HappyMap：一种广义多重校准方法

Zhun Deng ！ 美国哥伦比亚大学计算机科学系**Cynthia Dwork** ！ 美国哈佛大学约翰·A·保尔森工程与应用科学学院**Linjun Zhang** [1] ！ 美国罗格斯大学统计学系

arXiv:2303.04379v1 [cs.LG] 8 Mar 2023

摘要

多重校准是一个强大且不断发展的概念，起源于算法公平性领域。对于一个预测器 f，它根据协变量 x 估计结果 y；给定函数类 $\mathcal{C}$，多重校准要求，在 $\mathcal{C}$ 中的审计器看来，预测器 $f(x)$ 与结果 y 无法区分。通过将人口统计学子群体纳入函数类 $\mathcal{C}$，即可体现公平性。近期研究表明，通过扩充函数类 $\mathcal{C}$，使其纳入适当的倾向得分重新加权函数，多重校准还可以实现目标无关学习：在源域上训练的模型能够在未见过的未来目标域上取得良好表现，而这些目标域可由重新加权函数近似刻画。

形式上，相对于 $\mathcal{C}$ 的多重校准对所有$c \in \mathcal{C}$，都能限制 $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x)\ x)\cdot(f(x)\ -y)]\right|$。在这项工作中，我们将项 $(f(x)-y)$ 视为一种特定映射，并探究扩充后的映射类所蕴含的能力。我们提出 *s-Happy* 多重校准，这是一种多重校准的推广，可带来广泛的新应用，包括一种用于不确定性量化（保形预测）的新公平性概念、一种在协变量偏移下进行保形预测的新技术，以及一种分析缺失数据的不同方法；同时，它还统一阐释了若干现有的、看似彼此无关的算法公平性概念和目标无关学习方法。

我们提出一个统一的$HappyMap$ 元算法，以涵盖所有这些结果，并给出其成功的充分性条件。

2012年 ACM 学科分类 计算理论 → 算法设计与分析；计算理论 → 面向应用领域的理论与算法

关键词与短语 算法公平性、目标无关学习、迁移学习

数字对象标识符 10.4230/LIPIcs.ITCS.2023.54

经费支持 本研究部分由阿尔弗雷德·P·斯隆基金会 G-2020-13941、西蒙斯基金会协作研究资助 733782，以及美国国家科学基金会 DMS-2015378 资助。

A致谢 我们感谢所有审稿人的评论和建议。我们深表感激
t也感谢 Aaron Roth 提供富有见地且宝贵的反馈，极大地改进了本文。

1 引言

预测算法对个人或个体实例进行评分，为每个对象赋予一个 $[0,1]$ 通常解释为概率的分数，例如明天下雨的概率。如果对于所有 $v \in [0,1]$，在被赋予数值 v的实例中，有 v 比例获得正向结果，则这些预测经过了校准。几十年来，校准一直被视为预测的必要条件 [7]。

* 作者姓名按字母顺序排列。

© Zhun Deng、Cynthia Dwork 和 Linjun Zhang；采用知识共享许可协议 CC-BY 4.0 授权。第14届理论计算机科学创新会议（ITCS 2023）。编辑：Yael Tauman Kalai；文章编号：54；页码：54:1–54:26。莱布尼茨国际信息学论文集，达格施图尔城堡——莱布尼茨信息学中心，达格施图尔出版，德国

LIPICS

要求预测器同时在两个或更多不相交群体上经过校准，即分别单独考察每个群体时，预测器都经过校准，这一要求最初由 Kleinberg、Mullainathan 和 Raghavan 提出，作为一项公平性条件 [31]。受 [31] 启发，为弥合个体公平性 [14],与群体公平性之间的鸿沟，前者要求以相似方式对待相似个体，但需要针对具体任务定义相似性度量；后者则可能只是貌似合理的 [14], ，Hébert-Johnson、Kim、Reingold 和 Rothblum 提出了多重校准，要求在一组（可能规模很大）、预先指定的任意相交的集合上进行校准，这些集合可以在指定的计算类中被识别 [22]。Kearns、Neel、Roth 和 Wu 另行开展了相关研究，考察了针对布尔值分类器的类似情形。他们认为，纳入相交群体可以防止公平性选区操纵，并提出了多重均等 [28]。

这一领域在理论与应用方面都蓬勃发展。例如，多重校准已被用于公平排序 [15], ，也用于为个体概率提供基于不可区分性的解释，即针对不可重复事件的概率 [16]。研究还表明，多重校准可以实现全预测，这意味着对于每个“良好”的损失函数 ℓ，都可以对多重校准预测器赋予的分数进行后处理，无需额外训练，即可在训练分布上达到与$\mathcal{C}$ 中最佳预测器 [19]相媲美的表现。

▶ **定义 1 (**多重校准 [22] 如 [30]所述。令 $\mathcal{C} \subseteq \{[0,1] \times X \to \mathbb{R}\}$ 为一个函数集合。对于支撑集位于 $X \times Y$ 上的给定分布，若预测器 $f: X \mapsto [0,1]$ 满足 $\forall c \in \mathcal{C}$，则称其在 D 上是$(\mathcal{C}, \alpha)$-多重校准的：

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x), x) \cdot (f(x) - y)]\right| \leq \alpha. \tag{1}$$

通过将人口统计学子群体纳入函数类 $\mathcal{C}$，我们刻画了公平性。通过扩充函数类 $\mathcal{C}$ ，使其包含适当的倾向得分重新加权函数，多重校准还能实现目标无关学习：在源域上训练的模型，能够在由这些重新加权函数（近似）刻画的、未见过的未来目标域上表现良好。 [30]。

在本文中，我们将方程（1）中的项（$f(x) - y$）视为一种特定的映射$s(f, y)$: $\mathbb{R} \times \mathcal{Y} \to \mathbb{R}$，并探讨扩充后的映射类所具备的能力。为此，我们提出 *HappyMap*，这是一种多重校准的推广，可带来广泛的新应用，包括一种用于公平不确定性量化的新算法、一种在分布偏移（又称协变量偏移）下进行保形预测的新技术，以及一种分析缺失数据的不同方法；同时，它还为若干现有的、表面上彼此无关的算法公平性概念和目标无关学习方法提供了统一的理解。

我们给出一个统一的 *HappyMap* 元算法，涵盖所有这些结果，并给出其成功的充分性条件。粗略地说，要求映射具有一个满足类似光滑性假设的原函数（见第 3 节）。我们称这样的映射为快乐的。宽泛地说，原函数充当势函数，由此可得算法迭代次数的上界。

贡献总结。

1. 我们提出*HappyMap*，这是一种多重校准的泛化方法，通过丰富映射类（即公式 2 中术语 $(f(x) - y)$ 的替代项），如上文所述；并提出一种 *HappyMap* 元算法，其运行时间和样本复杂度与文献中的其他多重校准算法相当，同时给出该算法成功的充分条件（第 3 节）。
2. 我们首先将 *HappyMap* 用于获得广义版本的公平不确定性量化 [37] ，以此展示

3. 此外，我们将$HappyMap$ 应用于目标无关学习中的问题，这些问题超出了 [30], 所考虑的统计估计问题，旨在获得目标无关的统计推断和不确定性量化。我们的方法还为分析缺失数据提供了富有成效的新视角，并为这一问题带来了新的解决方案（第5节）。

高层次视角。 Hébert-Johnson等人 [22]的开创性工作长期以来一直让我们对一种机器学习新范式的设想心驰神往。我们将微学习 一词用于描述梯度下降：这是一种基于损失最小化、由一系列局部模型更新构成的学习范式。相比之下，多重准确性和多重校准算法通过与审计器（借助弱不可知学习）交互来生成预测器；审计器能够在没有损失函数介入的情况下发现大型问题区域，并据此进行幅度相应较大的模型更新。这启发我们提出了所谓的宏学习。这种视角有一个引人注目的透明性特点：“学习算法找到重大错误并加以纠正”；当审计器审查人口统计学子群体受到的待遇时，这一点尤其令人称道。宏学习可以用作后处理这一事实 [22] 告诉我们，这两种范式可以协同发挥作用。本文提出的 s-Happy 多重校准概念，推动了宏学习的宏大愿景。

2 预备知识

2.1 符号约定

对于 $d \in \mathbb{N}+$，任意凸集 $S \subseteq \mathbb{R}^d$ 和向量 $v \in \mathbb{R}^d$，我们用 $\Pi_S(v)$ 表示在欧几里得距离下 v 在 S 上的投影。有时我们也用概率记号 $\mathbb{P}$表示密度函数，例如，$\mathbb{P}(x)$ 表示连续随机向量 x在 x 处的密度函数值。我们将 $X \in \mathrm{X}$ 记作特征向量（通常 $\mathrm{X} = \mathbb{R}^d$），将 $Y \in \mathrm{Y}$（回归问题中通常 $\mathrm{Y} = \mathbb{R}$，分类问题中通常 $\mathrm{Y} = \{0, 1\}$）记作我们要预测的响应变量。对于 $(x, y) \in \mathrm{X} \times \mathrm{Y}$ 的联合分布，我们分别将 x 和 y 的边缘分布记作 $\mathrm{D}^{\mathcal{X}}$ 和 $\mathrm{D}^{\mathcal{Y}}$。对于两个正数列 $\{a_k\}$ 和 $\{b_k\}$，如果$\lim_{k\to\infty}(a_k/b_k) < \infty$ 且 $\lim_{k\to\infty}(a_k/b_k) = 0$，则记作 $a_k = \mathrm{O}(b_k)$（或 $a_n \lesssim b_n$）以及 $a_k = o(b_k)$。$\mathrm{O}(\cdot)$ 表示忽略对数因子的项。如果 $a_k \lesssim b_k$ 且 $b_k \lesssim a_k$，我们也记作 $a_k = \Theta(b_k)$（或 $a_k \asymp b_k$）。我们用 $\mathrm{O}_P(\cdot)$ 表示随机有界性：若对于某个实数列 a_k，随机变量 $Z_k = \mathrm{O}_P(a_k)$，则表示对于 $\forall \epsilon > 0$，存在 $M, N > 0$，使得当 $k > N$时，$\mathbb{P}(|Z_k/a_k| > M) \leq \epsilon$。对于两个数 $a < b$，我们用记号 $U(a, b)$ 表示在 $[a, b]$上的均匀分布。对于 $\mu \in \mathbb{R}$和 $\sigma > 0$，我们用 $N(\mu, \sigma^2)$ 表示均值为 μ、方差为 σ^2的正态分布。

3 s-快乐多重校准

现在，我们正式阐述多重校准的推广形式。

▶**定义 2** (s-快乐多重校准)。设 $C \subseteq \{\mathbb{R} \times X \to \mathbb{R}\}$ 为一个函数集合。对于一个支撑于 $X \times Y$ 上的给定分布，以及一个映射 $s: \mathbb{R} \times X \mapsto \mathbb{R}$，若预测器$f: X \mapsto \mathbb{R}$ 满足以下条件，则称其在 D 和 s 上关于 $c \in C$ 是(C, α)-s-快乐多重校准的：

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x), x) \cdot s(f(x), y)]\right| \leq \alpha. \tag{2}$$

当上下文已明确给出$\mathcal{D}$ 和 s 时，我们也会直接称 f 为（$\mathcal{C}$, α）-愉悦多重校准。有时我们会将 f 的取值范围限制在某个凸集 O中，例如

$[0,1]$；此时，我们称$\mathcal{X} \mapsto O$为（C, α）-s-愉悦多重校准，当它满足式（2）时即是如此。当 $c(f(x)\ x)$与 f无关且仅取决于 x时，我们将其简写为 $c(x)$。

s-Happy 多重校准（公式 (2)）涵盖了文献中的若干多群体公平性概念。例如，若令 $c(f(x)\ x)$ 与 f 独立，并取 $s(f(x)\ y) = f(x) - y$，即可得到多重准确性 [29, 22]；若对某些函数 $\tilde{c}$和 w，取 $c(f(x)\ x) = \tilde{c}(x) \cdot w(f(x))$，并令 $s(f(x)\ y) = f(x) - y$，即可得到 [20] 中的低阶多重校准概念；若令 $s(f(x)\ y)$ 为非负损失函数，并令 $c(x)$ 为不同群体的指示函数，即可得到 [10] 中的极小极大群体公平性。

***HappyMap* 元算法**。现在我们介绍*HappyMap* 元算法，并证明其关键性质。为简单起见，我们介绍算法的总体 版本，在该版本中，我们可以访问 $\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x)\ x)s(f(x)\ y)]$。当然，在实践中，我们需要根据训练数据估计这些量，因此我们会在附录 B.1 中介绍算法 1 的样本版本，并推导相应所需的样本复杂度。遵循多群体公平性文献 [22, 29], 的做法，我们可以在每次迭代中使用新的样本；或者在样本有限时，采用自适应数据分析[11, 12, 13] 中的技术，在每次迭代中重复使用相同的样本，如 [22] 所建议。

我们考虑一般情形，目标是针对给定的凸集 O，返回一个 $f: \mathcal{X} \mapsto O$，使其满足（$\mathcal{C}\ \alpha$）-快乐多重校准。我们不妨假设类$\mathcal{C}$是对称的，即 c 和 $-c$ 均属于$\mathcal{C}$（我们总可以通过将$-c$ 加入$\mathcal{C}$ 来扩充它，其中 $c \in \mathcal{C}$，因此这一假设不失一般性）。算法以初始预测器 f_0作为输入，该预测器可以是平凡的（对所有 $x \in \mathcal{X}$，$f_0(x) = c$），也可以是融入了丰富领域知识的定制预测器。

算法 1 HappyMap

输入：容差 $\alpha > 0$、步长 $\eta > 0$、迭代次数上限 $T \in \mathbb{N}_+$、初始预测器 $f_0(\cdot)$、分布 D、凸集 $O \subseteq \mathbb{R}$、映射s： $\mathbb{R} \times \mathsf{Y} \to \mathbb{R}$令 $t = 0$当 $t < T$ 且 $\exists c_t \in \mathsf{C}$： $\mathbb{E}_{(x,y)\sim\mathcal{D}}[c_t(f_t(x)\ x)s(f_t(x)\ y)] > \alpha$ 时，执行：令 c_t 为满足 while 语句中条件的 C 的任意元素 $\forall x \in \mathsf{X}$，$f_{t+1}(x) = \Pi_O\left[f_t(x) - \eta \cdot c_t(f(x)\ x)\right]t = t + 1$结束输出： $f_t(\cdot)$

与先前的工作一样， [22, 37, 30, 28], 我们不讨论弱学习器的计算复杂度。弱学习器负责搜索满足 while 循环条件的函数 $c_t \in \mathcal{C}$ s，或报告不存在这样的函数。该问题至少与无偏学习一样难 [22, 28]。关于此问题及其对公平性的影响，请参见 [5] 。下文中，我们以调用弱学习器的次数作为计算复杂度的单位，即 while 循环的一次迭代。

分析。下面的定理 3 总结了算法 1 的理论保证。与文献中的常见做法一样，终止性证明依赖于势函数论证（参见，例如， [4, 22]）。其中关键的新要素是 *happy* 映射的概念，它给出了我们能够找到所需势函数的条件。

我们需要一些假设，并将在陈述定理之前说明并讨论这些假设。三**个假设**。 (a) 对所有 $c \in \mathcal{C}\ x \in \mathcal{X}$， $\mathbb{E}_{x\sim\mathcal{D}^{\mathcal{X}}}[c^2(f(x)\ x)] \le B$，其中 $\mathcal{D}^{\mathcal{X}}$是$\mathcal{D}$ 在 x上的边缘分布；(b) 存在一个势函数 $\mathcal{L}$： $\mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}$，以及常数 $C^l_{\mathcal{L}}, C^u_{\mathcal{L}}$、正数常数 $\kappa_{\mathcal{L}} > 0$，使得 (1) 对所有 $f, \tilde{f}: \mathcal{X} \mapsto \mathbb{R}$，

$C^l_{\mathcal{L}} \le \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathrm{L}(f(x)\ y)]$, (2) 对初始化 $-f_0$, $C^l_{\mathcal{U}} \ge \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathrm{L}(f_0(x)\ y)]$，且 (3)~ ~ $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathrm{L}(f(x)\ y)\ \tilde{f}\ x\ ^2]$ $y \in$ $v \in \mathbb{R}$ $\Pi_O[v]\ y\ \le\ v\ y$ L(f $(x)\ y)] \ge \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x)\ -\ f(x))s(f(x)\ y)] - \kappa_{\mathcal{L}}\mathbb{E}_{x\sim\mathcal{D}^x}[(f(x)\ (\ ))$; (c) 对所有 Y 和，L(,) L(,)。

假设 (a) 是常规的，而假设 (c) 表明，关于第一坐标进行投影时，势函数会减小。一个具体例子是 $O = \mathcal{Y} = [0,1]$ and $\mathcal{L}(v,\ y) = |v-y|^2$ 的情形。

现在我们来讨论假设 (b)，重点关注如何构造势函数 $\mathcal{L}$。首先要指出的是，该假设与光滑性密切相关；光滑性是优化中广泛使用的概念。我们使用以下事实。

事实。若 $\mathcal{L}(\cdot,\ \cdot) : \mathbb{R}\times\mathcal{X} \mapsto \mathbb{R}$ 是关于第一坐标 $2\kappa_{\mathcal{L}}$-光滑的，即 $\mathcal{L}(\cdot,\ \cdot)$ 关于第一坐标连续可微，且相应的偏~导数是 $2\kappa_{\mathcal{L}}$-利普希茨的，那么，对于所有 $x\in\mathcal{X}$ $y\in\mathcal{Y}$ 和 f, $f\colon\mathcal{X}\mapsto\mathbb{R}$,

$$\mathcal{L}(f(x),y) - \mathcal{L}(\tilde{f}(x),y) \ge \partial_u\mathcal{L}(u,y)|_{u=f(x)}(f(x)-\tilde{f}(x)) - \kappa_{\mathcal{L}}(f(x)-\tilde{f}(x))^2. \tag{3}$$

因此，如果 $s(f(x)\ y)$ 关于其第一坐标可微，且 $|\partial_u s(u,y)| \le \kappa_{\mathcal{L}}$，我们可以将势函数取为 s 的原函数，具体而言，对任意 $g\in\mathbb{R}$，令 $\mathrm{L}(f(x)\ y) = \int^{f(x)\ g}\ ^s(u,y)du$，只要我们能够确保对所有 $f(x)$ 和 y，都有 $\int^{f}(x)\ g\ \ ^s(u,y)du \ge C^l_{\mathcal{L}}$。然后，对该等式两边以及 (3) 取期望，即可满足 (b)。事实上，即使函数 $s(u,y)$ 关于 u 处处不可微，我们仍然可以遵循上述一般思路构造 L，并且关于 L 的期望仍可满足假设 (b)。例如，在第 5.2 节中，当 $s(u,y) = \mathbb{1}\{u\le y\} - (1-\delta)$ 时，令 $\mathrm{L}(f(x)\ y) = (1-\delta)\cdot f(x) - \min(f(x) - y, 0)$，即可满足所需条件。

▶**定理 3.** 在上述假设 *(a)-(c)* 下，且若 C 是对称的[2]，，则算法 *1* 在适当选择 $\eta = O(\alpha/(\kappa_{\mathcal{L}}B))$ 后，经过 $T = O((C^u_{\mathcal{L}} - C^l_{\mathcal{L}})\kappa_{\mathcal{L}}B/\alpha^2)$ 次迭代收敛，并输出满足以下条件的函数 f

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x),x)s(f(x),y)]\right| \le \alpha.$$

定理 3 的证明. 首先，由于类别 $\mathcal{C}$ 是对称的，这意味着只要我们能证明对所有 $c\in\mathcal{C}$,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f_t(x),x)s(f_t(x),y)] \le \alpha,$$

对于算法 1 的输出 $f_t(\cdot)$，我们也有对所有 $c\in\mathcal{C}$,

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f_t(x),x)s(f_t(x),y)]\right| \le \alpha.$$

根据我们的假设，存在一个势函数 $\mathcal{L}$、一个常数 $C_{\mathcal{L}}$ 和一个非负常数 $\kappa_{\mathcal{L}}$，使得对于任意 $x\in\mathcal{X}$,

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_t(x),y) - \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_{t+1}(x),y) \ge\ & \mathbb{E}_{(x,y)\sim\mathcal{D}}(f_t(x)-f_{t+1}(x))s(f_t(x),y)\\ & - \mathbb{E}_{(x,y)\sim\mathcal{D}}\kappa_{\mathcal{L}}(f_t(x)-f_{t+1}(x))^2.\end{aligned}$$

因此，我们有

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_t(x),y) - \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_{t+1}(x),y)\\ \ge&\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_t(x),y) - \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}\big(f_t(x)-\eta c_t(f_t(x),x),y\big)\\ \ge&\mathbb{E}_{(x,y)\sim\mathcal{D}}\eta c_t(f_t(x),x)\cdot s(f_t(x),y) - \kappa_{\mathcal{L}}\mathbb{E}_{(x,y)\sim\mathcal{D}}(\eta c_t(f_t(x),x))^2.\end{aligned}$$

[2] 请注意，对称性假设不失一般性。

第一个不等式是由我们的假设中的 (c) 得出的，第二个不等式是由我们的假设中的 (b) 得出的。

给定 $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[c^2(f(x)\ x)] \le B$，若存在 $c_t \in \mathcal{C}$，则 $\mathbb{E}_{(x,y)\sim\mathcal{D}}[c_t(f_t(x)\ x)s(f_t(x)\ y)] > \alpha$

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_t(x),y) - \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_{t+1}(x),y)\\ \ge&\eta\mathbb{E}_{(x,y)\sim\mathcal{D}}c_t(f_t(x),x)\cdot s(f_t(x),y) - \kappa_{\mathcal{L}}\mathbb{E}_{(x,y)\sim\mathcal{D}}(\eta c_t(f_t(x),x))^2 \ge \eta\alpha - \kappa_{\mathcal{L}}\eta^2 B.\end{aligned}$$

取 $\eta = \alpha/(2\kappa_{\mathcal{L}}B)$，我们有

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_t(x),y) - \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_{t+1}(x),y) \ge \frac{\alpha^2}{2\kappa_{\mathcal{L}}B}.$$

由于 $C^u_{\mathcal{L}} \ge \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f_0(x)\ y)$ 且 $\mathbb{E}_{(x,y)\sim\mathcal{D}}\mathcal{L}(f(x)\ y \ge C^l_{\mathcal{L}}$ 对所有 f 均成立，根据假设，每次迭代若发生更新，都会至少取得 $\frac{\alpha^2}{2\kappa_{\mathcal{L}}B}$ 的进展。因此，我们知道更新次数至多为 $2\kappa_{\mathcal{L}}B(C^u_{\mathcal{L}} - C^l_{\mathcal{L}})/\alpha^2$。◀

在接下来的章节中，我们始终假设 $\mathcal{C}$ 是对称的。此外，为简单起见，我们假设 $\mathcal{C}$ 关于 L_2-范数是闭合的。

4 应用：预测区间中的算法公平性

保形预测是预测模型中一种广受欢迎的不确定性量化方法。沿着这一思路，Romano等人提出了一种新的群体公平性标准——均等覆盖 [37]，，其目标是构造一个预测区间$\mathcal{I}(x)$，使其在所有相关受保护群体中以相近的概率覆盖 y。更准确地说，给定一组互不相交的受保护人口群体$\mathcal{A} \subset 2^{\mathcal{X}}$，若集值函数$\mathcal{I}:\mathcal{X} \to \mathbb{R}$ 满足均等覆盖，则有

$$\mathbb{P}(y \in \mathcal{I}(x) \mid x \in A) \ge 1-\delta, \text{ for all } A \in \mathcal{A}. \tag{4}$$

本节为简化起见，聚焦于单侧预测区间；将我们的结果应用两次，或应用于非一致性分数，即可得到双侧区间（见下文的推论 5 和注释 6）。此外，与多重校准文献中的通常做法一致，我们的结果即使在人口子群体任意相交的情形下也成立。 [21, 26] 也研究了交叉群体情形的一种略有不同的版本；稍后我们将说明如何使用 HappyMap 来处理这种情形。

在式（4）的单侧版本中，我们令 $\mathcal{I}(x) = (l_\delta(x)\ \infty)$ 为一个单侧（$1-\delta$）预测区间，并研究以下覆盖率准则：

$$\mathbb{P}(x \in A)\cdot|\mathbb{P}(y \ge l_\delta(x) \mid x \in A) - (1-\delta)| \le \alpha, \text{ for all } A \in \mathcal{A}, \tag{5}$$

进而可将其改写为

$$|\mathbb{E}[\mathbb{1}(x \in A)\cdot(\mathbb{1}(y \ge l_\delta(x)) - (1-\delta))]| \le \alpha, \text{ for all } A \in \mathcal{A}.$$

关于问题定义，有几点说明。首先，对于式 (4) 中原始的均等覆盖问题，无论各组是否互不相交，都存在一个平凡的确定性解：只需设置 $\mathcal{I}(x) = \mathbb{R}$.[3] 类似地，虽然式 (5) 中我们的一侧或双侧版本排除了这种做法，但它允许采用如下平凡的随机化解：

[3] 这并不意味着以往的算法微不足道！

$l_\delta(x) = -\infty$ 以概率 $1-\delta$，以及 $l_\delta(x) = \infty$ 以概率 δ。排除这两种平凡解的一种方法，是要求满足一个更强的条件——{style id='9'>多重有效性，该条件由 [21, 26] 提出。沿用我们的记号，与式 (6) 中要求 $|\mathbb{P}(y \geq l_\delta(x) \mid x \in A) - (1-\delta)|$ 较小不同，多重有效性 要求 $|\mathbb{P}(y \geq l_\delta(x) \mid x \in A, l_\delta(x)) - (1-\delta)|$ 较小。与多重校准和多重准确性之间的关系类似，多重有效性比我们在式 (6) 中提出的要求更强。下文将讨论，我们也可以为此使用 HappyMap。在定理 9 中，我们还将提供另一种论证，说明即使不强制要求多重有效性，我们的算法也在做一些非平凡的事情。

下面我们将公式 (5) 推广为以下交叉群体均等覆盖要求：

$$\sup_{c \in \mathcal{C}} \left| \mathbb{E}[c(x)(\mathbb{1}\{y \geq l_\delta(x)\} - (1-\delta))] \right| \leq \alpha, \tag{6}$$

其中，$\mathcal{C}$ 表示任意预先指定的函数集合（包括预先指定子群体的指示函数，以及更一般的连续函数，这在多重校准文献中很常见）。例如，$\mathcal{C}$ 可以是能够通过固定深度的决策树计算的函数。对 $s(l, y) = (1-\delta) - \mathbb{1}\{l \leq y\}$ 应用 HappyMap，可得到以下结果。

▶ **定理 4**。 假设 $y \mid x$ 是连续随机变量[4]，给定 x 时 y 的条件密度上界为 K_p，并且对于某个通用常数 $C > 0$ 有 $|\mathbb{E}[y]| < C$。此外，假设对于所有 $c \in \mathcal{C}$，都有 $\mathbb{E}_{x \sim \mathcal{D}^x}[c^2(x)] \leq B$。那么，对于适当选取的 $\eta = O(\alpha/(K_p B))$，使用势函数 $L(l, y) = (1-\delta) \cdot l - \min(l - y, 0)$ 时，*HappyMap*（算法 *1*）在 $T = O(K_p B^2/\alpha^2)$ 步内收敛，并输出满足以下条件的函数 $l_\delta(\cdot)$

$$\sup_{c \in \mathcal{C}} \left| \mathbb{E}[c(x)(\mathbb{1}\{y \geq l_\delta(x)\} - (1-\delta))] \right| \leq \alpha.$$

定理 4 的证明。 为了应用定理 3，只需验证势函数 $\mathcal{L}(l、\ y) = (1-\delta)l - \min(l - y、0)$ 满足类平滑条件，即

$$\mathbb{E}[\mathcal{L}(l, y) - \mathcal{L}(l', y)] \geq \mathbb{E}[(l(x) - l'(x))s(f(x), y)] - \kappa_L \mathbb{E}[(l(x) - l'(x))^2].$$

特别地，我们有

$$\begin{aligned}
\mathcal{L}(l, y) - \mathcal{L}(l', y) =& (1-\delta)(l - l') + \min(l' - y, 0) - \min(l - y, 0) \\
=& (1-\delta)(l - l') + (l' - l)\mathbb{1}\{l - y, l' - y < 0\} - (l - y)\mathbb{1}\{1' - y > 0 > l - y\} \\
&+ (l' - y)\mathbb{1}\{1' - y < 0 < l - y\} \\
=& (1-\delta)(l - l') + (l' - l)\mathbb{1}\{l - y < 0\} - (l' - y)\mathbb{1}\{1' - y > 0 > l - y\} \\
&+ (l' - y)\mathbb{1}\{1' - y < 0 < l - y\} \\
=& (l - l')((1-\delta) - \mathbb{1}\{l - y < 0\}) + (l' - y)(\mathbb{1}\{1' < y < l\} - \mathbb{1}\{1' > y > l\}).
\end{aligned}$$

我们有

$$|(l' - y)(\mathbb{1}\{1' < y < l\} - \mathbb{1}\{1' > y > l\})| \leq |l - l'| \cdot (\mathbb{1}\{1' < y < l\} + \mathbb{1}\{1' > y > l\})).$$

上述不等式源于这样一个事实：如果一个 $|l - l'|$ 扰动能够改变 $l' - y$ 的符号，那么 $|l' - y| < \eta |c(x)|$。

[4] 我们的分析可以直接扩展到以下情况： y 以有限精度存储，且 α 取值大于精度误差

假设 y 在给定 x 条件下的条件密度以 K_p 为上界，则有

$$\mathbb{E}[|(l'-y)(\mathbb{1}\{\ell'<y<l\}+\mathbb{1}\{\ell'>y>l\})|]\le L\eta^2\mathbb{E}[c(x)^2]\le K_p|l-l'|^2.$$

此外，我们验证了 $\mathbb{E}[\mathrm{L}(l,y)]$ 具有均匀的下界。我们分情况讨论。若$l-y<0$（即 $l<y$），则有 $\mathrm{L}(l,y)=(1-\delta)\cdot l-\min(l-y,0)=(1-\delta)\cdot l-l+y=y-\delta\cdot l>(1-\delta)\cdot y$；若 $l-y>0$（即 $l>y$），则有 $\mathrm{L}(l,y)=(1-\delta)\cdot l-\min(l-y,0)=(1-\delta)\cdot l>(1-\delta)\cdot y$.由于根据假设，$|\mathbb{E}[y]|<C$ 对某个通用常数 $C>0$ 成立，因此我们有

$$\mathbb{E}[\mathcal{L}(l,y)]\ge-(1-\delta)C.$$

◀

以下推论可由定理 4 直接得出。

▶推论 *5.* 如果我们将上述算法（即针对当前设置专门化的 *HappyMap*）应用于两个不同的截断值 $\delta/2$ 和 $(1-\delta/2)$，并分别得到 $l_{\delta/2}(x)$ 和 $l_{1-\delta/2}(x)$，那么我们得到双侧预测区间 $I(x)=[l_{\delta/2}(x)\ l_{1-\delta/2}(x)]$，使得 $\sup_{c\in\mathcal{C}}\mathbb{E}[c(x)(\mathbb{1}\{y\in I(x)\}-(1-\delta))]\le 2\alpha$。

▶注释 6. 在保形预测文献中，还有另一种常用方法，可基于非一致性分数构造双侧预测区间。具体而言，非一致性分数 $m(x,\ y)$ 是一种度量，用于衡量响应变量 y 与预测结果 $h(x)$“不一致”的程度，其中 h 是任意固定的预测函数。例如，回归任务中一种常用的非一致性分数是 $m(x,\ y)=|y-h(x)|$。其他非一致性分数的选择可见 [40, 1]。给定非一致性分数$m(x,\ y)$，集合 $\{y\colon m(x,\ y)\le f(x)\}$ 将自然地给出 y的双侧区间。为此，直接将我们的方法应用于 $(x,\ \tilde{y})$，并令 $\tilde{y}:=m(x,\ y)$，即可得到有效的 $(1-\delta)$ 预测区间。

回想一下，多重有效性 [21, 26] 界定了 $|\mathbb{P}(y\ge l_\delta(x)\mid x\in A, l_\delta(x))-(1-\delta)|$，对所有$A$ 以及 $l_\delta(x)$ 的取值均成立。与多重校准和多重准确性的关系类似，多重有效性通常比公式 (6) 中的要求更强，因此可能会导致长得多的预测区间（更详细的讨论见 E.3 节）。考虑具有 $c(l_\delta(x)\ x)=\mathbb{1}\{l_\delta(x)\in I\}$ 形式的 c，其中 $I\in\mathrm{I}$，而 I 是小区间箱的集合 $\mathrm{I}=\{[-C,-C+\lambda],[-C+\lambda,-C+2\lambda],\ldots,[C-2\lambda,C-\lambda],[C-\lambda,C]\}$，对应某个离散化级别 λ，且 C 是 $|y|$ 的上界，s-Happy 多重校准即可恢复多重有效性；应用算法 1 可得到一种实现多重有效性的新算法。

[18] 中也考虑了将 (4) 扩展到相交集合的情形。他们在收集训练数据之后，将集合族 $\mathcal{A}$ 事后定义为训练集的所有足够大的子集，然后枚举所有这些（数量呈指数级增长的）集合。与此对比，按照多重校准文献中的惯例，我们预先指定这些集合，并依赖弱学习；当集合族$\mathcal{A}$具有特殊结构时，即使它是无限的，这种方法也可能比穷举搜索更高效。

HappyMap 会生成一个预测区间，该区间对于大量群体都具有公平性。接下来，我们将分析将 HappyMap 用作高质量初始保形映射的后处理时，其效用（即所构造预测区间的宽度）；该初始映射不一定具有群体公平性 l_0。我们不限定 l_0的来源：它可以由他人提供，也可以通过拆分样本并使用任何标准方法构建一个高质量但不考虑公平性的保形映射来获得。

为便于分析，我们首先介绍关于分位数的以下两个定义。这些定义是标准的，遵循分位数估计领域的相关文献，参见 [6, 41]及其中的参考文献。

▶ 定义 7. 对于一个分布 Q，其 *supp* $Q \subset [-C, C]$ 满足某个通用常数 $C > 0$，记 Q 的 τ-分位数为 $F_\tau(Q) := \{t \in \mathbb{R}: Q((-\infty, t]) \geq \tau$ 且 $Q([t, \infty)) \geq 1-\tau\}$。当 $F_\tau(Q)$ 是一个区间时，我们记 $t_{\min}(Q)$ 为 $= min\ F_\tau(Q)$，记 $t_{\max}(Q)$ 为 $= max\ F_\tau(Q)$。Q 称其具有类型为 $q \in (1, \infty)$ 的 τ-分位数，如果存在常数 $\alpha_Q > 0$ 和 $b_Q > 0$，使得对所有 $s \in [0, \alpha_Q]$，

$$Q(t_{\min} - s, t_{\min}) \geq b_Q s^{q-1}$$
$$Q(t_{\max}, t_{\max} + s) \geq b_Q s^{q-1}.$$

由于我们关注的不是 $\mathbb{R}$ 上的单个分布 Q，而是 $\mathcal{X} \times \mathbb{R}$上的多个分布 P，因此下面的定义将前面的定义扩展到此类 P。

▶ 定义 8. 设 $p \in (0, \infty]$, $q \in (1, \infty)$，且 P是 $X \times \mathbb{R}$ 上的一个分布。若存在集合 $\Omega_{\mathcal{X}} \subset X$，使得 $\mathbb{P}(x \in \Omega_{\mathcal{X}}) = 1$,$supp(P(\cdot|x)) \subset [-C, C]$ 对某个通用常数 $C > 0$成立，且对所有任意 $x \in \Omega_{\mathcal{X}}$， $P(\cdot|x)$ 都具有 τ-分位数类型 q，并且存在定义 7 中定义的 $\alpha_{P(\cdot|x)}$ 和 $b_{P(\cdot|x)}$，使得$\mathbb{E}[|b_P^{-p}(\cdot|x)(x) \cdot \alpha_P^{(1-q)p}(\cdot|x)(x)|] < \infty$，则称其具有 τ-分位数的 p-平均类型 q。

接下来，我们给出如下结果，表明应用 HappyMap 几乎不会增加预测区间的宽度：如果输入 l_0 接近最优分位数函数，那么 HappyMap 算法可以保留这一近似。

▶ **定理 9.** 假设定理 *4* 中的条件成立，且 P 是 $X \times Y$ 上的分布，具有 δ-分位数 p-平均类型 q。设 $pq/(p+1) \geq 2$。令 $l_\delta^*(x) \in F_\delta(P(\cdot \mid x))$ 为 $Y \mid X = x$ 的 δ-分位数中的任意元素。若输入 l_0满足 $\mathbb{E}_{x \sim P_X}(l_0(x) - l_\delta^*(x))^2 \leq \beta$，则存在一个通用常数 C，使得 *HappyMap* 算法 *1* 的输出 l_δ 满足

$$\mathbb{E}(l_\delta(x) - l_\delta^*(x))^2 \leq C\beta.$$

定理 9 的证明。 回顾一下，在将 HappyMap 应用于当前问题时，我们使用势函数 $\mathcal{L}(l, y) = (1-\delta)l - \min(l-y, 0)$。根据定理 4 的证明中的推导，我们有

$$\begin{aligned}
&|\mathbb{E}[\mathcal{L}(l, y)] - \mathbb{E}[\mathcal{L}(l', y)]| \\
&\leq |\mathbb{E}[(l - l')((1-\delta) - \mathbb{1}\{l - y < 0\})]| + |\mathbb{E}[(l' - y)(\mathbb{1}\{1' < y < l\} - \mathbb{1}\{1' > y > l\})]| \\
&\leq |\mathbb{E}[(l - l')((1-\delta) - \mathbb{1}\{l - y < 0\})]| + K_p \cdot \mathbb{E}[|l - l'|^2].
\end{aligned}$$

由于 $l_\delta^*(x) = \arg\min_l \mathbb{E}[\mathcal{L}(l, y) \mid x]$，且 $\mathbb{P}(l_\delta^*(x) < y \mid X = x) = 1 - \delta$，因此我们有

$$\mathbb{E}[\mathcal{L}(l_0, y)] - \mathbb{E}[\mathcal{L}(l_\delta^*, y)] \leq K_p \cdot \mathbb{E}|l_0(x) - l_\delta^*(x)|^2 \leq K_p\beta.$$

此外，根据定理 4 的证明，我们有 $\mathbb{E}[\mathcal{L}(l, y)] \leq \mathbb{E}[\mathcal{L}(l_0, y)]$,，这意味着

$$\mathbb{E}[\mathcal{L}(l, y)] - \mathbb{E}[\mathcal{L}(l_\delta^*, y)] \leq K_p\beta.$$

然后，根据 [41], 中的定理 2.7，并利用当 $pq/(p+1) \geq 2$ 时，函数 L_2 范数被函数 $L_{pq/(p+1)}$ 范数所支配这一事实，我们得到

$$\mathbb{E}[|l(x) - l_\delta^*(x)|^2] \leq C' \cdot \mathbb{E}[\mathcal{L}(l, y)] - \mathbb{E}[\mathcal{L}(l_\delta^*, y)] \leq C' K_p\beta.$$

◀

上述定理表明，当条件分布 $Y \mid X = x$ 是单峰且对称的[5]，，并且我们拥有一个近似误差为 $\beta = o(1)$的良好初始化函数 f_0 时，那么按照推论 5 或注释 6 中任一种方式构造的预测区间，其宽度都将收敛到最优（$1-\delta$）预测区间 $[l^*_{\delta/2}(x)\ l^*_{1-\delta/2}(x)]$ 的宽度，此时我们已知分布 $Y \mid X = x$ 的确切形式。我们注意到，可以先将数据拆分为两半，从而有可能实现 $\beta = o(1)$：在前一半数据上训练 l_0，再用 HappyMap 处理后一半数据。这种数据拆分在保形预测文献 [32, 38, 33]中很常见。

5 应用：目标无关学习

Kim 等人 [30] 证明，使用一组合适的倾向性重加权函数$\mathcal{C}$，在源数据$\mathcal{D}_{so}$ 上训练一个（$\mathcal{C}$ α）-多重校准预测器，可以在从此前未见过的目标分布$\mathcal{D}_{ta}$ [30]抽取的数据上，高效地估计感兴趣的统计量。本节中，我们使用 HappyMap 将这些结果扩展到统计估计问题 [30]之外的目标无关学习问题：目标无关预测和不确定性量化。我们的方法还为分析缺失数据带来了富有成效的新视角，并为这一问题提供了新的解决方案。

我们用$\mathcal{Z} = \{so, ta\}$ 分别表示从源分布和目标分布中抽样。如 [30], 中一样，我们考虑定义在（x, y, z）三元组上的联合分布$\mathcal{D}_{joint}$。源总体$\mathcal{D}_{so}$ 和目标总体$\mathcal{D}_{ta}$ 可以分别视为在 $z = ta$ 或 $z = so$ 条件下，定义在（x, y）上的联合分布。类似地，我们用 $D^{\mathcal{X}}_z$ 和 $D^{\mathcal{Y}}_z$ 表示通过对 $z \in \mathcal{Z}$进行条件化得到的边缘分布。如 [30],中一样，我们假设 x 与 y 之间的某种函数关系在分布$\mathcal{D}_{so}$ 和$\mathcal{D}_{ta}$上相同；这在因果推断文献中称为“可忽略性”，在机器学习中称为“协变量偏移”。形式化地说，我们有：

▶ **假设 10**（协变量偏移假设）。 对于一个从 $\mathcal{D}_{joint}$ 中抽取的三元组 (x, y, z)，

$$\mathbb{P}_{(x,y,z)\sim\mathcal{D}_{joint}}(x, y, z) = \mathbb{P}(x)\mathbb{P}(y|x)\mathbb{P}(z|x).$$

根据假设 10，我们有

$$\mathbb{P}_{(x,y)\sim\mathcal{D}_{so}}(y|x) = \mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(y|x).$$

按照惯例且不失一般性，在本节中，我们假设对$\mathcal{Z} = \{so, ta\}$采用均匀先验，即 $\mathbb{P}(z = ta) = \mathbb{P}(z = so)$。

进行协变量偏移研究的一种常用工具是倾向得分重加权。倾向得分定义为 $e(x) = \mathbb{P}(z = so|x)$，倾向评分比 $\frac{1-e(x)}{e(x)} = \frac{\mathbb{P}(z=ta|x)}{\mathbb{P}(z=so|x)}$ 可用于将对 $\mathrm{D}^{\mathcal{X}}_{so}$ 的期望转换为对 $\mathrm{D}^{\mathcal{X}}_{ta}$ 的期望。然而，在训练时未观察到来自 D_{ta} 的样本时，我们无法估计倾向评分比。Kim 等人 [30] 提出了针对函数类 $\mathsf{C} = \{\frac{1-\sigma(x)}{\sigma(x)} : \sigma \in \Sigma\}$进行多重校准，其中函数族 Σ（希望）能够涵盖所关注的倾向评分比。他们证明，针对该函数类进行多重校准（事实上，即使是针对该函数类进行多重准确性校准），也能确保所得预测器的估计精度不逊于函数类 C 中最佳的倾向得分重加权函数。这一概念被他们称为普适适应性。在可实现情形下，如果未见目标域的倾向评分比属于函数类 C，则保证 f 能够对目标域上所关注的统计量给出良好估计。

[5] 保形预测文献中也采用了类似的假设，参见，例如， [32, 33]

5.1 普适自适应预测器在 ℓ_2 损失下

作为一项热身练习，我们考虑 预测器 $f: \mathcal{X} \to [0,1]$ 在 ℓ_2 损失下的普适适应性；这并不需要 HappyMap 的全部能力。与协变量偏移下的统计估计相比，这一问题更为复杂，需要使用更复杂的函数类 $c(f(x), x)$。

形式上，我们的目标是获得一个预测 f，使估计误差较小

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}(f(x)-\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}[y|x])^2$$

对于 $\mathcal{Y} = [0,1]$。我们注意到，这个量通常用作衡量预测函数质量的指标。根据假设 10（协变量偏移），$\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x] = \mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}[y|x])$，从而得到

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}(f(x)-\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}[y|x])^2 &= \mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}(f(x)-\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}[y|x])(f(x)-y)\\ &= \mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}(f(x)-\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])(f(x)-y).\end{aligned}$$

利用倾向评分比，我们有

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}(f(x)-\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])(f(x)-y) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\left(\frac{1-e(x)}{e(x)}(f(x)-\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])\right)(f(x)-y).$$

训练时，我们看不到来自$\mathcal{D}_{ta}$的样本，因此无法像经典重加权方法那样从样本中估计 $e(\cdot)$。不过，按照 [30], ，我们可以使用一个类 $\{\frac{1-\sigma(x)}{\sigma\ x}() : \sigma \in \Sigma\}$ 来表示感兴趣的倾向评分比，并尝试找到一个误差较小的 $f: \mathbb{R} \mapsto [0,1]$：

$$\sup_{\sigma\in\Sigma}\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\left(\frac{1-\sigma(x)}{\sigma(x)}(f(x)-\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])\right)\cdot(f(x)-y).$$

这样基本上就得到了我们在式 (2) 中需要的形式；只需再处理 $\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x]$。为此，我们引入一个类 $\mathcal{P} = \{p : \mathcal{X} \mapsto [0,1]\}$，希望其中包含对 $\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x]$ 的良好近似，并尝试找到一个 $f: \mathbb{R} \mapsto [0,1]$，使得

$$\left|\sup_{\sigma\in\Sigma,p\in\mathcal{P}}\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\left(\frac{1-\sigma(x)}{\sigma(x)}(f(x)-p(x))\right)\cdot(f(x)-y)\right| \le \alpha, \tag{7}$$

对于某个较小的值 α。换言之，我们取 $\mathcal{C} = \{c(f(x), x) = \pm^{1-\sigma(x)}\sigma(x)(f(x) - p(x)) : \sigma \in \Sigma, p \in \mathcal{P}\}$，其中 $\sigma, p \in (0,1)$。我们将近似误差定义为

$$\beta_1(p) = \sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}(p(x)-\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])^4}$$

and

$$\beta_2(\sigma) = \sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}(\frac{1-\sigma(x)}{\sigma(x)}-\frac{1-e(x)}{e(x)})^4}.$$

尽管我们没有充分发挥 HappyMap 的全部能力，因为我们沿用了原始映射（$f(x) - y$），但仍然可以应用定理 3 得到：

▶ 定理 *11.* 假设 $\sigma(\cdot) \in (c_1, c_2)$，其中 $0 < c_1 < c_2 < 1$。那么，我们有 $|c(f(x), x)| \le 2(1-c_1)/c_1 := B$。令 $\beta = \inf_{\sigma\in\Sigma,p\in\mathcal{P}}\sqrt{2B^2\beta_1(p) + 2\beta_2(\sigma)}$。假设我们使用经过适当选择的 $\eta = O(\alpha/B)$ 运行算法 *1*，则该算法经过 $T = O(2B/\alpha^2)$ 次迭代后收敛，所用的势函数为 $L(f(x), y) = 1/2(f(x) - y)^2$ and $O = [0,1]$，从而得到

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[\frac{1-e(x)}{e(x)}(f_t(x)-\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[y|x])(f_t(x)-y)]\right| \le \alpha+\beta,$$

对于算法 *1* 的输出 $f_t(\cdot)$。

定理 11 的证明推迟至 D.1 节。

5.2 通用自适应保形预测

在本节中，我们应用 HappyMap 实现普适自适应的保形预测：在协变量偏移情形下，利用源数据训练保形预测模型，同时确保其在未见过的目标分布上保持有效性。

回想一下，标准保形预测的目标是为 y 获得一个 $(1-\delta)$ 预测区间。本节为简化表述，考虑单侧区间 $(l(x), \infty)$，并重点讨论响应变量 y 为连续变量的情形[6] 令 $\mathcal{D}_{ta}$ 表示未观测到的目标分布，我们的目标是构造 $l(\cdot)$，使得 $\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x) \le y)$ 接近期望水平 $(1-\delta)$：

$$|\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x) \le y) - (1-\delta)| \le \alpha.$$

注意，上述不等式意味着 $\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x) \le y) \ge 1-\delta-\alpha$，这是保形预测文献中更常见的要求。根据假设 10，可以将该概率改写为：

$$\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x) \le y) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\Big(\frac{1-e(x)}{e(x)}\mathbb{1}\{l(x) \le y\}\Big).$$

现在，本着与上一节（具体而言，式（7））相同的思路，我们寻求 $l(\cdot)$，使得

$$\Big|\sup_{\sigma\in\Sigma}\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[\frac{1-\sigma(x)}{\sigma(x)}\cdot(\mathbb{1}\{l(x) \le y\} - (1-\delta))]\Big| \le \alpha, \tag{8}$$

对于未见过的 $\mathcal{D}_{ta}$ 以及某个 $\alpha > 0$。为此，我们对$\mathcal{C} = \{c(x) = {}^{1-\sigma}{}_{(x)\,\sigma(x)}$ 应用 HappyMap（算法 1）：$\sigma \in \Sigma\}$，并使用映射 $s(l, y) = \mathbb{1}\{l \le y\} - (1-\delta)$。

▶ 定理*12.* 假设 $y \mid x$ 是一个连续随机变量，给定 x 时 y 的条件密度有上界 K_p，并且对于某个通用常数 $C > 0$，$|\mathbb{E}[y]| < C$。设 $\mathbb{E}_x[c^2(x)] \le B, \forall c \in \mathcal{C}$，并考虑满足 $s(l, y) = \mathbb{1}\{l \le y\} - (1-\delta)$ 的 *HappyMap* 元算法（算法 *1*）。那么，对于任意目标分布，其边缘密度函数满足 x, $p_{ta}(x)$，并令 β 满足 $\inf_{c\in\mathcal{C}}\mathbb{E}[(c(x) - \overline{p_{ta}(x)\,p_{so}(x)})^2] \le \beta^2$，存在一个 $\eta = O(\alpha/(K_pB))$ 的取值，使得算法 *1* 在 $T = O(LB/\alpha^2)$ 内终止；此时使用势函数 $L(l, y) = (1-\delta)\cdot l - \min(l-y, 0)$。所得函数 l_δ 满足

$$|\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l_\delta(x) \le y) - (1-\delta)| \le \alpha + \beta.$$

定理12的证明结合了普适自适应性论证与定理 4的分析，推迟至D.2节。

▶ 注释13。据我们所知，Tibshirani 等人 [42] 首次考虑了协变量偏移下的保形预测。他们的方法依赖于对（$1-e$（x））$/e$（x）的精确了解。相比之下，只要 Σ中存在与该分数接近的量，我们无需精确了解（$1-e$（x））$/e$（x），也能实现有效的渐近覆盖保证。

最后需要指出的是，如果扩充$\mathcal{C}$，我们就能将结果推广到普适自适应均等覆盖和普适多重有效性。例如，将$\mathcal{C}$ 中的$\{c(x) = {}^{1-\sigma}{}_{(x)\,\sigma(x)}$：$\sigma \in \Sigma\}$ 和子群体指示函数一并纳入，就能在协变量偏移下实现均等覆盖。

[6] 与第 4 节一样，我们可以通过两次应用我们的方法处理双侧情形（推论 5）。

5.3 缺失数据预测

s-Happy 多重校准也可用于研究缺失数据这一统计数据分析中的常见问题 [35, 43]。假设我们考虑基于特征 $x \in \mathcal{X}$来预测响应变量$y \in [0,1]$ 的学习任务。令 X 表示完整数据矩阵，其中第 i行 x_i 是第 i个观测的特征，响应向量为 $Y = (y_1, ..., y_n)^\top$。假设数据独立同分布，且 $(x_i, y_i) \overset{i.i.d.}{\sim} \mathcal{D}$。在缺失数据问题中，$X$ 中的某些条目缺失，因此我们将 $x_i^{(1)}$ 定义为观测i中 x_i 的已观测分量，将 $x_i^{(0)}$ 定义为 x_i 中缺失的分量。此外，我们令 R_i 表示完整案例的指示变量，也就是说，当且仅当 x_i 被完整观测（即 $x_i^{(1)} = x_i$）时， $R_i = 1$ 成立。

缺失数据机制主要有三种：完全随机缺失（MCAR）、随机缺失（MAR）和非随机缺失（MNAR）。如果缺失指示变量 R 的分布与 X 相互独立，我们就称数据集满足 MCAR。对于 MAR， R 仅通过其已观测分量依赖于 X，即 $R_i \perp\!\!\!\perp x_i^{(0)} | x_i^{(1)}$ 。对于 MNAR， R 的分布可能取决于 X 的缺失分量。接下来将介绍的方法适用于这三种机制。

假设我们的目标是学习一个预测函数 f，使测试均方误差（MSE） $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2]$ 较小。文献中有两种主要方法：加权和插补。加权方法 [34] 最小化以下在完整数据上评估的损失函数：

$$\arg\min_f \sum_{i=1}^{n} \mathbb{1}\{R_i = 1\} w(x_i)(y_i - f(x_i))^2, \tag{9}$$

其中 $w(x_i)$近似于 $\mathbb{P}_{(x_i|R_i=1)}\mathbb{P}(x_i)$。

我们将加权方法调整到我们的框架中，并将 C 用作一个函数类，其中某个 $c(f(x), x) \in$ C 近似于 $\mathbb{P}_{(x|R_i=1)}\mathbb{P}(x)(\mathbb{E}[y|x] - f(x))$。由于 $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2]$ 可以等价地写作 $\mathbb{E}_{(x,y)\sim\mathcal{D}|R=1}[\mathbb{P}_{(x|R_i=1)}\mathbb{P}(x)(\mathbb{E}[y|x] - f(x)) \cdot (y - f(x))]$。不同于像式（9）中那样最小化损失函数，我们旨在找到一个 f，使得 $\sup_{c\in\mathcal{C}} \mathbb{E}_{(x,y)\sim\mathcal{D}|R=1}[c(x)(y - f(x))] \le \alpha$，其中 C 是一个近似于 $\mathbb{P}_{(x|R_i=1)}\mathbb{P}(x)(\mathbb{E}[y|x] - f(x))$ 的函数类。使用这样的函数类 C 和 $s(f, y) = f - y$ 应用 HappyMap 后，我们得到以下结果。

▶ **定理 14.** 假设 $\inf_{c\in\mathcal{C}} \sqrt{\mathbb{E}[(c(f(x), x) - \mathbb{P}_{(x|R_i=1)}\mathbb{P}(x)(\mathbb{E}[y|x] - f(x))]^2} \le \beta$，并且

$$\sup_{c\in\mathcal{C}} \mathbb{E}[c^2(f(x), x)] \le B.$$

考虑 *HappyMap* 元算法（算法 *1*），其中 $s(f, y) = f - y$。那么，存在一个合适的 $\eta = O(\alpha/B)$ 选择，使得算法 *1* 在 $T = O(LB/\alpha^2)$步内终止，使用势函数 $L(f(x), y) = 1/2(f(x) - y)^2$ 和 $O = [0, 1]$。由此得到的函数 l_δ满足

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2] \le \alpha + \beta.$$

定理 14 的证明推迟至 D.3 节。

▶ 注 15. 上述分析考虑了在缺失数据情境下学习均方误差较小的预测器。同样的思路也可用于其他存在缺失数据的情境，包括在部分观测数据缺失时构造预测区间，以及在缺失数据下实现多群体公平性概念（见第 4 节）。

6 讨论与结论

本文提出了 s-Happy 多重校准，这是对多重校准的泛化，它统一了许多现有的算法公平性概念和目标无关学习方法，并带来了一系列新的应用，包括一种用于不确定性量化的新公平性概念、一种在协变量偏移下进行保形预测的新颖技术，以及一种分析缺失数据的不同方法。从更高层面来看，我们推动了宏学习领域的发展。

一个有趣的未来研究方向是将 s-Happy 多重校准扩展到（低维）表示函数。正如深度学习领域近期所指出的，预训练或表征学习是提升最终预测准确率[7]的有力工具。我们希望，将 s-Happy 多重校准恰当地扩展到表征学习领域，能够通过使用扩展后的 HappyMap（对预训练学得的表示进行后处理），进一步促进目标无关学习。此外，我们当前的目标无关学习要求满足协变量偏移假设（即，条件分布 $y \mid x$ 在不同环境之间保持不变）。一个有趣的问题是：当存在其他类型的不变性时，例如例如不变风险最小化（IRM） [2, 8, 9]中的情形，如何利用 s-Happy 多重校准进行目标无关学习。

“多重 X”/宏学习框架（包括 s-Happy 多重校准）的强大之处，源于选择了一个丰富的集合$\mathcal{C}$。然而，丰富的$\mathcal{C}$ 意味着弱无关学习器面临更困难的任务，且样本复杂度会增加。这促使我们研究如何降低函数类复杂度（第 F 节）。为了充分发挥宏学习的潜力，我们需要理解 $\mathcal{C}$ 的能力与弱学习器计算复杂度之间的权衡，即使只是从启发式方法的角度来理解。

参考文献

1 Anastasios N Angelopoulos 和 Stephen Bates。保形预测与分布无关不确定性量化简介。*arXiv* 预印本 *arXiv:2107.07511*，2021年。2 Martin Arjovsky、Léon Bottou、Ishaan Gulrajani 和 David Lopez-Paz。不变风险最小化。*arXiv* 预印本 *arXiv:1907.02893*，2019年。3 Osbert Bastani、Varun Gupta、Christopher Jung、Georgy Noarov、Ramya Ramalingam 和 Aaron Roth。实用的对抗式多重有效保形预测。*arXiv* 预印本 *arXiv:2206.01067*，2022年。4 Sébastien Bubeck 等。凸优化：算法与复杂性。《基础与趋势》® 机器学习，8(3-4):231–357，2015年。5 Maya Burhanpurkar、Zhun Deng、Cynthia Dwork 和 Linjun Zhang。脚手架集。*arXiv* 预印本 *arXiv:2111.03135*，2021年。6 Andreas Christmann 和 Ingo Steinwart。SVM 如何估计分位数和中位数。神经信息处理系统进展，20，2007年。7 A Philip Dawid。校准良好的贝叶斯学家。《美国统计学会杂志》，77(379):605–610，1982年。8 Zhun Deng、Frances Ding、Cynthia Dwork、Rachel Hong、Giovanni Parmigiani、Prasad Patil 和 Pragya Sur。通过表示实现表示：通过对抗式学习不变表示实现领域泛化。*arXiv* 预印本 *arXiv:2006.11478*，2020年。9 Zhun Deng、Linjun Zhang、Kailas Vodrahalli、Kenji Kawaguchi 和 James Zou。更好的表示有助于迁移学习的对抗训练。*NeurIPS 2021*，2021年。

[7] 例如，在 https://arxiv.org/pdf/2204.02937.pdf 的表 5 中，预训练使准确率从 60% 提高到 90+%。

10 Emily Diana、Wesley Gill、Michael Kearns、Krishnaram Kenthapadi 和 Aaron Roth。极小极大群体公平性：算法与实验。载于《2021年AAAI/ACM人工智能、伦理与社会会议论文集》，第66–76页，2021年。11 Cynthia Dwork、Vitaly Feldman、Moritz Hardt、Toni Pitassi、Omer Reingold 和 Aaron Roth。自适应数据分析与留出集复用中的泛化。神经信息处理系统进展，28，2015年。12 Cynthia Dwork、Vitaly Feldman、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Aaron Roth。可复用留出集：在自适应数据分析中保持有效性。《科学》，349(6248):636–638，2015年。13 Cynthia Dwork、Vitaly Feldman、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Aaron Leon Roth。在自适应数据分析中保持统计有效性。载于第四十七届年度ACM计算理论研讨会论文集，第117–126页，2015年。14 Cynthia Dwork、Moritz Hardt、Toniann Pitassi、Omer Reingold 和 Richard Zemel。基于认知的公平性。载于第三届理论计算机科学创新会议论文集，第214–226页，2012年。15 Cynthia Dwork、Michael P Kim、Omer Reingold、Guy N Rothblum 和 Gal Yona。从结果中学习：循证排名。载于2019年IEEE第60届年度计算机科学基础研讨会（FOCS），第106–125页。IEEE，2019年。16 Cynthia Dwork、Michael P Kim、Omer Reingold、Guy N Rothblum 和 Gal Yona。结果不可区分性。载于第53届ACM SIGACT计算理论年度研讨会论文集，第1095–1108页，2021年。17 Dean P Foster 和 Rakesh V Vohra。渐近校准。《生物计量学》，85(2):379–390，1998年。18 Rina Foygel Barber、Emmanuel J Candes、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):455–482，2021年。19 Parikshit Gopalan、Adam Tauman Kalai、Omer Reingold、Vatsal Sharan 和 Udi Wieder。全能预测器。arXiv 预印本 arXiv:2109.05389，2021年。20 Parikshit Gopalan、Michael P Kim、Mihir Singhal 和 Shengjia Zhao。低阶多重校准。arXiv 预印本 arXiv:2203.01255，2022年。21 Varun Gupta、Christopher Jung、Georgy Noarov、Mallesh M Pai 和 Aaron Roth。在线多重有效学习：均值、矩与预测区间。arXiv 预印本 arXiv:2101.01739，2021年。22 Ursula Hébert-Johnson、Michael Kim、Omer Reingold 和 Guy Rothblum。多重校准：针对（可计算识别的）大众的校准。载于国际机器学习会议，第1939–1948页。PMLR，2018年。23 Wenlong Ji、Zhun Deng、Ryumei Nakada、James Zou 和 Linjun Zhang。对比在特征学习中的力量：理论分析。arXiv 预印本 arXiv:2110.02473，2021年。24 Wenlong Ji、Yiping Lu、Yiliang Zhang、Zhun Deng 和 Weijie J Su。关于神经坍缩的无约束逐层剥离视角。arXiv，2110.02796，2021年。25 Christopher Jung、Changhwa Lee、Mallesh Pai、Aaron Roth 和 Rakesh Vohra。用于不确定性估计的矩多重校准。载于学习理论会议，第2634–2678页。PMLR，2021年。26 Christopher Jung、Georgy Noarov、Ramya Ramalingam 和 Aaron Roth。批量多重有效保形预测。arXiv 预印本 arXiv:2209.15145，2022年。27 Kenji Kawaguchi、Linjun Zhang 和 Zhun Deng。理解非线性表示学习的动态及其应用。《神经计算》，34(4):991–1018，2022年。28 Michael Kearns、Seth Neel、Aaron Roth 和 Zhiwei Steven Wu。防止公平性选区操纵：子群公平性的审计与学习。载于国际机器学习会议，第2564–2572页。PMLR，2018年。

54:16 HappyMap：一种广义多重校准方法

29 Michael P Kim、Amirata Ghorbani 和 James Zou。《多重准确性：用于分类公平性的黑盒后处理》。载于《2019年AAAI/ACM人工智能、伦理与社会会议论文集》，第247–254页，2019年。30 Michael P Kim、Christoph Kern、Shafi Goldwasser、Frauke Kreuter 和 Omer Reingold。《普适适应性：与倾向评分竞争的目标无关推断》。《美国国家科学院院刊》，119(4):e2108097119，2022年。31 Jon Kleinberg、Sendhil Mullainathan 和 Manish Raghavan。《公平确定风险评分中的内在权衡》。《arXiv》预印本 arXiv:1609.05807，2016年。32 Jing Lei 和 Larry Wasserman。《非参数回归的无分布假设预测带》。《英国皇家统计学会会刊：B辑（统计方法学）》，76(1):71–96，2014年。33 Lihua Lei 和 Emmanuel J Candès。《反事实与个体治疗效应的保形推断》。《英国皇家统计学会会刊：B辑（统计方法学）》，2021年。34 Lingling Li、Changyu Shen、Xiaochun Li 和 James M Robins。《缺失数据的加权方法》。《医学研究中的统计方法》，22(1):14–30，2013年。35 Roderick JA Little 和 Donald B Rubin。《缺失数据的统计分析》，第793卷。约翰·威立父子出版公司，2019年。36 Ryumei Nakada、Halil Ibrahim Gulluk、Zhun Deng、Wenlong Ji、James Zou 和 Linjun Zhang。《理解多模态对比学习并纳入未配对数据》。《arXiv》预印本 arXiv:2302.06232，2023年。37 Yaniv Romano、Rina Foygel Barber、Chiara Sabatti 和 Emmanuel J Candès。《对任何人都无恶意：通过均等覆盖评估不确定性》。《arXiv》预印本 arXiv:1908.05428，2019年。38 Yaniv Romano、Evan Patterson 和 Emmanuel Candes。《保形化分位数回归》。《神经信息处理系统进展》，32:3543–3553，2019年。39 Alvaro Sandroni、Rann Smorodinsky 和 Rakesh V Vohra。《基于多种检验规则的校准》。《运筹学数学》，28(1):141–153，2003年。40 Glenn Shafer 和 Vladimir Vovk。《保形预测教程》。《机器学习研究期刊》，9(3)，2008年。41 Ingo Steinwart 和 Andreas Christmann。《借助弹球损失估计条件分位数》。《伯努利》，17(1):211–225，2011年。42 Ryan J Tibshirani、Rina Foygel Barber、Emmanuel Candes 和 Aaditya Ramdas。《协变量偏移下的保形预测》。《神经信息处理系统进展》，32，2019年。43 T Tony Cai 和 Linjun Zhang。《高维线性判别分析：最优性、自适应算法与缺失数据》。《英国皇家统计学会会刊B辑：统计方法学》，81(4):675–705，2019年。44 Nilesh Tripuraneni、Michael Jordan 和 Chi Jin。《迁移学习理论：任务多样性的重要性》。《神经信息处理系统进展》，33:7852–7862，2020年。45 Huaxiu Yao、Long-Kai Huang、Linjun Zhang、Ying Wei、Li Tian、James Zou、Junzhou Huang 等。《通过任务增强提升元学习的泛化能力》。载于《国际机器学习会议》，第11887–11897页。PMLR，2021年。46 Huaxiu Yao、Yiping Wang、Linjun Zhang、James Zou 和 Chelsea Finn。《C-mixup：提升回归的泛化能力》。《arXiv》预印本 arXiv:2210.05775，2022年。47 Huaxiu Yao、Yu Wang、Sai Li、Linjun Zhang、Weixin Liang、James Zou 和 Chelsea Finn。《通过选择性增强提升分布外鲁棒性》。载于《国际机器学习会议》，第25407–25437页。PMLR，2022年。48 Huaxiu Yao、Linjun Zhang 和 Chelsea Finn。《通过任务插值以更少任务进行元学习》。《arXiv》预印本 arXiv:2106.02695，2021年。49 Haotian Ye、James Zou 和 Linjun Zhang。《先冻结再训练：迈向伪相关和特征噪声下可证明的表征学习》。《arXiv》预印本 arXiv:2210.11075，2022年。

A 附加文献综述

Hébert-Johnson 等人提出了在批处理设置中获得高效多重校准预测器的方法 [22], 多重校准在在线设置中的研究在统计学中历史悠久（不过并不使用这个名称；例如参见 [17, 39]及其中的参考文献）。与此同时， [22], [28]提出了多群体公平性的概念，包括统计均等子群公平性和假阳性子群公平性。 [25]将多重校准的概念扩展到批处理设置中的高阶矩，而 [21] 则提供了在线解决方案。具体而言，取$s(f(x)\ y) = f(x)^k - y^k$即可得到矩多重校准的原始矩版本；他们还采用一种新颖且不同的技术，研究了高阶中心矩的校准。此外，多重校准还被应用于批处理设置，以解决若干类似问题：公平排序 [15]；全能预测，即大致学习一个预测器，使其针对给定的一类损失函数进行后处理后，能够对该类中的任意函数最小化损失 [19]；以及为泛化到未来人群提供倾向评分的替代方法 [30]。我们在下表中总结了这些先前的概念，并展示它们与我们提出的新概念之间的关系。

先前概念与 s-快乐多重校准之间的关系		
概念名称	$c(f(x)\ x)$ 的选择	选择 $s(f(x)\ y)$
多重准确性 [29]	$c(x)$	$f(x) - y$
多重校准 [22]	$c(f(x)\ x)$	$f(x) - y$
低阶多重校准 [20]	$\tilde{c}(x)w(f(x))$，对于某个函数 $\tilde{c}$	$f(x) - y$
极小极大群体公平性 [10]	$\mathbb{1}\{x \in G\}$ 对于人口统计群体 G	非负损失
全能预测器 [19]（更多内容见附录 E.1）	$\mathbb{1}\{f(x) \in I_v\}$ 和 $c(x)\mathbb{1}\{f(x) \in I_v\}$	$f(x) - y$

表 1 与先前概念关系的概述

B 第 3 节省略的细节

B.1 样本版本算法的详细说明

本节将给出算法 1 对应的样本版本算法及其形式化理论保证。在此之前，与 [29]，中类似，我们先介绍一个**非正式**但实用的概念。对于数据集 D，我们用 $\mathbb{E}_{(x,y)\sim D}$ 表示在 D 上的经验期望。在本节中，我们假设 $\sup_{c\in\mathcal{C}} \|c\|_\infty \le B_0$，其中 $B_0 > 0$ 为某个通用常数。

▶ **定义 16** (函数类的维数)。 我们用 $d(\mathcal{C})$ 表示无关学习类 $\mathcal{C}$ 的维数，满足：若样本量 $m \ge C_1 \frac{d(\mathcal{C})+\log(1/\delta)}{\alpha^2}$，其中 $C_1 > 0$ 为某个通用常数，则来自 D 的随机样本 S_m 保证 $\mathcal{C}$ 上的一致收敛误差至多为 α，失败概率至多为 δ，即对于任意固定的 f 和固定的 s，若 $\|s\|_\infty \le C_2$，其中 $C_2 > 0$ 为某个通用常数：

$$\sup_{c(f(x),x)\in\mathcal{C}} \left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x),x)s(f(x),y)] - \mathbb{E}_{(x,y)\sim S_m}[c(f(x),x)s(f(x),y)]\right| \le \alpha.$$

此类维度上界的例子包括$\mathcal{C}$的度量熵。

基于以上准备，我们给出以下样本版本算法。

算法 2 HappyMap

输入： 步长 $\eta > 0$，迭代次数上限 $T \in \mathbb{N}_+$，初始预测器 $f_0(\cdot)$，分布 $\mathcal{D}$，凸集 $O \subseteq \mathbb{R}$，映射 $s: \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, T 验证数据集 $D_0, \cdots D_T$，每个数据集包含 m 个样本

设置 $t = 0$

当 $t < T$ 并 $\exists c_t \in \mathcal{C}$： $\mathbb{E}_{(x,y)\sim D_t}[c(f_t(x)\ x)s(f_t(x)\ y)] > 3\alpha/4$ **执行**

令 c_t 为 $\mathcal{C}$ 中的任意元素，满足 while 语句中的条件

$\forall x \in \mathcal{X}$， $f_{t+1}(x) = \Pi_O\left[f_t(x)\ - \eta \cdot c(f(x)\ x)\right]$

$t = t + 1$

end

输出： $f_t(\cdot)$

▶**定理 17.** 在我们的假设下，若 $\mathcal{C}$ 是对称的，并假设我们使用适当选取的 $\eta = O(\alpha/(\kappa_{\mathcal{L}} B))$ 和 $m = \Omega(T \cdot {}^{d_{(\mathcal{C})+\log(1/\delta)}}\ \alpha^2)$ 来运行算法 2，则该算法以至少 ~~1 — δ~~ 的概率在 $T = O((C^u_{\mathcal{L}} - C^l_{\mathcal{L}})\ \kappa_{\mathcal{L}} B/\alpha^2)$ 内收敛，从而得到

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x),x)s(f(x),y)]\right| \le \alpha$$

算法 2 的最终输出 f。

证明。 我们可以取一个适当选定的 $m = \Omega(T \cdot {}^{d_{(\mathcal{C})+\log(1/\delta)}}\ \alpha^2)$，使得对于所有 $t \in [T]$,

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f_t(x),x)s(f_t(x),y)] - \mathbb{E}_{(x,y)\sim D_t}[c(f_t(x),x)s(f_t(x),y)]\right| \le \alpha/4.$$

因此，每当算法 2 更新时，我们知道

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f_t(x),x)s(f_t(x),y)] \ge \alpha/2.$$

因此，底层势函数的进展至少为 ${}^{\alpha^2}_{8\kappa_{\mathcal{L}} B}$。按照算法 1 的类似证明，只要 T 满足 $(C^U_{\mathcal{L}} - C^l_{\mathcal{L}})$ $/\ {}^{\alpha^2}_{8\kappa_{\mathcal{L}} B} < T$，我们就知道算法 2 提供了一个解 f，使得 ——

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[c(f(x),x)s(f(x),y)]\right| \le \alpha.$$

◀

▶ 注释 18。第 4、5 节中所有定理的样本版本结果都可以自然地由上述定理得到。我们将在论文中不再赘述。

C 第 4 节的省略细节

C.1 评注 5 的证明

证明。 由于 $\mathbb{1}\{y \in \mathcal{I}(x)\} = \mathbb{1}\{y \le l_{\delta/2}(x)\} - \mathbb{1}\{y \le l_{1-\delta/2}(x)\}$ 且 $(1-\delta) = (1-\delta/2) - \delta/2$，我们有

$$\mathbb{1}\{y \in \mathcal{I}(x)\} - (1-\delta) = [\mathbb{1}\{y \le l_{\delta/2}(x)\} - \delta/2] - [\mathbb{1}\{y \le l_{1-\delta/2}(x)\} - (1-\delta/2)].$$

因此，$\sup_{c\in\mathcal{C}} \mathbb{E}[c(x)(\mathbb{1}\{y \in \mathcal{I}(x)\}\ - (1-\delta))] \le \sup_{c\in\mathcal{C}} \mathbb{E}[c(x)(\mathbb{1}\{y \le l_{\delta/2}(x)\}\ - \delta/2] + \sup_{c\in\mathcal{C}} \mathbb{E}[c(x)(\mathbb{1}\{y \le l_{1-\delta/2}(x)\} - (1-\delta/2))] \le 2\alpha$ ◀

D 第 5 节省略的细节

D.1 定理 11 的证明

证明。C $=\{c(f(x)\ x)\ =\ \pm^{1-\sigma}(x)\ \sigma(x)\ (f(x)\ -\ p(x))\colon\ \sigma\ \in\ \Sigma,p\ \in \mathrm{P}\}$，其中 $\sigma,p\ \in(0,$ 1)，且 $\mathrm{L}(f(x)\ y)\ =\ 1/2(f(x)\ -\ y)^2$ 且 $O\ =\ [0,1]$。然后，我们直接应用定理 3 的证明，可得对于所有 $\sigma\in\Sigma$ 和 $p\in\mathrm{P}$,

$$\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[\frac{1-\sigma(x)}{\sigma(x)}(f_t(x)-p(x))(f_t(x)-y)]\right|\leq\alpha.$$

为得到最终保证，对于任意 $\hat{}\ f$，记 $\delta_{\sigma、\hat{}f,p}=\phi^*_{\hat{f}}(x)-\phi_{\sigma、\hat{}f,p}(x)$，其中$\phi^*_{\hat{f}}$是底层真实值 $^{1-e}(x)\ e(x)$（$\hat{}f(x)\ -\mathbb{E}(y|x))$，而 $\phi_{\sigma、\hat{}f,p}(x)\ =\ ^{1-\sigma}(x)\ \sigma(x)$（$\hat{}f(x)\ -\ p(x)$）。

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}[\left(\hat{f}(x)-\mathbb{E}(y|x)\right)^2]&=\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[\frac{1-e(x)}{e(x)}(\hat{f}(x)-\mathbb{E}(y|x))(\hat{f}(x)-y)]\\&=\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}(\phi_{\sigma,\hat{f},p}(x)(\hat{f}(x)-y))+\mathbb{E}_s((\phi^*_{\hat{f}}(x)-\phi_{\sigma,\hat{f},p}(x))(\hat{f}(x)-y))\\&\leq\alpha+\sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}(\hat{f}-y)^2\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\delta^2_{\sigma,\hat{f},p}}.\end{aligned}$$

我们应注意，可以选择合适的 $\sigma\in\Sigma$，$p\in\mathcal{P}$，上述界可写为:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}[\left(\hat{f}(x)-\mathbb{E}(y|x)\right)^2]\leq\alpha+\min_{\sigma\in\Sigma,p\in\mathcal{P}}\sqrt{\mathbb{E}_s(\hat{f}-y)^2\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\delta^2_{\sigma,\hat{f},p}}$$

现在，唯一需要做的就是界定 $\min_{\sigma\in\Sigma,\ p\in\mathcal{P}}\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\delta^2_{\sigma\ ,\ \hat{}f,\ p}$。我们将偏差记为

$$\beta_{st}=\min_{\sigma\in\Sigma}\sqrt{E_{(x,y)\sim\mathcal{D}_{so}}(\frac{1-\sigma(x)}{\sigma(x)}-\frac{1-e(x)}{e(x)})^4}.$$

因此，我们有

$$\min_{\sigma\in\Sigma,p\in\mathcal{P}}\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\delta^2_{\sigma,\hat{f},p}\leq 2\beta_{st}\sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}(\hat{f}-\mathbb{E}(y|x))^4}+2\min_{p\in\mathcal{P}}\sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[\frac{1-\bar{\sigma}(x)}{\bar{\sigma}(x)}]^4}\sqrt{\mathbb{E}_s[p(x)-\mathbb{E}(y|x)]^4}$$

其中，$\bar{\sigma}$ 是使 β_{st}达到最小值的那个。因此，结合事实 f，p，$y\in[0,1]$，我们可以得到最终结果。◀

D.2 定理12 的证明

证明。 根据定理 3 和定理 4 的证明（其中证明了函数$s(l,\ y)\ =\mathbb{1}\{l\leq y\}\ -(1-\delta)$ 满足定理 3 的条件 (b)-(c)），我们有

$$\left|\sup_{c\in\mathcal{C}}\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[c(x)\cdot(\mathbb{1}\{l(x)\leq y\}-(1-\delta))]\right|\leq\alpha.$$

由于$\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x)\ \leq y)=\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}\left(^{1-e(x)}_{e\ x}()\ \mathbb{1}\{l(x)\ \leq y\}\right)$，因此我们有

$$\begin{aligned}|\mathbb{P}_{(x,y)\sim\mathcal{D}_{ta}}(l(x)\leq y)-(1-\delta)|\leq&\alpha+|\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(c(x)-\frac{1-e(x)}{e(x)})\cdot(\mathbb{1}\{l(x)\leq y\}-(1-\delta))]|\\\leq&\alpha+\sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(c(x)-\frac{1-e(x)}{e(x)})^2]\cdot\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(\mathbb{1}\{l(x)\leq y\}-(1-\delta))^2]}\\\leq&\alpha+\beta,\end{aligned}$$

其中，最后一个不等式用到了如下事实：$|\mathbb{1}\{l(x)\ \leq y\}-(1-\delta)|<1$。 ◀

D.3 定理 14 的证明

证明。根据定理 3，以及势函数 $\mathcal{L}(f(x)\ y) = 1/2(f(x) - y)^2$ 满足定理 3 的条件 (b)-(c)，我们有

$$\left|\sup_{c\in\mathcal{C}} \mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[c(x)\cdot(y - f(x))]\right| \le \alpha.$$

由于 $\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2]$ 可以等价地写成 $\mathbb{E}_{(x,y)\sim\mathcal{D}|R=1}[^{\mathbb{P}\ (x|R_i=1)}\ \mathbb{P}\ (x)\ (\mathbb{E}[y|x] - f(x))\cdot(y - f(x))]$,，因此我们有

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2] \le& \alpha + |\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(c(f(x),x) - \frac{\mathbb{P}(x \mid R_i = 1)}{\mathbb{P}(x)}(\mathbb{E}[y|x] - f(x)))\cdot(y - f(x))]| \\ \le& \alpha + \sqrt{\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(c(x) - \frac{\mathbb{P}(x \mid R_i = 1)}{\mathbb{P}(x)}(\mathbb{E}[y|x] - f(x)))^2]\cdot\mathbb{E}_{(x,y)\sim\mathcal{D}_{so}}[(y - f(x))^2]} \\ \le& \alpha + \beta,\end{aligned}$$

其中，最后一个不等式利用了这样一个事实：投影到 $O = [0,1]$ 后，$|f(x) - y| < 1$。◀

E 与现有文献的关系

E.1 与全能预测器的关系

在近期的一项工作中，[19] 提出利用多重校准这一概念来构造一个（$\mathcal{L}$，$\mathcal{C}$）全能预测器——一种可用于优化一族损失$\mathcal{L}$中任意损失的预测器；也就是说，这种预测器的输出可以经过后处理，从而获得比给定函数类$\mathcal{G}$中的任意假设都更小的损失。

我们首先陈述 [19] 中多重校准的定义。

▶ **定义19.** *([19])* 设 D 是 $X \times\{0,1\}$ 上的一个分布。若对于每个 $i \in [m]$ 和 $c \in$，条件分布 $D_i = D \mid x \in S_i$ 满足以下条件，则 X 的划分 $S = \{S_1, ..., S_m\}$ 称为关于 G、D 的 α-多重校准划分：

$$|\mathrm{Cov}_{\mathcal{D}_i}(c(x), y)| \le \alpha. \tag{10}$$

尽管这一概念看起来与定义 2 不同，[19] 证明了，当$c(x)$ 是布尔函数时，定义19 等价于均值多重校准 [25]。我们的下一个定理将这一关系进一步推广到实值函数 c，并表明我们在公式（2）中提出的广义概念实际上蕴含了（10）。

▶ **定理20.** 考虑一个函数类 G。记 $\Lambda = \{\lambda, 3\lambda, 5\lambda, ..., 1 - \lambda\}$和 $I_v = [v - \lambda, v + \lambda]$。取 $I_\Lambda = \{I_v\}_{v\in\Lambda}$，并令 C 包含 $\mathbb{1}\{f(x) \in I_v\}$ 和 $c(x)\mathbb{1}\{f(x) \in I_v\}$，其中 $c \in G$。假设 $|c(x)| < L$ 和 $\mathbb{P}(x \in I_v) \ge \gamma$。

如果预测器 f 满足相对于 C 的 α-多重校准（见公式 (2)），其中$s(f(x)\ y) = f(x) - y$，那么由该预测器 f定义的集合划分 S_Λ满足相对于 G 的 $(^{(L+1)_\gamma}\ ^{\alpha} + \lambda)$- 多重校准（见公式 (10)）。

证明。回顾一下，$\mathcal{G}$包含$\mathbb{1}\{f(x) \in I_v\}$ 以及 $c(x)\mathbb{1}\{f(x) \in I_v\}$。令 $S_v = \{x : f(x) \in I_v\}$。那么，由 (2) 可得

$$\begin{aligned}|\mathbb{E}[\mathbb{1}\{x \in S_v\}\cdot(f(x) - y)]| < \alpha, \\ |\mathbb{E}[\mathbb{1}\{x \in S_v\}c(x)\cdot(f(x) - y)]| < \alpha.\end{aligned}$$

Then

$$|\mathbb{E}[\mathbb{1}\{x \in S_v\}c(x) \cdot (y - \mathbb{E}[y \mid x \in S_v])]| \leq \alpha + |\mathbb{E}[\mathbb{1}\{x \in S_v\}c(x) \cdot (f(x) - \mathbb{E}[y \mid x \in S_v])]|$$
$$\leq \alpha + |\mathbb{E}[\mathbb{1}\{x \in S_v\}c(x) \cdot (\mathbb{E}[f(x) - y \mid x \in S_v])]| + |\mathbb{E}[\mathbb{1}\{x \in S_v\}c(x) \cdot (f(x) - \mathbb{E}[f(x) \mid x \in S_v])]|$$

对于第二项，我们有

$$|\mathbb{E}[\mathbb{1}\{x \in S_v\}c(x) \cdot (\mathbb{E}[f(x) - y \mid x \in S_v])]| \leq |\mathbb{E}[\mathbb{1}\{x \in S_v\}|c(x)| \cdot \frac{\alpha}{\mathbb{P}(x \in S_v)}]| \leq \alpha L.$$

对于第三项，由于对于任意 $x \in S_v$，$|f(x) - \mathbb{E}[f(x) \mid x \in S_v]| < \lambda$，我们有

$$|\mathbb{E}[\mathbb{1}\{x \in S_v\}c(x) \cdot (f(x) - \mathbb{E}[f(x) \mid x \in S_v])]| \leq L\lambda\mathbb{P}(x \in S_v).$$

综合所有部分，我们有

$$|\mathbb{E}[\mathbb{1}\{x \in S_v\}c(x) \cdot (y - \mathbb{E}[y \mid x \in S_v])]| \leq (L+1)\alpha + L\lambda\mathbb{P}(x \in S_v),$$

这意味着

$$\begin{aligned}|\mathrm{Cov}(c(x), y \mid x \in S_v)| =& |\mathbb{E}[c(x)(y - \mathbb{E}[y \mid x \in S_v]) \mid x \in S_v]| \\ =& \frac{|\mathbb{E}[c(x)(y - \mathbb{E}[y \mid x \in S_v]) \cdot \mathbb{1}\{x \in S_v\}]|}{\mathbb{P}(x \in S_v)} \\ \leq& \frac{(L+1)\alpha}{\gamma} + \lambda.\end{aligned}$$

◀

E.2 广义统计均等子群体公平性

Kearns 等人 [28] 考虑多重均等（也称为统计均等子群体公平性）：给定一组预先指定且可能彼此重叠的人口统计学子群体，被赋予正向结果的人群在人口统计特征上应与总体相同。本节将说明，通过在 HappyMap 中适当选择映射，可以表达统计均等以及假正率（和假负率）相等的约束。这些保证的多群体版本可通过常数函数轻易实现。与 [28], 不同，本节不对我们的算法给出保证。我们陈述这些观察，是为了展望将 HappyMap 扩展至约束优化。

按照 [28], ，我们的目标是找到满足这一公平性概念的随机分类器。我们考虑二元分类，其中 $\mathcal{Y} = \{0, 1\}$。对于均匀随机变量 $U \sim U(0, 1)$，令 $\mathcal{U}(x) = \mathbb{1}\{U < f(x)\}$，其中 $f(x) \in [0, 1]$ 表示拟合概率函数，满足 $\mathcal{U}(x) \sim Bern(f(x))$。假设存在一组人口统计群体指示函数 $\mathcal{G} = \{g : \mathcal{X} \to \{0, 1\}$，其中 $g(x) = 1$ 表示具有特征 x的个体属于群体 g。多重均等要求对所有 $g \in \mathcal{G}$ 均成立：

$$|\mathbb{P}(g(x) = 1) \cdot (\mathbb{P}(\mathcal{U}(x) = 1) - \mathbb{P}(\mathcal{U}(x) = 1 \mid g(x) = 1))| \leq \alpha. \tag{11}$$

此不等式可以改写为

$$|\mathbb{E}[(\mathbb{1}\{g(x) = 1\} - \mathbb{P}(g(x) = 1)) \cdot \mathbb{1}\{U < f(x)\}]| \leq \alpha.$$

这一概念仅考虑取值范围为 {0,1} 的指示函数型公平性审计器 g。接下来，我们将多重均等性推广到允许 g 取实值的情形：

$$\sup_{g\in\mathcal{G}}\Big|\mathbb{E}[(g(x)-\mathbb{E}[g(x)])\cdot\mathbb{1}\{U<f(x)\}]\Big|\le\alpha, \tag{12}$$

其中 $\mathcal{G}\subset[0,1]^{\mathcal{X}}$ 表示一组取值范围可能连续的审计器。正如 [29], 所指出的，这使得纳入更灵活的统计检验成为可能。令$c(f(x),x)=g(x)-\mathbb{E}[g(x)]$,和 $s(f,y)=\mathbb{1}\{U<f\}$，将 HappyMap 应用于此设置（稍作修改，使 **while** 条件检查 $\exists g\in\mathcal{G}:\mathbb{E}_{U,(x,y)\sim\mathcal{D}}[(g(x)-\mathbb{E}[g(x)])\cdot(\mathbb{1}\{f_t(x)>U\})]>\alpha$，其中期望是对数据和 U的随机性共同取平均）可得到以下保证。

▶ 定理 *21.* 假设对所有 $c\in\mathcal{C}$，均有 $\mathbb{E}_{x\sim\mathcal{D}^{\mathcal{X}}}[c^2(x)]\le B$。取适当选定的$\eta=O(\alpha/B)$，则 *HappyMap*（算法 *1*）在使用 $T=O(B/\alpha^2)$ 和势函数 $L(f(x),y)=(1-\delta)\cdot f(x)-\mathbb{E}_U[\min(f(x)-U,0)]$ 运行时，会输出满足下述条件的函数 $l(\cdot)$

$$\sup_{g\in\mathcal{G}}\mathbb{E}[(g(x)-\mathbb{E}[g(x)])\cdot\mathbb{1}\{U<f(x)\}]\le\alpha.$$

定理 21 的证明。 定理 21 的证明直接沿用定理 4 的证明，只需将 y 其中的 U替换为即可。同样的推导仍适用于这种情况，此处省略细节。◀

▶ 注 22. 我们也可以将 [28], 中考虑的假阳性子群体公平性概念类似地推广，也就是说，$|\mathbb{P}(g(x)=1,\ y=0)\cdot(\mathbb{P}(\mathcal{U}(x)=1)-\mathbb{P}(\mathcal{U}(x)=1\mid g(x)=1))|\le\alpha$，推广到考虑所有可由特定分类器识别的群体的情况：$\sup_{g\in\mathcal{G}}\Big|\mathbb{E}[(g(x)-\mathbb{E}[g(x)\mid y=0])\cdot\mathbb{1}\{U<f(x)\}\mid y=0]\Big|\le\alpha$. 将 HappyMap 应用于分布$\mathcal{D}\mid y=0$ 可得到一个满足此约束的分类器。

E.3 边缘覆盖率与校准覆盖率的比较

本节在$\mathcal{C}=\{1\}$的情形下，比较式（6）中的概念与多重有效性。在这一特殊情形下，我们分别称它们为边缘覆盖率和校准覆盖率。下面先介绍一些基础知识和定义。

假设 y 是一个连续随机变量。对于 $\delta>0$，我们的目标是基于 n 独立同分布样本（x_i,y_i）$\sim P_{x,y}$，即（x,y）的联合分布，构造（$1-\delta$）预测区间 $C_n(x)=(l(x),u(x))$。按照 [21, 3], 中的设定，预测区间 $C_n(x)$基于非一致性分数 $m(x,y)$构造，例如，对于任意预测函数 h，可取 $|y-h(x)|$；并使用一致性阈值函数 $q(x)$。那么，$C_n(x)=\{y:m(x,y)<q(x)\}$。

边际覆盖保证要求 $C_n(x)$满足

$$\mathbb{P}(y\in C_n(x))=1-\delta, \tag{13}$$

对所有 $P_{x,\ y}$，其中概率是对所有训练样本以及 $(x,\ y)\sim P_{x,\ y}$取的。

校准覆盖保证要求 $C_n(x)$满足

$$\mathbb{P}(y\in C_n(x)\mid q(x)=q)=1-\delta, \tag{14}$$

对所有 $P_{x,\ y}$ 以及几乎所有 q。于是我们得到以下定理。

▶ **定理 23**。 假设根据 n 个样本构造的区间 $C_n(x)$ 满足条件 (14)意义下的 $1-\delta$校准覆盖率。那么，对于任意 $P_{x,\ y}$，以概率 *1*，

$$|C_n(x)| = \infty,$$

在几乎所有满足 q *that* $q(x) = q$ 是 P_q的非原子点的位置，其中 P_q 表示 $q(x)$的边缘分布。

这里，如果点 q 满足： q 属于 P_q 的支撑集，且 $\mathbb{P}(B(q,\epsilon)) \to 0$ *as*$\epsilon \to 0$，则称其为 P_q 的非原子点。其中， $B(q,\epsilon)$ 是以 q 为中心、半径为 ϵ的欧几里得球。

我们注意到，该定理的关键部分在于“有限样本”，并将 $C_n(\cdot)$视为对训练样本应用某种方法后得到的区间。当某种方法根据 n 个样本生成一个区间，并且需要对一大类分布满足条件 (14) 时，定理 23 表明，这样的区间必定具有无限长度。相比之下，标准保形预测方法（例如分割保形预测） [40], 会生成有限长度的区间，并且对所有$P_{x,y}$都满足公式 (13)。

▶ 备注 24。在 [3], 的某些正则性条件下，会对 $P_{x,y}$ 施加限制，从而使分布类变小，因此定理 23 不适用。这表明，校准覆盖率需要某些正则性条件，而边缘覆盖率则不需要。此外，在某些额外的正则性条件下，校准覆盖率会产生短区间，并排除长区间。例如，假设 $y \in [-1,1]$。取非一致性分数 $m(x,y) = |y|$，则所有满足 (14) 的区间长度都必须小于 2。这是因为 $q(x)$ 必须小于 1，否则，以 $q(x) = q$ 为条件，针对 $q \geq 1$ 会使 (14) 左侧等于 1。相比之下，满足 (13) 的区间长度可能是无限的。例如，对于 x 的 $(1-\delta)\times 100\%$ 样本，取 $C_n(x) = \mathbb{R}$，其余情况下取 $C_n(x) = 0$。这样的 $C_n(x)$ 满足 (13)，但不满足更强的条件 (14)。

▶ 备注 25. 我们还想指出，定理 23 仅对 P_q 中所有非原子点成立。这表明，在弱假设下，要实现校准覆盖率（或多重有效），必须对 $q(x)$进行离散化。

定理 23 的证明。 该证明的核心思路遵循 [32] 和 [18]。我们来证明定理对任意给定的分布 P_q成立。为便于表述，我们省略下标 q，并将分布记为 P。首先固定任意的 $\epsilon, B > 0$。

对于一个非原子点 x_0，存在一个 $\delta_0 > 0$，使得 $\mathbb{P}_P(q(x) \in B(q(x_0)\ \delta_0) < \epsilon_n$其中 $\epsilon_n = 2\{1-(1-\epsilon^2/8)^{1/n}\}$。我们还令 $B_0 = {}_2{}^{B}{}_{\overline{(1-\delta)_\circ}}$

接下来，我们通过以下方式定义一个新的分布 Q:

$$\mathbb{P}_Q(A) = \mathbb{P}_P(A \cap S^c) + \mathbb{P}_U(A \cap S),$$

其中 $S = \{(x,y)\colon\ q(x) \in B(q(x_0)\ \delta_0)\}$，并且 U在 $\{(x,y)\colon\ q(x) \in B(q(x_0)\ \delta_0)\ |y| \leq B_0\}$ 上服从均匀分布，总质量为 $\mathbb{P}_P(S)$。

根据定义，我们可以验证 $TV(P,Q) \leq \epsilon_n$，因此 $TV(P^n, Q^n) \leq \epsilon_\circ$

由于 (14) 对 Q 也成立，因此我们有

$$\int_{C_n(x)} dQ(y \mid q(x)) \geq 1-\delta,$$

对于所有满足 $q(x) \in B(q(x_0)\ \delta_0)$ 的 x。对于这样的 x，我们有 $|dQ(y \mid q(x))| \leq \frac{1}{2B_0}$，这意味着$|C_n(x)| \geq 2B_0(1-\delta) = B$。因此，对于这样的 x

$$\mathbb{P}_Q(|C_n(x)| > B) = 1,$$

由此可得

$$\mathbb{P}_P(|C_n(x)| > B) \geq \mathbb{P}_Q(|C_n(x)| > B) - \epsilon = 1 - \epsilon.$$

由于 ϵ 和 B 是任意的，我们完成了证明。 ◀

F 函数类复杂度降低

在本节中，我们简要讨论在某些情况下如何降低类 $\mathcal{C}$ 的复杂度。我们将采用机器学习社区近期开发的表征学习工具 [44, 24, 23, 36, 49], ，这些工具曾用于提升泛化能力 [27, 47, 45, 48, 46]、鲁棒性 [9], 和公平性 [8, 5]。事实上，在我们的应用中，如果 $\mathcal{C}$ 是一个非常大的类，那么算法 1 的效率就会降低，因为我们需要在每次迭代时从一个丰富的类中搜索 c_t 。

具体而言，我们将重点讨论使用 $\mathcal{C}$ 来刻画倾向评分比的情形

$$\frac{1 - e(x)}{e(x)} = \frac{\mathbb{P}(z = ta|x)}{\mathbb{P}(z = so|x)}.$$

这里，$\mathcal{C}$取为 $\mathcal{C} = \{\frac{1-\sigma(x)}{\sigma\ x} \text{() } \sigma \in \Sigma\}$，且 $\sigma : \mathcal{X} \mapsto [0, 1]$。我们在此指出，如何选择 Σ存在一个两难问题。一方面，如果 Σ 不够丰富，那么该类可能缺乏足够的表达能力，并且 $e(x) \notin \Sigma$ 的可能性很高。另一方面，如果 Σ 过于丰富，算法 1 可能会面临效率问题。有没有办法让 Σ 具有表达能力，同时保持 Σ 的规模适中？本节将探讨使用预训练模型，在保持类规模适中的同时实现表达能力的可能性。

具体来说，使用预训练模型的核心思路是：对于函数类 $\mathcal{G} = \{g = w \circ h, w \in \mathcal{W}\ h \in \mathcal{H}\}$，其中 $\mathcal{W}$的复杂度较低，而 $\mathcal{H}$ 的复杂度较高，我们可以使用一些相关且丰富的数据集进行预训练 h ，并获得 h。然后，对于我们真正想要将 g 应用于的数据集 D ，只需从类$\mathcal{G}_{shrink} = \{g = w \circ \hat{h}, w \in \mathcal{W}\}$中选出最终的 g 即可。这是迁移学习中一种相当常见的方案 [44]。

预训练设置。现在，我们演示如何使用迁移学习来预训练 h。为简单起见，考虑 $g(x) = w^T h(x)$，也就是说， g 由一个线性变换和一个可能较复杂的函数 h组合而成。这在现代学习中十分常见，其中 h 通常称为表示。如果对 h 进行预训练，则称为固定特征学习。考虑 $\sigma(x) = e^{g(x)}/(1 + e^{g(x)})$，使得 $\Sigma = \{\frac{e^{g(x)}}{1+e^{g(x)}} : g = w^T \hat{h}, w \in \mathcal{W}\}$，其中$\hat{h}$是通过迁移学习得到的表示。

为了运用迁移学习的思想，我们假设

$$e(x) \in \Sigma_{h^*} := \{\frac{e^{g(x)}}{1 + e^{g(x)}} : g = w^T h^*, w \in \mathcal{W}\}.$$

为求得 $\hat{h}$，考虑存在 k 对相互独立的源域—目标域数据集，$\{(D_j^{so}, D_j^{ta})\}_{j=1}^k$，且 D_j^{so} 与 D_j^{ta} 也相互独立。 D_j^{so}由形式为 $(x,1)$ 的数据对组成，而 D_j^{ta}由形式为 $(x, 0)$ 的数据对组成，其中 0,1 是标签。将增广数据集 $\{(D_j^{so}, D_j^{ta})\}_{j=1}^k$ 的分布记为 $\{\mathcal{D}_i\}_{i=1}^k$。$\{(\mathcal{D}_i\}_{i=1}^k$ 的倾向评分比共享相同的表示 h^* ，即 $\mathcal{D}_j \sim Bern(\rho(w_j^T h^*))$，这意味着

$$\mathbb{P}(z = so|x) = \rho(w_j^T h^*(x)),$$

其中 $\rho(x) = e^x/(1 + e^x)$。

为了进一步理解这种设定为何合理，作为一个贴近现实的具体例子，我们假设源数据来自波士顿的一家医院，而目标是将

训练好的模型应用于旧金山的一家医院。我们可以选择波士顿其他一些医院的数据，以及旧金山一些医院的数据，并将它们配成源-目标对。通常，在元学习中，人们可以使用这些源-目标数据对来训练共享相同表示的模型，并为每一对数据额外设置简单的、特定于该数据对的结构。这种方法在实践中效果良好。这推动了一系列基于我们在此考虑的模型设定的理论研究。

不失一般性，我们假设所有数据集对（D_j^{so}，D_j^{ta}）总共包含 n 个样本，并将（D_j^{so}，D_j^{ta}）中的样本记为 $\{(x_{ji}, y_{ji})\}_{i=1}^{n}$。这里，标签 y_{ji} 是二元的，因此我们可以将其视为获取 $\rho(w_j^T h^*(x))$ 的分类问题，常用的损失是交叉熵损失 $\ell(z, y) = -y \log(\frac{e^z}{1+e^z})$。因此，我们可以通过以下优化从 $\mathcal{H}$ 中得到 $\hat{h}$：

$$(\{\hat{w}_j\}_{j=1}^{j}, \hat{h}) = \underset{\{w_j\}, h \in \mathcal{H}}{\arg\min} \frac{1}{nk} \sum_{j=1}^{k} \sum_{i=1}^{n} \ell(w_j \circ h(x_{ji}), y_{ji}).$$

推导。我们先介绍一些记号和定义。推导主要遵循 [44]，，此处我们给出所有细节以求完整。

▶ **定义 26** (高斯复杂度及其变体)。对于一个包含$q(\cdot)$的通用函数类 Q，以及 N 个数据点，$\bar{X} = (x_1, \cdots x_N)^T$，经验高斯复杂度定义为

$$\hat{\mathscr{O}}_{\bar{X}}(\mathcal{Q}) = \mathbb{E}_g[\sup_{q \in \mathcal{Q}} \frac{1}{N} \sum_{k=1}^{r} \sum_{i=1}^{N} g_{ki} q_k(x_i)], \; g_{ki} \sim \mathcal{N}(0,1) \; i.i.d.,$$

其中 $g = \{g_{ki}\}_{k \in [r], \; i \in [N]}$，且 $q_k(\cdot)$ 是向量值函数$\hat{q}(\cdot)$的第 k 个坐标。相应的总体高斯复杂度为 $\mathscr{O}_N(Q) = \mathbb{E}_{\bar{X}}[\mathscr{O}_{\bar{X}}(Q)]$，，其中期望是对数据样本 X 的分布取的。

▶ **定义 27**。对于函数类 W、k 向量 $w = (w_1, \cdots w_k)^T$以及数据 $(x_j, y_j) \sim Bern(\rho(w_j^T h^*(x_j)))$，我们定义

$$\bar{d}_{\mathcal{W},w}(h, h') = \frac{1}{k} \sum_{j=1}^{k} \inf_{w' \in \mathcal{W}} \mathbb{E}_{(x_j, y_j)}\{\ell(w'^T h'(x_j), y_j) - \ell(w_j^T h^*(x_j), y_j)\}.$$

我们首先给出 $\hat{h}$ 在 $\bar{d}_{\mathcal{W},w}(\hat{h}, h^*)$ 下的理论保证。为此，我们进一步定义关于 $\mathcal{W}$ 的最坏情况高斯复杂度：

$$\bar{\mathscr{O}}_n(\mathcal{W}) = \max_{Z \in \mathcal{Z}} \hat{\mathscr{O}}_Z(\mathcal{W}),$$

其中 $\mathcal{W} = \{h^*(x_1) \; h_2^*(x_2), \cdots \; h^*(x_n) | h^* \in \mathcal{H} \; x_i \in \mathcal{X}$ 对所有 $i \in [n]$ 均成立。此外，我们用$\mathscr{O}_{nk} = \mathbb{E}_X[\hat{\mathscr{O}}_X(Q)]$表示相对于数据矩阵 X 计算的总体高斯复杂度，该数据矩阵由 nk 训练数据点 $\{x_{ji}\}_{j,i}$构成。假设：对于所有 $w \in \mathcal{W}$，$\|w\| \leq L$。$D_i \sim Bern(w_i^T h^*)$，其中 $w_i \in \mathcal{W}$，$h^* \in \mathcal{H}$。$x_{ji} \in \mathcal{X}$。我们进一步假设，对于任意 $h \in \mathcal{H}$，$h(x)$ 对所有 $x \in \mathcal{X}$ 均有界，即 $\sup_{x \in \mathcal{X}} \|h(x)\| \leq B_h$。据此，我们知道 $|w^T h| \leq \|w\| \|h\| \leq L B_h$。

▶ **定理 28**。以 $1 - \delta$ 的概率，

$$\bar{d}_{\mathcal{W},w}(\hat{h}, h^*) \lesssim \mathcal{O}\Big(\frac{L B_h}{(nk)^2} + \log(nk)[L \cdot \mathscr{O}_{nk}(\mathcal{H}) + \bar{\mathscr{O}}_n(\mathcal{W})] + \sqrt{\frac{\log(2/\delta)}{nk}}\Big).$$

这里 $\mathscr{O}_{nk}$，$\mathscr{O}_n$ 是总体高斯复杂度和最坏情况高斯复杂度。

证明。 证明可由 [44]中的定理1得出。我们唯一需要做的是验证 [44]中的定理1所需的那些假设。

- 损失函数 $\ell(z,y) = -y\log(\frac{e^z}{1+e^z})$ 有界，并且对于所有 $y \in \mathcal{Y}$， $\ell(\cdot\ y)$ 都是利普希茨的。对
- 于所有 $w \in \mathcal{W}$，函数 $w^\top x$ 关于 x 在 ℓ_2 距离下是利普希茨的。对于所有 $w \in \mathcal{W}$ $h \in \mathcal{H}$，
- 复合函数 $w^T h$ 有界。

现在，让我们逐点证明。

首先，根据假设 $|w^T h| \leq LB_h$，以及 $y \in \{0,1\}$，我们知道对于$z = w^T h(x)$ 和 $y \in \{0,1\}$， $\ell(z,y)$ 有界。接着注意到 $\partial_z \ell(z,,y) = \frac{-y}{1+e^z}$，并且由于 z 有界且$y \in \{0,1\}$，我们有 $|\partial\ell(z,y)| \leq 1$。因此， $\ell(\cdot\ y)$ 是 1-利普希茨的。

其次，我们知道 $\|w\| \leq L$，因此知道 $w^\top x$ 是 L-利普希茨的。

最后，根据我们的假设， $w^T h$ 是有界的。

然后，我们直接应用 [44]中的定理1。 ◀

现在，我们给出最终保证。

▶ **定义29**。 对于函数类 W 和 W_0，满足 $w_0 \in W_0,$，且数据为 $(x,y) \sim \mathbb{P}_{w_0^T h}$， $h, h' \in H$ 之间的最坏情况表示差异为

$$d_{\mathcal{W},\mathcal{W}_0}(h';h) = \sup_{w_0 \in \mathcal{W}_0} \inf_{w' \in \mathcal{W}} \mathbb{E}_{x,y}\{\ell(w'^T h'(x),y) - \ell(w_0{}^T h(x),y)\}.$$

▶ **定义30** (任务(ν,ϵ)-多样)。 对于函数类 W，如果对于 h，对所有 $h' \in H$ 均满足均匀分布条件 $w = (w_1, \cdots\ w_k)$ 是(ν,ϵ)-多样的，则称其在 W_0 上是多样的：$d_{\mathcal{W},\mathcal{W}_0}(h';h) \leq d_{\mathcal{W},w}(h';h)/\nu + \epsilon$。

那么，根据定理 28，如果任务具有 (ν,ϵ)-多样性，就可以直接得到最终保证。

▶ **定理31.** 如果 $h^* \in H$，且满足先前的假设，并且训练任务具有 (ν,ϵ)-多样性，那么以 $1-\delta$ 的概率，对于未见任务 (w_0^*, h^*)

$$\inf_{w' \in \mathcal{W}} \mathbb{E}_{x,y}\{\ell(w'^T \hat{h}(x),y) - \ell(w_0^{*T} h^*(x),y)\} \leq$$

$$\mathcal{O}\Big(\frac{LB_h}{(nk)^2} + \log(nk)[L \cdot \mathscr{O}_{nk}(\mathcal{H}) + \bar{\mathscr{O}}_n(\mathcal{W})] + \sqrt{\frac{\log(2/\delta)}{nk}}\Big)\Big)/\nu + \epsilon.$$

ˆ 该定理展示了如何利用丰富的相关数据来促进 h 的学习。因此，对于 $w \in \mathcal{W}$，使用类 $w^T \hat{}h$ 就足以捕获 $w \in \mathcal{W}$ 的类 $w^T h^*$。