arXiv:2305.05248v2 [cs.LG] 2 Jun 2023

迈向理解　　多标签学习中宏平均AUC的泛化

吴国强 [1] 李崇轩 [2] 殷一龙 [1]

摘要

宏平均AUC是多标签学习中各类别AUC的算术平均值，在实践中得到了广泛应用。然而，对其理论理解仍远远不足。为解决这一问题，我们基于与宏平均AUC相对应的替代损失，刻画了各种学习算法的泛化性质。我们从理论上确定了影响泛化界的一个关键数据集因素：标签级类别不平衡。关于考虑不平衡因素的误差界，我们的结果表明，广泛使用的基于单变量损失的算法比本文提出的基于成对损失和重加权损失的算法对标签级类别不平衡更为敏感，这可能意味着其性能较差。此外，在多个数据集上的实证结果证实了我们的理论发现。为建立这一理论，我们在技术上提出了一种新的（且更一般的）McDiarmid 型集中不等式，该结果本身也可能具有独立研究价值。

1. 引言

多标签学习（MLC）（McCallum，1999）是 Machine Learning 中一项重要的学习任务，其中每个实例可能关联多个标签。它已广泛应用于各个领域，例如自然语言处理（Schapire& Singer，2000）、计算机视觉（Carneiro 等，2007）和生物信息学（Elisseeff& Weston，2001）。由于多标签分类的复杂性以及不同场景的多样化需求，人们开发了多种评估指标（Zhang& Zhou，2014；Wu & Zhou，2017），以进行全面评估，例如汉明损失、排序损失和子集准确率。其中，宏平均AUC（Zhang & Zhou，2014）是实践中广泛使用的一种指标。通俗地说，它是各类别 AUC 指标的算术平均值，这正是本文的关注重点。

多标签分类（MLC）中的宏平均AUC（以及许多其他评价指标）具有不连续性和非凸性，因此，直接对其进行优化可能会导致NP难问题（Arora& Barak，2009）。因此，实践中通常采用许多替代损失，以提高计算效率。从经验上看，许多基于代理损失的学习算法通常使用宏平均AUC进行评估，其中包括广泛使用的基于替代单变量损失的算法（Boutell 等人，2004；Wu & Zhu，2020），这些算法最初旨在优化汉明损失。然而，从理论上讲，目前对这一问题的理解仍远远不够。为推动这一问题的解决，本文尝试正式回答以下问题：

> 基于广泛使用的替代门控单变量损失的算法，在宏平均AUC方面能提供怎样的学习保证?

为回答上述问题，我们提出了一个分析框架，用于刻画学习算法在宏平均AUC方面的泛化性质。受理论分析启发，我们还针对宏平均AUC提出了一种成对代理损失和一种重加权单变量损失。从理论上讲，我们分析了采用这三种损失的算法的学习保证。我们从理论上指出，标签级类别不平衡是多标签分类（MLC）数据集的一个因素（Tarekegn 等人，2021年；Zhang 等人，2020b），在这些泛化界中起着关键作用。

具体而言，基于成对损失的学习算法 A^{pa} 具有能够感知标签级类别不平衡的学习保证，其界限为 $O(\frac{1}{\sqrt{n}}(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}))$（见表 1），其中 n 表示样本量，K 表示标签数量，$\tau_k \in [\frac{1}{n},\frac{1}{2}]$ 刻画第 k 个标签的类别不平衡程度。τ_k 越小，不平衡程度越高。相比之下，广泛使用的基于单变量损失的算法 A^{u_1} 的误差界为 $O(\frac{1}{\tau_S^*\sqrt{n}}(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}))$，其中 $\tau_S^* = \mathrm{argmin}_{k\in[K]}\tau_k$。因此，我们可以观察到，相较于 A^{pa}，A^{u_1} 对标签级类别不平衡更加敏感，这意味着在实际应用中，A^{pa} 的表现可能优于 A^{u_1}，尤其是在 $\frac{1}{\tau_S^*}$ 较大的情况下，而这种情况在多标签分类（MLC）的真实数据集中经常出现。需要注意的是，在计算方面，

[1]山东大学软件学院 [2]中国人民大学高瓴人工智能学院；北京大数据管理与分析方法重点实验室，中国北京。通讯作者：吴国强 <guoqiangwu@sdu.edu.cn>。

发表于 40th 国际机器学习会议论文集，美国夏威夷州檀香山。PMLR 202，2023 年。版权所有 © 2023，作者所有。

表1. 主要理论结果总结。本文的贡献以红色突出显示。

算法	代理损失	泛化界	计算
$\mathcal{A}^{pa}$	成对（L_{pa}）	$\widehat{R}_S^{pa}(f) + O(\frac{1}{\sqrt{n}}(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}))$	$O(n^2K)$
$\mathcal{A}^{u_1}$（Boutell 等人，2004）	单变量（L_{u_1}）	$\frac{1}{\tau_S^*}\widehat{R}_S^{u_1}(f) + O(\frac{1}{\tau_S^*\sqrt{n}}(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}))$	$O(nK)$
$\mathcal{A}^{u_2}$	重加权单变量（L_{u_2}）	$\widehat{R}_S^{u_2}(f) + O(\frac{1}{\sqrt{n}}(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}))$	$O(nK)$

A^{pa} 会导致复杂度达到 $O(n^2K)$，这高于 A^{u_1}（即 $O(nK)$）；当样本量 n 较大时，其计算成本可能高得令人难以承受。有趣的是，我们提出的重加权单变量损失算法 A^{u_2} 具有 $O(\frac{1}{\sqrt{n}}(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}))$ 的泛化界，与 A^{pa}几乎相同。这可能表明 A^{u_2} 的性能优于 A^{u_1}，且计算效率更高。最后，实证结果证实了我们的理论发现。

严格来说，由于优化宏平均AUC可能涉及使用相关样本进行学习，因此，针对多标签分类中其他度量的现有泛化分析技术（Wu &Zhu，2020；Wu 等人，2021年）无法应用，这使得问题更具挑战性。受二部排序（BR，或等价地，二元分类中的 AUC 最大化）技术的启发（Usunieret al.，2005；Amini & Usunier，2015年），我们将其扩展到多标签分类中的宏平均AUC最大化。需要注意的是，由于多标签分类具有多个标签（或任务）的特性，二部排序中的技术无法直接应用于多标签分类。[1] 因此，我们提出了通用技术，包括一种新的 McDiarmid 型集中不等式，以及一个针对图依赖样本的多任务学习的一般泛化界，这些技术本身也可能具有独立的研究价值。（详见附录 A 。）我们对多标签分类中宏平均AUC最大化的泛化分析，可视为这些通用技术的一种应用。

2. 预备知识

记号。用粗体小写字母和大写字母分别表示向量（例如 $\mathbf{a}$）和矩阵（例如 $\mathbf{A}$）。对于矩阵 $\mathbf{A}$，$\mathbf{a}_i$、$\mathbf{a}^j$ 和 a_{ij} 分别表示其第 i 行、第 j 列以及第（i，j）个元素。对于向量 $\mathbf{a}$，a_i 表示其第 i 个元素。令 $[K]$ 表示集合 $\{1,\ldots,K\}$。对于集合，$|\cdot|$ 表示其基数。$[\![\cdot]\!]$ 表示指示函数，即当命题成立时返回 1，否则返回 0。

[1]注意，可以使用并集界合并 BR 中原有的界，从而得到针对多标签分类中宏平均AUC的所需界，但这会导致一个包含项 $\log(\frac{K}{\delta})$ 的松散界。

2.1. 问题设定

令 $\mathbf{x} \in \mathbf{X} \subset \mathbb{R}^d$ 和 $\mathbf{y} \in \mathbf{Y} \subset \{-1,+1\}^K$ 分别表示输入和输出，其中 d 为特征维度，K 为标签数量。$y_k = 1$（或 -1）表示相应的第 k 个标签是相关的（或不相关的）。给定一个训练集 $S = \{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n$，其中包含从分布 P中抽取的 n 个独立同分布样本，且该分布定义在 $\mathbf{X} \times \mathbf{Y}$ 上，MLC 的原始目标是学习一个多标签分类器 $H : \mathbb{R}^d \to \{-1,+1\}^K$。

为了解决多标签分类问题，一种标准方法是首先学习一个基于向量的 评分函数 (或预测器) $f = [f_1,\ldots,f_K] : \mathcal{X} \to \mathbb{R}^K$，其来自假设空间 $\mathcal{F}$，然后通过阈值函数得到分类器H。多标签分类的一个典型目标是，根据某种基于排序的度量，从有限的训练数据中学习最佳预测器，这通常称为多标签排序（Dembczynski 等人，2012年；Wu 等人，2021年），这也是本文的研究重点。

2.2. 评估度量

人们开发了许多评估度量来评估不同算法的性能。这里我们重点关注常用的宏平均AUC，它对所有类别标签的AUC度量取宏平均。给定数据集 S 和预测器 $f \in \mathrm{F}$，宏平均AUC定义如下：[2]

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-}[\![f_k(\mathbf{x}_p) > f_k(\mathbf{x}_q)]\!],$$

其中，S_k^+（或 S_k^-）表示标签 k 的相关（或不相关）实例索引集。

最大化宏平均AUC等价于最小化以下目标函数（即 1 减去宏平均AUC）：

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-}L_{0/1}(\mathbf{x}_p,\mathbf{x}_q,f_k), \quad (1)$$

其中，0/1 损失函数定义为

$$L_{0/1}(\mathbf{x}^+,\mathbf{x}^-,f_k) = [\![f_k(\mathbf{x}^+) \le f_k(\mathbf{x}^-)]\!], \quad (2)$$

[2]请注意，这里我们不采用关于等式的另一种常见形式（即 $[\![f_k(\mathbf{x}_p) \ge f_k(\mathbf{x}_q)]\!]$），以避免试探性零假设 f。此外，在评估算法时，这两种形式在实际应用中几乎相同。

其中，$\mathbf{x}^+$（或 $\mathbf{x}^-$）表示标签 k的相关（或不相关）输入。

2.3. 风险

由于宏平均AUC（或0/1 损失函数）是不连续且非凸的，直接对其进行优化会导致 NP 难问题（Arora& Barak, 2009）。在实践中，为提高计算效率，通常会寻求（凸）替代损失。设 $L_\phi : \mathcal{X} \times \mathcal{X} \times \mathcal{F}_k \to \mathbb{R}_+$ 表示一个替代损失函数，其中$\mathcal{F}_k = \{f_k : \mathcal{X} \to \mathbb{R}\}$，其具体形式将在下一节讨论。对于预测器$f \in \mathcal{F}$，相对于宏平均AUC的真实（0/1）泛化（或期望）风险定义如下：

$$R_{0/1}(f) = \frac{1}{K}\sum_{k=1}^{K} \mathop{\mathbb{E}}_{\mathbf{x}_p \sim P_k^+, \mathbf{x}_q \sim P_k^-} \left[L_{0/1}(\mathbf{x}_p, \mathbf{x}_q, f_k) \right],$$

其中，条件分布为 $P_k^+ = P(\mathbf{x}|y_k = 1)$ 和$P_k^- = P(\mathbf{x}|y_k = -1)$。此外，相对于 L_ϕ 的替代经验风险和泛化风险分别定义如下：

$$\widehat{R}_S^\phi(f) = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{|S_k^+||S_k^-|} \sum_{(p,q)\in S_k^+ \times S_k^-} L_\phi(\mathbf{x}_p, \mathbf{x}_q, f_k),$$

$$R_\phi(f) = \mathop{\mathbb{E}}_{S}\left[\widehat{R}_S^\phi(f)\right]. \tag{3}$$

注意，我们并未将替代泛化风险定义为以下常见形式

$$\frac{1}{K}\sum_{k=1}^{K} \mathop{\mathbb{E}}_{\mathbf{x}_p \sim P_k^+, \mathbf{x}_q \sim P_k^-} L_\phi(\mathbf{x}_p, \mathbf{x}_q, f_k). \tag{4}$$

这是因为，公式(3)比公式(4)更一般，其中公式(3)可以涵盖依赖于训练数据集 L_ϕ 的代理损失 S，而公式(4)则不能。[3] 此外，对于某些与 S 无关的损失，两者相等。

3. 方法

本节介绍所考虑的替代损失及其对应的学习算法。

3.1. 替代损失

为了优化宏平均AUC，自然可以使用以下（替代）成对损失：

$$L_{pa}(\mathbf{x}^+, \mathbf{x}^-, f_k) = \ell\left(f_k(\mathbf{x}^+) - f_k(\mathbf{x}^-)\right), \tag{5}$$

其中，基础损失函数 $\ell(t)$可以是许多常见的（基于间隔的）损失函数，例如合页损失$\ell(t) = \max(0, 1-t)$、逻辑函数 $\ell(t) = \log_2(1+\exp(-t))$ 等。基础损失的一个自然性质是，它是原始 0/1 损失的上界，即$\ell(t) \geq [\![t \leq 0]\!]$。请注意，最小化这种基于成对损失的风险会导致计算复杂度为 $O(n^2)$，当样本量 n较大时，计算成本可能高得令人难以承受。[4]

最初旨在优化汉明损失指标的广泛使用的单变量损失（Boutell 等人，2004；Wu & Zhu，2020）也可以被视为一种代理损失L_{u_1}，用于宏平均AUC。其原始经验风险可写为：

$$\widehat{R}_S^{u_1}(f) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}\ell(y_{ik}f_k(\mathbf{x}_i)) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|S_k^+||S_k^-|}\times \sum_{(p,q)\in S_k^+\times S_k^-}\left(\frac{|S_k^+|}{n}\ell(f_k(\mathbf{x}_p)) + \frac{|S_k^-|}{n}\ell(-f_k(\mathbf{x}_q))\right).$$

因此，我们可以定义关于宏平均AUC的这种单变量代理损失 L_{u_1} 如下：

$$L_{u_1}(\mathbf{x}^+, \mathbf{x}^-, f_k) = \frac{|S_k^+|}{n}\ell\left(f_k(\mathbf{x}^+)\right) + \frac{|S_k^-|}{n}\ell\left(-f_k(\mathbf{x}^-)\right). \tag{6}$$

注意，这种代理损失无法严格上界0/1 损失，即 $L_{0/1} \nleq L_{u_1}$。代理损失关于0/1 损失的上界性质对于提供其关于0/1 损失的泛化分析至关重要，我们将在后文详细讨论这一点。此外，请注意，由于泛化风险依赖于数据集 L_{u_1}，我们无法通过公式（4）定义关于 S的泛化风险。

为了上界0/1 损失（即 $L_{0/1}$），我们在此提出一种具有计算效率的新型重加权单变量代理损失 L_{u_2}，其定义如下：

$$L_{u_2}(\mathbf{x}^+, \mathbf{x}^-, f_k) = \ell\left(f_k(\mathbf{x}^+)\right) + \ell\left(-f_k(\mathbf{x}^-)\right). \tag{7}$$

然后，我们可以将其经验风险写为

$$\begin{aligned}\widehat{R}_S^{u_2}(f) &= \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-} L_{u_2}(\mathbf{x}_p, \mathbf{x}_q, f_k) \\ &= \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{n}\left([\![y_{ik}=1]\!]\frac{1}{|S_k^+|}\ell(f_k(\mathbf{x}_i)) + [\![y_{ik}\neq 1]\!]\frac{1}{|S_k^-|}\ell(-f_k(\mathbf{x}_i))\right).\end{aligned}$$

我们可以看到，在计算上，关于 L_{u_2} 最小化经验风险可能会导致 $O(n)$ 的复杂度，这

[3]请注意，我们定义公式(3)的目的是便于对 L_{u_1} 进行分析，并最终得到仅依赖于数据集的界限。若采用常见形式定义，最终得到的界限将包含依赖于分布的项。

[4]请注意，对于某些基础损失，可以采用多种方法加速计算。

与 L_{u_1} 相同。直观而言，这种重新加权损失可以被视为一种针对平衡准确率进行优化的代价敏感损失。

这些损失之间存在一些关系，我们将在下一节中进行详细讨论。

3.2. 学习算法

在接下来的分析中，我们聚焦于基于核的学习算法。这类算法已广泛应用于多标签分类（MLC）的实践（Elisseeff & Weston, 2001; Boutell 等人, 2004; Hariharan 等人, 2010; Tan 等人, 2020; Wu 等人, 2020a）和理论研究（Wu & Zhu, 2020; Wu 等人, 2021）。需要指出的是，我们接下来的分析也可以扩展到其他形式的假设空间，例如神经网络（Anthony & Bartlett, 1999）。令 $\kappa : \mathrm{X} \times \mathrm{X} \to \mathbb{R}$ 为一个正定对称核，并将其诱导的再生核希尔伯特空间（RKHS）记为 $\mathbb{H}$。令 $\Phi : \mathrm{X} \to \mathbb{H}$ 为与 κ 相关联的特征映射。所考虑的基于核的假设类可定义为

$$\mathcal{F} = \left\{ \mathbf{x} \mapsto \mathbf{W}^\top \Phi(\mathbf{x}) : \mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)^\top, \|\mathbf{w}_k\| \le \Lambda \right\}, \tag{8}$$

其中，$\|\mathbf{w}_k\|$ 为方便起见表示 $\|\mathbf{w}_k\|_{\mathbb{H}}$。

这里，我们考虑以下三种正则化学习算法及其对应的上述替代损失：

$$\mathcal{A}^{pa}: \quad \min_{\mathbf{W}} \widehat{R}_S^{pa}(f) + \lambda \|\mathbf{W}\|^2,$$
$$\mathcal{A}^{u_j}: \quad \min_{\mathbf{W}} \widehat{R}_S^{u_j}(f) + \lambda \|\mathbf{W}\|^2, j = 1, 2,$$

其中，λ 表示一个权衡超参数，$\|\mathbf{W}\|$ 为方便起见表示 $\|\mathbf{W}\|_{\mathbb{H},\,2} = (\sum_{j=1}^{c} \|\mathbf{w}_j\|_{\mathbb{H}}^2)^{1/2}$。

4. 理论结果

本节主要介绍上述采用不同替代损失的学习算法关于宏平均曲线下面积度量的泛化结果，相关引理、定理和推论的证明见附录B。

具体而言，为了证明这一点，我们提出了包括一种新的McDiarmid 型不等式在内的新技术。（证明概要请参见第4.5 节，详细内容见附录A 。）

首先，我们给出以下定义，以刻画多标签分类中的标签级类别不平衡。

定义 1（标签级类别不平衡）。给定数据集 S，定义以下因子来刻画每个标签 $k \in [K]$ 的标签级类别不平衡程度：

$$\tau_k = \frac{\min\{|S_k^+|, |S_k^-|\}}{n},$$

where $\tau_k \in [\frac{1}{n}, \frac{1}{2}]$. *Besides, define* $\tau_S^* = \operatorname{argmin}_{k \in [K]} \tau_k$.

根据上述定义可以看出，τ_k 越小，标签级类别不平衡程度越高。此外，为便于后续讨论，我们给出如下定义。[5]

定义 2（按标签的类别平衡和极度不平衡数据集）。给定数据集 S，若满足 $\forall k \in [K], \tau_k = \frac{1}{2}$（或 $\tau_k = \frac{1}{n}$），则称其为按标签的类别平衡（或类别极度不平衡）数据集。[6]

接下来，我们介绍后续分析中常用的温和假设。

假设 1（常用假设）。

(1) 训练数据集 $S = \{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^n$ 是从分布 P 中独立同分布抽取的样本，其中 $\exists\, r > 0$，对于所有 $\mathbf{x} \in \mathrm{X}$，均满足 $\kappa(\mathbf{x}, \mathbf{x}) \le r^2$。(2) 假设类定义见式 (8)。(3) 基础（凸）损失 $\ell(z)$ 是 ρ-利普希茨连续的，且有界于 B。[7]

这里给出损失空间的分数拉德马赫复杂度的定义。

定义 3（损失空间的分数拉德马赫复杂度）。对于每个标签 $k \in [K]$，基于原始数据集 S_k 构造˜数据集 $S_k = \{(\tilde{\mathbf{x}}_{ki}, \tilde{y}_{ki})\}_{i=1}^{m_k} = \{((\tilde{\mathbf{x}}_{ki}^+, \tilde{\mathbf{x}}_{ki}^-), 1)\}_{i=1}^{m_k}$，其中 $(\tilde{\mathbf{x}}_{ki}^+, \tilde{\mathbf{x}}_{ki}^-) \in S_k^+ \times S_k^-$，并令 $\{(I_{kj}, \omega_{kj})\}_{j \in [J_k]}$ 为在 S_k 上构造的依赖图 G_k 的分数独立顶点覆盖，其中 $\sum_{j \in [J_k]} \omega_{kj} = \chi_f(G_k)$，而 $\chi_f(G_k)$ 是 G_k 的分数色数。对于假设空间 F 和损失函数 $L : \mathrm{X} \times \mathrm{X} \times \mathrm{F}_k \to \mathbb{R}_+$，损失空间的经验分数拉德马赫复杂度定义为

$$\widehat{\mathfrak{R}}_{\tilde{S}}^*(L \circ \mathcal{F}) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{\boldsymbol{\sigma}} \left[\frac{1}{m_k} \sum_{j \in [J_k]} \omega_{kj} \times \sup_{f \in \mathcal{F}} \left(\sum_{i \in I_{kj}} \sigma_{ki} L(\tilde{\mathbf{x}}_{ki}^+, \tilde{\mathbf{x}}_{ki}^-, f_k) \right) \right].$$

然后，我们给出后续泛化分析中使用的宏平均AUC基本定理。

[5]需要注意的是，多标签数据集也可能在标签间分布不平衡，也就是说，有些标签的正例很少，而另一些标签的正例很多。[6]为简洁起见，本文将其称为平衡或极度不平衡。[7]需要注意的是，广泛使用的合页损失和逻辑损失都是 1-利普希茨连续的。虽然指数损失和平方合页损失并非全局利普希茨连续，但它们是局部利普希茨连续的。

定理 1（Macro-AUC的基本定理）。假设损失函数 $L_\phi : \mathbf{X} \times \mathbf{X} \times \mathbf{F}_k \to \mathbb{R}_+$ 有界，且上界为 M。那么，对于任意 $\delta > 0$，以下泛化界以至少 $1-\delta$ 的概率成立，其中 S 是从大小为 n 的独立同分布样本中抽取的：

$$\forall f \in \mathcal{F},\ R_\phi(f) \le \widehat{R}_S^\phi(f) + 2\widehat{\mathfrak{R}}_{\tilde{S}}^*(L_\phi \circ \mathcal{F}) + 3M\sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}\left(\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k}}\right).$$

然后，我们分析替代损失与真实损失之间的关系，如下所示。

引理 1（替代损失与真实损失之间的关系）。假设基础损失函数是原始 0/1 损失的上界，即 $\ell(t) \ge [\![t \le 0]\!]$。那么，对于任意$f_k \in \mathbf{F}_k$和 $(\mathbf{x}^+, \mathbf{x}^-) \in S_k^+ \times S_k^-$，以下不等式成立：

$$L_{0/1}(\mathbf{x}^+, \mathbf{x}^-, f_k) \le L_{pa}(\mathbf{x}^+, \mathbf{x}^-, f_k),$$
$$L_{0/1}(\mathbf{x}^+, \mathbf{x}^-, f_k) \le L_{u_2}(\mathbf{x}^+, \mathbf{x}^-, f_k) \le \frac{1}{\tau_k} L_{u_1}(\mathbf{x}^+, \mathbf{x}^-, f_k)$$
$$\le \frac{1-\tau_k}{\tau_k} L_{u_2}(\mathbf{x}^+, \mathbf{x}^-, f_k).$$

备注。根据该引理，我们可以观察到，在最小化 L_{u_1} 时，也会最小化一个依赖于 $\frac{1}{\tau_k}$ 的 $L_{0/1}$ 的上界。此外，对于涉及 L_{u_1} 和 L_{u_2} 的第二个不等式，该界是紧的，因为当 $\tau_k = \frac{1}{2}$ 时等号成立。

根据引理1，我们可以得到替代损失与真实风险之间的关系，如下所示，这对于泛化分析至关重要。

引理 2（替代损失与真实风险之间的关系）。假设基础损失函数对原始 0/1 损失给出上界，即 $\ell(t) \ge [\![t \le 0]\!]$。那么，对于任意 $f \in \mathbf{F}$ 和任意样本 $S \overset{i.i.d.}{\sim} P$，以下不等式成立：

$$R_{0/1}(f) \le R_{pa}(f),$$
$$R_{0/1}(f) \le R_{u_2}(f) = \mathop{\mathbb{E}}_{S}\left[\widehat{R}_S^{u_2}(f)\right] \le \mathop{\mathbb{E}}_{S}\left[\frac{1}{\tau_S^*}\widehat{R}_S^{u_1}(f)\right]$$
$$\le \mathop{\mathbb{E}}_{S}\left[\frac{1-\tau_S^*}{\tau_S^*}\widehat{R}_S^{u_2}(f)\right].$$

备注。对于涉及 L_{u_1} 和 L_{u_2} 的泛化风险的第二个不等式，该界是紧的，因为当 $\tau_k = \frac{1}{2}$ 时等号成立。

接下来，为了便于清晰地讨论，我们介绍算法在标签级类别不平衡方面的泛化结果：一般、平衡和极度不平衡情形。

4.1. 一般情况

这里我们介绍算法在一般数据集上的泛化结果，这些数据集涵盖了后续讨论的平衡数据集和极度不平衡数据集。

定理 2 (一般情况下 $\mathcal{A}^{pa}$ 的学习保证)。假设损失 $L_\phi = L_{pa}$，其中 L_{pa} 定义见公式(5)。此外，假设条件 1 成立。那么，对于任意 $\delta > 0$，以至少 $1-\delta$的概率，在抽取大小为 n的独立同分布样本 S时，对于任意$f \in \mathcal{F}$，以下泛化界成立：

$$R_{0/1}(f) \le R_{pa}(f) \le \widehat{R}_{pa}(f) + \frac{4\rho r\Lambda}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right) + 3B\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\left(\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k}}\right). \quad (9)$$

根据该定理，我们可以观察到，A^{pa} 在 Macro-AUC 方面具有 $O(\frac{1}{\sqrt{n}}(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}})) \approx O(\frac{1}{\sqrt{n}}(\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k}}))$ 的标签级类别不平衡感知学习保证。

定理 3 (一般情况下的 $\mathcal{A}^{u_1}$ 学习保证)。假设损失函数 $L_\phi = \frac{1}{\tau_S^*}L_{u_1}$，其中 L_{u_1} 由公式(6)定义。此外，假设 1 成立。那么，对于任意 $\delta > 0$，以至少 $1-\delta$的概率，对大小为 S 的独立同分布样本 n进行抽样时，以下泛化界对任意 $f \in \mathcal{F}$均成立：

$$R_{0/1}(f) \le \frac{1}{\tau_S^*}\widehat{R}_{u_1}(f) + \frac{4\rho r\Lambda}{\tau_S^*\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right) + \frac{3B}{\tau_S^*}\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\left(\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k}}\right). \quad (10)$$

由此定理可知，A^{u_1} 针对宏平均AUC具有能够感知类别不平衡的学习保证 $O(\frac{1}{\tau_S^*\sqrt{n}}(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}))$。

定理 4 (一般情况下 $\mathcal{A}^{u_2}$ 的学习保证)。假设损失 $L_\phi = L_{u_2}$，其中 L_{u_2} 定义见公式(7)。此外，假设1 成立。则，对于任意 $\delta > 0$，以至少 $1-\delta$的概率，在抽取大小为 S的独立同分布样本 n时，对于任意$f \in \mathcal{F}$，以下泛化界成立：

$$R_{0/1}(f) \le R_{u_2}(f) \le \widehat{R}_{u_2}(f) + \frac{8\rho r\Lambda}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right) + 6B\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\left(\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k}}\right). \quad (11)$$

由该定理可知，A^{u_2} 在 Macro-AUC 方面具有 $O\left(\frac{1}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)\right)$ 的不平衡感知学习保证，这与 A^{pa} 几乎相同。

4.2. 平衡情形

这里我们考虑平衡数据集。请注意，在这种情况下，算法 A^{u_1} 和 A^{u_2} 完全相同，因此应具有相同的学习保证，以下推论对此进行了证实。

推论 1（平衡情况下 A^{u_1} 和 A^{u_2} 的学习保证）。假设损失 $L_\phi = 2L_{u_1} = L_{u_2}$，其中$L_{u_1}$和 L_{u_2} 分别在式 (6) 和式 (7) 中定义。此外，假设 1 成立，并且假设 S 是平衡的。那么，对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，对大小为 n 的独立同分布样本 S 进行抽样，可得对于任意 $f \in \mathrm{F}$，以下泛化界成立：

$$R_{0/1}(f) \le R_{u_2}(f) = 2R_{u_1}(f) \le \widehat{R}_{u_2}(f) + \frac{8\sqrt{2}\rho r\Lambda}{\sqrt{n}} + 6\sqrt{2}B\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}, \quad (12)$$

其中 $\widehat{R}_{u_2}(f) = 2\widehat{R}_{u_1}$({v4})。

请注意，在这种情况下，相同的误差界 $\mathcal{A}^{u_1}$ 和$\mathcal{A}^{u_2}$ 证实了我们分析的有效性。从该推$\mathcal{A}^{u_1}$ （或 $\mathcal{A}^{u_2}$）的误差界为$O\left(\sqrt{\frac{1}{n}}\right)$，这与 $\mathcal{A}^{pa}$ 相同（见附录 B.3.4中的推论 4）。

4.3. 极度不平衡情形

这里我们考虑极度不平衡的数据集。在这种情形下，泛化结果如下。

推论 2 (在极度不平衡情形下$\mathcal{A}^{pa}$ 的学习保证)。假设损失 $L_\phi = L_{pa}$，其中 L_{pa}定义见式(5)。此外，假设 1 成立，并假设 S 极度不平衡。那么，对于任意 $\delta > 0$，以至少 $1-\delta$的概率，在抽取大小为 n的独立同分布样本 S后，以下泛化界对任意$f \in \mathcal{F}$均成立：

$$R_{0/1}(f) \le R_{pa}(f) \le \widehat{R}_{pa}(f) + 4\rho r\Lambda + 3B\sqrt{\log(\frac{2}{\delta})}.$$

推论 3 (极度不平衡情形下 $\mathcal{A}^{u_1}$ 的学习保证)。 假设损失函数 $L_\phi = nL_{u_1}$，其中 L_{u_1}定义见式(6)。此外，假设 1 成立，并假设 S 处于极度不平衡情形。那么，对于任意 $\delta > 0$，以至少 $1-\delta$的概率，在抽取大小为 n的独立同分布样本 S后，对于任意$f \in \mathcal{F}$，以下泛化界成立：

$$R_{0/1}(f) \le n\widehat{R}_{u_1}(f) + 4n\rho r\Lambda + 3Bn\sqrt{\frac{\log(\frac{2}{\delta})}{2}}.$$

从上述推论中，我们可以看到 $\mathcal{A}^{pa}$ 的误差界为 $O(1)$ 关于 n，而 $\mathcal{A}^{u_1}$ 取决于 $O(n)$。此外，$\mathcal{A}^{u_2}$ 的误差界与 $\mathcal{A}^{pa}$ 相似（参见附录 5 中的推论B.4.7）。可以注意到，当 $n \to \infty$ 时，这些界都会发散。这可能有以下两个原因。一方面，极度不平衡情形下的学习确实很困难。另一方面，我们的分析技术可能并非关于 n最优，而先进的技术（例如局部 Rademacher 型复杂度（Bartlett 等人，2005））可能可以用来改进这一点。不过，这并非本文的重点；我们主要关注在同一框架下，按标签划分的类别不平衡因子的泛化效应，不同算法的阶仍然可以提供有价值的见解。

4.4. 比较与讨论

对于泛化分析，更紧的上界通常意味着可能有更好的性能（Mohri 等人，2018）。[8]在本文中，所有算法都在同一框架下进行分析，并且替代损失与真实风险（或损失）之间的不等式是紧的。因此，通过比较这些算法的上界来从理论上评估其性能是相对稳妥的。现在我们对这些算法进行如下比较。

- A^{pa} 与 A^{u_1}相比，A^{pa} 通常具有更紧的界，而 A^{u_1} 则不同。具体而言，在给定相同假设空间的情况下，通常更容易训练 $\widehat{R}_S^{pa}$，使得 $\widehat{R}_S^{pa}$ 小于 $\frac{1}{\tau_S^*}\widehat{R}_S^{u_1}$。[9] 此外，对于模型复杂度项（即最后两项），$\mathrm{A}^{pa}$ 的误差界为 $O\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)$，而 A^{u_1} 取决于 $O\left(\frac{1}{\tau_S^*}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)\right)$。
- A^{u_2} 与 A^{u_1}相比。类似地，我们认为 A^{u_2} 通常具有比 A^{u_1}更紧的界。对于第一个风险项，$\frac{1}{\tau_S^*}\widehat{R}_S^{u_1}$通常与$\widehat{R}_S^{u_2}$相当，甚至更大。[10] 对于模型复杂度项，$\mathrm{A}^{u_2}$ 的误差界为 $O\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)$，而 A^{u_1} 取决于$O\left(\frac{1}{\tau_S^*}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)\right)$。
- A^{pa} 与 A^{u_2}。A^{pa} 和 A^{u_2} 的学习保证相似或相当。具体而言，对于第一项风险项，$\widehat{R}_S^{pa}$ 通常与$\widehat{R}_S^{u_2}$相当。[11] 对于模型项，A^{pa} 和 A^{u_2} 几乎相同（

[8]需要注意的是，在比较界时，通常比较依赖变量的阶数比比较其绝对值更为合理。[9]尽管我们无法对这一论断进行形式化表达，但我们在实验中通过经验观察到了这一现象。[10]在某些情况下，第一个风险项可能大于 1，但我们仍然可以通过模型复杂度的依赖变量从误差界中获得启示。[11]需要注意的是，尽管我们无法对此进行形式化表达，但实验验证了这一点。

仅常数不同），就标签层面的类别不平衡而言。

总体而言，A^{pa} 和 A^{u_2} 相较于 A^{u_1} 的界限收紧程度很大程度上取决于 $\frac{1}{\tau_S^*}$。因此，当 $\frac{1}{\tau_S^*}$ 较大时，A^{pa} 和 A^{u_2} 的表现可能会显著优于 A^{u_1}。相反，当 $\frac{1}{\tau_S^*}$ 较小时，A^{pa} 和 A^{u_2} 的表现可能略优于 A^{u_1}，或与其大致相当。此外，A^{pa} 和 A^{u_2} 具有几乎相同的学习保证，因此其表现可能大致相当。表 3 中基准数据集上的实验结果证实了我们的分析。

请注意，在先前工作的分析框架下，可能还有另一种分析$\mathcal{A}^{u_1}$ 和$\mathcal{A}^{u_2}$ 关于宏平均AUC学习保证的方法（Wu &Zhu，2020；Wu 等人，2021年），而本文着重于在相同框架下分析三种算法，并将其留作未来工作。

该理论在现实应用中的意义。尽管现实应用中的多标签分类数据集通常存在严重的（标签级）类别不平衡问题，我们的理论在实践中仍具有重要意义。具体而言，我们关于不平衡感知界的理论结果表明，与使用原始单变量损失（例如交叉熵损失）的算法 $\mathcal{A}^{u_2}$ 相比，基于不平衡感知损失的算法 $\mathcal{A}^{u_1}$ 针对标签级类别不平衡具有更好的学习保证，这或许意味着其性能更优。这为解释现有不平衡感知重加权损失（Ridnik 等，2021年；Wu 等人，2020b）为何在实践中能在类似宏平均AUC的基于排序的度量（例如平均准确率均值（mAP））上取得良好表现，提供了有价值的见解。此外，如何针对特定度量设计更有效的不平衡感知损失，以及如何进一步收紧这些界，都将有助于开发更有效的算法。

4.5. 证明概要

这里我们主要概述第 4.1 节一般情况下泛化结果的证明概要。

证明概要：总体而言，证明主要分为以下两个步骤。

步骤 1：构建所需的通用技术（详情见附录 A ）。首先，我们提出一种新的（且更一般的）McDiarmid 型不等式（即定理 5）。然后，基于该不等式，我们针对利用图依赖样本学习多任务的问题设定，提出一个通用的泛化界（即定理 6），其中涉及损失空间的分数拉德马赫复杂度。

步骤 2：应用泛化界，获得多标签分类（MLC）中宏平均AUC 最大化（MaAUCM）的结果在步骤 1 中得到的结果（详见附录B ）。首先，我们在步骤 1 中将 MaAUCM 问题转化为图依赖样本下的多任务学习问题设定，随后得到宏平均 AUC 的基础误差界（即定理1）。接下来，我们分析替代损失与真实风险之间的关系。然后，对于每种算法，定义其假设空间特有的分数拉德马赫复杂度，并对基于核的复杂度进行上界估计（例如，引理 4），再得到特定的收缩不等式（例如，引理 5），以关联损失空间与假设空间的复杂度。最后，我们可以基于上述中间结果得到所需的结论。

4.6. 替代损失的一致性

除了有限样本泛化保证之外，替代损失的一致性也很重要。遵循（Gao 和 Zhou，2015年；Kotlowski 等，2011年），这里我们考虑 L_{pa}、 L_{u_1} 和 L_{u_2} 关于宏平均AUC 的一致性，其中 $L_{0/1}(\mathrm{x}^+, \mathrm{x}^-, f_k) = [\![f_k(\mathrm{x}^+) < f_k(\mathrm{x}^-)]\!] + \frac{1}{2}[\![f_k(\mathrm{x}^+) = f_k(\mathrm{x}^-)]\!]$。多标签分类的宏平均AUC 最大化任务可分解为 K 个二元分类的 AUC 最大化任务。因此，我们可以基于此前对二元分类（或等价地，二分排序）中 AUC 一致性结果的深入研究，考察这些替代损失的一致性（Gao 和 Zhou，2015年；Kotlowski 等，2011年）。具体而言，我们可以得到关于这些替代损失的以下结果：

- L_{pa}：基于二分类中关于 AUC 最大化的先前结果（Gao 和 Zhou，2015年）（即第 4 页的推论 1），我们可以得出 L_{pa} 对于宏平均AUC而言，使用（基础）逻辑损失和指数损失是一致的。此外，基于先前的结果（Gao 和 Zhou，2015年）（即第 3 页的引理 3），我们可以得出 L_{pa} 对于宏平均AUC而言，使用（基础）合页损失和绝对损失是不一致的。
- L_{u_1}：基于先前的结果（Gao 和 Zhou，2015年）（即第 6 页的定理 7），我们可以得出L_{u_1} 使用（基础）指数损失是一致的。
- L_{u_2}：作为一种重加权单变量损失， L_{u_2} 涉及依赖于数据集的重加权因子（即，$\frac{1}{|S_k^+|}$，$\frac{1}{|S_k^-|}$ 分别对应正例和负例）。因此，在无限样本（即总体）情形下，可以认为它依赖于分布，其中正例和负例的重加权因子分别与 $P^{1}{}_{(y_k=1)}$和 $1-P^{1}{}_{(y_k=1)}$ 成正比。在这种情况下，根据二部排序的先前研究结果（Kotlowski 等人，2011年）（即第 4 页上的定理 4.1），可以得到L_{u_2} 在逻辑损失和指数损失作为基础损失时，与宏平均AUC保持一致。

至于采用其他基础损失的替代损失 L_{u_1} 和 L_{u_2}，我们将其留作未来的研究工作。

5. 相关工作

一致性。Gao 和 Zhou（2013）一般性地研究了替代损失相对于汉明度量和（部分）排序度量的一致性。此外，Dembczynski 等人（2012）针对部分排序度量，为替代单变量损失给出了明确的遗憾界。值得注意的是，对于二元分类中的 F 度量，Ye 等人（2012）阐明并联系了经验效用最大化（EUM）框架与决策论方法（DTA），并将其应用于优化多标签分类（MLC）中的宏平均 F 度量。[12] 对于这两种方法框架，Dembczy´nski 等人（2017）重新审视了复杂度量下二元分类的一致性分析，并采用了更具描述性的名称：总体效用（PU）和期望测试效用（ETU）。此外，对于 MLC 中的 F 度量，Waegeman 等人（2014）；Zhang 等人（2020a）从 DTA 的角度，通过以不同方式估计条件分布 $P(\mathbf{y}|\mathbf{x})$，研究了一致性。此外，Koyejo 等人（2015）在 EUM 框架下研究了相对于各种度量的一致性 MLC 方法，而 Menon 等人（2019）则考察了各种归约方法相对于 precision@k 和 recall@k 度量的多标签一致性。

泛化。 Wu &Zhu (2020) 研究了旨在优化汉明损失和子集准确率这两项指标的替代损失学习算法的泛化，并发现标签数量对泛化界有重要影响，这解释了以下经验现象：当标签数量不大时，使用其替代损失优化汉明损失，在子集准确率方面可能表现良好。此外，Wu 等人(2021年) 重新审视了许多基于代理损失的算法在排序损失指标下的一致性和泛化，并指出数据集（或分布）的实例级类别不平衡对泛化界起着关键作用，这比一致性更能解释经验现象。

我们提到，Wu 和 Zhou （2017）也提出了一种成对损失（类似于公式（5））， 但该方法省略了重加权因子 $\frac{1}{|S_k^+||S_k^-|}$，且缺乏正式的泛化分析。此外，关于近期一种适用于图依赖数据的 McDiarmid 型集中不等式（Zhang 等人，2019）与我们的方法之间的比较，详见附录D 中的详细讨论。

[12]请注意，我们针对学习算法的宏平均AUC泛化分析是在 EUM 框架下进行的。

6. 实验

作为一项理论研究，实验的主要目的是验证我们的理论发现，而非展示所提方法的优越性能。因此，我们在第3.2 节中，使用涵盖不同领域、标签数量和数据规模的10个广泛使用的基准数据集，以宏平均AUC评估上述三种学习算法。数据集的详细统计信息汇总于表 2，其中包括四种标签级类别不平衡相关因子。[13] 此外，图1展示了三个代表性数据集的标签级类别不平衡程度。（所有数据集的结果见附录C.1 中的图2 。）为简化分析并确保公平比较，对于所有算法，我们均采用以逻辑损失为基础的线性模型。此外，我们使用相同的高效随机优化算法（即 SVRG-BB（Tan 等人，2016））求解这些凸优化问题。此外，我们在较大的范围内搜索所有算法在所有数据集上的超参数 λ，即 $\{10^{-6}, 10^{-5}, \ldots, 10_2\}$，并采用3折交叉验证。[14]

实验结果总结在表 3中。总体而言，我们可以观察到，算法$\mathcal{A}^{pa}$ 和 $\mathcal{A}^{u_2}$ 的表现优于算法 $\mathcal{A}^{u_1}$，这证实了我们的理论结果，即 $\mathcal{A}^{pa}$ 和 $\mathcal{A}^{u_2}$ 相较于 $\mathcal{A}^{u_1}$，在标签级类别不平衡方面具有更好的学习保证。此外，$\mathcal{A}^{pa}$ 的表现与 $\mathcal{A}^{u_2}$相当，这也验证了我们的理论结果：它们在标签级类别不平衡方面具有相同的学习保证。

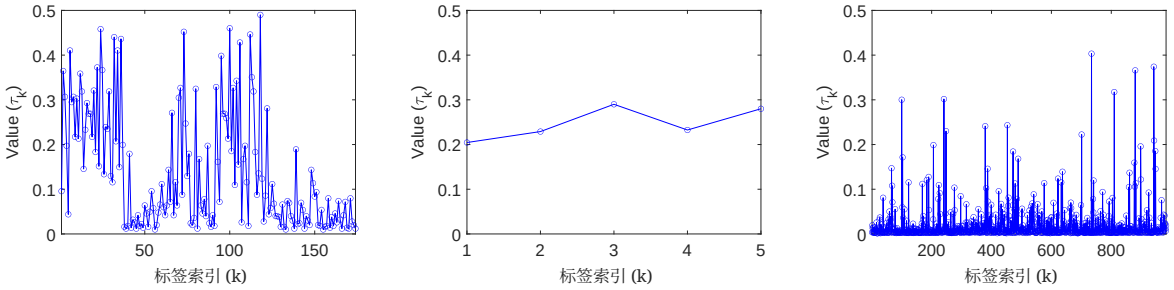
此外，通过表 2 和表 3，我们可以仔细研究标签级类别不平衡对性能的影响。回想一下，A^{pa} 和 A^{u_2} 的学习保证都取决于因子 Imb_1,，而 A^{u_1} 的学习保证取决于因子 Imb_4。对于数据集 CAL500、Enron、rcv-s1、BibTeX、corel5k 和 delicious，因子 Imb_1 和 Imb_4 的阶数差距较大（等价地说，Imb_3 较大），因此 A^{u_2}（或 A^{pa}）的表现显著优于 A^{u_1}。相反，对于数据集 emotions、图像和场景，因子 Imb_1 和 Imb_4 的差距较小（等价地说，Imb_3 较小），因此 A^{u_2} （或 A^{pa}）的表现略优于 A^{u_1}，或与其几乎相当。这也证实了我们关于这些算法在标签级类别不平衡方面的理论发现。

此外，与多标签分类中排序损失度量的先前理论结果类似（Wu 等人，2021年），我们的泛化上界绝对值可能无法合理地反映真实的泛化误差（即大于1）。不过，它们仍然可以提供有价值的

[13]这些数据集可从http://mulan.sourceforge.net/datasets-mlc.html和http://palm.seu.edu.cn/zhangml/下载。[14]我们的代码可在以下网址获取：https://github.com/GuoqiangWoodrowWu/M

表 2. 基准数据集的基本统计信息。分别用 $\text{Imb}_1 = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}$、$\text{Imb}_2 = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k}}$、$\text{Imb}_3 = \frac{1}{\tau_S^*}$ 和 $\text{Imb}_4 = \frac{1}{\tau_S^*}$（$\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}$）表示标签级类别不平衡相关因子。

数据集	#实例	#特征	#标签	领域	Imb_1	Imb_2	Imb_3	Imb_4
CAL500	502	68	174	音乐	4.2	4.8	100.4	421.1
情绪	593	72	6	音乐	1.8	1.8	4.0	7.3
图像	2000	294	5	图像	2.0	2.0	4.9	9.9
场景	2407	294	6	图像	2.4	2.4	6.6	15.7
酵母	2417	103	14	生物学	2.6	3.2	71.1	188.4
Enron	1702	1001	53	text	9.1	11.7	1702	15566
rcv1-s1	6000	944	101	text	11.8	15.4	3000	35267
BibTeX	7395	1836	159	text	9.2	9.4	7395	1332
corel5k	5000	499	374	图像	23.4	29.1	5000	117000
delicious	16105	500	983	文本（网页）	12.2	13.3	766.9	9344



(a) CAL500 (b) 图像 (c) delicious

图1. 三个代表性数据集的标签级类别不平衡示意图。

表3. 三种算法在基准数据集上的宏平均AUC（均值 ± 标准差，符号 . 表示 0.）。对于每个数据集，排名前两位的算法以粗体突出显示，排名第一的算法标记为 †。此外，“-”表示在相应数据集上使用16核CPU服务器运行A^{pa}需要超过一周。

数据集	$\mathcal{A}^{pa}$	$\mathcal{A}^{u_1}$	$\mathcal{A}^{u_2}$
CAL500	**.5735 ± .0186**†	.5571 ± .0102	**.5717 ± .0177**
情绪	**.8372 ± .0172**†	.8346 ± .0223	**.8348 ± .0189**
图像	**.8383 ± .0073**†	**.8359 ± .0121**	.8314 ± .0094
场景	**.9319 ± .0013**†	.9271 ± .0067	**.9285 ± .0035**
酵母	**.6872 ± .0100**	.6862 ± .0064	**.6892 ± .0088**†
Enron	**.7211 ± .0320**	.6908 ± .0105	**.7356 ± .0121**†
rcv1-s1	-	.8585 ± .0204	**.9097 ± .0068**†
BibTeX	-	.8693 ± .0156	**.9299 ± .0034**†
corel5k	-	.5703 ± .0092	**.6645 ± .0253**†
delicious	-	.7633 ± .0020	**.8044 ± .0040**†

在同一分析框架下深入理解这些学习算法。（详见表 4 附录 C.2 中的详细内容）。更先进的技术（例如局部拉德马赫型复杂度）可以进一步改进这些结果，这将作为未来的研究方向。

7. 结论

为了理解多标签分类中宏平均AUC的泛化，本文通过分析采用各种替代损失（包括广泛使用的单变量损失）的算法的泛化界，迈出了初步一步。我们的结果表明，数据集的标签级类别不平衡在这些泛化界中发挥着关键作用。采用所提出的成对损失和重加权单变量损失的算法，相较于原始的基于单变量损失的算法，具有更好的学习保证，这可能意味着它们具有更优越的性能。实验结果也证实了我们的理论发现。

社会影响：作为一项理论研究，这项工作有助于理解并有可能开发出更好的多标签学习算法，同时不会对社会造成明显的负面影响。

致谢

本研究得到中国国家自然科学基金（项目编号：62206159、62076145）、山东省自然科学基金（项目编号：ZR2022QF117、ZR2021ZD15）、北京优秀青年科学家计划（编号：BJJWZYJH012019100020098）以及基础研

究经费（山东大学）、「双一流」建设重大创新与规划跨学科平台（中国人民大学）、中央高校基本科研业务费以及中国人民大学科研基金（22XNKJ13）的资助。C. Li 还获得了北京新星计划的资助。

参考文献

Amini, M.-R. 和 Usunier, N. 基于部分标记且相互依赖的数据进行学习。Springer，2015年。

Anthony, M. 和 Bartlett, P. L. 神经网络学习：理论基础。剑桥大学出版社，1999年。

Arora, S. 和 Barak, B. 计算复杂性：现代方法。剑桥大学出版社，2009年。

Bartlett, P. L.、Bousquet, O.、Mendelson, S. 等。局部 Rademacher 复杂度。《统计学年鉴》，33 (4):1497–1537，2005年。

Boucheron, S., Lugosi, G. 和 Massart, P. 《集中不等式：独立性的非渐近理论》。牛津大学出版社，2013年。

Boutell, M. R., Luo, J., Shen, X. 和 Brown, C. M. 学习多标签场景分类。《Pattern Recognition》，37(9):1757–1771，2004年。

Carneiro, G., Chan, A. B., Moreno, P. J. 和 Vasconcelos, N. 用于图像标注与检索的语义类别监督学习。《IEEE Transactions on Pattern Analysis and Machine Intelligence》，29(3):394–410，2007年。

Dembczynski, K., Kotłowski, W. 和 Hüllermeier, E. 通过最小化单变量损失实现一致的多标签排序。载于第29届国际 Machine Learning 会议论文集，第1347–1354页，2012年。

Dembczy´nski, K., Kotłowski, W., Koyejo, O. 和 Natarajan, N. 二元分类一致性分析再探。载于 International Conference on Machine Learning，第 961–969 页。PMLR，2017年。

Elisseeff, A. 和 Weston, J. 一种用于多标签分类的核方法。《神经信息处理系统进展》，14，2001年。

Gao, W. 和 Zhou, Z.-H. 关于多标签学习的一致性。《人工智能》，199(1):22–44，2013年。

Gao, W. 和 Zhou, Z.-H. 关于 AUC 成对优化的一致性。载于第二十四届国际人工智能联合会议，2015年。

Hariharan, B.、Zelnik-Manor, L.、Varma, M. 和 Vish- wanathan, S. 基于先验的大规模最大间隔多标签分类。载于第 27 届国际机器学习会议论文集（ICML-10），第 423–430 页，2010 年。Janson, S. 部分相依随机变量之和的大偏差。随机结构与算法，24 (3):234–248，2004 年。Kotlowski, W.、Dembczynski, K. J. 和 Huellermeier, E. 通过最小化单变量损失实现二部排序。载于第 28 届国际机器学习会议论文集（ICML-11），第 1113–1120 页。Citeseer，2011 年。

Koyejo, O.、Natarajan, N.、Ravikumar, P. 和 Dhillon, I. S. 一致的多标签分类。载于《神经信息处理系统进展》，第 29 卷，第 3321–3329 页，2015年。

McCallum, A. K. 使用期望最大化训练的混合模型进行多标签文本分类。载于 AAAI 99文本学习研讨会。Citeseer，1999年。

McDiarmid, C. 等. 有界差分方法.《组合学综述》，141(1):148–188，1989.

Menon, A. K.、Rawat, A. S.、Reddi, S. 和 Kumar, S. 多标签归约：我的损失在优化什么？载于《神经信息处理系统进展》，第 10599–10610 页，2019.

Mohri, M.、Rostamizadeh, A. 和 Talwalkar, A. 机器学习基础. 麻省理工学院出版社，2018.

Ridnik, T.、Ben-Baruch, E.、Zamir, N.、Noy, A.、Friedman, I.、Protter, M. 和 Zelnik-Manor, L. 多标签分类的非对称损失. 载于 IEEE/CVF 国际计算机视觉大会论文集，第 82–91 页，2021年.

Schapire, R. E. 和 Singer, Y. Boostexter：一种基于提升的文本分类系统。Machine Learning，39(2):135–168，2000。

Tan, C., Ma, S., Dai, Y.-H. 和 Qian, Y. 用于随机梯度下降的 Barzilai-Borwein步长。载于《神经信息处理系统进展》，第 685–693 页，2016 年。

Tan, Z.-H., Tan, P., Jiang, Y. 和 Zhou, Z.-H. 多标签最优间隔分布机。Machine Learning，109(3):623–642，2020。

Tarekegn, A. N.、Giacobini, M. 和 Michalak, K. 关于不平衡多标签分类方法的综述。Pattern Recognition，118:107965，2021年。

Usunier、N.、Amini、M. R. 和 Gallinari、P.《使用相互依赖数据训练的分类器的泛化误差界》。神经信息处理系统进展，18，2005 年。

Waegeman, W., Dembczy´nski, K., Jachnik, A., Cheng, W., 和 Hüllermeier, E. 关于 F 度量最大化器的贝叶斯最优性。《Machine Learning 研究杂志》，15:3333–3388，2014年。Wu, G. 和 Zhu, J. 多标签分类：汉明损失和子集准确率真的相互冲突吗？《神经信息处理系统进展》，33:3130–3140，2020年。Wu, G., Zheng, R., Tian, Y., 和 Liu, D. 用于多标签分类的联合排序 SVM 与二元关联方法：结合稳健低秩学习。《神经网络》，122:24–39，2020a。Wu, G., Li, C., Xu, K., 和 Zhu, J. 重新审视并重新加权用于多标签排序的单变量损失：一致性与泛化。《神经信息处理系统进展》，34:14332–14344，2021年。Wu, T., Huang, Q., Liu, Z., Wang, Y., 和 Lin, D. 用于长尾数据集多标签分类的分布平衡损失。载于《计算机视觉——ECCV 2020：第 16 届欧洲会议》，2020年8月23日至28日，英国格拉斯哥，会议论文集第四部分，第 162–178 页，2020b。Wu, X.-Z. 和 Zhou, Z.-H. 多标签性能度量的统一视角。载于国际机器学习会议，第 3780–3788 页。PMLR，2017年。Ye, N., Chai, K. M., Lee, W. S., 和 Chieu, H. L. 优化 F 度量：两种方法的故事。载于国际机器学习会议，第 289–296 页。Omnipress，2012年。Zhang, M., Ramaswamy, H. G., 和 Agarwal, S. 用于多标签 F 度量的凸校准替代损失。载于国际机器学习会议，第 11246–11255 页。PMLR，2020a。Zhang, M.-L. 和 Zhou, Z.-H. 多标签学习算法综述。《IEEE 知识与数据工程汇刊》，26(8):1819–1837，2014年。Zhang, M.-L., Li, Y.-K., Yang, H., 和 Liu, X.-Y. 迈向类别不平衡感知的多标签学习。《IEEE 控制论汇刊》，2020b。Zhang, R.-R. 和 Amini, M.-R. 图依赖条件下学习的泛化界：综述。arXiv 预印本 arXiv:2203.13534，2022年。

Zhang, R. R.、Liu, X.、Wang, Y. 和 Wang, L. 图依赖变量的 McDiarmid 型不等式及稳定性界。《神经信息处理系统进展》，第 32 卷，2019 年。

附录目录

A 通用技术 **13**

A.1 一种新的 McDiarmid 型集中不等式. 13

A.1.1 背景 13

A.1.2 新 McDiarmid 型集中不等式的证明 . 14

A.2 利用图相关样本进行多任务学习 . 16

A.2.1 问题设定 16

A.2.2 损失空间的分数拉德马赫复杂度 . 17

A.2.3 图依赖样本下多任务学习的一般泛化界证明 17

B 多标签分类中的宏平均AUC最大化 **19**

B.1 定理 1 的证明 .19

B.1.1 问题转换 .19B.1.2 定理 1 的证明 .19

B.2 引理 1 和 2 的证明 .20

B.2.1 引理 1 的证明 .20B.2.2 引理 2 的证明 .20

B.3 定理 2、推论 4 和 2 的证明 .22

B.3.1 假设空间的分数拉德马赫复杂度.22B.3.2 收缩不等式.23B.3.3 定理 2 的证明.25B.3.4 推论 4 的证明.25B.3.5 推论 2 的证明.25B.4 定理 3 和 4、推论 1、3 和 5 的证明.25B.4.1 假设空间的分数拉德马赫复杂度.25B.4.2 收缩不等式.27B.4.3 定理 3 的证明.29B.4.4 定理 4 的证明.29B.4.5 推论 1 的证明.30B.4.6 推论 3 的证明.30B.4.7 推论 5 的证明.30

C 其他实验结果 **30**

C.1 基准数据集的标签级类别不平衡示意图.30C.2 关于界的绝对值的实验结果.30

D 其他相关工作 **31**

A. 通用技术

在本节中，我们介绍通用技术，主要包括一种新型的 McDiarmid 型集中不等式，以及一个适用于图依赖样本的多任务学习通用泛化界。

A.1. 一个新的 McDiarmid 型集中不等式

A.1.1. 背景

首先，我们引入有界差分性质，以及一个用于证明后续定理的引理 m.

定义 4 （有界差分性质 （McDiarmid 等人，1989年））。 设 x_1， x_2， ....., $x_m \in \mathcal{X}$， 以及函数 $f : \mathcal{X}^m \to \mathbb{R}$。则，$f$如果存在 c_1， ....., $c_m > 0$ 使得

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \le c_i,$$

对于所有 $i \in [m]$ 以及任意点 x_1， ....., x_m、 $x_i' \in \mathbf{X}$。

引理 3 （（McDiarmid 等人，1989年））。 设 $\mathbf{X} = (X_1, \ldots, X_m) \in \mathcal{X}^m$ 是由 m 个独立随机变量组成的向量，且函数 $f : \mathcal{X}^m \to \mathbb{R}$ 满足有界差分性质，其 c_i （$i \in [m]$） ，则对于任意 $s > 0$,

$$\mathbb{E}[\exp(s(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})]))] \le \exp\left(\frac{s^2 \sum_{i\in[m]} c_i^2}{8}\right).$$

本文将介绍一些图论中的必要概念，相关内容可参阅（Janson，2004年；Amini & Usunier，2015年）以及近期综述（Zhang& Amini，2022年）。

给定一个图 $G = (V, E)$，我们引入以下概念。

定义 5 (分数独立顶点覆盖和分数色数 (Zhang 和 Amini，2022))。

(1) 一族对 (F_j, ω_j)，其中 $F_j \subseteq V(G)$ 且 $\omega_j \in (0, 1]$ 是 G 的分数顶点覆盖，如果对于每个 $v \in V(G)$，都有 $\sum_{j:v\in F_j} \omega_j = 1$。

(2) G 的独立集是 G 中的一个顶点集合，其中任意两个顶点都不相邻。 G的独立集集合记作 $\mathrm{I}(G)$。

(3) 分数独立顶点覆盖 $\{(I_j, \omega_j)\}_j$ 是一种分数顶点覆盖，满足对于每个 j， $I_j \in \mathrm{I}(G)$。

(4) 图 G 的分数着色是一个映射 g，它从 $\mathrm{I}(G)$ 映射到 $(0, 1]$，满足对每个顶点 $v \in V(G)$ 都有 $\sum_{I\in\mathcal{I}(G):v\in I} g(I) \ge 1$。分数色数 $\chi_f(G)$ 是对 G 的所有分数着色中，值 $\sum_{I\in\mathcal{I}(G)} g(I)$ 的最小值。

注意，图 G 的分数色数 $\chi_f(G)$ 是所有分数独立顶点覆盖 $\{(I_j, \omega_j)\}_j$ 中 $\sum_j \omega_j$ 的最小值。

接下来，我们介绍如下的依赖图概念。

定义 6 (依赖图 (Janson，2004))。 无向图 $G = (V, E)$ 称为随机向量 $= \mathbf{X}(X_1$， ...， $X_m)$ 的依赖图，如果

(1) $V(G) = [m]$。

(2) 对于所有不相交的 I、 $J \in [m]$，如果 I、 J 在 G 中不相邻，则随机变量 $\{X_i\}_{i\in I}$ 和 $\{X_j\}_{j\in J}$ 相互独立。

于是，我们称随机向量 $\mathbf{X}$ 是具有依赖图 G 的 G-依赖随机向量。

依赖图的一个重要性质，结合分数独立覆盖的概念，即为 Janson 分解性质（Janson，2004）。具体而言，设相依随机变量 $(X_i)^{i\in[m]}$ 是 G-依赖的，且其依赖图为 G，并且 $\{(I_j, \omega_j)\}^{j\in[J]}$ 是 G的一个分数独立顶点覆盖。那么，我们可以将相依变量之和分解为若干个独立变量之和的加权和，即

$$\sum_{i=1}^{m} X_i = \sum_{i=1}^{m}\sum_{j=1}^{J} \omega_j [\![i \in I_j]\!] X_i = \sum_{j=1}^{J} \omega_j \sum_{i\in I_j} X_i. \tag{13}$$

A.1.2. 新型 McDiarmid 型集中不等式的证明

我们在此提出如下一个新的、更一般的 McDiarmid 型不等式，主要参考了相关研究（Janson，2004；Usunier et al.，2005；Amini & Usunier，2015年），并请参阅近期的一篇相关综述（Zhang& Amini，2022）。

定理 5（一个新的、更一般的 McDiarmid 型不等式）。设 $\mathbf{X}_1 = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{1m_1}) \in \mathbf{X}^{m_1}, \ldots, \mathbf{X}_K = (\mathbf{x}_{K1}, \ldots, \mathbf{x}_{Km_K}) \in \mathbf{X}^{m_K}$ 为随机变量向量，为方便起见，记 $\mathbf{X}$ 为 $(\mathbf{X}_1, \ldots, \mathbf{X}_K) = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{Km_K})$。设 $f_1 : \mathbf{X}^{m_1} \to \mathbb{R}, \ldots, f_K : \mathbf{X}^{m_K} \to \mathbb{R}$ 和 $f : \mathbf{X}^{m} \to \mathbb{R}$ 为满足 $\sum_{k=1}^{K} m_k = m$ 的函数。假设每个 $\mathbf{X}_k$ $(k \in [K])$ 都是 G_k-依赖的，且具有依赖图 G_k。[15] 此外，假设函数 f 满足以下约束：

(1) $f(\mathbf{X}) = \sum_{k\in[K]} f_k(\mathbf{X}_k)$；(2) $f_k(\mathbf{X}_k)$ 满足关于图 G_k 的具有有界差分性质的可分解性约束，即，对于所有 $\mathbf{x}_k \in \mathbf{X}^{m_k}$ 以及 G_k 的最小分数独立顶点覆盖 $\{(I_{kj}, \omega_{kj})\}^{j\in[J_k]}$，存在函数 $\{f_{kj} : \mathbf{X}^{|I_{kj}|} \to \mathbb{R}\}^{j\in[J_k]}$，使得 f_{kj} 关于 $c_{ki} (i \in I_{kj})$ 满足有界差分性质，并且

$$f_k(\mathbf{x}_k) = \sum_{j\in[J_k]} \omega_{kj} f_{kj}(\mathbf{x}_{I_{kj}}),$$

其中，$\mathbf{x}^{I_{kj}}$ 表示 $(\mathbf{x}_{ki})^{i\in I_{kj}}$。

那么，对于任意 $t > 0$,

$$\mathbb{P}(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})] \geq t) \leq \exp\left(-\frac{2t^2}{K\sum_{k\in[K]}\left(\chi_f(G_k)\sum_{i\in[m_k]} c_{ki}^2\right)}\right),$$

其中，$\chi_f(G_k)$ 是 G_k 的分数色数。

备注。 先前工作中的 McDiarmid 型不等式（*Usunieret al.*，2005年；*Amini & Usunier*，2015年）可以通过设置 $K = 1$而视为上述不等式的特例。因此，我们提出的新 McDiarmid 型不等式更为一般。

证明。遵循根据 Cramér-Chernoff 方法（Boucheron 等人，2013），对于任意 $s > 0$ 和 $t > 0$ ，

$$\mathbb{P}(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})] \geq t) \leq e^{-st}\,\mathbb{E}[\exp(s(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})]))]. \tag{14}$$

对于依赖图， k号子图 G_k 有 m_k 个顶点。此外，设 I_k 为 G_k 的顶点集，$\{(I_{kj}, \omega_{kj})\}^{j\in[J_k]}$ 为 G_k 的一个最小分数独立顶点覆盖，其满足 $\sum_{j\in[J_k]} \omega_{kj} = \chi_f(G_k)$。利用分解性质 $f(\mathbf{x}) = \sum_{k\in[K]} f_k(\mathbf{x}_k) = \sum_{k\in[K]}\sum_{j\in[J_k]} \omega_{kj} f_{kj}(I_{kj})$, 其中为简化记号，将$\mathbf{x}^{I_{kj}}$记为 I_{kj}，我们可以得到上述不等式右端期望项的表达式：

$$\mathbb{E}[\exp(s(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})]))] = \mathbb{E}\left[\exp\left(\sum_{k\in[K]}\sum_{j\in[J_k]} s\omega_{kj}\left(f_{kj}(I_{kj}) - \mathbb{E}\, f_{kj}(I_{kj})\right)\right)\right].$$

设 $\{p_1, p_2, \ldots, p_K\}$ 是任意一组总和为 1 的 K 严格正实数。同样地，对于每个 $k \in [K]$，设 $\{q_{k1}, q_{k2}, \ldots, q_{kJ_k}\}$ 是任意一组总和为 1 的 J_k 严格正实数。然后，基于指数函数的凸性，我们可以得到如下结果：

$$\begin{aligned}
\mathbb{E}[\exp(s(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})]))] &= \mathbb{E}\left[\exp\left(\sum_{k\in[K]}\sum_{j\in[J_k]} s\omega_{kj}\left(f_{kj}(I_{kj}) - \mathbb{E}\, f_{kj}(I_{kj})\right)\right)\right] \\
&= \mathbb{E}\left[\exp\left(\sum_{k\in[K]} p_k \sum_{j\in[J_k]} \frac{s\omega_{kj}}{p_k}\left(f_{kj}(I_{kj}) - \mathbb{E}\, f_{kj}(I_{kj})\right)\right)\right] \qquad \text{(definition of } p_k\text{)}
\end{aligned}$$

[15]请注意，这里我们只对每个 $\mathbf{X}_k$ 内部的依赖关系作出假设，而不对不同的 $\mathbf{X}_k$ 之间作出任何假设；$\mathbf{X}_k$ 可以相互独立，也可以相互依赖，不论其是否独立。

$$
\begin{aligned}
&\leq \mathbb{E}\left[\sum_{k\in[K]} p_k \exp\left(\sum_{j\in[J_k]} \frac{s\omega_{kj}}{p_k}\left(f_{kj}(I_{kj}) - \mathbb{E}\, f_{kj}(I_{kj})\right)\right)\right] && \text{(Jensen's inequality)}\\
&= \mathbb{E}\left[\sum_{k\in[K]} p_k \exp\left(\sum_{j\in[J_k]} q_{kj}\frac{s\omega_{kj}}{p_k q_{kj}}\left(f_{kj}(I_{kj}) - \mathbb{E}\, f_{kj}(I_{kj})\right)\right)\right] && \text{(definition of } p_{kj}\text{)}\\
&\leq \mathbb{E}\left[\sum_{k\in[K]} p_k \sum_{j\in[J_k]} q_{kj} \exp\left(\frac{s\omega_{kj}}{p_k q_{kj}}\left(f_{kj}(I_{kj}) - \mathbb{E}\, f_{kj}(I_{kj})\right)\right)\right] && \text{(Jensen's inequality)}\\
&= \sum_{k\in[K]} p_k \underbrace{\sum_{j\in[J_k]} q_{kj}\, \mathbb{E}\left[\exp\left(\frac{s\omega_{kj}}{p_k q_{kj}}\left(f_{kj}(I_{kj}) - \mathbb{E}\, f_{kj}(I_{kj})\right)\right)\right]}_{\stackrel{def}{=} \clubsuit_k} && \text{（期望的线性性）。}
\end{aligned}
$$

我们可以观察到，对于求和项$\clubsuit_k$，与每一项 $j \in [J_k]$ 相关联的随机变量彼此独立。因此，应用引理3，可得

$$
\clubsuit_k = \sum_{j\in[J_k]} q_{kj}\, \mathbb{E}\left[\exp\left(\frac{s\omega_{kj}}{p_k q_{kj}}\left(f_{kj}(I_{kj}) - \mathbb{E}\, f_{kj}(I_{kj})\right)\right)\right] \leq \sum_{j\in[J_k]} q_{kj} \exp\left(\frac{s^2\omega_{kj}^2}{8p_k^2 q_{kj}^2}\sum_{i\in I_{kj}} c_{ki}^2\right).
$$

通过重新整理上述不等式右侧指数中的各项，并令

$$
q_{kj} = \frac{\omega_{kj}\sqrt{\sum_{i\in I_{kj}} c_{ki}^2}}{\sum_{j\in[J_k]}\left(\omega_{kj}\sqrt{\sum_{i\in I_{kj}} c_{ki}^2}\right)},
$$

我们有：

$$
\begin{aligned}
\sum_{j\in[J_k]} q_{kj} \exp\left(\frac{s^2\omega_{kj}^2}{8p_k^2 q_{kj}^2}\sum_{i\in I_{kj}} c_{ki}^2\right) &= \sum_{j\in[J_k]} q_{kj} \exp\left(\frac{s^2}{8p_k^2}\left(\sum_{j\in[J_k]}\left(\omega_{kj}\sqrt{\sum_{i\in I_{kj}} c_{ki}^2}\right)\right)^2\right)\\
&= \exp\left(\frac{s^2}{8p_k^2}\left(\sum_{j\in[J_k]}\left(\omega_{kj}\sqrt{\sum_{i\in I_{kj}} c_{ki}^2}\right)\right)^2\right) \qquad \left(\sum_{j\in[m_k]} q_{kj} = 1\right).
\end{aligned}
$$

至此，我们得到：

$$
\mathbb{E}[\exp(s(f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})]))] \leq \sum_{k\in[K]} p_k \clubsuit_k \leq \sum_{k\in[K]} p_k \exp\left(\frac{s^2}{8p_k^2}\left(\sum_{j\in[J_k]}\left(\omega_{kj}\sqrt{\sum_{i\in I_{kj}} c_{ki}^2}\right)\right)^2\right).
$$

接下来，按照上述关于 q_{kj} 的证明思路，我们令 p_k 如下：

$$
p_k = \frac{\sum_{j\in[J_k]}\left(\omega_{kj}\sqrt{\sum_{i\in I_{kj}} c_{ki}^2}\right)}{\sum_{k\in[K]}\sum_{j\in[J_k]}\left(\omega_{kj}\sqrt{\sum_{i\in I_{kj}} c_{ki}^2}\right)}.
$$

然后，就出现了：

$$
\begin{aligned}
\sum_{k\in[K]} p_k \exp\left(\frac{s^2}{8p_k^2}\left(\sum_{j\in[J_k]}\left(\omega_{kj}\sqrt{\sum_{i\in I_{kj}} c_{ki}^2}\right)\right)^2\right) &= \sum_{k\in[K]} p_k \exp\left(\frac{s^2}{8}\left(\sum_{k\in[K]}\sum_{j\in[J_k]}\left(\omega_{kj}\sqrt{\sum_{i\in I_{kj}} c_{ki}^2}\right)\right)^2\right)\\
&= \exp\left(\frac{s^2}{8}\left(\sum_{k\in[K]}\sum_{j\in[J_k]}\left(\omega_{kj}\sqrt{\sum_{i\in I_{kj}} c_{ki}^2}\right)\right)^2\right) \quad \left(\sum_{k\in[K]} p_k = 1\right)
\end{aligned}
$$

$$\overset{①}{\leq} \exp\left(\frac{s^2K}{8}\sum_{k\in[K]}\left(\sum_{j\in[J_k]}\omega_{kj}\sqrt{\sum_{i\in I_{kj}}c_{ki}^2}\right)^2\right)$$
$$= \exp\left(\frac{s^2K}{8}\sum_{k\in[K]}\left(\sum_{j\in[J_k]}\left(\sqrt{\omega_{kj}}\sqrt{\omega_{kj}\sum_{i\in I_{kj}}c_{ki}^2}\right)\right)^2\right)$$
$$\overset{②}{\leq} \exp\left(\frac{s^2K}{8}\sum_{k\in[K]}\left(\sum_{j\in[J_k]}\omega_{kj}\right)\left(\sum_{j\in[J_k]}\omega_{kj}\sum_{i\in I_{kj}}c_{ki}^2\right)\right)$$
$$\overset{③}{=} \exp\left(\frac{s^2K}{8}\sum_{k\in[K]}\chi_f(G_k)\left(\sum_{i\in[m_k]}c_{ki}^2\right)\right).$$

对于①，其依据是不等式 $(\sum_{i=1}^n a_i)^2 \leq n\sum_{i=1}^n a_i^2$。对于 2，其依据是柯西-施瓦茨不等式。对于 3，其依据是分数色数的定义，即 $\sum_{j\in[J_k]}\omega_{kj} = \chi_f(G_k)$，以及依赖图中分数独立顶点覆盖的分解性质（Janson，2004），即对于 $\{(I_{kj},\ \omega_{kj})\}^{j\in[J_k]}$的一个分数独立顶点覆盖 G_k，相互依赖变量的和可以分解为若干独立变量之和的加权和，如下所示：

$$\sum_{i=1}^{m_k}\mathbf{x}_{ki} = \sum_{i=1}^{m_k}\sum_{j=1}^{J_k}\omega_{kj}[\![i\in I_{kj}]\!]\mathbf{x}_{ki} = \sum_{j=1}^{J_k}\omega_{kj}\sum_{i\in I_{kj}}\mathbf{x}_{ki}.$$

由于 $\mathbf{X}_k = [\mathbf{x}_{k1},\ \ldots,\ \mathbf{x}_{km_k}]$ 是一个随机向量，我们可以代入具体值，基于不等式 2 得到等式 3。具体而言，如果对每个 $i\in[m_k]$ 取 $\mathbf{x}_{ki} = c_{ki}^2$，则可以得到

$$\sum_{i=1}^{m_k}c_{ki}^2 = \sum_{j=1}^{J_k}\omega_{kj}\sum_{i\in I_{kj}}c_{ki}^2.$$

因此，我们得到

$$\mathbb{E}[\exp(s(f(\mathbf{X})-\mathbb{E}[f(\mathbf{X})]))] \leq \exp\left(\frac{s^2K}{8}\sum_{k\in[K]}\chi_f(G_k)\left(\sum_{i\in[m_k]}c_{ki}^2\right)\right).$$

结合不等式（14），我们可以得到

$$\mathbb{P}(f(\mathbf{X})-\mathbb{E}[f(\mathbf{X})]\geq t) \leq \exp\left(-st+\frac{s^2K}{8}\sum_{k\in[K]}\chi_f(G_k)\left(\sum_{i\in[m_k]}c_{ki}^2\right)\right).$$

对上述不等式右侧取最小值，即可得到最终结果 s. □

A.2. 使用图依赖样本进行多任务学习

A.2.1. 问题设定

这里我们考虑多任务学习，其中每个任务可能包含相互依赖的训练样本，依赖关系由依赖图刻画。形式化地，给定训练数据集 $S = \{(\tilde{\mathbf{x}}\ \tilde{y})\}_{i=1}^m$，它由 K 个数据块（或任务）组成，即 $S = (S_1, \ldots, S_K)$，其中每个 $S_k = \{(\tilde{\mathbf{x}}_{ki}, \tilde{y}_{ki})\}_{i=1}^{m_k}$均从分布 $D_k(k\in[K])$ 中抽取，该分布定义在 $\tilde{\mathbf{X}} \times \tilde{\mathbf{Y}}$ 上，并具有依赖图 G_k和 $\sum_{k\in[K]} m_k = m$。目标是学习映射 $\tilde{h} = (\tilde{h}_1, \ldots, \tilde{h}_K)$，其中每个 $k\in[K]$都有 $h_k : \mathbf{X} \to \mathbf{Y}$。

令 $\tilde{\mathbf{F}} = \{\tilde{f} = (\tilde{f}_1, \ldots, \tilde{f}_K) \mid \tilde{f}_k : \tilde{\mathbf{X}} \to \tilde{\mathbf{Y}}\ k\in[K]\}$ 为假设空间，并对每个 $k\in[K]$记 $\tilde{\mathbf{F}}_k = \{\tilde{f}_k \mid \tilde{f}_k : \tilde{\mathbf{X}} \to \tilde{\mathbf{Y}}\}$。考虑一个损失函数 $L : \mathbf{X} \times \mathbf{Y} \times \mathbf{F}_k \to \mathbb{R}_+$。对于假设 $f\in\mathbf{F}$ 和训练集 S，f 的经验风险定义为

$$\widehat{R}_{\tilde{S}}(\tilde{f}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m_k}\sum_{i=1}^{m_k}L(\tilde{\mathbf{x}}_{ki}, \tilde{y}_{ki}, \tilde{f}_k),$$

泛化（或期望）风险定义为

$$R(\tilde{f}) = \mathop{\mathbb{E}}_{\tilde{S}}\left[\widehat{R}_{\tilde{S}}(\tilde{f})\right]. \tag{15}$$

请注意，我们没有将泛化风险定义为以下常见形式。

$$\frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{(\tilde{\mathbf{x}},\tilde{y})\sim D_k}\left[L(\tilde{\mathbf{x}},\tilde{y},\tilde{f}_k)\right]. \tag{16}$$

这是因为方程（15）中的定义比方程（16）更一般。具体而言，方程（15）可以涵盖$\tilde{}$ $\tilde{}$依赖于训练集的损失函数 S，而方程（16）则不能。此外，对于某些独立于 S 的损失，两者相等。

A.2.2. 损失空间的分数拉德马赫复杂度

这里给出损失空间的分数拉德马赫复杂度定义如下。

定义 7（损失空间的分数拉德马赫复杂度）。对于每个 $k \in [K]$，令 $\{(I_{kj},\omega_{kj})\}_{j\in[J_k]}$ 为在 S_k 上构造的依赖图 G_k 的一个分数独立顶点覆盖，其中 $\sum_{j\in[J_k]}\omega_{kj} = \chi_f(G_k)$。令 $\widetilde{\mathrm{F}} = \{\tilde{f} = (\tilde{f}_1,\ldots,\tilde{f}_K) \mid \tilde{f}_k : \widetilde{\mathrm{X}} \to \widetilde{\mathrm{Y}}\ k \in [K]\}$ 为假设空间。那么，给定 S 时，F 的经验分数拉德马赫$\widetilde{}$ $\tilde{}$复杂度定义为

$$\widehat{\mathfrak{R}}^{*}_{\tilde{S}}(L\circ\widetilde{\mathcal{F}}) = \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_{kj}}\sigma_{ki}L(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki},\tilde{f}_k)\right)\right],$$

其中 $\boldsymbol{\sigma} = (\sigma_{ki})_{k\in[K],\ i\in[m_k]}$ 表示 m 独立的拉德马赫变量，也就是说，对于$\widetilde{}$所有变量，$\mathbb{P}(\sigma_{ki} = +1) = \mathbb{P}(\sigma_{ki} = -1) = 1/2$。此外，$\mathrm{F}$ 在所有大小为 m 的样本上的分数拉德马赫复杂度定义为

$$\mathfrak{R}^{*}_{m}(L\circ\widetilde{\mathcal{F}}) = \mathop{\mathbb{E}}_{\tilde{S}\sim D^{m}_{[K]}}\left[\widehat{\mathfrak{R}}^{*}_{\tilde{S}}(L\circ\widetilde{\mathcal{F}})\right],$$

$\tilde{}$ $\tilde{}$ $\tilde{}$其中 $S\sim D^{m}_{[K]}$ 表示 $S_1\sim D_1^{m_1}$、.....、$S_K\sim D_K^{m_K}$，为简洁起见。

A.2.3. 图依赖样本下多任务学习的一般泛化界证明

下面给出图依赖样本下多任务学习的一般泛化界。

定理 6（图依赖样本下多任务学习的一般泛化界）。给定一个 $\tilde{}$ $\tilde{}$ $\tilde{}$ $\tilde{}$ $\widetilde{}$ $\widetilde{}$ $\widetilde{}$样本 $S = \{S_1、.....、S_K\}$，其中每个 $S_{k\in[K]}$ 的大小为 m_k，其依赖图为 G_k，损失函数为 $L : \mathrm{X}\times\mathrm{Y}\times\mathrm{F}_k \to [0,M]$。那么，对于任意 $\delta\in(0,1)$，以至少 $1-\delta$ 的概率，有

$$\forall\tilde{f}\in\widetilde{\mathcal{F}},\ R(\tilde{f}) \le \widehat{R}_{\tilde{S}}(\tilde{f}) + 2\mathfrak{R}^{*}_{m}(L\circ\widetilde{\mathcal{F}}) + M\sqrt{\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\chi_f(G_k)}{2m_k}\right)\log\left(\frac{1}{\delta}\right)}, \tag{17}$$

and

$$\forall\tilde{f}\in\widetilde{\mathcal{F}},\ R(\tilde{f}) \le \widehat{R}_{\tilde{S}}(\tilde{f}) + 2\widehat{\mathfrak{R}}^{*}_{\tilde{S}}(L\circ\widetilde{\mathcal{F}}) + 3M\sqrt{\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\chi_f(G_k)}{2m_k}\right)\log\left(\frac{2}{\delta}\right)}. \tag{18}$$

证明。证明过程可分为以下三个主要步骤。

步骤 1：将 $R(\tilde{f}) - \widehat{R}_{\tilde{S}}(\tilde{f})$ 在 $\widetilde{\mathrm{F}}$ 上的上确界与其期望联系起来。

对于任意 $\tilde{f} \in \widetilde{\mathrm{F}}$， $\widehat{R}(\tilde{f})$ 是 $R(\tilde{f})$ 的无偏估计量，因为样本 $\tilde{S}_k$ 中的数据点被假设为 G-依赖，且具有相同的边缘分布。因此，考虑一个独立的幽灵样本 S'，其生成过程与 S 相同，我们有

$$\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(R(\tilde{f})-\widehat{R}_{\tilde{S}}(\tilde{f})\right)=\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\mathop{\mathbb{E}}_{\tilde{S}'}[\widehat{R}_{\tilde{S}'}(\tilde{f})]-\widehat{R}_{\tilde{S}}(\tilde{f})\right)=\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\mathop{\mathbb{E}}_{\tilde{S}'}\left[\widehat{R}_{\tilde{S}'}(\tilde{f})-\widehat{R}_{\tilde{S}}(\tilde{f})\right]\right).$$

对于每个 $k\in[K]$，令 $\{(I_{kj},\omega_{kj})\}_{j\in[J_k]}$ 为依赖图 G_k with $\sum_{j\in[J_k]}\omega_{kj}=\chi_f(G_k)$ 的分数独立顶点覆盖。由于期望的上确界低于上确界的期望，我们可以得到

$$\begin{aligned}
\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\mathop{\mathbb{E}}_{\tilde{S}'}\left[\widehat{R}_{\tilde{S}'}(\tilde{f})-\widehat{R}_{\tilde{S}}(\tilde{f})\right]\right)&\le\mathop{\mathbb{E}}_{\tilde{S}'}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\widehat{R}_{\tilde{S}'}(\tilde{f})-\widehat{R}_{\tilde{S}}(\tilde{f})\right)\right]\\
&=\mathop{\mathbb{E}}_{\tilde{S}'}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{1}{m_k}\sum_{i=1}^{m_k}\left(L(\tilde{\mathbf{x}}'_{ki},\tilde{y}'_{ki},\tilde{f}_k)-L(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki},\tilde{f}_k)\right)\right)\right]\\
&\overset{①}{=}\mathop{\mathbb{E}}_{\tilde{S}'}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in[J_k]}\frac{\omega_{kj}}{m_k}\sum_{i\in I_{kj}}\left(L(\tilde{\mathbf{x}}'_{ki},\tilde{y}'_{ki},\tilde{f}_k)-L(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki},\tilde{f}_k)\right)\right)\right]\\
&\overset{②}{\le}\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in[J_k]}\frac{\omega_{kj}}{m_k}\mathop{\mathbb{E}}_{\tilde{S}'_k}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_{kj}}\left(L(\tilde{\mathbf{x}}'_{ki},\tilde{y}'_{ki},\tilde{f}_k)-L(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki},\tilde{f}_k)\right)\right)\right],
\end{aligned}$$

其中，不等式 1 源于Janson 分解（Janson，2004），不等式 2 源于上确界函数的次可加性（即 $\sup(a+b)\le\sup(a)+\sup(b)$）以及期望的线性性。

通过定义 $g(\tilde{S})=\sum_{k=1}^{K}g_k(\tilde{S}_k)$，其中每个 $g_k:\tilde{S}_k\mapsto\sum_{j\in[J_k]}\omega_{kj}g_{kj}(I_{kj})$ 满足

$$g_{kj}:I_{kj}\mapsto\frac{1}{Km_k}\mathop{\mathbb{E}}_{\tilde{S}'_k}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_{kj}}\left(L(\tilde{\mathbf{x}}'_{ki},\tilde{y}'_{ki},\tilde{f}_k)-L(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki},\tilde{f}_k)\right)\right)\right]$$

在差异被 $\frac{M}{Km_k}$ 所界定的意义下，满足定理 5的条件；那么对于任意 $\delta\in(0,1)$，以至少 $1-\delta$ 的概率，我们有

$$\begin{aligned}
&\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(R(\tilde{f})-\widehat{R}_{\tilde{S}}(\tilde{f})\right)\\
&\le\underbrace{\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in[J_k]}\frac{\omega_{kj}}{m_k}\mathop{\mathbb{E}}_{\tilde{S}_k,\tilde{S}'_k}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_{kj}}\left(L(\tilde{\mathbf{x}}'_{ki},\tilde{y}'_{ki},\tilde{f}_k)-L(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki},\tilde{f}_k)\right)\right)\right]}_{\overset{def}{=}\bigstar}+M\sqrt{\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\chi_f(G_k)}{m_k}\right)\log\left(\frac{1}{\delta}\right)}.
\end{aligned}$$

步骤 2：根据分数拉德马赫复杂度界定 ☀。

接下来，通过引入拉德马赫变量采用对称化技术，我们得到

$$\begin{aligned}
\bigstar&=\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in[J_k]}\frac{\omega_{kj}}{m_k}\mathop{\mathbb{E}}_{\tilde{S}_k,\tilde{S}'_k}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_{kj}}\left(L(\tilde{\mathbf{x}}'_{ki},\tilde{y}'_{ki},\tilde{f}_k)-L(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki},\tilde{f}_k)\right)\right)\right]\\
&=\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in[J_k]}\frac{\omega_{kj}}{m_k}\mathop{\mathbb{E}}_{\tilde{S}_k,\tilde{S}'_k}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_{kj}}\sigma_{ki}\left(L(\tilde{\mathbf{x}}'_{ki},\tilde{y}'_{ki},\tilde{f}_k)-L(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki},\tilde{f}_k)\right)\right)\right]\\
&\overset{③}{\le}\frac{2}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\tilde{S}_k}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_{kj}}\sigma_{ki}L(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki},\tilde{f}_k)\right)\right]\\
&=2\mathfrak{R}^{*}_{m}(L\circ\widetilde{\mathcal{F}}).
\end{aligned}$$

对于固定的一对 $(k,\ i)$，$\sigma_{ki}=1$ 并未改变任何内容，只是 $\sigma_{ki}=-1$ 交换两个样本 $(\tilde{\mathbf{x}}'_{ki},\ \tilde{y}'_{ki})$ 和 $(\tilde{\mathbf{x}}_{ki},\ \tilde{y}_{ki})$。因此，对 S_k 和 S'_k 取期望时，引入拉德马赫变量不会改变其值。对于 3，这是由于上确界函数的次可加性以及期望的线性性。

因此，我们可以得到

$$\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(R(\tilde{f})-\widehat{R}_{\tilde{S}}(\tilde{f})\right)\leq 2\mathfrak{R}^{*}_{m}(L\circ\widetilde{\mathcal{F}})+M\sqrt{\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\chi_f(G_k)}{m_k}\right)\log\left(\frac{1}{\delta}\right)}.$$

此外，根据函数上确界的定义，我们有

$$\forall\,\tilde{f}\in\widetilde{\mathcal{F}},\ R(\tilde{f})-\widehat{R}_{\tilde{S}}(\tilde{f})\leq\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(R(\tilde{f})-\widehat{R}_{\tilde{S}}(\tilde{f})\right).$$

然后，我们可以得到所需的第一个界限（17）。

步骤 3：用经验拉德马赫复杂度界定分数拉德马赫复杂度。

通过定义 $g(\tilde{S})=\sum_{k=1}^{K}g_k(\tilde{S}_k)$，其中每个 $g_k:\tilde{S}_k\mapsto\sum_{j\in[J_k]}\omega_{kj}g_{kj}(I_{kj})$，其中每个

$$g_{kj}:I_{kj}\mapsto\frac{1}{Km_k}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum_{j\in[J_k]}\omega_{kj}\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_{kj}}\sigma_{ki}L(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki},\tilde{f}_k)\right)\right]$$

差异有界于 $\frac{M}{Km_k}$，其意义与定理5中的条件一致；那么，对于任意 $\delta\in(0,\ 1)$，以至少 $1-\delta$ 的概率，有

$$\mathfrak{R}^{*}_{m}(L\circ\widetilde{\mathcal{F}})\leq\widehat{\mathfrak{R}}^{*}_{\tilde{S}}(L\circ\widetilde{\mathcal{F}})+M\sqrt{\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\chi_f(G_k)}{m_k}\right)\log\left(\frac{1}{\delta}\right)}.$$

然后，利用并合界以及第一个界（17），即可得到所需的第二个界（18）。 □

B. 多标签分类中的宏平均AUC 最大化

B.1. 定理1的证明

B.1.1. 问题转换

对于多标签学习中的宏平均AUC 最大化问题，我们可以将其转换为一个多任务学习问题，其中样本依赖于图；该问题已在第A.2.1节中讨论。

具体而言，根据原始训练集 S 构造训练数据集 S，方法如下。对于每个标签 $k\in[K]$，基于原始数据集 S_k 构造数据集 $S_k=\{(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki})\}_{i=1}^{m_k}$，其中 $\tilde{\mathbf{x}}_{ki}=(\tilde{\mathbf{x}}^{+}_{ki},\tilde{\mathbf{x}}^{-}_{ki})$，$\tilde{y}_{ki}=1$，且 $(\tilde{\mathbf{x}}^{+}_{ki},\tilde{\mathbf{x}}^{-}_{ki})\in S^{+}_{k}\times S^{-}_{k}$，$m_k=|S^{+}_{k}||S^{-}_{k}|=n^2\tau_k(1-\tau_k)$，并令 $\{(I_{kj},\ \omega_{kj})\}^{j\in[J_k]}$ 为依赖图 G_k 上的分数独立顶点覆盖；该图在 S_k 上构造，满足 $\sum_{j\in[J_k]}\omega_{kj}=\chi_f(G_k)$，其中 $G_k(G_k)$ 是 G_k 的分数色数。根据二部排序领域先前的研究结果（Usunier et al., 2005；Amini & Usunier, 2015），我们知道

$$\forall k\in[K],\ \chi_f(G_k)=\max\{|S^{+}_{k}|,|S^{-}_{k}|\}=(1-\tau_k)n.$$

此外，对于每个标签 $k\in[K]$，$f_k(\tilde{\mathbf{x}}_i)=f_k(\mathbf{x}^{+}_{i})-f_k(\mathbf{x}^{-}_{i})$。

B.1.2. 定理的证明 1

定理 1（Macro-AUC的基本定理）。假设损失函数 $L_\phi:\mathrm{X}\times\mathrm{X}\times\mathrm{F}_k\to\mathbb{R}_+$ 被 M所界定。那么，对于任意 $\delta>0$，以至少 $1-\delta$ 的概率，在抽取大小为 n的独立同分布样本 S时，以下泛化界成立：

$$\forall f\in\mathcal{F},\ R_\phi(f)\leq\widehat{R}^{\phi}_{S}(f)+2\widehat{\mathfrak{R}}^{*}_{\tilde{S}}(L_\phi\circ\mathcal{F})+3M\sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}\left(\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k}}\right).$$

证明。 基于第 B.1.1 节中的问题转换，通过应用定理 6，可以直接得到该定理。

□

B.2. 引理的证明 1 和 2

B.2.1. 引理证明 引理 1

引理 1（替代损失与真实损失之间的关系）。假设基础损失函数是原始 0/1 损失的上界，即 $\ell(t) \geq [\![t \leq 0]\!]$。那么，对于任意 $f_k \in \mathrm{F}_k$ 和 $(\mathbf{x}^+, \mathbf{x}^-) \in S_k^+ \times S_k^-$，以下不等式成立：

$$\begin{aligned}
&L_{0/1}(\mathbf{x}^+, \mathbf{x}^-, f_k) \leq L_{pa}(\mathbf{x}^+, \mathbf{x}^-, f_k), \\
&L_{0/1}(\mathbf{x}^+, \mathbf{x}^-, f_k) \leq L_{u_2}(\mathbf{x}^+, \mathbf{x}^-, f_k) \leq \frac{1}{\tau_k} L_{u_1}(\mathbf{x}^+, \mathbf{x}^-, f_k) \\
&\leq \frac{1-\tau_k}{\tau_k} L_{u_2}(\mathbf{x}^+, \mathbf{x}^-, f_k).
\end{aligned}$$

证明。对于第一个不等式，以下结论成立：

$$L_{0/1}(\mathbf{x}^+, \mathbf{x}^-, f_k) = [\![f_k(\mathbf{x}^+) \leq f_k(\mathbf{x}^-)]\!] \leq \ell(f_k(\mathbf{x}_p) - f_k(\mathbf{x}_q)) = L_{pa}(\mathbf{x}^+, \mathbf{x}^-, f_k).$$

对于第二个不等式，有如下结论：

$$\begin{aligned}
L_{0/1}(\mathbf{x}^+, \mathbf{x}^-, f_k) &= [\![f_k(\mathbf{x}^+) \leq f_k(\mathbf{x}^-)]\!] \\
&\leq [\![\operatorname{sgn}(f_k(\mathbf{x}^+)) \leq \operatorname{sgn}(f_k(\mathbf{x}^-))]\!] \\
&= [\![\operatorname{sgn}(f_k(\mathbf{x}^+)) \neq +1]\!] + [\![\operatorname{sgn}(f_k(\mathbf{x}^-)) \neq -1]\!] - [\![\operatorname{sgn}(f_k(\mathbf{x}^+)) \neq +1]\!][\![\operatorname{sgn}(f_k(\mathbf{x}^-)) \neq -1]\!] \\
&\leq [\![\operatorname{sgn}(f_k(\mathbf{x}^+)) \neq +1]\!] + [\![\operatorname{sgn}(f_k(\mathbf{x}^-)) \neq -1]\!] \\
&\leq \ell(f_k(\mathbf{x}^+)) + \ell(-f_k(\mathbf{x}^-)) \\
&= L_{u_2}(\mathbf{x}^+, \mathbf{x}^-, f_k) \\
&= \frac{n}{\min\{|S_k^+|, |S_k^-|\}} \left(\frac{\min\{|S_k^+|, |S_k^-|\}}{n} \ell(f_k(\mathbf{x}^+)) + \frac{\min\{|S_k^+|, |S_k^-|\}}{n} \ell(-f_k(\mathbf{x}^-)) \right) \\
&\leq \frac{1}{\tau_k} \left(\frac{|S_k^+|}{n} \ell(f_k(\mathbf{x}^+)) + \frac{|S_k^-|}{n} \ell(-f_k(\mathbf{x}^-)) \right) \\
&= \frac{1}{\tau_k} L_{u_1}(\mathbf{x}^+, \mathbf{x}^-, f_k) \\
&\leq \frac{1}{\tau_k} \left(\frac{\max\{|S_k^+|, |S_k^-|\}}{n} \ell(f_k(\mathbf{x}^+)) + \frac{\max\{|S_k^+|, |S_k^-|\}}{n} \ell(-f_k(\mathbf{x}^-)) \right) \\
&\leq \frac{\max\{|S_k^+|, |S_k^-|\}}{\min\{|S_k^+|, |S_k^-|\}} \left(\ell(f_k(\mathbf{x}^+)) + \ell(-f_k(\mathbf{x}^-)) \right) \\
&= \frac{1-\tau_k}{\tau_k} L_{u_2}(\mathbf{x}^+, \mathbf{x}^-, f_k).
\end{aligned}$$

因此，不等式成立。

□

B.2.2. 引理的证明 2

引理 2（替代损失与真实风险之间的关系）。假设基础损失函数以上界形式界定原始 0/1 损失，即 $\ell(t) \geq [\![t \leq 0]\!]$。那么，对于任意 $f \in \mathrm{F}$ 和任意样本 $S \overset{i.i.d.}{\sim} P$，以下不等式成立：

$$\begin{aligned}
&R_{0/1}(f) \leq R_{pa}(f), \\
&R_{0/1}(f) \leq R_{u_2}(f) = \mathop{\mathbb{E}}_{S} \left[\widehat{R}_S^{u_2}(f) \right] \leq \mathop{\mathbb{E}}_{S} \left[\frac{1}{\tau_S^*} \widehat{R}_S^{u_1}(f) \right] \\
&\leq \mathop{\mathbb{E}}_{S} \left[\frac{1-\tau_S^*}{\tau_S^*} \widehat{R}_S^{u_2}(f) \right].
\end{aligned}$$

证明。对于第一个不等式，以下结论成立：

$$\begin{aligned}
R_{0/1}(f) &= \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\mathbf{x}_p\sim P_k^+,\mathbf{x}_q\sim P_k^-}[\![f_k(\mathbf{x}_p)\le f_k(\mathbf{x}_q)]\!] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\mathbf{x}_p\sim P_k^+,\mathbf{x}_q\sim P_k^-}\left[L_{0/1}(\mathbf{x}_p,\mathbf{x}_q,f_k)\right] \\
&\le \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\mathbf{x}_p\sim P_k^+,\mathbf{x}_q\sim P_k^-}\left[L_{pa}(\mathbf{x}_p,\mathbf{x}_q,f_k)\right] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\mathbf{x}_p\sim P_k^+,\mathbf{x}_q\sim P_k^-}\left[\ell(f_k(\mathbf{x}_p)-f_k(\mathbf{x}_q))\right] \\
&= \mathop{\mathbb{E}}_{S}\left[\widehat{R}_S^{pa}(f)\right] \\
&= R_{pa}(f)
\end{aligned}$$

对于第二个不等式，我们首先有如下结论

$$\begin{aligned}
R_{u_1}(f) &= \mathop{\mathbb{E}}_{S}\left[\widehat{R}_S^{u_1}(f)\right] \\
&= \mathop{\mathbb{E}}_{S}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-}L_{u_1}(\mathbf{x}_p,\mathbf{x}_q,f_k)\right] \\
&= \mathop{\mathbb{E}}_{S}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-}\left(\frac{|S_k^+|}{n}\ell(f_k(\mathbf{x}_p))+\frac{|S_k^-|}{n}\ell(-f_k(\mathbf{x}_q))\right)\right] \\
&\ge \mathop{\mathbb{E}}_{S}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{\tau_k}{|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-}\left(\ell(f_k(\mathbf{x}_p))+\ell(-f_k(\mathbf{x}_q))\right)\right].
\end{aligned}$$

由此可得

$$\begin{aligned}
R_{0/1}(f) &= \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\mathbf{x}_p\sim P_k^+,\mathbf{x}_q\sim P_k^-}\left[L_{0/1}(\mathbf{x}_p,\mathbf{x}_q,f_k)\right] \\
&\le \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\mathbf{x}_p\sim P_k^+,\mathbf{x}_q\sim P_k^-}\left[L_{u_2}(\mathbf{x}_p,\mathbf{x}_q,f_k)\right] \\
&= \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\mathbf{x}_p\sim P_k^+,\mathbf{x}_q\sim P_k^-}\left[\ell(f_k(\mathbf{x}_p))+\ell(-f_k(\mathbf{x}_q))\right] \\
&= \mathop{\mathbb{E}}_{S}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-}\left(\ell(f_k(\mathbf{x}_p))+\ell(-f_k(\mathbf{x}_q))\right)\right] \\
&= \mathop{\mathbb{E}}_{S}\left[\widehat{R}_S^{u_2}(f)\right] \\
&= R_{u_2}(f) \\
&\le \mathop{\mathbb{E}}_{S}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-}\left(\frac{|S_k^+|}{n}\ell(f_k(\mathbf{x}_p))+\frac{|S_k^-|}{n}\ell(-f_k(\mathbf{x}_q))\right)\right] \\
&\le \mathop{\mathbb{E}}_{S}\left[\frac{1}{K\tau_S^*}\sum_{k=1}^{K}\frac{1}{|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-}\left(\frac{|S_k^+|}{n}\ell(f_k(\mathbf{x}_p))+\frac{|S_k^-|}{n}\ell(-f_k(\mathbf{x}_q))\right)\right] \\
&= \mathop{\mathbb{E}}_{S}\left[\frac{1}{\tau_S^*}\widehat{R}_S^{u_1}(f)\right] \\
&\le \mathop{\mathbb{E}}_{S}\left[\frac{1}{K\tau_S^*}\sum_{k=1}^{K}\frac{\max\{S_k^+,S_k^-\}}{n|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-}\left(\ell(f_k(\mathbf{x}_p))+\ell(-f_k(\mathbf{x}_q))\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathop{\mathbb{E}}_{S}\left[\frac{1}{K\tau_S^*}\sum_{k=1}^{K}\frac{1-\tau_k}{|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-}\left(\ell(f_k(\mathbf{x}_p))+\ell(-f_k(\mathbf{x}_q))\right)\right]\\
&\leq \mathop{\mathbb{E}}_{S}\left[\frac{1-\tau_S^*}{K\tau_S^*}\sum_{k=1}^{K}\frac{1-\tau_k}{|S_k^+||S_k^-|}\sum_{(p,q)\in S_k^+\times S_k^-}\left(\ell(f_k(\mathbf{x}_p))+\ell(-f_k(\mathbf{x}_q))\right)\right]\\
&= \mathop{\mathbb{E}}_{S}\left[\frac{1-\tau_S^*}{\tau_S^*}\widehat{R}_S^{u_2}(f)\right].
\end{aligned}
$$

因此，第二个不等式成立。 □

B.3. 定理的证明 2、推论4 和 2

B.3.1. 假设空间的分数拉德马赫复杂度

定义 8（假设空间针对 L_{pa} 的分数拉德马赫复杂度）。给定一个数据集 S （及其 $\tilde{}$ 对应的构造数据集 S），定义假设空间 $\widetilde{\mathrm{F}}$ 关于 S 的经验分数拉德马赫复杂度如下:

$$
\begin{aligned}
\widehat{\mathfrak{R}}_{\tilde{S}}^{*}(\widetilde{\mathcal{F}}) &= \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_{kj}}\sigma_{ki}\tilde{f}_k(\tilde{\mathbf{x}}_{ki})\right)\right]\\
&= \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_{kj}}\sigma_{ki}(f_k(\tilde{\mathbf{x}}_{ki}^{+}-f_k(\tilde{\mathbf{x}}_{ki}^{-}))\right)\right].
\end{aligned}
$$

引理 4 (基于核的假设空间关于 L_{pa})。 假设 假设 1 中的 (1) 和 (2) 成立。那么，对于基于核的假设空间 (8)，其经验分数拉德马赫复杂度$\tilde{}$关于数据集 S，可按如下方式界定:

$$
\widehat{\mathfrak{R}}_{\tilde{S}}^{*}(\widetilde{\mathcal{F}}) \leq \frac{2Br}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)
$$

证明。根据 $\widehat{\mathfrak{R}}_{\tilde{S}}^{*}(\ \widetilde{}\mathrm{F})$ 的定义，我们有

$$
\begin{aligned}
\widehat{\mathfrak{R}}_{\tilde{S}}^{*}(\widetilde{\mathcal{F}}) &= \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_{kj}}\sigma_{ki}\tilde{f}_k(\tilde{\mathbf{x}}_{ki})\right)\right]\\
&= \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{\|\mathbf{w}_k\|\leq B}\left(\sum_{i\in I_{kj}}\sigma_{ki}\langle\mathbf{w}_k,\Phi(\tilde{\mathbf{x}}_{ki})\rangle\right)\right]\\
&\leq \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\|\sup_{\|\mathbf{w}_k\|\leq B}\|\mathbf{w}_k\|\left\|\sum_{i\in I_{kj}}\sigma_{ki}\Phi(\tilde{\mathbf{x}}_{ki})\right\|\right] \qquad \text{(Cauchy–Schwarz inequality)}\\
&= \frac{1}{K}\sum_{k=1}^{K}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left\|\sum_{i\in I_{kj}}\sigma_{ki}\Phi(\tilde{\mathbf{x}}_{ki})\right\|\right] \qquad \text{(the definition of sup)}\\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i\in I_{kj}}\sigma_{ki}\Phi(\tilde{\mathbf{x}}_{ki})\right\|\right] \qquad \text{(linearity of expectation)}\\
&\leq \frac{1}{K}\sum_{k=1}^{K}\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i\in I_{kj}}\sigma_{ki}\Phi(\tilde{\mathbf{x}}_{ki})\right\|^2\right]\right)^{\frac{1}{2}} \qquad \text{(Jensen's inequality)}\\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum_{p\in I_{kj},q\in I_{kj}}\sigma_{kp}\sigma_{kq}\left\langle\Phi(\tilde{\mathbf{x}}_{kp}),\Phi(\tilde{\mathbf{x}}_{kq})\right\rangle\right]\right)^{\frac{1}{2}}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{K}\sum_{k=1}^{K}\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\sum_{i\in I_{kj}}\langle\Phi(\tilde{\mathbf{x}}_{ki}),\Phi(\tilde{\mathbf{x}}_{ki})\rangle\right)^{\frac{1}{2}} && (\forall p\neq q, \mathbb{E}[\sigma_{kp}\sigma_{kq}]=0 \text{ and } \mathbb{E}[\sigma_{ki}\sigma_{ki}]=1)\\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\sum_{i\in I_{kj}}\langle\Phi(\mathbf{x}_{ki}^{+})-\Phi(\mathbf{x}_{ki}^{-}),\Phi(\mathbf{x}_{ki}^{+})-\Phi(\mathbf{x}_{ki}^{-})\rangle\right)^{\frac{1}{2}} && (\Phi(\tilde{\mathbf{x}})=\Phi(\mathbf{x}^{+})-\Phi(\mathbf{x}^{-}))\\
&\le \frac{1}{K}\sum_{k=1}^{K}\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\sum_{i\in I_{kj}}(\langle\Phi(\mathbf{x}_{ki}^{+}),\Phi(\mathbf{x}_{ki}^{+})\rangle+\langle\Phi(\mathbf{x}_{ki}^{-}),\Phi(\mathbf{x}_{ki}^{-})\rangle)\right)^{\frac{1}{2}} && (\|a+b\|^2\le 2\|a\|^2+2\|b\|^2)\\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\sum_{i\in I_{kj}}2\left(\kappa(\mathbf{x}_{ki}^{+},\mathbf{x}_{ki}^{+})+\kappa(\mathbf{x}_{ki}^{-},\mathbf{x}_{ki}^{-})\right)\right)^{\frac{1}{2}} && (\kappa(\mathbf{x},\mathbf{x})=\langle\Phi(\mathbf{x}),\Phi(\mathbf{x})\rangle)\\
&\le \frac{1}{K}\sum_{k=1}^{K}\frac{2Br}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sqrt{m_{kj}} && (\kappa(\mathbf{x},\mathbf{x})\le r^2 \text{ and let } |I_{kj}|=m_{kj})\\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{2Br\chi_f(G_k)}{m_k}\sum_{j\in[J_k]}\frac{\omega_{kj}}{\chi_f(G_k)}\sqrt{m_{kj}}\\
&\le \frac{1}{K}\sum_{k=1}^{K}\frac{2Br\sqrt{\chi_f(G_k)}}{m_k}\sqrt{\sum_{j\in[J_k]}\omega_{kj}m_{kj}} && (\sum_{j\in[J_k]}\frac{\omega_{kj}}{\chi_f(G_k)}=1 \text{ and Jensen's inequality})\\
&= \frac{1}{K}\sum_{k=1}^{K}2Br\sqrt{\frac{\chi_f(G_k)}{m_k}} && (\sum_{j\in[J_k]}\omega_{kj}m_{kj}=m_k)
\end{aligned}
$$

由于在多标签学习中优化宏平均AUC时，$\chi_f(G_k)=\max\{|S_k^+|,|S_k^-|\}$、$m_k=|S_k^+||S_k^-|$和$\tau_k=\min\{|S_k^+|,|S_k^-|\}$均成立，因此我们可以得到

$$
\frac{1}{K}\sum_{k=1}^{K}2Br\sqrt{\frac{\chi_f(G_k)}{m_k}}=\frac{2Br}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right).
$$

因此，结合上述等式和前面的不等式，即可得到所需结果。 □

B.3.2. 收缩不等式

˜引理 5（L_{pa}的收缩不等式）。假设损失函数 $L_\phi=L_1(y,f_k(\tilde{}\mathbf{x}))$ 关于第二个参数是 μ-利普希茨连续的，其中 L_1 为方便起见表示两个输入的损失函数。则有：

$$
\widehat{\mathfrak{R}}_{\tilde{S}}^{*}(L_\phi\circ\mathcal{F})\le\mu\widehat{\mathfrak{R}}^{*}(\widetilde{\mathcal{F}}).
$$

证明。由于 $\widehat{\mathfrak{R}}_{\tilde{S}}^{*}(L_\phi\bigcirc \mathrm{F})=\frac{1}{K}\sum_{k=1}^{K}\widehat{\mathfrak{R}}_{\tilde{S}_k}^{*}(L_\phi\bigcirc \mathrm{F}_k)$ 且 $\widehat{\mathfrak{R}}_{\tilde{S}}^{*}(\mathrm{F})=\frac{1}{K}\sum_{k=1}^{K}\widehat{\mathfrak{R}}_{\tilde{S}_k}^{*}(\mathrm{F}_k)$，我们首先证明 $\widehat{\mathfrak{R}}_{\tilde{S}_k}^{*}(L_\phi\bigcirc \mathrm{F}_k)\le\mu\widehat{\mathfrak{R}}_{\tilde{S}_k}^{*}(\mathrm{F}_k)$，随后即可得到所需结果。

下面我们按照 $\widehat{\mathfrak{R}}_{\tilde{S}_k}^{*}(L_\phi\circ\mathcal{F}_k)\le\mu\widehat{\mathfrak{R}}_{\tilde{S}_k}^{*}(\mathcal{F}_k)$的思路证明不等式Mohri 等人 (2018)（引理 5.7，第 93 页），˜并且为简化后续记号，我们省略指标 k 和符号 S_k。

首先固定一个样本（$\tilde{}\mathbf{x}_1,\ldots,\tilde{}\mathbf{x}_m$），然后根据定义，

$$
\begin{aligned}
\widehat{\mathfrak{R}}^{*}(L_\phi\circ f)&=\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{j\in[J]}w_j\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_j}\sigma_i L_1(y_i,\tilde{f}(\tilde{\mathbf{x}}_i))\right)\right]\\
&=\frac{1}{m}\sum_{j\in[J]}w_j\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_{n_j-1}}\left[\mathop{\mathbb{E}}_{\sigma_{n_j}}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}u_{n_j-1}(\tilde{f})+\sigma_{n_j}L_1(y_{n_j},\tilde{f}(\tilde{\mathbf{x}}_{n_j}))\right]\right], \qquad (\text{denote } n_j=|I_j| \text{ for simplicity})
\end{aligned}
$$

其中 $u_{n_j-1}(\tilde{f}) = \sum_{i=1}^{n_j} \sigma_i L_1(y_i, \tilde{f}(\tilde{\mathbf{x}}_i))$。根据上确界的定义，对于任意 $\epsilon > 0$，存在 $\tilde{f}^1, \tilde{f}^2 \in \widetilde{\mathbf{F}}$，使得

$$u_{n_j-1}(\tilde{f}^1) + L_1(y_{n_j}, \tilde{f}^1(\tilde{\mathbf{x}}_{n_j})) \le (1-\epsilon)\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}} u_{n_j-1}(\tilde{f}) + L_1(y_{n_j}, \tilde{f}(\tilde{\mathbf{x}}_{n_j}))\right],$$

and

$$u_{n_j-1}(\tilde{f}^2) - L_1(y_{n_j}, \tilde{f}^2(\tilde{\mathbf{x}}_{n_j})) \le (1-\epsilon)\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}} u_{n_j-1}(\tilde{f}) - L_1(y_{n_j}, \tilde{f}(\tilde{\mathbf{x}}_{n_j}))\right].$$

因此，对于任意 $\epsilon > 0$，根据 $\mathbb{E}$的定义$_{\sigma_{n_j}}$，

$$\begin{aligned}
&(1-\epsilon)\mathop{\mathbb{E}}_{\sigma_{n_j}}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}} u_{n_j-1}(\tilde{f}) + \sigma_{n_j} L_1(y_{n_j}, \tilde{f}(\tilde{\mathbf{x}}_{n_j}))\right] \\
&= (1-\epsilon)\left[\frac{1}{2}\sup_{\tilde{f}\in\widetilde{\mathcal{F}}} u_{n_j-1}(\tilde{f}) + L_1(y_{n_j}, \tilde{f}(\tilde{\mathbf{x}}_{n_j}))\right] + \left[\frac{1}{2}\sup_{\tilde{f}\in\widetilde{\mathcal{F}}} u_{n_j-1}(\tilde{f}) - L_1(y_{n_j}, \tilde{f}(\tilde{\mathbf{x}}_{n_j}))\right] \\
&\le \frac{1}{2}\left[u_{n_j-1}(\tilde{f}^1) + L_1(y_{n_j}, \tilde{f}^1(\tilde{\mathbf{x}}_{n_j}))\right] + \frac{1}{2}\left[u_{n_j-1}(\tilde{f}^2) - L_1(y_{n_j}, \tilde{f}^2(\tilde{\mathbf{x}}_{n_j}))\right].
\end{aligned}$$

令 $s = \operatorname{sgn}(\tilde{f}^1(\tilde{\mathbf{x}}^{n_j}) - \tilde{f}^2(\tilde{\mathbf{x}}^{n_j}))$。那么，前一个不等式意味着

$$\begin{aligned}
&(1-\epsilon)\mathop{\mathbb{E}}_{\sigma_{n_j}}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}} u_{n_j-1}(\tilde{f}) + \sigma_{n_j} L_1(y_{n_j}, \tilde{f}(\tilde{\mathbf{x}}_{n_j}))\right] \\
&\le \frac{1}{2}\left[u_{n_j-1}(\tilde{f}^1) + u_{n_j-1}(\tilde{f}^2) + s\mu(\tilde{f}^1(\tilde{\mathbf{x}}_{n_j}) - \tilde{f}^2(\tilde{\mathbf{x}}_{n_j}))\right] && \text{(Lipschitz property)} \\
&= \frac{1}{2}\left[u_{n_j-1}(\tilde{f}^1) + s\mu\tilde{f}^1(\tilde{\mathbf{x}}_{n_j})\right] + \frac{1}{2}\left[u_{n_j-1}(\tilde{f}^2) - s\mu\tilde{f}^2(\tilde{\mathbf{x}}_{n_j})\right] && \text{(rearranging)} \\
&\le \frac{1}{2}\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left[u_{n_j-1}(\tilde{f}) + s\mu\tilde{f}(\tilde{\mathbf{x}}_{n_j})\right] + \frac{1}{2}\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left[u_{n_j-1}(\tilde{f}) - s\mu\tilde{f}(\tilde{\mathbf{x}}_{n_j})\right] && \text{(definition of sup)} \\
&= \mathop{\mathbb{E}}_{\sigma_{n_j}}\left[u_{n_j-1}(\tilde{f}) + \sigma_{n_j}\mu\tilde{f}(\tilde{\mathbf{x}}^{n_j})\right]. && \text{(definition of } \mathop{\mathbb{E}}_{\sigma_{n_j}}\text{)}
\end{aligned}$$

由于该不等式对任意 $\epsilon > 0$都成立，因此我们有

$$\mathop{\mathbb{E}}_{\sigma_{n_j}}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}} u_{n_j-1}(\tilde{f}) + \sigma_{n_j} L_1(y_{n_j}, \tilde{f}(\tilde{\mathbf{x}}_{n_j}))\right] \le \mathop{\mathbb{E}}_{\sigma_{n_j}}\left[u_{n_j-1}(\tilde{f}) + \sigma_{n_j}\mu\tilde{f}(\tilde{\mathbf{x}}_{n_j})\right].$$

以同样的方式处理所有其他 $\sigma_i (i \in [I_j], i \ne n_j)$，即可证明

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_j}\sigma_i L_1(y_i, \tilde{f}(\tilde{\mathbf{x}}_i))\right)\right] \le \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_j}\sigma_i\mu\tilde{f}(\tilde{\mathbf{x}}_i)\right)\right].$$

通过采用其他 $j \in [J]$，我们可以得到以下结果

$$\widehat{\mathfrak{R}}^*(L_\phi \circ f) = \frac{1}{m}\sum_{j\in[J]} w_j \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_j}\sigma_i L_1(y_i, \tilde{f}(\tilde{\mathbf{x}}_i))\right)\right] \le \frac{1}{m}\sum_{j\in[J]} w_j \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_j}\sigma_i\mu\tilde{f}(\tilde{\mathbf{x}}_i)\right)\right] = \mu\widehat{\mathfrak{R}}^*(\tilde{f}).$$

□

B.3.3. 定理证明2

定理 2 (一般情况下的学习保证 $\mathcal{A}^{pa}$ 在一般情况下)。 假设损失 $L_\phi = L_{pa}$，其中 L_{pa} 定义于式(5)。此外，假设 1 成立。那么，对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，在抽取大小为 S 的独立同分布样本n上，对任意 $f \in \mathcal{F}$，以下泛化界成立：

$$R_{0/1}(f) \le R_{pa}(f) \le \widehat{R}_{pa}(f) + \frac{4\rho r\Lambda}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right) + 3B\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\left(\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k}}\right). \tag{9}$$

证明。由于基础损失 $\ell(z)$ 有界于 B，且损失 $L_{pa}(\mathbf{x}^+, \mathbf{x}^-, f_k) = \ell(f_k(\mathbf{x}^+) - f_k(\mathbf{x}^-))$，因此损失 $L_\phi = L_{pa}$ 有界于 B。然后，应用定理 1，并结合引理 5 和引理 4，可得对于任意 $\delta > 0$，以下泛化界以至少 $1-\delta$ 的概率成立，其中概率是针对抽取大小为 n 的独立同分布样本 S：

$$R_{pa}(f) \le \widehat{R}_{pa}(f) + \frac{4\rho r\Lambda}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right) + 3B\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\left(\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k}}\right).$$

最后，结合上述不等式和引理2 （即， $R_{0/1}(f) \le R_{pa}(f)$）。 □

B.3.4. 证明 推论 4

推论 4 (平衡情形下 $\mathcal{A}^{pa}$ 的学习保证)。 假设损失 $L_\phi = L_{pa}$，其中 L_{pa} 定义于式(5)。此外，假设 1 成立，并假设 S 是平衡的。那么，对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，在抽取大小为 S 的独立同分布样本 n时，对于任意 $f \in \mathcal{F}$，以下泛化界成立：

$$R_{0/1}(f) \le R_{pa}(f) \le \widehat{R}_{pa}(f) + \frac{4\sqrt{2}\rho r\Lambda}{\sqrt{n}} + 3\sqrt{2}B\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}. \tag{19}$$

证明。 将 定理 2 中的 $\tau_k = \frac{1}{2}$ 代入，即可直接得到该结果。 □

B.3.5. 推论 2 的证明

推论 2 (极度不平衡情形下 $\mathcal{A}^{pa}$ 的学习保证)。 设损失函数 $L_\phi = L_{pa}$，其中 L_{pa} 定义于公式(5)中。此外，假设 假设 1 成立，并设 S 为极度不平衡情形。则对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，在抽取大小为 S 的独立同分布样本 n后，对任意$f \in \mathcal{F}$，以下泛化界成立：

$$R_{0/1}(f) \le R_{pa}(f) \le \widehat{R}_{pa}(f) + 4\rho r\Lambda + 3B\sqrt{\log(\frac{2}{\delta})}.$$

证明。 将 定理 2 中的 $\tau_k = \frac{1}{n}$ 代入，即可直接得到该结果。 □

B.4. 定理 3 和 4、推论1、3 和 5的证明

B.4.1. 假设空间的分数拉德马赫复杂度

定义 9（关于（重新加权）单变量损失的假设空间的分数拉德马赫复杂度）。 给定一个数据集 S（及其对应的构造数据集 $\tilde{S}$），假设对于每个标签 $k \in [K]$，损失函数 $L_\phi = L_u(\mathbf{x}^+, \mathbf{x}^-, f_k) = a^+(S_k)\ell(f_k(\mathbf{x}^+)) + a^-(S_k)\ell(-f_k(\mathbf{x}^-))$，其中 $\ell(t)$ 是基础损失函数，重新加权

函数 $a^+(S_k)$（或 $a^-(S_k)$）表示其对 S_k 的依赖关系。然后，定义假设空间 F 关于 S 和 L_u 的经验分数拉德马赫复杂度˜如下：

$$\widehat{\mathfrak{R}}^*_{\tilde{S},L_u}(\mathcal{F}) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{f\in\mathcal{F}}\left(\sum_{i\in I_{kj}}(\sigma^+_{ki}a^+(S_k)f_k(\mathbf{x}^+_{ki})+\sigma^-_{ki}a^-(S_k)f_k(\mathbf{x}^-_{ki}))\right)\right].$$

引理 6（关于（重新加权的）单变量损失，基于核的假设空间的分数拉德马赫复杂度）。假设假设 1 中的 (1) 和 (2) 成立，且损失函数 $L_u(\mathbf{x}^+,\mathbf{x}^-,f_k) = a^+(S_k)\ell(f_k(\mathbf{x}^+)) + a^-(S_k)\ell(-f_k(\mathbf{x}^-))$。那么，对于基于核的假设空间 (8)，其关于数据集 S 和损失函数 L_u 的经验分数拉德马赫复杂度˜，可由下式界定：

$$\widehat{\mathfrak{R}}^*_{\tilde{S},L_u}(\mathcal{F}) \le \frac{Br}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\,(a^+(S_k)+a^-(S_k))\right).$$

证明。根据 $\widehat{\mathfrak{R}}^*_{\tilde{S},L_u}$(F) 的定义，并为简化记号，令 a^+_k (或 a^-_k) 表示 $a^+(S_k)$（或 $a^-(S_k)$），我们有

$$\begin{aligned}
\widehat{\mathfrak{R}}^*_{\tilde{S},L_u}(\mathcal{F}) &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{f\in\mathcal{F}}\left(\sum_{i\in I_{kj}}(\sigma^+_{ki}a^+(S_k)f_k(\mathbf{x}^+_{ki})+\sigma^-_{ki}a^-(S_k)f_k(\mathbf{x}^-_{ki}))\right)\right]\\
&= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{\|\mathbf{w}_k\|\le B}\left(\sum_{i\in I_{kj}}(\sigma^+_{ki}a^+_k\langle\mathbf{w}_k,\Phi(\mathbf{x}^+_{ki})\rangle+\sigma^-_{ki}a^-_k\langle\mathbf{w}_k,\Phi(\mathbf{x}^-_{ki})\rangle)\right)\right]\\
&\overset{①}{\le} \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\sup_{\|\mathbf{w}_k\|\le B}\left(\sum_{i\in I_{kj}}\sigma^+_{ki}a^+_k\langle\mathbf{w}_k,\Phi(\mathbf{x}^+_{ki})\rangle\right)+\sup_{\|\mathbf{w}_k\|\le B}\left(\sum_{i\in I_{kj}}\sigma^-_{ki}a^-_k\langle\mathbf{w}_k,\Phi(\mathbf{x}^-_{ki})\rangle\right)\right)\right]\\
&\overset{def}{=} \clubsuit+\spadesuit
\end{aligned}$$

其中不等式 1 源于上确界函数的次可加性（即 $\sup(a+b)\le\sup a+\sup b$），并且

$$\clubsuit = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{\|\mathbf{w}_k\|\le B}\left(\sum_{i\in I_{kj}}\sigma^+_{ki}a^+_k\langle\mathbf{w}_k,\Phi(\mathbf{x}^+_{ki})\rangle\right)\right],$$

$$\spadesuit = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{\|\mathbf{w}_k\|\le B}\left(\sum_{i\in I_{kj}}\sigma^-_{ki}a^-_k\langle\mathbf{w}_k,\Phi(\mathbf{x}^-_{ki})\rangle\right)\right].$$

接下来，我们将分别对 ♣ 和 ♠ 进行（上界）界定。

首先，对于♣，我们可以有

$$\begin{aligned}
\clubsuit &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{\|\mathbf{w}_k\|\le B}\left(\sum_{i\in I_{kj}}\sigma^+_{ki}a^+_k\langle\mathbf{w}_k,\Phi(\mathbf{x}^+_{ki})\rangle\right)\right]\\
&\le \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{a^+_k}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sup_{\|\mathbf{w}_k\|\le B}\|\mathbf{w}_k\|\left\|\sum_{i\in I_{kj}}\sigma^+_{ki}\Phi(\mathbf{x}^+_{ki})\right\|\right] \qquad \text{(Cauchy–Schwarz inequality)}\\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{Ba^+_k}{m_k}\sum_{j\in[J_k]}\omega_{kj}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i\in I_{kj}}\sigma^+_{ki}\Phi(\mathbf{x}^+_{ki})\right\|\right] \qquad \text{(the definition of the sup and linearity of expectation)}\\
&\le \frac{1}{K}\sum_{k=1}^{K}\frac{Ba^+_k}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i\in I_{kj}}\sigma^+_{ki}\Phi(\mathbf{x}^+_{ki})\right\|^2\right]\right)^{\frac{1}{2}} \qquad \text{(Jensen's inequality)}\\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{Ba^+_k}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{p\in I_{kj},q\in I_{kj}}\sigma^+_{kp}\sigma^+_{kq}\langle\Phi(\mathbf{x}^+_{kp}),\Phi(\mathbf{x}^+_{kq})\rangle\right]\right)^{\frac{1}{2}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{K}\sum_{k=1}^{K}\frac{Ba_k^+}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\sum_{i\in I_{kj}}\langle\Phi(\mathbf{x}_{ki}^+),\Phi(\mathbf{x}_{ki}^+)\rangle\right)^{\frac{1}{2}} && (\forall p\neq q,\mathbb{E}[\sigma_{kp}\sigma_{kq}]=0 \text{ and } \mathbb{E}[\sigma_{ki}\sigma_{ki}]=1)\\
&\le \frac{1}{K}\sum_{k=1}^{K}\frac{Bra_k^+}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sqrt{m_{kj}} && (\langle\Phi(\mathbf{x}_{ki}^+),\Phi(\mathbf{x}_{ki}^+)\rangle=\kappa(\mathbf{x}_{ki}^+,\mathbf{x}_{ki}^+)\le r^2 \text{ and let } m_{kj}=|I_{kj}|)\\
&= \frac{1}{K}\sum_{k=1}^{K}\frac{Bra_k^+\chi_f(G_k)}{m_k}\sum_{j\in[J_k]}\frac{\omega_{kj}}{\chi_f(G_k)}\sqrt{m_{kj}}\\
&\le \frac{1}{K}\sum_{k=1}^{K}\frac{Bra_k^+\sqrt{\chi_f(G_k)}}{m_k}\sqrt{\sum_{j\in[J_k]}\omega_{kj}m_{kj}} && (\sum_{j\in[J_k]}\frac{\omega_{kj}}{\chi_f(G_k)}=1 \text{ and Jensen's inequality}).
\end{aligned}
$$

由于 $\sum_{j\in[J_k]}\omega_{kj}m_{kj}=m_k\chi_f(G_k)=\max\{|S_k^+|,|S_k^-|\}$ $m_k=|S_k^+||S_k^-|$和 $\tau_k=\min\{|S_k^+|,|S_k^-|\}$ 成立，因此得出】【。

$$
\clubsuit \le \frac{1}{K}\sum_{k=1}^{K}\frac{Bra_k^+\sqrt{\chi_f(G_k)}}{m_k}\sqrt{\sum_{j\in[J_k]}\omega_{kj}m_{kj}} \le \frac{Br}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}a_k^+\sqrt{\frac{1}{\tau_k}}\right).
$$

与 ♣ 的证明类似，我们可以得到 ♠ 的上界：

$$
\spadesuit \le \frac{Br}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}a_k^-\sqrt{\frac{1}{\tau_k}}\right).
$$

因此，我们可以得到最终结果：

$$
\widehat{\mathfrak{R}}^*_{\tilde{S},L_u}(\mathcal{F}) \le \frac{Br}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\left(a^+(S_k)+a^-(S_k)\right)\right).
$$

□

命题 1 (基于核的假设空间相对于 L_{u_1} 和 L_{u_2}的分数拉德马赫复杂度)。 对于代理损失函数 L_{u_1} 和 L_{u_2}，它们分别定义在公式(6) 和公式(7) 中，我们有

$$
\begin{aligned}
\widehat{\mathfrak{R}}^*_{\tilde{S},L_{u_1}}(\mathcal{F}) &\le \frac{Br}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right),\\
\widehat{\mathfrak{R}}^*_{\tilde{S},L_{u_2}}(\mathcal{F}) &\le \frac{2Br}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right).
\end{aligned}
$$

证明。 根据引理 6 代入代理单变量损失函数的具体重加权取值（即 $a^+(S_k)$ 和$a^-(S_k)$），证明显而易见。

□

B.4.2. 收缩不等式

引理 7（（重新加权的）单变量损失 L_u 的收缩不等式）。对于数据集 S （及其对应的˜构造数据集 S），假设对每个$k\in[K]$，损失函数 $L_\phi=L_u(\mathrm{x}^+,\mathrm{x}^-,f_k)=a^+(S_k)\ell(f_k(\mathrm{x}^+))+a^-(S_k)\ell(-f_k(\mathrm{x}^-))$，其中基础损失函数 $\ell(t)$ 是 ρ-Lipschitz 的，重新加权函数 $a^+(S_k)$（或 $a^-(S_k)$）表示其对 S_k 的依赖关系。则以下不等式成立

$$
\widehat{\mathfrak{R}}^*_{\tilde{S}}(L_\phi\circ\mathcal{F}) \le 2\rho\widehat{\mathfrak{R}}^*_{\tilde{S},L_u}(\mathcal{F}).
$$

证明。由于 $\widehat{\mathfrak{R}}^*_{\tilde{S}}(L_\phi\bigcirc \mathrm{F})=\frac{1}{K}\sum_{k=1}^{K}\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(L_\phi\bigcirc \mathrm{F}_k)$ 且 $\widehat{\mathfrak{R}}^*_{\tilde{S}}(\mathrm{F}\ L_u)=\frac{1}{K}\sum_{k=1}^{K}\widehat{\mathfrak{R}}^*_{\tilde{S},\ L_u}(\mathrm{F})$，我们先证明 $\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(L_\phi\bigcirc \mathrm{F}_k)\le 2\rho\widehat{\mathfrak{R}}^*_{\tilde{S},\ L_u}(\mathrm{F})$，然后即可得到所需结果。

下面我们证明不等式 $\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(L_\phi\circ\mathcal{F}_k)\le\rho\widehat{\mathfrak{R}}^*_{\tilde{S}_k,\ L_u}(\mathcal{F}_k)$遵循 Mohri 等人 (2018)（引理 5.7，第 93 页）中的思路，˜为简化后续记号，我们省略下标 k 以及符号 S_k 或 S。

首先固定一个样本（$\tilde{\mathbf{x}}_1 = (\mathbf{x}_1^+, \mathbf{x}_1^-) \ldots,\ \tilde{\mathbf{x}}_m = (\mathbf{x}_m^+, \mathbf{x}_m^-)$），根据定义，

$$\begin{aligned}\widehat{\mathfrak{R}}^*(L_\phi \circ f) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{j\in[J]} w_j \sup_{f\in\mathcal{F}}\left(\sum_{i\in I_j}\sigma_i L_u(\mathbf{x}_i^+, \mathbf{x}_i^-, f)\right)\right] \\ &= \frac{1}{m}\sum_{j\in[J]} w_j \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_{n_j-1}}\left[\mathop{\mathbb{E}}_{\sigma_{n_j}}\left[\sup_{f\in\mathcal{F}} u_{n_j-1}(f) + \sigma_{n_j} L_u(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f)\right]\right], \qquad \text{(denote } n_j = |I_j| \text{ for simplicity)}\end{aligned}$$

其中 $u_{n_j-1}(f) = \sum_{i=1}^{n_j}\sigma_i L_u(\mathbf{x}_i^+, \mathbf{x}_i^-, f)$。根据上确界的定义，对于任意 $\epsilon > 0$，存在 $f^1, f^2 \in \mathrm{F}$，使得

$$u_{n_j-1}(f^1) + L_{u_2}(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f^1) \le (1-\epsilon)\left[\sup_{f\in\mathcal{F}} u_{n_j-1}(f) + L_u(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f)\right],$$

and

$$u_{n_j-1}(f^2) - L_{u_2}(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f^2) \le (1-\epsilon)\left[\sup_{f\in\mathcal{F}} u_{n_j-1}(f) - L_u(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f)\right].$$

因此，对于任意 $\epsilon > 0$，根据 $\mathbb{E}$ 的定义$_{\sigma_{n_j}}$，

$$\begin{aligned}&(1-\epsilon)\mathop{\mathbb{E}}_{\sigma_{n_j}}\left[\sup_{f\in\mathcal{F}} u_{n_j-1}(f) + \sigma_{n_j} L_u(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f)\right] \\ &= (1-\epsilon)\left[\frac{1}{2}\sup_{f\in\mathcal{F}} u_{n_j-1}(f) + L_u(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f)\right] + \left[\frac{1}{2}\sup_{f\in\mathcal{F}} u_{n_j-1}(f) - L_u(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f)\right] \\ &\le \frac{1}{2}\left[u_{n_j-1}(f^1) + L_u(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f^1))\right] + \frac{1}{2}\left[u_{n_j-1}(f^2) - L_u(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f^2)\right]. \\ &= \frac{1}{2}\left[u_{n_j-1}(f^1) + a^+\ell(f^1(\mathbf{x}_{n_j}^+)) + a^-\ell(-f^1(\mathbf{x}_{n_j}^-))\right] + \frac{1}{2}\left[u_{n_j-1}(f^2) - a^+\ell(f^2(\mathbf{x}_{n_j}^+)) - a^-\ell(-f^2(\mathbf{x}_{n_j}^-))\right].\end{aligned}$$

令 $s^+ = \mathrm{sgn}(f^1(\mathbf{x}_{n_j}^+) - f^2(\mathbf{x}_{n_j}^+))$，且 $s^- = \mathrm{sgn}(f^1(\bar{\mathbf{x}}^{n_j}) - f^2(\bar{\mathbf{x}}^{n_j}))$。那么，前一个不等式意味着

$$\begin{aligned}&(1-\epsilon)\mathop{\mathbb{E}}_{\sigma_{n_j}}\left[\sup_{f\in\mathcal{F}} u_{n_j-1}(f) + \sigma_{n_j} L_u(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f)\right] \\ &\le \frac{1}{2}\left[u_{n_j-1}(f^1) + u_{n_j-1}(f^2) + s^+a^+\rho(f^1(\mathbf{x}_{n_j}^+) - f^2(\mathbf{x}_{n_j}^+)) + s^-a^-\rho(f^1(\mathbf{x}_{n_j}^-) - f^2(\mathbf{x}_{n_j}^-))\right] \qquad \text{(Lipschitz property)} \\ &= \frac{1}{2}\left[u_{n_j-1}(f^1) + s^+a^+\rho f^1(\mathbf{x}_{n_j}^+) + s^-a^-\rho f^1(\mathbf{x}_{n_j}^-)\right] + \frac{1}{2}\left[u_{n_j-1}(f^2) - s^+a^+\rho f^2(\mathbf{x}_{n_j}^+) - s^-a^-\rho f^2(\mathbf{x}_{n_j}^-)\right] \text{(rearranging)} \\ &\le \frac{1}{2}\sup_{f\in\mathcal{F}}\left[u_{n_j-1}(f) + s^+a^+\rho f(\mathbf{x}_{n_j}^+) + s^-a^-\rho f(\mathbf{x}_{n_j}^-)\right] + \frac{1}{2}\sup_{f\in\mathcal{F}}\left[u_{n_j-1}(f) - s^+a^+\rho f(\mathbf{x}_{n_j}^+) - s^-a^-\rho f(\mathbf{x}_{n_j}^-)\right] \text{(def of sup)} \\ &= 2\mathop{\mathbb{E}}_{\sigma_{n_j}^+,\sigma_{n_j}^-}\left[u_{n_j-1}(\tilde{f}) + \sigma_{n_j}^+a^+\rho f(\mathbf{x}_{n_j}^+) + \sigma_{n_j}^-a^-\rho f(\mathbf{x}_{n_j}^-)\right]. \qquad \text{(definition of } \mathop{\mathbb{E}}_{\sigma_{n_j}^+,\sigma_{n_j}^-}\text{)}\end{aligned}$$

由于该不等式对任意 $\epsilon > 0$ 都成立，因此有

$$\mathop{\mathbb{E}}_{\sigma_{n_j}}\left[\sup_{f\in\mathcal{F}} u_{n_j-1}(f) + \sigma_{n_j} L_u(\mathbf{x}_{n_j}^+, \mathbf{x}_{n_j}^-, f)\right] \le 2\mathop{\mathbb{E}}_{\sigma_{n_j}^+,\sigma_{n_j}^-}\left[u_{n_j-1}(\tilde{f}) + \sigma_{n_j}^+a^+\rho f(\mathbf{x}_{n_j}^+) + \sigma_{n_j}^-a^-\rho f(\mathbf{x}_{n_j}^-)\right].$$

对所有其他 $\sigma_i (i \in [I_j], i \ne n_j)$ 进行相同的操作，可证明

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{j\in[J]} w_j \sup_{f\in\mathcal{F}}\left(\sum_{i\in I_j}\sigma_i L_u(\mathbf{x}_i^+, \mathbf{x}_i^-, f)\right)\right] \le \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{j\in[J]} w_j \sup_{f\in\mathcal{F}} 2\rho\left(\sum_{i\in I_j}(\sigma_i^+a^+f(\mathbf{x}_i^+) + \sigma_i^-a^-f(\mathbf{x}_i^-))\right)\right].$$

通过对其他 $j \in [J]$ 进行操作，我们可以得到如下结果

$$\begin{aligned}\widehat{\mathfrak{R}}^*(L_\phi \circ f) &= \frac{1}{m}\sum_{j\in[J]} w_j \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i\in I_j}\sigma_i L_u(\mathbf{x}_i^+, \mathbf{x}_i^-, f)\right)\right] \\ &\le \frac{1}{m}\sum_{j\in[J]} w_j \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i\in I_j} 2\rho(\sigma_i^+ a^+ f(\mathbf{x}_i^+) + \sigma_i^- a^- f(\mathbf{x}_i^-))\right)\right] \\ &= 2\rho\widehat{\mathfrak{R}}^*_{L_u}(f).\end{aligned}$$

□

B.4.3. 定理 3 的证明

定理 3 （一般情况下的 $\mathcal{A}^{u_1}$ 学习保证）。 假设损失 $L_\phi = \frac{1}{\tau_S^*} L_{u_1}$，其中 L_{u_1} 定义见公式（6）。此外，假设 1 成立。那么，对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，在抽取大小为 S 的独立同分布样本 n 后，对于任意 $f \in \mathcal{F}$，以下泛化界成立：

$$\begin{aligned}R_{0/1}(f) \le &\frac{1}{\tau_S^*}\widehat{R}_{u_1}(f) + \frac{4\rho r\Lambda}{\tau_S^*\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^K\sqrt{\frac{1}{\tau_k}}\right) + \\ &\frac{3B}{\tau_S^*}\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\left(\sqrt{\frac{1}{K}\sum_{k=1}^K\frac{1}{\tau_k}}\right).\end{aligned} \tag{10}$$

证明。由于基础损失 $\ell(z)$ 被 B 所界定，且损失 $L_{u_1}(\mathbf{x}^+, \mathbf{x}^-, f_k) = \frac{|S_k^+|}{n}\ell(f_k(\mathbf{x}^+)) + \frac{|S_k^-|}{n}\ell(-f_k(\mathbf{x}^-))$，因此损失 $L_\phi = \frac{1}{\tau_S^*}L_{u_1}$ 被 $\frac{B}{\tau_S^*}$ 所界定。然后，应用定理 1，并结合引理 7 和命题 1，可得对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，在抽取大小为 n 的独立同分布样本 S 后，以下泛化界成立：

$$\mathop{\mathbb{E}}_S\left[\frac{1}{\tau_S^*}\widehat{R}_{u_1}(f)\right] \le \frac{1}{\tau_S^*}\widehat{R}_{u_1}(f) + \frac{4\rho r\Lambda}{\tau_S^*\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^K\sqrt{\frac{1}{\tau_k}}\right) + \frac{3B}{\tau_S^*}\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\left(\frac{1}{K}\sum_{k=1}^K\sqrt{\frac{1}{\tau_k}}\right).$$

最后，结合上述不等式与引理 2 （即 $R_{0/1}(f) \le \mathop{\mathbb{E}}_S\left[\frac{1}{\tau_S^*}\widehat{R}_{u_1}(f)\right]$），即可得到所需结果。 □

B.4.4. 定理 4 的证明

定理 4 （一般情况下 $\mathcal{A}^{u_2}$ 的学习保证）。 假设损失 $L_\phi = L_{u_2}$，其中 L_{u_2} 定义见式(7)。此外， 假设 1 成立。那么，对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，在抽取大小为 S 的独立同分布样本 n 时，对于任意 $f \in \mathcal{F}$，以下泛化界成立：

$$\begin{aligned}R_{0/1}(f) \le R_{u_2}(f) \le &\widehat{R}_{u_2}(f) + \frac{8\rho r\Lambda}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^K\sqrt{\frac{1}{\tau_k}}\right) + \\ &6B\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\left(\sqrt{\frac{1}{K}\sum_{k=1}^K\frac{1}{\tau_k}}\right).\end{aligned} \tag{11}$$

证明。由于基础损失 $\ell(z)$ 有界于 B，且损失 $L_{u_2}(\mathbf{x}^+, \mathbf{x}^-, f_k) = \ell(f_k(\mathbf{x}^+)) + \ell(-f_k(\mathbf{x}^-))$，损失 $L_\phi = L_{u_2}$ 有界于 $2B$。然后，应用定理 1，并结合引理 7 和命题 1，可得对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，在抽取大小为 n 的独立同分布样本 S 时，以下泛化界成立：

$$R_{u_2}(f) = \mathop{\mathbb{E}}_S\left[\widehat{R}_{u_2}(f)\right] \le \widehat{R}_{u_2}(f) + \frac{8\rho r\Lambda}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^K\sqrt{\frac{1}{\tau_k}}\right) + 6B\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}\left(\sqrt{\frac{1}{K}\sum_{k=1}^K\frac{1}{\tau_k}}\right)$$

最后，结合上述不等式和引理，我们可以得到所需结果 2(即， $R_{0/1}(f) \le R_{u_2}(f)$)。 □

B.4.5. 证明 推论 1

推论 1（平衡情形下 A^{u_1} 和 A^{u_2} 的学习保证）。假设损失 $L_\phi = 2L_{u_1} = L_{u_2}$，其中 L_{u_1} 和L_{u_2} 分别在公式 (6) 和公式 (7) 中定义。此外，假设 1 成立，并假设 S 是平衡的。那么，对于任意$\delta > 0$，以至少 $1-\delta$ 的概率，对抽取大小为 n 的独立同分布样本 S，以下泛化界对任意 $f \in \mathrm{F}$ 成立：

$$R_{0/1}(f) \leq R_{u_2}(f) = 2R_{u_1}(f) \leq \widehat{R}_{u_2}(f) + \frac{8\sqrt{2}\rho r\Lambda}{\sqrt{n}} + 6\sqrt{2}B\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}, \tag{12}$$

其中 $\widehat{R}_{u_2}(f) = 2\widehat{R}_{u_1}(f)$。

证明。 通过代入 3 定理和定理 4 中的 $\tau_k = \frac{1}{2}$，即可直接得到该结果。 □

B.4.6. 推论 3的证明

推论 3 (极度不平衡情形下 $\mathcal{A}^{u_1}$ 的学习保证)。 设损失 $L_\phi = nL_{u_1}$，其中 L_{u_1}定义见公式(6)。此外， 假设 1 成立，并假设 S 处于极度不平衡情形。于是，对于任意 $\delta > 0$，在从大小为 n的独立同分布样本 S 中抽取样本时，以至少 $1-\delta$ 的概率，以下泛化界对任意$f \in \mathcal{F}$成立：

$$R_{0/1}(f) \leq n\widehat{R}_{u_1}(f) + 4n\rho r\Lambda + 3Bn\sqrt{\frac{\log(\frac{2}{\delta})}{2}}.$$

证明。 将 3 代入 $\tau_k = \frac{1}{n}$，即可直接得到该结果。 □

B.4.7. 推论 5的证明

推论 5 (极度不平衡情形下 $\mathcal{A}^{u_2}$ 的学习保证)。 假设损失函数 $L_\phi = L_{u_2}$，其中 L_{u_2} 在公式(7)中定义。此外，假设 1 成立，并且假设 S 极度不平衡。那么，对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，对于抽取的大小为 n 的独立同分布样本 S，以下泛化界对任意$\mathcal{F}$成立：

$$R_{0/1}(f) \leq R_{u_2}(f) \leq \widehat{R}_{u_2}(f) + 8\rho r\Lambda + 6B\sqrt{\frac{\log(\frac{2}{\delta})}{2}}.$$

证明。直接 即可通过应用定理 4 和定理 4，代入 $\tau_k = \frac{1}{n}$后得到该结果。 □

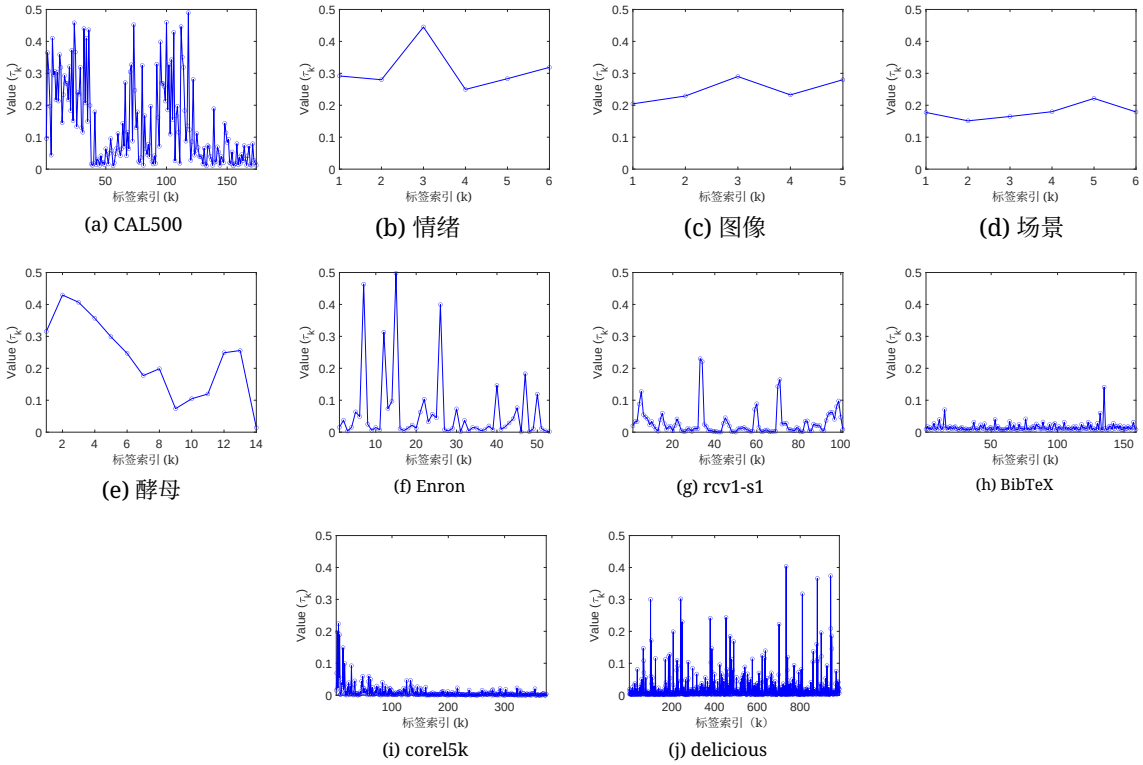
C. 补充实验结果

C.1. 基准数据集的标签级类别不平衡示意图

所有基准数据集的标签级类别不平衡程度如图 2所示 .

C.2. 界的绝对值的实验结果

此处我们报告表格中基准数据集上三种算法的平均上界值 4。从实验结果中，我们可以观察到，这些绝对值可能无法反映真实的泛化风险（即大于 1），但在相同框架下比较相关因素的排序，仍能为了解这些算法提供有价值的见解。



(a) CAL500 (b) 情绪 (c) 图像 (d) 场景

(e) 酵母 (f) Enron (g) rcv1-s1 (h) BibTeX

(i) corel5k (j) delicious

图2。每个基准数据集的标签级类别不平衡示意图。

表4。三种算法在基准数据集上的平均上界值。我们设定 $\delta = 0.01$。

数据集	$\mathcal{A}^{pa}$	$\mathcal{A}^{u_1}$	$\mathcal{A}^{u_2}$
CAL500	13.0	7082.4	36.3
情绪	13.6	56.8	33.7
图像	13.1	191.0	30.7
场景	7.4	146.2	19.6
酵母	3.6	375.2	12.1
Enron	-	37016.0	69.0
rcv1-s1	-	96284.0	20.2
BibTeX	-	207.4	55.2
corel5k	-	97359.0	3.1
delicious	-	293000.0	82.8

D. 其他相关工作

最近，在相当一般的设置下，针对具有图依赖的数据的 McDiarmid 型集中不等式，已有一些研究工作，例如：Zhang 等人 (2019)。然而，据我们所知，将现有工作的结论（Zhang 等人，2019）（即第 5 页针对一般依赖图的定理 3.6）应用于获得多标签分类的 Macro-AUC 最大化中的显式泛化界，目前并不明确，因为在这种情况下，依赖图的森林复杂度可能难以估计。相比之下，我们提出的新 McDiarmid 型集中不等式（即定理5）在这种情况下易于应用。

请注意，我们提出的集中不等式（即，定理 5）并非现有结果的推论（即，(Zhang 等人，2019) 中的定理 3.6）。它们在不同假设下互为补充。尽管 Zhang 等人(2019) 的结果适用于一般依赖图，但我们考虑的是附加假设下的特殊情形，即对函数和依赖图施加约束（即定理 5中的假设 (1) 和 (2)），而这无法由 Zhang 等人

（2019）。