arXiv:2302.08507v1 [cs.LG] 16 Feb 2023

多重校准的范围：通过性质诱导刻画多重校准

Georgy Noarov*[1] 和 Aaron Roth[1]

[1]宾夕法尼亚大学计算机与信息科学系

2023年2月17日

摘要

我们建立了多重校准与性质诱导之间的联系，并证明（在温和的技术条件下），当且仅当 Γ是 *elicitable*时，才能为连续标量性质 Γ 生成多重校准预测器。

在负面方面，我们证明，对于不可诱导的连续性质，存在一些简单的数据分布，使得即使是真实分布预测器也无法校准。在正面方面，对于可诱导的 Γ，我们为批处理设定和在线对抗设定给出简单的规范算法，以学习一个经过 Γ-多重校准的预测器。这推广了此前关于均值和分位数多重校准的工作，实际上还强化了现有的在线分位数多重校准结果。

为进一步平衡我们的负面结果，我们证明，如果性质 Γ^1 本身不可诱导，但 *is* 可由另一种可诱导性质 Γ^0 上的 *conditionally* 诱导，那么就存在一种规范算法，可以对 Γ^1 和 Γ^0 进行 *jointly* 多重校准；这推广了此前关于均值-矩多重校准的工作。

最后，作为我们理论的应用，我们为公平（多重校准的）风险评估提供了新颖的算法结果和不可能性结果。

*这项工作部分是在作者访问西蒙斯计算理论研究所期间完成的。

1 引言

考虑一个分布 D，它定义在由具有可观测特征$x \in \mathcal{X}$ 和标签 $y \in \mathbb{R}$的样例构成的带标签数据域 $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ 上。预测器 f： $\mathcal{X} \to \mathbb{R}$ 是（均值）校准的，通俗地说，就是即使以其自身的预测为条件，也能正确估计标签值的均值：即，对于所有预测 v， $\mathbb{E}_{(x,y)\sim D}[y|f(x) = v] = v$。校准是一种理想的性质，但较为薄弱，因为它只涉及标签的平均值，即对所有满足 $f(x) = v$ 的样例取平均；例如，可能存在一个结构化子集 $G \subset \mathcal{X}$，使得对于样例 $x \in G$， f 会系统性地低估标签均值——如果它通过高估 $x \notin G$ 的标签均值进行补偿，这样的预测器仍然可以是校准的。

多重校准由H´ebert-Johnson 等人提出，旨在强化校准的概念。多重校准预测器由群组集合 $\mathcal{G} \subseteq 2^{\mathcal{X}}$参数化；它不仅整体上是校准的，而且在条件于属于 G 中任意群组 $G \in \mathcal{G}$时也经过校准。也就是说，对于所有 v， G我们必须满足：

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[y|f(x) = v, x \in G] = v.$$

Jung 等人将多重校准从均值推广到矩。 [2021]引人深思的是，Jung 等人通过一个明确的反例 [2021] 证明，方差（以及其他高阶矩）本身无法实现多重校准，但可以实现多重校准，前提是与均值联合进行，即作为（均值，矩）对的一部分。后来，Gupta 等人 [2022] 和Jung 等人 [2023] 展示了如何得到多重校准的分位数类比形式；该形式要求，对于任意目标覆盖水平$\tau \in [0, 1]$、任意 v 预测器值域中的值 f 以及任意 $G \in \mathcal{G}$:

$$\mathbb{E}[\mathbb{I}[y \le f(x)]|f(x) = v, x \in G] = \tau.$$

因此，迄今为止，我们已经针对若干自然的分布性质（均值和分位数）提出了高效的批量[H´ebert-Johnson 等人，2018 年，Jung 等人， 2021, 2023] 和在线 [Gupta 等人，2022年，Bastani 等人， 2022] 多重校准算法；针对方差和高阶矩，得到了一个不可能性结果；此外，还有一项结果表明，尽管矩本身无法单独进行多重校准，但仍可将均值和矩一起进行多重校准 [Jung 等人， 2021]。不过，这些是彼此孤立的结果，还是说存在一种针对分布性质的多重校准一般理论——也就是说，针对任意泛函 Γ： $\mathcal{P} \to \mathbb{R}$将任意分布 $P \in \mathcal{P}$ 映射为标量统计量 $\Gamma(P)$？ 为研究这一问题，自然可以定义（参见 Dwork 等人 [2022]）： 若对于所有群组 $G \in \mathcal{G}$ 以及 f值域中的值 γ，均满足以下条件，则称预测器 f 对于性质 Γ 是（$\mathcal{G}$， Γ）-多重校准的:

$$\Gamma(D|(f(x) = \gamma, x \in G)) = \gamma,$$

其中， $D|(f(x) = \gamma, x \in G)$ 是在事件 $\{x: f(x) = \gamma, x \in G\}$ 条件下的数据分布。（当 $\mathcal{G} = \{\mathcal{X}\}$ 时，我们只需说 f 是 Γ-校准的。）

我们的动机 在发展性质多重校准理论时，我们致力于回答几个特别重要的问题:

1. 哪些我们感兴趣的分布性质可以实现多重校准，哪些不能？ 2. 对于那些在批处理设定下可以实现多重校准的性质，我们是否总能在在线对抗设定下也找到解决方案，还是存在在线-离线分离？ 3. 无论是在批处理设定还是在线设置下，我们能否提出一种自然的规范算法，它具有简单且通用的性能保证，能够接收（以某种简单格式给出的）任意可实现多重校准的相关性质的描述，并输出相应的多重校准预测器？ 4. 对于那些本身不可实现多重校准、但在实践中很重要的性质，是否存在足够通用的技术来解决这一问题，使我们仍能实现某种经过修改的多重校准概念？ （参见方差的例子：如 Jung 等人所示，将方差与经多重校准的均值打包在一起，就能使其重新成为可实现多重校准的性质。） [2021]。

1.1 我们的结果

我们通过将性质多重校准与久经研究的性质诱导理论联系起来，对所有这些问题给出了（几乎）完整的答案。

现代的性质诱导理论表述由 Lambert 等人 [2008], 提出，但该理论在早期和后续的研究中都得到了广泛发展。简而言之，如果对于任意数据分布，都能通过在数据集上最小化某个损失函数来直接求得某个性质的值，那么我们就称该性质是可诱导的。例如，均值和分位数都是可诱导的，因为它们分别可以通过最小二乘法和分位数回归求得。但方差就不是可诱导的，因此通常要先预测均值，再计算围绕该均值的方差。

一个等价的概念（在温和假设下）是可识别属性：识别函数（通常是关于该属性“损失函数”的一阶条件）能够根据数据集上的期望，告诉我们对属性值的估计是过高还是过低。例如，均值预测器 f_m 的期望识别函数就是 $\mathbb{E}_{(x,y)}[f_m(x) - y]$，而 τ-分位数预测器 f_τ的期望识别函数是 $\Pr_{(x,y)}[y \le f_\tau(x)] - \tau$，即 f_τ的平均超额覆盖率。下面给出一系列结果。

1. 可行性判据： Γ-多重校准当且仅当 Γ 是可诱导的。 我们给出一个非常一般的 当且仅当 刻画：在温和假设下，将各种感兴趣的分布性质归类为可能进行多重校准或不可能进行多重校准。我们证明，性质 Γ 具备校准合理性 （见定义 12）当且仅当它是可诱导的，当且仅当它是可识别的。见第 4 节第 3节。一项关键工具是 Steinwart 等人 [2014]的一个核心结果：（在温和条件下），性质 Γ 是可诱导的 $\iff$当且仅当它是可识别的 $\iff$，且其水平集是凸集；关键在于，我们将 Γ-校准的敏感性与 Γ的水平集的凸性紧密联系起来。

2. 通用的规范批处理算法和在线算法。 我们确定了两个“规范的” Γ-多重校准算法，用于可诱导的性质 Γ：批处理算法 1 和在线对抗式算法 5，并证明了它们的收敛保证。分别见第 4 节和第 6节。我们的批处理算法自然扩展了关于均值和分位数的已知方法 [H´ebert-Johnson 等人，2018 年、Jung 等人、2023]。我们的在线算法推广了关于均值和分位数的现有在线算法 [Gupta 等人、2022年、Bastani 等人、2022年以及Lee 等人， 2022] （并且在分位数方面取得了强于以往工作的结果）。

3. 条件可诱导性质的联合多重校准。我们证明，如果某个性质Γ^0 是可诱导的，并且 Γ^1 在给定条件下是可诱导的 Γ^1 （非正式地说，这意味着在已知 Γ^0确切值的条件下，存在一种回归程序来学习 Γ^1），那么（在满足技术条件时），这一对（Γ^0， Γ^1）可以通过规范算法 3 进行联合多重校准。这推广了 Jung 等人 [2021]提出的（均值，矩）多重校准。参见第5节。

4. 应用：公平风险评估的正面与负面结果。此前关于均值、联合均值-矩以及分位数多重校准的研究，为将多重校准应用于风险估计的前景提供了理论和实证依据。然而，过去对于分位数和方差之外的大量其他风险度量的多重校准，一无所知。在第7节中，我们开始填补这一空白，将我们的理论应用于金融风险评估中一系列具有核心意义的风险度量，并推导出相关结果。例如，我们证明了一个一般性的否定结果：除均值、分位数（以及另外两种技术性分位数变体）外，大类扭曲风险度量均无法进行多重校准。另一方面，我们证明，所谓的贝叶斯风险可以与其所度量风险的可诱导性质进行联合多重校准。条件风险价值（CVaR），又称预期损失（ES），就是一个例子。它在理论和实践上都具有核心意义。正如我们所证明的，条件风险价值本身无法进行多重校准，但可以与分位数进行联合多重校准。

1.2 其他相关工作

本文的主要概念性贡献是将多重校准领域的文献与相关领域的文献联系起来。

性质诱导，两者都有许多相关的研究脉络。

多重校准 （均值）多重校准在批处理设定中由 H´ebert-Johnson 等人[2018], 提出，随后又以多种方式得到推广。正如前文所述，Jung 等人[2021] 研究了批处理设定中的均值条件矩多重校准，Gupta 等人 [2022]研究了序贯设定中的均值、分位数和均值条件矩多重校准，而Jung 等人 [2023] 研究了批处理设定中的分位数多重校准。这些推广可以被视为要求相对于不同分布性质进行校准——也就是说，它们属于我们在本文中刻画的那类推广。Dwork 等人 [2021] 研究了结果不可区分性，将二元标签设定中的批处理多重校准推广为允许 区分器 评估三元组的任意谓词的期望（x， f（x） y），并考虑了区分器可以进一步访问 f 的强化版本——例如，能够在任意点对其进行查询，或能够访问其代码。特别是，他们表明，在没有对 f额外访问权限的情况下，结果不可区分性可以归约为（均值）多重校准。请注意，许多分布性质（例如方差、分位数等）只有在标签空间从二元值变为实值时才会变得有意义。

与本文最密切相关的研究探讨了多重校准的抽象推广。Dwork 等人[2022] 研究了结果不可区分性向实值标签的推广，并在此过程中考察了相对于线性化统计量的批量多重校准。用我们的术语来说，这些是分布性质 Γ ，它们在混合分布上呈线性变化——直观地说，对于任意两个分布 D_1、 D_2 以及任意 $\alpha \in [0,1]$、 $\Gamma(\alpha D_1 + (1-\alpha) D_2) = \alpha\Gamma(D_1) + (1-\alpha)\Gamma(D_2)$ 。我们采用了他们关于相对于一般分布性质进行多重校准的其中一种定义。所有线性化统计量都具有凸水平集（因而是可诱导的），但并非所有可诱导性质都具有这种意义上的线性——例如，分位数无法线性化。因此，我们对多重校准的刻画表明，相较于 Dwork 等人 [2022]研究的性质，可以针对更广泛的一类性质实现多重校准。Lee 等人 [2022] 研究了一般的在线学习问题，并将其称为“在线极小极大多目标优化”，进而在这一框架下推导出用于多重校准及其他若干应用的算法。我们利用这一框架推导序贯多重校准的界，并采用了一个源自我们与性质诱导之间联系的识别函数。Deng 等人 [2023] 的近期研究探讨了多重校准的一维推广，要求满足如下条件：对于抽象函数 c和 s， $\mathbb{E}[c(f(x) \ x) s(f(x) \ y)] = 0$；该研究还推导出了在批处理设定下实现这一条件的充分（但非必要）条件。我们为批量多重校准推导的算法与他们的算法类似，其中识别函数取代了他们的 s 函数，而评分函数取代了他们的势函数；他们并未考虑序贯问题或多维问题。相较于这条研究脉络，我们的结果首次刻画了何时能够实现性质多重校准，并为批量和序贯多重校准提供了统一的结果。

此外，多重校准也存在沿正交方向的推广。Gopalan 等人 [2022b] 定义了“低次多重校准”，它是一个关于预测器的性质层级：这些预测器仍试图预测均值，但放宽了对 $f(x) = v$ 的条件限制。层级的底部是多重准确性（H´ebert-Johnson 等人 [2018], Kim 等人 [2019] ），它完全不以 $f(x)$ 为条件。多重校准位于层级顶端；介于两者之间的条件仅通过一个次数为 k 的多项式平滑地依赖于 $f(x)$。Gopalan 等人 [2022b] 证明，这一层级的中间层次具有多重校准的一些理想性质，而且更容易实现。若干研究[Kim 等人，2019年，Gopalan 等人，2022a，Kim 等人，2022年，Globus-Harris 等人，2023]探讨了均值多重校准的推广：其中，代表数据域子集的“群组”被放宽为任意实值函数，并由此给出了一些应用。在这一框架下，Globus-Harris 等人[2023] 刻画了均值多重校准何时蕴含贝叶斯最优性。关于这方面工作的入门介绍，可参见 Roth[2022] 。

性质诱导 Brier [1950], Good [1992], 和 Savage [1971] 研究适当评分规则。这类规则是一种契约，根据专家的预测和实际结果向其支付报酬；如果专家如实报告结果事件的真实概率，就能在期望意义上获得最高奖励。由于评分规则可以直接诱导概率，原则上可以通过诱导支撑集中每个事件的概率来诱导整个概率分布，但这通常并不可行。因此，Lambert 等人 [2008] 提出了 性质诱导问题，其目标是设计契约，激励专家如实报告大规模或无限支撑集分布的某种 性质 ——例如均值、方差、中位数等。通俗地说，如果存在某个函数，使得 .....

报告与结果的函数在对结果取期望后，于该性质的取值处达到最小值，那么这个性质就是可诱导的。如今，性质诱导领域已有大量文献；本文将运用其中若干关键结果。 Osband[1985] 和 Gneiting [2011] 定义了性质的识别函数。识别函数与评分规则一样，是报告与结果的函数；当报告等于该性质的取值时，识别函数对结果取期望后的值为 0。 Steinwart 等人 [2014] 证明了一个核心刻画定理（在满足一些较弱的技术条件时）：一个连续性质可诱导，当且仅当它具有识别函数；也当且仅当它具有凸水平集。 Steinwart 等人 [2014] 的刻画定理普遍适用于连续结果空间。当结果空间有限时，Finocchiaro 和 Frongillo [2018] 证明（在满足技术条件时），可诱导的性质可以通过凸评分规则来诱导。

2 预备知识

我们研究 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$上带标签数据点的预测问题，其中 $\mathcal{X}$ 是特征向量空间，而 $\mathcal{Y} \subseteq \mathbb{R}$ 是标签空间。我们研究批处理（离线）和序列（在线）预测问题。在线设置将在第6节讨论，序列预测的定义和预备知识也留待该节介绍。非正式地说，性质多重校准在序贯设定中的定义与此处在批处理设定中的定义完全相同，只是将这里所考虑的底层数据集分布替换为已实现数据序列上的经验分布。

在批处理设定中，带标签样本服从分布 $D \in \Delta\mathcal{Z}$ 。该分布诱导出特征上的分布 X 和标签上的分布 Y。将 D 隐去不写，我们令 $Y_x := (Y|\{X = x\})$ 表示给定特征向量 $x \in \mathcal{X}$ 时的条件标签分布；更一般地，给定子集 $G \subseteq \mathcal{X}$ 时，我们将以 $x \in G$ 为条件的条件标签分布记为 $Y_G := (Y|\{x \in G\})$。预测器或模型就是一个实值映射 $f\colon \mathcal{X} \to \mathbb{R}$。

无论离线还是在线，我们的目标都是针对某种分布性质作出（多重）校准的预测。现在，我们介绍这两个概念所需的背景知识。

2.1 分布性质

一维（分布）性质是一个泛函 $\Gamma\colon \mathcal{P} \to \mathbb{R}$，其中 $\mathcal{P}$ 是实值随机变量的某个概率分布空间。我们用 Range_Γ 表示 Γ 的值域，并且不失一般性，在本文中将性质缩放至使 $\text{Range}_\Gamma \subseteq [0, 1]$。

分布性质的例子包括均值、中位数、 τ分位数，或分布的方差（更多示例请参见第 7 节）。这些概念的共同之处在于，它们都将一个给定的分布与一个实数对应起来。

形式化定义 $\mathcal{P}$ 在本文中，我们对分布空间 $\mathcal{P}$ 所作的基本假设是：其定义域 Γ 是某个向量空间的凸子空间，因此，对 $\mathcal{P}$ 中的分布取凸组合是一个定义良好的运算。

分布 $P \in \mathcal{P}$ 定义在标签空间 $\mathcal{Y}$上。根据 $\mathcal{Y}$ 是否为有限集合，我们有两种便捷的方式来定义 $\mathcal{P}$ 的结构。如果 $\mathcal{Y}$ 是有限集合（$|\mathcal{Y}| = d < \infty$），那么每个分布 $P \in \mathcal{P}$ 都可以嵌入到 $\mathbb{R}^d$ 中，作为 d 维单纯形 $\Delta(d)$ 的一个元素。因此，我们只需定义 $\mathcal{P} \subseteq \Delta(d) \subset \mathbb{R}^d$ 是单纯形 $\Delta(d)$ 的凸子集，并赋予其通常的欧几里得范数。

如果 $\mathcal{Y}$ 是无限的，我们假设 $\mathcal{P} \subseteq W_{\text{TV}}$，其中 W_{TV} 是一个巴拿赫空间（即完备赋范向量空间；参见Diestel 和 Uhl [1977]，了解关于巴拿赫空间的通俗入门介绍），它由定义在 $\mathcal{Y}$ 上的概率分布构成，这些分布几乎处处具有有界密度，并赋予全变差范数 $||\cdot||_{\text{TV}}$。特别地， W_{TV} 是一个度量空间，其中任意两个概率分布 P_1， $P_2 \in \mathcal{P}$之间的距离按全变差距离 $||P_1 - P_2||_{\text{TV}}$计算。

我们经常对 Γ 作出的一项假设是连续性：也就是说， Γ 在非常相似的分布上不能取相差悬殊的值。（许多性质，包括均值、方差、分位数、熵等，确实都是连续的。）

定义 1（连续性质）。如果泛函 $\Gamma\colon \mathcal{P} \to \mathbb{R}$ 连续（相对于 $\mathcal{P}$ 上的度量拓扑和 $\mathbb{R}$ 上的标准拓扑），则称 Γ 为连续性质。

数据集上的性质预测。在我们的批处理设定中，我们处理定义在 $\mathcal{X} \times \mathcal{Y}$ 上的数据集，并且希望训练预测器 $f_\Gamma : \mathcal{X} \to \mathbb{R}$，以预测 $\mathcal{Y}$ 上条件标签分布Y_x 的性质 Γ。非正式地说，对于任意 $x \in \mathcal{X}$，性质 Γ 的良好预测器应满足 $f_\Gamma(x) \approx \Gamma(Y_x)$。

由于我们研究的是在数据集的标签空间 $\mathcal{Y}$上的某个分布族 $\mathcal{P}$ 上定义的性质 $\Gamma : \mathcal{P} \to \mathbb{R}$，因此需要将注意力限定在这样的数据集分布 $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ 上：对于所有 $x \in \mathcal{X}$，其诱导的标签分布 Y_x 都属于 $\mathcal{P}$，从而至少能确保该性质 Γ 在给定任意 $x \in \mathcal{X}$的每个条件分布上都有定义良好的取值。我们将其形式化如下。

定义 2（$\mathcal{P}$-兼容数据集分布）。给定标签空间 $\mathcal{Y}$上的一个分布族 $\mathcal{P} \subseteq \Delta\mathcal{Y}$，如果对于所有$x \in \mathcal{X}$，给定 x 时诱导的标签分布都属于 $\mathcal{P}$，即满足 $Y_x \in \mathcal{P}$，则称数据集空间 $\mathcal{X} \times \mathcal{Y}$ 上的数据集分布 $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ 是 $\mathcal{P}$-兼容的。

2.2 性质诱导与识别

现在，我们可以正式定义性质诱导文献所研究的三个相关概念：分布性质的可诱导性、可识别性和水平集凸性（CxLS）。对于任何关注的性质而言，这三者都是理想属性，并且彼此密切相关——事实上，正如我们即将讨论的，在关于分布性质的一些技术性假设下，它们彼此等价。

可诱导性 简而言之，若定义在分布族 $\mathcal{P}$ 上的性质，其在任意分布 $P \in \mathcal{P}$ 上的取值都可以通过最小化对来自 P 的样本的某个损失函数的期望来获得，那么该性质就称为可诱导的；换言之，用统计学习的术语来说，就是通过求解一个回归问题来获得。按照诱导性文献中的惯例，我们将这类损失函数称为评分函数：从数学上说，评分函数只是一个函数 $S : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$。

定义 3（严格一致评分函数）。固定一个概率分布空间 $\mathcal{P}$。若评分函数 $S : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ 对性质 $\Gamma : \mathcal{P} \to \mathbb{R}$ 是严格 $\mathcal{P}$-一致的，则：

$$\Gamma(P) = \operatorname*{argmin}_{\gamma \in \mathbb{R}} \mathop{\mathbb{E}}_{y \sim P}[S(\gamma, y)] \quad \textit{for all } P \in \mathcal{P}.$$

我们也称 S 诱导 Γ。

为简洁起见，我们记作：$S(\gamma, P) = \mathbb{E}_{\gamma \sim P}[S(\gamma, y)]$。

若对于所有满足 $|\gamma_1 - \Gamma(P)| \lhd \gamma_2 - \Gamma(P)|$的 $P \in \mathcal{P}$和 $\gamma_1, \gamma_2 \in \mathbb{R}$，都有 $S(\gamma_1, P) < S(\gamma_2, P)$，则称 S对 Γ具有 $\mathcal{P}$ 阶敏感性。

定义 4（可诱导性质）。固定一个概率分布空间 $\mathcal{P}$。若性质 $\Gamma : \mathcal{P} \to \mathbb{R}$具有严格 $\mathcal{P}$一致的评分函数，则称其为可诱导的。

上述定义的一个基本例子是，由 $S(\gamma, y) = (\gamma - y)^2$定义的评分函数能够诱导分布均值；因此，均值是一种可诱导性质。

然而，诱导性文献的一项重要启示是，可诱导性不应被想当然地视为理所当然。例如，分布的方差无法直接通过最小化某种损失来诱导，除非先估计例如均值；因此，方差并非可诱导性质。我们将在第7节讨论更多此类例子。

水平集凸性（CxLS）可诱导性一个简单却深刻的必要条件最早由Osband [1985], 提出：可诱导性质必须具有凸水平集（也称为$CxLS$）。这一点对于我们通过可诱导性刻画校准的敏感性至关重要。这个论断简单却在概念上意义重大，因此我们完整列出证明。

事实 1（Osband [1985]）。设 $\mathcal{P}$ 是一个由概率分布构成的凸空间，且 $\Gamma : \mathcal{P} \to \mathbb{R}$ 是一个可诱导性质。那么对于所有 $\gamma \in$ Ranger，水平集 $\{P \in \mathcal{P} : \Gamma(P) = \gamma\}$ 都是凸的：对于任意 P_1、P_2，若 $\Gamma(P_1) = \Gamma(P_2) = \gamma$，则 $\Gamma(\lambda P_1 + (1-\lambda)P_2) = \gamma$ 对所有 $\lambda \in [0, 1]$均成立。

证明。固定任意两个分布 P_1、$P_2 \in \mathcal{P}$，使得 $\Gamma(P_1) = \Gamma(P_2) = \gamma$，并令 S 为针对 Γ 的严格$\mathcal{P}$一致评分函数。由于 S 是严格一致评分函数，我们有

$\gamma = \arg\min_{\gamma'} S(\gamma', P_1) = \arg\min_{\gamma'} S(\gamma', P_2)$。对于任意 $\alpha \in [0,1]$，考虑 $\hat{P} = \alpha P_1 + (1-\alpha)P_2$。根据 $\mathcal{P}$ 的 $\hat{}$ 凸性，我们有 $P \in \mathcal{P}$；根据期望的线性性，对于任意 $\gamma' \in \mathbb{R}$，我们有：

$$\begin{aligned} S(\gamma, \hat{P}) &= \alpha S(\gamma, P_1) + (1-\alpha) S(\gamma, P_2) \\ &< \alpha S(\gamma', P_1) + (1-\alpha) S(\gamma', P_2) \\ &= S(\gamma', \hat{P}) \end{aligned}$$

因此，$\gamma = \arg\min_{\gamma'} S(\gamma', \hat{P})$，从而 $\Gamma(\hat{P}) = \gamma$，这就证明了该结论。 □

可识别性 与上述两个概念相关的一个概念是可识别性，由Osband [1985]和Gneiting [2011]提出。它要求一项性质具有所谓的识别函数：

定义 5（识别（Id）函数）。函数 $V\colon \mathbb{R} \times \mathcal{Y}$是性质 $\Gamma : \mathcal{P} \to \mathbb{R}$ 的 $\mathcal{P}$-识别函数（或简称 id 函数），如果对于每个 $P \in \mathcal{P}$：

$$\mathop{\mathbb{E}}_{y \sim P}[V(\gamma, y)] = 0 \Leftrightarrow \Gamma(P) = \gamma.$$

如果 $\mathcal{P}$-识别函数对应于 Γ，且满足以下条件，则称其为定向识别函数：$\mathbb{E}_{y \sim P}[V(\gamma, y)] > 0 \Leftrightarrow \gamma > \Gamma(P)$。为简化记号，我们写作：$V(\gamma, P) = \mathbb{E}_{\gamma \sim P}[V(\gamma, y)]$。

换言之，可以将性质值识别为其恒等函数在该分布上的期望值的零点。对于定向识别函数，我们还可以得出：高估（分别是低估）性质值会导致正（分别是负）的期望识别函数值。

定义 6（可识别属性）。若存在一个用于 Γ 的 $\mathcal{P}$ 识别函数，则称性质 $\Gamma : \mathcal{P} \to \mathbb{R}$ 可识别。

直观而言，与可诱导性的联系在于，例如，若某个性质的评分函数是凸且可微的，那么它关于第一个参数的导数便是同一性质的识别函数。

可诱导性、可识别性与 CxLS 之间的联系Steinwart 等人证明，在温和假设下，连续分布属性的可诱导性、可识别性与 CxLS 性质 Γ 实际上是等价的。Steinwart 等人 [2014]。从本质上说，Osband 的结果（事实1）指出（除 $\mathcal{P}$ 为凸之外，不作其他假设），CxLS 是可诱导性的必要条件；而 Steinwart 等人 [2014] 的刻画则表明，它也是充分条件，但需满足其他技术假设（并证明在相同假设下，可识别性与可诱导性等价）。

形式上，Steinwart 等人 [2014] 证明了他们对 $\mathcal{P}$ 以下两种定义之一的刻画：

定义 7。我们将 $\mathcal{P}_0$ 定义为巴拿赫空间 W_{TV} 的子空间，其中包含所有密度几乎处处有上界的概率分布。

我们将 $\mathcal{P}_{>0}$ 定义为 $\mathcal{P}_0$ 的子空间，其中包含所有密度处处都被某个 $\epsilon_P > 0$ 从下方界定的概率分布 $P \in \mathcal{P}_0$。

定理 1（Steinwart 等人 [2014]）。考虑一个概率分布空间 $\mathcal{P} \in \{\mathcal{P}_0, \mathcal{P}_{>0}\}$。$\Gamma$： $\mathcal{P} \to \mathbb{R}$ 为任意连续、严格局部非恒定的[1] 性质。那么，以下陈述等价：

1. Γ 是可诱导的。
2. Γ 具有一个有界非负序敏感评分函数 S。
3. Γ 可识别，且具有有界的定向识别函数 V。
4. Γ 具有凸水平集：对于 Γ 值域中的任意 γ，$\{P \in \mathcal{P} : \Gamma(P) = \gamma\}$ 是凸的。

[1]“严格局部非恒定”是一项（较弱的）要求：对于 $\mathcal{P}$内部满足 $\Gamma(P) = \gamma$ 的每个 P，以及 $\mathcal{P}$ 上度量拓扑中 P 的任意 ϵ-邻域 U_ϵ，都存在分布 P'、$P'' \in U_\epsilon$，使得 $\Gamma(P'') < \gamma < \Gamma(P')$。

此外，如果 Γ 是可诱导的，那么它具有一个规范有界识别函数 V^*，使得对于 Γ 的每个局部利普希茨连续序敏感评分函数 S，都可以写成：

$$S(\gamma, y) = \int_{\gamma_0}^{\gamma} V^*(r, y) w(r) dr + \kappa(y)$$

对于某个 $\gamma_0 \in$值域Γ、某个有界非负加权函数 w，以及一个仅依赖于标签的函数 κ。

2.3 性质预测器的校准与多重校准

现在，我们给出批处理设定下针对任意分布性质的预测器的校准和多重校准的一般定义。序贯多重校准的定义推迟到第 6 节介绍。这些定义的一种变体最早见于 Dwork 等人 [2022]，当时称为校准混合下的一致性。这是对批处理均值和分位数校准误差的推广，相关研究见H´ebert-Johnson 等人 [2018], Jung 等人 [2023]。

固定一个定义在 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ 上的数据集，其数据分布为 $D \in \Delta\mathcal{Z}$。假设对于任意特征 $x \in \mathcal{X}$，我们都希望预测性质 Γ 在 Y_x 上的取值，即给定 x 时的条件标签分布。为此，我们构造一个 Γ-预测器 $f: \mathcal{X} \to \mathbb{R}$。如果对于所有 $\gamma \in$值域f，给定预测值 $f(x) = \gamma$ 的条件标签分布确实具有性质值 γ，我们就称该预测器为 Γ-校准的。

定义 8（性质 Γ的校准预测器）。若一个 Γ-预测器 $f: \mathcal{X} \to \mathbb{R}$ 满足以下条件，则称其在数据集分布 $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ 上是 Γ-校准的：对于每个 $\gamma \in$ 值域f：

$$\Gamma(Y_{f,\gamma}) = \gamma,$$

其中 $Y_{f,\gamma} = Y_{\{x: f(x)=\gamma\}}$ 是由 D 在给定 $f(x) = \gamma$ 条件下所诱导的条件标签分布。

现在，我们将这一定义扩展为多重校准：对于特征空间 $\mathcal{X}$ 中任意子集（“群组”）集合都成立的校准保证。

定义 9（性质 Γ 的 $\mathcal{G}$-多重校准预测器）。固定一个群组集合 $\mathcal{G} \subseteq 2^{\mathcal{X}}$。若一个 Γ-预测器 $f: \mathcal{X} \to \mathbb{R}$ 满足以下条件，则称其在数据集分布 $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ 上是 $(\mathcal{G}, \Gamma)$-多重校准的：对于每个$\gamma \in$ 值域f 和 $G \in \mathcal{G}$：

$$\Gamma(Y_{f,\gamma,G}) = \gamma,$$

其中 $Y_{f,\gamma,G} = Y_{\{x: f(x)=\gamma, x \in G\}}$ 是在给定 $f(x) = \gamma$ 和 $x \in G$ 的条件下由 D 诱导的条件标签分布。换言之，一个 $(\mathcal{G}, \Gamma)$-多重校准预测器具有这样的性质：在预测为 $f(x) = \gamma$ 的条件下，同时在 x 属于群组 G 这一事件的条件下，条件标签分布确实具有性质值 γ。

接下来我们需要使用一个关于近似有限值域预测器多重校准的定义。我们采用一种 ℓ_2校准误差的概念，它推广了 Jung 等人 [2023]所定义的近似分位数多重校准。该保证比文献中更常见的 ℓ_∞校准误差概念更强，例如 Hebert-Johnson 等人´、Jung 等人 [2018], 、Deng 等人 [2021], 所研究的概念；正如 Roth [2023], 所讨论的，它还可以通过柯西-施瓦茨不等式与 ℓ_1 多重校准误差的变体建立联系 Roth [2022]。

定义10（关于性质 Γ的α-近似 $\mathcal{G}$-多重校准预测器）。固定一个分布$D \in \Delta\mathcal{Z}$ 和一个群组集合 $\mathcal{G} \subseteq 2^{\mathcal{X}}$。对于每个 $G \in \mathcal{G}$，令 $\mu(G) = \Pr_{(x,y)\sim D}[x \in G]$ 表示群组 G上的概率质量。有限值域预测器 $f: \mathcal{X} \to$ 值域f 在 D 上是 α-近似的（$\mathcal{G}, \Gamma$）多重校准预测器，当且仅当对所有 $G \in \mathcal{G}$：

$$\sum_{\gamma \in Range_f} \Pr_{(x,y)\sim D}[f(x) = \gamma | x \in G] \left(\gamma - \Gamma(Y_{f,\gamma,G})\right)^2 \leq \frac{\alpha}{\mu(G)}.$$

请注意，0-近似的（$\mathcal{G}, \gamma$）-多重校准等价于我们对（$\mathcal{G}, \Gamma$）-多重校准的定义。

3 校准与可诱导性的敏感性

现在，我们可以将性质诱导与（多重）校准联系起来。首先，我们定义性质 Γ 对校准而言合理这一概念。

定义 11（性质的真实分布预测器）。固定一个分布性质 $\Gamma: \mathcal{P} \to \mathbb{R}$和一个与 $\mathcal{P}$相容的数据集分布 D。性质 Γ 在 D 上的真实分布预测器 f_Γ^D定义为：对 $x \in \mathcal{X}$， $f_\Gamma^D(x) = \Gamma(Y_x)$；也就是说，该预测器对每个 $x \in \mathcal{X}$ 都给出给定 x时条件标签分布上的性质 Γ 的正确值。

定义 12（适宜校准的性质）。固定一个性质 $\Gamma: \mathcal{P} \to \mathbb{R}$，以及一组 $\mathcal{P}$相容的数据集分布 $\mathcal{D}$。如果对所有 $D \in \mathcal{D}$，真实分布预测器 f_Γ^D在 D 上都是 Γ-校准的，我们就称 Γ 在 $\mathcal{D}$ 上具有校准合理性。

多重校准的一项关键动机（Dwork 等人 [2021]中对此有详细阐述）是，我们希望生成一个预测器 f，它相对于由 $\mathcal{G}$参数化的一类校准检验而言，无法与 f_Γ^D区分开来——而这只有在 Γ 具有校准合理性时才有意义。一般而言，对于不具有校准合理性的性质，可能根本不存在校准预测器（即使不考虑 f_Γ^D也是如此）。

Jung 等人 [2021] 指出（用我们的术语来说），方差对于校准并不合理。下面是他们的例子。假设 $\mathcal{X} = \{x_0, x_1\}$，其中 $\mathbf{Pr}[X = x_1] = \mathbf{Pr}[X = x_2] = \frac{1}{2}$，且 $Y(x_0) = 0$和 $Y(x_1) = 1$的概率为 1。于是，真实分布预测器满足 $f_{\mathrm{Var}}^D(x_0) = f_{\mathrm{Var}}^D(x_1) = 0$（因为标签是确定的）。然而，$\mathbf{Var}(Y | f_{\mathrm{Var}}^D(x) = 0) = \mathbf{Var}(Y) = 0.25 \neq 0$。换句话说，对于每个点上标签方差均为 0 的点集，真实标签方差却非零。现在，我们将这一观察大幅推广并加以完善，得到如下刻画：在温和假设下，一个性质对于校准合理，当且仅当它是可诱导的（或者等价地，是可识别的／具有凸水平集）。

我们首先证明：如果性质 $\Gamma: \mathcal{P} \to \mathbb{R}$ 在 $\mathcal{P}$上不具有凸水平集，那么对于任何包含所有支撑于 $\mathcal{X}$ 中两个点、且这两个点各自的标签分布属于 $\mathcal{P}$ 的可能数据集的 $\mathcal{P}$-兼容数据集族而言，该性质都不具备校准合理性。

定义 13（双点数据集分布）。如果数据集分布 D 关于特征—标签对 $(X, Y) \in \mathcal{X} \times \mathcal{Y}$，则称其为双点分布：存在两个特征向量 $x_1 \neq x_2 \in \mathcal{X}$，使得 $\mathbf{Pr}_D[X \notin \{x_1, x_2\}] = 0$ 且 $\mathbf{Pr}_D[X = x_1] \neq 0$，$\mathbf{Pr}_D[X = x_2] \neq 0$ ——换言之，在分布 D 下，恰好有两个特征向量具有非零出现概率。

定理 2（无 CxLS $\implies$ 不具备校准合理性）。考虑性质 $\Gamma: \mathcal{P} \to \mathbb{R}$，其中 $\mathcal{P}$ 是凸的；再考虑任意由 $\mathcal{P}$-兼容数据集分布构成的数据集分布族 $\mathcal{D}$，该族包含所有 $\mathcal{P}$-兼容的双点数据集分布。那么，如果 Γ 在 $\mathcal{P}$ 上不具有凸水平集，则其在 $\mathcal{D}$ 上不具备校准合理性。

证明。假设 Γ 违反了分布族 $\mathcal{P}$ 的凸水平集假设。那么存在某个值 $\gamma \in$ 值域$_\Gamma$，使得 $\{P \in \mathcal{P}: \Gamma(P) = \gamma\}$ 不是凸集。等价地，存在分布 Y_1、$Y_2 \in \mathcal{P}$，满足 $\Gamma(Y_1) = \Gamma(Y_2) = \gamma$，但对于某个 $\lambda \in [0, 1]$，$\Gamma(\lambda Y_1 + (1 - \lambda) Y_2) \neq \gamma$。现在，我们需要找出一个数据集分布 $D \in \mathcal{D}$，使得真实分布预测器 f_Γ^D 不是 Γ 校准的。我们通过令 $Y_{x_1} = Y_1$、$Y_{x_2} = Y_2$ 和 $\mathbf{Pr}[X = x_1] = \lambda = 1 - \mathbf{Pr}[X = x_2]$，构造一个支撑集包含任意两个特征向量 $x_1 \neq x_2 \in \mathcal{X}$ 的数据集分布。于是立即可得 $\Gamma(Y_{f_\Gamma^D,\ \gamma}) \neq \gamma$，证明完毕。 □

为了证明逆命题，我们对数据集分布 D施加一个弱且自然的正则性假设：我们要求从特征 x 到由 D 诱导的相应 Y_x 的映射具有足够良好的性质，使得任意子集 $G \subseteq \mathcal{X}$ 上的标签分布 Y_G 都可以明确定义为各个 $x \in G$对应的标签分布 Y_x 的混合。

在 $|\mathcal{Y}| < \infty$ 的情形下，映射 $x \to Y_x$ 良好性质可通过要求其为勒贝格可测来形式化。当 $|\mathcal{Y}| = \infty$ 时，每个 Y_x 所属的空间 W_{TV} 是巴拿赫空间——在此情形下，勒贝格可测性和可积性的概念（它们仅在有限维欧几里得空间中定义）由相应的博赫纳可测性和博赫纳可积性概念取代（有关这些概念的介绍，请参阅 Diestel 和 Uhl [1977] ）。因此，当 $|\mathcal{Y}| = \infty$ 时，我们要求映射 $x \to Y_x$ 为博赫纳可测。

定义 14（$\mathcal{P}$-正则数据集分布）。固定特征空间 $\mathcal{X}$、标签空间 $\mathcal{Y}$ 和一族定义在 $\mathcal{Y}$ 上的概率分布 $\mathcal{P}$。考虑定义在 $\mathcal{X} \times \mathcal{Y}$ 上的 $\mathcal{P}$-兼容数据集分布 D。令$\xi_D : \mathcal{X} \to \mathcal{P}$ 由 $\xi_D(x) = Y_x$ 定义，即 ξ_D 是由 D诱导的、从特征向量到其标签分布的映射 $x \to Y_x$。

数据集分布 D 称为 $\mathcal{P}$-正则的，当以下任一条件成立时：

1. $\mathcal{X}$ 是有限的；

2. $\mathcal{Y}$是 有限的（$|\mathcal{Y}| < \infty$），并且当将 $\mathcal{P}$ 视为 $\mathbb{R}^{|}$的子集时， ξ_D 是勒贝格可测的 $\mathcal{Y}|$；3. $\mathcal{X}$

$\mathcal{Y}$ 是无限的，并且当将 $\mathcal{P}$ 视为 W_{TV} 的子集时， ξ_D 是博赫纳可测的 .

对于任意 $\mathcal{P}$-正则数据集 D，我们现在证明：在 $\mathcal{P}$ 上具有凸水平集的连续性质 Γ 意味着真实分布预测器 f_Γ^D 实际上是 Γ-校准的。

定理 3（CxLS $\implies$ 校准合理性）。考虑连续性质 Γ： $\mathcal{P} \to \mathbb{R}$，以及任意一族 $\mathcal{D}$ $\mathcal{P}$-正则数据集分布。则如果 Γ 在 $\mathcal{P}$ 上具有凸水平集，那么它对于 $\mathcal{D}$ 的校准是合理的。

证明。假设 Γ 在 $\mathcal{P}$上具有凸水平集。我们现在证明，真实分布预测器 f_Γ^D 在每个数据集 $D \in \mathcal{D}$ 上都是校准的：也就是说，对于所有 $\gamma \in \text{Range}_\Gamma (= \text{Range}^{f_\Gamma^D})$，我们有 $\Gamma(Y_{f_\Gamma^D,\gamma}) = \gamma$。在证明的其余部分，我们固定任意 $\gamma \in \text{Range}_\Gamma$ 以及任意数据集分布$D \in \mathcal{D}$（根据假设，该分布是 $\mathcal{P}$-正则的），并证明 $\Gamma(Y_{f_\Gamma^D,\gamma}) = \gamma$。

令 $\text{LS}_\Gamma(\gamma) = \{P \in \mathcal{P} : \Gamma(P) = \gamma\}$ 为性质 Γ 的 γ-水平集，并令 $Q_\gamma := \{x \in \mathcal{X} : Y_x \in \text{LS}_\Gamma(\gamma)\}$ 为由所有满足其标签分布属于 $\text{LS}_\Gamma(\gamma)$ 的点 $x \in \mathcal{X}$ 组成的特征空间区域。为了证明 $\Gamma(Y_{f_\Gamma^D,\ \gamma}) = \gamma$，我们需要证明 $Y_{f_\Gamma^D,\ \gamma} \in \text{LS}_\Gamma(\gamma)$。

为此，我们将 $Y_{f_\Gamma^D,\ \gamma}$ 解释为各个标签分布 Y_x（对应于 $x \in Q_\gamma$）上的混合分布。令 μ_γ 为数据集分布 D 在 Q_γ 上诱导的概率测度。那么，我们可以写出如下形式化表达式：

$$Y_{f_\Gamma^D,\gamma} = \mathop{\mathbb{E}}_{x \sim \mu_\gamma}[Y_x] := \int_{Q_\gamma} Y_x d\mu_\gamma.$$

$\mathcal{X}$ 是有限的：在这种情况下，我们可以直接将 $Y_{f_\Gamma^D,\ \gamma}$ 写成对应于 $x \in Q_\gamma$ 的组成分布 Y_x 的凸组合：

$$Y_{f_\Gamma^D,\gamma} = \sum_{x \in Q_\gamma} \mu_\gamma(x) \cdot Y_x.$$

根据水平集 $\text{LS}_\Gamma(\gamma)$ 的凸性，由于对于每个 $x \in Q_\gamma$， $Y_x \in \text{LS}_\Gamma(\gamma)$，因此上述 $x \in Q_\gamma$ 的 Y_x 凸组合属于 $\text{LS}_\Gamma(\gamma)$，从而可得 $Y_{f_\Gamma^D,\ \gamma} \in \text{LS}_\Gamma(\gamma)$。

$\mathcal{Y}$ 是有限的：在这种情况下，期望 $\mathbb{E}_{x \sim \mu}[Y_x]$被定义为随机变量 Y_x 在单纯形 $\Delta(d) \subset \mathbb{R}^d$上的勒贝格积分。根据数据集是 $\mathcal{P}$-正则的假设，我们有 Y_x 是一个有界的（例如在 ℓ_∞ 范数下）且勒贝格可测的随机变量。因此，Y_x 实际上是勒贝格可积的，所以期望 $\mathbb{E}_{x \sim \mu}[Y_x]$ 是良定义的，并且其值为某个点 $u \in \mathbb{R}^d$。接下来只需证明 $u \in \text{LS}_\Gamma(\gamma)$。

为此，引入示性函数 $\mathbb{I}^{\text{LS}_\Gamma(\gamma)}$： $\Delta(d) \to \{0\} \cup \{+\infty\}$，其定义为：当$Y_x \in \text{LS}_\Gamma(\gamma)$时取0，否则取 ∞。由于集合$\text{LS}_\Gamma(\gamma)$是凸集，其示性函数 $\mathbb{I}^{\text{LS}_\Gamma(\gamma)}$ 是凸函数。因此，我们可以对 $\mathbb{I}^{\text{LS}_\Gamma(\gamma)}$ 应用詹森不等式，从而得到：

$$\mathbb{I}^{\text{LS}_\Gamma(\gamma)}(u) = \mathbb{I}^{\text{LS}_\Gamma(\gamma)}\left(\mathop{\mathbb{E}}_{x \sim \mu_\gamma}[Y_x]\right) \leq \mathop{\mathbb{E}}_{x \sim \mu_\gamma}\left[\mathbb{I}^{\text{LS}_\Gamma(\gamma)}(Y_x)\right] = 0,$$

这意味着 $\mathbb{I}^{\text{LS}_\Gamma\ (\gamma)}(u) = 0$。根据 $\mathbb{I}^{\text{LS}_\Gamma\ (\gamma)}$ 的定义，这证明了 $u \in \text{LS}_\Gamma(\gamma)$，正如所需。

$\mathcal{X}$，$\mathcal{Y}$无穷情形：在此情况下，我们将 $\mathbb{E}_{x\sim\mu_\gamma}[Y_x]$ 定义为 博赫纳积分 （其定义和性质参见 Diestel 和 Uhl [1977]）作用于博赫纳可测映射 ξ_D 所得到的结果。根据标准的博赫纳可积性判据（参见Diestel 和 Uhl [1977]第 45 页的定理 2），该积分确实存在，并且其值属于环境空间 W_{TV}；这是因为很容易验证 $\mathbb{E}_{x\sim\mu_\gamma}[||Y_x||_{\mathrm{TV}}] < \infty$ （事实上，任意概率分布的全变差范数均为 1，因此 $\mathbb{E}_{x\sim\mu_\gamma}[||Y_x||_{\mathrm{TV}}] = 1 < \infty$）。同样，我们希望证明 $Y_{f^D_\Gamma,\ \gamma} = \mathbb{E}_{x\sim\mu_\gamma}[Y_x] \in \mathrm{LS}_\Gamma(\gamma)$。为此，我们使用以下结果，它可以被理解为博赫纳积分的中值定理：

事实 2 （Diestel 和 Uhl [1977]第 48 页上的推论 8）。设 （Ω $\mathcal{A}$，μ） 为有限测度空间， E为巴拿赫空间，且 $f:\Omega \to E$ 为博赫纳 μ可积函数。对于 $G \subseteq E$，令 co（G） 表示 G的凸包的闭包。则，对于每个 $A \in \mathcal{A}$ 满足 μ（A） >0的，均有

$$\frac{1}{\mu(A)}\int_A f d\mu \in \overline{co}(f(A)).$$

为应用这一事实，我们令：（1）Ω：$=\mathcal{X}$，并配备由数据集在 $\mathcal{X}$ 上诱导的概率测度；（2）巴拿赫空间 E：$=W_{\mathrm{TV}}$；（3）博赫纳可积映射 f：$=\xi_D$；以及（4）可测事件 A：$=Q_\gamma \subseteq \mathcal{X}$。

根据定义， f（A） $=f$（Q_γ） $=\mathrm{LS}_\Gamma$（γ）。注意，根据假设，LS_Γ（γ）是凸的；它在 W_{TV} 上的标准度量拓扑中也是闭的，因为它是连续映射 Γ 的闭单点集 $\{\gamma\} \in$ 值域$_\Gamma$ 的原像。因此， f（A）是闭凸集，从而 co（f（A））$=f$（A）。

最后，回顾我们对 μ_γ 的定义，即给定 $x \in Q_\gamma$ 时的条件特征向量分布 $_i$由 D 诱导，积分 $\frac{1}{\mu(A)}\int_A f d\mu$ 便简化为 $\mathbb{E}_{x\sim\mu_\gamma}[Y_x]$。

因此，根据事实2，我们得出 $Y_{f^D_\Gamma,\ \gamma} = \mathbb{E}_{x\sim\mu_\gamma}[Y_x] \in \mathrm{LS}_\Gamma(\gamma)$，正如所期望的那样。 □

定理 2 和 3 共同表明，在较弱的正则性条件下，对于连续性质，校准合理性与具有凸水平集等价。为了最终将校准合理性与可诱导性和可识别性联系起来，我们现在可以援引上文讨论的 Steinwart 等人1定理 。考虑定理1所要求的关于 [2014]的额外假设（见定义 7），以及关于 $\mathcal{P}$ 的额外假设（脚注 1 中讨论的处处非局部常值假设），我们便得到最终的刻画结果：

定理 4（校准合理性 $\iff$ 可诱导性 $\iff$ 可识别性 $\iff$ CxLS）。设 Γ ： $\mathcal{P} \to \mathbb{R}$ 是定义在分布凸空间 $\mathcal{P}$ 上的连续且严格局部非恒定性质，其中 $\mathcal{P} \in \{\mathcal{P}_0, \mathcal{P}_{>0}\}$。设 $\mathcal{D}$ 是数据域 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ 上一族 $\mathcal{P}$-正则数据集分布，并且包含所有 $\mathcal{P}$-相容的双点数据集分布。

那么， Γ 在 $\mathcal{D}$ 上校准合理，当且仅当 Γ 是可诱导的（等价地说，它是可识别的，或具有凸水平集）。

证明。根据我们对 Γ、 $\mathcal{P}$和 $\mathcal{D}$的假设，定理2 和3 证明了 Γ 在 $\mathcal{P}$上校准合理，当且仅当它具有凸水平集。根据这些假设应用Steinwart 等人的定理 1 [2014] ，还可得出 Γ 在 $\mathcal{P}$ 上具有凸水平集，当且仅当 Γ 是可诱导的，且当且仅当 Γ是可识别的。这足以证明我们所需的等价关系链。 □

4 批量多重校准

本节将介绍一种适用于可引导属性的通用批量 Γ-多重校准算法 Γ。它是以往针对特定属性（如均值）设计的多重校准算法的推广，且与之非常相似[H´ebert-Johnson 等人，2018 年、Gopalan 等人，2022a] 以及针对分位数的算法 [Jung 等人， 2023]。这些算法在具体做法上有所不同（例如它们界定的校准度量、是否对预测进行离散化等）；我们最接近于Jung 等人 [2023]提出的分位数多重校准算法，该算法使用离散化预测器界定一种 ℓ_2 校准误差。

提醒一下，以下我们将聚焦于经过 Γ 重新缩放、使其值域为$_\Gamma = [0,1]$的有界性质。我们的算法将输出值域有限的多重校准 Γ-预测器 f。对于任意整数 $m \geq 1$，我们记作[2]

[2]我们本可以 将 $[1/m]$ 定义为 $[0,1]$ 中任意其他一组 m 点，只要任意两个相邻 $\frac{1}{m}$ 之间的间隔至多为 e 点。

$[1/m]:=\{\frac{1}{m+1},\frac{2}{m+1},\ldots,\frac{m}{m+1}\}$。给定任意 $m\geq 1$（这实际上是我们唯一的超参数），我们的算法可以生成一个 $O(\frac{1}{m})$-近似多重校准的 Γ-预测器 f，其值域为 $_f=[1/m]$。

根据定理 1，任意连续可诱导性质 Γ 都有一个严格一致评分函数和一个识别函数。现在我们还需要作出一个额外假设：

假设 1。假设 $\Gamma:\ \mathcal{P}\to$ 值域$_\Gamma$ 有一个识别函数 V，使得 $V(\cdot,Y)$对每个标签分布 $Y\in\mathcal{P}$都严格递增且为 L-利普希茨：$|V(\gamma,Y)-V(\gamma',Y)|\leq L|\gamma-\gamma'|$ 对所有γ、γ'均成立。

这个假设可以说并不苛刻。令 S 为 V的一个原函数，则 $V(\gamma,y)=\ ^{\partial S_{(\gamma,y)}}{}_{\partial\gamma}$。那么，$V$严格递增等价于 S 是一个凸的严格一致评分函数，适用于 Γ。在优化背景下，这种凸性假设相当自然；此外，Finocchiaro 和 Frongillo [2018] 证明，对于有限标签空间上的可诱导性质 $|\mathcal{Y}|<\infty$，这一假设不失一般性。额外的利普希茨假设则使我们能够量化算法的收敛速率。

为了表述算法 1，我们方便地将定义 10 中关于 Γ-多重校准的定义重新参数化为关于恒等函数 V 对于 Γ的定义。（根据 V的性质，随着这一更新后的校准误差概念趋近于 0，定义 10中的校准误差也将趋近于 0。）下文中，令 $Y_{(G,\gamma)}$ 表示在事件 $\{x\in\mathcal{X}$条件下的标签分布：$f(x)=\gamma,\ x\in G\}$。

定义 15（近似（$\mathcal{G}$，V）-多重校准）。固定群组 $\mathcal{G}$、分布 D，以及针对性质 Γ的恒等函数 V。有限值域 Γ-预测器 $f:\ \mathcal{X}\to[0,1]$，若对于所有 $G\in\mathcal{G}$均满足以下条件，则称其为 α-近似（$\mathcal{G}$，V）-多重校准：

$$\sum_{\gamma\in Range_f}\Pr_{x\sim X}[f(x)=\gamma|x\in G]\cdot(V(\gamma,Y_{(G,\gamma)}))^2\leq\frac{\alpha}{\Pr_{x\in X}[x\in G]}.$$

算法 1 十分自然。在可能的情况下，它会寻找群组 Q_t 与当前预测器 f的某个水平集 $G\in\mathcal{G}$ 的交集 f，使得 Q_t在 Q_t 上的预测值与真实情况相差过大，这可通过 f在 Q_t 上的期望识别函数值的幅度来衡量；然后通过将 $\gamma\in[1/m]$的值移至最佳网格点来解决这一问题。

算法 1：BatchMulticalibration($\Gamma,\mathcal{G},m,f,L$)

初始化 $t=1$ 和 $f_1=f$。令 $\alpha=\frac{4L^2}{m}$，并令 V：值域$_\Gamma\times\mathcal{Y}\to\mathbb{R}$ 是一个满足假设 -利普希茨恒等函数，适用于 Γ，且满足假设 1。当 f_t 尚未达到 α-近似（$\mathcal{G}$，V）多重校准时，执行：令 $Q_t=\{x:\ f_t(x)=\gamma,\ x\in G\}$，其中

$$(\gamma,G)\in\operatorname*{argmax}_{(\gamma',G')\in[\frac{1}{m}]\times\mathcal{G}}\Pr_{x\in\mathcal{X}}[f_t(x)=\gamma',x\in G'](V(\gamma',Y_{(\gamma',G')}))^2.$$

令：$\gamma'=\operatorname{argmin}_{\gamma''\in[\frac{1}{m}]}|V(\gamma'',Y_{Q_t})|$更新：$f_{t+1}(x):=\ _{\mathbb{I}}[x\notin Q_t]\cdot f_t(x)+\ _{\mathbb{I}}[x\in Q_t]\cdot\gamma'$ 对所有 $x\in\mathcal{X}$，且 $t\leftarrow t+1$。输出 f_t。

定理 5（算法 1 的保证）。固定数据分布 $D\in\Delta\mathcal{Z}$ 和群组 $\mathcal{G}\subseteq 2^{\mathcal{X}}$。固定一个性质 Γ，其评分函数为 S，恒等函数为 $V=\frac{\partial S}{\partial\gamma}$，并满足假设 1，使得$V(\cdot,Y_Q)$ 在由 D 导出的所有标签分布 Y_Q（对于 $Q\subseteq\mathcal{X}$）上都是 L-利普希茨的。令离散化为$m\geq 1$。若算法 1 初始化时使用预测器 $f_1:\ \mathcal{X}\to\mathbb{R}$，其评分为 $\mathbb{E}_{(x,y)\sim D}[S(f_1(x),y)]=C_{init}$，且 $C_{opt}=\mathbb{E}_{(x,y)\sim D}[S(f_\Gamma^D(x),y)]$是真实分布预测器 f_Γ^D的评分，则算法 1 至多经过 $(C_{init}-C_{opt})\frac{m^2}{L}$次更新后，会生成一个 $\frac{4L^2}{m}$-近似的（$\mathcal{G},V$）多重校准 Γ预测器 f。

定理的证明5 见附录A，其分析与此前针对各种属性的多重校准算法分析类似 [H´ebert-Johnson 等人，2018 年，Jung 等人，2023 年，Deng 等人，2023]。我们将期望评分 $\mathbb{E}_D[S]$（其中$V=\frac{\partial S}{\partial\gamma}$）用作算法的势函数：它在算法的每一步都会下降。此前关于均值多重校准的分析 [H´ebert-Johnson 等人，

2018] 和分位数多重校准 [Jung 等人， 2023] 分别使用平方损失和针球损失作为势函数——它们分别是针对均值和分位数的严格适当评分规则。我们的分析是将这一方法自然推广到任意可诱导属性。通过证明 $\mathbb{E}[S]$ 在每次迭代时都会大幅下降，即可得到收敛速率。

请注意，我们已将算法1 描述为能够直接在真实数据分布 V 上查询预期识别函数 D。实际上，我们会在独立同分布 样本 $D \sim D^n$ 的 n 经验分布上运行该算法，这些样本点来自 D。附录 B 为这种情况提供了有限样本保证。

5 联合多重校准

现在，我们开始着手理解两个相互关联的问题：（1）如何对向量值性质进行多重校准；（2）如何将多重校准的概念恰当地扩展到一些实践中重要的标量性质，这些性质具有非凸水平集，因此根据我们的定理 2 既不可诱导，也不适用于校准合理性（例如方差）。这两个问题密切相关，因为显然，如果我们有一个向量值性质，且其每个坐标都可以单独诱导，那么我们只需使用第 4 节中的算法，分别对该性质的每个坐标进行多重校准；向量值性质之所以具有挑战性，恰恰在于其坐标无法分别诱导。

具体而言，我们研究二维性质 $\Gamma = (\Gamma^0, \Gamma^1)$ 这一重要情形，其中 Γ^0 是可诱导的，而 Γ^1 本身不可诱导，但在给定 Γ^0任意固定值的条件下可诱导，并给出一种能够为这类性质对生成联合多重校准估计量的算法。下面列出我们对性质所需的假设，并定义联合多重校准。

定义 16（条件可诱导性）。若性质 $\Gamma^1 : \mathcal{P} \to \mathbb{R}$ 在性质 $\Gamma^0 : \mathcal{P} \to \mathbb{R}$的条件下可诱导，我们称之为：对每个 $\gamma^0 \in \text{Range}_{\Gamma^0}$， Γ^1 在 Γ^0: $\mathcal{P}_{\gamma^0} = \{P \in \mathcal{P} : \Gamma^0(P) = \gamma^0\}$ 的每个水平集上均可诱导。

对于可诱导的分量 Γ^0，我们将其评分函数和识别函数分别记为 S^0、V^0。对于在 Γ^0条件下可诱导的性质 Γ^1 ，对每个 $\gamma^0 \in \text{Range}_{\Gamma^0}$ ，我们用 $V^1_{\gamma^0} : \text{Range}_{\Gamma^1} \times \mathcal{Y} \to \mathbb{R}$ 表示一个函数，该函数能在满足 $\Gamma^0(P) = \gamma^0$的每个分布 P 上识别 Γ^1 ；用 $S^1_{\gamma^0} : \text{Range}_{\Gamma^1} \times \mathcal{Y} \to \mathbb{R}$ 表示一个评分函数，该函数在满足$\Gamma^0(P) = \gamma^0$的每个分布 P 上对 Γ^1 严格一致。

假设 与第 4 节中一样，我们假设可诱导组成部分 Γ^0 满足假设 1，其中 $V^0(\gamma^0, \cdot)$ 严格递增，且关于 L^0 是 γ^0- 利普希茨的。这里，我们还需要相反的（同样温和的）假设：

假设 2。 $V^0(\cdot, Y)$ 对所有 $Y \in \mathcal{P}$ 而言，在 $\Gamma^0(Y)$ 附近是 L^0_a-反利普希茨的：对所有 γ^0， $|\gamma^0 - \Gamma^0(Y)| \le L^0_a |V^0(\gamma^0, Y)|$。

Γ^1 的情形更为复杂：对于 Γ^0 的不同水平集，它具有不同的识别函数 $V^1_{\gamma^0}$ ，而不是对所有 $P \in \mathcal{P}$ 使用单一函数。一般而言，没有任何因素能阻止这些函数V_{γ^0} 在 γ^0 的不同取值下彼此完全无关（甚至在彼此的水平集上未定义）。然而，对于大多数感兴趣的性质，我们可以预期 $V^1_{\gamma^0}(\gamma^1, P)$ 会随参数 γ^0 连续变化，并且即使对于 $P \notin \{P' \text{ 也有良好定义：} \Gamma^0(P') = \gamma^0\}$。为反映这一点，并使我们的条件多重校准算法的保证成立，我们作出以下（稍强但仍温和的）假设：

假设 3。假设对于所有 $P \in \mathcal{P}$， $V^1_{\gamma^0}(\gamma^1, P)$ 均有定义。此外，假设 $V^1_{\gamma^0}$ 作为 γ^0 的函数是 L_c-利普希茨的：也就是说，对于任意固定的 γ^1 和 $P \in \mathcal{P}$，以及任意 γ^0_1 和 γ^1_1，都有 $|V^1_{\gamma^0_0}(\gamma^1, P) - V^1_{\gamma^0_1}(\gamma^1, P)| \le L_c |\gamma^0_0 - \gamma^0_1|$。

此外，我们假设条件识别函数 $V^1_{\gamma^0}$ 对于 Γ^1 在 Γ^0的水平集上 $\{\Gamma^0 = \gamma^0\}$ 保持其“形状”，即对于来自 Γ^0其他水平集的分布，仍然严格递增且满足利普希茨条件。虽然这是一个非平凡的假设，但在第 7 节中，我们验证了当（Γ^0， Γ^1）是一个 贝叶斯对时，该假设成立；贝叶斯对是 Embrechts 等人 [提出的一类重要且一般性的性质 ， 2021], 也是联合多重校准理论的一个主要应用场景。

假设4. 对于所有[3] $\gamma^0 \in$ 值域Γ^0，假设 $V^1_{\gamma^0}(\cdot,\ P)$ 对于所有 L^1 均为 -利普希茨且严格递增 $P \in \mathcal{P}$。

现在我们可以定义本节的核心概念：针对二维性质的联合多重校准预测器 $\Gamma = (\Gamma^0,\ \Gamma^1)$。与标准多重校准的定义10、15 类似，我们以两种形式定义这一概念：第一种由 $(\Gamma^0,\ \Gamma^1)$ 参数化，第二种则由识别函数 $(V^0,\ V^1)$ 参数化。与第 4 节一样，后一种定义有助于简化算法分析中的记号（并且，当这种多重校准误差趋于 0 时，由 $(\Gamma^0,\ \Gamma^1)$ 参数化的误差也会趋于 0）。

在下面的定义中，我们使用以下简写（对于 $i = 0,\ 1$）：

$$\mu_f(\gamma^i|G,\gamma^{1-i}) := \Pr_x[f^i(x) = \gamma^i | x \in G, f^{1-i}(x) = \gamma^{1-i}],$$

and

$$\mu_f(G,\gamma^i) := \Pr_x[x \in G, f^i(x) = \gamma^i].$$

另外，令$Y_{(G,\gamma^0,\gamma^1)}$表示在事件 $\{x \in \mathcal{X}$ 条件下的标签分布：$f(x) = (\gamma^0,\gamma^1)\ x \in G\}$。

定义17（近似联合多重校准）。固定分布 $D \in \Delta\mathcal{Z}$ 和群组族 $\mathcal{G}$。给定性质 $\Gamma = (\Gamma^0,\Gamma^1)$，若有限值域预测器 $f = (f^0, f^1) : \mathcal{X} \to \mathbb{R}^2$ 满足对每个 $G \in \mathcal{G}$，均有：(1) 对所有 $\gamma^1 \in$ 值域f^1，

$$\sum_{\gamma^0 \in Range_{f^0}} \mu_f(\gamma^0|G,\gamma^1) \cdot (\gamma^0 - \Gamma^0(Y_{(G,\gamma^0,\gamma^1)}))^2 \le \frac{\alpha^0}{\mu_f(G,\gamma^1)},$$

并且 (2) 对所有 $\gamma^0 \in$ 值域f^0，类似地有：

$$\sum_{\gamma^1 \in Range_{f^1}} \mu_f(\gamma^1|G,\gamma^0) \cdot (\gamma^1 - \Gamma^1(Y_{(G,\gamma^0,\gamma^1)}))^2 \le \frac{\alpha^1}{\mu_f(G,\gamma^0)}.$$

类似地，给定识别函数 $V^0, \{V^1_{\gamma^0}\}_{\gamma^0 \in Range_{\Gamma^0}}$，如果对于 $G \in \mathcal{G}$, $\gamma^1 \in$ 值域f^1：预测器 $f = (f^0, f^1)$ 是 (α^0, α^1)- 近似 $(\mathcal{G}, V^0, V^1)$- 联合多重校准的，

$$\sum_{\gamma^0 \in Range_{f^0}} \mu_f(\gamma^0|G,\gamma^1) \cdot \left(V^0(\gamma^0, Y_{(G,\gamma^0,\gamma^1)})\right)^2 \le \frac{\alpha^0}{\mu_f(G,\gamma^1)}, \tag{1}$$

并且对于所有 $G \in \mathcal{G}$ 以及 $\gamma^0 \in$ 值域f^0：

$$\sum_{\gamma^1 \in Range_{f^1}} \mu_f(\gamma^1|G,\gamma^0) \cdot (V^1_{\Gamma^0(Y_{(G,\gamma^0,\gamma^1)})}(\gamma^1, Y_{(G,\gamma^0,\gamma^1)}))^2 \le \frac{\alpha^1}{\mu_f(G,\gamma^0)}. \tag{2}$$

算法概述 现在我们介绍 联合多重校准(算法 3)，这是一种用于学习联合多重校准预测器的标准算法 $f = (f^0,\ f^1)$（针对 $\Gamma^0,\ \Gamma^1$）。为了处理二维性质，我们采用一种两阶段结构：交替在 f^0的当前水平集上 f^1进行多重校准，并在 f^1 的当前水平集上进行 f^0，直至根据上述定义 17 中的公式 1 和 2 达到所需的联合多重校准误差水平。

与第 4 节一样，我们的预测器经过离散化：值域f^0 =值域$_{f^1} = [\frac{1}{m}]$。对 f^0和 f^1 的更新，均通过调用子程序 `BatchMulticalibration`^{V}来执行，该子程序与`BatchMulticalibration`（算法 1）非常相似，但有两点不同：（1）为简化记号，它直接接受识别函数 V，而不是性质 Γ；（2）为满足联合多重校准的额外要求，它采用了更严格的停止条件（对于第 4 节所定义的批量多重校准而言，该条件是充分但非必要的）。我们在算法 `BatchMulticalibration`^{V} 2中给出其伪代码，并将其（非常相似的）分析放到附录 C 的引理 3中。

[3]事实上，对于下面的算法 3，它生成的预测器是在 $[1/m] \times [1/m]$上离散化的，我们只需要该条件对 $\gamma^0 \in [1/m]$成立即可，这一要求较为宽松：$[1/m]$ 是一个有限集合，因此无需在整个值域上统一满足利普希茨性Γ^0。

在算法 3的整个执行过程中，该子程序会在辅助群族上被调用；这些辅助群族由 $\mathcal{G}$ 中的群组与 f^0 或 f^1的水平集的两两交集构成。由于预测器会被更新，在这些调用过程中，辅助群组始终处于漂移状态，因此需要仔细记账，以验证这不会阻碍整体收敛。为此，我们证明的一个关键事实是：在算法 3的整个过程中，针对 V^0 的所有调用中，$\texttt{BatchMulticalibration}^V$ 子程序对 f^0执行的更新次数有界，这也意味着 f^1 最多会被重新校准相同次数。

算法 3 大幅推广了 Jung 等人的（均值、矩）多重校准算法[2021], ，因此在分析中存在一些关键差异。值得注意的是，按照我们的术语，在他们的特定情形下，给定 f^0 时对 f^1 进行重新校准，可以表述为一个单独的均值多重校准子程序，并采用他们所称的“伪标签”技术。在我们这种一般性下，这种做法不再奏效，因为我们不得不在每个水平集 $\{f^0=\gamma^0\}$ 上，针对 Γ^1 使用不同的恒等函数 $V^1_{\gamma^0}$。这就是为什么我们的内层 for 循环要遍历 f^0 的水平集，并通过分别调用 m 次子程序，使用 f^1 进行重新校准（幸运的是，这些调用实际上可以并行执行，因为 f^0 的水平集彼此不相交）。即便有了这一构造，我们也无法轻易将第 4 节中的势函数论证移植过来：每个水平集 $\{f^0=\gamma^0\}$ 都可能与 Γ^0 的多个水平集重叠，因此真实性质 Γ^1 通常不会在 $\{f^0=\gamma^0\}$ 上对应一个可用作势函数的单一评分函数。此时，我们关于 $V^1_{\gamma^0}$ 相对于 γ^0 的行为所作的假设，就起到了关键作用：这些假设使我们能够证明，只要 f^0 已得到充分的多重校准，在水平集 $\{f^0=\gamma^0\}$ 上使用代理标识函数 $V^1_{\gamma^0}$，就不会导致针对 Γ^1 的多重校准子程序无法收敛。

我们首先正式定义子程序 $\texttt{BatchMulticalibration}^V$，然后在算法 联合多重校准中给出完整的算法 2。

算法 2： $\text{BatchMulticalibration}^V(V,\mathcal{G},m,f,\alpha)$

初始化 $t=1$ 和 $f_1=f$。

当 $\exists(\gamma,G)\in[1/m]\times\mathcal{G}$ 使得 $\Pr_{x\in\mathcal{X}}[f_t(x)=\gamma,\ x\in G]\left(V(\gamma,Y_{(\gamma,G)})\right)^2\geq\alpha/m$ 时，执行

令 $Q_t=\{x:\ f_t(x)=\gamma,\ x\in G\}$

Let:

$$\gamma'=\operatorname*{argmin}_{\gamma''\in[1/m]}|V(\gamma'',Y_{Q_t})|$$

更新：对所有 $x\in\mathcal{X}$及 $t\leftarrow t+1$，$f_{t+1}(x):=\mathbb{I}[x\notin Q_t]\cdot f_t(x)+\mathbb{I}[x\in Q_t]\cdot\gamma'$。输出 f_t。

算法 3： 联合多重校准$((\Gamma^0,\Gamma^1)\ \mathcal{G},m,(f^0,f^1))$

令 V^0 和 $\{V^1_{\gamma^0}\}_{\gamma^0\in[1/m]}$ 分别为 Γ^0 和 Γ^1 的恒等函数，满足假设 1、2、3、4。

初始化 $t=1$和 $f_1=(f^0,f^1)$。

当 $\exists(\gamma^0,\gamma^1,G)\in[\frac{1}{m}]\times[\frac{1}{m}]\times\mathcal{G}$时 $\Pr_{x\in\mathcal{X}}[f_t(x)=(\gamma^0,\gamma^1)\ x\in G]\left(V^0(\gamma^0,Y_{(\gamma^0,\gamma^1,G)})\right)^2\geq\frac{\alpha}{m}$ **do**

令 $\mathcal{G}^0_t\leftarrow\{G\cap\{x\in\mathcal{X}:\ f^1_t(x)=\gamma^1\}:\ G\in\mathcal{G},\gamma^1\in[\frac{1}{m}]\}$

更新 $f^0_{t+1}\leftarrow\text{BatchMulticalibration}^V(V^0,\mathcal{G}^0_t,m,f^0_t,\alpha^0)$

对于 $\gamma^0\in[1/m]$ 执行

令 $\mathcal{G}^{1,\gamma^0}_t\leftarrow\{G\cap\{x\in\mathcal{X}:\ f^0_{t+1}(x)=\gamma^0\}:\ G\in\mathcal{G}\}$

令 $f^{1,\gamma^0}_{t+1}\leftarrow\text{BatchMulticalibration}^V(V^1_{\gamma^0},\mathcal{G}^{1,\gamma^0}_t,m,f^1_t,\alpha^1)$

更新 $f^1_{t+1}(x)\leftarrow\sum_{\gamma^0\in[1/m]}\mathbb{I}[f^0_{t+1}(x)=\gamma^0]\cdot f^{1,\gamma^0}_{t+1}(x)\ \forall x\in\mathcal{X}$

更新 $t\leftarrow t+1$。

输出 $f_t=(f^0_t,f^1_t)$。

下面的定理给出了算法 3 的收敛保证。完整的证明严格阐述了上文讨论的思想，见附录 C。

定理 6（算法 3 的保证）。设 $\alpha^0=\alpha^0(L^0,\cdot)$ 和 $\alpha^1=\alpha^1(L^1,\cdot)$。令 $\alpha^1_*=\alpha^1_*((L^0L^0_*L_\circ)^2+(L^1)^2,m)$。对于任意 $\mathcal{G}\subseteq2^{\mathcal{X}}$, $m\geq1$, $\textit{JointMulticalibration}^m$，算法 3 至多通过对 f进行 $\frac{B^0B^1m^4}{L^0L^1}$次更新，输出一个关于性质 (Γ^0,Γ^1) 的 (α^0,α^1_*)-近似 $(\mathcal{G},V^0,V^1)$-联合多重校准预测器 $f=(f^0,f^1)$。这里，$B^0:=\sup_{\gamma,y\in[0,1]}S^0(\gamma,y)-\inf_{\gamma,y\in[0,1]}S^0(\gamma,y)$，其中 S^0是 V^0的一个原函数，并且

$B^1 := \max_{\gamma^0 \in [1/m]} \left(\sup_{\gamma, y \in [0,1]} S^1_{\gamma^0}(\gamma, y) - \inf_{\gamma, y \in [0,1]} S^1_{\gamma^0}(\gamma, y) \right)$ 对每个 $S^1_{\gamma^0}$ 均为 $V^1_{\gamma^0}$的一个原函数。

6 序贯多重校准

现在，我们转向序贯对抗设定。在这种设定中，不存在底层分布，我们的目标是在学习者与对手交互所产生的交互记录 π 所定义的经验分布上，实现近似的（$\mathcal{G}$, V）多重校准（定义 15）。这将 Gupta 等人 [2022] 研究的均值和分位数的序贯多重校准推广到了任意可诱导属性。事实上，即使对于分位数，我们也强化了Gupta 等人 [2022] 的结果——他们给出了一种 ℓ_∞ 校准变体，其条件事件采用了“分桶”；而我们针对用于批量校准的同一 ℓ_2校准概念给出了一个界，无需任何分桶，且保证严格更强。Garg 等人 [2023] 也为序贯均值多重校准得到了这一更强的保证。

6.1 设定与预备知识

6.1.1 序贯学习设定

在序贯设定中，学习者与对手按轮次 $t = 1$ 到 T 进行交互，过程如下：

1. 对手选择一个特征向量 $x_t \in \mathcal{X}$ 和一个分布 $Y_t \in \Delta Y$（可能受某些限制），并向学习者揭示 x_t。
2. 学习者作出预测 $p_t \in \mathbb{R}$。3. 对手抽取 $y_t \sim Y_t$，并向学习者揭示 y_t。

交互记录累积形成交互记录 $\pi = \{(x_t, p_t, y_t)\}_{t=1}^T$。对于任意 $s \le T$ 和交互记录 π，交互记录 $\pi^{<s}$ 的前缀定义为 $\pi^{<s} = \{(x_t, p_t, y_t)\}_{t=1}^{s-1}$。我们用 $\Pi^{<s}$ 表示长度为 $< s$的所有交互记录的定义域。学习者由一组映射组成（每个轮次 $t \le T$对应一个映射）$\mathcal{L}_t$：$\Pi^{<t} \times \mathcal{X} \to \Delta\mathbb{R}$；对手则由一组映射 $\mathcal{A}_t$：$\Pi^{<t} \to \mathcal{X} \times \Delta Y$组成，这些映射规定了双方根据迄今为止的观察所采取的行为。

现在，我们可以给出在线多重校准的强定义 ℓ_2，并随后说明如何实现它。

定义 18（在线多重校准）。固定交互记录 $\pi = \{(x_t, p_t, y_t)\}_{t=1}^T$。令 $n(\pi, G) = |\{t : x_t \in G\}|$ 表示在 π中包含群组 G 成员的轮次数，令 $n(\pi, \gamma, G) = |\{t : x_t \in G, p_t = \gamma\}|$ 表示包含群组 G 且预测 p_t 为 γ的轮次数。

固定 π、群组集合 $\mathcal{G}$、性质 Γ以及针对 Γ的识别函数 V。我们称交互记录 π 是 α-近似 $(\mathcal{G}, V)$-多重校准的，当且仅当对所有 $G \in \mathcal{G}$：

$$\sum_{\gamma} \frac{n(\pi, \gamma, G)}{n(\pi, G)} \left(\sum_{t : p_t = \gamma, x_t \in G} \frac{V(\gamma, y_t)}{n(\pi, \gamma, G)} \right)^2 \le \alpha \frac{T}{n(\pi, G)}.$$

备注 1。 请注意，这恰好是我们在定义*15*中给出的近似多重校准定义，其中 π 上的经验分布取代了分布 $\mathcal{D}$。

通过消去一些项，我们可以在一定程度上简化多重校准的概念：

观察 1. 固定一个交互记录 π、一个群组集合 $\mathcal{G}$、一个性质 Γ，以及一个用于 Γ 的识别函数 V。对于每个群组 $G \in \mathcal{G}$，定义如下量：

$$K_2(G, \pi) = \sum_{\gamma} \frac{1}{n(\pi, \gamma, G)} \left(\sum_{t : p_t = \gamma, x_t \in G} V(\gamma, y_t) \right)^2.$$

那么，对于所有 $G \in \mathcal{G}$、K_2，若满足（G, π）$\le \alpha T$，则 π 是 α-近似（$\mathcal{G}$, V）多重校准的。

在线设置中，对于每个 $G \in \mathcal{G}$，我们的目标是在生成交互记录的过程中控制 $K_2(G, \pi)$ 的增长。以下引理将起到关键作用：

引理 1. 固定一个部分转录 $\pi^{<s} = \{(x_t, p_t, y_t)\}_{t=1}^{s-1}$ 以及一个单轮延续 (x_s, p_s, y_s)。将延长一轮后的交互记录记为 $\pi^{\leq s} = \pi^{<s} \circ (x_s, p_s, y_s)$。定义：

$$R(\pi^{<s}, G, \gamma) = \sum_{t<s: p_t=\gamma, x_t \in G} V(\gamma, y_t).$$

那么对于每个 $G \in \mathcal{G}$，若 $x_s \notin G$，则有：

$$K_2(G, \pi^{\leq s}) - K_2(G, \pi^{<s}) = 0.$$

若 $x_s \in G$ 且 $p_s = \gamma$，则有：

$$K_2(G, \pi^{\leq s}) - K_2(G, \pi^{<s}) \leq \frac{1}{n(\pi^{<s}, \gamma, G)} \left(2V(\gamma, y_s) R(\pi^{<s}, G, \gamma) + V(\gamma, y_s)^2\right).$$

证明。若 $x_s \notin G$，则根据定义，$K_2(G, \pi^{\leq s}) = K_2(G, \pi^{<s})$，因此结论成立。否则，若 $x_s \in G$，我们可以计算：

$$\begin{aligned}
& K_2(G, \pi^{\leq s}) - K_2(G, \pi^{<s}) \\
= & \frac{1}{n(\pi^{<s}, \gamma, G) + 1} \left(\left(\sum_{t<s: p_t=\gamma, x_t \in G} V(\gamma, y_t) \right) + V(\gamma, y_s) \right)^2 - \frac{1}{n(\pi^{<s}, \gamma, G)} \left(\sum_{t<s: p_t=\gamma, x_t \in G} V(\gamma, y_t) \right)^2 \\
\leq & \frac{1}{n(\pi^{<s}, \gamma, G)} \left(\left(\sum_{t<s: p_t=\gamma, x_t \in G} V(\gamma, y_t) \right) + V(\gamma, y_s) \right)^2 - \frac{1}{n(\pi^{<s}, \gamma, G)} \left(\sum_{t<s: p_t=\gamma, x_t \in G} V(\gamma, y_t) \right)^2 \\
\leq & \frac{1}{n(\pi^{<s}, \gamma, G)} \left(2V(\gamma, y_s) R(\pi^{<s}, G, \gamma) + V(\gamma, y_s)^2\right).
\end{aligned}$$

证明至此完成。 □

6.1.2 一个关键工具：在线极小极大多目标优化框架

我们将通过 Lee 等人 [2022] 引入的在线极小极大多目标优化框架推导我们的算法。

定义 19（在线极小极大多目标优化设定）。学习者在轮次 $t \in [T] := \{1, \ldots, T\}$ 中与对手进行博弈。在这些轮次中，学习者累积一个 d 维损失向量（$d \geq 1$），其中每轮的损失向量都属于 $[-C, C]^d$，对于某个 $C > 0$。在每轮 t，学习者与对手按如下方式互动：

1. 在 轮次 t 之前，对手选择并向学习者揭示一个环境，其中包含 ：

   （a）学习者和对手各自的凸紧行动集 $\mathcal{X}^t$、$\mathcal{Y}^t$，嵌入于有限维欧几里得空间中；（b）连续向量值损失函数 $\ell^t(\cdot, \cdot)$：$\mathcal{X}^t \times \mathcal{Y}^t \to [-C, C]^d$，其中每个 $\ell_j^t(\cdot, \cdot)$：$\mathcal{X}^t \times \mathcal{Y}^t \to [-C, C]$（对于 $j \in [d]$）关于第一个参数是凸的，关于第二个参数是凹的。

2. 学习者选择某个 $x^t \in \mathcal{X}^t$。

3. 对手观察学习者的选择 x^t，并作出回应，选择某个 $y^t \in \mathcal{Y}^t$.

4. 学习者承受（并观察到）损失向量 $\ell^t(x^t, y^t)$。

学习者的目标是最小化经过 T 轮次后累积损失向量的最大维度值，换言之，最小化：$\max_{j \in [d]} \sum_{t \in [T]} \ell_j^t(x^t, y^t)$。

分析学习者在在线极小极大多目标优化情境下的表现时，一个关键量是每轮互动中阶段博弈的先手对手（AMF）值 t ——也就是说，如果学习者（假设性地）预先知道对手的行动，她能做得多好。

定义 20（第 t 轮的先手对手（AMF）值）。由环境 $(\mathcal{X}^t, \mathcal{Y}^t, \ell^t)$ 定义的博弈在第 t 轮的先手对手（AMF）值为：

$$w_A^t := \sup_{y^t \in \mathcal{Y}^t} \min_{x^t \in \mathcal{X}^t} \left(\max_{j \in [d]} \ell_j^t(x^t, y^t) \right).$$

我们可以将学习者的表现与基准进行比较来衡量；该基准由每轮定义的博弈中对手先行动时的取值确定。

定义 21（先手对手（AMF）遗憾）。对于交互记录 $\pi^t = \{(\mathcal{X}^s, \mathcal{Y}^s, \ell^s)\ x^s, y^s\}_{s=1}^t$，我们将学习者在时间 t 、针对 j^{th} 维度的先手对手（AMF）遗憾定义为：

$$R_j^t(\pi^t) := \sum_{s=1}^t \ell_j^s(x^s, y^s) - \sum_{s=1}^t w_A^s.$$

整体 AMF 遗憾定义如下： $R^t(\pi^t) = \max_{j \in [d]} R_j^t$.

Lee 等人 [2022] 证明，在任何在线极小极大多目标优化设定中，算法 4都能实现递减的 AMF 遗憾。

算法 4：实现次线性 AMF 遗憾的学习者通用算法

对于轮次 $t = 1, \ldots, T$ **do**

学习对手选择的 $\mathcal{X}^t$、 $\mathcal{Y}^t$ 和损失函数 $\ell^t(\cdot, \cdot)$。

Let $$\chi_j^t := \frac{\exp\left(\eta \sum_{s=1}^{t-1} \ell_j^s(x^s, y^s)\right)}{\sum_{i \in [d]} \exp\left(\eta \sum_{s=1}^{t-1} \ell_i^s(x^s, y^s)\right)} \text{ for } j \in [d].$$

Play $$x^t \in \operatorname*{argmin}_{x \in \mathcal{X}^t} \max_{y \in \mathcal{Y}^t} \sum_{j \in [d]} \chi_j^t \cdot \ell_j^t(x, y).$$

观察对手对 $y^t \in \mathcal{Y}^t$ 的选择。

定理 7（算法 4 [Lee 等人、 2022] 的 AMF 遗憾保证）。对于任意 $T \geq \ln d$，学习率为 $\eta = \sqrt{\ln \frac{d}{4TC^2}}$ 的算法4，面对任意对手时，AMF 遗憾有如下界：$R^T \leq 4C\sqrt{T \ln d}$.

6.2 典型的序贯多重校准算法

在本节余下部分中，我们将说明，对于任何具有利普希茨识别函数 Γ 的可诱导性质 V，以及任何有限群组结构 $\mathcal{G}$，在序贯对抗设定中实现递减的（$\mathcal{G}$， V）多重校准误差的问题可以转化为在线极小极大多目标优化的一个实例，因此可以通过适当实例化算法 4 来求解；多重校准误差的界则可由适当实例化定理 7 得出。

本节不需要假设 1 的全部条件——我们只需要识别函数 V 是利普希茨的。我们回顾一下，识别函数的利普希茨条件取决于标签分布 Y，因此，我们所需的假设不仅针对性质，也针对对手在每一轮选择的分布。利普希茨常数可以随轮次变化；我们的假设仅针对其平均值。

假设 5。固定一个可诱导性质 Γ 及其识别函数 V。假设在每一轮 t，对手选择一个标签分布 Y_t ，使得 V 是 L_t-利普希茨的：

$$|V(\gamma, Y_t) - V(\gamma', Y_t)| \leq L_t |\gamma - \gamma'| \quad \textit{for all } \gamma, \gamma'.$$

我们不作任何关于个体 L_t 的假设，但假设其平均值有界于 L ：

$$\frac{1}{T} \sum_{t=1}^T L_t \leq L.$$

现在我们介绍规范算法，并证明在假设 5下的保证。

算法 5：在线多重校准$(\mathcal{G}, V, m)$

初始化一个空的交互记录 $\pi^{\leq 0}$.对于轮次 $t = 1, \ldots, T$，执行以下操作：观察对手选择的特征向量 x_t。定义损失函数 $\ell^t : [1/m] \times \mathcal{G} \to \mathbb{R}^{m \times |\mathcal{G}|}$，使得对于每个 $G \in \mathcal{G}$ 和 $\gamma \in [1/m]$:

$$\ell^t_{G,\gamma}(\gamma_t, y_t) = \mathbb{I}[x_t \in G, \gamma_t = \gamma] \cdot \frac{1}{n(\pi^{<t}, \gamma, G)} \left(2V(\gamma, y_t)R(\pi^{<t}, G, \gamma) + V(\gamma, y_t)^2\right),$$

其中:

$$R(\pi^{<t}, G, \gamma) = \sum_{s<t : p_s = \gamma, x_s \in G} V(\gamma, y_s).$$

Let $$\chi^t_{G,\gamma} := \frac{\exp\left(\eta \sum_{s=1}^{t-1} \ell^s_{G,\gamma}(p^s, y^s)\right)}{\sum_{(G',m') \in \mathcal{G} \times [1/m]} \exp\left(\eta \sum_{s=1}^{t-1} \ell^s_{G',m'}(p^s, y^s)\right)} \text{ for } (G, \gamma) \in \mathcal{G} \times [1/m].$$

Let $$P^t \in \operatorname*{argmin}_{P \in \Delta[1/m]} \max_y \sum_{(G,\gamma) \in \mathcal{G} \times [1/m]} \mathbb{E}_{p \sim P^t} \left[\chi^t_{G,\gamma} \cdot \ell^t_{G,\gamma}(p, y)\right].$$

抽取样本 $p_t \sim P^t$ 并进行预测 p_t。
观察对手对 y_t的选择。
更新交互记录 $\pi^{\leq t} = \pi^{\leq t-1} \circ (x_t, p_t, y_t)$。

定理8 （序贯多重校准的算法保证）。 固定任意有限群组集合 $\mathcal{G}$ 以及任意具有有界识别函数的可诱导性质 Γ V，满足 $|V$（γ， y）$| \leq C$。假设序贯对抗设定中的对手选择一系列分布，该分布与 V 满足假设 5 ，且利普希茨常数为 L。那么对于任意 $m > 0$ ，存在一个供学习者使用的随机化算法（算法 5），该算法在每一轮从 m 个离散预测中进行选择，并与对手共同诱导出一种交互记录分布，使其生成的交互记录满足以下 α-近似 （$\mathcal{G}$， V）-多重校准:

$$\mathbb{E}_{\pi}[\alpha] \leq \frac{2CL}{m} + \frac{2C^2 \log(T)}{T} + 12C^2\sqrt{\frac{\ln(|\mathcal{G}|m)}{T}}.$$

证明。 我们将学习问题嵌入在线极小极大优化设定中，以便应用定理 7。首先，学习者和对手的策略空间是什么？在每一轮，我们令学习者的策略空间为 $\Delta[1/m]$，即预测上的概率分布单纯形γ_t ，其离散粒度为 $1/m$。对手的策略空间是 $\mathcal{Y}$上的所有利普希茨分布的集合。这两个策略空间都是所要求的凸集。接下来，我们需要定义每一轮使用的损失函数 ℓ^t 。我们将损失函数的维度设为 $d = |\mathcal{G}|m$ —— 为每一对（$G \in \mathcal{G}$， $i \in [m]$）分配一个坐标。假设在第 t轮，对手选择了特征向量 x_t （回顾一下，在学习者必须做出预测之前，该向量会展示给她）。然后，我们将损失向量 ℓ^t 定义如下：对于每个 $G \in \mathcal{G}$， $i \in [m]$ 我们引入损失函数

$$\ell^t_{G,i}(Alg_t, Y_t) = \mathbb{E}_{\gamma_t \sim Alg_t, y_t \sim Y_t} \left[\mathbb{I}\left[x_t \in G, \gamma_t = \frac{i}{m}\right] \cdot \frac{1}{n(\pi^{<t}, \frac{i}{m}, G)} \left(2V(\frac{i}{m}, y_t)R(\pi^{<t}, G, \frac{i}{m}) + V(\frac{i}{m}, y_t)^2\right)\right],$$

其中， $Alg_t \in \Delta[1/m]$ 是学习者选择的预测分布， Y_t 是对手选择的标签分布。根据期望的线性性，该损失函数关于双方的行动均为线性，因此特别满足所要求的凸凹性。根据 V的有界性以及 R的定义，该损失函数的取值落在所要求的 $[-C', C']$ 内，其中 $C' \leq 3C^2$。N

[e]接下来，我们需要给出第 t 轮博弈中对手先行动值的上界:

$$w^A_t = \sup_{Y_t} \min_{Alg_t \in \Delta[\frac{1}{m}]} \max_{G \in \mathcal{G}, i \in [m]} \mathbb{E}_{\gamma_t \sim Alg_t, y_t \sim Y_t} \left[\mathbb{I}\left[x_t \in G, \gamma_t = \frac{i}{m}\right] \cdot \frac{1}{n(\pi^{<t}, \frac{i}{m}, G)} \left(2V(\frac{i}{m}, y_t)R(\pi^{<t}, G, \frac{i}{m}) + V(\frac{i}{m}, y_t)^2\right)\right].$$

为了界定 w_t^A，考虑一下：如果对手先承诺并揭示真实标签分布 Y_t，学习者应该怎么做。学习者可以计算真实属性值 $\gamma_t^* = \Gamma(Y_t)$。如果她可以选择 $\gamma_t = \gamma_t^*$，就能确保 $V(\gamma_t, Y_t) = 0$，这意味着

$$w_t^A = \frac{(V(\gamma_t^*, y_t))^2}{n(\pi^{<t}, \gamma_t^*, G)} \leq \frac{C^2}{n(\pi^{<t}, \gamma_t^*, G)}.$$

学习者通常不能采取 γ_t^*（因为它可能不是 $1/m$ 的倍数，因此不在其策略空间中），但可以选择最接近 γ_t^* 的离散点 $\gamma_t \in [1/m]$ ——尤其是，该点会满足 $|\gamma_t^* - \gamma_t| \leq \frac{1}{m}$。采取这一行动后，对于满足 $x_t \notin G$ 的群组 G 所对应的所有坐标，以及满足 $\gamma_t \neq \frac{i}{m}$ 的预测 $\frac{i}{m}$ 所对应的坐标，学习者的损失均为 0（因为对于这些坐标中的每一个，指示函数 $\mathbb{I}[x_t \in G, \gamma_t = \frac{i}{m}] = 0$）。因此，只需考虑满足 $x_t \in G$ 和 $\gamma_t = \frac{i}{m}$ 的数对 (G, i) 对应的坐标。对于任意这样的数对，指示函数 $\mathbb{I}[x_t \in G, \gamma_t = \frac{i}{m}] = 1$，因此该坐标中的损失值可界定为：

$$\mathop{\mathbb{E}}_{y_t \sim Y_t}\left[\frac{1}{n(\pi^{<t}, \frac{i}{m}, G)}\left(2V(\frac{i}{m}, y_t)R(\pi^{<t}, G, \frac{i}{m}) + V(\frac{i}{m}, y_t)^2\right)\right] \leq \frac{2CL_t}{m} + \frac{C^2}{n(\pi^{<t}, \frac{i}{m}, G)},$$

根据定义，我们使用了 $\mathbb{E}_{y_t \sim Y_t}[V(\gamma_t^*, y_t)] = 0$，并且 $|\gamma_t - \gamma_t^*| \leq \frac{1}{m}$，还使用了 $V(\cdot, Y_t)$ 是 L_t-利普希茨的，这是由假设 5 得出的。

这为 AMF 值 w_t^A 提供了上界，因此我们可以应用定理 7，得出算法 4 满足以下 AMF 遗憾界：

$$\begin{aligned}
&\max_{G \in \mathcal{G}, i \in [m]} \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\gamma_t \sim Alg_t, y_t \sim Y_t}\left[\mathbb{I}\left[x_t \in G, \gamma_t = \frac{i}{m}\right] \cdot \frac{1}{n(\pi^{<t}, \frac{i}{m}, G)}\left(2V(\frac{i}{m}, y_t)R(\pi^{<t}, G, \frac{i}{m}) + V(\frac{i}{m}, y_t)^2\right)\right] \\
\leq\ & \frac{1}{T}\sum_{t=1}^{T}\left(\frac{2CL_t}{m} + \frac{\mathbb{I}[x_t \in G, \gamma_t = \frac{i}{m}] \cdot C^2}{n(\pi^{<t}, \frac{i}{m}, G)}\right) + 12C^2\sqrt{\frac{\ln(|\mathcal{G}|m)}{T}} \\
\leq\ & \frac{2CL}{m} + \frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{I}[x_t \in G, \gamma_t = \frac{i}{m}] \cdot C^2}{n(\pi^{<t}, \frac{i}{m}, G)} + 12C^2\sqrt{\frac{\ln(|\mathcal{G}|m)}{T}} \\
\leq\ & \frac{2CL}{m} + \frac{1}{T}\sum_{t=1}^{T}\frac{C^2}{t} + 12C^2\sqrt{\frac{\ln(|\mathcal{G}|m)}{T}} \\
\leq\ & \frac{2CL}{m} + \frac{2C^2\log(T)}{T} + 12C^2\sqrt{\frac{\ln(|\mathcal{G}|m)}{T}},
\end{aligned}$$

在第二个不等式中，我们使用了关于对手的利普希茨假设；在倒数第二个不等式中，我们使用了这样一个事实：在满足 $x_t \in G$ 和 $\gamma_t = \frac{i}{m}$ 的任意轮次中，必有 $n(\pi^{\leq t}, \frac{i}{m}, G) = n(\pi^{<t}, \frac{i}{m}, G) + 1$。

根据我们对损失函数的选择和引理 1，这意味着对于所有 G：

$$\mathbb{E}[K_2(G, \pi)] = \sum_{t=1}^{T}\mathbb{E}[K_2(G, \pi^{\leq t}) - K_2(G, \pi^{<t})] \leq \frac{2CL}{m} + \frac{2C^2\log(T)}{T} + 12C^2\sqrt{\frac{\ln(|\mathcal{G}|m)}{T}},$$

由此完成证明。 □

7 应用

通过将我们的理论与诱导性文献中的已知结果以基本上黑箱的方式结合起来，我们现在得到了若干新的正面和负面结果，揭示了一个重要问题：何时能够针对各种 风险度量生成多重校准预测器？我们首先非正式地概述结果，然后在第 7.1 节给出正式陈述。

贝叶斯对与风险的联合多重校准 任何可诱导性质 Γ 按定义都会使其评分函数 S 最小化；如前所述，该函数可以解释为一种损失，在数据集上对其取期望并将其最小化后，便可得到针对 Γ 的预测器。例如，如果 Γ 是均值，我们会针对 $S(\gamma, y) = (\gamma - y)^2$ 最小化期望评分 $\mathbb{E}_{(x,y)\sim D}[S(\gamma, y)]$——这就是熟悉的最小二乘回归。再如，针对 τ-分位数的一种自然评分 S_τ 是广为人知的针球损失，其定义为 $S_\tau(\gamma, y) := (1-\tau)\gamma + \max\{y - \gamma, 0\}$；将其最小化的过程称为分位数回归。

在损失最小化的语境下，人们关注的可能不仅是最小化者，还包括损失的实际大小，从而提出这样的问题：真实属性值处的预期损失值，即 $\min_\gamma \mathbb{E}_{(x,y)\sim D}[S(\gamma, y)]$，究竟有多高？对于可诱导性质 Γ 及其严格一致损失 S，这个问题的答案由贝叶斯风险 Γ^B 的概念刻画；二维性质 (Γ, Γ^B) 随后被称为关于损失 S 的贝叶斯对。

定义22（贝叶斯风险与贝叶斯对）。固定一个可诱导性质 $\Gamma: \mathcal{P} \to \mathbb{R}$，以及一个严格一致的 Γ-评分函数 S。Γ 在 S 上的贝叶斯风险是一个性质 $\Gamma^B: \mathcal{P} \to \mathbb{R}$，由 $\Gamma^B(P) := S(\Gamma(P), P)$（对于 $P \in \mathcal{P}$）给出。性质 $\Gamma^{BP} := (\Gamma, \Gamma^B)$ 随后被称为关于 S 的贝叶斯对。

例如，（均值，方差）是关于平方损失的贝叶斯对 S；另一个例子是贝叶斯对（分位数，条件风险价值），稍后将对此进行讨论。在一些自然的假设下，大多数 相关的贝叶斯风险本身并不可诱导[Embrechts 等人， 2021]。不过，贝叶斯风险 Γ^B 显然总是可以在其底层性质的条件下被诱导 Γ（因为知道Γ 的值便能完全确定 Γ^B的值）。因此，贝叶斯对是应用我们联合多重校准理论的理想场景：

定理 9 （非正式表述）。在温和假设下，所有关于利普希茨损失的贝叶斯对 $\Gamma^{BP} := (\Gamma, \Gamma^B)$ S 均可使用算法 3实现联合多重校准。

条件风险价值（预期损失）多重校准 条件风险价值（CVaR），亦称预期损失（ES），是金融风险文献中至关重要的尾部风险度量。它最初由 Artzner 等人 [1999], 提出，并由 Rockafellar和Uryasev[2000], 引入凸优化文献，此后一直是近期诸多研究的核心。其定义适用于任意 $\tau \in [0, 1]$，如下：

$$\mathrm{CVaR}_\tau(P) := \mathop{\mathbb{E}}_{Y\sim P}[Y | Y > q_\tau(P)]$$

（其中 $q_\tau(P)$ 是 τ-分位数，属于 P），条件风险价值衡量的是随机变量最高值中占比为 $(1-\tau)$ 的那部分的均值。因此，它能够提供相应分位数以上尾部风险行为的有用信息。加上它还具有许多其他非常有用的性质——例如，它是 相干的 风险度量，定义和证明见 Artzner 等人 [1999] ——条件风险价值由此成为一种广受欢迎且重要的金融风险度量。预期损失在现实世界中的重要性也体现在这样一个事实上：在过去十年中，预期损失被纳入称为巴塞尔协议的国际银行监管规定，用以取代分位数（在金融领域称为风险价值（VaR）），用于市场风险资本计算；详情见 Embrechts 等人 [2014] 。

因此，从理论和实践角度而言，探究条件风险价值是否满足校准合理性十分重要——这将使我们能够采用标准的批处理和在线多重校准算-法来训练条件风险价值的多重校准预测器，从而补充近期关于分位数的算法多重校准研究，参见 Jung 等人 [2023] 和 Bastani 等人 [2022]。

这个问题的答案其实颇为微妙。从否定的一面来看，我们将证明条件风险价值不具备校准合理性，因此无法直接对其训练多重校准预测器。幸运的是，我们证明了可以通过对条件风险价值$_\tau$和相应的分位数 q_τ进行联合多重校准来解决这一问题，而不是单独对条件风险价值进行多重校准。

定理 10 （非正式表述）。 对于 $\tau \in [0, 1]$， τ-条件风险价值 不具备校准合理性。然而， τ-条件风险价值 可以与 对应的 τ-分位数联合实现多重校准，只需实例化算法 3。

对于该定理的否定部分，我们只需结合 2 与 Gneiting[2011] 关于条件风险价值具有非凸水平集的结果，直接调用我们的定理。对于肯定部分，我们回顾一个经典结果（例如参见 Frongillo 和 Kash [2021]）：（q_τ，条件风险价值$_\tau$）是 S作为（重新缩放的） τ-针球损失的贝叶斯对——这使我们能够通过直接将上述关于一般贝叶斯对的结果（定理 9）实例化，为（q_τ，条件风险价值$_\tau$）建立联合多重校准算法。

扭曲风险度量的不可能性结果 扭曲风险度量 [Wang 等人，1997] 是一大类在理论和实践上都具有重要意义的风险度量。扭曲风险度量可以理解为：首先（通过所谓的扭曲函数）对给定分布重新加权，从而为某些关注的结果赋予更高权重；随后评估修改后分布的期望值。均值、分位数、条件风险价值、谱风险度量类别，以及其他许多在理论和实践上具有重要意义的风险度量，都是扭曲风险度量的实例；更多示例和详细信息请参见 Kou 和 Peng [2016] 以及 Gzyl 和 Mayoral [2008] 。

定义 23（扭曲风险度量）。给定扭曲函数 $h:\ [0,1]\to[0,1]$ （即，h是非递减的，并满足 $h(0)=0$ 且 $h(1)=1$），相应的扭曲风险度量$\Gamma^h:\ \mathcal{P}\to\mathbb{R}$ 定义为：

$$\Gamma^h(P):=\int_{-\infty}^{0}(h(1-F_P(x))-1)dx+\int_0^{\infty}h(1-F_P(x))dx$$

对于 $P\in\mathcal{P}$，其中 F_P 是 P的累积分布函数。（我们假设对于所有 $P\in\mathcal{P}$，积分均存在。）

例如，令 $h(x)=x$ （对于 $x\in[0,1]$），则 Γ^h 为分布均值；而选择$h_\tau(x)=\mathbb{I}[x>1-\tau]$，则 Γ^{h_τ}为 τ分位数。

正如 Kou 和 Peng [2016] 以及 Wang 和 Ziegel [2015] 所表明的，然而，均值和分位数基本上是[4] 唯一的 在有限支撑分布上具有凸水平集的扭曲风险度量。通过援引我们的定理2 （无凸水平集 $\Longrightarrow$ 对校准不合理），我们由此可以得出以下全面的否定性结论：

定理 11（非正式）。除了（本质上）均值和分位数之外，对于任何允许标签分布具有有限支撑集的数据集族 $\mathcal{D}$，没有其他扭曲风险度量满足校准合理性。

这一结果表明，对于任何扭曲风险度量，都不会有其他多重校准算法：现有的均值和分位数多重校准算法（本质上）就是仅有的算法。

7.1 正式陈述

7.1.1 贝叶斯风险的联合多重校准

考虑任意一个相对于严格一致评分函数 Γ，Γ^B 的贝叶斯对 $S(\gamma,\ y)$。与第 5 节一样，我们假设值域$_\Gamma\subseteq[0,1]$ 和值域$_{\Gamma^B}\subseteq[0,1]$。为证明贝叶斯对可以联合多重校准，我们需要对与 Γ，Γ^B) 相关联的评分函数和识别函数作出若干假设，以确保满足假设 1、2、3和 4，第 5 节的一般联合多重校准结果依赖于这些假设。

为了满足假设1，我们假设性质 Γ 具有一个识别函数 V，该函数严格递增，且关于其第一个参数是 L-利普希茨的。为了满足假设2，我们还假设 $V(\cdot,\ P)$ 是 L_a-反利普希茨的，适用于 $P\in\mathcal{P}$。

现在注意，对所有 $\gamma\in$ 值域$_\Gamma$，根据定义，贝叶斯风险 Γ^B满足 $\Gamma^B(P)=S(\gamma,y)$，在水平集 $\{P\in\mathcal{P}$ 上有：$\Gamma(P)=\gamma\}$ 属于 Γ。因此，贝叶斯风险 Γ^B 在水平集 $\{P\in\mathcal{P}$ 上的识别函数可以简单地取为：$\Gamma(P)=\gamma\}$

$$V_\gamma^B(\gamma^B,y):=\gamma^B-S(\gamma,y)$$

对所有 $\gamma^B,y\in[0,1]$都成立。因此，对任意 $P\in\mathcal{P}$ 取期望，我们可以将以 $\Gamma=\gamma$ 为条件的 Γ^B 的期望条件识别函数写为 $V_\gamma^B(\gamma^B,P):=\gamma^B-S(\gamma,P)$。

为满足假设 3，我们需要确保 V_γ^B 关于其下标 γ 满足利普希茨条件。为此，我们假设贝叶斯对 (Γ,Γ^B) 的评分函数 S 关于其第一个参数是 L_S- 利普希茨的。对于任意 γ^B 和任意 P，我们可以据此写出 $|V_\gamma(\gamma^B,P)-V_{\gamma'}(\gamma^B,P)|=|S(\gamma',P)-S(\gamma,P)|\le L_S|\gamma-\gamma'|$，这意味着 V_γ^B 关于 γ 是 L_S-利普希茨的。

最后，我们验证第 5 节的假设4。请注意，贝叶斯风险 Γ^B 的识别函数 $V_\gamma^B(\cdot,P)$ 对每个 $\gamma\in[0,1]$ 和 $P\in\mathcal{P}$ 都定义良好，即使在 $\Gamma(P)\neq\gamma$ 时也是如此。此外，$V_\gamma^B(\gamma^B,P)$ 关于 γ^B 是斜率为 1 的线性函数。因此，$V_\gamma^B(\cdot,P)$ 严格递增；事实上，对于 $\gamma\in[0,1]$ 和 $P\in\mathcal{P}$，它还是 1-利普希茨的，正如所需。

[4]参见定义 24 和定理 14，见第 7.1 节，以了解其结果的精确表述。

对于 Γ 和 Γ^B 的评分函数和识别函数，所有必要的假设均已满足，我们现在可以调用定理 6，从而得到以下针对贝叶斯对的联合多重校准保证：

定理 12（贝叶斯对可联合多重校准）。考虑相对于严格一致评分函数 S 的任意贝叶斯对 (Γ, Γ^B)。令 V 为 Γ 的识别函数。假设：(1) 评分函数 S 关于其第一个参数是 L_S-利普希茨的；(2) V 严格递增，关于其第一个参数是 L-利普希茨且是 L_a-反利普希茨的。

选取一个离散化因子 $m \geq 1$。设定 $\alpha^0 = \frac{4L^2}{m}$ 和 $\alpha^1 = \frac{4}{m}$。令 $\alpha^1_* = \frac{8}{m}((LL_aL_S)^2 + 1)$。给定任意 $\mathcal{G} \subseteq 2^{\mathcal{X}}$，使用针对 Γ 的恒等函数 V，以及针对 Γ^B 的恒等函数集合 V^B，实例化 *JointMulticalibration*（算法 3），使得对于所有 γ、$\gamma^B \in [0,1]$、$P \in \mathcal{P}$，$V^B_\gamma(\gamma^B, P) := \gamma^B - S(\gamma, P)$。

然后，算法 3将在最多 $\left(\frac{4L^2}{m}, \frac{8}{m}((LL_aL_S)^2 + 1)\right)$-近似地输出一个 $(\mathcal{G}, V, V^B)$-联合多重校准预测器 $f = (f^0, f^1)$，适用于 (Γ, Γ^B)，并对联合预测器 f进行更新[5]。

7.1.2 联合（分位数，条件风险价值）多重校准

单独来看，条件风险价值并不符合校准合理性。根据我们的定理2，这可直接由 Gneiting [2011], 的经典否定性结果推出；他证明了条件风险价值$_\tau$ 不可诱导，因为它在多个分布族上具有非凸水平集 $\mathcal{P}$。

事实 3 (条件风险价值$_\tau$ 具有非凸水平集 [Gneiting, 2011])。对于任意 $\tau \in [0,1]$，条件风险价值$_\tau$ 相对于以下任何一类分布都具有非凸水平集：$\mathcal{P}$ 定义在某个区间上的分布 $I \subseteq \mathbb{R}$，该分布类包含有限支撑分布，或具有良好定义的概率密度函数的紧支撑分布的有限混合。

积极的一面是，作为定理12的一个直接推论，我们得到下一个结果：对（分位数，条件风险价值）这一对可以进行联合多重校准。为了能够应用定理12，只需确定一个严格一致评分函数 S_τ，使得对于任意τ，（τ-分位数，条件风险价值）这一对都是一个贝叶斯对；然后求得 $\tau \in [0,1]$ 的利普希茨常数，以及用于 S_τ-分位数的严格递增识别函数 V_τ 的利普希茨常数和反利普希茨常数。 τ

事实上，众所周知（例如参见Embrechts 等人 [2021]中的例1），（τ-分位数，条件风险价值$_\tau$）是评分函数 S_τ的一个贝叶斯对，该评分函数是经$\frac{1}{1-\tau}$倍缩放的针球损失：

事实 4（（τ-分位数，条件风险价值$_\tau$）是一个贝叶斯对）。固定任意 $\tau \in [0,1]$，令 $\Gamma := q_\tau$为 τ-分位数，且令 $\Gamma^B :=$ 条件风险价值$_\tau$为 τ-条件风险价值。则（Γ, Γ^B）是关于严格 Γ-一致评分函数 S_τ的贝叶斯对，该函数对所有 γ、$y \in [0,1]$定义如下：

$$S_\tau(\gamma, y) := \gamma + \frac{1}{1-\tau}(y - \gamma)_+,$$

其中，我们将（u）$_+ =$ 记为最大值$\{u, 0\}$。

为了界定 S_τ 的利普希茨常数，注意到其关于第一个参数的导数为 $\frac{\partial S_\tau(\gamma, y)}{\partial \gamma} = \mathbb{I}[y \leq \gamma] - \frac{\tau}{1-\tau}\mathbb{I}[y \geq \gamma]$。因此，$S_\tau$ 的利普希茨常数为 $L_{S_\tau} \leq$上确界$_{\gamma^*, y^*} \left|\frac{\partial S_\tau}{\partial \gamma}(\gamma^*, y^*)\right| =$，）最大值$\{1, \frac{\tau}{1-\tau}\}$。

现在，我们需要确定一个关于第一个参数严格递增的识别函数 V_τ，用于 τ-分位数 q_τ，并研究其利普希茨性质。具体而言，我们采用标准的分位数识别函数，定义为 $V_\tau(\gamma, P) := \Pr_{y\sim P}[y \leq \gamma] - \tau$，对所有 γ 和所有 $P \in \mathcal{P}$ 均成立。显然，$V_\tau(\cdot, P)$ 就是将 P 的累积分布函数平移 τ 后所得的函数。因此，假设 $\mathcal{P}$ 中的所有分布都具有严格递增的累积分布函数，就能确保 V_τ 关于 γ 严格递增。

为方便量化 V_τ 的利普希茨性，假设它关于 γ 可微：这等价于所有 $P \in \mathcal{P}$ 都具有定义良好的概率密度函数 pdf_P，而它就是 $V_\tau(\cdot, P)$ 的导数，即 $\frac{\partial V_\tau(\gamma}{\partial \gamma} = pdf_{P,\ P)}(\gamma)$。因此，对 V_τ 施加利普希茨常数和反利普希茨常数的约束，只需假设底层分布族 $\mathcal{P}$ 中的分布的概率密度函数具有上界和下界。事实上，若现在假设对所有 $P \in \mathcal{P}$，概率密度函数都满足 $0 < M_1 \leq pdf_P(y) \leq M_2 < \infty$，对所有 $y \in [0,1]$ 均成立，那么 V_τ 就是 M_2-利普希茨的，也是 M_1-反利普希茨的。

将上述关于 S_τ 和 V_τ 的利普希茨界和反Lipschitz界代入定理12，由此得到以下联合（分位数、CVaR）多重校准结果：

[5]具体而言，算法 3 对预测器 f至多执行 $R^- R^+ \frac{m^4}{L}$ 次更新，其中记作$R^- = \sup_{\gamma, y\in[0,1]} S(\gamma, y) - \inf_{\gamma, y\in[0,1]} S(\gamma, y)$，以及 $R^+ = \frac{1}{2}\max_{\gamma\in[1/m]}\left(\sup_{\gamma^B, y\in[0,1]}(\gamma^B - S(\gamma, y))^2 - \inf_{\gamma^B, y\in[0,1]}(\gamma^B - S(\gamma, y))^2\right)$.

定理13（(τ-分位数，CVaR$_\tau$) 的联合多重校准）。固定任意常数 $0 < M_1 < M_2$，并取任意一个概率分布族 $\mathcal{P}$，其定义在 $[0,1]$ 上，且其中每个 $P \in \mathcal{P}$ 都具有严格递增的累积分布函数和定义良好的密度函数 pdf$_P$，满足 $M_1 \leq \text{pdf}_P(y) \leq M_2$ 对所有 $y \in [0,1]$均成立。

固定任意目标覆盖水平 $\tau \in [0,1]$，以及数据集上的任意群组结构 $\mathcal{G} \subseteq 2^{\mathcal{X}}$。选取一个离散化 $m \geq 1$。设定 $\alpha^0 = \frac{4M_2^2}{m}$ 和 $\alpha^1 = \frac{4}{m}$。令 $\alpha_*^1 = \frac{8}{m}((M_1 M_2 \max\{1, \tau/(1-\tau)\})^2 + 1)$。

然后，b 恰当地实例化 *JointMulticalibration*(算法 3)，我们可以计算 a

$$\left(\frac{4M_2^2}{m}, \frac{8}{m}\left(\left(M_1 M_2 \max\left\{1, \frac{\tau}{1-\tau}\right\}\right)^2 + 1\right)\right) - \textit{approximately jointly } \mathcal{G}\textit{-multicalibrated predictor}$$

对于 (τ-分位数, 条件风险价值$_\tau$) 这一对，在对联合预测器 f 至多进行 $O\left(\frac{m^4}{M_2}\right)$ 次更新后，得到 $f = (f^0, f^1)$。

7.1.3 扭曲风险度量校准的敏感性

我们首先正式陈述 Kou 和 Peng [2016] 和 Wang 和 Ziegel [2015] 的结果，后文将会用到。该结果表明，在所有扭曲风险度量中，只有以下几类在所有有限支撑分布构成的分布族上具有凸水平集：(1) 均值，(2) 分位数，以及 (3) 另外两种分位数变体风险度量；以下给出相应的定义。

定义24。考虑任意概率分布族 $\mathcal{P}$。对于任意分布 $P \in \mathcal{P}$，记其累积分布函数为 F_P（它不一定严格递增或连续）。我们定义 $\mathcal{P}$ 上的以下分布性质：

1. 对于任意 $\tau \in [0,1]$，τ 分位数定义为：

$$q_\tau(P) = \inf\{y : F_P(y) \geq \tau\} \quad \textit{for } P \in \mathcal{P}.$$

2. 对于任意 $\tau \in [0,1]$ 和 $c \in [0,1]$，定义性质：

$$q^1_{\tau,c}(P) := c \cdot \inf\{y : F_P(y) \geq \tau\} + (1-c) \cdot \inf\{y : F_P(y) > \tau\} \quad \textit{for } P \in \mathcal{P}.$$

3. 对于任意 $\tau \in [0,1]$ 和 $c \in [0,1]$，定义性质：

$$q^2_{\tau,c}(P) := c \cdot \inf\{y : F_P(y) > 0\} + (1-c) \cdot \inf\{y : F_P(y) = 1\} \quad \textit{for } P \in \mathcal{P}.$$

注意：(1) $q^2_{\tau,c}$ 只是分布的 0% 分位数与 100% 分位数的凸组合；(2) $q^1_{\tau,c}$ 实际上（对于所有 $c \in [0,1]$）是累积分布函数 F_P 严格递增时的 τ 分位数。

Kou 和 Peng [2016] 证明，分布均值以及定义24中列出的三种（参数化）性质，是仅有的具有凸水平集的扭曲风险度量。后来，Wang 和 Ziegel [2015], 简化并完善了这一结果的证明，指出即使将范围限制在支撑集至多包含 3 个点的分布族内，这一否定性结论仍然成立。

定理14 （具有凸水平集的扭曲风险度量的刻画 [Kou 和 Peng，2016年，Wang 和 Ziegel，2015]）。设 $\mathcal{P}_3$ 为支撑集至多包含3 个实数值点的所有概率分布构成的集合。设 $\mathcal{P}_{bd}$ 为实数上具有定义良好的概率密度函数的所有有界分布构成的集合。设 $\mathcal{P}$ 为实数上的任意分布族，且满足 $\mathcal{P} \supseteq \mathcal{P}_3$，或 $\mathcal{P} \supseteq \mathcal{P}_{bd}$。

考虑任意扭曲风险度量 $\Gamma : \mathcal{P} \to \mathbb{R}$。那么，$\Gamma$ 不满足关于 $\mathcal{P}$ 的凸水平集假设，除非它属于以下性质之一：

1. 分布均值；
2. 对于某个 $\tau \in [0,1]$，τ-分位数 q_τ；
3. 性质 $q^1_{\tau,c}$，其中 $\tau, c \in [0,1]$；4. 性质 $q^2_{\tau,c}$，其中 $\tau, c \in [0,1]$。

现在，根据我们的定理2，我们可以立即得出结论：对于定理 14 中的任意 $\mathcal{P}$，除均值、分位数或这两个参数性质 $q^1_{\tau,c}$ 或 $q^2_{\tau,c}$ 之外，没有任何扭曲风险度量适用于在任何 $\mathcal{P}$-兼容的数据集分布族 $\mathcal{D}$ 上进行校准，只要该分布族包含所有 $\mathcal{P}$-兼容的双点数据集分布。正式地重述如下：

定理 15（扭曲风险度量的校准合理性）。设 $\mathcal{P}_3$ 为所有支撑集至多包含 3 个实数值点的概率分布的集合。设 $\mathcal{P}_{bd}$ 为所有定义在实数域上且有界、概率密度函数（PDF）有定义的分布的集合。设 $\mathcal{P}$ 为实数域上任意一个分布凸空间，且满足 $\mathcal{P} \supseteq \mathcal{P}_3$ 或 $\mathcal{P} \supseteq \mathcal{P}_{bd}$。

考虑一个 任意扭曲风险度量 $\Gamma : \mathcal{P} \to \mathbb{R}$，以及任意一个由 $\mathcal{P}$ 兼容的数据集 t 分布构成的族，其中包含所有与 $\mathcal{P}$ 兼容的双点数据集分布。

Th则 Γ 在 $\mathcal{D}$ 上不具备校准合理性，除非 Γ 是以下属性之一 $:$

1. 分布均值；2. 某个 $\tau \in [0,1]$ 的 τ 分位数 q_τ；3. 对于某个 $\tau, c \in [0,1]$，性质 $q^1_{\tau,c}$；
4. 对于某个 $\tau, c \in [0,1]$，性质 $q^2_{\tau,c}$。

致谢

我们衷心感谢 Christopher Jung 和 Arpit Agarwal 在本研究早期进行的启发性交流。本研究部分由西蒙斯算法公平性理论合作项目以及美国国家科学基金会的 FAI-2147212 和 CCF-2217062 项目资助。

参考文献

Philippe Artzner、Freddy Delbaen、Jean-Marc Eber 和 David Heath。相干风险度量。《数学金融》，9(3):203–228，1999年。

Osbert Bastani、Varun Gupta、Christopher Jung、Georgy Noarov、Ramya Ramalingam 和 Aaron Roth。实用对抗性多重有效保形预测。载于《神经信息处理系统（NeurIPS）》，2022年。

Glenn W Brier。《月度天气评论》中以概率形式表示的预测验证，78 (1):1–3，1950年。

Zhun Deng、Cynthia Dwork 和 Linjun Zhang。Happymap：一种广义多重校准方法。载于《理论计算机科学创新会议》(ITCS)，2023 年。

Joseph Diestel 和 John Jerry Uhl。向量测度，数学调查丛书第 15 卷。美国数学学会，美国罗得岛州普罗维登斯，1977年。

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计算理论研讨会论文集，第 1095–1108 页，2021年。

Cynthia Dwork、Michael P Kim、Omer Reingold、Guy N Rothblum 和 Gal Yona。超越伯努利：生成无法与自然区分的随机结果。载于国际算法学习理论会议，第 342–380 页。PMLR，2022年。

Paul Embrechts、Giovanni Puccetti、Ludger R¨uschendorf、Ruodu Wang 和 Antonela Beleraj。《回应巴塞尔协议3.5：一份学术意见》。《风险》，2(1):25–48，2014。

Paul Embrechts、Tiantian Mao、Qiuqi Wang 和 Ruodu Wang。贝叶斯风险、可诱导性与预期损失。《数学金融》，31(4):1190–1217，2021年。

Jessica Finocchiaro 和 Rafael Frongillo。《连续性质的凸引出》。《神经信息处理系统进展》，31，2018 年。

Rafael M Frongillo 和 Ian A Kash。《统计性质的引 elicitation 复杂度》。《生物计量学》，108（4）：857–879，2021年。Sumegha Garg、Christopher Jung、Omer Reingold 和 Aaron Roth。《在线全预测器》。手稿，2023 年。Ira Globus-Harris、Declan Harrison、Michael Kearns、Aaron Roth 和 Jessica Sorrell。《多重校准作为回归的提升》。arXiv 预印本 arXiv:2301.13767，2023 年。Tilmann Gneiting。《制作和评估点预测》。《美国统计学会杂志》，106（494）：746–762，2011 年。Irving John Good。《理性决策》。载于《统计学突破》，第 365–377 页。施普林格，1992年。Parikshit Gopalan、Adam Tauman Kalai、Omer Reingold、Vatsal Sharan 和 Udi Wieder。《全预测器》。载于 ITCS，2022a。Parikshit Gopalan、Michael P Kim、Mihir A Singhal 和 Shengjia Zhao。《低次多重校准》。载于《学习理论会议》，第 3193–3234 页。PMLR，2022b。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》。载于第 13 届理论计算机科学创新会议（ITCS 2022）。Schloss Dagstuhl-Leibniz-Zentrum f¨ur Informatik，2022年。Henryk Gzyl 和 Silvia Mayoral。《扭曲风险度量与谱风险度量之间的关系》。《经济与金融评论》，2008年。Ursula H´ebert-Johnson、Michael Kim、Omer Reingold 和 Guy Rothblum。《多重校准：面向（计算可识别）群体的校准》。载于《国际机器学习会议》，第 1939–1948 页。PMLR，2018 年。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。Christopher Jung、Georgy Noarov、Ramya Ramalingam 和 Aaron Roth。《批量多重有效保形预测》。载于《国际学习表征会议》（ICLR），2023 年。Michael P Kim、Amirata Ghorbani 和 James Zou。《多重准确性：用于分类公平性的黑箱后处理》。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。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》。《论文集》

f

美国国家科学院，119(4):e2108097119，2022年。Steven Kou 和 Xianhua Peng。《经济尾部风险的度量》。Operations Research，64 (5):1056–1072，2016年。Nicolas S Lambert、David M Pennock 和 Yoav Shoham。《概率分布属性的引出》。载于第9届ACM电子商务会议论文集，第129–138页，2008年。Daniel Lee、Georgy Noarov、Mallesh M. Pai 和 Aaron Roth。《在线极小极大多目标优化：Multicalibeating及其他应用》。载于神经信息处理系统（NeurIPS），2022年。Kent Harold Osband。《为更好的成本预测提供激励（预测、不确定性引出）》。博士论文，加利福尼亚大学伯克利分校，1985年。R Tyrrell Rockafellar 和 Stanislav Uryasev。《条件风险价值的优化》。Journal o

f

风险，2:21–42，2000年。

Aaron Roth。《不确定性：不确定性估计的现代主题》。https://www.cis.upenn.edu/aaroth/uncertainty-notes.pdf，2022年。Leonard J Savage。《个人概率与期望的引出》。《美国统计学会杂志》，66(336):783–801，1971年。Ingo Steinwart、Chlo´e Pasin、Robert Williamson 和 Siyu Zhang。《性质的引出与识别》。载于《学习理论会议》，第482–526页。PMLR，2014。Ruodu Wang 和 Johanna F Ziegel。《可引出的扭曲风险度量：简明证明》。《统计与概率快报》，100:172–175，2015年。Shaun S Wang、Virginia R Young 和 Harry H Panjer。《保险价格的公理化刻画》。《保险：数学与经济学》，21(2):173–183，1997年。

批量多重校准的收敛保证（定理5的证明）

我们的c 算法 1 的收敛分析将利用以下自然的势函数：

定义 25。 算法 1 在第 t 轮次的势函数为：

$$\Phi_t := \mathop{\mathbb{E}}_{(x,y)\sim D}[S(f_t(x), y)] = \mathop{\mathbb{E}}_{x\sim X}[S(f_t(x), Y_x)],$$

其中 $f_t : \mathcal{X} \to \mathbb{R}$ 是算法第 t 次迭代开始时的性质预测器，且 S 是满足假设1的性质 Γ 严格一致评分函数。

首先，我们证明以下辅助引理，该引理界定了在标签分布 Y上，将性质值 $\Gamma(Y)$ 的错误预测 γ 修正时， S——算法 1 的势函数——所发生的变化。稍后，我们将利用这一事实，界定预测器 f 在每个更新步骤中针对 Γ的算法进展。

引理 2. 考虑任意性质 $\Gamma: \mathcal{P} \to \mathbb{R}$。假设 $S: \text{Range}_\Gamma \times \mathcal{Y} \to \mathbb{R}$ and $V: \text{Range}_\Gamma \times \mathcal{Y} \to \mathbb{R}$with $V(\gamma, y) = {}^{\partial S}(\gamma,y)\,{}_{\partial\gamma}$ 是~~Γ 的~~严格一致评分函数及其对应的识别函数，并且满足假设 1。那么，对于任意 $\gamma \in \text{Range}_\Gamma$ 以及任意标签分布 $Y \in \mathcal{P}$，令 L_Y 为 $V(\cdot, Y)$ 的利普希茨常数，则有：

$$\frac{(V(\gamma,Y))^2}{2L_Y} \le S(\gamma,Y) - S(\Gamma(Y),Y) \le V(\gamma,Y)(\gamma-\Gamma(Y)) - \frac{(V(\gamma,Y))^2}{2L_Y}.$$

证明。我们将借助以下命题证明这一结果。

命题1. 对于定义在任意区间 $[a,b]$ 上的任意 L-利普希茨递增函数 h ，有：

$$h(a)(b-a) + \frac{(h(b)-h(a))^2}{2L} \le \int_a^b h(t)dt \le h(b)(b-a) - \frac{(h(b)-h(a))^2}{2L}.$$

证明。在对 h 的这些约束下，积分 $\int_a^b h(t)dt$ 的最大值将在以下情况下取得：$h(t)$ 以尽可能快的速率（即速率为 L）从 $h(a)$ 增至 $h(b)$，其中 $t' \in [a,b]$ 由 $(t'-a)L = h(b)-h(a)$ 定义；随后在 $t \in [t',b]$ 上保持为常数 $h(b)$。该分段线性函数在 $[a,b]$ 上的积分给出了上界。

反过来，如果 $h(t)$ 首先在 $t \in [a,t']$ 内保持为常数 $h(a)$，其中 t' 的定义满足 $(b-t')L = h(b) - h(a)$，随后再以可能的最快速率（即速率 L）在 $t \in [t',b]$ 内从 $h(a)$ 增加到 $h(b)$，则积分 $\int_a^b h(t)dt$ 将取得最小值。在 $[a,b]$ 上对该函数积分，即可得到所述下界。 □

由于 V 是 S 的导数，根据微积分基本定理，我们得到： $S(\gamma,Y) - S(\Gamma(Y)\ Y) = \int_{\Gamma(Y)}^{\gamma} V(t,Y)dt$。

首先，假设 $\gamma \ge \Gamma(Y)$。根据假设 1， V 从 $\Gamma(Y)$ 连续增加至 γ，且利普希茨常数为 L_Y。于是，根据结论 1，并利用 $V(\Gamma(Y)\ Y) = 0$，我们得到

$$\frac{(V(\gamma,Y))^2}{2L_Y} \le \int_{\Gamma(Y)}^{\gamma} V(t,Y)dt \le V(\gamma,Y)(\gamma-\Gamma(Y)) - \frac{(V(\gamma,Y))^2}{2L_Y}.$$

现在假设 $\gamma < \Gamma(Y)$。那么，我们有：

$$S(\gamma,Y) - S(\Gamma(Y),Y) = \int_{\Gamma(Y)}^{\gamma} V(t,Y)dt = -\int_{\gamma}^{\Gamma(Y)} V(t,Y)dt.$$

根据假设 1， V从 γ连续增加至 $\Gamma(Y)$，且其利普希茨常数为 L_Y。根据命题1，我们有： $-V(\Gamma(Y)\ Y)(\Gamma(Y) - \gamma) + (V(\Gamma(Y)\ Y) - V(\gamma,Y))^2 \le -\int_{\gamma}^{\Gamma(Y)} V(t,Y)dt \le -V(\gamma,Y)(\Gamma(Y) - \gamma) - (V(\Gamma(Y)\ Y) - V(\gamma,Y))^2$。由 $V(\Gamma(Y)\ Y) = 0$ 可简化为：$(V(\gamma\ Y))^2 \le -\int_{\gamma}^{\Gamma(Y)} V(t,Y)dt \le V(\gamma,Y)(\gamma-\Gamma(Y)) - (V(\gamma\ Y))^2$。因此，我们证明了在两种情形 $\gamma \ge \Gamma(Y)$ 和 $\gamma < \Gamma(Y)$ 下的界。 □

现在，我们可以证明定理 5，该定理给出了算法 1 的收敛速率。为方便起见，我们在此重述该定理。

定理 5（算法 1 的保证）。固定数据分布 $D \in \Delta\mathcal{Z}$ 和群组 $\mathcal{G} \subseteq 2^{\mathcal{X}}$。固定一个性质 Γ 及其评分函数 S 和恒等函数 $V = \frac{\partial S}{\partial \gamma}$，并令其满足假设 1，使得$V(\cdot, Y_Q)$ 在由 D 诱导的所有标签分布 Y_Q（针对 $Q \subseteq \mathcal{X}$）上均为 L-利普希茨函数。设离散化$m \geq 1$。若算法 1 以预测器 f_1 初始化：$\mathcal{X} \to \mathbb{R}$ 的评分为 $\mathbb{E}_{(x,y)\sim D}[S(f_1(x)\ y)] = C_{init}$，且 $C_{opt} = \mathbb{E}_{(x,y)\sim D}[S(f_\Gamma^D(x)\ y)]$ 是真实分布预测器 f_Γ^D的评分，那么算法 1 至多经过 $(C_{init} - C_{opt})\frac{m^2}{L}$ 次更新，就会生成一个 $\frac{4L^2}{m}$-近似（$\mathcal{G}, V$）多重校准的 Γ-预测器 f。

证明。假设算法在第 t 轮尚未停止。因此，f_t 尚不满足 α-近似的（$\mathcal{G}, V$）-多重校准；由抽屉原理，存在一对（G, γ）$\in \mathcal{G} \times [1/m]$，使得在集合 $Q_t := \{x \in \mathcal{X} : x \in g, f_t(x) = \gamma\}$: 上：

$$|V(\gamma, Y_{Q_t})| \geq \sqrt{\frac{\alpha/m}{\Pr_{x\sim X}[x \in Q_t]}}. \tag{3}$$

现在，令 $\gamma' = \operatorname{argmin}_{\gamma'' \in [1/m]} |V(\gamma'', Y_{Q_t})|$，算法将依据以下规则更新 $f_t \to f_{t+1}$：

$$f_{t+1}(x) := \mathbb{I}[x \notin Q_t] \cdot f_t(x) + \mathbb{I}[x \in Q_t] \cdot \gamma'.$$

根据势函数值 Φ_t 和 Φ_{t+1} 的定义，我们有

$$\begin{aligned}
\Phi_{t+1} &= \Pr_{x\sim X}[x \in Q_t] \mathop{\mathbb{E}}_{x\in X}[S(f_{t+1}(x), Y_x)|x \in Q_t] + \Pr_{x\sim X}[x \notin Q_t] \mathop{\mathbb{E}}_{x\in X}[S(f_{t+1}(x), Y_x)|x \notin Q_t] \\
&= \Pr_{x\sim X}[x \in Q_t] \mathop{\mathbb{E}}_{x\in X}[S(f_{t+1}(x), Y_x)|x \in Q_t] + \Pr_{x\sim X}[x \notin Q_t] \mathop{\mathbb{E}}_{x\in X}[S(f_t(x), Y_x)|x \notin Q_t] \\
&= \Phi_t + \Pr_{x\sim X}[x \in Q_t] \left(\mathop{\mathbb{E}}_{x\in X}[S(f_{t+1}(x), Y_x) - S(f_t(x), Y_x)|x \in Q_t] \right) \\
&= \Phi_t + \Pr_{x\sim X}[x \in Q_t] \left(\mathop{\mathbb{E}}_{x\in X}[S(\gamma', Y_x) - S(\gamma, Y_x)|x \in Q_t] \right) \\
&= \Phi_t + \Pr_{x\sim X}[x \in Q_t] \left(S(\gamma', Y_{Q_t}) - S(\gamma, Y_{Q_t}) \right).
\end{aligned}$$

这里，步骤 2 成立是因为对于 Q_t 之外的所有 x，$f_t(x) = f_{t+1}(x)$；步骤 5 则利用了 f_t 和 f_{t+1} 在 Q_t 上均为常数这一事实，从而可以将 f_t, f_{t+1} 在 $x \sim X$ 上的期望评分，直接改写为预测器取值 γ, γ' 相对于区域 Q_t 上标签的混合分布 Y_{Q_t} 的评分。

由此可得：

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= \Pr_{x\sim X}[x \in Q_t]\Big((S(\gamma', Y_{Q_t}) - S(\Gamma(Y_{Q_t}), Y_{Q_t})) - (S(\gamma, Y_{Q_t}) - S(\Gamma(Y_{Q_t}), Y_{Q_t})) \Big) \\
&\leq \Pr_{x\sim X}[x \in Q_t]\Big(V(\gamma', Y_{Q_t})(\gamma' - \Gamma(Y_{Q_t})) - \frac{(V(\gamma', Y_{Q_t}))^2}{2L} - \frac{(V(\gamma, Y_{Q_t}))^2}{2L} \Big) \\
&\leq \Pr_{x\sim X}[x \in Q_t]\Big(V(\gamma', Y_{Q_t})(\gamma' - \Gamma(Y_{Q_t})) - \frac{(V(\gamma', Y_{Q_t}))^2}{2L} \Big) - \frac{\alpha}{2Lm} \\
&\leq \Pr_{x\sim X}[x \in Q_t]\Big(V(\gamma', Y_{Q_t})(\gamma' - \Gamma(Y_{Q_t})) \Big) - \frac{\alpha}{2Lm} \\
&\leq V(\gamma', Y_{Q_t})(\gamma' - \Gamma(Y_{Q_t})) - \frac{\alpha}{2Lm} \\
&\leq L|\gamma' - \Gamma(Y_{Q_t})| \cdot |\gamma' - \Gamma(Y_{Q_t})| - \frac{\alpha}{2Lm} \\
&\leq \frac{L}{m^2} - \frac{\alpha}{2Lm}.
\end{aligned}$$

通过引入一个加上又减去的项 $S(\Gamma(Y_{Q_t})\ Y_{Q_t})$，即可得到该等式。第 1 个不等式将引理 2 中的上界和下界分别应用于两个评分差。第 2 个不等式由方程 3的 α-失准校准条件得出。第 3 个不等式舍去了非正项 $-\Pr_{x\sim X}[x \in Q_t](V(\gamma', Y_{Q_t})^2))$。第 4 个不等式舍去了因子 $\Pr_{x\sim X}[x \in Q_t] \leq 1$。第 5 个

个不等式是因为根据 V 的利普希茨性，有 $|V(\gamma', Y_{Q_t})| = |V(\gamma', Y_{Q_t}) - V(\Gamma(Y_{Q_t}) \; Y_{Q_t})| \leq L|\gamma' - \Gamma(Y_{Q_t})|$。第 6 个不等式是因为 γ' 与真实属性值$\Gamma(Y_{Q_t})$ 在 Q_t 上的距离至多为 $\frac{1}{m}$：更远的网格点 γ'' 会导致更差的 $|V(\gamma'', Y_{Q_t})|$（因此会与算法选择 γ' 相矛盾）；这是由 m-离散化的结构以及 $|V(\cdot, Y_{Q_t})|$ 在远离 $\Gamma(Y_{Q_t})$ 的两个方向上单调递增所致。

现在令 $\alpha = \frac{4L^2}{m}$，得到 $\Phi_{t+1} - \Phi_t \leq \frac{L}{m^2} - \frac{\alpha}{2Lm} = -\frac{L}{m^2} = -\frac{L}{m^2}$。对轮次 $t = 1, \ldots, T$ 求和并进行望远镜求和，其中 T是收敛前的总迭代次数，我们得到：

$$\Phi_T - \Phi_1 \leq -T\frac{L}{m^2}.$$

根据假设 $\Phi_1 = C_{\text{init}}$ 和 $\Phi_T \geq C_{\text{opt}}$，因此有 $T\frac{L}{m^2} \leq \Phi_1 - \Phi_T \leq C_{\text{init}} - C_{\text{opt}}$，从而

$$T \leq (C_{\text{init}} - C_{\text{opt}})\frac{m^2}{L},$$

由此完成证明。 □

B 批量多重校准的有限样本保证

我们对算法 1 的描述仿佛它可以直接访问底层分布 D （因为它会计算识别函数 V 在底层分布上的期望）。一般而言，我们无法直接访问 D，而只能获得一个样本 $D \sim D^n$，其中包含从 D中独立同分布抽取的 n 个点。实际应用中，我们会在 D中这 n 个点所构成的经验分布上运行算法，而算法的保证也会推广到底层分布 D，即这些点的抽样来源。Jung 等人 [2023] 针对分位数这一特殊情况（此时识别函数 V 是针球损失）证明了这一点；但事实上，他们的证明只用到了假设 1 中的条件。我们在此陈述该定理更一般的版本（Jung 等人 [2023]的工作中隐含了这一版本），并简要概述其论证。我们指出，这一论证旨在说明，当算法 1 被用作经验风险最小化算法时，它能够实现泛化。另一种获得泛化界的方法是沿用 H´ebert-Johnson 等人 [2018] 的策略，利用自适应数据分析技术来实现统计查询预言机，然后修改算法，使其仅通过该预言机计算量 $V(\gamma, Q_t)$——这种方法同样可以得到类似的界。

定理 16（Jung 等人 [2023] 的工作中隐含）。固定一个分布 $D \in \Delta\mathcal{Z}$ 和一个性质 Γ，以及一个有界识别函数 V。假设算法 1 使用数据集 $D \sim D^n$ 上的经验分布 ˆ 运行，该数据集由来自 D 的 n 个独立同分布样本组成。那么，如果算法 1 在 T 轮后停止并返回模型 f_T，则以关于数据分布随机性的概率 $1-\delta$， f^T 满足相对于 D 的 α-近似 $(\mathcal{G}, V)$-多重校准，适用于：

$$\alpha = \frac{4L^2}{m} + O\left(\sqrt{\frac{\ln(1/\delta) + T\ln(m^2|\mathcal{G}|)m}{n}} + \frac{m\ln(1/\delta) + T\ln(m^2|\mathcal{G}|)}{n}\right)$$

证明的结构很简单。由于 V 有界，关于 D 的 V 的期望（即定义近似 $(\mathcal{G}, V)$-多重校准时出现的形如 $V(\gamma, Y_{G,\gamma})$ 的量），在经验分布 D 上计算时，会以高概率集中在其期望附近。因此，对于任意固定模型 f_t，我们可以对每个 $G \in \mathcal{G}$ 和 $\gamma \in$ 值域 $_{f_t} = [1/m]$ 使用并合界，从而证明其 $(\mathcal{G}, V)$-多重校准误差在样本内外相近。当然，算法输出的模型 f_T ˆ 在抽取 D 之前并非固定，因此要证明这一结论，必须对所有可能输出的模型 f_T 使用并合界。不过这是可行的：由于算法 1 生成的模型的值域被限制在 $[1/m]$，固定一个模型 f_t 后，我们可以计算下一步可能产生的模型 f_{t+1} 的数量——对于 $G \in \mathcal{G}$、 $\gamma \in [1/m]$ 和 $\gamma' \in [1/m]$ 的每一种选择，至多产生一个，因此模型总数至多为 $|G|m^2$ 。因此，固定某个初始模型 $f_1 = f$ 后，如果算法在 T 步后停止，则可能输出的模型 f_T 数量不超过 $(|G|m^2)^T$。随后对所有此类模型使用并合界，即可推出该定理。

定理 16 根据轮次数量给出了算法 1 在停止前的泛化误差上界T。因此，结合对 T的上界，就能得到泛化误差的最坏情况界

。定理 5 将轮复杂度上界设为 $T \leq O\left(\frac{m^2}{L}\right)$，但有一个问题： L 是针对真实底层分布 V 计算的期望的利普希茨常数 D，而它通常无法在经验分布 $\hat{}D$上保持不变。尽管如此，Jung 等人 [2023] 证明，在 $\hat{}D$ 上运行时，同样的收敛界依然成立（相差常数因子）——他们指出，在 $\hat{}D$ 上运行的算法每一轮都会降低在 D 上测得的势函数（此处已作出利普希茨假设）。这是因为算法每轮都通过测量形如 $V(\gamma, Q)$ 的量来决定更新，这些量是有界函数 V的期望，因此会集中在其真实值附近。

定理 17（隐含于 Jung 等人 [2023]）。固定一个分布 $D \in \Delta\mathcal{Z}$（它为每个 $Q \subset \mathcal{X}$ 导出一组条件标签分布 Y_Q）以及一个性质 Γ 和一个识别函数 V。假设 Γ 和 V连同标签分布集 $\mathcal{P} = \{Y_Q : Q \subset \mathcal{X}\}$满足假设 1，且利普希茨常数为 L。设算法 1 使用数据集 $D \sim D^n$ 上的经验分布 $\hat{}$ 运行，该数据集包含来自 D的 n 个独立同分布样本。那么，对于任意 $\delta > 0$，若：

$$n \geq \Omega\left(\ln\left(\frac{m^2}{L\delta}\right) + \frac{m^2}{L}\ln\left(\frac{|G|m}{L}\right)\frac{m^4}{L^2}\right)$$

以 $1-\delta$ 的概率， D算法 *1* 在至多 $T = O\left(\frac{m^2}{L}\right)$ 步后终止。

结合定理 16，这为批量性质多重校准建立了一个最坏情况泛化界，该界关于问题的所有参数以及所假设的利普希茨常数 L（即该性质的识别函数 V的利普希茨常数）。

C 联合多重校准保证

我们在此陈述算法 2 所具备的保证。该陈述比类似算法 1 的保证更强，体现在两个方面：1）在收敛时所达到的多重校准误差概念（公式 4）比算法 1 的更强。2）我们证明，即使输入群组族 $\mathcal{G}$ 事先并未固定（因此可能随时间变化），算法 2 对预测器 f 所执行的更新次数也不会超过某个确定的数量；并且如果确实执行了这么多次更新，那么它将关于 所有 可测子集 $\mathcal{X}$ 近似校准（而不仅仅是关于它明确执行过更新的那些子集）。

引理 3。 设 $\alpha = \frac{4L^2}{m}$。 $\mathit{BatchMulticalibration}^V$（算法 *2*）在函数 V 上运行，该函数单调递增，并且关于其第一个参数是 L-利普希茨的，输出一个 $[1/m]$-离散化预测器f，满足：

$$\Pr_{x\in\mathcal{X}}[f(x) = \gamma, x \in G]\left(V(\gamma, Y_{(\gamma,G)})\right)^2 \leq \frac{\alpha}{m} \quad \textit{for all } \gamma \in [1/m], G \in \mathcal{G}. \tag{4}$$

此外，算法 *2* 至多经过 $\frac{Bm^2}{L}$次迭代即可终止，其中 $B =$上确界$_{\gamma,\ y\in[0,1]} S(\gamma,\ y)-$下确界$_{\gamma,\ y\in[0,1]} S(\gamma,\ y)$，其中 S 是 V的一个原函数；如果运行这么长时间，则所得预测器将满足（*4*）对于 所有（可测）区域 $G \subseteq \mathcal{X}$。

证明。 令 S 为 V的一个原函数，并仿照定理 5的证明，将算法 t 2第 t 次迭代时的势函数值定义为：

$$\Phi_t := \mathop{\mathbb{E}}_{(x,y)\sim D}[S(f_t(x), y)] = \mathop{\mathbb{E}}_{x\sim X}[S(f_t(x), Y_x)].$$

假设在第 t轮，算法发现其 `while` 循环条件对某些 $G \in \mathcal{G}$和 $\gamma \in [1/m]$不成立。令 $Q_t = \{x \in \mathcal{X} : x \in \mathcal{G},\ f\ (x)\ = \gamma\}$。通过与定理5证明中相同的计算，我们得到

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &\leq \Pr_{x\sim X}[x \in Q_t]\left(V(\gamma', Y_{Q_t})(\gamma' - \gamma_t^*) - \frac{(V(\gamma, Y_{Q_t}))^2}{2L}\right) \\
&\leq \Pr_{x\sim X}[x \in Q_t]\left(|V(\gamma', Y_{Q_t})||\gamma' - \gamma_t^*| - \frac{(V(\gamma, Y_{Q_t}))^2}{2L}\right) \\
&\leq \Pr_{x\sim X}[x \in Q_t]\left(L|\gamma' - \gamma_t^*|^2 - \frac{(V(\gamma, Y_{Q_t}))^2}{2L}\right),
\end{aligned}$$

其中，我们用 γ_t^*表示满足 $V(\gamma_t^*, Y_{Q_t}) = 0$ 的唯一点（它类似于定理 $\Gamma(Y_{Q_t})$ 证明中的 5）。这一论证仍然成立，因为它基于引理 2，而该引理所要求的 V 和 S 的性质在此处仍然满足。

现在，与上述证明一样，由于 $V(\cdot, Y_{Q_t})$ 的单调性，并且算法选择了 $\gamma' = \text{argmin}_{\gamma'' \in [1/m]} |V(\gamma'', Y_{Q_t})|$，因此必有 $|\gamma' - \gamma_t^*| \le \frac{1}{m}$，从而得到

$$\Phi_{t+1} - \Phi_t \le \frac{L}{m^2} - \frac{1}{2L} \Pr_{x \sim X}[x \in Q_t] |V(\gamma, Y_{Q_t})|^2.$$

由于 while 循环的条件要求 $\Pr_{x \sim X}[x \in Q_t] |V(\gamma, Y_{Q_t})|^2 \ge \alpha/m$，我们得到

$$\Phi_{t+1} - \Phi_t \le \frac{L}{m^2} - \frac{\alpha}{2Lm}.$$

令 $\alpha = \frac{4L^2}{m}$，则有 $\Phi_{t+1} - \Phi_t \le -\frac{L}{m^2}$，因此通过望远镜求和可得 $\Phi_T - \Phi_0 \le -\frac{TL}{m^2}$。根据定义，$B = \sup_{\gamma,\ y \in [0,1]} S(\gamma, y) - \inf_{\gamma,\ y \in [0,1]} S(\gamma, y)$，因此还有 $\Phi_T - \Phi_0 \ge -B$，从而得到$T \le \frac{Bm^2}{L}$，这给出了算法迭代次数的上界。

重要的是，在这一论证中，我们从未用到 Q_t 的实际定义（即 $Q_t = \{x \in \mathcal{X} : x \in \mathcal{G}, f(x) = \gamma\}$）——我们只用到了它满足 while 循环条件这一点，即 $\Pr_{x \sim X}[x \in Q_t] |V(\gamma, Y_{Q_t})|^2 \ge \alpha/m$。因此，关于迭代总次数的上界 $\frac{Bm^2}{L}$ 实际上适用于任意区域序列 $Q_1, Q_2, \ldots$，其中每个 $Q_i \subseteq \mathcal{X}$ 都相对于 $\mathcal{X}$ 上的边缘数据分布可测。由此可知，如果 BatchMulticalibration^V 至少运行 $\frac{Bm^2}{L}$ 次迭代，那么在完成第 $t = \frac{Bm^2}{L}$ 次迭代后，就不会存在任何违反条件 (4) 的可测 $Q \subseteq \mathcal{X}$。因此，无论 BatchMulticalibration^V 针对哪个群组族 $\mathcal{G}$ 运行（即使群组族在执行过程中任意变化），它更新预测器 f 的次数也绝不会超过 $\frac{Bm^2}{L}$ 次。证明完毕。 □

定理 6（算法 3 的保证）。设 $\alpha^0 = 4(L^0)^2$ 和 $\alpha^1 = 4(L^1)^2$。令 $\alpha_*^1 = 8((L^0 L_*^0 L_c)^2 + (L^1)^2)$。对于任意 $\mathcal{G} \subseteq 2^{\mathcal{X}}$、$m \ge 1$、*JointMulticalibration*^{m}，算法 3 输出一个针对性质 (Γ^0, Γ^1) 的 (α^0, α_*^1)-近似 $(\mathcal{G}, V^0, V^1)$-联合多重校准预测器 $f = (f^0, f^1)$，最多对 f 进行 $\frac{B^0 B^1 m^4}{L^0 L^1}$ 次更新。在此，$B^0 := \sup_{\gamma, y \in [0,1]} S^0(\gamma, y) - \inf_{\gamma, y \in [0,1]} S^0(\gamma, y)$，其中 S^0 是 V^0 的一个原函数；并且 $B^1 := \max_{\gamma^0 \in [1/m]} \left(\sup_{\gamma, y \in [0,1]} S_{\gamma^0}^1(\gamma, y) - \inf_{\gamma, y \in [0,1]} S_{\gamma^0}^1(\gamma, y) \right)$，其中对于每个 $S_{\gamma^0}^1$，$V_{\gamma^0}^1$ 是其一个原函数。

证明。运行时间：首先，注意如果我们设置 $\alpha^0 = 4(L^0)^2$，算法 3 中的 while 循环将在至多 $\frac{B^0 m^2}{L^0}$ 次迭代后停止。事实上，所有在 f^0 上使用识别函数$^m V^0$ 对 BatchMulticalibration^V 的调用，都可以整合为一个过程：先针对 $\mathcal{G}_1^0$ 对 f^0 进行多重校准，然后取所得的预测器并针对 $\mathcal{G}_2^0$ 对其进行多重校准，如此继续，直到满足联合多重校准中 while 循环的停止条件。这等价于运行一次 BatchMulticalibration^V，其中群组族会不时地从外部更新：$\mathcal{G}_1^0 \to \mathcal{G}_2^0 \to \ldots \to \mathcal{G}_t^0 \to \ldots$。但是，根据引理 3，该过程对预测器 f^0 的更新总次数不会超过 $\frac{B^0 m^2}{L^0}$。由于在联合多重校准的 while 循环每次迭代中，预测器 f^0 至少会更新一次，因此这也限定了 while 循环的迭代次数。

现在，对于 while 循环的每次迭代，我们都会 m 次调用 BatchMulticalibration^V，并将其应用于 $\gamma^0 \in [1/m]$的所有识别函数 $V_{\gamma^0}^1$。再次根据引理 3，每次调用最多经过$\frac{B^1 m^2}{L^1}$次更新即可收敛。这直接源于假设4：对于所有 $P \in \mathcal{P}$（而不仅仅是满足 $\Gamma^0(P) = \gamma^0$的 P），$V_{\gamma^0}(\cdot, P)$ 都是L^1-利普希茨且单调递增的。直接的做法是针对 f_{t+1}^0的每个水平集各运行一次子程序，共运行 m 次，这将总计进行 $m \cdot \frac{B^1 m^2}{L^1} = \frac{B^1 m^3}{L^1}$ 次迭代。但实际上，所有这些 m 次调用都可以视为对 BatchMulticalibration^V 的一次调用：它使用一个识别函数 V_*^1更新 Γ^1 的预测器 f^1，并在每个水平集 $\{f_{t+1}^0 = \gamma^0\}$ 上将该函数定义为 $V_{\gamma^0}^1$（这是良定义的，因为这些水平集构成了域 $\mathcal{X}$的一个划分）。因此，在所有这些 BatchMulticalibration^{V}调用中，最多只需进行 $\frac{B^1 m^2}{L^1}$次更新。

综合以上观察，算法 3 因此将在至多 $mL^0 \cdot \frac{B^1 m^2}{L^1} = \frac{B^0 B^1 m^4}{L^0 L^1}$次更新后终止，对 f^0和 f^1进行更新，正如所述。

多重校准保证： 现在，我们证明在终止时输出的预测器 f_T^0、 f_T^1满足近似联合多重校准定义 17 中的条件。

根据 while 循环的停止条件，在终止时，对于所有 γ^0、 γ^1、 G，均有 $\mathbf{Pr}_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1)$ $x \in G]\left(V^0(\gamma^0, Y_{(\gamma^0,\gamma^1,G)})\right)^2 \le \frac{\alpha^0}{m}$，两边除以 $\mathbf{Pr}_{x\in\mathcal{X}}[x \in G, f_T^1(x) = \gamma^1]$ 后可得

$$\Pr_{x\in\mathcal{X}}[f_T^0(x) = \gamma^0 | x \in G, f_T^1(x) = \gamma^1]\left(V^0(\gamma^0, Y_{(\gamma^0,\gamma^1,G)})\right)^2 \le \frac{\alpha^0/m}{\Pr[x \in G, f_T^1(x) = \gamma^1]} \text{ for all } G \in \mathcal{G}, \gamma^1 \in \text{Range}_{f_T^1}. \tag{5}$$

对于每个 $G \in \mathcal{G}$和 $\gamma^1 \in$值域f_T^1，将该不等式对值域f_T^0中至多 m 个值 $\gamma^0 \in$求和，我们得到：

$$\sum_{\gamma^0\in\text{Range}_{f_T^0}} \Pr_{x\sim X}[f_T^0(x) = \gamma^0 | x \in G, f_T^1(x) = \gamma^1] \cdot \left(V^0(\gamma^0, Y_{(G,\gamma^0,\gamma^1)})\right)^2 \le \frac{\alpha^0}{\Pr_{x\in X}[x \in G, f_T^1(x) = \gamma^1]},$$

因此，预测器 f_T^0 满足定义 1中的联合多重校准条件（17）。

现在我们证明，对于 Γ^1 ，预测器 f_T^1 满足其联合多重校准条件（2）。根据构造，对于每个 $\gamma^0 \in [1/m]$，函数 f_T^1在区域 $\{x \in \mathcal{X}$中等于 f^1、$_T{}^{\gamma^0}$： $f_T^0(x) = \gamma^0\}$。由于 f^1、$_T{}^{\gamma^0}$由对 BatchMulticalibration 的相应调用输出，根据引理 3，这保证了对于每个 $G' \in \mathcal{G}^1$、 $_T{}^{\gamma^0} = \{G \cap \{x \in \mathcal{X}: f_T^0(x) = \gamma^0\}: G \in \mathcal{G}\}$，并且对于每个 $\gamma^1 \in [1/m]$，有 $\mathbf{Pr}_{x\in\mathcal{X}}[f_T^1(x) = \gamma^1$、 $x \in G'] \left(V_{\gamma^0}^1(\gamma^1, Y_{(\gamma^1,G')})\right)^2 \le \frac{\alpha^1}{m}$，从而对于所有 $G \in \mathcal{G}$，有 $\mathbf{Pr}_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1)$ $x \in G] \left(V_{\gamma^0}^1(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})\right)^2 \le \frac{\alpha^1}{m}$。

因此，对于所有 γ^0、 γ^1、 G ，我们有如下界：

$$\left|V_{\gamma^0}^1(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})\right| \le \sqrt{\frac{\alpha^1/m}{\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1), x \in G]}}. \tag{6}$$

但请注意，我们实际想要界定的是 $\left|V_{\Gamma^0(Y_{(G,\gamma^0,\gamma^1)})}^1(\gamma^1, Y_{(G,\gamma^0,\gamma^1)})\right|$在该集合上的真实识别函数的绝对值。这时我们可以利用假设3，它给出了 L_c关于 $V_{\gamma^0}^1(\cdot, \cdot)$作为 γ^0函数的利普希茨性；此外，假设2给出了 L_a^0关于 $V^0(\gamma^0, \cdot)$作为 γ^0函数的反利普希茨性：由此可得

$$\begin{aligned}
|V_{\Gamma^0(Y_{(G,\gamma^0,\gamma^1)})}^1(\gamma^1, Y_{(G,\gamma^0,\gamma^1)})| &\le |V_{\Gamma^0(Y_{(G,\gamma^0,\gamma^1)})}^1(\gamma^1, Y_{(G,\gamma^0,\gamma^1)}) - V_{\gamma^0}^1(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})| + |V_{\gamma^0}^1(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})| \\
&\le L_c|\gamma^0 - \Gamma^0\left(Y_{(G,\gamma^0,\gamma^1)}\right)| + |V_{\gamma^0}^1(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})| \\
&\le L_c L_a^0 |V^0(\gamma^0, Y_{(G,\gamma^0,\gamma^1)})| + |V_{\gamma^0}^1(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})| \\
&\le L_c L_a^0 \sqrt{\frac{\alpha^0/m}{\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1), x \in G]}} + \sqrt{\frac{\alpha^1/m}{\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1), x \in G]}},
\end{aligned}$$

其中第四步是代入不等式5 和6。

由此，对于所有 γ^0、 γ^1、 G，我们得到以下界

$$\left|V_{\Gamma^0(Y_{(G,\gamma^0,\gamma^1)})}^1(\gamma^1, Y_{(G,\gamma^0,\gamma^1)})\right| \le \frac{(L_c L_a^0\sqrt{\alpha^0} + \sqrt{\alpha^1})/\sqrt{m}}{\sqrt{\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1), x \in G]}},$$

对不等式两边平方后，我们得到：

$$\left(V_{\Gamma^0(Y_{(G,\gamma^0,\gamma^1)})}^1(\gamma^1, Y_{(G,\gamma^0,\gamma^1)})\right)^2 \le \frac{(L_c L_a^0\sqrt{\alpha^0} + \sqrt{\alpha^1})^2/m}{\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1), x \in G]} \quad \text{for all } \gamma^0, \gamma^1, G.$$

现在，将两边乘以 $\mathbf{Pr}_{x\in\mathcal{X}}[f_T^1(x) = \gamma^1 | x \in G, f_T^0(x) = \gamma^0]$，并注意到

$$(L_c L_a^0\sqrt{\alpha^0}+\sqrt{\alpha^1})^2 \le 2((L_c L_a^0)^2\alpha^0+\alpha^1) = 2((L_c L_a^0)^2\cdot 4(L^0)^2+4(L^1)^2)/m = 8((L^0 L_a^0 L_c)^2+(L^1)^2)/m = \alpha_*^1,$$

我们得到：

$$\Pr_{x\in\mathcal{X}}[f_T^1(x)=\gamma^1|x\in G,f_T^0(x)=\gamma^0]\left(V^1_{\Gamma^0\left(Y_{(G,\gamma^0,\gamma^1)}\right)}(\gamma^1,Y_{(G,\gamma^0,\gamma^1)})\right)^2\leq\frac{\alpha_*^1/m}{\Pr_{x\in\mathcal{X}}[f_T^0(x)=\gamma^0,x\in G]}\quad\text{for all }\gamma^0,\gamma^1,G.$$

对于每个 $G\in\mathcal{G}$ 和 $\gamma^0\in$ 值域f_T^0，将该不等式对所有至多 m 个值 $\gamma^1\in$ 值域f_T^1 求和，可得：

$$\sum_{\gamma^1\in\text{Range}_{f_T^1}}\Pr_{x\in\mathcal{X}}[f_T^1(x)=\gamma^1|x\in G,f_T^0(x)=\gamma^0]\left(V^1_{\Gamma^0\left(Y_{(G,\gamma^0,\gamma^1)}\right)}(\gamma^1,Y_{(G,\gamma^0,\gamma^1)})\right)^2\leq\frac{\alpha_*^1}{\Pr_{x\in\mathcal{X}}[f_T^0(x)=\gamma^0,x\in G]},$$

因此，预测器 f_T^1 满足其联合多重校准条件（2），该条件见定义 17，证明至此完成。 □