arXiv:2304.06733v1 [cs.LG] 13 Apr 2023

贝叶斯网络的近最优度测试

Vipul Arora*[1]、Arnab Bhattacharyya†[1]、Cl´ement L.Canonne[2], 和 Joy Qiping Yang‡[1,2]

[1]新加坡国立大学。 {vipul, arnab, jyang}@comp.nus.edu.sg。[2]悉尼大学。
clement.canonne@sydney.edu.au, qyan6238@uni.sydney.edu.au

摘要

本文研究了如何检验未知概率分布 P所对应贝叶斯网络的最大入度问题；给定该分布 P的样本访问。我们证明，该问题的样本复杂度为 $\tilde{\Theta}(2^{n/2}/\varepsilon^2)$。我们的算法依赖于一个通过学习进行检验框架，该框架此前曾用于构造样本最优测试器；为应用这一框架，我们开发了用于贝叶斯网络“近似适当”学习以及在 χ^2散度下进行高概率学习的新算法，这些算法本身也具有独立的研究价值。

1 引言

对高维分布建模，最自然且广泛使用的方法之一是将其表示为贝叶斯网络（简称贝叶斯网络） [Pea88]。具体而言，定义在 $\{0,1\}^n$ 上的贝叶斯网络由一个定义在 n 个顶点上的有向无环图（DAG） G ，以及概率分布 p_i、π 构成；对于所有 $i \in [n]$，以及图中第 i 个节点的所有父节点赋值 π ，这些分布均定义在 $\{0,1\}$ 上。要生成一个 n 维样本，需按 G 的拓扑序对节点进行采样，其中第 i 个节点根据 p_i、π 进行采样，赋值 π 已由 i 的父节点样本确定。贝叶斯网络能够自然地编码因果信息 [Pea95], , 而学习概率分布的贝叶斯网络描述被视为统计学和机器学习中的一个基础问题 [Hec98]。例如，Sachs 等人 [SPP+05] 使用这种方法，根据基因表达数据发现了蛋白质调控网络。

足够复杂的贝叶斯网络可以编码任意分布，因此通常不可能学习一般的贝叶斯网络。相反，人们通常会将范围限制在底层有向无环图的入度有界的贝叶斯网络。具有稀疏贝叶斯网络描述的分布自然出现在机器学习、机器人学、自然语言处理 [WJ08], 医学和计算生物学中 [FLNP00]。此外，从信息论角度而言，已知入度不超过 d 的贝叶斯网络可以通过 ε 个 $\tilde{O}(n2^d/\varepsilon^2$样本学习到全变差距离（TV） [CDKS17]以内。因此，贝叶斯网络的最大入度是一个值得考虑的重要建模参数。

在这项工作中，我们研究检验一个分布是否属于度数为 d 的贝叶斯网络概念类的问题，也就是说，其入度至多为 d。具体而言，给定对定义在 $\{0,1\}^n$ 上的分布 P 的样本访问，我们考虑如下性质检验问题：P 是否由度数为 d 的贝叶斯网络描述，或者它与所有此类贝叶斯网络的全变差距离是否至少为 P ε?

Theorem 1.1. (Main Theorem, Informal, See Theorem 5.1) 给定定义在 $\{0,1\}^n$ 上的未知分布 P 和度参数 $d \ll n$，检验 P 是否相对于任意一个度为 d 的贝叶斯网络满足马尔可夫性，其样本复杂度为 $\tilde{\Theta}\left(\frac{2^{n/2}}{\varepsilon^2}\right)$。

*Vipul 部分获得 NRF-AI Fellowship R-252-100-B13-281 的资助。
†A. Bhattacharyya 部分获得 MOE 二级奖项和亚马逊教师研究奖的资助。
‡Joy 部分获得 NRF-AI Fellowship R-252-100-B13-281 的资助。

我们的结果要求 $d < n/2 - \Omega(\log n)$，这与我们检验贝叶斯网络稀疏性的动机一致。本工作的主要贡献在于确立该定理的上界；我们指出，（几乎）匹配的下界可由 [BCY22，推论 B2]。

在推导这一上界的过程中，我们得到了三个新的学习结果，我们认为它们本身也具有独立的研究价值。（1）首先，是一个关于 χ^2 散度的高概率学习结果（命题 4.1）。众所周知，以恒定成功概率将离散定义域 Σ 上的任意分布 P 学习到 χ^2 散度 ε，所需的样本复杂度为 $\Theta(|\Sigma|/\varepsilon)$；然而，如何将这一成功概率提升至 $1-\delta$（其中 $\delta > 0$ 可任意小），此前仍是一个悬而未决的问题。特别是，能否像全变差距离以及（近期证明的）KL散度那样，使其对 δ 呈对数依赖，此前尚不明确。我们证明这确实可行：只需 $O\left(\frac{|\Sigma|}{\varepsilon}\log\frac{|\Sigma|}{\delta}\right)$ 个独立同分布样本，即可在 χ^2 散度下实现高概率学习。有趣的是，达到这一上界的估计量既不是经验估计量，也不是常用的拉普拉斯加一估计量，而是一个针对合适的 $K = K(\delta)$ 的加 K 估计量。

(2) 第二，对于度数为 d 的贝叶斯网络，我们给出了一个在 χ^2-散度下的近适当学习算法，其样本复杂度为 $\tilde{O}(2^d n^2/\varepsilon)$（定理 4.1）。值得注意的是，以往针对贝叶斯网络的学习算法，要么是在较弱的距离度量（全变差距离或 KL 散度）下进行学习，要么是不适当的，也就是说，它们输出的假设本身并不是度数为 d 的贝叶斯网络。相比之下，我们的算法输出一个真正的、度数为 d 的贝叶斯网络 P，以及定义域的一个子集 $S \subseteq \{0,1\}^n$，使得 (i) P, P 在 S 上的概率质量总和不低于 1-$O(\varepsilon)$，且 (ii) P 和 $\widehat{P}$ 限制在该子集 S 上时，其 χ^2 散度不超过 ε。

这种混合保证使我们的学习算法成为近适当学习算法，而非适当学习算法，这或许显得有些人为。然而，我们的第三个结果表明，从非常强的意义上说，这确实是必要的：

(3) 我们在 命题 4.2 中证明了适当学习在 χ^2散度下的样本复杂度下界，表明 任何 学习算法，只要其 χ^2 保证适用于整个定义域，就必须具有 $\Omega(2^{n/2}/\varepsilon)$ 的样本复杂度，即使对于一阶贝叶斯网络也是如此。

2 相关工作

尽管在各种设定下，使用不同距离度量学习图模型的研究历史悠久，在统计学和机器学习领域皆是如此，但直到最近，人们才开始研究对由（简洁的）图模型表示的未知分布进行性质检验这一任务。 [DDK19] 率先提出了对依赖结构由无向图建模的高维分布（即伊辛模型，或更一般地说，马尔可夫随机场）进行恒等性（拟合优度）检验和独立性检验的问题；随后，[DDK17、GNS18、NL19、BBC+20]等延续了这一研究方向，取得了一系列正面（算法）和负面（下界）结果。最近， [CDDK22] 提出了对潜在伊辛模型进行拟合优度检验的问题，其中只有树的叶节点可观测。

聚焦于另一类受到广泛研究的图模型，并行开展的工作 [CDKS17],和 [DP17] 研究了贝叶斯网络中类似的检验问题，其中底层依赖结构被建模为有向图。[ABDK18] 研究了 因果 贝叶斯网络中的相关任务，前提是能够对网络进行干预。最后，与我们的工作最为接近的是， [BCY22] 研究了贝叶斯网络的独立性检验 问题，并为此任务获得了近乎最优的样本复杂度界：判断一个给定的高维分布（已知它是稀疏贝叶斯网络）是否实际上是乘积分布。

我们工作的另一项贡献，即针对任意离散分布在 χ^2 散度下进行高概率学习，

[1]即，输出 $\widehat{P}$使得 $d_{\chi^2}(P,\widehat{P}) \le \varepsilon$的概率至少为 $9/10$。

延续了密度估计领域一系列悠久的研究成果，这些成果涉及在各种距离度量下进行密度估计（例如，[KOPS15，DL01，Dia16], 以及 [Can20], 及其中的参考文献）。尽管在 χ^2 距离度量 下以恒定成功概率学习任意分布早已得到充分理解，甚至极小极大估计率的最优首项常数也已确定 [KOPS15], ，但要获得高概率界一直颇为困难。即使是较弱的、在 Kullback–Leibler（KL）散度下学习这一概念，也直到最近才取得这样的高概率学习结果 [BGPV21]。

最后，据我们所知，此前尚无关于在χ^2 散度下学习贝叶斯网络的结果，即使是对于常数成功概率的情形也是如此；除了将贝叶斯网络视为定义在 Σ^n上的无结构概率分布所得到的平凡界之外。这再次与 KL 散度的情形形成对比：对于 KL 散度，最优的常数成功概率学习界、以及高概率学习界都是已知的（参见 [BGPV21], 及其中的参考文献）。

3 概述

我们的算法遵循 [ADK15], 提出、并已被数项工作采用的通过学习进行检验框架（例如， [DKW18、CDKL22]）。具体而言，如果要在全变差距离下检验的性质，在某种“更强”的距离概念下相对容易学习，例如 χ^2，那么可以先假设分布具有该性质，在 χ^2下学习该分布，然后使用“χ^2-TV 容忍型检验器” [ADK15] 来检验学习算法输出的假设是否接近真实分布，从而构造出高效且可能达到样本最优的检验器。关键在于，这些容忍型检验器不仅必须拒绝与目标在全变差距离下相距甚远的分布，还必须接受那些在 χ^2 下足够接近的分布（即“容忍度”）：现在，如果未知分布 P 确实具有该性质，那么学习算法会按预期工作，$\widehat{}$其输出 P 接近 P，因此检验器将予以接受。然而，如果 P 与该性质相距甚远，那么$\widehat{}$要么学习算法失败，且 P 与 P 相距甚远（检验器将拒绝）；要么学习算法仍然成功，但根据三角不等式，$\widehat{}$必然与该性质本身相距甚远（这可以被检验并检测出来，$\widehat{}$因为此时我们已经有了 P的显式描述）。

关键在于，这种“χ^2容错”检验器针对定义域为 $O(\sqrt{|\Sigma|}/\varepsilon^2$ 的概率分布，其样本复杂度为 Σ ——对于许多检验任务而言，这已是最优的（差一个常数因子），并且与“非容错”检验器的样本复杂度相当。因此，只要学习阶段所需的样本远少于 $O(\sqrt{|\Sigma|}/\varepsilon^2)$，整体方法就能得到一种样本高效的检验算法。（此外，这种方法还可以通过多种方式扩展，例如检验海灵格距离而非全变差距离 [DKW18、BCY22]。）

尽管通过学习进行检验的思路看起来相对直接，但本文的主要技术贡献在于获得了应用这一思路所需的学习算法：即一种高效的学习算法，能够以高概率适当地学习贝叶斯网络（两者均以 χ^2 衡量）。确实，令人颇为意外的是，在我们的工作之前，人们仍不清楚能否在 χ^2 散度下学习离散分布，同时将失败概率控制在 δ，而样本复杂度中只需付出 $O(\log(1/\delta))$ 的依赖。具体而言，基于应用 McDiarmid不等式或某种“中位数技巧”的“显而易见”方法，在 χ^2 散度下均会失效：前者是由于估计量高度敏感，后者则是因为三角不等式不成立。我们指出，通过马尔可夫不等式，可以直接得到差 指数级 更$O(1/\delta)$ 的依赖 [KOPS15]；然而，在需要将失败概率降至指数级小的场景中，例如本文的情形，这种代价会变得难以承受——因为我们需要同时良好地学习 $2^d \cdot n$ 许多条件分布，因此必须对学习算法的多次运行取并集界。

有趣的是，仅需在样本复杂度中以 $O(\log(1/\delta))$ 的形式依赖即可实现高概率学习，而以 $\tilde{O}(2^d \cdot n/\varepsilon)$ 学习贝叶斯网络，对于相对容易的d_{KL} 散度（即使如此，该高概率结果也只是最近才得到证明 [BGPV21]）。遗憾的是，在 KL 散度下进行学习所提供的保证要弱得多，而尝试将上述

“通过学习进行检验”框架中的 KL-TV 替代 χ^2-TV，只能得到宽松得多的$\tilde{O}(2^n/(n\cdot\varepsilon^2))$ 检验上界。

下面，我们给出一系列关于相对于 χ^2 散度的学习结果：(1) 使用 $O(|\Sigma|\cdot\log(1/\delta)/\varepsilon)$ 个样本进行高概率学习；(2) 一种近似适当的贝叶斯网络学习算法，其样本复杂度为 $\tilde{O}(2^d n^2/\varepsilon)$；以及 (3) 学习贝叶斯网络的指数级样本复杂度下界 $\Omega(2^{n/2}/\varepsilon)$。稍后在 第5节中，我们将以第二个结果为基础，迈出实现主要结果的第一步，即定理 定理 1.1中的最大入度检验算法。

预备知识。我们使用标准的渐近记号 $O(\cdot)$、$\Omega(\cdot)$ $o(\cdot)$ $\Theta(\cdot)$，以及（半）标准的~ ~ ~$O(\cdot)$、$\Omega(\cdot)$ 和 $\Theta(\cdot)$ 来隐藏论证中的多对数因子。由于我们完全关注贝叶斯网络的入度，除非另有说明，所有关于度的表述均默认指入度。我们用 $A\lesssim B$ 表示对于某个绝对常数 C，有 $A\leqslant C\cdot B$。

对于一个支撑集为 $\{0,1\}^n$ 的多元随机变量 X，我们用 X_i 表示其第 i 个分量（坐标）；对于贝叶斯网络，我们用 Π_i 表示 X_i 的父节点集合。

4 在 χ^2 中的学习

4.1 在 χ^2 中的高概率学习。我们的分析大体遵循[BGPV21定理 6.1 和命题 4 中对4] KL散度的分析，但有几处关键差异，使我们能够将其推广到 χ^2 散度，并获得严格优于先前工作的样本复杂度。具体而言，我们不再局限于标准的加1拉普拉斯估计器，而是分析更一般的加K 估计器，其中 取适当的非恒定值 K。回顾一下，给定K 来自定义域 N 上某个概率分布的独立同分布样本（ Σ 其经验计数为 N_1、...、$N_{|\Sigma|}$），加

$$\widehat{P}(i)=\frac{N_i+K}{N+K|\Sigma|},\qquad i\in\Sigma\,. \tag{4.1}$$

直观而言，参数 K 控制了我们估计量的平滑程度。由于 χ^2 散度是一种非常严格的距离度量（远比KL散度严格，更不用说全变差距离），要实现高概率学习保证，一个思路是增加平滑程度，以抵消“概率很低但后果灾难性”的事件所带来的风险；这类事件可能会使 χ^2 散度急剧增大。随后我们可以证明，在高概率情形下，令 $K=\Theta$ 为 $\log(1/\delta)$ 可达到所需的样本复杂度，且远优于拉普拉斯估计器（对应于 $K=1$）。

Proposition 4.1. 任取 $\delta\in(0,1]$, 并令 $K=\Theta(\log(1/\delta))$。给定来自字母表 Σ 上某个未知概率分布 P 的 N 个独立同分布样本，以至少 $1-\delta$ 的概率，加-K 估计器^会得到一个假设 P，使得

$$d_{\chi^2}(P,\widehat{P})\lesssim\frac{|\Sigma|\log(|\Sigma|/\delta)}{N}.$$

特别地，只需 $N=O\left(\frac{|\Sigma|}{\varepsilon}\log\frac{|\Sigma|}{\delta}\right)$ 个样本，就能以 $1-\delta$ 的概率将 P 学习到 χ^2 散度。

相比之下，对拉普拉斯估计器进行类似分析，只能得到更大的样本复杂度：$O\big((|\Sigma|/\varepsilon)\log^2(|\Sigma|/\delta)\big)$，在上相差一个对数因子。我们将在第 4.2 节中使用这一结果。

证明。 [对加法K 估计量] 设 Q 为该加法K 估计量在 N 个来自 P 的独立同分布样本下的输出。根据[BGPV21 的命题 4 的证明，4], 对于每个 i，我们有如下结论，其中 T_i 是元素 i 在 N 个样本中的计数，且该结论成立的概率至少为 $1-\delta$,

$$\left|P_i-\frac{T_i}{N}\right|\leqslant\sqrt{\frac{3P_i\log\left(\frac{2}{\delta}\right)}{N}}+\frac{3\log\left(\frac{2}{\delta}\right)}{N}.$$

下文中，我们用 $A \lesssim B$ 表示：对于某个绝对常数 C，$A \leqslant C \cdot B$。

$$
\begin{aligned}
|P_i - Q_i| &\leqslant \left|P_i - \frac{T_i}{N}\right| + \left|\frac{T_i}{N} - Q_i\right| \\
&= \left|P_i - \frac{T_i}{N}\right| + \left|\frac{T_i}{N} - \frac{T_i + K}{N + K|\Sigma|}\right| \\
&= \left|P_i - \frac{T_i}{N}\right| + \left|\frac{T_i|\Sigma|/N - 1}{N + K|\Sigma|}\right| K \\
&\leqslant \sqrt{\frac{3P_i \log\left(\frac{2}{\delta}\right)}{N}} + \frac{3\log\left(\frac{2}{\delta}\right)}{N} + \frac{T_i|\Sigma|/N}{N + K|\Sigma|}K + \frac{K}{N + K|\Sigma|} \\
&\leqslant \sqrt{\frac{3P_i \log\left(\frac{2}{\delta}\right)}{N}} + \frac{3\log\left(\frac{2}{\delta}\right)}{N} + \frac{K|\Sigma|}{N + K|\Sigma|}\frac{T_i}{N} + \frac{K}{N} \\
&\leqslant \sqrt{\frac{3P_i \log\left(\frac{2}{\delta}\right)}{N}} + \frac{3\log\left(\frac{2}{\delta}\right)}{N} + \frac{K}{N} \\
&\quad + \frac{K|\Sigma|}{N + K|\Sigma|}\left(\sqrt{\frac{3P_i \log\left(\frac{2}{\delta}\right)}{N}} + \frac{3\log\left(\frac{2}{\delta}\right)}{N} + P_i\right) \\
&\leqslant 2\sqrt{\frac{3P_i \log\left(\frac{2}{\delta}\right)}{N}} + \frac{6\log\left(\frac{2}{\delta}\right)}{N} + \frac{K|\Sigma|}{N + K|\Sigma|}P_i + \frac{K}{N}.
\end{aligned}
$$

我们希望证明，当最多取 $\frac{|\Sigma|}{\varepsilon^2}\log(|\Sigma|\delta^{-1})$ 个样本时，$\Pr\left[d_{\chi^2}(P,Q) = \sum_i \frac{(P_i - Q_i)^2}{Q_i} \leqslant C\varepsilon^2\right] \geqslant 1 - \delta$。

我们将分析分为两种情形：

如果 $P_i \leqslant \frac{C'}{N}\log\left(\frac{2}{\delta}\right)$，那么 $\sqrt{\frac{3P_i}{N}\log\left(\frac{2}{\delta}\right)} \leqslant \sqrt{3C'}\log\left(\frac{2}{\delta}\right)$ 且 $\frac{K|\Sigma|}{N+K|\Sigma|}P_i \leqslant \frac{C'}{N}\log\left(\frac{2}{\delta}\right)$，因此 $2\sqrt{\frac{3P_i}{N}\log\left(\frac{2}{\delta}\right)} + \frac{6\log\left(\frac{2}{\delta}\right)}{N} + \frac{K|\Sigma|}{N+K|\Sigma|}P_i + \frac{K}{N} \leqslant \left(2\sqrt{3C'} + 6 + C'\right)\log\left(\frac{2}{\delta}\right)\frac{1}{N} + \frac{K}{N}$。于是我们有

$$
\frac{(P_i - Q_i)^2}{Q_i} \lesssim \frac{\left(\log\left(\frac{1}{\delta}\right)\right)^2}{N^2 Q_i} + \frac{K^2}{N^2 Q_i} \lesssim \frac{\left(\log\left(\frac{1}{\delta}\right)\right)^2}{NK} + \frac{K}{N},
$$

由于 $Q_i \geqslant \frac{K}{N+K|\Sigma|} \geqslant \frac{K}{2N}$，对于 $N \geqslant C'K \cdot |\Sigma|$。[2]

如果 $P_i > \frac{C'}{N}\log\left(\frac{1}{\delta}\right)$，$|P_i - Q_i| \lesssim P_i\left(\frac{1}{\sqrt{C'}} + \frac{1}{C'} + \frac{1}{\frac{N}{K|\Sigma|}+1}\right) \lesssim P_i\left(\frac{1}{\sqrt{C'}} + \frac{1}{C'} + \frac{1}{C'+1}\right)$ 因此，对于足够大的 C'，$Q_i \geqslant P_i/2$，从而得到

$$
\begin{aligned}
\frac{(P_i - Q_i)^2}{Q_i} &\lesssim \frac{\left(\sqrt{\frac{P_i \log\left(\frac{1}{\delta}\right)}{N}} + \frac{K|\Sigma|}{N+K|\Sigma|}P_i + \frac{K}{N}\right)^2}{P_i} \\
&\lesssim \frac{\log\left(\frac{1}{\delta}\right)}{N} + \left(\frac{K|\Sigma|}{N + K|\Sigma|}\right)^2 P_i + \frac{K^2}{N^2 P_i} \\
&\lesssim \frac{\log\left(\frac{1}{\delta}\right)}{N} + \left(\frac{K|\Sigma|}{N + K|\Sigma|}\right)^2 P_i + \frac{K^2}{N \log\frac{1}{\delta}}.
\end{aligned}
$$

[2] 如果 $N < C'K|\Sigma|$，那么 $Q_i > \frac{1}{(1+C')|\Sigma|}$ 并且我们可以轻松构造一个反例，使得对于足够小的 ε，$|P_i - Q_i| > \varepsilon$。

综合这两种情形，并应用并集界，我们以至少 $1-|\Sigma|\delta$ 的概率得到：

$$\sum_i \frac{(P_i-Q_i)^2}{Q_i} \lesssim |\Sigma| \left(\frac{\left(\log\left(\frac{1}{\delta}\right)\right)^2}{NK} + \frac{K}{N} + \frac{\log\left(\frac{1}{\delta}\right)}{N} + \frac{K^2}{N\log\frac{1}{\delta}} \right) + \left(\frac{K|\Sigma|}{N+K|\Sigma|} \right)^2.$$

为分析各个单独项，使其总和不超过 ε^2，我们需要 $N \geqslant \frac{|\Sigma|}{\varepsilon^2} \cdot \max\left\{ \log^2(1/\delta)\, K, K, \log(1/\delta) \quad K^2 \log(1/\delta) \right.$ $\left. \varepsilon \cdot K \right\} \geqslant \frac{|\Sigma|}{\varepsilon^2} \cdot \max\left\{ \log^2(1/\delta)\, K, K, \log(1/\delta) \quad K^2 \log(1/\delta) \right\} \geqslant \frac{|\Sigma|}{\log(1/\delta)\,\varepsilon^2}$；最优值在 $K = \Theta(\log(1/\delta))$ 时取得。将 δ 重新缩放，以调整并集界 $|\Sigma|\delta$，即可得证。

4.2 贝叶斯网络的近适当学习算法 对于在布尔超立方体 $\{0,1\}^n$ 上学习一般分布，已知需要且只需 $\Theta(2^n/\varepsilon^2)$ 个样本，即可在 χ^2 散度意义下完成学习 ε^2 （成功概率为常数）。由此可直接得到一种在整个支撑集上学习贝叶斯网络的非适当算法，其样本开销至多为 $O(2^n/\varepsilon^2)$。然而，在我们的研究背景下，这种方法并不值得关注：学习所需的样本比测试还多，所得分布不是贝叶斯网络，而且由于 χ^2 中缺乏三角不等式，我们无法在贝叶斯网络空间中找到足够接近的分布，使其成为适当学习算法。事实上，我们（尚）不清楚能否在 χ^2 中以 $O(2^n/\varepsilon^2)$ 的复杂度适当学习贝叶斯网络。

尽管人们可能希望，由于贝叶斯网络的描述较为稀疏（$2^d \cdot n$ 相较于 2^n），我们或许可以绕过 $2^n/\varepsilon^2$中提出的 Ω([KOPS15] 下界，并可能得到好得多的上界。然而，我们将在第 4.6 节证明，与 KL 散度学习形成鲜明对比的是，在 χ^2 中学习稀疏贝叶斯网络的样本复杂度仍然是指数级困难的。

粗略地说，我们的下界通过一个高度偏置的公共父节点隐藏了一个“特殊节点”。按照构造，除非学习算法获取指数级数量的样本，否则永远无法观测到这个“罕见”的特殊节点，因此会产生估计误差，而该误差完全源于这个高度偏置、出度很大的父节点。虽然在全变差距离甚至 KL 散度下，这种估计误差尚可接受，但 χ^2 散度对低概率元素的脆弱性会导致整体估计误差非常大。因此，任何针对整个支撑集的 χ^2 学习算法都不能容忍在这些罕见节点上出现不准确，因而必须获取指数级数量（关于 n）的样本。

尽管如此，鉴于检验才是我们的最终目标，我们只需充分学习支撑集中的大多数部分即可继续进行，从而绕过这一先前的下界。具体而言，对于全变差距离检验，只需保证在支撑集的大多数部分上，$\widehat{P}$ 与 P 在 χ^2中接近 S:

$$d_{\chi^2}(P_S, \widehat{P}_S) = d_{\chi^2}(P, \widehat{P}, S) = \sum_{i\in S} \frac{(P_i - \widehat{P}_i)^2}{\widehat{P}_i},$$

其中 $\sum_{i\in S} P_i = P(S) \geqslant 1 - O(\varepsilon)$。对于海灵格距离，我们需要稍强一些的条件，但思路相同。这样的框架已见于 [ADK15、DKW18], ，不过这仅隐含在 [DKW18] 对海灵格距离的分析中。因此，剩下的主要问题是能否获得这样一种样本效率高的贝叶斯网络近适当学习算法。

回到度-1下界，主要困难在于偏置父节点造成的信息瓶颈。通过将问题放宽为近适当学习，我们可以干脆放弃学习那一小部分罕见父节点（只要其规模足够小），因为这些质量之和不会超过 $O(\varepsilon)$。由此，一个自然的想法是从支撑集中排除一个子集，其中子节点及其父节点的质量至多为 $O(\varepsilon/(2^d \cdot n))$。[3] 这一保证足以满足我们的

[3]最多有 $2^d \cdot n$ 个父节点配置；按此阈值将它们排除后，仍可剩下质量为 $1 - O(\varepsilon)$的一部分 .

主要结果，即相对于全变差距离进行检验；然而，要将其扩展到更严格的海林格检验保证，就需要进一步加强这一条件，将移除的子集质量限制为至多 $O(\varepsilon^2)$，而不是$O(\varepsilon)$。由于这一扩展几乎不改变证明，却能得到更强的结果，因此在接下来的分析中，我们将重点实现这一目标。

为节省篇幅并便于排版，本节使用以下简写：

$$P(x_i, \pi_i(x)) := P_{X_i,\Pi_i}(x_i, \pi_i(x)). \tag{4.2}$$

我们定义一个感兴趣的支撑集，以便在其上进行学习：

$$S'_k := \left\{ x \in \{0,1\}^k \mid \forall i \in [k], P(x_i, \pi_i(x)) \geqslant 4c\frac{\varepsilon^2}{2^{d+1}n} \right\}. \tag{4.3}$$

我们还定义 S_k 的一个超集 S'_k：

$$S_k := \left\{ x \in \{0,1\}^k \mid \forall i \in [k], P(x_i, \pi_i(x)) \geqslant c\frac{\varepsilon^2}{2^{d+1}n} \right\}. \tag{4.4}$$

~ ~下文中，我们假设算法获得一个集合 S_k 满足 $S'_k \subseteq S_k \subseteq S_k$。事实上，引理 4.2在 第 4.3 节 中通过算法 1进行了分析，并保证我们能够以高概率高效地学习这样一个集合。

算法 1： ε^2-多数支持集识别

输入：对分布 P 的样本访问、精度参数 ε 和一个有向无环图 G。1 从 P 中抽取 m 个样本，构成多重集 S，其中 $m = 3 \cdot 2^{d+1}n\log(6 \cdot 2^{d+1}n)/(c \cdot \varepsilon^2)$。2 令 N_{x_i}、π_i 分别表示 (x_i, π_i) 在 S 中出现的次数。3 令 Z_{x_i}、$\pi_i = {}^{N_{x_i}}$、$_{m}{}^{\pi_i}$ 和 $\bar{S} :=\varnothing$。4 对于 $i = 1;\quad i \leqslant n;$ $i = i + 1$，执行以下操作。5 对于 $(x_i, \pi_i) \in \{0,1\}^{|\Pi_i|+1}$，执行以下操作。6 如果 Z_{x_i}、$\pi_i \leqslant 2c\frac{\varepsilon^2}{2^{d+1}n}$，则 $\bar{S} \leftarrow \bar{S} \cup \{i, (x_i, \pi_i)\}$。7 结束。8 结束。9 将 S 中所有满足 $X_i = x_i$、$\Pi_i = \pi_i$ 的配对标记为排除项（从 $\{0,1\}^n$ 中），并返回剩余的支撑集。

一种直接解决这个问题的方法是考虑对所有条件分布的学习保证，即 $d_{\chi^2}(P_{X_i|\Pi_i}, Q_{X_i|\Pi_i}) \leqslant \frac{\varepsilon^2}{n}$对任意 $\Pi_i = \pi_i$都成立，因此[4]

$$1 + d_{\chi^2}(P, Q) \leqslant \prod_{i=1}^{n}(1 + \max_{\pi_i} d_{\chi^2}(P_{X_i|\Pi_i=\pi_i}, Q_{X_i|\Pi_i=\pi_i})) \leqslant \left(1 + \frac{\varepsilon^2}{n}\right)^n \leqslant 1 + \varepsilon^2.$$

结合父节点上的概率质量下界为$\frac{\varepsilon}{2^d n}$这一事实，我们可以通过为每个 $O\left(\frac{n}{\varepsilon^2}\right)$ 额外付出 $O\left(\frac{2^d n}{\varepsilon}\right)$，获得足够的样本来学习每个条件分布（再加上高概率学习和并集界所需的一些对数因子），从而得到样本复杂度 $\tilde{O}\left(\frac{2^d n^2}{\varepsilon^3}\right)$；在 $1 - O(\varepsilon^2)$ 的情况下，样本复杂度为 $\tilde{O}\left(\frac{2^d n^2}{\varepsilon^4}\right)$。正如我们稍后将看到的，我们可以稍作改进，从而收紧对 ε 的依赖。

我们将把证明 延后 引理 4.1 至 第 4.4 节 和 第 4.5 节，并在此处给出证明概要。

[4]虽然 1 $+ d_{\chi^2}(P, Q)$ 应为 $2P(S) - Q(S) + d_{\chi^2}(P_S, Q_S)$；如果 $P(S) \geqslant 1 - O(\varepsilon^2$，则不会影响分析).

算法 2：在 χ^2 中进行 ε^2-近似适当学习输入：对分布 P 的样本访问，精度参数 ε。1/* 获取多数支持集 $\mathcal{A}$。 */2 $\mathcal{A} \leftarrow$ 使用 P 和 ε 调用算法 1。 3 从 P 中抽取一个包含 m 个样本的多重集 S，其中 $m = O(2^d n^2 \log(2^d n)/\varepsilon^2)$。4 $K \leftarrow \Theta(\log(2^{d+1} \cdot n))$ 5 令 N_{x_i,π_i} 为 m 个样本中 $X_i = x_i, \Pi_i = \pi_i$ 的计数。 6 对于 $i = 1;\quad i \leqslant n;\quad i = i+1$，执行 7 对于 $x_i, \pi_i \in \{0,1\}^{|\Pi_i|+1}$，执行8 $Q(x_i|\pi_i) \leftarrow \frac{K+N_{x_i,\pi_i}}{\sum_{x_i' \in \{0,1\}} K+N_{x_i',\pi_i}}$9 结束 10 结束 11 /* 在（4.8）中的质量平移步骤对于后续的海林格检验是必要的。 */12 $Q(x_1, \ldots, x_n) \leftarrow \prod_{i=1}^{n} Q(x_i|\pi_i)$

LEMMA 4.1. 当 $m = O\left(\frac{2^d n^2}{c\varepsilon^2} \log\left(2^d n\right)\right)$ 时，由 Q 输出的 算法 2 满足以下递推关系：

$$d_{\chi^2}(P_{X_1,\ldots,X_k}, Q_{X_1,\ldots,X_k}, S_k) \leqslant \left(1 + \frac{1}{n}\right) d_{\chi^2}(P_{X_1,\ldots,X_{k-1}}, Q_{X_1,\ldots,X_{k-1}}, S_{k-1}) + O\left(\frac{\varepsilon^2}{n}\right),$$

由递归可得

$$d_{\chi^2}(P_{X_1,\ldots,X_n}, Q_{X_1,\ldots,X_n}, S_n) \leqslant O(\varepsilon^2).$$

证明。 [概要] 通过对所有父节点质量取下界，我们可以保证该假设的所有条件分布上的 χ^2 误差都有界，为，

$$d_{\chi^2}(P_{X_i|\Pi_i}, Q_{X_i|\Pi_i}) \leqslant \frac{O(1)}{m \cdot P(\Pi_i)}$$

结合切尔诺夫界即可得到；相比之下，朴素方法只是直接取 $P(\Pi_i) \geqslant \frac{\varepsilon^2}{2^d n}$。基于此，并通过对残差项进行一些细致的重新整理，我们能够证明这一递推关系。这样的整理颇为棘手，因为标准的 χ^2 形式会变成 $-2P(S) + Q(S) + \sum_i \frac{P_i^2}{Q_i}$，而不是标准的 $-1 + \sum_i P_i^2/Q_i$ 形式。

□

˜对于在 d_H 中的检验，我们需要做如下调整：从得到的 Q 出发，我们移动质量以得到一个 Q，使得 $\tilde{Q}(\tilde{S}_n) \geqslant 1 - O(\varepsilon^2)$，同时保持 Q 的图结构以及 P˜在 S_n 上的 d_{χ^2} 接近性；尽管这可能导致整个支撑集上的 d_{χ^2} 无界（由于分母中存在 0，其值将为无穷大），但这不会影响我们的后续检验。正如我们稍后在分析中将看到的，这是检验 d_H 中的最大入度时额外的（且似乎必需的）一步；˜但对于 d_{TV}，有趣的是，Q 和 S_n 就足够了。

THEOREM 4.1. 给定 $O\left(2^d n^2 \log(2^d n)/\varepsilon^2\right)$ 个样本，我们可以得到一个关于 d 阶贝叶斯网络 P 的适当假设 Q，以及一个 $\tilde{S}_n \subset \{0,1\}^n$，在其上 $d_{\chi^2}(P, Q, \tilde{S}_n) \leqslant O(\varepsilon^2)$，且 $P(\tilde{S}_n) \geqslant 1 - O(\varepsilon^2)$。

此外，通过一些后处理（无需额外样本），我们可以获得另一个适当假设 $\tilde{Q}$，其保证涵盖上述保证，且：$\tilde{Q}(\tilde{S}_n) = 1 > 1 - O(\varepsilon^2)$。

˜证明。 根据 引理 4.1，以及 $S_n \subseteq S_n$,

$$d_{\chi^2}(P_{X_1,\ldots,X_n}, Q_{X_1,\ldots,X_n}, \tilde{S}_n) \leqslant d_{\chi^2}(P_{X_1,\ldots,X_n}, Q_{X_1,\ldots,X_n}, S_n),$$

即为 $O(\varepsilon^2)$。证明至此结束。 □

4.3 高效估计 $O(\varepsilon^2)$-有效支撑集 我们将 $\tilde{S}_n$在第 4.2 节 中的假设形式化表述为 引理 4.2。

LEMMA 4.2. 存在一个例程（算法 1），至多使用 $O\left(2^{d+1}n\log(2^{d+1}n)/(c\cdot\varepsilon^2)\right)$˜个来自 P 的样本，并返回一个近似值 S_n，满足以下保证：对于所有 $k \in [n]$，且概率至少为 $5/6$,

$$S'_k \subseteq \tilde{S}_k \subseteq S_k,$$

其中 S'_k，S_k 如 方程 (4.3) 和 (4.4)中所定义。

证明。我们通过分析算法 1来证明。该算法获取 $m = \frac{3}{c}\cdot 2^{d+1}n\log(6\cdot 2^{d+1}n)/\varepsilon^2$ 个样本，并检查每个 X_i, Π_i 的比值是否超过 $2c\frac{\varepsilon^2}{2^{d+1}n}$– 对于 $x \in \{0,1\}^n$，如果所有 $i \in [n], (x_i, \pi_i) := (x_i(x)\ \pi_i(x)) \in \{0,1\}^{|\Pi_i|+1}$，$N_{x_i,\ \pi_i} \geqslant 2c\frac{\varepsilon^2\cdot m}{2^{d+1}n}$那么将 x 添加到 $\tilde{S}_n$ 中，其中 $N_{x_i,\ \pi_i}$是 $X_i = x_i, \Pi_i = \pi_i$ 在这 m 个样本中出现的次数。

这里的论证旨在确保质量小于 $c\frac{\varepsilon^2}{2^{d+1}n}$ 的 $a \in \{0,1\}^{|\Pi_i|+1}$ 无法通过该过程，并且对于每个 $i \in [n]$, 至多有 $O(\varepsilon^2/n)$ 的质量被剔除。我们通过切尔诺夫界、随机占优论证和并集界来实现这一点。首先，考虑伯努利分布 $T \sim \mathsf{Bern}(m,p)$，其中 $p = c\frac{\varepsilon^2}{2^{d+1}n}$。根据 Chernoff 不等式,

$$\Pr\left[T \geqslant 2\cdot c\frac{\varepsilon^2}{2^{d+1}n}\right] = \Pr[T \geqslant 2mp] \leqslant \exp(-mp/3) = \exp\left(-mc\frac{\varepsilon^2}{2^{d+1}n}/3\right) = \frac{1}{6\cdot 2^{d+1}n}.$$

由于当 $p' \leqslant p$ 时，任意 $T' \sim \mathsf{Bern}(m,p')$ 都被 $T \sim \mathsf{Bern}(m,p)$ 一阶随机占优，因此对于 $p' < \frac{c}{2}\frac{\varepsilon^2}{2^{d+1}n}$，$\Pr\left[T' \geqslant 2c\frac{\varepsilon^2}{2^{d+1}n}\right] \leqslant \Pr\left[T \geqslant 2c\frac{\varepsilon^2}{2^{d+1}n}\right] \leqslant \frac{1}{2^{d+1}n}$。
因此，对于 $p \leqslant c\frac{\varepsilon^2}{2^{d+1}n}, T \sim \mathsf{Bern}(m,p)$，$\Pr\left[T \geqslant 2c\frac{\varepsilon^2}{2^{d+1}n}\right] \leqslant \frac{1}{2^{d+1}n}$。

通过类似的论证，所有 $p \geqslant 4c\frac{\varepsilon^2}{2^{d+1}n}$都将以高概率通过测试：令 $T'' \sim \mathsf{Bern}(m,p'')$，其中 $p'' \geqslant 4c\frac{\varepsilon^2}{2^{d+1}n}$，由 Chernoff 可得,

$$\Pr\left[T'' \leqslant \frac{1}{2}\cdot 4c\frac{\varepsilon^2}{2^{d+1}n}\right] = \Pr\left[T'' \leqslant \frac{1}{2}mp''\right] \leqslant \exp(-mp''/8) \leqslant \exp\left(-\frac{1}{2}mc\frac{\varepsilon^2}{2^{d+1}n}\right) \leqslant \frac{1}{6\cdot 2^{d+1}n}.$$

最后，对支撑集 $\mathsf{Support}(X_i, \Pi_i)$ 中的所有元素 $a \in \{0,1\}^{\Pi_i+1}$ = 应用并集界可知，（以高概率）所有 $i \in [n]$ 和 $a \in \{0,1\}^{\Pi_i+1}$: $P_{X_i,\ \Pi_i}(a) \geqslant 4c\frac{\varepsilon^2}{2^{d+1}n}$ 将通过测试；$P_{X_i,\ \Pi_i}(a) \leqslant c\frac{\varepsilon^2}{2^{d+1}n}$ 将无法通过测试。这说明，对于 $k \in [n]$, $\left\{x \in \{0,1\}^k | \forall i \in [k], P_{X_i,\ \Pi_i}(x_i, \pi_i(x)) \geqslant 4c\frac{\varepsilon^2}{2^{d+1}n}\right\} \subset \tilde{S}_n \subset \left\{x \in \{0,1\}^k | \forall i \in [k], P_{X_i,\ \Pi_i}(x_i, \pi_i(x)) \geqslant c\frac{\varepsilon^2}{2^{d+1}n}\right\}$。 □

4.4 引理 4.1 的一些有用高概率事件 引理 4.1 在本小节中，我们分析若干关键的高概率事件，其形式为 公式 (4.7) 和 (4.9)，以便开展主要证明。但首先，我们需要建立一些技术工具。我们将使用术语 构型 来指代

某个二进制字符串集合，它是 $\{0,1\}^n$ 的子集。首先，我们定义质量“足够大”的父节点的配置集合，

$$A_{S_k} := \left\{ a \in \{0,1\}^{|\Pi_k|} \mid P_{\Pi_k}(a) \geqslant c \cdot \frac{\varepsilon^2}{2^d n} \right\}.$$

我们定义另一个密切相关的配置集合：令 $C(a,\ S_k)$ 表示在 S_k，[5] 给定 $\Pi_k = a$ 且与 Π_k 和最后一个节点 X_k无关时，剩余的配置集合（不包括当前父节点 Π_k 和最后一个节点 X_k）；并且该集合不取决于 X_k 所取的值。形式化地说：

$$\to C(a, S_k) = \{x \in \{0,1\}^{k-1-|\Pi_k|} \mid \exists x' \in \{0,1\}\{(X_1, \ldots, X_{k-1} \setminus \Pi_k) = x, \Pi_k = a, X_k = x'\} \in S_k\}$$

或者等价地，$C(a, S_k) = \{x \in \{0,1\}^{k-1-|\Pi_k|} \mid \{(X_1, \ldots, X_{k-1} \setminus \Pi_k) = x, \Pi_k = a\} \in S_{k-1}\}$——将 Π_k 固定为 a，从 $X_1, \ldots, X_k$ 个变量中选取，并检查 $x, a \in S_{k-1}$ 是否成立。同样，

$$\to B(a, S_k) = \{x \in \{0,1\} \mid \forall x' \in \{0,1\}^{k-1-|\Pi_k|}\{(X_1, \ldots, X_{k-1} \setminus \Pi_k) = x', \Pi_k = a, X_k = x\} \in S_k\}$$

或者我们也可以等价地定义 $B(a, S_k) = \left\{ x \in \{0,1\} \mid P(\Pi_k = a, X_k = x) \geqslant c \cdot \frac{\varepsilon^2}{2^{d+1} n} \right\}$.

其他配置集合包括：$C_k := \{0,1\}^{k-1-|\Pi_k|}$，不受任何限制时 $\{X_1, \ldots, X_{k-1} \setminus \Pi_k\}$ 的全部配置集合；正如 $B_k := \{0,1\}$ 与 X_k 相关；$A_k := \{0,1\}^{|\Pi_k|}$；$A_{S_k^c} := A_k \setminus A_{S_k}$；$B(a, S_k^c) := B_k \setminus B(a, S_k)$；$C(a, S_k^c) := C_k \setminus C(a, S_k)$。注意，$x$ 在 (x_k, π_k) 的情况下，$\in A_{S_k}$ 并不一定意味着 $x \in S_k$（其逆命题成立）；而且已知 $\Pi_k = a$ 时，为了满足 $x \in S_k$，必须同时满足两个约束——一个来自 $\{X_1, \ldots, X_{k-1} \setminus \Pi_k\}$，另一个来自 $\{X_k\}$。

令 $m_{\Pi_k = a}$ 为计数满足 $\Pi_k = a$ 的样本数量的随机变量。对于任意 $x \in S_n$，每个 $k \in [n]$，以及 $a \in \{0,1\}^{|\Pi_k|}$ 满足 $P(\Pi_k = a) = \sum_{x \in X_k} P(x,\ \Pi_k = a) \geqslant 2 \cdot c \frac{\varepsilon^2}{2^{d+1} n} = c \frac{\varepsilon^2}{2^d n}$，因此根据 Chernoff 不等式，$m_{\Pi_k = a} \geqslant \frac{1}{2} m P(\Pi_k = a)$ 以极高概率成立。我们以该事件为条件，并结合命题 4.1 中的学习结果，令 命题 4.1，通过设置 $K = \log(6 \cdot 2^{d+1} n)$，可以推出：对于所有 $a \in A_{S_k}$，

$$\Pr\left[-1 + \sum_{b \in B} \frac{P^2_{X_k|\Pi_k}(b|a)}{Q_{X_k|\Pi_k}(b|a)} \geqslant \frac{c' \log(2^{d+1} n)}{m_{\Pi_{k=a}}} \right] \leqslant \frac{1}{6} \frac{1}{2^{d+1} n}, \tag{4.5}$$

其中 c' 为某个常数。

因此，以高概率，对所有 $a \in A_{S_k}$，

$$-1 + \sum_{b \in B} \frac{P^2_{X_k|\Pi_k}(b|a)}{Q_{X_k|\Pi_k}(b|a)} \leqslant \frac{c' \log(2^{d+1} n)}{m_{\Pi_{k=a}}} \leqslant \frac{2c' \cdot \log(2^{d+1} n)}{m P(\Pi_k = a)}. \tag{4.6}$$

接着，我们再次以该事件发生为条件；由于

$$\sum_{b \in B(a, S_k)} \frac{P^2_{X_k|\Pi_k}(b|a)}{Q_{X_k|\Pi_k}(b|a)} \leqslant \sum_{b \in B} \frac{P^2_{X_k|\Pi_k}(b|a)}{Q_{X_k|\Pi_k}(b|a)},$$

从而得到，

$$-1 + \sum_{b \in B(a, S_k)} \frac{P^2_{X_k|\Pi_k}(b|a)}{Q_{X_k|\Pi_k}(b|a)} \leqslant \frac{2c' \cdot \log(2^{d+1} n)}{m P(\Pi_k = a)}. \tag{4.7}$$

[5] 这个术语用于（分别）分析在该马尔可夫链中以父节点 Π_k 为条件所生成的两个项：$\{X_1, \ldots, X_{k-1}\} \setminus \Pi_k \to \Pi_k \to X_k$。

在 引理 4.1 的证明后半部分，我们需要对密度估计施加一个更强的条件（通过移动至多 Q $O(\varepsilon^2)$ 的质量，得到一个略有不同的估计）——我们将其称为 $\tilde{Q}$，并且可以˜直接由 Q 和 S_n 得到：

$$\tilde{Q}_{X_k|\Pi_k}(b|a) := \begin{cases} \frac{Q_{X_k|\Pi_k}(b|a)}{\sum_{b\in B(a,\tilde{S}_k)} Q_{X_k|\Pi_k}(b|a)} & \text{if } b \in B(a,\tilde{S}_k), \\ 0 & \text{otherwise.} \end{cases} \tag{4.8}$$

˜ ˜此外， Q在 S_n与 Q在 S_n上具有非常相似的保证。对于每个$a \in A_{\tilde{S}_k} \subseteq A_{S_k}$，由于˜$B(a, S_k) \subseteq B(a, S_k)$（以公式 4.6为条件），

$$\sum_{b\in B(a,\tilde{S}_k)} \frac{P^2_{X_k|\Pi_k}(b|a)}{Q_{X_k|\Pi_k}(b|a)} \leqslant \sum_{b\in B(a,S_k)} \frac{P^2_{X_k|\Pi_k}(b|a)}{Q_{X_k|\Pi_k}(b|a)} \leqslant 1 + \frac{2c' \cdot \log(2^{d+1}n)}{mP(\Pi_k = a)}.$$

由于将质量从 $B(a, \tilde{S}_k^c)$ 移至 $B(a, \tilde{S}_k)$ 只会进一步减小量 $P^2_{X_k|\Pi_k}(b|a)\tilde{Q}_{X_k|\Pi_k}(b|a)$，因此我们有

$$-1 + \sum_{b\in B(a,\tilde{S}_k)} \frac{P^2_{X_k|\Pi_k}(b|a)}{\tilde{Q}_{X_k|\Pi_k}(b|a)} \leqslant \frac{2c' \cdot \log(2^{d+1}n)}{mP(\Pi_k = a)}. \tag{4.9}$$

最后，根据并集界，以至少 5/6的概率，对所有 $k \in [n]$ 和 $\Pi_k = a$， 公式 (4.7) 和(4.9) 中的两个陈述都成立。在附录余下部分中，我们将始终以这两个陈述为条件。

4.5 引理 4.1 的证明 我们将子集 $S_k \subset \{0,1\}^k$ 上 χ^2 在 $P_{X_1,\ldots,X_k}$ 与 $Q_{X_1,\ldots,X_k}$ 之间的部分和写作：

$$\begin{aligned} d_{\chi^2}(P_{X_1,\ldots,X_k}, Q_{X_1,\ldots,X_k}, S_k) &= \sum_{x\in S_k} \frac{(P_{X_1,\ldots,X_k}(x) - Q_{X_1,\ldots,X_k}(x))^2}{Q_{X_1,\ldots,X_k}(x)} \\ &= \sum_{x\in S_k} -2P_{X_1,\ldots,X_k}(x) + Q_{X_1,\ldots,X_k}(x) + \frac{P_{X_1,\ldots,X_k}(x)^2}{Q_{X_1,\ldots,X_k}(x)} \\ &= -2P_{X_1,\ldots,X_k}(S_k) + Q_{X_1,\ldots,X_k}(S_k) + \sum_{x\in S_k} \frac{P_{X_1,\ldots,X_k}(x)^2}{Q_{X_1,\ldots,X_k}(x)}, \end{aligned}$$

根据.....的定义 S_k，我们可以将对 $x \in S_k$ 的求和写成 $a \in A_{S_k}, b \in B(a, S_k)$ 和 $g \in C(a, S_k)$ ：

$$\begin{aligned} & d_{\chi^2}(P_{X_1,\ldots,X_k}, Q_{X_1,\ldots,X_k}, S_k) \\ = & \left\{ \sum_{a\in A_{S_k}} \frac{P^2_{\Pi_k}(a)}{Q_{\Pi_k}(a)} \cdot \sum_{g\in C(a,S_k)} \frac{P^2_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g \mid a)}{Q_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g \mid a)} \cdot \sum_{b\in B(a,S_k)} \frac{P^2_{X_k|\Pi_k}(b \mid a)}{Q_{X_k|\Pi_k}(b \mid a)} \right\} \\ & -(2P_{X_1,\ldots,X_k}(S_k) - Q_{X_1,\ldots,X_k}(S_k)) \\ = & \left\{ \sum_{a\in A_{S_k}} \frac{P^2_{\Pi_k=a}}{Q_{\Pi_k=a}} \cdot \sum_{g\in C(a,S_k)} \frac{P^2_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g \mid a)}{Q_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g \mid a)} \cdot \left(-1 + \sum_{b\in B(a,S_k)} \frac{P^2_{X_k|\Pi_k}(b \mid a)}{Q_{X_k|\Pi_k}(b \mid a)} \right) \right\} \\ & - \left(2P_{X_1,\ldots,X_{k-1}}(S_{k-1}) - Q_{X_1,\ldots,X_{k-1}}(S_{k-1})\right) \end{aligned}$$

$$
\begin{aligned}
&+\sum_{a\in A_{S_k}}\sum_{g\in C(a,S_k)}\frac{P^2_{\Pi_k=a}}{Q_{\Pi_k=a}}\cdot\frac{P^2_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g\mid a)}{Q_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g\mid a)}\\
&+(2P_{X_1,\ldots,X_{k-1}}(S_{k-1})-Q_{X_1,\ldots,X_{k-1}}(S_{k-1}))\\
&-(2P_{X_1,\ldots,X_k}(S_k)-Q_{X_1,\ldots,X_k}(S_k)) \qquad (4.10)\\
=\;&\left\{\sum_{a\in A_{S_k}}\frac{P^2_{\Pi_k}(a)}{Q_{\Pi_k}(a)}\sum_{g\in C(a,S_k)}\frac{P^2_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g\mid a)}{Q_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g\mid a)}\cdot\left(-1+\sum_{b\in B(a,S_k)}\frac{P^2_{X_k|\Pi_k}(b\mid a)}{Q_{X_k|\Pi_k}(b\mid a)}\right)\right\}\\
&+d_{\chi^2}(P_{X_1,\ldots,X_{k-1}},Q_{X_1,\ldots,X_{k-1}},S_{k-1})\\
&+(2P_{X_1,\ldots,X_{k-1}}(S_{k-1})-Q_{X_1,\ldots,X_{k-1}}(S_{k-1}))\\
&-(2P_{X_1,\ldots,X_k}(S_k)-Q_{X_1,\ldots,X_k}(S_k)) \qquad (4.11)\\
\leqslant\;&2c\frac{\varepsilon^2}{n}+4c'\frac{2^d\log(2^dn)}{m}+\left(1+\frac{2c'2^dn\log(2^dn)}{cm\varepsilon^2}\right)\times d_{\chi^2}(P_{X_1,\ldots,X_{k-1}},Q_{X_1,\ldots,X_{k-1}},S_{k-1}). \qquad (4.12)
\end{aligned}
$$

通过加上并减去相同的项，我们得到（4.10），从而得到关于 S_{k-1} 的递推关系（4.11）。由此，利用两个技术性结果，我们得到（4.12）：引理 4.3 和4.4。最后，令 $m=\frac{4c2^dn^2}{\log(2^dn)\varepsilon^2}$，即可得到 $d_{\chi^2}(P_{X_1,\ldots,X_k},Q_{X_1,\ldots,X_k},S_k)$ 的所需递归表述，从而完成 引理 4.1 的证明。

现在我们证明上述技术引理。

Lemma 4.3.

$$
(2P_{X_1,\ldots,X_{k-1}}(S_{k-1})-Q_{X_1,\ldots,X_{k-1}}(S_{k-1}))-(2P_{X_1,\ldots,X_k}(S_k)-Q_{X_1,\ldots,X_k}(S_k))\leqslant 2c\frac{\varepsilon^2}{n}.
$$

证明。

$$
\begin{aligned}
&(2P_{X_1,\ldots,X_{k-1}}(S_{k-1})-Q_{X_1,\ldots,X_{k-1}}(S_{k-1}))-(2P_{X_1,\ldots,X_k}(S_k)-Q_{X_1,\ldots,X_k}(S_k))\\
=\;&\sum_{x_1,\ldots,x_{k-1}\in S_{k-1}}\sum_{x_k\in\{0,1\}}2P_{X_1,\ldots,X_k}(x_1,\ldots,x_k)-\sum_{x_1,\ldots,x_k\in S_k}2P_{X_1,\ldots,X_k}(x_1,\ldots,x_k)\\
&-(Q_{X_1,\ldots,X_{k-1}}(S_{k-1})-Q_{X_1,\ldots,X_k}(S_k))\\
=\;&\sum_{x_1,\ldots,x_{k-1}\in S_{k-1},x_k\in S^c(x_1,\ldots,x_{k-1})}(2P(x_1,\ldots,x_k)-Q(x_1,\ldots,x_k))\\
\leqslant\;&\sum_{x_1,\ldots,x_{k-1}\in\{0,1\}^{k-1}}\sum_{x_k\in S^c(x_1,\ldots,x_{k-1})}2P(x_1,\ldots,x_k)\\
=\;&2\sum_{\{x_1,\ldots,x_{k-1}\setminus\pi_k\}\in\{0,1\}^{k-1-|\Pi_k|}}\sum_{\pi_k}\sum_{x_k\in S^c(\pi_k)}P(x_1,\ldots,x_{k-1}\setminus\pi_k|\pi_k)P(x_k,\pi_k)\\
\leqslant\;&2\sum_{x_k,\pi_k\in S^c(X_k,\Pi_k)}P_{X_k,\Pi_k}(x_k,\pi_k)\leqslant 2c\frac{\varepsilon^2}{n}.
\end{aligned}
$$

□

Lemma 4.4.

$$
\sum_{\substack{a\in A_{S_k}\\ g\in C(a,S_k)}}\frac{P^2_{\Pi_k=a}}{Q_{\Pi_k=a}}\frac{P^2_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g|a)}{Q_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g|a)}\left(-1+\sum_{b\in B(a,S_k)}\frac{P^2_{X_k|\Pi_k}(b|a)}{Q_{X_k|\Pi_k}(b|a)}\right)
$$

$$\leqslant \frac{2c'\cdot 2^d n\log(2^d n)}{cm\varepsilon^2} d_{\chi^2}(P_{X_1,\ldots,X_{k-1}}, Q_{X_1,\ldots,X_{k-1}}, S_{k-1}) + \frac{4c\cdot 2^d\log(2^d n)}{m}.$$

证明。

$$\sum_{\substack{a\in A_{S_k}\\ g\in C(a,S_k)}} \frac{P^2_{\Pi_k=a}}{Q_{\Pi_k=a}} \frac{P^2_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g|a)}{Q_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g|a)} \left(-1+\sum_{b\in B(a,S_k)} \frac{P^2_{X_k|\Pi_k}(b|a)}{Q_{X_k|\Pi_k}(b|a)}\right) \tag{4.13}$$

$$\leqslant \sum_{\substack{a\in A_S\\ g\in C(a,S_k)}} \frac{P^2_{\Pi_k=a}}{Q_{\Pi_k=a}} \frac{P^2_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g|a)}{Q_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g|a)} \frac{2c'\cdot\log(2^d n)}{mP_{\Pi_k}(a)}$$

$$= \sum_{\substack{a\in A_S\\ g\in C(a,S_k)}} \left(-2P_{X_1,\ldots,X_{k-1}}(a,g) + Q_{X_1,\ldots,X_{k-1}}(a,g) + \frac{P^2_{X_1,\ldots,X_{k-1}}(a,g)}{Q_{X_1,\ldots,X_{k-1}}(a,g)}\right)\cdot\frac{2c'\cdot\log(2^d n)}{mP_{\Pi_k}(a)}$$

$$+ \sum_{\substack{a\in A_S\\ g\in C(a,S_k)}} (2P_{X_1,\ldots,X_{k-1}}(a,g) - Q_{X_1,\ldots,X_{k-1}}(a,g))\cdot\frac{2c'\cdot\log(2^d n)}{mP_{\Pi_k}(a)}$$

$$= \sum_{\substack{a\in A_S\\ g\in C(a,S_k)}} \frac{(P_{X_1,\ldots,X_{k-1}}(a,g) - Q_{X_1,\ldots,X_{k-1}}(a,g))^2}{Q_{X_1,\ldots,X_{k-1}}(a,g)}\cdot\frac{2c'\cdot\log(2^d n)}{mP_{\Pi_k}(a)}$$

$$+ \sum_{a\in A_S}\sum_{g\in C(a,S_k)} (2P_{X_1,\ldots,X_{k-1}}(a,g) - Q_{X_1,\ldots,X_{k-1}}(a,g))\cdot\frac{2c'\cdot\log(2^d n)}{mP_{\Pi_k}(a)} \tag{4.14}$$

$$\leqslant \frac{2c'2^d n\log(2^d n)}{cm\varepsilon^2}\cdot \sum_{\substack{a\in A_S\\ g\in C(a,S_k)}} \frac{(P_{X_1,\ldots,X_{k-1}}(a,g) - Q_{X_1,\ldots,X_{k-1}}(a,g))^2}{Q_{X_1,\ldots,X_{k-1}}(a,g)}$$

$$+ \sum_{a\in A_S}\sum_{g\in C(a,S_k)} P_{X_1,\ldots,X_{k-1}}(a,g)\cdot\frac{4c'\cdot\log(2^d n)}{mP_{\Pi_k}(a)}$$

$$= \frac{2c'2^d n\log(2^d n)}{cm\varepsilon^2}(P_{X_1,\ldots,X_{k-1}}, Q_{X_1,\ldots,X_{k-1}}, S_{k-1})$$

$$+ \sum_{a\in A_S} P_{\Pi_k}(a)\cdot\frac{4c'\cdot\log(2^d n)}{mP_{\Pi_k}(a)} \underbrace{\sum_{g\in C(a,S_k)} P_{X_1,\ldots,X_{k-1}\setminus\Pi_k|\Pi_k}(g|a))}_{\leqslant 1}$$

$$\leqslant \frac{2c'2^d n\log(2^d n)}{cm\varepsilon^2} d_{\chi^2}(P_{X_1,\ldots,X_{k-1}}, Q_{X_1,\ldots,X_{k-1}}, S_k) + \frac{4c'\cdot 2^d\log(2^d n)}{m}. \tag{4.15}$$

□

4.6 在 χ^2 中学习贝叶斯网络的下界我们的下界依赖于一族一阶贝叶斯网络，其中所有 $n-1$ 节点共享同一个单节点父节点。我们将把父节点的概率设得极不均衡，以至于即使取 $\exp(O(n))$ 个样本，我们仍然无法从概率较低的一侧获得一个样本。在这种情况下，我们不可能观测到 $n-1$ 条件分布某一侧的任何样本，因此，从信息论角度而言，很难得到这些条件密度的良好估计值。由于这些“隐藏”的条件密度中的误差会以乘法方式累积，

尽管父节点概率很小， χ^2 距离仍然很大。需要注意的是，这对 KL 散度而言并不是问题：以 KL 散度表示的误差是线性累积的，而 χ^2 则是以乘法方式（因而呈指数级）累积。

DEFINITION 4.1. 我们定义一个生成困难实例的过程，以便在命题 4.2中进行分析：

1. 令先验分布 $P \sim \pi$ 为满足 $P = P_{X_1} \cdot P_{X_2|X_1} \cdots P_{X_n|X_1}$ 且 $P(X_1 = 1) = \varepsilon_0$ 的概率分布，其中参数由我们选择。
2. 从 $\{0,1\}^{n-1}$ 中均匀随机抽取 x^h 。
3. 根据 x^h，设定 $P_{X_i|X_1=1} = \delta_{x_i^h}$，因此 $P_{X_2,\ldots,X_n|X_1=1} = \prod_{i=2}^n P_{X_i|X_1=1} = \delta_{x^h}$，其中

$$\delta_y(x) = \begin{cases} 1, & x = y \\ 0, & otherwise \end{cases}.$$

4. 在另一种情况下，令 $P_{X_2,\ldots,X_n|X_1=0} = U_{n-1}$ 在 $\{0,1\}^{n-1}$ 上服从均匀分布。

直观地说，我们只是在向学习者隐藏一个特定点 x^h ；当以 $X_1 = 1$ 为条件时， π 中的每个分布都会将其概率质量集中在单纯形中的一个点上（即确定性的）。虽然只需一个样本（带有 $x_1 = 1$）即可学会，但少于 $O(1/\varepsilon_0)$ 个样本时，任何学习者都无法做到。我们将在下面陈述主要结果。

PROPOSITION 4.2. 估计一个度数为1 贝叶斯网络 P 在 χ^2 中的极小极大风险至少为 $\Omega(\varepsilon)$ 当样本数 $m \leqslant O(2^{n/2}/\varepsilon)$时。特别地，定义 4.1 中的概率分布族至少需要 $\Omega(2^{n/2}/\varepsilon)$ 个样本才能学到 ε。

受 [KOPS15 引理 5], 中的证明启发，通过标准的拉格朗日乘子计算，我们得到以下有用事实。我们将用它们来证明 命题 4.2。

FACT 4.1. 设 $q_i, a_i \geqslant 0, i \in [k]$，满足 $\sum_{i=1}^k q_i \leqslant 1$。那么，当$q_i \propto \sqrt{a_i}$时，量 $\sum_{i=1}^k \frac{a_i}{q_i}$ 取得最小值。

FACT 4.2. 当 $Q_i^2 = \frac{1}{2^{n-1}}, i = 1, \cdots, n-1$ 时，可得到如下形式的最优解：

$$\min_{Q \in \mathbb{R}^{n-1}} \sum_{i=1}^{2^{n-1}} \frac{1}{Q_i} \cdot \frac{1}{2^{n-1}}, s.t. \sum_{i=1}^{2^{n-1}} Q_i^2 = 1.$$

证明。 为此，我们只需验证 KKT 条件 [BV14] 是否满足该约束优化问题：

$$\sum_{i=1}^{2^{n-1}} \frac{1}{Q_i} \cdot \frac{1}{2^{n-1}} + \lambda \left(\sum_{i=1}^{2^{n-1}} Q_i^2 - 1 \right).$$

必要条件：

$$-\frac{1}{Q_i^2} \cdot \frac{1}{2^{n-1}} + 2\lambda Q_i = 0, \forall i; \quad \sum_{i=1}^{2^{n-1}} Q_i^2 - 1 = 0.$$

解第一个方程得出 $Q_i = \frac{1}{2^{n/3}\lambda^{1/3}}$；由于所有 Q_is 都相等， Q^* 应当是均匀的。由于 $\frac{1}{Q_i}$ 和 Q_i^2 都是凸的，因此必要条件也是取得全局最小值的充分条件。

□

证明。 [命题 4 的证明。2] 设 $s_i = (X_{1,\ i},\ \ldots,\ X_{n,\ i})$ 为样本，记事件为 $S = \{X_{1,i} = 0, i \in [m]\}$，因此概率为 $\Pr[S] = (1-\varepsilon_0)^m \geqslant e^{-\frac{1}{2}\varepsilon_0 m} \geqslant 1 - \frac{1}{2}\varepsilon_0 m$。在计算过程中，我们将以 S 为真作为条件。令 $\mathcal{A}_m$ 表示从 P 中取 m 个样本的确定性算法集合。那么， $R_{\chi^2}(m)$ 是 $\mathcal{A}_m$ 上的极小极大风险，

$$\begin{aligned}
R_{\chi^2}(m) &= \inf_{Q\in\mathcal{A}_m} \sup_{P\in\mathcal{P}} \mathbb{E}_{s_1,\ldots,s_m\sim P}\left[d_{\chi^2}\left(P, Q_{\boldsymbol{s}}\right)\right] \\
&\geqslant \inf_{Q\in\mathcal{A}_m} \mathbb{E}_{P\sim\pi}\mathbb{E}_{\boldsymbol{s}\sim P^{\otimes m}}\left[d_{\chi^2}\left(P, Q_{\boldsymbol{s}}\right)\right] \\
&\geqslant \inf_{Q\in\mathcal{A}_m} \mathbb{E}_{P\sim\pi}\mathbb{E}_{\boldsymbol{s}\sim P^{\otimes m}}\left[d_{\chi^2}\left(P, Q_{\boldsymbol{s}}\right)|S\right]\Pr[S] \\
&= \inf_{Q\in\mathcal{A}_m} \mathop{\mathbb{E}}_{P\sim\pi}\mathop{\mathbb{E}}_{\boldsymbol{s}\sim P^{\otimes m}}\left[-1 + \sum_{x_1,\ldots,x_n} \frac{P^2(x_1,\ldots,x_n)}{Q_{\boldsymbol{s}}(x_1,\ldots,x_n)} \mid S\right]\Pr[S]. \qquad (4.16)
\end{aligned}$$

为方便起见，$P_0 = P_{X_1,\ldots,X_{n-1}|X_n=0}$；$P_1 = P_{X_1,\ldots,X_{n-1}|X_n=1}$；$Q_{0,\boldsymbol{s}} = Q_{\boldsymbol{s}}(X_1,\ldots, X_{n-1}|X_n = 0)$；$Q_{1,\boldsymbol{s}} = Q_{\boldsymbol{s}}(X_1,\ldots,X_{n-1}|X_n = 1)$。我们分别对内层求和进行分析并求其下界；在此之前，我们需要一个单独的辅助工具——利用上面的事实 4.1，可以证明，对于任意固定的 P 和 $Q_{\boldsymbol{s}}$，

$$\begin{aligned}
&\frac{P^2(x_1=0)}{Q_{\boldsymbol{s}}(x_1=0)} + \frac{P^2(x_1=1)}{Q_{\boldsymbol{s}}(x_1=1)}\left(1 + d_{\chi^2}\left(P_1, Q_{1,\boldsymbol{s}}\right)\right) \\
\geqslant &\frac{P^2(x_1=0)}{\frac{P(x_1=0)}{P(x_1=0)+P(x_1=1)\sqrt{1+d_{\chi^2}(P_1,Q_1)}}} + \frac{P^2(x_1=1)(1+d_{\chi^2}(P_1,Q_1))}{\frac{P(x_1=1)\sqrt{1+d_{\chi^2}(P_1,Q_1)}}{P(x_1=0)+P(x_1=1)\sqrt{1+d_{\chi^2}(P_1,Q_1)}}} \\
\geqslant &\left(P(x_1=0) + P(x_1=1)\sqrt{1+d_{\chi^2}(P_1,Q_1)}\right)^2 \qquad (4.17)
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow &-1 + \sum_{x_1,\ldots,x_n} \frac{P^2(x_1,\ldots,x_{n-1},x_n)}{Q_{\boldsymbol{s}}(x_1,\ldots,x_{n-1},x_n)} \\
= &-1 + \sum_{x_1,\ldots,x_{n-1}} \frac{P^2(x_n=0)P^2(x_1,\ldots,x_{n-1}|x_n=0)}{Q_{\boldsymbol{s}}(x_n=0)Q_{\boldsymbol{s}}(x_1,\ldots,x_{n-1}|x_n=0)} \\
&+ \sum_{x_1,\ldots,x_{n-1}} \frac{P^2(x_n=1)P^2(x_1,\ldots,x_{n-1}|x_n=1)}{Q_{\boldsymbol{s}}(x_n=1)Q_{\boldsymbol{s}}(x_1,\ldots,x_{n-1}|x_n=1)} \\
= &-1 + \frac{P^2(x_n=0)}{Q_{\boldsymbol{s}}(x_n=0)}\left(1 + d_{\chi^2}\left(P_0, Q_{0,\boldsymbol{s}}\right)\right) + \frac{P^2(x_n=1)}{Q_{\boldsymbol{s}}(x_n=1)}\left(1 + d_{\chi^2}\left(P_1, Q_{1,\boldsymbol{s}}\right)\right) \\
\geqslant &-1 + \frac{P^2(x_n=0)}{Q_{\boldsymbol{s}}(x_n=0)} + \frac{P^2(x_n=1)}{Q_{\boldsymbol{s}}(x_n=1)}\left(1 + d_{\chi^2}\left(P_1, Q_{1,\boldsymbol{s}}\right)\right) \\
\geqslant &-1 + \left(P(x_n=0) + P(x_n=1)\sqrt{1+d_{\chi^2}\left(P_1,Q_{1,\boldsymbol{s}}\right)}\right)^2 \qquad (4.18) \\
= &-1 + \left((1-\varepsilon_0) + \varepsilon_0\sqrt{1+d_{\chi^2}\left(P_1,Q_{1,\boldsymbol{s}}\right)}\right)^2 \\
= &-1 + \left(1 + \varepsilon_0\left(\sqrt{1+d_{\chi^2}\left(P_1,Q_{1,\boldsymbol{s}}\right)} - 1\right)\right)^2 \\
= &\ 2\varepsilon_0\left(\sqrt{1+d_{\chi^2}\left(P_1,Q_{1,\boldsymbol{s}}\right)} - 1\right) + \left(\varepsilon_0\left(\sqrt{1+d_{\chi^2}\left(P_1,Q_{1,\boldsymbol{s}}\right)} - 1\right)\right)^2
\end{aligned}$$

$$\geqslant 2\varepsilon_0\left(\sqrt{1+d_{\chi^2}\left(P_1,Q_{1,\boldsymbol{s}}\right)}-1\right). \tag{4.19}$$

对于任意固定的 $Q_{\boldsymbol{s}}(x_1,\ldots,x_n)$，可以计算出 $Q_{\boldsymbol{s}}(x_n=0)$ $Q_{\boldsymbol{s}}(x_n=1)$，以及 $d_{\chi^2}(P_1,\ Q_{1,\ \boldsymbol{s}})$；换言之，给定 $\boldsymbol{s}$ 后，它们也是固定的。随后可通过对（4.18）和（4.17）进行变分论证，得到一个下界。结合（4.16）和（4.19），我们继续得到以下表达式：

$$\begin{aligned}
\frac{R_{\chi^2}(m)}{\Pr[S]} &\geqslant \inf_{Q\in\mathcal{Q}} \mathop{\mathbb{E}}_{\substack{P\sim\pi\\ \boldsymbol{s}\sim P^{\otimes m}}} \left[2\varepsilon_0(\sqrt{1+d_{\chi^2}\left(P_1,Q_{1,\boldsymbol{s}}\right)}-1)|S\right] \\
&= \inf_{Q\in\mathcal{Q}} \mathop{\mathbb{E}}_{P_1\sim\pi} \mathop{\mathbb{E}}_{\boldsymbol{s}\sim P^{\otimes m}|S}\left[2\varepsilon_0\left(\sqrt{1+d_{\chi^2}\left(P_1,Q_{1,\boldsymbol{s}}\right)}-1\right)\right] \\
&= \inf_{Q\in\mathcal{Q}} \mathop{\mathbb{E}}_{x^h\sim U_{n-1}} \mathop{\mathbb{E}}_{\boldsymbol{s}\sim\tilde{U}_{n-1}^{\otimes m}}\left[2\varepsilon_0\left(\sqrt{1+d_{\chi^2}\left(P_1,Q_{1,\boldsymbol{s}}\right)}-1\right)\right] \qquad (4.20)\\
&= \inf_{Q\in\mathcal{Q}} \mathop{\mathbb{E}}_{\boldsymbol{s}\sim\tilde{U}_{n-1}^{\otimes m}} \mathop{\mathbb{E}}_{x^h\sim U_{n-1}}\left[2\varepsilon_0\left(\sqrt{1+d_{\chi^2}\left(P_1,Q_{1,\boldsymbol{s}}\right)}-1\right)\right] \qquad (4.21)\\
&\geqslant \left(\mathop{\mathbb{E}}_{\boldsymbol{s}\sim\tilde{U}_{n-1}^{\otimes m}} \inf_{Q\in\mathcal{Q}} \mathop{\mathbb{E}}_{x^h\sim U_{n-1}}\left[2\varepsilon_0\left(\sqrt{\sum_x \frac{P_1^2(x)}{Q_{1,\boldsymbol{s}}(x)}}-1\right)\right]\right) \qquad (4.22)\\
&\geqslant \left(\inf_{Q^*\in\Delta^{2^{n-1}}} \mathop{\mathbb{E}}_{x^h\sim U_{n-1}}\left[2\varepsilon_0\left(\sqrt{\sum_x \frac{P_1^2(x)}{Q^*(x)}}-1\right)\right]\right) \qquad (4.23)\\
&= \inf_{Q^*\in\Delta^{2^{n-1}}} \mathop{\mathbb{E}}_{\boldsymbol{\alpha}\triangleq(\alpha_1,\ldots,\alpha_{n-1})\sim U_{n-1}}\left[2\varepsilon_0\left(\sqrt{\frac{1}{Q^*(\boldsymbol{\alpha})}}-1\right)\right] \\
&= 2\varepsilon_0\left(2^{\frac{n-1}{2}}-1\right). \qquad (4.24)
\end{aligned}$$

由于我们只在构造中改变 P_1 （或 x^h），因此可以用 $P_1\sim\pi$ 替换 $P\sim\pi$ （或用 $x^h\sim U_{n-1}$ 替换 $P\sim\pi$）；注意，在 $\boldsymbol{s}$ 中不存在满足 $X_1=1$ 的样本，因此 $\boldsymbol{s}$ 只是从均匀分布中抽取的样本及其对应的所有 $X_1=0$，这就是我们在 (4.20) 中所说的 $\boldsymbol{s}\sim\tilde{U}_{n-1}^{\otimes m}$。因此，$P_1$ 或 x^h 与 $\boldsymbol{s}$ 相互独立，我们可以交换 (4.21) 中的期望；在 (4.22) 中，我们将期望下界化，因为学习者可以先观察样本 $\boldsymbol{s}$，再从 $\mathcal{Q}$ 中选择算法。但无论如何，在最后一个期望之前它都是固定的，因此 (4.23) 得证。由于我们假设学习算法 Q 是确定性的，对于固定的 $\boldsymbol{s}$，Q_1 和 $\boldsymbol{s}$ 也都是固定的。由事实 4.2 得到 (4.24)。

最终，我们得到

$$R_{\chi^2}(m)\geqslant 2\varepsilon_0\left(2^{\frac{n-1}{2}}-1\right)\Pr[S]\geqslant \varepsilon_0 2^{\frac{n}{2}}(1-\varepsilon_0 m).$$

通过设置 $\varepsilon_0=\frac{2\varepsilon}{2^{n/2}}$，我们可以看出，如果 $m\leqslant\frac{1}{4\varepsilon}2^{\frac{n}{2}}$，那么 $R_{\chi^2}(m)\geqslant 2\varepsilon-\frac{4\varepsilon^2}{2^{n/2}}m\geqslant\varepsilon$。 □

5 贝叶斯网络最大入度检验

THEOREM 5.1. 给定一个未知分布 P，以及一个支撑在$\{0,1\}^n$上的最大度数为d 的图 G，检验 $d_H(P,G)=0$ 或 $d_H(P,G)\geqslant\varepsilon$，最多需要 $O\left(\max\left(\frac{2^{n/2}}{\varepsilon^2},\frac{2^d n^2 d\log(n)}{\varepsilon^2}\right)\right)$个独立同分布样本，检验的正确率为 $\geqslant 2/3$。

此外，以至少 $2/3$ 的成功概率检验 P 是否相对于任意最大度为 d 的图满足马尔可夫性，最多需要 $O\left(\max\left(\frac{2^{n/2}}{\varepsilon^2},\frac{2^d n^2 d\log(n)}{\varepsilon^2}\right)\right)\log(n^{dn})$ 个样本。

算法 3：检验 P 是否为度数-d 有向无环图 G

输入：对分布 P 的样本访问、精度参数 ε 和一个度数为 d 的有向无环图 G。1 通过算法 2 使用 ε、G 和 $2^d n^2 \log(2^d n)/\varepsilon^2$ 个样本学习 P：获得估计值 $\tilde{Q}$，并通过算法 1 获得一个 $O(\varepsilon^2)$-有效支撑集 $\mathcal{A}$。2 从 P 中抽取多重集 S，其中包含 m 个泊松分布(m)样本，其中 $m = \frac{2^{n/2}}{\varepsilon^2}$。3 调用 [DKW18、算法 1]，并返回 $P_{\mathcal{A}}, Q_{\mathcal{A}}, S, \varepsilon$。

证明。我们通过分析 算法 3 来证明。根据 [DKW18 中底层检验器的保证，算法 1], 只需验证以下内容：

- 可靠性：如果 $d_H(P,G) \geqslant \varepsilon$（它与任何图结构为 G 的贝叶斯网络都相距甚远），那么 $d_H(P_{\mathcal{A}}, \tilde{Q}_{\mathcal{A}}) \geqslant \Omega(\varepsilon)$;
- 正确性：如果。$d_H(P,G) = 0$, then $d_{\chi^2}(P_{\mathcal{A}}, \tilde{Q}_{\mathcal{A}}) \leqslant O(\varepsilon^2)$; and we have this from Theorem 4.1

粗略地说，我们可以假设 P 和 $\tilde{Q}$ 的支撑集仅包含在 $\mathcal{A}$ 中，并且由于 $|\mathcal{A}| \leqslant 2^n$，$O\left(\sqrt{2^n}\right)/\varepsilon^2$个样本足以进行检验。根据 定理 4.1，对于可靠性，我们有

$$\tilde{Q}(\tilde{S}) \geqslant 1 - O(\varepsilon^2), \text{ and } P(\tilde{S}) \geqslant 1 - O(\varepsilon^2).$$

由于

$$d_H^2(P,Q) = d_H^2(P_{\mathcal{A}}, Q_{\mathcal{A}}) + d_H^2(P_{\bar{\mathcal{A}}}, Q_{\bar{\mathcal{A}}}), \text{ and } d_H^2(P_{\bar{\mathcal{A}}}, Q_{\bar{\mathcal{A}}}) \leqslant d_{\mathrm{TV}}(P_{\bar{\mathcal{A}}}, Q_{\bar{\mathcal{A}}}) \leqslant \frac{1}{2}(P(\bar{\mathcal{A}}) + Q(\bar{\mathcal{A}})) \leqslant O(\varepsilon^2),$$

我们有 $d_H^2(P_{\mathcal{A}}, Q_{\mathcal{A}}) \geqslant \Omega(\varepsilon^2)$。

由于将每个测试的成功概率提升至 $1-\delta$ 需要额外花费 $O(\log(1/\delta))$，我们将在所有 $n^{O(dn)}$ 个可能的最大入度-d 图上运行经过放大的精确测试，然后应用 $1 - \delta \cdot n^{O(dn)}$的并集界。具体而言，我们令 $\delta = \frac{1}{n^{dn}}$，这会使总体样本复杂度额外增加一个 $O(\log(n^{dn}))$ 因子，因此，我们得到一个最大入度-d 图测试器，其样本复杂度为 $O\left(\max\left(\frac{2^{n/2}}{\varepsilon^2}, \frac{2^d n^2 d \log(n)}{\varepsilon^2}\right) \cdot \log(n^{dn})\right)$。

5.1 将定理 5.1 扩展至全变差距离 尽管我们的定理 5.1 已经推出了一个 d_{TV}中的测试器，但借助我们针对有界度贝叶斯网络的 d_{χ^2} 近似适当学习器，它还推出了一个类似于定理 5.1 的图模型测试器 d_{TV}，其中无需移动质量（见 [ADK15，备注 1]）；也就是说，在 ˜ ˜全变差距离的情形下，不再需要额外的要求 $Q(\tilde{S}) \geqslant 1 - O(\varepsilon^2)$；此外，我们还可以放宽对 $P(S)$ 的要求：$P(S) \geqslant 1 - O(\varepsilon)$。

为说明这一点，在可靠性情形下，我们只需验证 $d_{\mathrm{TV}}(P_{\mathcal{A}}, Q_{\mathcal{A}}) \geqslant \Omega(\varepsilon)$。假设 $d_{\mathrm{TV}}(P,Q) > 10\varepsilon$ 且 $P(S) > 1 - \varepsilon$，我们分析以下两种情形，

– 当 $Q(S) < 1 - 2\varepsilon$ 时，我们有 $P(S^c) \leqslant \varepsilon, Q(S^c) > 2\varepsilon$，因此

$$\frac{1}{2}\sum_{i \in S} |P_i - Q_i| \geqslant \frac{1}{2}\left|\sum_{i \in S}(P_i - Q_i)\right| = \frac{1}{2}(P(S) - Q(S)) > \frac{1}{2}(1 - \varepsilon - (1 - 2\varepsilon)) = \frac{\varepsilon}{2}.$$

– 当 $Q(S) < 1 - 2\varepsilon$ 时，同理，

$$\frac{1}{2}\sum_{i \in S} |P_i - Q_i| = \frac{1}{2}\sum_{i \in \Omega} |P_i - Q_i| - \frac{1}{2}\sum_{i \notin S} |P_i - Q_i| > \frac{1}{2}(10\varepsilon - (\varepsilon + 2\varepsilon)) = \frac{7}{2}\varepsilon.$$

在这两种情形下，我们都有 $d_{\mathrm{TV}}(P_S, Q_S) \geqslant \Omega(\varepsilon)$。不过，我们注意到，相同的技术不适用于海林格，因此需要稍强一些的保证。

6 结论与未来方向

在本文中，我们给出了检验未知贝叶斯网络最大入度所需的（近乎）紧的样本复杂度界。在此过程中，我们还获得了若干具有独立意义的结果，包括在 χ^2 散度下针对贝叶斯网络的近似适当学习器，以及一个高概率 χ^2 学习算法（适用于任意离散分布）。

我们的结果提出了两个有趣的未来研究方向。首先，将我们的检验结果推广到更一般的问题：在最大度为k 的假设下，对最大度为d 的情况进行检验，其中$k > d$ 均为输入参数；特别地，我们的结果对应于 $k = n$。其次，要么加强我们的高概率 χ^2 学习下界，以获得对 $\log(1/\delta)$ 的加性依赖（全变差距离学习已知此类结果），而不是乘性依赖；要么证明 $\log(1/\delta)$ 的这种乘性依赖是必要的。我们注意到，即使对于较弱的 KL散度学习，目前也尚无此类结果。

致谢

Yang 谨向 Philips George John 致谢，感谢他进行的有益讨论，并感谢他在下界分析中建议使用 δ 函数记号。

参考文献

[ABDK18]Jayadev Acharya、Arnab Bhattacharyya、Constantinos Daskalakis 和 Saravanan Kandasamy。利用干预学习和检验因果模型。载于 NeurIPS，第 9469–9481 页，2018 年。2[ADK15] Jayadev Acharya、Constantinos Daskalakis 和 Gautam Kamath。分布性质的最优检验。载于 NIPS，第 3591–3599 页，2015 年。3, 6, 17[BBC+20] Ivona Bez´akov´a、Antonio Blanca、Zongchen Chen、Daniel Stefankovic 和 Eric Vigoda。图模型检验的下界：着色与反铁磁伊辛模型。《机器学习研究杂志》，21:25:1–25:62，2020 年。2[BCY22] Arnab Bhattacharyya、Cl´ement L. Canonne 和 Joy Qiping Yang。有界度贝叶斯网络的独立性检验。CoRR，abs/2204.08690，2022 年。2, 3[BGPV21] Arnab Bhattacharyya、Sutanu Gayen、Eric Price 和 N. V. Vinodchandran。通过 Chow–Liu 近乎最优地学习树结构分布。载于 STOC，第 147–160 页。ACM，2021 年。3, 4[BV14] Stephen P. Boyd 和 Lieven Vandenberghe。《凸优化》。剑桥大学出版社，2014 年。14[Can20] Cl´ement L. Canonne。关于离散分布学习的简短札记，2020 年。3[CDDK22] Davin Choo、Yuval Dagan、Constantinos Daskalakis 和 Anthimos Vardis Kandiros。高效学习和检验潜在树伊辛模型。CoRR，abs/2211.13291，2022 年。2[CDKL22] Cl´ement L. Canonne、Ilias Diakonikolas、Daniel M. Kane 和 Sihan Liu。直方图分布检验的近乎最优界。CoRR，abs/2207.06596，2022 年。3[CDKS17] Cl´ement L Canonne、Ilias Diakonikolas、Daniel M Kane 和 Alistair Stewart。检验贝叶斯网络。载于学习理论会议，第 370–448 页。PMLR，2017 年。1, 2[DDK17]Constantinos Daskalakis、Nishanth Dikkala 和 Gautam Kamath。伊辛模型多线性函数的集中性及其在网络数据中的应用。载于 NIPS，第 12–23 页，2017 年。2[DDK19] Constantinos Daskalakis、Nishanth Dikkala 和 Gautam Kamath。检验伊辛模型。IEEE《信息论汇刊》，65(11):6829–6852，2019 年。2[Dia16] Ilias Diakonikolas。学习结构化分布。载于《大数据手册》，Chapman & Hall/CRC《现代统计方法手册》，第 267–283 页。CRC 出版社，美国佛罗里达州博卡拉顿，2016 年。3[DKW18]Constantinos Daskalakis、Gautam Kamath 和 John Wright。哪些分布距离可以用次线性样本进行检验？载于 SODA，第 2747–2764 页。SIAM，2018 年。3, 6, 17[DL01] Luc Devroye 和 G´abor Lugosi。《密度估计中的组合方法》。Springer统计学丛书。Springer-Verlag，纽约，2001 年。3[DP17] Constantinos Daskalakis 和 Qinxuan Pan。贝叶斯网络的平方海灵格距离次可加性及其

在恒等性检验中的应用。载于 COLT，《机器学习研究论文集》第 65 卷，第 697–703 页。PMLR，2017 年。2[FLNP00] Nir Friedman、Michal Linial、Iftach Nachman 和 Dana Pe'er。利用贝叶斯网络分析表达数据。载于第四届计算分子生物学年度国际会议论文集，第 127–135 页，2000 年。1[GNS18] Aditya Gangrade、Bobak Nazer 和 Venkatesh Saligrama。伊辛模型中的双样本检验可能与结构学习一样困难：极小极大下界。载于 ICASSP，第 6931–6935 页。IEEE，2018 年。2[Hec98] David Heckerman。贝叶斯网络学习教程。Springer，1998 年。1[KOPS15]Sudeep Kamath、Alon Orlitsky、Dheeraj Pichapati 和 Ananda Theertha Suresh。关于从样本中学习分布。载于 COLT，《JMLR研讨会与会议论文集》第 40 卷，第 1066–1100 页。JMLR.org，2015 年。3, 6, 14[NL19] Matey Neykov 和 Han Liu。高维伊辛模型中的性质检验。《统计年刊》，47(5):2472– 2503，2019 年。2[Pea88] Judea Pearl。《智能系统中的概率推理：合理推断网络》。Morgan kaufmann，1988 年。1[Pea95] Judea Pearl。从贝叶斯网络到因果网络。《人工智能中处理部分知识的数学模型》，第 157–182 页，1995 年。1[SPP+05] Karen Sachs、Omar Perez、Dana Pe'er、Douglas A Lauffenburger 和 Garry P Nolan。基于多参数单细胞数据推导因果蛋白质信号网络。《科学》，308(5721):523–529，2005 年。1[WJ08] Martin J Wainwright 和 Michael I Jordan。图模型、指数族与变分推断。《机器学习基础与趋势》，1(1–2):1–305，2008 年。1