的样本量。设 $m_2(k,\epsilon)$ 是在分布无关的基于样本模型中，检验某个函数是否为 k-交替函数所需的样本量。那么 $m_1(k,\epsilon)=\Theta(m_2(k,\epsilon))$.

这是针对非对称性质的分布测试与性质测试之间第一个紧确关系；它只是分布测试与密度性质测试之间更一般等价关系的一个特例，详见下文。要在迹重构模型中得到我们的结果，就需要这一等价关系。编辑距离的出现凸显了它对于奇偶迹应用的重要性。作者们对正确的样本量 $m_1(k,\epsilon)$

在该定理中应为[6], 持何种看法；我们将在未来的工作中研究这一问题。本文聚焦于更简单的均匀性检验问题。即便如此，与标准模型相比，在奇偶迹模型中对这一问题进行分析仍然困难得多。不过，即使不知道 $m_1(k,\ \epsilon)$，我们仍利用定理 1.3 以不同的证明方法恢复了 [BFH21] 中的许多界，而且该方法对检验与学习问题具有更强的推论。我们给出了关于 k-交替函数和半空间的界，同时也恢复了关于半空间交和决策树的界[7]。（参见 [MORS10、BBBY12、Har19、CP22]，了解其他关于半空间性质测试的早期工作。）

定理 1.4 (参见 [BFH21])。对定义域 $\mathbb{R}$ 上的 k-交替函数进行分布无关的基于样本的检验需要$\Omega(\frac{k}{\epsilon}\log k)$个样本；对定义域 $\mathbb{R}^n$ 上的半空间进行检验需要$\Omega(\frac{n}{\epsilon}\log n)$个样本。 ——

与以下技术不同， [BFH21], 我们的技术有可能为检验与学习 k交替函数、半空间及其他对象带来更好的答案。关于奇偶迹下分布测试的更好下界，将意味着针对 k交替函数、半空间、半空间交和决策树的更好下界。另一方面，关于（例如）半空间检验的 $o(\mathsf{VC})$上界 将意味着类似于令人惊讶的 $o(n)$ 界的结果[VV11，WY19] 用于区分支撑集大小 $\leq \alpha n$ 与 $\geq n$，而且这一结果 即使在奇偶迹下也成立，此时测试器并不知道样本点的身份。

为了阐明分布测试与分布无关的基于样本的性质测试之间的联系，我们将分布无关的性质测试的研究范围扩展到允许检验带标签分布的性质。上的带标签分布是一个二元组（f，$\mathcal{D}$），其中函数为 f，分布为 $\mathcal{D}$。其思想是，我们可能不仅想检验函数 f的某种性质，还想检验函数 f 与概率分布 $\mathcal{D}$的联合性质。（我们也请读者参阅 [RV23、GKK23]中另一类有趣的函数与分布联合检验。）

例如，我们可能不仅想检验 f 是否为 k-交替函数，还想检验它是否将 $\mathcal{D}$ 均匀划分为均匀单色区间。我们称这类函数为均匀 k-交替函数，并对均匀 $2k$-交替函数的测试得到了紧确结果（假设输入保证是 $2k$ 交替的）。这为分布无关的基于样本的测试中为数不多的正面结果增添了新例 [GR16、BFH21、RR20、RR21]。

定理 1.5 (非正式表述；见 定理 6.33)。设 f 是 $2k$-交替的。那么，要检验它是否相对于全变差距离均匀地 $2k$-交替，需要 $\widetilde{\Theta}((k/\epsilon)^{4/5}+\sqrt{k}/\epsilon^2)$ 个样本。

(对于这个问题更具挑战性的非承诺版本，我们通过为带标签分布定义一种合适的“通过学习进行测试”归约，并使用 [VV17a]的容错均匀性检验器，得到 引理 6.35所述的 $O(\frac{k}{\epsilon}+\epsilon^2\frac{k}{\log k})$ 界；见此引理。）定理 1.3 和 1.5的证明利用了与奇偶迹下分布测试之间的一种等价关系。对于一种自然的带标签分布类，这种等价关系普遍成立，我们称之为 密度性质。

密度性质。 非正式地说[8], 每个布尔函数 $f:\ \mathbb{R}\to\{0,1\}$ 都有唯一的交替点 集合 $\mathbb{R}$，在这些点处，函数值会从 0 变为 1，或从 1 变为 0。密度性质 是带标签分布的集合，其中（f，$\mathcal{D}$）是否属于该集合，取决于其密度序列：概率质量序列 $\mathcal{D}(a,\ b]$ 其中 a，b 是 f的连续交替点。

[6]事实上，即使在 *standard* 模型中，这一点似乎也尚不为人所知：我们能找到的最好上界是$O\left(\frac{n}{\epsilon^2\log n}\right)$和 $O(n/\epsilon)$，而下界为 $\Omega\left(\frac{n}{\epsilon\log n}\right)$。[7]我们的表述还包含对 ϵ的依赖，而 [BFH21] 没有这一项。我们的结果可由分布测试问题的界推出，但也可以用标准方法将对 ϵ的依赖补充到 [BFH21] 的结果中（如 [ES20]所示）。[8]为简洁起见，本讨论忽略了交替点可能有无穷多个的情况。

k-交替函数和均匀 k-交替函数都可以定义为密度性质，但还有许多其他有趣的例子。测试密度性质的困难在于，测试器不知道某个样本属于哪个交替点区间。给定两个样本点 x、$y \in \mathbb{R}$，测试器不知道 x、y属于同一个区间还是不同区间，除非 $f(x) \neq fy$，或者 $f(x) = fy$，并且在 x和 y之间还有另一个样本点 z，使得$f(z) \neq fx$。以往的工作使用查询来克服这一困难 [CGG+19]，但这在基于样本的模型中是不可能的。

奇偶迹下的分布测试刻画了这一困难：测试密度性质本质上等价于在奇偶迹下测试分布。对于任意密度性质Ξ，令 $\Pi(\Xi)$ 表示定义 Ξ 的密度序列（即概率分布）集合。利用拉姆齐理论（受 [Fis04、DKN15a], 的启发，另见 [CW20]），我们证明：

引理 1.6 （非正式；参见 引理6.26）。 在带标签分布模型中对 Ξ 进行性质测试（相对于 编辑距离的适当类比）等价于 $\Pi(\Xi)$ 在奇偶迹下相对于编辑距离进行性质测试。

技术。 这里的关键贡献是编辑距离和奇偶迹的定义，它们借助拉姆齐理论在引理 1.6中的应用，使得基于样本的性质测试能够与分布测试联系起来。本节的主要结果（包括恢复 [BFH21] 的结果）随后通过一些归约得出，这些归约主要依赖于编辑距离的性质。

1.4 迹重构的性质测试

现在我们转而讨论轨迹重建的属性测试。与本文中的其他问题相比，我们对这个问题的兴趣是独立产生的，而本文提出的形式化联系则出乎意料。在轨迹重建问题中（采用删除信道），给定一个字符串 $x \in \{0,1\}^N$ 和删除率 $\delta \in (0, 1)$。通过以概率 δ 相互独立地删除 x 的每个字符，并取所得的子字符串，可以得到 x 的一条迹 $\boldsymbol{T}$。例如， $x = 110011001100$的一条迹可能是 11110000 或 101010。目标是尽可能少地使用独立轨迹来重建 x （参见上文的参考文献）。

我们希望在不 x 完全重建它的情况下作出判断，因此我们提出在迹重建模型中进行性质测试。对于字符串的某个性质 Ψ $\{0,1\}^N$，算法应当接受（概率为 $2/3$）字符串 $x \in \Psi$，并拒绝（概率为 $2/3$）那些远离该性质的 Ψ字符串。自然的度量选择是字符串上的（相对[9]）编辑距离，这是近似迹重建 [CDL+22]的标准选择。字符串上的编辑距离与我们对概率分布的编辑距离的定义密切相关。

为了衡量迹测试器的复杂度，我们同时考察迹的数量以及每条迹的期望大小。轨迹重建通常在常数删除率 δ下进行，对应的迹的期望大小为 $\Theta(N)$。对于性质测试，我们希望允许极高的删除率，使迹的期望大小为 $\rho N = o(N$，其中 $\rho = 1 - \delta$ 是保留率（这使我们的研究与近期独立研究[CDL+23] 在轨迹重建中处于相同的低保留率范围）。这与传统的性质测试问题一致：其目标是在仅观察到输入中少于一个常数比例的内容时作出判断。我们将此问题与奇偶迹下的分布测试和带标签分布测试联系起来，并给出三个轨迹测试问题的结果。这似乎是此类问题的首批研究成果（先前工作中与之最相近的问题，是区分两个任意字符串x，y 它们在汉明距离或编辑距离意义下彼此接近 [GSZ22]）。

[9]两个长度为 N 的字符串之间的相对编辑距离是编辑距离的 $\frac{1}{N}$ 倍。

结果。为开启轨迹重建的性质测试研究，我们证明了关于字符串三种基本性质测试的界。对于 $n \in \mathbb{N}$，我们称 $x \in \{0,1\}^N$ 为 n-块字符串，如果 x由至多 n 个连续的块组成，其中一个块是全为 1 或全为 0 的（极大）字符串。均匀 n-块字符串是指由 n 个长度相等的块组成的字符串。我们的结果如下：

1. 测试一个 n-块字符串是否为均匀分布的 n-块字符串；2. 测试任意字符串是否为均匀分布的 n-块字符串；以及 3. 测试任意字符串是否为 n-块字符串。

这些结果利用了轨迹测试与奇偶迹下的分布测试之间的一般双向归约（引理 7.9 和 7.24）。直接应用我们的归约可得出使用¨单条迹进行性质测试的上界，这对应于单条迹近似重构问题；该问题的研究由同期独立开展的工作 [CDL+23]发起。我们的主要应用使用了一个额外技巧来应用该归约，从而得到适用于任意数量迹的界。注意，即使块数 n很大，例如 $n = \Theta(N$，我们仍然可以用一条次线性大小的迹（即删除率 $\delta = 1 - o(1)$）来测试该性质。

定理 1.7 （非正式表述；参见定理 7.19 和7.25）。假设 $x \in \{0,1\}^N$ 已知是一个 n-块字符串。对于任意 k，判断 x 是否为均匀分布的 n-块字符串，或在相对编辑距离下与均匀分布的 ϵ-块字符串相距 n，可以通过使用 k 条迹完成，其期望大小为 $\rho N = \widetilde{O}\left(\frac{n^{4/5}}{k^{1/5}\epsilon^{4/5}} + \frac{\sqrt{n}}{\sqrt{k}\epsilon^2}\right)$。与此同时，对于足够大的 N，必有 $k\rho N = \widetilde{\Omega}\left(\frac{n^{4/5}}{\epsilon^{4/5}} + \frac{\sqrt{n}}{\epsilon^2}\right)$。

我们发现，用每条迹的期望大小来衡量复杂度很方便；不过，也可以用更常见的迹重构术语重述我们的结果：对于固定的 k，如果$\rho_k \cdot N$ 是期望迹大小的界，那么对于所有保留率 $\rho \geq \rho_k$，测试所需的迹数至多为 k。增加迹的数量 k ，可以让测试器处理更低的保留率，但观测到的比特总数 $k\rho N$ 也会增加。

对于最后两个结果，我们尚未得到奇偶迹下相应分布测试问题的紧界，但借助已建立的理论，我们可以几乎“免费”地得到超越优惠券收集器论证的非平凡界。对于带标签分布测试模型（第1.3 节），我们通过定义一个“恰当的学习器与验证器对”，证明“通过学习进行测试”的归约成立；这一做法与 [GGR98]，的标准归约类似，只不过它使用分布测试任务，而非 [GGR98] 中的“验证步骤”。随后，我们利用引理 1.6以及它与轨迹测试之间的关系，得到一种适用于轨迹测试的一般性“通过学习进行测试”技术。

定理 1.8 (非正式表述；参见 定理 7.15)。 只需一条期望大小为 $\rho N = O\left(\frac{n}{\epsilon} + \frac{n \log n}{\epsilon^2}\right)$ 的迹，即可测试 $x \in \{0,1\}^N$ 是否为均匀分布的 n 块字符串，或与均匀分布的 n 块字符串的相对编辑距离至少为 ϵ。

定理 1.9 (非正式表述；参见 定理 7.13 和 7.26。)。 测试 $x \in \{0,1\}^N$ 是否为 n 块字符串，或与所有 n 块字符串的相对编辑距离至少为 ϵ，只需一条期望大小为 $\rho N = O(n/\epsilon)$ 的迹；而当 N 足够大时，任何使用 k 条迹的迹测试器都必须满足$k\rho N = \Omega(n/\log n)$。

定理 1.8 使用了 [VV17b] 中的均匀性容错测试器，作为通过学习进行测试的归约中的“验证步骤”，而 定理 1.9 使用了 $O(k/\epsilon)$ 对 k-交替函数进行性质测试的上界，该上界源自 VC 维数。我们认为这些界颇为神秘，因为我们针对轨迹测试的通过学习进行测试的归约经过非构造性的

拉姆齐理论 引理 1.6 论证，因此，轨迹测试器并不知道轨迹中字符的位置，却能从带标签分布学习器和验证器中以非构造性的方式得到；而这些学习器和验证器强烈依赖于知道样本点的绝对位置。

1.5 证明概述

我们简要介绍在奇偶迹下以及使用混淆收集器进行均匀性检验的证明，定理 1.1 和 1.2。

上界。 我们回顾一下标准均匀性检验器 [GR00, DGPP19]（另见 [Can22]）。设 p 为定义域 $[n]$ 上的输入分布。对于大小为 m的样本 S，设 X_i 为元素 i 在 S中的重数。测试器统计样本中的“碰撞”数量：计算 $Y := \frac{1}{(m-1)} \sum_{i=1}^{n} X_i(X_i-1)$，若该值过大则拒绝。此方法有效，是因为 $\mathbb{E}[\boldsymbol{Y}] = p^\top p = \|p\|_2^2$在 p 远离均匀分布时会很大。现在我们介绍混淆收集器的测试器。对于定义域 $\mathbb{Z}_n$ 上的输入分布 p（定义域中的元素是路径或环的顶点），我们采用标准的简化处理：元素 j 在样本中的出现次数服从参数为 mp_j的泊松分布 $\mathrm{Poi}(mp_j)$，且与其他元素相互独立。现在重新定义 $\boldsymbol{X}_i$，将其视为包含在 i^{th} 的连通分量中的样本点数量；测试器无法区分这些样本点：$\boldsymbol{X}_i$ 变量仍服从泊松分布，但它们并不相互独立。测试器按照标准算法计算“碰撞计数”：

$$\boldsymbol{Y} := \frac{1}{m}\sum_i \boldsymbol{X}_i(\boldsymbol{X}_i - 1) = \frac{1}{m}\left(\boldsymbol{T}^\top \boldsymbol{\Phi}\boldsymbol{T} - \|\boldsymbol{T}\|_1\right),$$

其中 $\boldsymbol{\Phi}$ 是随机布尔矩阵，满足 $\boldsymbol{\Phi}_{i,\,j} = 1$ 当且仅当 $i,\ j$ 属于 $\boldsymbol{H}$的同一个连通分量。其期望值为 $\mathbb{E}[\boldsymbol{Y}] = m \cdot p^\top \phi p$ 其中 $\phi = \mathbb{E}[\boldsymbol{\Phi}]$，我们证明当 p 远离均匀分布时，该值也会再次变大。证明使用矩阵 ϕ 的谱分析；该矩阵在路径情形下是托普利兹矩阵，在环情形下是循环矩阵。为了完成分析，我们需要对随机二次型 $\boldsymbol{T}^\top\boldsymbol{\Phi}\boldsymbol{T} - \|\boldsymbol{T}\|_1$建立集中不等式，只要 p 在任何区间内都不至于“高度集中”即可（算法可以单独测试这一点）；见 引理 3.21：

$$\mathbb{P}\left[\left|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\right| \geq t\right] \leq \frac{\|p\|_2^2}{\eta t^2} \cdot \mathrm{poly}\log n\,. \tag{1}$$

将结果推广到奇偶迹更具挑战性。在定义域 $[2n]$上，我们将输入分布 π 分为“奇数部分” p和“偶数部分” q，因此 $\pi = \pi(p,\ q) = (p_1,\ q_1,\ p_2,\ q_2,\ \ldots\ldots,\ p_n,\ q_n)$。测试器接收到如下形式的迹

$$\mathsf{trace}(S) = 1^{X_1}0^{Z_1}1^{X_2}0^{Z_2}\ldots 1^{X_t}0^{Z_t},$$

其中 $X_i,\ Z_i$ 表示迹中连续若干个 1 或 0 所构成的“游程”的长度（即 $X_i,\ Z_i > 0$ 不过我们可能有 $X_0 = 0$ 或 $Z_t = 0$）。类比标准测试器，自然的尝试是计算 “ 碰撞 ” 的数量 $\sum_{i=1}^{n} X_i(X_i-1)$ $+\sum_{i=1}^{n} Z_i(Z_i-1)$，并祈祷它能奏效，而事实上它基本奏效了。我们的测试器分别考虑 1 和 0 的游程：首先，我们将 p 视为环的顶点上的一个分布，其中 q 为边赋予权重。如果 q 是均匀分布的，那么混淆收集器的分析现在便适用，但它可能并不是。

为此，我们定义 q 的均匀共轭分布，并将其记为 $\tilde{p}$。直观而言，$\tilde{p}$ 是 p 的“最坏情况”实例，它使 $\boldsymbol{H}$ 的每个连通分量（按照由 q 确定的权重进行采样）具有相同的期望质量 τ，从而使 $\mathbb{E}[\boldsymbol{Y}]$最小化。我们通过近似计算出 $\tilde{p}$ 在 $\tau = \frac{1-\|q\|_1}{\sum_{i=1}^{n} \text{双曲正切}(mq_i/2)}$ 下的闭式解。

对 $\boldsymbol{H}$ 的分量进行采样的过程，使用马尔可夫过程（此处使用环而非路径会有所帮助）。然后我们写作 $p = \tilde{p} + z$，并且关键的是，使用 z 与均匀共轭分布之间的偏差来控制 $\boldsymbol{Y}$ 的均值和方差。我们得到方程（1）的一个类似结果，该结果在 p 的某些条件下成立，q 算法可以单独检验这些条件；见 引理 4.21：

$$\mathbb{P}\left[|\boldsymbol{Y} - \mathbb{E}\left[\boldsymbol{Y}\right]| \geq t\right] \leq \frac{\frac{1}{m} + z^{\top}\phi z}{t^2} \cdot \operatorname{poly}\log n\,. \tag{2}$$

算法必须单独检验的主要条件是：相对于 q，p 不能过于“集中”，也就是说，不存在某个区间使得 p 和 q 都足够大，而 p 却远大于 q。算法会交换 p 和 q 的角色，重复这些检验。

下界。为了证明定理 1.1 中的下界，考虑一个对手：它翻转一个随机比特 $\boldsymbol{Z}$，并向算法提供一个从“元分布” $\mathcal{D}_{\boldsymbol{Z}}$ 中采样得到的输入分布，其中 $\mathcal{D}_0$ 和 $\mathcal{D}_1$ 由多米诺骨牌按如下方式构造。一个多米诺骨牌是分布 $\pi(p,q) = (p_1, q_1, p_2, q_2, \ldots, p_n, q_n)$ 中的一个四元组 $(p_i, q_i, p_{i+1}, q_{i+1})$，因此定义域为 $[2n]$ 的 $\pi(p,q)$ 由 $n/2$ 个多米诺骨牌组成。我们使用多米诺骨牌 $(\frac{1}{2n}, \frac{1}{2n}, \frac{1}{2n}, \frac{1}{2n})$、$(\frac{1-\epsilon}{2n}, \frac{1}{2n}, \frac{1+\epsilon}{2n}, \frac{1}{2n})$ 和 $(\frac{1+\epsilon}{2n}, \frac{1}{2n}, \frac{1-\epsilon}{2n}, \frac{1}{2n})$。$\mathcal{D}_0$ 只包含均匀分布（第一个多米诺骨牌的 $n/2$ 个副本），而 $\mathcal{D}_1$ 则通过对最后两个多米诺骨牌进行 $n/2$ 次随机选择而得到。

我们使用受以下内容启发的信息论论证 [DK16]，来证明，除非算法接收到 $\boldsymbol{Z}$，否则它关于 $\widetilde{\Omega}((n/\epsilon)^{4/5})$ 的信息不足。对于从少于 3 个样本点接收样本的任何多米诺骨牌，测试器都无法获得关于 $\boldsymbol{Z}$ 的信息。我们对足够小的多米诺骨牌序列使用信息链式法则，并利用接收 3 个样本点的多米诺骨牌数量的界，来界定每个足够小的序列所提供的信息。

1.6 讨论与开放问题

读者可能会注意到本文有三个令人遗憾的缺点：上界为 $\widetilde{O}(\cdot)$，而非 $O(\cdot)$；检验算法不止 1 步；页数似乎也过多。我们猜想，如果能够证明关于二次型 $\boldsymbol{T}^{\top}\boldsymbol{\Phi}\boldsymbol{T}$ 的更紧的无条件集中界，或许就能一石三鸟。

关于检验与学习的问题，下一步是证明奇偶迹下测试支持集大小的紧界，这将为 k-交替函数（以及因此的半空间和半空间交）给出更好的下界，或者可能为 k-交替函数给出一个令人意外的 $O(k/\log k)$ 上界。我们计划在后续工作中研究这一问题。

由于与奇偶迹的联系，我们关于混淆收集器的结果仅限于路径和环，但我们猜想，类似的上界也适用于树。鉴于树结构数据无处不在，我们认为，在这个模型中，这将是下一个最自然的问题。

密度性质是一类自然的性质，在奇偶迹下，性质测试等价于分布测试。将其他分布测试结果（例如单调性测试（[BKR04，CDGR18]））推广到奇偶迹模型，将为带标签分布的分布无关的基于样本的检验以及迹重构模型带来新的结果。

我们认为，在迹重构模型中进行性质测试是本文在概念上的主要贡献之一。我们已证明，对 n-块字符串进行性质测试，与奇偶迹下的分布测试以及带标签分布测试模型中的密度性质测试有关。其他值得研究的有趣性质包括子序列无关性（采用非二进制字母表）；这或许可以借鉴近期关于基于样本的测试的研究 [RR21]，，或对正则语言进行测试；正则语言在标准查询模型中是可测试的（例如 [AKNS01，BS21]），且已知与迹重构相关 [Cha21b]。

2 预备知识和上界通用框架

本节给出分布测试中奇偶迹模型和混淆收集器模型的正式定义，并引入一套通用术语和框架，用于分析这些模型中的算法。本节安排如下：第 2 节.1 介绍全文通用的记号。第 2 节.2 和2.3 分别定义分布测试中的混淆收集器模型和奇偶迹模型。第 2 节.4 介绍统一的术语，将这两个模型视为某些路径结构和环结构定义域上的泊松随机变量的结果。随后，第 2 节.5运用这套术语建立通用结果，并在后续各节将其具体化为混淆收集器模型和奇偶迹模型的上界。

2.1 记号

在本文中，$\log x$ 表示 x的自然对数。 $\mathbb{N}$ 表示正整数集合，即不包括 0。对于任意 x，我们将大于 x的整数集合记作 $\mathbb{Z}_{>x}$ ，而$\mathbb{Z}_{<x}, \mathbb{Z}_{\geq x}, \mathbb{Z}_{\leq x}$ 也类似定义。我们用粗体符号表示随机变量，例如 $\boldsymbol{X}$。我们用 $x = a \pm b$ 作为 $a - b \leq x \leq a + b$的简写。对于一个事件 E， $\mathbb{1}[E]$ 是 E的指示变量，当且仅当 E 发生时取值为 1。

对于距离函数 a 集合 $\mathcal{X}$ 上的度量 dist$(\cdot,\cdot)$、一个元素 $y \in \mathcal{X}$ 和一个集合 $X \subseteq \mathcal{X}$，我们记作 wri te

$$\mathsf{dist}(y, X) := \inf_{x \in X} \mathsf{dist}(y, x)\,.$$

对于定义域 $\mathcal{X}$ 上的概率分布 $\mathcal{D}$，以及任意集合 $S \subseteq \mathcal{X}$，我们记 $\mathcal{D}(S) = \sum_{x \in S} \mathcal{D}(\{$v8$\}$。

给定概率分布 π 和 $m \in \mathbb{N}$，我们将用 $\boldsymbol{S} \sim \mathsf{samp}(\pi, m)$ 表示通过从 π中抽取 m 个独立样本而得到的多重集 S 上的分布。

对于固定定义域 $\mathcal{X}$ 以及定义在 $\mathcal{X}$上的概率分布集合 Π，我们用 $\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Pi)$ 表示定义在 $\mathcal{X}$ 上的分布集合 π，满足 $\mathsf{dist}_{\mathsf{TV}}(\pi, \Pi) > \epsilon$。对于其他定义域（例如带标签分布类 Ξ和字符串 Ψ）以及其他适用的（伪）度量（例如，表示与 Π 的编辑距离较远的分布的 $\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi)$），我们也会使用类似的记号。

在证明中，我们经常会使用 ?$\geq$、?$\leq$、?= 等记号，来表示将在证明后文中确立的（不）等式。

2.2 混淆收集器：定义与术语

我们将介绍一般的混淆收集器模型，不过本文只关注路径结构和环结构的定义域。在这些定义域上的混淆收集器模型也可作为奇偶迹的预备内容，因此我们先对其进行介绍和分析。分布测试中的标准做法是分析算法的“泊松化”版本：算法不是从输入分布中接收 m 个独立随机样本点 π，而是先选择 $\boldsymbol{m} \sim$ 泊松分布$(m$，然后从 $\boldsymbol{m}$ 中抽取 π个独立随机点 x；这意味着定义域中的每个点 泊松分布$(m \cdot \pi(x))$-次出现在样本中，且各点之间相互独立。为简单起见，我们将定义混淆收集器的泊松化版本（更多细节请参见附录 A及其中的参考文献）。

定义 2.1 (混淆收集器采样)。 设 $G = (V, E)$ 是一个图， $w\colon E \to [0,1]$ 是一个非负权重向量。我们定义如下采样过程。通过以概率 $1 - w(e)$ 相互独立地纳入每条边 e，从 G 中选取一个随机子图 $\boldsymbol{H}$。设 $\boldsymbol{C}_1$, ……， $\boldsymbol{C}_k$ 是 $\boldsymbol{H}$ 的连通分量；为每个 $\boldsymbol{C}_i$ 任意指定一个代表顶点 $\boldsymbol{c}_i$，该顶点属于 $\boldsymbol{C}_i$。

对于概率分布 π（或者实际上，对于任意非负向量 $\pi: V \to \mathbb{R}_{\geq 0}$）和样本规模参数 m，我们定义从 π 得到的困惑收集器样本 $\boldsymbol{S}$ 如下。按上述方式选取 $\boldsymbol{H}$。对于每个顶点 $v \in V$，我们独立采样一个泊松随机变量 $\boldsymbol{s}(v) \sim \mathrm{Poi}(m\pi(v))$，并将 $\boldsymbol{c}(v)$ 以重数 $\boldsymbol{s}(v)$ 加入样本 $\boldsymbol{S}$，其中 $\boldsymbol{c}(v)$ 是包含顶点 v 的连通分量 $\boldsymbol{C}_i$ 的代表元。

目前，我们只关注权重为常数的情形，即存在某个 $\eta \in [0,1]$，使得对于所有边 e，都有 $w(e) = \eta$。我们将 η 称为分辨率。

定义 2.2（混淆收集器下的分布检验）。固定一个图 $G = (V, E)$ 和一个分辨率参数 η。设 Π_1, Π_2 是定义在 V 上的概率分布性质，并设$\alpha \in (0, 1)$。具有分辨率 η 和样本复杂度 m 的（Π_1, Π_2, α）-分布测试器是一个算法 A，它接收来自输入分布 π 的困惑收集器样本 $\boldsymbol{S}$，并满足：

1. 若 $\pi \in \Pi_1$，则 $\mathbb{P}[A(\boldsymbol{S}) \text{ 接受 }] \geq \alpha$；并且
2. 若 $\pi \in \Pi_2$，则 $\mathbb{P}[A(\boldsymbol{S}) \text{ 拒绝 }] \geq \alpha$。

在假设 $\alpha = 2/3$ 的情况下，我们会在记号中省略 α。不过，我们指出，由于采样过程有所改变，混淆收集器无法像标准分布测试器那样提高成功概率。

2.3 奇偶迹：定义与术语

本节将正式定义奇偶迹下的分布测试，并介绍用于分析测试器和证明定理 1.1的记号和术语。对于多重集 $S \subset \mathbb{N}$，回顾引言中的迹定义迹(S。接下来，我们定义测试模型：

定义 2.3。设 Π_1 和 Π_2 是定义域 $\mathbb{N}$上的任意分布性质。一个样本复杂度为 m的、在奇偶迹下的（Π_1, Π_2, α）分布测试器，是满足以下条件的算法 A。

1. 若 $\pi \in \Pi_1$，则 $\mathbb{P}_{\boldsymbol{S} \sim \mathsf{samp}(\pi, m)}[A(\text{迹}(\boldsymbol{S})) \text{ 接受 }] \geq \alpha$。
2. 若 $\pi \in \Pi_2$，则 $\mathbb{P}_{\boldsymbol{S} \sim \text{采样}(\pi,\ m)}[A(\mathsf{trace}(\boldsymbol{S})) \text{ 拒绝 }] \geq \alpha$。

这个问题的规范版本将具有 $\Pi_2 = \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi_1)$，或者在某些情况下具有 $\Pi_2 = \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Pi_1)$。

我们称向量 $r \in \mathbb{R}^{\mathbb{N}}$ 为**部分分布**，如果它的所有分量均为非负数，并且$\sum_i r_i \leq 1$。

在奇偶迹模型中，我们认为，定义在 $\mathbb{N}$ 上的概率分布 π 由两个部分分布 p，$q \in \mathbb{R}_{\geq 0}^{\mathbb{N}}$构成，因此 $\pi = \pi(p, q)$，其中

$$\pi(p, q) := (p_1, q_1, p_2, q_2, p_3, q_3, \dots),$$

因此，p 定义了分布在奇数元素上的部分，而 q 定义了分布在偶数元素上的部分。我们将始终分别用字母 p 和 q 表示奇数元素和偶数元素上的部分分布。

在奇偶迹模型中，算法接收一个包含 1 和 0 的迹 $\mathsf{trace}(S)$，并分别考察 1 和 0 的统计量。在分析中，我们只考虑

1 的统计量，因为 0 的统计量可以用对称的方式处理。为了分析 1，我们可以将算法接收的迹（用正则表达式记法）写成如下形式

$$\mathsf{trace}(S) = 1^{Z_1}\ 0^{+}\ 1^{Z_2}\ 0^{+} \cdots 1^{Z_t}\ 0^{*}\,,$$

对于某个 t，其中 $Z_2, \ldots, Z_t > 0$ 并且我们允许 $Z_1 = 0$。一段连续的 1 称为“游程”，我们将数值 Z_i 称为“游程长度”。

对于我们的测试器来说，实际使用“圆形轨迹”会更方便：将字符串 $\mathsf{trace}(S)$ 的两端连接起来，形成一条项链。如果轨迹以相同的符号开始和结束，那么首尾两个“游程”就会合并。因此，算法看到的是如下形式的**圆形轨迹**

$$1^{X_1}\ 0^{+}\ 1^{X_2}\ 0^{+} \cdots 1^{X_b}\ 0^{+}\,,$$

其中可能有 $X_1 = Z_1 + Z_t$。（这里，σ^{+} 表示符号 σ 至少出现一次。）在均匀性检验中，我们只关注定义域 $[2n]$，部分分布 p, q 定义在 $[n]$ 上，因此可以将定义域本身视为连接成一条项链。等价地，我们将 p 的定义域 $[n]$视为一个环的顶点。

更准确地说，我们考虑顶点为 $\mathbb{Z}_n$、顶点上带有部分分布 p 的一个环，并在边上定义权重向量 w，其中顶点 i 与 $i+1$ 之间的边权重为 $1 - e^{-mq_i}$。然后，按照混淆收集器采样过程对一个子图 $\boldsymbol{H}$ 进行采样，我们看到环中的顶点 i 和 $i+1$ 在 $\boldsymbol{H}$ 中相邻的概率为 $e^{-mq_i} = \mathbb{P}[\,\text{泊松分布}(mq_i) = 0]$；这就是这些顶点对迹中同一游程长度 X_j 有贡献的概率。

2.4 路径与环结构的泊松随机变量。

为分析环和路径上的泊松随机变量，引入一套通用术语会很方便。我们将环的 n 个顶点标记为模 n 整数的集合 $\mathbb{Z}_n$，并将环的边 也标记为集合 $\mathbb{Z}_n$，使得边 i 连接顶点 i 和 $i+1$（按模 n 运算）。我们将含有 n 个顶点的路径视为环的一个子图，该子图不包含连接标记为 0 和 $n-1$ 的顶点的边 $n-1$。当子图 H 包含边 e 时，我们有时会滥用记号，写作 $e \in H$。

圆周区间 是一个元组 $\langle\!\langle i, d \rangle\!\rangle$，其中 $i \in \mathbb{Z}_n$ 且 $d \in \mathbb{Z}$。如果 $d \geq 0$，我们将元素$\mathcal{E}\langle\!\langle i, d \rangle\!\rangle$ 定义为一个多重集：从顶点 $i \in \mathbb{Z}_n$ 开始，包含从 i起沿“顺时针”方向的 $d-1$ 个元素，即多重集 $\{i, i+1, i+2, \ldots, i+d-1\}$，其中加法按模 n进行。注意，对于 $d = 1$，其中只包含 i；而对于 $d > n$，其中某些元素的重数大于 1。如果 $d < 0$，我们将元素 $\mathcal{E}\langle\!\langle i, d \rangle\!\rangle$ 定义为一个多重集：从顶点 $i \in \mathbb{Z}_n$开始，包含沿“逆时针”方向的 $|d|-1$ 个元素，即多重集 $\{i, i-1, i-2, \ldots, i-|d|+1\}$。

$d \geq 0$时，$\langle\!\langle i, d \rangle\!\rangle$ 的**端点**为整数 i 和 $i+d-1$；$d < 0$时，端点为 $i+d+1$ 和 i。我们通常会在记号中省略 $\mathcal{E}$，并将元组 $\langle\!\langle i, d \rangle\!\rangle$ 与其元素多重集视为等同，因此我们写作 $x \in \langle\!\langle i, d \rangle\!\rangle$，而不写作 $\mathcal{E}\langle\!\langle i, d \rangle\!\rangle$。然而，圆周区间不与其元素多重集等同；例如，圆周区间 $\langle\!\langle i, n \rangle\!\rangle$ 和 $\langle\!\langle i+1, n \rangle\!\rangle$都包含相同的元素 $\mathbb{Z}_n$，但它们的端点不同。

对于圆周区间 I 和向量 $u\colon \mathbb{Z} \to \mathbb{R}$，我们定义

$$u[I] := \sum_{s \in I} u_s\,,$$

wh此处注意，s 可能在 I 中出现多次，而 u_s 每次都会计数。

对于圆周区间 $\langle\!\langle i, d \rangle\!\rangle$，我们将圆周区间 $\langle\!\langle i, d \rangle\!\rangle^*$ 定义为与顶点 $\langle\!\langle i, d \rangle\!\rangle$所诱导的边 对应的整数；具体而言

$$\langle\!\langle i, d \rangle\!\rangle^* = \begin{cases} \emptyset & \text{if } d = 0 \\ \langle\!\langle i, d-1 \rangle\!\rangle & \text{if } d \geq 1 \\ \langle\!\langle i-1, 1-|d| \rangle\!\rangle & \text{if } d \leq -1 \,. \end{cases}$$

对于任意 $s \in \mathbb{Z}_n$，我们称圆周区间 I **跨越** s，如果 $s \in I^*$；也就是说，s 是 I中两个顶点之间的一条边。

固定环（或路径）的任意子图 H，并假设 H 有 b 个连通分量；注意，每个连通分量都是一个圆周区间。我们将由 H 诱导的**桶** 定义为 Γ_1、……、Γ_n，使得 Γ_1、……、Γ_b是 H的连通分量，而 Γ_{b+1}、……、$\Gamma_n = \emptyset$。对于每个顶点 $i \in \mathbb{Z}_n$，我们定义

$$\gamma(i) := t \text{ such that } i \in \Gamma_t \,.$$

如果 $\gamma(i) = \gamma(j)$，我们称两个顶点 i、j 相**连**，并将**连接矩阵** $\Phi = \Phi(H)$定义为

$$\Phi_{i,j} := \begin{cases} 1 & \text{if } \gamma(i) = \gamma(j) \\ 0 & \text{otherwise.} \end{cases}$$

我们定义一个**连接函数** J，使得对于任意圆周区间 $I = \langle\!\langle i, d \rangle\!\rangle$，

$$J(I) := \mathbb{1}\left[\forall e \in \langle\!\langle i, d \rangle\!\rangle^* : e \in H\right] \,.$$

因此，如果 $J(I) = 1$，那么对于每个 $i,\ j \in \mathcal{E}(I)$，都有 $\Phi_{i,\ j} = 1$。

对于固定样本（即多重集）$S \subset \mathbb{Z}_n$ 和 $i \in \mathbb{Z}_n$，记 T_i 为元素i 在 S中的重数。接着，对每个 $i \in [n]$ 定义变量

$$X_i := \sum_{j \in \mathbb{Z}_n : \gamma(j) = i} T_j \,,$$

它表示桶 Γ_i 中出现在 S里的元素的总重数。

注意，上述变量取决于子图 H 和样本 S。对于固定的权重向量 w 、按照混淆收集器采样过程选取的随机子图 $\boldsymbol{H}$，以及顶点的随机样本 $\boldsymbol{S}$，我们将上述变量加粗，以表示它们是依赖于 $\boldsymbol{H}$ 和 $\boldsymbol{S}$的随机变量。于是我们记

$$\phi := \mathbb{E}\left[\boldsymbol{\Phi}\right] \,,$$

并注意到

$$\phi_{i,j} = \mathbb{P}\left[\boldsymbol{\gamma}(i) = \boldsymbol{\gamma}(j)\right] \,.$$

在分析混淆收集器和奇偶迹时，我们有一个样本量参数 m和一个输入（部分）分布 p： $\mathbb{Z}_n \to [0, 1]$。于是有

$$\boldsymbol{T}_j \sim \mathsf{Poi}(mp_j)$$

对每个 $j \in \mathbb{Z}_n$均成立，因此

$$\boldsymbol{X}_i \sim \mathsf{Poi}(m \cdot p[\boldsymbol{\Gamma}_i])$$

对每个 $i \in [n]$。我们还将使用随机布尔矩阵 $\boldsymbol{\Phi}$，它表示 $\boldsymbol{H}$ 的连通分量。我们的检验算法将基于对以下<替换 id='4'>检验统计量</替换>的分析。

<替换 id='1'>定义 2.4 </替换>（检验统计量）<替换 id='3'>。 </替换>对于固定参数 m 和权重向量 w，以及如上定义的随机变量，我们定义检验统计量

$$\boldsymbol{Y} := \frac{1}{m}\sum_{i=1}^{n} \boldsymbol{X}_i(\boldsymbol{X}_i - 1)\,.$$

将变量 $\boldsymbol{X}_i$ 展开后，检验统计量可写成如下二次型

$$\boldsymbol{Y} = \frac{1}{m}\left(\boldsymbol{T}^\top \boldsymbol{\Phi}\boldsymbol{T} - \|\boldsymbol{T}\|_1\right)\,.$$

2.5 共享分析

现在，我们将讨论一部分分析内容，这部分内容在混淆收集器模型和奇偶迹结果中是共通的，反映了这两种模型各自带来的共同挑战。随后各节应用这些结果时，将利用每种模型的特殊性质——本质上，在混淆收集器模型中，分辨率 η 是固定的；而在奇偶迹模型中，部分分布和所选样本量会影响顶点和边的抽样率。

2.5.1 圆周区间

我们的分析将处理 G 为环或路径的情形。对于路径，跨越顶点 0 与 $n-1$ 之间那条边的圆形区间无关紧要，因此将 $\mathcal{I}^{\mathsf{cycle}}$ 定义为所有圆形区间的集合，并将 $\mathcal{I}^{\mathsf{path}}$ 定义为所有不跨越边 $n-1$ 的圆形区间的集合，会比较方便。

我们用 $\mathcal{I} \in \{\mathcal{I}^{\mathsf{cycle}}, \mathcal{I}^{\mathsf{path}}\}$ 表示与分析相关的圆形区间集合。在 $\mathcal{I} = \mathcal{I}^{\mathsf{cycle}}$ 的情形下，每对顶点 $i \le j$ 之间都有两条互不相交的路径相连，因此可以通过两种方式将它们连接起来。我们将 $\mathsf{small}(i,\ j)$ 和 $\mathsf{large}(i,\ j)$ 定义为如下所述的两个圆形区间。设

$$I_1 := \langle\!\langle i, j-i+1 \rangle\!\rangle \qquad \text{and} \qquad I_2 := \langle\!\langle j, n-(j-i)+1 \rangle\!\rangle$$

作为连接 i 和 j的两条不同路径所对应的圆周区间。然后我们定义

$$\mathsf{small}(i,j) := \arg\max_{I\in\{I_1,I_2\}} \mathbb{E}\left[\boldsymbol{J}[I]\right] \qquad \text{and} \qquad \mathsf{large}(i,j) := \arg\min_{I\in\{I_1,I_2\}} \mathbb{E}\left[\boldsymbol{J}[I]\right]\,,$$

以任意方式打破平局。请注意，在路径的情况下，连接 i和 j只有一种方式，因此在这种情况下 $\mathbb{E}\left[\boldsymbol{J}[\mathsf{large}(i,\ j)]\right] = 0$。对称地，当 $i > j$时，我们定义 $\mathsf{small}(i,\ j) := \mathsf{small}(j,\ i)$ 和 $\mathsf{large}(i,\ j) := \mathsf{large}(j,\ i)$。

对于 $\mathcal{I} \in \{\mathcal{I}^{\mathsf{cycle}}, \mathcal{I}^{\mathsf{path}}\}$，我们定义

$$\zeta(\mathcal{I}) := \max_{i,j} \mathbb{E}\left[\boldsymbol{J}[\mathsf{large}(i,j)]\right]\,.$$

分析分为两种情况。对于混淆收集器，我们假设权重向量为常数，因此对于每条边 j，$w(j) = \eta$，其中 η 是分辨率参数；那么边 j 出现在 $\boldsymbol{H}$ 中的概率为 $1-w(j) = 1-\eta$。对于奇偶迹，我们有两个部分分布，p 和 q，分别是输入分布中对应于定义域的奇数元素和偶数元素的部分。我们将 p 视为环顶点上的分布，并定义权重向量 $w(j) := 1 - e^{-mq_j}$，使得边 j 出现在 $\boldsymbol{H}$ 中的概率为 $1 - w(j) = e^{-mq_j} = \mathbb{P}\left[\text{泊松分布}(mq_j) = 0\right]$。

命题 2.5. $\zeta(\mathcal{I}^{\mathsf{path}}) = 0$. 对于常数权重 $w(j) = \eta$, 我们有

$$\zeta(\mathcal{I}^{\mathsf{cycle}}) \le (1-\eta)^{n/2},$$

对于权重 $w(j) = 1 - e^{-mq_j}$, 我们有

$$\zeta(\mathcal{I}^{\mathsf{cycle}}) \le e^{-\frac{m\|q\|_1}{2}}.$$

证明。 对于不同的 i, $j \in \mathbb{Z}_n$, 定义 I_1 和 I_2 如上，并注意对于 $a \in \{1,2\}$,

$$\mathbb{E}[\boldsymbol{J}[I_a]] = \prod_{j \in I_a^*} (1 - w(j)).$$

在 $w(j) = \eta$ 的情况下，这是 $(1-\eta)^{|I_a^*|}$; 而在 $w(j) = 1 - e^{-mq_j}$ 的情况下，这是 $e^{-mq[I_a^*]}$。注意， I_1, I_2 分割 $\mathbb{Z}_n$, 因此在第一种情况下，我们有 $|I_1| \ge n/2$ 或 $|I_2| \ge n/2$ 中的一个，所以最小值至多为 $(1-\eta)^{n/2}$。在第二种情况下，我们有 $q[I_1^*] \ge \|q\|_1/2$ 或 $q[I_2^*] \ge \|q\|_1/2$ 中的一个，所以最小值至多为 $e^{-\frac{m\|q\|_1}{2}}$。

□

2.5.2 检验统计量的期望

我们首先给出统计量 $\boldsymbol{Y}$ 的期望表达式。回顾一下，我们将 p 写作： $\mathbb{Z}_n \to [0,1]$ ，表示顶点（无论是路径还是环上的顶点）上的（部分）分布， m 是样本量参数，且 $\phi = \mathbb{E}[\boldsymbol{\Phi}]$。

命题 2.6. 统计量 $\boldsymbol{Y}$ 满足

$$\mathbb{E}[\boldsymbol{Y}] = mp^\top \phi p.$$

证明. 我们利用以下事实： $\boldsymbol{T}$ 与 $\boldsymbol{\Phi}$ 相互独立，并且对于 $i \ne j$, $\boldsymbol{T}_i$ 与 $\boldsymbol{T}_j$ 相互独立。我们还将用到如下性质：对于 $\boldsymbol{Z} \sim$ 泊松分布(λ)，有 $\mathbb{E}[\boldsymbol{Z}] = \mathsf{Var}[\boldsymbol{Z}] = \lambda$，因此 $\mathbb{E}[\boldsymbol{Z}^2] = \mathbb{E}[\boldsymbol{Z}] + \mathbb{E}[\boldsymbol{Z}]^2$。我们得到:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Y}] &= \frac{1}{m}\left(\mathbb{E}\left[\boldsymbol{T}^\top \boldsymbol{\Phi} \boldsymbol{T}\right] - \mathbb{E}[\|\boldsymbol{T}\|_1]\right) = \frac{1}{m}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \mathbb{E}[\boldsymbol{T}_i \boldsymbol{T}_j]\,\mathbb{E}[\boldsymbol{\Phi}_{i,j}] - \frac{1}{m}\sum_{i=0}^{n-1}\mathbb{E}[\boldsymbol{T}_i] \\
&= \frac{1}{m}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \mathbb{E}[\boldsymbol{T}_i]\,\mathbb{E}[\boldsymbol{T}_j]\,\phi_{i,j} + \frac{1}{m}\sum_{i=0}^{n-1}\mathbb{E}[\boldsymbol{T}_i] - \frac{1}{m}\sum_{i=0}^{n-1}\mathbb{E}[\boldsymbol{T}_i] = \frac{1}{m}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}(mp_i)(mp_j)\phi_{i,j} \\
&= mp^\top \phi p.
\end{aligned}$$

□

有时，写作 $p = p^* + z$ 会很有用，其中 p^* 是参考部分分布，此时我们要求:

命题 2.7. 令 $p = p^* + z$。则 $\boldsymbol{Y}$满足

$$\mathbb{E}[\boldsymbol{Y}] = m(p^*)^\top \phi p^* + 2m(p^*)^\top \phi z + mz^\top \phi z.$$

证明。 这可由 命题 2.6 直接得出：展开二次型，并回想 $\boldsymbol{\Phi}$ 始终是对称矩阵，因此 $\phi = \mathbb{E}[\boldsymbol{\Phi}]$也如此。

□

2.5.3 检验统计量的方差：第一项

本节将为 $\boldsymbol{Y}$ 的方差建立适用于奇偶迹模型和混淆收集器模型的一般上界。后续关于各个具体模型的章节将以此为基础展开。

回顾一下，将顶点随机划分到桶 $\boldsymbol{\Gamma}_1$、……、$\boldsymbol{\Gamma}_n$ 取决于随机子图 $\boldsymbol{H}$。首先注意，根据全方差公式，我们可以将 $\boldsymbol{Y}$ 的方差分解为两项：

$$\operatorname{Var}[\boldsymbol{Y}] = \operatorname*{Var}_{\boldsymbol{H}}\left[\operatorname*{\mathbb{E}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right] + \operatorname*{\mathbb{E}}_{\boldsymbol{H}}\left[\operatorname*{Var}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right].$$

我们将在此处处理第一项，并在第 2.5.4 节处理第二项。

回顾一下，权重向量 w 要么是常数向量 η，要么是 $w(j) = 1 - e^{-mq_j}$。

命题 2.8。 设 $\mathcal{I} \in \{\mathcal{I}^{\mathsf{cycle}}, \mathcal{I}^{\mathsf{path}}\}$。对于每个 i、j、k、$\ell \in \mathbb{Z}_n$，以下结论成立：

1. $\mathbb{E}[\Phi(\boldsymbol{J})_{i,\,j}] \geq \mathbb{E}[\boldsymbol{J}[\mathsf{small}(i,\;j)]]$；
2. $\mathbb{E}[\Phi(\boldsymbol{J})_{i,\,j} \cdot \Phi(\boldsymbol{J})_{k,\,\ell}] \leq \mathbb{E}[\boldsymbol{J}[小(i,\;j)] \cdot \boldsymbol{J}[小(k,\;\ell)]] + 4 \cdot \zeta(\mathcal{I})$。

证明。回忆一下，$\Phi(\boldsymbol{J})_{i,\,j} = 1$ 当且仅当 $\boldsymbol{\gamma}(i) = \boldsymbol{\gamma}(j)$。当小$(i,\;j) \subseteq \boldsymbol{\Gamma}_{\gamma(i)}$ 时，这种情况就会发生；而当 $\boldsymbol{J}[小(i,\;j)] = 1$ 时，上述条件成立，从而得出第一个结论。接下来，注意到

$$\boldsymbol{\Phi}_{i,j} = \max\{\boldsymbol{J}[\mathsf{small}(i,j)], \boldsymbol{J}[\mathsf{large}(i,j)]\} \leq \boldsymbol{J}[\mathsf{small}(i,j)] + \boldsymbol{J}[\mathsf{large}(i,j)].$$

为证明第二个结论，展开乘积，并利用 $\boldsymbol{J}$ 是一个布尔向量这一事实：

$$\begin{aligned}\boldsymbol{\Phi}_{i,j}\boldsymbol{\Phi}_{k,\ell} &\leq (\boldsymbol{J}[\mathsf{small}(i,j)] + \boldsymbol{J}[\mathsf{large}(i,j)])\,(\boldsymbol{J}[\mathsf{small}(k,\ell)] + \boldsymbol{J}[\mathsf{large}(k,\ell)]) \\ &\leq \boldsymbol{J}[\mathsf{small}(i,j)]\boldsymbol{J}[\mathsf{small}(k,\ell)] + 2\,(\boldsymbol{J}[\mathsf{large}(i,j)] + \boldsymbol{J}[\mathsf{large}(k,\ell)])\,.\end{aligned}$$

根据定义，我们有 $\mathbb{E}[\boldsymbol{J}[大(i,\;j)]]$，$\mathbb{E}[\boldsymbol{J}[大(k,\;\ell)]] \leq \zeta(\mathcal{I})$，因此取期望即可得出结论。

□

引理 2.9。 设 $\mathcal{I} \in \{\mathcal{I}^{\mathsf{cycle}}, \mathcal{I}^{\mathsf{path}}\}$，并设 w 为边上的权重。存在一个绝对常数 $c > 0$，使得 $\boldsymbol{Y}$ 的方差的第一部分满足

$$\operatorname*{Var}_{\boldsymbol{H}}\left[\operatorname*{\mathbb{E}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right] \leq 5m^2\zeta(\mathcal{I})\|p\|_1^4 + cm^2 \cdot \sum_{\substack{I=\langle\langle i,d\rangle\rangle \in \mathcal{I} \\ 1 \leq d \leq n}} p_i p_{i+d-1} p[I]^2 \mathbb{E}[\boldsymbol{J}(I)]\,.$$

证明。任取子图 H。给定 $\boldsymbol{H} = H$，

$$\begin{aligned}\operatorname*{\mathbb{E}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H} = H] &= \frac{1}{m}\mathbb{E}\left[\boldsymbol{T}^\top \Phi \boldsymbol{T}\right] - \frac{1}{m}\mathbb{E}[\|\boldsymbol{T}\|_1] = \frac{1}{m}\sum_{i,j \in \mathbb{Z}_n} \mathbb{E}[\boldsymbol{T}_i\boldsymbol{T}_j]\,\Phi_{i,j} - \frac{1}{m}\mathbb{E}[\mathsf{Poi}(m)] \\ &= \frac{1}{m}\left(\sum_{i \in \mathbb{Z}_n} \mathbb{E}[\boldsymbol{T}_i] + \sum_{i,j \in \mathbb{Z}_n} \mathbb{E}[\boldsymbol{T}_i]\,\mathbb{E}[\boldsymbol{T}_j]\,\Phi_{i,j}\right) - 1 = \frac{1}{m}\left(m + m^2 p^\top \Phi p\right) - 1 \\ &= mp^\top \Phi p\,,\end{aligned}$$

因此，所需的方差为

$$\operatorname*{Var}_{\boldsymbol{H}}\left[\operatorname*{\mathbb{E}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right] = m^2 \operatorname*{Var}_{\boldsymbol{H}}\left[p^\top \boldsymbol{\Phi} p\right].$$

然后，回忆起 $\phi = \mathbb{E}[\boldsymbol{\Phi}]$，我们将 $\mathrm{Var}_{\boldsymbol{H}}\left[p^\top \boldsymbol{\Phi} p\right]$ 展开如下：

$$\begin{aligned}
\mathrm{Var}\left[p^\top \boldsymbol{\Phi} p\right] &= \mathbb{E}\left[\left(p^\top \boldsymbol{\Phi} p\right)^2\right] - \left(\mathbb{E}\left[p^\top \boldsymbol{\Phi} p\right]\right)^2 = \mathbb{E}\left[\left(p^\top \boldsymbol{\Phi} p\right)^2\right] - \left(p^\top \phi p\right)^2 \\
&= \mathbb{E}\left[\left(\sum_{i,j\in\mathbb{Z}_n} p_i p_j \boldsymbol{\Phi}_{i,j}\right)^2\right] - \left(\sum_{i,j\in\mathbb{Z}_n} p_i p_j \phi_{i,j}\right)^2 \\
&= \mathbb{E}\left[\sum_{i,j,k,\ell\in\mathbb{Z}_n} p_i p_j p_k p_\ell \boldsymbol{\Phi}_{i,j}\boldsymbol{\Phi}_{k,\ell}\right] - \sum_{i,j,k,\ell\in\mathbb{Z}_n} p_i p_j p_k p_\ell \phi_{i,j}\phi_{k,\ell} \\
&= \sum_{i,j,k,\ell\in\mathbb{Z}_n} p_i p_j p_k p_\ell \left(\mathbb{E}\left[\boldsymbol{\Phi}_{i,j}\boldsymbol{\Phi}_{k,\ell}\right] - \phi_{i,j}\phi_{k,\ell}\right) .
\end{aligned}$$

现在我们利用 命题 2.8 来简化量 $\mathbb{E}[\boldsymbol{\Phi}_{i,j}\boldsymbol{\Phi}_{k,\ell}] - \phi_{i,\ j}\phi_{k,\ \ell}$：

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{\Phi}_{i,j}\boldsymbol{\Phi}_{k,\ell}] - \phi_{i,j}\phi_{k,\ell} \\
&\quad \le \mathbb{E}[\boldsymbol{J}[\mathsf{small}(i,j)]\boldsymbol{J}[\mathsf{small}(k,\ell)]] - \mathbb{E}[\boldsymbol{J}[\mathsf{small}(i,j)]]\,\mathbb{E}[\boldsymbol{J}[\mathsf{small}(k,\ell)]] + 4\cdot\zeta(\mathcal{I}) .
\end{aligned}$$

如果区间 $\mathsf{small}(i,\ j)$ 和 $\mathsf{small}(k,\ \ell)$ 不相交，那么

$$\mathbb{E}[\boldsymbol{J}[\mathsf{small}(i,j)]\boldsymbol{J}[\mathsf{small}(k,\ell)]] - \mathbb{E}[\boldsymbol{J}[\mathsf{small}(i,j)]]\,\mathbb{E}[\boldsymbol{J}[\mathsf{small}(k,\ell)]] = 0 .$$

另一方面，如果这些区间相交，我们将使用简单的上界

$$\mathbb{E}[\boldsymbol{J}[\mathsf{small}(i,\ j)]\boldsymbol{J}[\ \mathsf{small}(k,\ \ell)]] - \mathbb{E}[\boldsymbol{J}[\mathsf{small}(i,\ j)]]\,\mathbb{E}[\boldsymbol{J}[\mathsf{small}(k,\ \ell)]] \le \mathbb{E}[\boldsymbol{J}[\mathsf{small}(i,\ j)]\boldsymbol{J}[\mathsf{small}(k,\ \ell)]] .$$

接下来我们考虑两种情况。

首先，假设对于每条边 $s \in \mathbb{Z}_n$，$\mathsf{small}(i,\ j)$ 与 s or $\mathsf{small}(k,\ \ell)$ 与 s 相交。那么$\mathbb{Z}_n \subseteq \mathsf{small}(i,\ j)^* \cup \mathsf{small}(k,\ \ell)^*$，因此 $\boldsymbol{J}(\mathsf{small}(i,\ j))\cdot\boldsymbol{J}(\mathsf{small}(k,\ \ell)) = 1$ 仅当每条边都出现在$\boldsymbol{H}$中时成立，而这种情况发生的概率至多为 $\zeta(\mathcal{I})$（因为该事件意味着每个大区间都相连）。在这种情况下，$\mathbb{E}[\boldsymbol{J}[\mathsf{small}(i,\ j)]\boldsymbol{J}[\mathsf{small}(k,\ \ell)]] \le \zeta(\mathcal{I})$。

至于第二种情况，令 $s \in \mathbb{Z}_n$ 满足 $\mathsf{small}(i,j)$ 和 $\mathsf{small}(k,\ell)$ 均不跨越 s。由于 $\mathsf{small}(i,j)$ 和 $\mathsf{small}(k,\ell)$ 不相离，因此存在一个区间 $I = I_{i,j,k,\ell} \in \mathcal{I}$满足以下条件：

1. 集合 $\mathsf{small}(i,j) \cup \mathsf{small}(k,\ell)$ 等于 $I_{i,j,k,\ell}$ 中元素组成的集合，这里我们取的是集合的并集（而非多重集的并集）；
2. $I_{i,\ j,\ k,\ \ell}$ 的端点是索引 i、j、k、ℓ 中的两个；并且，
3. $|I_{i,\ j,\ k,\ \ell}| \le n$ （因为，特别是，$I_{i,\ j,\ k,\ \ell}$ 不会跨越 s）。

由此可知，仅当 $\boldsymbol{J}[I] = 1$ 时，$\boldsymbol{J}[\mathsf{small}(i,\ j)]\boldsymbol{J}[\mathsf{small}(k,\ \ell)] = 1$，因此我们有上界 $\mathbb{E}[\boldsymbol{J}[\mathsf{small}(i,\ j)]\boldsymbol{J}[\ \mathsf{small}(k,\ \ell)]] - \mathbb{E}[\boldsymbol{J}[\mathsf{small}(i,\ j)]]\,\mathbb{E}[\boldsymbol{J}[\mathsf{small}(k,\ \ell)]] \le \mathbb{E}[\boldsymbol{J}[I]]$。因此，

$$\sum p_i p_j p_k p_\ell \left(\mathbb{E}[\boldsymbol{\Phi}_{i,j}\boldsymbol{\Phi}_{k,\ell}] - \phi_{i,j}\phi_{k,\ell}\right)_{i,j,k,\ell\in\mathbb{Z}_n} \le \sum p_i p_j p_k p_\ell \left(\mathbb{E}[\boldsymbol{J}[\mathit{small}(i,j)]\boldsymbol{J}[\mathit{small}(k,\ell)]] - \mathbb{E}[\boldsymbol{J}[\ \mathit{small}(i,j)]]\,\mathbb{E}[\boldsymbol{J}[\mathit{small}(k,\ell)]] + 4\cdot\zeta\right)_{i,j,k,\ell\in\mathbb{Z}_n} \le 4\cdot\|p\|_1^4\cdot\zeta + \sum p_i p_j p_k p_\ell \left(\mathbb{1}[\mathbb{Z}_n \subseteq \mathit{small}(i,j)^* \cup \mathit{small}(k,\ell)^*]\cdot\zeta(\mathcal{I}) + \mathbb{1}[I_{i,j,k,\ell}\text{存在}]\cdot\mathbb{E}[\boldsymbol{J}[I_{i,j,k,\ell}]]\right)_{i,j,k,\ell\in\mathbb{Z}_n} \le 5\zeta(\mathcal{I})\|p\|_1^4 + \sum p_i p_j p_k p_\ell \mathbb{1}[I_{i,j,k,\ell}\text{存在}]\cdot\mathbb{E}[\boldsymbol{J}[I_{i,j,k,\ell}]]_{i,j,k,\ell\in\mathbb{Z}_n}$$

后者通过对所有满足 $d \le n$ 的区间 $I = \langle\langle i, d\rangle\rangle \in \mathcal{I}$ 求和，并对每个区间取表达式 $c \cdot p_i p_{i+d-1} \sum_{j,\ k\in I} p_j p_k \cdot \mathbb{E}[\boldsymbol{J}[I]]$ 来界定，其中 i 和 $i+d-1$ 是 I 的端点，而 c 是一个常数，用于计数端点位于 i，$(i+d-1)$ k，ℓ 的相交区间的构造方式数。现在，利用 $\sum_{j,\ k\in I} p_j p_k = p[I]^2$，我们得到

$$\operatorname{Var}\left[p^\top \Phi(\boldsymbol{J}) p\right] \le 5\zeta(\mathcal{I})\|p\|_1^4 + c \cdot \sum_{\substack{I=\langle\langle i,d\rangle\rangle\in\mathcal{I}\\ 1\le d\le n}} \mathbb{E}[\boldsymbol{J}[I]]\, p_i p_{i+d-1} p[I]^2 . \qquad \square$$

2.5.4 检验统计量的方差：第二个分量

我们引入一些记号，用于描述由 p 诱导的桶上的部分分布 Γ_i（即 H 的连通分量）。

定义 2.10（分桶向量）。设 $\Gamma = (\Gamma_1, \dots, \Gamma_b)$ 为子图 H 得到的桶，且 $u\colon \mathbb{Z}_n \to \mathbb{R}$。那么 Γ-分桶后的 u 是由以下定义的向量 $u_{|\Gamma} \in \mathbb{R}^b$

$$\left(u_{|\Gamma}\right)_i := u[\Gamma_i] = \sum_{j\in\Gamma_i} u_j \quad \text{for all } i \in [b] .$$

现在，我们证明方差的第二个分量可以通过分桶向量 p 的 2-范数和 3-范数来刻画。回顾一下，$\boldsymbol{T}_i \sim \mathrm{Poi}(mp_i)$ 表示顶点 $i \in \mathbb{Z}_n$ 在样本中出现的次数。我们首先计算构成检验统计量的各项 $\boldsymbol{X}_i(\boldsymbol{X}_i - 1)$ 的方差：

命题 2.11. 若 $\boldsymbol{X} \sim$ 泊松分布(λ)，则 方差$[\boldsymbol{X}(\boldsymbol{X}-1)] = 4\lambda^3 + 2\lambda^2$。

证明。该 泊松随机变量 $\boldsymbol{X}$ 具有以下原始矩（例如参见 [Rio37]）：

$$\begin{aligned}
\mathbb{E}[X] &= \lambda ,\\
\mathbb{E}\left[X^2\right] &= \lambda + \lambda^2 ,\\
\mathbb{E}\left[X^3\right] &= \lambda + 3\lambda^2 + \lambda^3 ,\\
\mathbb{E}\left[X^4\right] &= \lambda + 7\lambda^2 + 6\lambda^3 + \lambda^4 .
\end{aligned}$$

因此，我们有

$$\begin{aligned}
\operatorname{Var}[X(X-1)] &= \mathbb{E}\left[(X(X-1))^2\right] - \mathbb{E}[X(X-1)]^2 = \mathbb{E}\left[(X^2-X)^2\right] - (\mathbb{E}\left[X^2\right] - \mathbb{E}[X])^2\\
&= \mathbb{E}\left[X^4\right] - 2\mathbb{E}\left[X^3\right] + \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X^2\right]^2 + 2\mathbb{E}\left[X^2\right]\mathbb{E}[X] - \mathbb{E}[X]^2\\
&= (\lambda + 7\lambda^2 + 6\lambda^3 + \lambda^4) - 2(\lambda + 3\lambda^2 + \lambda^3) + (\lambda + \lambda^2) - (\lambda+\lambda^2)^2 + 2(\lambda+\lambda^2)\lambda - \lambda^2\\
&= 4\lambda^3 + 2\lambda^2 .
\end{aligned}$$

$\square$

引理 2.12。设 H 为一个具有诱导桶 $\Gamma = (\Gamma_1, \dots, \Gamma_b)$ 的子图，并设 p 为 $\mathbb{Z}_n$ 上的测度，则给定 $\boldsymbol{H} = H$ 时，$\boldsymbol{Y}$ 的条件方差满足

$$\operatorname*{Var}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H} = H] = 2\|p_{|\Gamma}\|_2^2 + 4m\|p_{|\Gamma}\|_3^3 .$$

证明。利用命题 2.11，所需的方差为

$$\begin{aligned}
\operatorname*{Var}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H} = H] &= \operatorname{Var}\left[\frac{1}{m}\sum_{i=1}^{b} \boldsymbol{X}_i(\boldsymbol{X}_i - 1)\right] = \frac{1}{m^2}\sum_{i=1}^{b}\left[4\left(mp[\Gamma_i]\right)^3 + 2\left(mp[\Gamma_i]\right)^2\right]\\
&= 4m\|p_{|\Gamma}\|_3^3 + 2\|p_{|\Gamma}\|_2^2 .
\end{aligned}$$

$\square$

2.5.5 相对集中度

我们分析中的一个主要工具是"相对集中度"，它将 p 在圆形区间 I 内的概率质量与边上的另一种测度 q 进行比较。

定义 2.13（相对集中度）。设 $\mathcal{I} \in \{\mathcal{I}^{\mathsf{cycle}}, \mathcal{I}^{\mathsf{path}}\}$，且 $p,\ q\colon \mathbb{Z}_n \to \mathbb{R}_{\geq 0}$ 是部分分布。设 $t \in \mathbb{R}$。于是我们定义

$$\rho_t(p \parallel q) := \max_{I \in \mathcal{I}:|I| \leq n} \frac{p[I]}{\max\{q[I^*], t\}}.$$

我们需要以下引理，它允许我们在假设相对集中度 $\rho_t(p \parallel q)$ 较高的情况下，找到一个区间 I，使得 $p[I]$ 与 $q[I^*]$ 之间存在较大差异。

引理 2.14. 设 $\mathcal{I} \in \{\mathcal{I}^{\mathsf{cycle}}, \mathcal{I}^{\mathsf{path}}\}$，且 p、$q\colon \mathbb{Z}_n \to \mathbb{R}_{\geq 0}$ 是部分分布。那么存在一个大小至多为 n 的 $I \in \mathcal{I}$，满足以下条件：

1. $q[I^*] \leq t$；且 *2.* $p[I] \geq \frac{1}{2} \cdot \underline{t} \cdot \rho_t(p \parallel q)$。

证明. 根据相对集中度的定义，存在一个大小至多为 n 的区间 $I \in \mathcal{I}$，使得以下条件之一成立

1. $q[I^*] \leq t$ 和 $p[I] = t\rho_t$（$p \parallel q$）；或
2. $q[I^*] \geq t$ 和 $p[I] = q[I^*]\rho_t$（$p \parallel q$）。

在前一种情况下，I 满足所需条件，因此得证。

因此，我们可以假设第二个条件成立。设 I 是满足 $q[I^*] \geq t$ 和 $p[I] \geq q[I^*]\rho_t(p \parallel q)$ 的最小区间（特别地，等号成立）。注意，必有 $|I| \geq 2$，否则 I^* 将为空，这与假设 $q[I^*] \geq t > 0$ 矛盾。现在我们考虑两种情况。

情况 1. 假设我们可以将 $I = L \cup R$ 分割为 $L,\ R$ 这些都是非空圆形区间，且满足以下两个条件之一[10]，将这对条件称为（$*$）：

1. $q[L^*], q[I^* \setminus L^*] \leq t$; or
2. $q[L^*], q[I^* \setminus L^*] \geq t$.

如果第一个条件成立，我们按如下方式完成证明。由于 $p[I] = p[L] + p[R]$ 和 $p[I] = q[I^*]\rho_t(p \parallel q) \geq t\rho_t(p \parallel q)$，因此必有 $p[L] \geq \frac{1}{2}t\rho_t(p \parallel q)$ 或 $p[R] \geq \frac{1}{2}t\rho_t(p \parallel q)$ 成立。在第一种情况下，L 满足所要求的条件。在第二种情况下，由于 $q[R^*] \leq q[I^* \setminus L^*] \leq t$，我们得出 R 满足所要求的条件，证明完毕。

如果第二个条件成立（这尤其意味着 $L^*,\ I^* \setminus L^*$ 均非空），注意到由于 $p[I] = p[L] + p[R]$ 和 $q[I^*] = q[L^*] + q[I^* \setminus L^*]$，我们有

$$\rho_t(p \parallel q) = \frac{p[I]}{q[I^*]} = \frac{p[L] + p[R]}{q[L^*] + q[I^* \setminus L^*]} \leq \max\left\{\frac{p[L]}{q[L^*]}, \frac{p[R]}{q[I^* \setminus L^*]}\right\}.$$

[10] 对于两个区间 $I_1 = \langle\!\langle i, d_1 \rangle\!\rangle$ 和 $I_2 = \langle\!\langle i, d_2 \rangle\!\rangle$，若满足 $d_1 > d_2 > 0$（即 I_2 是 I_1的非空前缀），我们用 $I_1 \setminus I_2$ 表示区间 $\langle\!\langle i + d_2, d_1 - d_2 \rangle\!\rangle$。

如果 $\frac{p[L]}{q[L^*]} \geq \rho_t(p \parallel q)$，那么由于 $q[L^*] \geq t$，我们得出 L 与 I的极小性矛盾。因此必有 $\frac{p[R]}{q[R^*]} \geq \frac{p[R]}{q[I^* \setminus L^*]} \geq \rho_t(p \parallel q)$。现在，如果 $q[R^*] \leq t$，那么 R 满足所需条件，因为我们有 $p[R] \geq q[I^* \setminus L^*]\rho_t(p \parallel q) \geq t\rho_t(p \parallel q)$。否则，如果 $q[R^*] > t$，那么 R与 I的极小性矛盾。这就完成了第一种情形下的证明。

情况 2。在第二种情况中，我们有：每个具有非空 L 和 R 的分割 $I = L \cup R$ 都不满足 (∗) 中的两个条件。记 $I = \langle\langle i, d \rangle\rangle$，其中如前所述，$d \geq 2$。考虑由 $L_j := \langle\langle i, j \rangle\rangle$ 给出的圆形区间序列 L_1、……、L_{d-1}，并记 $R_j := \langle\langle i + j, d - j \rangle\rangle = I \setminus L_i$，于是 $L_j^* = \langle\langle i, j - 1 \rangle\rangle$ 且 $R_j^* = \langle\langle i + j, d - j - 1 \rangle\rangle$。那么，对 $j \in [d - 1]$，每个 $L_j \cup R_j$ 都是 I 的一个分割，且其 L_j 和 R_j 非空，因此必定不满足 (∗) 中的两个条件。

现在注意到 $q[L_j^*]$ 随 j单调不减，且 $q[I^* \setminus L_j^*]$，$q[R_j^*]$ 随 j单调不增，而 $q[L_1^*] = q[\emptyset] = 0$，因此 $q[L_1^*] \leq q[I^* \setminus L_1^*]$。固定满足$q[L_j^*] \leq q[I^* \setminus L_j^*]$的最大指标 $j \in [d - 1]$。我们断言 $q[L_j^*]$，$q[R_j^*] \leq t$。

反设 $q[L_j^*] > t$。根据 j 的选取，我们有 $q[I^* \setminus L_j^*] > t$。但这样一来，分割 $I = L_j \cup R_j$ 满足 (∗) 中的第二个条件，矛盾。因此，我们已验证 $q[L_j^*] \leq t$。

现在反设 $q[R_j^*] > t$。那么 R_j^* 非空，这意味着 $j < d - 1$，并且可以验证 $R_j^* = I^* \setminus L_{j+1}^*$。因此 $q[I^* \setminus L_{j+1}^*] > t$。如果 $q[L_{j+1}^*] \geq t$，那么分割 $I = L_{j+1} \cup R_{j+1}$ 满足（∗）中的第二个条件，这再次导致矛盾。因此，我们可以假设 $q[L_{j+1}^*] < t$。但这意味着 $q[L_{j+1}^*] < q[I^* \setminus L_{j+1}^*]$，与我们对 j 的极大性选择相矛盾。因此，我们已经验证了 $q[R_j^*] \leq t$。

最后，回顾一下，$p[L_j] + p[R_j] = p[I] \geq q[I^*]\rho_t(p \parallel q) \geq t\rho_t(p \parallel q)$，因此$p[L_j] \geq \frac{1}{2}t\rho_t(p \parallel q)$或 $p[R_j] \geq \frac{1}{2}t\rho_t(p \parallel q)$成立。由于 $q[L_j^*]$，$q[R_j^*] \leq t$，可知 L_j或 R_j满足所需条件。 □

3 混淆收集器模型中的均匀性检验 1

沿用第 2 节中的设置，我们考虑在混淆收集器模型中，对基图 p 的顶点 $\mathbb{Z}_n$ 上的未知分布 $G = (\mathbb{Z}_n, E)$ 进行均匀性检验的任务。这里，G 是路径或环，并且每条边 $(i, i + 1) \in E$ 的权重为 $w(e) = \eta$，其中 η 是分辨率参数。

我们的分析将以统一的方式讨论环和路径两种情况。对于环，E 包含连接每个顶点 i 与 $i + 1$（模 n）的所有 n 边，相关区间集合为 $\mathcal{I} = \mathcal{I}^{\mathsf{cycle}}$。我们将用 $\mathbf{\Phi} = \mathbf{\Phi}^{\mathsf{cycle}}$ 表示相应的连接矩阵。对于路径，E 不包含连接 0 和 $n - 1$的边，相关区间集合为$\mathcal{I} = \mathcal{I}^{\mathsf{path}}$，连接矩阵为 $\mathbf{\Phi} = \mathbf{\Phi}^{\mathsf{path}}$。当结果取决于定义域的选择时，我们将明确说明所考虑的定义域。

测试器是算法 1，由以下两个步骤组成：

1. 集中度检验：检查样本中是否有任何计数过大；这种情况对应于高度集中的分布，可以予以拒绝。2. 基于碰撞的检验：根据检验统计量 $\boldsymbol{Y}$ 是否低于某个阈值来决定接受还是拒绝。

算法 1 困惑收集器模型中的均匀性检验器。令 $m \leftarrow c \cdot \frac{\sqrt{n}}{\epsilon^2} \cdot \log^2_{\eta^{3/2}} n$。常数：$\alpha$、$\beta$、$L$、$c > 0$ 将在后文定义。要求：$\eta \geq L \log^{4/5}_{n^{1/5}\epsilon^{4/5}} n$。1：过程 UniformityTester-ConfusedCollector (p, n, ϵ, η) 2：令 $X = (X_1, \ldots, X_n)$ 为第 2 节中针对从p获取的大小为 Poi(m) 的样本所定义的变量。3：如果 $\max_i X_i \geq \alpha \log n$，则拒绝。4：$Y \leftarrow \frac{1}{m} \sum_i X_i(X_i - 1)$。5：如果 $Y \geq \frac{m}{n^2} \sum_{i,j} \phi_i$, $_j + \beta \frac{m}{n} \epsilon^2 \eta$，则拒绝。6：接受。

关于基于碰撞的测试器最优性的说明。考虑到我们给出了一个泊松化测试器，其主要统计量 $\boldsymbol{Y}$ 等价于 [GR00] 在$\eta = 1$时的基于碰撞的统计量，因此我们声称样本复杂度为 $\widetilde{O}\left(\sqrt{n}/\epsilon^2\right)$ —而非 $O\left(\sqrt{n}/\epsilon^4\right)$—或许会令人惊讶；毕竟众所周知，对于基于界定 $\boldsymbol{Y}$ 的方差并应用切比雪夫不等式的分析，只有使用不同的检验统计量（例如修正卡方统计量 [CDVV14、DKN15b、VV17a]），或对非泊松化测试器 [DGPP19] 进行谨慎分析（另见其中第 2 节的说明），才能确立最优样本复杂度。我们的分析通过相对集中度检验隐式地避开了这个问题，该检验给出了$\|p\|_\infty$的上界；不过，化解这一表面矛盾的另一种方式是注意到，省去这项额外检验并将我们的证明专门用于 $\eta = 1$ 这一情形，只会引入对 $1/\epsilon^4$的依赖，而非$\sqrt{n}/\epsilon^4$；并且由于我们的分析只处理 $\epsilon \geq \widetilde{\Omega}\left(n^{-1/4}\right)$这一情形，$\sqrt{n}/\epsilon^2$ 项占主导地位 $1/\epsilon^4$。

3.1 简单情形：高度集中的分布

如果分布 p 在顶点的一个小连续区间 V内包含过多质量，我们就称其为“高度集中”。测试器会检测高度集中的分布并予以拒绝，而非高度集中的分布在我们对主检验统计量方差的分析中表现良好。具体而言，我们通过对相对集中度 $\rho_t(p \parallel w)$ 设定阈值来定义高度集中的分布，相关定义见第 2 节，其中 w 是由以下式子给出的边权重常数向量 $w(e) = \eta$。

定义 3.1（高度集中的分布）。给定常数 $C > 0$、正整数 m、分辨率参数 η，以及定义在 $\mathbb{Z}_n$ 上的概率分布 p，若 p 在关于 m 的分辨率 η 下满足 $\rho_t(p \parallel w) \geq C \log^2_m n$，其中 $t = 1 \log n$，则称其为 C-高度集中（分辨率为 η，关于 m）。

可以这样理解这个定义：分辨率为 η 的桶（连通分量）的平均大小为 $\approx 1/\eta$。在这样一个满足 $I^* = \langle\langle i, d-1 \rangle\rangle$ 的区间 $I = \langle\langle i, d \rangle\rangle$ 中，我们得到 $w[I^*] \approx 1$。如果 $p[I]/w[I^*] \geq \text{poly}\log(n)/m$，那么样本 Poi($mp[I]$) 应当会产生一个较大的条目 $\boldsymbol{X}_i$，因此算法应当能够拒绝这样的分布。另一方面，对于非常小的区间 I，若要确保算法很可能会拒绝，唯一的办法就是无论 $w[I^*]$ 有多小，$p[I]$ 仍然很大，这也促使我们在定义中选择 t。

注 3.2. 如果 p 不是 C-高度集中的，那么特别有 $\|p\|_\infty < C^{\text{对数}}_m n$；取区间 $I = \langle\langle i, 1 \rangle\rangle$ 分别对应每个 $i \in \mathbb{Z}_n$即可看出这一点。

现在我们证明，测试器的第一步能够以较高概率正确接受均匀分布，并拒绝高度集中的分布。因此，在分析测试器的第二步时，我们可以假设 p 不是高度集中的。我们需要以下辅助结果。

命题 3.3（桶几乎总是很小）。设 $K \geq 2$ 为一个常数，并假设 $n \geq 3$。那么由 $\boldsymbol{H}$ 诱导的桶 $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_{\boldsymbol{b}})$ 满足

$$|\boldsymbol{\Gamma}_i| \leq \frac{2K \log n}{\eta} \qquad \forall i \in [\boldsymbol{b}]$$

除概率至多为 $1/n^K$ 的情况外。

证明。令 d 为满足 $d \geq {}^{2K}\log_{\eta}{}^{n}$ 的最小整数，并固定任意区间 $I = \langle\langle i, d \rangle\rangle$。该区间内所有顶点都相连的概率是

$$\mathbb{P}[\boldsymbol{J}(I) = 1] = (1-\eta)^{d-1} \leq (1-\eta)^{\frac{2K \log n}{\eta} - 1} \leq (1-\eta)^{\frac{(2K-1)\log n}{\eta}} \leq e^{-(2K-1)\log n} = 1/n^{2K-1} \leq 1/n^{K+1},$$

其中我们利用了 $n \geq 3 \implies {}^{n}\log_{\eta} \geq 1$ 和 $K \geq 2 \implies 2K - 1 \geq K + 1$ 这一事实。现在，如果任意一个桶的大小至少为 d，那么存在某个区间 $I = \langle\langle i, d \rangle\rangle$ 满足 $J(I) = 1$。由于此类区间至多有 n 个，根据并集界，这一事件的概率至多为 $1/n^K$。 □

我们将 需要如下所述的泊松分布尾界，见 [Can17].

事实 3.4。令 $\boldsymbol{X} \sim$ 泊松分布(λ) 其中 $\lambda > 0$。那么对于任意 $t > 0$,

$$\mathbb{P}[\boldsymbol{X} \leq \lambda - t], \mathbb{P}[\boldsymbol{X} \geq \lambda + t] \leq e^{-\frac{t^2}{2(\lambda + t)}}.$$

以下结果假设 $m \leq n\eta$，这简化了分析，并且直观上对应于标准均匀性检验模型中的次线性样本复杂度区间。事实证明，对于我们考虑的参数值域，这一假设成立，但在更极端的情形下未必成立（见注 3.24）。

引理 3.5。对于充分大的常数 $\alpha > 0$ 以及所有充分大的 n，以下结论成立。假设 $m \leq n\eta$。那么对于定义在 $\mathbb{Z}_n$ 上的任意分布 p，有：

1. 若 p 是均匀的形式下，测试器的第一步仅以至多 $1/100$ 的概率拒绝；并且
2. 如果 p 是 4α-高度集中的，测试器的第一步以至少 $99/100$ 的概率拒绝.

证明。**完备性**。假设 p 是 $\mathbb{Z}_n$上的均匀分布。命题 3.3表明，除概率为 o 的情形外 **(1)**，每个桶的大小至多为 ${}^{4}\log_{\eta}{}^{n}$。假设这一点成立，并固定某个特定的桶 Γ_i。从该桶中采样到的元素数量服从 $\boldsymbol{X}_i \sim$泊松分布 $(mp[\Gamma_i])$ $=$泊松分布 $(m|\Gamma_i|/n)$。接着，利用 事实 3.4，对于 $\alpha > 16$，$\boldsymbol{X}_i$ 大到足以使测试器拒绝的概率为

$$\begin{aligned}
\mathbb{P}[\boldsymbol{X}_i \geq \alpha \log n] &\leq \mathbb{P}\left[\mathsf{Poi}\left(m \cdot \frac{4\log n}{\eta} \cdot \frac{1}{n}\right) \geq \alpha \log n\right] \\
&\leq \mathbb{P}[\mathsf{Poi}(4\log n) \geq \alpha \log n] && \text{（由于 } m \leq n\eta\text{）} \\
&= \mathbb{P}[\mathsf{Poi}(4\log n) - 4\log n \geq (\alpha - 4)\log n] \\
&\leq e^{-\frac{(\alpha-4)^2 \log^2 n}{2(4\log n + (\alpha-4)\log n)}} \leq e^{-\frac{(\alpha/2)^2 \log n}{2\alpha}} \leq e^{-2\log n} = 1/n^2.
\end{aligned}$$

因此，对于任意 $\boldsymbol{X}_i$，这种情况发生的概率至多为 $1/n = o(1)$。

可靠性。 假设 p 是 4α-高度集中。根据 引理 2.14和高集中性的定义，我们可以选取某个区间 $I = \langle\langle i, d\rangle\rangle \in \mathcal{I}$ 使其满足

1. $(d-1)\eta = w[\langle\langle i, d-1\rangle\rangle] \leq \frac{1}{\log n}$, and thus $|I| = d \leq 1 + \frac{1}{\eta \log n}$; and
2. $p[I] \geq \frac{1}{2} \cdot \frac{1}{m} \cdot \frac{4\alpha \log n}{m} = \frac{2\alpha \log n}{m}$.

我们首先声称，I 中的所有元素都将以高概率连接起来，也就是说，I^*中的每条边都会被采样到 $\boldsymbol{H}$中。事实上，根据并集界，我们有

$$\mathbb{P}[J(I) = 0] = \mathbb{P}[\exists e \in I^* : e \notin \boldsymbol{H}] \leq (|I| - 1) \cdot \eta \leq \frac{1}{\eta \log n} \cdot \eta = \frac{1}{\log n} = o(1).$$

现在，假设 I 中的每个元素都属于同一个桶，比如说 Γ_i。回想一下，随机变量 $\boldsymbol{X}_i \sim$ 泊松分布 $(mp[\Gamma_i])$ 表示从这个桶中抽取的元素数量，并且根据我们对 I的假设，有

$$mp[\Gamma_i] \geq m \cdot \frac{2\alpha \log n}{m} = 2\alpha \log n.$$

我们现在声称，以高概率，$\boldsymbol{X}_i > \alpha \log n$ 因此测试器将拒绝。事实上，根据事实 3.4，此情况不发生的概率为

$$\begin{aligned}\mathbb{P}[\boldsymbol{X}_i \leq \alpha \log n] &\leq \mathbb{P}[\mathsf{Poi}(2\alpha \log n) \leq \alpha \log n] = \mathbb{P}[\mathsf{Poi}(2\alpha \log n) \leq 2\alpha \log n - \alpha \log n] \\ &\leq e^{-\frac{\alpha^2 \log^2 n}{2(2\alpha \log n + \alpha \log n)}} = e^{-\frac{\alpha \log n}{6}} = o(1).\end{aligned}$$

□

现在我们继续讨论测试器的第二步，并分析检验统计量 $\boldsymbol{Y}$。

3.2 检验统计量的期望值

记号。 令 μ 表示 $\mathbb{Z}_n$上的均匀分布。当关于路径和环的陈述均成立时，我们将用 $\boldsymbol{\Phi}$ 表示随机连接矩阵，用 ϕ 表示其期望值。否则，我们会分别注明 $\boldsymbol{\Phi}^{\mathsf{path}}$ 或 $\boldsymbol{\Phi}^{\mathsf{cycle}}$。

先回想 到，如命题 2.6所示，我们可以将 $\boldsymbol{Y}$ 的期望值写作

$$\mathbb{E}[\boldsymbol{Y}] = mp^\top \phi p. \tag{3}$$

当 G 是路径时，期望连接矩阵 $\phi^{\mathsf{path}} := \mathbb{E}\left[\boldsymbol{\Phi}^{\mathsf{path}}\right]$ 可以用 η 简洁地表示。我们将 $1 - \eta$ 这一量命名为每条边包含在 $\boldsymbol{H}$ 中的概率：

$$\nu := 1 - \eta.$$

命题 *3.6。 </style}*矩阵 ϕ^{path} 为

$$\phi^{\mathsf{path}}_{i,j} = \nu^{|i-j|}$$

对每个 $i, j \in \mathbb{Z}_n$。

证明。 这里，相关区间为 $\mathcal{I} = \mathcal{I}^{\mathsf{path}}$。因此，对于任意 $i < j$， i 和 j 位于同一个桶中，当且仅当它们之间的每条边都在 $\boldsymbol{H}$ 中：

$$\phi^{\mathsf{path}}_{i,j} = \mathbb{P}\left[\boldsymbol{\Phi}^{\mathsf{path}}_{i,j} = 1\right] = \mathbb{P}[\forall e \in \langle\langle i, j - i + 1\rangle\rangle^* : e \in \boldsymbol{H}] = \nu^{j-i}.$$

□

当 G 是环时，因此期望连接矩阵为 $\phi^{\mathsf{cycle}} := \mathbb{E}\left[\mathbf{\Phi}^{\mathsf{cycle}}\right]$，我们需要考虑连接 i 和 j的小区间和大区间（采用 第 2.5.1 节中的记号。

命题 3.7。 矩阵 ϕ^{cycle} 由下式给出

$$\phi^{\mathsf{cycle}}_{i,j} = \nu^{|i-j|} + \nu^{n-|i-j|} - \nu^n \,.$$

证明。 集合 $\mathsf{small}(i,\ j)^*$ 和集合 $\mathsf{large}(i,\ j)^*$ 的大小分别为 $\min\{|i-j|, n-|i-j|\}$ 和 $\max\{|i-j|, n-|i-j|\}$（回想一下，它们构成了环的边的一个分割）。根据容斥原理，

$$\begin{aligned}
\phi^{\mathsf{cycle}}_{i,j} &= \mathbb{P}\left[\mathbf{\Phi}^{\mathsf{cycle}}_{i,j} = 1\right] \\
&= \mathbb{P}\left[\boldsymbol{J}(\mathsf{small}(i,j)) = 1\right] + \mathbb{P}\left[\boldsymbol{J}(\mathsf{large}(i,j)) = 1\right] - \mathbb{P}\left[\forall e \in E : e \in \boldsymbol{H}\right] \\
&= \nu^{|i-j|} + \nu^{n-|i-j|} - \nu^n \,.
\end{aligned}$$

□

我们希望证明，当 $\mathbb{E}[\boldsymbol{Y}]$ 远离均匀分布时，它的值较大。令 $p = \mu + z$，其中$z \in \mathbb{R}^{\mathbb{Z}_n}$。如 命题 2.7所示，我们可以将期望分解为

$$\mathbb{E}[\boldsymbol{Y}] = m\mu^\top \phi \mu + 2m\mu^\top \phi z + mz^\top \phi z \,. \tag{4}$$

令 $\boldsymbol{Y}^{(\mu)}$ 表示当 p 服从均匀分布时的检验统计量。我们注意到，上式的第一项正是算法 1 为检验统计量设定阈值时所依据的基准：

命题 3.8（均匀情形下 $\boldsymbol{Y}$ 的期望）。 当 $p = \mu,\ \boldsymbol{Y} = \boldsymbol{Y}^{(\mu)}$ 满足

$$\mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] = m\mu^\top \phi \mu = \frac{m}{n^2} \sum_{i,j} \phi_{i,j} \,.$$

证明。 该结论由（3）以及 $p = \mu = \vec{1}/n$ 的假设得出。 □

因此，我们的策略是证明：1） ϕ 的最小特征值较大，因此当 $\|z\|_2^2$ 较大时， $z^\top \phi z$ 也会较大；2）项 $\mu^\top \phi z$ 的绝对值较小（事实上，当 G为环时，该项为零），因此它不会对总和产生太大影响。我们首先着手实现第一个目标。 ϕ^{path} 和ϕ^{cycle} 都具有良好的性质（它们分别是托普利兹矩阵和循环矩阵），我们将依次对每个矩阵的最小特征值进行界定。令 $\lambda_{\min}(\cdot)$和 $\lambda_{\max}(\cdot)$分别表示一个（实对称）矩阵的最小特征值和最大特征值。

引理 3.9（ ϕ^{path} 的最小特征值）。 设 $\eta \in (0, 1]$。那么 $\lambda_{\min}(\phi^{\mathsf{path}}) > \eta/2$。

证明。 矩阵 ϕ^{path} 是一个对称托普利兹矩阵，其逆矩阵可按照 [Sra]中的方法求得。回顾 $\nu = 1 - \eta$，因此 $0 \le \nu < 1$ 以及 $\phi^{\mathsf{path}}_{i,\ j} = \nu^{|i-j|}$ 根据 命题 3.6。于是，ϕ^{path} （简记为 ϕ^{-1} ）的逆矩阵如下，是一个三对角矩阵：

$$\phi^{-1} = \frac{1}{1-\nu^2} \cdot \begin{bmatrix}
1 & -\nu & 0 & 0 & \cdots & 0 \\
-\nu & 1+\nu^2 & -\nu & 0 & \cdots & 0 \\
0 & -\nu & 1+\nu^2 & -\nu & \cdots & 0 \\
 & \ddots & \ddots & \ddots & \ddots & \\
0 & \dots & \dots & -\nu & 1+\nu^2 & -\nu \\
0 & \dots & \dots & 0 & -\nu & 1
\end{bmatrix} .$$

现在，我们可以使用 Gershgorin 圆盘定理得到 ϕ^{-1} 的最大特征值的上界：

$$\lambda_{\max}\left(\phi^{-1}\right) \le \max_{i\in\mathbb{Z}_n}\left\{\phi^{-1}_{i,i} + \sum_{j\neq i}\left|\phi^{-1}_{i,j}\right|\right\} \le \left(\frac{1}{1-\nu^2}\right)\cdot\left(1+\nu^2+2\nu\right) = \frac{1+\nu}{1-\nu}\,.$$

因此，我们得到

$$\lambda_{\min}(\phi^{\mathsf{path}}) = \frac{1}{\lambda_{\max}\left(\phi^{-1}\right)} \ge \frac{1-\nu}{1+\nu} > \frac{1-\nu}{2} = \eta/2\,. \qquad \square$$

当 G 是环时，使用 ϕ^{cycle}的简化紧近似会更方便。本质上，我们希望忽略大区间，转而使用由下式给出的矩阵 ϕ^{small}

$$\phi^{\mathsf{small}}_{i,j} := \nu^{|\mathsf{small}(i,j)|-1}\,.$$

我们需要注意到，在参数 η的值域内， $\zeta(\mathcal{I})$可以忽略不计。

命题 3.10. 设 $\eta \ge \Omega(n^{-1/5})$，且 $K>0$ 为任意常数。则对所有充分大的 n，

$$\zeta(\mathcal{I}) \le \nu^{n/2} = o(n^{-K})\,.$$

证明. 第一个不等式来自命题 2.5。第二个不等式容易验证：

$$\nu^{n/2} = (1-\eta)^{n/2} \le e^{-\eta\cdot n/2} \le e^{-\Omega(n^{-1/5}\cdot n)} = e^{-\Omega(n^{4/5})} = o(n^{-K})\,. \qquad \square$$

现在，我们可以对 ϕ^{small} 和 ϕ^{cycle}的特征值给出下界了。我们先对$\lambda_{\min}(\phi^{\mathsf{small}})$给出下界，然后证明近似误差可以忽略不计。

事实 3.11 （循环矩阵的特征值；参见 [Gra06]）。 设 c_0，c_1，……，$c_{n-1}\in\mathbb{R}$。则矩阵

$$M = \begin{bmatrix} c_0 & c_1 & c_2 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & \cdots & c_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_2 & c_3 & c_4 & \cdots & c_1 \\ c_1 & c_2 & c_3 & \cdots & c_0 \end{bmatrix}$$

由 $M_{j,\,k} = c_{(k-j)\ \text{模}\ n}$ 给出，其特征值为

$$\lambda_\ell = \sum_{k=0}^{n-1} c_k\omega^{\ell k} \qquad \ell = 0,1,\ldots,n-1\,,$$

其中 $\omega = e^{-\frac{2\pi i}{n}}$是单位根的一个原始 n次根。

引理 3.12 （的最小特征值 ϕ^{small}）。 设 $\eta\in(0,1]$ 满足 $\eta \ge \Omega(n^{-1/5})$。则对于所有充分大的 n，$\lambda_{\min}\left(\phi^{\mathsf{small}}\right) > \eta/3$。

证明。 首先假设 n 为奇数。对于每个 $k=0,1,\ldots,n-1$，令 $c_k := \phi^{\mathsf{small}}_{0\,k}$，因此 ϕ^{small}是一个具有 事实 3.11所述形式的对称循环矩阵。特别地，为方便起见，令 $h := \lfloor n/2\rfloor$，则有

$$c_k = \begin{cases} \nu^k & \text{if } k\le h \\ \nu^{n-k} & \text{if } k > h\,. \end{cases}$$

因此，对于每个 $\ell = 0, 1, \ldots, n-1$，特征值 λ_ℓ 为

$$\begin{aligned}\lambda_\ell &= \sum_{k=0}^{n-1} c_k \omega^{\ell k} = c_0 + \sum_{k=1}^{h} \nu^k \omega^{\ell k} + \sum_{k=h+1}^{n-1} \nu^{n-k}\omega^{\ell k} = 1 + \sum_{k=1}^{h} \nu^k \omega^{\ell k} + \sum_{k=1}^{h} \nu^{n-(n-k)}\omega^{\ell(n-k)} \\ &= 1 + \sum_{k=1}^{h} \nu^k \omega^{\ell k} + \sum_{k=1}^{h} \nu^k \omega^{-\ell k} = 1 + \frac{\nu\omega^\ell - \nu^{h+1}\omega^{\ell(h+1)}}{1-\nu\omega^\ell} + \frac{\nu\omega^{-\ell} - \nu^{h+1}\omega^{-\ell(h+1)}}{1-\nu\omega^{-\ell}} \\ &= \frac{1 - \nu(\omega^\ell + \omega^{-\ell}) + \nu^2 + \nu\omega^\ell - \nu^2 - \nu^{h+1}\omega^{\ell(h+1)} + \nu^{h+2}\omega^{\ell h} + \nu\omega^{-\ell} - \nu^2 - \nu^{h+1}\omega^{-\ell(h+1)} + \nu^{h+2}\omega^{-\ell h}}{1 - \nu(\omega^\ell + \omega^{-\ell}) + \nu^2} \\ &= \frac{1 - \nu^2 - \nu^{h+1}(\omega^{\ell(h+1)} + \omega^{-\ell(h+1)}) + \nu^{h+2}(\omega^{\ell h} + \omega^{-\ell h})}{1 - \nu(\omega^\ell + \omega^{-\ell}) + \nu^2} = \frac{(1-\nu)(1+\nu) \pm O(\nu^{n/2})}{1 - 2\nu\cos(2\pi\ell/n) + \nu^2}\,,\end{aligned}$$

其中最后一步用到了恒等式 $e^{i\theta} + e^{-i\theta} = 2\cos(\theta)$。因此，回顾 $\nu = 1 - \eta \in [0, 1)$，我们得出 λ_ℓ 的下界为

$$\frac{(1-\nu)(1+\nu) \pm O(\nu^{n/2})}{1 - 2\nu\cos(2\pi\ell/n) + \nu^2} \geq \frac{(1-\nu)(1+\nu)}{(1+\nu)^2} - \frac{O(\nu^{n/2})}{(1-\nu)^2} \geq \frac{\eta}{2} - \frac{O(\nu^{n/2})}{(1-\nu)^2}\,.$$

然后，利用 命题 3.10，

$$\frac{O(\nu^{n/2})}{(1-\nu)^2} \leq \frac{o(n^{-3/5})}{\eta^2} \leq o(n^{-3/5} \cdot n^{2/5}) = o(n^{-1/5}) = o(\eta)\,,$$

因此 $\lambda_\ell > \eta/3$。当 n 为偶数时，同样的论证适用，但会多出一项阶为 $O(\nu^{n/2})$的项，这不影响渐近分析。 □

引理 3.13。 设 $\eta \in (0, 1]$ 满足 $\eta \geq \Omega(n^{-1/5})$。那么，对所有充分大的 n， $\lambda_{\min}(\phi^{\mathsf{cycle}}) > \eta/4$。

证明。设 $\phi^{\mathsf{err}} := \phi^{\mathsf{cycle}} - \phi^{\mathsf{small}}$。容易验证， $\lambda_{\min}(\phi^{\mathsf{cycle}}) \geq \lambda_{\min}(\phi^{\mathsf{small}}) + \lambda_{\min}(\phi^{\mathsf{err}})$。由于根据引理 3.12， $\lambda_{\min}(\phi^{\mathsf{small}}) > \eta/3$，只需证明 $\lambda_{\min}(\phi^{\mathsf{err}}) > -o(\eta)$。由于根据命题 3.7， $\phi^{\mathsf{cycle}}_{i,j} = \nu^{|\mathsf{small}(i,j)|-1} + \nu^{|\mathsf{large}(i,j)|-1} - \nu^n$，我们得到

$$\phi^{\mathsf{err}}_{i,j} = \nu^{|\mathsf{large}(i,j)|-1} - \nu^n$$

对所有 $i, j \in \mathbb{Z}_n$。根据 $\zeta(\mathcal{I})$ 的定义，并回顾 命题 2.5，我们得出

$$\|\phi^{\mathsf{err}}\|_\infty \leq \zeta(\mathcal{I}) \leq \nu^{n/2}\,.$$

根据 Gershgorin 圆盘定理和 命题 3.10，

$$\lambda_{\min}(\phi^{\mathsf{err}}) \geq \min_{i \in \mathbb{Z}_n} \left\{ \phi^{\mathsf{err}}_{i,i} - \sum_{j \neq i} |\phi^{\mathsf{err}}_{i,j}| \right\} > -n \cdot \nu^{n/2} \geq -n \cdot o(n^{-6/5}) = -o(n^{-1/5}) \geq -o(\eta)\,. \qquad \square$$

我们使用 ϕ 的最小特征值来证明：如果 $\|p - \mu\|_2^2$ 很大，那么 $\mathbb{E}[\boldsymbol{Y}]$ 也很大。以下关于期望值的中间表述将很有用。

命题 3.14. 对于所有充分大的 n，以下结论成立。设 p 是 $\mathbb{Z}_n$上的一个分布，使得 $\mathsf{dist}_{\mathsf{TV}}(p, \mu) > \epsilon$，并记 $p = \mu + z$。则

$$\mathbb{E}[\boldsymbol{Y}] > \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] + m\frac{\eta}{8}\|z\|_2^2 + \frac{m}{2n}\epsilon^2\eta + 2m\mu^\top\phi z\,.$$

证明. 首先，结合（4）和 命题 3.8，再利用 $x^\top Mx \geq \lambda_{\min}(M)\|x\|_2^2$对任意对称矩阵 M 和向量 x 以及引理 3.9 和 3.13 给出的特征值界，得到

$$\begin{aligned}\mathbb{E}[\boldsymbol{Y}] &= m\mu^\top\phi\mu + 2m\mu^\top\phi z + mz^\top\phi z \\ &\geq \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] + m\lambda_{\min}(\phi)\|z\|_2^2 + 2m\mu^\top\phi z \\ &\geq \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] + m\frac{\eta}{4}\|z\|_2^2 + 2m\mu^\top\phi z \\ &= \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] + m\frac{\eta}{8}\|z\|_2^2 + m\frac{\eta}{8}\|z\|_2^2 + 2m\mu^\top\phi z\,.\end{aligned}$$

那么，由于 $\|z\|_1 = 2\mathsf{dist}_{\mathsf{TV}}(p,\ \mu) > 2\epsilon$，我们有 $\|z\|_2^2 \geq \left(\frac{2\epsilon}{n}\right)^2 \cdot n = 4\epsilon^2/n$，证明得证。 □

现在我们证明交叉项 $\mu^\top\phi z$ 很小。当 G 是环时，这一项实际上为零；当 G 是路径时，由于靠近端点或中点的顶点之间存在不对称性，交叉项会产生影响。不过，只要 $\|z\|_\infty$ 不太大，就不会有问题；而当 p 没有高度集中时，这一条件确实成立。

命题 3.15。 设 $\eta \in (0,1]$，并设 $\phi = \mathbb{E}[\boldsymbol{\Phi}]$ 为相应的期望连接矩阵。令$\delta > 0$ 为任意正实数。那么，对于任意满足以下条件的 $z \in \mathbb{R}^n$

1. $\sum_i z_i = 0$; *and*
2. $\|z\|_\infty \leq \delta$,

事实是

$$\left|\mu^\top\phi z\right| \leq \frac{2\delta}{n\eta^2}\,.$$

证明。 当 $\phi = \phi^{\mathsf{cycle}}$ 时，我们有 $\mu^\top\phi$ 是一个常数向量（这对任何循环矩阵都成立），因此 $\mu^\top\phi z = 0$ （因为 $\sum_i z_i = 0$）。所以现在可以假设 $\phi = \phi^{\mathsf{path}}$。

注意，由于左端表达式中 z 和 $-z$ 具有对称性，只需给出 $\mu^\top\phi z$ 的上界。我们将此表达式展开如下:

$$\mu^\top\phi z = \sum_{i,j\in\mathbb{Z}_n}\mu_i z_j\phi_{i,j} = \frac{1}{n}\sum_{j=0}^{n-1} z_j\left(\sum_{i=0}^{n-1}\phi_{i,j}\right).$$

因此，我们的目标是证明

$$\sum_{j=0}^{n-1} z_j S_j \overset{?}{\leq} \frac{2\delta}{\eta^2}\,, \tag{5}$$

其中 $S_j := \left(\sum_{i=0}^{n-1}\phi_{i,j}\right)$ 是 j 的第 ϕ 列中各项之和。注意，$(S_j)_{j=0,\ \ldots,\ n-1}$是一个对称单峰序列（先递增后递减），除指标 $\lfloor (n-1)/2\rfloor$ 和 $\lceil (n-1)/2\rceil$ 在 n 为偶数时外，其余位置的相邻项均严格不等。我们将采用“重排与饱和”论证，构造一个向量 z^*，使其成为式 (5) 左端（下文称为目标值）的上界。

设 z' 是满足题述条件（下文称为可行解）且使目标值最大的向量。设 σ 是 $\{0,\ldots,n-1\}$ 的一个排列，使列和序列按非递减顺序排列： $S_{\sigma(0)} \leq \cdots \leq S_{\sigma(n-1)}$。那么我们也可以假设 z' 遵循这一顺序： $z'_{\sigma(0)} \leq \cdots \leq z'_{\sigma(n-1)}$；否则，重新排列 z' 的条目，就会得到另一个目标值相等或更大的可行解。

现在我们来论证：可以假设在 z' 的所有非零条目中，除了至多一个正条目和一个负条目外，其余条目的绝对值都等于 δ （我们称之为饱和项）。事实上，如果两个在 σ 下相邻且符号相同的非零条目未饱和，即它们满足 $\left|z'_{\sigma(i)}\right|, \left|z'_{\sigma(i+1)}\right| < \delta$，那么我们可以通过“饱和”这对条目，即减小 $z'_{\sigma(i)}$ 并增大 $z'_{\sigma(i+1)}$，直到其中一个达到 $\{-\delta, 0, \delta\}$ 中的某个值，从而得到另一个目标值相等或更大的可行解。

我们声称，也可以假设 z' 中正项的值构成的多重集等于 z' 中负项绝对值构成的多重集。设 z' 中有 N^+ 个等于 δ 的项， N^- 个等于 $-\delta$ 的项， $M^+ \in \{0,1\}$ 个项位于区间 $(0, \delta)$ 内，且有 $M^- \in \{0,1\}$ 个项位于区间 $(-\delta, 0)$ 内。若 $N^+ = N^-$，则由于 $\sum_j z'_j = 0$，我们还必须有 $M^+ = M^-$；并且若该值为 1，则 z' 中对应的项必须具有相同的绝对值，才能使它们的和为零。另一方面，若 $N^+ \neq N^-$，不妨设 $N^+ > N^-$，则 $\sum_i z'_i > 0$，因为 z' 中饱和值之和至少为 δ，而未饱和值之和必须位于区间 $(-\delta, \delta)$ 内。这与 z' 是一个可行解这一事实矛盾。

现在，我们通过将 z' 中剩余的（零个或两个）项饱和，来构造 z^*：

$$z_i^* := \begin{cases} \delta, & \text{if } z'_i > 0 \\ -\delta, & \text{if } z'_i < 0 \\ 0, & \text{if } z'_i = 0. \end{cases}$$

然后，依据与上文相同的论证， z^* 的目标值不小于 z'。现在，我们用式（5）的右端来给出该目标值的上界，从而完成论证。

设 z^* 中正项的个数为 N。根据构造，我们有

$$N = |\{i \in \mathbb{Z}_n : z_i^* = \delta\}| = |\{i \in \mathbb{Z}_n : z_i^* = -\delta\}|.$$

那么我们的目标值为

$$\sum_j z_j^* S_j = \delta \left[-\sum_{i=0}^{N-1} S_{\sigma(i)} + \sum_{i=0}^{N-1} S_{\sigma(n-1-i)} \right]. \tag{6}$$

设 $h := (n-1)/2$。由于 $(S_j)_{j=0,\ \ldots,\ n-1}$ 是一个在中点处达到最大值的对称单峰序列，因此不妨设式 (i) 右端第一个求和项中的指标6为 $\{0, \ldots, \lceil N/2 \rceil - 1\} \cup \{n-1, \ldots, n - \lfloor N/2 \rfloor\}$。对于指标 $\sigma(n-1-i)$，精确的计数取决于 n的奇偶性，但只需使用指标 $\{\lceil h \rceil, \ldots, \lceil h \rceil + \lceil N/2 \rceil - 1\} \cup \{\lfloor h \rfloor, \ldots, \lfloor h \rfloor - \lfloor N/2 \rfloor + 1\}$就能增大目标值。注意，当 n为奇数时，这种选择会因两次使用最大值 S_h 而略微高估目标值，但这个较松的界已足以满足我们的需要。

因此，我们终于可以给出并计算任意可行向量 z 的目标值上界。回想一下，$\phi_{i,\ j} = \nu^{|i-j|}$。在 $\eta = 1$，因而 $\nu = 0$ 的边界情况下，我们有 ϕ 是单位矩阵，因此对于每个 $j \in \mathbb{Z}_n$，都有 $S_j = 1$。因此我们得到

$$\sum_{j=0}^{n-1} z_j S_j = \sum_{j=0}^{n-1} z_j = 0,$$

它满足（5），至此证明完毕。现在，假设 $0 < \nu < 1$。那么

$$\sum_{j=0}^{n-1} z_j S_j \le \delta \left[-\sum_{i=0}^{N-1} S_{\sigma(i)} + \sum_{i=0}^{N-1} S_{\sigma(n-1-i)} \right]$$
$$\le \delta \left[-\sum_{i=0}^{\lceil N/2 \rceil - 1} S_i - \sum_{i=0}^{\lfloor N/2 \rfloor - 1} S_{n-1-i} + \sum_{i=0}^{\lceil N/2 \rceil - 1} S_{\lceil h \rceil + i} + \sum_{i=0}^{\lfloor N/2 \rfloor - 1} S_{\lfloor h \rfloor - i} \right] \le \frac{2\delta}{\eta^2},$$

其中我们定义 了最后一个不等式中繁琐的几何级数计算，以命题 B .1. □

引理 3.16. 设 $C > 0$，且设 p 是 $\mathbb{Z}_n$ 上的一个分布，它并非 C-高度集中。那么只要 $\frac{8C \log}{\eta^3 \epsilon^2}^n \le m \le Cn \log n$，以下结论成立：

$$\left| 2m\mu^\top \phi z \right| \le \frac{m}{2n} \epsilon^2 \eta .$$

证明。 记 $p = \mu + z$。由于 $\|p\|_1 = \|\mu\|_1 = 1$，可知 $\sum_i z_i = 0$，满足命题 3.15的第一个条件。我们将证明 z 也满足参数为 δ 的第二个条件：$= C \log nm$ 。

事实上，由于 $0 \le \frac{1}{n} + z_i \le \frac{C}{\log_m} n$ (第二个不等式由 注 3.2 得出），我们得到，一方面，

$$z_i \ge -\frac{1}{n} \ge -\frac{C \log n}{m},$$

这里我们用到了假设 $m \le Cn \log n$，另一方面，

$$z_i \le \frac{C \log n}{m} - \frac{1}{n} \le \frac{C \log n}{m},$$

因此 $\|z\|_\infty \le \frac{C}{\text{对数}_m} n$ 如所需。 命题 3.15 蕴含

$$\left| \mu^\top \phi z \right| \le \frac{2C \log n}{mn\eta^2} .$$

最后，只需将这个不等式与我们假设的 m 下界结合起来。我们得到

$$\left| 2m\mu^\top \phi z \right| \le 2m \cdot \frac{2C \log n}{mn\eta^2} = \frac{8C \log n}{\eta^3 \epsilon^2} \cdot \frac{\epsilon^2 \eta}{2n} \le \frac{m}{2n} \epsilon^2 \eta . \qquad \square$$

我们结合前面的结果，证明 $\boldsymbol{Y}$ 的期望值差异满足所需条件：

引理 3.17 ($\boldsymbol{Y}$ 的期望值差异)。设 C、$c > 0$ 为常数，且设 $m = c \cdot \frac{\sqrt{n}}{\epsilon^2 \eta^{3/2}}$ 对数$^2 n$。那么，对于所有充分大的 n 以及所有满足 $\eta \ge (c/C)^{2/3}$ 对数$\frac{2/3}{n^{1/3} \epsilon^{4/3}} n$ 的 ϵ、$\eta \in (0, 1]$，以下结论成立。设 p 是 $\mathbb{Z}_n$ 上的一个分布，且该分布并非 C-高度集中，并满足$\text{dist}_{\text{TV}}(p, \mu) > \epsilon$。记 $p = \mu + z$。那么，检验统计量 $\boldsymbol{Y}$ 满足

$$\mathbb{E}[\boldsymbol{Y}] > \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] + \frac{m\eta}{8} \|z\|_2^2 .$$

证明。 由 命题 3.14，我们有

$$\mathbb{E}[\boldsymbol{Y}] > \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] + \frac{m\eta}{8} \|z\|_2^2 + \frac{m}{2n} \epsilon^2 \eta + 2m\mu^\top \phi z .$$

因此，只要我们能证明 $2m\mu^\top \phi z \geq -\frac{m}{2n}\epsilon^2\eta$，就完成了。这可直接由引理 3.16 得出，前提是我们能验证 m上的前置条件。我们先检查下界：

$$m \geq \frac{8C\log n}{\eta^3\epsilon^2} \iff c\frac{\sqrt{n}}{\epsilon^2}\cdot\frac{\log^2 n}{\eta^{3/2}} \geq \frac{8C\log n}{\eta^3\epsilon^2} \iff \eta \geq (8C/c)^{2/3}\frac{1}{n^{1/3}\log^{2/3} n},$$

根据我们对 η 的假设，当 n 足够大时，该结论成立。至于上界，

$$m \leq Cn\log n \iff c\frac{\sqrt{n}}{\epsilon^2}\cdot\frac{\log^2 n}{\eta^{3/2}} \leq Cn\log n \iff \eta \geq (c/C)^{2/3}\frac{\log^{2/3} n}{n^{1/3}\epsilon^{4/3}},$$

根据假设，该结论成立。因此引理 3.16 适用，证毕。 □

注 3.18. 上述陈述中的条件 $\eta = \Omega\left(\log^{2/3}_{n^{1/3}\epsilon^{4/3}} n\right)$ 在本论证所考虑的参数值域内成立。具体而言，当 $\eta = \Omega\left(\log^{4/5}_{n^{1/5}\epsilon^{4/5}} n\right)$ 且 $\epsilon = \Omega\left(\frac{1}{n^{1/4}}\right)$ 时，该条件成立，因为

$$\frac{\left(\frac{\log^{4/5} n}{n^{1/5}\epsilon^{4/5}}\right)}{\left(\frac{\log^{2/3} n}{n^{1/3}\epsilon^{4/3}}\right)} = (\log n)^{2/15} n^{2/15}\epsilon^{8/15} \geq (\log n)^{2/15} n^{2/15}\Omega(n^{-2/15}) = \omega(1).$$

3.3 检验统计量的集中

我们应用第 2.5.3 节中给出的一般结果，来给 $\boldsymbol{Y}$的方差取上界。

引理 3.19 (方差的第一部分)。 设 $C > 0$ 为一个常数， $n \in \mathbb{N}$ 足够大，并假设 p 是 $\mathbb{Z}_n$ 上的一个概率分布，且不是 C-高度集中的。假设 $m \leq \mathrm{poly}(n)$ 且 $\eta \geq \Omega(n^{-1/5})$。则

$$\operatorname*{Var}_{\boldsymbol{H}}\left[\mathop{\mathbb{E}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right] \leq \frac{\|p\|_2^2}{\eta}\cdot O\left(\log^4 n\right).$$

证明。 根据 引理 2.9，存在某个常数 $c > 0$ 使得

$$\operatorname*{Var}_{\boldsymbol{H}}\left[\mathop{\mathbb{E}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right] \leq 5m^2\zeta(\mathcal{I})\|p\|_1^4 + cm^2\cdot\sum_{\substack{I=\langle\langle i,d\rangle\rangle\in\mathcal{I}\\ 1\leq d\leq n}} p_i p_{i+d-1} p[I]^2\mathbb{E}[\boldsymbol{J}(I)].$$

我们从右端的第二个分量开始。回想一下，对于任意 $I = \langle\langle i,d\rangle\rangle$, $\mathbb{E}[\boldsymbol{J}(I)] \leq \nu^{d-1}$

（这个值可能是　如果 $\mathcal{I}=\mathcal{I}^{\mathsf{path}}$ 和 I 跨越顶点 0 与 $n-1$ 之间的边，则值为零。我们有。　ve

$$\sum_{\substack{I=\langle\!\langle i,d\rangle\!\rangle\in\mathcal{I}\\1\le d\le n}} p_i p_{i+d-1} p[I]^2 \mathbb{E}\left[\boldsymbol{J}(I)\right] \le \sum_{i=0}^{n-1}\sum_{d=1}^{n} p_i p_{i+d-1} p[\langle\!\langle i,d\rangle\!\rangle]^2 \nu^{d-1}$$

$$\le \sum_{i=0}^{n-1}\sum_{d=1}^{n} p_i p_{i+d-1}\nu^{d-1}\left[\frac{C\log^2 n}{m}\cdot\max\left\{\eta(d-1),\frac{1}{\log n}\right\}\right]^2 \qquad (p \text{ is not } C\text{-highly-concentrated})$$

$$\le \sum_{i=0}^{n-1}\sum_{d=1}^{n} p_i p_{i+d-1}\nu^{d-1}\left[\frac{C^2\log^4 n}{m^2}\cdot\left(\eta^2 d^2+\frac{1}{\log^2 n}\right)\right]$$

$$= \frac{C^2\log^2 n}{m^2}\left(\eta^2\log^2(n)\sum_{d=1}^{n} d^2\nu^{d-1}\sum_{i=0}^{n-1} p_i p_{i+d-1}+\sum_{d=1}^{n}\nu^{d-1}\sum_{i=0}^{n-1} p_i p_{i+d-1}\right)$$

$$\le \frac{C^2\|p\|_2^2\log^2 n}{m^2}\left(\eta^2\log^2(n)\sum_{d=1}^{n} d^2\nu^{d-1}+\sum_{d=1}^{n}\nu^{d-1}\right) \qquad (\text{by Cauchy-Schwarz})$$

$$\le \frac{C^2\|p\|_2^2\log^2 n}{m^2}\left(\eta^2\log^2(n)\frac{1+\nu}{\eta^3}+\frac{1}{\eta}\right) \qquad (\text{since } \eta=1-\nu)$$

$$\le \frac{3C^2\|p\|_2^2\log^4 n}{m^2\eta} \qquad (\text{since } \nu<1)$$

$$= \frac{1}{m^2}\cdot\frac{\|p\|_2^2}{\eta}\cdot O\left(\log^4 n\right),$$

如所需。于是，只需证明项 $5m^2\zeta(\mathcal{I})\|p\|_1^4$ 是 $O(1/n)$，因为 $\|p\|_2^2\ge\|\mu\|_2^2=1/n$。令 $K>0$ 为一个常数，使得 $m\le n^K$ 对于充分大的 n成立，根据 $m\le\mathrm{poly}(n)$ 的假设。于是我们有 $\|p\|_1^4=1$ 以及 $m^2\zeta(\mathcal{I})=o(m^2/n^{2K+1})\le o(1/n)$，根据 命题 3.10，正如所需。 □

引理 3.20 (方差的第二个分量)。 设 $C>0$ 为一个常数，设 $n\in\mathbb{N}$ 足够大，并假设 p 是定义在 $\mathbb{Z}_n$ 上的概率分布，且不是 C-高度集中的。假设$m\le$poly(n)。那么

$$\mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\mathop{\mathrm{Var}}_{\boldsymbol{T}}\left[\boldsymbol{Y}\;\middle|\;\boldsymbol{H}\right]\right]\le\frac{\|p\|_2^2}{\eta}\cdot O(\log^4 n).$$

证明。 令 $K\ge 2$ 为某个常数，使得 $m\le O(n^{K-1})$。回顾一下，根据命题 3.3，由 $\boldsymbol{H}$ 导出的桶 $\boldsymbol{\Gamma}=(\boldsymbol{\Gamma}_1,\dots,\boldsymbol{\Gamma}_n)$ 满足：对所有 i，都有 $|\boldsymbol{\Gamma}_i|\le{}^{2K}\log_\eta{}^{n}$，除非发生概率至多为 $1/n^K$的情况。我们将证明，当该条件成立时，方差很小；而该条件不成立这一低概率情况对期望的贡献也不会太大。

情形 1。 假设 H 使其导出的桶 $\Gamma=(\Gamma_1,\dots,\Gamma_b)$ 满足 $|\Gamma_i|\le{}^{2K}\log^{\eta}{}^{n}$ 对每个 $i\in[n]$均成立。我们希望证明 $\mathop{\mathrm{Var}}_{\boldsymbol{T}}[\boldsymbol{Y}\;|\;\boldsymbol{H}=H]$ 满足陈述中的上界。我们从引理 2.12的结果开始：

$$\mathop{\mathrm{Var}}_{\boldsymbol{T}}\left[\boldsymbol{Y}\;\middle|\;\boldsymbol{H}=H\right]=2\|p_{|\Gamma}\|_2^2+4m\|p_{|\Gamma}\|_3^3.$$

我们首先对右端的第一项进行界定。对于每个桶 Γ_i，我们有　e

$$\left(p_{|\Gamma}\right)_i^2=\left(\sum_{j\in\Gamma_i}p_j\right)^2=\sum_{j,k\in\Gamma_i}p_j p_k\le\sum_{j,k\in\Gamma_i}\frac{p_j^2+p_k^2}{2}=|\Gamma_i|\sum_{j\in\Gamma_i}p_j^2\le O\left(\frac{\log n}{\eta}\right)\sum_{j\in\Gamma_i}p_j^2.$$

因此，我们得到

$$\|p_{|\Gamma}\|_2^2 = \sum_i^b \left(p_{|\Gamma}\right)_i^2 \leq O\left(\frac{\log n}{\eta}\right)\|p\|_2^2,$$

如所需。接下来处理第二项，首先注意到 $\|p_{|\Gamma}\|_3^3 = \sum_i p[\Gamma_i]^3 \leq \sum_i (\max_j p[\Gamma_j]) p[\Gamma_i]^2 = \|p_{|\Gamma}\|_\infty \|p_{|\Gamma}\|_2^2$。我们声称 $\|p_{|\Gamma}\|_\infty \leq O\left(\frac{\log^3 n}{m}\right)$。固定任意 $i \in [b]$，并考虑条目 $\left(p_{|\Gamma}\right)_i = p[\Gamma_i]$。$p$ 的反集中性给出

$$p[\Gamma_i] \leq \frac{C\log^2 n}{m} \cdot \max\left\{\eta\left(|\Gamma_i| - 1\right), \frac{1}{\log n}\right\}.$$

结合 $|\Gamma_i| \leq O\left(\frac{\log n}{\eta}\right)$ 这一假设，我们得到

$$p[\Gamma_i] \leq \frac{C\log^2 n}{m} \cdot O(\log n),$$

这就证明了该结论。我们已经证明了 $\|p_{|\Gamma}\|_2^2 \leq \frac{\|p\|_2^2}{\eta} \cdot O(\log n)$，因此

$$m\|p_{|\Gamma}\|_3^3 \leq m \cdot O\left(\frac{\log^3 n}{m}\right) \cdot \frac{\|p\|_2^2}{\eta} \cdot O(\log n) = \frac{\|p\|_2^2}{\eta} \cdot O(\log^4 n),$$

由此得出案例 1。

案例 2。 在罕见情况下，$\boldsymbol{H} = H$ 是一个不满足案例 1 的小桶条件的子图，我们将退而采用一个更宽松的条件方差上界，该上界对每个H都成立。我们再次从以下结果开始：引理 2.12：

$$\operatorname*{Var}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H} = H] = 2\|p_{|\Gamma}\|_2^2 + 4m\|p_{|\Gamma}\|_3^3.$$

结合 $\|p_{|\Gamma}\|_1 = 1$ 和 ℓ^p 范数的单调性可得

$$\operatorname*{Var}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H} = H] \leq 2\|p_{|\Gamma}\|_1^2 + 4m\|p_{|\Gamma}\|_1^3 \leq 6m = O(n^{K-1}).$$

总结论证。 现在我们结合这两种情况，对期望方差给出上界。根据 命题 3.3，可得

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\operatorname*{Var}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right] &\leq \left[\begin{array}{l}\mathbb{P}\left[\boldsymbol{H} \text{ satisfies } \|\boldsymbol{\Gamma}\|_\infty \leq \frac{2K\log n}{\eta}\right] \mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\operatorname*{Var}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}] \;\middle|\; \boldsymbol{H} \text{ satisfies } \|\boldsymbol{\Gamma}\|_\infty \leq \frac{2K\log n}{\eta}\right] \\ +\mathbb{P}\left[\boldsymbol{H} \text{ does not satisfy } \|\boldsymbol{\Gamma}\|_\infty \leq \frac{2K\log n}{\eta}\right] \max_H\left\{\operatorname*{Var}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H} = H]\right\}\end{array}\right] \\ &\leq \frac{\|p\|_2^2}{\eta} \cdot O(\log^4 n) + \frac{1}{n^K} \cdot O(n^{K-1}).\end{aligned}$$

由于 $\|p\|_2^2 \geq \|\mu\|_2^2 = 1/n$，第一项大于第二项，证明得证。 □

我们得出结论：$\boldsymbol{Y}$ 满足以下集中界：

引理 3.21（检验统计量的集中性）。设 $C > 0$ 为一个常数，设 $n \in \mathbb{N}$ 足够大，并假设 p 是定义在自然数集 $\mathbb{Z}_n$ 上的概率分布，且不是 C-高度集中的。假设 $\eta \geq \Omega(n^{-1/5})$ 且 $m \leq \mathrm{poly}(n)$。那么对于所有 $t > 0$，

$$\mathbb{P}\left[|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]| \geq t\right] \leq \frac{\|p\|_2^2}{\eta t^2} \cdot O(\log^4 n).$$

证明。结合 引理 3.19和 3.20，运用全方差公式和切比雪夫不等式。

□

3.4 测试器的正确性

结合上述分离与集中结果，我们可以证明 $\boldsymbol{Y}$ 集中在测试器阈值的正确一侧。

引理 *3.22.* 设 $\alpha, L > 0$ 为常数。则存在常数 $\beta > 0$ 和 $c = c_{\alpha,\beta} > 0$，使得对所有充分大的 n，以下结论成立。令 $\epsilon, \eta \in (0, 1]$ 满足 $\eta \geq {}^{L}\log_{n^{1/5}\epsilon^{4/5}} n$。假设 p 是 $\mathbb{Z}_n$ 上的概率分布，且相对于 $m = c \cdot \frac{\sqrt{n}}{\epsilon^2\eta^{3/2}} \log^2 n$ 并非 4α-高度集中。令 $T := \frac{m}{n^2}\sum_{i,j}\phi_{i,j} + \beta\frac{m}{n}\epsilon^2\eta$ 为算法 *1* 第二步所使用的阈值。则检验统计量 $\boldsymbol{Y}$ 满足以下条件：

1. （完备性）若 $p = \mu$，则 $\boldsymbol{Y} < T$ 的概率至少为 $99/100$；*2*. （可靠性）如果 $\mathsf{dist}_{\mathsf{TV}}(p, \mu) > \epsilon$，则 $\boldsymbol{Y} > T$，其概率至少为 $99/100$。

证明。**完备性**。假设 $p = \mu$。根据 命题 3.8，$\boldsymbol{Y}$ 满足 $\mathbb{E}[\boldsymbol{Y}] = \frac{m}{n^2}\sum_{i,\,j}\phi_{i,\,j}$。因此，对于任意固定的 β(将在下文选定)，只需证明 $\boldsymbol{Y} < \mathbb{E}[\boldsymbol{Y}] + \beta\frac{m}{n}\epsilon^2\eta$ 以较高的概率成立。根据 引理 3.21 并利用 $\|\mu\|_2^2 = 1/n$这一事实，，

$$\mathbb{P}\left[\boldsymbol{Y} \geq \mathbb{E}[\boldsymbol{Y}] + \beta\frac{m}{n}\epsilon^2\eta\right] \leq \mathbb{P}\left[|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]| \geq \beta\frac{m}{n}\epsilon^2\eta\right] \leq \frac{\|p\|_2^2}{\eta\left(\beta\frac{m}{n}\epsilon^2\eta\right)^2} \cdot O(\log^4 n) = \frac{O(n\log^4 n)}{\beta^2 m^2\epsilon^4\eta^3},$$

并且我们有

$$\frac{O(n\log^4 n)}{\beta^2 m^2\epsilon^4\eta^3} \leq 1/100 \iff m \geq \frac{1}{\beta} \cdot O\left(\frac{\sqrt{n}}{\epsilon^2\eta^{3/2}}\log^2 n\right), \tag{7}$$

如所需。因此，存在常数 $c^{(1)} = c^{(1)}_{\alpha,\beta}$，使得如果 $m \geq c^{(1)}\frac{\sqrt{n}}{\epsilon^2\eta^{3/2}}\log^2 n$，那么对于所有充分大的 n，$\boldsymbol{Y} < T$ 以至少 $99/100$ 的概率成立。

可靠性。我们采用类似的方法。根据 引理 3.17 和 命题 3.8，我们有

$$\mathbb{E}[\boldsymbol{Y}] > \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] + \frac{m\eta}{8}\|z\|_2^2 = \frac{m}{n^2}\sum_{i,j}\phi_{i,j} + \frac{m\eta}{8}\|z\|_2^2.$$

因此，只需证明对于适当选择的 β，有

$$\boldsymbol{Y} \overset{?}{>} \mathbb{E}[\boldsymbol{Y}] - \frac{m\eta}{8}\|z\|_2^2 + \beta\frac{m}{n}\epsilon^2\eta.$$

回顾一下，当 $\mathsf{dist}_{\mathsf{TV}}(p, \mu) > \epsilon$ 时，我们有 $\|z\|_1 > 2\epsilon$，这意味着 $\|z\|_2^2 > 4\epsilon^2/n$，因此

$$\frac{m\eta}{8}\|z\|_2^2 > \frac{m}{2n}\epsilon^2\eta,$$

从而

$$\mathbb{E}[\boldsymbol{Y}] - \frac{m\eta}{8}\|z\|_2^2 + \beta\frac{m}{n}\epsilon^2\eta < \mathbb{E}[\boldsymbol{Y}] - \frac{m\eta}{8}\|z\|_2^2 + 2\beta\frac{m\eta}{8}\|z\|_2^2 = \mathbb{E}[\boldsymbol{Y}] - (1 - 2\beta)\frac{m\eta}{8}\|z\|_2^2.$$

因此，对于 $\beta \leq 1/3$，我们有 $1 - 2\beta \geq \beta$，只需证明以下结论以至少 $99/100$ 的概率成立：

$$\boldsymbol{Y} \overset{?}{>} \mathbb{E}[\boldsymbol{Y}] - \beta\frac{m\eta}{8}\|z\|_2^2.$$

我们再次应用 引理 3.21，并结合 $\|p\|_2^2 = \|\mu\|_2^2 + \|z\|_2^2 = \frac{1}{n} + \|z\|_2^2$ 和 $\|z\|_2^2 \geq \frac{4\epsilon^2}{n}$。

$$\begin{aligned}
\mathbb{P}\left[\boldsymbol{Y} \leq \mathbb{E}[\boldsymbol{Y}] - \beta\frac{m\eta}{8}\|z\|_2^2\right] \leq \mathbb{P}\left[|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]| \geq \beta\frac{m\eta}{8}\|z\|_2^2\right] &\leq \frac{\|p\|_2^2}{\eta\left(\beta\frac{m\eta}{8}\|z\|_2^2\right)^2} \cdot O(\log^4 n) \\
&= \frac{(\frac{1}{n} + \|z\|_2^2)O(\log^4 n)}{\beta^2 m^2 \eta^3 \|z\|_2^4} \leq \frac{\frac{1}{n}O(\log^4 n)}{\beta^2 m^2 \eta^3 (\epsilon^2/n)^2} + \frac{O(\log^4 n)}{\beta^2 m^2 \eta^3 (\epsilon^2/n)} \\
&= \frac{O(n\log^4 n)}{\beta^2 m^2 \eta^3 \epsilon^4}.
\end{aligned}$$

该失败概率在渐近意义上与完备性情形中得到的结果相同。因此，存在一个常数 $c^{(2)} = c_\alpha^{(2)}$，$\beta > 0$ 使得，对于 $m \geq c^{(2)} \frac{\sqrt{n}}{\epsilon^2\eta^{3/2}} \log^2 n$ 以及所有充分大的n，$\boldsymbol{Y} \to T$以至少 $99/100$ 的概率成立。令 $c = \max\left\{c_{\alpha,\beta}^{(1)}, c_\alpha^{(2)}, \beta\right\}$，即可完成证明。 □

最后，我们结合测试器两个步骤的结果来证明其正确性：

定理 *3.23*（定理 *1.2* 的细化）。存在常数 $\alpha > 0$、$\beta > 0$、$c = c_\alpha$、$\beta > 0$ 和 $L = L_c > 0$，使得对于所有充分大的 n，下述结论成立。假设 ϵ、$\eta \in (0, 1])$ 满足 $\eta \geq L\log_{n^{1/5}\epsilon^{4/5}}^{4/5} n$。令 G 为顶点集为 $V = \mathbb{Z}_n$ 的环或路径，并令 p 为 $\mathbb{Z}_n$ 上的概率分布。将常数设为 α、β、L 和 c 后，算法 *1* 的样本复杂度为 $\Theta\left(\frac{\sqrt{n}}{\epsilon^2\eta^{3/2}}\log^2 n\right)$，且其在 p 上的输出满足

1.（完备性）如果 p 是 $\mathbb{Z}_n$ 上的均匀分布，那么算法以至少 $9/10$ 的概率接受；*2.*（可靠性）如果 p 与均匀分布的全变差距离至少为 ϵ，那么算法以至少 $9/10$ 的概率拒绝。

证明。我们首先按照 $\alpha > 0$ 将其取为足够大，使其满足引理 3.5。该引理还要求 $m \leq n\eta$，现在我们来验证这一点。任取常数 $c > 0$，并假设 $m = c \cdot \frac{\sqrt{n}}{\epsilon^2\eta^{3/2}} \log^2 n$。于是

$$m \leq n\eta \iff c \cdot \frac{\sqrt{n}}{\epsilon^2} \cdot \frac{\log^2 n}{\eta^{3/2}} \leq n\eta \iff \eta^{5/2} \geq c \cdot \frac{\log^2 n}{n^{1/2}\epsilon^2} \iff \eta \geq \frac{c^{2/5}\log^{4/5} n}{n^{1/5}\epsilon^{4/5}}.$$

因此，对于任意的 c，令 $L = L_c \geq c^{2/5}$ 即可保证 $m \leq n\eta$。

因此，按照 β，$c > 0$ 取值，并采用引理 3.22给出的 L_c，如上所述。现在，我们可以利用引理 3.5和引理 3.22来证明测试器整体上的正确性。（注意，样本复杂度的结论可由算法的规格说明直接推出。）

完备性。根据 引理 3.5，测试器的第一步拒绝的概率至多为 $1/100$。同样，根据 引理 3.22，测试器的第二步拒绝的概率至多为 $1/100$. 因此，总拒绝概率至多为 $2/100 < 1/10$。

可靠性。根据 p的集中程度，分两种情况讨论。首先，假设 p 是4α-高度集中的。根据 $99/100$，由 引理 3.5，测试器的第一步以至少 p 的概率拒绝。另一方面，如果 4α不是 $99/100$ -高度集中的，那么根据 引理 3.22，测试器的第二步以至少 $99/100 > 9/10$的概率拒绝。无论哪种情况，测试器都以至少 $99/100 > 9/10$的概率拒绝。 □

注 3.24. 在引言中（见 定理 1.2），我们指出，我们给出的样本复杂度上界适用于以下情形：$\epsilon \geq \widetilde{\Theta}\left(n^{-1/4}\right)$ 且 $\eta \geq \widetilde{\Theta}\left(n^{-1/5}\epsilon^{-4/5}\right)$。尽管

关于 ϵ 的条件虽未在上文明确说明，但它可由关于 η 的条件以及在我们的问题定义中 $\eta \le 1$ 这一事实推出：

$$\frac{L \log^{4/5} n}{n^{1/5}\epsilon^{4/5}} \le \eta \le 1 \implies \epsilon \ge \frac{L^{5/4} \log n}{n^{1/4}}.$$

一个有趣的问题是，能否处理范围更广的参数；具体而言，我们的分析使用了不等式 $m \le n\eta$，但如果允许 ϵ 任意小，那么必然有 $m \gg n$。还要注意，无法处理参数的全部范围：当 ϵ 足够小且（例如）$\eta = 1/10$ 时，可以将偏离均匀性放在两个相邻顶点上，而测试器以至少 $9/10$ 的概率无法察觉这种偏离。

4 ParityTraceModel 中的均匀性检验

本节陈述主要定理 1.1中关于上界的部分，并在此给出其正式表述：

定理 4.1。 固定定义域 $[2n]$。设 Π 仅包含均匀分布。则在奇偶迹下进行$(\Pi, \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Pi))$分布测试的样本复杂度为 $\widetilde{\Theta}\left(\binom{n}{\epsilon}^{4/5} + \frac{\sqrt{n}}{\epsilon^2}\right)$。

遵循第 2 节中的设置，我们考虑在奇偶迹模型中对未知分布 $\pi = \pi(p, q)$ 进行均匀性检验。回想一下，我们将迹的两端缝合成一条项链，并研究由此得到的圆形轨迹。因此，我们的基图 $G = (\mathbb{Z}_n, E)$ 是环，并将 p 视为顶点上的部分分布，而 q 将决定边的权重：如果 $e \in E$ 连接顶点 i 和 $i+1$（模 n），那么 $w(e) = 1 - e^{-mq_i}$，并且该边以 $1 - w(e) = \mathbb{P}[\text{泊松分布}(mq_i) = 0]$ 的概率被采样到 $\boldsymbol{H}$ 中。

该检验算法分为两种情况：当 ϵ 非常小时，此时我们可以将问题归约为标准的均匀性检验算法（在 第 4.1 节中处理）；当 ϵ 不太小时，则适用我们的主要分析。在主要分析中，测试器执行 3 个步骤：

1. 偏差检验：检查迹中取值为 1 和 0 的符号数量是否过于不平衡；若是，则该分布必定远离均匀分布；2. 集中度检验：检查取值为 1 或 0 的符号的游程长度是否过大。我们将证明，这种情况也可以安全地拒绝。3. 基于碰撞的检验：根据检验统计量 $\boldsymbol{Y}$ 是否低于某个阈值来决定接受或拒绝。这一步需要最多的技术工作。

形式上，该算法为算法 2。它由绝对常数 α, β, γ, K 参数化，这些常数将在后文定义，并要求 $\epsilon \ge {}^{K}\log_{n^{1/4}}{}^{3} n$；注意，该条件与算法设定 $m = O\left(\binom{n}{\epsilon}^{4/5} \log^{7/5} n\right)$ 这一事实结合后，意味着 $m = O\left({}^{n}\log n\right) = o(n)$。在整个分析过程中，我们都会用到这一事实。

我们将 μ 记为满足 $\pi(\mu, \mu)$ 为均匀分布的部分分布，即对每个 $i \in \mathbb{Z}_n$，都有 $\mu_i = 1/2n$ 。如第 2.3 节所述，我们只分析迹中 1 的统计量（即算法 2 中的情形 $b = 1$），因为 0 的情形与之对称。第 4.2 节将证明算法 2 的前两个步骤是正确的。第 4.3 节将证明，当输入分布为均匀分布时，$\mathbb{E}[\boldsymbol{Y}]$ 很小；当输入分布远离均匀分布时，$\mathbb{E}[\boldsymbol{Y}]$ 很大。第 4.4 节将给出 $\boldsymbol{Y}$ 的方差上界。第 4.5 节将结合这些结果证明算法 2 的正确性。再结合第 4.1 节中针对较小 ϵ 情形的算法，即可证明定理 4.1 的上界。

算法 2 当 $\epsilon \geq K_{\alpha,\beta,\gamma} \cdot \sqrt[3]{n}$ 对数时的均匀性检验器。

集合 $m \leftarrow \Theta_{\alpha,\beta,\gamma}\left(\left(\frac{n}{\epsilon}\right)^{4/5} \log^{7/5} n\right)$。常数：$\alpha, \beta, \gamma > 0$ 和 $K = K_{\alpha,\beta,\gamma} > 1$，稍后定义。输入：对于定义域 $\mathbb{Z}_n$ 上的 $\pi = \pi(p, q)$，接收样本 $S \leftarrow$ 的迹 trace(S)，采样 samp(π, m)。要求：$\epsilon \geq K \log_{n^{1/4}} \sqrt[3]{n}$。1：过程 UniformityTester-ParityTrace (trace(S)) 2：根据 ~~trace~~(S) 构造圆形轨迹。3：对 $b \in \{0, 1\}$ 执行 4：令 $X_1, \ldots, X_n$ 为第 2.3 节中定义的 "b" 游程长度。5：$N \leftarrow \sum_i X_i$。6：$Y \leftarrow \frac{1}{m} \sum_i X_i(X_i - 1)$。7：如果 $\frac{N}{m} \geq \frac{1}{2} + \frac{\gamma}{\sqrt{m}}$，则拒绝。8：如果 $\max_i X_i \geq \alpha \log n$，则拒绝。9：如果 $Y \geq \frac{m}{4n^2} \sum_{i,j} \phi_{i,j}^{(\mu)} + \beta \frac{\epsilon^2 m^2}{n^2}$，则拒绝。10：接受。

4.1 小 ϵ 情形

任取一个常数 $K > 0$。在 $\epsilon < K \log_{n^{1/4}} \sqrt[3]{n}$ 的情形下，测试器将模拟标准均匀性测试器（例如参见 [VV17a、DGPP19]）。奇偶迹下的测试器将使用大小为 $O\left(\frac{\sqrt{n}}{\epsilon^2} + n \log n\right) = \widetilde{O}\left(\frac{\sqrt{n}}{\epsilon^2}\right)$ 的样本。该模拟之所以可行，是因为使用 $O(n \log n$ 个样本时，测试器要么会收到来自每个定义域元素的样本（从而能够正确区分支撑集中的所有元素），要么可以安全地拒绝。这个证明并无特别启发性，因此我们将其推迟到附录 C.1。

引理 4.2。假设 $\epsilon < K \log_{n^{1/4}} \sqrt[3]{n}$。~~那么存~~在一个奇偶迹下的分布测试器，其样本复杂度为 $\widetilde{O}\left(\frac{\sqrt{n}}{\epsilon^2}\right)$；对于输入分布 $\pi = \pi(p, q)$：

1. 如果 $p = q = \mu$，则算法将以至少 $2/3$ 的概率接受；*2.* 如果 π 与均匀分布的差距为 ϵ，则算法将以至少 $2/3$ 的概率拒绝。

4.2 简单情形：不平衡且高度集中的分布

现在我们继续进行主要分析，其中 $\epsilon \geq K \log_{n^{1/4}} \sqrt[3]{n}$，~~其中~~ K 是某个常数。本节处理测试器的两种"简单"拒绝情形：检测取值为 1 的元素和取值为 0 的元素上的总概率质量是否显著不平衡，或检测部分分布 p 或 q 相对于另一个是否高度集中。首先，我们考虑部分分布不平衡的情形：

命题 4.3。对于充分大的绝对常数 $\gamma > 0$ 和充分大的 n，测试器满足以下条件：

- 当 $\pi(p, q)$ 服从均匀分布时，第 7 行以不超过 $1/100$ 的概率拒绝。

- 当 $\|p\|_1 \notin \frac{1}{2} \pm \frac{2\gamma}{\sqrt{m}}$ ~~(等价地，$\|q\|_1 \notin \frac{1}{2} \pm \frac{2\gamma}{\sqrt{m}}$)~~ 时，第 *7* 行以至少 $99/100$ 的概率拒绝。

证明。固定测试器的第 $b = 1$次迭代，使得 $\boldsymbol{N} \sim$ 泊松分布$(m\|p\|_1)$ 表示在迹中观察到的**"1"**符号数量。首先，假设 $\pi(p,q)$ 均匀分布，因此特别有 $\|p\|_1 = 1/2$。那么根据事实 3.4，此轮迭代中测试拒绝的概率为

$$\mathbb{P}\left[\frac{\boldsymbol{N}}{m} \geq \frac{1}{2} + \frac{\gamma}{\sqrt{m}}\right] = \mathbb{P}\left[\boldsymbol{N} \geq m/2 + \gamma\sqrt{m}\right] \leq e^{-\frac{m\gamma^2}{2(m/2+\gamma\sqrt{m})}} \leq e^{-\frac{\gamma^2}{2}} \leq 1/200\,.$$

现在假设 $\|p\|_1 > \frac{1}{2} + \frac{2\gamma}{\sqrt{m}}$。那么，本轮测试未能拒绝的概率为

$$\mathbb{P}\left[\frac{\boldsymbol{N}}{m} < \frac{1}{2} + \frac{\gamma}{\sqrt{m}}\right] \leq \mathbb{P}\left[\boldsymbol{N} \leq m\|p\|_1 - \gamma\sqrt{m}\right] \leq e^{-\frac{m\gamma^2}{2(m\|p\|_1+\gamma\sqrt{m})}} \leq e^{-\frac{\gamma^2}{4}} \leq 1/200\,.$$

由对称性可知，对于迭代 $b = 0$ 相对于 $\|q\|_1$，同样结论成立。因此，除概率至多为 $1/100$ 的情况外，此步骤都能正确地接受或拒绝。 □

接下来，我们处理 p 或 q 相对于另一个高度集中的情况，定义如下。

定义 4.4（高度集中的部分分布）。给定常数 $C > 0$、正整数 m 和部分分布 $\pi = \pi(p,q)$，如果 p 相对于 q（关于 m）是 C-高度集中[11]，则称其为此性质，当且仅当 $\rho_t(p \parallel q) \geq C\log^2 n$，其中 $t = \frac{1}{m}\log n$。

备注 4.5。如果 p 相对于 q并非 C高度集中，那么特别地，$\|p\|_\infty < \frac{C}{m}$对数$n$，这可通过对每个 $i \in \mathbb{Z}_n$取区间 $I = \langle\langle i, 1\rangle\rangle$ 看出。

现在，我们可以将此定义与引理 2.14 结合起来，证明测试器的第二步会按预期运行：

命题 4.6。对于足够大的绝对常数 $\alpha > 0$ 以及足够大的 n，测试器满足以下条件：

1. 当 $\pi(p,q)$ 服从均匀分布时，第 *8* 行以至多 $1/100$的概率拒绝。
2. 当 p、q 中至少有一个相对于另一个关于 m 4α高度集中时，第 *8* 行以至少 $99/100$的概率拒绝。

证明。**完备性**。假设 $\pi(p,q)$服从均匀分布，并固定迭代 $b = 1$，此时测试器在寻找由**"1"**符号组成的长游程。因此，我们有桶 $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1, \ldots\ldots, \boldsymbol{\Gamma}_b)$，它们对应于 $\boldsymbol{H}$的连通分量，并且每个桶 $\boldsymbol{\Gamma}_i$贡献一个游程长度 $\boldsymbol{X}_i \sim$泊松分布 $(mp[\boldsymbol{\Gamma}_i])$。

首先，我们断言，以高概率有 $\max_i q[\boldsymbol{\Gamma}_i^*] \leq \frac{2}{m}\log n$。对于每个 $i \in \mathbb{Z}_n$，令 I_i为满足 $q[I_i^*] > \frac{2}{m}\log n$ 的最小圆周区间 $I_i = \langle\langle i, d\rangle\rangle$。如果 $\max_i q[\boldsymbol{\Gamma}_i^*] > \frac{2}{m}\log n$，那么这些区间中至少有一个被连接起来，即对于某个 $i \in \mathbb{Z}_n$，有 $\boldsymbol{J}(I_i) = 1$。回顾一下，每条边 $e = (i, i+1)$ 以 $1 - w(e) = e^{-mq_i} = \mathbb{P}[$泊松分布$(mq_i) = 0]$ 的概率出现在 $\boldsymbol{H}$ 中。因此

$$\begin{aligned}\mathbb{P}[\boldsymbol{J}(I_i) = 1] &= \mathbb{P}[\forall e \in I_i^* : e \in \boldsymbol{H}] = \mathbb{P}[\mathsf{Poi}(mq[I_i^*]) = 0] \\ &< \mathbb{P}\left[\mathsf{Poi}\left(m \cdot \frac{2\log n}{m}\right) = 0\right] = e^{-2\log n} = 1/n^2\,.\end{aligned}$$

[11]我们会适当调整这个定义，以检验 0 的统计量。

根据并集界，$\mathbb{P}\left[\max_i q\left[\boldsymbol{\Gamma}_i^*\right] > 2^{\log_m n}\right] < 1/n = o(1)$。

现在，假设 $\max_i q\left[\boldsymbol{\Gamma}_i^*\right] \le 2^{\log_m n}$。由于 π 是均匀分布，因此 $p = q$，可得 $\max_i p\left[\boldsymbol{\Gamma}_i\right] \le 4^{\log_m n}$（系数 2 是因为 Γ_i 通常比 Γ_i^* 多一个元素；如果后者为空且 Γ_i 的大小为 1，则根据 π 是均匀分布这一事实，结论成立）。因此，我们使用 事实 3.4 来给任意固定的桶 Γ_i 产生游程长度且会被第 8 行拒绝的概率提供上界：当 α 足够大时，

$$\begin{aligned}\mathbb{P}\left[\boldsymbol{X}_i \ge \alpha \log n \mid \boldsymbol{\Gamma}_i = \Gamma_i\right] &= \mathbb{P}\left[\text{泊松分布}\left(mp[\Gamma_i]\right) \ge \alpha \log n\right] \le \mathbb{P}\left[\text{泊松分布}(4\log n) \ge \alpha \log n\right] \\ &\le \mathbb{P}\left[\mathsf{Poi}(4\log n) - 4\log n \ge (\alpha-4)\log n\right] \le e^{-\frac{(\alpha-4)^2 \log^2 n}{2((\alpha-4)\log n + 4\log n)}} \\ &\le e^{-\frac{(\alpha/2)^2 \log n}{2\alpha}} \le e^{-2\log n} \le 1/n^2 .\end{aligned}$$

因此，该事件对任意桶发生的概率至多为 $1/n = o(1)$；根据对称性，对于测试器在 $b = 0$ 下的迭代也同样如此。因此，当 $\pi(p,\ q)$ 服从均匀分布时，第 8 行仅以 $o(1)$ 的概率拒绝。

可靠性。不失一般性，假设 p 是 4α 相对于 q 高度集中。根据引理 2.14，令 I 为满足以下条件的圆周区间

1. $q[I^*] \le \frac{1}{m \log n}$；
以及 2. $p[I] \ge 2\alpha \frac{\log n}{m}$。

首先，我们声称，以高概率，$\boldsymbol{J}(I) = 1$。事实上，我们有

$$\begin{aligned}\mathbb{P}\left[\boldsymbol{J}(I) = 0\right] &= 1 - \mathbb{P}\left[\forall e \in I^* : e \in \boldsymbol{H}\right] = 1 - \mathbb{P}\left[\mathsf{Poi}(mq[I^*]) = 0\right] \\ &= \mathbb{P}\left[\mathsf{Poi}(mq[I^*]) > 0\right] \le \mathbb{P}\left[\mathsf{Poi}\left(\frac{1}{\log n}\right) > 0\right] = 1 - e^{-\frac{1}{\log n}} = o(1) .\end{aligned}$$

因此，从 I 中的顶点采样得到的任何“1”符号都属于同一个游程，并且该游程至少包含 $\mathsf{Poi}(mp[I])$ 个符号，其中 $mp[I] \ge 2\alpha \log n$。因此，该游程长度未能超过拒绝阈值的概率至多为

$$\begin{aligned}\mathbb{P}\left[\text{泊松分布}(mp[I]) < \alpha \log n\right] &\le \mathbb{P}\left[\text{泊松分布}(2\alpha \log n) \le \alpha \log n\right] = \mathbb{P}\left[\text{泊松分布}(2\alpha \log n) \le 2\alpha \log n - \alpha \log n\right] \\ &\le e^{-\frac{\alpha^2 \log^2 n}{2(\alpha \log n + 2\alpha \log n)}} = e^{-\frac{\alpha \log n}{6}} = o(1) .\end{aligned}$$

因此，第 8 行会拒绝，除非发生概率为 $o(1)$ 的情况，证明完毕。 □

这两个步骤使我们可以在需要时作如下假设：1) $\|p\|_1, \|q\|_1 \in \frac{1}{2} \pm \frac{2\gamma}{\sqrt{m}}$；以及 2) 就 m 而言，p 和 q 均未相对于对方呈现 4α-高度集中。现在我们可以开始分析主要检验统计量 $\boldsymbol{Y}$。

4.3 检验统计量的期望值

我们的首要目标是证明：当 $\pi(p,q) = \pi(\mu,\mu)$ 时，与 $\mathsf{dist}_{\mathsf{TV}}(\pi(p,q)\ \pi(\mu,\mu)) > \epsilon$ 时的情形相比，$\mathbb{E}\left[\boldsymbol{Y}\right]$ 有明显区别。注意，后一种情形意味着 $\|p-\mu\|_1 > \epsilon$ 或 $\|q-\mu\|_1 > \epsilon$，并且我们的测试器对于取值为 0 和 1 的符号是对称的，因此不失一般性，可以假设 $\|q-\mu\|_1 > \epsilon$。

回顾一下我们对 $\boldsymbol{Y}$ 的期望值的表述，该表述来自命题 2.6：

$$\mathbb{E}\left[\boldsymbol{Y}\right] = mp^{\top}\phi p\,. \tag{8}$$

我们首先证明，当 $\pi(p,\ q) = (\mu,\ \mu)$ 时，这个值恰好是 算法 2 为检验统计量设定阈值时所依据的基准。令 $\phi^{(\mu)} := \mathbb{E}\left[\boldsymbol{\Phi}^{(\mu)}\right]$，其中 $\boldsymbol{\Phi}^{(\mu)} = \boldsymbol{\Phi}(\mu)$ 是由均匀部分分布 $q = \mu$生成的随机连接矩阵。令随机变量 $\boldsymbol{Y}^{(\mu)}$ 表示当 $\pi(p,\ q)$ 服从均匀分布时检验统计量的取值。于是有：

命题 4.7 (均匀情形下 $\boldsymbol{Y}$的期望)。 当 $\pi(p, q)$ 服从均匀分布时，统计量 $\boldsymbol{Y} = \boldsymbol{Y}^{(\mu)}$ 满足

$$\mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] = \frac{m}{4n^2}\sum_{i,j}\phi_{i,j}^{(\mu)}\,.$$

证明。 由于 $p = q = \mu = \vec{1}/2n$，式（8）给出

$$\mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] = m\mu^{\top}\phi^{(\mu)}\mu = \frac{m}{4n^2}\sum_{i,j}\phi_{i,j}^{(\mu)}\,. \qquad \square$$

我们希望证明，当 q 与 μ相差甚远时，无论如何选择 p，二次型 $p^{\top}\phi p$ 都很大（假设相对集中度和偏差满足我们的条件）。我们的策略是证明，即使采用专门设计来尽可能减小 $\boldsymbol{Y}$的“最坏情况”部分分布 $p = p(q)$，与均匀情形相比仍会产生很大的差距；然后再论证，如果 p 偏离p，检验任务只会变得更容易。

我们将这种最坏情况部分分布称为 q的均匀共轭分布，记作 $\overline{p}$（因为它起到 p的作用）。形式上，若部分分布 p 是 q 的 τ-均匀共轭分布，则

$$\left(\sum_{i=0}^{n-1}\overline{p}_i\right) + \left(\sum_{i=0}^{n-1}q_i\right) = 1 \qquad \text{and} \qquad \phi\overline{p} = \tau\cdot\vec{1}\,.$$

注意， ϕp 的第 i个条目是

$$(\phi p)_i = \sum_{j=0}^{n-1}\phi_{i,j}p_j = \mathop{\mathbb{E}}_{\boldsymbol{\Phi}}\left[\sum_{j=0}^{n-1}\boldsymbol{\Phi}_{i,j}p_j\right]\,,$$

它是包含元素 $i\in\mathbb{Z}_n$的桶的期望总和 $p[\boldsymbol{\Gamma}_{\boldsymbol{\gamma}(i)}]$。因此，当 p 是 q的 τ-均匀共轭分布时，每个桶的期望总和都相同。这是“最坏情况”，因为从 p 中采样得到的奇数元素的分布，在期望意义下将是均匀分布。

如果允许 p含有负条目，可以证明，对于某个值 τ，每个 q 都有一个均匀共轭分布，因为 ϕ 是半正定的[12] （断言 4.10）。但我们要求一个显式的 τ （且条目非负）。我们将给出一个近似均匀共轭分布的闭式解：

定义 4.8 （近似均匀共轭分布）。 对于 q 上满足 $\|q\|_1 > 0$ 的部分分布 $\mathbb{Z}_n$，以及样本量参数 m，给定其期望连接矩阵 $\phi = \phi(q)$，令 $\xi(m, q) :=$

[12]如果 q 的所有条目均非零，那么如断言 4.10 ϕ 中所述， p 是正定的，因此可逆，所以对于适当的 τ， p 存在。如果 q 存在一些零条目，可以先通过删除这些条目来缩小定义域，求解逆问题，然后在由零 p 密度分隔的每个值域区间内任意分配密度 q。

$\frac{e^{-m\|q\|_1}}{(1-e^{-m\|q\|_1})^2}$。我们称 $\widetilde{p} \in \mathbb{R}_{\geq 0}^{\mathbb{Z}_n}$ 是 q 的近似均匀共轭分布，如果 $\|\widetilde{p}\|_1 = 1 - \|q\|_1$，并且对于 $\tau = \tau(m,q) := \frac{1-\|q\|_1}{\sum_{i=0}^{n-1} \text{双曲正切}\binom{mq_i}{2}}$，有

$$\max_{i \in \mathbb{Z}_n} |(\phi\widetilde{p})_i - \tau| \leq 4n \cdot \xi(m,q)\,.$$

为方便起见，以下记 $\xi := \xi(m,\ q)$。我们用近似均匀共轭分布来表示检验统计量的期望值：

命题 4.9。 假设 $\widetilde{p}$ 是 q 的一个近似均匀共轭分布，并记 $p = \widetilde{p} + z$。则

$$\mathbb{E}[\boldsymbol{Y}] = m\widetilde{p}^\top \phi\widetilde{p} + mz^\top \phi z \pm 8mn^2\xi$$

and

$$\widetilde{p}^\top \phi\widetilde{p} = \|p\|_1(\tau \pm 4n\xi)\,.$$

此外，若 $z = \vec{0}$，则我们只需考虑

$$\mathbb{E}[\boldsymbol{Y}] = m\widetilde{p}^\top \phi\widetilde{p}\,.$$

证明。我们有

$$\begin{aligned}\mathbb{E}[\boldsymbol{Y}] &= mp^\top \phi p = m(\widetilde{p}+z)^\top \phi(\widetilde{p}+z) = m\left(\widetilde{p}^\top \phi\widetilde{p} + 2z^\top \phi\widetilde{p} + z^\top \phi z\right) \\ &= m\widetilde{p}^\top \phi\widetilde{p} + mz^\top \phi z + 2m\sum_i z_i(\phi\widetilde{p})_i = m\widetilde{p}^\top \phi\widetilde{p} + mz^\top \phi z + 2m\sum_i z_i(\tau \pm 4n\xi)\,.\end{aligned}$$

由于 $\sum_j z_j = 0$（因为 $\|p\|_1 = \|\widetilde{p}\|_1 = 1 - \|q\|_1$），且对每个 $i \in \mathbb{Z}_n$ 都有 $|z_i| \leq 1$，因此项 $2m\sum_i z_i(\tau \pm 4n\xi)$ 的绝对值不超过 $8mn^2\xi$，由此得到第一个结论。考察 $\vec{}$ 且 $z = 0$ 的情形，也可得到最后一个结论。对于第二个陈述，注意到

$$\widetilde{p}^\top \phi\widetilde{p} = \widetilde{p}^\top(\phi\widetilde{p}) = \sum_{i=0}^{n-1} \widetilde{p}_i(\tau \pm 4n\xi)\,.$$

由于 $\sum_{i=0}^{n-1} \widetilde{p}_i = \|\widetilde{p}\|_1 = \|p\|_1 = 1 - \|q\|_1$，该值的上界为 $(\tau + 4n\xi)\|p\|_1$，下界为 $(\tau - 4n\xi)\|p\|_1$。

□

ϕ 的半正定性（如下所示）已经说明 $z^\top \phi z$ 是非负的，我们还将证明近似误差项 $8mn^2\xi$ 可以忽略不计。因此，我们的任务是证明 $\widetilde{p}^\top \phi\widetilde{p} = \|p\|_1(\tau \pm 4n\xi)$足够大，即证明当 q 与均匀分布相距较远时，τ 足够大。

断言 4.10。 对于任意部分分布 $q \in \mathbb{Z}_n$，矩阵 $\phi = \phi(q)$ 是半正定的。若对每个 $i \in \mathbb{Z}_n$ 都有 $q_i > 0$，则 ϕ 是正定的。

证明。 对于任意向量 $u \in \mathbb{R}^{\mathbb{Z}_n}$，我们有

$$u^\top \phi u = \mathbb{E}\left[u^\top \boldsymbol{\Phi} u\right] = \mathbb{E}\left[\sum_{i,j \in \mathbb{Z}_n} \boldsymbol{\Phi}_{i,j} u_i u_j\right] = \mathbb{E}\left[\sum_{i=1}^{n} \sum_{j,j' \in \boldsymbol{\Gamma}_i} u_j u_{j'}\right] = \mathbb{E}\left[\sum_{i=1}^{n} \left(\sum_{j \in \boldsymbol{\Gamma}_i} u_j\right)^2\right].$$

这是非负的，因此我们得出 ϕ 是半正定的。如果对于每个 $i \in \mathbb{Z}_n$都有 $q_i > 0$，那么对于每条边 e， $w(e) > 0$，因此以正概率， $\kappa > 0$ 子图 $\boldsymbol{H}$ 将是一个独立集。在这种情况下， $\boldsymbol{\Phi} = \Phi$ 将对每个 $i \in \mathbb{Z}_n$都有单元素桶 $\Gamma_i = (i,\ 1)$，并且

$$u^\top \phi u \geq \kappa \cdot u^\top \Phi u = \kappa \cdot \sum_{i=1}^{n} \left(\sum_{j,j' \in \Gamma_i} u_j u_{j'} \right)^2 = \kappa \cdot \sum_{i=0}^{n-1} u_i^2 > 0\,,$$

每当 $u \neq \vec{0}$。因此 ϕ 是正定的。 □

在继续利用近似均匀共轭分布的良好性质之前，我们先证明这种对象确实存在：

引理 4.11. 对于任意部分分布 q，使得 $\|q\|_1 > 0$，以及样本量参数 m，都存在 q的一个近似均匀共轭分布 $\widetilde{p}$。若 $q = \mu$，则可取 $\widetilde{p} = \mu$。

证明. 回想一下，我们将环的顶点与模 n 的整数等同起来，即 $\mathbb{Z}_n$。

令 E表示事件 $\boldsymbol{H} = G$，也就是说，每条边都被采样到 $\boldsymbol{H}$ 中。定义 $\varepsilon := \mathbb{P}[E]$，并注意到 $\varepsilon = e^{-m\|q\|_1}$，因为每条边 $e = (i, i+1)$ 被采样的概率为 $1 - w(e) = e^{-mq_i}$。

令 $u \in \mathbb{R}^{\mathbb{Z}_n}$，我们将其视为 $\widetilde{p}$的候选对象。对于每个 $i \in \mathbb{Z}_n$，定义三个随机变量：

- $\boldsymbol{N}_i^{\boldsymbol{R}}$ 是在 $\boldsymbol{H}$ 中沿顺时针方向与 i 相连的顶点数（包括 i 本身）。形式上， $\boldsymbol{N}^{\boldsymbol{R}}$: $= \sum_{d=1}^{n} \boldsymbol{J}(\langle\!\langle i, d \rangle\!\rangle)$。然后定义 $\boldsymbol{R}_i := u[\langle\!\langle i, \boldsymbol{N}_i^{\boldsymbol{R}} \rangle\!\rangle] = \sum_{t=0}^{\boldsymbol{N}_i^{\boldsymbol{R}}-1} u_{i+t}$。

- $\boldsymbol{N}_i^{\boldsymbol{L}}$是沿逆时针方向与 i 相连的 $\boldsymbol{H}$ 中的顶点数（包括i 本身），除非所有边都被采样到 $\boldsymbol{H}$ 中（即事件 E 发生），此时我们定义$\boldsymbol{N}_i^{\boldsymbol{L}} = 1$。形式化地， $\boldsymbol{N}_i^{\boldsymbol{L}} := \mathbb{1}[E] + (1 - \mathbb{1}[E]) \sum_{d=1}^{n} \boldsymbol{J}(\langle\!\langle i, -d \rangle\!\rangle)$。然后定义 $\boldsymbol{L}_i := u[\langle\!\langle i, -\boldsymbol{N}_i^{\boldsymbol{L}} \rangle\!\rangle] = \sum_{t=0}^{\boldsymbol{N}_i^{\boldsymbol{L}}-1} u_{i-t}$。

- 定义 $\boldsymbol{D}_i := u[\boldsymbol{\Gamma}_{\boldsymbol{\gamma}(i)}] = u[\langle\!\langle i - \boldsymbol{N}^{\boldsymbol{L}} + 1, \boldsymbol{N}^{\boldsymbol{L}} + \boldsymbol{N}^{\boldsymbol{R}} - 1 \rangle\!\rangle]$。

注意，对于每个 $i \in \mathbb{Z}_n$， $\boldsymbol{D}_i = \boldsymbol{L}_i + \boldsymbol{R}_i - u_i$；当 E 不出现时，这是成立的，因为 $\boldsymbol{N}_i^{\boldsymbol{L}}$ 和 $\boldsymbol{N}_i^{\boldsymbol{R}}$ 分别计算 i 左右相连元素的数量，并且由于两者都会在某处遇到一个不相连的元素，因此只有 i 本身被重复计算了一次。另一方面，当 E 出现时，根据构造这也是成立的，因为我们得到 $\boldsymbol{D}_i = \boldsymbol{R}_i = \vec{1}^\top u$，而 $\boldsymbol{L}_i = u_i$。

完美的均匀共轭分布 u 会使 $\mathbb{E}[\boldsymbol{D}_i]$ 处处相等。注意，随机变量 $\boldsymbol{R}_i$ 和 $\boldsymbol{L}_i$ 并非相互独立，且 $\boldsymbol{N}_i^{\boldsymbol{R}}$， $\boldsymbol{N}_i^{\boldsymbol{L}}$ 有界，这使分析变得复杂。我们通过定义相互独立的随机变量，并要求得到一个近似的 均匀共轭分布，来放宽问题。

我们引入由马尔可夫过程生成的相互独立的随机变量 $\boldsymbol{N}_i^{\boldsymbol{R}'}$ 和 $\boldsymbol{N}_i^{\boldsymbol{L}'}$，并定义依赖于 $\boldsymbol{N}_i^{\boldsymbol{R}'}$ 和 $\boldsymbol{N}_i^{\boldsymbol{L}'}$ 的新变量 $\boldsymbol{L}_i'$、 $\boldsymbol{R}_i'$和 $\boldsymbol{D}_i'$，其依赖方式与之前相同：

- 生成$\boldsymbol{N}_i^{\boldsymbol{R}'}$的过程如下：将 $\boldsymbol{N}_i^{\boldsymbol{R}'}$初始化为 1。对于按递增顺序排列的每个 $t \geq 0$，从$\boldsymbol{P}_t \sim$泊松分布(mq_{i+t})中采样。如果 $\boldsymbol{P}_t > 0$，则停止；否则将 $\boldsymbol{N}_i^{\boldsymbol{R}'}$递增。定义 $\boldsymbol{R}_i'$: $= u[\langle\!\langle i, \boldsymbol{N}_i^{\boldsymbol{R}'} \rangle\!\rangle] = \sum_{t=0}^{\boldsymbol{N}_i^{\boldsymbol{R}'}-1} u_{i+t}$。

生成• $\boldsymbol{N}_i^{\boldsymbol{L}'}$的过程如下：将$\boldsymbol{N}_i^{\boldsymbol{L}'}$初始化为 1。对于按递增顺序排列的每个 $t \geq 1$，从$\boldsymbol{P}_t \sim$泊松分布(mq_{i-t})中采样。如果 $\boldsymbol{P}_t > 0$，则停止；否则将 $\boldsymbol{N}_i^{\boldsymbol{L}'}$递增。

Define $\boldsymbol{L}_i' := u[\langle\!\langle i, -\boldsymbol{N}_i^{\boldsymbol{L}'} \rangle\!\rangle] = \sum_{t=0}^{\boldsymbol{N}_i^{\boldsymbol{L}'}-1} u_{i-t}$.

• 定义 $\boldsymbol{D}_i' := \boldsymbol{L}_i' + \boldsymbol{R}_i' - u_i$，与上述原始过程中的定义相同。

回顾一下，$\langle\!\langle i, k \rangle\!\rangle$ 是一个多重集，因此 $u[\langle\!\langle i, k \rangle\!\rangle]$ 可以将元素 u_t 计数多次。令$f := 1 - \|q\|_1$。回顾定义 4.8，我们希望 u 满足三个要求：1) $u_i \geq 0$对所有 $i \in \mathbb{Z}_n$都成立；2) $\sum_i u_i = f$；以及 3) 对所有 $i \in \mathbb{Z}_n$，$\mathbb{E}[\boldsymbol{D}_i'] = \tau = \frac{1-\|q\|_1}{\sum_{i=0}^{n-1} \text{双曲正切}(\frac{mq_i}{2})}$。如果我们得到这样的 u并证明 $\mathbb{E}[\boldsymbol{D}_i] = \mathbb{E}[\boldsymbol{D}_i'] \pm 4n\xi$，就找到了我们的近似均匀共轭分布 $\widetilde{p}$。

我们给出一个显式解，然后对其进行验证。令

$$u_i := \tau \left(\frac{1}{1 + e^{-mq_i}} + \frac{1}{1 + e^{-mq_{i-1}}} - 1 \right)$$

对每个 $i \in \mathbb{Z}_n$ 都成立。显然，$u_i \geq 0$ 满足第一个要求。第二个要求也满足：

$$\begin{aligned} \sum_{i=0}^{n-1} u_i &= \tau \sum_{i=0}^{n-1} \left[\frac{1}{1 + e^{-mq_i}} + \frac{1}{1 + e^{-mq_{i-1}}} - 1 \right] = \tau \sum_{i=0}^{n-1} \left[\frac{2}{1 + e^{-mq_i}} - 1 \right] = \tau \sum_{i=0}^{n-1} \left[\frac{1 - e^{-mq_i}}{1 + e^{-mq_i}} \right] \\ &= \frac{f}{\sum_{i=0}^{n-1} \tanh(mq_i/2)} \cdot \sum_{i=0}^{n-1} \tanh(mq_i/2) = f \,. \end{aligned}$$

现在验证第三个要求。为方便记号，定义 $r_i' := \mathbb{E}[\boldsymbol{R}_i']$, $l_i' := \mathbb{E}[\boldsymbol{L}_i']$，以及 $d_i' := \mathbb{E}[\boldsymbol{D}_i']$。令 R 为由下式给出的 $n \times n$ 矩阵

$$\overline{R}_{i,i+d} = \mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}'} > d \right] = e^{-mq[\langle\!\langle i,d \rangle\!\rangle]}$$

对所有 $i \in \mathbb{Z}_n$ 和 $0 \leq d \leq n-1$ 均成立。则 $R_{i,\ i+d}$ 是生成 $\boldsymbol{R}_i'$ 的马尔可夫过程至少一次计数 u_{i+d} 的概率。注意，$\varepsilon = \mathbb{P}[E] = e^{-m\|q\|_1}$ 是该过程循环回到同一元素 i 一次的概率。根据马尔可夫性质，$\boldsymbol{R}_i'$ 的期望为

$$\begin{aligned} r_i' &= \sum_{t=0}^{\infty} \sum_{d=0}^{n-1} \varepsilon^t \overline{R}_{i,i+d} u_{i+d} = \sum_{d=0}^{n-1} \overline{R}_{i,i+d} u_{i+d} + \sum_{t=1}^{\infty} \sum_{d=0}^{n-1} \varepsilon^t \overline{R}_{i,i+d} u_{i+d} \\ &= \sum_{d=0}^{n-1} \overline{R}_{i,i+d} u_{i+d} + \varepsilon \sum_{t=0}^{\infty} \sum_{d=0}^{n-1} \varepsilon^t \overline{R}_{i,i+d} u_{i+d} = \sum_{d=0}^{n-1} \overline{R}_{i,i+d} u_{i+d} + \varepsilon r_i' \,, \end{aligned}$$

因此

$$\begin{aligned} (1-\varepsilon) r_i' &= \sum_{d=0}^{n-1} \overline{R}_{i,i+d} u_{i+d} = \sum_{d=0}^{n-1} \overline{R}_{i,i+d} \cdot \tau \left(\frac{1}{1 + e^{-mq_{i+d}}} + \frac{1}{1 + e^{-mq_{i+d-1}}} - 1 \right) \\ &= \tau \left(\sum_{d=0}^{n-1} \frac{\overline{R}_{i,i+d}}{1 + e^{-mq_{i+d}}} - \overline{R}_{i,i+d} \right) + \tau \left(\sum_{d=0}^{n-1} \frac{\overline{R}_{i,i+d}}{1 + e^{-mq_{i+d-1}}} \right) \\ &= -\tau \left(\sum_{d=0}^{n-1} \overline{R}_{i,i+d} \frac{e^{-mq_{i+d}}}{1 + e^{-mq_{i+d}}} \right) + \tau \left(\sum_{d=0}^{n-1} \overline{R}_{i,i+d} \frac{1}{1 + e^{-mq_{i+d-1}}} \right) . \end{aligned}$$

现在观察到，对于 $1 \le d \le n-1$, $R_{i,\ i+d} = e^{-mq_{i+d-1}} R_{i,\ i+d-1}$。另请注意，$R_{i,\ i} = 1$。对上面的第二个求和式进行变量代换后，我们得到

$$\begin{aligned}
\frac{1-\varepsilon}{\tau} r'_i &= -\left(\sum_{d=0}^{n-2} \overline{R}_{i,i+d} \frac{e^{-mq_{i+d}}}{1+e^{-mq_{i+d}}} \right) - \overline{R}_{i,i+n-1} \frac{e^{-mq_{i+n-1}}}{1+e^{-mq_{i+n-1}}} \\
&\qquad + \overline{R}_{i,i} \frac{1}{1+e^{-mq_{i-1}}} + \left(\sum_{d=1}^{n-1} e^{-mq_{i+d-1}} R_{i,i+d-1} \frac{1}{1+e^{-mq_{i+d-1}}} \right) \\
&= -\left(\sum_{d=0}^{n-2} \overline{R}_{i,i+d} \frac{e^{-mq_{i+d}}}{1+e^{-mq_{i+d}}} \right) - \overline{R}_{i,i+n-1} \frac{e^{-mq_{i+n-1}}}{1+e^{-mq_{i+n-1}}} \\
&\qquad + \frac{1}{1+e^{-mq_{i-1}}} + \left(\sum_{d=0}^{n-2} \overline{R}_{i,i+d} \frac{e^{-mq_{i+d}}}{1+e^{-mq_{i+d}}} \right) \\
&= -\overline{R}_{i,i+n-1} \frac{e^{-mq_{i+n-1}}}{1+e^{-mq_{i+n-1}}} + \frac{1}{1+e^{-mq_{i-1}}} \,.
\end{aligned}$$

另请注意，$\varepsilon = e^{-mq_{i+n-1}} R_{i,\ i+n-1}$，以及 $q_{i-1} = q_{i+n-1}$，因此

$$\frac{1-\varepsilon}{\tau} r'_i = -\frac{\varepsilon}{1+e^{-mq_{i+n-1}}} + \frac{1}{1+e^{-mq_{i-1}}} = (1-\varepsilon) \frac{1}{1+e^{-mq_{i-1}}} \,.$$

由于 $\varepsilon < 1$（因为 $\|q\| > 0$），我们得出结论：

$$r'_i = \frac{\tau}{1+e^{-mq_{i-1}}} \,.$$

对确定 $\boldsymbol{L}'_i$ 的对称过程进行同样的分析，可得

$$l'_i = \frac{\tau}{1+e^{-mq_i}} \,.$$

现在验证第三个条件：对于每个 $i \in \mathbb{Z}_n$,

$$\mathbb{E}\left[\boldsymbol{D}'_i\right] = d'_i = l'_i + r'_i - u_i = \frac{\tau}{1+e^{-mq_i}} + \frac{\tau}{1+e^{-mq_{i-1}}} - \tau \left(\frac{1}{1+e^{-mq_i}} + \frac{1}{1+e^{-mq_{i-1}}} - 1 \right) = \tau \,,$$

正如所需。

还需证明 $\mathbb{E}[\boldsymbol{D}_i]$ 与 $\mathbb{E}[\boldsymbol{D}'_i] = \tau$ 的差不超过 $4n\xi$。令 $l_i := \mathbb{E}[\boldsymbol{L}_i]$，并令 $r_i := \mathbb{E}[\boldsymbol{R}_i]$。我们将证明 $|r_i - r'_i|$ 和 $|l_i - l'_i|$ 都很小。

固定某个 $i \in \mathbb{Z}_n$。回想一下，$\boldsymbol{N}_i^{\boldsymbol{R}'}$ 统计了 $\boldsymbol{R}'_i$ 的马尔可夫过程在停止前访问了多少个状态，也就是说，该过程在停止前恰好连续 $\boldsymbol{N}_i^{\boldsymbol{R}'} - 1$ 次对 $\boldsymbol{P}_t = 0$ 进行了采样并前进到下一个状态（顶点）。令 $\boldsymbol{K} := \left\lfloor \frac{\boldsymbol{N}_i^{\boldsymbol{R}'} - 1}{n} \right\rfloor$，那么 $\boldsymbol{K}$ 就是该过程“绕回”并再次到达顶点 i 的次数。

然后，回想事件 E 的定义，注意到 $\mathbb{P}[\boldsymbol{K} \ge 1] = \mathbb{P}[E] = \varepsilon = e^{-m\|q\|_1}$。更一般地，根据马尔可夫性质，对于每个非负整数 k，我们都有 $\mathbb{P}[\boldsymbol{K} \ge k] \le e^{-km\|q\|_1}$。

现在，我们可以如下界定 $|r_i - r'_i|$。首先，注意到 $r_i \le r'_i$，因为 u 是一个非负向量；尽管两个过程的合并（或转移）概率相同，马尔可夫过程在访问 n 元素后仍可能继续。事实上，我们有 $\mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}} = t\right] = \mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}'} = t\right]$

对每个 $1 \le t \le n-1$ 都成立，因此 $\mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}} = n\right] \ge \mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}'} = n\right]$。现在，只需对 $r_i' - r_i$ 给出上界，具体如下：

$$\begin{aligned}
r_i' - r_i &= \sum_{t=1}^{\infty} \mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}'} = t\right] \sum_{d=0}^{t-1} u_{i+d} - \sum_{t=1}^{n} \mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}} = t\right] \sum_{d=0}^{t-1} u_{i+d} \\
&= \sum_{t=1}^{n} \left(\mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}'} = t\right] - \mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}} = t\right]\right) \sum_{d=0}^{t-1} u_{i+d} + \sum_{t=n+1}^{\infty} \mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}'} = t\right] \sum_{d=0}^{t-1} u_{i+d} \\
&\le \sum_{k=1}^{\infty} \sum_{t=1}^{n} \mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}'} = kn + t\right] \sum_{d=0}^{kn+t-1} u_{i+d} \\
&\le \sum_{k=1}^{\infty} n\mathbb{P}\left[\boldsymbol{N}_i^{\boldsymbol{R}'} \ge kn + 1\right] (k+1) f = nf \sum_{k=1}^{\infty} \mathbb{P}\left[\boldsymbol{K} \ge k\right] (k+1) \\
&\le nf \sum_{k=1}^{\infty} (k+1) e^{-km\|q\|_1} \le nf \cdot 2 \sum_{k=1}^{\infty} k e^{-km\|q\|_1} \le 2nf \cdot \frac{e^{-m\|q\|_1}}{(1 - e^{-m\|q\|_1})^2},
\end{aligned}$$

其上界为 $2n\xi(m,q)$，其中 ξ 的定义与定义 4.8 中相同。类似的分析表明，$|l_i' - l_i| \le 2n\xi(m,q)$。因此 $|d_i' - d_i| \le 4n\xi(m,q)$，从而 $\mathbb{E}\left[\boldsymbol{D}_i\right] = \mathbb{E}\left[\boldsymbol{D}_i'\right] \pm 4n\xi(m,q) = \tau \pm 4n\xi(m,q)$。因此，$\widetilde{p} = u$ 是具有所需参数的近似均匀共轭分布。此外，可以验证我们得到的解 $\widetilde{p}$ 在 $q = \mu$ 时给出 μ。 □

为了分析 $\tau = \frac{1 - \|q\|_1}{\sum_{i=0}^{n-1} \text{双曲正切}\binom{mq_i}{2}}$，我们需要以下关于双曲正切$(x)$的界。

事实 4.12。当 $x > 0$ 足够小时，

$$\frac{x}{2} \le \tanh(x) \le 2x.$$

证明。这可由泰勒展开 $\tanh(x) = x - \frac{x^3}{3} + \frac{2x^5}{15} + O(x^7)$ 得出。 □

记号 对于向量 $u \in \mathbb{R}^{\mathbb{Z}_n}$，我们将 u^+ 记为对每个$i \in \mathbb{Z}_n$由 $u_i^+ = \max(0, u_i)$ 给出的向量，并将 u^- 记为对每个 $i \in \mathbb{Z}_n$由 $u_i^- = -\min(0, u_i)$ 给出的向量。

命题 4.13（双曲正切在零点右侧附近的二次上界）。对于所有充分小的实数 $r > 0$ 以及所有 $0 \le x \le \frac{1}{2 \text{双曲正切}(r)}$，有

$$\tanh(r + x) \le \tanh(r) + (1 - \tanh^2(r))x - \tanh(r)(1 - \tanh^2(r))x^2.$$

证明。定义函数 f、g：$\mathbb{R}_{\ge 0} \to \mathbb{R}$ 为所需不等式两边的量：

$$\begin{aligned}
f(x) &:= \tanh(r + x), \\
g(x) &:= \tanh(r) + (1 - \tanh^2(r))x - \tanh(r)(1 - \tanh^2(r))x^2.
\end{aligned}$$

因此，我们希望证明：对于充分小的 r，对所有 $0 \le x \le \frac{1}{2 \text{双曲正切}(r)}$，$f(x) \le g(x)$。

回想一下，在非负定义域内，双曲正切的值介于 0 和 1 之间；当 $r \to 0$ 时，双曲正切$(r) = r \pm O(r^3)$（这可由其泰勒级数推出）；当 $y \to \infty$ 时，双曲正切$(y) \to 1$。由于 g 是

开口向下的抛物线，我们首先确定 x^*，使 g 取得最大值。我们可以令导数 g' 等于零来确定该点：

$$g'(x^*) = 0 \implies (1 - \tanh^2(r)) - 2\tanh(r)(1 - \tanh^2(r))x^* = 0 \implies x^* = \frac{1}{2\tanh(r)}\,.$$

现在，我们的策略是定义一个"断点" $c\log(1/r)$（其中c 是一个足够小的常数，具体取值稍后给出），并证明 $g(c\log(1/r)) \to \infty$，从而将论证分为两部分：$c\log(1/r) \le x \le x^*$，此时 g 递增，因此结论立即成立；以及$0 \le x \le c\log(1/r)$，这部分还需要进一步论证。

我们首先证明，对每个 $c > 0$，当 $r \to 0$时，$g(c\log(1/r)) \to \infty$：

$$\begin{aligned} g(c\log(1/r)) &\ge (1 - \tanh^2(r))c\log(1/r)(1 - \tanh(r)c\log(1/r)) \\ &\ge \frac{1}{2}c\log(1/r)(1 - 2cr\log(1/r)) \\ &\ge \frac{c\log(1/r)}{4} \\ &= \omega(1)\,, \end{aligned}$$

在最后一个不等式中，我们用到了这样一个事实：$r\log(1/r) \to 0$。

注意，对所有足够小的r，$c\log_{\text{双曲正切}(r)}$。这意味着 $g(x)$在 $[c\log(1/r)\ x^*]$ 上递增，因此在此值域内 $g(x) > 1$。由于对所有 x都有 $f(x) < 1$，我们已证明，当 $c\log(1/r) \le x \le x^*$时，$f(x) \le g(x)$。

现在我们继续考虑值域 $0 \le x \le c\log(1/r)$。根据泰勒定理的中值形式，我们有

$$f(x) = f(0) + f'(0)x + \frac{1}{2}f''(\eta)x^2$$

对于某个 $0 \le \eta \le x$。代入 f 的定义，得到：

$$f(x) = \tanh(r) + (1 - \tanh^2(r))x - \tanh(r + \eta)(1 - \tanh^2(r + \eta))x^2\,.$$

因此，要证明 $f(x) \le g(x)$，只需证明

$$\tanh(y)(1 - \tanh^2(y)) \ge \tanh(r)(1 - \tanh^2(r))$$

for all $r \le y \le r + x \le r + c\log(1/r)$. We will show this for the l更大 interval $r \le y \le 2c\log(1/r)$。定义 $h\colon\ \mathbb{R}_{\ge 0} \to \mathbb{R}$，令 $h(x) :=$ 双曲正切$(x)(1 -$ 双曲正切$^2(x))$，因此我们希望证明，对于 $r \le y \le 2c\log(1/r)$，有 $h(y) \ge h(r)$。导数 h' 满足以下性质：

1. $h'(x) = (1 - 3$双曲正切$^2(x))(1 -$ 双曲正切$^2(x))$;
2. $h'(0) = 1$;
3. h' 唯一的正实根为 $\nu :\ = \frac{1}{2}\cosh^{-1}(2) =\ \frac{1}{2}\log(2 + \sqrt{3})$; 一个 $\quad^{-}\quad$ d
4. h' 在 $[0, \nu)$ 上为正，在 $(\nu,\ \infty)$ 上为负。

由此可知，h 在 $[r, \nu]$ 上递增，并在 $[\nu, 2c\log(1/r)]$上递减。由于 $h(r) \ge h(r)$ 显然成立，我们得到，对于 $r \le y \le \nu$，$h(y) \ge h(r)$。因此，只需证明 $h(2c\log(1/r)) \ge$

只要 r 足够小，$h(r)$ 就成立。事实上，我们有

$$\begin{aligned}
& h(2c\log(1/r)) \geq h(r) \\
\iff & \tanh(2c\log(1/r))(1-\tanh^2(2c\log(1/r))) \geq \tanh(r)(1-\tanh^2(r)) \\
\iff & \frac{\tanh(2c\log(1/r))}{\tanh(r)} \geq \frac{1-\tanh^2(r)}{1-\tanh^2(2c\log(1/r))} \\
\impliedby & \frac{1/2}{2r} \geq \frac{1}{1-\left(\frac{1-e^{-4c\log(1/r)}}{1+e^{-4c\log(1/r)}}\right)^2} \\
\impliedby & \frac{1}{4r} \geq \frac{1}{1-(1-e^{-4c\log(1/r)})} \\
\iff & \frac{1}{4r} \geq \frac{1}{r^{4c}} \\
\iff & r^{1-4c} \leq \frac{1}{4},
\end{aligned}$$

只要 $c < 1/4$，该结论对所有充分小的 r 都成立，因为此时 $r^{1-4c} \to 0$。证明至此完成。

□

引理 4.14（双曲正切在零点附近的定量詹森不等式）。对于所有充分小的 $r > 0$，以下结论成立。设 $u \in \mathbb{R}^{\mathbb{Z}_n}$ 是一个满足以下条件的向量：对每个 $i \in \mathbb{Z}_n$，都有 $u_i \in \left[0, r + \frac{1}{2}^{1_{\text{双曲正切}(r)}}\right]$，且其平均值为 $\frac{1}{n}\sum_i u_i = r$。则有

$$\frac{1}{n}\sum_{i=0}^{n-1}\tanh(u_i) \leq \tanh(r)\left(1 - \frac{1}{n}\left(1 - O(r^2)\right)\|(u - r\cdot\vec{1})^+\|_2^2\right).$$

证明。写作 $u_i = r + x_i$，于是对每个 $i \in \mathbb{Z}_n$，都有 $x_i \leq 2^{1_{\text{双曲正切}(r)}}$，并且 $\sum_i x_i = \sum_i(u_i - r) = 0$。由于双曲正切在其非负定义域上是凹函数，在$r$处的一阶泰勒级数，即双曲正切$(r+y) \approx$ 双曲正切$(r) + (1 -$ 双曲正切$^2(r))y$，对每个 $i \in \mathbb{Z}_n$ 都给出了双曲正切(u_i) 的上界。

因此，我们的策略是通过一阶级数对 $x_i < 0$ 对应的各项取上界，并通过 $x_i \geq 0$ 利用命题 4.13对相应各项取上界。这样，一阶项将相互抵消，而二阶项则会给出所需的界。具体而言，我们有：

$$\begin{aligned}
\sum_{i=0}^{n-1}\tanh(u_i) &= \sum_{i=0}^{n-1}\tanh(r+x_i) \\
&\leq \sum_{i=0}^{n-1}\Big[\tanh(r) + (1-\tanh^2(r))x_i\Big] - \sum_{i\in\mathbb{Z}_n : x_i \geq 0}\tanh(r)(1-\tanh^2(r))x_i^2 \\
&= n\tanh(r) + (1-\tanh^2(r))\sum_{i=1}^{n}x_i - \tanh(r)(1-\tanh^2(r))\sum_{i:x_i\geq 0}x_i^2 \\
&\leq \tanh(r)\Big(n - \left(1 - O\left(r^2\right)\right)\|(u-r)^+\|_2^2\Big).
\end{aligned}$$

□

现在，我们可以结合上述结果来证明，只要 q 相对于 p没有高度集中，$\mathbb{E}[\boldsymbol{Y}]$ 就存在分离：

引理 4.15（检验统计量期望值的分离）。设 C、$\gamma > 0$ 为常数，设 $n \in \mathbb{N}$ 充分大，设 $\epsilon \geq \frac{1}{n^{1/4}}$，并设 $m = m(n, \epsilon)$ 满足

$$\left(2^8\gamma^2 + (2^{18}\gamma)^{2/5}\right)\left(\frac{n}{\epsilon}\right)^{4/5} \leq m \leq \frac{4n}{3C\log n}.$$

设 $\pi = \pi(p, q)$，其中 p, q 是满足 $\|p\|_1, \|q\|_1 = \frac{1}{2} \pm \frac{2\gamma}{\sqrt{m}}$，$\|q - \mu\|_1 > \epsilon$ 的部分分布，并且 q 相对于 p 不是 C-高度集中分布。记 $p = \widetilde{p} + z$，其中 $\widetilde{p}$ 是 q 的一个近似均匀共轭分布。则检验统计量 $\boldsymbol{Y}$ 的期望值满足

$$\mathbb{E}[\boldsymbol{Y}] \geq \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] + \Omega\left(\frac{\epsilon^2 m^2}{n^2}\right) + mz^\top \phi z.$$

证明。由引理 4.11，μ 是其自身的近似均匀共轭分布。由于我们将同时讨论 μ 作为其自身的近似均匀共轭分布，以及讨论 $\widetilde{p}$ 作为 q的近似均匀共轭分布，因此令 $\xi := \max(\xi(m, \mu)\ \xi(m, q))$。首先，利用 命题 4.9，

$$\mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] = m\mu^\top \phi^{(\mu)}\mu = m\|\mu\|_1(\tau(m,\mu) \pm 4n\xi) = \frac{m\|\mu\|_1^2}{\sum_{i=0}^{n-1}\tanh\left(\frac{m}{4n}\right)} \pm 4mn\|\mu\|_1\xi = \frac{m}{4n\tanh\left(\frac{m}{4n}\right)} \pm 2mn\xi. \tag{9}$$

我们现在考虑 $\mathbb{E}[\boldsymbol{Y}]$。根据 $\|p\|_1 \geq \frac{1}{2} - \frac{2\gamma}{\sqrt{m}}$ 这一假设，我们得到

$$\|p\|_1^2 \geq \left(\frac{1}{2} - \frac{2\gamma}{\sqrt{m}}\right)^2 \geq \frac{1}{4} - \frac{2\gamma}{\sqrt{m}}.$$

令 $\tau := \tau(m, q)$。利用 命题 4.9，

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Y}] &= m\|p\|_1(\tau \pm 4n\xi) + mz^\top\phi z \pm 8mn^2\xi \\
&\geq \frac{m\|p\|_1^2}{\sum_{i=0}^{n-1}\tanh\left(\frac{mq_i}{2}\right)} + mz^\top\phi z - 4mn\xi\|p\|_1 - 8mn^2\xi \\
&\geq \frac{m - 8\gamma\sqrt{m}}{4\sum_{i=0}^{n-1}\tanh\left(\frac{mq_i}{2}\right)} + mz^\top\phi z - 12mn^2\xi.
\end{aligned} \tag{10}$$

现在对每个 $i \in \mathbb{Z}_n$ 将 $q_i = \frac{1}{n}\|q\|_1 + x_i$ 记为，因此

$$\sum_{i=0}^{n-1}\tanh\left(\frac{mq_i}{2}\right) = \sum_{i=0}^{n-1}\tanh\left(\frac{m}{2n}\|q\|_1 + \frac{mx_i}{2}\right),$$

以及 $\sum_{i=0}^{n-1} x_i = 0$。记 $r := \frac{m}{2n}\|q\|_1$ 和 $u_i := \frac{mq_i}{2}$，我们有 $u = \frac{m}{2}q$ 以及

$$0 \leq u_i = r + \frac{mx_i}{2}.$$

由于相对于 p，q 并非 C-高度集中，根据备注 4.5 中的观察，有 $\|q\|_\infty < \frac{C\log n}{m}$，因此我们有

$$x_i < \frac{C\log n}{m} - \frac{1}{n}\|q\|_1 < \frac{C\log n}{m}.$$

此外，由于 $\|q\|_1 \leq \frac{1}{2} + \frac{2\gamma}{\sqrt{m}}$，我们有 r 满足

$$r \leq \frac{m}{4n} + \frac{\gamma\sqrt{m}}{n} \leq \frac{3m}{8n} \implies \tanh(r) \leq \frac{3m}{4n},$$

其中，在第二个不等式中我们用到了这一事实： $\sqrt{m/n} = o(m/n)$，而在最后一个不等式中，我们用到了 事实 4.12 以及假设 $m/n \le 3C^4 \log n$ 且 n 充分大。因此我们得到

$$x_i < \frac{C\log n}{m} = \frac{2}{m}\cdot\frac{C\log n}{2} \le \frac{2}{m}\cdot\frac{2n}{3m} = \frac{2}{m}\cdot\frac{1}{2\cdot 3m/(4n)} \le \frac{2}{m}\cdot\frac{1}{2\tanh(r)}\,,$$

其中，第二个不等式用到了假设 $m \le 3C^{4n}\log n$。因此， u, r 满足这些条件

$$0 \le u_i = r + \frac{mx_i}{2} \le r + \frac{1}{2\tanh(r)}\,, \qquad \frac{1}{n}\sum_{i=0}^{n-1} u_i = \frac{m}{2n}\|q\|_1 = r\,.$$

令 $q' := \frac{1}{n}\|q\|_1 \cdot \vec{1}$，它是一个部分分布，其均匀分布的总质量等于 q的总质量。注意到 $r\cdot\vec{1} = m\|q\|_1\cdot\mu = (m/2)\cdot q'$，应用 引理 4.14，得到

$$\begin{aligned}\frac{1}{n}\sum_{i=0}^{n-1}\tanh\left(\frac{mq_i}{2}\right) = \frac{1}{n}\sum_{i=0}^{n-1}\tanh(u_i) &\le \tanh\left(\frac{m}{2n}\|q\|_1\right)\left(1 - \frac{1}{n}(1-O(r^2))\|((m/2)q - (m/2)q')^+\|_2^2\right)\\ &\le \tanh\left(\frac{m}{2n}\|q\|_1\right)\left(1 - \frac{m^2}{8n}\|(q-q')^+\|_2^2\right)\,,\end{aligned}$$

其中，最后一个不等式用到了 $r^2 = \left(\frac{m}{2n}\|q\|_1\right)^2 = o(1)$ 这一事实。我们还将使用关于 $\|(q-q')^+\|_2^2$ 的以下上界：

$$\|(q-q')^+\|_2^2 \le \|q\|_2^2 < \left(\frac{C\log n}{m}\right)^2 \cdot \frac{1}{\left(\frac{C\log n}{m}\right)} = \frac{C\log n}{m}\,,$$

其中，根据反集中假设，我们用到了 $\|q\|_\infty < C\log_m n$ 这一事实，因此在满足此约束的前提下，尽可能集中部分分布时， ℓ^2-范数达到最大值。由此可得

$$\frac{m^2}{8n}\|(q-q')^+\|_2^2 < \frac{Cm\log n}{8n} \le \frac{1}{6}\,, \tag{11}$$

后者成立是因为 $m \le 3C^{4n}\log n$。因此，我们使用不等式 $\frac{1}{1-x} \ge 1+x$，该不等式对 $x<1$ 有效，如下所示：

$$\begin{aligned}\frac{m - 8\gamma\sqrt{m}}{4\sum_{i=0}^{n-1}\tanh\left(\frac{mq_i}{2}\right)} &\ge \frac{m-8\gamma\sqrt{m}}{4n\tanh\left(\frac{m}{2n}\|q\|_1\right)\left(1-\frac{m^2}{8n}\|(q-q')^+\|_2^2\right)}\\ &\ge \frac{m-8\gamma\sqrt{m}}{4n\tanh\left(\frac{m}{2n}\|q\|_1\right)}\left(1+\frac{m^2}{8n}\|(q-q')^+\|_2^2\right)\,.\end{aligned} \tag{12}$$

由（9）、（10）和（12），现在可得

$$\mathbb{E}[\boldsymbol{Y}] - \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] \ge \frac{m}{4n}\cdot F + G - H + mz^\top\phi z - 14mn^2\xi\,, \tag{13}$$

其中

$$\begin{aligned}F &= \frac{1}{\tanh\left(\frac{m}{2n}\|q\|_1\right)} - \frac{1}{\tanh\left(\frac{m}{4n}\right)}\,,\\ G &= \frac{m}{4n\tanh\left(\frac{m}{2n}\|q\|_1\right)}\cdot\frac{m^2}{8n}\|(q-q')^+\|_2^2\,,\\ H &= \frac{8\gamma\sqrt{m}}{4n\tanh\left(\frac{m}{2n}\|q\|_1\right)}\left(1+\frac{m^2}{8n}\|(q-q')^+\|_2^2\right)\,.\end{aligned}$$

我们将证明， G 足够大，能够实现所需的分离 $\Omega(\epsilon^2 m^2/n^2)$，而 F 和 H 在渐近意义下足够小。我们首先给出 F的下界。利用 $\frac{m}{2n}\|q\|_1 \le \frac{m}{2n}\left(\frac{1}{2}+\frac{2\gamma}{\sqrt{m}}\right) = \frac{m}{4n}+\frac{\gamma\sqrt{m}}{n}$ 这一事实，以及上界 $\tanh(r+x) \le \tanh(r)+x(1-\tanh^2(r))$，可得下式。该上界在所有参数均非负时由双曲正切的泰勒展开得到。

$$F \ge \frac{1}{\tanh\left(\frac{m}{4n}+\frac{\gamma\sqrt{m}}{n}\right)}-\frac{1}{\tanh\left(\frac{m}{4n}\right)} = \frac{\tanh\left(\frac{m}{4n}\right)-\tanh\left(\frac{m}{4n}+\frac{\gamma\sqrt{m}}{n}\right)}{\tanh\left(\frac{m}{4n}\right)\tanh\left(\frac{m}{4n}+\frac{\gamma\sqrt{m}}{n}\right)} \ge -\frac{\frac{\gamma\sqrt{m}}{n}\left(1-\tanh^2\left(\frac{m}{4n}\right)\right)}{\tanh\left(\frac{m}{4n}\right)\tanh\left(\frac{m}{4n}+\frac{\gamma\sqrt{m}}{n}\right)}\,.$$

当 n 足够大，因此 m/n 足够小时，我们有 $\tanh(m/4n+\gamma\sqrt{m}/n) > \tanh(m/4n) \ge m/8n$，这由事实 4.12 得出。我们得到

$$\frac{m}{4n}\cdot F \ge -\frac{m}{4n}\cdot\frac{\gamma\sqrt{m}/n}{(m/8n)^2} = -\frac{16\cdot\gamma}{\sqrt{m}}\,.$$

我们验证这个负因子不会压倒所需的分离 $\frac{\epsilon^2 m^2}{n^2}$，如下所示：

$$\frac{16\gamma}{\sqrt{m}} \le \frac{\epsilon^2 m^2}{2^{12}\cdot n^2} \iff m^{5/2} \ge \frac{2^{16}\cdot\gamma\cdot n^2}{\epsilon^2} \iff m \ge (2^{16}\cdot\gamma)^{2/5}(n/\epsilon)^{4/5}\,,$$

这由假设成立。

对于 G，我们利用界 $\tanh\left(\frac{m}{2n}\|q\|_1\right) \le \tanh(m/2n) \le m/n$ 以及柯西-施瓦茨不等式，得到

$$G \ge \frac{m^2}{32n}\|(q-q')^+\|_2^2 \ge \frac{m^2}{32n^2}\|(q-q')^+\|_1^2 = \frac{m^2}{2^7\cdot n^2}\|q-q'\|_1^2\,,$$

这里等号成立是因为，既然 $\|q\|_1 = \|q'\|_1$，我们有 $\|(q-q')^+\|_1 = \|(q-q')^-\|_1 = \frac{1}{2}\|q-q'\|_1$。

因此，我们的目标是证明 $\|q-q'\|_1^2$ 不会比 ϵ^2小很多。根据三角不等式，我们有

$$\epsilon < \|q-\mu\|_1 \le \|q-q'\|_1+\|q'-\mu\|_1 = \|q-q'\|_1+\sum_{i=0}^{n-1}\left|\frac{\|q\|_1}{n}-\frac{1}{2n}\right| = \|q-q'\|_1+\left|\|q\|_1-\frac{1}{2}\right| \le \|q-q'\|_1+\frac{2\gamma}{\sqrt{m}}\,,$$

因此，利用始终成立的 $\epsilon \le 2$，

$$G \ge \frac{m^2}{2^7\cdot n^2}\left(\epsilon-\frac{2\gamma}{\sqrt{m}}\right)^2 > \frac{m^2}{2^7\cdot n^2}\left(\epsilon^2-\frac{8\gamma}{\sqrt{m}}\right) \ge \frac{\epsilon^2 m^2}{2^8\cdot n^2}\,,$$

其中，最后一个不等式由下述推导得到，利用假设 $m \ge 2^8\cdot\gamma^2(n/\epsilon)^{4/5}$：

$$\frac{8\gamma}{\sqrt{m}} \le \frac{\epsilon^2}{2} \iff m \ge \frac{2^8\gamma^2}{\epsilon^4}$$
$$\impliedby 2^8\gamma^2(n/\epsilon)^{4/5} \ge \frac{2^8\gamma^2}{\epsilon^4} \iff \epsilon^{16/5} \ge \frac{1}{n^{4/5}} \iff \epsilon \ge \frac{1}{n^{1/4}}\,,$$

根据假设，这一结论成立。

我们还证明了 H 不会压过这一项。对于足够大的 n，因此 m，我们有不等式 双曲正切 $\left(\frac{m}{2n}\|q\|_1\right) \ge$ 双曲正切 $\left(\frac{m}{2n}\left(\frac{1}{2}-\frac{2\gamma}{\sqrt{m}}\right)\right) \ge$ 双曲正切 $\left(\frac{m}{8n}\right) \ge m/16n$。结合（11），我们得到

$$H < \frac{8\gamma\sqrt{m}}{4n\cdot(m/16n)}\left(1+\frac{1}{6}\right) < \frac{2^6\cdot\gamma}{\sqrt{m}} \le \frac{\epsilon^2 m^2}{2^{12}\cdot n^2}\,,$$

其中最后一个不等式成立是因为

$$\frac{2^6\cdot\gamma}{\sqrt{m}}\le\frac{\epsilon^2m^2}{2^{12}\cdot n^2}\iff m^{5/2}\ge\frac{2^{18}\gamma n^2}{\epsilon^2}\iff m\ge(2^{18}\gamma)^{2/5}(n/\epsilon)^{4/5}\,.$$

最后，我们考察误差项 $-14mn^2\xi$。回忆一下，$\xi=\max(\xi(m,\mu)\ \xi(m,q))$，其中 $\xi(m,\mu)=e^{-m\|\mu\|_1}(1-e^{-m\|\mu\|_1})^2=\frac{e^{-m/2}}{(1-e^{-m/2})^2}$，且 $\xi(m,q)=\frac{e^{-m\|q\|_1}}{(1-e^{-m\|q\|_1})^2}$。利用界 $\|q\|_1\ge\frac12-\frac{2\gamma}{\sqrt{m}}\ge\frac14$ as$n$ 并令 m 增大，我们得出 $\xi\le2e^{-m/4}$；因此，利用（简化后的）假设 $\Omega(n^{4/5})\le m\le n$，我们得出

$$mn^2\xi\le n^3e^{-\Omega(n^{4/5})}=o(n^{-5/2})=o\left(\frac{\epsilon^2}{n^2}\right)=o\left(\frac{\epsilon^2m^2}{n^2}\right),$$

其中最后第二步用到了假设 $\epsilon\ge n^{-1/4}$。回到（13），我们得到

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{Y}\right]-\mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right]&\ge\frac{m}{4n}\cdot F+G-H+mz^\top\phi z-14mn^2\xi\ge-3\cdot\frac{\epsilon^2m^2}{2^{12}n^2}+\frac{\epsilon^2m^2}{2^8\cdot n^2}+mz^\top\phi z\\&=\Omega\left(\frac{\epsilon^2m^2}{n^2}\right)+mz^\top\phi z\,.\end{aligned}$$

□

4.4 检验统计量的集中

在本节中，我们从第 2.5.3 节 中确立的一般结果出发，并在当前情形下得出关于 $\boldsymbol{Y}$ 方差的具体界。

引理 4.16 (方差的第一部分)。设 $C>0$ 为常数，$n\in\mathbb{N}$ 充分大，且 m 满足 $m\le\mathrm{poly}(n)$。设 $\pi=\pi(p,q)$，其中 p、q 为部分分布，满足$\|p\|_1$、$\|q\|_1\ge1/4$，且 p 相对于 q并非 C-高度集中。那么

$$\operatorname*{Var}_{\boldsymbol{H}}\left[\operatorname*{\mathbb{E}}_{\boldsymbol{T}}[\boldsymbol{Y}\ \mid\ \boldsymbol{H}]\right]\le O\left(\log^6n\right)\cdot p^\top\phi p+O(m^2e^{-m/8})\,.$$

证明。根据 引理 2.9，存在某个绝对常数 $c>0$,

$$\operatorname*{Var}_{\boldsymbol{H}}\left[\operatorname*{\mathbb{E}}_{\boldsymbol{T}}[\boldsymbol{Y}\ \mid\ \boldsymbol{H}]\right]\le5m^2\zeta(\mathcal{I})\|p\|_1^4+cm^2\cdot\sum_{\substack{I=\langle\!\langle i,d\rangle\!\rangle\in\mathcal{I}\\1\le d\le n}}p_ip_{i+d-1}p[I]^2\mathbb{E}\left[\boldsymbol{J}(I)\right]\,.\tag{14}$$

Our 的假设表明，p 相对于 q 并非高度集中，因此有如下不等式

$$p[\langle\!\langle i,d+1\rangle\!\rangle]^2\le\left[C\log^2(n)\cdot\max\left\{q[\langle\!\langle i,d\rangle\!\rangle],\frac{1}{m\log n}\right\}\right]^2\le C^2\log^4(n)q[\langle\!\langle i,d\rangle\!\rangle]^2+\frac{C^2\log^2n}{m^2}\,,$$

因此，对 d 做一个小的变量代换,

$$\begin{aligned}&\sum_{\substack{I=\langle\!\langle i,d\rangle\!\rangle\in\mathcal{I}\\1\le d\le n}}p_ip_{i+d-1}p[I]^2\mathbb{E}\left[\boldsymbol{J}[I]\right]\\&\quad=\sum_{i=0}^{n-1}\sum_{d=0}^{n-1}p_ip_{i+d}\mathbb{E}\left[\boldsymbol{J}(\langle\!\langle i,d+1\rangle\!\rangle)\right]p[\langle\!\langle i,d+1\rangle\!\rangle]^2\\&\quad\le C^2\log^4(n)\sum_{i=0}^{n-1}\sum_{d=0}^{n-1}p_ip_{i+d}\mathbb{E}\left[\boldsymbol{J}(\langle\!\langle i,d+1\rangle\!\rangle)\right]q[\langle\!\langle i,d\rangle\!\rangle]^2+\frac{C^2\log^2n}{m^2}\sum_{i=0}^{n-1}\sum_{d=0}^{n-1}p_ip_{i+d}\mathbb{E}\left[\boldsymbol{J}(\langle\!\langle i,d+1\rangle\!\rangle)\right]\,.\end{aligned}$$

我们证明上面的两项都满足所需的渐近界。对于第二项，注意到 $\boldsymbol{J}(\langle\!\langle i, d+1\rangle\!\rangle) = 1 \implies \boldsymbol{\Phi}_{i,\ i+d} = 1$，因此 $\mathbb{E}\left[\boldsymbol{J}(\langle\!\langle i, d+1\rangle\!\rangle)\right] \le \phi_{i,\ i+d}$。于是

$$\frac{C^2 \log^2 n}{m^2} \sum_{i=0}^{n-1} \sum_{d=0}^{n-1} p_i p_{i+d} \mathbb{E}\left[\boldsymbol{J}(\langle\!\langle i, d+1\rangle\!\rangle)\right] \le \frac{C^2 \log^2 n}{m^2} \sum_{i=0}^{n-1} \sum_{d=0}^{n-1} p_i p_{i+d} \phi_{i,i+d} = O\left(\frac{\log^2 n}{m^2}\right) \cdot p^\top \phi p\,, \tag{15}$$

如所需。至于第一项，回顾

$$\mathbb{E}\left[\boldsymbol{J}(\langle\!\langle i, d+1\rangle\!\rangle)\right] = \mathbb{P}\left[\forall e \in \langle\!\langle i, d\rangle\!\rangle : e \in \boldsymbol{H}\right] = \prod_{e \in \langle\!\langle i,d\rangle\!\rangle} (1 - w(e)) = e^{-mq[\langle\!\langle i,d\rangle\!\rangle]}\,.$$

令 $K \ge 1$ 为一个常数，使得对于所有充分大的 n，都有 $m \le n^K$；这是基于假设$m \le \mathsf{poly}(n)$。现在，令 $x := q[\langle\!\langle i, d\rangle\!\rangle]$，它的取值介于 0 和 1 之间，我们考虑两种情况。首先，假设 $x \ge {}^{2}{}_{\log_m}{}^{(K+2)}{}^{n}$。于是得到

$$x^2 e^{-mx} \le 1 \cdot e^{-2(K+2)\log n} = n^{-2K-4}\,.$$

另一方面，若 $x \le {}^{2}{}_{\log_m}{}^{(K+2)\log n}{}^{n}$，则

$$x^2 e^{-mx} \le O\left(\frac{\log^2 n}{m^2}\right) e^{-mx}\,.$$

因此，第一项为

$$\begin{aligned}
&C^2 \log^4(n) \sum_{i=0}^{n-1} \sum_{d=0}^{n-1} p_i p_{i+d} e^{-mq[\langle\!\langle i,d\rangle\!\rangle]} q[\langle\!\langle i, d\rangle\!\rangle]^2 \\
&\quad \le O\left(\log^4 n\right) \sum_{i=0}^{n-1} \sum_{d=0}^{n-1} p_i p_{i+d} \left(n^{-2K-4} + O\left(\frac{\log^2 n}{m^2}\right) e^{-mq[\langle\!\langle i,d\rangle\!\rangle]}\right) \\
&\quad \le O(n^{-2K-1}) + O\left(\frac{\log^6 n}{m^2}\right) \cdot p^\top \phi p\,,
\end{aligned}$$

最后一步中，我们再次使用了不等式 $e^{-mq[\langle\!\langle i,d\rangle\!\rangle]} = \mathbb{E}\left[\boldsymbol{J}(\langle\!\langle i, d+1\rangle\!\rangle)\right] \le \phi_{i,i+d}$。

为了给项 $O(n^{-2K-1})$ 取一个上界，我们回顾 $m \le n^K$，并注意到，由于 ϕ 在对角线上为 1，且 $\|p\|_1 \ge 1/4$，因此有 $p^\top \phi p \ge \|p\|_2^2 \ge \Omega(1/n)$。于是

$$n^{-2K-1} = \frac{1}{n^{2K}} \cdot \frac{1}{n} \le \frac{1}{m^2} \cdot O(p^\top \phi p)\,.$$

由此可得

$$C^2 \log^4(n) \sum_{i=0}^{n-1} \sum_{d=0}^{n-1} p_i p_{i+d} e^{-mq[\langle\!\langle i,d\rangle\!\rangle]} q[\langle\!\langle i, d\rangle\!\rangle]^2 \le O\left(\frac{\log^6 n}{m^2}\right) \cdot p^\top \phi p\,. \tag{16}$$

至于误差项 $5m^2 \zeta(\mathcal{I}) \|p\|_1^4$，我们将 $\|p\|_1$ 上界取为 1，并回顾 $\zeta(\mathcal{I}) \le e^{-\frac{m\|q\|_1}{2}}$由 命题 2.5。结合假设 $\|q\|_1 \ge 1/4$，可得

$$5m^2 \zeta(\mathcal{I}) \|p\|_1^4 \le O(m^2 e^{-m/8})\,, \tag{17}$$

按需。将 (15)、(16) 和 (17) 代入 (14)，我们得出

$$\operatorname*{Var}_{\boldsymbol{H}}\left[\mathop{\mathbb{E}}_{\boldsymbol{T}}\left[\boldsymbol{Y} \mid \boldsymbol{H}\right]\right] \le O(\log^6 n) p^\top \phi p + O(m^2 e^{-m/8})\,. \qquad \square$$

为了使上述结果发挥作用，我们需要用一个与 $p^\top \phi p$ 中所示分离量级相当的量来给二次型引理 4.15一个上界。回顾按近似均匀共轭分布进行的分解， $p = \widetilde{p} + z$，我们的首要任务是为 $\widetilde{p}^\top \phi \widetilde{p}$求上界。

命题4.17。 设 $C > 0$ 为一个常数， $n \in \mathbb{N}$ 充分大，并假设 m 满足$m = \omega(\log n)$， $m = o(n)$。令 $\pi = \pi(p, q)$，其中 p, q 是满足 $\|q\|_1 \geq 1/4$ 的部分分布，且 q 相对于 p 并非 C- 高度集中。设 $\widetilde{p}$ 是 q的近似均匀共轭分布。那么

$$\widetilde{p}^\top \phi \widetilde{p} = O\left(\frac{\log n}{m}\right).$$

证明。 由 命题 4.9， $\widetilde{p}^\top \phi \widetilde{p} = \|p\|(\tau \pm 4n\xi)$。我们的主要任务是证明 $\tau = O\left(\frac{\log n}{m}\right)$，但我们先检查 $4n\xi$ 是否足够小。事实上，根据 定义 4.8 ，且由于 $\|q\|_1 \geq 1/4$，我们有

$$n\xi = n \cdot \frac{e^{-m\|q\|_1}}{(1 - e^{-m\|q\|_1})^2} \leq \frac{n \cdot e^{-\Omega(m)}}{(1 - e^{-\Omega(m)})^2} = o(1/n) \leq o(1/m),$$

最后两个步骤分别是 $m = \omega$（对数 n）和 $m = o$（n）。现在我们研究 τ。回顾

$$\tau = \frac{\|p\|_1}{\sum_{i=0}^{n-1} \tanh\left(\frac{mq_i}{2}\right)}.$$

由于双曲正切在非负定义域上是凹函数，我们的目标是使用詹森不等式对 τ 给出上界。令 $q = \mu + y$。令 $S := \{i \in \mathbb{Z}_n : \ y_i \geq 0\}$ 以及 $S := \{i \in \mathbb{Z}_n : \ y_i < 0\} = \overline{\mathbb{Z}}_n \setminus S$。注意，向量 y^+ 仅在 S中取非零值，而 y^- 仅在S中取非零值。对于每个 $i \in S$，我们有 $\frac{1}{2n} \leq q_i \leq {}^{C}\!/_{\text{对数}_m}{}^{n}$，这是上界，因为 $\|q\|_\infty \leq {}^{C}\!/_{\text{对数}_m}{}^{n}$，由反集中假设可知（备注 4.5）。于是我们可以写成

$$q_i = \lambda_i \cdot \frac{C \log n}{m} + (1 - \lambda_i) \cdot \frac{1}{2n}, \qquad \text{where} \qquad \lambda_i = \frac{y_i}{\frac{C \log n}{m} - \frac{1}{2n}} \in [0, 1].$$

对于 $i \in \overline{S}$，我们有 $0 \leq q_i \leq \frac{1}{2n}$，因此可以写成

$$q_i = \lambda_i \cdot \frac{1}{2n}, \qquad \text{where} \qquad \lambda_i = 2nq_i.$$

应用詹森不等式，并利用双曲正切 $\left(\frac{m}{4n}\right) \geq \frac{m}{8n}$ 在足够小的 $m/4n$(事实 4.12),

$$\begin{aligned}
\sum_{i=0}^{n-1} \tanh\left(\frac{mq_i}{2}\right) &= \sum_{i \in S} \tanh\left(\lambda_i \cdot \frac{C \log n}{2} + (1 - \lambda_i) \cdot \frac{m}{4n}\right) + \sum_{i \in \overline{S}} \tanh\left(\lambda_i \cdot \frac{m}{4n}\right) \\
&\geq \sum_{i \in S} \left[\lambda_i \tanh\left(\frac{C \log n}{2}\right) + (1 - \lambda_i) \tanh\left(\frac{m}{4n}\right)\right] + \sum_{i \in \overline{S}} \lambda_i \tanh\left(\frac{m}{4n}\right) \\
&\geq \frac{1}{2} \sum_{i \in S} \lambda_i + \sum_{i \in S} (1 - \lambda_i) \cdot \frac{m}{8n} + \sum_{i \in \overline{S}} \lambda_i \cdot \frac{m}{8n} \\
&= \frac{1}{2} \sum_{i \in S} \lambda_i + \frac{m}{8n} \left(\sum_{i \in S} (1 - \lambda_i) + \sum_{i \in \overline{S}} \lambda_i\right),
\end{aligned}$$

这里我们用到了这样一个事实：对于足够大的 n，双曲正切 $\left(C_{\text{对数}_2\ n}\right) \geq \frac{1}{2}$。我们考虑两种情况。首先假设 $\|y^+\|_1 \geq \frac{1}{8}$。那么

$$\frac{1}{2}\sum_{i\in S}\lambda_i = \frac{1}{2}\cdot\frac{m}{C\log n - \frac{m}{2n}}\|y^+\|_1 \geq \frac{m}{16C\log n}\,,$$

因此

$$\tau = \frac{\|p\|_1}{\sum_{i=0}^{n-1}\tanh\left(\frac{mq_i}{2}\right)} \leq \frac{16C\log n}{m}\,,$$

正如所需。接下来假设$\|y^+\|_1 < \frac{1}{8}$。注意，对于 $i \in S$,

$$\lambda_i = \frac{y_i}{\frac{C\log n}{m} - \frac{1}{2n}} = \frac{my_i}{C\log n - \frac{m}{2n}} \leq \frac{2my_i}{C\log n}\,,$$

因为 $\frac{m}{2n} \leq \frac{1}{2}C\log n$。所以

$$\begin{aligned}\sum_{i\in S}(1-\lambda_i) + \sum_{i\in\overline{S}}\lambda_i &\geq \sum_{i\in S}\left(1 - \frac{2my_i}{C\log n}\right) + 2n\sum_{i\in\overline{S}}q_i = \sum_{i\in S}\left(1-\frac{2my_i}{C\log n}\right) + \sum_{i\in\overline{S}}(1+2ny_i)\\ &= |S| - \frac{2m}{C\log n}\|y^+\|_1 + |\overline{S}| - 2n\|y^-\|_1\\ &= n - \frac{2m}{C\log n}\|y^+\|_1 - 2n\|y^-\|_1\,.\end{aligned}$$

回想假设 $\|q\|_1 \geq 1/4$，现在注意到

$$\|q\|_1 = \sum_{i\in S}\left(\frac{1}{2n} + |y_i|\right) + \sum_{i\in\overline{S}}\left(\frac{1}{2n} - |y_i|\right) = \frac{1}{2} + \|y^+\|_1 - \|y^-\|_1 \geq \frac{1}{4}\,,$$

因此$\|y^-\|_1 \leq \|y^+\|_1 + \frac{1}{4}$。接着

$$\begin{aligned}\sum_{i\in S}(1-\lambda_i) + \sum_{i\in\overline{S}}\lambda_i &\geq n - \frac{2m}{C\log n}\|y^+\|_1 - 2n\|y^-\|_1\\ &\geq n - \frac{2m}{C\log n}\|y^+\|_1 - 2n\left(\|y^+\|_1 + \frac{1}{4}\right)\\ &= n - \|y^+\|_1\left(\frac{2m}{C\log n} + 2n\right) - \frac{n}{2}\\ &\geq \frac{n}{2} - \frac{1}{8}\left(\frac{2m}{C\log n} + 2n\right) = \left(\frac{1}{4} - o(1)\right)n\,.\end{aligned}$$

我们得出结论：

$$\sum_{i=0}^{n-1}\tanh\left(\frac{mq_i}{2}\right) \geq \frac{m}{8n}\left(\sum_{i\in S}(1-\lambda_i) + \sum_{i\in\overline{S}}\lambda_i\right) \geq \frac{m}{8n}\left(\frac{1}{4} - o(1)\right)n \geq \frac{m}{33}\,,$$

对于足够大的 n，此时 $\tau \leq \frac{33}{m}$。证明至此结束。 □

推论 4.18. 在引理 *4.16* 和命题*4.17*的假设下，并记 $p = \widetilde{p} + z$，方差的第一部分满足

$$\operatorname*{Var}_{\boldsymbol{H}}\left[\mathop{\mathbb{E}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right] \le O\left(\frac{\log^7 n}{m}\right) + O\left(\log^6 n\right) z^\top \phi z\,.$$

证明。注意 $m^2 e^{-m/8} = o(1/m)$，因为 $m = \omega(\log n) = \omega(1)$。因此 引理 4.16结合分解 $p = \widetilde{p} + z$，得到

$$\begin{aligned}\operatorname*{Var}_{\boldsymbol{H}}\left[\mathop{\mathbb{E}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right] &\le O\left(\log^6 n\right) \cdot (\widetilde{p} + z)^\top \phi(\widetilde{p} + z) + O(m^2 e^{-m/8}) \\ &= O\left(\log^6 n\right)\left(\widetilde{p}^\top \phi \widetilde{p} + z^\top \phi z + 2 z^\top \phi \widetilde{p}\right) + o(1/m)\,.\end{aligned}$$

通过与 命题 4.9 证明中类似的论证，可得 $z^\top \phi \widetilde{p} \le 4n^2 \xi = o(1/m)$，最后一步与 命题4.17 证明中相同。应用命题4.17于项 $\widetilde{p}^\top \phi \widetilde{p}$，得到

$$\operatorname*{Var}_{\boldsymbol{H}}\left[\mathop{\mathbb{E}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right] \le O\left(\frac{\log^7 n}{m}\right) + O\left(\log^6 n\right) z^\top \phi z\,. \qquad \square$$

现在我们对方差的第二个分量给出上界。关键步骤是证明：以高概率，没有任何桶包含过多的概率质量：

命题4.19。设 $C, K > 0$ 为常数。设 $\pi = \pi(p, q)$ 其中 p, q 为部分分布，且 p 相对于 q不是 C-高度集中。则由 $\boldsymbol{H}$ 诱导的随机分桶 $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_b)$满足

$$\mathbb{P}\left[\max_j p[\boldsymbol{\Gamma}_j] \ge 2(K+1)\frac{C \log^3 n}{m}\right] < 2/n^K\,.$$

证明。固定任意 $i \in \mathbb{Z}_n$，令 $\boldsymbol{\Gamma}_j$ 为包含 i的（随机）桶。我们希望界定 $p[\boldsymbol{\Gamma}_j] > 2(K+1)C\log_m^3 n$ 的概率。令 $R = \langle\!\langle i, d_R \rangle\!\rangle$ 为从 i开始、沿顺时针方向满足 $p[R] \ge (K+1)C\log_m^3 n$ 的最小圆周区间。同理，令 $L = \langle\!\langle i, -d_L \rangle\!\rangle$为从 i 开始、沿逆时针方向满足 $p[L] \ge (K+1)C\log_m^3 n$ 的最小圆周区间。注意，若 $p[\boldsymbol{\Gamma}_j] \ge 2(K+1)C\log_m^3 n$，则该桶至少包含这两个区间之一：$\mathcal{E}(L) \subseteq \mathcal{E}(\boldsymbol{\Gamma}_j)$ 或 $\mathcal{E}(R) \subseteq \mathcal{E}(\boldsymbol{\Gamma}_j)$，因此 $\boldsymbol{J}(L) = 1$ 或 $\boldsymbol{J}(R) = 1$。

由于 p相对于 q并非 C-高度集中，我们有 $(K+1)C\text{对数}_m^3 n \le p[R] < C\text{对数}^2(n) \cdot \text{最大值}\{q[R^*], m^{-1\text{对数}n}\}$。因此，必然有

$$q[R^*] > \frac{p[R]}{C\log^2 n} \ge (K+1)\frac{\log n}{m}\,.$$

Now

$$\mathbb{P}\left[\boldsymbol{J}(R) = 1\right] = \mathbb{P}\left[\forall z \in R^* : z \in \boldsymbol{H}\right] = \prod_{z \in R^*}(1 - w(z)) = \prod_{z \in R^*} e^{-m q_z} = e^{-m q[R^*]} < e^{-(K+1)\log n} = \frac{1}{n^{K+1}}\,.$$

对 $\mathbb{P}\left[\boldsymbol{J}(L) = 1\right]$，结论同样成立。然后，对 $i \in \mathbb{Z}_n$ 应用并集界，

$$\mathbb{P}\left[\max_j p[\boldsymbol{\Gamma}_j] \ge 2(K+1)\frac{C \log^3 n}{m}\right] < 2/n^K\,. \qquad \square$$

引理 4.20 (方差的第二个分量)。 设 $C > 0$ 为一个常数。设 n 足够大，并假设 m 满足 $m \le$ 多项式(n)。假设 $\pi = \pi(p, q)$，其中 p, q 是部分分布，且相对于 q， p 并非 C-高度集中。那么我们有

$$\mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\mathop{\mathrm{Var}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right] \le O\left(\frac{\log^6 n}{m}\right).$$

证明。 我们从以下一般结果出发：引理 2.12：对于某个绝对常数 $c > 0$，对于每个 H，只要它属于 $\boldsymbol{H}$ 的支撑集，且诱导出的桶为 $\Gamma = (\Gamma_1, \ldots, \Gamma_b)$，

$$\mathop{\mathrm{Var}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H} = H] \le c\|p_{|\Gamma}\|_2^2 + cm\|p_{|\Gamma}\|_3^3.$$

令 $K > 0$ 为一个常数，使得对于所有充分大的 n，都有 $m \le n^{K/2}$；根据假设 $m \le$ poly(n)，这样的常数存在。首先，假设子图 H 诱导出的分桶为 $\Gamma = (\Gamma_1, \ldots, \Gamma_b)$，且满足 $\max_j p[\Gamma_j] \le 2(K+1)C\frac{\log^3 n}{m}$。由于 $\|p_{|\Gamma}\|_1 = \|p\|_1 \le 1$，我们可以通过将总权重 1 以尽可能集中的方式分配，来给出 $\|p_{|\Gamma}\|_2^2$ 和 $\|p_{|\Gamma}\|_3^3$ 可取值的上界，也就是说，满足我们刚才假设的每个桶的上界。因此，我们得到

$$\|p_{|\Gamma}\|_2^2 \le \left(2(K+1)C\frac{\log^3 n}{m}\right)^2 \cdot \frac{1}{\left(2(K+1)C\frac{\log^3 n}{m}\right)} = O\left(\frac{\log^3 n}{m}\right)$$

and

$$\|p_{|\Gamma}\|_3^3 \le \left(2(K+1)C\frac{\log^3 n}{m}\right)^3 \cdot \frac{1}{\left(2(K+1)C\frac{\log^3 n}{m}\right)} = O\left(\frac{\log^6 n}{m^2}\right).$$

因此，在这种情况下，我们有

$$\mathop{\mathrm{Var}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H} = H] \le O\left(\frac{\log^6 n}{m}\right).$$

另一方面，由于 $\|p_{|\Gamma}\|_1 \le 1$，每个 H 都满足更简单的界

$$\mathop{\mathrm{Var}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H} = H] \le c\|p_{|\Gamma}\|_2^2 + cm\|p_{|\Gamma}\|_3^3 = O(m).$$

利用命题 4.19，我们写作

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\mathop{\mathrm{Var}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}]\right] \\
&\quad\le \mathbb{P}\left[\max_j p[\boldsymbol{\Gamma}_j] \le \frac{2(K+1)C\log^3 n}{m}\right] \cdot \mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\mathop{\mathrm{Var}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}] \;\middle|\; \max_j p[\boldsymbol{\Gamma}_j] \le \frac{2(K+1)C\log^3 n}{m}\right] \\
&\qquad+ \mathbb{P}\left[\max_j p[\boldsymbol{\Gamma}_j] > \frac{2(K+1)C\log n}{m}\right] \cdot \mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\mathop{\mathrm{Var}}_{\boldsymbol{T}}[\boldsymbol{Y} \mid \boldsymbol{H}] \;\middle|\; \max_j p[\boldsymbol{\Gamma}_j] > \frac{2(K+1)C\log^3 n}{m}\right] \\
&\quad\le O\left(\frac{\log^6 n}{m}\right) + O(m) \cdot \frac{2}{n^K},
\end{aligned}$$

并且由于 $m \le n^{K/2}$，我们有 $O(m) \cdot \frac{2}{n^K} \le O(1/n^{K/2}) \le O(1/m)$，如所需。 □

现在，我们可以利用全方差公式将这些结果结合起来，得到检验统计量的集中界：

引理 4.21 (检验统计量的集中性)。 设 $C > 0$ 为一个常数，且 $n \in \mathbb{N}$ 足够大。假设 $m = m(n, \epsilon)$ 满足 $m \leq$ 多项式(n)。令 $\pi = \pi(p, q)$，其中 p, q 是满足 $\|p\|_1, \|q\|_1 \geq 1/4$ 的部分分布，且 p 相对于q并非 C-高度集中。那么对于所有 $t > 0$，

$$\mathbb{P}\left[|\boldsymbol{Y} - \mathbb{E}[Y]| \geq t\right] \leq \frac{\frac{1}{m} + p^\top \phi p}{t^2} \cdot O(\log^6 n).$$

此外，假设 q 相对于 p 并非 C高度集中，且 m 满足 $m = \omega(\log n)$，$m = o(n)$。那么，写作 $p = \widetilde{p} + z$，其中 $\widetilde{p}$ 是 q的近似均匀共轭分布，我们还有

$$\mathbb{P}\left[|\boldsymbol{Y} - \mathbb{E}[Y]| \geq t\right] \leq \frac{\frac{1}{m} + z^\top \phi z}{t^2} \cdot O(\log^7 n).$$

证明。 根据全方差公式，

$$\mathrm{Var}\left[\boldsymbol{Y}\right] = \mathop{\mathrm{Var}}_{\boldsymbol{H}}\left[\mathop{\mathbb{E}}_{\boldsymbol{T}}\left[\boldsymbol{Y} \mid \boldsymbol{H}\right]\right] + \mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\mathop{\mathrm{Var}}_{\boldsymbol{T}}\left[\boldsymbol{Y} \mid \boldsymbol{H}\right]\right].$$

第一项的界为

$$\mathop{\mathrm{Var}}_{\boldsymbol{H}}\left[\mathop{\mathbb{E}}_{\boldsymbol{T}}\left[\boldsymbol{Y} \mid \boldsymbol{H}\right]\right] \leq O\left(\log^6 n\right) \cdot p^\top \phi p + O(m^2 e^{-m/8})$$

根据 引理 4.16，第二项的界为

$$\mathop{\mathbb{E}}_{\boldsymbol{H}}\left[\mathop{\mathrm{Var}}_{\boldsymbol{T}}\left[\boldsymbol{Y} \mid \boldsymbol{H}\right]\right] \leq O\left(\frac{\log^6 n}{m}\right)$$

根据 引理 4.20。此外，对于任意常数 $c > 0$，函数 $m \mapsto m^3 e^{-cm}$ 具有全局最大值 $\frac{3^3}{e^3 c^3}$，因此 $m^2 e^{-m/8} = O(1/m)$。第一个结论由切比雪夫不等式得出。

同时作出第二组假设时，推论 4.18 意味着

$$\mathop{\mathrm{Var}}_{\boldsymbol{H}}\left[\mathop{\mathbb{E}}_{\boldsymbol{T}}\left[\boldsymbol{Y} \mid \boldsymbol{H}\right]\right] \leq O\left(\frac{1}{m} + z^\top \phi z\right) \cdot O(\log^7 n),$$

因此，第二个结论再次由切比雪夫不等式推出。 □

4.5 ϵ较大时测试器的正确性

我们可以利用上述分离和集中性结果，证明 $\boldsymbol{Y}$ 集中在测试器阈值的正确一侧。再结合有偏分布和高度集中的分布这两种简单情形，即可得到正确性结论。

引理 4.22. 设 α、$\gamma > 0$ 为常数。存在常数$\beta = \beta_{\alpha,\gamma} > 0$ 和$K = K_{\alpha,\beta,\gamma} > 1$，使得对所有充分大的 n，以下结论成立。假设 $\epsilon \geq {}^{K}\log_{n^{1/4}}{}^{3} n$。设 $\pi = \pi(p、q)$，其中p、q 为部分分布，满足 $\|p\|_1$、$\|q\|_1 = \frac{1}{2} \pm \frac{2\gamma}{\sqrt{m}}$，并假设 p、q 相对于对方都不是 4α-高度集中的。

令 $T := \frac{m}{4n^2} \sum_{i,\ j} \phi_i^{(\mu)}{}_{,\ j} \pm \beta \frac{\epsilon^2 m^2}{n^2}$ 为算法 算法 2 使用的阈值。令 $\boldsymbol{Y}(0)$ 和 $\boldsymbol{Y}(1)$ 分别表示该算法第 $b = 0$ 次和第 $b = 1$ 次迭代中检验统计量 $\boldsymbol{Y}$的值的随机变量。那么，当 $m = \Theta_{\alpha,\ \beta,\ \gamma}\left(\binom{n}{\epsilon}^{4/5} \log^{7/5} n\right)$ 时，以下结论成立：

1.（完备性）如果 $\pi(p,q)$ 是 $[2n]$ 上的均匀分布，那么 $\max\{\boldsymbol{Y}_{(0)}\boldsymbol{Y}_{(1)}\} < T$ 的概率至少为 $99/100$；并且

2.（可靠性）如果 $\mathsf{dist}_{\mathsf{TV}}(\pi(p,q)\ \pi(\mu,\mu)) > \epsilon$，那么 $\max\{\boldsymbol{Y}_{(0)}\boldsymbol{Y}_{(1)}\} > T$的概率至少为$99/100$。

证明。 注意，根据 命题 4.7，我们可以直接写成 $T = \mathbb{E}\left[\boldsymbol{Y}_{(\mu)}\right] + \beta\frac{\epsilon^2 m^2}{n^2}$。

完备性。 在这种情况下， $\mathbb{E}\left[\boldsymbol{Y}_{(0)}\right] = \mathbb{E}\left[\boldsymbol{Y}^{(1)}\right] = \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right]$。此外，在这种情况下，由于根据 引理 4.11， $p = \widetilde{p} + z$ 对于 $\widetilde{p} = \mu$ 和 $z = \vec{0}$，因为 μ 是其自身的均匀共轭分布，因此可以写成 于是，引理 4.21 给出

$$\mathbb{P}\left[\boldsymbol{Y}^{(1)} \geq T\right] \leq \mathbb{P}\left[\left|\boldsymbol{Y}^{(1)} - \mathbb{E}\left[\boldsymbol{Y}^{(1)}\right]\right| \geq \beta\frac{\epsilon^2 m^2}{n^2}\right] \leq \frac{\frac{1}{m} + z^\top\phi z}{\left(\beta\frac{\epsilon^2 m^2}{n^2}\right)^2} \cdot O(\log^7 n) = \frac{n^4}{\beta^2\epsilon^4 m^5} \cdot O(\log^7 n)\,.$$

因此，对于任意常数 β （将在下文选定），当 $m = \Omega((n/\epsilon)^{4/5}\log^{7/5} n)$ 时，该概率至多为（例如） $1/200$ 。由对称性，对于 $\boldsymbol{Y}_{(0)}$ 也同样成立，因此 $\max\{\boldsymbol{Y}_{(0)}\boldsymbol{Y}_{(1)}\} < T$ 不成立的概率至多为 $1/100$，正如所需。

可靠性。 不失一般性，只需考虑 $\|q-\mu\|_1 > \epsilon$ 的情形，并证明以至少 $99/100$ 的概率有 $\boldsymbol{Y}^{(1)} > T$。

由于 $\epsilon \geq \frac{K}{\log^{3} n}$ 且 $m = \Theta_{\alpha,\beta,\gamma}((n/\epsilon)^{4/5}\log^{7/5} n)$，对于 β 的任意取值，只要令 $K = K_{\alpha,\beta,\gamma}$ 足够大，就能确保 $n/m^{\log n}$ 小于任意常数。事实上，对于某个常数$A = A_{\alpha,\beta,\gamma} > 0$，且当 n 足够大时，我们有

$$m \leq A\left(\frac{n}{\epsilon}\right)^{4/5}\log^{7/5} n \leq \frac{An\log^{7/5} n}{K^{4/5}\log^{12/5} n} = \frac{A}{K^{4/5}} \cdot \frac{n}{\log n}\,,$$

通过使 K 充分大，可以使其足够小。因此，引理 4.15 的条件得到满足，从而我们得到

$$\mathbb{E}\left[\boldsymbol{Y}^{(1)}\right] \geq \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] + \Omega\left(\frac{\epsilon^2 m^2}{n^2}\right) + mz^\top\phi z\,.$$

具体而言，令 $L = L_{\alpha,\ \gamma} > 0$ 为一个常数，使得当 n充分大时，我们有

$$\mathbb{E}\left[\boldsymbol{Y}^{(1)}\right] \geq \mathbb{E}\left[\boldsymbol{Y}^{(\mu)}\right] + L\frac{\epsilon^2 m^2}{n^2} + mz^\top\phi z\,.$$

那么只要 $\beta < L/2$，引理 4.21 便可推出

$$\begin{aligned}
\mathbb{P}\left[\boldsymbol{Y}^{(1)} \leq T\right] &\leq \mathbb{P}\left[\boldsymbol{Y}^{(1)} - \mathbb{E}\left[\boldsymbol{Y}^{(1)}\right] \leq \beta\frac{\epsilon^2 m^2}{n^2} - L\frac{\epsilon^2 m^2}{n^2} - mz^\top\phi z\right] \\
&\leq \mathbb{P}\left[\left|\boldsymbol{Y}^{(1)} - \mathbb{E}\left[\boldsymbol{Y}^{(1)}\right]\right| \geq (L-\beta)\frac{\epsilon^2 m^2}{n^2} + mz^\top\phi z\right] \\
&\leq \mathbb{P}\left[\left|\boldsymbol{Y}^{(1)} - \mathbb{E}\left[\boldsymbol{Y}^{(1)}\right]\right| \geq \frac{L}{2}\cdot\frac{\epsilon^2 m^2}{n^2} + mz^\top\phi z\right] \\
&\leq \frac{\frac{1}{m} + z^\top\phi z}{\left(\frac{L}{2}\cdot\frac{\epsilon^2 m^2}{n^2} + mz^\top\phi z\right)^2} \cdot O(\log^7 n) \\
&\leq \frac{\frac{1}{m} + z^\top\phi z}{\frac{L^2}{4}\cdot\frac{\epsilon^4 m^4}{n^4} + m^2\left(z^\top\phi z\right)^2} \cdot O(\log^7 n)\,,
\end{aligned}$$

最后一步成立是因为 $z^\top \phi z \geq 0$，这归因于 ϕ 的半正定性。现在我们考虑两种情况。首先，假设 $z^\top \phi z \leq 1/m$。那么

$$\mathbb{P}\left[\boldsymbol{Y}^{(1)} \leq T\right] \leq \frac{2/m}{\frac{L^2}{4} \cdot \frac{\epsilon^4 m^4}{n^4}} \cdot O(\log^7 n) = \frac{n^4}{L^2 \epsilon^4 m^5} \cdot O(\log^7 n)\,,$$

这再次至多为 $1/200$。另一方面，假设 $z^\top \phi z \geq 1/m$。那么

$$\mathbb{P}\left[\boldsymbol{Y}^{(1)} \leq T\right] \leq \frac{1/m}{\frac{L^2}{4} \cdot \frac{\epsilon^4 m^4}{n^4}} \cdot O(\log^7 n) + \frac{z^\top \phi z}{m^2 (z^\top \phi z)^2} \cdot O(\log^7 n) \leq \frac{n^4}{L^2 \epsilon^4 m^5} \cdot O(\log^7 n) + \frac{1}{m} \cdot O(\log^7 n)\,.$$

我们已经证明第一项至多为 $1/200$，而第二项显然为 $o(1)$。因此，$\mathbb{P}\left[\boldsymbol{Y}^{(1)} \leq T\right] \leq 1/100$，证明得证。 □

现在，我们可以结合前面的结果，得出测试器的正确性：

定理 4.23。 存在常数 α、β、$\gamma > 0$ 和 $K > 1$，使得对所有充分大的 n，以下结论成立。假设 $\epsilon \geq {}^{K}_{\text{对数}}{}_{n^{1/4}}{}^{3}n$。令 $\pi = \pi$（p, q）~~，~~其中 p、q 是部分分布。

然后算法 2 采用常数 α，β，以及 γ，其样本复杂度为 $\Theta\left(\binom{n}{\epsilon}^{4/5} \log^{7/5} n\right)$，并满足以下条件：

1. *(完备性)* 若 $\pi(p, q)$ 是 $[2n]$ 上的均匀分布，则算法以至少 $9/10$ 的概率接受；并且

2. *(可靠性)* 若 $f\,\mathsf{dist}_{\mathsf{TV}}(\pi(p, q)\ \pi(\mu, \mu)) > \epsilon$，则算法以至少 $9/1$ 的概率拒绝 $0.$

证明。 我们首先依次取足够大的 $\alpha, \gamma > 0$、足够小的 $\beta > 0$ 以及足够大的$K > 1$，以满足命题 4.3 和 4.6 以及 引理 4.22的条件。样本复杂度由算法的定义可得；现在我们来证明该算法能够正确地接受或拒绝。

完备性。 根据命题 4.3和 4.6以及 引理 4.22，算法以至多 $1/100 + 1/100 + 1/100 < 1/10$的概率拒绝。

可靠性。 我们分三种情况讨论。首先，假设$\|p\|_1 \notin \frac{1}{2} \pm \frac{2\gamma}{\sqrt{m}}$。由~~命题 4.3~~可知，算法以至少 $99/100$ 的概率拒绝。

其次，假设 p 或 q 相对于另一个 4α-高度集中。于是根据 命题 4.6，算法以至少 $99/100$的概率拒绝。

最后，假设 $\|p\|_1$，$\|q\|_1 = \frac{1}{2} \pm \frac{2\gamma}{\sqrt{m}}$~~且~~ p 和 q 都没有 4α-高度集中。于是根据引理 4.22，算法以至少 $99/100$的概率拒绝，符合要求。 □

结合 定 理 4.23 和引理 4.2给出了定理 4.1的上界部分 .

5 奇偶迹模型中的均匀性性质测试下界

记号 在本节中，令 μ 表示定义域 $[2n]$ 上的部分分布，其总质量为$1/2$，并在其支撑集上均匀分布，因此 $\pi(\mu, \mu)$ 是 $[2n]$上的均匀分布。

我们希望证明以下结果：

定理 5.1 （定理 4.1的下界部分。 设 Π_1 仅包含 $[2n]$ 上的均匀分布，并设 Π_2 是 $[2n]$ 上的分布集合，且这些分布在全变差距离下与均匀分布相差 ϵ。那么 $(\Pi_1, \Pi_2, 51/100)$-性质测试在奇偶迹下所需的样本复杂度至少为 $\widetilde{\Omega}\left(\binom{n}{\epsilon}^{4/5} + \frac{\sqrt{n}}{\epsilon^2}\right)$，其中 $\widetilde{\Omega}$ 记号中仅省略了关于 n 的多对数因子。此外，即使保证输入分布 π 的质量均匀分布在取值为零（即偶数）的坐标上，此下界仍然成立。

我们将分析分为两部分：一是从标准均匀性检验模型进行归约，从而得到 $\Omega(\sqrt{n}/\epsilon^2)$ 下界；二是更精妙的论证，适用于$\epsilon \geq n^{-1/4}$的情形。幸运的是，这恰好是 $(n/\epsilon)^{4/5} \geq \sqrt{n}/\epsilon^2$ 所处的范围。首先，给出较简单的下界：

命题 5.2。 设 Π_1 仅包含 $[2n]$ 上的均匀分布，并设 Π_2 为 $[2n]$ 上与均匀分布的全变差距离至少为 ϵ 的分布集合。那么，在奇偶迹下进行 $(\Pi_1, \Pi_2, 51/100)$- 性质测试，所需的样本复杂度至少为 $\Omega(\sqrt{n}/\epsilon^2)$。此外，即使保证输入分布 π 在值为零（即偶数）的坐标上均匀分布的质量为 $1/2$，这一界仍然成立。

证明。我们从对 $[n]$ 上分布进行均匀性检验的问题归约而来；该问题的下界为 $\Omega(\sqrt{n}/\epsilon^2)$ [Pan08]。对于 π 上的输入分布，令 π' 为定义在 $[2n]$ 上的分布，定义为：对每个 $i \in [n]$，令 $\pi'(2i-1) = \pi(i)/2$；并令 $\pi(2i) = \frac{1}{2n}$，其中 $i \in [n]$，因此 π' 在偶数元素上均匀分布。注意，我们可以通过对 $\boldsymbol{x} \sim \pi$ 采样并以 $1/2$ 的概率取 $2\boldsymbol{x} - 1$，否则取 $[2n]$ 中一个均匀随机的偶数元素，来模拟从 π' 中抽取一个样本。于是有以下结论：

1. 若 π 在 $[n]$ 上均匀分布，则 π' 在 $[2n]$ 上均匀分布；并且
2. 若 π 是 ϵ与均匀分布相差至少 π'，那么 $\epsilon/2$ 与均匀分布相差至少（以全变差距离衡量 ）.

因此 均匀性测试器可以用参数 $\epsilon/2$模拟奇偶迹测试器。 □

现在我们给出主要技术论证，以证明情形 $\epsilon \geq n^{-1/4}$ 下的 $\widetilde{\Omega}\left((n/\epsilon)^{4/5}\right)$ 界。

5.1 论证概要

我们的方法受 [DK16], 启发，旨在构造 YES 输入和 NO 输入上的分布，使得当 $\boldsymbol{Z}$ 是一个指示 YES/NO 情形的随机变量，且 $\boldsymbol{T}$是算法的输入（从 YES 或 NO 分布中抽取的奇偶迹）时，互信息 $I(\boldsymbol{Z} : \boldsymbol{T})$ 很小，因此没有算法能以较高概率根据 $\boldsymbol{Z}$ 来预测 $\boldsymbol{T}$。具体而言，我们沿用 [DK16]的方法，并使用以下 Fano 不等式的简单推论：

事实 5.3 (Fano 不等式)。 假设 $\boldsymbol{Z}$ 是一个均匀随机比特，$\boldsymbol{T}$是一个随机变量，并且存在一个函数 f，使得 $f(\boldsymbol{T}) = \boldsymbol{Z}$ 的概率至少为 51%。那么 $I(\boldsymbol{Z} : \boldsymbol{T}) \geq 2 \cdot 10^{-4}$。

因此，我们的目标是构造“分布上的分布”（以下简称分布）$\mathcal{D}_0$ （是情形）和 $\mathcal{D}_1$ （否情形），它们支撑于分布 π，这些分布定义在 $[2n]$ 上，并满足以下条件：令 $m = m(n, \epsilon)$ 为测试器的样本复杂度，并假设处于泊松化设定中（这将在后文带来便利）。那么，我们希望满足以下条件：

1. $\mathcal{D}_0$ 支撑于单个元素 $\pi(\mu, \mu)$，即 $[2n]$ 上的均匀分布 ；
2. 支撑集中的每个 π $\mathcal{D}_1$ 都满足 $\mathrm{dist}_{\mathsf{TV}}(\pi, \pi(\mu, \mu)) \geq \Omega(\epsilon)$；并且

3. 令 $\boldsymbol{Z} \sim \mathrm{Ber}(1/2)$，且 $\boldsymbol{\pi} \sim \mathcal{D}_{\boldsymbol{Z}}$。令 $\boldsymbol{\mathcal{T}}$ 按如下方式分布：抽取 $\boldsymbol{S} \sim \boldsymbol{\pi}$，$\mathrm{Poi}(m)$，并令 $\boldsymbol{\mathcal{T}} =$ 迹$(\boldsymbol{S})$。那么当 $m = o\left(\binom{n}{\epsilon}^{4/5} \quad {}^{1}\!/_{\log^4 n}\right)$ 时，我们有 $\cancel{I(\boldsymbol{Z}:\boldsymbol{\mathcal{T}})} = o(1)$。

现在我们概述证明的主要要点，然后给出完整论证。为简单起见，我们假设 n 是偶数。

YES 和 NO 分布。 回顾一下，定义域 $[2n]$ 上的概率分布 $\pi = \pi(p, q)$ 由两个部分分布组成：p 定义在取值为 1 的元素（奇数索引）上，q 定义在取值为 0 的元素（偶数索引）上。我们将把定义域 $[2n]$分割为 $n/2$ 个连续的长度为 4 的区间，称为多米诺骨牌，使得第 i 个多米诺骨牌决定条目 $(p_j,q_j,p_{j+1},q_{j+1})$，其中 $j = 2i - 1$；并为迹贡献一个字符串（称为子迹），其分布为

$$1^{\boldsymbol{A}_j}0^{\boldsymbol{B}_j}1^{\boldsymbol{A}_{j+1}}0^{\boldsymbol{B}_{j+1}},$$

其中，对每个 $k \in [n]$，$\boldsymbol{A}_k \sim$ 服从 $\mathrm{Poi}(mp_k)$，$\boldsymbol{B}_k \sim$ 服从 $\mathrm{Poi}(mq_k)$，且二者相互独立。

我们始终令 $q = \mu$，也就是说，定义在取值为 0 的元素上的部分分布是均匀分布的，其总质量为 $1/2$。在 YES 分布中，$\mathcal{D}_0$，$p = \mu$ 也如此。在 NO 分布中，对于 $\mathcal{D}_1$，我们将以相同概率设置为 $(p_j, p_{j+1}) = \left(\frac{1+\epsilon}{2n}, \frac{1-\epsilon}{2n}\right)$ 或 $(p_j, p_{j+1}) = \left(\frac{1-\epsilon}{2n}, \frac{1+\epsilon}{2n}\right)$，并且对每个多米诺骨牌相互独立地进行选择。因此，每个多米诺骨牌都是“平衡的”，且不同多米诺骨牌产生的子迹在给定 $\boldsymbol{Z}$ 的条件下相互独立。此外，我们将证明，从一个多米诺骨牌中至多采样 2 个符号，不会泄露有关 $\boldsymbol{Z}$ 的信息；也就是说，只有 3 路或更多元素的碰撞才包含有用信息。

部分指纹。 由于每块多米诺骨牌中抽取的符号不超过 2 个时都不会提供信息，我们将研究从中抽取了较多符号的那些多米诺骨牌的分布性质——算法可能正是从这些骨牌中揭示出关于 $\boldsymbol{Z}$ 的信息。受标准分布检验理论的启发，我们将研究这些多米诺骨牌上的部分指纹，它本质上衡量了抽取到多少个可揭示信息的符号。

粗略地说，我们将证明，部分指纹出现的概率会随着可揭示信息的样本元素数量（即来自抽取了 3 个或更多符号的多米诺骨牌的样本元素数量）的增加而呈指数级下降，这就为算法能从这些元素中学到多少信息设定了上界。我们指出，[DKN15a]对结构化分布进行接近性测试时所用的下界论证，也采用了类似思路：构造一个构造组件，使其在 YES 和 NO 条件下至多两个样本的分布完全相同。

定义域划分与互信息链式法则。 根据上述观测结果，我们或许可以通过以下步骤结束论证：1）对 $\boldsymbol{Z}$ 与来自每个多米诺骨牌的子迹之间的互信息给出上界；2）利用互信息链式法则，将这一数量对所有多米诺骨牌求和。（如果随机变量 $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_k$ 在给定 $\boldsymbol{Z}$ 的条件下相互独立，那么互信息链式法则意味着 $I(\boldsymbol{T}_1, \ldots, \boldsymbol{T}_k : \boldsymbol{Z}) \leq \sum_{i=1}^{k} I(\boldsymbol{T}_i : \boldsymbol{Z})$。）遗憾的是，这种策略无法得到良好的界；直观地说，它假定算法“知道”得太多了——具体而言，它知道自己看到的迹中所有多米诺骨牌的边界，而实际上，要预测这些边界应该非常困难。[13]

[13]另一个有意思的尝试是，依据迹中出现的值为 0 的符号的具体身份来对分析施加条件，然后考察每个由这些零值划定的值域内，值为 1 的符号的分布。由于类似的原因，这种方法似乎也行不通：根据生日悖论，从 $[2n]$ 中抽取 $n^{4/5}$ 个样本时，由值为 0 的符号划定的许多区间都会非常小，这也相当于“泄露”了过多信息。

事实证明，一种解决方法是考虑 $\Theta(m)$ 个连续区间，每个区间由 $\Theta(n/m)$ 个多米诺骨牌组成。由于我们总共采样了 $\mathrm{Poi}(m)$ 个符号，因此从每个这样的区间中采样到的符号数量的期望为 $\Theta(1)$，这使得分析变得可行；将这些区间各自对互信息的贡献相加，即可得到所需的界。

5.2 YES 和 NO 分布的构造

We 现在正式定义多米诺骨牌、子迹以及 YES 和 NO 分布。

定义 5.4（多米诺骨牌）。对于任意整数 $i \in [n/2]$，令 $j = 2i-1$ 和 $j' = 4i-3$。我们将定义域 $[2n]$的值域 $\{j', j'+1, j'+2, j'+3\}$，以及与这些位置相关联的 p和 q的概率质量（即 p_j,q_j,p_{j+1},q_{j+1}）称为第 i 个多米诺骨牌。

特别地，我们根据其 p 条目的概率质量，将多米诺骨牌分为三类（在 YES 和 NO 分布下，这些概率质量的取值方式不同）：

1. 无偏：当 $p_j = p_{j+1} = \frac{1}{2n}$。2. 左侧 ϵ-偏：当 $p_j = \frac{1}{2n}(1+\epsilon)$ 且 $p_{j+1} = \frac{1}{2n}(1-\epsilon)$。3. 右侧 ϵ- 偏：当 $p_j = \frac{1}{2n}(1-\epsilon)$ 且 $p_{j+1} = \frac{1}{2n}(1+\epsilon)$。

定义 5.5（子迹）。给定关于 $[2n]$ 的概率分布 $\pi(p,\ q)$，对于每个 $i \in [n/2]$，我们称由 i号多米诺骨牌产生的子迹为随机二进制字符串

$$\boldsymbol{t}_i := 1^{\boldsymbol{A}_j} 0^{\boldsymbol{B}_j} 1^{\boldsymbol{A}_{j+1}} 0^{\boldsymbol{B}_{j+1}},$$

其中 $j = 2i-1$ 和 $\boldsymbol{A}_k \sim$ 泊松分布(mp_k) $\boldsymbol{B}_k \sim$ 泊松分布(mq_k) 相互独立。

给定由 $\{i, i+1, \ldots, i+r-1\}$ 索引的 r 个多米诺骨牌构成的连续区间，该区间内这些多米诺骨牌产生的子迹为

$$\boldsymbol{T}_{i,r} := \boldsymbol{t}_i \circ \ldots \circ \boldsymbol{t}_{i+r-1},$$

其中 $\circ$ 表示拼接。

观察 5.6。回想一下，任意多米诺骨牌都有 $q_j = q_{j+1} = 1/2n$，也就是说，取值为 *0* 的元素上的部分分布是均匀分布的，总质量为 $1/2$。因此，每个多米诺骨牌都满足

$$p_j + q_j + p_{j+1} + q_{j+1} = 2/n,$$

因此，每个多米诺骨牌产生的子迹长度相互独立地服从泊松分布$(2m/n)$，而与 $\boldsymbol{Z}$ 的值无关。

根据上述定义，我们可以看到，完整轨迹 $\mathcal{T}$的分布为

$$\boldsymbol{\mathcal{T}} = \boldsymbol{t}_1 \circ \cdots \circ \boldsymbol{t}_{n/2}.$$

或者，如果我们将所有多米诺骨牌的集合分割为连续区间 $\{i_1, \ldots, i_1 + r_1 - 1\}, \ldots, \{i_k, \ldots, i_k + r_k - 1\}$，那么

$$\boldsymbol{\mathcal{T}} = \boldsymbol{T}_{i_1,r_1} \circ \cdots \circ \boldsymbol{T}_{i_k,r_k}.$$

现在我们定义 YES 和 NO 分布。

定义 5.7（YES 和 NO 分布）。令 $\epsilon > 0$。YES 分布 $\mathcal{D}_0$ 是一个支撑于单个元素 $\pi(\mu,\ \mu)$ 的分布，即在 $[2n]$ 上的均匀分布。NO 分布 $\mathcal{D}_1$ 是一个支撑于 $[2n]$ 上各分布的分布，按如下方式抽取：对于每个 $i \in [n/2]$，以相等的概率将第 i 个多米诺骨牌设为左偏 ϵ或右偏 ϵ，且每个多米诺骨牌的选择相互独立。

当我们将分布 $\pi(p,q)$ 视为从这些分布中抽取的随机变量时，相应地记作 $\boldsymbol{\pi} = \pi(\boldsymbol{p},\boldsymbol{q})$。

观察 5.8. 支撑集 $\mathcal{D}_1$ 中的每个 $\pi(p,q)$ 都满足 分布$_{\mathsf{TV}}(\pi(p,q)\ \pi(\mu,\mu)) = \epsilon/4$.

因此，我们要证明以下结果：

断言 5.9。设 $\boldsymbol{Z}$ ~伯努利分布$(1/2)$，并设 $\boldsymbol{\mathcal{T}}$为一个大小为 泊松分布(m) 的奇偶迹样本，从定义在 $[2n]$ 上、分布为$\boldsymbol{\pi}$的分布中采样得到，其中 $\boldsymbol{\pi} \sim \mathcal{D}_{\boldsymbol{Z}}$。那么，如果 $m = o\left(\binom{n}{\epsilon}^{4/5} \quad {}^{1}/_{\text{对数}^4\, n}\right)$，则有 $I(\boldsymbol{Z}{:}\boldsymbol{\mathcal{T}}) = o(1)$。

5.3 部分指纹及其概率

在分布测试的标准模型中，样本的指纹完整描述了测试离散分布的对称性质所需的相关信息 [Bat01]。指纹就是“直方图的直方图”：对于每个正整数 k，记录在样本中恰好出现 k 次的元素数量。

在我们的构造中，我们希望按照如下方式分析多米诺骨牌上的指纹。对于每个 i，令 d_i 表示由第 i 个多米诺骨牌产生的迹符号数量。于是， d 就是我们的直方图，而相应的指纹对于每个正整数 k统计：来自满足 $d_i = k$ 的多米诺骨牌的迹符号有多少个。

研究多米诺骨牌上的指纹很有用，因为正如我们的分析将表明的，当从每个多米诺骨牌中最多采样 2 个符号时，它不会提供关于 $\boldsymbol{Z}$ 的信息；而当采样 3 个或更多符号时，所揭示的信息量会随着符号数量的增加而增长。这一现象表明，我们应考虑一个部分指纹：它通过将对应于所有整数 $k \geq 3$ 的计数合并为一个类别“3^{+}”，从而由指纹得到。下面给出一种等价表述，它更便于我们的分析：

定义 5.10（$(h,\ k,\ s)$-碰撞）。考虑将 b 个相同小球投入 r 个箱子的过程，每个小球都相互独立地以均匀分布的随机方式选择一个箱子。向量 $d = (d_1,\ \ldots,\ d_r) \in \mathbb{Z}^r_{\geq 0}$，满足对于每个 $i \in [r]$，d_i 表示第 i 个箱子中的小球数量，称为该过程的一个结果。

我们称结果 d 是一个 $(h,\ k,\ s)$-碰撞，如果在 r 个箱子中，恰好有 h 个箱子各包含两个球，恰好有 k 个箱子至少包含三个球，并且此外，这 k 个箱子中的球总数为 $k+s$（换句话说， s 是包含至少 3 个球的箱子中的“多余”球的数量）。我们将 $\mathcal{C}_{r,\ b}(h,\ k,\ s)$，即 $(h,\ k,\ s)$-碰撞结果的集合，定义为

$$\mathcal{C}_{r,b}(h,k,s) := \left\{ \begin{array}{l} (d_1,\ldots,d_r) \in \mathbb{Z}^r_{\geq 0} : \\ \qquad \sum_{i=1}^{r} d_i = b, \sum_{i=1}^{r} \mathbb{1}\left[d_i = 2\right] = h, \sum_{i=1}^{r} \mathbb{1}\left[d_i \geq 3\right] = k, \sum_{i=1}^{r} \mathbb{1}\left[d_i \geq 3\right] \cdot (d_i - 1) = s \end{array} \right\}.$$

Note that $\mathcal{C}_{r,b}(h,k,s) = \emptyset$ 当 $s < 2k$ 时成立， $b < 2h$ 或 $b < k+$ 时也类似 $s.$

注意，表示上述随机过程结果的随机向量 $\boldsymbol{d}$ 服从 $\boldsymbol{d} \sim$ 多项分布$(b,\ r,\ (1/r,\ \ldots,\ 1/r))$。我们将证明，对于适当范围的取值，$(h,\ k,\ s)$-碰撞在 h 和 s 中发生的概率呈指数级小：

命题 5.11。 设 $\boldsymbol{d} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_r) \sim$服从多项分布$(b, r, (1/r, \ldots, 1/r))$。那么对于每个 $h \geq 0$ 和 $k, s > 0$，都有

$$\mathbb{P}\left[\boldsymbol{d} \in C_{r,b}(h,k,s)\right] \leq \left(\frac{(bh)^2}{r}\right)^h \left(\frac{(bk)^{3/2}}{r}\right)^s,$$

其中 0^0 解释为1。

证明。 我们可以通过枚举哪些箱子会恰好装有两个球（如果存在），称为“二重碰撞”，以及哪些箱子会装有三个或更多球，称为“3^+-碰撞”，再枚举落入这些箱子的球，并对决定每种情况的组合数量取一个粗略上界，从而得到该概率的上界。记 $\binom{[n]}{m}$ 为 $[n]$ 中大小为 m的子集所组成的集合。我们有

$$\mathbb{P}\left[\boldsymbol{d} \in C_{r,b}(h,k,s)\right] \leq \sum_{I_2 \in \binom{[r]}{h}} \sum_{I_3 \in \binom{[r]\setminus I_2}{k}} \sum_{J_2 \in \binom{[b]}{2h}} \sum_{J_3 \in \binom{[b]\setminus J_2}{k+s}} \Big[\mathbb{P}\big[\text{个球在箱子中形成-碰撞 } J_2 \cup J_3\big] \cdot \mathbb{P}\big[\text{个球在箱子中形成-碰撞 } I_3\big] \cdot \mathbb{P}\big[\text{个球落入箱子 } I_2 \cup I_3 \text{ 且彼此不发生碰撞}\big]\Big]$$

$$\leq \sum_{I_2 \in \binom{[r]}{h}} \sum_{I_3 \in \binom{[r]\setminus I_2}{k}} \sum_{J_2 \in \binom{[b]}{2h}} \sum_{J_3 \in \binom{[b]\setminus J_2}{k+s}} \mathbb{P}\big[\text{个球落在箱子内 } I_2\big]\, \mathbb{P}\big[\text{个球落在箱子内 } I_3\big] \leq r^{h+k} b^{2h+k+s} h^{2h} k^{k+s} r^{-2h-k-s} \leq (bh)^{2h} r^{-h} \cdot (bk)^{\frac{3}{2}s} r^{-s} = \left(\frac{(bh)^2}{r}\right)^h \left(\frac{(bk)^{3/2}}{r}\right)^s, \qquad (\text{Since } k \leq s/2)$$

其中我们使用了 $k \leq s/2$ ，该结论成立，除非概率为零；在这种情况下，结论显然成立。

□

我们还需要以下简单的“生日问题”界：

命题 5.12. 设 $\boldsymbol{d} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_r) \sim$多项分布$(b, r, (1/r, \ldots, 1/r))$。那么，出现无碰撞情形的概率满足

$$\mathbb{P}\left[\boldsymbol{d} \in \mathcal{C}_{r,b}(0,0,0)\right] \geq 1 - \frac{b^2}{r}.$$

证明. 该概率为

$$\mathbb{P}\left[\boldsymbol{d} \in \mathcal{C}_{r,b}(0,0,0)\right] = \frac{r \cdot (r-1) \cdots (r-b+1)}{r^b} \geq \left(\frac{r-b}{r}\right)^b = \left(1 - \frac{b}{r}\right)^b \geq 1 - \frac{b^2}{r}.$$

□

5.4 YES 和 NO 多米诺骨牌的行为相似

现在我们证明，每个长度不太长的子迹，在 YES 分布和 NO 分布下由多米诺骨牌生成的概率必然相近。

我们首先需要以下简单的界，它以非正式的方式概括了这样一个性质：尽管多米诺骨牌的相对概率质量存在 $\Theta(\epsilon)$ 偏差，但它所揭示的“信息”会以 $O(\epsilon^2)$ 的速度衰减。

命题 5.13. 对于所有非负整数 x 和 y，以及所有 $0 < \epsilon < 1$ ，

$$\frac{1}{2}(1+\epsilon)^x(1-\epsilon)^y + \frac{1}{2}(1-\epsilon)^x(1+\epsilon)^y = 1 \pm \epsilon^2 \cdot 2^{x+y}\,.$$

证明。 根据二项式定理，我们有

$$\begin{aligned}
&\frac{1}{2}(1+\epsilon)^x(1-\epsilon)^y + \frac{1}{2}(1-\epsilon)^x(1+\epsilon)^y \\
&\quad= \frac{1}{2}\left(\sum_{i=0}^{x}\binom{x}{i}\epsilon^i\right)\left(\sum_{j=0}^{y}\binom{y}{j}(-1)^j\epsilon^j\right) + \frac{1}{2}\left(\sum_{i=0}^{x}\binom{x}{i}(-1)^i\epsilon^i\right)\left(\sum_{j=0}^{y}\binom{y}{j}\epsilon^j\right) \\
&\quad= \sum_{i=0}^{x}\sum_{j=0}^{y}\binom{x}{i}\binom{y}{j}\epsilon^{i+j}\left(\frac{(-1)^i+(-1)^j}{2}\right) \\
&\quad= 1 \pm \epsilon^2\left(\sum_{i=0}^{x}\binom{x}{i}\right)\left(\sum_{j=0}^{y}\binom{y}{j}\right) \\
&\quad= 1 \pm \epsilon^2 \cdot 2^{x+y}\,,
\end{aligned}$$

其中，第三个（不）等式成立，是因为当 $i = j = 0$ 时，整个内部表达式等于 1；当 $i + j = 1$ 时，它等于 0，因为 i 和 j 的奇偶性不同；其他情况下，我们有 $\epsilon^{i+j} \le \epsilon^2$。 □

引理 5.14。 设 $\epsilon \in (0,\ 1)$。令 $\boldsymbol{t} = \boldsymbol{t}_i$ 表示多米诺骨牌所产生的子迹对应的随机变量。那么对于任意二进制字符串 t，有以下两种情况：

1. 如果 t 至少包含一个*"0"*符号和两个*"1"*符号，那么

$$\mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{Z} = 1\right] = \Big(\mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{Z} = 0\right]\Big)\left(1 \pm \epsilon^2 \cdot 2^{|t|-1}\right)\,.$$

2. 否则，

$$\mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{Z} = 1\right] = \mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{Z} = 0\right]\,.$$

注意，概率是针对分布 $\boldsymbol{\pi} \sim \mathcal{D}_{\boldsymbol{Z}}$ 的选择以及与来自 $\boldsymbol{\pi}$的（泊松化）迹相对应的随机向量$\boldsymbol{A}$、$\boldsymbol{B}$ 而言的。

证明。 注意，如果 t 不属于正则语言 $1^*0^*1^*0^*$，那么上述所有概率都为零（因为多米诺骨牌无法产生这样的迹），因此结论成立。假设 t 的形式为 $1^*0^*1^*0^*$。

不失一般性，设 $\boldsymbol{t}$ 是第一个多米诺骨牌产生的迹，因此对于 $\boldsymbol{A}_1 \sim$， $\boldsymbol{t} = 1^{\boldsymbol{A}_1}0^{\boldsymbol{B}_1}1^{\boldsymbol{A}_2}0^{\boldsymbol{B}_2}$，且 $\boldsymbol{A}_1 \sim$ 服从 Poi($m\boldsymbol{p}_1$)， $\boldsymbol{A}_2 \sim$ 服从 Poi($m\boldsymbol{p}_2$)， $\boldsymbol{B}_1 \sim$ 服从 Poi($m\boldsymbol{q}_1$)， $\boldsymbol{B}_2 \sim$ 服从 Poi($m\boldsymbol{q}_2$)；其中 $\boldsymbol{p}$，$\boldsymbol{q}$ 是 $\boldsymbol{\pi}$ 的部分分布，因此 $\boldsymbol{q}_1 = \boldsymbol{q}_2 = 1/2n$。注意， $|\boldsymbol{t}| = \boldsymbol{A}_1 + \boldsymbol{B}_1 + \boldsymbol{A}_2 + \boldsymbol{B}_2$。

根据标准论证，在以 $|\boldsymbol{t}| = |t|$ 为条件后，随机变量 $\boldsymbol{A}_i$ 和 $\boldsymbol{B}_i$ 服从由 $|t|$ 次试验和 4 个箱子构成的多项分布，各箱子的概率对应于每个位置上概率质量的相对权重：

$$\begin{aligned}
&\mathbb{P}\left[\boldsymbol{A}_1 = a_1, \boldsymbol{B}_1 = b_1, \boldsymbol{A}_2 = a_2, \boldsymbol{B}_2 = b_2 \;\middle|\; \boldsymbol{A}_1 + \boldsymbol{B}_1 + \boldsymbol{A}_2 + \boldsymbol{B}_2 = |t|\right] \\
&\qquad= \mathbb{P}\left[(\boldsymbol{X}_1, \boldsymbol{Y}_1, \boldsymbol{X}_2, \boldsymbol{Y}_2) = (a_1, b_1, a_2, b_2)\right]\,, \\
&(\boldsymbol{X}_1, \boldsymbol{Y}_1, \boldsymbol{X}_2, \boldsymbol{Y}_2) \sim \mathsf{Multinomial}\left(|t|, \left(\frac{n}{2}\boldsymbol{p}_1, \frac{1}{4}, \frac{n}{2}\boldsymbol{p}_2, \frac{1}{4}\right)\right)\,.
\end{aligned}$$

为方便起见，令 $\boldsymbol{p}'_1 := \frac{n}{2}\boldsymbol{p}_1$ 和 $\boldsymbol{p}'_2 := \frac{n}{2}\boldsymbol{p}_2$。注意，在是情形（$\boldsymbol{Z} = 0$）中，我们有 $\boldsymbol{p}'_1 = \boldsymbol{p}'_2 = \frac{1}{4}$；而在否情形（$\boldsymbol{Z} = 1$）中，我们以相同的概率出现以下情形之一：

1. $\boldsymbol{p}'_1 = \frac{1}{4}(1+\epsilon)$ and $\boldsymbol{p}'_2 = \frac{1}{4}(1-\epsilon)$;

2. $\boldsymbol{p}'_1 = \frac{1}{4}(1-\epsilon)$ and $\boldsymbol{p}'_2 = \frac{1}{4}(1+\epsilon)$.

现在我们来证明这一结论。我们先讨论第二种情况，因为它比较简单。首先，假设 t 中不含“0”符号。那么，由于值为 1 的位置的总权重为 $\boldsymbol{p}'_1 + \boldsymbol{p}'_2 = 1/2$，与 $\boldsymbol{Z}$ 的取值无关，因此我们有

$$\mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{Z} = 1\right] = \mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{Z} = 0\right] = \left(\frac{1}{2}\right)^{|t|}.$$

另一方面，假设 t 中至多包含一个“1”符号。如果其中不包含“1”符号，则同样的逻辑仍然适用，因此我们可以假设 t 中恰好包含一个“1”符号。我们可以将 t 的概率写成所有能生成二进制字符串 t 的（a_1，b_1，a_2，b_2）的概率之和，即 $t = 1^{a_1}0^{b_1}1^{a_2}0^{b_2}$。令 $\mathcal{S}$ 表示所有能生成 t的此类元组的集合。利用多项式公式，对于 $\boldsymbol{p}'_1$，$\boldsymbol{p}'_2$的任意可能取值 p'_1，p'_2，可得：

$$\mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{p}'_1 = p'_1, \boldsymbol{p}'_2 = p'_2\right] = \sum_{(a_1,b_1,a_2,b_2)\in\mathcal{S}} \frac{|t|!}{a_1!b_1!a_2!b_2!} \left(p'_1\right)^{a_1} \left(\frac{1}{4}\right)^{b_1} \left(p'_2\right)^{a_2} \left(\frac{1}{4}\right)^{b_2}.$$

由于 t 恰好包含一个“1”符号——设该符号为 $t = 0^x10^y$，其中某个 x，$y \geq 0$——，我们有：当且仅当$1^{a_1}0^{b_1}1^{a_2}0^{b_2} = t$

1. $a_1 = 0$、 $a_2 = 1$、 $b_1 = x$和 $b_2 = y$；或者，互斥地，

2. $a_1 = 1$, $a_2 = 0$, $b_1 + b_2 = y$, and $x = 0$.

因此，我们可以将 $t = 0^x10^y$ 的概率写作

$$\mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{p}'_1 = p'_1, \boldsymbol{p}'_2 = p'_2\right] = \frac{(x+y+1)!}{x!y!}\left(\frac{1}{4}\right)^{x+y} p'_2 + \mathbb{1}\left[x = 0\right]\frac{(y+1)!}{y!}\left(\frac{1}{2}\right)^{y} p'_1.$$

我们验证，当 $\boldsymbol{Z} = 1$ 时， $\boldsymbol{p}'_1$ 和 $\boldsymbol{p}'_2$ 分别正偏和负偏的情况会相互抵消，因此得到与 $\boldsymbol{Z} = 0$ 时相同的概率：

$$\begin{aligned}
&\mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{Z} = 1\right] \\
&\quad= \frac{1}{2}\mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{p}'_1 = \frac{1}{4}(1-\epsilon), \boldsymbol{p}'_2 = \frac{1}{4}(1+\epsilon)\right] \\
&\qquad+ \frac{1}{2}\mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{p}'_1 = \frac{1}{4}(1+\epsilon), \boldsymbol{p}'_2 = \frac{1}{4}(1-\epsilon)\right] \\
&\quad= \frac{1}{2}\left[\frac{(x+y+1)!}{x!y!}\left(\frac{1}{4}\right)^{x+y}\frac{1}{4}(1+\epsilon) + \mathbb{1}\left[x = 0\right]\frac{(y+1)!}{y!}\left(\frac{1}{2}\right)^{y}\frac{1}{4}(1-\epsilon)\right] \\
&\qquad+ \frac{1}{2}\left[\frac{(x+y+1)!}{x!y!}\left(\frac{1}{4}\right)^{x+y}\frac{1}{4}(1-\epsilon) + \mathbb{1}\left[x = 0\right]\frac{(y+1)!}{y!}\left(\frac{1}{2}\right)^{y}\frac{1}{4}(1+\epsilon)\right] \\
&\quad= \frac{(x+y+1)!}{x!y!}\left(\frac{1}{4}\right)^{x+y}\left(\frac{1}{4}\right) + \mathbb{1}\left[x = 0\right]\frac{(y+1)!}{y!}\left(\frac{1}{2}\right)^{y}\left(\frac{1}{4}\right) \\
&\quad= \mathbb{P}\left[\boldsymbol{t} = t \;\middle|\; |\boldsymbol{t}| = |t|, \boldsymbol{Z} = 0\right],
\end{aligned}$$

由此完成该命题第二种情况的证明。

让我们回到第一种情况。假设 t 中至少包含一个“0”符号和两个“1”符号；设为 $t = 1^x0^z1^y0^w$、 $x+y \geq 2$ 和 $z+w \geq 1$。我们再次从一般的多项式公式开始：令 $\mathcal{S}$ 为满足 $t = 1^{a_1}0^{b_1}1^{a_2}0^{b_2}$的元组 $(a_1,\ b_1,\ a_2,\ b_2)$ 的集合。我们有

$$\mathbb{P}\left[\boldsymbol{t} = t \ \middle|\ |\boldsymbol{t}| = |t|, \boldsymbol{p}'_1 = p'_1, \boldsymbol{p}'_2 = p'_2\right] = \sum_{(a_1,b_1,a_2,b_2)\in\mathcal{S}} \frac{|t|!}{a_1!b_1!a_2!b_2!}\left(\frac{1}{4}\right)^{b_1}\left(\frac{1}{4}\right)^{b_2}\cdot\left(p'_1\right)^{a_1}\left(p'_2\right)^{a_2}.$$

定义 $F_{a_1,b_1,a_2,b_2} := \frac{|t|!}{a_1!\ b_1!\ a_2!\ b_2!}\left(\frac{1}{4}\right)^{a_1}\left(\frac{1}{4}\right)^{b_1}\left(\frac{1}{4}\right)^{a_2}\left(\frac{1}{4}\right)^{b_2}$，使得

$$\mathbb{P}\left[\boldsymbol{t} = t \ \middle|\ |\boldsymbol{t}| = |t|, \boldsymbol{Z} = 0\right] = \sum_{(a_1,b_1,a_2,b_2)\in\mathcal{S}} F_{a_1,b_1,a_2,b_2}$$

and

$$\begin{aligned}\mathbb{P}\left[\boldsymbol{t} = t \ \middle|\ |\boldsymbol{t}| = |t|, \boldsymbol{Z} = 1\right] = &\frac{1}{2}\sum_{(a_1,b_1,a_2,b_2)\in\mathcal{S}} \frac{|t|!}{a_1!b_1!a_2!b_2!}\left(\frac{1}{4}\right)^{b_1}\left(\frac{1}{4}\right)^{b_2}\cdot\left(\frac{1}{4}(1+\epsilon)\right)^{a_1}\left(\frac{1}{4}(1-\epsilon)\right)^{a_2}\\ &+\frac{1}{2}\sum_{(a_1,b_1,a_2,b_2)\in\mathcal{S}} \frac{|t|!}{a_1!b_1!a_2!b_2!}\left(\frac{1}{4}\right)^{b_1}\left(\frac{1}{4}\right)^{b_2}\cdot\left(\frac{1}{4}(1-\epsilon)\right)^{a_1}\left(\frac{1}{4}(1+\epsilon)\right)^{a_2}\\ &=\sum_{(a_1,b_1,a_2,b_2)\in\mathcal{S}} F_{a_1,b_1,a_2,b_2}\left(\frac{1}{2}(1+\epsilon)^{a_1}(1-\epsilon)^{a_2}+\frac{1}{2}(1-\epsilon)^{a_1}(1+\epsilon)^{a_2}\right).\end{aligned}$$

因此，只需证明对于每个 $(a_1, b_1, a_2, b_2) \in \mathcal{S}$，

$$\frac{1}{2}(1+\epsilon)^{a_1}(1-\epsilon)^{a_2}+\frac{1}{2}(1-\epsilon)^{a_1}(1+\epsilon)^{a_2} \stackrel{?}{=} 1 \pm \epsilon^2\cdot 2^{|t|-1},$$

并且由于 $a_1 + a_2 \leq |t| - 1$（因为 t 至少包含一个“0”符号），这由命题 5.13得出，证明完成。

□

5.5 信息界

回顾一下，我们希望对互信息 $I(\boldsymbol{Z} : \boldsymbol{T})$ 求上界。根据互信息链式法则，我们将对各个量 $I(\boldsymbol{Z} : \boldsymbol{T}_{i_j,\ r_j})$ 求和，其中每个$\{i_j, i_j + 1, \ldots, i_j + r_j - 1\}$ 都是多米诺骨牌的一个连续区间（构成一个分割），而 $\boldsymbol{T}_{i_j,\ r_j}$ 是该区间产生的子迹。为简便起见，令 $\boldsymbol{T} = \boldsymbol{T}_{i_j,\ r_j}$ 表示这样一个变量。令 P_0和 P_1 表示在 $\boldsymbol{Z}$的每个取值下 $\boldsymbol{T}$ 的条件分布：对于每个二进制字符串 T 和$z \in \{0, 1\}$， $P_z(T)$:
$= \mathbb{P}[\boldsymbol{T} = T \ \middle|\ \boldsymbol{Z} = z]$。

下面的事实表明，如果 P_1 与 P_0 之间的逐点比值接近 1，那么互信息 $I(\boldsymbol{Z} : \boldsymbol{T})$就很小。由于完整论证将需要一个经过细化的版本，以处理低概率结果（对于这些结果，比值上界可能不成立），因此这里不加证明地陈述这一事实，仅用于直观理解。该表述的灵感来自 [DK16]。

事实 5.15。 设 P_0 和 P_1 是定义域 $\mathcal{X}$ 上的离散概率分布。令 $\xi > 0$，并假设对于每个 $T \in \mathcal{X}$，都有

$$\frac{P_1(T)}{P_0(T)} = 1 \pm \xi.$$

那么我们有

$$\chi^2(P_1\|P_0) \le \xi^2,$$

其中 $\chi^2(P_1\|P_0) = \mathbb{E}_{\boldsymbol{T}\sim P_0}\left[\left(\frac{P_1}{P_0}(\boldsymbol{T})\right)^2 - 1\right]$ 是 *Pearson* χ^2-散度。此外，若 $\boldsymbol{Z}$ 是均匀随机比特，且 $\boldsymbol{T}$ 服从 $P_{\boldsymbol{Z}}$，则

$$I(\boldsymbol{Z}:\boldsymbol{T}) \le \frac{1}{2}\chi^2(P_1\|P_0) \le \frac{1}{2}\xi^2.$$

因此，我们的任务是给 $\left|\frac{P_1}{P_0}(T) - 1\right|$ 求上界。以下结果适用于任意值域的 r 个多米诺骨牌，以及长度相对于 r不过长的字符串 T。稍后我们会看到，过长的字符串 T 出现的概率很低，因此对互信息的影响很小。

引理 5.16. 存在一个普适常数 $c > 0$，使得以下结论成立。令 $\epsilon \in (0, 1)$，并令 $r \ge 1$ 为一个整数。令 $\boldsymbol{T}$ 表示由一段包含 r 个连续多米诺骨牌的范围所产生的子迹，并令 P_z 表示在上述条件下 $\boldsymbol{T}$ 的条件概率分布。那么，对于任意满足 $|T|^4 \le r/100$的二进制字符串 T**s**，有

$$\frac{P_1(T)}{P_0(T)} = 1 \pm c\cdot\epsilon^2\cdot\frac{|T|^6}{r^2}.$$

证明。记 $g_z(t)$ 为引理 5.14中给出的概率：给定 $\boldsymbol{Z} = z$ 以及子迹的长度，任意给定的多米诺骨牌产生子迹 t 的概率为：$g_z(t) := \mathbb{P}[\boldsymbol{t} = t \mid |\boldsymbol{t}| = |t|, \boldsymbol{Z} = z]$
其中，$\boldsymbol{t}$ 是与所考虑的多米诺骨牌产生的子迹相对应的随机变量。回想一下，根据多米诺骨牌的定义，概率 $g_z(t)$对每个多米诺骨牌都相同。

对于每个 $i \in [r]$，令 $\boldsymbol{D}_i$ 为值域中第 i个多米诺骨牌产生的子迹长度所对应的随机变量。正如观察 5.6中所述，泊松随机变量的可加性以及多米诺骨牌的构造意味着 $\boldsymbol{D}_i \sim$泊松分布$(2m/n)$，对所有 i均相互独立。为方便起见，令 $\lambda := 2m/n$。

要使 T 成为所考虑的值域产生的迹，该值域中的每个多米诺骨牌都必须产生一个子迹，并满足：1）所有子迹的总长度为 $|T|$；2）每个多米诺骨牌产生的子迹都等于 T中相应的子字符串。为此，令 $\mathcal{M}$表示所有分量之和为 $|T|$的子迹长度向量集合：

$$\mathcal{M} := \left\{ d \in \mathbb{Z}^r_{\ge 0} : \sum_{i=1}^{r} d_i = |T| \right\}.$$

回顾定义 5.10，我们可以将 $\mathcal{M}$ 写成不相交并

$$\mathcal{M} = \biguplus_{h,k,s\ge 0} \mathcal{C}_{r,|T|}(h,k,s).$$

我们将使用以下记号来表示 T 的子串。对于索引 $1 \le a, b \le |T|$，令 $T[a..b]$ 表示 T 中索引 a 到 b（含端点）之间的子字符串，其中 $a \le b$；当 $a > b$ 时，令 $T[a..b] := \varnothing$。对于直方图 $d \in \mathcal{M}$，以及每个 $i \in [r]$，令

$$T(d,i) := T\left[\left(1 + \sum_{j=1}^{i-1} d_j\right) .. \left(1 + \sum_{j=1}^{i-1} d_j\right) + d_i - 1\right],$$

那么，对于所有 $d \in \mathcal{M}$，T 等于拼接 $T(d,1) \circ \cdots \circ T(d,r)$。

现在，对于每个 $z \in \{0,1\}$，我们都有

$$P_z(T) = \sum_{d\in\mathcal{M}} \prod_{i=1}^{r} \mathbb{P}\left[\boldsymbol{D}_i = d_i\right] g_z(T(d,i)) = \sum_{d\in\mathcal{M}} \prod_{i=1}^{r} \frac{e^{-\lambda}\lambda^{d_i}}{d_i!} g_z(T(d,i))$$
$$= \frac{e^{-r\lambda}(r\lambda)^{|T|}}{|T|!} \sum_{d\in\mathcal{M}} \left[\left(\frac{|T|!}{d_1!\cdots d_r!} \left(\frac{1}{r}\right)^{|T|} \right) \left(\prod_{i=1}^{r} g_z(T(d,i)) \right) \right] .$$

注意到求和式中的第一个因子是一个多项概率：令 $\boldsymbol{d} = (\boldsymbol{d}_1, \ldots\ldots, \boldsymbol{d}_r) \sim$多项分布$(|T|, r, (1/r, \ldots\ldots, 1/r))$，第一个因子恰好是 $\mathbb{P}[\boldsymbol{d} = d]$。这就是 定义 5.10 中引入的“球与箱子”过程。[14]

至于第二个因子，对于每个 d 和 i，定义

$$\delta(d,i) := \frac{g_1(T(d,i))}{g_0(T(d,i))} - 1 ,$$

从而 $g_1(T(d,i)) = (1+\delta(d,i)) g_0(T(d,i))$。为简单起见，通过略微放宽 引理 5.14，我们可以将每个 $\delta(d,i)$ 上界为：

$$d_i \le 2 \implies \delta(d,i) = 0 , \text{ and}$$
$$d_i \ge 3 \implies |\delta(d,i)| \le \epsilon^2 \cdot 2^{d_i - 1} .$$

于是得到

$$P_z(T) = \frac{e^{-r\lambda}(r\lambda)^{|T|}}{|T|!} \sum_{d\in\mathcal{M}} \mathbb{P}\left[\boldsymbol{d} = d\right] \prod_{i=1}^{r} g_0(T(d,i))(1 + z\delta(d,i))$$
$$= \frac{e^{-r\lambda}(r\lambda)^{|T|}}{|T|!} \sum_{h,k,s\ge 0} \sum_{d\in\mathcal{C}_{r,|T|}(h,k,s)} \mathbb{P}\left[\boldsymbol{d} = d\right] \left(\prod_{i=1}^{r} g_0(T(d,i)) \right) \left(\prod_{i=1}^{r} \left(1 \pm z\mathbb{1}\left[d_i \ge 3\right] \epsilon^2 2^{d_i - 1} \right) \right) .$$

对于任意 $d \in \mathcal{C}_{r,|T|}(h,k,s)$，项 $\prod_{i=1}^{r} \left(1 \pm z\mathbb{1}\left[d_i \ge 3\right] \epsilon^2 2^{d_i-1}\right)$ 是一个乘积，其中除 k 项之外的所有项都只是 1，因为根据 (h,k,s)-碰撞的定义，只有 k 个条目 d_i 可能满足 $d_i \ge 3$。因此，展开该乘积后，我们得到 2^k 项；其中一项为 1，另有 $2^k - 1$ 项分别至少包含一个 ϵ^2 因子和一个 2^x 因子，对某个 $x \le \sum_{i=1}^{r} \mathbb{1}\left[d_i \ge 3\right](d_i - 1)$ 而言成立。于是，利用 (h,k,s)-碰撞定义中的恒等式 $\sum_{i=1}^{r} \mathbb{1}\left[d_i \ge 3\right](d_i - 1) = s$，我们得到

$$\prod_{i=1}^{r} \left(1 \pm z\mathbb{1}\left[d_i \ge 3\right] \epsilon^2 2^{d_i - 1} \right) = 1 \pm z\epsilon^2 (2^k - 1) 2^s .$$

因此，我们可以写成

$$P_z(T) = \frac{e^{-r\lambda}(r\lambda)^{|T|}}{|T|!} \sum_{h,k,s\ge 0} \left(1 \pm z\epsilon^2 (2^k - 1) 2^s \right) \sum_{d\in\mathcal{C}_{r,|T|}(h,k,s)} \mathbb{P}\left[\boldsymbol{d} = d\right] \left(\prod_{i=1}^{r} g_0(T(d,i)) \right) .$$

[14] 对于这部分分析，我们实际上已经“剥离”了泊松化。

回想一下，我们要证明 $P_1(T)/P_0(T) = 1 \pm \epsilon^2 \cdot O\left(|T|^6/r^2\right)$。代入上述表述后，我们得到

$$\begin{aligned}\frac{P_1(T)}{P_0(T)} &= \frac{\sum_{h,k,s\geq 0}\left(1\pm\epsilon^2(2^k-1)2^s\right)\sum_{d\in\mathcal{C}_{r,|T|}(h,k,s)}\mathbb{P}\left[\boldsymbol{d}=d\right]\left(\prod_{i=1}^r g_0(T(d,i))\right)}{\sum_{h,k,s\geq 0}\sum_{d\in\mathcal{C}_{r,|T|}(h,k,s)}\mathbb{P}\left[\boldsymbol{d}=d\right]\left(\prod_{i=1}^r g_0(T(d,i))\right)} \\ &= 1\pm\sum_{h,k,s\geq 0}\left[\epsilon^2(2^k-1)2^s\cdot\frac{\sum_{d\in\mathcal{C}_{r,|T|}(h,k,s)}\mathbb{P}\left[\boldsymbol{d}=d\right]\left(\prod_{i=1}^r g_0(T(d,i))\right)}{\sum_{d\in\mathcal{M}}\mathbb{P}\left[\boldsymbol{d}=d\right]\left(\prod_{i=1}^r g_0(T(d,i))\right)}\right].\end{aligned}$$

因此，我们的目标是证明以下结论：

$$\sum_{h\geq 0,s\geq 2k\geq 2}\epsilon^2(2^k-1)2^s\cdot\frac{\sum_{d\in\mathcal{C}_{r,|T|}(h,k,s)}\mathbb{P}\left[\boldsymbol{d}=d\right]\prod_{i=1}^r g_0(T(d,i))}{\sum_{d\in\mathcal{M}}\mathbb{P}\left[\boldsymbol{d}=d\right]\prod_{i=1}^r g_0(T(d,i))}\overset{?}{\leq} c\cdot\epsilon^2\cdot\frac{|T|^6}{r^2}, \tag{18}$$

其中，我们利用了当 $k=0$ 时，$\epsilon^2(2^k-1)2^s=0$ 这一事实，将求和中 k 的值域限制为 $k\geq 1$；随后又利用了对任意非空$\mathcal{C}_r$，$s\geq 2k$(h，k，s)这一事实，将 s的值域限制为 $k\geq 1$。

首先请注意，对于任意单字符二进制字符串 t （即字符串“0”和“1”），我们有g_0（t）$=1/2$。因此，我们可以通过只计算完全没有碰撞的那些18 $d\in\mathcal{M}$，来给（18）的分母提供下界（即 $d\in\mathcal{C}_{r|T|}$（0，0，0），其总概率根据命题 5.12可得下界：

$$\sum_{d\in\mathcal{M}}\mathbb{P}\left[\boldsymbol{d}=d\right]\prod_{i=1}^r g_0(T(d,i))\geq\sum_{d\in\mathcal{C}_{r,|T|}(0,0,0)}\mathbb{P}\left[\boldsymbol{d}=d\right]\left(\frac{1}{2}\right)^{|T|}\geq\left(\frac{1}{2}\right)^{|T|}\left(1-\frac{|T|^2}{r}\right)>\left(\frac{1}{2}\right)^{|T|+1},$$

其中，在最后一个不等式中，我们使用了 $|T|^4\leq r/100$ 这一假设。

我们采用类似的方法来给（18）的分子设定上界。对于任意 $d\in\mathcal{C}_{r,|T|}(h,k,s)$，$\prod_{i=1}^r g_0(T(d,i))$ 中满足 $d_i=1$ 的项再次等于 $1/2$，而所有其他项都显然至多为 1。此外，根据（h,k,s）-碰撞的定义，我们有 $\sum_{i=1}^r\mathbb{1}\left[d_i=1\right]=|T|-(2h+k+s)$。因此，对于任意 $h\geq 0$ 和 $s\geq 2k\geq 2$，有

$$\prod_{i=1}^r g_0(T(d,i))\leq\left(\frac{1}{2}\right)^{|T|-(2h+k+s)}\leq\left(\frac{1}{2}\right)^{|T|}\cdot 2^{2h+\frac{3}{2}s},$$

因此，利用 命题 5.11，

$$\begin{aligned}\sum_{d\in\mathcal{C}_{r,|T|}(h,k,s)}\mathbb{P}\left[\boldsymbol{d}=d\right]\prod_{i=1}^r g_0(T(d,i)) &\leq\left(\frac{1}{2}\right)^{|T|}\cdot 2^{2h+\frac{3}{2}s}\cdot\sum_{d\in\mathcal{C}_{r,|T|}(h,k,s)}\mathbb{P}\left[\boldsymbol{d}=d\right] \\ &\leq\left(\frac{1}{2}\right)^{|T|}\cdot\left(\frac{(2h|T|)^2}{r}\right)^h\cdot\left(\frac{(2k|T|)^{3/2}}{r}\right)^s.\end{aligned}$$

结合上述结果，以及以下观察：对于任意非零项，h，k，$s\leq|T|$

在式（18）的分子中，并使用记号 $\lesssim$ 吸收常数因子，可得

$$\begin{aligned}
&\sum_{h\geq 0, s\geq 2k\geq 2}(2^k-1)2^s\cdot\frac{\sum_{d\in\mathcal{C}_{r,|T|}(h,k,s)}\mathbb{P}\left[\boldsymbol{d}=d\right]\prod_{i=1}^r g_0(T(d,i))}{\sum_{d\in\mathcal{M}}\mathbb{P}\left[\boldsymbol{d}=d\right]\prod_{i=1}^r g_0(T(d,i))}\\
&\quad<\sum_{\substack{|T|\geq h\geq 0,\\ |T|\geq s\geq 2k\geq 2}}2^{\frac{3}{2}s}\cdot\frac{\left(\frac{1}{2}\right)^{|T|}\cdot\left(\frac{(2h|T|)^2}{r}\right)^h\cdot\left(\frac{(2k|T|)^{3/2}}{r}\right)^s}{(1/2)^{|T|+1}}\\
&\quad=2\cdot\sum_{\substack{|T|\geq h\geq 0,\\ |T|\geq s\geq 2k\geq 2}}\left(\frac{(2h|T|)^2}{r}\right)^h\cdot\left(\frac{(4k|T|)^{3/2}}{r}\right)^s\\
&\quad\leq 2\cdot\sum_{\substack{|T|\geq h\geq 0,\\ |T|\geq s\geq 2k\geq 2}}\left(\frac{4|T|^4}{r}\right)^h\left(\frac{8|T|^3}{r}\right)^s\\
&\quad<2\cdot\sum_{h\geq 0}\left(\frac{4|T|^4}{r}\right)^h\left[\sum_{s\geq 2}\sum_{1\leq k\leq s/2}\cdot\left(\frac{8|T|^3}{r}\right)^s\right]\\
&\quad\leq 2\cdot\sum_{h\geq 0}\left(\frac{4|T|^4}{r}\right)^h\left[\sum_{s\geq 2}\frac{s}{2}\cdot\left(\frac{8|T|^3}{r}\right)^s\right]\\
&\quad\lesssim\left(\frac{|T|^3}{r}\right)^2\sum_{h\geq 0}\left(\frac{4|T|^4}{r}\right)^h\\
&\quad\lesssim\frac{|T|^6}{r^2},
\end{aligned}$$

其中，我们利用假设 $|T|^4\leq r/100$ 证明了两个几何级数的收敛性，[15] 从而完成证明。

□

现在我们使用这一结果，来给 $\boldsymbol{Z}$ 与一段由 Θ 个（n/m）连续多米诺骨牌产生的子迹之间的互信息给出上界。

引理 5.17. 设 $n^{-1/4}<\epsilon<1$，并令 $\frac{n}{m}\leq r\leq 2\frac{n}{m}$ 为一个正整数。设$\boldsymbol{Z}\sim\mathsf{Ber}(1/2)$，并令 $\boldsymbol{T}$ 表示根据分布 $\mathcal{D}_{\boldsymbol{Z}}$ 由一段包含 r 个连续多米诺骨牌的值域所生成的子迹。设 $m=o\left(\binom{n}{\epsilon}^{4/5}\frac{1}{\log^4 n}\right)$。那么当 $n\to\infty$ 时，我们有

$$I(\boldsymbol{Z}:\boldsymbol{T})=O\left(\frac{\epsilon^4\log^{12}n}{r^4}\right).$$

证明。与之前一样，令 P_z 表示给定 $\boldsymbol{Z}=z$ 时 $\boldsymbol{T}$ 的条件概率。令 $Q:=(P_0+P_1)/2$ 表示 $\boldsymbol{T}$ 的（边缘）分布。

我们的策略是根据以下条件对可能的子迹集合进行分解：T 判断是否 $|T|\lesssim\log n$（典型情况），或者 $|T|\gtrsim\log n$。在前一种情况下，引理 5.16 将给出比值$P_1(T)/P_0(T)$ 接近 1；而在后一种情况下，我们将使用泊松集中界来说明如此长的迹对互信息的贡献不会太大。

[15] Namely, we used the formulas $\sum_{i\geq 0}x^i=\frac{1}{1-x}$ and $\sum_{i\geq 2}ix^i=\frac{(2-x)x^2}{(1-x)^2}$ for $|x|<1$.

具体来说，我们首先将 $I(\boldsymbol{Z}:\boldsymbol{T})$ 上界为两项之和：一项是针对 $|T| \leq 20\log n$ 的 χ^2 类型表达式，另一项是针对 $|T| > 20\log n$ 的尾概率：

$$\begin{aligned}
I(\boldsymbol{Z}:\boldsymbol{T}) &= H(\boldsymbol{T}) - H(\boldsymbol{T}|\boldsymbol{Z}) \\
&= -\sum_T Q(T)\log(Q(T)) + \sum_{z\in\{0,1\}} \mathbb{P}[Z=z]\sum_T P_z(T)\log(P_z(T)) \\
&= -\sum_T \frac{P_0(T)+P_1(T)}{2}\log(Q(T)) + \sum_T \left[\frac{P_0(T)}{2}\log(P_0(T)) + \frac{P_1(T)}{2}\log(P_1(T))\right] \\
&= \frac{1}{2}\sum_T \left[P_0(T)\log\left(\frac{P_0(T)}{Q(T)}\right) + P_1(T)\log\left(\frac{P_1(T)}{Q(T)}\right)\right] \\
&\leq \frac{1}{2}\sum_T \left[P_0(T)\left(\frac{P_0(T)}{Q(T)}-1\right) + P_1(T)\left(\frac{P_1(T)}{Q(T)}-1\right)\right] \\
&= \frac{1}{2}\sum_T \left[P_0(T)\left(\frac{P_0(T)-P_1(T)}{P_0(T)+P_1(T)}\right) + P_1(T)\left(\frac{P_1(T)-P_0(T)}{P_0(T)+P_1(T)}\right)\right] \\
&= \frac{1}{2}\sum_T \left[\frac{(P_1(T)-P_0(T))^2}{P_0(T)+P_1(T)}\right] \\
&< \frac{1}{2}\sum_{T:|T|\leq 20\log n} \frac{(P_1(T)-P_0(T))^2}{P_0(T)} + \frac{1}{2}\sum_{T:|T|>20\log n}(P_0(T)+P_1(T)) \\
&= \frac{1}{2}\sum_{T:|T|\leq 20\log n} P_0(T)\left(\frac{P_1(T)}{P_0(T)}-1\right)^2 \\
&\quad + \frac{1}{2}\mathbb{P}[|\boldsymbol{T}| > 20\log n \mid \boldsymbol{Z}=0] + \frac{1}{2}\mathbb{P}[|\boldsymbol{T}| > 20\log n \mid \boldsymbol{Z}=1]\,.
\end{aligned}$$

我们先来看上一个表达式中的第一项。我们要证明，当$|T| \leq 20\log n$时，有 $|T|^4 = o(r$，这足以满足 引理 5.16 的条件。回顾假设 $\frac{n}{m} \leq r \leq 2\frac{n}{m}$, $m = o\left(\left(\frac{n}{\epsilon}\right)^{4/5} \frac{1}{\log^4 n}\right)$ 和 $\epsilon \geq \frac{1}{n^{1/4}}$，可得

$$\frac{|T|^4/20^4}{r} \leq \frac{\log^4 n}{r} \leq \frac{m\log^4 n}{n} \ll \frac{\left(\frac{n}{\epsilon}\right)^{4/5}\frac{1}{\log^4 n}\log^4 n}{n} \leq 1\,,$$

因此条件 对足够大的 $|T|^4 \leq r/100$ 成立 n。因此引理 5.16 给出

$$\frac{1}{2}\sum_{T:|T|\leq 20\log n} P_0(T)\left(\frac{P_1(T)}{P_0(T)}-1\right)^2 \leq \frac{1}{2}\sum_{T:|T|\leq 20\log n} P_0(T)\left(c\cdot\epsilon^2\cdot\frac{|T|^6}{r^2}\right)^2 = O\left(\frac{\epsilon^4\log^{12} n}{r^4}\right).$$

现在我们处理第二个部分。回顾（见 观察 5.6）， $|\boldsymbol{T}|$ 服从一个完全由值域中的多米诺骨牌数量决定的泊松分布：

$$|\boldsymbol{T}| \sim \mathsf{Poi}\left(r\cdot\frac{2m}{n}\right),$$

相互独立于 $\boldsymbol{Z}$。令 $\lambda := r\cdot\frac{2m}{n}$ 并注意到 $2 \leq \lambda \leq 4 \leq$ 对数 n。 事实 3.4 给出，对于 $z \in \{0,1\}$,

$$\mathbb{P}[|\boldsymbol{T}| > 20\log n \mid \boldsymbol{Z}=z] \leq \mathbb{P}[|\boldsymbol{T}| - \lambda > 19\log n] \leq e^{-\frac{(19\log n)^2}{2(\lambda+19\log n)}} \leq e^{-\frac{361\log n}{40}} \leq \frac{1}{n^9}\,.$$

最后，我们有

$$\frac{1}{n^9} \le \frac{\epsilon^4}{r^4} \impliedby \frac{1}{n^9} \le \frac{1/n}{16n^4/m^4} \iff \frac{16}{n^4} \le m^4 \iff \frac{2}{n} \le m\,,$$

这一点显然成立，证明完毕。 □

由于给定 $\boldsymbol{Z}$ 时，由不相交范围产生的子迹条件独立，应用链式法则和数据处理不等式即可完成证明。

引理 5.18 (断言 5.9的改进)。设 $n^{-1/4} < \epsilon < 1$。令 $\boldsymbol{Z} \sim$ 服从$(1/2)^n$，$\boldsymbol{\pi} \sim \mathcal{D}_{\boldsymbol{\pi}}$，并令 $\boldsymbol{\mathcal{T}}$为从 π 中抽取的大小为 泊松分布(m) 的迹。则若 $m = o\left(\left(\frac{n}{\epsilon}\right)^{4/5} \frac{1}{\log^4 n}\right)$，我们有

$$I(\boldsymbol{Z} : \boldsymbol{\mathcal{T}}) = O\left(\frac{\epsilon^4 m^5}{n^4} \log^{12} n\right) = o(1)\,,$$

因此，任何以至少 51% 的概率（对 $\boldsymbol{Z}$ 和 $\boldsymbol{\mathcal{T}}$而言）成功区分 *YES* 和 *NO* 情形的算法，其样本复杂度至少为

$$\Omega\left(\left(\frac{n}{\epsilon}\right)^{4/5} \cdot \frac{1}{\log^4 n}\right)\,.$$

证明。固定一个将定义域划分为连续范围的任意分割，$[2n]$ 其中连续范围为 R_1，……，$R_{m'}$，满足：1）每个 R_i 都是长度为 4 的倍数的连续区间，因此由 $|R_i|/4$ 个连续多米诺骨牌组成（回忆一下，为简单起见，我们假设 n 为偶数）；2）每个 R_i 满足 $\frac{n}{m} \le |R_i|/4 \le 2\frac{n}{m}$。由此可得 $m' = \Theta(m)$，并令 $r_i := |R_i|/4$。对于每个 $i \in [m']$，每个 r_i 均满足引理 5.17的要求。

设 $\boldsymbol{T}_i$ 为由值域 R_i 生成的子迹，因此最终的迹由所有子迹拼接而成：

$$\boldsymbol{\mathcal{T}} = \boldsymbol{T}_1 \circ \boldsymbol{T}_2 \cdots \circ \boldsymbol{T}_{m'}\,.$$

数据处理不等式给出

$$I(\boldsymbol{Z} : \boldsymbol{\mathcal{T}}) \le I(\boldsymbol{Z} : \boldsymbol{T}_1, \ldots, \boldsymbol{T}_{m'})\,.$$

注意，在条件 $\boldsymbol{Z}$下，$\boldsymbol{\pi}$ 在不同多米诺骨牌中的各项相互独立，正如 定义 5.7所述。再回想 定义 5.5中对子迹分布的描述，可知 $\boldsymbol{T}_i$ 在给定条件 $\boldsymbol{Z}$下相互独立。因此，互信息链式法则和引理 5.17给出

$$I(\boldsymbol{Z} : \boldsymbol{\mathcal{T}}) \le \sum_{i=1}^{m'} I(\boldsymbol{Z} : \boldsymbol{T}_i) = \Theta(m) \cdot O\left(\frac{\epsilon^4 m^4 \log^{12} n}{n^4}\right) = O\left(m^5 \cdot \frac{\epsilon^4}{n^4} \log^{12} n\right) = o(1)\,,$$

如所需。最后，应用 事实 5.3即可得出第二个结论。 □

结合 命题 5.2 和引理 5.18 即可证明定理 5.1：

推论 5.19 (对 定理 5.1 的改进)。设 Π_1 仅包含$[2n]$上的均匀分布，并设 Π_2 为所有在全变差距离意义下与均匀分布相距至少 ϵ的、定义在 $[2n]$ 上的分布所构成的集合。那么，在奇偶迹下对$(\Pi_1, \Pi_2, 51/100)$进行性质测试所需的样本复杂度至少为$\Omega\left(\left(\frac{n}{\epsilon}\right)^{4/5} \frac{1}{\log^4 n} + \frac{\sqrt{n}}{\epsilon^2}\right)$ 个样本。此外，即使保证输入分布 π 在取值为零（即偶数）的坐标上均匀分配了 $1/2$ 质量，这一下界仍然成立。

证明。 $\Omega(\sqrt{n}/\epsilon^2)$的下界由命题 5.2得出。此外，我们有

$$\left(\frac{n}{\epsilon}\right)^{4/5}\frac{1}{\log^4 n} \geq \frac{\sqrt{n}}{\epsilon^2} \iff \epsilon \geq \frac{\log^{10/3} n}{n^{1/4}},$$

在这种情况下引理 5.18给出了界。 □

6 分布无关的基于样本的性质测试

在本节中，我们将奇偶迹下的分布检验与分布无关的基于样本的性质测试联系起来。本节的主要思路如下:

1. 我们将带标签分布测试定义为对分布无关的基于样本的性质测试的一种广义重新表述，以更明确地体现其与分布测试之间的联系。2. 有一种自然的带标签分布性质，我们称之为密度性质，它涵盖了一些性质测试问题，以及标准分布测试问题的一些更具挑战性的版本。我们利用编辑距离和拉姆齐理论证明，测试这些性质等价于在奇偶迹下测试分布。3. 利用这一等价关系，我们通过应用定理 1.1，在分布无关的基于样本的检验（更一般的带标签分布定义下）中获得新的紧确正面结果。4. 对带标签分布测试存在一种通过学习进行测试的归约，类似于标准的通过学习进行测试的归约 [GGR98]，，它能够给出奇偶迹下分布测试的非构造性上界。我们将在第 7 节中利用这一方法，为迹重构模型中的一些测试问题给出上界。

本节安排如下:

第6.1节： 带标签分布的定义。

第6.2节： 编辑距离的定义。编辑距离与带标签分布密切相关，也是本文余下部分所有应用所必需的。

第6.3节： 带标签分布测试的定义，以及它如何推广传统分布测试和分布无关的基于样本的性质测试模型。

第6.4节： 用于带标签分布测试的通过学习进行测试的归约。

第6.5节： 密度性质的定义，以及密度性质测试与奇偶迹下的分布检验的等价性。

第 6.6 节： 带标签分布测试主要结果的证明，定理 1.5，这是定理 1.1的一个应用。

第 6.7 节： 奇偶迹下检验均匀分布与不受限分布的上界（定理 6.34）。

第6.8节： 在奇偶迹下测试支持集大小与测试 k 分布无关的基于样本模型中的k-交替函数之间的等价性（定理 6.38），以及 [BFH21] 关于测试半空间下界的另一种证明，等等。

6.1 带标签分布

We s现在定义带标签分布和编辑距离，这两者密切相关 .

定义 6.1（带标签分布）。 在任意固定域 $\mathcal{X}$ 上，带标签分布 是一个由函数 $f\colon \mathcal{X} \to \{0,1\}$ 和定义在 $\mathcal{X}$ 上的概率分布 $\mathcal{D}$ 构成的二元组（f， $\mathcal{D}$）。对于定义在 $\mathcal{X} \times \{0,1\}$ 上的概率分布，我们记作 $\mathcal{D}_f$，其中任意（x， b）$\in \mathcal{X} \times \{0,1\}$ 的密度定义为

$$\mathcal{D}_f(x,b) := \begin{cases} \mathcal{D}(x) & \text{if } b = f(x) \\ 0 & \text{otherwise.} \end{cases}$$

换句话说，通过对 $x \sim \mathcal{D}$ 进行采样并取 $(x,\ f(x))$，即可获得来自 $\mathcal{D}_f$ 的样本。

我们研究 $\mathcal{X} = \mathbb{Z}$ 这一情形。对于 $\mathbb{Z}$ 上的带标签分布 $(f, \mathcal{D})$，f 可能会无限次地“交替”。我们将研究范围限制在“适当的”带标签分布上，即 f 在“左侧”具有有限次交替，其定义如下。

定义 6.2（适当的带标签分布）。 如果存在 $t \in \mathbb{Z}$ 使得对于所有 $x < t$， $f(x) = 1$，则带标签分布 $(f, \mathcal{D})$ 是 *1*-适当 的。如果对于所有 $x < t$， $f(x) = 0$，则称其为 *0*-适当 。如果 $(f, \mathcal{D})$ 是 0-适当或 1-适当，我们称其为适当。

注 6.3. 研究带标签分布测试时，只需考虑适当的带标签分布。这是因为，对于每个带标签分布 $\mathcal{D}_f$ 和每个 $\delta > 0$，都存在一个适当的带标签分布 $\mathcal{D}_g$，使得 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_g) < \delta$。因此，对于任何样本量有界的算法而言，每个分布都与某个适当分布不可区分。

定义 6.4（交替序列）。 对于任意适当的带标签分布 $(f, \mathcal{D})$，交替序列 是唯一的序列 $a_1 < a_2 < a_3 < \cdots$，使得 f 在各区间 $(-\infty, a_1]$、$(a_1, a_2]$、$\ldots$ 上为常数，并且 $f(a_{i-1}) \neq f(a_i)$。注意，如果 $(f, \mathcal{D})$ 是 1-适当的，那么 f 在区间 $(a_{i-1}, a_i]$ 上取值 $\mathsf{par}(i)$，在 $(-\infty, a_1]$ 上取值 1。如果 $(f, \mathcal{D})$ 是 0-适当的，则取相反的值。注意，当 $(f, \mathcal{D})$ 适当时，该序列总是存在，并且可能是无限序列。

定义 6.5（密度序列）。 对于任意适当的带标签分布 $(f, \mathcal{D})$，其交替序列为 $a_1 < a_2 < a_3 < \cdots$，我们将密度序列 $\pi_{f,\mathcal{D}}$ 定义如下。如果 $(f, \mathcal{D})$ 是 1-适当的，我们定义

$$\pi_{f,\mathcal{D}}(i) := \begin{cases} \mathcal{D}(-\infty, a_1] & \text{if } i = 1 \\ \mathcal{D}(a_{i-1}, a_i] & \text{if } i > 1\,. \end{cases}$$

如果 $(f, \mathcal{D})$ 是 0-适当的，我们定义

$$\pi_{f,\mathcal{D}}(i) := \begin{cases} 0 & \text{if } i = 1 \\ \mathcal{D}(-\infty, a_1] & \text{if } i = 2 \\ \mathcal{D}(a_{i-2}, a_{i-1}] & \text{if } i > 2\,. \end{cases}$$

注意，$\pi_{f,\mathcal{D}}$ 是一个概率分布，因为 $\sum_{i=1}^{\infty} \pi_{f,\mathcal{D}}(i) = \sum_{x \in \mathbb{Z}} \mathcal{D}(x) = 1$。对于任意由适当的带标签分布构成的 Ξ 集合，我们记作

$$\Pi(\Xi) := \{\pi_{f,\mathcal{D}} : (f, \mathcal{D}) \in \Xi\}$$

对于与 Ξ 中的适当带标签分布相关联的密度序列（概率分布）集合。

以下用于带标签分布的全变差距离的简单公式通常很有用。

命题 6.6. 设 $(f, \mathcal{D})$ 和 $(g, \mathcal{E})$ 是定义在 $\mathbb{Z}$ 上的带标签分布。这 n

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g) = \frac{1}{2} \sum_{i \in \mathbb{Z}} \mathbb{1}\left[f(i) \neq g(i)\right] \left(\mathcal{D}(i) + \mathcal{E}(i)\right) + \mathbb{1}\left[f(i) = g(i)\right] \left|\mathcal{D}(i) - \mathcal{E}(i)\right|.$$

证明。 根据定义，

$$\begin{aligned}
\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g) &= \frac{1}{2} \sum_{(i,b) \in \mathbb{Z} \times \{0,1\}} \left|\mathcal{D}_f(i,b) - \mathcal{E}_g(i,b)\right| \\
&= \frac{1}{2} \left(\sum_{\substack{(i,b) \in \mathbb{Z} \times \{0,1\} \\ f(i) = g(i) = b}} \left|\mathcal{D}(i) - \mathcal{E}(i)\right| + \sum_{\substack{(i,b) \in \mathbb{Z} \times \{0,1\} \\ f(i) = b, g(i) \neq b}} \mathcal{D}(i) + \sum_{\substack{(i,b) \in \mathbb{Z} \times \{0,1\} \\ f(i) \neq b, g(i) = b}} \mathcal{E}(i) \right) \\
&= \frac{1}{2} \sum_{i \in \mathbb{Z}} \mathbb{1}\left[f(i) \neq g(i)\right] \left(\mathcal{D}(i) + \mathcal{E}(i)\right) + \mathbb{1}\left[f(i) = g(i)\right] \left|\mathcal{D}(i) - \mathcal{E}(i)\right|. \qquad \square
\end{aligned}$$

6.2 编辑距离

我们定义两种编辑距离：一种适用于定义在 $\mathbb{Z}$ 上的带标签分布，另一种适用于定义在 $\mathbb{N}$ 上的分布。

定义 6.7 （带标签分布的编辑距离）。 对于定义在 $\mathbb{Z}$ 上的任意两个适当的带标签分布 $(f, \mathcal{D})$ 和 $(g, \mathcal{E})$，定义

$$\mathsf{dist}_{\mathsf{edit}}((f, \mathcal{D}), (g, \mathcal{E})) := \inf \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_{f'}, \mathcal{E}'_{g'}),$$

其中，下确界取遍所有与原分布具有相同密度序列的适当的带标签分布 $(f', \mathcal{D}')$ 和 $(g', \mathcal{E}')$，即满足 $\pi_{f', \mathcal{D}'} = \pi_{f, \mathcal{D}}$和 $\pi_{g', \mathcal{E}'} = \pi_{g, \mathcal{E}}$ 的分布。

接下来，我们将定义 $\mathbb{N}$ 上分布的编辑距离。回顾一下，全变差距离并非奇偶迹下分布测试的自然距离度量，因为分布之间的全变差距离可能达到最大值 1，但在奇偶迹下却不可区分。编辑距离取代全变差距离，成为奇偶迹下自然的（伪）度量。我们首先引入分数字符串的概念。

定义 6.8 （分数字符串）。 分数字符串 是一个有限序列 $\sigma_1^{p_1} \sigma_2^{p_2} \ldots \sigma_n^{p_n}$，其中每个分数字符 $\sigma_i^{p_i}$由一个符号 $\sigma_i \in \{0, 1\}$ 和一个值 $p_i \in \mathbb{R}_{\geq 0}$组成。

现在我们定义分数字符串的编辑距离，它类似于字符串的标准编辑距离。

定义 6.9 （分数字符串的编辑距离）。 设 $a = a_1^{p_1} a_2^{p_2} \cdots a_n^{p_n}$ 是一个分数字符串。我们定义对 a 进行的以下允许的编辑操作及其相应代价：

插入： 对于 $i \in [n+1]$ 和 $b \in \{0, 1\}$，$\mathsf{ins}_{i,\, b}(a)$ 是通过在 $a_i^{p_i}$ 前立即插入分数字符 b^0 得到的分数字符串。此操作的代价为 0。

删除： 对于满足 $p_i = 0$ 的 $i \in [n]$，$\mathsf{del}_i(a)$ 是删除分数字符 $a_i^{p_i} = a_i^0$ 后得到的分数字符串。此操作的成本为 0。

重排： 对于满足 $a_i = a_{i+1}$ 且 $-p_i \le \delta \le p_{i+1}$ 的 $i \in [n-1]$，$\mathsf{rearr}_{i,\,\delta}(a)$ 是将 $a_i^{p_i} a_{i+1}^{p_{i+1}}$ 替换为 $a_i^{p_i+\delta} a_{i+1}^{p_{i+1}-\delta}$ 后得到的分数字符串。此操作的成本为 0。

调整： 对于 $i \in [n]$ 和 $\delta \ge -p_i$，$\mathsf{adj}_{i,\,\delta}(a)$ 是将 p_i 替换为$p_i + \delta$后得到的分数字符串，因此分数字符 $a_i^{p_i}$变为 $a_i^{p_i+\delta}$。此操作的代价为 $|\delta|/2$。

对于分数字符串 a，我们称操作序列 $O_1, \ldots, O_k$是允许的，如果对于每个$i \in [k]$，O_i都是作用于分数字符串 $(O_{i-1} \circ O_{i-2} \circ \cdots \circ O_1)(a)$ 的允许操作。

对于两个分数字符串 a 和 b，我们将 编辑距离 $\mathsf{dist}_{\mathsf{fr-edit}}(a,\ b)$ 定义为满足以下条件的最小c：存在一系列允许的操作 $O_1, \ldots, O_k$，使得 $O_k \circ O_{k-1} \circ \cdots \circ O_1(a) = b$，且操作 O_i 的代价之和为 c。

令 $\pi : \mathbb{N} \to \mathbb{R}_{\ge 0}$ 为任意有限支持概率分布。我们按如下方式定义分数字符串 $\mathsf{str}(\pi)$。由于 π 有限支撑，因此存在某个 $k \in \mathbb{N}$，使得对所有$i > k$，$\pi(i) = 0$。然后我们定义

$$\mathsf{str}(\pi) := 1^{\pi(1)} 0^{\pi(2)} 1^{\pi(3)} 0^{\pi(4)} \cdots (\mathsf{par}(k))^{\pi(k)} .$$

最后，我们可以定义分布的编辑距离。

定义 6.10（分布的编辑距离）。对于两个有限支持分布 π, π'，我们定义

$$\mathsf{dist}_{\mathsf{edit}}(\pi, \pi') := \mathsf{dist}_{\mathsf{fr-edit}}(\mathsf{str}(\pi), \mathsf{str}(\pi')) .$$

下面给出的编辑距离的另一种刻画很有帮助。证明将在附录 D.2中给出。

引理 6.11. 设 π 和 π' 是 $\mathbb{N}$上的有限支持分布。那么

$$\mathsf{dist}_{\mathsf{edit}}(\pi, \pi') = \inf \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g) ,$$

下确界是在满足 $\pi = \pi_{f,\mathcal{D}}$ 和$\pi' = \pi_{g,\mathcal{E}}$的带标签分布$(f, \mathcal{D})$ 与$(g, \mathcal{E})$ 上取的。

从这个引理可以看出，分布的编辑距离与带标签分布的编辑距离本质上是等价的：对于两个适当的带标签分布（$f, \mathcal{D}$）和（$g, \mathcal{E}$），该引理意味着

$$\mathsf{dist}_{\mathsf{edit}}((f, \mathcal{D}), (g, \mathcal{E})) = \mathsf{dist}_{\mathsf{edit}}(\pi_{f,\mathcal{D}}, \pi_{g,\mathcal{E}}) .$$

很容易看出，以下不等式通常成立：

$$\mathsf{dist}_{\mathsf{edit}}(\pi, \pi') \le \mathsf{dist}_{\mathsf{TV}}(\pi, \pi') . \tag{19}$$

取 $(f, \mathcal{D})$，其中 $f(i) = \mathsf{par}(i)$ 对于 $i \in \mathbb{N}$和 $f(i) = 1$对于 $i \le 0$，并且$\mathcal{D}(i) = \pi(i)$ 对于 $i \in \mathbb{N}$ 和 $\mathcal{D}(i) = 0$ 对于 $i \le 0$，即可验证。对 π' 类似地定义 $(g, \mathcal{E})$。这满足 $\pi_{f,\mathcal{D}} = \pi$和 $\pi_{g,\mathcal{E}} = \pi'$，并且 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g) = \mathsf{dist}_{\mathsf{TV}}(\pi, \pi')$。

6.3 带标签分布测试

现在我们介绍带标签分布测试。对于定义域为固定域 $\mathcal{X}$ 上的带标签分布 $(f, \mathcal{D})$ 和 $(g, \mathcal{E})$，我们滥用记号，将其写作

$$\mathsf{dist}_{\mathsf{TV}}((f, \mathcal{D}), (g, \mathcal{E})) := \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g) ,$$

因此，对于带标签分布的一种性质 Ξ，我们有

$$\mathrm{FAR}_{\epsilon}^{\mathsf{TV}}(\Xi) := \{(f,\mathcal{D}) : \forall (g,\mathcal{E}) \in \Xi, \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\mathcal{E}_g) > \epsilon\}\,.$$

定义 6.12（带标签分布测试）。设 Ξ_1 和 Ξ_2 是固定定义域 $\mathcal{X}$ 上带标签分布的性质。样本复杂度为 m 的（Ξ_1，Ξ_2，α）带标签分布测试器是一个算法 A，对于 $\mathcal{X}$ 上的每个带标签分布（f，$\mathcal{D}$），它都满足以下条件：

1. 如果（f，$\mathcal{D}$）满足 $\in \Xi_1$，则 $S_f \sim$ $\mathbb{P}_{\text{采样}(\mathcal{D}_f,\ m)}[A(S_f)$ 接受 $] \geq \alpha$；并且 2. 如果（f，$\mathcal{D}$）满足 $\in \Xi_2$，则 $S_f \sim$ $\mathbb{P}_{\text{采样}\ (\mathcal{D}_f,\ m)}[A(S_f)$ 拒绝 $] \geq \alpha$。

该问题的规范形式为：对于某个 $\epsilon > 0$，有 $\Xi_2 := \mathrm{FAR}_{\epsilon}^{\mathsf{TV}}(\Xi_1)$。

我们在此证明，可以从更一般的带标签分布测试模型中得到标准分布测试模型和分布无关的基于样本的性质测试模型。为了从带标签分布测试模型得到标准分布测试模型（其目标是测试定义域 $\mathcal{X}$ 上分布的性质 Π），只需选择性质 $\Xi = \Lambda \times \Pi$，其中 Λ 仅包含定义域 $\mathcal{X}$ 上的常零函数。

如何从带标签分布测试模型获得分布无关的基于样本的性质测试，略微不那么显而易见。分布无关的基于样本的性质测试定义如下。

定义 6.13（分布无关的基于样本的性质测试）。设 Λ 是定义在某个固定定义域 $\mathcal{X}$ 上的函数 $\mathcal{X} \to \{0,1\}$ 的一个性质，且设 $\epsilon > 0$。一个样本复杂度为 m 的（$\Lambda, \epsilon, \alpha$）分布无关的基于样本的性质测试器，是满足以下条件的算法 A：对于每个函数 f：$\mathcal{X} \to \{0,1\}$，以及定义在 $\mathcal{X}$ 上的分布 $\mathcal{D}$：

1. 如果 $f \in \Lambda$，那么 $S_f \sim$ $\mathbb{P}_{\text{采样}(\mathcal{D}_f,m)}[A(S_f)$ 接受 $] \geq \alpha$；并且

2. 如果对于所有 $g \in \Lambda$，$\mathbb{P}_{x\sim\mathcal{D}}[f(x) \neq g(x)] > \epsilon$，那么 $S_f \sim$ $\mathbb{P}_{\text{采样}(\mathcal{D}_f,m)}[A(S_f)$ 拒绝 $] \geq \alpha$。

这个问题无法简洁地表述为区分函数 $\mathcal{X} \to \{0,1\}$ 的两个性质Λ_1 和 Λ_2 的问题，因为应当被拒绝的函数集合取决于分布 $\mathcal{D}$。但是，我们可以将其表述为区分带标签分布的两个性质 Ξ_1和 Ξ_2 的问题，如接下来的两个命题所示。

命题 **命题 6.14.** 固定一个定义域 $\mathcal{X}$，并设 $(f,\mathcal{D})$、$(g,\mathcal{E})$ 为带标签分布。 n

$$\mathbb{P}_{x\sim\mathcal{D}}[f(x) \neq g(x)] \leq 2 \cdot \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\mathcal{E}_G)\,.$$

证明。对于任意事件 $X \subseteq \mathcal{X}\times\{0,1\}$，记 $\mathcal{D}_f(X) := \mathbb{P}_{x\sim\mathcal{D}}[(x, f(x)) \in X]$。记 $\epsilon := \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\mathcal{E}_g)$，使得对于任意事件 $X \subseteq \mathcal{X} \times \{0,1\}$，都有 $|\mathcal{D}_f(X) - \mathcal{E}_g(X)| \leq \epsilon$。

令 $\Delta := \{x \in \mathcal{X} : f(x) \neq g(x)\}$，并定义事件 $E := \{(x, b) : x \in \Delta, b = g(x)\}$。那么$\mathcal{D}_f(E) = 0$ 根据定义，并且 $|\mathcal{D}_f(E) - \mathcal{E}_g(E)| \leq \epsilon$，所以 $\mathcal{E}_g(E) \leq \epsilon$。那么同样 $\mathcal{E}(\Delta) = \mathcal{E}_g(E) \leq \epsilon$。

定义事件 $F := \{(x, b) : x \in \Delta, b \in \{0,1\}\}$，它满足 $\mathcal{D}_f(F) = \mathcal{D}(\Delta)$ 和 $\mathcal{E}_g(F) = \mathcal{E}(\Delta)$。因此 $|\mathcal{D}(\Delta) - \mathcal{E}(\Delta)| \leq \epsilon$，所以 $\mathcal{D}(\Delta) \leq \mathcal{E}(\Delta) + \epsilon \leq 2\epsilon$。我们得出结论：

$$\mathbb{P}_{x\sim\mathcal{D}}[f(x) \neq g(x)] = \mathcal{D}(\Delta) \leq 2\epsilon\,. \qquad \square$$

命题 6.15。 固定任意定义域 $\mathcal{X}$。令 Λ 为函数 $\mathcal{X} \to \{0,1\}$的任意函数性质，并令 Π为 $\mathcal{X}$上的所有分布构成的集合。那么，对于 $\Xi = \Lambda \times \Pi$,

1. 如果存在一个$(\Xi, \mathrm{FAR}^{\mathsf{TV}}_{\epsilon/2}(\Xi)\ \alpha)$-带标签分布测试器，其样本复杂度为 m，那么就存在一个$(\Lambda, \epsilon, \alpha)$-无分布假设的基于样本的测试器，其样本复杂度为 m。 *2.* 如果存在一个$(\Lambda, \epsilon, \alpha)$-无分布假设的基于样本的测试器，其样本复杂度为 m，那么就存在一个$(\Xi, \mathrm{FAR}^{\mathsf{TV}}_{\epsilon}(\Xi))$-带标签分布测试器，其样本复杂度为 m。

证明。 对于第一个结论，分布无关的基于样本的性质测试器的输入是一个函数 $f\colon \mathcal{X} \to \{0,1\}$ 以及定义在 $\mathcal{X}$ 上的一个分布 $\mathcal{D}$。该算法将获取一个带标签的样本 S_f，其中 $S \sim \mathsf{samp}(\mathcal{D}, m)$，并在 S_f 上模拟带标签分布测试器。假设 $f \in \Lambda$。那么 $\mathcal{D}_f \in \Lambda \times \Pi = \Xi$，因此带标签分布测试器将以至少 α 的概率接受。

现在假设 f 相对于 $\mathcal{D}$ 与 Λ 的距离至少为 ϵ。那么对于所有 $g\colon \mathcal{X} \to \{0,1\}$，以及所有定义在 $\mathcal{X}$ 上的分布 $\mathcal{E}$，根据命题 6.14，有 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g) \geq \frac{1}{2}\mathbb{P}_{x\sim\mathcal{D}}[f(x) \neq g(x)] > \epsilon/2$。因此，$(f, \mathcal{D}) \in \mathrm{FAR}^{\mathsf{TV}}_{\epsilon/2}(\Xi)$，所以测试器以至少 α 的概率拒绝。

对于第二个结论，我们通过以下方式得到 Ξ 的带标签分布测试器：输入 $(f, \mathcal{D})$ 后，获取带标签的样本 $S_f \sim \mathsf{samp}(\mathcal{D}_f, m)$，并在 S_f 上运行基于样本的无分布假设性质测试器。

如果 $(f, \mathcal{D}) \in \Xi$，那么 $f \in \Lambda$，因此性质测试器以至少 α 的概率接受。

如果对于所有 $(g, \mathcal{E}) \in \Xi$，都有 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g) > \epsilon$，那么特别地，对于所有 $g \in \Lambda$，都有 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_g) > \epsilon$。于是存在一个事件 $E \subseteq \mathcal{X} \times \{0,1\}$，使得 $|\mathcal{D}_f(E) - \mathcal{D}_g(E)| > \epsilon$。不失一般性，我们可以假设，对于每个 $(x,b) \in E$，$f(x) \neq g(x)$（因为我们可以移除所有满足 $f(x) = g(x)$ 的点对 (x,b)，而不改变这个差值）。然后我们可以将 $E = E_f \cup E_g$ 划分为其中 $E_f := \{(x,b) \in E\colon b = f(x)\}$，以及 $E_g := \{(x,b) \in E\colon b = g(x)\}$。于是 $\mathcal{D}_f(E) = \mathcal{D}_f(E_f)$，并且 $\mathcal{D}_g(E) = \mathcal{D}_g(E_g)$。

必定有 $\mathcal{D}_f(E_f) > \epsilon$ 或 $\mathcal{D}_g(E_g) > \epsilon$。若 $\mathcal{D}_f(E_f) > \epsilon$，则我们可以选择$\Delta := \{x \in \mathcal{X}\colon (x, f(x)) \in E_f\}$，因此

$$\mathbb{P}_{x\sim\mathcal{D}}[f(x) \neq g(x)] \geq \mathcal{D}(\Delta) = \mathcal{D}_f(E_f) > \epsilon\,.$$

若 $\mathcal{D}_g(E_g) > \epsilon$，则类似的结论也成立。由于这对所有 $g \in \Lambda$ 都成立，我们看到性质测试器应以至少 α 的概率拒绝。 □

6.4 通过学习进行测试

[GGR98] 观察到，假设类 Π 的适当学习算法可以通过加入一个额外的“验证”步骤用作性质测试器。对我们而言，将同样的技术应用于另一种适用于带标签分布类的学习算法很方便；在这种情况下，“假设类”不只是一组函数 $\mathcal{X} \to \{0,1\}$，而是一组函数—分布对。

对于固定域 $\mathcal{X}$ 和分布在 $\mathcal{X}$ 上的性质 Π，用 $\mathrm{CLOSE}^{\mathsf{TV}}_{\epsilon}(\Pi)$ 表示满足 $\mathsf{dist}_{\mathsf{TV}}(\pi, \Pi) \leq \epsilon$ 的、定义在 $\mathcal{X}$ 上的分布集合 π。

定义 6.16（带标签分布学习）。 设 Ξ 是定义在某个固定域 $\mathcal{X}$ 上的带标签分布的一个性质。对于 Ξ，成功概率为 α、误差为 $\epsilon > 0$、样本复杂度为 m 的带标签分布学习算法 是指这样一种算法 A：对于任意输入 $(f, \mathcal{D})$，它接收一个带标签的样本 $S_f \sim \mathsf{samp}(\mathcal{D}_f, m)$，并输出一个函数 $g\colon \mathcal{X} \to \{0,1\}$；该算法在 S_f 和算法随机性的共同作用下，以至少 α 的概率成功 。成功 事件定义如下：

成功： 如果 $(f, \mathcal{D}) \in \Xi$，则 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_g) < \epsilon$。

我们称该 算法*proper* 是恰当的，当成功事件还满足以下附加条件时 s:

成功*：
若 $(f, \mathcal{D}) \notin \Xi$，则存在 $\mathcal{E}$ 使得 $\mathcal{E}_g \in \Xi$。若 $(f, \mathcal{D}) \in \Xi$，则存在 $\mathcal{E}$ 使得 $\mathcal{E}_g \in \Xi$，且 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{E}) < \epsilon$。

定义 6.17（学习器-验证器对）。设 Ξ 是某个固定域 $\mathcal{X}$上的带标签分布性质，设 ϵ，$\delta > 0$，并设 A是针对 Ξ的恰当带标签分布学习算法，其成功概率为 $1 - \delta/3$、误差为 $\epsilon/4$、样本复杂度为 m_A。设 $\mathcal{G}_A \subseteq \{0,1\}^{\mathcal{X}}$是 A的值域。对于每个 $g \in \mathcal{G}_A$，令 Π_g 为性质

$$\Pi_g := \{\mathcal{E} : (g, \mathcal{E}) \in \Xi\}.$$

假设 $B = \{B_g\}_{g \in \mathcal{G}_A}$ 是一个算法族，使得对于每个 $g \in \mathcal{G}_A$，算法 B_g 都是一个 $(\textsc{close}^{\mathsf{TV}}_{\epsilon/4}(\Pi_g)\ \textsc{far}^{\mathsf{TV}}_{\epsilon/2}(\Pi_g)\ 1-\delta/3)$-分布测试器，样本复杂度为 m_B。我们称 (A, B) 是针对 Ξ的学习器-验证器对，其成功概率为 $1-\delta$、误差为 ϵ、样本复杂度为 $m_A + m_B$。

命题 6.18。设 Ξ 是带标签分布的一个性质，使得存在一个学习器—验证器对 (A, B)，其成功概率为 $1-\delta$，误差为 $\epsilon > 0$，样本复杂度为 $m = m_A + m_B$。那么存在一个 $(\Xi, \textsc{far}^{\mathsf{TV}}_{\epsilon}(\Xi))$—带标签分布测试器，其样本复杂度为 $m + O(1/\epsilon)$。

证明。对于输入 $(f, \mathcal{D})$，测试器执行以下步骤。

1. 使用 m_A 个样本运行适当学习器 A，并得到一个输出函数 g。
2. 从 $\mathcal{D}$ 中使用 $O(1/\epsilon)$ 个样本，计算 $\mathbb{P}_{x \sim \mathcal{D}}[f(x) \neq g(x)$ 的估计值 z，若该值大于 $\frac{3}{8}\epsilon$，则拒绝。
3. 使用 m_B 个样本运行分布测试器 B_g。

假设 $(f, \mathcal{D}) \in \Xi$。假设算法 A 成功，这一事件发生的概率至少为 $1 - \delta/3$。那么存在 $\mathcal{E}$ 使得 $\mathcal{E}_g \in \Xi$，并且

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_g) \leq \epsilon/4, \qquad \text{and} \qquad \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{E}) \leq \epsilon/4,$$

根据算法 A的条件（它是一个适当学习器），由乘法型切尔诺夫界可知，使用 $O(1/\epsilon)$个样本后，$z < \frac{3}{8}\epsilon$ 以至少 $1 - \delta/3$的概率成立，因此第二步通过。最后，假设算法 B_g 成功，其发生概率为 $1 - \delta/3$。由于 $\mathcal{E}_g \in \Xi$，且 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{E}) \leq \epsilon/4$，我们有

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}, \Pi_g) \leq \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{E}) \leq \epsilon/4.$$

因此$\mathcal{D} \in \textsc{close}^{\mathsf{TV}}_{\epsilon/4}(\Pi_g)$，算法 B_g 将接受。这些步骤中任一步骤失败的概率至多为 δ，由并集界可得。

现在假设$(f, \mathcal{D}) \in \textsc{far}^{\mathsf{TV}}_{\epsilon}(\Xi)$。反设

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}, \Pi_g) < \epsilon/2 \qquad \text{and} \qquad \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_g) < \epsilon/2.$$

那么存在 $\mathcal{E} \in \Pi_g$ 使得 $\mathcal{E}_g \in \Xi$ 且 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{E}) < \epsilon/2$。接着

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g) \leq \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_g) + \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_g, \mathcal{E}_g) = \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_g) + \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{E}) < \epsilon,$$

这就产生了矛盾。因此，必定是 $\mathrm{dist}_{\mathsf{TV}}(\mathcal{D}, \Pi_g) \geq \epsilon/2$，此时第三步以至少 $1-\delta/3$ 的概率拒绝；或者 $\mathrm{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_g) \geq \epsilon/2$，此时第二步以至少 $1-\delta/3$ 的概率拒绝。这里再次使用乘法型切尔诺夫界。

□

注记 6.19。当 Ξ 是通过选取函数 Λ_1 的性质 $\mathcal{X} \to \{0,1\}$，令 Λ_2 为所有定义在 $\mathcal{X}$上的分布的集合，并令 $\Xi = \Lambda_1 \times \Lambda_2$时，这种形式化刻画了标准的通过学习进行测试的归约。在这种情况下，学习器 A 就是标准 PAC 学习算法（见引理 6.35的证明），而验证器B 则直接接受所有输入。

6.5 密度性质与奇偶迹下的分布测试

带标签分布测试是对无分布假设、基于样本的性质测试的一种更一般的重新表述，它允许定义更丰富的性质类。在第 6 节的其余部分，我们关注一类特定的带标签分布性质，称为密度性质。

定义 6.20 （密度性质）。适当的带标签分布的性质 Ξ 称为密度性质，如果存在一个 Π 概率分布在 $\mathbb{N}$ 上的集合，使得 $\Pi = \Pi(\Xi)$；也就是说，对于任意适当的带标签分布（f，$\mathcal{D}$），当且仅当$\pi_{f,\ \mathcal{D}} \in \Pi$时，($f$，$\mathcal{D}$) $\in \Xi$。

在本小节中，我们旨在确立密度性质测试与奇偶迹下的分布测试之间的关系，这些关系如图 1所示。这里我们证明 $\to$ 关系；展示 $\not\to$ 关系的示例见附录 E。

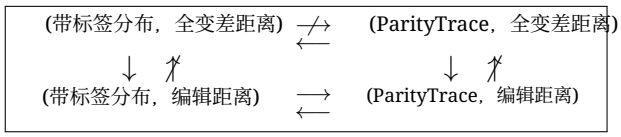


图 1：测试模型与距离度量相对强度的总结。箭头（$\mathcal{M}_1$，d_1）$\to$（$\mathcal{M}_2$，d_2）表示，在模型 $\mathcal{M}_1$ 下针对距离 d_1 的测试器意味着存在一个在模型 $\mathcal{M}_2$ 下针对距离 d_2的测试器，而 $\not\to$ 表示这种蕴含通常并不成立。

引理 6.21。若 t Ξ 是任意密度性质，且令 $\epsilon > 0$。则 $\Pi(\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)) \subseteq \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Pi(\Xi))$。

证明。设 $\pi \in \Pi(\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi))$，因此 $\pi = \pi_{f,\ \mathcal{D}}$ 对于某个 $\mathcal{D}_f \in \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)$。反设 $\mathrm{dist}_{\mathsf{TV}}(\pi,\ \pi')$ $\leq \epsilon$ 对于某个 $\pi' = \pi_{g,\ \mathcal{E}} \in \Pi(\Xi)$。令 $a_1 < a_2 < a_3 < \cdots$为 f 的交替序列，并定义 $a_0 = -\infty$。然后我们如下定义带标签分布 $\mathcal{F}_f$。下文中，不妨设 $\mathcal{D}_f$ 是 1-适当的；若其为 0-适当的，则需要将某些指标调整 1。

对于每个 $i \in \mathbb{N}$，令 $\delta_i := \pi'(i) - \sum_{x \in (a_{i-1},\ a_i]} \mathcal{D}(x)$。如果 $\delta_i \geq 0$，对于 $x \in (a_{i-1},\ a_i]$，我们可以通过设定 $\mathcal{F}(z) = \mathcal{D}(z) + \delta$，得到区间 $(a_{i-1},\ a_i]$ 上的 $\mathcal{F}$：任意选取一个元素 $z \in (a_{i-1},\ a_i]$，然后为该区间内其余的 $x \neq z$ 选择 $\mathcal{F}(x) = \mathcal{D}(x)$。

如果 $\delta_i < 0$，我们可以通过从区间内的密度 $\mathcal{D}(x)$ 中总共减去 $|\delta_i|$，得到区间 $(a_{i-1},\ a_i]$ 上的 $\mathcal{F}$；这是可行的，因为必有 $\sum_{a_{i-1} < x \leq a_i} \mathcal{D}(x) \geq |\delta_i|$。

容易验证，该构造满足 $\pi_{f,\ \mathcal{F}} = \pi' = \pi_{g,\ \mathcal{E}}$。由于 Ξ 是一种密度性质，

因此有 $\mathcal{F}_f \in \Xi$。我们可以看到

$$\begin{aligned}\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{F}_f) &= \frac{1}{2}\sum_x |\mathcal{D}(x) - \mathcal{F}(x)| = \frac{1}{2}\sum_i \left(\sum_{a_{i-1}<x\le a_i} |\mathcal{D}(x) - \mathcal{F}(x)| \right) \\ &= \frac{1}{2}\sum_i |\delta_i| = \mathsf{dist}_{\mathsf{TV}}(\pi, \pi') \le \epsilon\,.\end{aligned}$$

但这与假设 $\mathcal{D}_f \in \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)$ 矛盾。因此 $\Pi(\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)) \subseteq \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Pi(\Xi))$。 □

密度性质在布尔运算下封闭；特别地，我们需要下面这个事实，它可由定义直接推出。

事实 6.22。 设 Ξ_1 和 Ξ_2 是密度性质。那么 $\Xi_1 \cap \Xi_2$ 是一个密度性质 .

我们将证明，对于密度性质而言，带标签分布测试与奇偶迹下的分布测试本质上等价。首先要证明，针对密度性质的带标签分布测试器总能转换为一种受限类型的测试器：它忽略样本点的绝对位置，只保留其标签和顺序。这个证明的思路源自 [DKN15a]。我们需要一些记号。

对于任意大小为 m 的多重集 $S \subseteq \mathbb{Z}$，记作 $S = \{x_1, x_2, \cdots, x_m\}$，其中我们令 $x_1 \le x_2 \le \cdots \le x_m$。那么对于任意 $f\colon\ \mathbb{Z} \to \{0,1\}$，我们将 S 中由 f 标记的点按顺序组成的有序序列写作

$$S_f := ((x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_m, f(x_m)))\ .$$

于是我们定义

$$\mathsf{trace}^*(S_f) := (f(x_1), f(x_2), \ldots, f(x_k))\,.$$

事实 6.23。 设 $(f, \mathcal{D})$ 是任意适当的带标签分布。对于任意 m，令 $T \sim$ 采样$(\pi_{f,\mathcal{D}}, m)$ 和$S \sim$ 采样$(\mathcal{D}, m)$。那么，随机变量 迹$^*(S_f)$ 与 迹(T) 同分布。

引理 6.24。 设 Ξ_1 和 Ξ_2 是任意密度性质，令 $\alpha \in (0,1)$，并假设存在一个算法 A 以及一个数 m，使得：

1. *If* $\mathcal{D}_f \in \Xi_1$ *then* $\mathop{\mathbb{P}}_{S\sim\mathsf{samp}(\mathcal{D},m)}[A(S_f)\ \textit{accepts}\ ] \ge \alpha$*; and,*
2. *If* $\mathcal{D}_f \in \Xi_2$ *then* $\mathop{\mathbb{P}}_{S\sim\mathsf{samp}(\mathcal{D},m)}[A(S_f)\ \textit{rejects}\ ] \ge \alpha$*.*

那么，对于任意 $\delta > 0$，存在一个算法 A' 满足

1. *If* $\mathcal{D}_f \in \Xi_1$ *then* $\mathop{\mathbb{P}}_{S\sim\mathsf{samp}(\mathcal{D},m)}[A'(\mathsf{trace}^*(S_f))\ \textit{accepts}\ ] \ge \alpha - \delta$*; and,*
2. *If* $\mathcal{D}_f \in \Xi_2$ *then* $\mathop{\mathbb{P}}_{S\sim\mathsf{samp}(\mathcal{D},m)}[A'(\mathsf{trace}^*(S_f))\ \textit{rejects}\ ] \ge \alpha - \delta$*.*

为了证明这个引理，我们需要用到无限 Ramsey 定理。对于任意集合 X 和 $n \in \mathbb{N}$，令$\binom{X}{n}$表示 X的大小为 n 的子集所构成的集合。

定理 6.25 (无限 Ramsey 定理 [Ram09])。 固定任意 c，$n \in \mathbb{N}$ 并令 X 为任意可数无限集。对于 $\binom{X}{n}$ 的任意染色 c，其中使用 $Y \subseteq X$ 种颜色，存在一个无限集 $\binom{Y}{n}$，使得其中的集合具有相同的颜色。

现在我们可以证明引理 6.24。

引理 *6.24* 的证明。 由于 Ξ_1和Ξ_2 是密度性质，它们分别有相关的密度序列集合 $\Pi(\Xi_1)$ 和 $\Pi(\Xi_2)$。

对于输入 $(f, \mathcal{D})$，算法 A 接收 S_f，其中 $S \sim \mathsf{samp}(\mathcal{D}, m)$。对于每个大小为 m 的多重集$S \subseteq \mathbb{N}$，算法对 S_f 的判定可以写作 $A(S_f) = A_S(\text{迹}^*(S_f))$，其中$A_S\colon \{0,1\}^m \to \{0,1\}$。可能的判定函数至多有 $b = 2^{2^m}$ 个。我们将每个可能的判定函数 $\Delta\colon \{0,1\}^m \to \{0,1\}$ 与 $[b]$ 中的一个元素对应起来。

对于每个大小为 $S \subset \mathbb{N}$ 的子集 m，我们用函数 S 进行着色 A_S，并将其对应到 $[b]$中的一个元素。根据 定理 6.25，存在一个无限集 $N \subset \mathbb{N}$，使得所有大小为 $S \subset \mathbb{N}$ 的 m 具有相同颜色。于是存在一个判定函数 $\Delta\colon \{0,1\}^m \to \{0,1\}$ 使得对于每个 $S \in \binom{N}{m}$，$A_S = \Delta$。

现在定义算法 A'如下。对于输入迹$^*(S_f) = (f(x_1)\ f(x_2) \ldots, f(x_m))$，$A'$只需输出 $\Delta(f(x_1)\ f(x_2)\ldots, f(x_m))$。剩下只需证明该算法是正确的。

任取输入 $(f, \mathcal{D})$；不失一般性，假设 $(f, \mathcal{D}) \in \Xi_1$，因为对 Ξ_2 也有类似的论证。令 $a_1 < a_2 < a_3 < \cdots$为 f 的交替序列。我们还假设 $(f, \mathcal{D})$ 是 1-适当的，因为当 $(f, \mathcal{D})$ 是 0-适当时也有类似的论证（区别在于，我们将有 $\pi_{f,\mathcal{D}}(1) = 0$ 和 $\pi_{f,\mathcal{D}}(i+1) = \mathcal{D}(a_{i-1}, a_i]$，而不是$\pi_{f,\mathcal{D}}(i) = \mathcal{D}(a_{i-1}, a_i])$。

选择 $C = \lceil m^2/\delta \rceil$。由于 N 是无限的，我们可以选取一个单射 $\gamma\colon \mathbb{N} \to N \cup \{\infty\}$，满足

$$1 < \gamma(a_1) < \gamma(a_2) < \cdots < \gamma(a_k) < \cdots,$$

为方便记号，定义 $\gamma(a_0) = 1$。我们可以选择 γ 使其满足 $|N \cap (\gamma(a_{i-1})\ \gamma(a_i)]| \geq C$对每个 $i \geq 1$都成立。定义一个分布 $\mathcal{D}'$，为前 C个元素 $x \in N \cap (\gamma(a_{i-1})\ \gamma(a_i)]$,赋值 $\mathcal{D}'(x) = \frac{1}{C} \cdot \mathcal{D}(a_{i-1},\ a_i]$，并定义函数 $f'\colon \mathbb{Z} \to \{0,1\}$，使其成为交替序列为 $\gamma(a_1) < \gamma(a_2) < \cdots$的唯一函数。注意，对每个 $i \in \mathbb{N}$,

$$\pi_{f',\mathcal{D}'}(i) = \mathcal{D}(\gamma(a_{i-1}), \gamma(a_i)] = C \cdot \frac{1}{C}\mathcal{D}(a_{i-1}, a_i] = \pi_{f,\mathcal{D}}(i)\,,$$

因此 $(f',\ \mathcal{D}') \in \Xi_1$，因为 Ξ_1是一个密度性质。所以

$$\alpha \leq \mathop{\mathbb{P}}_{S' \sim \mathsf{samp}(\mathcal{D}', m)} \left[A_{S'}(\mathsf{trace}^*(S'_{f'}))\ \text{accepts}\ \right].$$

我们有 $\mathsf{supp}(\mathcal{D}') \subseteq N$，因此 $S' \subseteq N$以概率 1 成立。所以，如果 S'是一个大小为 m的集合（即多重集 S 中的每个元素都以重数 1 出现），那么 $A_{S'} = \Delta$。令 $F(S')$ 表示 S' 是一个大小为 m的集合这一事件。由于 N 中的每个元素的密度至多为 $1/C$，由并集界可得

$$\mathop{\mathbb{P}}_{S'} \left[\neg F(S')\right] < \frac{m^2}{C} \leq \delta\,.$$

Then

$$\begin{aligned}\alpha \leq \mathop{\mathbb{P}}_{S'} \left[A_{S'}(\mathsf{trace}^*(S'_{f'}))\ \text{accepts}\ \right] &\leq \mathop{\mathbb{P}}_{S'} \left[\Delta(\mathsf{trace}^*(S'_{f'}))\ \text{accepts}\ \right] + \mathop{\mathbb{P}}_{S'} \left[\neg F(S')\right] \\ &< \mathop{\mathbb{P}}_{S'} \left[\Delta(\mathsf{trace}^*(S'_{f'}))\ \text{accepts}\ \right] + \delta\,.\end{aligned}$$

现在注意，对于 $S \sim$采样$(\mathcal{D},\ m)$，变量迹$^*(S'_{f'})$ 和 迹$^*(S_f)$ 同分布。因此我们有

$$\alpha - \delta < \mathop{\mathbb{P}}_{S'} \left[\Delta(\mathsf{trace}^*(S'_{f'}))\ \text{accepts}\ \right] = \mathop{\mathbb{P}}_{S} \left[\Delta(\mathsf{trace}^*(S_f))\ \text{accepts}\ \right].$$

由于我们可以对 Ξ_2重复类似的论证，证明至此完成。 □

现在我们可以证明带标签分布测试与奇偶迹下的分布测试之间的一般等价性。（注意，下面引理的第二部分不要求 Ξ_1 和 Ξ_2 是密度性质。）

引理 6.26. 设 Ξ_1 和 Ξ_2 是带标签分布的性质。则：

1. 假设 Ξ_1 和 Ξ_2 是密度性质。如果存在一个(Ξ_1, Ξ_2)-带标签分布测试器，其样本复杂度为 m，成功概率为 α，那么对于任意 $\delta > 0$，都存在一个$(\Pi(\Xi_1)\ \Pi(\Xi_2))$-分布测试器，可在奇偶迹下进行测试，其样本复杂度为 m，成功概率为 $\alpha - \delta$。

2. 如果存在一个在奇偶迹下的$(\Pi(\Xi_1)\ \Pi(\Xi_2))$-分布测试器，其样本复杂度为m，成功概率为 α，那么就存在一个(Ξ_1, Ξ_2)-带标签分布测试器，其样本复杂度为 m，成功概率为 α。

证明。 假设存在一个（Ξ_1，Ξ_2，α）带标签分布测试器 A，其样本复杂度为 m。根据引理 6.24，对于任意 $\delta > 0$，存在一个测试器 A' 使得

$$\mathcal{D}_f \in \Xi_1 \implies \mathop{\mathbb{P}}_{S \sim \mathsf{samp}(\mathcal{D}, m)} \left[A'(\mathsf{trace}^*(S_f)) \text{ accepts } \right] > \alpha - \delta$$
$$\mathcal{D}_f \in \Xi_2 \implies \mathop{\mathbb{P}}_{S \sim \mathsf{samp}(\mathcal{D}, m)} \left[A'(\mathsf{trace}^*(S_f)) \text{ rejects } \right] > \alpha - \delta\,.$$

假设 $\pi \in \Pi(\Xi_1)$。那么存在 $\mathcal{D}_f \in \Xi_1$ 使得 $\pi = \pi_{f,\,\mathcal{D}}$。利用以下事实：当 迹$(T)$ 与 迹$^*(S_f)$ 同分布时，$T \sim$ 采样$(\pi,\ m)$ 与 $S \sim$ 采样$(\mathcal{D},\ m)$（事实 6.23），我们有

$$\mathop{\mathbb{P}}_{T \sim \mathsf{samp}(\pi, m)} \left[A'(\mathsf{trace}(T)) \text{ accepts } \right] = \mathop{\mathbb{P}}_{S \sim \mathsf{samp}(\mathcal{D}, m)} \left[A'(\mathsf{trace}^*(S_f)) \text{ accepts } \right] > \alpha - \delta\,.$$

类似的 论证在 $\pi \in \Pi(\Xi_2)$ 的情形下也成立，由此完成证明的第一部分。

现在假设存在一个在奇偶迹下的（$\Pi(\Xi_1)\ \Pi(\Xi_2)$）分布测试器，其样本复杂度为 m，成功概率为 α，因此存在一个算法 A 使得

$$\pi \in \Pi(\Xi_1) \implies \mathop{\mathbb{P}}_{T \sim \mathsf{samp}(\pi, m)} \left[A(T) \text{ accepts } \right] > \alpha$$
$$\pi \in \Pi(\Xi_2) \implies \mathop{\mathbb{P}}_{T \sim \mathsf{samp}(\pi, m)} \left[A(T) \text{ rejects } \right] > \alpha\,.$$

假设 $\mathcal{D}_f \in \Xi_1$。那么 $\pi_{f,\,\mathcal{D}} \in \Pi(\Xi_1)$。再次利用以下事实：当 迹$^*(S_f)$ 与 迹(T) 同分布时，$S \sim$ 采样$(\mathcal{D},\ m)$ 与 $T \sim$ 采样$(\pi_{f,\,\mathcal{D}},\ m)$（事实 6.23），我们有

$$\mathop{\mathbb{P}}_{S \sim \mathsf{samp}(\mathcal{D}, m)} \left[A(\mathsf{trace}^*(S_f)) \text{ accepts } \right] = \mathop{\mathbb{P}}_{T \sim \mathsf{samp}(\pi, m)} \left[A(\mathsf{trace}(T)) \text{ accepts } \right] > \alpha\,.$$

当 $\mathcal{D}_f \in \Xi_2$ 时，类似的论证同样成立，由此证明结束。 □

为了将距离度量引入等价关系，我们需要：

命题 6.27。 设 Ξ 是任意一种密度性质，且设 $\epsilon > 0$。那么 $\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi)$ 是一种密度性质，并且

$$\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi)) = \Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi))\,.$$

证明。 显然， $\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi)$ 是一种密度性质，因为对于任意带标签分布 $\mathcal{D}_f$ 和 $\mathcal{E}_g$，$\mathsf{dist}_{\mathsf{edit}}((f,\mathcal{D}),(g,\mathcal{E}))$ $=\mathsf{dist}_{\mathsf{edit}}(\pi_{f,\mathcal{D}},\pi_{g,\mathcal{E}})$，所以 $\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi)$ 完全由密度序列定义。

我们首先证明 $\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi)) \subseteq \Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi))$。令 $\pi \in \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi))$，使得对于所有 $\pi' \in \Pi(\Xi)$，$\mathsf{dist}_{\mathsf{edit}}$ $(\pi,\pi') > \epsilon$。反设 $\pi \notin \Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi))$。令 $(f,\mathcal{D})$ 为满足 $\pi_{f,\mathcal{D}} = \pi$ 的任意带标签分布。则 $\mathsf{dist}_{\mathsf{edit}}((f,\mathcal{D})\ \Xi) \le \epsilon$，因此存在 $(g,\mathcal{E}) \in \Xi$，以及满足 $\pi_{f,\mathcal{D}} = \pi_{f',\mathcal{D}'}$ 的 $\mathcal{D}'_{f'}$，使得 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_{f'},\mathcal{E}_g) \le \epsilon$。但随后

$$\mathsf{dist}_{\mathsf{edit}}(\pi,\Pi(\Xi)) \le \mathsf{dist}_{\mathsf{edit}}(\pi,\pi_{g,\mathcal{E}}) = \mathsf{dist}_{\mathsf{edit}}(\pi_{f',\mathcal{D}'},\pi_{g,\mathcal{E}}) \le \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_{f'},\mathcal{E}_g) \le \epsilon\,,$$

这就产生了矛盾。由此可得 $\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi)) \subseteq \Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi))$。

接下来，我们证明 $\Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi)) \subseteq \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi))$。令 $\pi \in \Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi))$，使得 $\pi = \pi_{f,\ \mathcal{D}}$ 对于某个满足 $\mathsf{dist}_{\mathsf{edit}}((f,\ \mathcal{D})\ \Xi) > \epsilon$ 的 $(f,\ \mathcal{D})$。反设 $\pi \notin \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi))$，则存在 $\pi' \in \Pi(\Xi)$ 使得 $\mathsf{dist}_{\mathsf{edit}}(\pi,\ \pi') \le \epsilon$。于是 $\pi' = \pi_{g,\ \mathcal{E}}$对于某个 $(g,\ \mathcal{E}) \in \Xi$，因此

$$\mathsf{dist}_{\mathsf{edit}}((f,\mathcal{D}),\Xi) \le \mathsf{dist}_{\mathsf{edit}}((f,\mathcal{D}),(g,\mathcal{E})) = \mathsf{dist}_{\mathsf{edit}}(\pi,\pi') \le \epsilon\,,$$

这与假设矛盾。至此，证明完成。 □

我们可以现在确立从图 1 得到箭头（带标签分布，编辑距离）$\to$（ParityTrace，编辑距离）图 1 .

引理 6.28。 设 Ξ 是任意密度性质，并假设存在一个样本复杂度为 m 的 $(\Xi,\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi)\ \alpha)$-带标签分布测试器。那么对于任意 $\delta > 0$，都存在一个在奇偶迹下样本复杂度为 m 的 $(\Pi(\Xi)\ \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi))$ $\alpha-\delta)$-分布测试器。

证明。根据命题 6.27， $\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi)$ 是一种密度性质。因此，引理 6.26 给出了一个在奇偶迹下的 $(\Pi(\Xi)\ \Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi))\ \alpha-\delta)$ 分布测试器，适用于任意 $\delta > 0$，其样本复杂度为 m。根据命题 6.27，我们得到一个在奇偶迹下的 $(\Pi(\Xi)\ \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi))\ \alpha-\delta)$ 分布测试器。 □

下面这个简单事实给出了图 1 中（带标签分布，全变差距离）$\to$ (带标签分布，编辑距离）之间的关系。图 1。

事实 6.29。 令 Ξ 为（适当的）带标签分布上的任意性质，并令 $\epsilon > 0$。则 $\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi)\ \subseteq\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)$。

证明。 这可由不等式 $\mathsf{dist}_{\mathsf{edit}}((f,\ \mathcal{D}),\ (g,\ \mathcal{E})) \le \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\ \mathcal{E}_g)$ 得出，该不等式对任意两个适当的带标签分布 $\mathcal{D}_f$ 和 $\mathcal{E}_g$均成立。 □

现在我们陈述一个便于后续使用的引理。

引理 6.30. 令 Ξ为任意密度性质，并假设存在一个$(\Xi,\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)\ \alpha)$带标签分布测试器，其样本复杂度为 m。那么，对于任意 $\delta > 0$，存在一个$(\Pi(\Xi)\ \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi))\ \alpha-\delta)$-奇偶迹下的分布测试器，其样本复杂度为 m。

箭头 (ParityTrace, 编辑距离) $\to$ (带标签分布, 编辑距离) 的证明如下。假设我们有一个 $(\Pi,\mathrm{FAR}_\epsilon^{\mathsf{edit}}$ $(\Pi)\ \alpha)$-奇偶迹下的分布测试器。令 $\Xi = \Xi(\Pi)$ 为相应的密度性质，因此 $\Pi = \Pi(\Xi)$。由命题 6.27，$\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi)) = \Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi))$，因此由引理 6.26 可得一个 $(\Xi,\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi)\ \alpha)$-带标签分布测试器。

映射 (ParityTrace，TV) $\to$(带标签分布，TV) 是类似的。假设我们有一个在奇偶迹下的 $(\Pi,$ $\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Pi)\ \alpha)$-分布测试器。令 $\Xi = \Xi(\Pi)$ 为对应的密度性质，那么

满足 $\Pi = \Pi(\Xi)$。由引理 6.21， $\Pi(\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)) \subseteq \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Pi(\Xi))$，因此我们得到一个在奇偶迹下的 $(\Pi(\Xi)\ \Pi(\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi))\ \alpha)$-分布测试器。由引理 6.26，我们有一个 $(\Xi, \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)\ \alpha)$-带标签分布测试器。

6.6 均匀测试 k-交替函数

现在，我们证明带标签分布测试的主要结果。为方便起见，下面重述该结果；它是我们的主要定理 1.1的一个推论。首先，我们注意到，当其中一个分布是均匀分布时，编辑距离与全变差距离相等；证明留待附录 D.4给出。

引理 6.31。 存在一个绝对常数 $c > 0$，使得以下结论成立。设 π 是 $\mathbb{N}$ 上的分布，且在 $[k]$ 上均匀支撑；设 π' 是 $\mathbb{N}$ 上的另一个概率分布，其支撑集包含于 $[k]$。那么 $\mathsf{dist}_{\mathsf{edit}}(\pi, \pi') \geq c \cdot \mathsf{dist}_{\mathsf{TV}}(\pi, \pi')$。

注 6.32。 我们的主要定理 1.1 的表述并未排除这样一种可能：针对奇偶迹、以编辑距离为度量的均匀性检验器，能够突破该定理的下界。这是因为编辑距离测试器弱于全变差距离测试器，这是由不等式（19）导致的。上述引理表明，我们可以加强定理 1.1中的下界，使其同样适用于编辑距离测试器。

定理 6.33。 令 Ξ_1 为均匀 $2k$ 交替的带标签分布，令 Ξ_2 为与 Ξ_1 的全变差距离至少为 ϵ、且 $2k$ 交替的带标签分布。则 $(\Xi_1$、Ξ_2、$2/3)$-带标签分布测试的样本复杂度为 $\widetilde{\Theta}\left(\frac{k^{4/5}}{\epsilon^{4/5}} + \frac{\sqrt{k}}{\epsilon^2}\right)$。

上界的证明。令 K 为 $2k$ 交替带标签分布的性质，这是一个密度性质，其中 $\Pi(K)$ 表示支撑于 $[2k]$ 上的密度序列的性质。则 $\Xi_2 = \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi_1) \cap K$，且 $\Pi(\Xi_2) = \Pi(\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi_1) \cap K) = \Pi(\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi_1)) \cap \Pi(K)$；而 $\Pi(\Xi_1)$ 仅包含支撑于 $[2k]$ 上的均匀分布 μ。根据引理 6.21， $\Pi(\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi_1)) \subseteq \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Pi(\Xi_1))$。因此，在奇偶迹下，具有样本复杂度 m 的 $(\Pi(\Xi_1)\ \mathrm{FAR}_\epsilon^{\mathsf{TV}}\Pi(\Xi_1) \cap \Pi(K)\ \alpha)$-分布测试器，也是一个在奇偶迹下、样本复杂度为 m 的 $(\Pi(\Xi_1)\ \Pi\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi_1) \cap K)\ \alpha)$-分布测试器。结论由引理 6.26 得出。 □

下界的证明。我们从引理 6.28 中论证的一种专门变体开始。假设存在一个 $(\Xi_1, \Xi_2, 2/3)$-带标签分布测试器，其样本复杂度为 m，并回顾 $\Xi_2 = \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi_1) \cap K$。根据事实 6.29，我们有 $\mathrm{FAR}^{\mathsf{edit}}(\Xi_1) \subseteq \mathrm{FAR}^{\mathsf{TV}}(\Xi_1)$，因此这也是一个 $(\Xi_1, \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi_1) \cap K, 2/3)$-带标签分布测试器。根据命题 6.27，$\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi_1)$ 是一个密度性质，因此根据事实 6.22， $\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi_1) \cap K$ 也是一个密度性质。由引理 6.26，我们进而得到一个在奇偶迹下的 $(\Pi(\Xi_1)\ \Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi_1) \cap K)\ 2/3 - \delta)$-分布测试器，对任意 $\delta > 0$均成立，其样本复杂度为 m。注意到 $\Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi_1) \cap K) = \Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi_1)) \cap \Pi(K)$。根据命题 6.27，我们得到一个在奇偶迹下的 $(\Pi(\Xi_1)\ \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi_1)) \cap \Pi(K)\ 2/3 - \delta)$-分布测试器。

注意， $\Pi(\Xi_1) = \{\mu\}$，其中 μ 是支撑于 $[2k]$上的均匀分布。令 $0 < c < 1$为 引理 6.31中的常数，并考虑任意分布 $\pi \in \mathrm{FAR}_{\epsilon/c}^{\mathsf{TV}}(\Pi(\Xi_1))$。于是，根据引理 6.31，有

$$\mathsf{dist}_{\mathsf{edit}}(\pi, \mu) \geq c \cdot \mathsf{dist}_{\mathsf{TV}}(\pi, \mu) > c \cdot (\epsilon/c) = \epsilon\,,$$

因此 $\pi \in \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi_1))$。那么 $\mathrm{FAR}_{\epsilon/c}^{\mathsf{TV}}(\Pi(X_1)) \subseteq \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi(\Xi_1))$。

因此，我们的测试器也是一个$\Pi(\Xi_1)$ $\mathrm{FAR}_{\epsilon/c}^{\mathsf{TV}}(\Pi(\Xi_1)) \cap \Pi(K)$ $2/3-\delta$-奇偶迹下的分布测试器，其样本复杂度为 m。根据定理 5.1，我们必然得到所需的下界

$$m = \widetilde{\Omega}\left(\left(\frac{n}{\epsilon}\right)^{4/5} + \frac{\sqrt{n}}{\epsilon^2}\right). \qquad \square$$

6.7 无承诺测试 k-交替函数和均匀 k-交替函数

定理 1.1 证明了在奇偶迹下，测试支撑于 $[k]$ 上的分布是否为均匀分布的紧确界。我们使用通过学习进行测试，证明了一个更难问题的界：测试一个分布是否在 $[k]$上服从均匀分布，而无需承诺输入支撑于 $[k]$。

定理 6.34。 固定定义域 $\mathbb{N}$。令 Π 仅包含 $[k]$上的均匀分布。在奇偶迹下，存在一个$(\Pi, \mathrm{FAR}_{\epsilon}^{\mathsf{edit}}(\Pi)$ $2/3)$-分布测试器，其样本复杂度为 $O\left(\frac{k}{\epsilon} + {}_{\epsilon^2}{}^{k}\log k\right)$。 – ——

由下一个引理即可推出这一点，利用 $\Pi = \Pi(\Xi)$，其中 Ξ 是均匀 k-交替带标签分布的性质，再结合 引理 6.30。

引理 6.35。 令 Ξ 为均匀 k-交替函数。则存在一个 $(\Xi, \mathrm{FAR}_{\epsilon}^{\mathsf{TV}}(\Xi)$ $2/3)$ 带标签分布测试器，其样本复杂度为 $O\left(\frac{k}{\epsilon} + {}_{\epsilon^2}{}^{k}\log k\right)$。 – ——

证明。 令 $c \in (0,1)$ 为 引理 6.31中的普适常数。我们将构造一个学习器-验证器对。令 A 为针对 k-交替函数类、误差为 $c\epsilon/4$的标准 PAC 学习算法。该算法使用大小为 $m_A = O(k/\epsilon$的样本，输出一个 k-交替函数g： $\mathbb{Z} \to \{0,1\}$，且以至少 $8/9$ 的概率，

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_g) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}[f(x) \neq g(x)] < c\epsilon/4\,,$$

其中等式由 事实 D.1 得出。显然，存在一个分布 $\mathcal{E}$ 使得 $\mathcal{E}_g \in \Xi$。

假设（f， $\mathcal{D}$） $\in \Xi$。我们必须证明存在一个分布 $\mathcal{E}$ 使得 $\mathcal{E}_g \in \Xi$并且$\mathsf{dist}_{\mathsf{TV}}$（$\mathcal{D}$， $\mathcal{E}$） $\leq \epsilon/4$。令 μ 为 $[k+1]$上的均匀分布。根据命题 D.2，我们得到 $\mathcal{E}$ 使得

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_g, \mathcal{E}_g) = \mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{D}}, \pi_{g,\mathcal{E}}) = \mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{D}}, \mu)\,.$$

根据$\pi_{g、}$ $\mathcal{E} = \mu = \pi_{f、}$ $\mathcal{D}$ 以及引理 6.31，我们有

$$\mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{D}}, \pi_{g,\mathcal{E}}) = \mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{D}}, \pi_{f,\mathcal{D}}) \leq \frac{1}{c}\cdot \mathsf{dist}_{\mathsf{edit}}(\pi_{g,\mathcal{D}}, \pi_{f,\mathcal{D}}) \leq \frac{1}{c}\cdot \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_g, \mathcal{D}_f)\,,$$

其中，最后一个不等式根据定义成立。于是

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{E}) = \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_g, \mathcal{E}_g) \leq \frac{1}{c}\cdot \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_g, \mathcal{D}_f) \leq \epsilon/4\,.$$

因此，算法 A 满足学习器-验证器对的条件。接下来需要构造验证器 B。令 g： $\mathbb{Z} \to \{0,1\}$ 为学习器的任意可能输出，它必须是一个 k-交替函数。于是 Π_g 是所有满足 $\pi_{g,}$ $\mathcal{E}$ $=$ μ的分布 $\mathcal{E}$ 的集合，其中 μ 是 $[k+1]$上的均匀分布。将算法 B_g 定义如下：

1. 采样 $S \sim \mathsf{samp}(\mathcal{D},\ m_B)$，并通过对每个 $x \in S$ 取样，将数字 $i \in [k+1]$ 加入 S'来构造多重集 S'，其中 i 是包含 x的唯一区间（$a_{i-1},\ a_i$]。于是S' 服从 $\mathsf{samp}(\pi_{g,\ \mathcal{D}},\ m_B)$。

2. 在样本 $\epsilon/4, \epsilon/2$ 上使用一个容错均匀性检验器，以检验 S' 是否满足 $\mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{D}}, \mu) < \epsilon/4$ 或 $\mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{D}}, \mu) > \epsilon/2$。此步骤需要 $m_B = O\left(\epsilon^{-2} k \log k\right)$ 个样本 [VV17b]。

假设 $\mathcal{E} \in \textsc{close}^{\mathsf{TV}}_{\epsilon/4}(\Pi_g)$，则存在 $\mathcal{F}$ 使得 $\pi_{g,\mathcal{F}} = \mu$ 且 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{E}_g, \mathcal{F}_g) < \epsilon/4$。于是

$$\mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{E}}, \mu) = \mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{E}}, \pi_{g,\mathcal{F}}) \le \mathsf{dist}_{\mathsf{TV}}(\mathcal{E}_g, \mathcal{F}_g) < \epsilon/4\,,$$

因此，容错均匀性检验器将接受。

现在假设 $\mathcal{E} \in \textsc{far}^{\mathsf{TV}}_{\epsilon/2}(\Pi_g)$。为导出矛盾，假设

$$\mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{E}}, \mu) \le \epsilon/2\,.$$

根据 命题 D.2，我们得到 $\mathcal{F}$ 使得 $\pi_{g,\mathcal{F}} = \mu$ （因此 $\mathcal{F} \in \Pi_g$），并且

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{E}_g, \mathcal{F}_g) = \mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{E}}, \pi_{g,\mathcal{F}}) = \mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{E}}, \mu) \le \epsilon/2\,.$$

这与 $\mathcal{E} \in \textsc{far}^{\mathsf{TV}}_{\epsilon/2}(\Pi_g)$ 矛盾，因此必定有 $\mathsf{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{E}}, \mu) > \epsilon/2$。那么容错均匀性检验器将正确地拒绝。

□

我们还注意到，由于等价于分布无关的基于样本的检验，检验 k-交替带标签分布的无承诺问题继承了上界 $O(k/\epsilon)$:

引理 6.36。令 Ξ 为满足以下条件的带标签分布集合：$(f, \mathcal{D})$，其中 f 是一个 k-交替函数，且 $\mathcal{D}$ 是 $\mathbb{Z}$ 上的任意分布。令 $m(k, \epsilon)$ 为用于测试 k-交替函数的无分布假设的基于样本的测试器的最优样本复杂度。那么，对$(\Xi, \textsc{far}^{\mathsf{TV}}_{\epsilon}(\Xi))$-带标签分布进行测试的最优样本复杂度为 $\Theta(m(k, \epsilon))$。特别地，存在一个这样的测试器，其样本复杂度为 $O(k/\epsilon)$。

证明。该陈述的第一部分由 命题 6.15 得出。第二部分由标准 PAC 学习理论得出，因为 k-交替函数类的 VC 维数为 $k+1$，再结合通过学习进行检验的归约 [GGR98]。

□

与前一个结果一样，这个引理为奇偶迹下分布的支持集大小 测试给出了一个界；这也是我们接下来讨论奇偶迹下的分布测试与分布无关的基于样本的测试之间联系的起点。

6.8 分布无关的基于样本的性质测试

我们现在证明，在奇偶迹下测试支持集大小 k，等价于在标准的分布无关的基于样本模型中测试 k-交替函数（其最优样本复杂度尚不明确 [BFH21]）。这一结果有一个有趣的推论：在 引理 6.43 中，若能改进奇偶迹下测试支持集大小的下界，就有可能得到分布无关的基于样本模型中测试半空间的更优下界。

我们需要以下关于编辑距离的命题，其证明见附录 D.5。

命题 6.37. 设 Ξ 是适当的带标签分布 $(g, \mathcal{E})$ 的性质，其中 $\pi_{g,\mathcal{E}}$ 的支持集大小至多为 k。那么，对于任意适当的带标签分布 $(f, \mathcal{D})$，$\mathsf{dist}_{\mathsf{TV}}((f, \mathcal{D})\ \Xi) \le \mathsf{dist}_{\mathsf{edit}}((f, \mathcal{D})\ \Xi)$。

定理 6.38. 设 Π 是定义域为 $\mathbb{N}$ 且支持集大小至多为 k的分布类。设 $m_1(k,\epsilon)$ 是奇偶迹下一个($\Pi, \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi)$ $2/3$)-分布测试器的最优样本复杂度，而 $m_2(k,\epsilon)$ 是针对 k-交替函数的无分布假设的基于样本的测试器的最优样本复杂度。那么 $m_1(k,\epsilon) = \Theta(m_2(k,\epsilon))$。

第一方向的证明。我们希望构造一个针对奇偶迹的（$\Pi, \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi)$ $2/3$）分布测试器，其样本复杂度为 $O(m_2(k,\epsilon))$。令 Ξ为带标签分布（$f,\mathcal{D}$）的集合，其中 f 是一个（$k-1$）-交替函数，且 $\mathcal{D}$ 是 $\mathbb{Z}$上的任意分布。令 Ξ' 为定义域为 $\mathbb{Z}$ 的带标签分布（$f,\mathcal{D}$）的集合，其中 $\pi_{f,\mathcal{D}}$ 的支持集大小至多为 k，因此 $\Pi(\Xi') = \Pi$。引理 6.36 给出了一个（$\Xi, \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)$ $3/4$）带标签分布测试器，其样本复杂度为 $O(m_2(k,\epsilon))$。我们将构造一个（$\Xi', \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi')$ $3/4$）带标签分布测试器，其样本复杂度为 $O(m_2(k,\epsilon))$；由引理 6.30 可知，结论随之成立。

注意到 $\Xi \subset \Xi'$，因为对于任意 $(k-1)$ 交替函数 f 和任意分布 $\mathcal{D}$，π_f，$\mathcal{D}$ 的支持集大小至多为 k。我们证明，对于任意带标签分布 $(f,\mathcal{D}) \in \Xi'$，都存在带标签分布 $(g,\mathcal{D}) \in \Xi$，使得 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\mathcal{D}_g) = 0$，从而 $\mathcal{D}_f = \mathcal{D}_g$。令 $a_1 < a_2 < \cdots$为 f 的交替序列，并采用约定 $a_0 = -\infty$。由于 π_f，$\mathcal{D}$ 的支持集大小至多为 k，因此至多存在 k 个区间 $(a_{i-1},a_i]$，使得 $\mathcal{D}(a_{i-1},a_i] > 0$。通过对所有属于任一满足 $\mathcal{D}(a_{i-1},a_i] = 0$ 的区间 $(a_{i-1},a_i]$ 的 x，赋值 $g(x) = 1 - f(x)$，构造 g。由事实 D.4 可得 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\mathcal{D}_g) = 0$。所得函数至多有 $k-1$ 个交替点，因此 $(g,\mathcal{D}) \in \Xi$。

假设 $(f,\mathcal{D}) \in \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)$，并反设存在 $(g,\mathcal{E}) \in \Xi'$使得 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\mathcal{E}_g) \le \epsilon$。那么存在 $(g',\mathcal{E}') \in \Xi$ 使得 $\mathcal{E}'_{g'} = \mathcal{E}_g$，因此 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\mathcal{E}'_{g'}) = \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\mathcal{E}_g) \le \epsilon$，且 $(f,\mathcal{D}) \notin \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)$，矛盾。因此 $(f,\mathcal{D}) \in \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi')$。于是$\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi) = \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi')$。

那么，任何 $(\Xi, \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)$ $3/4)$-带标签分布测试器也是 $(\Xi', \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi')$ $3/4)$-带标签分布测试器，因为来自 Ξ中元素的样本与来自 Ξ'中元素的样本不可区分。

□

第二个方向的证明。我们希望构造一个（Ξ，$\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)$ $2/3$）-带标签分布测试器；那么结论便可由命题 6.15得出。

如上界论证所示，这等价于一个（Ξ'，$\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi')$ $2/3$）-带标签分布测试器，其中 Ξ' 是带标签分布类（f，$\mathcal{D}$），其中 $\pi_{f,\mathcal{D}}$ 的支持集大小至多为 k。注意，$\Pi = \Pi(\Xi')$。

根据命题 6.37，$\mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi') = \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi')$，因此这等价于一个（$\Xi'$，$\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi')$ $2/3$）-带标签分布测试器。由于 Ξ' 和 $\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi')$是密度性质，根据引理 6.26，只需在奇偶迹下获得一个（$\Pi(\Xi')$ $\Pi(\mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Xi'))$ $2/3$）-分布测试器。最后，应用命题 6.27。

□

上述定理将测试 k-交替函数的样本复杂度与关于编辑距离、在奇偶迹下测试支持集大小的复杂度联系起来。接下来，我们将重述 [BFH21], 的结果：测试 k-交替函数需要 $\Omega(k/\log k)$ 个样本。这可由$n/\log n)$ 下界推出；该下界针对支持集大小估计，由[VV11、WY19]给出。我们采用以下形式表述 [VV11]的结果：

定理 6.39 ([VV11])。对于任意足够小的常数 $\delta > 0$，存在一对分布 π^+, π^-，其非零密度至少为 $\frac{1}{n}$，使得 π^+ 的支持集大小至少为 $(1-\delta)n$，π^- 的支持集大小至多为 $(1+\delta)\frac{n}{2}$，而区分二者需要 $\Omega\left(\frac{n}{\log n}\right)$ 个样本。

他们的结果也适用于分布熵的估计；在这种情况下，他们通过构造分布 $\pi_\epsilon^+, \pi_\epsilon^-$ 得到 $\Omega(n/\epsilon \log n)$ 的下界：这些分布分别以概率 ϵ 从 π^+, π^- 中采样，否则采样元素 $\perp$；这会将熵差缩小到原始差值的 ϵ，而区分 π_ϵ^+ 和 π_ϵ^- 则需要多出 $\frac{1}{\epsilon}$ 的样本。虽然这一论证不适用于支持集大小估计问题（该问题要求密度以下界为 $1/n$），但适用于针对全变差距离测试支持集大小：

推论 6.40。令 Π 为支撑集大小至多为 n 的 $\mathbb{N}$ 上的分布集合，并令 $\epsilon > 0$。那么，任何$(\Pi, \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Pi))$-分布测试器所需的样本量至少为$\Omega\left(\frac{n}{\epsilon \log n}\right)$。

为应用这个下界，我们将相对于全变差距离的测试归约为相对于编辑距离的测试。我们需要以下引理，其证明推迟到附录 D.6。

引理 6.41. 令 $k \in \mathbb{N}$。令 Π_k 为支撑于 $\mathbb{N}$ 且至多包含 k 个元素的分布集合，并令 Π_{2k} 为支撑于 $\mathbb{N}$ 且至多包含 $2k$ 个元素的分布集合。令 π 为 $\mathbb{N}$上的有限支撑概率分布，并令 π' 为 $\mathbb{N}$ 上的概率分布，其定义为$\pi'(2i-1) = \pi'(2i) := \pi(i)/2$，对每个 $i \in \mathit{supp}(\pi)$ 均成立。那么 $\mathit{dist}_{\mathsf{edit}}(\pi', \Pi_{2k}) \geq \frac{1}{4} \cdot \mathit{dist}_{\mathsf{TV}}(\pi, \Pi_k)$。

现在可以给出这个简单的归约。

引理 6.42. 令 Π_{2k} 为定义域 $\mathbb{N}$ 上支撑集大小至多为 $2k$的分布集合，并令$\epsilon > 0$。那么，$(\Pi_{2k}, \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi_{2k}))$-分布测试器的样本复杂度至少为 $\Omega\left(\frac{k}{\epsilon \log k}\right)$。

证明. 令 Π_k 为定义域 $\mathbb{N}$ 上支撑集大小至多为 k的分布集合。我们将把 $(\Pi_k, \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Pi_k))$-分布测试归约为 $(\Pi_{2k}, \mathrm{FAR}_{\epsilon/4}^{\mathsf{edit}}(\Pi_k))$-分布测试，由此即可得出结论，这归功于 推论 6.40。

对于输入分布 π，算法按如下方式进行。我们定义分布 π'，其中对每个 $i \in \mathbb{N}$，$\pi'(2i) := \pi(i)/2$ 且 $\pi'(2i-1) := \pi(i)/2$。算法可以通过对 $i \sim \pi$ 采样，然后以相同概率取 $2i$ 或 $2i-1$，来模拟从 π' 中抽取一个样本。随后，算法在输入 π'上模拟 $(\Pi_{2k}, \mathrm{FAR}_{\epsilon/4}^{\mathsf{edit}}(\Pi_{2k}))$ -分布测试器。如果 $\pi \in \Pi_k$，显然$\pi' \in \Pi_{2k}$，因此算法将正确接受（概率至少为 $2/3$）。如果 $\pi \in \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Pi_k)$，那么根据引理 6.41，$\pi' \in \mathrm{FAR}_{\epsilon/4}^{\mathsf{edit}}(\Pi_{2k})$，因此算法将正确拒绝（概率至少为 $2/3$）。

□

现在，我们可以还原 [BFH21] 关于在定义域 $\mathbb{R}^n$上测试若干性质的结果，包括半空间、半空间交和决策树。其思路是：通过以恰当的方式将一维空间嵌入 $\mathbb{R}^n$，将对 k-交替函数的性质测试归约为对所关注性质的测试。这一技术也用于 [ES20、BFH21]。我们对半空间给出形式化证明；关于半空间交和决策树的详细内容，请参见 [BFH21]。

引理 6.43。设 Π 是 $\mathbb{N}$ 上支持集大小至多为 k 的分布性质。对于任意 d 和 $\epsilon > 0$，令 $h(d, \epsilon)$ 表示在分布无关的基于样本模型中对定义域 $\mathbb{R}^d$ 上半空间进行性质测试的样本复杂度。那么，存在一个在奇偶迹下的 $(\Pi, \mathrm{FAR}_\epsilon^{\mathsf{edit}}(\Pi)\ 2/3)$-分布测试器，其样本复杂度为 $O(h(k+1, \epsilon))$。

证明。 这由 定理 6.38 以及以下从测试 k-交替函数到测试半空间的归约得出。

对于输入 $f:\ \mathbb{Z}\to\{0,1\}$ 以及定义在 $\mathbb{Z}$ 上的分布 $\mathcal{D}$，考虑单射函数 $\psi(x):=(1,\ x,\ x^2,\ \ldots\ldots,\ x^k)$，以及定义为 $\psi(x)$ 上的分布的分布 $\psi\mathcal{D}$，其中 $x\sim\mathcal{D}$。测试器将在样本 $(\psi(x)\ f(x))$ 上模拟半空间测试器，其中 $(x,\ f(x))\ \sim\mathcal{D}_f$。

注意，函数 $f:\ \mathbb{Z}\to\{0,1\}$ 是 k-交替的，当且仅当存在一个次数为 k的多项式 $p:\ \mathbb{Z}\to\mathbb{R}$，使得 $f(x)\ =\ \mathrm{sign}(p(x))$；其中我们定义 $\mathrm{sign}(z)\ =\ 0$ of $z\ \le\ 0$，且当 $z>0$时，$\mathrm{sign}(z)\ =1$。那么，f是 k-交替的，当且仅当存在 $w=(w_0,w_1,\ldots,w_k)\ \in\mathbb{R}^{k+1}$，使得 $f(x)\ =\mathrm{sign}(\langle w,\psi(x)\rangle)$。因此，$f$是 k-交替的，当且仅当存在一个半空间g：$\mathbb{R}^{k+1}\to\{0,1\}$，使得对所有 x，$f(x)\ =g(\phi(x))$。

记 $\psi(\mathcal{D}_f)$ 为当 $x\sim\mathcal{D}$ 时 $(\psi(x)\ f(x))$ 的分布。因此，对于任意 k-交替函数f，都存在一个半空间 g，使得 $\psi(\mathcal{D}_f)\ =(\psi\mathcal{D})_g$。另一方面，对于任意半空间 g，都存在一个 k-交替函数 f，使得 $\psi(\mathcal{D}_f)\ =(\psi\mathcal{D})_g$。

因此，对于任意 k-交替函数 f 和分布 $\mathcal{D}$，来自 $\psi(\mathcal{D}_f)$ 的样本与来自 $(\psi\mathcal{D})_g$ 的样本无法区分，其中 g 是一个半空间，因此半空间测试器会接受。另一方面，对于任意在分布 $\mathcal{D}$ 下与 k-交替函数相距 ϵ-远的函数 f，考虑任意函数 f'：$\mathbb{R}^{k+1}\to\{0,1\}$，使得在所有 $x\in\mathbb{Z}$ 上，$f'(\psi(x))\ =f(x)$，因此 $(\psi\mathcal{D})_{f'}=\psi(\mathcal{D}_f)$，所以半空间测试器在模拟样本 $(\psi(x)\ f(x))$ 上的表现与在样本 $(z,\ f'(z))\ \sim(\psi\mathcal{D})_{f'}$ 上完全相同，因此测试器的表现就如同它收到输入 f'和 $\psi\mathcal{D}$ 一样。

如果存在一个半空间 g，使得 $\mathbb{P}_{x\sim\mathcal{D}}[f'(\psi(x))\ =g(\psi(x))]\le\epsilon$，那么 $\mathbb{P}_{x\sim\mathcal{D}}[f(x)\ =g(\psi(x))]\le\epsilon$，其中 $g(\psi(\cdot))$ 是 k-交替的，这就产生了矛盾。因此，相对于 $\psi\mathcal{D}$，f' 与半空间的距离必定大于 ϵ，所以半空间测试器会拒绝 f'。 □

结合这一归约和引理 6.42给出的下界，我们得到以下界。请注意，[BFH21] 只给出了常数 ϵ 对应的界，但上述放大论证也可以直接应用于他们的结果。

推论 6.44 (参见 [BFH21])。以下是分布无关的基于样本模型中性质测试样本复杂度的下界：

1. k 上的交替函数：$\Omega\left(\epsilon^{k}{}_{\log k}\right)$；2. $\mathbb{R}^n$上的半空间：$\Omega\left(\epsilon^{n}{}_{\log n}\right)$；3. $\mathbb{R}^n$ 上 k 个半空间的交集：$\Omega\left(\epsilon^{nk}{}_{\log(nk)}\right)$；4. $\mathbb{R}^n$ 上大小为 k 的决策树：$\Omega\left(\epsilon^{k}{}_{\log k}\right)$。

---

7 迹重构模型中的性质测试

我们首先正式定义迹重构模型中的性质测试。接着讨论字符串上的（相对）编辑距离与我们引入的分布间编辑距离之间的联系，这将是我们结果中的一个关键组成部分。

对于字符串 $x\in\{0,1\}^N$ 和保留率 $\rho\in(0,1)$，$\mathrm{del}(x,\rho)$ 表示通过以概率 $1-\rho$ 相互独立地删除 x 的每个字符而从 x 得到的子串分布。样本 $\boldsymbol{T}\sim\mathrm{del}(x,\rho)$ 称为来自 x 、删除率为 $1-\rho$ 的迹 。

定义 7.1 （轨迹测试）。设 Ψ_1和 Ψ_2是 $\{0,1\}^N$中字符串的性质，且$\alpha,\ \rho\in(0,\ 1)$，分别称为成功概率和保留率。使用 m 条迹的$(\Psi_1,\ \Psi_2,\ \rho,\ \alpha)$-轨迹测试算法是一个算法 A，满足：对于从 x 以删除率 $1-\rho$获得的 m 条独立迹$\boldsymbol{T}_1(x)$ ..... $\boldsymbol{T}_m(x)$,

1. 若 $x \in \Psi_1$，则 $\mathbb{P}[A(\boldsymbol{T}_1(x)\ldots\ldots,\ \boldsymbol{T}_m(x))$ 接受 $] \geq \alpha$；并且，2. 若 $x \in \Psi_2$，则 $\mathbb{P}[A(\boldsymbol{T}_1(x)\ldots\ldots,\ \boldsymbol{T}_m(x))$ 拒绝 $] \geq \alpha$。

我们的许多结果都涉及 n 块字符串和均匀的 n 块字符串，现定义如下。

定义 7.2。固定 $N \in \mathbb{N}$。如果 $x \in \{0,1\}^N$ 是一个 n-块字符串，则称其为：x 至多由 n 个块组成，其中块是全为 1 的字符串或全为 0 的字符串。对于整除 N 的整数 n，*1*-均匀 n-块字符串为 $1^{N/n}0^{N/n}1^{N/n}\cdots\mathsf{par}(n)^{N/n}$，而 *0*- 均匀 n- 块字符串为 $0^{N/n}1^{N/n}0^{N/n}\cdots(1-\mathsf{par}(n))^{N/n}$。如果 x 是 1-均匀或 0-均匀的 n-块字符串，则称其为均匀 n-块字符串。

定义 7.3（相对编辑距离）。记 $\mathsf{dist}^{\mathsf{string-edit}}\colon \{0,1\}^* \times \{0,1\}^* \to \mathbb{Z}$ 为字符串上的编辑距离，我们将字符串 $x \in \{0,1\}^N$ 和 $y \in \{0,1\}^M$ 之间的相对字符串编辑距离定义为

$$\mathsf{dist}_{\mathsf{rel-edit}}(x,y) = \frac{2}{N+M}\mathsf{dist}_{\mathsf{string-edit}}(x,y)\,.$$

我们定义字符串与概率分布之间的一种对应关系，使我们能够将轨迹重建的性质测试与奇偶迹下的分布测试联系起来。

定义 7.4（字符串到分布的对应关系）。对于任意固定的 $N \in \mathbb{N}$ 和定义在 $\mathbb{N}$ 上、密度为 $1/N$ 的整数倍的概率分布 π，我们定义字符串 $\psi(\pi) \in \{0,1\}^N$ 如下：

$$\psi(\pi) := 1^{\pi(1)\cdot N}0^{\pi(2)\cdot N}1^{\pi(3)\cdot N}0^{\pi(3)\cdot N}\cdots,$$

其中 b^k 表示重复 k 次的字符 b。此映射不是一一对应的。但是，对于字符串 $x \in \{0,1\}^N$，我们定义概率分布 $\psi^{-1}(x)$ 如下。定义函数 $f_x\colon \mathbb{N} \to \{0,1\}$，令其对每个 $i \in [N]$ 满足 $f_x(i) = x_i$（在其他情况下取 1），并令 $\mathcal{D}$ 为定义在 $[N]$ 上的均匀分布。于是

$$\psi^{-1}(x) := \pi_{f_x,\mathcal{D}}\,.$$

可以验证，对于任意字符串 $x \in \{0,1\}^N$，都有 $\psi(\psi^{-1}(x)) = x$。对于字符串集合 $\{0,1\}^N$ 中的每个性质 Ψ，我们关联一个概率分布的性质 $\Pi = \Pi(\Psi) := \{\psi^{-1}(x)\colon\ x \in \Psi\}$，其中 ψ^{-1} 如定义 7.4 所定义。对于任意这样的 Ψ，令 $\mathrm{FAR}_\epsilon^{\mathsf{rel-edit}}(\Psi)$ 表示满足 $\mathsf{dist}_{\mathsf{rel-edit}}(x,\Psi) > \epsilon$ 的字符串 $x \in \{0,1\}^N$ 的集合，其中

$$\mathsf{dist}_{\mathsf{rel-edit}}(x,\Psi) := \min_{y\in\Psi}\mathsf{dist}_{\mathsf{rel-edit}}(x,y)\,.$$

观察 7.5。如果 x 是一个 n-块字符串，那么 $\psi^{-1}(x)$ 支撑集至多包含 n 个元素。如果 x 是 *1*-均匀的 n-块字符串，那么 $\psi^{-1}(x)$ 是 $\{1,2,\ldots,n\}$ 上的均匀分布。如果 x 是 *0*-均匀的 n-块字符串，那么 $\psi^{-1}(x)$ 是 $\{2,3,\ldots,n+1\}$ 上的均匀分布。

回顾一下，我们将分布间编辑距离定义为奇偶迹下分布测试的自然度量。下一个引理表明，在字符串到分布的对应关系下，分布的编辑距离与字符串上的相对编辑距离基本等价。这将使我们能够得到奇偶迹下的分布测试与轨迹重构的性质测试之间的等价关系。证明见 附录 D.3。

引理 7.6。固定任意 N，令 π、π' 为定义在 $\mathbb{N}$ 上的概率分布，其密度为 $1/N$ 的整数倍。那么 $\frac{1}{2}\cdot\mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi)\ \psi(\pi')) \leq \mathsf{dist}_{\mathsf{edit}}(\pi、\ \pi') \leq \mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi)\ \psi(\pi'))$。

7.1 单迹上界

我们旨在通过将奇偶迹下的分布性质测试（关于分布间编辑距离）归约为字符串的性质测试（关于相对编辑距离），从而获得字符串性质测试的算法。我们将使用以下简单技术：它可以将字符串的一个迹（即由删除信道产生的结果）转换为相关概率分布的泊松化样本的奇偶迹（即有放回抽样的结果）。

命题 7.7。 固定 $\rho \in (0,1)$。存在一个算法 $\textsc{Poissonize}$，它接收一个二进制字符串并输出另一个二进制字符串，满足以下性质。令 $N \in \mathbb{N}$ 和 $x \in \{0,1\}^N$，并假设输入是来自 x、删除率为 $1-\rho$ 的随机轨迹。那么输出服从 迹$(\boldsymbol{S})$，其中 $\boldsymbol{S} \sim$ 采样$(\psi^{-1}(x)\ \boldsymbol{m})$ 且 $\boldsymbol{m} \sim$ 泊松分布 $\left(N\text{对数}\left(\frac{1}{1-\rho}\right)\right)$。

证明。 其思路是将输入中的每个符号视为对应的泊松随机变量非零这一事件的指示，然后对该符号进行过采样，以获得相应的条件分布，从而得到泊松化样本。

令 $\lambda := \log\left(\frac{1}{1-\rho}\right)$，并令泊松分布$_{>0}(\lambda)$表示一个参数为$\lambda$的泊松分布随机变量在非零条件下的分布。

算法如下进行：对于输入字符串 s，从左到右处理每个符号 s_j，相互独立地采样 $\boldsymbol{z}_j \sim$Poi$_{>0}(\lambda)$，并将 $\boldsymbol{z}_j$ 个 s_j 追加到输出中。

F对于每个 $i \in [N]$，令 $\boldsymbol{X}_i \sim \mathsf{Ber}(\rho)$，且相互独立。于是输入的分布为 a s

$$x_1^{\boldsymbol{X}_1} x_2^{\boldsymbol{X}_2} \cdots x_{N-1}^{\boldsymbol{X}_{N-1}} x_N^{\boldsymbol{X}_N}\,.$$

令 $\pi := \psi^{-1}(x)$，称其支撑于 $[n]$。对于每个 $i \in [n]$，令 $\boldsymbol{Y}_i \sim \mathsf{Poi}(N\lambda\pi(i))$，且相互独立。于是目标输出分布与以下分布相同

$$1^{\boldsymbol{Y}_1} 0^{\boldsymbol{Y}_2} \cdots \mathsf{par}(n-1)^{\boldsymbol{Y}_{n-1}} \mathsf{par}(n)^{\boldsymbol{Y}_n}\,.$$

根据泊松分布的可加性以及 ψ^{-1} 的定义，该分布与以下分布相同

$$x_1^{\boldsymbol{Z}_1} x_2^{\boldsymbol{Z}_2} \cdots x_{N-1}^{\boldsymbol{Z}_{N-1}} x_N^{\boldsymbol{Z}_N}\,.$$

其中对于每个 $i \in [N]$，$\boldsymbol{Z}_i \sim \mathsf{Poi}(\lambda)$，且相互独立。

考虑从 x 生成迹的随机过程以及算法的随机过程，我们可以将算法输出中的每个符号与迹中出现 x_i 的位置 $i \in [N]$ 对应起来。对于每个 $i \in [N]$，令 $\boldsymbol{K}_i$ 为表示 x_i 被追加到输出中的次数的随机变量。于是 $\boldsymbol{K}_i$ 按照以下随机过程分布：如果 $\boldsymbol{X}_i = 0$，则 $\boldsymbol{K}_i \leftarrow 0$；否则为 $\boldsymbol{K}_i \leftarrow$ Poi$_{>0}(\lambda)$。注意，$\boldsymbol{K}_i$ 相互独立，且算法的输出为

$$x_1^{\boldsymbol{K}_1} x_2^{\boldsymbol{K}_2} \cdots x_{N-1}^{\boldsymbol{K}_{N-1}} x_N^{\boldsymbol{K}_N}\,.$$

因此，只要对每个 $i \in [N]$、$\boldsymbol{Z}_i$ 和 $\boldsymbol{K}_i$，它们的分布都相同，我们就完成了。下面来验证这一点。我们有 $\mathbb{P}[\boldsymbol{K}_i = 0] = \mathbb{P}[\boldsymbol{X}_i = 0] = 1-\rho$ 和 $\mathbb{P}[\boldsymbol{Z}_i = 0] = e^{-\lambda} = 1-\rho$，并且对每个 $k \geq 1$，

$$\begin{aligned}\mathbb{P}[\boldsymbol{Z}_i = k] &= \mathbb{P}[\boldsymbol{Z}_i > 0]\,\mathbb{P}[\boldsymbol{Z}_i = k \mid \boldsymbol{Z}_i > 0] = \rho \cdot \mathbb{P}[\mathsf{Poi}_{>0}(\lambda) = k] \\ &= \mathbb{P}[\boldsymbol{X}_i = 1] \cdot \mathbb{P}[\boldsymbol{K}_i = k \mid \boldsymbol{X}_i = 1] = \mathbb{P}[\boldsymbol{K}_i = k \text{ and } \boldsymbol{X}_i = 1] = \mathbb{P}[\boldsymbol{K}_i = k]\,. \qquad \square\end{aligned}$$

注记 7.8。 虽然我们假设 ρ 是显式已知的，以便得到 命题 7.7，但这一假设并非关键：如果我们只知道 ρ_0 在 ρ上的一个下界，就可以通过从迹中子采样 $\mathsf{Bin}(N,\ \rho')$ 个元素，得到基本等价的结果，其中 $\rho' = \rho_0/C$，这里 C是某个较大的常数。这样，除非发生可忽略不计的失败概率，否则样本的分布就等同于删除率已知的迹 $1-\rho'$。

借助这一结果，我们可以通过黑盒归约，从奇偶迹模型中的测试器得到迹重构模型中的测试器。

引理 7.9. 设 $N, m \in \mathbb{N}$。设 Ψ_1, Ψ_2 是 $\{0,1\}^N$中字符串的性质，且设 $\alpha > 0$。如果奇偶迹下存在一个泊松化 $(\Pi(\Psi_1)\ \Pi(\Psi_2)\ \alpha)$-分布测试器，其样本复杂度为$m$，那么就存在一个$(\Psi_1, \Psi_2, \rho, \alpha)$-迹测试器，仅使用一条迹，且满足 $\rho = 1 - e^{-m/N}$。

证明. 设 A 是一个奇偶迹下的泊松化 $(\Pi(\Psi_1)\ \Pi(\Psi_2)\ \alpha)$-分布测试器，其样本复杂度为 m。我们的迹测试器 B 按如下方式工作：

1. 接收一条删除率为 $1-\rho$的迹 x。
2. 令 $y \leftarrow \textsc{Poissonize}(x)$。
3. 返回 $A(y)$。

令 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 分别表示输入到 B 和 A的随机变量。根据 命题 7.7，$\boldsymbol{y}$服从 迹$(\boldsymbol{S})$，其中 $\boldsymbol{S} \sim$采样$(\psi^{-1}(x)\ \boldsymbol{m})$，且 $\boldsymbol{m} \sim$泊松分布$(N\lambda)$，其中 $\lambda =$对数 $\left(\frac{1}{1-\rho}\right) =$ —— m/N。换言之，$\boldsymbol{y}$ 服从从 $\psi^{-1}(x)$ 中抽取大小为 $\mathsf{Poi}(m)$ 的样本所得的奇偶迹。此外，根据 $\Pi(\Psi_1)$ 的定义，如果 $x \in \Psi_1$，那么 $\psi^{-1}(x) \in \Pi(\Psi_1)$，$\Psi_2$ 也同样如此。因此，B 的正确性可由 A 的正确性推出。 □

现在，我们利用 第 1.4 节推出每个单迹上界，这直接源于字符串与分布之间编辑距离的等价性。

命题 7.10。 令 $N \in \mathbb{N}$和 $\epsilon > 0$，并设 Ψ 是 $\{0,1\}^N$中字符串的一个性质。那么$\Pi(\textsc{far}_\epsilon^{\mathsf{rel-edit}}(\Psi)) \subseteq \textsc{far}_{\epsilon/2}^{\mathsf{edit}}(\Pi(\Psi))$。

证明。设 $\pi \in \Pi(\textsc{far}_\epsilon^{\mathsf{rel-edit}}(\Psi))$。根据 $\Pi(\textsc{far}_\epsilon^{\mathsf{rel-edit}}(\Psi))$ 的定义，对于某个 $x \in \textsc{far}_\epsilon^{\mathsf{rel-edit}}(\Psi)$，有 $\pi = \psi^{-1}(x)$。因此 $x \in \{0,1\}^N$，并且对于每个 $y \in \Psi$，$\mathsf{dist}_{\mathsf{rel-edit}}(x,y) > \epsilon$；根据引理 7.6，$\mathsf{dist}_{\mathsf{edit}}(\psi^{-1}(x)\ \psi^{-1}(y)) > \epsilon/2$。

根据 $\Pi(\Psi)$ 的定义，对于每个 $\pi' \in \Pi(\Psi)$，存在某个 $y' \in \Psi$ 使得 $\pi' = \psi^{-1}(y')$。但于是 $\mathsf{dist}_{\mathsf{edit}}(\pi, \pi') = \mathsf{dist}_{\mathsf{edit}}(\psi^{-1}(x)\ \psi^{-1}(y')) > \epsilon/2$。因此 $\pi \in \textsc{far}_{\epsilon/2}^{\mathsf{edit}}(\Pi(\Psi))$。 □

我们对性质测试 n-块字符串的结果需要以下等价关系：字符串到 n-块字符串性质的相对编辑距离，与适当概率分布到支撑集至多包含 n 个元素的分布性质的编辑距离之间的等价关系。证明见附录 D.7。

命题 7.11. 设 $N, n \in \mathbb{N}$。设 Ψ 为 $\{0,1\}^N$ 中所有 n-块字符串构成的集合，Π 为支撑集大小至多为 n 的 $\mathbb{N}$ 上的概率分布集合。那么，对于密度为 $1/N$ 整数倍的 $\mathbb{N}$ 上任意分布 π，以及 $x = \psi(\pi)$，

$$\mathsf{dist}_{\mathsf{edit}}(\pi, \Pi) \le \mathsf{dist}_{\mathsf{rel-edit}}(x, \Psi) \le 2 \cdot \mathsf{dist}_{\mathsf{edit}}(\pi, \Pi)\,.$$

这意味着：

命题 7.12. 设 $N \in \mathbb{N}$。设 Ψ 是 $\{0,1\}^N$ 中字符串的一个性质， Π 是 $\mathbb{N}$上的概率分布的一个性质。假设对于每个 $\mathbb{N}$ 上的分布 π，其密度均为 $1/N$ 的整数倍，并且对于 $x = \psi(\pi)$，都有 $\mathsf{dist}_{\mathsf{rel-edit}}(x, \Psi) \leq 2 \cdot \mathsf{dist}_{\mathsf{edit}}(\pi, \Pi)$。那么

$$\Pi(\mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)) \subseteq \mathrm{FAR}_{\epsilon/4}^{\mathsf{edit}}(\Pi)\,.$$

证明. 设 I 为 $\mathbb{N}$ 上的分布集合，其中各分布的密度都是 $1/N$的整数倍。我们断言 $\mathrm{FAR}_{\epsilon/2}^{\mathsf{edit}}(\Pi(\Psi)) \cap I \subseteq \mathrm{FAR}_{\epsilon/4}^{\mathsf{edit}}(\Pi)$。

任取 $\pi \in \mathrm{FAR}_{\epsilon/2}^{\mathsf{edit}}(\Pi(\Psi)) \cap I$。对于每个 $y \in \Psi$，都有 $\mathsf{dist}_{\mathsf{edit}}(\pi, \psi^{-1}(y)) > \epsilon/2$；又由于根据 $\pi \in I$，π 中的所有密度都是 $1/N$ 的整数倍，令 $x := \psi(\pi)$，由引理 7.6 可得 $\mathsf{dist}_{\mathsf{rel-edit}}(x, y) > \epsilon/2$。因此 $\mathsf{dist}_{\mathsf{rel-edit}}(x, \Psi) > \epsilon/2$，再利用假设可得 $\mathsf{dist}_{\mathsf{edit}}(\pi, \Pi) > \epsilon/4$，从而 $\pi \in \mathrm{FAR}_{\epsilon/4}^{\mathsf{edit}}(\Pi)$，这就证明了第一个断言。

根据命题 7.10， $\Pi(\mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)) \subseteq \mathrm{FAR}_{\epsilon/2}^{\mathsf{edit}}(\Pi(\Psi))$，因此 $\Pi(\mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)) \cap I \subseteq \mathrm{FAR}_{\epsilon/2}^{\mathsf{edit}}(\Pi(\Psi)) \cap I \subseteq \mathrm{FAR}_{\epsilon/4}^{\mathsf{edit}}(\Pi)$。最后，注意到 $\Pi(\mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)) \cap I = \Pi(\mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi))$，因为后者中的每个元素都具有某个 $x \in \{0,1\}^N$ 对应的 $\pi = \psi^{-1}(x)$ 形式，故证毕。 □

以下结果给出了定理 1.9的上界部分。

定理 7.13。 设 N、$n \in \mathbb{N}$ 和 $\epsilon > 0$，并令 Ψ 为 $\{0,1\}^N$ 中所有 n-块字符串组成的集合。存在一个$(\Psi$、$\mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)$ ρ、$2/3)$-迹测试器，只需使用一条迹，其期望轨迹大小为 $\rho N = O(n/\epsilon)$。

证明。 设 Π 为定义在 $\mathbb{N}$上且支持集大小至多为 n 的分布类。根据 定理 6.38，存在一个在奇偶迹下的（Π, $\mathrm{FAR}_{\epsilon/4}^{\mathsf{edit}}(\Pi)$ $2/3$）分布测试器。根据 命题 A.1，不妨假设该测试器经过泊松化，其样本复杂度为 $O(n/\epsilon)$。

由于对每个 $x \in \Psi$，都有 $\psi^{-1}(x)$ 的支持集大小至多为 n (观察 7.5)，因此可得 $\Pi(\Psi) \subseteq \Pi$。命题 7.12 结合 命题 7.11 可得 $\Pi(\mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)) \subseteq \mathrm{FAR}_{\epsilon/4}^{\mathsf{edit}}(\Pi)$。

因此，我们得到一个在奇偶迹下的（$\Pi(\Psi)$ $\Pi(\mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi))$ $2/3$）分布测试器，其样本复杂度为 $m = O(n/\epsilon)$。接着，引理 7.9 给出一个（$\Psi, \mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)$ $\rho, 2/3$）迹测试器，对 $\rho = 1 - e^{-m/N} \leq m/N$使用一条迹，即期望轨迹大小为 $\rho N = O(n/\epsilon)$。 □

在奇偶迹下进行分布测试，更容易得到我们关于迹测试均匀分布的 n-块字符串的结果，因为此时相应的概率分布性质 Π仅包含与均匀分布的 n-块字符串相对应的分布。

记号。 对于固定的 $N \in \mathbb{N}$ 和满足 n 整除条件的 N，令 $u^{(1)}$ 表示 1-均匀分布的 n-块字符串，令 $u^{(0)}$ 表示 0-均匀分布的 n-块字符串。

命题 7.14。 设 N、$n \in \mathbb{N}$ 满足 n 整除 N。设 Ψ 仅包含 *1*-均匀分布的n-块字符串 $u^{(1)}$，且设 Π 仅包含 $[n]$上的均匀分布。对于任意在 $\mathbb{N}$ 上的分布π，其密度均为 $1/N$ 的整数倍，并且对于 $x = \psi(\pi)$，

$$\mathsf{dist}_{\mathsf{edit}}(\pi, \Pi) \leq \mathsf{dist}_{\mathsf{rel-edit}}(x, \Psi) \leq 2 \cdot \mathsf{dist}_{\mathsf{edit}}(\pi, \Pi)\,.$$

证明。令 π^* 为 $[n]$上的均匀分布，因此 $\pi^* = \psi^{-1}(u^{(1)})$（观察 7.5）。这两个不等式直接由引理 7.6 得出。首先，我们有 $\mathsf{dist}_{\mathsf{edit}}(\pi, \Pi) = \mathsf{dist}_{\mathsf{edit}}(\pi, \pi^*) \leq \mathsf{dist}_{\mathsf{rel-edit}}(x, u^{(1)}) = \mathsf{dist}_{\mathsf{rel-edit}}(x, \Psi)$。类似地，$\mathsf{dist}_{\mathsf{edit}}(\pi, \Pi) = \mathsf{dist}_{\mathsf{edit}}(\pi, \pi^*) \geq 1\ 2\mathsf{dist}_{\mathsf{edit}}(x, u^{(1)}) = \frac{1}{2}\mathsf{dist}_{\mathsf{edit}}(x, \Psi)$。 □

我们首先给出一个用于 1-均匀 n块字符串的测试器，然后将其推广到两类均匀字符串，从而得到定理 1.8，重述如下：

定理 7.15. 设 $N, n \in \mathbb{N}$ 满足 n 整除 N，并设 $\epsilon > 0$。设 Ψ 只包含 $\{0,1\}^N$ 中均匀分布的 n-块字符串。存在一个 $(\Psi, \mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)\ \rho, 2/3)$-迹测试器，仅使用一条迹，其期望轨迹大小为 $\rho N = O\left(\frac{n}{\epsilon} + \epsilon^2 n \log n\right)$。

引理 7.16. 设 $N, n \in \mathbb{N}$ 满足 n 整除 N，并设 $\epsilon > 0$。设 Ψ 只包含 $u^{(1)} \in \{0,1\}^N$ 中的 *1*-均匀 n-块字符串。存在一个 $(\Psi, \mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)\ \rho, 2/3)$-迹测试器，仅使用一条迹，其期望轨迹大小为 $\rho N = O\left(\frac{n}{\epsilon} + \epsilon^2 n \log n\right)$。

证明。令 Π 仅包含 $[n]$ 上的均匀分布 $\pi = \psi^{-1}(u^{(1)})$。根据定理 6.34，存在一个针对奇偶迹的 $(\Pi, \mathrm{FAR}_{\epsilon/4}^{\mathsf{edit}}(\Pi)\ 2/3)$-分布测试器。根据命题 A.1，我们可以假设该测试器经过泊松化，其样本复杂度为 $O\left(\frac{n}{\epsilon} + \epsilon^2 n \log n\right)$。

注意到 $\Pi = \{\pi\} = \{\psi^{-1}(u^{(1)})\} = \Pi(\Psi)$。此外，命题 7.12 与命题 7.14 共同推出 $\Pi(\mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)) \subseteq \mathrm{FAR}_{\epsilon/4}^{\mathsf{edit}}(\Pi)$，因此我们得到一个针对奇偶迹的 $(\Pi(\Psi)\ \Pi(\mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)\ 2/3)$-分布测试器，其样本复杂度为 $m = O\left(\frac{n}{\epsilon} + \epsilon^2 n \log n\right)$。接着引理 7.9 给出一个使用单条迹测试 $\rho = 1 - e^{-m/N}$ 的 $(\Psi, \mathrm{FAR}_{\epsilon}^{\mathsf{rel-edit}}(\Psi)\ \rho, 2/3)$-迹测试器，即期望迹大小为 $\rho N = O\left(\frac{n}{\epsilon} + \epsilon^2 n \log n\right)$。 □

证明定理 *7.15*。使用 n-块字符串的一个测试器版本，该版本来自引理 7.16，其成功概率为 $5/6$，再将所有符号取反后重复测试；只要任一次执行接受，就接受。若输入为 $x = u^{(1)}$，第一次执行以至少 $5/6$的概率接受；若 $x = u^{(0)}$，第二次执行以至少 $5/6$的概率接受。若$\mathsf{dist}_{\mathsf{rel-edit}}(x, \Psi) > \epsilon$，则 x 与二者都相距较远 $u^{(1)}$ 和 $u^{(0)}$，因此每次执行接受的概率至多为 $1/6$；根据并集界，x 被接受的概率至多为 $1/3$。 □

为了证明定理 1.7中的上界部分，我们引入以下定义。若 $x \in \{0,1\}^N$是一个类型-*1* n-块字符串，且 $x = 1^{t_1}0^{t_2}1^{t_3}\cdots \mathsf{par}n^{t_n}$对某些非负整数 t_1、……、t_n，则称 x是一个类型-*0* n-块字符串，若 $x = 0^{t_1}1^{t_2}0^{t_3}\cdots(1-\mathsf{par}n)^{t_n}$对某些非负整数 t_1、……、t_n。

注 7.17。字符串 x 可能同时是类型-1 n-块字符串和类型-0 n-块字符串。此外，x 是一个- n-块字符串，当且仅当它是类型-1 n-块字符串或类型-0 n-块字符串。

如果 x 是类型-1 n-块字符串，那么 $\psi^{-1}(x)$ 支撑于 $[n]$ 内；如果 x 是类型-0 n-块字符串，那么 $\psi^{-1}(x)$ 支撑于 $\{2, 3, \ldots, n+1\}$ 内。

引理 7.18。设 N、$n \in \mathbb{N}$ 满足：n 为偶数且整除 N，并设 $\epsilon > 0$。令 Ψ_1 仅包含 *1*-均匀的 n-块字符串 $u^{(1)} \in \{0,1\}^N$，并令 Ψ_2 包含 $\{0,1\}^N$ 中所有与 Ψ_1 的（相对）编辑距离至少为 ϵ 的类型-*1* n-块字符串。存在一个 $(\Psi_1$、Ψ_2、ρ、$2/3)$-迹测试器，仅使用一条期望大小为 $\rho N = \widetilde{O}((n/\epsilon)^{4/5} + \sqrt{n/\epsilon^2})$ 的迹。

证明。令 Π_1 仅包含均匀分布 $\pi^* = \psi^{-1}(u^{(1)})$ 上的 $[n]$，并令 Π_2 包含定义在 $[n]$ 上且 $\epsilon/2$与均匀分布的编辑距离至少为 的分布。根据 定理 4.1 和 事实 6.29，存在一个在奇偶迹下的 $(\Pi_1, \Pi_2, 2/3)$-分布测试器。根据 命题 A.1，不妨设它经过泊松化，其样本复杂度为 $\widetilde{O}((n/\epsilon)^{4/5} + \sqrt{n/\epsilon^2})$。

我们断言 $\Pi(\Psi_1) \subseteq \Pi_1$ 且 $\Pi(\Psi_2) \subseteq \Pi_2$。首先，令 $\pi \in \Pi(\Psi_1)$，因此必有$\pi = \psi^{-1}(u^{(1)})$。于是，π 在 $[n]$上均匀分布，故 $\pi \in \Pi_1$。现在，假设 $\pi \in \Pi(\Psi_2)$，于是$\pi = \psi^{-1}(x)$，其中存在某个 $x \in \{0,1\}^N$，使得 x 是一个类型为1 n的块字符串，且 $\mathsf{dist}_{\mathsf{rel-edit}}(x, u^{(1)}) > \epsilon$。由注 7.17 可知，$\pi$ 支撑于 $[n]$ 之内；再由引理 7.6，$\mathsf{dist}_{\mathsf{edit}}(\pi, \pi^*) \geq 1$ $2\mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi)\ \psi(\pi^*)) = \frac{1}{2}\mathsf{dist}_{\mathsf{rel-edit}}(x, u^{(1)}) > \epsilon/2$，因此 $\pi \in \Pi_2$。

因此，我们得到一个样本复杂度为 $\Pi(\Psi_1)$ $\Pi(\Psi_2)$ $2/3$ 的奇偶迹分布测试器 $m = \widetilde{O}((n/\epsilon)^{4/5} + \sqrt{n}/\epsilon^2)$。于是，引理 7.9 给出了一个使用一条迹、适用于 Ψ_1，Ψ_2，ρ，$2/3$ 的迹测试器 $\rho = 1 - e^{-m/N}$，即期望迹大小为 $\rho N = \widetilde{O}(($ $n/\epsilon)^{4/5} + \sqrt{n}/\epsilon^2$。 □

现在，我们得到定理 1.7中单迹上界的部分：

定理 7.19。令 $N, n \in \mathbb{N}$ 满足：为偶数且整除，并令 。令 仅包含均匀分布的 -块字符串 ,；令 包含 中所有与 的（相对）编辑距离至少为 的 -块字符串。存在一个使用一条迹的 $(\Psi_1, \Psi_2, \rho, 2/3)$ 迹测试器，其期望迹大小为 $\rho N = \widetilde{O}((n/\epsilon)^{4/5} + \sqrt{n}/\epsilon^2)$。

证明。关键观察是，算法 A 在引理 7.18 中得到的算法对输入中所有符号取反保持不变：它由定理 4.1中的分布测试器（它仅取决于游程长度）以及 Poissonize 来自 命题 7.7的算法组合而成；后者无论符号的取值如何，都会以相同方式转换输入。形式上，对于 $x \in \{0,1\}^N$ 并令 x 表示将 x 中的每个符号取反后得到的字符串，则输出 $A(x)$ 和 $A(x)$ 同分布。

因此，我们声称 A 也是一个（Ψ_1，Ψ_2，ρ，$2/3$）迹测试器。事实上，如果输入 $x \in \Psi_1$，那么要么 $x = u^{(1)}$ 和 A 以至少 $2/3$ 的概率接受（由引理 7.18），要么 $x = u^{(0)}$，因此 $x = u^{(1)}$，所以 A 也会接受。另一方面，如果 $x \in \Psi_2$，那么 $\mathsf{dist}_{\mathsf{rel-edit}}(x, u^{(1)}) > \epsilon$，且 $\mathsf{dist}_{\mathsf{rel-edit}}(x, u^{(1)}) = \mathsf{dist}_{\mathsf{rel-edit}}$ $(x, u^{(0)}) > \epsilon$。此外，x 或 x 是一个类型为 1 n 的块字符串，因此 A 以至少 $2/3$的概率拒绝。 □

我们指出，通过使用具有相应更高成功概率的奇偶迹下分布测试器，上述结果中的成功概率 $2/3$可以替换为任意更高的常数，而不影响渐近界。或者，也可以使用多个独立迹，将结果提升至高概率范围。

7.2 多迹上界

从多个独立轨迹中进行推断的能力是迹重构模型的核心组成部分。因此，我们希望测试均匀分布的 n-块字符串类，所用的迹比我们的单迹结果所需的更短，代价是需要获取更多轨迹。主要思路是通过将 k 迹拼接在一起，归约到单迹情形，并将结果视为输入字符串复制 k 次后得到的一条迹。

对于任意字符串 x，$y \in \{0,1\}^*$和整数 $k > 0$，记 $x \circ y$ 为 x 与 y 的拼接，记 $x^{\circ k}$ 为包含 k 个项的拼接 $x \circ \cdots \circ x$。

命题 7.20。存在一个普适常数 $c > 0$，使得下述结论成立。设 N、n、$k \in \mathbb{N}$满足：n 为偶数且整除 N。设 $u = u^{(1)} \in \{0,1\}^N$为 *1*-均匀的 n-块字符串，$x \in \{0,1\}^N$ 为类型-*1* n的块字符串。那么

$$\mathsf{dist}_{\mathsf{rel-edit}}(u^{\circ k}, x^{\circ k}) \geq c \cdot \mathsf{dist}_{\mathsf{rel-edit}}(u, x)\,.$$

证明。令 $\pi_u := \psi^{-1}(u)$ 和 $\pi_x := \psi^{-1}(x)$，使得 π_u 和 π_x 支撑于 $[n]$和$\psi(\pi_u) = u$，$\psi(\pi_x) = x$ 之内。将 π_u^k 定义为 $\mathbb{N}$上的如下分布：对于所有 $t \in \mathbb{Z}_{\geq 0}$ 和 $i \in [n]$，

$$\pi_u^k(tn+i) := \begin{cases} \pi_u(i)/k, & \text{if } t \leq k-1 \\ 0, & \text{if } t \geq k. \end{cases}$$

以类似的方式定义 π_x^k。由于前者的各项在后者的 k 个缩放副本中分别对齐，因此 $\mathsf{dist}_{\mathsf{TV}}(\pi_u^k, \pi_x^k) = \mathsf{dist}_{\mathsf{TV}}(\pi_u, \pi_x)$。此外，$\pi_u^k$，$\pi_x^k$ 的各项都是 $1/kN$ 的整数倍，并且它们在 $\{0,1\}^{kN}$ 上对应的字符串满足 $\psi(\pi_u^k) = u^{\circ k}$ 和 $\psi(\pi_x^k) = x^{\circ k}$。

由引理 7.6 可得 $\mathsf{dist}_{\mathsf{rel-edit}}(u^{\circ k}, x^{\circ k}) = \mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi_u^k)\ \psi(\pi_x^k)) \geq \mathsf{dist}_{\mathsf{edit}}(\pi_u^k, \pi_x^k)$。由于 n 是偶数，$u^{\circ k}$ 是 1-均匀的 kn-块字符串，因此 π_u^k 在 $[kn]$上均匀分布。显然，π_x^k 的支撑集包含于 $[kn]$。由引理 6.31 可知，对于某个普适常数 $c' > 0$，有 $\mathsf{dist}_{\mathsf{edit}}(\pi_u^k, \pi_x^k) \geq c' \cdot \mathsf{dist}_{\mathsf{TV}}(\pi_u^k, \pi_x^k) = \mathsf{dist}_{\mathsf{TV}}(\pi_u, \pi_x)$。根据编辑距离的定义，很容易看出 $\mathsf{dist}_{\mathsf{TV}}(\pi_u, \pi_x) \geq \mathsf{dist}_{\mathsf{edit}}(\pi_u, \pi_x)$。最后，再次应用引理 7.6，得到 $\mathsf{dist}_{\mathsf{edit}}(\pi_u, \pi_x) \geq \frac{1}{2}\mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi_u)\ \psi(\pi_x))$。回顾 $\psi(\pi_u) = u$ 和 $\psi(\psi_x) = x$，证明得证。 □

我们首先利用上述结果，证明对类型 1 n 块字符串中的 1-均匀字符串进行测试的多迹上界，然后将该结果推广到两类（均匀）字符串，从而得到定理 1.7的上界部分。

引理 7.21。设 N、n、$k \in \mathbb{N}$ 满足 n 为偶数且整除 N，并设 $\epsilon > 0$。设 Ψ_1 仅包含 *1*-均匀的 n-块字符串 $u^{(1)} \in \{0,1\}^N$，并设 Ψ_2 为与 Ψ_1 的（相对）编辑距离至少为 ϵ 的类型-*1* n-块字符串集合。则存在一个 $(\Psi_1, \Psi_2, \rho, 2/3)$-迹测试器，使用 k 条迹，其期望大小为 $\rho N = \widetilde{O}\left(\frac{n^{4/5}}{k^{1/5}\epsilon^{4/5}} + \frac{\sqrt{n}}{\sqrt{k}\epsilon^2}\right)$。

证明。为便于记号，令 $u := u^{(1)}$。令 Ψ_1' 仅包含 1-均匀 kn-块字符串$u^{\circ k} \in \{0,1\}^{kN}$，并令 Ψ_2'为 1 kn 中类型为 $\{0,1\}^{kN}$-块的字符串集合，其中这些字符串与 $c\epsilon$ 的相对编辑距离至少为 Ψ_1'，其中 c 是命题 7.20中的常数。

根据 引理 7.18 可得到算法 A，这是一个（Ψ_1'，Ψ_2'，ρ'，$2/3$）迹测试器，使用一条期望大小为 $\rho'(kN) = \widetilde{O}((kn)^{4/5}/\epsilon^{4/5} + \sqrt{kn}/\epsilon^2)$。我们的算法 B，这是一个（Ψ_1，Ψ_2，ρ，$2/3$）迹测试器，使用 k 条迹，工作方式如下：

1. 获取 k 条期望大小为 $\rho N = \widetilde{O}\left(\frac{n^{4/5}}{k^{1/5}\epsilon^{4/5}} + \frac{\sqrt{n}}{\sqrt{k}\epsilon^2}\right)$ 的独立迹 x_1、……、x_k。

2. 返回 A 在 $x_1 \circ \cdots \circ x_k$上的输出。

设 $x \in \{0,1\}^N$ 为 B的未知输入。若 $x \in \Psi_1$，则 $x = u$，因此 $x^{\circ k} \in \Psi_1'$。另一方面，若 $x \in \Psi_2$，则 $\mathsf{dist}_{\mathsf{rel-edit}}(x, u) > \epsilon$，我们利用命题 7.20 得出 $\mathsf{dist}_{\mathsf{rel-edit}}(x^{\circ k}, u^{\circ k}) > c\epsilon$，所以 $x^{\circ k} \in \Psi_2'$。此外，A的输入 $\boldsymbol{x}_1 \circ \cdots \circ \boldsymbol{x}_k$服从一个来自 $x^{\circ k}$的迹的分布，其期望大小为 $k(\rho N)$，即删除率为 ρ。因此，只要满足以下条件，A就会以至少 $2/3$ 的概率产生正确的输出（ B也一样）：

$$k\rho N \geq \widetilde{O}\left(\frac{(kn)^{4/5}}{\epsilon^{4/5}} + \frac{\sqrt{kn}}{\epsilon^2}\right),$$

该条件在以下情况下成立

$$\rho N \geq \widetilde{O}\left(\frac{n^{4/5}}{k^{1/5}\epsilon^{4/5}} + \frac{\sqrt{n}}{\sqrt{k}\epsilon^2}\right). \qquad \square$$

我们现在得到定理 1.7的上界部分。

定理 7.22。 设 N、n、$k \in \mathbb{N}$ 满足： n 为偶数且整除 N，并且设 $\epsilon > 0$。设 Ψ_1 仅包含 $\{0,1\}^N$ 中均匀分布的 n-块字符串，并设 Ψ_2 为与 Ψ_1 的（相对）编辑距离为 ϵ-远的 n-块字符串集合。那么，存在一个使用 k 条迹、期望大小为 ρN 的 $\widetilde{O}\left(\frac{n^{4/5}}{k^{1/5}\epsilon^{4/5}} + \frac{\sqrt{n}}{\sqrt{k}\epsilon^2}\right)$-迹测试器。

证明。该论证与定理 7.19 的证明相同。令 A 为引理 7.21 中的算法，我们观察到输出 $A(x)$ 与 $A(x)$ 的分布完全相同。因此， A 也是一个使用 k 条迹的 $(\Psi_1, \Psi_2, \rho, 2/3)$-迹测试器：如果 $x \in \Psi_1$，则 $x = u^{(1)}$ 或 $x = u^{(1)}$，所以 A 接受；而如果 $x \in \Psi_2$，则 x 和 x 都远离 $u^{(1)}$，此外 x 或 x 是一个类型-1n-块字符串，所以 A 拒绝。 □

7.3 下界

我们希望将奇偶迹下的分布检验归约为迹重构模型中的字符串性质测试。我们定义一个“迹拆分”过程，该过程接收来自分布 π 的奇偶迹，并生成 k 个字符串，这些字符串看起来像是字符串 $\psi(\pi)$ 的独立轨迹。

命题 7.23（泊松化轨迹拆分）。 存在一个随机化算法 SPLIT，满足以下性质。设 N、$k \in \mathbb{N}$，且 $\rho > 0$ 满足 $\rho < \frac{1}{20\cdot\sqrt{kN}}$。设 π 是 $\mathbb{N}$ 上任意一个概率分布，其密度是 $1/N$ 的整数倍。那么，对于输入 N、k 和 ρ， SPLIT 从 π 抽取大小为 泊松分布 $\left(k \cdot \frac{\rho}{1-\rho} \cdot N\right)$ 的奇偶迹，并输出一个由 k 个二进制字符串组成的序列，满足以下性质。设 $\boldsymbol{x}_1$、……、$\boldsymbol{x}_k$ 是表示 SPLIT 输出的随机变量（关于奇偶迹的随机性和算法的内部随机性），并设 $\boldsymbol{y}_1$、……、$\boldsymbol{y}_k$ 满足：每个 $\boldsymbol{y}_i$ 都是从 $\psi(\pi)$ 独立抽取的迹，且期望大小为 ρN。那么

$$\mathsf{dist}_{\mathsf{TV}}((\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k), (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k)) < 1/100\,.$$

证明。 令 $\lambda := \rho/(1-\rho)$。算法 SPLIT 按如下方式进行：

1. 从 π 中抽取大小为泊松分布$(k\lambda N)$的奇偶迹 T；
2. 初始化空字符串 x_1、……、x_k；
3. 对于 T 中从左到右的每个符号 b，将 b 追加到 x_i，其中 i 是从 $[k]$ 中均匀随机抽取的；
4. 返回字符串 $(x_1$、……、$x_k)$。

对于每个 $i \in [k]$ 和 $j \in [N]$，令 $\boldsymbol{X}_{i,\,j} \sim$ 相互独立且服从泊松分布(λ)。定义随机变量 $\boldsymbol{x}'_1$、……、$\boldsymbol{x}'_k$ 如下：对于每个 $i \in [k]$，

$$\boldsymbol{x}'_i = \psi(\pi)_1^{\boldsymbol{X}_{i,1}} \psi(\pi)_2^{\boldsymbol{X}_{i,2}} \cdots \psi(\pi)_N^{\boldsymbol{X}_{i,N}}\,.$$

我们声称，$(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ 与 $(\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_k)$ 的分布完全相同。事实上，首先回顾步骤 1 中的奇偶迹 $\boldsymbol{T}$ 具有如下分布：

$$\boldsymbol{T} = \psi(\pi)_1^{\boldsymbol{A}_1} \psi(\pi)_2^{\boldsymbol{A}_2} \cdots \psi(\pi)_N^{\boldsymbol{A}_N}\,,$$

其中，对每个 $j \in [N]$， $\boldsymbol{A}_j \sim$ 都服从参数为 $k\lambda$ 的泊松分布，且相互独立。然后，步骤 3 等价于将每个$\boldsymbol{A}_j$ 拆分为随机变量 $(\boldsymbol{A}_{1,j}, \dots, \boldsymbol{A}_{k,j})$，其服从参数为 $\boldsymbol{A}_j, (1/k, \dots, 1/k)$ 的多项分布，并通过拼接生成每个$\boldsymbol{x}_i$：

$$\boldsymbol{x}_i = \psi(\pi)_1^{\boldsymbol{A}_{i,1}} \psi(\pi)_2^{\boldsymbol{A}_{i,2}} \cdots \psi(\pi)_N^{\boldsymbol{A}_{i,N}} .$$

根据标准论证可知，$\boldsymbol{A}_{1,\,j}, \dots,\ \boldsymbol{A}_{k,\,j}$是独立同分布的随机变量，对每个 $i \in [k]$，其分布为$\boldsymbol{A}_{i,\,j} \sim$ 泊松分布(λ)。因此，$\boldsymbol{A}_{i,\,j}$ 的分布与$\boldsymbol{X}_{i,\,j}$相同，所以 $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_k)$ 的分布与 $(\boldsymbol{x}'_1, \dots, \boldsymbol{x}'_k)$ 相同。

现在，对于每个 $i \in [k]$ 和 $j \in [N]$，令 $\boldsymbol{Y}_{i,\,j} \sim$ 独立服从 $\mathrm{Ber}(\rho)$。根据迹的定义，我们有

$$\boldsymbol{y}_i = \psi(\pi)_1^{\boldsymbol{Y}_{i,1}} \psi(\pi)_2^{\boldsymbol{Y}_{i,2}} \cdots \psi(\pi)_N^{\boldsymbol{Y}_{i,N}} .$$

因此，只要证明 $\boldsymbol{X}_{i,\,j}$ 与$\boldsymbol{Y}_{i,\,j}$ 足够相似，我们就完成了。

具体而言，固定某个 $i \in [k]$ 和 $j \in [N]$，为方便起见，令 $\boldsymbol{X} = \boldsymbol{X}_{i,\,j}$ 和 $\boldsymbol{Y} = \boldsymbol{Y}_{i,\,j}$。我们声称 $\mathsf{dist}_{\mathsf{TV}}(\boldsymbol{X},\ \boldsymbol{Y}) \le \lambda^2$。事实上，首先注意到，在给定 $\boldsymbol{X} \le 1$ 的条件下， $\boldsymbol{X}$ 的分布与 $\boldsymbol{Y}$ 的分布相同：

$$\mathbb{P}\left[\boldsymbol{X} = 1 \ \mid \ \boldsymbol{X} \le 1\right] = \frac{\mathbb{P}\left[\boldsymbol{X} = 1\right]}{\mathbb{P}\left[\boldsymbol{X} = 0\right] + \mathbb{P}\left[\boldsymbol{X} = 1\right]} = \frac{\lambda e^{-\lambda}}{e^{-\lambda} + \lambda e^{-\lambda}} = \frac{\lambda}{1+\lambda} = \rho = \mathbb{P}\left[\boldsymbol{Y} = 1\right] ,$$

因此

$$\begin{aligned}\mathsf{dist}_{\mathsf{TV}}(\boldsymbol{X}, \boldsymbol{Y}) &= \mathbb{P}\left[\boldsymbol{X} \ge 2\right] = 1 - \mathbb{P}\left[\boldsymbol{X} = 0\right] - \mathbb{P}\left[\boldsymbol{X} = 1\right] = 1 - e^{-\lambda} - \lambda e^{-\lambda} = 1 - e^{-\lambda}(1+\lambda) \\ &\le 1 - (1-\lambda)(1+\lambda) = \lambda^2 .\end{aligned}$$

最后，由于 $\boldsymbol{X}_{i,j}$ 和 $\boldsymbol{Y}_{i,j}$ 都相互独立，我们得到

$$\mathsf{dist}_{\mathsf{TV}}((\boldsymbol{X}_{i,j})_{i\in[k],j\in[N]}, (\boldsymbol{Y}_{i,j})_{i\in[k],j\in[N]}) \le kN \cdot \lambda^2 < kN(2\rho)^2 < 1/100 ,$$

并且 t 因此 $\mathsf{dist}_{\mathsf{TV}}((\boldsymbol{x}_1, \dots, \boldsymbol{x}_k), (\boldsymbol{y}_1, \dots, \boldsymbol{y}_k)) < 1/100$ 由数据处理不等式 . □

我们用这一过程给出轨迹测试的一般下界：

引理 7.24。 设 $\alpha > 0$。设 $N, k \in \mathbb{N}$，并设 Π_1, Π_2 是 $\mathbb{N}$ 上的概率分布性质，其密度是 $1/N$ 的整数倍，并且在奇偶迹下进行 (Π_1, Π_2, α)-性质测试所需的样本复杂度至少为 m。那么，任何使用 k 条迹、期望大小为 ρN 的 $(\Psi(\Pi_1)\ \Psi(\Pi_2)\ \rho, \alpha + 1/100)$-迹测试器都必须满足 $k\rho N = \Omega\left(\text{最小值}\ \left(m, \sqrt{kN}\right)\right)$。

证明。 假设 B 是一个使用 k 条迹、期望大小为 ρN 的 $(\Psi(\Pi_1)\ \Psi(\Pi_2)\ \alpha+1/100)$-迹测试器，并假设 $k\rho N < \sqrt{kN}/20$。那么 $\rho < \frac{1}{20\sqrt{kN}}$，我们的目标是证明 $k\rho N = \Omega(m)$。为此，我们通过在奇偶迹模型中构造一个算法 A，并证明 A 是奇偶迹下的 (Π_1, Π_2, α)-测试器，其样本复杂度为 $O(k\rho N)$。算法如下：

1. 取一个大小为 $\mathrm{Poi}\left(k \cdot \frac{\rho}{1-\rho} \cdot N\right)$ 的奇偶迹。
2. 应用 SPLIT，得到 k 个二进制字符串 $x_1, \dots, x_k$。
3. 在输入 $x_1, \dots, x_k$ 上返回 B 的输出。

注意，A 的样本复杂度为 $O(k\rho N)$。令 π 为输入分布；回忆一下，若 $\pi \in \Pi_1$，则 $\psi(\pi) \in \Psi(\Pi_1)$；若 $\pi \in \Pi_2$，则 $\psi(\pi) \in \Psi(\Pi_2)$。令 $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k$ 为馈送给 B 的输入，并令 $\boldsymbol{y}_1, \dots, \boldsymbol{y}_k$ 为来自 $\psi(\pi)$ 的相互独立的迹，每条迹的期望大小为 ρN。根据命题 7.23，$\mathsf{dist}_{\mathsf{TV}}((\boldsymbol{x}_1, \dots, \boldsymbol{x}_k), (\boldsymbol{y}_1, \dots, \boldsymbol{y}_k)) < 1/100$；并且我们知道，如果向 B 提供输入 $\boldsymbol{y}_1, \dots, \boldsymbol{y}_k$，它以至少 $\alpha + 1/100$ 的概率成功（即，若 $\psi(\pi) \in \Psi(\Pi_1)$ 则接受，若 $\psi(\pi) \in \Psi(\Pi_2)$ 则拒绝）。因此，A 以至少 α 的概率成功，并且 $k\rho N = \Omega(m)$。

□

现在我们得到了定理 1.7和1.9中所述的下界。

定理 7.25。存在一个普适常数 $C > 0$，使得以下结论成立。设 n、$N \in \mathbb{N}$、ρ 和 $\epsilon > 0$ 满足：n 为偶数且整除 N，ϵ 小于某个足够小的普适常数，并且 $N \geq C \cdot \max\left\{(n/\epsilon)^{8/5}, n/\epsilon^4\right\}$。设 Ψ_1 仅包含 $\{0,1\}^N$ 中均匀分布的 n-块字符串，并设 Ψ_2 为所有与 Ψ_1 的（相对）编辑距离至少为 ϵ 的 n-块字符串构成的集合。那么，任何使用 k 条迹、期望大小为 ρN 的$(\Psi_1, \Psi_2, \rho, 2/3)$-迹测试器都必须满足 $k\bar{\rho}N = \tilde{\Omega}((n/\epsilon)^{4/5} + \sqrt{n/\epsilon^2})$。

证明。令 Π_1 仅包含 $[n]$上的均匀分布 $\pi^* = \psi^{-1}(u^{(1)})$。为方便起见，令 $\epsilon^* := 8\epsilon$。使用下文定义的值 $\epsilon' \in (\epsilon^*, 2\epsilon^*]$，令 Π_2 为满足以下条件的$[n]$ 上的分布集合：1）在全变差距离下与均匀分布相距(ϵ'/c)（其中 c 是引理 6.31 中的常数）；2）所有密度均属于集合 $\left\{\frac{1}{n}, \frac{1-4\epsilon'/c}{n}, \frac{1+4\epsilon'/c}{n}\right\}$。注意，根据引理 6.31，$\Pi_2$ 中的每个分布在编辑距离下都与均匀分布相距 ϵ'。

我们将 ϵ' 定义为满足以下条件的最小 $\epsilon' > \epsilon^*$：$\frac{4\epsilon'/c}{n}$ 是 $1/N$的整数倍，并断言$\epsilon' \leq 2\epsilon^*$。事实上，我们有

$$\frac{4\epsilon^*/c}{n} \geq \frac{1}{N} \iff N \geq \frac{cn}{4\epsilon^*},$$

根据假设，当 C充分大时该不等式成立，因此在 $\frac{4\epsilon^*/c}{n}$ 和 $\frac{8\epsilon^*/c}{n}$ 之间存在 $1/N$的整数倍。

由于定理 5.1 的证明仅使用形如 Π_2 的分布，因此在奇偶迹下，对（Π_1，Π_2，$51/100$）-分布进行性质测试需要 $\tilde{\Omega}\left((n/\epsilon)^{4/5} + \sqrt{n/\epsilon^2}\right)$ 个样本。因此，注意到根据假设我们有 $\sqrt{kN} = \Omega\left((n/\epsilon)^{4/5} + \sqrt{n/\epsilon^2}\right)$，引理 7.24 表明，任何使用 k 条期望大小为 ρN 的迹的（$\Psi(\Pi_1)$ $\Psi(\Pi_2)$ $52/100$）-迹测试器都必须满足 $k\rho N = \tilde{\Omega}((n/\epsilon)^{4/5} + \sqrt{n/\epsilon^2})$。只要证明 $\Psi(\Pi_1) \subseteq \Psi_1$ 且 $\Psi(\Pi_2) \subseteq \Psi_2$，结论便可得出。

首先，假设 $x \in \Psi$ (Π_1). Th在 $x = \psi(\pi^*) = \psi(\psi^{-1}(u^{(1)})) = u^{(1)}$ and h性 $x \in \Psi_1$。

其次，假设 $x \in \Psi(\Pi_2)$，因此对于某个 $\pi \in \Pi_2$，有 $x = \psi(\pi)$。我们声称 $x \in \Psi_2$。显然，x 是一个 n-块字符串，因此只需证明 $\mathsf{dist}_{\mathsf{rel-edit}}(x, \Psi_1) > \epsilon$。由于 $\pi \in \Pi_2$，我们有 $\mathsf{dist}_{\mathsf{TV}}(\pi, \pi^*) \geq \epsilon'/c$，因此，如前所述，$\mathsf{dist}_{\mathsf{edit}}(\pi, \pi^*) \geq \epsilon' > \epsilon^*$。由引理 7.6，$\mathsf{dist}_{\mathsf{rel-edit}}(x, u^{(1)}) = \mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi)\ \psi(\pi^*)) \geq \mathsf{dist}_{\mathsf{edit}}(\pi, \pi^*) > \epsilon^* > \epsilon$。我们还需要证明 $\mathsf{dist}_{\mathsf{rel-edit}}(x, u^{(0)}) > \epsilon$。分两种情况讨论。

首先，假设 $\epsilon^* \geq 4/n$。根据三角不等式，$\mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi)\ u^{(1)}) \leq \mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi)\ u^{(0)}) + \mathsf{dist}_{\mathsf{rel-edit}}(u^{(0)}, u^{(1)}) = \mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi)\ u^{(0)}) + 2/n$。因此 $\mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi)\ u^{(0)}) > \epsilon^* - 2/n > \epsilon$。

其次，假设 $\epsilon^* < 4/n$，于是 $1/2n > \epsilon^*/8$。根据构造，$\psi(\pi)$ 的第一个块长度至少为 $N \cdot \frac{1-4\epsilon'/c}{n}$，对于足够小的 ϵ，这至少为 $N \cdot \frac{1}{2n}$。此外，$\psi(\pi)$ 的这个第一个块是 1-块，而 $u^{(0)}$ 的第一个块是长度为 $N \cdot \frac{1}{n}$ 的 0 块。因此 $\mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi)\ u^{(0)}) \geq 1/2n > \epsilon^*/8 = \epsilon$，如所需。因此 $x \in \Psi_2$，证明完毕。

□

定理 7.26. 对于所有充分小的常数 $\epsilon > 0$，以下结论成立。存在一个函数 $N(n) = \Theta(n^2)$，使得对于所有 n、$k \in \mathbb{N}$ 和 $N \geq N(n)$，以下结论成立。令 Ψ 为

n-块字符串在 $\{0,1\}^N$ 中组成的集合。那么，任何使用 k 条迹、期望大小为 ρN 的 $(\Psi, \mathrm{FAR}_\epsilon^{\mathsf{rel-edit}}(\Psi)$ $\rho, 2/3)$-迹测试器都必须满足 $k\rho N = \Omega(n/\log n)$。

证明。令 Π_1 为支撑集大小至多为 n的 $\mathbb{N}$ 上的概率分布集合。根据定理 6.38 以及 [BFH21], 中关于测试 n-交替函数的下界，在奇偶迹下，任何 $(\Pi_1, \mathrm{FAR}_{2\epsilon}^{\mathsf{edit}}(\Pi_1)$,$51/100)$-分布测试器的样本复杂度都必须为 $\Omega(n/\log n)$。事实上，更强的结论成立：对于某个足够大的普适常数 $C > 0$，令 Π_2为 $\mathrm{FAR}_{2\epsilon}^{\mathsf{edit}}(\Pi_1)$在支撑集大小至多为 Cn的那些分布上的限制。那么，在奇偶迹下，任何 $(\Pi_1, \Pi_2,$ $51/100)$-分布测试器的样本复杂度都必须为 $\Omega(n/\log n)$。这是因为 [BFH21] 中关于测试 n-交替函数的下界，是通过从 [VV11, WY19], 中的支持集大小区分问题进行归约证明的，而该问题的困难样例的支撑集大小与 n成线性关系。

应用引理 7.24 时，我们需要所有密度都是 $1/N$ 的整数倍。设 Π_1' 是通过对每个分布 $\pi \in \Pi_1$ 的所有密度进行舍入得到的性质：将其舍入为 $1/N$ 的倍数，并确保得到另一个概率分布 π'。设 Π_2' 是以同样方式从 Π_2 得到的性质。于是，每个 $\pi \in \Pi_1$ 都满足 $\mathsf{dist}_{\mathsf{TV}}(\pi, \Pi_1')$ $\leq \frac{1}{N} \cdot O(n)$，且每个 $\pi \in \Pi_2$ 都满足 $\mathsf{dist}_{\mathsf{TV}}(\pi, \Pi_2') \leq \frac{1}{N} \cdot O(n)$。

我们声称，在奇偶迹下，任何 $(\Pi_1', \Pi_2', 60/100)$-分布测试器的样本复杂度都必须为 $\Omega(n/\log n)$。设 A 是一个在奇偶迹下的 $(\Pi_1', \Pi_2', 60/100)$-分布测试器，其样本复杂度为 $m \leq n/\log n$。接下来我们证明， A 是一个在奇偶迹下的 $(\Pi_1, \Pi_2, 55/100)$-分布测试器。对于任意输入 $\pi \in \Pi_1$，都存在 $\pi' \in \Pi_1'$ 使得 $\mathsf{dist}_{\mathsf{TV}}(\pi, \pi') = O(n/N)$。于是，随机变量 $\boldsymbol{S} \sim \mathsf{samp}(\pi, m)$ 和 $\boldsymbol{S'} \sim \mathsf{samp}(\pi', m)$ 满足 $\mathsf{dist}_{\mathsf{TV}}(\boldsymbol{S}, \boldsymbol{S'}) \leq m \cdot \mathsf{dist}_{\mathsf{TV}}(\pi, \pi') = O(n^2/N \log n) = o(1)$。由于 A 接受 π' 的概率至少为 $60/100$，因此它接受 π 的概率至少为 $60/100 - o(1) \geq 55/100$。对 Π_2 和 Π_2' 也可用同样的论证，因此 A 是一个在奇偶迹下的 $(\Pi_1, \Pi_2, 55/100)$-分布测试器。由此得到 $m = \Omega(n/\log n)$，证毕。

现在，注意到根据假设，我们有 $\sqrt{Nk} = \Omega(n)$，由引理 7.24 可知，任何使用 k 个迹、期望大小为 ρN 的 $(\Psi$ (Π_1'),$\Psi(\Pi_2')$ $61/100)$-迹测试器都必须满足 $k\rho N = \Omega(n/\log n)$。只要证明 $\Psi(\Pi_1') \subseteq \Psi$ 且 $\Psi(\Pi_2') \subseteq \mathrm{FAR}_\epsilon^{\mathsf{rel-edit}}(\Psi)$, 即可得到结论。

首先，令 $x \in \Psi(\Pi_1')$，则对于某个 $\pi_1' \in \Pi_1'$，有 $x = \psi(\pi_1')$。由于从 Π_1得到 Π_1' 的过程不会向分布的支撑集中添加任何新元素，因此每个 $\pi_1' \in \Pi_1'$ 的支持集大小至多为 n，所以 x 是一个 n-块字符串。因此 $\Psi(\Pi_1') \subseteq \Psi$。

其次，令 $x \in \Psi(\Pi_2')$，于是对某个 $\pi_2' \in \Pi_2'$ 有 $x = \psi(\pi_2')$。我们需要证明 $\mathsf{dist}_{\mathsf{rel-edit}}(x, \Psi) > \epsilon$。首先，我们声称 $\mathsf{dist}_{\mathsf{edit}}(\pi_2', \Pi_1) > \epsilon$，也就是说， π_2' 与任何支撑集大小至多为 n 的分布（不一定经过舍入）之间的编辑距离至少为 ϵ。反设存在 $\pi_1 \in \Pi_1$ 使得 $\mathsf{dist}_{\mathsf{edit}}(\pi_1, \pi_2') \leq \epsilon$。令 $\pi_2 \in \Pi_2$ 为某个满足 $\mathsf{dist}_{\mathsf{TV}}(\pi_2, \pi_2') = O(n/N)$ 的分布；根据 Π_2' 的构造，这样的分布确实存在。于是 $\mathsf{dist}_{\mathsf{edit}}$ $(\pi_2, \pi_2') = O(n/N)$，并且由于 ϵ 是常数，我们可以假设 $\mathsf{dist}_{\mathsf{edit}}(\pi_2, \pi_2') < \epsilon/2$。那么

$$\mathsf{dist}_{\mathsf{edit}}(\pi_1, \pi_2) \leq \mathsf{dist}_{\mathsf{edit}}(\pi_1, \pi_2') + \mathsf{dist}_{\mathsf{edit}}(\pi_2', \pi_2) < \epsilon + \epsilon/2 < 2\epsilon\,,$$

这与 Π_2 的定义矛盾。因此 $\mathsf{dist}_{\mathsf{edit}}(\pi_2', \Pi_1) > \epsilon$。现在，令 $y \in \Psi$。我们声称 $\mathsf{dist}_{\mathsf{rel-edit}}(x, y) > \epsilon$。由于 y 是一个 n-块字符串，我们有 $\psi^{-1}(y) \in \Pi_1$，因此 $\mathsf{dist}_{\mathsf{edit}}(\pi_2', \psi^{-1}(y)) > \epsilon$。接着，由引理 7.6 可得 $\mathsf{dist}_{\mathsf{rel-edit}}(x, y) =$ $\mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi_2')\ \psi(\psi^{-1}(y))) \geq \mathsf{dist}_{\mathsf{edit}}(\pi_2', \psi^{-1}(y)) > \epsilon$。因此 $\mathsf{dist}_{\mathsf{rel-edit}}(x, \Psi) > \epsilon$，进而 $\Psi(\Pi_2') \subseteq \mathrm{FAR}_\epsilon^{\mathsf{rel-edit}}(\Psi)$, 证明得证。 □

致谢

我们感谢 Eric Blais 就本文的呈现所提出的宝贵讨论和意见，并感谢 Maryam Aliakbarpour 提供有关不完美信息条件下性质测试的参考文献。我们也感谢匿名审稿人提出的意见，并提供相关工作的参考文献。

参考资料

[ACF+21] Jayadev Acharya、Cl´ement L Canonne、Cody Freitag、Ziteng Sun 和 Himanshu Tyagi。信息约束下的推断 III：局部隐私约束。*IEEEJournal on Selected Areas in Information Theory*，2(1):253–267，2021年。[ACFT19] Jayadev Acharya、Cl´ement Canonne、Cody Freitag 和 Himanshu Tyagi。无需信任的测试：最优的局部隐私分布检验。载于人工智能与统计国际会议论文集（*AISTATS*），第 2067–2076 页。PMLR，2019年。[ACT19] Jayadev Acharya、Cl´ement L Canonne 和 Himanshu Tyagi。信息约束下的推断：基于卡方收缩的下界。载于学习理论会议论文集（*COLT*），第 3–17 页。PMLR，2019年。[ACT20] Jayadev Acharya、Cl´ement L Canonne 和 Himanshu Tyagi。信息约束下的推断 II：通信约束与共享随机性。*IEEE* 信息论汇刊，66(12):7856–7877，2020年。[ADK15] Jayadev Acharya、Constantinos Daskalakis 和 Gautam Kamath。分布性质的最优检验。神经信息处理系统进展，28，2015年。[AFCK23] Sepehr Assadi、Martin Farach-Colton 和 William Kuszmaul。单调最小完美哈希的紧界。载于*2023*年 *ACM-SIAM* 离散算法年度研讨会（*SODA*）论文集，第 456–476 页。SIAM，2023年。[AKNS01] Noga Alon、Michael Krivelevich、Ilan Newman 和 Mario Szegedy。正则语言可通过常数次查询进行测试。*SIAM* 计算学杂志，30(6):1842–1862，2001年。[Bat01] Tugkan Batu。分布性质测试。康奈尔大学，2001年。[BBBY12] Maria-Florina Balcan、Eric Blais、Avrim Blum 和 Liu Yang。主动性质测试。载于*IEEE* 计算机科学基础研讨会（*FOCS*）论文集，第 21–30 页。IEEE，2012年。[BFH21] Eric Blais、Renato Ferreira Pinto Jr 和 Nathaniel Harms。VC 维数与分布无关的基于样本的检验。载于第 53 届 ACM SIGACT 计算理论研讨会论文集，第 504–517 页，2021 年。[BH18] Avrim Blum 和 Lunjia Hu。主动容错测试。载于第 31 届学习理论会议论文集，2018年。

[BKKM04] Tu˘gkan Batu、Sampath Kannan、Sanjeev Khanna 和 Andrew McGregor。从随机迹重构字符串。载于 ACM-SIAM 离散算法研讨会（SODA）论文集，第 910–918 页，2004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。载于第三十六届 ACM 计算理论研讨会论文集，第 381–390 页，2004年。[BPRS20] Vinnu Bhardwaj、Pavel A Pevzner、Cyrus Rashtchian 和 Yana Safonova。计算生物学中的迹重构问题。IEEE 信息论汇刊，67(6):3295–3314，2020年。[BS21] Gabriel Bathie 和 Tatiana Starikovskaya。正则语言的性质测试及其在可见下推语言流式性质测试中的应用。载于 ICALP 2021，2021年。[Can17] Cl´ement Canonne。泊松尾界简札，2017年。http://www.cs. columbia.edu/~ccanonne/files/misc/2017-poissonconcentration.pdf。[Can22] Cl´ement Canonne。分布测试中的主题与技术。预印本，https://ccanonne. github.io/files/misc/main-survey-fnt. pdf，2022年。[CDGR18] Cl´ement L Canonne、Ilias Diakonikolas、Themis Gouleakis 和 Ronitt Rubinfeld。离散分布形状约束的测试。计算理论系统，62(1):4–62，2018年。[CDL+22] Xi Chen、Anindya De、Chin Ho Lee、Rocco A Servedio 和 Sandip Sinha。利用少量迹实现近乎最优的平均情形近似迹重构。载于2022年 ACM-SIAM 离散算法研讨会（SODA）论文集，第 779–821 页。SIAM，2022年。[CDL+23] Xi Chen、Anindya De、Chin Ho Lee、Rocco A Servedio 和 Sandip Sinha。单条轨迹的近似迹重构。载于2023年 ACM-SIAM 离散算法研讨会（SODA）论文集，第 605–637 页。SIAM，2023年。[CDVV14] Siu-On Chan、Ilias Diakonikolas、Paul Valiant 和 Gregory Valiant。检验离散分布接近度的最优算法。载于第二十五届 ACM-SIAM 离散算法研讨会论文集，第 1193–1203 页。SIAM，2014年。[CFG+22] Sourav Chakraborty、Eldar Fischer、Arijit Ghosh、Gopinath Mishra 和 Sayantan Sen。巨型对象模型中的索引不变性质测试。arXiv 预印本 arXiv:2207.12514，2022年。[CGG+19] Cl´ement L Canonne、Elena Grigorescu、Siyao Guo、Akash Kumar 和 Karl Wimmer。测试 k-单调性：布尔函数的兴衰。计算理论，15(1):1–55，2019年。[Cha21a] Zachary Chase。迹重构的新下界。载于《亨利·庞加莱研究所年鉴：概率与统计》，第 57 卷，第 627–643 页。亨利·庞加莱研究所，2021年。[Cha21b] Zachary Chase。分离词与迹重构。载于 ACM SIGACT 计算理论研讨会（STOC）论文集，第 21–31 页，2021年。

[CP22] Xi Chen 和 Shyamal Patel。半空间的分布无关测试（几乎）需要 PAC 学习。载于《2022年 ACM-SIAM 离散算法年度研讨会（SODA）论文集》，第 1715–1743 页。SIAM，2022年。[CW20] Cl´ement L Canonne 和 Karl Wimmer。数据分箱测试。载于 APPROX/RANDOM 论文集。Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl 出版社，2020年。[CW21] Cl´ement L Canonne 和 Karl Wimmer。标签不匹配下的恒等性检验。载于第 32 届国际算法与计算研讨会（ISAAC 2021）。Schloss Dagstuhl-Leibniz-Zentrum f¨ur Informatik，2021年。[DGPP18] Ilias Diakonikolas、Themis Gouleakis、John Peebles 和 Eric Price。高概率下样本最优的恒等性检验。载于自动机、语言与编程国际研讨会（ICALP）论文集。Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik，2018年。[DGPP19] Ilias Diakonikolas、Themis Gouleakis、John Peebles 和 Eric Price。基于碰撞的测试器在均匀性和接近性检验中是最优的。《Chicago Journal of Theoretical Com- puter Science》，1：1–21，2019年。[DK16] Ilias Diakonikolas 和 Daniel M Kane。离散分布性质测试的新方法。载于 IEEE 计算机科学基础研讨会（FOCS）论文集，第 685–694 页。IEEE，2016年。[DKN15a] Ilias Diakonikolas、Daniel M Kane 和 Vladimir Nikishkin。结构化分布接近性检验的最优算法与下界。载于 IEEE 计算机科学基础研讨会（FOCS）论文集，第 1183–1202 页。IEEE，2015年。[DKN15b] Ilias Diakonikolas、Daniel M Kane 和 Vladimir Nikishkin。结构化分布的恒等性检验。载于 ACM-SIAM 离散算法研讨会（SODA）论文集，第 1841–1854 页。SIAM，2015年。[DOS17] Anindya De、Ryan O’Donnell 和 Rocco A Servedio。用于迹重构的最优基于均值的算法。载于 ACM SIGACT计算理论研讨会（STOC）论文集，第 1047–1056 页，2017年。[ES20] Rogers Epstein 和 Sandeep Silwal。LP 型问题的性质测试。载于 Artur Czumaj、Anuj Dawar 和 Emanuela Merelli 编，第 47 届自动机、语言与编程国际研讨会（ICALP 2020），Leibniz 国际信息学论文集（LIPIcs）第 168 卷，第 98:1–98:18 页，德国达格施图尔，2020年。Schloss Dagstuhl–Leibniz-Zentrum f¨ur Informatik。[Fis04] Eldar Fischer。性质测试中比较的作用。《信息与计算》，189(1)：107–116，2004年。[FKKT21] Dimitris Fotakis、Alkis Kalavasis、Vasilis Kontonis 和 Christos Tzamos。利用粗粒度标签进行学习的高效算法。载于学习理论会议论文集，第 2060–2079 页。PMLR，2021年。[GGR98] Oded Goldreich、Shari Goldwasser 和 Dana Ron。性质测试及其与学习和近似的联系。《ACM 学报（JACM）》，45(4)：653–750，1998年。

[GKK23] Aravind Gollakota、Adam R Klivans 和 Pravesh K Kothari。用于可检验学习的矩匹配方法及 Rademacher 复杂度的新刻画。2023年。即将发表。[GR00] Oded Goldreich 和 Dana Ron。有界度图扩张性的测试。载于计算复杂性电子讨论会（ECCC），第 20 卷，2000年。[GR16] Oded Goldreich 和 Dana Ron。关于基于样本的测试器。《ACM 计算理论汇刊》，8(2)：1–54，2016年。[GR18] Marco Gaboardi 和 Ryan Rogers。局部隐私假设检验：卡方检验。载于国际机器学习会议（ICML）论文集，第 1626–1635 页。PMLR，2018年。[GR22] Oded Goldreich 和 Dana Ron。巨型对象分布的测试。载于理论计算机科学创新会议（ITCS）论文集。Schloss Dagstuhl-Leibniz-Zentrum f¨ur Informatik，2022年。[Gra06] Robert M Gray。托普利兹矩阵与循环矩阵：综述。《通信与信息论的基础与趋势》，2(3)：155–239，2006年。[GSS21] Friedrich G¨otze、Holger Sambale 和 Arthur Sinulis。关于 α次指数随机变量多项式的集中不等式。《电子概率期刊》，26(48)，2021年。[GSZ22] Elena Grigorescu、Madhu Sudan 和 Minshen Zhu。小编辑距离下基于均值的迹重构算法的局限性。《 IEEE 信息论汇刊》，2022 年。[Har19] Nathaniel Harms。旋转不变分布上的半空间测试。载于第三十届 ACM-SIAM 离散算法年会论文集，第 694–713 页。SIAM，2019年。[HL20] Nina Holden 和 Russell Lyons。迹重构的下界。《应用概率年刊》，30(2)：503–525，2020年。[HMPW08] Thomas Holenstein、Michael Mitzenmacher、Rina Panigrahy 和 Udi Wieder。常数删除概率下的迹重构及相关结果。载于 ACM-SIAM 离散算法研讨会（SODA）论文集，第 389–398 页。Citeseer，2008年。[KMMP21] Akshay Krishnamurthy、Arya Mazumdar、Andrew McGregor 和 Soumyabrata Pal。迹重构：广义化与参数化。《IEEE 信息论汇刊》，67(6)：3233–3250，2021年。[KR00] Michael Kearns 和 Dana Ron。具有子学习样本复杂度的测试问题。《计算机与系统科学期刊》，61(3)：428–456，2000年。[Lev01] Vladimir I Levenshtein。序列的高效重构。《IEEE 信息论汇刊》，47(1)：2–22，2001年。[MORS10] Kevin Matulef、Ryan O’Donnell、Ronitt Rubinfeld 和 Rocco A Servedio。半空间测试。《SIAM 计算学杂志》，39(5)：2004–2047，2010年。

[Nee14] Joe Neeman. 任意准确度下的表面积检验。载于《计算理论研讨会》（STOC 2014），第393–397页。ACM，2014年。[NP17] Fedor Nazarov 和 Yuval Peres。使用 exp(o(n1/3)) 个样本进行迹重构。载于《第49届 ACM SIGACT 年度计算理论研讨会论文集》，第1042–1046页，2017年。[Pan08] Liam Paninski。针对极稀疏采样离散数据的基于重合的均匀性检验。IEEE 信息论汇刊，54(10):4750–4755，2008年。[Ram09] Frank P Ramsey。形式逻辑中的一个问题。载于《组合学经典论文》，第1–24页。Springer，2009年。[Rio37] John Riordan。二项、泊松和超几何频数分布的矩递推关系。《数学统计年刊》，8(2):103–111，1937年。[RR20] Dana Ron 和 Asaf Rosin。k-模态性的近乎最优分布无关的基于样本的检验。载于 Jaros-law Byrka 和 Raghu Meka 编，《近似、随机化与组合优化：算法与技术（APPROX/RANDOM 2020）》，Leibniz 国际信息学论文集（LIPIcs）第176卷，第27:1–27:19页，德国达格施图尔，2020年。Schloss Dagstuhl–Leibniz 信息学中心。[RR21] Dana Ron 和 Asaf Rosin。子序列无关性的最优分布无关基于样本检验。载于 ACM-SIAM 离散算法研讨会论文集（SODA），第337–256页。SIAM，2021年。[RRSS09] Sofya Raskhodnikova、Dana Ron、Amir Shpilka 和 Adam Smith。近似分布支持集大小和不同元素问题的强下界。SIAM 计算学杂志，39(3):813–842，2009年。[Rub22] Ittai Rubinstein。迹重构问题从平均情形到（平移后的）最坏情形的归约。arXiv 预印本 arXiv:2207.11489，2022年。[RV23] Ronitt Rubinfeld 和 Arsen Vasilyan。学习算法分布假设的检验。载于计算理论研讨会（STOC 2023）。ACM，2023年。即将发表。[She18] Or Sheffet。局部隐私假设检验。载于国际机器学习会议（ICML）论文集，第4605–4614页。PMLR，2018年。[Sra] Suvrit Sra。将某个实对称托普利兹矩阵对角化。MathOverflow。https://mathoverflow.net/q/68471（版本：2011-07-05）。[VV11] Gregory Valiant 和 Paul Valiant。未见部分估计：一种用于估计熵和支持集大小的 $n/\log(n)$ 样本估计量，并通过新的中心极限定理证明其最优性。载于 ACM SIGACT 计算理论研讨会论文集（STOC），第685–694页，2011年。[VV17a] Gregory Valiant 和 Paul Valiant。自动不等式证明器与实例最优的恒等性检验。SIAM 计算学杂志，46(1):429–455，2017年。[VV17b] Gregory Valiant 和 Paul Valiant。未见部分估计：熵和其他性质的改进估计量。ACM 学报（JACM），64(6):1–41，2017年。

[WY19] Yihong Wu 和 杨鹏坤。切比雪夫多项式、矩匹配与未见部分的最优估计。《统计年刊》，47(2):857–883，2019年。

A 泊松化与增强

A.0.1 泊松化

通常（例如参见 [VV11、VV17a、WY19]）会在“泊松化”设置下分析分布测试算法：算法不直接获取 m 独立样本，而是先采样$\boldsymbol{m} \sim \mathrm{Poi}(m)$，然后再获取 $\boldsymbol{m}$ 独立样本。我们稍微滥用记号，直接称测试器获取 $\mathrm{Poi}(m)$ 个样本。这一技术的优点在于，每个定义域元素在样本中出现的次数变得相互独立。从分布 π 上定义域 $\mathcal{X}$ 中的分布获取 $\mathrm{Poi}(m)$ 个独立样本，等价于从每个 $x \in \mathcal{X}$中分别相互独立地获取 $\mathrm{Poi}(m\pi(x))$ 个样本。由于 $\mathrm{Poi}(m)$ 高度集中在 m 附近，因此可以在保留样本复杂度上界和下界的同时，在泊松化模型与非泊松化模型之间来回转换。下面简要陈述与本文相关的转换，并请读者参阅例如 [Can22 附录 C] 及其中的参考文献以了解详细信息：

命题 A.1。 我们称一个算法为 泊松化的（Π_1， Π_2， α）奇偶迹下的分布测试器，其样本复杂度为 m，如果它满足与定义 2.3相同的条件，但它抽取的是大小为 泊松分布（m）的样本（的奇偶迹），而不是 m。那么对所有 $\delta > 0$，以下结论成立：

1. 如果存在一个样本复杂度为 m 的（标准）（$\Pi_1, \Pi_2, 1-\delta/2$）-奇偶迹下的分布测试器，那么就存在一个样本复杂度为 $\max\{2m, 12\log(4/\delta)\}$ 的泊松化（$\Pi_1, \Pi_2, 1-\delta$）-奇偶迹下的分布测试器。

2. 如果存在一个样本复杂度为 m 的泊松化（$\Pi_1, \Pi_2, 1-\delta/2$）-奇偶迹下的分布测试器，那么就存在一个样本复杂度为 $\max\{\frac{3}{2}m, 18\log(4/\delta)\}$ 的（标准）（$\Pi_1, \Pi_2, 1-\delta$）-奇偶迹下的分布测试器。

我们同样可以通过调整定义 2.2，在标准版本和泊松化版本之间转换混淆收集器模型，并注意到该模型也具有类似的结果。

A.0.2 增强成功概率

在标准分布测试中，通常可以通过放大将算法的成功概率提升到任意所需的水平：重复运行算法多次，并采取多数表决。

在奇偶迹模型中，算法只能接收单个样本的迹，因此无法简单地多次重复测试。因此，我们需要一种不同的技术来增强成功概率。通过采用更大的原始样本，测试器可以执行“样本拆分”，生成若干相互独立的迹，然后分别对其进行测试，具体如下所述。

回顾一下，在（泊松化的）奇偶迹模型中，当测试器从π中抽取 $\mathrm{Poi}(m)$个样本时，它会接收到一条迹

$$\boldsymbol{\mathcal{T}} = 1^{\boldsymbol{A}_1}0^{\boldsymbol{B}_1}\dots 1^{\boldsymbol{A}_n}0^{\boldsymbol{B}_n},$$

其中，每个 $\boldsymbol{A}_i \sim$泊松分布(mp_i) 和每个 $\boldsymbol{B}_i \sim$ 泊松分布(mq_i) 都相互独立。

事实 A.2. 设 $\lambda > 0$ 和 $k \in \mathbb{N}$。通过以下概率过程定义随机变量 $\boldsymbol{X}_1$、$\ldots$、$\boldsymbol{X}_k$：

1. 抽取 $X \leftarrow$泊松分布$(k\lambda)$；

2. 抽取 $X_1, \ldots, X_k \leftarrow$ 多项分布$(X, (1/k, \ldots, 1/k))$。

那么，$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ 是相互独立的随机变量，并且对每个 $i \in [k]$，$\boldsymbol{X}_i \sim$ 泊松分布(λ)。

因此，我们可以通过以下方式模拟大小为 Poi(m) 的 k 个奇偶迹：

1. 抽取一个大小为 Poi(mk) 的奇偶迹 $\boldsymbol{\mathcal{T}}$；以及
2. 从左到右，将 $\boldsymbol{\mathcal{T}}$中的每个符号依次分配给 $\boldsymbol{\mathcal{T}}_{\boldsymbol{j}}$，其中 $\boldsymbol{j}$ 是从 $[k]$ 中相互独立地均匀随机选取的。

那么，上述事实意味着 j^{th} 迹的分布为

$$\boldsymbol{\mathcal{T}}_j = 1^{\boldsymbol{A}_{j,1}} 0^{\boldsymbol{B}_{j,1}} \ldots 1^{\boldsymbol{A}_{j,n}} 0^{\boldsymbol{B}_{j,n}}$$

其中，对于所有 i 和 j，$\boldsymbol{A}_{j,i} \sim$Poi($mp_i$) 与 $\boldsymbol{B}_{j,i} \sim$Poi($mq_i$) 相互独立，正如所需。

因此，在奇偶迹下，测试器的成功概率可以提升到任意水平 $1-\delta$，代价是样本复杂度增加一个 Θ(log($1/\delta$)) 的乘法因子。

注释 A.3. 在困惑收集器模型中，无法以通常意义上的方式提升概率，因为算法无法控制其分辨率参数 η。

B 第 3 节中缺失的证明 第 3 节

命题 B.1. 设 $\eta \in (0,1)$。令 S_i 表示 ϕ^{path}中第 i 列各项之和（对每个 $i \in \mathbb{Z}_n$），并令 $h := (n-1)/2$。那么，对于任意非负整数 $N \le n/2$,

$$-\sum_{i=0}^{\lceil N/2 \rceil - 1} S_i - \sum_{i=0}^{\lfloor N/2 \rfloor - 1} S_{n-1-i} + \sum_{i=0}^{\lceil N/2 \rceil - 1} S_{\lceil h \rceil + i} + \sum_{i=0}^{\lfloor N/2 \rfloor - 1} S_{\lfloor h \rfloor - i} < \frac{2}{\eta^2}.$$

证明。回顾一下，$\phi^{\mathsf{path}}_{i,j} = \nu^{|i-j|}$ 其中 $\nu = 1-\eta$。我们将列和明确写出，并

化简由此产生的几何级数：

$$
\begin{aligned}
&\left[-\sum_{i=0}^{\lceil N/2\rceil-1} S_i-\sum_{i=0}^{\lfloor N/2\rfloor-1} S_{n-1-i}+\sum_{i=0}^{\lceil N/2\rceil-1} S_{\lceil h\rceil+i}+\sum_{i=0}^{\lfloor N/2\rfloor-1} S_{\lfloor h\rfloor-i}\right] \\
&=\left[\begin{array}{l}-\sum_{i=0}^{\lceil N/2\rceil-1}\sum_{j=0}^{n-1}\nu^{|i-j|}-\sum_{i=0}^{\lfloor N/2\rfloor-1}\sum_{j=0}^{n-1}\nu^{|n-1-i-j|} \\ +\sum_{i=0}^{\lceil N/2\rceil-1}\sum_{j=0}^{n-1}\nu^{|\lceil h\rceil+i-j|}+\sum_{i=0}^{\lfloor N/2\rfloor-1}\sum_{j=0}^{n-1}\nu^{|\lfloor h\rfloor-i-j|}\end{array}\right] \\
&=\left[\begin{array}{l}-\sum_{i=0}^{\lceil N/2\rceil-1}\left(\sum_{j=0}^{i}\nu^{i-j}+\sum_{j=i+1}^{n-1}\nu^{j-i}\right) \\ -\sum_{i=0}^{\lfloor N/2\rfloor-1}\left(\sum_{j=0}^{n-1-i}\nu^{n-1-i-j}+\sum_{j=n-i}^{n-1}\nu^{j+i+1-n}\right) \\ +\sum_{i=0}^{\lceil N/2\rceil-1}\left(\sum_{j=0}^{\lceil h\rceil+i}\nu^{\lceil h\rceil+i-j}+\sum_{j=\lceil h\rceil+i+1}^{n-1}\nu^{j-i-\lceil h\rceil}\right) \\ +\sum_{i=0}^{\lfloor N/2\rfloor-1}\left(\sum_{j=0}^{\lfloor h\rfloor-i}\nu^{\lfloor h\rfloor-i-j}+\sum_{j=\lfloor h\rfloor-i+1}^{n-1}\nu^{j+i-\lfloor h\rfloor}\right)\end{array}\right] \\
&=\left[\begin{array}{l}-\sum_{i=0}^{\lceil N/2\rceil-1}\left(\frac{\nu^i-\nu^{-1}}{1-\nu^{-1}}+\frac{\nu-\nu^{n-i}}{1-\nu}\right)-\sum_{i=0}^{\lfloor N/2\rfloor-1}\left(\frac{\nu^{n-1-i}-\nu^{-1}}{1-\nu^{-1}}+\frac{\nu-\nu^{i+1}}{1-\nu}\right) \\ +\sum_{i=0}^{\lceil N/2\rceil-1}\left(\frac{\nu^{\lceil h\rceil+i}-\nu^{-1}}{1-\nu^{-1}}+\frac{\nu-\nu^{n-i-\lceil h\rceil}}{1-\nu}\right)+\sum_{i=0}^{\lfloor N/2\rfloor-1}\left(\frac{\nu^{\lfloor h\rfloor-i}-\nu^{-1}}{1-\nu^{-1}}+\frac{\nu-\nu^{n+i-\lfloor h\rfloor}}{1-\nu}\right)\end{array}\right] \\
&=\frac{1}{\eta}\left[\begin{array}{l}-\sum_{i=0}^{\lceil N/2\rceil-1}\left(1-\nu^{i+1}+\nu-\nu^{n-i}\right)-\sum_{i=0}^{\lfloor N/2\rfloor-1}\left(1-\nu^{n-i}+\nu-\nu^{i+1}\right) \\ +\sum_{i=0}^{\lceil N/2\rceil-1}\left(1-\nu^{\lceil h\rceil+1+i}+\nu-\nu^{n-i-\lceil h\rceil}\right)+\sum_{i=0}^{\lfloor N/2\rfloor-1}\left(1-\nu^{\lfloor h\rfloor+1-i}+\nu-\nu^{n+i-\lfloor h\rfloor}\right)\end{array}\right] \\
&=\frac{1}{\eta}\left[\begin{array}{l}\sum_{i=0}^{\lceil N/2\rceil-1}\nu^{i+1}+\sum_{i=0}^{\lceil N/2\rceil-1}\nu^{n-i}+\sum_{i=0}^{\lfloor N/2\rfloor-1}\nu^{n-i}+\sum_{i=0}^{\lfloor N/2\rfloor-1}\nu^{i+1} \\ -\sum_{i=0}^{\lceil N/2\rceil-1}\nu^{\lceil h\rceil+1+i}-\sum_{i=0}^{\lceil N/2\rceil-1}\nu^{n-i-\lceil h\rceil}-\sum_{i=0}^{\lfloor N/2\rfloor-1}\nu^{\lfloor h\rfloor+1-i}-\sum_{i=0}^{\lfloor N/2\rfloor-1}\nu^{n+i-\lfloor h\rfloor}\end{array}\right] \\
&=\frac{1}{\eta}\left[\begin{array}{l}\frac{\nu-\nu^{\lceil N/2\rceil+1}}{1-\nu}+\frac{\nu^n-\nu^{n-\lceil N/2\rceil}}{1-\nu^{-1}}+\frac{\nu^n-\nu^{n-\lfloor N/2\rfloor}}{1-\nu^{-1}}+\frac{\nu-\nu^{\lfloor N/2\rfloor+1}}{1-\nu} \\ -\frac{\nu^{\lceil h\rceil+1}-\nu^{\lceil h\rceil+1+\lceil N/2\rceil}}{1-\nu}-\frac{\nu^{n-\lceil h\rceil}-\nu^{n-\lceil h\rceil-\lceil N/2\rceil}}{1-\nu^{-1}}-\frac{\nu^{\lfloor h\rfloor+1}-\nu^{\lfloor h\rfloor+1-\lfloor N/2\rfloor}}{1-\nu^{-1}}-\frac{\nu^{n-\lfloor h\rfloor}-\nu^{n+\lfloor N/2\rfloor-\lfloor h\rfloor}}{1-\nu}\end{array}\right] \\
&=\frac{1}{\eta^2}\left[\begin{array}{l}\nu-\nu^{\lceil N/2\rceil+1}+\nu^{n+1-\lceil N/2\rceil}-\nu^{n+1} \\ +\nu^{n+1-\lfloor N/2\rfloor}-\nu^{n+1}+\nu-\nu^{\lfloor N/2\rfloor+1} \\ -\nu^{\lceil h\rceil+1}+\nu^{\lceil h\rceil+1+\lceil N/2\rceil}-\nu^{n+1-\lceil h\rceil-\lceil N/2\rceil}+\nu^{n+1-\lceil h\rceil} \\ -\nu^{\lfloor h\rfloor+2-\lfloor N/2\rfloor}+\nu^{\lfloor h\rfloor+2}-\nu^{n-\lfloor h\rfloor}+\nu^{n+\lfloor N/2\rfloor-\lfloor h\rfloor}\end{array}\right] \\
&=\frac{1}{\eta^2}\left[\begin{array}{l}(\nu+\nu)+(\nu^{n+1-\lceil N/2\rceil}-\nu^{\lceil N/2\rceil+1})+(\nu^{n+1-\lfloor N/2\rfloor}-\nu^{\lfloor N/2\rfloor+1})-(\nu^{n+1}+\nu^{n+1}) \\ +(\nu^{\lceil h\rceil+1+\lceil N/2\rceil}-\nu^{\lceil h\rceil+1})+(\nu^{n+1-\lceil h\rceil}-\nu^{n+1-\lceil h\rceil-\lceil N/2\rceil}) \\ +(\nu^{\lfloor h\rfloor+2}-\nu^{\lfloor h\rfloor+2-\lfloor N/2\rfloor})+(\nu^{n+\lfloor N/2\rfloor-\lfloor h\rfloor}-\nu^{n-\lfloor h\rfloor})\end{array}\right] \\
&<\frac{2}{\eta^2},
\end{aligned}
$$

其中我们用到了以下事实： $0<\nu<1$ ，并且在最后一步用到了 $2\lceil N/2\rceil\le n$ （这是因为根据假设，$2N\le n$ ）。 □

C 省略证明第4节

C.1 均匀性检验：小 ϵ 情形

我们在此证明引理 4.2。用于均匀分布检验的标准算法仅依据样本的直方图 作出判断。直方图是元组 $(X_1,\ldots,X_{2n})$，其中 X_i 表示元素 $i\in[2n]$ 在样本中出现的次数。对于样本 S，我们将 $H(S)$ 记作其直方图。著名的均匀性检验结果可表述如下：

定理 C.1 ([VV17a, DGPP19])。存在一个常数 $C > 0$ 和一个算法 UniformityHistogramTester（简称 UHT），使得对于任意 $n \in \mathbb{N}$、$\epsilon > 0$ 和 $m \geq C \cdot \frac{\sqrt{n}}{\epsilon^2}$：

1. 若 $\pi = \pi(\mu, \mu)$，$\mathbb{P}[\mathrm{UHT}(H(S)) \text{ 接受 }] \geq 3/4$；并且，2. 若 $\pi = \pi(p, q)$ 与均匀分布相差 ϵ，$\mathbb{P}[\mathrm{UHT}(H(S) \text{ 拒绝 }] \geq 3/4$。

命题 C.2. 设 $\pi = \pi(\mu, \mu)$ 为 $[2n]$ 上的均匀分布，并设 $m \geq 2 \cdot n \log(100n)$。令 $\boldsymbol{S} \sim$ 采样(π, m)。那么

$$\mathbb{P}[\exists i \in [2n] : i \notin \boldsymbol{S}] < 1/50 .$$

证明。根据并集界，并利用以下事实：对于任意 $i \in [2n]$，$\mathbb{P}[i \notin \boldsymbol{S}] = \left(1 - \frac{1}{2n}\right)^m \leq e^{-\frac{m}{2n}} = e^{-\log(100n)} = \frac{1}{100n}$。□

算法 3 当 $\epsilon < \frac{K}{\log^3_{n^{1/4}} n}$ 时的均匀性检验器。

设定 $m \leftarrow \Theta\left(\frac{\sqrt{n}}{\epsilon^2} \log^7 n\right)$。**常数：** $K = K_{\alpha, \beta, \gamma} > 1$ 如 算法 2。**输入：** 对于定义域 $[n]$ 上的 $\pi = \pi(p, q)$，接收样本 $S \leftarrow \mathsf{samp}(\pi, m)$ 的迹(S)**要求：** $\epsilon < \frac{K}{\log^3_{n^{1/4}} n}$。

p过程 UniformityTesterSmall(p, q, n, ϵ) 设 $X_1, \ldots, X_n$ 为迹中 1 的游程长度，如第 2.3 节所定义。设 $X'_1, \ldots, X'_n$ 为迹中 0 的游程长度，如第 2.3 节所述。**如果** $\exists i \in [n]$ 使得 $X_i = 0$ 或 $X'_i = 0$ **则** **拒绝** **否则** **输出** UniformityHistogramTester$(X_1, X'_1, \ldots, X_n, X'_n)$

我们需要以下关于全变差距离的事实。

事实 C.3。设 $\mathcal{D}$ 是一个概率分布，E 是同一概率空间中的一个事件。记 $\mathcal{D}_{|E}$ 为一个随机变量在给定 E 的条件下服从 $\mathcal{D}$ 时的概率分布。那么

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{D}_{|E}) \leq \mathbb{P}[\neg E] .$$

下面给出 引理 4.2 的证明。

Le引理 C.4。假设 $\epsilon < \frac{K}{\log^3_{n^{1/4}} n}$，那么 *算法 3* 满足以下条件：

1. 如果 $\pi = \pi(\mu, \mu)$，该算法将以至少 $2/3$ 的概率接受。
2. 如果 $\pi = \pi(p, q)$ 与均匀分布的距离至少为 ϵ，那么该算法将以至少 $2/3$ 的概率拒绝。

证明。根据 定理 C.1，我们知道，对于适当选择的常数 $C > 0$，如果 $m \geq C \cdot \frac{\sqrt{n}}{\epsilon^2}$，那么该 UHT 算法正确的概率至少为 $3/4$。在我们的情形中，m 满足此条件。

令 $\boldsymbol{S} \sim$ 采样(π, m) 和 $\boldsymbol{T} =$ 迹$(\boldsymbol{S})$。定义

$$(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{2n}) := (\boldsymbol{X}_1, \boldsymbol{X}'_1, \boldsymbol{X}_2, \boldsymbol{X}'_2, \ldots, \boldsymbol{X}_n, \boldsymbol{X}'_n)\,,$$

使得 $\boldsymbol{Z}$ 是 $\boldsymbol{T}$ 中游程长度的向量。

将 A 记作事件 $\boldsymbol{Z}_i > 0$ 对所有 $i \in [2n]$。注意，如果事件 A 发生，那么 $\boldsymbol{Z}$ 就是直方图 $H(\boldsymbol{S})$。假设 $\pi = \pi(\mu, \mu)$。我们首先论证，即使其输入来自于在事件 A条件下的样本，UHT 被拒绝的概率也很小。令 $\boldsymbol{S}'$ 为服从在 $\boldsymbol{S}$ 发生条件下的样本 A 的分布的随机变量。于是，根据 事实 C.3 和 命题 C.2,

$$\mathsf{dist}_{\mathsf{TV}}(\boldsymbol{S}, \boldsymbol{S}') \le \mathbb{P}\left[\neg A\right] < \frac{1}{50}\,.$$

因此 $\mathbb{P}\left[\mathrm{UHT}(H(\boldsymbol{S})) \neq \mathrm{UHT}(H(\boldsymbol{S}'))\right] < 1/50$ 并且

$$\begin{aligned}
\mathbb{P}\left[\mathrm{UHT}(\boldsymbol{Z}) \text{ rejects} \;\middle|\; A\right] &= \mathbb{P}\left[\mathrm{UHT}(H(\boldsymbol{S})) \text{ rejects} \;\middle|\; A\right] \\
&= \mathbb{P}\left[\mathrm{UHT}(H(\boldsymbol{S}')) \text{ rejects}\right] \\
&< \mathbb{P}\left[\mathrm{UHT}(H(\boldsymbol{S})) \text{ rejects}\right] + \frac{1}{50} < \frac{1}{4} + \frac{1}{50}\,.
\end{aligned}$$

那么，算法 3 拒绝的概率为

$$\begin{aligned}
&\mathbb{P}\left[\textsc{UniformityTesterSmall}(\boldsymbol{Z}) \text{ rejects}\right] \\
&\quad = \mathbb{P}\left[A\right]\mathbb{P}\left[\mathrm{UHT}(\boldsymbol{Z}) \text{ rejects} \;\middle|\; A\right] + \mathbb{P}\left[\neg A\right] < \frac{1}{4} + \frac{2}{50} < \frac{1}{3}\,.
\end{aligned}$$

现在假设 $\pi = \pi(p, q)$ 与均匀分布相距 ϵ。由于当 A 发生时 $\boldsymbol{Z} = H(\boldsymbol{S})$，我们有

$$\begin{aligned}
\mathbb{P}\left[\textsc{UniformityTesterSmall}\text{（}\boldsymbol{Z}\text{）拒绝} \;\middle|\; A\right] &= \mathbb{P}\left[\mathrm{UHT}\text{（}\boldsymbol{Z}\text{）拒绝} \;\middle|\; A\right] \\
&= \mathbb{P}\left[\mathrm{UHT}(H(\boldsymbol{S})) \text{ rejects} \;\middle|\; A\right].
\end{aligned}$$

此外，由于算法 3 在 A 未出现时总是拒绝，因此我们有

$$1 = \mathbb{P}\left[\textsc{UniformityTesterSmall}(\boldsymbol{Z}) \text{ rejects} \;\middle|\; \neg A\right] \ge \mathbb{P}\left[\mathrm{UHT}(H(\boldsymbol{S})) \text{ rejects} \;\middle|\; \neg A\right].$$

因此，算法 3拒绝的概率为

$$\begin{aligned}
&\mathbb{P}\left[\textsc{UniformityTesterSmall}(\boldsymbol{Z}) \text{ rejects}\right] \\
&\quad = \mathbb{P}\left[A\right]\mathbb{P}\left[\textsc{UniformityTesterSmall}(\boldsymbol{Z}) \text{ rejects} \;\middle|\; A\right] \\
&\qquad + \mathbb{P}\left[\neg A\right]\mathbb{P}\left[\textsc{UniformityTesterSmall}(\boldsymbol{Z}) \text{ rejects} \;\middle|\; \neg A\right] \\
&\quad \ge \mathbb{P}\left[A\right]\mathbb{P}\left[\mathrm{UHT}(H(\boldsymbol{S})) \text{ rejects} \;\middle|\; A\right] + \mathbb{P}\left[\neg A\right]\mathbb{P}\left[\mathrm{UHT}(H(\boldsymbol{S})) \text{ rejects} \;\middle|\; \neg A\right] \\
&\quad = \mathbb{P}\left[\mathrm{UHT}(H(\boldsymbol{S})) \text{ rejects}\right] > 3/4 > 2/3\,,
\end{aligned}$$

由此证明得证。 □

D 编辑距离证明

D.1 关于编辑距离和带标签分布的事实

事实 D.1. 设 $\mathcal{D}_f$ 和 $\mathcal{D}_g$ 是定义域 $\mathcal{X}$上的带标签分布。那么

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_g) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}\left[f(x) \neq g(x)\right].$$

证明。为简化记号，假设 $\mathcal{X}$ 是可数的。利用命题6。 6,

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_g) = \frac{1}{2}\sum_x \mathbb{1}\left[f(x) \neq g(x)\right](2\mathcal{D}(x)) = \sum_x \mathbb{1}\left[f(x) \neq g(x)\right]\cdot\mathcal{D}(x) = \mathop{\mathbb{P}}_{x\sim\mathcal{D}}\left[f(x) \neq g(x)\right].$$

□

命题 D.2。 对于定义域为 $\mathbb{Z}$ 的任意适当的带标签分布$(f, \mathcal{D})$，以及 $\mathbb{N}$ 上的任意分布 π，存在 $\mathbb{Z}$ 上的一个分布 $\mathcal{E}$，使得 $\pi = \pi_{f,\mathcal{E}}$ 并且

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{E}) = \mathsf{dist}_{\mathsf{TV}}(\pi_{f,\mathcal{D}}, \pi).$$

证明。令 $a_1 < a_2 < \cdots$ 为 f 的交替序列，并采用约定 $a_0 = -\infty$；若该序列的有限长度为 t，还定义 $a_{t+1} = \infty$。对于每个区间 $I = (a_{i-1}, a_i]$，我们按如下方式定义点 $x \in I$ 上的 $\mathcal{E}$。

- 如果 $\mathcal{D}(I) \leq \pi(i)$，任取一点 $x^* \in I$。对于 $x \in I \setminus \{x^*\}$，令 $\mathcal{E}(x) \leftarrow \mathcal{D}(x)$。然后令 $\mathcal{E}(x^*) \leftarrow \pi(i) - \sum_{x\in I\setminus\{x^*\}} \mathcal{D}(x)$。注意到 $\mathcal{E}(I) = \pi(i)$，如所需，并且

$$\begin{aligned}\sum_{x\in I}|\mathcal{D}(x) - \mathcal{E}(x)| &= \mathcal{E}(x^*) - \mathcal{D}(x^*) + \left(\sum_{x\in I\setminus\{x^*\}}\mathcal{E}(x)\right) - \left(\sum_{x\in I\setminus\{x^*\}}\mathcal{D}(x)\right)\\ &= |\mathcal{E}(I) - \mathcal{D}(I)|.\end{aligned}$$

- 如果 $\mathcal{D}(I) > \pi(i)$，令 $\mathcal{E}(x) \leftarrow \mathcal{D}(x) - \delta_x$ 任意取一组满足 $0 \leq \delta_x \leq \mathcal{D}(x)$ 和 $\sum_{x\in I}\delta_x = \mathcal{D}(I) - \pi(i)$ 的值 δ_x；容易验证这样的取值存在。注意到，如所需，

$$\mathcal{E}(I) = \mathcal{D}(I) - \sum_{x\in I}\delta_x = \pi(i),$$

and

$$\sum_{x\in I}|\mathcal{E}(x) - \mathcal{D}(x)| = \sum_{x\in I}\delta_x = |\mathcal{D}(I) - \pi(i)| = |\mathcal{D}(I) - \mathcal{E}(I)|.$$

现在我们有 $\pi_{f,\ \mathcal{E}} = \pi$，并且

$$\begin{aligned}\mathsf{dist}_{\mathsf{TV}}(\pi_{f,\mathcal{D}}, \pi) = \mathsf{dist}_{\mathsf{TV}}(\pi_{f,\mathcal{D}}, \pi_{f,\mathcal{E}}) &= \frac{1}{2}\sum_{i=1}^{k+1}|\pi_{f,\mathcal{D}}(i) - \pi_{f,\mathcal{E}}(i)| = \frac{1}{2}\sum_{i=1}^{k+1}|\mathcal{D}(a_{i-1}, a_i] - \mathcal{E}(a_{i-1}, a_i]|\\ &= \frac{1}{2}\sum_x|\mathcal{D}(x) - \mathcal{E}(x)| = \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_f).\end{aligned}$$

□

事实 D.3. 设 a_1、…、$a_m \geq 0$ 和 $b \geq 0$。那么存在 b_1、…、b_m，使得 $\sum_i b_i = b$ 且 $\sum_i |a_i - b_i| = |b - \sum_i a_i|$。

证明. 首先假设 $\sum_i a_i \le b$。然后令 $b_i = a_i$ 为 $i < m$ 和 $b_m = b - \sum_{i<m} a_i$。于是 $\sum_i b_i = b$ 和 $\sum_i |a_i - b_i| = b_m - a_m = b - \left(\sum_{i<m} a_i\right) - a_m = b - \sum_i a_i$，如所需。

现在假设 $\sum_i a_i > b$。令 j 为满足 $\sum_{i\le j} a_i > b$的最小数。对$i < j$赋值 $b_i = a_i$，对 $i > j$赋值 $b_i = 0$，$b_j = b - \sum_{i<j} a_i$，于是 $\sum_i b_i = b$ 并且

$$\sum_i |a_i - b_i| = (a_j - b_j) + \sum_{i>j} a_i = a_j - b + \sum_{i<j} a_i + \sum_{i>j} a_i = \left(\sum_i a_i\right) - b\,. \qquad \square$$

事实 D.4. 设 $(f, \mathcal{D})$ 和 $(g, \mathcal{E})$ 是任意两个适当的带标签分布，并设 g' 是满足以下条件的任意函数：当$\min(\mathcal{D}(x)\ \mathcal{E}(x)) > 0$时，$g'(x) = g(x)$。那么

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_{g'}) = \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g)\,.$$

证明。这可由 命题 6.6得出。 $\square$

事实 D.5. 设 $(f, \mathcal{D})$ 和 $(g, \mathcal{E})$ 是任意两个适当标记分布。那么存在 $(f', \mathcal{D}')$ 和 $(g', \mathcal{E}')$，满足 $\pi_{f',\mathcal{D}'} = \pi_{f,\mathcal{D}}$，$\pi_{g',\mathcal{E}'} = \pi_{g,\mathcal{E}}$，且$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_{f'}, \mathcal{E}'_{g'}) \le \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g)$，并满足以下条件：

1. 如果 I 是任意一个区间，使得 f' 和 g' 在 I 上都是常数，并且对所有 $x \in I$ 都有 $f'(x) \ne g'(x)$，那么要么 $\mathcal{D}'(I) = 0$，要么 $\mathcal{E}'(I) = 0$。
2. f' 和 g' 没有共同的交替点。

证明。设 $a_1 < a_2 < \cdots$ 和 $b_1 < b_2 < \cdots$ 分别是 f 和 g 的交替序列。不妨设 f 和 g 没有共同的交替点。这是因为如果 $a_i = b_j$，我们可以将所有值 $b_{j'} \ge b_j$ 和 $a_{i'} > a_i$ 都加 1，将所有密度 $\mathcal{D}(x)$ 和 $\mathcal{E}(x)$ 分别向右移动一个位置，用于 $x > a_i$ 和 $y > b_j$，并重新定义 $\mathcal{D}(a_i + 1) = \mathcal{E}(b_j + 1) = 0$。

任何区间 I，若 f 和 g 在 I 上均为常数，且 $f(x) \ne g(x)$ 在所有 $x \in I$上都成立，则必须满足 $I \subset I^*$，其中 $I^* := (a_{i-1}, a_i] \cap (b_{j-1}, b_j]$，对应某些交替点 a_i, b_j。由于 f和 g没有任何共同的交替点，因此，要么存在 $z \in (a_{i-1}, a_i]$，使得$f(x) = g(x)$；要么存在 $z \in (b_{j-1}, b_j]$，使得 $f(z) = g(z)$。在第一种情况下，定义 $\mathcal{D}'$，使其除 $I^* \cup \{z\}$外与 $\mathcal{D}$ 相同；并定义 $\mathcal{D}'(x) = 0$，其中 $x \in I^*$，以及 $\mathcal{D}'(z) = \mathcal{D}(z) + \mathcal{D}(I^*)$。由于$\{z\} \cup I^* \subset (b_{j-1}, b_j]$，我们有 $\mathcal{D}'(b_{j-1}, b_j] = \mathcal{D}(b_{j-1}, b_j]$，因此 $\pi_{f,\mathcal{D}'} = \pi_{f,\mathcal{D}}$。注意，

$$|\mathcal{D}'(z) - \mathcal{E}(z)| + \sum_{x \in I^*} (\mathcal{D}'(x) + \mathcal{E}(x)) = |\mathcal{D}(z) + \mathcal{D}(I^*) - \mathcal{E}(z)| + \mathcal{E}(I^*) \le |\mathcal{D}(z) - \mathcal{E}(z)| + \mathcal{D}(I^*) + \mathcal{E}(I^*)\,,$$

因此 距离$_{\mathsf{TV}}(\mathcal{D}'_f, \mathcal{E}_g) \le$距离$_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g)$ 根据 命题 6.6。在第二种情况中，其中 $z \in (b_{j-1}, b - j]$，我们对 $\mathcal{E}$执行类似的调整，从而得到 $\mathcal{E}'$。 $\square$

事实 D.6. 设$(f, \mathcal{D})$ 和$(g, \mathcal{E})$ 是任意两个适当的带标签分布。设 $a_1 < a_2 < \cdots$为 f的交替序列，$b_1 < b_2 < \cdots$为 g的交替序列。那么存在分布 $\mathcal{D}'$ 和 $\mathcal{E}'$，满足以下条件：

1. $\pi_{f,\ \mathcal{D}} = \pi_{f,\ \mathcal{D}'}$ 和 $\pi_{g,\ \mathcal{E}'} = \pi_{g,\ \mathcal{E}}$；2. $\mathcal{D}'$ 和 $\mathcal{E}'$ 支撑于集合 $C = \{a_i\} \cup \{b_j\}$；3. 距离$_{\mathsf{TV}}(\mathcal{D}'_f, \mathcal{E}'_g) \le$距离$_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g)$。

证明。定义 $\mathcal{D}'$和 $\mathcal{E}'$如下。记 $C = \{c_1, c_2, \dots\}$，其中 $c_1 \le c_2 \le c_3 \le \cdots$。对于每个区间 $(c_{i-1}, c_i] \ne \emptyset$，定义 $\mathcal{E}'(c_i) = \mathcal{E}(c_{i-1}, c_i]$ 和 $\mathcal{D}'(c_i) = \mathcal{D}(c_{i-1}, c_i]$。很容易验证所需的性质。

$\square$

D.2 编辑距离定义的等价性

我们必须证明第6.2节中的以下引理。

引理 6.11. 设 π 和 π' 是 $\mathbb{N}$ 上的有限支持分布。那么

$$\mathsf{dist}_{\mathsf{edit}}(\pi,\pi') = \inf \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\mathcal{E}_g)\,,$$

其中，下确界取遍满足 $\pi = \pi_{f,\mathcal{D}}$ 和 $\pi' = \pi_{g,\mathcal{E}}$ 的带标签分布 $(f,\mathcal{D})$ 和 $(g,\mathcal{E})$。

我们将使用以下事实。这些事实很容易验证，只需交换相邻的允许操作（并相应地调整索引）即可。

事实 D.7. 设 a 是任意分数字符串，O_1、……、O_k 是对 a 执行的任意一列允许操作。那么，存在一列对 a 执行的允许操作 O'_1、……、O'_k，使得

$$(O_k \circ O_{k-1} \circ \cdots \circ O_1)(a) = (O'_k \circ O'_{k-1} \circ \cdots \circ O'_1)(a)$$

并且，对于某个 $0 \le i \le j \le k+1$，对所有 $\ell \le i$，O_ℓ 都是 插入 或 重排 操作；对所有 $i < \ell < j$，O_ℓ 都是 调整 操作；对所有$\ell \ge j$，O_ℓ 都是 删除 或 重排 操作。

事实 D.8. 令 a 为任意分数字符串，令 $O_1,\ldots,O_k$ 为任意一列允许的操作，其中每个操作都是插入 或重排 操作。则存在一列 $O'_1,\ldots,O'_k$ 允许的操作，使得$(O'_k \circ \cdots \circ O'_1)(a) = (O_k \circ \cdots \circ O_1)(a)$，并且满足以下条件。存在某个 $s \le k$，使得对于所有 $j > s$，每个 O_j 都是插入 操作；对于所有 $j \le s$，每个 O_j 都是重排 操作，或是形如插入 的$\mathsf{ins}_{i,b}$ 操作，其中 $b = a_{i-1}$ 或 $b = a_i$。

引理的证明 引理 *6.11*。对于一个带标签分布（f，$\mathcal{D}$），其中 $\mathcal{D}$ 是有限支撑的，存在 $k \in \mathbb{N}$使得 $\mathcal{D}(i) = 0$ 对于所有满足 $i \in \mathbb{Z}$ 且 $|i| > k$的情况，我们定义

$$\mathsf{str}(\mathcal{D}_f) := (f(-k))^{\mathcal{D}(-k)}(f(1-k))^{\mathcal{D}(1-k)}\cdots(f(k))^{\mathcal{D}(k)}\,.$$

上界。设（f，$\mathcal{D}$）和（g，$\mathcal{E}$）是任意两个带标签分布，满足 $\mathcal{D}$和 $\mathcal{E}$支撑于有限集，且 $\pi_{f,\mathcal{D}} = \pi$、$\pi_{g,\mathcal{E}} = \pi'$。（由于 π、π' 支撑于有限集，这样的带标签分布总是存在。）我们将分两步证明 $\mathsf{dist}_{\mathsf{edit}}(\pi,\pi') \le \|\mathcal{D}_f - \mathcal{E}_g\|_{\mathsf{TV}}$。首先，我们证明 $\mathsf{dist}_{\mathsf{fr-edit}}(\mathsf{str}(\pi),\mathsf{str}(\mathcal{D}_f)) = \mathsf{dist}_{\mathsf{fr-edit}}(\mathsf{str}(\pi'),\mathsf{str}(\mathcal{E}_g)) = 0$。其次，我们证明 $\mathsf{dist}_{\mathsf{fr-edit}}(\mathsf{str}(\mathcal{D}_f),\mathsf{str}(\mathcal{E}_g)) \le \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\mathcal{E}_g)$。由此，根据三角不等式即可得到结论。

步骤 *1.* 设 $a_1,\ldots,a_t$ 是 f 的交替点，其中不妨假设对所有 $x \le a_1$，$f(x) = 1$；并且由于 $\pi_{f,\mathcal{D}} = \pi$ 和 π 均为有限支撑的，不妨假设交替点的数量有限。令 $a_0 = -\infty$ 和 $a_{t+1} = \infty$，使得对所有 $i \in [t+1]$，$\mathcal{D}(a_{i-1},a_i] = \pi(i)$。对每个 $i \in [t+1]$，注意到对所有 $x \in (a_{i-1},a_i]$，$f(x) = \mathsf{par}(i)$。我们将 $\mathsf{str}(\pi)$ 中的每个字符 $(\mathsf{par}(i))^{\pi(i)}$ 替换为分数字符串

$$\begin{aligned}&(\mathsf{par}(i))^{\mathcal{D}(a_{i-1}+1)}(\mathsf{par}(i))^{\mathcal{D}(a_{i-1}+2)}\cdots(\mathsf{par}(i))^{\mathcal{D}(a_i)}\\ =&(f(a_{i-1}+1))^{\mathcal{D}(a_{i-1}+1)}(f(a_{i-1}+2))^{\mathcal{D}(a_{i-1}+2)}\cdots(f(a_i))^{\mathcal{D}(a_i)}\,,\end{aligned}$$

通过有限次插入和重排操作完成。对每个 $i \in [t+1]$ 重复此过程，便得到分数字符串 $\mathsf{str}(\mathcal{D}_f)$，且仅使用成本为 0 的操作。对 π' 和 $\mathcal{E}_g$ 重复相同的论证，可得类似结论，从而完成证明的第一步。

步骤 2。 存在 k, k' 使得

$$a := \mathsf{str}(\mathcal{D}_f) = (f(-k))^{\mathcal{D}(-k)}(f(1-k))^{\mathcal{D}(1-k)}\dots(f(k))^{\mathcal{D}(k)}$$
$$b' := \mathsf{str}(\mathcal{E}_g) = (g(-k'))^{\mathcal{E}(-k')}(g(1-k'))^{\mathcal{E}(1-k')}\dots(g(k'))^{\mathcal{E}(k')}.$$

不失一般性，假设 $k' \le k$ 并定义

$$b := (g(-k))^{\mathcal{E}(-k)}(g(1-k))^{\mathcal{E}(1-k)}\dots(g(k))^{\mathcal{E}(k)}.$$

容易看出， b 可以仅通过插入操作从 $b' = \mathsf{str}(\mathcal{E}_g)$得到，因为 $\mathcal{E}(x) = 0$ 对于$|x| > k'$。由命题 6.6 可得

$$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g) = \frac{1}{2}\sum_{-k\le i\le k}\left(\mathbb{1}\left[f(i)=g(i)\right]|\mathcal{E}(i)-\mathcal{D}(i)| + \mathbb{1}\left[f(i)\neq g(i)\right](\mathcal{E}(i)+\mathcal{D}(i))\right).$$

对于每个 $-k \le i \le k$，我们按如下方式编辑 a：

- 如果 $f(i) = g(i)$，则使用一次调整操作，将分数字符 $(f(i))^{\mathcal{D}(i)}$ 替换为 $(g(i))^{\mathcal{E}(i)}$，代价为 $\frac{1}{2}|\mathcal{D}(i) - \mathcal{E}(i)|$。
- 如果 $f(i) \neq g(i)$，则使用一次调整操作，将分数字符 $(f(i))^{\mathcal{D}(i)}$ 替换为 $(f(i))^0$，代价为 $\mathcal{D}(i)/2$；随后使用一次删除操作和一次插入操作，将 $(f(i))^0$ 替换为 $(g(i))^0$；最后使用一次调整操作，将 $(g(i))^0$ 替换为 $(g(i))^{\mathcal{E}(i)}$，代价为 $\mathcal{E}(i)/2$。总代价为 $\frac{1}{2}(\mathcal{D}(i) + \mathcal{E}(i))$。

所得字符串为 b，其生成成本为

$$\frac{1}{2}\sum_{-k\le i\le k}\left(\mathbb{1}\left[f(i)=g(i)\right]|\mathcal{E}(i)-\mathcal{D}(i)| + \mathbb{1}\left[f(i)\neq g(i)\right](\mathcal{E}(i)+\mathcal{D}(i))\right) = \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g).$$

结合三角不等式和步骤 1，我们现已证明

$$\mathsf{dist}_{\mathsf{edit}}(\pi, \pi') \le \inf \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g),$$

其中，下确界取遍所有满足 $\mathcal{D}$ 且 $\mathcal{E}$有限支撑的带标签分布 $(f, \mathcal{D})$ 和 $(g, \mathcal{E})$，并且 $\pi_{f,\mathcal{D}} = \pi$ 以及 $\pi_{g,\mathcal{E}} = \pi'$。为完成证明，我们必须允许带标签分布不是有限支撑的。注意到，对于任意带标签分布 $(f, \mathcal{D})$ 和任意 $\epsilon > 0$，我们都可以找到一个有限支撑的 $\mathcal{D}'$，使得 $\pi_{f,\mathcal{D}'} = \pi_{f,\mathcal{D}}$ 且 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}'_f) < \epsilon$。

下界。考虑任意一列允许的编辑操作 $O_1, \dots, O_k$，满足 $\mathsf{str}(\pi') = (O_k \circ O_{k-1} \circ \cdots \circ O_1)(\mathsf{str}(\pi))$。根据事实 D.7，我们不妨假设 $O_1, \dots, O_s$ 是插入和重排操作，而 $O_t, \dots, O_k$ 是删除和重排操作，其中某个 $s < t$。记 $a = (O_s \circ \cdots \circ O_1)(\mathsf{str}(\pi))$ 和 $b = (O_{t-1} \circ \cdots \circ O_1)(\mathsf{str}(\pi))$。于是，不失一般性，我们可以假设该序列具有事实 D.8 所述的形式，并将其中所述的索引记为 s'。对于每个 $j \in [s]$，记 $a^{(j)} = (O_j \circ \cdots \circ O_1)(\mathsf{str}(\pi))$。

我们将归纳地定义一个带标签分布序列 $(f^{(0)}, \mathcal{D}^{(0)})$，$(f^{(1)}, \mathcal{D}^{(1)})\dots$，$(f^{(s)}, \mathcal{D}^{(s)})$，使得 $\pi_{f^{(j)},\mathcal{D}^{(j)}} = \pi$ 对于每个 j 都成立；并且对于每个 $a^{(j)} = (O_j \circ \cdots \circ O_1)(\mathsf{str}(\pi))$，若$a^{(j)} = (a_1^{(j)})^{p_1^{(j)}} \cdots (a_n^{(j)})^{p_n^{(j)}}$，我们还将有 $\mathcal{D}^{(j)}(i) = p_i^{(j)}$，且除非 $p_i^{(j)} = 0$，否则有 $f^{(j)}(i) = a_i^{(j)}$。

对所有 $i \in \mathbb{N}$，定义 $f^{(0)}(i) = \mathrm{par}(i)$；并对 $i \le 1$，定义 $f^{(0)}(i) = 1$。对 $i \in \mathbb{N}$，定义 $\mathcal{D}^{(0)}(i) = \pi(i)$；其他情况下，定义 $\mathcal{D}^{(0)}(i) = 0$。根据定义，$\pi_{f^{(0)},\ \mathcal{D}^{(0)}} = \pi$。

对于每个 $j \in [s']$，其中 $a^{(j)} = (a_1^{(j)})^{p_1^{(j)}} \cdots (a_n^{(j)})^{p_n^{(j)}}$，我们将 $(f^{(j)}, \mathcal{D}^{(j)})$ 简单定义为 $f^{(j)}(i) = a_i^{(j)}$ 和 $\mathcal{D}^{(j)}(i) = p_i^{(j)}$。考虑操作 O_j。如果 O_j 是一个 重排 操作，那么 $f^{(j)} = f^{(j-1)}$，因为符号均未发生变化。如果 O_j 是一个 插入 操作，那么它会插入一个与其前一个或后一个符号相等的符号。无论哪种情况，$f^{(j)}$ 的交替点数量都与 $f^{(j-1)}$的交替点数量相同，并且 i^{th} 和 $(i+1)^{th}$ 各自交替点之间的 $\mathcal{D}^{(j)}$ 和 $\mathcal{D}^{(j-1)}$ 的质量不变。因此 $\pi_{f^{(j)},\mathcal{D}^{(j)}} = \pi_{f^{(j-1)},\mathcal{D}^{(j-1)}} = \pi$。

对于剩余的操作 O_j with $s' < j \le s$，我们知道 O_j 是一个插入 操作。在紧邻 i^{th} 分数字符之前插入一个新的分数字符时，我们通过将 $i' \ge i$ 的所有值 $f^{(j-1)}(i')$ 和 $\mathcal{D}^{(j-1)}(i')$ 向右移动一位，将 $f^{(j-1)}$更改为 $f^{(j)}$，并将 $\mathcal{D}^{(j)}$ 更改为 $\mathcal{D}^{(j-1)}$。然后我们定义 $\mathcal{D}^{(j)}(i) = 0$ 并设置 $f^{(j)}(i) = f^{(j-1)}(i-1)$，这不会增加交替点的数量。我们再次得到 $\pi_{f^{(j)},\mathcal{D}^{(j)}} = \pi_{f^{(j-1)},\mathcal{D}^{(j-1)}} = \pi$。在这种情况下，我们可能有 $f^{(j)}(i) \ne a_i^{(j)}$，但我们有 $p_i^{(j)} = 0$，因为这是一个插入 操作。

现在，我们有一个带标签分布 $(f, \mathcal{D}) := (f^{(s)}, \mathcal{D}^{(s)})$，使得 $\pi_{f,\mathcal{D}} = \pi$，并且对于所有 $i \in \mathbb{N}$，都有 $f(i) = a_i^{(s)} = a_i$，除非 $p_i^{(s)} = p_i = 0$；且对于所有 $i \in \mathbb{N}$，$\mathcal{D}(i) = p_i^{(s)} = p_i$（这进一步推出，对于 $i \notin \mathbb{N}$，$\mathcal{D}(i) = 0$）。注意，分数字符串 $b = (O_{t-1} \circ \cdots \circ O_1)(\mathrm{str}(\pi))$ 只能通过插入和重排操作从 $\mathrm{str}(\pi')$ 得到，因此，应用相同的论证可得 $(g, \mathcal{E})$，使得 $\pi_{g,\mathcal{E}} = \pi'$，并且对于所有 $i \in \mathbb{N}$，$g(i) = b_i$，除非 $q_i = 0$；且 $\mathcal{E}(i) = p_i$。

现在，对于所有 i，我们必须有 $a_i = b_i$，因为 b 仅通过 调整 操作由 a 得到。这些 调整 操作的代价至少为 $\frac{1}{2}\sum_i |p_i - q_i|$。另一方面，我们有

$$\begin{aligned}
\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g) &= \frac{1}{2}\sum_i \left(\mathbb{1}\left[f(i) = g(i)\right]|\mathcal{D}(i) - \mathcal{E}(i)| + \mathbb{1}\left[f(i) \ne g(i)\right](\mathcal{D}(i) + \mathcal{E}(i))\right) \\
&= \frac{1}{2}\sum_i \left(\mathbb{1}\left[f(i) = g(i)\right]|p_i - q_i| + \mathbb{1}\left[f(i) \ne g(i)\right](p_i + q_i)\right) \\
&= \frac{1}{2}\sum_{i:p_i=0 \text{ or } q_i=0} \left(\mathbb{1}\left[f(i) = g(i)\right]|p_i - q_i| + \mathbb{1}\left[f(i) \ne g(i)\right]|p_i - q_i|\right) \\
&\quad + \frac{1}{2}\sum_{i:p_i>0 \text{ and } q_i>0} |p_i - q_i| \\
&= \frac{1}{2}\sum_i |p_i - q_i| .
\end{aligned}$$

因此

$$\inf \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_g, \mathcal{E}_f) \le \mathsf{dist}_{\mathsf{edit}}(\pi, \pi') ,$$

如所期望的那样。 □

D.3 字符串与分布的编辑距离等价性

记 $\mathsf{dist}_{\mathsf{ham}}(x, y)$ 表示两个长度相同的字符串 x，y 之间的汉明距离。

定义 D.9。对于字符串 $x \in \{0,1\}^*$，将 $\mathsf{ext}(x)$ 记为所有字符串 $z \in \{0,1,\perp\}^*$ 的集合，其中包含非$\perp$ 字符的唯一（不一定连续的）z 子序列为 x。

事实 D.10。 给定字符串 $u \in \{0,1\}^N$ 和 $v \in \{0,1\}^M$，有

$$\mathsf{dist}_{\mathsf{string-edit}}(u,v) = \min_{x,y} \mathsf{dist}_{\mathsf{ham}}(x,y)\,,$$

其中，最小值取遍所有等长的字符串 $x \in$文本(u) 和 $y \in$文本(v)。

引理 7.6。 任取 N，并令 π、π' 为定义在 $\mathbb{N}$ 上的概率分布，其密度是 $1/N$的整数倍。则 $\frac{1}{2}\cdot$距离$_{\mathsf{rel-edit}}(\psi(\pi)\ \psi(\pi')) \le$距离$_{\mathsf{edit}}(\pi,\pi') \le$距离$_{\mathsf{rel-edit}}(\psi(\pi)\ \psi(\pi'))$。

证明。 我们通过证明两个断言来展开论证。

断言 D.11。 距离$_{\mathsf{edit}}(\pi,\pi') \le$距离$_{\mathsf{rel-edit}}(\psi(\pi)\ \psi(\pi'))$。

命题证明。 令 $x \in$文本$(\psi(\pi))$ 和 $y \in$ 文本$(\psi(\pi'))$ 为达到最小值的字符串

$$\mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi),\psi(\pi')) = \frac{1}{N}\mathsf{dist}_{\mathsf{ham}}(x,y)\,,$$

并令 M 为它们的长度。注意，对于某个 $n \le N$，$\psi(\pi)$ 是一个 n-块字符串；对于某个 $n' \le N$，$\psi(\pi')$ 是一个 n'-块字符串。于是存在一个序列 $0 = a_0 \le a_1 < a_2 < \cdots < a_{n-1} \le a_n$，使得对每个 $j \in [n]$ 和每个 $i \in (a_{j-1},\ a_j]$，都有 $x_i \in \{\perp,\mathsf{par}(j)\}$。类似地，存在一个序列 $0 = b_0 \le b_1 < b_2 < \cdots < b_{n'-1} \le b_{n'}$，使得对每个 $j \in [n']$ 和每个$i \in (b_{j-1},\ b_j]$，都有 $y_i \in \{\perp,\mathsf{par}(j)\}$。

随后，我们可以将 $f\colon \mathbb{N} \to \{0,1\}$ 定义为具有交替序列 (a_j) 的函数，并将 $g\colon \mathbb{N} \to \{0,1\}$ 定义为具有交替序列 (b_j) 的函数。注意，对于每个 $i \in [N]$，当 $x_i \neq \perp$ 时，我们有 $f(i) = x_i$；当 $y_i \neq \perp$ 时，我们有 $g(i) = y_i$。

现在，定义概率分布 $\mathcal{D}$ 使其在每个 $1/N$ 上的密度为 $i \in [N]$，其中 $x_i \neq \perp$，并定义概率分布 $\mathcal{E}$ 使其在每个 $1/N$ 上的密度为 $i \in [N]$，其中 $y_i \neq \perp$。由此可得 $\pi_{f,\ \mathcal{D}} = \pi$ 并且 $\pi_{g,\ \mathcal{E}} = \pi'$。利用 命题 6.6：

$$\begin{aligned}
\mathsf{dist}_{\mathsf{edit}}(\pi,\pi') &\le \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\mathcal{E}_g) \\
&= \frac{1}{2}\sum_{i=1}^{N}\left(\mathbb{1}\left[f(i) \neq g(i)\right](\mathcal{D}(i)+\mathcal{E}(i)) + \mathbb{1}\left[f(i) = g(i)\right]|\mathcal{D}(i)-\mathcal{E}(i)|\right) \\
&= \frac{1}{2}\sum_{i=1}^{N}\mathbb{1}\left[f(i) \neq g(i)\right](\mathcal{D}(i)+\mathcal{E}(i))\,.
\end{aligned}$$

如果 $f(i) \neq g(i)$，那么，要么：

1. $x_i = \perp$ 和 $y_i \neq \perp$，或 $x_i \neq \perp$ 和 $y_i = \perp$，在这种情况下，$\mathbb{1}\left[f(i) \neq g(i)\right](\mathcal{D}(i)+\mathcal{E}(i)) = \mathbb{1}\left[x_i \neq y_i\right]\cdot\frac{1}{N}$；或
2. $x_i \neq \perp$ 和 $y_i \neq \perp$，在这种情况下，$\mathbb{1}\left[f(i) \neq g(i)\right](\mathcal{D}(i)+\mathcal{E}(i)) = \mathbb{1}\left[x_i \neq y_i\right]\cdot\frac{2}{N}$；或
3. $x_i = y_i = \perp$，在这种情况下，$\mathbb{1}\left[f(i) \neq g(i)\right](\mathcal{D}(i)+\mathcal{E}(i)) = 0$。

Then

$$\mathsf{dist}_{\mathsf{edit}}(\pi,\pi') \le \frac{1}{2}\sum_{i=1}^{N}\mathbb{1}\left[x_i \neq y_i\right]\cdot\frac{2}{N} = \frac{1}{N}\cdot\mathsf{dist}_{\mathsf{ham}}(x,y) = \mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi),\psi(\pi'))\,,$$

这就证明了该命题。 □

命题 D.12. $\mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi)\ \psi(\pi')) \le 2 \cdot \mathsf{dist}_{\mathsf{edit}}(\pi, \pi')$。

命题证明。 设 $(f,\ \mathcal{D})$ 和 $(g,\ \mathcal{E})$ 是任意带标签分布，其中 $\pi_{f,\ \mathcal{D}} = \pi$ 且 $\pi_{g,\ \mathcal{E}} = \pi'$。我们希望证明

$$\mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi), \psi(\pi')) \overset{?}{\le} 2 \cdot \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g)\,.$$

根据 事实 D.5，再结合 事实 D.6，不失一般性，我们可以假设 $(f,\ \mathcal{D})$ 和 $(g,\ \mathcal{E})$ 满足以下条件：

1. 若 I 是任意区间，使得 f 和 g 在 I 上均为常数，且对所有 $x \in I$，$f(x) \neq g(x)$，则 $\mathcal{D}(I) = 0$ 或 $\mathcal{E}(I) = 0$。
2. f 和 g 没有共同的交替点。
3. $\mathcal{D}$ 且 $\mathcal{E}$ 支撑于集合 C，该集合包含 f 和 g 的交替点 .

我们将把 $\mathcal{D}$ 和 $\mathcal{E}$ 转换为满足以下性质的 $\mathcal{D}'$ 和 $\mathcal{E}'$：

1. $\pi_{f,\ \mathcal{D}'} = \pi_{f,\ \mathcal{D}} = \pi$ 和 $\pi_{g,\ \mathcal{E}'} = \pi_{g,\ \mathcal{E}} = \pi'$；2. $\mathcal{D}'$ 和 $\mathcal{E}'$ 支撑于 C；
3. $\forall i \in \mathbb{Z}$，$\mathcal{D}'(i)$ 和 $\mathcal{E}'(i)$ 是 $1/N$ 的整数倍（包括 0）；
4. $\forall i \in \mathbb{Z}$，若 $f(i) \neq g(i)$，则 $\min\{\mathcal{D}'(i)\ \mathcal{E}'(i)\} = 0$；
且 5. $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_f,\ \mathcal{E}'_g) \le \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f,\ \mathcal{E}_g)$。

令 $a_1 < a_2 < \cdots$ 为 f 的交替点，令 $b_1 < b_2 < \cdots$ 为 g 的交替点。记 $A = \{a_1, a_2, \dots\}$ 和 $B = \{b_1, b_2, \dots\}$；我们有 $A \cap B = \emptyset$，并且 $\mathcal{D}$ 和 $\mathcal{E}$ 支撑于 $C = A \cup B$。

我们通过在每个区间 $(a_{i-1}, a_i]$ 和 $(b_{j-1}, b_j]$ 内，按照端点 a_i 和 b_j 的顺序执行以下变换，来定义 $\mathcal{D}'$ 和 $\mathcal{E}'$；由于 $A \cap B = \emptyset$，这种排序是良定义的。我们针对区间 $(a_{i-1}, a_i]$ 定义这一过程；区间 $(b_{j-1}, b_j]$ 则以对称的方式处理。对于每个区间 $(a_{i-1}, a_i]$，在过程的每次迭代中，记 $\mathcal{D}$ 和 $\mathcal{E}$ 为迭代前的分布，记 $\mathcal{D}'$ 为调整 $(a_{i-1}, a_i]$ 内质量后的分布。我们将保证

$$\begin{aligned}&\sum_{x \in (a_{i-1}, a_i]} \left(\mathbb{1}\left[f(x) = g(x)\right] \left|\mathcal{D}'(x) - \mathcal{E}(x)\right| + \mathbb{1}\left[f(x) \neq g(x)\right] \left(\mathcal{D}'(x) + \mathcal{E}(x)\right)\right) \\ &\quad \le \sum_{x \in (a_{i-1}, a_i]} \left(\mathbb{1}\left[f(x) = g(x)\right] \left|\mathcal{D}(x) - \mathcal{E}(x)\right| + \mathbb{1}\left[f(x) \neq g(x)\right] \left(\mathcal{D}(x) + \mathcal{E}(x)\right)\right)\,.\end{aligned} \tag{20}$$

根据 命题 6.6，这足以保证性质（5）。

1. Let $C_i := C \cap (a_{i-1}, a_i]$. Let $C_i^+ := \{x \in C_i : f(x) = g(x)\}$ and $C_i^- := \{x \in C_i : f(x) \neq g(x)\}$.
2. 如果 $C_i^+ = \emptyset$，我们可以确定 $(a_{i-1}, a_i] \cap B = \emptyset$，并且以下两种情况之一成立：$\mathcal{D}(a_{i-1}, a_i] = 0$，或 $\mathcal{E}(a_{i-1}, a_i] = 0$，因为 $(a_{i-1}, a_i]$ 是一个区间，在此区间上 f, g 为常数且彼此不相等。在这种情况下，对所有 $x \in (a_{i-1}, a_i]$，令 $\mathcal{D}'(x) = \mathcal{D}(x)$，于是 $\mathcal{D}'(a_i) = \mathcal{D}(a_{i-1}, a_i]$，它是 $1/N$ 的整数倍。这保证了 $(a_{i-1}, a_i]$ 内满足性质（3），并且保证条件 (20) 显然成立。

3. 否则，写作 $C_i^+ = \{c_1, \ldots, c_m\}$，并考虑序列 $u_1, \ldots, u_m$，其中 $u_j := \mathcal{E}(c_j)$。注意，对于 $c_j < a_i$，$\mathcal{E}(c_j)$ 已在此过程中较早定义，因为 c_j 是一个区间的端点，因此 $\mathcal{E}(c_j)$ 是 $1/N$ 的整数倍。接着，根据事实 D.3，将质量 $\mathcal{D}(a_{i-1}, a_i]$ 按照 $c_1, \ldots, c_m$ 这些点进行分配，以定义 $\mathcal{D}'$。该事实保证 $\mathcal{D}'(a_{i-1}, a_i] = \mathcal{D}(a_{i-1}, a_i]$；并且通过检查证明可知，每个 $\mathcal{D}'(c_j)$ 都是 $1/N$ 的整数倍；因此我们得到保证条件 (20)，因为：

$$
\begin{aligned}
&\sum_{x \in (a_{i-1}, a_i]} \left(\mathbb{1}\left[f(x) = g(x)\right] \left|\mathcal{D}'(x) - \mathcal{E}(x)\right| + \mathbb{1}\left[f(x) \neq g(x)\right] \left(\mathcal{D}'(x) + \mathcal{E}(x)\right) \right) \\
&\quad = \sum_{x \in C_i^+} \left|\mathcal{D}'(x) - \mathcal{E}(x)\right| + \sum_{x \in C_i^-} \left(\mathcal{D}'(x) + \mathcal{E}(x)\right) \\
&\quad = \sum_{x \in C_i^+} \left|\mathcal{D}'(x) - \mathcal{E}(x)\right| + \sum_{x \in C_i^-} \mathcal{E}(x) = \left| \mathcal{D}(a_{i-1}, a_i] - \sum_{x \in C_i^+} \mathcal{E}(x) \right| + \sum_{x \in C_i^-} \mathcal{E}(x) \\
&\quad \leq \sum_{x \in C_i^+} \left|\mathcal{D}(x) - \mathcal{E}(x)\right| + \sum_{x \in C_i^-} \left(\mathcal{D}(x) + \mathcal{E}(x)\right) \\
&\quad = \sum_{x \in (a_{i-1}, a_i]} \left(\mathbb{1}\left[f(x) = g(x)\right] \left|\mathcal{D}(x) - \mathcal{E}(x)\right| + \mathbb{1}\left[f(x) \neq g(x)\right] \left(\mathcal{D}(x) + \mathcal{E}(x)\right) \right) .
\end{aligned}
$$

得到所需的带标签分布（$f, \mathcal{D}'$）和（$g, \mathcal{E}'$）后，我们如下完成证明。令 $C = \{c_1, c_2, \cdots\}$，使得 $c_1 < c_2 < \ldots$。由于 $\mathcal{D}'$ 和 $\mathcal{E}'$ 的密度都是 $1/N$ 的整数倍，因此存在某个 m，使得 $\mathcal{D}'$ 和 $\mathcal{E}'$ 支撑于 $c_1, \ldots, c_m$。对于每个 $t \in [m]$，定义 $Z_t := [(t-1)N + 1, tN]$，使得 $|Z_t| = N$。令 $p_t := N \cdot \mathcal{D}'(c_t)$ 和 $q_t := N \cdot \mathcal{E}'(c_t)$，它们都是非负整数。然后我们如下定义字符串 $x, y \in \{0, 1, \perp\}^{m \cdot N}$。对于每个 $t \in [m]$，将 $x_i = f(c_t)$ 定义为 $i \in Z_t$ 的前 p_t 个值，并将 $y_i = g(c_t)$ 定义为 $i \in Z_t$ 的前 q_t 个值，再令 Z_t 中剩余的字符为 $\perp$。

容易验证，$x \in \mathsf{ext}(\psi(\pi))$ 和 $y \in \mathsf{ext}(\psi(\pi'))$，因此

$$
\begin{aligned}
&\mathsf{dist}_{\mathsf{rel-edit}}(\psi(\pi), \psi(\pi')) = \frac{1}{N} \mathsf{dist}_{\mathsf{string-edit}}(\psi(\pi), \psi(\pi')) \\
&\quad \leq \frac{1}{N} \cdot \mathsf{dist}_{\mathsf{ham}}(x, y) \\
&\quad = \frac{1}{N} \sum_{t=1}^{m} \left(\mathbb{1}\left[f(c_t) = g(c_t)\right] \cdot |p_t - q_t| + \mathbb{1}\left[f(c_t) \neq g(c_t)\right] \max\{p_t, q_t\} \right) \\
&\quad = \sum_{t=1}^{m} \left(\mathbb{1}\left[f(c_t) = g(c_t)\right] \cdot \left|\mathcal{D}'(c_t) - \mathcal{E}'(c_t)\right| + \mathbb{1}\left[f(c_t) \neq g(c_t)\right] \cdot \max\{\mathcal{D}'(c_t) + \mathcal{E}'(c_t)\} \right) \\
&\quad = \sum_{x} \left(\mathbb{1}\left[f(x) = g(x)\right] \cdot \left|\mathcal{D}'(x) - \mathcal{E}'(x)\right| + \mathbb{1}\left[f(x) \neq g(x)\right] \cdot \max\{\mathcal{D}'(x) + \mathcal{E}'(x)\} \right) \\
&\quad = \sum_{x} \left(\mathbb{1}\left[f(x) = g(x)\right] \cdot \left|\mathcal{D}'(x) - \mathcal{E}'(x)\right| + \mathbb{1}\left[f(x) \neq g(x)\right] \cdot \left(\mathcal{D}'(x) + \mathcal{E}'(x)\right) \right) \\
&\quad = 2 \cdot \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_f, \mathcal{E}'_g) ,
\end{aligned}
$$

这就证明了该结论。 □

这两项断言完成了证明。 □

D.4 均匀分布的编辑距离

引理 6.31. 存在一个绝对常数 $c > 0$，使得以下结论成立。设 π 是定义在 $\mathbb{N}$ 上、均匀支撑于 $[k]$ 的分布，π' 是另一个定义在 $\mathbb{N}$ 上且支撑于 $[k]$ 内的概率分布。那么 $\mathsf{dist}_{\mathsf{edit}}(\pi, \pi') \geq c \cdot \mathsf{dist}_{\mathsf{TV}}(\pi, \pi')$。

证明。回想一下，对每个 $i \in [k]$，$\pi(i) = 1/k$。令 $z \in \mathbb{R}^k$ 为满足对每个 $i \in [k]$，$\pi'(i) = \pi(i) + z_i$ 的向量。注意，由于 π 和 π' 是概率分布，因此我们有 $\sum_i z_i = 0$ 和

$$\mathsf{dist}_{\mathsf{TV}}(\pi, \pi') = \frac{1}{2}\left(\sum_{i:z_i>0} z_i + \sum_{i:z_i<0} |z_i|\right) = \sum_{i:z_i>0} z_i \,.$$

令 $S := \{i \in [k] \colon z_i > 0\}$，我们可以假设其非空，否则结论显然成立。现在，我们的目标是证明 $\mathsf{dist}_{\mathsf{edit}}(\pi, \pi') \geq c \cdot \sum_{i\in S} z_i$。

设 $\mathcal{D}_f$ 和 $(g, \mathcal{E})$ 是两个 1-适当标记分布，满足 $\pi_{f,\mathcal{D}} = \pi$ 且 $\pi_{g,\mathcal{E}} = \pi'$。我们的目标是证明 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g) \geq c \cdot \sum_{i\in S} z_i$。我们可以假设 g（也）恰好交替 $k-1$ 次，因为如果交替次数更少，我们可以从一个足够大的坐标开始引入额外交替，使得受影响的 $\mathcal{E}_g$ 质量任意小。

令 $a_1 < a_2 < \cdots < a_{k-1}$ 为 f 的交替序列，令 $b_1 < b_2 < \cdots < b_{k-1}$ 为 g 的交替序列。为方便记号，记 $a_0 := -\infty$, $a_k = \infty$, $b_0 := -\infty$ 和 $b_k := \infty$，于是对每个 $t \in [k]$，都有 $1/k = \pi(t) = \mathcal{D}(a_{t-1}, a_t]$ 且 $\pi'(t) = \mathcal{E}(b_{t-1}, b_t]$。

任取 $t \in S$。根据 命题 6.6，只需证明以下内容：

$$\sum_{i\in[b_{t-1},b_t)} \mathbb{1}\left[f(i) \neq g(i)\right]\left(\mathcal{D}(i) + \mathcal{E}(i)\right) + \mathbb{1}\left[f(i) = g(i)\right]\left|\mathcal{D}(i) - \mathcal{E}(i)\right| \overset{?}{\geq} 2c \cdot z_t \,.$$

为方便起见，令 F_t 表示该待证不等式左端的表达式。

令 $h := \lfloor \frac{\mathcal{E}(b_{t-1},b_t]}{1/k} \rfloor$ 且 $\epsilon := \mathcal{E}(b_{t-1}, b_t] - h/k$，于是 $0 \leq \epsilon < 1/k$ 且 $\mathcal{E}(b_{t-1}, b_t] = h/k + \epsilon$。注意，由于 $t \in S$，因此 $h \geq 1$，这意味着 $\mathcal{E}(b_{t-1}, b_t] > 1/k$。此外，回想 $\mathcal{E}(b_{t-1}, b_t] = 1/k + z_t$，可得 $z_t = (h-1)/k + \epsilon$。现在我们分情况讨论。

情形 1. 假设 $\mathcal{D}(b_{t-1}, b_t] \leq \frac{1}{k}\left(1 + \frac{h-1}{2}\right) + \frac{\epsilon}{2}$。由此可得

$$\begin{aligned} F_t &\geq \sum_{i\in(b_{t-1},b_t]} |\mathcal{D}(i) - \mathcal{E}(i)| \geq |\mathcal{E}(b_{t-1}, b_t] - \mathcal{D}(b_{t-1}, b_t]| \geq \frac{h}{k} + \epsilon - \frac{1}{k} - \frac{(h-1)/2}{k} - \frac{\epsilon}{2} \\ &= \frac{(h-1)/2}{k} + \frac{\epsilon}{2} = \frac{z_t}{2} \,, \end{aligned}$$

因此，这种情形下的证明就完成了。

情形 2. 假设 $\mathcal{D}(b_{t-1}, b_t] > \frac{1}{k}\left(1 + \frac{h-1}{2}\right) + \frac{\epsilon}{2}$。根据 h 的取值，我们进一步讨论若干子情形。在接下来的分析中，我们使用这样一个事实：g 在 $(b_{t-1}, b_t]$ 上为常数。

情形 2A。$h = 1$。令 t' 为满足 $a_{t'} \in (b_{t-1}, b_t)$ 的最小指标；这样的指标必定存在，因为 $\mathcal{D}(b_{t-1}, b_t] > 1/k$，因此 f 在此区间内必定交替。若 $f(a_{t'}) = g(a_{t'})$，则从 $a_{t'} + 1$ 到下一个交替点 $a_{t'+1}$ 或 b_t（以先出现者为准）之前，f 与 g 不一致。此外，由于 t' 的最小性意味着 $a_{t'-1} \leq b_{t-1}$，从而 $\mathcal{D}(b_{t-1}, a_{t'}] \leq 1/k$，我们有 $\mathcal{D}(a_{t'}, b_t] = \mathcal{D}(b_{t-1}, b_t] - \mathcal{D}(b_{t-1}, a_{t'}] > (h-1)/k + \frac{\epsilon}{2} = \epsilon/2$。因此，回想起 $\epsilon < 1/k$，我们得到

$$F_t \geq \sum_{i\in(a_{t'},\min\{a_{t'+1},b_t\}]} (\mathcal{D}(i) + \mathcal{E}(i)) \geq \mathcal{D}(a_{t'}, \min\{a_{t'+1}, b_t\}] \geq \min\{1/k, \epsilon/2\} = \epsilon/2 = z_t/2 \,,$$

如所需。

否则，假设 $f(a_{t'}) \neq g(a_{t'})$。论证思路类似，但现在我们要说明，在 $(b_{t-1}, b_t]$ 中且满足不超过 $a_{t'}$ 或位于 $(a_{t'+1}, a_{t'+2}]$ 的区域内，必定存在相当大的 $\mathcal{D}$-质量；也就是说，这些区域中 f 和 g 不一致。确实，假设 $\mathcal{D}(b_{t-1}, a_{t'}] < \epsilon/4$。那么 $\mathcal{D}(b_{t-1}, a_{t'+1}] = \mathcal{D}(b_{t-1}, a_{t'}] + \mathcal{D}(a_{t'}, a_{t'+1}] < \epsilon/4 + 1/k$，而 $\mathcal{D}(b_{t-1}, b_t] > 1/k + \epsilon/2$，这意味着 $\mathcal{D}(a_{t'+1}, \min\{a_{t'+2}, b_t\}] \geq \min\{1/k, \epsilon/4\} = \epsilon/4$。因此，在 $(b_{t-1}, b_t]$ 内，f 和 g 至少在 $\epsilon/4$ $\mathcal{D}$-质量上不一致，所以 $F_t \geq \epsilon/4 = z_t/4$，这个情形就证明完毕。

包含 $h \geq 2$ 的情形遵循类似的逻辑，但现在，z_t 中的 $(h-1)/k$ 项占主导地位，因此我们必须相应地调整论证。

情况 2B。$2 \leq h \leq 13$。如上所述，令 t' 为满足 $a_{t'} \in (b_{t-1}, b_t)$ 的最小指标。若 $f(a_{t'}) = g(a_{t'})$，则按如下方式即可完成证明。注意，由于 t' 的最小性，我们有 $a_{t'-1} \leq b_{t-1}$，从而 $\mathcal{D}(b_{t-1}, a_{t'}] \leq 1/k$)。因此可得 $\mathcal{D}(a_{t'}, b_t] = \mathcal{D}(b_{t-1}, b_t] - \mathcal{D}(b_{t-1}, a_{t'}] > (h-1)_k{}^{/2} \geq 1/2k$)。因此，我们得到———

$$F_t \geq \sum_{i \in (a_{t'}, \min\{a_{t'+1}, b_t\}]} (\mathcal{D}(i) + \mathcal{E}(i)) \geq \mathcal{D}(a_{t'}, \min\{a_{t'+1}, b_t\}] \geq \min\{1/k, 1/2k\} = 1/2k\,.$$

于是，由于 $\epsilon < 1/k$ 和 $1 \leq h-1 \leq 12$，我们得到 $F_t > \frac{1}{4k} + \frac{\epsilon}{4} \geq \frac{(h-1)_{4k}{}^{/12}}{} + \frac{\epsilon}{4} \geq \frac{z_t}{48}$，如所需。

否则，假设 $f(a_{t'}) \neq g$({v4)。我们沿用前面几种情况的论证，说明必然存在相当大的 $\mathcal{D}$质量，同时位于 $(b_{t-1}, b_t]$，且满足以下条件之一：至多为 $a_{t'}$，或位于 $(a_{t'+1}, a_{t'+2}]$；也就是说，这些区域中 f 和 g 不一致。确实，假设 $\mathcal{D}(b_{t-1}, a_{t'}] < 1/4k$。那么 $\mathcal{D}(b_{t-1}, a_{t'+1}] = \mathcal{D}(b_{t-1}, a_{t'}] + \mathcal{D}(a_{t'}, a_{t'+1}] < 1/4k + 1/k$，而根据假设，$\mathcal{D}(b_{t-1}, b_t] > 1/k + 1/2k$，这意味着$\mathcal{D}(a_{t'+1}, \min\{a_{t'+2}, b_t\}] \geq \min\{1/k, 1/4k\} = 1/4k$。然后，再次利用 $\epsilon < 1/k$和 $1 \leq h-1 \leq 12$，可得 $F_t \geq \frac{1}{4k} > \frac{1}{8k} + \frac{\epsilon}{8} \geq (h-1)_{8k}{}^{/12} + \frac{\epsilon}{8} \geq \frac{z_t}{96}$，如所需。[16]

情况 2C. $h \geq 14$。设 ℓ 为 f 在 $(b_{t-1}, b_t]$ 中的交替点数；设这些点为$a_{t'}, a_{t'+1}, \ldots, a_{t'+\ell-1}$。我们声称 $\ell \geq (h-1)/2$。事实上，假设 $\ell < (h-1)/2$。那么 $(b_{t-1}, b_t]$ 中的总$\mathcal{D}$质量至多为

$$\mathcal{D}(\max\{a_{t'-1}, b_{t-1}\}, a_{t'}] + \sum_{j=1}^{\ell-1} \mathcal{D}(a_{t'+j-1}, a_{t'+j}] + \mathcal{D}(a_{t'+\ell-1}, \min\{a_{t'+\ell}, b_t\}] \leq \frac{1}{k}(\ell+1) < \frac{1}{k}\left(1 + \frac{h-1}{2}\right),$$

这与我们关于 $\mathcal{D}$ 的假设矛盾。因此 $\ell \geq (h-1)/2$。

现在，考虑由 $(a_{t'+j-1}, a_{t'+j}]$ 形式的 $\ell - 1$ 值域组成的区间，其中每个区间包含 $(b_{t-1}, b_t]$ 内一对相邻的 f交替。由于 f 在每个这样的区间内都是常数，因此其中至少有$\lfloor(\ell-1)/2\rfloor \geq \lfloor(h-3)/4\rfloor \geq (h-7)/4 \geq h/8$ 个区间与 g 不一致；最后一个不等式成立，因为$h \geq 14$。因此，f 与 g 在足够大的 $\mathcal{D}$ 质量上不一致：回想 $\epsilon < 1/k$，我们有

$$F_t \geq \frac{h}{8} \cdot \frac{1}{k} > \frac{1}{8}\left(\frac{h-1}{k} + \epsilon\right) = \frac{z_t}{8}\,,$$

至此证明完毕。 □

D.5 带标签分布的支持集大小的编辑距离

命题 6.37。令 Ξ 为适当的带标签分布的性质 $(g, \mathcal{E})$，其中 $\pi_{g,\ \varepsilon}$ 的支撑集大小至多为 k。那么，对于任意适当的带标签分布 $(f, \mathcal{D})$，$\mathsf{dist}_{\mathsf{TV}}$（$(f, \mathcal{D})$ Ξ）$\leq \mathsf{dist}_{\mathsf{edit}}$（$(f, \mathcal{D})$ Ξ）。

[16]我们没有尝试优化常数 c。

证明。 令 $(f', \mathcal{D}')$ 和 $(g, \mathcal{E}) \in \Xi$ 满足 $\pi_{f',\mathcal{D}'} = \pi_{f,\mathcal{D}}$ 且 $\mathsf{dist}_{\mathsf{edit}}((f, \mathcal{D}) \Xi) = \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_{f'}, \mathcal{E}_g)$。

步骤 1. 我们将证明存在 $(g', \mathcal{E}') \in \Xi$，使得 $\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_{f'}, \mathcal{E}'_{g'}) \leq \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_{f'}, \mathcal{E}_g)$，且 g' 的交替序列是 f'的交替序列的子集。根据事实 D.5，我们可以假设，对于任意区间 I，若 f' 和 g 均为常数且彼此不相等，则要么$\mathcal{D}(I) = 0$，要么 $\mathcal{E}(I) = 0$。

令 $a'_1 < a'_2 < \cdots$和 $b_1 < b_2 < \cdots$分别为 f'和 g的交替序列。假设存在 $(a'_{i-1}, a'_i]$，使得存在 $b_j \in (a'_{i-1}, a'_i)$。定义 g'，使得对于所有 $x \in (a'_{i-1}, a'_i]$， $g'(x) = f'(x)$；否则， $g'(x) = g(x)$。根据事实 D.4，$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_{f'}, \mathcal{E}_{g'}) = \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_{f'}, \mathcal{E}_g)$。

我们声称，这不会增加交替点的数量，因此 g' 的交替点数量至多与 g相同。令 z为满足以下条件的常数：对所有 $x \in (a'_i,\ a'_{i-1}]$， $f'(x) = g'(x) = z$。由于存在交替点 $b_j \in (a'_i,\ a'_{i-1})$，因此存在一个区间 $I \subseteq (a'_i,\ a'_{i-1}]$，使得 g 和g' 在 I上的取值均为常数 z。当我们将$(a'_i,\ a'_{i-1}]$ 中 g 的值替换为 z 以得到 g'时，交替点的数量不会增加，因为我们只是扩展了区间 I。

在每个区间中执行此操作（$(a'_{i-1}, a'_i]$ 其中存在一个交替点$b_j \in (a'_{i-1}, a'_i)$），并简单地设定 $\mathcal{E}' = \mathcal{E}$（为使步骤 2 中的记号清晰），我们便得到具有所需性质的 $(g', \mathcal{E}')$。

步骤 2。现在我们有 $(f', \mathcal{D}')$ 和 $(g', \mathcal{E}')$，其中 g' 的交替序列 $b'_1 < b'_2 < \cdots$ 是 f' 的交替序列的子集。令 $a_1 < a_2 < \cdots$ 为原函数 f 的交替序列。我们将如下定义 $(h, \mathcal{H})$。对于每个区间 $B'_j :=$（$(b'_{j-1}, b'_j]$ 属于 g' 的交替序列），令 $a'_i = b'_{j-1} < a'_{i+1} < \cdots < a'_{i+t} = b'_j$ 为包含在 $[b'_{j-1}, b'_j]$ 中的 f' 的交替点，并令 $A_j = (a_i, a_{i+t}]$。令 $T'_j \subseteq B'_j$ 为满足 $f'(x) = z_j$ 的点 x。我们定义 $h(x) = z_j$，对所有 $x \in A_j$ 均成立。注意，区间 A_j 对定义域构成一个分割，因此这就完整定义了 h。

固定一个区间 B'_j。如果对所有 $x \in A_j$ 都有 $f(x) \neq z_j$，则令 $\mathcal{H}(a_{i+t}) = \mathcal{E}'(B'_j)$，并令剩余的 $x \in (a_i, a_{i+t})$ 满足 $\mathcal{H}(x) = 0$。于是

$$
\begin{aligned}
&\sum_{x \in A_j} \mathbb{1}\left[f(x) \neq h(x)\right](\mathcal{D}(x) + \mathcal{H}(x)) + \mathbb{1}\left[f(x) = h(x)\right]|\mathcal{D}(x) + \mathcal{H}(x)| \\
&\quad = \sum_{x \in A_j} (\mathcal{D}(x) + \mathcal{H}(x)) \\
&\quad = \mathcal{D}(a_{i+t}) + \mathcal{H}(a_{i+t}) + \sum_{x \in A_j \setminus \{a_{j+t}\}} \mathcal{D}(x) \\
&\quad = \mathcal{E}'(B'_j) + \mathcal{D}(A_j) = \mathcal{E}'(B'_j) + \mathcal{D}'(B'_j) \\
&\quad = \sum_{x \in B'_j} \mathbb{1}\left[f'(x) \neq g'(x)\right](\mathcal{D}'(x) + \mathcal{E}'(x)) + \mathbb{1}\left[f'(x) = g'(x)\right]|\mathcal{D}'(x) - \mathcal{E}'(x)|\,.
\end{aligned}
$$

否则，令 $T_j \subseteq A_j$ 表示满足 $f(x) = z_j$ 的坐标 $x \in T_j$。首先注意到

$$
\begin{aligned}
&\sum_{x \in B'_j} \mathbb{1}\left[f'(x) \neq g'(x)\right](\mathcal{D}'(x) + \mathcal{E}'(x)) + \mathbb{1}\left[f'(x) = g'(x)\right]|\mathcal{D}'(x) - \mathcal{E}'(x)| \\
&\quad = \mathcal{D}'(B'_j \setminus T'_j) + \mathcal{E}'(B'_j \setminus T'_j) + |\mathcal{E}'(T'_j) - \mathcal{D}'(T'_j)| \\
&\quad \geq \mathcal{D}'(B'_j \setminus T'_j) + |\mathcal{E}'(B'_j) - \mathcal{D}'(T'_j)| \\
&\quad = \mathcal{D}(A_j \setminus T_j) + |\mathcal{E}'(B'_j) - \mathcal{D}'(T'_j)|\,.
\end{aligned}
$$

我们为 $\mathcal{H}$ 中的坐标赋值，使得 T_j 满足 $\mathcal{H}(T_j) = \mathcal{E}'(B'_j)$， $\mathcal{H}(x) = 0$ 对所有

$x \in A_j \setminus T_j$，并且

$$\sum_{x \in T_j} |\mathcal{H}(x) - \mathcal{D}(x)| = |\mathcal{H}(T_j) - \mathcal{D}(T_j)| = |\mathcal{H}(A_j) - \mathcal{D}(T_j)| = |\mathcal{E}'(B'_j) - \mathcal{D}'(T'_j)| .$$

这由于事实 D.3是可能的。于是

$$\begin{aligned}
&\sum_{x \in T_j} |\mathcal{H}(x) - \mathcal{D}(x)| + \sum_{x \in A_j \setminus T_j} (\mathcal{H}(x) + \mathcal{D}(x)) \\
&\quad = |\mathcal{E}'(B'_j) - \mathcal{D}'(T'_j)| + \mathcal{D}(A_j \setminus T_j) \\
&\quad \le \sum_{x \in B'_j} \mathbb{1}\left[f'(x) \neq g'(x)\right] (\mathcal{D}'(x) + \mathcal{E}'(x)) + \mathbb{1}\left[f'(x) = g'(x)\right] |\mathcal{D}'(x) - \mathcal{E}'(x)| .
\end{aligned}$$

对每个区间应用相同的论证$B'_j = (b'_{j-1},\ b'_j]$，我们得到具有所需性质的（h，$\mathcal{H}$），这是由命题 6.6得出的：

$$\begin{aligned}
&\sum_x (\mathbb{1}\left[f(x) = h(x)\right] \cdot |\mathcal{D}(x) - \mathcal{H}(x)| + \mathbb{1}\left[f(x) \neq h(x)\right] (\mathcal{D}(x) + \mathcal{H}(x))) \\
&= \sum_j \left(\sum_{x \in T_j} |\mathcal{H}(x) - \mathcal{D}(x)| + \sum_{x \in A_j \setminus T_j} (\mathcal{H}(x) + \mathcal{D}(x)) \right) \\
&\le \sum_j \left(\sum_{x \in B'_j} \mathbb{1}\left[f'(x) \neq g'(x)\right] (\mathcal{D}'(x) + \mathcal{E}'(x)) + \mathbb{1}\left[f'(x) = g'(x)\right] |\mathcal{D}'(x) - \mathcal{E}'(x)| \right) \\
&= \mathsf{dist}_{\mathsf{TV}}(\mathcal{D}'_{f'}, \mathcal{E}'_{g'}) \le \mathsf{dist}_{\mathsf{edit}}((f, \mathcal{D}), \Xi) .
\end{aligned}$$

□

D.6 分布支持大小的编辑距离

引理 6.41。设 $k \in \mathbb{N}$。令 Π_k 为支撑于 $\mathbb{N}$ 上至多 k 个元素的分布集合，并令 Π_{2k} 为支撑于 $\mathbb{N}$ 上至多 $2k$ 个元素的分布集合。令 π 为 $\mathbb{N}$ 上的有限支撑概率分布，并令 π' 为 $\mathbb{N}$ 上由 $\pi'(2i-1) = \pi'(2i) := \pi(i)/2$ 对每个 $i \in \mathsf{supp}(\pi)$ 给出的概率分布。则 $\mathsf{dist}_{\mathsf{edit}}(\pi', \Pi_{2k}) \ge \frac{1}{4} \cdot \mathsf{dist}_{\mathsf{TV}}(\pi, \Pi_k)$。

证明。 通过极限论证，并利用编辑距离与全变差距离的三角不等式，只需证明当 π（因而 π'）仅具有有理密度时，该结论成立。这使我们能够通过使用字符串上的标准编辑距离来减少技术细节，因为这种距离更容易分析，具体如下。我们可以固定一个整数 N，使得 π' 的所有密度都是 $1/N$ 的整数倍（例如，我们可以取 $N = \prod_i b_i$，其中我们记 $\pi'(i) = a_i/b_i$）。

现在应用命题 7.11：令 Ψ 为 $\{0,1\}^N$ 中 $2k$-块字符串的集合，并令 $x' := \psi(\pi')$；则 $\mathsf{dist}_{\mathsf{rel-edit}}(x', \Psi) \le 2 \cdot \mathsf{dist}_{\mathsf{edit}}(\pi', \Pi_{2k})$。令 $\delta := \mathsf{dist}_{\mathsf{rel-edit}}(x', \Psi)$。根据相对编辑距离的定义，存在一系列 $N\delta$ 编辑操作 $O_1, \ldots, O_{N\delta}$，使得 $(O_{N\delta} \circ O_{N\delta-1} \circ \cdots \circ O_1)(x') \in \Psi$，其中每个 O_j 都是对单个字符执行的插入、删除或替换。

我们声称存在一个操作序列 $O'_1, \ldots, O'_\ell$，满足 $\ell \le N\delta$，使得

1. $O'_1, \ldots, O'_\ell$ 中的每个操作都是删除；
2. $(O'_\ell \circ O'_{\ell-1} \circ \cdots \circ O'_1)(x')$ 是一个 $2k$-块字符串。

要看出这一点为何成立，选择 $s \in \Psi$ 使得 $\mathsf{dist}_{\mathsf{rel-edit}}(x', s) = \delta$，并利用事实 D.10，固定长度相等的 $u \in \mathsf{ext}(x')$ $v \in \mathsf{ext}(s)$，使得 $\mathsf{dist}_{\mathsf{ham}}(u, v) = \mathsf{dist}^{\mathsf{string-edit}}(x', s) = N\delta$。我们按如下方式构造字符串 $v' \in \{0, 1, \perp\}^M$：对于每个 $i \in [M]$，

1. 如果 $u_i = \perp$，令 $v'_i = \perp$。2. 如果 $u_i \neq \perp$ 且 $u_i = v_i$，令 $v'_i = v_i$。3. 如果 $u_i \neq \perp$ 且 $u_i \neq v_i$，令 $v'_i = \perp$。

我们做出三个观测结果。首先，存在一个 $2k$-块字符串 $s' \in \{0,1\}^*$，使得 $v' \in \mathsf{ext}(s')$；这是因为对于每个 $i \in [M]$，$v'_i = v_i$或 $v'_i = \perp$必有其一。其次，对于每个 $i \in [M]$，都有蕴含关系 $u_i \neq v'_i \implies v'_i = \perp$；这是由构造方式直接得到的。第三，$\mathsf{dist}_{\mathsf{ham}}(u, v') \leq \mathsf{dist}_{\mathsf{ham}}(u, v) = N\delta$，这同样由构造方式显然成立。令 $\ell := \mathsf{dist}_{\mathsf{ham}}(u, v')$。

我们按如下方式进行删除操作。令 $u^* \in \{0,1,\perp\}^M$ 在$u_i = v'_i$时由 $u^*_i = u_i$ 给出，否则由 $u^*_i = \perp$ 给出。由第二个观测结果，$\mathsf{dist}_{\mathsf{ham}}(u^*, v') = 0$。令 $x^* \in \{0,1\}^*$由 x' 通过删除与上文情形 $u_i \neq v'_i$ 相对应的每个 ℓ 字符得到（即，若 $u_i \neq v'_i$，则该字符是 u_i中第 j 个非-$\perp$ 字符，因此删除 x'中的第 j 个字符）。于是，$\mathsf{dist}^{\mathsf{string-edit}}(x', x^*) = \ell$。此外，$u^* \in \mathsf{ext}(x^*)$，因为我们从 x' 中删除字符得到x^*，并相应地将 u 中的字符集合设为 $\perp$，从而得到 u^*。因此，$\mathsf{dist}^{\mathsf{string-edit}}(x^*, s') \leq \mathsf{dist}_{\mathsf{ham}}(u^*, v') = 0$，所以 $x^* = s'$ 是一个 $2k$- 块字符串。因此，只需 $\ell = \mathsf{dist}^{\mathsf{string-edit}}(x', x^*)$ 次删除操作，就能将 x' 变成一个 $2k$-块字符串。这就得到了所需的 $O'_1, \ldots, O'_\ell$。

现在，我们使用这些操作将 π 转换为一个至多支撑于 k 个元素的向量 ν，具体如下。我们将 ν 在 π 的支撑集之外的所有位置都设为零。对于每个 $i \in \in \mathsf{supp}(\pi)$,

1. 令 $x'_p \cdots x'_{p+\frac{1}{2}N\pi(i)-1} = 1^{N\pi'_{(2i-1)}}$为 $x' = \psi(\pi')$ 中与条目 $\pi'(2i-1)$ 对应的 1 块。类似地，令 $x'_q \cdots x'_{q+\frac{1}{2}N\pi(i)-1} = 0^{N\pi'_{(2i)}}$为 x' 中与条目 $\pi'(2i)$ 对应的 0 块。

2. 如果这两个块中至少有一个块里的所有字符都被 $O'_1, \ldots, O'_\ell$ 中的操作删除，则将 $\nu(i)$ 设为 $= 0$。否则，将 $\nu(i)$ 设为 $= \pi(i)$。

首先，注意到

$$\begin{aligned}\|\pi - \nu\|_1 &= \sum_{i \in \mathrm{supp}(\pi)} \pi(i) \cdot \mathbb{1}\left[\text{all 0s or all 1s corresponding to } \pi(i) \text{ deleted}\right] \\ &\leq 2 \sum_{i' \in \mathrm{supp}(\pi')} \pi'(i') \cdot \mathbb{1}\left[\text{entire block corresponding to } \pi'(i') \text{ deleted}\right].\end{aligned}$$

如果删除了一个块 $b^{N\pi'_{(i')}}$，那么删除该块中的字符需要 $N\pi'(i')$ 次删除。每个被删除的块都对应一个互不相交的删除操作集合，而删除操作总数为 ℓ。因此，必有

$$\sum_{i' \in \mathrm{supp}(\pi')} (N\pi'(i')) \cdot \mathbb{1}\left[\text{entire block corresponding to } \pi'(i') \text{ deleted}\right] \leq \ell\,.$$

因此，

$$\|\pi - \nu\|_1 \leq 2\ell/N \leq 2\delta\,.$$

现在，令 ν^* 为从 ν 得到的概率分布，即将所需的概率质量$1 - \|\nu\|_1$ 添加到任意一个质量非零的元素上。注意到 $1 - \|\nu\|_1 = \|\pi\|_1 - \|\nu\|_1 \leq 2\delta$，所以 $\|\pi - \nu^*\|_1 \leq \|\pi - \nu\|_1 + 2\delta \leq 4\delta$。由此可得 $\mathsf{dist}_{\mathsf{TV}}(\pi, \nu^*) \leq 2\delta$。剩下只需证明 ν^*至多支撑于 k 个元素。

设 $K := |\mathsf{supp}(\pi)|$；则 $|\mathsf{supp}(\pi')| = 2K$，且 x'恰好包含 $2K$ 个块。由于 $(O'_\ell \circ \ldots \circ O'_1)(x')$ 至多包含 $2k$ 个块，因此至少有 $2(K-k)$ 个块被完全删除

这些操作。因此，根据上述构造，至少有 $K-k$ 个不同的指标 $i \in \mathrm{supp}(\pi)$ 满足 $\nu(i)=0$。因此 $|\mathrm{supp}(\nu^*)| = |\mathrm{supp}(\nu)| \le |\mathrm{supp}(\pi)| - (K-k) = k$，证毕。

综合以上各项，我们得到 $\mathrm{dist}_{\mathsf{TV}}(\pi, \Pi_k) \le \mathrm{dist}_{\mathsf{TV}}(\pi, \nu^*) \le 2\delta \le 4 \cdot \mathrm{dist}_{\mathsf{edit}}(\pi', \Pi_{2k})$。

□

D.7 支持集大小的字符串编辑距离

命题 7.11。设 $N, n \in \mathbb{N}$。设 Ψ 为 $\{0,1\}^N$ 中所有 n-块字符串构成的集合，Π 为 $\mathbb{N}$ 上支撑集大小至多为 n 的概率分布集合。那么，对于任意一个定义在 $\mathbb{N}$ 上、其密度为 $1/N$ 的整数倍的分布 π，以及 $x = \psi(\pi)$，

$$\mathrm{dist}_{\mathsf{edit}}(\pi, \Pi) \le \mathrm{dist}_{\mathsf{rel-edit}}(x, \Psi) \le 2 \cdot \mathrm{dist}_{\mathsf{edit}}(\pi, \Pi)\,.$$

证明。我们首先证明 $\mathrm{dist}_{\mathsf{edit}}(\pi, \Pi) \le \mathrm{dist}_{\mathsf{rel-edit}}(x, \Psi)$。取某个满足 $\mathrm{dist}_{\mathsf{rel-edit}}(x, y) = \mathrm{dist}_{\mathsf{rel-edit}}(x, \Psi)$ 的 $y \in \Psi$。那么，$\psi^{-1}(y) \in \Pi$，因为 $\psi^{-1}(y)$ 至多支撑于 n 个元素（根据观察 7.5），因此，利用引理 7.6，$\mathrm{dist}_{\mathsf{edit}}(\pi, \Pi) \le \mathrm{dist}_{\mathsf{edit}}(\pi, \psi^{-1}(y)) \le \mathrm{dist}_{\mathsf{rel-edit}}(x, y) = \mathrm{dist}_{\mathsf{rel-edit}}(x, \Psi)$。

现在我们证明 $\mathrm{dist}_{\mathsf{edit}}(\pi, \Pi) \ge \frac{1}{2}\mathrm{dist}_{\mathsf{rel-edit}}(x, \Psi)$。如果我们能够证明，对于某些 $\pi^* \in \Pi$，其密度是 $1/N$ 的整数倍，有 $\mathrm{dist}_{\mathsf{edit}}(\pi, \Pi) = \mathrm{dist}_{\mathsf{edit}}(\pi, \pi^*)$，那么就完成了：这样的 π^* 将满足 $\psi(\pi^*) \in \Psi$，因此结论将由引理 7.6 得出。

设 $\mathcal{D}$ 为 $\mathbb{Z}$ 上与 π 具有相同密度的概率分布（即，我们只需将定义域从 $\mathbb{N}$ 扩展到 $\mathbb{Z}$），并设 $f: \mathbb{Z} \to \{0,1\}$ 为满足 $\pi = \pi_{f,\,\mathcal{D}}$ 的函数，该函数必定存在。

令 Ξ 为带标签分布 $(h, \mathcal{F})$ 的性质，使得 $\pi_{h,\mathcal{F}}$ 的支持集大小至多为 n；等价地，使得 π_h，$\mathcal{F} \in \Pi$。根据命题 6.37，$\mathrm{dist}_{\mathsf{edit}}((f,\mathcal{D}) \Xi) \ge \mathrm{dist}_{\mathsf{TV}}((f,\mathcal{D}) \Xi)$。我们声称存在 $(g, \mathcal{E}) \in \Xi$，使得 $\mathrm{dist}_{\mathsf{TV}}((f,\mathcal{D}) \Xi) = \mathrm{dist}_{\mathsf{TV}}((f,\mathcal{D}), (g,\mathcal{E}))$，此外，$\mathcal{E}$ 的所有密度都是 $1/N$ 的整数倍。

设 $(g, \mathcal{E}) \in \Xi$ 使得 $\mathrm{dist}_{\mathsf{TV}}((f,\mathcal{D}), (g,\mathcal{E})) = \mathrm{dist}_{\mathsf{TV}}((f,\mathcal{D}) \Xi)$成立。我们首先断言，$\mathrm{supp}(\mathcal{D}) \cap \mathrm{supp}(\mathcal{E}) \neq \emptyset$，此外，存在 $i^* \in \mathrm{supp}(\mathcal{D}) \cap \mathrm{supp}(\mathcal{E})$，使得 $f(i^*) = g(i^*)$。事实上，假设情况并非如此。则 $\mathrm{dist}_{\mathsf{TV}}((f,\mathcal{D}), (g,\mathcal{E})) = 1$。另一方面，带标签分布 $(f, \mathcal{D}')$ 满足 $\mathrm{dist}_{\mathsf{TV}}((f,\mathcal{D}), (f,\mathcal{D}')) < 1$，其中 $\mathcal{D}'$ 是一个支撑于任意 $i \in \mathrm{supp}(\mathcal{D})$ 的单点分布，这与我们对 $(g, \mathcal{E})$ 的选择矛盾，因为 $\pi_{f,\mathcal{D}'}$ 支撑于单个元素，因此 $(f, \mathcal{D}') \in \Xi$。固定任意一个这样的 i^*。

我们首先从 $\mathcal{E}$ 构造一个满足三个条件的分布 $\mathcal{E}'$：

1. $(g, \mathcal{E}') \in \Xi$. 2. 对每个 $i \in \mathrm{supp}(\mathcal{E}')$，$g(i) = f(i)$；以及 3. $\mathrm{dist}_{\mathsf{TV}}((f,\mathcal{D}), (g,\mathcal{E}')) \le \mathrm{dist}_{\mathsf{TV}}((f,\mathcal{D}), (g,\mathcal{E}))$;

我们按如下方式构造 $\mathcal{E}'$：对于每个 $i \in \mathbb{N}$,

1. 若 $i \neq i^*$ 和 $g(i) = f(i)$，集合 $\mathcal{E}'(i) = \mathcal{E}(i)$。 2. 若 $g(i) \neq f(i)$，集合 $\mathcal{E}'(i) = 0$。
3. 集合 $\mathcal{E}'(i^*) = 1 - \mathcal{E}'(\mathrm{supp}(\mathcal{E}) \setminus \{i^*\})$】【。

根据构造，$\mathcal{E}'$ 是一个概率分布。第一个条件很容易验证：注意到 $\mathrm{supp}(\mathcal{E}') \subseteq \mathrm{supp}(\mathcal{E})$，因此我们有 $|\mathrm{supp}(\pi_{g,\mathcal{E}'})| \le |\mathrm{supp}(\pi_{g,\mathcal{E}})| \le n$，从而 $(g, \mathcal{E}') \in \Xi$。

第二个条件由 $\mathcal{E}'$ 的构造以及 i^*的选择可知成立。最后，我们验证第三个条件。根据命题 6.6，我们有

$$
\begin{aligned}
&\mathsf{dist}_{\mathsf{TV}}((f,\mathcal{D}),(g,\mathcal{E}')) \\
&= \frac{1}{2}\sum_{i\in\mathbb{N}} \mathbb{1}\left[f(i)\neq g(i)\right](\mathcal{D}(i)+\mathcal{E}'(i)) + \mathbb{1}\left[f(i)=g(i)\right]\left|\mathcal{D}(i)-\mathcal{E}'(i)\right| \\
&= \frac{1}{2}\sum_{i\neq i^*}\left[\mathbb{1}\left[f(i)\neq g(i)\right]\mathcal{D}(i) + \mathbb{1}\left[f(i)=g(i)\right]\left|\mathcal{D}(i)-\mathcal{E}(i)\right|\right] \\
&\quad + \frac{1}{2}\left[\left|\mathcal{D}(i^*) - \left(1-\sum_{i\neq i^*}\mathbb{1}\left[f(i)=g(i)\right]\mathcal{E}(i)\right)\right|\right] \\
&= \frac{1}{2}\sum_{i\neq i^*}\left[\mathbb{1}\left[f(i)\neq g(i)\right]\mathcal{D}(i) + \mathbb{1}\left[f(i)=g(i)\right]\left|\mathcal{D}(i)-\mathcal{E}(i)\right|\right] \\
&\quad + \frac{1}{2}\left[\left|\mathcal{D}(i^*) - \left(\mathcal{E}(i^*)+\sum_{i\in\mathbb{N}}\mathbb{1}\left[f(i)\neq g(i)\right]\mathcal{E}(i)\right)\right|\right] \\
&\leq \frac{1}{2}\sum_{i\in\mathbb{N}}\left[\mathbb{1}\left[f(i)\neq g(i)\right](\mathcal{D}(i)+\mathcal{E}(i)) + \mathbb{1}\left[f(i)=g(i)\right]\left|\mathcal{D}(i)-\mathcal{E}(i)\right|\right] \\
&= \mathsf{dist}_{\mathsf{TV}}((f,\mathcal{D}),(g,\mathcal{E}))\,,
\end{aligned}
$$

该不等式即为三角不等式。

现在，我们构造 $\mathcal{E}''$，使其从 $\mathcal{E}'$ 满足三个条件：

1. $(g,\ \mathcal{E}'') \in \Xi$。
2. $\mathcal{E}''$ 的每个密度都是 $1/N$ 的整数倍；并且
3. $\mathsf{dist}_{\mathsf{TV}}((f,\mathcal{D}),(g,\mathcal{E}'')) \leq \mathsf{dist}_{\mathsf{TV}}((f,\mathcal{D}),(g,\mathcal{E}'))$;

我们按如下方式构造 $\mathcal{E}''$：对于每个 $i \in \mathsf{supp}(\mathcal{E}') \setminus \{i^*\}$，令 $\mathcal{E}''(i) = \mathcal{D}(i)$；并令 $\mathcal{E}''(i^*) = 1 - \mathcal{E}''(\mathsf{supp}(\mathcal{E}') \setminus \{i^*\})$（其他位置取 0）。同样，根据构造，$\mathcal{E}''$ 是一个概率分布。可以验证，$\mathsf{supp}(\mathcal{E}'') \subseteq \mathsf{supp}(\mathcal{E})$，因此 $(g,\ \mathcal{E}'') \in \Xi$，从而满足第一个条件。第二个条件成立，因为每个 $\mathcal{D}(i)$ 都是 $1/N$ 的倍数（回想一下，$\mathcal{D}$ 与 π 具有相同的密度）。还可以验证

$$
\mathsf{dist}_{\mathsf{TV}}((f,\mathcal{D}),(g,\mathcal{E}'')) = \sum_{i\in\mathbb{N}\setminus(\mathsf{supp}(\mathcal{E}')\cup\{i^*\})} \mathcal{D}(i) \leq \mathsf{dist}_{\mathsf{TV}}((f,\mathcal{D}),(g,\mathcal{E}'))\,,
$$

满足第三个条件。

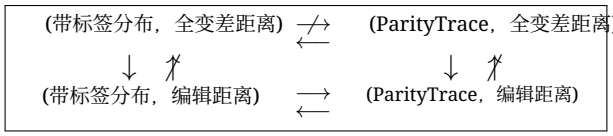
由此可得 $\mathsf{dist}_{\mathsf{TV}}((f,\mathcal{D})\ \Xi) = \mathsf{dist}_{\mathsf{TV}}((f,\mathcal{D}),\ (g,\mathcal{E}''))$。令 $\pi^* := \pi_{g,\mathcal{E}''}$。那么 π^* 的密度都是 $1/N$ 的倍数，并且回想起 $\pi = \pi_{f,\mathcal{D}}$，

$$
\begin{aligned}
\mathsf{dist}_{\mathsf{TV}}((f,\mathcal{D}),(g,\mathcal{E}'')) &\geq \mathsf{dist}_{\mathsf{edit}}(\pi,\pi^*) && \text{（编辑距离的定义）} \\
&= \mathsf{dist}_{\mathsf{edit}}((f,\mathcal{D}),(g,\mathcal{E}'')) && \text{（编辑距离的定义）} \\
&\geq \mathsf{dist}_{\mathsf{edit}}((f,\mathcal{D}),\Xi) && \text{（由于 } (g,\ \mathcal{E}'') \in \Xi\text{）} \\
&\geq \mathsf{dist}_{\mathsf{TV}}((f,\mathcal{D}),\Xi) && \text{（命题 6.37）} \\
&= \mathsf{dist}_{\mathsf{TV}}((f,\mathcal{D}),(g,\mathcal{E}'')) && \text{（以上结论）。}
\end{aligned}
$$

因此等式成立，且 $\mathrm{dist}_{\mathrm{edit}}(\pi, \pi^*) = \mathrm{dist}_{\mathrm{edit}}((f, \mathcal{D})\ \Xi)$。我们实际上要证明，$\mathrm{dist}_{\mathrm{edit}}((f, \mathcal{D})\ \Xi) = \mathrm{dist}_{\mathrm{edit}}(\pi, \Pi)$。事实上，对于任意的 $(h, \mathcal{F}) \in \Xi$，有 $\pi_{h,\mathcal{F}} \in \Pi$，因此 $\mathrm{dist}_{\mathrm{edit}}(\pi, \Pi) \le \mathrm{dist}_{\mathrm{edit}}(\pi_{f,\mathcal{D}}, \pi_{h,\mathcal{F}}) = \mathrm{dist}_{\mathrm{edit}}((f, \mathcal{D}), (h, \mathcal{F}))$，所以 $\mathrm{dist}_{\mathrm{edit}}(\pi, \Pi) \le \mathrm{dist}_{\mathrm{edit}}((f, \mathcal{D})\ \Xi)$。类似地，对于任意 $\pi' \in \Pi$，我们可以构造 $(h, \mathcal{F})$，使得 $\pi_{h,\mathcal{F}} = \pi'$，从而 $(h, \mathcal{F}) \in \Xi$，因此 $\mathrm{dist}_{\mathrm{edit}}((f, \mathcal{D})\ \Xi) \le \mathrm{dist}_{\mathrm{edit}}((f, \mathcal{D}), (h, \mathcal{F})) = \mathrm{dist}_{\mathrm{edit}}(\pi_{f,\mathcal{D}}, \pi_{h,\mathcal{F}}) = \mathrm{dist}_{\mathrm{edit}}(\pi, \pi')$，进而 $\mathrm{dist}_{\mathrm{edit}}((f, \mathcal{D})\ \Xi) \le \mathrm{dist}_{\mathrm{edit}}(\pi, \Pi)$。我们构造出了所需的 π^*，证明完毕。 □

E 带标签分布测试与奇偶迹的比较

我们仍需证明 $\not\to$ 图 1 所示的图 1中的关系，为方便起见，我们在此重述：



(带标签分布，全变差距离) $\not\to$ (ParityTrace，全变差距离) 是 命题 E.1，而 (带标签分布，编辑距离) $\not\to$ (带标签分布，全变差距离) 是 命题 E.2。剩余的箭头可由传递性推出。

回顾一下，任意密度性质 Ξ 都对应一个分布性质 Π，反之亦然。

命题 E.1。对于每个足够小的 $\epsilon > 0$ 以及每个 $m \in \mathbb{N}$，存在一个关于 $\mathbb{N}$ 上分布的性质 Π及其对应的密度性质 $\Xi = \Xi(\Pi)$，使得

1. 存在一个 $(\Xi,\ \mathrm{FAR}_{\epsilon}^{\mathsf{TV}}(\Xi)\ 3/4)$-带标签分布测试器，其样本复杂度为 $O(1/\epsilon^2)$；
2. 不存在样本复杂度为 m 的、针对奇偶迹的 $(\Pi,\ \mathrm{FAR}_{1/2}^{\mathsf{TV}}(\Pi)\ 2/3)$-分布测试器。

证明概要。令 $\delta = O(1/m)$，并令 $\Pi := \{\pi^*\}$，其中 π^* 是定义在 $\mathbb{N}$ 上的性质，其密度为 $\left(\frac{1-\delta}{2}, \frac{\delta}{2}, \frac{1-\delta}{2}, 0, 0, \ldots\right)$。现在我们证明，$\Pi$和 $\Xi = \Xi(\Pi)$ 满足陈述中的两个性质。

高效带标签分布测试器。我们概述如何使用命题 6.18中的通过学习进行测试的方法，构造一个 $(\Xi,\ \mathrm{FAR}_{\epsilon}^{\mathsf{TV}}(\Xi)\ 3/4)$-带标签分布测试器。为此，我们需要为 Ξ 给出一个样本复杂度为 $O(1/\epsilon^2)$ 的学习器-验证器对。

学习器 A 从 $\mathcal{D}_f$ 中抽取大小为 $O(1/\epsilon^2)$ 的样本，并用它来获得 $\mathcal{D}$中位数的估计值 $\boldsymbol{y}$。然后，它生成一个函数 $\boldsymbol{g} : \mathbb{Z} \to \{0, 1\}$，定义为 $\boldsymbol{g}(x) = 1$，其中 $x \neq \boldsymbol{y}$，且 $\boldsymbol{g}(\boldsymbol{y}) = 0$。注意，对于适当的分布 $\mathcal{E}$，有 $(\boldsymbol{g}, \mathcal{E}) \in \Xi$。接下来只需证明，当 $(f, \mathcal{D}) \in \Xi$时，输出 $\boldsymbol{g}$ 也满足定义 6.17 中的其他条件；具体而言，以较高的常数概率，$\mathrm{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{D}_{\boldsymbol{g}}) < \epsilon/4$，并且对于某个满足 $(\boldsymbol{g}, \mathcal{E}) \in \Xi$ 的 $\mathcal{E}$，有 $\mathrm{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{E}) < \epsilon/4$。

主要思路是，根据霍夫丁不等式，$\boldsymbol{y}$ 将以 $O(\epsilon)$的精度接近$\mathcal{D}$的真实中位数；按照 Ξ的定义，这个点应被赋予零值和 δ 的质量。更准确地说，通过尽可能选择接近样本中位数的 $\boldsymbol{y}$，同时注意不要选择样本中任何值为 1 的元素，我们就能以足够高的概率保证以下条件成立：1）左侧和右侧的总 $\mathcal{D}$质量与 $\boldsymbol{y}$ 之差仅为 $O(\epsilon)$；2）$\mathcal{D}(\boldsymbol{y}) = O(\epsilon)$。第一个条件足以确保，对于某个满足（$\boldsymbol{g}$，$\mathcal{E}$）$\in \Xi$的 $\mathcal{E}$，$\mathrm{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{E}) < \epsilon/4$。这是因为 $\boldsymbol{y}$ 左右两侧的 $\mathcal{D}$质量足够接近

所需值$\frac{1-\delta}{2} \approx \frac{1}{2}$；这也是为了满足$(g, \mathcal{E}) \in \Xi$，任一值域中需要包含的$\mathcal{E}$质量。由命题 D.2 可知，只要$y$左右两侧的质量在$O(\epsilon)$的误差范围内，就可以选取特定的$\mathcal{E}$，使得$\mathrm{dist}_{\mathsf{TV}}(\mathcal{D}, \mathcal{E})$很小。随后，第二个条件确保，如果$f(y) = 1$（否则$y$就会选择不同的值），那么$\mathcal{D}(y)$就足够小于$\epsilon$，因此赋值$g(y) = 0$不会使$\mathrm{dist}_{\mathsf{TV}}(\mathcal{D}_f, \mathcal{E}_g)$过大（若$\mathcal{D}_f$和$\mathcal{E}_g$在某个具有$\Omega(\epsilon)$ $\mathcal{D}$质量的元素上取值不一致，就会发生这种情况）。综上可知，A是一个适当学习器。

现在我们概述验证器B_g。设$g = g_y$。那么，对于输入$\mathcal{D}$（它是定义在$\mathbb{Z}$上的一个分布），B_g必须区分以下两种情况：$\mathcal{D} \in \mathrm{CLOSE}^{\mathsf{TV}}_{\epsilon/4}(\Pi_g)$和$\mathcal{D} \in \mathrm{FAR}^{\mathsf{TV}}_{\epsilon/2}(\Pi_g)$。考虑分布$\pi_{g,\mathcal{D}}$在$\mathbb{N}$上。注意，$B_g$可以通过抽取一个样本$x \sim \mathcal{D}$，并根据$x < y$、$x = y$还是$x > y$，将其映射到$\{1, 2, 3\}$中的某个索引，从而从$\pi_{g,\mathcal{D}}$中采样。

B_g通过从π_g中抽取$O(1/\epsilon^2)$个点，并利用它们学习π_g，将误差控制在足够小的加性误差$O(\epsilon)$内。由此可知，B_g以较高的常数概率区分$\mathrm{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{D}}, \pi^*) \le \epsilon/4$与$\mathrm{dist}_{\mathsf{TV}}(\pi_{g,\mathcal{D}}, \pi^*) > \epsilon/2$。接下来可以证明，这等价于区分$\mathcal{D} \in \mathrm{CLOSE}^{\mathsf{TV}}_{\epsilon/4}(\Pi_g)$和$\mathcal{D} \in \mathrm{FAR}^{\mathsf{TV}}_{\epsilon/2}(\Pi_g)$。直观地说，这是因为决定$\mathcal{D}$与$\Pi_g$之间距离的唯一因素，是它在$y$附近的密度与目标向量$\left(\frac{1-\delta}{2}\right.$的差距。形式上，其中一个方向需要再次应用命题 D.2。

因此，(A, B)是一个针对Ξ的学习器-验证器对，成功概率为$3/4$，误差为ϵ，样本复杂度为$O(1/\epsilon^2)$。根据 命题 6.18，存在一个$(\Xi, \mathrm{FAR}^{\mathsf{TV}}_{\epsilon}(\Xi))$ 3/4-带标签分布测试器，其样本复杂度为$O(1/\epsilon^2)$。

奇偶迹下高效测试器的不存在性。考虑分布π，其定义为$\pi(1) = 1$，并且$\mathrm{dist}_{\mathsf{TV}}(\pi, \pi^*) > 1/2$。然而，在奇偶迹下，唯一能够区分$\pi$和$\pi^*$的事件是从$\pi^*$采样到一个值为 0 的符号；对每个采样元素而言，该事件发生的概率为$\pi^*(2) = \delta$。因此，任何使用$o(1/\delta)$个样本的测试器，都无法以非可忽略的概率区分π和π^*。 □

命题 E.2. 对于任意充分小的$\epsilon > 0$以及任意$m \in \mathbb{N}$，存在一个关于$\mathbb{N}$上分布的性质Π，以及相应的密度性质$\Xi = \Xi(\Pi)$，使得

1. 存在一个$(\Xi, \mathrm{FAR}^{\mathsf{edit}}_{\epsilon}(\Xi))$ 3/4 带标签分布测试器，其样本复杂度为$O(1/\epsilon^2)$；2. 不存在$(\Xi, \mathrm{FAR}^{\mathsf{TV}}_{\epsilon}(\Xi))$ 2/3 带标签分布测试器，其样本复杂度为m。

证明概要。令$n = \Theta(m^2\epsilon^4)$为一个整数。令$\Pi$为所有满足以下条件、支撑于$\mathbb{N}$上的分布$\pi$所构成的类：1）奇数上的总密度恰好为$1/2$；2）对于每个$i \in \mathbb{N}$，$\pi(i) \le 1/n$。令$\Xi = \Xi(\Pi)$为相应的密度性质。

高效编辑距离测试器的存在性。我们观察到，在编辑距离下，Π的第二个要求基本没有影响，具体如下所述。令$\mathcal{O}$表示正奇数集合。则$\mathrm{dist}_{\mathsf{edit}}(\pi, \Pi) > \epsilon$意味着$\left|\pi(\mathcal{O}) - \frac{1}{2}\right| > \epsilon$。原因是，给定一个带标签分布$(f, \mathcal{D})$，满足$\pi_{f,\mathcal{D}} = \pi$具有足够小的逐点质量，且非零项之间有足够的间隔（这总能做到而不影响$\pi_{f,\mathcal{D}}$），我们可以在$\mathcal{D}$中将$\left|\pi(\mathcal{O}) - \frac{1}{2}\right|$质量移至偶数元素和奇数元素之间，从而满足$\Pi$的第一个条件；然后利用非零项之间的足够间隔，改变$f$在零质量点处的取值，以打破任何质量超过$1/n$的交替，从而无需额外代价即可满足$\Pi$的第二个条件。

因此，以下算法A以较高的常数概率区分Ξ与$\mathrm{FAR}^{\mathsf{edit}}_{\epsilon}(\Xi)$：采集$O(1/\epsilon^2)$个样本，并将值为 1 的样本点的经验频率$\hat{o}$用作

$\pi(\mathcal{O})$ 的估计值，误差为 $\epsilon/4$ 加性误差。当且仅当 $\left|\widehat{o} - \frac{1}{2}\right| < \epsilon/2$ 时接受_ .

不存在高效的全变差距离测试器。我们将标准模型下对 $[n]$ 上分布的均匀性测试问题，归约为 $(\Xi, \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi))$-带标签分布测试。假设算法 A 是一个 $(\Xi, \mathrm{FAR}_\epsilon^{\mathsf{TV}}(\Xi)\ 2/3)$-带标签分布测试器。那么，我们的算法 B 在标准模型下区分 $[n]$ 上的均匀分布与 n 上那些在全变差距离意义下与均匀分布相差至少 ϵ 的分布，具体如下。

对于输入分布 π 上的 $[n]$，设 $f, \mathcal{D}$ 为由 $f(x) = \mathsf{par}(x)$ 在 $x \geq 1$ 上以及 $f(x) = 1$ 在 $x \leq 0$ 上给出的带标签分布，且 $\mathcal{D} = \pi$。由此可得 $\pi = \pi_{f,\mathcal{D}}$。因此， B 在输入 π 上，可以通过采样 $\boldsymbol{x} \sim \pi$ 并在 A 请求从 $(f, \mathcal{D})$ 中获取一个样本时，生成 $(\boldsymbol{x}, \mathsf{par}(\boldsymbol{x}))$，从而模拟 A 在输入 $(f, \mathcal{D})$ 上的行为。

如果 π 在 $[n]$上均匀分布，则可得 $(f,\ \mathcal{D}) \in \Xi$。另一方面，如果 π 支撑于 $[n]$，且与均匀分布的全变差距离为 ϵ，那么其超出 $1/n$ 的总密度为 $\sum_i \mathsf{max}\{0, \pi(i) - 1/n\} > \epsilon$。因此，$\mathsf{dist}_{\mathsf{TV}}(\mathcal{D}_f, \Xi) > \epsilon$。由此可知， B 以至少 $2/3$的概率正确接受或拒绝。由于在标准模型中进行均匀性检验需要 $\Omega(\sqrt{n}/\epsilon^2)$ 个样本，因此 A 的样本复杂度至少为 $\Omega(\sqrt{n}/\epsilon^2)$。 □