arXiv:2304.09167v1 [cs.LG] 18 Apr 2023

无一致收敛下的最优 PAC 界

Ishaq Aden-Ali* Yeshwanth Cherapanamjeri* Abhishek Shetty*
Nikita Zhivotovskiy†

2023年4月19日

摘要

在统计学习理论中，确定 VC 类的可实现二分类的样本复杂度，长期以来一直是一个开放问题。Simon [Sim15] 和 Hanneke [Han16a]的研究结果在这一设定下建立了紧确的上界。然而，他们的论证依赖一致收敛原理，这限制了其在更一般的学习设定中的适用性，例如多分类。在本文中，我们通过提供一个超越一致收敛论证局限性的框架，给出了最优的高概率风险界，从而解决这一问题。

我们的框架将置换不变预测器的留一法误差转化为高概率风险界。作为应用，通过改造 Haussler、Littlestone 和 Warmuth [HLW94], 的单包含图算法，我们提出了一种能够为二分类达到最优 PAC 界的算法。具体而言，我们的结果表明，某些单包含图算法的聚合是最优的，从而回应了 Warmuth [War04] 提出的一个经典问题的变体。

我们进一步在三种一致收敛可证明地并非最优的情形下实例化了我们的框架。对于多分类，我们证明了一个最优风险界，其随该类的单包含超图密度变化，从而解决了 Daniely 和 Shalev- Shwartz [DS14]的分析并非最优的问题。对于部分假设分类，我们确定了最优样本复杂度界，解决了 Alon、Hanneke、Holzman 和 Moran [AHHM22]提出的一个问题。对于采用绝对损失的可实现有界回归，我们推导出一个最优风险界，该界基于尺度敏感维度的一个修正版，从而改进了 Bartlett 和 Long [BL98]的结果。由于我们的风险界采用了更小的复杂度度量，因此我们得到的速率优于基于标准一致收敛的结果。

1 引言

预测的统计复杂度研究是统计学习理论中的一个核心问题。在最简单的可实现预测情形下，预测任务如下：给定实例空间$\mathcal{X}$、标签空间 $\mathcal{Y}$和假设类 $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$，我们可以访问 n个独立同分布（i.i.d.）的带标签训练样本点$S = ((X_1, Y_1) \ldots, (X_n, Y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$，这些样本点来自分布 P，并且已知存在一个目标假设 $f^* \in \mathcal{H}$，能够完美地为数据标注。我们将 S 称为训练样本。目标$\widehat{}$是利用训练样本 S设计一个预测器，记为 $f(\cdot; S)$，使其在从相同分布中抽取的留出样本上的预测误差（我们称之为风险）尽可能小，并且该结果以关于样本的高概率成立。具体而言，我们希望最小化

$$\mathop{\mathbb{E}}_{(X,Y)\sim P}\left[\ell(\widehat{f}(X; S), Y)\right]$$

对于合适的损失函数 $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0, 1]$，以至少 $1 - \delta$ 的概率，在 S 上。

*加州大学伯克利分校电气工程与计算机科学系。邮箱：adenali@berkeley.edu, yesh- wanth@berkeley.edu, shetty @berkeley.edu
†加州大学伯克利分校统计学系。邮箱：zhivotovskiy@berkeley.edu

大量研究致力于刻画最优可达风险如何随数据点数量而变化 n、失败概率 δ以及假设类的复杂度 $\mathcal{H}$。即便是在最简单的二分类设定中，其中 $\mathcal{Y}=\{0,1\}$，$\ell(\widehat{y},\ y)=\mathbf{1}\{\widehat{y}\neq y\}$，且函数类的复杂度由其 VC维刻画 d，这仍然是一个具有挑战性的开放问题。在这一特定设定下，早期的奠基性研究 [VC68a、VC74、BEHW89] 证明，经验风险最小化器（ERM），即任何在训练样本上使损失最小化的预测器，都能达到如下风险界[1]

$$\mathop{\mathbb{E}}_{(X,Y)\sim P}\left[\ell(\widehat{f}_{\mathrm{ERM}}(X;S),Y)\right]\leqslant C\left(\frac{d}{n}\log\left(\frac{n}{d}\right)+\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right).$$

这种形式的高概率风险界通常称为可能近似正确（PAC）界，这一概念由 Valiant 在其经典论文中提出 [Val84]。

与此同时，关于直推式设定及相关学习的预测模型[HLW94], 的一系列互补研究，推动了具有优异留一法（LOO）性能的预测器的发展，并对分布设定有直接影响。在这类设定的典型例子中，给定一个固定的可实现样本 $S=((x_i,\ y_i))_{i=1}^n$，学习器可以访问其中随机选取的 y_i 个样本的标签 $n-1$，任务是对剩余的点进行准确预测。记 $S^{-i}=((x_j,\ y_j))_{j\neq i}$，^目标是设计一个预测器 f，使其留一法误差最小

$$\sum_{i=1}^{n}\ell\left(\widehat{f}\left(x_i;S^{-i}\right),y_i\right).$$

对于二分类，Haussler、Littlestone 和 Warmuth 著名的单包含图策略[HLW88，HLW94] 实现的留一法误差至多为 d。[2] 通过可交换性论证，可以得到一个预测器 $\widehat{f}_{\mathrm{OIG}}$，其界为

$$\mathop{\mathbb{E}}_{\substack{(X,Y)\sim P\\ S\sim P^n}}\left[\ell\left(\widehat{f}_{\mathrm{OIG}}(X;S),Y\right)\right]\leqslant\frac{d}{n}$$

对于分布设定 在期望意义下 在训练样本上。值得注意的是，这种方法消除了经验风险最小化界中通常存在的额外对数因子，从而引出了这样的猜想：即使在 PAC 设置中，对数因子也可能不是必需的。此外，这表明单包含图预测器可能达到最优 [War04]。此后，这些预测器被应用于各种一致收敛这一分析经验风险最小化预测器所必需的关键性质表现欠佳或不成立的场景。其中包括多分类学习 [RBR09、SS10、DS14], 部分假设学习 [AHHM22], 以及有界回归 [BL98]。

在一项突破性成果中，Hanneke [Han16a], 在 Simon [Sim15], 的工作基础上，构造了一个^最优 PAC预测器 f_{OPT}，满足

$$\mathop{\mathbb{E}}_{(X,Y)\sim P}\left[\ell(\widehat{f}_{\mathrm{OPT}}(X;S),Y)\right]\leqslant C\left(\frac{d}{n}+\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right),$$

从而证实了猜想中的最优速率。值得注意的是，该分析并未利用留一法预测器在期望意义下的最优表现，而是对次优的经验风险最小化预测器进行了谨慎的聚合。不幸的是，对经验风险最小化的使用阻碍了该方法应用于一致收敛不成立的场景；然而，正如留一法预测器的成功所表明的那样，在这些场景中仍然可能实现准确的预测。

在此背景下，我们的主要贡献（定理 2.1）是一种通用技术，可将一大类具有最优留一法性能且满足强有力的期望意义下保证的预测器，转化为具有最优 PAC 界的预测器。从概念层面而言，我们的工作回应了

[1]在整篇论文中，我们使用 log 表示自然对数。[2]即使在首项常数意义下，他们的界也是最优的[LLS01]。

Warmuth [War04] 提出的猜想：最优速率可以通过单包含图算法实现。尽管严格意义上的猜想并不成立，因为存在一些单包含策略，它们在期望意义下是最优的，但在 PAC模型中表现不佳 [ACSZ22], ，我们证明，只需对基于训练数据前缀构造的单包含策略进行简单聚合，即可恢复 PAC 设置下的最优性能。我们的主要结果 定理 2.1 对此进行了形式化表述。

更具体地说，由于我们依赖最优留一法预测器的存在，因此可以完全绕过一致收敛，从而将应用扩展到一系列一致收敛可能^不成立的设定。考虑一般的可实现预测设置，其中存在一个预测器 f_{LOO}，满足

$$\sum_{i=1}^{n} \ell\left(\widehat{f}_{\mathrm{LOO}}(x_i; S^{-i}), y_i\right) \leqslant M_n$$

^对于任意可实现样本 S。现在，我们的框架可以得到一个预测器 f_{PAC}，满足

$$\mathop{\mathbb{E}}_{(X,Y)\sim P}\left[\ell\left(\widehat{f}_{\mathrm{PAC}}\left(X; S\right), Y\right)\right] \leqslant C\left(\frac{M_n}{n} + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right).$$

对于分布设定。

在若干值得关注的情形中，存在留一法预测器，其第一项的期望性能达到最优。我们的结果表明，可以将这些预测器扩展到 PAC 设定，且只需在 $\log(1/\delta)$ 上增加加性项；已知这也是最优的。下面我们将展示本结果的一些应用。

1. 多分类：存在留一法预测器，其错误率随相关单包含超图的图密度 $\mu_{\mathcal{H}}(n)$ 而变化，从而实现最优的期望性能。我们的结果立即给出该设定下的最优 PAC 界（定理 2.2）。此外，在二分类的背景下，我们通过简化的分析重新得到了 Hanneke 的最优 PAC 界 [Han16a]，并将常数因子提升了数个数量级。

2. 部分假设学习： 在此， M_n 的最优界随部分假设类的 VC 维变化；该假设类的 VC 维是对二元设定中 VC 维的一种推广。我们的结果给出了最优 PAC 界，解决了文献中的一个开放问题 [AHHM22]。参见推论 2.10。

3. 有界回归： 在有界 ℓ_1 回归设定中，我们得到了 PAC 设置下的首个预测器，其误差取决于尺度敏感复杂度度量 $\mathrm{V}_{\mathcal{H}}(\gamma)$；在某些情况下，相较于经典的一致收敛论证所得到的结果，这是一个显著的改进。经典论证依赖的则是*fat-shattering* 维度 $\mathrm{P}_{\mathcal{H}}(\gamma)$。参见第 2.5 节 和定理 2.3。

本文其余部分安排如下。本文的主要结果，包括对我们框架的正式描述及其在若干感兴趣设定中的应用，将在第 2 节中讨论。我们将在第3节介绍相关工作。随后，我们将在第 4 节证明主要结果。

2 主要结果

2.1 符号说明

我们使用 $\mathcal{X}$ 表示我们的实例空间，使用 $\mathcal{Y}$ 表示标签空间，有时也使用 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$。令 $\mathcal{U} = \bigcup_{n=1}^{\infty} \mathcal{Z}^n$ 表示所有可能的可观测训练样本的集合。预测器 $\widehat{f}\colon \mathcal{X} \times \mathcal{U} \to \mathcal{Y}$以测试点和训练样本作为输入，并输出一个预测。对于训练样本 $S = ((X_1, Y_1) \ldots, (X_n, Y_n))$，我们将使用 S^{-i} 表示移除了第 i^{th} 个数据点的训练样本，即对于所有 $i \in [n]$， $S^{-i} = ((X_1, Y_1) \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}) \ldots, (X_n, Y_n))$。对于 $j \in [n]$，我们将使用$S_{\leqslant j}$ 表示训练样本 $((X_1, Y_1) \ldots, (X_j, Y_j))$。有时，我们还会方便地将训练样本 S中的第 i^{th} 个点写作$Z_i = (X_i, Y_i)$。在整篇论文中，我们使用大写字母 X、Y 和 Z 表示随机变量，使用小写字母 x、 y 和 z 表示它们的实现值。

在学习设定中，算法可以访问来自未知分布 P 的样本，这些样本取自 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$。学习算法的目标是生成一个假设 f，使其在 P 下具有较低的风险，其中 f 相对于损失函数 ℓ: $\mathcal{Y} \times \mathcal{Y} \to [0,1]$，我们将风险定义为 $\mathrm{R}_P(f) := \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)]$。在本文中，我们始终假设未知分布 P 可由 $\mathcal{H}$实现。这意味着存在某个 $f^* \in \mathcal{H}$，使得对于 $Y = f^*(X)$ 几乎处处成立。类似地，如果存在某个 $f^* \in \mathcal{H}$ 使得对于所有 $(x_i, y_i) \in S$，$f^*(x_i) = y_i$，则称训练样本 $S = ((x_1, y_1) \ldots, (x_n, y_n))$ 可由 $\mathcal{H}$ 实现（或简称为可实现）。给定训练样本 $S = ((x_1, y_1) \ldots, (x_n, y_n))$，我们将方便地定义 $\mathcal{X}_S \subseteq \mathcal{X}$ 为训练样本 S中唯一的 x_i 项的集合。

2.2 假设

我们的主要结果适用于一大类满足两个自然假设的预测器。主要假设涉及留一法误差。如第1节所述，首先使用在训练样本中移除该点后的样本上训练得到的预测器，逐点对训练样本中的每个点进行评估。将训练样本中所有点对应的误差相加，即得到留一法误差。我们假设预测器的留一法误差有界。

假设 2.1。对于任意可实现样本，预测器 f 的留一法误差有界。形式化地，对于任意可实现的 $S = ((x_i, y_i))_{i=1}^n$，有

$$\sum_{i=1}^{n} \ell(\widehat{f}(x_i; S^{-i}), y_i) \leqslant M_n.$$

此外，我们还假设 M_n 关于样本量 n是单调的，即对于所有 $n \in \mathbb{N}$，都有 $M_n \leqslant M_{n+1}$。[3]

注释2.2。注意，对一个从 S 中抽取的大小为 n 的样本取平均后， P，假设 2.1 中的留一法误差可转化为如下期望风险界:

$$\mathop{\mathbb{E}}_{S\sim P^n} \mathrm{R}_P(\widehat{f}(\cdot; S)) \leqslant \frac{M_{n+1}}{n+1}. \tag{1}$$

第二个假设要求我们的分类器关于训练数据是对称的。在批量设定中，这一点很直观：训练数据是从某个分布中独立同分布抽取的，而且许多理论和实际场景中的算法都满足这一要求。我们将其正式表述为以下假设。

假设 2.3。预测器 f 关于训练数据是对称的：对于所有 $n \in \mathbb{N}$，以及任意排列 π,$S \in \mathcal{Z}^n$,$x \in \mathcal{X}$，都有

$$\widehat{f}(x; S) = \widehat{f}(x; S_\pi),$$

其中 S_π 是经 π排列后的样本 S。

考虑这些假设的主要动机是单包含图算法，该算法是对称的，并且如前所述，在各种设定下具有最优的留一法误差。例如，在假设类具有有限 VC维的可实现二分类问题中，该算法的 M_n 与样本量无关。我们将在后文讨论其应用，并在附录A中详细介绍单包含图算法。

2.3 主要结果

我们给出了主要结果，该结果涉及通过在训练数据的前缀（$S_{\leqslant t}$）上训练 $\widehat{f}$ 预测器 f所生成的假设序列的性能。当 f 是对称的，且具有有界的留一误差时，我们的结果断言，该序列后缀（$f(\cdot;\ S_{\leqslant t})$）中预测器的平均风险$_{t=1}^n$

[3]这在很大程度上不失一般性，因为在大多数自然情形下，我们可以将 M_n 替换为 $\max_{i\leqslant n} M_i$，以强制满足单调性。

以高概率被 M_n所界定。为简便起见，在本文其余部分中，我们假设$n/4$ 是一个整数。

定理 2.1。 固定一个假设类 $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$、损失函数 $\ell\colon \mathcal{Y}\times\mathcal{Y} \to [0,1]$以及一个预测器 $\widehat{f}\colon \mathcal{X}\times\mathcal{U} \to \mathcal{Y}$满足 假设 2.1 和 2.3。那么，对于任意定义在 P 上的可实现分布 $\mathcal{X} \times \mathcal{Y}$ 和置信参数 $\delta \in (0, 1)$，给定一个训练样本 $S \sim P^n$，我们有

$$\frac{4}{3n} \cdot \sum_{t=n/4}^{n-1} \mathrm{R}_P(\widehat{f}(\cdot; S_{\leqslant t})) \leqslant 4.82 \left(\frac{M_n}{n} + \frac{1}{n} \log\left(\frac{2}{\delta}\right) \right),$$

关于 S的随机性，以至少 $1-\delta$ 的概率成立。

我们的证明技巧基于在线到批量转换，它将任何在线学习算法——该算法在有限样本上犯下 M 次错误——转换为一种在统计设定中利用整个样本的批量算法。我们工作的一个关键技术方面是证明，即使 M 是依赖于训练样本具体实现的随机变量，这种技巧仍然适用。

我们证明中的另一个技术要素是使用后缀平均。为说明这一点，考虑一个算法：处理大小为 t的独立同分布可实现样本后，它在新抽取的点上出错的概率至多为 d/t，其中 d 例如可以是 VC 维数。那么，大小为 n 的样本上的预期错误总数至多为 $\sum_{t=1}^{n} d/t = O(d \log\ n)$，这无法达到所需的准确度。有趣的是，如果只使用样本的后缀，我们有 $\sum_{t=n/2}^{n} d/t = O(d)$，从而有效地消除了对数因子。通过鞅论证，我们基本上将最后的计算转化为一个高概率风险界，其尺度为

$$O\left(\frac{d}{n} + \frac{1}{n} \log\left(\frac{1}{\delta}\right) \right).$$

值得注意的是，上述两项观察几乎在 20 世纪 80 年代末同时出现。Haussler、Littlestone 和 Warmuth [HLW88] 发现，后缀平均不会在序贯错误中引入对数因子，而 Littlestone 证明了对于错误数量确定的算法，可以实现最优的在线到批量转换 [Lit89]。 定理 2.1 将这两项洞见联系起来。

2.4 分类

定理 2.1 的第一个重要应用是在多分类设置中。在这一设置下，我们假设标签空间 $\mathcal{Y}$ 是有限的，并将损失函数取为 $0/1$ 损失。在本节中，我们使用 $\mathrm{err}_P(f)$ 表示 f的预测风险。在陈述主要结果之前，我们先介绍单包含超图密度，这是我们所使用的核心复杂度度量。为此，我们先介绍一些基本定义和符号说明。

设 $G = (V, E)$ 是一个超图。给定顶点子集 $U \subseteq V$，我们将诱导超图$G[U]$定义为顶点集为 U且边集为 $E' = \{e \cap U \text{ 的超图}\colon\ e \in E,\ |e \cap U| \geqslant 1\}$。对于超图 $G = (V, E)$，我们将 G 的密度定义为

$$\mathrm{Dens}(G) = \frac{1}{|V|} \sum_{e \in E} (|e| - 1).$$

我们将 G 的最大密度定义为

$$\mu(G) = \max_{U \subseteq V} \mathrm{Dens}(G[U]).$$

换句话说， $\mu(G)$ 是 G 的诱导超图中最大的密度。我们还将定义超图 $G = (V, E)$ 的平均度为

$$\mathrm{Avg}(G) = \frac{1}{|V|} \sum_{e \in E : |e| > 1} |e|.$$

给定一个假设类 $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ 和实例空间 $\mathcal{X}$ 的子集 $U = \{x_1, \ldots, x_n\}$，我们将 $\mathcal{H}$ 在 U 上的投影定义为 $\mathcal{H}|_U = \{(f(x_1) \ldots, f(x_n)) \colon f \in \mathcal{H}\}$。对于投影到一个含 n 个元素的集合 U 上的投影 $\mathcal{H}|_U$，我们通常记作 $f(i) = f(x_i)$，其中 $i \in [n]$。

现在我们可以定义单包含超图。Haussler、Littlestone 和 Warmuth 首先在二分类设定下研究了单包含超图 [HLW94],[4] ，随后 Rubinstein、Bartlett 和 Rubinstein 将其推广到多分类设定[RBR09]。Daniely 和 Shalev-Shwartz [DS14] 开始研究这些对象在无限 $\mathcal{Y}$。

定义 2.4（单包含超图）。固定一个假设类 $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ 和定义域 $U \subseteq \mathcal{X}$ 的一个 n 元子集。单包含超图 $G(\mathcal{H}|_U) = (V, E)$ 的顶点集为 $V \colon = \mathcal{H}|_U$。对于每个$i \in [n]$ 和 $f \colon [n] \setminus \{i\} \to \mathcal{Y}$，令 $e_{i,\, f}$ 为所有满足以下条件的 $f' \in \mathcal{H}|_U$ 的集合：对于所有 $j \in [n] \setminus \{i\}$，$f'(j) = f(j)$。单包含图 $G(\mathcal{H}|_U)$ 的边集定义为

$$E = \{e_{i,f} : i \in [n],\ f : [n] \setminus \{i\} \to \mathcal{Y},\ |e_{i,f}| > 0\}.$$

换言之，对于除一个坐标 $i \in [n]$ 外其他坐标都相同的所有顶点（投影后的假设），我们创建一条超边，并对每个坐标都这样做。

The 假设类 $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ 的单包含超图密度定义如下。[5]

定义 2.5 ($\mathcal{H}$ [DS14]的单包含超图密度)。固定一个假设类 $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$。我们将该类 的单包含超图密度 $\mathcal{H}$ 在样本量 n 下定义为

$$\mu_{\mathcal{H}}(n) = \max_{S \in \mathcal{Z}^n} \mu(G(\mathcal{H}|_{\mathcal{X}_S})).$$

换言之， $\mu_{\mathcal{H}}(n)$ 是由 $\mathcal{H}$ 和大小为 n 的训练样本 S （可能包含重复样本）构成的所有单包含图中最大密度的最大值。

Haussler、Littlestone 和 Warmuth [HLW94] 为二分类设定定义了单包含图算法，而 Rubinstein、Bartlett 和 Rubinstein [RBR09] 则将该算法推广到多分类设定，并将其称为单包含超图算法。单包含超图算法满足假设 2.1 和2.3，并且是我们在结果中使用的预测器。不过，证明本文的新结果并不需要该算法的具体细节，但我们仍在附录A 中对其进行了介绍，供感兴趣的读者参阅。最后，我们注意到，当 $\mathcal{Y} = \{0, 1\}$时，单包含超图算法会退化为原始的单包含图算法。

我们的多分类主要 PAC 风险界随 $\mu_{\mathcal{H}}(n)$ 变化。

定理 2.2。固定一个假设类 $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$。存在一个预测器 $\widehat{f} \colon \mathcal{X} \times \mathcal{U} \to \mathcal{Y}$，对于任意 $\delta \in (0, 1)$，以及从任意可实现分布 P 中抽取的 $S \sim P^n$，均满足

$$\mathrm{err}_P(\widehat{f}(\cdot; S)) \leqslant 9.64 \left(\frac{\lceil \mu_{\mathcal{H}}(n) \rceil}{n} + \frac{1}{n} \log\left(\frac{2}{\delta}\right) \right),$$

以至少 $1 - \delta$ 的概率（关于 S 的随机性）。

我们将该结果的证明收录于附录A中。该界根据 $\mu_{\mathcal{H}}$（n）对预测风险给出了精确刻画，从而将 Daniely 和 Shalev-Shwartz 的最优期望风险界提升为 [DS14] 标准PAC设定下的结果。特别地，我们的最优 PAC 界没有表现出对类别数的显式依赖，而基于一致收敛得到的经验风险最小化分类器的界则并非如此。更多细节请参见第3节。

[4]在二元设定中，单包含超图就是一个图。[5]我们对这一复杂度度量的定义与 Daniely 和 Shalev- Shwartz 最初提出的定义略有不同[DS14]。在他们的论文中，他们使用 *average degree* 而非 *edge density*来定义这一复杂度度量，两者相差至多 2 倍。他们的定义也更为一般，能够处理无限的 $\mathcal{Y}$。

关于相对于序列 $(\mu_{\mathcal{H}}(n))_{n\in\mathbb{N}}$ 的可学习性刻画，由于其对 n的隐式依赖，略显令人不满意。这促使 Daniely 和 Shalev-Shwartz 研究是否存在一个单一数值，能够在这些风险界中替代这一复杂度度量序列。他们提出了维度的概念，后来被命名为 Daniely-Shalev-Shwartz（DS）维度[BCD+22]。[6]

定义 2.6 (DS 维度 [DS14])。固定一个假设类 $\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{X}}$。Daniely-Shalev-Shwartz 维度 $\mathcal{H}$，记作 d_{DS}，是满足下列条件的最大整数 d：存在一个子集 $\mathcal{H}'\subseteq\mathcal{H}$ 满足

$$\max_{S\in\mathcal{X}^d}\mathrm{Avg}(G(\mathcal{H}'|_{\mathcal{X}_S}))=d.$$

为了更好地理解这个定义，回顾一下二元设定中的 VC 维数定义。固定一个二元假设类 $\mathcal{H}\subseteq\{0,1\}^{\mathcal{X}}$。我们称一个含有 n 个元素的集合 $U\subseteq\mathcal{X}$被 $\mathcal{H}$打散，当且仅当 $\mathcal{H}|_U=\{0,1\}^n$。$\mathcal{H}$ 的 VC 维数是满足以下条件的最大整数 d：存在 $\mathcal{X}$ 的一个含有 d 个元素的子集，且该子集被 $\mathcal{H}$打散。当 $\mathcal{Y}=\{0,1\}$ 时，DS 维度等价于 VC 维数。Daniely 和 Shalev-Shwartz 提出了以下猜想，该猜想将 DS 维度与序列 $(\mu_{\mathcal{H}}(n))_{n\in\mathbb{N}}$ 联系起来。

猜想 2.7 ([DS14])。固定一个假设类 $\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{X}}$，其 DS 维度为 d_{DS}。存在一个绝对常数 $c>0$，使得对于任意 $n\geqslant d_{\mathrm{DS}}$，都有

$$\mu_{\mathcal{H}}(n)\leqslant c\cdot d_{\mathrm{DS}}.$$

若这一组合猜想得到肯定解，则通过 定理 2.2，可推出 PAC 界

$$\mathrm{err}_P(\widehat{f})=O\left(\frac{d_{\mathrm{DS}}}{n}+\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right),$$

这将显著改进 Brukhim、Carmon、Dinur、Moran 和 Yehuday- off 最近的工作 [BCD+22]，他们证明了如下界[7]

$$\widetilde{O}\left(\frac{d_{\mathrm{DS}}^{3/2}}{n}+\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right).$$

提出 猜想 2.7 的主要动机，是 Haussler、Littlestone 和 Warmuth 的一个优雅结果[HLW94] (另见 [Hau95])，该结果表明，对于每个二元假设类 $\mathcal{H}$，若其 VC 维数为 d，则单包含超图的密度至多为 d。

定理 2.8 ([HLW94，Hau95])。固定一个假设类 $\mathcal{H}\subseteq\{0,1\}^{\mathcal{X}}$，其 VC 维数为 d。对于每个$n\geqslant 1$，都有

$$\mu_{\mathcal{H}}(n)\leqslant d.$$

事实上，在二元设定下结合 定理 2.2 和 定理 2.8，我们可以得到一个紧确的 PAC 界，它仅随一个参数——VC 维数——变化。

推论 2.9。固定一个 VC 维数为 d 的假设类 $\mathcal{H}\subseteq\{0,1\}^{\mathcal{X}}$。存在一个预测器 $\widehat{f}:\ \mathcal{X}\times\mathcal{U}\to\{0,1\}$，对于任意 $\delta\in\in$ (0, 1)，以及从任意可实现分布 P 中采样得到的 $S\sim P^n$，均满足

$$\mathrm{err}_P(\widehat{f}(\cdot;S))\leqslant 9.64\left(\frac{d}{n}+\frac{1}{n}\log\left(\frac{2}{\delta}\right)\right),$$

以至少 $1-\delta$ 的概率（关于 S 的随机性）。

[6]此维度最初定义时允许 $\mathcal{Y}$ 为无限。[7]这里，波浪号记号隐藏了参数 n 和 d_{DS} 中的多对数因子。

推论 2.9 重新得到了 Hanneke 首先证明的二分类最优风险上界 [Han16a]。最近，Larsen [Lar22] 证明，自然的装袋启发式方法的一种实现也能达到最优风险界。我们对这一最优上界的证明，比他们两人的证明都简单得多。此外，Hanneke 和 Larsen 都指出，他们上界中的常数因子非常大，并明确提出这些常数能否缩小。我们的新分析将常数因子缩小了几个数量级。

此外，我们的结果也可以视为对 Warmuth 提出的问题 [War04] 的部分回答。他曾询问单包含图算法能否达到最优 PAC 风险界。近期工作[ACSZ22]表明，这一猜想严格来说并不成立。也就是说，在 PAC 设置中，单包含图算法可能具有较大的风险界。我们的结果表明，尽管单包含预测器本身并非最优，但对这类预测器进行简单聚合便可达到最优 PAC 界。

我们将结果应用于的最后一种分类设置是部分假设分类。在此设置中，标签集为 $\mathcal{Y} = \{0, 1, \star\}$，其中 $\star$ 标签被解释为“我不知道”响应。损失函数 ℓ 仍然是 $0/1$ 损失。我们可以将二元假设类所用的打散和 VC 维数概念扩展到部分假设类。若一个部分假设类 $\mathcal{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$ 满足 $\{0, 1\}^n \subseteq \mathcal{H}|_U$，则称其打散了一个 n 元集合 $U = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$。部分假设类 $\mathcal{H}$ 的 VC 维数现定义为最大的整数 d，使得存在一个 d 元子集 $\mathcal{X}$ 被 $\mathcal{H}$ 打散。[8] 若分布 P 定义在 $\mathcal{X} \times \{0, 1\}$ 上（重要的是，不是定义在 $\mathcal{X} \times \{0, 1, \star\}$ 上），且存在一个目标假设 $f^* \in \mathcal{H}$ 使得 $\mathrm{err}_P(f^*) = 0$，则称该分布可由 $\mathcal{H}$ 实现。请注意，这一定义与我们上文使用的可实现性定义略有不同。

在这种设定下，我们使用的预测器是对单包含超图算法的简单修改，以排除 $\star$ 标签。与二分类情形（定理 2.8）一样，可以证明，单包含超图密度 $\mu_{\mathcal{H}}(n)$ 不超过 $\mathcal{H}$ 的 VC 维数 d，其中 $\mathcal{H}$。由此可得以下风险界，它蕴含的样本复杂度上界与 [AHHM22] 中的下界仅相差一个普适常数因子。这解答了 [AHHM22] 提出的一个问题。[9] 我们在 附录 A.1 中进一步阐述对单包含超图的这一简单修改。

推论 2.10。给定一个部分假设类 $\mathcal{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$，其 VC 维数为 d。存在一个预测器 $\wedge f\colon \mathcal{X} \times \mathcal{U} \to \{0, 1, \star\}$，对于任意 $\delta \in (0, 1)$，以及从可实现分布 P 中抽取的 $S \sim P^n$，均满足

$$\mathrm{err}_P(\widehat{f}(\cdot; S)) \leqslant 9.64 \left(\frac{d}{n} + \frac{1}{n} \log\left(\frac{2}{\delta} \right) \right),$$

关于 S 的随机性，以至少 $1 - \delta$ 的概率。

除了回答有关最优速率的问题外，我们的技术还可以视为对这些分类概念的一种统一视角。相比之下，二分类情形下以往的算法和最优性证明 [Han16a、Lar22] 都是基于经验风险最小化，而在更一般的设定中，可以证明这种方法会失效。由于我们的主要结果具有普适性，本节列出的潜在应用并不全面。例如，可以将推论 2.10扩展到多分类部分假设设定。这将改进 [KVK22] 中的一些近期界。

2.5 有界回归

从一个离散集合 $\mathcal{Y}$出发，我们现在探讨具有 $\mathcal{Y} = [0, 1]$ 和绝对损失的有界回归$\ell(\widehat{y}, y) = |\widehat{y} - y|$。本节的关键参考文献是 Bartlett 和 Long 的研究 [BL98], ，他们指出单包含图算法适用于这一情形。我们先回顾一些标准

[8]注意，一个类别的复杂度并不直接取决于 $\star$ 标签，这使得该设定有别于仅包含三个标签的多分类设定实例。[9]AHHM[中使用的可实现学习定义，比我们在此采用的标准概念更一般。这个差异对我们的分析没有影响；我们的界在他们更一般的定义下同样成立。

定义。设 $\mathcal{X}$ 为一个实例空间，标签空间为 $\mathcal{Y}=[0,1]$，且 $\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{X}}$ 表示一类从 $\mathcal{X}$ 映射到 $\mathcal{Y}$的实值函数。设 $\gamma\geqslant 0$ 为一个间隔参数。定义 $\mathrm{V}_{\mathcal{H}}(\gamma)$ 为满足以下条件的最大整数：存在 $\tau\in[0,1]$ 和一个子集 $S\subseteq X$，使得 $|S|=\mathrm{V}_{\mathcal{H}}(\gamma)$，并且对于任意 $A\subseteq S$，都存在一个$f_A\in\mathcal{H}$ 满足 $f_A(x)\geqslant\tau+\gamma$ 对所有 $x\in A$ 成立，且 $f_A(x)\leqslant\tau-\gamma$ 对 $x\in S\setminus A$成立。这一复杂度度量是 Vapnik 和 Chervonenkis 针对实值函数提出的原始打散概念 [VC71] 的尺度敏感版本，在 [ABCH97] 中有明确的定义，称为 V_γ-维度。关于这一复杂度度量的详细讨论，请参阅 [KS23]。

下面的引理构造了一个使用单包含图算法的实值预测器。遗憾的是，作者无法验证 [BL98] 中定理 1 证明的正确性，该文最初提出了这种构造。我们改为设计了另一种预测器，其期望风险界甚至比 [BL98] 中声称的更紧。引理的证明见 附录 B.2。

引理 2.1。 固定一个假设类 $\mathcal{H}\subseteq[0,1]^{\mathcal{X}}$。存在一个预测器 $\widehat{f}\colon\mathcal{X}\times\mathcal{U}\to\mathcal{Y}$，对于任意可实现样本 $S=(x_i,y_i)_{i=1}^n$，均满足 假设 2.1 和 *2.3*，其留一法误差界为

$$M_n\leqslant(n+1)\gamma+\mathrm{V}_{\mathcal{H}}(\gamma).$$

本节的主要结果是以下 PAC 界。

定理 2.3。固定假设类 $\mathcal{H}\subseteq[0,1]^{\mathcal{X}}$ 和间隔 $\gamma\in(0,1)$。存在一个预测器 $\widehat{f}\colon\mathcal{X}\times\mathcal{U}\to[0,1]$，对于任意 $\delta\in(0,1)$，以及从任意可实现分布 P 中抽取的 $S\sim P^n$ 样本，均满足

$$\mathop{\mathbb{E}}_{(X,Y)\sim P}\left[|\widehat{f}(X;S)-Y|\right]\leqslant 4.82\left(\gamma+\frac{\mathrm{V}_{\mathcal{H}}(\gamma)}{n}+\frac{1}{n}\log\left(\frac{2e}{\delta}\right)\right),$$

以至少 $1-\delta$ 的概率（关于 S 的随机性）。

定理 2.3 给出了如下意义下的最优界：Bartlett 和 Long [BL98] 证明了存在某个假设类 $\mathcal{H}$，对于该假设类，定理 2.3 的期望版本在相差一个乘法常数因子的意义下是最优的。

2.5.1 与一致收敛结果的关系

实值函数集的一致收敛刻画一直是经验过程理论中的核心问题，其动机来源于统计学习理论和凸几何中的问题。Vapnik 和 Chervonenkis [VC71] 以复杂度度量 V 的一种与尺度无关的版本为基础，开启了对实值函数一致收敛的研究$_{\mathcal{H}}(\gamma)$。Alon 等人 [ABCH97]研究了复杂度度量 $\mathrm{V}_{\mathcal{H}}(\gamma)$，并证明对于所有 $\gamma>0$，其有限性是函数类成为一致 Glivenko-Cantelli（GC）类的必要且充分条件。类似结论也出现在 [Tal96、Tal03] 中。另请参见 [VC82]中的相关结果。

在一致 GC 性质的背景下，复杂度度量 $\mathrm{V}_{\mathcal{H}}(\gamma)$ 的主要问题在于，它无法很好地刻画相应函数类的覆盖数。根据 Talagrand [Tal03]，在定量界中使用这种复杂度度量会导致“精确速率下精度的损失是灾难性的”。事实上，大多数文献采用另一种复杂度度量，我们称之为 P_γ-维度。设 $\mathcal{H}$ 是定义域集合 $\mathcal{X}$ 上的一类取值于 $[0,1]$ 的函数，γ 是一个间隔。若存在函数 $s:S\to[0,1]$，使得对每个 $A\subseteq S$，都存在某个 $f_A\in\mathcal{H}$ 满足：对每个 $x\in S\setminus A$，$f_A(x)\leqslant s(x)-\gamma$；并且对每个$x\in A$，$f_A(x)\geqslant s(x)+\gamma$，则称 $\mathcal{H}$ P_γ将集合 $S\subseteq\mathcal{X}$ 打散。令 $\mathcal{H}$ 的 P_γ- 维度记为 $\mathrm{P}_{\mathcal{H}}(\gamma)$，它是满足以下条件的最大整数：大小为 $\mathrm{P}_{\mathcal{H}}(\gamma)$ 的 $S\subseteq X$ 被 $\mathcal{H}$以 P_γ打散。这种复杂度度量通常称为脂肪打散维数 [KS94、BL98、AB99]。复杂度度量 $\mathrm{V}_{\mathcal{H}}(\gamma)$ 与 $\mathrm{P}_{\mathcal{H}}(\gamma)$ 的主要区别在于打散发生的层级。对于 $\mathrm{V}_{\mathcal{H}}(\gamma)$，打散发生在单一

层级，记为 τ。另一方面，在第二个定义中引入函数 s，使得可以在不同尺度 τ_i上打散不同的点 $x \in S$。这些复杂度度量之间的直接关系可通过以下结果表述。

引理 2.11 ([ABCH97])。 对于任意假设类 $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ 和 $\gamma > 0$，我们有

$$\mathrm{V}_{\mathcal{H}}(\gamma) \leqslant \mathrm{P}_{\mathcal{H}}(\gamma) \leqslant \left(\frac{1}{\gamma} + 1\right) \mathrm{V}_{\mathcal{H}}(\gamma/2).$$

给定训练样本 $S = ((x_i, y_i))_{i=1}^n$，回顾一下，经验风险最小化算法选择一个假设 f_{ERM} 作为

$$\widehat{f}_{\mathrm{ERM}} \in \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - y_i|.$$

为了更好地理解定理 2.3 的结果，我们通过一致收敛论证，给出一个适用于任意经验风险最小化分类器的更标准的风险界。该界用 $\mathrm{P}_{\mathcal{H}}(\gamma)$ 表示。我们定义关于经验 L_1 距离的覆盖数。给定样本 $x_1, \ldots, x_n$，对于任意 $f, g \in \mathcal{H}$, $\rho_n(f,g) = \frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)|$，我们定义。令 $\mathcal{N}_1(t, \mathcal{H}, n)$ 表示关于 $x_1, \ldots, x_n$ 的最大覆盖数，即 $\mathcal{H}$ 关于 ρ_n 在尺度 t 下的覆盖数。由于我们未能在文献中找到明确的参考文献，[10] 我们在附录 B.1 中给出一个标准证明。为简化该界，我们对 $\mathrm{P}_{\mathcal{H}}(\cdot)$ 的行为引入一个温和的正则性假设，该假设使我们能够使用 Rudelson 和 Vershynin [RV06] 的强大结果。

命题 2.12。存在绝对常数 $c_1, c_2 > 0$，使得以下结论成立。固定一个假设类$\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$和间隔参数 $\gamma \in (0,1)$。假设存在常数 $c_3 > 1$，使得对所有 $\theta > 0$，均有 $\mathrm{P}_{\mathcal{H}}(c_3\theta) \leqslant \mathrm{P}_{\mathcal{H}}(\theta)/2$。对于任意 $\delta \in (0,1)$，任何以可实现分布 P中的训练样本 $S \sim P^n$ 为输入、并输出假设 f_{ERM} 的经验风险最小化算法，都满足

$$\mathop{\mathbb{E}}_{(X,Y)\sim P}\left[|\widehat{f}_{\mathrm{ERM}}(X) - Y|\right] \leqslant c_1 \left(\gamma + \frac{\mathrm{P}_{\mathcal{H}}(c_2\gamma)}{n} + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right),$$

以至少 $1-\delta$ 的概率（关于 S 的随机性）。

引理 2.11 表明，定理 2.3 的界至多比命题 2.12的界差常数乘法因子。然而，二者对应的复杂度度量可能有显著差异，以下关于经典保序回归问题的例子便说明了这一点。我们的例子还展示了 [RV06], 中的结果，该结果强调了对于一般假设类，覆盖数的对数与 P_γ-维度之间的等价性。

例 2.13（绝对损失下的保序回归）。设 $\mathcal{H} \subseteq [0,1]^{[0,1]}$ 为所有单调函数组成的集合。那么，对于任意 $\gamma, t \in (0,1]$,

$$\mathrm{V}_{\mathcal{H}}(\gamma) \leqslant 2, \quad \text{but} \quad \mathrm{P}_{\mathcal{H}}(\gamma) = \Theta\left(\frac{1}{\gamma}\right), \quad \text{and} \quad \log \mathcal{N}_1(t, \mathcal{H}, n) = \Theta\left(\frac{1}{t}\right).$$

具体而言，在这种情况下，定理 2.3 的上界随 $O\left(\frac{1}{n}\text{对数}\left(\frac{1}{\delta}\right)\right)$变化，而 命题 2.12 基于覆盖数的上界随 $O\left(\frac{1}{\sqrt{n}} + \frac{1}{n}\text{对数}\left(\frac{1}{\delta}\right)\right)$变化。

本例中的界基于若干已有结果。对于 $\mathrm{P}_{\mathcal{H}}(\gamma)$ 的上界由 引理 2.11 得出，而下界可直接根据定义构造。对于 $\log \mathcal{N}_1(t, \mathcal{H}, n)$ 的上界由 [VdVW96，定理 2.7 得出。5], 而下界则可由 [AB99，定理 12 推出。10]。

[10] 在 [AB99第 21 节中有密切相关的上界。4]。

有人可能会认为，采用一个结合局部覆盖数的 命题 2.12 版本，可以改善 例 2.13 中的经验风险最小化收敛速率。不幸的是，这并无帮助，因为对于非参数类，全局覆盖数与局部覆盖数的表现相同 [YB99、MN06]。最后，我们提及，Kot lowski、Koolen 和 Malek 的近期研究 [KKM17] 在类似的学习设定下，探讨了 例 2.13 中所考虑类别的期望风险界。

3 相关工作

二分类的最优样本复杂度。 确定可实现（无噪声）二分类的最优样本复杂度，或等价地，最优风险界，是统计学习理论中的经典问题。Vapnik 和 Chervonenkis 的早期工作[VC64] 提出了本文所称的 PAC 模型（这一术语由 Valiant 提出 [Val84], ，他还从计算角度研究了这些问题），并为有限假设类给出了匹配的风险上界和下界$\mathcal{H}$。随后，Vapnik 和 Chervonenkis 首次为无限类，具体而言是半空间类，提出了错误率上界 $\mathbb{R}^p$ [VC68a]。他们接下来的开创性工作 [VC68b] 将半空间的打散性质推广到一般类，并引入了如今所称的 VC 维数这一复杂度度量。专著 [VC74] 和论文 [BEHW89] 给出了适用于任何与训练样本一致的分类器的典型风险界：

$$O\left(\frac{d}{n}\log\left(\frac{n}{d}\right)+\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right). \tag{2}$$

已知，对于某些样本一致分类规则和假设类，这一收敛速率无法进一步改进 [HLW94、AO07、Sim15、Han16b]。然而，现有的下界 [EHKV89] 适用于任何预测器

$$\Omega\left(\frac{d}{n}+\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right) \tag{3}$$

与风险界（2）相差一个乘法对数因子。Haussler、Littlestone 和 Warmuth 的开创性工作[HLW94] （最初以 [HLW88]的形式发表）朝着证明与（3）速率相匹配的上界迈出了重要一步。他们提出了单包含图算法，并给出了一种简单的置信度提升策略，该策略偶尔能够提供更优的风险界

$$O\left(\frac{d}{n}\log\left(\frac{1}{\delta}\right)\right), \tag{4}$$

由此引出了这样的问题：PAC 学习的最优样本复杂度是多少，以及单包含图算法本身是否最优 [War04]。此后，许多作者针对特定类 [BL89, Lit89, Aue97, AO07, Dar15, Han16b] 和/或在关于数据分布的额外假设下 [Lon03, BLL09, BL13, Han16b]给出了与下界（3）速率相匹配的界。最终，Hanneke [Han16a], 解决了获得与（3）相匹配且仅相差常数因子的速率这一问题；他改进了 Simon [Sim15] 此前的结果，后者包含一个缓慢增长的额外乘法项。

留一法/可交换性论证。 留一分析是控制分类规则的期望风险的标准方法。一些早期应用可以追溯到 Vapnik 和 Chervonenkis 的专著[VC74], ，他们利用该方法为感知机算法和硬间隔支持向量机推导出了最优风险界。然而，这些结果仅在期望意义下成立，因此 Vapnik 和 Chervonenkis[VC74，第六章第 7] 节提出了这样的问题：对于一些具有实际意义的算法，是否可以以高概率达到同样的结果。Lunts 和 Brailovsky [LB67] 通过给出一个留一法论证无法产生低方差风险界的例子，证明了一般情况下这是不可能的，从而排除了在没有

额外假设的情况下获得高概率上界的可能性。留一法论证也是单包含图算法分析的核心，该算法由[HLW94]引入。Warmuth [War04] 猜想该算法能够始终提供最优 PAC上界。然而，最近在 [ACSZ22] 中证明，在没有额外假设的情况下，单包含图算法无法达到最优 PAC 界。最后，我们关注最近对稳定压缩方案的分析，该分析见于 [Zhi17、BHMZ20和HK21], ，其表明，在纳入稳定性假设后，留一法论证确实能够得到精确的高概率界。

置信度提升和在线到批量转换。 我们首先讨论 置信度提升 方法，这是一种从期望意义下的风险界推导高概率风险界的自然方法。该技术涉及在 $O(\log(1/\delta))$ 个原始样本的独立分区上运行期望意义下的最优算法，并根据独立的留出样本选择表现最佳的算法。Haussler、Littlestone 和 Warmuth [HLW94], 采用了置信度提升方法，但这种方法只能得到带有乘法因子 $\log(1/\delta)$ 的风险界，如式（4）所示。

另一个更为相关的方法是在线到批量转换，其最早的相关应用可以追溯到以下研究 [ABR65、VC74]。Littlestone [Lit89] 证明，如果一个保守在线算法（即仅在某个点被误分类时才改变其状态的算法）在任意样本上产生的错误次数有限，那么它就可以被转换为统计设定下的分类器，并具有风险界 M

$$O\left(\frac{M}{n}+\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right). \tag{5}$$

尽管给出了一个包含 $\log(1/\delta)$ 的正确加性项，Littlestone 的界仍有两个主要局限，而本文解决了这两个问题。第一个局限是要求错误次数有限，这一要求由所谓的 Littlestone维数 [Lit88], 刻画；对于大多数实际相关的假设类，该维数都是无穷的。Haussler、Littlestone 和 Warmuth [HLW94] 试图通过考察以在线方式在样本上运行的单包含图算法所犯错误的期望次数来解决这一问题，由此得到的期望错误数按 $O(d\log\, n)$ 的量级增长。该界 (5) 的第二个局限是算法必须是保守的。尽管任何能够达到确定性错误界的算法都可以转换为保守算法 [FW95], 但对于能够达到期望错误界的算法（如本文所研究的算法），目前尚不知这一结论是否成立。

一个相关的参考文献是 Wu、Heidari、Grama 和 Szpankowski 的工作[WHGS22]。他们通过反向鞅论证证明，可以得到与 Haussler 等人 [HLW94] 的期望界类似的高概率在线界，并表明以至少 $O(d\log\, n+\log(1/\delta))$ 的概率，错误次数不超过相应界。另一项观察来自 [HLW94], : 他们指出，以在线方式运行单包含图算法，并且仅在样本的后半部分（称为后缀）上衡量其错误时，期望错误数会降至 $O(d)$。利用样本后缀这一思路具有相当的普适性，并为随机梯度下降带来了近期更紧的高概率界 [HLPR19]。

多分类与部分假设。多分类是二分类情形的自然扩展。相关研究 [NT88、Nat89] 将 PAC 模型扩展到多分类问题，并引入了 VC 维数的一种类似概念，即如今通常所说的 Natarajan维度。后续研究探讨了这一维度能否刻画可学习性，并指出标准的一致收敛技术不足以回答这一问题。此后关于 Natarajan维度的研究包括[HL95、BCHL95]。虽然当类别数有限时，Natarajan维度能够保证一致收敛性质，但当类别数无界时，经验风险最小化器可能无法学会该类别[DSBS15]。文献 [RBR09] 将单包含图算法推广到多分类，而 Daniely 和 Shalev-Shwartz [DS14] 则通过单包含超图密度刻画了多分类问题的期望最优风险

。关于多分类中复杂度度量的近期结果和详细综述可参见[BCD+22]。

部分假设类由 $\{0,1,\star\}$值假设组成。与多分类的一个关键区别在于对 $\star$值的独特处理，这些值不会影响打散的概念或相应的复杂度度量。这一思想可追溯至 Haussler 和 Long [HL95] 的工作，并在 Bartlett 和 Long [BL98]的工作中得到进一步研究。文献 [BL98] 认识到，不应将一致收敛用于部分假设，并据此调整了单包含图算法。随后，Long [Lon01] 将部分假设的分析扩展到不可知设置，并推导出相应的风险界。Alon、Hanneke、Holzman 和 Moran [AHHM22]的工作重新激发了人们对这一主题的兴趣。他们的贡献包括为可实现情形和不可知情形确定更紧的界，同时还将这些思想扩展到在线学习。

4 主要结果的证明

在本节中，我们证明论文的主要技术结果，定理 2.1。回顾一下，我们的目标是界定平均风险

$$\frac{4}{3n}\cdot\sum_{t=n/4}^{n-1}\mathrm{R}_P(\widehat{f}(\cdot;S_{\leqslant t})) \tag{6}$$

由数据构造的一系列预测器的平均风险，并以高概率保证这一界。作为第一步，我们分析经验量

$$\frac{4}{3n}\cdot\sum_{t=n/4}^{n-1}\ell(\widehat{f}(X_{t+1};S_{\leqslant t}),Y_{t+1}). \tag{7}$$

直观而言，(7)近似于(6)，因为测试点 X_{t+1} 与用于 $S_{\leqslant t}$ 构造$\widehat{\ }$ $\widehat{\ }$预测器的训练样本 $\ell(f(X_{t+1};\ S_{\leqslant t})\ Y_{t+1})$是 $\mathrm{R}_P(f(\cdot;\ S_{\leqslant t}))$的无偏估计量。随后，我们在引理 4.2中使用鞅集中论证，将期望意义下的界转化为 PAC 界。

主要的技术挑战在于以高概率界定(7)。这一步关键地利用了所考虑的预测器具有置换不变性这一事实，假设 2.3 对此作了刻画。至关重要的是，我们的分析以输入的取值和标签为条件，即以$((X_i,Y_i))_{i=1}^n$ 为条件，但不以它们在输入序列中的排列为条件。证明中的关键洞见是以下条件独立结构。对于$\widehat{\ }$ $\widehat{\ }$任意 $i<j$, $\ell(f(X_{i+1};\ S_{\leqslant i})\ Y_{i+1})$，即使在以 $\ell(f(X_{j+1};\ S_{\leqslant j})\ Y_{j+1})$为条件时，其期望值仍然很小。$\widehat{\ }$这是因为 $\ell(f(x_{j+1};\ S_{\leqslant j})\ y_{j+1})$由训练样本集合 $S_{\leqslant j}$和 $\widehat{\ }(X_{j+1},Y_{j+1})$决定，而 $\ell(f(x_{i+1};\ S_{\leqslant i})\ y_{i+1})$则取决于训练样本集合 $S_{\leqslant i}$和 (X_{i+1},Y_{i+1})。由于即使给定集合 $S_{\leqslant j}$，$S_{\leqslant i}$仍具有可交换性，我们可以将 $\ell(f(X_{j+1};\ S_{\leqslant j})\ Y_{j+1})$期望值很小这一结论与假设 2.1 联系起来。引理 4.3 通过构造适当的滤过并使用一个简单的鞅集中界，将这一想法形式化。

为证明我们的结果，我们需要以下鞅切尔诺夫界。为完整起见，我们在引理4.1 的附录 C.1 中给出证明。

引理 4.1. 设 $W_1,\ \ldots,\ W_T$ 是关于滤过 $(\mathcal{F}_i)_{i\leqslant T}$ 适应的随机过程。假设 0 6$W_t\leqslant 1$ 几乎处处成立。那么，对于任意 δ、λ和 $\eta\in(0,1)$，有

$$\mathbf{Pr}\left[\sum_{t=1}^{T}W_t\geqslant\frac{e^{\lambda}-1}{\lambda}\sum_{t=1}^{T}\mathbb{E}[W_t|\mathcal{F}_{t-1}]+\frac{\log(1/\delta)}{\lambda}\right]\leqslant\delta, \tag{8}$$

and

$$\mathbf{Pr}\left[\sum_{t=1}^{T}\mathbb{E}[W_t|\mathcal{F}_{t-1}]\geqslant\frac{\eta e^{\eta}}{e^{\eta}-1}\sum_{t=1}^{T}W_t+\frac{e^{\eta}\log(1/\delta)}{e^{\eta}-1}\right]\leqslant\delta. \tag{9}$$

我们首先正式考察将经验损失与风险联系起来的论证。

引理 4.2（正向鞅风险界）。固定一个可实现分布 P，并令 ℓ 为有界于 1 的损失。令 $\widehat{f}\colon \mathcal{X}\times\mathcal{U}\to\mathcal{Y}$ 为一个预测器。给定训练样本 $S=((X_1,Y_1)\ldots,(X_T,Y_T))\sim P^T$，对于任意 δ、$\eta\in(0,1)$，有

$$\sum_{t=T/4}^{T-1}\mathrm{R}_P(\widehat{f}\,(\cdot;S_{\leqslant t}))\leqslant\frac{\eta e^{\eta}}{e^{\eta}-1}\sum_{t=T/4}^{T-1}\ell\left(\widehat{f}\,(X_{t+1};S_{\leqslant t}),Y_{t+1}\right)+\frac{e^{\eta}\log(1/\delta)}{e^{\eta}-1},$$

以至少 $1-\delta$ 的概率（关于 S 的随机性）。

证明。对于 $t=1,\ldots,T$，定义通常的滤过 $(\mathcal{F}_t)_{t=1}^T$，其中 $\mathcal{F}_t=\sigma(X_1,\ldots,X_t)$。这里，$\sigma(\cdot)$ 表示使这些参数可测的最小 σ-代数。稍微滥用符号，我们写作 $\widehat{f}_t(\cdot)=\widehat{f}(\cdot;S_{\leqslant t})$。根据定义，$\mathrm{R}_P(\widehat{f}_t)=\mathbb{E}[\ell(\widehat{f}_t(X_{t+1}),Y_{t+1})\mid\mathcal{F}_t]$，因此将引理 4.1 应用于随机变量

$$\ell(\widehat{f}_{T/4}(X_{T/4+1}),Y_{T/4+1}),\ldots,\ell(\widehat{f}_{T-1}(X_T),Y_t)$$

可得

$$\sum_{t=T/4}^{T-1}\mathrm{R}_P(\widehat{f}_t)\leqslant\frac{\eta e^{\eta}}{e^{\eta}-1}\sum_{t=T/4}^{T-1}\ell(\widehat{f}_t(X_{t+1}),Y_{t+1})+\frac{e^{\eta}\log(1/\delta)}{e^{\eta}-1},$$

以至少 $1-\delta$ 的概率（关于 S 的随机性）。 □

如前文所述，引理 4.2 告诉我们，要界定预测器 $\widehat{f}\,(\cdot,S_{\leqslant t})$ 的平均风险，只需界定在点 X_t 上评估时的损失[。为此，我们使用反向鞅论证。自 Pollard81]以来，这些论证在经验过程中已成为标准方法，并且最近已应用于序贯估计和不确定性量化。Manole 和 Ramdas [MR23 的命题 5] 建立了留一法论证与反向鞅之间的联系。Wu、Heidari、Grama 和 Szpankowski[WHGS22] 使用了类似的论证，专门研究如何利用单包含图算法处理序贯覆盖数。下面的证明可以通过将其转换为反向鞅和可交换滤过的语言而缩短，但为清晰起见，我们选择使用随机排列进行初等表述。

引理 4.3（反向鞅界）。固定一个可实现分布 P。令 $f\colon\mathcal{X}\times\mathcal{U}\to\mathcal{Y}$ 为满足假设 2.1 和 2.3 且留一法误差为 M_n 的预测器。给定一个训练样本 $S=(Z_1、\ldots、Z_T)=((X_1,Y_1)\ldots,(X_T,Y_T))\sim P^T$，对于任意 $\delta,\lambda\in(0,1)$，我们有

$$\sum_{t=T/4}^{T-1}\ell\left(\widehat{f}\,(X_{t+1};S_{\leqslant t}),Y_{t+1}\right)\leqslant\left(\ln(4)+\frac{1}{2}\right)\frac{e^{\lambda}-1}{\lambda}M_T+\frac{\log(1/\delta)}{\lambda},$$

以至少 $1-\delta$ 的概率（关于 S 的随机性）。

证明。令 $I_t=\ell(f(X_{T-t+1};S_{\leqslant T-t}),Y_{T-t+1})$。注意，$I_t$ 是 $Z_1、\ldots、Z_{T-t+1}$ 的函数；在需要时，我们将通过写作 $I_t(Z_1\ldots,Z_{T-t+1})$ 来明确这种依赖关系。此外，由于 $Z_1\ldots,Z_{T-t+1}$ 是可交换随机变量，因此 $Z_{\pi(1)}\ldots,Z_{\pi(T)}$ 的联合分布不依赖于排列 π。考虑一个与 $Z_1、\ldots、Z_T$ 独立的均匀随机置换 π。也就是说，在整个证明过程中，我们可以假设值 $Z_1、\ldots、Z_T$ 是固定的。考虑滤过

$$\mathcal{F}_t=\sigma\left(\left\{Z_{\pi(1)},\ldots,Z_{\pi(T-t)}\right\},Z_{\pi(T-t+1)},\ldots,Z_{\pi(T)}\right).$$

这里，$\left\{Z_{\pi(1)},\ldots,Z_{\pi(T-t)}\right\}$ 表示 $Z_{\pi(1)}、\ldots\ldots、Z_{\pi(T-t)}$ 所取值组成的集合（不考虑顺序），而 $\sigma(\cdot)$表示使各个自变量可测的最小 σ-代数。令 $S_\pi=((X_{\pi(1)},Y_{\pi(1)})\ldots,(X_{\pi(n)},Y_{\pi(n)}))$表示相对于 π对训练样本进行的排列。定义

$$I'_t=I_t(Z_{\pi(1)},\ldots,Z_{\pi(T-t+1)})=\ell(\widehat{f}(X_{\pi(T-t+1)};(S_\pi)_{\leqslant T-t}),Y_{\pi(T-t+1)}).$$

注意， I'_t 相对于滤过 $(\mathcal{F}_t)_{t=1}^T$ 是适应的，因为根据假设 2.3， $\widehat{f}$ 在训练样本 $(S_\pi)_{\leqslant T-t}$ 上是置换不变的。应用 (8) 可得，以至少 $1-\delta$ 的概率，有

$$\sum_{t=1}^{3T/4} I'_t \leqslant \frac{e^\lambda - 1}{\lambda} \sum_{t=1}^{3T/4} \mathbb{E}\left[I'_t | \mathcal{F}_{t-1}\right] + \frac{\log(1/\delta)}{\lambda}. \tag{10}$$

为了使用这一界，我们需要控制

$$\mathbb{E}\left[I'_t | \mathcal{F}_{t-1}\right] = \mathbb{E}\left[I(Z_{\pi(1)}, \ldots, Z_{\pi(T-t+1)}) | \mathcal{F}_{t-1}\right]. \tag{11}$$

注意，在时刻 $t-1$（相对于滤过（$\mathcal{F}_t$）$_{t=1}^T$），随机索引 π（1） ...， π（$T-t+1$）对应于集合 $[T] \setminus \{\pi$（i）$\}_{i=T-t+2}^T$中元素的均匀随机置换。一个标准计算将对均匀随机置换的平均与留一法误差联系起来，由此结合（11）和假设 2.1 可得该界

$$\mathbb{E}\left[I'_t | \mathcal{F}_{t-1}\right] \leqslant \frac{M_{T-t+1}}{T-t+1}.$$

将上述结果代入（10），再结合 Z_1，....., Z_T的可交换性，可得

$$\begin{aligned}
\sum_{t=T/4}^{T-1} \ell\left(\widehat{f}(X_{t+1}; S_{\leqslant t}), Y_{t+1}\right) &\leqslant \frac{e^\lambda - 1}{\lambda} \sum_{t=1}^{3T/4} \frac{M_{T-t+1}}{T-t+1} + \frac{\log(1/\delta)}{\lambda} \\
&\leqslant \frac{e^\lambda - 1}{\lambda} M_T \sum_{t=T/4+1}^{T} \frac{1}{i} + \frac{\log(1/\delta)}{\lambda} \qquad (12) \\
&\leqslant \frac{e^\lambda - 1}{\lambda} M_T \left(\ln 4 + \frac{1}{2}\right) + \frac{\log(1/\delta)}{\lambda}. \qquad (13)
\end{aligned}$$

不等式（12）由函数 M_t的单调性得到。最后，（13）由对所考虑的调和级数进行简单近似得到。

□

现在，我们已经可以将上述两个结果结合起来证明主要结果。为方便起见，我们重新陈述该定理。

定理 2.1。 固定一个假设类 $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$、损失函数 $\ell\colon \mathcal{Y}\times\mathcal{Y} \to [0,1]$，以及一个预测器 $\widehat{f}\colon \mathcal{X}\times\mathcal{U} \to \mathcal{Y}$，满足 假设 2.1 和 2.3。那么，对于任意关于 P 的可实现分布 $\mathcal{X}\times\mathcal{Y}$ 和置信参数 $\delta \in$（0，1），给定训练样本 $S \sim P^n$，我们有

$$\frac{4}{3n} \cdot \sum_{t=n/4}^{n-1} \mathrm{R}_P(\widehat{f}(\cdot; S_{\leqslant t})) \leqslant 4.82 \left(\frac{M_n}{n} + \frac{1}{n}\log\left(\frac{2}{\delta}\right)\right),$$

以至少 $1-\delta$ 的概率，关于 S 的随机性。

定理 2.1的证明。 利用 引理 4.2 和 4.3 以及置信参数 $\delta/2$，结合并集界可知，以至少 $1-\delta$ 的概率，关于 S 的随机性，有

$$\frac{4}{3n} \sum_{t=n/4}^{n-1} \mathrm{R}_P(\widehat{f}(\cdot; S_{\leqslant t})) \leqslant \frac{4}{3n}\left(\frac{\eta e^\eta}{e^\eta - 1}\left((\ln(4) + 1/2)\frac{e^\lambda - 1}{\lambda} M_n + \frac{\log(1/\delta)}{\lambda}\right) + \frac{e^\eta \log(2/\delta)}{e^\eta - 1}\right).$$

设置参数 为 [11] 和 $\lambda = 0.82$，并令 $\eta = 0.78$，经过简单计算可得上界

$$\frac{4}{3n} \sum_{t=n/4}^{n-1} \mathrm{R}_P(\widehat{f}(\cdot; S_{\leqslant t})) \leqslant 4.82 \left(\frac{M_n}{n} + \frac{1}{n}\log\left(\frac{2}{\delta}\right)\right).$$

证明至此完成。 □

[11]参数的选择 已经过优化，以获得较小的复杂度 nstant。

致谢。作者衷心感谢 Peter Bartlett 和 Phil Long 就有界回归场景所进行的有益讨论。IA 和 AS 谨此感谢 Prasad Raghavendra 开设相关课程，使本研究得以作为期末项目启动。

参考文献

[AB99] Martin Anthony 和 Peter Bartlett。《神经网络学习：理论基础》，第9卷。剑桥大学出版社，1999年。9，10，23[ABCH97] Noga Alon、Shai Ben-David、Nicolo Cesa-Bianchi 和 David Haussler。尺度敏感维度、一致收敛与可学习性。《ACM学报》，44(4)：615–631，1997年。9，10[ABR65] Mark Aizerman、Emmanuil Braverman 和 Lev Rozonoer。模式识别学习中势函数方法的理论基础。《自动化与远程控制》，25：821–837，1965年。12[ACSZ22] Ishaq Aden-Ali、Yeshwanth Cherapanamjeri、Abhishek Shetty 和 Nikita Zhivotovskiy。单包含图算法并不总是最优，2022年。3，8，12[AHHM22] Noga Alon、Steve Hanneke、Ron Holzman 和 Shay Moran。部分概念类的PAC可学习性理论。载于2021年IEEE第62届计算机科学基础年度研讨会（FOCS），第658–671页。IEEE，2022年。1，2，3，8，13[AO07] Peter Auer 和 Ronald Ortner。交封闭概念类的新PAC 界。《机器学习》，66(2)：151–163，2007年。11[AT92] Noga Alon 和 Michael Tarsi。图的着色与定向。《组合数学》，12(2)：125–134，1992年。20[Aue97] Peter Auer。恶意噪声下嵌套差分的学习。《理论计算机科学》，185(1)：159–175，1997年。11[BCD+22] Nataly Brukhim、Daniel Carmon、Irit Dinur、Shay Moran 和 Amir Yehudayoff。多分类可学习性的刻画。arXiv预印本 arXiv:2203.01550，2022年。7，13[BCHL95] Shai Ben-David、Nicol´o Cesa-Bianchi、David Haussler 和 Philip M Long。$\{0,\ldots,n\}$值函数类可学习性的刻画。《计算机与系统科学学报》，50(1)：74–86，1995年。12[BEHW89] Anselm Blumer、Andrzej Ehrenfeucht、David Haussler 和 Manfred K Warmuth。可学习性与Vapnik-Chervonenkis维数。《ACM学报》，36(4)：929–965，1989年。2，11[BHMZ20] Olivier Bousquet、Steve Hanneke、Shay Moran 和 Nikita Zhivotovskiy。适当学习、Helly数与最优 SVM界。载于学习理论会议，第582–609页，2020年。12[BL89] Anselm Blumer 和 Nick Littlestone。比Vapnik-Chervonenkis维数所承诺的更快学习。《离散应用数学》，24(1-3)：47–53，1989年。11[BL98] Peter Bartlett 和 Philip Long。预测、学习、一致收敛与尺度敏感维度。《计算机与系统科学学报》，56(2)：174–190，1998年。1，2，8，9，13, 23, 24[BL13] Maria-Florina Balcan 和 Phil Long。对数凹分布下线性分隔器的主动学习与被动学习。载于学习理论会议，第288–316页，2013年。11

[BLL09] Nader H Bshouty、Yi Li 和 Philip M Long。利用倍增维数分析学习算法的泛化。《计算机与系统科学学报》，75(6)：323–335，2009年。11[Dar15] Malte Darnst¨adt。交封闭概念类的最优 PAC 界。《信息处理快报》，115(4)：458–461，2015年。11[DS14] Amit Daniely 和 Shai Shalev-Shwartz。多分类问题的最优学习器。载于学习理论会议，第287–316页，2014年。1，2，6，7，12[DSBS15] Amit Daniely、Sivan Sabato、Shai Ben-David 和 Shai Shalev-Shwartz。多分类可学习性与经验风险最小化原则。《机器学习研究杂志》，16：2377–2404，2015年。12[EHKV89] Andrzej Ehrenfeucht、David Haussler、Michael J. Kearns 和 Leslie G. Valiant。学习所需样本数量的一般下界。《信息与计算》，82(3)：247–261，1989年。11[FW95]Sally Floyd 和 Manfred Warmuth。样本压缩、可学习性与 Vapnik-Chervonenkis维数。《机器学习》，21(3)：269–304，1995年。12[Han16a] Steve Hanneke。PAC 学习的最优样本复杂度。《机器学习研究杂志》，17(1)：1319–1333，2016年。1，2，3，8，11[Han16b] Steve Hanneke。若干学习算法的精细误差界。《机器学习研究杂志》，17(1)：4667–4721，2016年。11[Hau95] David Haussler。具有有界 Vapnik-Chervonenkis维数的布尔 n维超立方体子集的球装填数。《组合理论杂志A辑》，69(2)：217–232，1995年。7[HK21] Steve Hanneke 和 Aryeh Kontorovich。稳定样本压缩方案：新应用与最优 SVM间隔界。载于算法学习理论，第697–721页，2021年。12[HL95] David Haussler 和 Philip M Long。Sauer引理的推广。《组合理论杂志A辑》，71(2)：219–240，1995年。12，13[HLPR19] Nicholas Harvey、Christopher Liaw、Yaniv Plan 和 Sikander Randhawa。非光滑随机梯度下降的紧致分析。载于学习理论会议，第1579–1613页，2019年。12[HLW88] David Haussler、Nick Littlestone 和 Manfred Warmuth。在随机抽取的点上预测 $\{0,1\}$函数（扩展摘要）。载于第29届计算机科学基础年度研讨会（FOCS），第100–109页，1988年。2，5，11[HLW94] David Haussler、Nick Littlestone 和 Manfred K Warmuth。在随机抽取的点上预测 $\{0,1\}$函数。《信息与计算》，115(2)：248–292，1994年。1，2，6，7，11，12，19，20，24[KKM17] Wojciech Kot lowski、Wouter M Koolen 和 Alan Malek。随机排列在线保序回归。《神经信息处理系统进展》，30，2017年。11[KS94] Michael J Kearns 和 Robert E Schapire。概率概念的高效分布无关学习。《计算机与系统科学学报》，48(3)：464–497，1994年。9[KS23] Pieter Kleer 和 Hans Simon。原始与对偶组合维数。《离散应用数学》，327：185–196，2023年。9

[KVK22] Alkis Kalavasis、Grigoris Velegkas 和 Amin Karbasi。《超越 PAC 框架的多分类可学习性：普适速率与部分概念类》。arXiv 预印本 arXiv:2210.02297，2022年。8[Lar22] Kasper Green Larsen。《装袋法是一种最优 PAC 学习器》。arXiv 预印本 arXiv:2212.02264，2022年。8[LB67] Alexander Lunts 和 Victor Brailovsky。《统计决策规则中获得的属性的评估》。工程控制论，3:98–109，1967年。11[Lit88] Nick Littlestone。《无关属性大量存在时的快速学习：一种新的线性阈值算法》。《机器学习》，2:285–318，1988年。12[Lit89] Nick Littlestone。《从在线学习到批量学习》。载于第二届计算学习理论年度研讨会论文集，第269–284页。Morgan Kaufmann，1989年。5, 11, 12[LLS01] Yi Li、Philip M Long 和 Aravind Srinivasan。《单包含图算法在学习预测模型中近乎最优》。IEEE 信息论汇刊，47(3):1257–1261，2001年。2[Lon01] Philip M Long。《关于以 $\{0, \star, 1\}$值和实值假设进行无关学习》。载于~~计算学习理论会议~~，第289–302页，2001年。13[Lon03] Philip M Long。《关于相对于均匀分布进行 PAC 半空间学习的样本复杂度上界》。信息处理快报，87(5):229–234，2003年。11[MN06] P~~ascal Massart 和 Élodie Nédelec。~~《统计学习的风险界》。《统计年刊》，34(5)，2006年。11[MR23] Tudor Manole 和 Aaditya Ramdas。《用于凸泛函与散度序贯估计的鞅方法》。IEEE 信息论汇刊，2023年。14[MV03] Shahar Mendelson 和 Roman Vershynin。《熵与组合维数》。数学发明，152(1):37–55，2003年。23[Nat89] Balas K Natarajan。《关于集合与函数的学习》。《机器学习》，4:67–97，1989年。12[NT88] Balas K Natarajan 和 Prasad Tadepalli。《两种新的学习框架》。载于《机器学习论文集 1988》，第402–415页。Elsevier，1988年。12[Pol81] David Pollard。《经验过程的极限定理》。《概率论及相关领域》，57(2):181–195，1981年。14[RBR09] Benjamin Rubinstein、Peter Bartlett 和 Hyam Rubinstein。《平移：单包含错误界与样本压缩》。《计算机与系统科学学报》，75(1):37–59，2009年。2, 6, 12, 19, 20[RV06] Mark Rudelson 和 Roman Vershynin。《随机过程的组合学与凸体的截面》。《数学年刊》，第603–648页，2006年。10, 23[Sim15] H~~ans U Simon。《一种~~近乎最优的 PAC 算法》。载于学习理论会议，第1552–1563页，2015年。1, 2, 11[SS10] Hans U Simon 和 Bal´ázs Sz¨ör´enyi。《重新审视单包含超图密度》。信息处理快报，110(8-9):341–344，2010年。2[Tal96] Michel Talagrand。~~《十年后的 Glivenko-Cantelli问题》。~~《理论概率论杂志》，9:371–384，1996年。9

[Tal03] Michel Talagrand。《Vapnik–Chervonenkis 型条件与一致 Donsker 类函数》。《概率年刊》，31(3):1565–1582，2003年。9[Val84] Leslie G Valiant。《可学习性理论》。《ACM 通讯》，27(11):1134–1142，1984年。2, 11[VC64] 弗拉基米尔·瓦普尼克和阿列克谢·切尔沃年基斯。《一类用于模式识别学习的算法》。《自动化与遥测技术》，25(6):937–945，1964年。11[VC68a] 弗拉基米尔·瓦普尼克和阿列克谢·切尔沃年基斯。《模式识别学习问题中的完全记忆算法与递归算法》。《自动化与遥测技术》，第95–106页，1968年。2, 11[VC68b] 弗拉基米尔·瓦普尼克和阿列克谢·切尔沃年基斯。《事件相对频率向其概率一致收敛的问题》。载于《苏联科学院通报》，第181卷，第781–787页，1968年。11[VC71] 弗拉基米尔·瓦普尼克和阿列克谢·切尔沃年基斯。《属性出现频率向其概率一致收敛的理论，以及利用经验数据确定最优解的问题》。《自动化与遥测技术》，2:42–53，1971年。9[VC74] 弗拉基米尔·瓦普尼克和阿列克谢·切尔沃年基斯。《模式识别理论》。科学出版社，莫斯科，1974年。2, 11, 12[VC82] 弗拉基米尔·瓦普尼克和阿列克谢·切尔沃年基斯。《均值向其期望一致收敛的充要条件》。《概率论及其应用》，26(3):532–553，1982年。9[VdVW96] 阿德·范德法特和乔恩·韦尔纳。《弱收敛与经验过程》。施普林格，1996年。10[War04]Manfred K Warmuth。《最优 PAC 算法》。载于国际计算学习理论会议，第641–642页。施普林格，2004年。1, 2, 3, 8, 11, 12[WHGS22] 吴长隆、莫森·海达里、阿南特·格拉马和沃伊切赫·什潘科夫斯基。《通过随机序贯覆盖求期望最坏情况遗憾》。arXiv 预印本 arXiv:2209.04417，2022年。12, 14[YB99] 杨玉红和安德鲁·巴伦。《极小极大收敛速率的信息论确定》。《统计年刊》，第1564–1599页，1999年。11[Zhi17] Nikita Zhivotovskiy。《通过局部熵与样本压缩实现最优学习》。载于学习理论会议，第2023–2065页，2017年。12

来自第 2.4 节的证明：分类

本节将介绍单包含超图算法的相关细节 [HLW94、RBR09]。我们还将证明第 2.4 节中所声称的结果。

我们首先给出超图的一些预备定义。我们将超图$G=(V,E)$的一个定向定义为函数 $\sigma_G:\ E\to V$，使得对于任意 $e\in E$, $\sigma_G(e)\ \in e$。给定一个超图 $G=(V,E)$ 及其定向 σ_G，定义顶点 $v\in V$ 的出度为

$$\mathrm{out}(v;\sigma_G)=|\{e\in E: v\in e, \sigma_G(e)\neq v\}|.$$

对于使用 σ_G定向的超图 G， G 的最大出度定义为 $\mathrm{out}(\sigma_G)\ =\max_{v\in V}\mathrm{out}(v;\sigma_G)$。

我们现在 准备将单包含超图算法以伪代码的形式写在算法 1.

算法1 单包含超图算法。

输入：训练样本 S。

输出：假设 $\widehat{f}_{\mathrm{OIG}}(\cdot; S)$： $\mathcal{X} \to \mathcal{Y}$。

对于任意点 $x \in \mathcal{X}$， $\widehat{f}_{\mathrm{OIG}}(\cdot; S)$ 的预测如下：

1: 构造超图 $G(\mathcal{H}|_{\mathcal{X}_S \cup \{x\}})$，并选取使 out($\sigma_{\mathcal{X}_S \cup \{x\}}$) 最小的定向 $\sigma_{\mathcal{X}_S \cup \{x\}}$ 。

2: 如果存在唯一一个与 S 和 $\mathcal{H}$ 一致、适用于 x 的标签 y，则预测为 y。

3: 否则，令 e 为 $G(\mathcal{H}|_{\mathcal{X}_S \cup \{x\}})$ 中与 S 一致、但不在 x 上的那条边。

4: 按照 $\sigma_{\mathcal{X}_S \cup \{x\}}(e)$ 进行预测，也就是说，预测定向 e 中箭头所指向的假设。

单包含（超）图算法 [HLW94、RBR09] 最初是针对转导式学习模型定义的，即它会为单个测试点预测一个标签 x。上述算法是其标准扩展版本，用于为整个定义域 $\mathcal{X}$定义预测器。当对测试点 x进行预测时，该算法会最小化 out($G(\mathcal{H}|_{\mathcal{X}_S \cup \{x\}})$)，它是留一法误差的上界（引理 A.2）。在探讨该算法的留一法误差之前，我们先将 out($G(\mathcal{H}|_{\mathcal{X}_S \cup \{x\}})$) 与 $\mu_{\mathcal{H}}(n)$联系起来。

每个超图 G 都有一个定向 σ_G 使得最大出度 out(σ_G) 至多为 $\lceil \mu(G) \rceil$。对于图而言，这是一个标准事实（例如，参见 [AT92，引理 3。1], [HLW94，定理 2。2]。我们预计，针对超图的推广结论也已为人所知，但由于未能找到明确的参考文献，我们给出一个基于最大流论证的证明，该论证推广了图情形下的一个已知证明 [HLW94]。

引理 A.1. 对于任意超图 $G = (V, E)$ 以及任意整数 $d \geqslant \mu(G)$，存在一个定向 σ_G 使得最大出度满足 out(σ_G) $\leqslant d$。

证明。我们构造一个辅助图，其中包含一个源点 s、一个汇点 t，以及位于 s 和 t 之间的两个顶点集 V_1 和 V_2。 V_1 包含 E 中每条边对应的一个顶点，即 $V_1 = E$。 V_2 等于 G 的顶点集，即 $V_2 = V$。从 s 到每个 $e \in V_1$ 都有一条带权有向边。对于每个 $e \in V_1$，从 s 到 e 的边的权重等于 $|e| - 1$。对于每个 $e \in V_1$，都有一条从 e 到参与 G 中超边 e 的每个 $v \in V_1$ 的带权有向边。每条此类边的权重均为 1。最后，从每个 $v \in V_2$ 到 t 都有一条边，每条边的权重为 d。我们将证明，该图从 s 到 t 的最大流等于m :
$= \sum_{e \in V_1} |e| - 1$。

由于所有权重都是整数，因此存在整数最大流。考虑这样一个流。我们可以用一种简单的方法从该流中得到所需的 G 定向。固定一个 $e \in V_1$。由于 e 获得了 $|e| - 1$ 单位流量，因此其 $|e|$ 个邻点中有 $|e| - 1$ 个获得了单位流量。因此， e 只有一个邻点 $v \in V_2$ 没有获得单位流量，于是我们在原超图 G 中将边 e 定向至该 v。对每个 $e \in V_1$重复此操作。 $v \in V_2$ 接收的每单位流量对应于 G 中的一条边，而它未被选为该边的头部。由于从每个 $v \in V_2$ 到 t 的边的容量为 d，我们可以得出，该定向满足每个 $v \in V$ 的出度至多为 d。

现在我们证明最大流为 m。容易看出，流至多为 m ，因为从 s 到 V_1 的边权之和等于 m。剩下需要证明最大流至少为 m。回顾一下，最大流等于将 s 与 t分隔开的最小割（边集），因此只需证明最小割的一个下界。

下面是一个有用的观察，我们将会用到。注意，从 s 到 t 的每条路径都由 3 条边组成，并经过某些 $e \in V_1$ 和 $v \in V_2$。固定其中一条这样的路径。由于我们考虑的是最小割，如果

第一条边（连接 s 到 e）被割断，那么就没必要割断 e 和 v 之间的边，因为这样只会增加割值。同理，如果第三条边（连接 v 到 t）被割断，就没必要割断 e 和 v之间的边。

令 $U \subseteq V_1$ 为这样一组顶点：它们与 s 之间的边未被割断。令 $W \subseteq V_2$ 为这样一组顶点：它们与 t 之间的边已被割断。固定一个 $e \in U$。从 e 到 t 有 $|e|$ 条路径，而这些路径中的每一条都必须在边的中间层或第三层被割断，但不能同时在两层被割断。令 $W_{e,2} \subseteq V_2$ 为这样一组顶点：从 e 到 t 的路径在第二层被割断；令 $W_{e,3} \subseteq V_2$ 为这样一组顶点：从 e 到 t 的路径在第三层被割断。显然 $W_{e,3} \subseteq W$。由于我们考虑的是最小割，根据先前的观察可得 $W_{e,3} \cap W_{e,2} = \emptyset$ 。

现在我们对从 s 到 U 的边的权重之和给出上界。将这个和记为 $a = \sum_{e\in U} |e| - 1$，它可以写成 $\sum_{e\in U} |e \cap W_{e,3}| + |e \cap W_{e,2}| - 1$。令 $W' = \cup_{e\in U} W_{e,3}$，并注意到

$$\sum_{e\in U} |e \cap W'| - 1 \leqslant \sum_{e\in V_1} |e \cap W'| - 1 \leqslant \mu(G)|W'| \leqslant d|W'|.$$

因此，$a \leqslant d|W'| + \sum_{e\in V_1} |e \cap W_{e,\ 2}|$，所以第一层被切割的边的权重之和至少为 $m - a \geqslant m - d|W'| - \sum_{e\in V_1} |e \cap W_{e,\ 2}|$。由于第二层中每条边的权重均为 1，根据前面的观察，第二层被切割的边的权重之和可写为 $\sum_{e\in U} |e \cap W_{e,\ 2}|$。最后，由于最终层中每条边的权重为 d，第三层被切割的边的总权重等于 $d|W|$。由于 $W' \subseteq W$，我们可以得出结论：被切割的边的总权重至少为

$$m - d|W'| - \sum_{e\in U} |e \cap W_{e,2}| + d|W| + \sum_{e\in U} |e \cap W_{e,2}| \geqslant m.$$

因此，最大流确实为 m。 □

现在我们可以使用 引理 A.1 来证明单包含图算法的留一法误差上界。

引理 A.2。固定一个假设类 $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$。单包含图算法 $\curvearrowright f_{\mathrm{OIG}}$ 的留一法误差： $\mathcal{X} \times \mathcal{U} \to \mathcal{Y}$ 对于任意可实现的训练样本 $S = ((x_1, y_1) \ldots, (x_n, y_n))$，满足

$$\sum_{i=1}^{n} \mathbf{1}\left\{\widehat{f}_{\mathrm{OIG}}(x_i; S^{-i}) \neq y_i\right\} \leqslant \lceil \mu_{\mathcal{H}}(n) \rceil.$$

证明。令 $U = \{x_1, \ldots, x_n\}$、 $G(\mathcal{H}|_U) = (V,\ E)$ 为使用 U构造的单包含超图，并令 σ_U为使 out($G(\mathcal{H}|_U)$) 最小的一个定向。此外，令 $f^* \in V$为能正确标记U 中所有点的假设，即 $f^*(i) = y_i$ 对所有 $i \in [n]$均成立。注意，算法1对于每种留出 x_i的选择，都会构造相同的单包含超图，并以相同方式对其定向。因此，预测器犯错的次数等于它未选择 f^*的次数。于是

$$\sum_{i=1}^{n} \mathbf{1}\left\{\widehat{f}_{\mathrm{OIG}}(x_i; S^{-i}) \neq y_i\right\} = \sum_{i=1}^{n} \mathbf{1}\left\{\sigma_U(e_{i,f^*}) \neq f^*\right\} = \mathrm{out}(f^*; \sigma_U) \leqslant \lceil \mu_{\mathcal{H}}(n) \rceil,$$

其中不等式由引理 A.1得出。 □

对于一组假设 $\{f_1, \ldots, f_t\} \subseteq \mathcal{H}$，我们将它们的多数表决 Maj($f_1, \ldots, f_t$) 定义为这样一个函数：输入 $x \in \mathcal{X}$时，输出 $y \in \mathcal{Y}$在 $f_1(x) \ldots, f_t(x)$ 中出现次数最多的标签。现在我们来证明主要的多类 PAC 界。为便于阅读，我们先在证明之前重述定理 2.2 。

定理 2.2。固定一个假设类 $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$。存在一个预测器 $\widehat{f}$： $\mathcal{X} \times \mathcal{U} \to \mathcal{Y}$对于任意 $\delta \in (0, 1)$，以及从任意可实现分布 P中采样得到的 $S \sim P^n$，满足

$$\mathrm{err}_P(\widehat{f}(\cdot; S)) \leqslant 9.64 \left(\frac{\lceil \mu_{\mathcal{H}}(n) \rceil}{n} + \frac{1}{n} \log\left(\frac{2}{\delta}\right)\right),$$

以至少 $1-\delta$ 的概率（关于 S 的随机性）。

证明。 给定完整的训练样本 S，令 $(\widehat{f}_{\mathrm{OIG}}(\cdot;\ S_{\leqslant t}))_{t=1}^{n-1}$表示在不断增长的训练样本序列（$S_{\leqslant t}$）$_{t=1}^{n-1}$上训练得到的单包含超图预测器。由于这些预测器关于其训练样本是对称的，并且具有留一法误差保证（引理 A.2），我们可以将定理 2.1应用于该序列的后缀，从而得到

$$\frac{4}{3n}\sum_{t=n/4}^{n-1}\mathrm{err}_P(\widehat{f}_{\mathrm{OIG}}(\cdot;S_{\leqslant t}))\leqslant 4.82\left(\frac{\lceil\mu_{\mathcal{H}}(n)\rceil}{n}+\frac{1}{n}\log\left(\frac{2}{\delta}\right)\right),$$

以至少 $1-\delta$ 的概率（关于 S 的随机性）。只要该界成立，后缀中的单包含超图预测器的多数表决，即 $\widehat{f}$ (·) $=$Maj $\left(\widehat{f}_{\mathrm{OIG}}(\cdot;S_{\leqslant\frac{n}{4}})\ldots\ldots,\widehat{f}_{\mathrm{OIG}}(\cdot;\ S_{\leqslant n-1})\right)$，满足

$$\begin{aligned}\mathrm{err}_P(\widehat{f}) &= \mathop{\mathbb{E}}_{(X,Y)\sim P}[\mathbf{1}\{\widehat{f}(X)\neq Y\}]\\ &\leqslant \mathop{\mathbb{E}}_{(X,Y)\sim P}\left[2\left(\frac{4}{3n}\sum_{t=n/4}^{n-1}\mathbf{1}\{\widehat{f}_{\mathrm{OIG}}(X;S_{\leqslant t})\neq Y\}\right)\right]\\ &= 2\cdot\frac{4}{3n}\sum_{t=n/4}^{n-1}\mathrm{err}_P(\widehat{f}_{\mathrm{OIG}}(\cdot;S_{\leqslant t}))\\ &\leqslant 9.64\left(\frac{\lceil\mu_{\mathcal{H}}(n)\rceil}{n}+\frac{1}{n}\log\left(\frac{2}{\delta}\right)\right).\end{aligned}$$

在上面的第一个不等式中，我们利用了这样一个事实：当多数表决错误时，正确的预测器数量不可能超过总数的一半。这就完成了证明。 □

A.1 部分假设类

如第 2.4 节所述，只需稍作修改单包含超图算法，就可以将多分类风险界应用于部分假设分类。下面将对此作进一步阐述。固定一个部分假设类 $\mathcal{H}\subseteq\{0,1,\star\}^{\mathcal{X}}$。若按照单包含超图的定义，对于定义域 $U\subseteq\mathcal{X}$的任意子集，我们首先计算投影 $\mathcal{H}|_U$，并构造相应的超边集合。对于部分假设类，我们稍微修改单包含超图的定义：从投影 $\mathcal{H}|_U$ 中移除所有会给训练点 $x\in U$标上 $\star$ 标签的顶点（假设），然后按照通常方式，从这个截断后的顶点集合构造边集。这样做的依据是，在从可实现分布 P中采样的点上，最优假设 $f^*\in\mathcal{H}$ 绝不会给点标上 $\star$ 标签。显然，这一修改使我们的单包含超图成为一个图。我们也可以在做出这一修改后重新定义 $\mu_{\mathcal{H}}$（n）的单包含超图密度。 $\mathcal{H}$ 根据二分类情形中的论证（定理 2.8），可立即得出 $\mu_{\mathcal{H}}$（n）的单包含超图密度 $\mathcal{H}$ 至多为 d，即 $\mathcal{H}$的 VC 维数。

B 第 2.5 节的证明：有界回归

B.1 命题 2.12 的证明

命题 2.12。存在绝对常数 c_1、$c_2>0$，使得以下结论成立。固定一个假设类$\mathcal{H}\subseteq[0,1]^{\mathcal{X}}$和间隔参数 $\gamma\in(0,1)$。假设存在常数 $c_3>1$，使得对于所有 $\theta>0$，均有 $\mathrm{P}_{\mathcal{H}}(c_3\theta)\leqslant\mathrm{P}_{\mathcal{H}}(\theta)/2$。对于任意 $\delta\in(0,1)$，每个接收来自可实现分布 P 的训练样本 $S\sim P^n$ 作为输入并输出假设 f_{ERM} 的经验风险最小化算法都满足

$$\mathop{\mathbb{E}}_{(X,Y)\sim P}\left[|\widehat{f}_{\mathrm{ERM}}(X)-Y|\right]\leqslant c_1\left(\gamma+\frac{\mathrm{P}_{\mathcal{H}}(c_2\gamma)}{n}+\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right),$$

以至少 $1-\delta$ 的概率成立，其中概率取自 S 的随机性。

证明。 不失一般性，我们假设在可实现情形下，有 $Y = f^*(X)$，其中$f^* \in \mathcal{H}$。首先，我们对经验过程使用一个标准的比率型界。 中的定理 19.799] 表明，对于任意 $t \geqslant 0$,

$$\mathbf{Pr}\left(\sup_{f\in\mathcal{H}} \mathop{\mathbb{E}}_{X\sim P}[|f(X)-f^*(X)|] - \frac{2}{n}\sum_{i=1}^{n}|f(X_i)-f^*(X_i)| \geqslant t\right) \leqslant 2\mathcal{N}_1\left(\frac{t}{12}, \mathcal{H}', 2n\right)\exp\left(-\frac{2nt}{9}\right), \quad (14)$$

其中 $\mathcal{H}' = \{x \mapsto |f(x) - f^*(x)| : f \in \mathcal{H}\}$. 对于任意分布 P'，我们有

$$\mathop{\mathbb{E}}_{X\sim P'}[||f(X)-f^*(X)| - |g(X)-f^*(X)||] \leqslant \mathop{\mathbb{E}}_{X\sim P'}[|f(X)-g(X)|].$$

因此，只需用 $\mathcal{N}_1\left(\frac{t}{12}, \mathcal{H}, 2n\right)$-维度来界定 P_γ。 中定理 1.3 的结果06] 恰好做到了这一点。[12] 事实上，在我们陈述中的正则性假设下，该结果意味着

$$\log 2\mathcal{N}_1\left(\frac{t}{12}, \mathcal{H}, 2n\right) \leqslant c_4 \mathrm{P}_{\mathcal{H}}(c_2 t),$$

其中 c_2, $c_4 > 0$ 是绝对常数。注意，$\mathrm{P}_{\mathcal{H}}(c_2\gamma)$ 只取整数值，并且随着γ 的增大而减小。因此，可以找到 $\gamma^* \in [0,1]$，使得 $\mathrm{P}_{\mathcal{H}}(c_2\gamma^*) \leqslant \frac{\gamma^* n}{9c_4} \leqslant \mathrm{P}_{\mathcal{H}}(c_2\gamma^*) + 1$。固定 $t = \gamma^*$。根据经验风险最小化器的定义和（14），我们以至少 $1-\exp(-n\gamma^*/9)$ 的概率有，

$$\mathop{\mathbb{E}}_{X\sim P}\left[|\widehat{f}_{\mathrm{ERM}}(X) - f^*(X)|\right] \leqslant \gamma^*.$$

如果 $n\gamma^*/9 \geqslant \log(1/\delta)$，则上述不等式以至少 $1-\delta$ 的概率成立。否则，我们可以改为考虑 $t = \frac{9}{n}\log(1/\delta)$，它满足 $t \geqslant \gamma^*$，因此根据单调性， $c_4\mathrm{P}_{\mathcal{H}}(c_2 t) \leqslant c_4\mathrm{P}_{\mathcal{H}}(c_2\gamma^*)\, 6\frac{\gamma^* n}{9} \leqslant \log(1/\delta)$。这意味着，只要 $n\gamma^*/9 < \log(1/\delta)$，我们就有至少 $1-\delta$ 的概率满足，

$$\mathop{\mathbb{E}}_{X\sim P}\left[|\widehat{f}_{\mathrm{ERM}}(X) - f^*(X)|\right] \leqslant \frac{9}{n}\log\left(\frac{1}{\delta}\right).$$

结合这些界，我们证明以至少 $1-\delta$ 的概率，

$$\mathop{\mathbb{E}}_{X\sim P}\left[|\widehat{f}_{\mathrm{ERM}}(X) - f^*(X)|\right] \leqslant \gamma^* + \frac{9}{n}\log\left(\frac{1}{\delta}\right).$$

现在任取一个 $\gamma \in [0,1]$。如果 $\gamma > \gamma^*$，那么显然有 $\gamma^* \leqslant \gamma + \mathrm{P}_{\mathcal{H}(c_2\gamma)}\, n$，结论即得。否则，如果 $\gamma \leqslant \gamma^*$，则 $\gamma^* \leqslant \frac{9c_4}{n}(\mathrm{P}_{\mathcal{H}}(c_2\gamma^* + 1 \leqslant 9c_4))\,(\mathrm{P}_{\mathcal{H}}(c_2\gamma)+1)\, n$。通过调整常数即可得到结论。 □

B.2 引理 2.1 的证明

在深入细节之前，我们先给出一些定义。由$\mathcal{H}(\gamma)$ 的另一种等价定义（由 [BL98], 提出）自然引出了部分假设类的概念，并有助于建立与单包含图算法的联系。对于固定的间隔 γ 和阈值τ，我们将阈值算子定义如下：

$$\psi_{\gamma,\tau}(z) = \begin{cases} 0 & \text{if } z \leqslant \tau - \gamma, \\ 1 & \text{if } z \geqslant \tau + \gamma, \\ \star & \text{otherwise.} \end{cases}$$

现在，我们从这一视角正式重新引入复杂度度量。

[12]定理 1.3 的结果针对 L_2 距离进行了表述，但根据[RV06], 第 607 页的讨论，它同样适用于所有 L_p, $1 \leqslant p < \infty$ 距离，只需对绝对常数作适当调整。在 [MV03]中也进行了类似的推广。

定义 B.1. 对于 $\gamma > 0$，如果存在某个 $\tau \in [0,1]$，使得序列 $(x_i)_{i=1}^d \subseteq \mathcal{X}$ 被函数类 $\mathcal{H}$ V_γ-打散，则称该序列为 V_γ-打散的。

$$\{0,1\}^d \subseteq \{(\psi_{\gamma,\tau}(f(x_1)), \ldots, \psi_{\gamma,\tau}(f(x_d))) : f \in \mathcal{H}\}.$$

我们将 $\mathrm{V}_{\mathcal{H}}(\gamma)$ 定义为由 $\mathcal{F}$ 打散的最长 V_γ-打散序列的长度。

与部分假设设置相比，我们在离散预测 $\{0,1,\star\}$ 上选择一个特定的损失函数，该函数对在 $\star$ 上的误分类不敏感。我们定义

$$\ell_b(\widehat{y}, y) = \mathbf{1}\{\widehat{y} \neq y \text{ and } y \neq \star\}. \tag{15}$$

现在，我们可以按照 $\mathcal{G} \subseteq \{0,1,\star\}^{\mathcal{X}}$中的相同方式定义部分假设类的 VC 维数第 2.4 节。我们回顾 [BL98], 中的单包含预测器的性质，该预测器扩展了 Haussler、Littlestone 和 Warmuth 的经典单包含图算法[HLW94]。

引理 B.2 ([BL98])。 固定一个假设类 $\mathcal{H} \subseteq \{0,1,\star\}^{\mathcal{X}}$，其 VC 维数为 d。令 $f^* \in \mathcal{G}$，并考虑损失函数(15)。存在一个预测器 $\widehat{f}$，使得对于任意可实现样本 $S = (x_i,\ f^*(x_i))_{i=1}^n$，其满足假设 2.1 和2.3，其中 $M_n \leqslant d$。

我们不深入探讨细节，而是非正式地说明预测器 $\widehat{f}$ （见引理 B.2 ）的运作方式。观察到一个大小为 $n-1$的可实现样本后，我们只关注满足$f^*(x_i) \neq \star$的索引。我们仅将假设类投影到这些索引上。因此，问题转化为一个取值于 $\{0,1\}$的分类问题，并使用 Haussler、Littlestone 和 Warmuth n提出的标准单包含图预测策略 [（HLW）94]来预测第 $y_n = \star$个观测的标签。注意，若 $y_n = \star$，我们的损失始终为零，因此我们确实将问题限制在二分类问题范围内。

现在我们重新陈述并证明引理 2.1。

引理 2.1。给定一个假设类 $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$。存在一个预测器 $\widehat{f}\colon\ \mathcal{X} \times \mathcal{U} \to \mathcal{Y}$对于任意可实现样本 $S = (x_i, y_i)_{i=1}^n$，满足假设 2.1 和2.3，且其留一法误差界为

$$M_n \leqslant (n+1)\gamma + \mathrm{V}_{\mathcal{H}}(\gamma).$$

证明。 定义 $m = \lceil 4n/\gamma \rceil$和阈值 $\tau_i = i\rho$ 对于 $i \in [m]$ 和 $\rho = 1/m$。现在，令 $\widehat{g}_i$ 表示满足 引理 B.2 关于阈值化函数类 $\mathcal{G}_i = \{\psi_{\gamma,\tau_i}(f)$ 的条件的预测器: $f \in \mathcal{H}\}$因此，对于可实现样本 $S = ((x_i, f^*(x_i)))_{i=1}^n$，对于任意 $i \in [m]$，我们有

$$\sum_{j=1}^n \ell_b(\widehat{g}_i(x_j; S^{-j}), \psi_{\gamma,\tau_i}(f^*(x_j))) \leqslant \mathrm{V}_{\mathcal{H}}(\gamma).$$

我们的预测器定义为

$$\widehat{f}(x; S) = \rho \sum_{i=1}^m \mathbf{1}\{\widehat{g}_i(x; S) = 1\}.$$

注意，对于任意 $z \in [0,1]$，我们有

$$\rho \sum_{i=1}^m \mathbf{1}\{z \geqslant \tau_i\} \leqslant z \leqslant \rho + \rho \sum_{i=1}^m \mathbf{1}\{z \geqslant \tau_i\}$$

现在，固定 $x \in \mathcal{X}$，并令 $\mathcal{I} = \{i \in [m] : \psi_{\gamma,\tau_i}(f^*(x)) = \star\}$，并令 $i_l = \min \mathcal{I}$ 和 $i_h = \max \mathcal{I}$。注意，对任意 $i \notin \mathcal{I}$，都有

$$\mathbf{1}\{f^*(x) \geqslant \tau_i\} = \mathbf{1}\{\psi_{\gamma,\tau_i}(f^*(x)) = 1\} \text{ and } \mathcal{I} = \{i : i_l \leqslant i \leqslant i_h\},$$

我们有

$$
\begin{aligned}
f^*(x) - \widehat{f}(x;S) &\leqslant \rho + \rho\left(\sum_{i=1}^{i_l-1} \mathbf{1}\left\{f^*(x) \geqslant \tau_i\right\} - \mathbf{1}\left\{\widehat{g}_i(x;S) = 1\right\}\right) \\
&\quad + \rho\left(\sum_{i=i_l}^{i_h} \mathbf{1}\left\{f^*(x) \geqslant \tau_i\right\} - \mathbf{1}\left\{\widehat{g}_i(x;S) = 1\right\}\right) \\
&\quad + \rho\left(\sum_{i=i_h+1}^{m} \mathbf{1}\left\{f^*(x) \geqslant \tau_i\right\} - \mathbf{1}\left\{\widehat{g}_i(x;S) = 1\right\}\right) \\
&\leqslant \rho + \rho\sum_{i=1}^{m} \mathbf{1}\left\{\widehat{g}_i(x;S) \neq \psi_{\gamma,\tau_i}(f^*(x)) \text{ and } \psi_{\gamma,\tau_i}(f^*(x)) \neq \star\right\} \\
&\quad + \rho\left(\sum_{i=i_l}^{i_h} \mathbf{1}\left\{f^*(x) \geqslant \tau_i\right\} - \mathbf{1}\left\{\widehat{g}_i(x) = 1\right\}\right) \\
&\leqslant \rho + \rho\sum_{i=1}^{m} \ell_b(\widehat{g}_i(x;S), \psi_{\gamma,\tau_i}(f^*(x))) + \rho\sum_{i=i_l}^{i_h} \mathbf{1}\left\{f^*(x) \geqslant \tau_i\right\} \\
&\leqslant 2\rho + \gamma + \rho\sum_{i=1}^{m} \ell_b(\widehat{g}_i(x;S), \psi_{\gamma,\tau_i}(f^*(x))).
\end{aligned}
$$

其中，最后一个不等式是根据这一观察得出的

$$
\tau_{i_l} \geqslant f^*(x) - \gamma \implies \sum_{i=i_l}^{i_h} \mathbf{1}\left\{f^*(x) \geqslant \tau_i\right\} \leqslant \frac{\gamma}{\rho} + 1.
$$

同理，对于下界，定义 $i_m = \max\{i : f^*(x) \geqslant \tau_i\}$，我们有

$$
\begin{aligned}
f^*(x) - \widehat{f}(x;S) &\geqslant \rho\left(\sum_{i=1}^{i_l-1} \mathbf{1}\left\{f^*(x) \geqslant \tau_i\right\} - \mathbf{1}\left\{\widehat{g}_i(x;S) = 1\right\}\right) \\
&\quad + \rho\left(\sum_{i=i_l}^{i_h} \mathbf{1}\left\{f^*(x) \geqslant \tau_i\right\} - \mathbf{1}\left\{\widehat{g}_i(x;S) = 1\right\}\right) \\
&\quad + \rho\left(\sum_{i=i_h+1}^{m} \mathbf{1}\left\{f^*(x) \geqslant \tau_i\right\} - \mathbf{1}\left\{\widehat{g}_i(x;S) = 1\right\}\right) \\
&\geqslant -\rho - \rho\sum_{i=1}^{m} \mathbf{1}\left\{\widehat{g}_i(x;S) \neq \psi_{\gamma,\tau_i}(f^*(x)) \text{ and } \psi_{\gamma,\tau_i}(f^*(x)) \neq \star\right\} \\
&\quad + \rho\left(\sum_{i=i_l}^{i_h} \mathbf{1}\left\{f^*(x) \geqslant \tau_i\right\} - \mathbf{1}\left\{\widehat{g}_i(x) = 1\right\}\right) \\
&\geqslant -\rho - \rho\sum_{i=1}^{m} \ell_b(\widehat{g}_i(x;S), \psi_{\gamma,\tau_i}(f^*(x))) - \rho(i_h - i_m) \\
&\geqslant -3\rho - \gamma - \rho\sum_{i=1}^{m} \ell_b(\widehat{g}_i(x;S), \psi_{\gamma,\tau_i}(f^*(x))),
\end{aligned}
$$

其中最后一个不等式由此得出

$$
\tau_{i_h} - \tau_{i_m} \leqslant \gamma + \rho \implies i_h - i_m \leqslant \frac{\gamma}{\rho} + 2.
$$

根据上述两个界以及 引理 B.2，可得

$$\sum_{i=1}^{n} \ell(\widehat{f}(x; S^{-i}), f^*(x_i)) \leqslant n\gamma + 3n\rho + \rho m \mathrm{V}_{\mathcal{H}}(\gamma) \leqslant (n+1)\gamma + \mathrm{V}_{\mathcal{H}}(\gamma).$$

证明至此完成。 □

B.3 定理 2.3 的证明定理 2.3

定理 2.3。固定假设类 $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ 和间隔 $\gamma \in (0,1)$。存在一个预测器 $\widehat{f}\colon \mathcal{X} \times \mathcal{U} \to [0,1]$，对于任意 $\delta \in (0,1)$，以及从任意可实现分布 P 中采样的 $S \sim P^n$，满足

$$\mathop{\mathbb{E}}_{(X,Y)\sim P}\left[\left|\widehat{f}(X;S) - Y\right|\right] \leqslant 4.82\left(\gamma + \frac{\mathrm{V}_{\mathcal{H}}(\gamma)}{n} + \frac{1}{n}\log\left(\frac{2e}{\delta}\right)\right),$$

以至少 $1-\delta$ 的概率（关于 S 的随机性）。

证明。不失一般性，我们假设在可实现情形下，有 $Y = f^*(X)$，其中$f^* \in \mathcal{H}$。利用绝对损失的凸性、定理 2.1 和引理 2.1，我们以至少 $1-\delta$的概率得到，

$$\begin{aligned}\mathop{\mathbb{E}}_{X}\left[\left|\frac{4}{3n}\sum\nolimits_{t=n/4}^{n-1}\widehat{f}(X;S_{\leqslant t}) - f^*(X)\right|\right] &\leqslant \frac{4}{3n}\sum\nolimits_{t=n/4}^{n-1}\mathop{\mathbb{E}}_{X}\left[\left|\widehat{f}(X;S_{\leqslant t}) - f^*(X)\right|\right] \\ &\leqslant 4.82\left(\gamma + \frac{\gamma}{n} + \frac{\mathrm{V}_{\mathcal{H}}(\gamma)}{n} + \frac{1}{n}\log\left(\frac{2}{\delta}\right)\right).\end{aligned}$$

选择预测器 $\widehat{f}(\cdot) = \frac{4}{3n}\sum_{t=n/4}^{n-1}\widehat{f}(\cdot; S_{\leqslant t})$，并注意到 $\gamma < 1$，即可得到该结论。 □

C 第 4 节的证明：主要结果

C.1 引理 4.1的证明：鞅集中不等式

引理 4.1. 设 $W_1, \ldots, W_T$ 是一个适应于滤过 $(\mathcal{F}_i)_{i\leqslant T}$ 的随机过程。假设 0 6$W_t \leqslant 1$ 几乎处处成立。那么，对于任意 $\delta, \lambda, \eta \in (0,1)$，有

$$\mathbf{Pr}\left[\sum_{t=1}^{T} W_t \geqslant \frac{e^\lambda - 1}{\lambda}\sum_{t=1}^{T}\mathbb{E}[W_t|\mathcal{F}_{t-1}] + \frac{\log(1/\delta)}{\lambda}\right] \leqslant \delta, \tag{8}$$

and

$$\mathbf{Pr}\left[\sum_{t=1}^{T}\mathbb{E}[W_t|\mathcal{F}_{t-1}] \geqslant \frac{\eta e^\eta}{e^\eta - 1}\sum_{t=1}^{T} W_t + \frac{e^\eta \log(1/\delta)}{e^\eta - 1}\right] \leqslant \delta. \tag{9}$$

证明。 我们首先证明（8）。固定 $\lambda \in (0,1]$。我们有

$$\begin{aligned}\mathbb{E}\left[e^{\lambda W_t} \mid \mathcal{F}_{t-1}\right] &\leqslant \mathbb{E}\left[1 + (e^\lambda - 1)W_t \mid \mathcal{F}_{t-1}\right] &(16)\\ &= 1 + (e^\lambda - 1)\,\mathbb{E}\left[W_t \mid \mathcal{F}_{t-1}\right] \\ &\leqslant \exp((e^\lambda - 1)\,\mathbb{E}[W_t \mid \mathcal{F}_{t-1}]), &(17)\end{aligned}$$

其中我们使用了以下事实： $e^{\lambda x} \leqslant 1 + (e^\lambda - 1)x$ 对于任意 $0 \leqslant x \leqslant 1$ 属于（16）的 $1 + x \leqslant e^x$ 对于任意 x 属于（17）。令 $\alpha = (e^\lambda - 1)$。现在我们可以轻松得出

$$\mathbb{E}\left[\prod_{t=1}^{T} e^{\lambda W_t - \alpha\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]}\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{t=1}^{T} e^{\lambda W_t - \alpha\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]} \,\middle|\, \mathcal{F}_{T-1}\right]\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\prod_{t=1}^{T-1} e^{\lambda W_t - \alpha\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]} \cdot \mathbb{E}\left[e^{\lambda W_T - \alpha\,\mathbb{E}[W_T|\mathcal{F}_{T-1}]} \;\middle|\; \mathcal{F}_{T-1}\right]\right] \\
&= \mathbb{E}\left[\prod_{t=1}^{T-1} e^{\lambda W_t - \alpha\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]} \cdot e^{-\alpha\,\mathbb{E}[W_T|\mathcal{F}_{T-1}]}\,\mathbb{E}\left[e^{\lambda W_T} \;\middle|\; \mathcal{F}_{T-1}\right]\right] \\
&\leqslant \mathbb{E}\left[\prod_{t=1}^{T-1} e^{\lambda W_t - \alpha\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]}\right] && \text{(using (17))} \\
&\leqslant \mathbb{E}\left[e^{\lambda W_1 - \alpha\,\mathbb{E}[W_1|\mathcal{F}_0]} \cdot 1\right] && \text{(using induction)} \\
&\leqslant 1.
\end{aligned}
$$

因此，我们可以对随机变量 $\prod_{t=1}^{T} e^{\lambda W_t - \alpha\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]}$ 应用马尔可夫不等式，从而得出

$$
\mathbf{Pr}\left[\sum_{t=1}^{T} \lambda W_t - \alpha\,\mathbb{E}[W_t \mid \mathcal{F}_{t-1}] \geqslant \log(1/\delta)\right] \leqslant \delta.
$$

（8）重新整理后即可得到。

现在证明（9）。固定 $\eta \in (0, 1]$。我们有

$$
\begin{aligned}
\mathbb{E}\left[e^{-\eta W_t} \mid \mathcal{F}_{t-1}\right] &\leqslant \mathbb{E}\left[1 - (1 - e^{-\eta}) W_t \mid \mathcal{F}_{t-1}\right] && (18) \\
&= 1 - (1 - e^{-\eta})\,\mathbb{E}\left[W_t \mid \mathcal{F}_{t-1}\right] \\
&\leqslant \exp(-(1 - e^{-\eta})\,\mathbb{E}[W_t \mid \mathcal{F}_{t-1}]), && (19)
\end{aligned}
$$

其中我们使用了以下事实：对于任意 $e^{-\eta x} \leqslant 1 - (1 - e^{-\eta})\, x$，均有 $0 \leqslant x \leqslant 1$ 属于（18）；并且对于任意 $1 + x \leqslant e^{x}$，均有 x 属于（19）。令 $\beta = (1 - e^{-\eta})$。现在可以得出

$$
\begin{aligned}
\mathbb{E}\left[\prod_{t=1}^{T} e^{-\eta W_t + \beta\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]}\right] &= \mathbb{E}\left[\mathbb{E}\left[\prod_{t=1}^{T} e^{-\eta W_t + \beta\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]} \;\middle|\; \mathcal{F}_{T-1}\right]\right] \\
&= \mathbb{E}\left[\prod_{t=1}^{T-1} e^{-\eta W_t + \beta\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]} \cdot \mathbb{E}\left[e^{-\eta W_T + \beta\,\mathbb{E}[W_T|\mathcal{F}_{T-1}]} \;\middle|\; \mathcal{F}_{T-1}\right]\right] \\
&= \mathbb{E}\left[\prod_{t=1}^{T-1} e^{-\eta W_t + \beta\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]} \cdot e^{\beta\,\mathbb{E}[W_T|\mathcal{F}_{T-1}]}\,\mathbb{E}\left[e^{-\eta W_T} \;\middle|\; \mathcal{F}_{T-1}\right]\right] \\
&\leqslant \mathbb{E}\left[\prod_{t=1}^{T-1} e^{-\eta W_t + \beta\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]}\right] && \text{(using (19))} \\
&\leqslant \mathbb{E}\left[e^{-\eta W_1 + \beta\,\mathbb{E}[W_1|\mathcal{F}_0]} \cdot 1\right] && \text{(using induction)} \\
&\leqslant 1.
\end{aligned}
$$

因此，我们可以对随机变量 $\prod_{t=1}^{T} e^{-\eta W_t + \beta\,\mathbb{E}[W_t|\mathcal{F}_{t-1}]}$ 应用马尔可夫不等式，从而得到

$$
\mathbf{Pr}\left[-\sum_{t=1}^{T} \eta W_t + \beta\,\mathbb{E}[W_t \mid \mathcal{F}_{t-1}] \geqslant \log(1/\delta)\right] \leqslant \delta.
$$

（9）重新整理后即可得到。 □