arXiv:2305.03712v2 [stat.ME] 8 Jun 2023

公平性审计的统计推断

John J. Cherian*和 Emmanuel J. Cand`es†

2023年6月9日

摘要

在高风险问题中部署黑箱模型之前，评估模型在敏感子总体上的性能非常重要。例如，在累犯预测任务中，我们可能希望识别出预测模型假阳性率高得不可接受的人口统计群体，或认证不存在此类群体。在本文中，我们将这一通常被称为“公平性审计”的任务表述为多重假设检验问题。我们展示了如何使用自助法，在统计保证下同时界定一组群体中的性能差异。我们的方法可用于标记受模型性能不佳影响的子群体，并认证模型表现充分的子群体。重要的是，我们的审计与模型无关，几乎适用于任何性能指标或群体公平性标准。我们的方法还可处理极其丰富的——甚至无限的——子群体集合。此外，我们通过展示如何评估某些分布偏移下的性能，将分析推广到超越子群体的情形。我们在预测推断和算法公平性的基准数据集上测试了所提出的方法，发现我们的审计能够提供可解释且可信赖的保证。

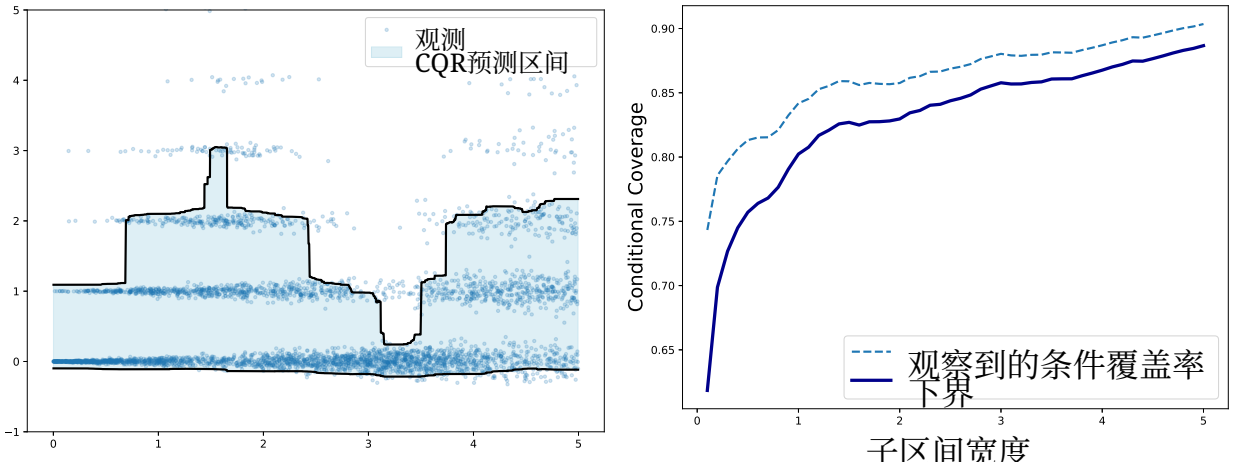
1 引言

尽管黑盒模型的平均准确率可能十分出色，但其性能在不同子总体之间仍可能存在显著差异。例如，用于预测再犯风险的算法在非裔美国人假释人员中的假阳性率显著高于白人假释人员 [1]。在面部识别和招聘等其他高风险应用中，也已记录到类似的性能差异 [5, 7]。

出于对这一问题的关注，许多利益相关方要求采用能够发现并量化此类差异的方法，这些方法通常被称为“公平性审计” [4, 33]。尽管该领域已有大量研究 [30, 42, 8, 37, 36, 34, 43, 40]，，公平性审计的定义仍存在诸多争议。公平性审计通常被描述为一项统计检验：只要在有限的一组敏感子总体中出现任何性能差异，检验便会拒绝原假设 [8, 37, 34, 30, 36, 42, 32]。人们希望（也往往会）开展后续调查，以定位差异，但这项任务也带来了新的挑战。例如，在有限的一组子群体中观察到经验上的平等，并不能排除较小子群体之间存在显著差异 [25]。此外，如果我们针对丰富多样的子总体考察多项性能指标，那么发现两个子群体之间存在某种差异也就不足为奇了。遗憾的是，现有用于识别局部（不）平等的方法鲜有统计保证 [40, 43, 33]。

我们提出一系列统计方法，能够严谨地实现两个目标：（1）“认证”模型表现良好的子群体；（2）“标记”

*斯坦福大学统计系
†斯坦福大学统计与数学系



(a) (b)

图 1：在图 1a中，我们展示了使用分位数随机森林在 [31]研究的合成数据集上构造的 90% 保形预测集。图 1b 展示了预测集在所有给定宽度的子区间上的条件覆盖率。虚线表示观测到的覆盖率，实线则绘制了通过 算法 2 ($\alpha = 0.05$, $p_* = 0.01$, $w_0 = \infty$) 获得的同时下界。

遭受有害性能差异的子群体。所提出的方法只需访问所谓的“审计轨迹”，即模型对一组未用于训练的数据集所做的预测， [4],但无需以白盒方式访问模型本身。我们提供了一个 Python 软件包，fairaudit，用于实现这些方法，地址为 github.com/jjcherian/fairaudit。

1.1 贡献概述

为阐明并概述本文的主要贡献，我们先介绍两个应用。

1.1.1 认证条件覆盖率

考虑训练集 $\{(X_i, Y_i)\}_{i=1}^n$ 和测试点 $\{X_{n+1}, Y_{n+1}\}$，它们均从某个未知分布 P 中独立同分布地抽取。以 $\{X_i, Y_i\}_{i=1}^n \cup \{X_{n+1}\}$ 为输入，共形预测会生成一个集合值函数，记为 $C(\cdot)$，并保证 $\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$，该保证对于训练点和测试点的随机性是边际成立的。然而，这种边际保证并不能排除在给定 C 和 $X \in G$ 的条件下覆盖率下降。可能存在一些子集 $G \subseteq \mathcal{X}$，使得 $\mathbb{P}(Y \in \hat{C}(X) \mid \hat{C}, X \in G) \ll 1 - \alpha$。众所周知，若不对数据生成分布作出强假设，要在丰富的子总体集合上保证条件覆盖率是很困难的 [18]。

我们重新审视 Romano 等人提出的合成数据实验 [31] ，以了解为什么人们可能希望对大量群组的条件覆盖率进行认证。查看图 1a ($n = 4000$) 中绘制的预测集后，我们可能会觉得， $\mathcal{X}$ 大多数子区间的条件覆盖率与 90% 的边际保证相差不大。但我们能否严谨地证明这一观察?

给定任意且可能依赖于数据的子区间集合，我们展示了如何针对每个子区间给出条件覆盖率的下置信界。为平衡计算成本与所考虑群组集合的复杂性，我们将分析限制

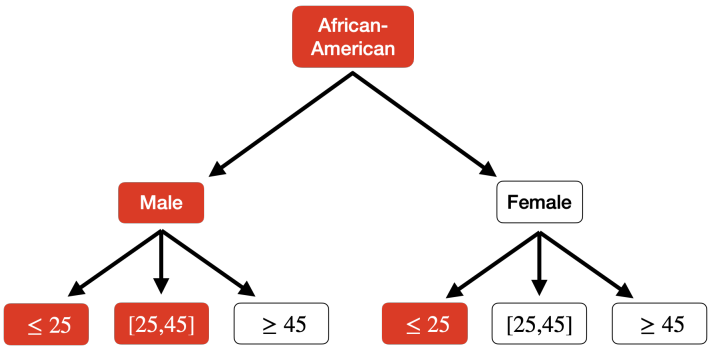


图 2：对于 COMPAS 算法，我们审计了由种族、性别和年龄交叉形成的 58 个子总体。此处绘制的是非裔美国人子总体中的子群体。红色阴影框表示被标记为假阳性率至少高于平均水平 5% 的群体。

在端点属于 $\{0, 0.1, 0.2, \ldots, 5\}$的子区间上。随后，以高概率，所给出的界将在该集合中的所有子区间上同时成立。利用绘制的审计轨迹，我们对给定宽度的所有子区间中的最小条件覆盖率给出下界，即 $\hat{} \hat{} \min_{G:\text{width}(G)=w} \mathbb{P}(Y \in C(X) \mid C,\ X \in G)$。图 1b 绘制了这一同时有效的下界以及观测到的条件覆盖率，$\min^{G:\text{width}(G)=w} \hat{} \mathbb{P}_n(Y \in \hat{} C(X) \mid \hat{} C,\ X \in G)$。例如，我们可以以 95% 的置信度断言，长度为 1 的子区间中没有任何一个的条件覆盖率低于 80。2%，长度为 2 的子区间中也没有任何一个的条件覆盖率低于 83。0%。至关重要的是，我们的保证是精确的：在大样本中，任一下界无效的概率收敛于 5%。

在本例中，我们同时为 1275 个子区间给出了有效的下界。不过，我们将范围限制在有限集合的子区间内并非必要，审计员可酌情放宽这一限制。有关生成下界的完整流程，请参见第 3 节。

1.1.2 标记假阳性率差异

设想某地方法院正在评估 COMPAS 再犯预测算法是否存在偏见。尽管以种族定义的子总体是以往研究中最受关注的对象， [1], 法院很可能会关注针对由受法律保护属性交叉形成的任何群体的歧视，例如年龄、性别和族裔。在这个范围更广的子群体集合中，发现存在某种差异已不再是关注重点，而准确定位严重差异则至关重要。

遵循 Angwin 等人 [1], 的做法，我们将审计方法应用于 ProPublica 于 2016年 获得的一个数据集，该数据集包含 COMPAS 风险评分（$f(X) \in \{$低风险，高风险$\}$）、被告人口统计信息（X_d）以及两年后的真实再犯结果（$Y \in \{0, 1\}$），涵盖 $n = 6781$ 名个体。在图2中，我们用红色阴影标出被标记为存在不成比例的假阳性率的人口统计群体，即 G 那些

$$\mathbb{P}(f(X) = \text{high-risk} \mid Y = 0, X_d \in G) - \mathbb{P}(f(X) = \text{high-risk} \mid Y = 0) > 0.05.$$

为了应对在审计许多群体时不可避免地会发现某些差异，我们的方法会对所发出标记无效的比率提供正式保证。错误标记性能差异的后果不如错误认证公平性严重，因此，我们提供的保证比前例中的同时有效性更宽松。我们改为，

提供“错误发现率”的渐近有效上界，即错误发出的标记所占比例的期望值。对于图示示例，我们采用第 4 节所述的标记方法，将这一比例控制在 10%。

1.2 概述

本文其余部分安排如下。我们首先在第 2 节介绍问题设定及相关符号。之后的每一节都专门介绍一种审计方法。在第 3 节中，我们介绍如何对一组子总体中的性能差异进行认证。在第 4 节中，我们展示如何标记性能差异。最后，在第5 节中，我们展示如何将方法从子总体扩展到一组分布偏移。

2 设置与符号说明

我们称预测规则 f 在子总体 G 上表现出性能差异，如果某项指标 $L(f(X), Y)$ 的均值，在条件为 (X, Y) 属于 G 时，与某个目标 $\theta_P \in \mathbb{R}$ 存在显著差异。

我们的统计审计通过检验和/或构建群体间性能差异的界限来进行，

$$\underbrace{\epsilon(G)}_{\text{disparity}} := \underbrace{\mathbb{E}_P[L(f(X), Y) \mid (X, Y) \in G]}_{\text{group-specific}} - \underbrace{\theta_P}_{\text{target}}. \tag{1}$$

子群成员资格可以由协变量的一个子集（例如某些敏感属性）来定义，而预测规则不会直接使用这些属性。不过，为了简化记号，在定义预测规则和群体成员资格时，我们都将使用同一个协变量向量 X。

几乎所有群体公平性定义都可以用 $\epsilon(G)$ 来表示。对于假阳性率示例，我们审计了一个由 $f(x)$ 给出的二元预测器。随后，我们针对所有交叉子群体实例化了标记方法，其中 $Y = 0$ 通过检验原假设 $H_0(G)$: $\epsilon(G) \leq 0.05$

$$L(f(x), y) = \mathbf{1}\{f(x) = 1\} \quad \text{and} \quad \theta_P = \mathbb{P}(f(X) = 1 \mid Y = 0).$$

ˆ 在条件覆盖率示例中，我们考虑了一个固定的集合值预测器 $C(x)$。然后，我们通过对 $\{0, 0.1, \ldots, 5\}$ 中端点构成的所有子区间，将 $\epsilon(G)$ 取下界，从而认证条件覆盖率

$$L(\hat{C}(x), y) = \mathbf{1}\{y \in \hat{C}(x)\} \quad \text{and} \quad \theta_P = 0.9.$$

参见附录A 以获取更多示例。

为了评估这些差异，我们设计的审计只需要访问一个留出数据集$\mathcal{D} = \{X_i, Y_i\}_{i=1}^n \overset{\text{iid}}{\sim} P$；这有时被称为“审计轨迹” [4]。

如果事先未知 θ_P，我们将假设可以利用这条审计轨迹构造一个一致估计量 θ。我们还将假设，对于某个可微函数 g 和已知特征 $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$，$\theta(\mathcal{D}) := g(\sum_i h(x_i, y_i)/n)$。在上下文明确时，我们省略 $\hat{\theta}$ 的参数。此假设的一个重要推论是渐近线性，即存在某个 $\psi(X, Y)$，使得 $\sqrt{n}(\hat{\theta} - \theta_P) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(X_i, Y_i) + o_P(1)$。通常，$\theta_P = \mathbb{E}_P[L(f(X), Y)]$，，因此 $\hat{\theta}(\mathcal{D}) = \frac{1}{n} \sum_{i=1}^n L(f(x_i), y_i)$ 将满足这一假设。

为方便记号，我们定义以下术语。在含义明确时，我们将 $(X,\ Y)\ \in\ G$ 替换为 G。例如，我们令 $\mathbb{P}_n(G) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{(x_i,\ y_i)\ \in G\}$ 表示留出集里 $(X,\ Y)$ 属于 G 的经验概率，并令 $\mathbb{P}(G) := \mathbb{P}((X,\ Y)\ \in G)$ 表示 $(X,\ Y)$ 属于 G 的总体概率。为进一步简化记号，我们还将 $L(f(X)\ Y)$ 替换为缩写 L。因此，我们定义差异的插件估计量 $\hat{}\epsilon(G) := |G|^{-1}\sum_{i\in G} L_i - \hat{}\theta$。

3 性能认证

3.1 方法

3.1.1 界认证

我们首先考虑认证子总体的问题，为 $\epsilon(G)$ 提供一个同时有效的置信集。为简化说明，我们构造组间差异的同时有效下界，即定义 $\epsilon_{\text{lb}}(G)$，使得

$$\lim_{n\to\infty} \mathbb{P}(\epsilon_{\text{lb}}(G) \le \epsilon(G) \text{ for all } G \in \mathcal{G}) = 1-\alpha. \tag{2}$$

尽管对性能差异取下界看起来可能有悖直觉，但请回想第1.1.1节预先介绍的条件覆盖率示例。通过附录 $\epsilon(G)$ 中所述的简单修改，可以得到其置信上界和区间B.3。

直观地，我们可以利用性能差异估计量与真实差异之间最大偏差的上界来定义 $\epsilon_{\text{lb}}(G)$，即 $\sup_{G\in\mathcal{G}}\{\hat{}\epsilon(G)\ -\epsilon(G)\}$ 的 $(1-\alpha)$ 分位数。形式化地定义 $\text{Quantile}(\alpha; X) := \inf\{x :\ \alpha \le \mathbb{P}(X \le x)\}$，便可通过以下方式得到 $\epsilon(G)$ 的同时有效下界：

$$\hat{\epsilon}(G) - \text{Quantile}\left(1-\alpha; \sup_{G\in\mathcal{G}}\{\hat{\epsilon}(G) - \epsilon(G)\}\right). \tag{3}$$

尽管证明所提出的下界满足（2）很直接，但仍有两个关键问题。首先，无法准确估计（3）中的分位数：小组中的 $\hat{}\epsilon(G)$ 会发散，因此当 $\mathcal{G}$ 较大时，$\hat{}\epsilon(G)$ 不会一致收敛到 $\epsilon(G)$。因此，标准渐近方法（例如自助法）会失效。其次，即使我们能够估计这个分位数，对于所有组而言，我们得到的也只是对朴素估计量的一个常数修正。由于要同时保证小组上的有效性，这种修正必须足够大，因此该方法会产生保守程度过高、难以实际应用的下界。

为绕过第一个障碍，我们证明可以一致地估计以下分布

$$\sup_{G\in\mathcal{G}}\{\mathbb{P}(G)\cdot\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G) - \epsilon(G))\}. \tag{4}$$

直观而言，将朴素过程乘以 $\mathbb{P}(G)\cdot\mathbb{P}_n(G)$ 后，小组中的取值会趋于稳定。接着，我们应用自助法 [14, 21] 来估计该过程的（$1-\alpha$）分位数。要严格证明自助法一致性，需要进行技术性论证；详细说明请参见定理 3.1 的证明。

仿照我们最初的方法，我们使用 $1-\alpha$ 的（4）分位数估计值，构造真实差异的同时有效下界。对于所有 $G\in\mathcal{G}$，我们定义 $\epsilon_{\text{lb}}(G)$，使得

算法1 自助法下置信界临界值

1：输入：子总体 $\mathcal{G}$，审计轨迹 $\mathcal{D}$，水平 α，自助样本数量 B2：对于 $b = 1, \ldots,$ B，执行 3：令 $\mathcal{D}_b^*$ 为从 $\mathcal{D}$中有放回抽取的大小为 n 的样本；4：定义 $\epsilon_b^*(G)$: $= \frac{1}{|G|^*} \sum_{(x_i^*, y_i^*) \in \mathcal{D}_b^*} L_i^* - \hat{}\theta(\mathcal{D}^*)$；5：定义 $\mathbb{P}_b^*(G) := \frac{1}{n} \sum_{(x_i^*, y_i^*) \in \mathcal{D}_b^*} \mathbf{1}\{(x_i^*, y_i^*) \in G\}$；6：$t^{(b)} = \max_{G \in \mathcal{G}} \{\mathbb{P}_n(G) \cdot \mathbb{P}_b^*(G) \cdot (\epsilon_b^*(G) - \hat{}\epsilon(G))\}$；7：循环结束 8：返回：$t^* =$ 分位数$(1 - \alpha; \{t^{(b)}\}_{b=1}^B)^-$

$\mathbb{P}_n(G)^2 \cdot (\hat{}\epsilon(G) - \epsilon_{\text{lb}}(G))$ 等于 $(1-\alpha)$ 分位数的 (4。令 t^* 表示该分位数的自助法估计，即 算法1的输出，我们可以得到 $\epsilon_{\text{lb}}(G)$ 的简化定义：

$$\epsilon_{\text{lb}}(G) := \hat{\epsilon}(G) - \frac{t^*}{\mathbb{P}_n(G)^2}. \tag{5}$$

给定 t^* 的有效估计， $\epsilon_{\text{lb}}(G)$ 的同时有效性是我们定义的直接推论：

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(\epsilon_{\text{lb}}(G) \le \epsilon(G) \text{ for all } G \in \mathcal{G}) &= \lim_{n\to\infty} \mathbb{P}(\hat{\epsilon}(G) - t^*/\mathbb{P}_n(G)^2 \le \epsilon(G) \text{ for all } G \in \mathcal{G}) \\ &= \lim_{n\to\infty} \mathbb{P}\left(\sup_{G\in\mathcal{G}} \mathbb{P}_n(G)^2 \cdot (\hat{\epsilon}(G) - \epsilon(G)) \le t^*\right).\end{aligned}$$

第一个等式由我们对 $\epsilon_{\text{lb}}(G)$的定义得出，第二个等式则由重新排列得出。将 $\mathbb{P}_n(G)$替换为 $\mathbb{P}(G)$（根据斯卢茨基引理），再应用 t^* 的定义，便完成了我们对同时有效性的论证：

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{P}\left(\sup_{G\in\mathcal{G}} \mathbb{P}_n(G)^2 \cdot (\hat{\epsilon}(G) - \epsilon(G)) \le t^*\right) \\ &\qquad = \lim_{n\to\infty} \mathbb{P}\left(\sup_{G\in\mathcal{G}} \mathbb{P}(G) \cdot \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon(G)) \le t^*\right) = 1 - \alpha.\end{aligned}$$

尽管该界满足（2）给出的有效性条件，但直接应用（5）会导致在小群体上得到的 $\epsilon(G)$ 下界在实际应用中几乎无法使用。这是由（5）中对群体规模的次优依赖关系造成的。从渐近意义上看，我们的界收敛于

$$\epsilon_{\text{lb}}(G) = \hat{\epsilon}(G) - C_0 \cdot \left(\sqrt{\frac{1}{|G| \cdot \mathbb{P}_n(G)^3}}\right),$$

其中， C_0 是某个与群体无关的常数， $|G|$ 表示属于 G的观测数据点数量。尽管界限修正项以预期的 $1/\sqrt{|G|}$ 速率收敛至 0，但与群体相关的因子 $\mathbb{P}_n(G)^3$ 会使我们的置信集急剧膨胀，即便群体规模适中也是如此。图3 复现了图1b ，使用了（5），并显示该定义对除最大的子区间外的所有子区间都给出了无信息的下界。

为说明我们修订后的方法，我们指出， $\epsilon(G)$的经典 Wald 置信界具有以下渐近行为，

$$\epsilon_{\text{lb}}^{\text{ideal}}(G) = \hat{\epsilon}(G) - C_1 \cdot \frac{\sigma_G}{\sqrt{|G|}}; \tag{6}$$

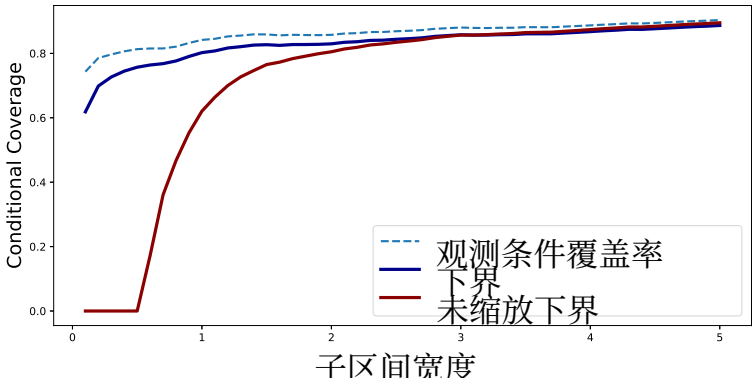


图 3：我们重现了图1b中的绘图。这里，红色曲线表示通过（5）得到的条件覆盖率下界。对于较小的群体，它明显比蓝色曲线宽松，即图1b中的界。蓝色曲线通过使用 $p_* = 0.01$ 和 $w_0 = \infty$ 对自助法过程进行缩放得到。

C_1 与群体独立， σ_G 表示$\sqrt{|G|}(\hat{\epsilon}(G) - \epsilon(G))$的渐近标准差。

通过对自助法过程进行重新缩放，我们可以为所有足够大的群组构造 Wald 型置信界。为了理解如何定义这一缩放因子，我们首先研究重新缩放自助法过程对最终置信界的影响。令 $s(G)$ 和 $\hat{s}(G)$ 分别表示缩放因子的总体值和估计量。然后，我们不再对 $\sup_{G\in\mathcal{G}}\{\mathbb{P}(G) \cdot \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon(G))\}$ 的 $(1-\alpha)$ 分位数进行自助法估计，而是估计以下表达式的 $(1-\alpha)$ 分位数。

$$\sup_{G\in\mathcal{G}} \left\{ \frac{1}{s(G)} \cdot \mathbb{P}(G) \cdot \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon(G)) \right\}. \tag{7}$$

令 t^* 表示（$1-\alpha$）分位数的（7，我们仿照前面的推导，得到关于 ϵ_{lb}（G）的闭式解，其表达式由 $\hat{\epsilon}$（G）和 $\hat{s}$（G）给出，

$$\epsilon_{\text{lb}}(G) := \hat{\epsilon}(G) - t^* \cdot \frac{\hat{s}(G)}{\mathbb{P}_n(G)^2}. \tag{8}$$

由于 $t^* = O_P(1/\sqrt{n})$，我们可以通过选择 $\hat{s}(G)$ 来估计$\mathbb{P}(G)^{3/2} \cdot \sigma_G$，从而得到 (6) 中的 Wald 型界。

直接对该过程进行朴素的重新缩放，会再次出现我们在第一种方法中遇到的小群体问题所导致的发散。具体而言，自助法一致性要求 $\{1/\hat{s}(G)\}_{G\in\mathcal{G}}$ 是 $\{1/s(G)\}_{G\in\mathcal{G}}$ 的一个一致一致估计量。为此，当 $\mathbb{P}_n(G)$ 低于某个阈值 p_* 时，我们不再估计 $\mathbb{P}(G)^{3/2} \cdot \sigma_G$。取而代之的是，我们采用更为朴素的估计量 $p_*^{3/2} \cdot \sqrt{\widehat{\text{Var}}(L)}$ 进行缩放，其中 $\widehat{\text{Var}}(L)$ 表示样本方差。

为了在这两种群体规模情形之间进行插值，我们设定

$$\hat{s}(G) = \max\{\mathbb{P}_n(G), p_*\}^{3/2} \cdot \left(\frac{\mathbb{P}_n(G)}{\mathbb{P}_n(G) + w_0} \cdot \hat{\sigma}_G + \frac{w_0}{\mathbb{P}_n(G) + w_0} \cdot \widehat{\text{Var}}(L)^{1/2} \right), \tag{9}$$

其中 $w_0 > 0$ 是用户指定的超参数，用于控制收缩程度；而 $\hat{\sigma}_G$是 σ_G的某个逐点估计量，但不一定一致相合。通过将 $\hat{\epsilon}(G)$ 的渐近方差估计收缩至一个与群体无关的量，$\hat{s}(G)$ 可获得所需的一致相合性。在 算法 2中，我们展示如何定义 $\hat{\sigma}_G$ ，并计算（8）中使用的 t^* 。

Al算法 2 对（重缩放后的）下置信界临界值进行自助法

1: 输入：子群体 $\mathcal{G}$，审计轨迹 $\mathcal{D}$，水平 α，阈值 p_*，权重 w_0，自助样本数量 B2: 对于 $b = 1, \ldots, B$，执行 3: 令 $\mathcal{D}_b^*$ 为从 $\mathcal{D}$ 中有放回抽取的、样本量为 n 的样本；4: 定义 $\mathbb{P}_b^*(G) := \frac{1}{n} \sum_{(x_i^*, y_i^*) \in \mathcal{D}_b^*} \mathbf{1}\{(x_i^*, y_i^*) \in G\}$；5: 定义 $\epsilon_b^*(G) := \frac{1}{\mathbb{P}_b^*(G) \cdot n} \sum_{(x_i^*, y_i^*) \in G} L_i^* - \hat{}\theta(\mathcal{D}_b^*)$；6: 结束循环 7: 将渐近方差估计量定义为_

$$\hat{\sigma}_G^2 := \widehat{\mathrm{Var}}(L \mid G) + \mathbb{P}_n(G) \left(\widehat{\mathrm{Var}}(\psi) - 2 \cdot \widehat{\mathrm{Cov}}(L, \psi \mid G) \right)$$

其中，Var(·) 和 Cov(·) 分别表示样本（条件）方差和协方差；8: 根据式 (9) 定义 $\hat{}s(G)$；9: 对于 $b = 1, \ldots, B$，执行 10: $t^{(b)} = \max_{G \in \mathcal{G}} \{ \frac{1}{\hat{}s(G) \cdot \mathbb{P}_n(G)} \cdot \mathbb{P}_b^*(G) \cdot (\epsilon_b^*(G) - \hat{}\epsilon(G))\}$；11: 结束循环 12: 返回：$t^* = \mathrm{Quantile}(1 - \alpha; \{t^{(b)}\}_{b=1}^B)$

实践中我们发现，估计 $\hat{}\epsilon(G)$ 的渐近方差可能会损害 $\epsilon_{\mathrm{lb}}(G)$ 在有限样本下的有效性。因此，除非认为适应 $\hat{}\epsilon(G)$ 的方差至关重要，否则我们建议设定 $w_0 = \infty$ 。

使用算法2在（8）中的输出，我们得到了更实用的置信界。我们使用 $p_* = 0.01$ 和 $w_0 = \infty$绘制了图3 中的蓝色曲线。鱼与熊掌不可兼得：可以看到，红色曲线（未缩放）在最大的子区间宽度下给出了更紧的下置信界。尽管如此，显然，重新缩放后的过程能在宽得多的群体规模范围内给出可用的下界。

定理3.1 给出了自助法一致性的充分条件，因此也保证了式（5）或（8）中定义的下界具有同时有效性。

$\hat{}$**定理 3.1**（同时有效性）。假设 L 有界，且 $L - \theta$ 在至少一个非空群体上不是常数。进一步假设 $\mathcal{G}$ 的 *Vapnik-Chernovenkis*（*VC*）维数有限。那么，$\epsilon_{\mathrm{lb}}(G)$ 是 $\epsilon(G)$ 的一个渐近 $(1 - \alpha)$ 下置信界，且对所有 $G \in \mathcal{G}$ 同时有效，即，

$$\lim_{n \to \infty} \mathbb{P}\left(\epsilon_{\mathrm{lb}}(G) \leq \epsilon(G) \text{ for all } G \in \mathcal{G}\right) = 1 - \alpha.$$

关于我们的假设，有两点说明。首先，我们要求 $\mathcal{G}$ 具有有限的 VC 维数，而大多数可解释的群集合都满足这一条件，例如：区间、矩形、半空间等。其次，在典型的公平性应用中，L 取值于 $\{0, 1\}$，并满足有界性假设；对于无界度量，审计员可以截断 L，或者在适当情况下假设定义域具有紧致性。此外，如果 $\mathcal{G}$ 是有限集合，我们可以将这一假设放宽为 $\mathrm{Var}(L) < \infty$。

尽管定理 3.1 是一个渐近结果，但有限样本方法（例如通过经验过程集中）只能满足保守的覆盖保证。相比之下，$\epsilon_{\mathrm{lb}}(\cdot)$ 的同时覆盖率在任何数据生成分布下都收敛于 精确的 $1 - \alpha$。即使样本量较小，我们也在第 3.2 节证明，名义覆盖率与实际覆盖率之间的差距很小。

算法3 自助法计算布尔证书临界值

1: **输入：** 子群体 $\mathcal{G}$、差异 ϵ、审计轨迹 $\mathcal{D}$、水平 α、自助样本数量 B 2: **对于** $b = 1, \ldots, B$ **执行** 3: 令 $\mathcal{D}_b^*$ 为从 $\mathcal{D}$ 中进行有放回抽样得到的大小为 n 的样本； 4: $t^{(b)} = \text{最大值}_{G \in \mathcal{G}} \{\mathbb{P}_b^*(G) \cdot (\hat{\epsilon}_b^*(G) - \epsilon) - \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon)\}$； 5: **结束循环** 6: **返回：** $t^* = \text{分位数}(1-\alpha; \{t^{(b)}\}_{b=1}^B)$

3.1.2 布尔认证

接下来，我们考虑在以下情况下为其颁发布尔证书 G 如果 $\epsilon(G)$ 高于某个预先指定的容差ϵ。我们的方法在认证 $\epsilon(G) < \epsilon$ 和 $|\epsilon(G)| < \epsilon$ 方面的直接扩展见附录 B.3。尽管与置信界证书相比，这种方法向审计员提供的信息严格更少，但它具有计算优势，并且在某些情形下能够提供更高的功效。

形式上，认证 $\epsilon(G) > \epsilon$ 对应于检验原假设 $H_0(G)$: $\epsilon(G) \leq \epsilon$。我们要求在大样本下，所有颁发的证书都以概率 $1-\alpha$同时有效。等价地，

$$\lim_{n \to \infty} \mathbb{P}\left(\text{there exists any falsely certified } G \in \mathcal{G}\right) \leq \alpha.$$

用多重检验的术语来说，这一目标是控制（渐近）强家族错误率（FWER）。

为构造这样的检验，我们证明自助法可以保守地估计 $\sup_{G \in \mathcal{G}} \{\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon)\}$ 的 $(1-\alpha)$-分位数。令 t^* 表示算法3给出的估计值；当 $\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) \geq t^*$ 时，我们认证 $\epsilon(G) > \epsilon$。简化后，当

$$\hat{\epsilon}(G) \geq \epsilon + \frac{t^*}{\mathbb{P}_n(G)}. \tag{10}$$

例如，如果 L 是覆盖率指示量，那么当 G上的经验条件覆盖率足够高时，我们就认证 G。

尽管（10）中拒绝阈值的缩放方式对小群体而言并非最优，但对此进行修正会消除布尔认证程序的一项独特优势：能够高效地针对某些无限群体集合（例如区间、带状区域）进行优化。注意，

$$\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) = \frac{1}{n} \sum_{i=1}^{n} (L_i - \hat{\theta} - \epsilon) \mathbf{1}\left\{(x_i, y_i) \in G\right\},$$

也就是说，该过程关于群指示变量是线性的。按 $s(G)$ 重新缩放会消除这种线性，因此 算法 3 的第 4 行需要对所有 $G \in \mathcal{G}$进行穷举搜索。为完整起见，我们在 算法7中描述了这一重新缩放的布尔认证程序变体。

定理 3.2 指出，在定理 10 的温和假设下，式（3.1）会生成有效的证书。

定理 3.2 （证书认证的家族错误率控制）。 保留定理 *3.1* 的假设。则，由 (10) 颁发的证书满足

$$\lim_{n \to \infty} \mathbb{P}\left(\text{there exists any falsely certified } G \in \mathcal{G}\right) \leq \alpha.$$

我们的自助法程序考虑了子总体之间的重叠情况，并优于邦弗罗尼检验等较为简单的 FWER 控制程序。然而，仅当所有 G 均满足 $\epsilon(G) = \epsilon$ 时，渐近 FWER 才恰好为 α。与此相比，同时置信集保证的表现较弱，因为后者承诺在任意数据生成分布下均能实现精确的第一类错误控制。接下来，我们将比较并分析基于布尔值和基于置信集的认证方法在有限样本有效性方面的异同。

3.2 实证结果

3.2.1 合成数据验证

首先，我们验证定理3.1 和定理3.2 的渐近结论在有限样本中是否准确。我们考察三个合成数据实验；各实验的结果见表1。

我们首先考察一个同方差线性模型。从该分布中抽取 $(X_i,\ Y_i)$

$$X_i \overset{\text{iid}}{\sim} \text{Unif}(0,1), \quad Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\beta_0 X_i, 1). \tag{11}$$

ˆ然后，我们使用从该分布中抽取的 1000 个训练样本点，通过普通最小二乘法得到 $f(x) = \beta x$。我们关注的性能指标是平方误差损失，即 $L(f(X)\ Y) = (Y - f(X))^2$，目标为 $\theta_P = 0$。我们使用不同规模的留出数据集，对 $\epsilon(G) < 1$ 所涵盖的子区间 $G \subseteq [0,1]$ 颁发布尔认证。

在我们的 在第二个试验中，我们使用异方差线性模型验证我们的审计，

$$X_i \overset{\text{iid}}{\sim} \text{Unif}(0,1), \quad Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\beta_0 X_i, X_i). \tag{12}$$

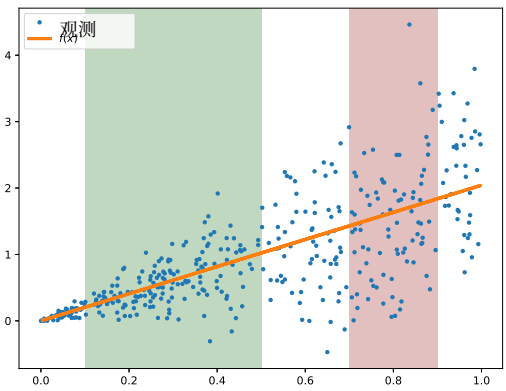
模型 $f(X)$ 和指标 $L(f(X)\ Y)$ 的获取方式与前一个合成试验完全相同。随后，我们针对满足 $G \subseteq [0,1]$ 且 $\epsilon(G) < \epsilon$ 对于$\epsilon \in \{0.4, 0.5\}$成立的子区间 G 颁发布尔认证。为简化对已颁发证书的验证，我们考虑端点属于 $\{0, 0.1, \ldots, 1\}$的 情形。图 4a 展示了一次试验，其审计轨迹的样本量为 400。在此设置下，我们观察到，名义错误率高估了错误认证的实际概率。

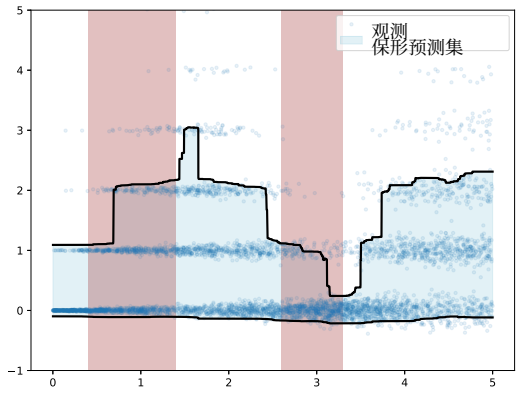
最后，我们考虑 [31] 引入并在图 4b中展示的合成数据集。在这里，共形化分位数回归（CQR）保证具有 90% 的边际覆盖率，因此我们考虑认证覆盖率超过 90% 的子区间，例如 $\theta_P = 0.9$ 以及 $L = \mathbf{1}\{Y \in \hat{C}(X)\}$。

表1 展示了认证审计在200次试验中的家族错误率。我们将名义错误率设为0.1，并改变样本量 n（100、200、400、800、1600）。结果证实了理论预测。当 n较大时，名义家族错误率与实际水平相当或高于实际水平。当原假设在边界处不成立时，即 $\epsilon(G)$与阈值 ϵ不近似相等时，名义家族错误率可能会大幅高估真实水平。表1还展示了功效，即达到认证阈值的子区间中，实际被我们的方法认证的比例。

相比之下，请回顾置信界方法能够在任何数据生成分布下（渐近地）保证精确覆盖率。为验证这一论断，我们针对上述每个合成实验构建了同时90%置信界。

表 2 总结了我们的结果：我们观察到 算法 1 及其重新缩放的变体 算法 2($p_* = 0.01, w_0 = \infty$) 在大样本中达到名义覆盖率。对于固定阈值 ϵ，我们将置信集的功效定义为真实错误率低于 ϵ





(a) (b)

图4：图4a 展示了一个线性模型在由（12）生成的400个留出点上的表现。为期望均方误差构造90%置信界后，绿色和红色子区间的上界分别为0。5（真实均方误差：0。3）和1。15（真实均方误差：0。80）。图4b展示了保形化分位数随机森林对来自 Romano 等人的合成数据生成过程的400个留出点所得到的预测区间。 [31]。为条件覆盖率构造90%置信界后，两个红色子区间的下界分别为81。1%（真实覆盖率：92。4%）和80。1%（真实覆盖率：89。5%）。这些置信界对于任意多个此类查询均同时有效。

时置信集排除 ϵ 的子区间比例。请注意，尽管未经缩放的置信界功效较低，也就是说，它们不太可能排除目标错误率，但重新缩放缓解了这一问题。

3.2.2 认证 COMPAS

接下来，我们重新审视此前对 COMPAS 再犯预测工具（RPI）的分析，并展示如何运用我们的方法建立严格保证。COMPAS 算法根据被告再次犯罪的估计可能性，为其分配 1 至 10 分的风险评分。先前的研究表明，与白人被告相比，非裔美国被告更有可能被错误地归类为高风险 [1]。针对这一发现，COMPAS 的创建者 Northpointe Inc. 辩称，根据预测平等标准评估时，该算法是公平的[9, 17]。尽管他们提供了统计证据，表明按照这一衡量标准不存在显著的种族偏见，但我们的方法能够构建一个关于真实差异的显式界。

当 G 这一单一群体满足预测平等标准时，意味着 f 对于 G 的阳性预测值（阳性预测值）与 G 的补集的阳性预测值相匹配，即，

$$\mathbb{P}(Y = 1 \mid f(X) = 1, X \in G) = \mathbb{P}(Y = 1 \mid f(X) = 1, X \in G^c).$$

直观而言，阳性预测值（PPV）衡量的是阳性预测所包含的信息量。例如，如果 COMPAS 将一名被告归类为高风险（$f(X) = 1$），那么阳性预测值对应的就是该被告实际再次犯罪的概率（$Y = 1$）。

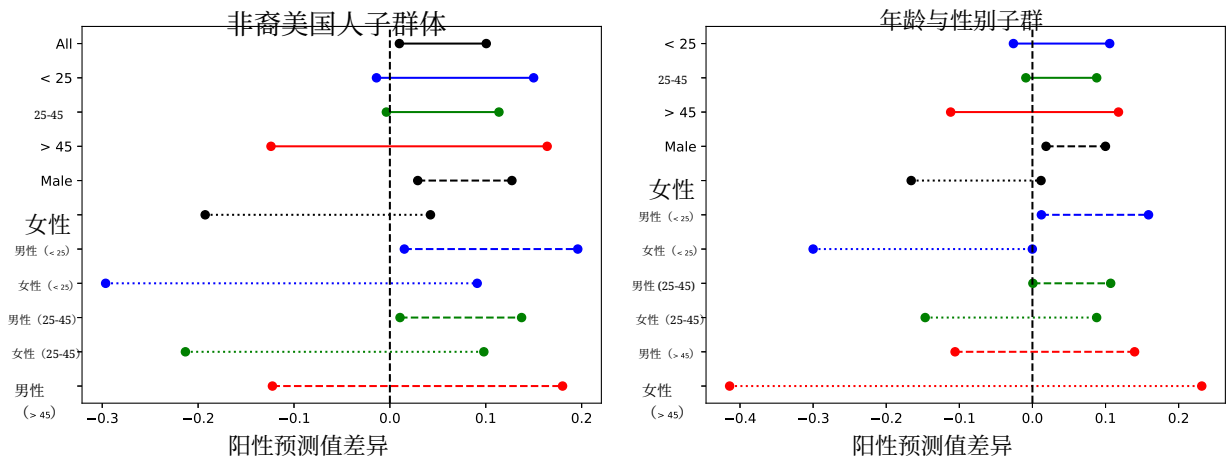
按照先前的研究，我们将 COMPAS 评分二值化，定义 $f(X)$ 在 RPI 评分为 5 或更高时取值为 1。然后，为了认证非裔美国人与白人被告阳性预测值差距的下界，我们考虑审计轨迹中满足 $f(X) = 1$ 的子集（$n = 2525$）

		样本量（n）				
		100	200	400	800	1600
模型（11）（$\epsilon = 1$）	FWER	0.145	0.145	0.17	0.125	0.095
模型（12）（$\epsilon = 0.5$）	FWER	0.105	0.105	0.1	0.075	0.065
	功效	0.145	0.268	0.411	0.562	0.709
模型（12）（$\epsilon = 0.4$）	FWER	0.045	0.04	0.015	0.025	0.02
	功效	0.036	0.079	0.21	0.375	0.57
CQR ($\epsilon = 0.9$)	FWER	0.081	0.034	0.019	0.012	0.003
	功效	0.019	0.008	0.009	0.02	0.062

表1：由 算法 3 和 $B = 500$ 及 $\alpha = 0.1$颁发的证书的家族错误率和功效。所有结果均基于200次试验。我们看到，对于较大的 n，同时有效性保证成立。

		样本量 (n)				
		100	200	400	800	1600
模型 (12)	覆盖率	0.84	0.83	0.86	0.895	0.905
	功效（$\epsilon = 0.5$）	0.043	0.081	0.15	0.256	0.378
	功效（$\epsilon = 0.4$）	0.003	0.005	0.012	0.058	0.187
模型（12）（重新缩放）	覆盖率	0.84	0.81	0.845	0.885	0.88
	功效（$\epsilon = 0.5$）	0.149	0.270	0.427	0.608	0.743
	功效（$\epsilon = 0.4$）	0.038	0.091	0.24	0.457	0.633
CQR（重新缩放）	覆盖率	0.815	0.87	0.845	0.865	0.91
	功效（$\epsilon = 0.9$）	0.025	0.011	0.007	0.017	0.055

表2：由 算法 1 和算法 2 在$B = 500$、 $\alpha = 0.1$和 $p_* = 0.01$下给出的置信界的同时覆盖率。所有结果均基于200次试验。我们看到，对于较大的 n，认证程序满足同时有效性保证。



(a) (b)

图 5：在图 5[5a]中，我们绘制了每个非裔美国人子群与整个白人子群之间 COMPAS 阳性预测值差异的 90% 置信区间。在图 [5b]中，我们绘制了由年龄和性别交叉划分形成的各个子群与整个白人子群之间 COMPAS 阳性预测值差异的 90% 置信区间。

以及 $X_{\text{race}} \in \{$非裔美国人，白人$\}$。我们将审计实例化为 L，它对应于指示变量，表示 Y 与 f（X）匹配； $\mathcal{G}$ 仅包含一个群体（非裔美国被告），而θ_P 表示白人被告的阳性预测值：

$$L(f(X), Y) = Y, \quad \mathcal{G} = \{\{(X, Y) \mid X_{\text{race}} = \text{African-American}, f(X) = 1\}\},$$
$$\theta_P = \mathbb{E}[Y = 1 \mid X_{\text{race}} = \text{Caucasian}].$$

^^ 我们使用经验条件期望 $\theta = \mathbb{E}_n[Y = 1 \mid X_{\text{race}} =$ 白人$]$ 来估计 θ_P。

为验证 [9] 所提出的关于非裔美国人被告相较于白人被告的阳性预测值没有降低这一主张，我们为 $\epsilon(G)$ 构造一个 $=\mathbb{P}(Y = 1 \mid f(X) = 1, \ X \in G) - \mathbb{P}(Y = 1 \mid f(X) = 1, \ X \in G^c)$。采用这种方法，我们可以认证先前的主张：在获得高风险预测的被告中，非裔美国人被告再次犯罪的可能性至少比白人被告高 1。87%。

由于我们仅考虑单个群体，因此本方法并非确立这一结果的必要条件。不过，借助这些工具，我们可以确立对所有受保护子群体同时成立的阳性预测值差异界限。我们再次以白人被告的 COMPAS 阳性预测值为目标，为由种族、性别和年龄交叉构成的每个子总体，构建同时有效的 90% 阳性预测值差异置信区间。在图 [5]中，我们绘出了非裔美国人子总体各子群体，以及仅由性别和年龄交叉构成的子群体所对应的区间。我们的结果证实，Northpointe Inc. 关于阳性预测值平等的主张适用于若干（但并非全部）非裔美国人子总体。更普遍地说，与白人子总体相比，年轻男性子总体的 COMPAS 阳性预测值似乎更高。

4 标记性能差异

4.1 方法

接下来，我们考虑标记差异超过某个容差的子总体这一问题，即

$$H_0(G) : \epsilon(G) \leq \epsilon. \tag{13}$$

不成立。错误标记的危害小于错误的性能证明，因此，只需采用弱于同时有效性的错误控制概念。事先，我们预计大多数群体都满足式（13）给出的原假设。因此，我们考虑一种相对于发出的标记数量而非所检验子群体数量定义的错误率；具体而言，我们控制渐近错误发现率（FDR）。在我们的设定中，控制错误发现率意味着将被错误标记的子总体的预期比例限制在 α以内。例如，在第1.1.2节中，我们将 FDR 控制在 10%，这意味着被标记的子群体中约有 90% 的假阳性率至少比总体平均值高 5%。

将这一标准形式化，我们要求

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{|\{\text{falsely flagged } G \in \mathcal{G}\}|}{|\{\text{flagged } G \in \mathcal{G}\}| \vee 1}\right] \leq \alpha.$$

我们指出，同时错误控制（家族错误率控制）意味着错误发现率也受到控制，因此后者是一个严格较弱的标准。

当子总体集合为有限集时，我们建议对一组边际p值应用 Benjamini-Hochberg程序（以下简称 BH(α）。算法 4描述了这些p值的计算过程。

算法4 为 $G \in \mathcal{G}$ 构造 p 值

1: 输入：子群体 $\mathcal{G}$，审计轨迹 $\mathcal{D}$，自助样本 B，容差 ϵ2: 对于 $b = 1, \ldots, B$ 执行 3: 令 $\mathcal{D}_b^*$ 为从 $\mathcal{D}$ 中有放回抽取的大小为 n 的样本； 4: 对于 $G \in \mathcal{G}$ 执行 5: $t^{(b)}(G) = \epsilon_b^*(G) - \hat{\epsilon}(G)$; 6: 结束循环 7: 结束循环 8: 对于 $G \in \mathcal{G}$ 执行 9: $s^*(G) = \Phi^{-1}(3/4) \cdot$ 分位数(0.5; $\{|t^{(b)}(G)|\}_{b=1}^{B}$); 10: $p(G) = 1 - \Phi((\hat{\epsilon}(G) - \epsilon)/s^*(G))$; 11: 结束循环 12: 返回： $\{p(G)\}_{G\in\mathcal{G}}$。

以下命题指出，BH(α) 程序（应用于 算法 4的输出）在两个具有实际意义的情形下控制渐近错误发现率：第一，群组两两不相交的情形；第二，二元结果且群组结构任意重叠的情形。第一种情形下该程序的有效性并不令人意外，因为不相交群组的 p 值相互独立。第二种情形的证明则更为微妙：我们证明，二元度量意味着 p 值之间存在某种正依赖[1] 。在这种相关结构下，BH 程序已知是有效的 [3]。

[1]正依赖 我们所识别的依赖关系在形式上称为子集上的正回归依赖（PRDS） [3] .

命题 4.1 (错误发现率控制)。 假设对于所有 $G \in \mathcal{G}$， $\mathbb{P}(G)$ 和 $\mathrm{Var}(L \mid G)$ 均有正的下界， θ_P 是先验已知的，并且以下条件至少有一项成立：

(i) $\{G\}_{G\in\mathcal{G}}$ 互不相交；*(ii)* L 取值于 $\{0,1\}$。

如果我们将 $BH(\alpha)$ 程序对 $\{p(G)\}_{G\in\mathcal{G}}$ 的拒绝标记，则错误发现率在渐近意义下受控于水平 α。

我们预计命题 4.1 在违反所述假设的情况下仍然成立。即使不属于命题 4.1 中所述的两种情况（例如，当我们必须估计 θ_P时），此前使用 BH 算法的实验以及我们自己的实证结果表明，该程序不会违反错误发现率控制 [15]。

如果希望在无限个子总体上进行错误发现率控制的标记，我们建议采用两阶段程序。首先，将审计轨迹拆分为两部分，并使用其中一部分发现一个有限的、具有可解释性的子总体子集，例如，拟合一棵回归树，并将每个叶节点定义为一个感兴趣的子总体。然后，使用另一部分执行由命题 4.1 验证的标记程序。

4.2 实证结果

4.2.1 Folktables

我们在一个收入预测数据集上评估这种标记方法。该数据集源自 2018年人口普查美国社区调查公共使用微观数据，并通过 Folktables软件包提供 [10, 16]。我们使用经过筛选的加利福尼亚州数据集，其中仅包含年龄超过 16 岁且过去一年收入至少为 100 美元的个体，旨在预测某个人的收入是否超过 $50,000。我们将年龄、出生地、受教育程度、种族、婚姻状况、职业、性别、种族和工作时长纳入拟合的预测规则中。

为了验证命题 4.1 中的（渐近）错误发现率控制结果，我们使用 1000 个数据点的训练集拟合逻辑回归和线性回归模型，然后从剩余数据中抽取不同大小的审计轨迹。我们将满足以下条件的子总体标记出来：（1）误分类率高于固定阈值 0。5；（2）误分类率高于总体平均错误率；（3）预测收入的均方误差（MSE）高于总体均方误差。这些任务各自揭示了命题 4.1 的相关性。第一个标记任务满足命题 4.1 的假设，而另外两个则违反了所述假设。

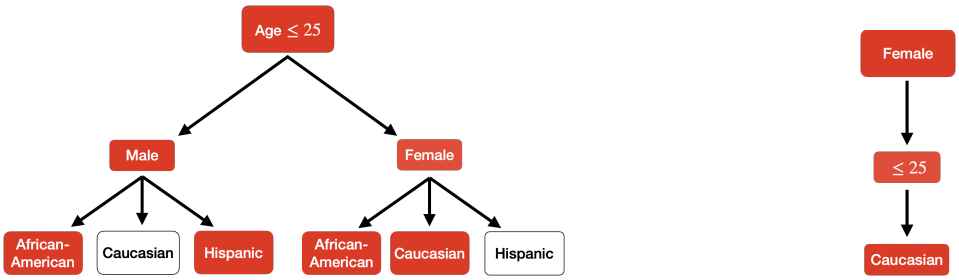
由于审计轨迹是有放回抽样的， P 因此，数据生成分布是模型拟合时留出的有限总体上的均匀分布。因此，如果被标记子组在整个留出数据集上的错误率未超过给定阈值，则该标记就是误报。表3 显示，在 1000 次试验中，每项任务的估计错误发现率都远低于所有测试样本量下 0.1 的名义上限。这是因为原假设 p 值在实践中较为保守，因为原假设（13）很少恰好成立。

4.2.2 COMPAS

此前对theCOMPAS RPI的分析表明，对于非裔美国被告，高风险判定的假阳性率明显高于白人被告 [1]。我们重新审视这个常被研究的公平性审计案例，以确定是否能发现其他假阳性率至少比被告平均水平高出 5% 的人口统计群体。具体而言，我们审计种族、性别和年龄组的所有交叉组合（$n = 6781$,

		样本量（n）					
	Task	100	200	400	800	1600	3200
Folktables	1	0.045	0.038	0.032	0.025	0.021	0.014
	2	0.003	0.005	0.003	0.004	0.003	0.004
	3	0.0	0.0	0.0	0.0	0.0	0.002

表 3：针对第 $B = 500$ 和 $\alpha = 0.1$ 节所述任务发出的标记的错误发现率：4.2.1。所有结果均基于 1000 次试验。我们看到，对于任何 n，错误发现率保证都是保守的。



(a) (b)

图6：在图 6a中，红色方框表示被标记为假阳性率显著高于平均水平的群体。大多数 25 岁以下子群体都受到这种差异的影响。在图 6b中，红色方框表示被标记为阳性预测值显著低于平均水平的群体。虽然我们只能标记三个嵌套子群体（女性、25 岁以下女性和 25 岁以下的高加索裔女性），但这些结果表明，某些子群体在这一指标上仍面临差异。

$|\mathcal{G}| = 48$）。图2 展示了非裔美国人子群体各子集所标记的对象；我们可以进一步将非裔美国被告中的假阳性率（FPR）差异定位到较年轻的非裔美国被告。如图6a所示，我们的方法还标记了几乎所有25岁以下子群体，这表明这种差异更普遍地影响年轻被告。

在上一节中，我们认证了 与白人被告相比，非裔美国被告的 COMPAS RPI 阳性预测值（PPV）更高。由于 COMPAS开发者声称 PPV 是一种更恰当的公平性衡量指标 [9]，我们将调查其他人口统计群体是否也遭受有害的阳性预测值差异。如图 6b所示，我们仍能标记出某些阳性预测值至少比平均值低 5% 的子群体。

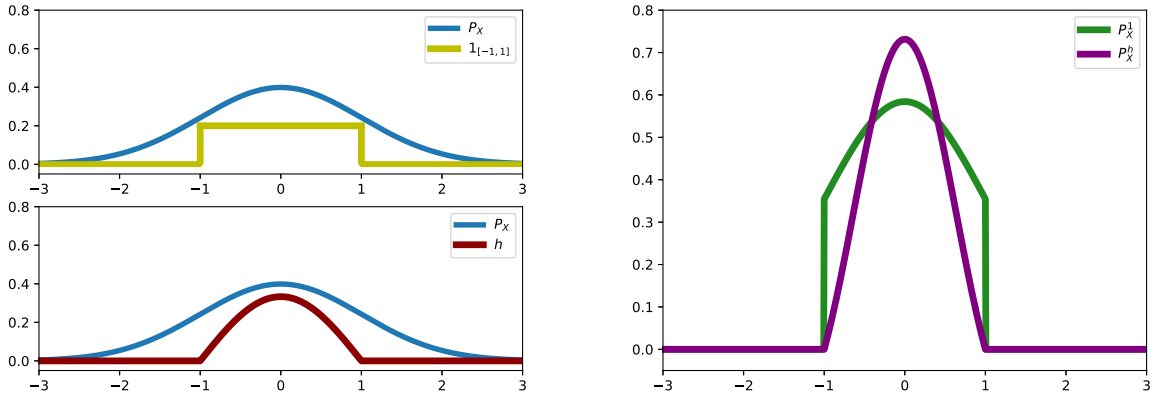
5 超越子群体

5.1 方法

对各子总体进行审计，等价于评估由倾斜 $\{\mathbf{1}\{(X,Y) \in G\}\}_{G\in\mathcal{G}}$ 所索引的分布偏移类别上的性能。具体而言，

$$\epsilon(G) = \mathbb{E}_P[L \mid G] - \theta_P = E_{P_G}[L] - \theta_P,$$

其中 $dP_G(x,\ y) \propto \mathbf{1}\{(x,\ y) \in G\}\, dP(x,\ y)$。在本节中，我们考虑将



(a) (b)

图 7：在图 7a中，我们绘制了一个指示函数倾斜，它对应于对 X 属于区间 $[-1,1]$进行条件化的情况，以及一个与该指示函数相似的非负倾斜。图 7b 绘制了这两种倾斜所产生的、关于 X 的（相似）概率密度。

这些倾斜从取值为 0 到 1 的函数推广到非负函数集合的自然方式。对于这样的集合，$\mathcal{H}$，我们可能希望对每个 h 界定以下差异，

$$\epsilon(h) = \mathbb{E}_{P_h}[L(f(X), Y)] - \theta_P,$$

其中 $dP_h(x,\ y) \propto h(x,\ y)dP(x,\ y)$。图 7 展示了如何用逼近指示函数的非负函数，在 $\mathcal{X}$上产生相似的倾斜分布。

为说明这一审计任务，考虑如何评估自动驾驶汽车在各种不同环境中的表现。尽管大多数关于分布鲁棒性的研究都侧重于评估模型在某个集合中的最坏情况分布偏移下的表现，但这种对模型鲁棒性的评估本质上存在局限 [11]。在某个易于处理的倾斜集合中寻找最坏情况，很难成为大多数现实世界分布偏移的有效替代指标。相比之下，我们的方法允许建模者查询属于已审计集合的任意分布偏移。随着遇到新的环境，建模者可以可靠地评估哪些环境会对表现产生良性影响，哪些环境可能尤其棘手。

当 $\mathcal{H}$ 为有限值时，我们用于认证和标记性能的方法保持不变。对于无限的 $\mathcal{H}$，先前的结果仍然适用：我们可以在任意 VC 类倾斜函数上认证性能差异。虽然这绝非详尽无遗，但此类函数包括非负多项式集合，以及针对单个协变量 [38]的任意有界单调倾斜函数集合。

从二元倾斜函数推广到非负倾斜函数，可将审计扩展至具有无限 VC 维数的某些函数类。我们在附录 $\mathcal{F}$ 中介绍了对一般函数类进行基于自助法的审计的充分条件，D，但此处重点介绍其中一个特别的函数集合：再生核希尔伯特空间（RKHS）的单位球。

令 h 表示属于某个再生核希尔伯特空间单位球的任意非负函数；我们用 $\mathcal{H}_1^+$ 表示此类函数的集合。定义 $\epsilon(h)$ 为在此倾斜下的差异，即：

$$\epsilon(h) := \mathbb{E}_{P_h}[L] = \frac{\mathbb{E}_P\left[(L-\theta_P)h(X)\right]}{\mathbb{E}_P[h(X)]}.$$

算法 5 自助法计算再生核希尔伯特空间置信集的临界值

1：输入：核函数 k，审计轨迹 $\mathcal{D}$，水平 α，自助样本 B2：
定义 L $:= \{L(f(x_i)\ y_i)\}_{i=1}^n$；3：定义 K $:= \{k(x_i, x_j)\}_{i,j=1}^n$；4：对于 $b = 1, \ldots, B$，执行 5：从 $\sim$Mult $\left(n \text{ 中抽取 w}; \frac{1}{n}, \ldots, \frac{1}{n}\right)$；6：A $= \frac{1}{n^2}\left((\text{w} \odot \text{L})\ \mathbf{1}_{-}^{\top} - \text{w}\underline{\text{L}}^{\top}\right)$；7：$t^{(b)} = \lambda_{\max}\left(\text{K}^{1/2}\left(\frac{\mathbf{A}+\mathbf{A}^{\top}}{2}\right)\text{K}^{1/2}\right)$；8：结束循环 9：返回：$t^* =$ 分位数$(1-\alpha; \{t^{(b)}\}_{b=1}^B)$

ˆ令 $\mathbb{E}_n[f(X)]$ 表示在经验分布下期望的插件估计量。然后，我们定义 ˆ$\epsilon(h)$: $=(\hat{}\mathbb{E}_n[h(X)])^{-1}\hat{}\mathbb{E}_n[(L-\hat{}\theta)h(X)]$。下文中我们将假设 $\theta_P = 0$，但我们在附录 D 中给出了将我们的方法推广到估计目标的情形。

我们强调基于再生核希尔伯特空间的审计任务的两个重要特点。首先，对于适当选取的再生核希尔伯特空间， h 可以近似任意定义在紧子集 P_X上的协变量分布的平滑倾斜， [29]。有了充足的数据，我们便能为几乎任意群体和协变量偏移下的模型性能提供保证。

在再生核希尔伯特空间单位球上进行审计还有另一项优势：我们可以构造 $\epsilon(h$的置信界，而无需在算法 1的每一步执行繁重的优化。回想一下，在算法 1中，我们使用自助法估计 sup$1-\alpha$ 的 $(_{G\in\mathcal{G}}\{\mathbb{P}_n)G$ 分位数 $\cdot\mathbb{P}(G)\cdot(\hat{}\epsilon(G) - \epsilon(G))\}$。每次迭代都需要针对 $\mathcal{G}$ 求解一个具有挑战性的组合优化问题。乍看之下，基于再生核希尔伯特空间的审计中类似的任务似乎更加困难。对于每个自助样本，我们必须计算

$$\max_{h\in\mathcal{H}_1^+}\{\hat{\mathbb{E}}_n[h(X)]\cdot\hat{\mathbb{E}}_b^*[h(X)]\cdot(\epsilon_b^*(h)-\hat{\epsilon}(h))\}.$$

从直接的角度来看，这个优化问题难以求解。

在算法5中，我们可以将此任务简化为计算一个低秩矩阵的最大特征值。对此简化有两个关键观测。首先，以 $h \in \mathcal{H}_1^+$为索引的过程的上确界，小于等于以不受限制的单位球 $\mathcal{H}_1$为索引的同一过程的上确界，即

$$t^* \doteq \text{Quantile}\left(1-\alpha;\ \sup_{h\in\mathcal{H}_1}\{\mathbb{E}_P[h(X)]\cdot\hat{\mathbb{E}}_n[h(X)]\cdot(\hat{\epsilon}(h)-\epsilon(h))\}\right). \tag{14}$$

因此，我们可以利用 RKHS 函数的有限维表示定理，简化自助算法每次迭代中的最大化问题。该定理指出，对于任意有限集合 $\{x_i\}_{i=1}^n$ 和 $h \in \mathcal{H}_1$， $\{h(x_i)\}_{i=1}^n = \mathbf{Kw}$ 对于 $\{\mathbf{K}^{ij}\}_{i,\ j=1}^n = \{k($ $x_i,\ x_j)\}_{i,\ j=1}^n$ 以及某个 $\mathbf{w}\ \in \mathbb{R}^n$ [35]。这一化简的其他步骤虽具技术性但缺乏启发性，故留待定理 5.1的证明中详述。

使用 算法 5的输出，我们为 $\epsilon(h)$ 构造下置信界

$$\epsilon_{\text{lb}}(h) := \hat{\epsilon}(h) - \frac{t^*}{\left(\frac{1}{n}\sum_{i=1}^n h(x_i)\right)^2}. \tag{15}$$

我们指出，类似于（7）的重新缩放也可以应用于此情形。然而，对于重新缩放后的再生核希尔伯特空间过程进行自助法计算，成本高得难以承受。尽管如此，我们仍在附录 D 中详细说明了适当的重新缩放和自助算法。

		样本量（n）				
	σ	100	200	400	800	1600
模型（16）	1	0.92	0.945	0.94	0.89	0.91
	0.5	0.93	0.955	0.935	0.885	0.915
	0.1	0.95	0.95	0.935	0.93	0.95

表 4：由 t^* 通过 算法 5 在 $B = 500$ 和 $\alpha = 0.1$条件下输出的实际百分位数。名义百分位数为 $1 - \alpha = 0.9$。所有结果均基于 200 次试验。在小样本中，自助法近似可能较为保守。

定理 5.1指出了 $\epsilon_{\mathrm{lb}}(h)$ 成为同时有效的置信界所需满足的假设。

定理 5.1 （同时有效的再生核希尔伯特空间置信界）。 假设$\mathrm{Var}(L)$ 有严格正下界， $\|L\|_\infty$ 且 $\|k(X, X)\|_\infty$ 有限， $k(\cdot, x)$ 对所有 x 连续，并且 $k(\cdot, \cdot)$ 是一个正定核。那么，

$$\lim_{n\to\infty} \mathbb{P}\left(\epsilon_{\mathrm{lb}}(h) \le \epsilon(h) \text{ for all } h \in \mathcal{H}_1^+\right) \ge 1 - \alpha.$$

常用核函数，例如高斯核和拉普拉斯核，满足定理5.1给出的假设。

5.2 实证结果

我们首先通过一项已知真实值的合成实验，验证基于再生核希尔伯特空间的置信集的覆盖保证。具体而言，我们采用与（12）相同的数据生成过程，但将 $\mathcal{X}$ 离散化如下：

$$X_i \overset{\text{iid}}{\sim} \mathrm{Unif}(\{0, 0.01, 0.02, \ldots, 1\}), \quad Y_i = \mathcal{N}(\beta_0 X_i, X_i). \tag{16}$$

在每次试验中，我们抽取 β_0 独立同分布$\sim$ $\mathcal{N}(0,\ 1)$，并使用 1000 个训练样本点（X_i，Y_i）拟合最小二乘预测器 $f(x) = \hat{\beta}^\top x$。随后，我们针对一系列带宽 $\sigma \in \{0.1, 0.5, 1\}$，对对应于高斯再生核希尔伯特空间单位球内非负函数的协变量偏移进行审计。

由于无法评估再生核希尔伯特空间中所有非负函数下的覆盖率，我们转而检验（14）的自助法近似是否有效。请回忆， t^* 估计的是$1 - \alpha$分位数：

$$\sup_{h\in\mathcal{H}_1} \{\mathbb{E}_P[h(X)] \cdot \hat{\mathbb{E}}_n[h(X)] \cdot (\hat{\epsilon}(h) - \epsilon(h))\}. \tag{17}$$

在表4中，我们计算合成实验中由 t^* 实现的（17）百分位数，即 t^* $\hat{}$$\mathbb{P}(\sup_{h\in\mathcal{H}_1}\{\mathbb{E}_P[h(X)] \cdot \mathbb{E}_n[h(X)] \cdot (\hat{}\epsilon(h) - \epsilon(h))\} \le t^*)$ 。结果表明，在小样本情况下，再生核希尔伯特空间审计可能较为保守：实际百分位数有时会高于名义水平。

我们还重新考察了 [31] 的合成数据集，以实际展示 (15) 如何为复杂子群体提供覆盖保证。图 8 展示了如何用带宽为[2] 的高斯再生核希尔伯特空间函数，近似任意选取的 3 个子区间的并集 $\sigma = 0.5$。给定包含 $n = 1000$ 个点的审计轨迹，我们给出的条件覆盖率下界为 84.0%；对于无限多个这样的倾斜，所给出的界以至少 90% 的概率同时有效。

[2]该子群并不属于我们先前考虑的子区间集合。

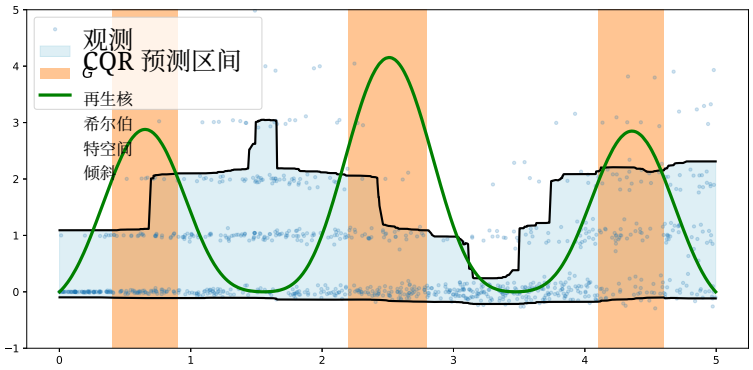


图 8：为了在三个突出显示的区间的并集上提供近似覆盖保证，我们使用图示函数对协变量分布进行倾斜，该函数属于带宽为 $\sigma = 0.5$ 的高斯再生核希尔伯特空间。对于所显示的审计轨迹，我们可以将“倾斜后”的覆盖率下界定为 84.0%。

致谢 感谢 LihuaLei、Jonathan Taylor、Isaac Gibbs、Tim Morrison 和 Anav Sood 提供宝贵的讨论。我们尤其感谢 Kevin Guo 分享他关于自助法和经验过程理论的笔记。

J.J.C. 获得 John and Fannie Hertz Foundation 的资助。E.J.C. 获得 Office of Naval Research 项目 N00014-20-1-2157、National Science Foundation 项目 DMS- 2032014、Simons Foundation 奖项 814641 以及 ARO 项目 2003514594 的资助。

参考文献

[1] Julia Angwin、Jeff Larson、Surya Mattu 和 Lauren Kirchner。机器偏见。载于数据与分析伦理，第 254–264 页。Auerbach Publications，2016年。[2] Miguel A Arcones 和 Evarist Gine。关于 U 统计量和 V 统计量的自助法。统计年刊，第 655–674 页，1992 年。[3] Yoav Benjamini 和 Daniel Yekutieli。在依赖条件下多重检验中对错误发现率的控制。统计年刊，第 1165–1188 页，2001 年。[4] Miles Brundage、Shahar Avin、Jasmine Wang、Haydn Belfield、Gretchen Krueger、Gillian Hadfield、Heidy Khlaaf、Jingying Yang、Helen Toner、Ruth Fong 等。迈向可信赖的人工智能开发：支持可验证声明的机制。*arXiv* 预印本 *arXiv:2004.07213*，2020 年。[5] Joy Buolamwini 和 Timnit Gebru。性别阴影：商业性别分类中的交叉性准确率差异。载于公平性、问责制与透明度会议，第 77–91 页。PMLR，2018 年。[6] Sam Corbett-Davies、Emma Pierson、Avi Feller、Sharad Goel 和 Aziz Huq。算法决策与公平性的代价。载于第 *23* 届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，第 797–806 页，2017 年。[7] Jeffrey Dastin。亚马逊弃用曾显示出对女性存在偏见的秘密 AI 招聘工具。载于数据与分析伦理，第 296–299 页。Auerbach Publications，2018 年。[8] Cyrus DiCiccio、Sriram Vasudevan、Kinjal Basu、Krishnaram Kenthapadi 和 Deepak Agarwal。使用置换检验评估公平性。载于第 *26* 届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，KDD '20，第 1467–1477 页，美国纽约州纽约市，2020 年。计算机协会。ISBN 9781450379984。[9] William Dieterich、Christina Mendoza 和 Tim Brennan。COMPAS 风险量表：展示准确性公平性与预测均等性。*Northpointe Inc*，7(7.4):1，2016年。[10] Frances Ding、Moritz Hardt、John Miller 和 Ludwig Schmidt。弃用 Adult：用于公平机器学习的新数据集。神经信息处理系统进展，34，2021 年。[11] John C Duchi 和 Hongseok Namkoong。通过分布鲁棒优化学习具有统一性能的模型。统计年刊，49(3):1378–1406，2021 年。[12] Richard M Dudley。经验过程课程。载于概率论研讨会 ´第 ´十二届圣弗卢尔概率论研讨会 ´（*1982*），第 1–142 页。施普林格，1984 年。[13] Cynthia Dwork、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Richard Zemel。通过认知实现公平性。载于第 *3* 届理论计算机科学创新会议论文集，ITCS '12，第 214–226 页，美国纽约州纽约市，2012 年。计算机协会。ISBN 9781450311151。doi: 10.1145/2090236.2090255。[14] Bradley Efron。自助法：重新审视刀切法。统计年刊，7(1):1–26，1979 年。

[15] William Fithian 和 Lihua Lei。依赖条件下控制错误发现率的条件校准。统计年鉴，50(6):3091–3118，2022年。[16] Sarah Flood。综合公共使用微观数据系列。当前人口调查：*4.0*版，2015年。[17] Anthony W Flores、Kristin Bechtel 和 Christopher T Lowenkamp。假阳性、假阴性和错误分析：对《机器偏见：全国各地都在使用预测软件预测未来罪犯，而该软件对黑人存在偏见》的回应。联邦缓刑，80:38，2016年。[18] Rina Foygel Barber、Emmanuel J Cand`es、Aaditya Ramdas 和 Ryan J Tibshirani。分布无关条件预测推断的局限。《信息与推断：*IMA*期刊》，10(2):455–482，2020年8月。ISSN 2049-8772。doi: 10.1093/imaiai/iaaa017。[19] Peter Gaenssler、P´eter Moln´ar 和 Daniel Rost。关于高斯过程上确界泛函的累积分布函数的连续性与严格递增性及其在统计学中的应用。数学研究成果，51:51–60，2007年。[20] Evarist Gin´e 和 Joel Zinn。经验过程的一些极限定理。概率年鉴，第929–989页，1984年。[21] Evarist Gin´e 和 Joel Zinn。一般经验测度的自助法。概率年鉴，第851–869页，1990年。[22] Moritz Hardt、Eric Price 和 Nati Srebro。监督学习中的机会平等。神经信息处理系统进展，29，2016年。[23] Ursula H´ebert-Johnson、Michael Kim、Omer Reingold 和 Guy Rothblum。多重校准：针对（计算上可识别的）群体的校准。载于国际机器学习大会，第1939–1948页。PMLR，2018年。[24] Marie Huskova 和 Paul Janssen。广义自助法对退化U统计量的一致性。统计年鉴，第1811–1823页，1993年。[25] Michael Kearns、Seth Neel、Aaron Roth 和 Zhiwei Steven Wu。防止公平性选区划分操纵：针对子群公平性的审计与学习。载于国际机器学习大会，第2564–2572页。PMLR，2018年。[26]Michael R Kosorok。经验过程与半参数推断导论。施普林格，2008年。[27] Erich Leo Lehmann、Joseph P Romano 和 George Casella。统计假设检验，第3卷。施普林格，2005年。[28] David J Marcus。Donsker类与Sobolev空间之间的关系。概率论及相关领域杂志 ¨，69(3):323–330，1985年。[29] Charles A Micchelli、Yuesheng Xu 和 Haizhang Zhang。通用核函数。机器学习研究杂志，7(12)，2006年。[30] Giulio Morina、Viktoriia Oliinyk、Julian Waton、Ines Marusic 和 Konstantinos Georgatzis。分类问题中的交叉公平性审计与实现。*arXiv*预印本 *arXiv:1911.01468*，2019年。

[31] Yaniv Romano、Evan Patterson 和 Emmanuel Candes。共形化分位数回归。神经信息处理系统进展，32:3543–3553，2019年。[32] Abhishek Roy 和 Prasant Mohapatra。公平性不确定性量化：你有多确定模型是公平的？arXiv预印本 arXiv:2110.01052，2023年。[33] Marietje Schaake 和 Jack Clark。斯坦福启动人工智能审计挑战，2022年7月。网址：https://hai.stanford.edu/news/stanford-launches-ai-audit-challenge。[34] Nian Si、Karthyek Murthy、Jose Blanchet 和 Viet Anh Nguyen。通过最优传输投影检验群体公平性。载于国际机器学习大会，第9649–9659页。PMLR，2021年。[35] Ingo Steinwart 和 Andreas Christmann。支持向量机。施普林格科学与商业媒体，2008年。[36] Bahar Taskesen、Jose Blanchet、Daniel Kuhn 和 Viet Anh Nguyen。概率公平性的统计检验。载于2021年ACM公平、问责与透明度大会论文集，第648–665页，2021年。[37]Florian Tramer、Vaggelis Atlidakis、Roxana Geambasu、Daniel Hsu、Jean-Pierre Hubaux、Mathias Humbert、Ari Juels 和 Huang Lin。Fairtest：发现数据驱动应用中的不当关联。载于2017年IEEE欧洲安全与隐私研讨会（EuroS&P），第401–416页。IEEE，2017年。[38] Aad W van der Vaart。渐近统计学，第3卷。剑桥大学出版社，2000年。[39] Aad W van der Vaart 和 Jon A Wellner。弱收敛与经验过程。施普林格，1996年。[40] Moritz von Zahn、Oliver Hinz 和 Stefan Feuerriegel。定位机器学习中的差异。arXiv预印本 arXiv:2208.06680，2022年。[41] Blake Woodworth、Suriya Gunasekar、Mesrob I Ohannessian 和 Nathan Srebro。学习非歧视性预测器。载于学习理论大会，第1920–1953页。PMLR，2017年。[42] Songkai Xue、Mikhail Yurochkin 和 Yuekai Sun。审计机器学习模型中的个体偏见与不公平性。载于国际人工智能与统计大会，第4552–4562页。PMLR，2020年。[43] Tom Yan 和 Chicheng Zhang。主动公平性审计。载于国际机器学习大会，第24929–24962页。PMLR，2022年。

A 连接到公平性

回顾一下，群体公平性定义要求 L在所有受保护的子群体中 [13, 22, 6, 25]近似均等。下面我们展示如何使用我们的方法审计两种流行的群体公平性定义。

多重校准 我们称预测器 $f(X)$ 为校准的，如果在给定 $f(X) = v$ 与 v 相匹配的条件下，二元标签 Y 的条件期望满足 $\mathbb{E}[Y \mid f(X) = v] = v$。为了衡量不同群体中的分类质量，我们可能要求二元分类器在许多子群体上均经过校准˜。因此，对于所有子群体 $G \in \mathcal{G}$ 以及（可能经过分箱的）预测值 v， γ-多重校准公平性准则 [23] 要求满足

$$|\mathbb{E}[Y \mid f(X) = v, X \in G] - v| \leq \gamma.$$

正式的多重校准定义排除了规模小于某个审计员设定阈值的群体。˜尽管用户可以根据自己的判断筛选 $\mathcal{G}$，但先验地说，我们的方法不会根据群体规模排除群体。不过，我们可以在重缩放认证方法（算法 2）中设置一个类似的阈值。该阈值不会阻止审计员查询任何群体以获取关于 γ的界限，但它可以确保阈值以上群体的置信集更窄。

仔细选择 L， θ_P 和 $\mathcal{G}$，我们可以应用这些方法来认证并标记多重校准˜在 $\mathcal{G}$ 上的情况。令 $\mathcal{V}$ 表示 $f(X)$ 可能取到的唯一值集合，我们设定

$$L := Y - f(X) \qquad \theta_P := 0$$
$$\mathcal{G} := \{G \cap \{(X, Y) \mid f(X) = v\} \mid G \in \tilde{\mathcal{G}}, v \in \mathcal{V}\}.$$

令 G_v 表示 $\mathcal{G}$中的每个群体。

对 G 进行 γ-多重校准认证，等价于证明 $\max_v |\epsilon(G_v)| \leq \gamma$。使用我们的审计方法，我们可以通过为所有 $\{\epsilon(G_v)\}_{G_v \in \mathcal{G}}$ 构造同时有效的置信区间来估计 γ。对于特定的 G，我们对 γ 的界等于针对 $\{\epsilon(G_v)\}_{v \in \mathcal{V}}$ 所发布的所有区间中的最大（绝对）值。对于固定阈值 γ，我们也可以使用 $H_0(G_v) : |\epsilon(G_v)| \geq \gamma$ 运行布尔认证方法；如果所有 $\{G_v\}_{v \in \mathcal{V}}$ 的原假设都被拒绝，则认证 G。

或者，我们也可以通过检验 G 是否违反 γ-多重校准，检验 $H_0(G_v)$: $|\epsilon(G_v)| > \gamma + \epsilon$，差异容忍度为 ϵ，从而对其进行标记。使用第 4 节概述的方法，我们可以为每个原假设构造 p 值。然后，对计算得到的所有 p 值执行 BH 程序。如果任何 $H_0(G_v)$ 被拒绝，我们就标记 G。

均等赔率 给定一组子集 $\mathcal{X}$记为 $\mathcal{G}_X$，我们称一个二元预测器f 满足均等赔率准则 [22, 41] 如果对于所有 $G \in \mathcal{G}_X$，其真阳性率均相等，

$$\mathbb{P}(f(X) = 1 \mid Y = 1, X \in G) = \mathbb{P}(f(X) = 1 \mid Y = 1),$$

且其假阳性率已实现均等化，

$$\mathbb{P}(f(X) = 1 \mid Y = 0, X \in G) = \mathbb{P}(f(X) = 1 \mid Y = 0).$$

关注这一公平性标准的从业者，可能希望审计性能差异 $\epsilon(G) := \max(|\epsilon(G_0)|, |\epsilon(G_1)|)$，其中

$$\epsilon(G_i) := \mathbb{P}(f(X) = 1 \mid Y = i, X \in G) - \mathbb{P}(f(X) = 1 \mid Y = i).$$

我们可以通过运行两次方法，对 $\epsilon(G)$ 进行认证审计：第一次用于界定$\epsilon(G_0)$，第二次用于界定 $\epsilon(G_1)$。对于第一次审计，令

$$L := \mathbf{1}\{f(x)=1\} \qquad \theta_P := \mathbb{P}(f(X)=1 \mid Y=0) \qquad \mathcal{G} = \{G \times \{0\} \mid G \in \mathcal{G}_X\}.$$

对于第二次审计，令

$$L := \mathbf{1}\{f(x)=1\} \qquad \theta_P := \mathbb{P}(f(X)=1 \mid Y=1) \qquad \mathcal{G} = \{G \times \{1\} \mid G \in \mathcal{G}_X\}.$$

对于每项审计任务，我们使用名义第一类错误阈值 $\alpha/2$来构造置信集或布尔认证。我们指出，由于与这两次审计对应的两个数据集互不相交，因此这个并集界在实际中几乎是紧的。于是，我们关于 ϵ（G）的最终证书由两个审计员输出中“较差”的那个构成。例如，我们可以将 ϵ（G）的上界设为针对 ϵ（G_0）和 ϵ（G_1）所给出的两个置信区间中包含的最大（绝对）值。

对于标记任务，我们检验 H_0（G_i）： $|\epsilon$（G_i）$| \le \epsilon$。使用第 4 节所述的方法，我们可以为这些原假设分别构造 p 值。随后，可以直接对计算出的所有 p 值执行 BH 程序。如果 G 满足以下任一条件，我们就将其标记：拒绝 H_0（G_0），或拒绝 H_0（G_1）。

个体公平性 其他公平性标准属于一类称为“个体公平性”度量的标准。这类度量量化了一种直觉：相似的输入 x 和 x'应由 f以相似的方式 [13]处理。尽管无法利用抽样得到的模型预测来检验这一公平性定义，但我们可以审计某种个体公平性是否以较高概率适用于受保护子群体。这里不再详细讨论这种扩展，但我们指出，关于 U 统计量及其过程的自助法的既有研究，使我们的审计程序能够自然地扩展至基于数据点对定义的公平性度量 [2, 24]。

B 认证审计

B.1 符号与回顾

我们首先定义一些相关符号，并回顾随机过程收敛的一些基本结果。给定 n 独立同分布样本， $\mathbb{P}_n := n^{-1}\sum_{i=1}^n \delta_{(X_i,Y_i)}$ 表示它们的经验分布。如果 $\{(X_i^*, Y_i^*)\}_{i=1}^n$ 是来自给定 $\{(X_i, Y_i)\}_{i=1}^n$条件下的 $\mathbb{P}_n$ 的独立同分布样本，那么 $\mathbb{P}_n^* := n^{-1}\sum_{i=1}^n \delta_{(X_i^*,Y_i^*)}$ 表示它们的经验分布。

对于函数 f: $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$, $P[f]$是 $\mathbb{E}_P[f(X,Y)]$的简写， $\mathbb{P}_n[f]$是$n^{-1}\sum_{i=1}^n f(X_i,Y_i)$的简写，而 $\mathbb{P}_n^*[f]$是 $n^{-1}\sum_{i=1}^n f(X_i^*,Y_i^*)$的简写。我们也用 $(\mathbb{P}_n - P)[f]$代替 $\mathbb{P}_n[f] - P[f]$。给定一个函数类 $\mathcal{F}$，我们将 $f \mapsto \sqrt{n}(\mathbb{P}_n - P)[f]$视为属于 $\ell_\infty(\mathcal{F})$的映射。

我们通常会要求和/或证明 $\mathcal{F}$ 是 P-Donsker类。这意味着以 $f \in \mathcal{F}$ 为指标的经验过程依分布收敛于 $\ell_\infty(\mathcal{F})$中的紧高斯极限。形式上， $\sqrt{n}(\mathbb{P}_n - P)[\cdot] \xrightarrow{d} \mathbb{G}[\cdot]$，其中极限过程 $f \mapsto \mathbb{G}[f]$是一个高斯过程，同时也是 $\ell_\infty(\mathcal{F})$中的紧Borel可测元素。如果 $\mathcal{F}$是 P-Donsker类，那么它也是 P-Glivenko-Cantelli [38],，即 $\sup_{f\in\mathcal{F}} |(\mathbb{P}_n - P)[f]| \xrightarrow{p} 0$。

接下来，我们回顾一些关于 Donsker类与 VC 维数以及自助法一致性的相关结果。当我们考虑由子总体 $\mathcal{G}$ 索引的示性函数类，即 $\mathcal{F} = \{\mathbf{1}\{(X, Y) \in G\} \mid G \in \mathcal{G}\}$ 时，Donsker 性质等价于假设 $\mathcal{G}$ 是 VC 类。

引理 B.1 （ [12]中的定理11.4.1）。 在适当的可测性假设下， $\mathcal{F} := \{(x、y) \mapsto \mathbf{1}\{(x、y) \in G\} : G \in \mathcal{G}\}$ 是 *Donsker* 的，当且仅当 *VC* （$\mathcal{G}$） $< \infty$。

最后，我们回顾一下：如果 $\mathcal{H}$ 是 P-Donsker 的，则自助法近似 $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)[h]$是有效的。令 $\mathbb{G}_n^*$ 表示自助经验过程 $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)$，并令 $\mathbb{G}$ 表示极限经验过程。

引理 B.2 ([38] 中的定理 23.7。对于每个由具有有限包络函数 F 的可测函数组成的 *Donsker*类 $\mathcal{H}$，$\sup_{g\in BL_1(\ell^\infty(\mathcal{H}))} |\mathbb{E}^*[g(\mathbb{G}_n^*)] - \mathbb{E}[g(\mathbb{G})]| \xrightarrow{p} 0$ 其中 $BL_1(\ell_\infty(\mathcal{H}))$ 是取值于 $\mathbb{R}$、定义在 $\ell_\infty(\mathcal{H})$ 上的有界 *1-Lipschitz* 函数集合。

B.2 认证定理的证明

ˆ估计 θ_P。 在正文中，我们假设 θ 是样本均值的可微函数。这里，我们将推广这一假设，以适用于其他可以通过自助法进行估计的估计量。

我们假设近似性和抽样分布的彩票论坛

$$\sqrt{n}(\hat{\theta}(\mathcal{D}^*) - \hat{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i^*, Y_i^*) + o_{P_n}(1)$$
$$\sqrt{n}(\hat{\theta} - \theta_P) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i, Y_i) + o_P(1).$$

这里， ψ 是一个均值为零且具有有限方差的影响函数。很容易验证，正文中给出的 θ 满足这一条件。我们的推广使得即使θ 更为复杂，也能够进行审计，例如，θ 是某个最大似然估计问题的解。

接下来，我们证明第 3 节的主要定理。如正文所述，主要的技术难点在于证明 t^* 能够一致地估计某个随机过程的（$1-\alpha$）分位数。要确立这一结果，需要两个关键的预备结果。首先，我们必须证明自助法一致性意味着分位数估计一致性。其次，我们必须证明所提出的重缩放估计量 ˆ$s(G)$ 对某个目标估计量是一致的。

分位数一致性 首先，为了对分位数作出唯一的定义，我们将随机变量 X 的 α-分位数定义为

$$\text{Quantile}(\alpha; X) := \inf_x\{x \mid \alpha \le \mathbb{P}(X \le x)\}.$$

主要定理的证明要求极限过程的上确界在其 $(1-\alpha)$ 分位数处连续且严格递增。以下结果给出了满足这一性质的温和条件。

引理 B.3。 设 $\mathbb{Z}(G)$ 表示某个高斯过程，它是以某个可数类 $\mathcal{G}$为指标的经验过程的极限。那么，如果 $\text{Var}(\mathbb{Z}(G)) > 0$ 对某个 $G \in \mathcal{G}$成立，则 $\sup_{G\in\mathcal{G}}\mathbb{Z}(G)$ 的分布函数在 $\mathbb{R}_+$ 上连续且严格递增，而 $\inf_{G\in\mathcal{G}}\mathbb{Z}(G)$ 的分布函数在 $\mathbb{R}_-$上连续且严格递增。

证明：由于 $\mathbb{Z}(G)$ 是一个中心化过程，我们注意到 $\sup_G \mathbb{Z}(G) \overset{d}{=} -\inf_G \mathbb{Z}(G)$。首先，我们证明对于所有 $x > 0$，$\sup_{G\in\mathcal{G}}\mathbb{Z}(G)$ 的分布函数在 x 处连续。在所述假设下，Gaenssler 等人 [19, Corollary 1.3] 证明了 $\sup_G|\mathbb{Z}(G)|$ 在 x 处具有连续的分布函数，即 $\mathbb{P}(\sup_G|\mathbb{Z}(G)| = x) = 0$。接下来，

$$0 = \mathbb{P}(\sup_G|\mathbb{Z}(G)| = x) = \mathbb{P}(\{\sup_G \mathbb{Z}(G) = x\} \cup \{\inf_G \mathbb{Z}(G) = -x\}),$$

这意味着 $\mathbb{P}(\sup_G \mathbb{Z}(G) = x) = \mathbb{P}(\inf_G \mathbb{Z}(G) = -x) = 0$。这等价于所需的连续性结论。

接下来，为了证明在 x 处严格递增，我们采用反证法。若 $\sup_G \mathbb{Z}(G)$ 在 $\mathbb{R}_+$ 上的分布函数不是严格递增的，则存在 $0 < x_1 < x_2$，使得

$$\mathbb{P}(\sup_G \mathbb{Z}(G) \le x_1) = \mathbb{P}(\sup_G \mathbb{Z}(G) \le x_2) \Leftrightarrow \mathbb{P}(\inf_G \mathbb{Z}(G) \ge -x_1) = \mathbb{P}(\inf_G \mathbb{Z}(G) \ge -x_2).$$

Gaenssler 等人 [19, 推论 1。3] 还证明了 $\sup_G |\mathbb{Z}(G)|$ 满足 $\mathbb{P}(\sup_G |\mathbb{Z}(G)| \le x_1) < \mathbb{P}(\sup_G |\mathbb{Z}(G)| \le x_2$。由于

$$\mathbb{P}(\sup_G |\mathbb{Z}(G)| \le x) = \mathbb{P}(\sup_G \mathbb{Z}(G) \le x) + \mathbb{P}(\inf_G \mathbb{Z}(G) \ge -x) - \mathbb{P}(\{\sup_G \mathbb{Z}(G) \le x\} \cap \{\inf_G \mathbb{Z}(G) \ge -x\}),$$

由此我们得出结论：

$$\mathbb{P}(\{\sup_G \mathbb{Z}(G) \le x_1\} \cap \{\inf_G \mathbb{Z}(G) \ge -x_1\}) > \mathbb{P}(\{\sup_G \mathbb{Z}(G) \le x_2\} \cap \{\inf_G \mathbb{Z}(G) \ge -x_2\}).$$

这是一个矛盾，因为左侧（LHS）的事件是右侧（RHS）事件的子集。由此我们得出结论：$\sup_G \mathbb{Z}(G)$ 和 $\inf_G \mathbb{Z}(G)$ 的分布函数分别在正实数和负实数上严格递增。

□

我们对引理 B.3 作出两点说明。首先，Gaenssler 等人 [19] 所引用的结果是针对由 VC 类索引的经验过程给出的，但他们的论证同样适用于由 P-Donsker 类索引的任意极限高斯过程。其次，可数性假设并非关键，它仅仅是为了避免验证某些技术性的可测性条件。在本文中，我们预期 van der Vaart 和 Wellner [39] 所提出的“逐点可测性”（大致等价于能够通过一个稠密的可数子集进行良好逼近）条件成立，但这一条件必须针对每个函数类分别进行检验。因此，为了简化论述，我们还将在全文中假设所研究的函数类均为可数类。

一致相合性 $\hat{s}$ 回想一下，我们定义

$$\hat{s}(G) := \max(\mathbb{P}_n(G), p_*)^{3/2} \cdot \hat{\sigma}(G, w_0)$$

for

$$\hat{\sigma}(G, w_0) := \left(\frac{\mathbb{P}_n(G)}{\mathbb{P}_n(G) + w_0}\right) \cdot \hat{\sigma}_G + \left(\frac{w_0}{\mathbb{P}_n(G) + w_0}\right) \cdot \sqrt{\widehat{\mathrm{Var}}(L)}$$

and

$$\hat{\sigma}_G^2 := \widehat{\mathrm{Var}}(L \mid G) + \mathbb{P}_n(G) \left(\widehat{\mathrm{Var}}(\psi) - 2 \cdot \widehat{\mathrm{Cov}}(L, \psi \mid G)\right).$$

引理 B.4。 假设 $0 < \mathrm{Var}(L) < \infty$ 且 $VC(\mathcal{G}) < \infty$。若 $p_*, w_0 > 0$，则 $\sup_{G \in \mathcal{G}} |1/\hat{s}(G) - 1/s(G)| \xrightarrow{p} 0$，其中

$$s(G) = \max(\mathbb{P}(G), p_*)^{3/2} \cdot \left[\left(\frac{\mathbb{P}(G)}{\mathbb{P}(G) + w_0}\right) \cdot \sigma_G + \left(\frac{w_0}{\mathbb{P}(G) + w_0}\right) \cdot \sqrt{\mathrm{Var}(L)}\right].$$

Pr哎为简化记号，我们将在全文中用 $\mathbf{1}_G$ 替换 $\mathbf{1}\{(X,Y) \in G\}$ with t.

我们首先通过线性化该随机量，计算 $\sqrt{|G|}(\hat{\epsilon}(G) - \epsilon(G))$ 的渐近方差：

$$\begin{aligned}
\sqrt{|G|}\left(\hat{\epsilon}(G) - \epsilon(G)\right) &= \sqrt{|G|}\left[\left(\frac{1}{|G|}\sum_{(X_i,Y_i)\in G} L_i - \hat{\theta}\right) - \left(\mathbb{E}[L \mid G] - \theta_P\right)\right] \\
&= \sqrt{|G|}\left[\frac{1}{|G|}\sum_{i=1}^{n}(L_i - \mathbb{E}[L \mid G]) \cdot \mathbf{1}_G - \frac{1}{n}\sum_{i=1}^{n}\psi_i\right] \\
&= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{L_i - \mathbb{E}[L \mid G]}{\sqrt{\mathbb{P}(G)}} \cdot \mathbf{1}_G - \sqrt{\mathbb{P}(G)} \cdot \psi_i\right] + o_P(1).
\end{aligned}$$

因此，我们得出结论：

$$\begin{aligned}
\sigma_G^2 &:= \mathrm{Var}\left(\frac{L - \mathbb{E}[L \mid G]}{\sqrt{\mathbb{P}(G)}} \cdot \mathbf{1}_G - \sqrt{\mathbb{P}(G)} \cdot \psi\right) \\
&= \mathrm{Var}\left(\frac{L - \mathbb{E}[L \mid G]}{\sqrt{\mathbb{P}(G)}} \cdot \mathbf{1}_G\right) + \mathrm{Var}\left(\sqrt{\mathbb{P}(G)} \cdot \psi\right) - 2 \cdot \mathrm{Cov}\left(\frac{L - \mathbb{E}[L \mid G]}{\sqrt{\mathbb{P}(G)}} \cdot \mathbf{1}_G, \sqrt{\mathbb{P}(G)} \cdot \psi\right) \\
&= \mathrm{Var}(L \mid G) + \mathbb{P}(G) \cdot \left(\mathrm{Var}(\psi) - 2 \cdot \mathrm{Cov}(L, \psi \mid G)\right).
\end{aligned}$$

接下来，经过一些重新整理，可以观察到 $1/\hat{s}^2(G) - 1/s^2(G)$ 等于

$$\frac{\mathbb{P}_n(G) + w_0}{\max(\mathbb{P}_n(G), p_*)^3 \cdot \left(\mathbb{P}_n(G) \cdot \hat{\sigma}_G + w_0 \cdot \sqrt{\widehat{\mathrm{Var}}(L)}\right)} - \frac{\mathbb{P}(G) + w_0}{\max(\mathbb{P}(G), p_*)^{3/2} \cdot \left(\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}\right)}.$$

合并这两个分数得到的分子为

$$\begin{aligned}
&(\mathbb{P}_n(G) + w_0)\left(\max(\mathbb{P}(G), p_*)^{3/2} \cdot \left(\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}\right)\right) \\
&\qquad - (\mathbb{P}(G) + w_0)\left(\max(\mathbb{P}_n(G), p_*)^{3/2} \cdot \left(\mathbb{P}_n(G) \cdot \hat{\sigma}_G + w_0 \cdot \sqrt{\widehat{\mathrm{Var}}(L)}\right)\right)
\end{aligned}$$

以及一个分母为

$$\begin{aligned}
&\max(\mathbb{P}_n(G), p_*)^{3/2} \cdot \max(\mathbb{P}(G), p_*)^{3/2} \cdot \left(\mathbb{P}_n(G) \cdot \hat{\sigma}_G + w_0 \cdot \sqrt{\widehat{\mathrm{Var}}(L)}\right) \\
&\qquad \cdot \left(\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}\right).
\end{aligned}$$

为证明一致相合性，只需证明

$$\begin{aligned}
&\lim_{n\to\infty} \inf_{G\in\mathcal{G}} \max(\mathbb{P}_n(G), p_*)^{3/2} \cdot \max(\mathbb{P}(G), p_*)^{3/2} \cdot \left(\mathbb{P}_n(G) \cdot \hat{\sigma}_G + w_0 \cdot \sqrt{\widehat{\mathrm{Var}}(L)}\right) \\
&\qquad \cdot \left(\mathbb{P}(G) \cdot \sigma_G + w_0 \cdot \sqrt{\mathrm{Var}(L)}\right) > 0
\end{aligned}$$

and

$$\lim_{n\to\infty}\sup_{G\in\mathcal{G}}\Big|(\mathbb{P}_n(G)+w_0)\left(\max(\mathbb{P}(G),p_*)^{3/2}\cdot\left(\mathbb{P}(G)\cdot\sigma_G+w_0\cdot\sqrt{\mathrm{Var}(L)}\right)\right)$$
$$-(\mathbb{P}(G)+w_0)\left(\max(\mathbb{P}_n(G),p_*)^{3/2}\cdot\left(\mathbb{P}_n(G)\cdot\hat{\sigma}_G+w_0\cdot\sqrt{\widehat{\mathrm{Var}}(L)}\right)\right)\Big|=0.$$

对于这些任务中的第一项，注意到分母的下界为

$$p_*^3\cdot\left(w_0^2\cdot\sqrt{\widehat{\mathrm{Var}}(L)}\cdot\sqrt{\mathrm{Var}(L)}\right)\to p_*^3\cdot w_0^2\cdot\mathrm{Var}(L)>0.$$

分子需要更仔细的分析。我们分配 $\mathbb{P}_n(G)+w_0$ 因子，应用三角不等式，并分析分子中的第一项：

$$\Big|\mathbb{P}_n(G)\cdot\max(\mathbb{P}(G),p_*)^{3/2}\cdot\left(\mathbb{P}(G)\cdot\sigma_G+w_0\cdot\sqrt{\mathrm{Var}(L)}\right)$$
$$-\ \mathbb{P}(G)\cdot\max(\mathbb{P}_n(G),p_*)^{3/2}\cdot\left(\mathbb{P}_n(G)\cdot\hat{\sigma}_G+w_0\cdot\sqrt{\widehat{\mathrm{Var}}(L)}\right)\Big|.$$

加上并减去 $\mathbb{P}(G)\cdot\max(\mathbb{P}(G)\ p_*)^{3/2}\cdot(\mathbb{P}(G)\ \cdot\sigma_G+w_0\cdot\sqrt{\mathrm{Var}(L)})$，并应用三角不等式，我们得到：

$$|\mathbb{P}_n(G)-\mathbb{P}(G)|\cdot\max(\mathbb{P}(G),p_*)^{3/2}\cdot(\mathbb{P}(G)\cdot\sigma_G+w_0\cdot\sqrt{\mathrm{Var}(L)})$$
$$+\left|\max(\mathbb{P}(G),p_*)^{3/2}-\max(\mathbb{P}_n(G),p_*)^{3/2}\right|\cdot\mathbb{P}(G)\cdot(\mathbb{P}(G)\cdot\sigma_G+w_0\cdot\sqrt{\mathrm{Var}(L)})$$
$$+\mathbb{P}(G)\cdot\max(\mathbb{P}_n(G),p_*)^{3/2}\cdot|\mathbb{P}(G)\cdot\sigma_G-\mathbb{P}_n(G)\cdot\hat{\sigma}_G|$$
$$+\mathbb{P}(G)\cdot\max(\mathbb{P}_n(G),p_*)^{3/2}\left|w_0\cdot(\sqrt{\mathrm{Var}(L)}-\sqrt{\widehat{\mathrm{Var}}(L)})\right|.$$

由于 L 的样本方差不依赖于 G，且在给定假设下是一致的，因此很容易看出第四项一致收敛于0。为了证明第一项和第二项一致收敛，我们必须证明该 n-独立项有界，即，

$$\begin{aligned}(\mathbb{P}(G)\cdot\sigma_G+w_0\cdot\sqrt{\mathrm{Var}(L)})&\le\mathbb{P}(G)^{1/2}\sqrt{\mathbb{E}[(L-\mathbb{E}[L\mid G])^2\cdot\mathbf{1}_G]}+w_0\cdot\sqrt{\mathrm{Var}(L)})\\&\le\mathbb{P}(G)^{1/2}\sqrt{\mathbb{E}[(L-\mathbb{E}[L])^2\cdot\mathbf{1}_G]}+w_0\cdot\sqrt{\mathrm{Var}(L)})\\&\le\mathbb{P}(G)^{1/2}\sqrt{\mathbb{E}[(L-\mathbb{E}[L])^2]}+w_0\cdot\sqrt{\mathrm{Var}(L)})\\&\le(1+w_0)\cdot\sqrt{\mathrm{Var}(L)}=C<\infty.\end{aligned}$$

因此，由于 $\mathrm{VC}(\mathcal{G})<\infty$，$(\mathbb{P}_n-P)[\mathbf{1}_G]$ 是 P-Glivenko-Cantelli，且

$$\sup_{G\in\mathcal{G}}\Big|(\mathbb{P}_n(G)-\mathbb{P}(G))\cdot\max(\mathbb{P}(G),p_*)^{3/2}\cdot(\mathbb{P}(G)\cdot\sigma_G+w_0\cdot\sqrt{\mathrm{Var}(L)})\Big|$$
$$\le C\cdot\sup_{G\in\mathcal{G}}|\mathbb{P}_n(G)-\mathbb{P}(G)|\xrightarrow{p}0.$$

由于 P-Glivenko Cantelli 类在截断下保持不变（ [39]中的例2.10.11），因此我们还可以得出

$$\sup_{G\in\mathcal{G}}|\max(\mathbb{P}(G),p_*)^{3/2}-\max(\mathbb{P}_n(G),p_*)^{3/2}|\cdot\mathbb{P}(G)\cdot\left(\mathbb{P}(G)\cdot\sigma_G+w_0\cdot\sqrt{\mathrm{Var}(L)}\right)$$
$$\le C\cdot\sup_{G\in\mathcal{G}}|\max(\mathbb{P}(G),p_*)^{3/2}-\max(\mathbb{P}_n(G),p_*)^{3/2}|\xrightarrow{p}0.$$

最后，我们证明第三项收敛于 0。与 n 独立的部分有界于 1，因此我们可以忽略它。因此，我们需要证明 $\sup_{G\in\mathcal{G}} |\mathbb{P}(G) \cdot \sigma_G - \mathbb{P}_n(G) \cdot \hat{}\sigma_G| \xrightarrow{p} 0$。为避免书写平方根，我们暂时将这两项平方。然后，

$$\begin{aligned}
\mathbb{P}(G)^2 \cdot \sigma_G^2 &= \mathbb{P}(G)^2 \left[\mathrm{Var}(L \mid G) + \mathbb{P}(G) \cdot (\mathrm{Var}(\psi) - 2 \cdot \mathrm{Cov}(L, \psi \mid G))\right] \\
&= \mathbb{P}(G)^2(\mathbb{E}[L^2 \mid G] - \mathbb{E}[L \mid G]^2) \\
&\quad + \mathbb{P}(G)^3 \cdot (\mathrm{Var}(\psi) - 2 \cdot (\mathbb{E}[L \cdot \psi \mid G] - \mathbb{E}[L \mid G]\mathbb{E}[\psi \mid G])) \\
&= \mathbb{P}(G)\mathbb{E}[L^2 \cdot \mathbf{1}_G] - \mathbb{E}[L \cdot \mathbf{1}_G]^2 + \mathbb{P}(G)^3 \cdot \mathrm{Var}(\psi) \\
&\quad - 2 \cdot \mathbb{P}(G)^2 \cdot \mathbb{E}[(L \cdot \psi)\mathbf{1}_G] \\
&\quad - 2 \cdot \mathbb{P}(G) \cdot \mathbb{E}[L \cdot \mathbf{1}_G]\mathbb{E}[\psi \cdot \mathbf{1}_G]
\end{aligned}$$

我们可以得到一个类似的 $\mathbb{P}_n(G)^2 \hat{}\sigma_G^2$的展开式。接下来，我们需要证明每一对对应项都一致收敛。我们将仅详细推导其中一对项的论证，其余项的论证应当是显而易见的。首先，注意到 $\mathcal{F}_g = \{g \cdot \mathbf{1}_G : G \in \mathcal{G}\}$ 是一个P-Glivenko-Cantelli 类，只要满足 $P[|g|] < \infty$ (推论 3，见 [20]) 。然后，

$$\begin{aligned}
&\sup_{G\in\mathcal{G}} \left|\mathbb{P}_n(G)^2 \cdot \mathbb{P}_n[(L \cdot \psi) \cdot \mathbf{1}_G] - \mathbb{P}(G)^2 \cdot P[(L \cdot \psi) \cdot \mathbf{1}\{X, Y) \in G\}]\right| \\
&\qquad \leq \sup_{G\in\mathcal{G}} \left|\mathbb{P}_n(G)^2 \cdot (\mathbb{P}_n[(L \cdot \psi) \cdot \mathbf{1}_G] - P[(L \cdot \psi) \cdot \mathbf{1}_G])\right| \\
&\qquad\qquad + \sup_{G\in\mathcal{G}} \left|(\mathbb{P}_n(G)^2 - \mathbb{P}(G)^2) \cdot P[(L \cdot \psi) \cdot \mathbf{1}_G]\right|
\end{aligned}$$

我们还可以进一步将右侧（RHS）上界化为

$$\underbrace{\sup_{G\in\mathcal{G}} |(\mathbb{P}_n - P)[(L \cdot \psi) \cdot \mathbf{1}_G]|}_{o_P(1)} + 2 \cdot \underbrace{P[|L \cdot \psi|]}_{\leq P[L^2]P[\psi^2] < \infty} \cdot \underbrace{\sup_{G\in\mathcal{G}} |\mathbb{P}_n(G) - \mathbb{P}(G)|}_{o_P(1)}.$$

对于该表达式中的其他各项，也可以作出类似的论证。

上述推导仅针对通过乘以 $\mathbb{P}_n(G)$ 得到的项。对于另一项，证明其一致收敛至 0 的过程基本相同，因此我们不再重复这一推导。由此，我们得出结论：分子在所有 G 上一致收敛至 0。这便得到了所声称的 $1/\hat{}s(G)$ 的一致相合性。

□

界认证 首先，我们证明所构造的下置信界是有效的。算法6 重新阐述了定义临界值的方法 t^*。然后，利用算法 6的输出，我们通过设置 $G \in \mathcal{G}$ ，为任意

$$\epsilon_{\mathrm{lb}}(G) = \hat{\epsilon}(G) - t^* \cdot \frac{\hat{s}(G)}{\mathbb{P}_n(G)^2}.$$

定理B.1 重新阐述了定理3.1。

˙定理 B.1。 假设 L 有界，且 $L - \theta$ 在至少一个非空群体上非常数。进一步假设 $\mathcal{G}$ 具有有限的 *Vapnik-Chernovenkis (VC)* 维数。那么，对于所有 $G \in \mathcal{G}$， $\epsilon_{\mathrm{lb}}(G)$是 $\epsilon(G)$ 的渐近 $1-\alpha$)-下置信界，并且同时有效。

Al算法 6 （重缩放）下置信界临界值的自助法

1: **输入：** 子群体 $\mathcal{G}$，审计轨迹 $\mathcal{D}$，水平 α，阈值 p_*，权重 w_0，自助样本数量 B

2：对于 $b = 1, \ldots, B$，执行 3：令 $\mathcal{D}_b^*$ 为从 $\mathcal{D}$ 中进行的大小为 n 的有放回抽样；4：定义 $\mathbb{P}_b^*(G) := \frac{1}{n} \sum_{(x_i^*, y_i^*) \in \mathcal{D}_b^*} \mathbf{1}\{(x_i^*, y_i^*) \in G\}$；5：定义 $\epsilon_b^*(G) := \frac{1}{\mathbb{P}_b^*(G) \cdot n} \sum_{(x_i^*, y_i^*) \in G} L_i^* - \hat{\theta}(\mathcal{D}_b^*)$；6：结束循环 7：将渐近方差估计量定义为

$$\hat{\sigma}_G^2 := \widehat{\mathrm{Var}}(L \mid G) + \mathbb{P}_n(G) \left(\widehat{\mathrm{Var}}(\psi) - 2 \cdot \widehat{\mathrm{Cov}}(L, \psi \mid G) \right)$$

其中 $\widehat{\mathrm{Var}}(\cdot)$ 和 $\widehat{\mathrm{Cov}}(\cdot)$ 分别表示样本（条件）方差和协方差；8：按 (9) 定义 $\hat{s}(G)$；9：对于 $b = 1, \ldots, B$ 执行 10：$t^{(b)} = \max_{G \in \mathcal{G}} \{ \frac{1}{\hat{s}(G) \cdot \mathbb{P}_n(G) \cdot \mathbb{P}_b^*(G)} \cdot (\epsilon_b^*(G) - \hat{\epsilon}(G)) \}$；11：结束循环 12：返回：$t^* = \text{分位数}(1 - \alpha; \{t^{(b)}\}_{b=1}^B)$

证明为简化记号，我们将全文中的 $\mathbf{1}\{(X, Y) \in G\}$ 替换为 $\mathbf{1}_G$。首先重述我们要证明的结论。

$$\lim_{n \to \infty} \mathbb{P}\left(\exists G \in \mathcal{G} \text{ s.t. } \epsilon_{\mathrm{lb}}(G) > \epsilon(G)\right) = \lim_{n \to \infty} \mathbb{P}\left(\exists G \in \mathcal{G} \text{ s.t. } \hat{\epsilon}(G) - t^* \cdot \frac{s^*(G)}{\mathbb{P}_n(G)^2} > \epsilon(G)\right)$$

重新整理后一个事件，可得

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{G \in \mathcal{G}} \left\{ \left(\frac{\mathbb{P}_n(G)^2}{\hat{s}(G)} \right) \cdot (\hat{\epsilon}(G) - \epsilon(G)) \right\} > t^*\right)$$
$$= \lim_{n \to \infty} \mathbb{P}\left(\sup_{G \in \mathcal{G}} \left\{ \left(\frac{\mathbb{P}_n(G)}{\hat{s}(G)} \right) \cdot \mathbb{P}_n[(L - \hat{\theta} - \epsilon(G)) \cdot \mathbf{1}_G] \right\} > t^*\right).$$

接着，我们断言 $\sup_{G \in \mathcal{G}} |\mathbb{P}_n(G)/\hat{s}(G) - \mathbb{P}(G)/s(G)| \xrightarrow{p} 0$。注意到

$$\sup_{G \in \mathcal{G}} \left| \frac{\mathbb{P}_n(G)}{\hat{s}(G)} - \frac{\mathbb{P}(G)}{s(G)} \right| \le \sup_{G \in \mathcal{G}} \left| \frac{\mathbb{P}_n(G)}{\hat{s}(G)} - \frac{\mathbb{P}_n(G)}{s(G)} \right| + \sup_{G \in \mathcal{G}} \left| \frac{\mathbb{P}_n(G)}{s(G)} - \frac{\mathbb{P}(G)}{s(G)} \right|$$
$$\le \sup_{G \in \mathcal{G}} |\mathbb{P}_n(G)| \cdot \sup_{G \in \mathcal{G}} \left| \frac{1}{\hat{s}(G)} - \frac{1}{s(G)} \right| + \left| \frac{1}{\inf_{G \in \mathcal{G}} s(G)} \right| \cdot \sup_{G \in \mathcal{G}} |\mathbb{P}_n(G) - \mathbb{P}(G)|.$$

注意，$\sup_{G \in \mathcal{G}} |\mathbb{P}_n(G)| \le 1$ 和 $\inf_{G \in \mathcal{G}} s(G) \ge p_*^{3/2} \cdot (w_0/(1 + w_0)) \cdot \sqrt{\mathrm{Var}(L)} > 0$。然后，根据引理B.4 以及 $\mathcal{G}$ 是 P-Donsker，因此也是 P-Glivenko-Cantelli这一假设，我们得出所需的一致相合性结论。

为了应用 Slut ky 引理应用于一致估计量时，我们必须证明 $\sqrt{n} \cdot \mathbb{P}_n[(L -$

$\hat{\theta} - \epsilon(G)) \cdot \mathbf{1}_G] = O_P(1)$。请注意，

$$\begin{aligned}
&\left|\sup_{G\in\mathcal{G}} \sqrt{n} \cdot \mathbb{P}_n[(L - \hat{\theta} - \epsilon(G)) \cdot \mathbf{1}_G]\right| \\
&\quad = \left|\sup_{G\in\mathcal{G}} \left\{\sqrt{n} \cdot (\mathbb{P}_n - P)[(L - \theta_P - \epsilon(G)) \cdot \mathbf{1}_G] - \sqrt{n}(\hat{\theta} - \theta_P) \cdot \mathbb{P}_n[\mathbf{1}_G]\right\}\right| \\
&\quad \leq \left|\sup_{G\in\mathcal{G}} \sqrt{n}(\mathbb{P}_n - P)[(L - \theta_P - \epsilon(G)) \cdot \mathbf{1}_G]\right| + \left|\sqrt{n}(\hat{\theta} - \theta_P)\right| \cdot \sup_{G\in\mathcal{G}} \mathbb{P}_n(G).
\end{aligned}$$

根据标准的 Donsker 保持结果 [39, Section 2.10],，函数类 $\{(L - \theta_P - \epsilon(G)) \cdot \mathbf{1}_G \mid G \in \mathcal{G}\}$ 是 P-Donsker 类，因此 $\sup_{G\in\mathcal{G}} \sqrt{n}(\mathbb{P}_n - P)[(L - \theta_P - \epsilon(G))\mathbf{1}_G] = O_P(1)$。与此同时，根据假设，$\hat{} \hat{} |\sqrt{n}(\theta - \theta_P)| \cdot \sup_{G\in\mathcal{G}}]\mathbb{P}_n(G) \leq |\sqrt{n}(\theta - \theta_P)| = O_P(1)$。因此，这个上界为$O_P(1)$。

然后，应用斯卢茨基引理，我们得到

$$\begin{aligned}
&\left(\frac{\mathbb{P}_n[\mathbf{1}_G]}{\hat{s}(G)}\right) \cdot \sqrt{n} \cdot \mathbb{P}_n\left[\left(L - \hat{\theta} - \epsilon(G)\right)\mathbf{1}_G\right] \\
&\quad = \frac{1}{s(G)} \cdot \sqrt{n} \cdot \mathbb{P}_n\left[\left((L - \hat{\theta})P[\mathbf{1}_G] - P[(L - \theta_P)\mathbf{1}_G]\right)\mathbf{1}_G\right] + o_P(1) \\
&\quad = \frac{1}{s(G)} \cdot \sqrt{n}\left(\mathbb{P}_n[(L - \hat{\theta})\mathbf{1}_G] \cdot P[\mathbf{1}_G] - P[(L - \theta_P)\mathbf{1}_G] \cdot \mathbb{P}_n[\mathbf{1}_G]\right) + o_P(1).
\end{aligned}$$

接下来，我们的目标是证明该过程的自助法对应过程是一致的，即必须证明该过程由一个 P-Donsker 类索引。为此，注意到我们可以将该过程改写为

$$\frac{P[\mathbf{1}_G]}{s(G)} \cdot \sqrt{n}\left(\mathbb{P}_n[(L - \hat{\theta}) \cdot \mathbf{1}_G] - P[(L - \theta_P) \cdot \mathbf{1}_G]\right) - \frac{P[(L - \theta_P)\mathbf{1}_G]}{s(G)} \cdot \sqrt{n}(\mathbb{P}_n - P)[\mathbf{1}_G] + o_P(1)$$

接着，我们通过观察下式对第一项进行线性化

$$\begin{aligned}
\sqrt{n}\left(\mathbb{P}_n[(L - \hat{\theta}) \cdot \mathbf{1}_G] - P[(L - \theta_P) \cdot \mathbf{1}_G]\right) &= \sqrt{n}(\mathbb{P}_n - P)[L \cdot \mathbf{1}_G] - \theta_P \cdot \sqrt{n}(\mathbb{P}_n - P)[\mathbf{1}_G] \\
&\quad - P[\mathbf{1}_G] \cdot \sqrt{n}(\hat{\theta} - \theta_P) + \underbrace{\frac{1}{\sqrt{n}} \cdot \sqrt{n}(\mathbb{P}_n - P)[\mathbf{1}_G] \cdot \sqrt{n}(\hat{\theta} - \theta_P)}_{o_P(1)}.
\end{aligned}$$

将其代入前面的展开式并化简后，我们得到

$$\sqrt{n}(\mathbb{P}_n - P)\left[\frac{P[\mathbf{1}_G]}{s(G)} \cdot L \cdot \mathbf{1}_G\right] - \sqrt{n}(\mathbb{P}_n - P)\left[\frac{P[L \cdot \mathbf{1}_G]}{s(G)} \cdot \mathbf{1}_G\right] - \sqrt{n}(\mathbb{P}_n - P)\left[\frac{P[\mathbf{1}_G]^2}{s(G)} \cdot \psi\right] + o_P(1)。$$

我们声称，此处展示的三个经验过程均由 Donsker 类索引。以第一个过程为例，我们注意到 $\{P[\mathbf{1}_G]/s(G) \mid G \in \mathcal{G}\}$ 是一个普适一致有界Donsker类，而 $L \cdot \mathbf{1}_G$ 是一个 P-Donsker类，因为有界可测函数与 Donsker 类的乘积仍是 Donsker 类。两个一致有界 Donsker 类的两两乘积仍是 Donsker 类。显然，$\{P[\mathbf{1}_G]/s(G) \cdot L \cdot \mathbf{1}_G \mid G \in \mathcal{G}\}$ 是这种乘积的子集，因此根据 [39], 中的定理 2.10.1，第一个过程由一个 Donsker 类索引。

通过几乎相同的论证，可以证明最后两个过程也都是 P-Donsker。回想一下，Donsker 性质在加法下保持不变。因此，我们得出

$$\sqrt{n}(\mathbb{P}_n - P)\left[\frac{P[\mathbf{1}_G]}{s(G)} \cdot L \cdot \mathbf{1}_G - \frac{P[L \cdot \mathbf{1}_G]}{s(G)} \cdot \mathbf{1}_G - \frac{P[\mathbf{1}_G]^2}{s(G)} \cdot \psi\right] + o_P(1) \tag{18}$$

由 Donsker类的一个子集索引，即由一个 Donsker类索引。

通过相同的推导，我们还证明，自助法过程 算法 6, $1\hat{s}(G)\cdot\mathbb{P}_n(G)\ \cdot\mathbb{P}_n^*(G)\cdot(\epsilon^*(G)\ -\hat{}\epsilon(G))$ 等于

$$\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)\left[\frac{P[\mathbf{1}_G]}{s(G)} \cdot L \cdot \mathbf{1}_G - \frac{P[L \cdot \mathbf{1}_G]}{s(G)} \cdot \mathbf{1}_G - \frac{P[\mathbf{1}_G]^2}{s(G)} \cdot \psi\right] + o_P(1).$$

应用引理B.2后，我们得出自助法是一致的。根据连续映射定理，我们可以对 $G \in \mathcal{G}$ 取上确界，并得出在 算法 6 中抽取的过程，与所关注概率中所述的过程渐近等价，

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{G\in\mathcal{G}}\left\{\sqrt{n}\cdot\left(\frac{\mathbb{P}_n(G)}{\hat{s}(G)}\right)\cdot\mathbb{P}_n[(L-\hat{\theta}-\epsilon(G))\cdot\mathbf{1}_G]\right\} \geq t^*\cdot\sqrt{n}\right).$$

最后，我们必须证明 $t^*\cdot\sqrt{n}$ 能一致地估计（$1-\alpha$）分位数，即式（18）上确界的分位数。回顾 Lehmann等人 [27, 引理11.2.1(ii)] 证明了，只要分布函数在所关注点连续且严格递增，分位数一致性就成立。我们的假设 $\hat{}\ L-\theta$ 对某个非空的 G 非常量，这意味着 $s\ \overset{1}{G}(\ )\cdot\mathbb{P}_n(G)\ \cdot\mathbb{P}(G)\cdot(\hat{}\epsilon(G)\ \underline{\ }\ \epsilon(G))$ 的渐近方差对某个 G非零。因此，上确界过程的（$1-\alpha$）分位数严格大于 0，而引理 B.3 推出所需结果。

完成证明，

$$\lim_{n\to\infty}\mathbb{P}\left(\exists\, G\in\mathcal{G}\ \text{s.t.}\ \epsilon_{\text{lb}}(G) > \epsilon(G)\right)$$
$$= \lim_{n\to\infty}\left(\sup_{G\in\mathcal{G}}\left\{\frac{1}{\hat{s}(G)}\cdot\mathbb{P}_n(G)\cdot\mathbb{P}_n^*(G)\cdot(\epsilon^*(G)-\hat{\epsilon}(G))\right\} > t^*\right) = \alpha.$$

□

布尔 **认证** 接下来我们考虑布尔认证，在这种认证中，我们检验 nu ll

$$H_0(G) : \epsilon(G) \leq \epsilon.$$

在 算法 7中，我们定义 算法 3的重新缩放的变体。回顾一下，当G 时，我们会为其颁发证书

$$\hat{\epsilon}(G) \geq \epsilon + \frac{t^*}{\mathbb{P}_n(G)}.$$

定理 B.2 随后证明，所颁发的证书以 $1-\alpha$ 的概率同时有效。

ˆ**定理 B.2.** 假设 L 具有有限方差，且 $L-\theta$ 在至少一个非空组上不是常数。进一步假设 $\mathcal{G}$ 具有有限 *VC* 维数。那么，

$$\lim_{n\to\infty}\mathbb{P}(\text{there exists any falsely certified } G\in\mathcal{G}) \leq \alpha.$$

算法 7 对布尔证书临界值进行自助法（重新缩放）

输入： 子群体 $\mathcal{G}$，差异 ϵ，审计轨迹 $\mathcal{D}$，水平 α，阈值 p_*，权重 w_0，自助样本数量 B

对于 $b = 1, \ldots, B$，执行以下操作 从 $\mathcal{D}$ 中进行一次大小为 n 的有放回抽样，得到样本 $\mathcal{D}_b^*$； 定义 $\mathbb{P}_b^*(G) := \frac{1}{n}\sum_{(x_i^*, y_i^*) \in \mathcal{D}_b^*} \mathbf{1}\{(x_i^*, \bar{y}_i^*) \in G\}$； 定义 $\epsilon_b^*(G) := \frac{1}{\mathbb{P}_b^*(G) \cdot n}\sum_{(x_i^*, y_i^*) \in G} L_i^* - \hat{\theta}(\mathcal{D}_b^*)$; 循环结束 将渐近方差估计量定义为

$$\hat{\sigma}_G^2 := \widehat{\mathrm{Var}}(L \mid G) + \mathbb{P}_n(G)\left(\widehat{\mathrm{Var}}(\psi) - 2 \cdot \widehat{\mathrm{Cov}}(L, \psi \mid G)\right)$$

其中 $\widehat{\mathrm{Cov}}(\cdot)$ 和 $\widehat{\mathrm{Var}}(\cdot)$ 分别为样本（条件）方差和协方差； 定义 $\hat{s}(G)$ 如下（9）；

对于 $b = 1, \ldots, B$ **执行** $t^{(b)} = \max_{G \in \mathcal{G}} \left\{ \frac{1}{s \hat{s}(G)} \cdot (\mathbb{P}_b^*(G) \cdot (\epsilon_b^*(G) - \epsilon) - \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon)) \right\}$; **结束循环** **返回：** $t^* = \text{分位数}(1-\alpha; \{t^{(b)}\}_{b=1}^B)$

证明 为简化记号，我们将在全文中用 $\mathbf{1}_G$ 替换 $\mathbf{1}\{(X, Y) \in G\}$。首先，重述我们需要证明的结论。

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{G \in \{G : H_0(G) \text{ holds}\}} \hat{\epsilon}(G) \geq \epsilon + t^* \cdot \frac{\hat{s}(G)}{\mathbb{P}_n(G)} \right) \leq \alpha.$$

重新排列后，我们得到

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{G \in \{G : H_0(G) \text{ holds}\}} \hat{\epsilon}(G) \geq \epsilon + t^* \cdot \frac{\hat{s}(G)}{\mathbb{P}_n(G)} \right) = \lim_{n \to \infty} \mathbb{P}\left(\sup_{G \in \{G : H_0(G) \text{ holds}\}} \frac{1}{\hat{s}(G)} \cdot \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) \geq t^* \right).$$

根据假设，在 $H_0(G)$ 下，$\mathbb{P}(G) \cdot (\epsilon(G) - \epsilon) \leq 0$，因此可得

$$\begin{aligned}
&\lim_{n \to \infty} \mathbb{P}\left(\sup_{G \in \{G : H_0(G) \text{ holds}\}} \left\{ \frac{1}{\hat{s}(G)} \cdot \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) \right\} \geq t^* \right) \\
&\leq \lim_{n \to \infty} \mathbb{P}\left(\sup_{G \in \{\mathcal{G} : H_0(G) \text{ holds}\}} \left\{ \frac{1}{\hat{s}(G)} \cdot (\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) - \mathbb{P}(G) \cdot (\epsilon(G) - \epsilon)) \right\} \geq t^* \right) \\
&\leq \mathbb{P}\left(\sup_{G \in \mathcal{G}} \left\{ \frac{1}{\hat{s}(G)} \cdot (\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) - \mathbb{P}(G) \cdot (\epsilon(G) - \epsilon)) \right\} \geq t^* \right)
\end{aligned}$$

对于 $B = \infty$，回想一下，t^* 是 $\sup_{G \in \mathcal{G}}\{(1/\hat{s}(G)) \cdot (\mathbb{P}_n^*(G) \cdot (\epsilon^*(G) - \epsilon) - \mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon))\}$ 的 $(1-\alpha)$ 分位数。我们必须证明，t^* 是 $\sup_{G \in \mathcal{G}}\{(1/\hat{s}(G)) \cdot (\mathbb{P}_n(G) \cdot (\hat{\epsilon}(G) - \epsilon) - \mathbb{P}(G) \cdot (\epsilon(G) - \epsilon))\}$ 的 $(1-\alpha)$ 分位数的一致估计。

在定理 B.1 的证明中，我们已经证明 $1/\hat{s}(G)$ 对 $1/s(G)$ 一致收敛，因此，只要 $\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon)-\mathbb{P}(G(G)-\epsilon)$ 是 $O_P(1)$，我们就可以应用斯卢茨基引理，用 $\frac{1}{s}_G()$ 替换 $1/\hat{s}(G)$。我们通过证明它在渐近上等价于一个以某个 P-Donsker 函数类为指标的经验过程来建立这一点。为此，我们将 $\sqrt{n}[\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon)-\mathbb{P}(G)\cdot(\epsilon(G)-\epsilon)]$ 改写为

$$\sqrt{n}(\mathbb{P}_n-P)[(L-\epsilon)\cdot\mathbf{1}_G]-\sqrt{n}\left(\hat{\theta}\cdot\mathbb{P}_n[\mathbf{1}_G]-\theta_P\cdot P[\mathbf{1}_G]\right).$$

我们可以进一步展开 $\sqrt{n}\left(\hat{\theta}\cdot\mathbb{P}_n[\mathbf{1}_G]-\theta_P\cdot P[\mathbf{1}_G]\right)$，得到

$$\theta_P\cdot\left(\sqrt{n}(\mathbb{P}_n-P)[\mathbf{1}_G]\right)+\mathbb{P}(G)\cdot\left(\sqrt{n}(\hat{\theta}-\theta_P)\right)+\underbrace{\frac{1}{\sqrt{n}}\left(\sqrt{n}(\mathbb{P}_n-P)[\mathbf{1}_G]\right)\left(\sqrt{n}(\hat{\theta}-\theta_P)\right)}_{o_P(1)}.$$

结合这些结果，并应用 $\sqrt{n}(\theta-\theta_P)$ 的渐近线性性质，我们现在可以将 $\sqrt{n}\left[\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon)-\mathbb{P}(G)\cdot(\epsilon(G)-\epsilon)\right]$ 线性化，从而得到（渐近）等价的过程：

$$\sqrt{n}(\mathbb{P}_n-P)[(L-\epsilon)\cdot\mathbf{1}_G]-\sqrt{n}(\mathbb{P}_n-P)[\theta_P\cdot\mathbf{1}_G]-\sqrt{n}(\mathbb{P}_n-P)[\mathbb{P}(G)\cdot\psi]+o_P(1).$$

以相同的方式，我们也可以将该过程的自助法对应形式线性化，即

$$\sqrt{n}(\mathbb{P}_n^*-\mathbb{P}_n)[(L-\epsilon)\cdot\mathbf{1}_G]-\sqrt{n}(\mathbb{P}_n^*-\mathbb{P}_n)[\theta_P\cdot\mathbf{1}_G]-\sqrt{n}(\mathbb{P}_n^*-\mathbb{P}_n)[\mathbb{P}(G)\cdot\psi]+o_{P_n}(1).$$

这三个项分别是经验过程或自助经验过程，它们以如下函数类为指标：

$$\mathcal{F}_1:=\{(L-\epsilon)\cdot\mathbf{1}_G\mid G\in\mathcal{G}\}\qquad\mathcal{F}_2:=\{\theta_P\cdot\mathbf{1}_G\mid G\in\mathcal{G}\}\qquad\mathcal{F}_3:=\{\mathbb{P}(G)\cdot\psi\mid G\in\mathcal{G}\}.$$

应用 [26] 的引理 9.9(vi)，取 $g(x,y)=L-\epsilon$ 和 $\mathcal{F}=\{\mathbf{1}\{x\in G\}:\ G\in\mathcal{G}\}$，可得 $\mathcal{F}_1$ 是 VC 类。接着，根据我们的假设 $\mathbb{E}_P[L^2]<\infty$，$\mathcal{H}_1$ 是 P-Donsker 类（参见 [39] 中的定理 2.10.20）。根据引理 B.1，$\mathcal{F}_2$ 是 P-Donsker 类。最后，注意到 $\mathcal{F}=\{\mathbb{P}(G)\mid G\in\mathcal{G}\}$ 的一致熵积分有限。由相同的论证可知，根据 [39] 中的定理 2.10.20，$\mathcal{F}_3$ 是 P-Donsker 类。

由于 Donsker 性质在逐点相加（ [39]中的例 2.10.7）以及取子集（ [39]中的定理 2.10.1）时均得以保持，因此我们得出，$\mathcal{F}_1+\mathcal{F}_2+\mathcal{F}_3$ 也具有 P-Donsker 性质。因此，应用斯卢茨基引理和引理 B.2，我们证明了自助法分布 $\sqrt{n}\left[(1/\hat{s}(G))\cdot(\mathbb{P}_n^*(G)\cdot(\epsilon^*(G)-\epsilon)-\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon))\right]$ 能够一致地估计 $\sqrt{n}[(1/\hat{s}(G))\cdot(\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon)-\mathbb{P}(G)\cdot(\epsilon(G)-\epsilon))]$的极限分布。由于上确界是连续的，根据连续映射定理，我们可以立即断言

$$\sup_{G\in\mathcal{G}}\sqrt{n}\left[(1/\hat{s}(G))\cdot(\mathbb{P}_n^*(G)\cdot(\epsilon^*(G)-\epsilon)-\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon))\right]$$

它能够一致地估计其极限分布。

为证明 t^* 是有效的临界值，我们需要证明，$\sup_{G\in\mathcal{G}}\{\sqrt{n}\left[(1/\hat{s}(G))\cdot(\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon)-\mathbb{P}(G)\cdot(\epsilon(G)-\epsilon))\right]\}$ 的极限分布函数在其 $(1-\alpha)$-分位数处严格递增且连续。

存在某个 G 使得 L 特别是，该假设意味着，在 G 处求值的过程的渐近方差为

大于 0。因此，极限过程的（$1-\alpha$）分位数严格大于 0。然后，引理 B.3 即可得到所需结果。

总结我们的论证，我们已经证明，错误认证的渐近概率可以按以下方式得到上界，

$$\begin{aligned}
&\lim_{n\to\infty}\mathbb{P}\left(\sup_{G\in\{G:H_0(G)\text{ holds}\}}\hat{\epsilon}(G)\geq\epsilon+t^*\cdot\frac{\hat{s}(G)}{\mathbb{P}_n(G)}\right)\\
&\qquad\leq\lim_{n\to\infty}\mathbb{P}\left(\sup_{G\in\mathcal{G}}\left\{\frac{1}{\hat{s}(G)}\cdot(\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon)-\mathbb{P}(G)\cdot(\epsilon(G)-\epsilon))\right\}\geq t^*\right)\\
&\qquad=\lim_{n\to\infty}\mathbb{P}\left(\sup_{G\in\mathcal{G}}\left\{\frac{1}{\hat{s}(G)}\cdot(\mathbb{P}_n^*(G)\cdot(\epsilon^*(G)-\epsilon)-\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon))\right\}\geq t^*\right)=\alpha.
\end{aligned}$$

□

定理 3.2 由定理 B.2 在 $\hat{s}(G)$ 时立即推出：$=1$。我们还要指出，如果 $\epsilon(G)=\epsilon$ 对所有 G均成立，则我们的第一类错误上界中的任何松弛都会消失。正如正文所述，这对应于 t^* 实现精确第一类错误控制的条件。

B.3 替代认证目标

在此，我们将说明如何把上一节和正文中所述的认证程序扩展到性能差异的其他定义。

在构造上界 置信界或认证 $\epsilon(G)<\epsilon$ 时，我们只需将 t^* 乘以-1。因此，现在 $\epsilon_{\text{ub}}(G)=\hat{\epsilon}(G)+t^*\hat{}$ $s(G)/\mathbb{P}_n(G)^2,$，并且当 $\hat{\epsilon}(G)\leq\epsilon-t^*/\mathbb{P}_n(G)$ 时进行认证。

注意，如果将上确界过程的（$1-\alpha$）分位数替换为下确界过程的 α 分位数，上置信界／证书有效性的论证仍然成立。然而，后者只是前者的 -1 倍。回想一下，中心化高斯过程的下确界在分布上等于该过程上确界的 -1 倍。

接下来，我们考虑构造置信区间的问题。我们提议对绝对值过程进行自助法，

$$\sup_{G\in\mathcal{G}}\left|\frac{1}{\hat{s}(G)}\cdot\mathbb{P}(G)\cdot\mathbb{P}_n(G)\cdot(\hat{\epsilon}(G)-\epsilon(G))\right|.$$

然后，若 t^* 表示该过程（$1-\alpha$）分位数的自助法估计，则置信集通过以下方式构造：

$$\left[\hat{\epsilon}(G)-t^*\cdot\frac{\hat{s}(G)}{\mathbb{P}_n(G)^2},\hat{\epsilon}(G)+t^*\cdot\frac{\hat{s}(G)}{\mathbb{P}_n(G)^2}\right].$$

最后，对于区间认证任务，我们检验原假设 $H_0(G):|\epsilon(G)|\geq\epsilon_\circ$ 等价地，我们检验：

$$\bar{H}_0(G):\{\epsilon(G)\geq\epsilon\}\bigcup\{\epsilon(G)\leq-\epsilon\}.$$

这也称为“生物等效性”零假设，可通过执行针对 $H_0(G):\epsilon(G)\geq\epsilon$ 和 $H_0(G):\epsilon(G)\leq-\epsilon$ 的单侧布尔认证程序，并在两个零假设都被拒绝时进行认证来检验。形式上，我们会在显著性水平 α 下检验第一个零假设，若 $\hat{\epsilon}(G)\leq\epsilon-t_1^*/\mathbb{P}_n(G)$ 则拒绝；并在显著性水平 α 下检验第二个零假设，若 $\hat{\epsilon}(G)$ $\geq-\epsilon+t_2^*/\mathbb{P}_n(G)$ 则拒绝。然后，只要这两个不等式同时成立，我们就认证 G 满足 $|\epsilon(G)|<\epsilon$ 。

C 标记审计

为了证明我们的标记程序在渐近意义下有效，我们首先证明所提出的检验统计量渐近服从正态分布。回忆一下，$\hat{\epsilon}(G) := |G|^{-1}\sum_{i\in G} L_i - \hat{\theta}$。我们将统计量向量 $\{\hat{\epsilon}(G)\}_{G\in\mathcal{G}}$ 记为 $\hat{\epsilon}$，并将相应的真实差异向量记为 ϵ。

命题 C.1。 假设对于所有 $G \in \mathcal{G}$， $\mathbb{P}(G)$ 和 $\mathrm{Var}(L \mid G)$ 均与 0 保持有界距离。如果$m = |\mathcal{G}| < \infty$，则有 $\sqrt{n}(\hat{\epsilon} - \epsilon) \rightsquigarrow \mathcal{N}_m(\mathbf{0}_m, \bar{\Sigma})$。

证明 将中心极限定理应用于引理B.4中推导出的渐近线性展开。 □

对于任意组 $G_j \in \mathcal{G}$，命题 C.1 表明，p 值

$$p_1(G_j;\epsilon) := \Phi\left(\frac{\sqrt{n}\,(\hat{\epsilon}(G_j) - \epsilon)}{\sqrt{\Sigma_{jj}}}\right), \quad p_2(G_j;\epsilon) := 1 - \Phi\left(\frac{\sqrt{n}\,(\hat{\epsilon}(G_j) + \epsilon)}{\sqrt{\Sigma_{jj}}}\right),$$
$$p_3(G_j) := p_1(G_j;-\epsilon) \wedge p_2(G_j;-\epsilon)$$

对于原假设 $H_1(G_j) : \epsilon(G_j) \geq \epsilon$、 $H_2(G_j) : \epsilon(G_j) \leq -\epsilon$、 $H_3(G_j) : |\epsilon(G_j)| \leq \epsilon$，分别具有边际渐近有效性，即 $\limsup_{n\to\infty} \mathbb{P}_{H_i}(p_i(G_j) \leq u) \leq u$。

Σ_{jj} 未知，但根据斯卢茨基引理，我们可以用任意一致估计量替代 Σ_{jj}。我们可以解析地计算出这样一个估计量，但下面我们依靠自助法来构造这样的估计量。具体而言，令

$$s_j^* := \mathrm{Quantile}\left(0.5, \sqrt{n}\,|\epsilon^*(G) - \hat{\epsilon}(G)|\right).$$

那么，引理 ma C.1 给出了以下条件，在这些条件下， s_j^* 是 $\sqrt{\Sigma_j}$的一致估计量 $\overline{}_j$.

引理 C.1. 保留命题 *C.1* 的假设。则， $s_j^*/\Phi^{-1}(3/4) \xrightarrow{p} \sqrt{\Sigma_{jj}}$ $\overline{}$.

证明 应用自助法德尔塔方法和连续映射定理，我们可以得出，自助法分布 $\sqrt{n}\,|\epsilon^*(G_j) - \hat{\epsilon}(G_j)|$ 一致地估计了 $\sqrt{n}\,|\hat{\epsilon}(G_j) - \epsilon(G_j)|$ 的分布。由于后者的极限分布是连续的，并且在处处严格递增（命题 C.1），我们可以应用 van der Vaart [38, 引理 5。10]从而得出，自助法估计的中位数绝对偏差对于渐近中位数绝对偏差是一致的。然后，回顾众所周知的事实：方差为 σ^2 的高斯分布的中位数绝对偏差等于 $\Phi^{-1}(3/4)\cdot\sigma$，即可得到该结果。

□

引理 C.1 因此意味着，我们可以将 $\sqrt{\Sigma_{jj}}$ 在 p 值定义中替换为由以下公式给出的蒙特卡罗估计 s_j^*。

有了这些结果，我们可以证明命题4.1，但首先需要回顾一些定义以及关于BH 程序有效性的众所周知的条件。

定义 C.1。 我们称 X 在子集 I_0 上具有正回归依赖（在 I_0 上具有正回归依赖），若对于任意递增集合 D，以及每个 $i \in I_0$， $\mathbb{P}(X \in D \mid X_i = x)$ 关于 x 单调递增。

例 1 (情形 3.1，见 [3])**：** 通过检验 $H_0 : \mu \leq \mu^*$或 $H_0 : \mu \geq \mu^*$，使用 $T \sim \mathcal{N}(\mu, \Sigma)$ 得到的单侧高斯 p 值在 I_0 上满足正回归依赖，当且仅当 $\Sigma_{ij} \geq 0$ 对所有 $i \in I_0$ 和 $j \in [m]$均成立。 ♣

定理 C.1 (定理 1.2，见 [3])。 如果检验统计量的联合分布在对应于 m_0 真原假设的检验统计量子集上满足正回归依赖，则 *BH(α)* 程序控制的错误发现率不超过水平 $\frac{m_0}{m}\alpha$。

—

为方便读者，我们在完成证明之前重述命题4.1 。

命题 C.2。 假设 $\mathbb{P}((X,Y) \in G)$ 以及 $\mathrm{Var}(L(X,Y) \mid (X,Y) \in G)$ 对所有 $G \in \mathcal{G}$, θ_P 均有一个已知的先验正下界，并且以下条件之一成立：

(i) $\{G\}_{G\in\mathcal{G}}$ 两两不相交；

(ii) L 取值于 $\{0,1\}$。

如果我们将 *BH(α)* 程序在 $\{p(G)\}_{G\in\mathcal{G}}$ 上的拒绝结果标记，那么错误发现率渐近地控制在水平 α。

证明 为使索引更清晰，我们令 $\mathcal{G} = \{G_j\}_{j=1}^m$。不失一般性，假设 $\theta_P = 0$。那么，命题 C.1 表明 $\sqrt{n}\{\hat{\epsilon}(G_j) - \epsilon(G_j)\}_{j=1}^m \rightsquigarrow \mathcal{N}(\bar{\mathbf{0}}_m, \Sigma)$。 Σ 的非对角元素等于

$$\frac{\mathbb{E}\left[(L-\epsilon(G_j))(L-\epsilon(G_k))\mathbf{1}\left\{G_j \cap G_k\right\}\right]}{\mathbb{P}(G_j)\mathbb{P}(G_k)}.$$

在条件 (i) 下，所有非对角元素均为 0，因为分子中的指示函数始终等于 0。

考虑 条件 (ii)。在这种情况下，我们可以将协方差表达式改写为

$$\left(\epsilon(G_j \cap G_k)\left(1-\epsilon(G_j)-\epsilon(G_k)\right)+\epsilon(G_j)\epsilon(G_k)\right)\frac{\mathbb{P}(G_j \cap G_k)}{\mathbb{P}(G_j)\mathbb{P}(G_k)}.$$

然后，为了导出矛盾，假设该表达式为负。那么，

$$\begin{aligned} 0 &> \epsilon(G_j \cap G_k)\left(1-\epsilon(G_j)-\epsilon(G_k)\right)+\epsilon(G_j)\epsilon(G_k) \\ &\geq 1-\epsilon(G_j)-\epsilon(G_k)+\epsilon(G_j)\epsilon(G_k) \\ &= (1-\epsilon(G_j))(1-\epsilon(G_k)). \end{aligned}$$

但由于最后一个表达式在条件 (ii) 下非负，我们得到矛盾，因此必须得出结论：渐近协方差是非负的。

在这两种情况下，我们都证明了这些检验统计量的渐近分布是协方差非负的多元高斯分布。因此，算法 4 输出的单侧 p 值在零假设集合上渐近满足正回归依赖。然后，应用有界收敛定理和定理 C.1即可得到所需结果。

□

虽然我们尚未证明在检验双侧原假设时能够控制错误发现率，但人们普遍推测 BH 程序在这种情况下能够控制错误发现率 [15]。此外，即使不满足正回归依赖（即不符合命题 4.1所给出的条件），实践中将 BH程序应用于渐近高斯 p 值时，也相当难以显著违反错误发现率控制。

D 审计分布偏移

在考虑对属于再生核希尔伯特空间单位球的偏移进行审计这一问题之前，我们注意到，定理 B.1 的证明和定理 B.2 的证明均不依赖于 $\mathcal{G}$ 是 VC 类。由于 $\mathcal{G}$ 是 VC 类，$\mathcal{F} = \{\mathbf{1}\{(X, Y) \in G\} \mid G \in \mathcal{G}\}$ 是一个 P-Donsker 类。这意味着，即使将 $\mathcal{G}$ 替换为某个一般的 Donsker 函数类，上述证明仍然成立。

然而，将我们的结果扩展到再生核希尔伯特空间（RKHS）情形，还需要做一些额外工作。为了证明我们在 $\epsilon(h)$上构造置信集的方法是有效的，我们必须先证明一些关于再生核希尔伯特空间单位球的初步结果。

引理 D.1。 假设 $\|k(X, X)\|_\infty$ 是有限的，且 $k(\cdot, x)$ 是连续的。那么，由 k诱导的再生核希尔伯特空间的单位球（记为 $\mathcal{H}_1$）是一个 *P-Donsker* 类。

证明 文献 [35] 中的引理 4.28 和引理 4.33 表明，这些假设足以保证 $\mathcal{H}_1$ 是一个可分希尔伯特空间，且是有界连续函数空间的子集。因此，结论可由文献 [28] 中的定理 1.1 得出，其中令 T 为恒等映射。恒等映射显然是线性的，并且满足上确界范数下的连续性假设，因为前者受再生核希尔伯特空间范数控制。 □

引理 D.2。 保留引理 *D.1*的假设。则，对于一致有界函数 $\tilde{L}$和 M，$\mathcal{H}_1 := \{P[L \cdot h] \cdot h \mid h \in \mathcal{H}_1\}$ 是一个 P-*Donsker*类。

证明 首先注意到，对于所有 $h \in \mathcal{H}_1$，$P[L \cdot h]$ 都是一致有界的：

$$P[L \cdot h] \leq \|L\|_\infty \|k(X, X)\|_\infty \|h\|_{\mathcal{H}} \leq \|L\|_\infty \|k(X, X)\|_\infty =: C.$$

因此，$\tilde{\mathcal{H}}_1$是 $\mathcal{H}_C$的子集，即它是 $\mathcal{H}_1$的伸缩集。由于 $C \cdot \mathcal{H}_1$是 P-Donsker（[39]中的定理 2.10.6 P），且 P-Donsker类的任何子集都是 [39]-Donsker（中的定理 2.10.1 $\tilde{\mathcal{H}}_1$），我们得出 P是 P-Donsker。 □

这里我们将主定理推广到包含相对于一个$\hat{}$ 估计阈值定义的差异； θ假设该阈值满足附录 B.2 中所述的渐近线性和自助法一致性假设。 算法 8 修改了算法 5 ，使得$\hat{}$ 自助法能够将 θ中的估计误差纳入考虑。

算法 8 使用估计阈值自助法计算再生核希尔伯特空间置信集临界值

1：输入：核函数 k、审计轨迹 $\mathcal{D}$、水平 α、自助样本 B2：定义 $\mathbf{L} := \{L(f(x_i)\ y_i)\}_{i=1}^n$；3：定义 $\mathbf{K} := \{k(x_i, x_j)\}_{i,j=1}^n$；4：对于 $b = 1$，…，B，执行 5：从 $\sim \text{Mult}\left(n \text{ 中抽取 } \mathbf{w};\ \frac{1}{n},\ \ldots,\ \frac{1}{n}\right)$；6：估计自助法阈值偏差 $t = \frac{1}{n}\sum_{i=1}^n (\mathbf{w}_i - 1) \cdot \psi_i$；7：$\mathbf{A} = \frac{1}{n^2}\left((\mathbf{w} \odot \mathbf{L})\ \mathbf{1}^\top - \mathbf{w}\mathbf{L}^\top - t \cdot \mathbb{I}_n\right)$；8：$t^{(b)} = \lambda_{\max}\left(\mathbf{K}^{1/2}\left(\frac{\mathbf{A}+\mathbf{A}^\top}{2}\right)\mathbf{K}^{1/2}\right)$；9：结束循环 10：返回：$t^* = \text{分位数}(1 - \alpha/4; \{t^{(b)}\}_{b=1}^B)$

引理 D.3。 对于 $B=\infty$，由 t^* 算法 8 输出的结果 等于 $(1-\alpha)$-分位数

$$\sup_{h\in\mathcal{H}_1}(\mathbb{P}_n^*-\mathbb{P}_n)[(L-\hat{\theta}(\mathcal{D}^*))\cdot\mathbb{P}_n[h]-\mathbb{P}_n[(L-\hat{\theta})\cdot h])h].$$

证明 首先，我们观察到，所关注的经验过程等于[3] to

$$\mathbb{P}_n[h]\mathbb{P}_n^*[L\cdot h]-\mathbb{P}_n^*[h]\mathbb{P}_n[L\cdot h]-\mathbb{P}_n[h]^2\cdot(\mathbb{P}_n^*-\mathbb{P}_n)[\psi]+o_P(1).$$

我们可以用一个多项式变量重写所关注的（线性化）过程，

$$\sup_{h\in\mathcal{H}_1}\mathbb{P}_n[W\cdot(L\cdot\mathbb{P}_n[h]-\mathbb{P}_n[L\cdot h])h]-\mathbb{P}_n[h]^2\cdot\mathbb{P}_n[(W-1)\cdot\psi],$$

对于 $W\sim\text{Mult}(n,\ 1/n)$】【。

我们可以将该过程改写为未知函数 $h\in\mathcal{H}$ 与核函数之间的内积，

$$\sup_{h\in\mathcal{H}_1}(\langle\mathbb{P}_n[W\cdot L\cdot k(X,\cdot)],h\rangle\langle\mathbb{P}_n[k(X,\cdot)],h\rangle-\langle\mathbb{P}_n[L\cdot k(X,\cdot)],h\rangle\langle\mathbb{P}_n[W\cdot k(X,\cdot)],h\rangle)$$
$$-\mathbb{P}_n[(W-1)\cdot\psi]\langle\mathbb{P}_n[k(X,\cdot)],h\rangle^2.$$

由于我们知道，最优的 h^* 必须具有 $\sum_{i=1}^n\alpha_k k(\cdot,x_i)$ 的形式（回忆一下任意再生核希尔伯特空间的直和分解），我们可以将上述上确界改写为

$$\sup_{\alpha:\alpha^\top\mathbf{K}\alpha\le1}\frac{1}{n^2}\left(\alpha^\top\mathbf{K}(\mathbf{w}\odot\mathbf{L})\mathbf{1}^\top\mathbf{K}\alpha-\alpha^\top\mathbf{K}\mathbf{w}\mathbf{L}^\top\mathbf{K}\alpha-t\cdot\alpha^T\mathbf{K}\mathbf{K}\alpha\right),$$

其中，$\mathbf{K}=\{k(x_i,x_j)\}_{i,j=1}^n$、$\mathbf{w}=(W_1,\ldots,W_n)^\top$、$\mathbf{L}=(L_1,\ldots,L_n)^\top$以及 $t=\mathbb{P}_n[(W-1)\cdot\psi]$。令$\mathbf{A}=\frac{1}{n^2}[(\mathbf{w}\odot\mathbf{L})\mathbf{1}^\top-\mathbf{w}\mathbf{L}^\top-t\mathbb{I}]$，即可得到等价的目标函数，

$$\begin{aligned}\sup_{\alpha:\alpha^\top\mathbf{K}\alpha\le1}\frac{1}{2}\left(\alpha^\top\mathbf{K}\left(\mathbf{A}+\mathbf{A}^\top\right)\mathbf{K}\alpha\right)&=\sup_{\beta:\|\beta\|_2\le1}\frac{1}{2}\left(\beta^\top\mathbf{K}^{1/2}\left(\mathbf{A}+\mathbf{A}^\top\right)\mathbf{K}^{1/2}\beta\right)\\&=\frac{1}{2}\lambda_{\max}\left(\mathbf{K}^{1/2}\left(\mathbf{A}+\mathbf{A}^\top\right)\mathbf{K}^{1/2}\right).\end{aligned}$$

□

我们指出，如果不估计 θ_P，则可以忽略 A 定义中的 $t=0$ 和单位矩阵。这大大简化了最大特征值所需的计算，因为 $A+A^T$ 现在是低秩的。因此，对于相对于未知阈值定义的性能指标，我们通常不建议使用基于再生核希尔伯特空间的置信集。

有了这些初步结果，我们现在可以证明关于再生核希尔伯特空间置信集有效性的主要定理。给定 t^* 由算法8输出的结果，回顾一下，我们通过设置

$$\epsilon_{\text{lb}}(h):=\hat{\epsilon}(h)-\frac{t^*}{\left(\frac{1}{n}\sum_{i=1}^n h(x_i)\right)^2}.$$

为方便记号，我们用 $\mathcal{H}_1^+$ 表示属于 $\mathcal{H}_1$ 的所有非负函数集合。

[3]为避免重复相同的推导，请参阅定理 D.1 的证明，其中完整阐述了这一等价关系

.

定理 D.1。 假设$\mathrm{Var}(L) > 0$, $\|L\|_\infty$和 $\|k(X,X)\|_\infty$均有限， $k(\cdot, x)$是连续的，并且 $k(\cdot,\cdot)$ 是正定定核。那么，

$$\lim_{n\to\infty} \mathbb{P}\left(\epsilon_{\mathrm{lb}}(h) \le \epsilon(h) \text{ for all } h \in \mathcal{H}_1^+\right) \ge 1-\alpha.$$

证明 所需结果等价于

$$\lim_{n\to\infty} \mathbb{P}\left(\exists\, h \in \mathcal{H}_1^+ \text{ s.t. } (\mathbb{P}_n[h]\cdot\mathbb{P}_n[h]\cdot(\hat{\epsilon}(h)-\epsilon(h)) > t^*\right) \le \alpha.$$

两边同乘 $\mathbb{P}_n[h]$，左侧的过程可改写为

$$\mathbb{P}_n[h]\cdot\mathbb{P}_n[(L-\hat{\theta}-\epsilon(h))\cdot h].$$

我们声称 $\mathbb{P}_n[h]$ 对于 $\mathbb{P}[h]$具有一致相合性。为此，首先回顾引理D.1意味着 $\mathcal{H}_1$ 是 P-Donsker。Donsker性质对任意子集都保持不变，因此若 $\mathcal{H}_1$ 是P-Donsker，则 $\mathcal{H}_1^+$也是如此。一致相合性由 P-Donsker $\implies$ P-Glivenko-Cantelli 推出。

ˆ 要应用斯卢茨基引理并用 $\mathbb{P}[h]$ 替换 $\mathbb{P}_n[h]$，我们还必须证明 $\sqrt{n}\cdot\mathbb{P}_n[(L-\theta-\epsilon(h))\dot{-}h]$ 是 O_P(1)。为此，请注意，

$$\left|\sup_{h\in\mathcal{H}_1^+}\sqrt{n}\cdot\mathbb{P}_n[(L-\hat{\theta}-\epsilon(h))\cdot h]\right| \le \left|\sup_{h\in\mathcal{H}_1^+}\sqrt{n}(\mathbb{P}_n-P)[(L-\theta_P-\epsilon(h))\cdot h]\right| + \left|\sup_{h\in\mathcal{H}_1^+}\sqrt{n}(\mathbb{P}_n-P)[\psi]\cdot\mathbb{P}_n[h]\right|$$

˜我们可以证明，右侧的两项均为 O_P(1)。首先，我们断言 $\mathcal{H}_1 = \{(L-\epsilon(h))h \,|\, h\in\mathcal{H}_1^+\}$ 是 P-Donsker。首先，注意到

$$\tilde{\mathcal{H}}_1 \subseteq \{L\cdot h \mid h\in\mathcal{H}_1^+\} - \{\epsilon(h)\cdot h \mid h\in\mathcal{H}_1^+\}.$$

回想一下， P-Donsker类的任意子集也是 P-Donsker类（ [39]中的定理 2.10.1），我们需要证明该式右侧是 P-Donsker类。右侧的每个函数类都是 P-Donsker类。第一个成立是因为 $\|L\|_\infty < \infty$ （ [39]中的例 2.10.10）。接着，第二个是 P-Donsker类，因为它是两个一致有界的 P-Donsker类的逐元素乘积的子集（ [39]中的例 2.10.8）。最后，两个 P-Donsker类的逐元素加法会得到一个 P-Donsker类（ [39]中的例 2.10.7）。因此，我们得出$\sqrt{n}(\mathbb{P}_n-P)[(L-\epsilon(h))h] = O_P$(1)。第二项可以上界为 $|\sqrt{n}(\mathbb{P}_n-P)[\psi]|$，因此该项也是 O_P(1)。

现在，我们应用斯卢茨基引理，得到:

$$\begin{aligned} P[h]\cdot\sqrt{n}\cdot\mathbb{P}_n[(L-\hat{\theta}-\epsilon(h))h] &= \sqrt{n}\cdot(\mathbb{P}_n[(L-\hat{\theta})\cdot P[h] - P[(L-\theta_P)\cdot h])h]) \\ &= P[h]\cdot\sqrt{n}(\mathbb{P}_n[(L-\hat{\theta})\cdot h] - P[(L-\theta_P)\cdot h]) \\ &\quad - P[(L-\theta_P)\cdot h]\cdot\sqrt{n}(\mathbb{P}_n-P)[h] \end{aligned}$$

接下来，我们证明该过程是 P-Donsker，并收敛到紧高斯极限。为此，做】【。

因此，我们对该过程的第一项进行线性化：

$$\begin{aligned}P[h]\cdot\sqrt{n}(\mathbb{P}_n[(L-\hat{\theta})\cdot h]-P[(L-\theta_P)\cdot h]) &= P[h]\cdot\sqrt{n}(\mathbb{P}_n-P)[L\cdot h]\\ &\quad -P[h]\cdot\sqrt{n}(\hat{\theta}\cdot\mathbb{P}_n[h]-\theta_P\cdot P[h]))\\ &= P[h]\cdot\sqrt{n}(\mathbb{P}_n-P)[L\cdot h]\\ &\quad -P[h]\cdot\theta_P\sqrt{n}(\mathbb{P}_n-P)[h]\\ &\quad -P[h]^2\cdot\sqrt{n}(\mathbb{P}_n-P)[\psi]\\ &\quad +\underbrace{\frac{1}{\sqrt{n}}\left(\sqrt{n}(\hat{\theta}-\theta_P)\cdot\sqrt{n}(\mathbb{P}_n-P)[h]\right)}_{o_P(1)}.\end{aligned}$$

结合这两项，我们得出该过程等价于

$$\begin{aligned}&\sqrt{n}(\mathbb{P}_n-P)[P[h]\cdot L\cdot h]-\sqrt{n}(\mathbb{P}_n-P)[P[L\cdot h]\cdot h]-\sqrt{n}(\mathbb{P}_n-P)[P[h]^2\cdot\psi]\\ &\qquad=\sqrt{n}(\mathbb{P}_n-P)[P[h]\cdot L\cdot h-P[L\cdot h]\cdot h-P[h]^2\cdot\psi]\end{aligned}$$

注意，为该过程建立索引的函数类可以写成三个类的逐元素和的子集，

$$\begin{aligned}&\{P[h]\cdot L\cdot h-P[L\cdot h]\cdot h-P[h]^2\cdot\psi\mid h\in\mathcal{H}_1^+\}\\ &\qquad\subseteq\{P[h]\cdot L\cdot h\mid h\in\mathcal{H}_1^+\}-\{P[L\cdot h]\cdot h\mid h\in\mathcal{H}_1^+\}-\{P[h]^2\cdot\psi\mid h\in\mathcal{H}_1^+\},\end{aligned}$$

其中每个类都是 Donsker 类。

我们以第一个类为例验证它是 Donsker 类；另外两个证明的论证过程完全相同。注意，$\{L\cdot h\mid h\in\mathcal{H}_1^+\}$ 是一个 P-Donsker 类，因为 L 一致有界，且$\{h\mid h\in\mathcal{H}_1\}$ 是一个一致有界的 Donsker 类（见 [39]中的例 2.10.10）。于是，$\{P[h]\cdot L\cdot h\mid h\in\mathcal{H}_1^+\}$ 是一个 Donsker 类，因为两个一致有界 Donsker 类的逐元素乘积的子集也是 Donsker 类（见 [39]中的例 2.10.8）。

因此，根据 Donsker 类的定义和连续映射定理，

$$\sup_{h\in\mathcal{H}_1^+}\sqrt{n}(\mathbb{P}_n-P)[P[h]\cdot L\cdot h-P[L\cdot h]\cdot h-P[h]^2\cdot\psi]$$

收敛到紧极限。

然后，我们可以给出.....的上界

$$\begin{aligned}&\lim_{n\to\infty}\mathbb{P}\left(\sup_{h\in\mathcal{H}_1^+}\sqrt{n}(\mathbb{P}_n-P)[P[h]\cdot L\cdot h-P[L\cdot h]\cdot h-P[h_+]^2\cdot\psi]>t^*\cdot\sqrt{n}\right)\\ &\qquad\le\lim_{n\to\infty}\mathbb{P}\left(\sup_{h\in\mathcal{H}_1}\sqrt{n}(\mathbb{P}_n-P)[P[h]\cdot L\cdot h-P[L\cdot h]\cdot h-P[h]^2\cdot\psi]>t^*\cdot\sqrt{n}\right)\end{aligned}$$

自助法对于.....是一致的

$$\sup_{h\in\mathcal{H}_1}\sqrt{n}(\mathbb{P}_n-P)[P[h]\cdot L\cdot h-P[L\cdot h]\cdot h-P[h]^2\cdot\psi]$$

因为为该过程编索引的函数类是 P-Donsker。我们可以重复用于证明以 $h\in\mathcal{H}_1^+$为索引的过程是 Donsker 的相同论证。

通过与上述推导完全相同的推导，我们观察到，我们从算法8(参见引理D.3) 中抽取的自助法过程，

$$\sqrt{n} \cdot \left(\mathbb{P}_n^*[(L - \hat{\theta}(\mathcal{D}^*)) \cdot h] \cdot \mathbb{P}_n[h] - \mathbb{P}_n[(L - \hat{\theta}) \cdot h])] \cdot \mathbb{P}_n^*[h] \right),$$

等于

$$\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)[P[h] \cdot L \cdot h - P[L \cdot h] \cdot h - P[h]^2 \cdot \psi] + o_P(1).$$

因此，根据引理B.2和连续映射定理，自助法过程的上确界可一致估计极限过程上确界的分布。

最后，我们需要证明自助法分位数 t^* 是真实极限分位数的一致估计量。我们通过验证极限分布的累积分布函数在其（$1 - \alpha$）分位数 [27, 引理11.2.1(ii)处连续且严格递增，来确立自助法分位数的一致性。 L 上的方差假设保证至少存在某个 $h \in \mathcal{H}$，使得其极限分布

$$\sqrt{n} \left(\mathbb{P}_n[(L - \hat{\theta}) \cdot h] \cdot P[h] - P[(L - \theta_P) \cdot h] \cdot \mathbb{P}_n[h] \right)$$

是一个非平凡高斯分布，进而推出所需的累积分布函数性质。引理B.3 蕴含 t^* 的一致性，因此也得到了关于同时覆盖率的预期结论。 □

我们或许可以针对再生核希尔伯特空间调整自助法过程，使置信界宽度能够更自然地随所选偏移的“复杂度”变化。这里我们不考虑定义 Wald 型置信界，而只是希望调整该过程，使置信界宽度能够更自然地随所查询偏移的“复杂度”变化。例如，当前界的尺度为

$$\epsilon_{\text{lb}}(h) = \hat{\epsilon}(h) - \frac{C}{\sqrt{n} \cdot \mathbb{P}_n[h]^2},$$

尽管我们可能希望该界按以下方式缩放

$$\epsilon_{\text{lb}}(h) = \hat{\epsilon}(h) - \frac{C}{\sqrt{n_{\text{eff}}}}$$

其中 n_{eff} 量化了重加权指标的“有效样本量”。

我们通过观察到，对于 $z_i \overset{\text{iid}}{\sim} (0,1)$，$(\sum_{i=1}^n z_i \cdot h_i)/(\sum_{i=1}^n h_i)$ 的方差按 $1/\sqrt{n_{\text{eff}}}$ 缩放，来说明我们选择 $n_{\text{eff}} = (\sum_i h(x_i))^2 / \sum_i h(x_i)^2$ 的理由。 ——

为说明我们选择 $\hat{s}(h)$ 的理由，请注意，将过程按 $1/\hat{s}(h)$ 缩放后，可得到如下形式的界：

$$\epsilon_{\text{lb}}(h) = \hat{\epsilon}(h) - C \frac{\hat{s}(h)}{\mathbb{P}_n[h]^2 \cdot \sqrt{n}}.$$

因此，若求解使分母满足所需条件的 $\hat{s}(\cdot)$，便得到 $|\mathbb{P}_n[h]| \cdot \sqrt{\mathbb{P}_n[h^2]}$。为确保一致相合性，我们对其进行截断，并定义

$$\hat{s}(h) := \max \left(|\mathbb{P}_n[h]| \cdot \sqrt{\mathbb{P}_n[h^2]},\, h_* \right),$$

对于某个阈值 h_*。与 p_* （9）中的情况不同，请注意 h_* 无法解释。

除了缺乏可解释性之外，重新缩放后的再生核希尔伯特空间过程也无法再通过自助法高效计算。即使我们假设 θ_P 已知，现在也必须在 算法 8的第 8 行进行计算，

$$t^{(b)} = \sup_{h \in \mathcal{H}_1} \frac{\mathbb{P}_n^b[L \cdot h] \cdot \mathbb{P}_n[h] - \mathbb{P}_n[L \cdot h] \cdot \mathbb{P}_n^b[h]}{\max\left(|\mathbb{P}_n[h]| \cdot \sqrt{\mathbb{P}_n[h^2]},\, h_*\right)}.$$

尽管我们可能希望仿照之前的方法，将这一计算化简为某个特征值问题，但对再生核希尔伯特空间函数应用有限维表示，只能得到

$$\sup_{\beta : \|\beta\|_2 \le 1} \frac{\left(\beta^\top \mathbf{K}^{1/2} \left(\mathbf{A} + \mathbf{A}^\top\right) \mathbf{K}^{1/2} \beta\right)}{2 \cdot \max\left(|\mathbf{1}^\top K^{1/2} \beta| \cdot \sqrt{\beta^\top K \beta},\, h_*\right)}.$$

分母中的 $\sqrt{\beta^\top K \beta}$ 项使得优化 β 极具挑战性。由于对于正定定核， K 是满秩的，因此我们无法依靠 A 中的任何低秩结构来简化这个问题。我们至多只能对 K 作秩为 m 的近似，但这会得到一个难以处理（且不准确）的优化问题，其定义域是一个（$m+4$）维超球面的表面。如果~~我们~~直接从 $\hat{s}(G)$ 中删去 $\sqrt{\beta^\top K \beta}$ 项，自助法步骤便可化简为一个定义域为四维超球面表面的优化问题。我们可以通过穷举搜索求解该问题，但在实践中发现，这样得到的置信界并没有改善。因此，我们不建议对再生核希尔伯特空间过程进行重新缩放。