Towards Understanding Generalization of Macro-AUC in Multi-label Learning

Guoqiang Wu [1] **Chongxuan Li** [2] **Yilong Yin** [1]

Abstract

Macro-AUC is the arithmetic mean of the class-wise AUCs in multi-label learning and is commonly used in practice. However, its theoretical understanding is far lacking. Toward solving it, we characterize the generalization properties of various learning algorithms based on the corresponding surrogate losses w.r.t. Macro-AUC. We theoretically identify a critical factor of the dataset affecting the generalization bounds: *the label-wise class imbalance*. Our results on the imbalance-aware error bounds show that the widely-used univariate loss-based algorithm is more sensitive to the label-wise class imbalance than the proposed pairwise and reweighted loss-based ones, which probably implies its worse performance. Moreover, empirical results on various datasets corroborate our theory findings. To establish it, technically, we propose a new (and more general) McDiarmid-type concentration inequality, which may be of independent interest.

1. Introduction

Multi-Label Learning (MLC) (McCallum, 1999) is an important learning task in machine learning where each instance might be associated with multiple labels. It has been widely applied in various areas, e.g., natural language processing (Schapire & Singer, 2000), computer vision (Carneiro et al., 2007), and bioinformatics (Elisseeff & Weston, 2001). Due to the complexity of MLC and the diverse demands of different scenarios, various measures (Zhang & Zhou, 2014; Wu & Zhou, 2017) have been developed for a comprehensive evaluation, e.g., Hamming loss, ranking loss, and subset accuracy. Among them, Macro-AUC (Zhang & Zhou, 2014) is a widely-used measure in practice. Informally, it is the arithmetic mean of the class-wise AUC measures, which is the focus of this paper.

Macro-AUC (and many other measures) in MLC are discontinuous and non-convex, which makes that optimizing them directly can lead to NP-hard problems (Arora & Barak, 2009). Thus, many surrogate losses are used in practice for computational efficiency. Empirically, many surrogate loss-based learning algorithms are commonly evaluated in terms of Macro-AUC, including the widely-used surrogate univariate loss-based algorithms (Boutell et al., 2004; Wu & Zhu, 2020) that originally aim to optimize the Hamming loss. Theoretically, however, the understanding is far lacking. To take a step towards solving it, this paper attempts to formally answer the following question:

> *What is the learning guarantee of the widely-used surrogate univariate loss-based algorithms w.r.t. the Macro-AUC?*

To answer the above question, we propose an analytical framework to characterize the generalization properties of learning algorithms w.r.t. the Macro-AUC. Inspired by the theory analyses, we also propose one pairwise surrogate loss and one reweighted univariate loss for Macro-AUC. Theoretically, we analyze the learning guarantees of algorithms with all three losses. We theoretically identify the *label-wise class imbalance*, which is a factor of the dataset in MLC (Tarekegn et al., 2021; Zhang et al., 2020b), plays a critical role in these generalization bounds.

Specifically, the pairwise loss-based learning algorithm $\mathcal{A}^{pa}$ has a label-wise class imbalance-aware leaning guarantee of $O\left(\frac{1}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)\right)$ (see Table 1), where n is the sample size, K is the label size, and $\tau_k \in [\frac{1}{n}, \frac{1}{2}]$ characterizes the k-th label class imbalance level. The smaller τ_k, the higher the imbalance level. In contrast, the widely-used univariate loss-based algorithm $\mathcal{A}^{u_1}$ has an error bound of $O\left(\frac{1}{\tau_S^*\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)\right)$, where $\tau_S^* = \operatorname{argmin}_{k\in[K]}\tau_k$. Thus, we can observe that $\mathcal{A}^{u_1}$ is more sensitive to the label-wise class imbalance than $\mathcal{A}^{pa}$, which implies that $\mathcal{A}^{pa}$ would probably perform better than $\mathcal{A}^{u_1}$ practically, especially when $\frac{1}{\tau_S^*}$ is large, which often occurs in real datasets of MLC. Note that, computationally, $\mathcal{A}^{pa}$ can lead to a complexity of $O(n^2K)$, which is worse than $\mathcal{A}^{u_1}$ (i.e., $O(nK)$), and it can be prohibitively costly when the sample size n

[1]School of Software, Shandong University [2]Gaoling School of AI, Renmin University of China; Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing, China. Correspondence to: Guoqiang Wu <guoqiangwu@sdu.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Table 1. Summary of the main theoretical results. The contributions of this paper are highlighted in red.

Algorithm	Surrogate loss	Generalization bound	Computation
$\mathcal{A}^{pa}$	pairwise (L_{pa})	$\widehat{R}_S^{pa}(f) + O\left(\frac{1}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^K \sqrt{\frac{1}{\tau_k}}\right)\right)$	$O(n^2K)$
$\mathcal{A}^{u_1}$ (Boutell et al., 2004)	univariate (L_{u_1})	$\frac{1}{\tau_S^*}\widehat{R}_S^{u_1}(f) + O\left(\frac{1}{\tau_S^*\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^K \sqrt{\frac{1}{\tau_k}}\right)\right)$	$O(nK)$
$\mathcal{A}^{u_2}$	reweighted univariate (L_{u_2})	$\widehat{R}_S^{u_2}(f) + O\left(\frac{1}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^K \sqrt{\frac{1}{\tau_k}}\right)\right)$	$O(nK)$

is large. Interestingly, our proposed reweighted univariate loss-based algorithm $\mathcal{A}^{u_2}$ has a generalization bound of $O\left(\frac{1}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^K \sqrt{\frac{1}{\tau_k}}\right)\right)$, which is nearly the same as $\mathcal{A}^{pa}$. This probably implies the performance superiority of $\mathcal{A}^{u_2}$ over $\mathcal{A}^{u_1}$, as well as the computational efficiency. Finally, empirical results corroborate our theory findings.

Technically, since optimizing Macro-AUC potentially involves learning with dependent examples, the existing generalization analytical techniques (Wu & Zhu, 2020; Wu et al., 2021) for other measures in MLC cannot be applied, making it more challenging. Following the technique in Bipartite Ranking (BR, or equivalently AUC maximization in binary classification) (Usunier et al., 2005; Amini & Usunier, 2015), we extend it to Macro-AUC maximization in MLC. Note that the technique in BR cannot be trivially applied in MLC due to the multiple labels (or tasks) property of MLC.[1] Thus, we propose general techniques that include a new McDiarmid-type concentration inequality and a general generalization bound of learning multiple tasks with graph-dependent examples, which may be of independent interest. (See Appendix A for details). Our generalization analyses on the Macro-AUC maximization in MLC can be viewed as an application of these general techniques.

2. Preliminaries

Notations. Let boldfaced lower and upper letters denote the vector (e.g., $\mathbf{a}$) and matrix (e.g., $\mathbf{A}$), respectively. For a matrix $\mathbf{A}$, $\mathbf{a}_i$, $\mathbf{a}^j$, and a_{ij} denote its i-th row, j-th column and (i,j)-th element, respectively. For a vector $\mathbf{a}$, a_i denotes its i-th element. Let $[K]$ denote the set $\{1,\dots,K\}$. For a set, $|\cdot|$ denotes its cardinality. $[\![\cdot]\!]$ denotes the indicator function, i.e., it returns 1 if the proposition holds and 0 otherwise.

2.1. Problem Setting

Let $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ and $\mathbf{y} \in \mathcal{Y} \subset \{-1,+1\}^K$ denote the input and output respectively, where d is the feature dimension, and K is the label size. $y_k = 1$ (or -1) indicates that the associated k-th label is relevant (or irrelevant). Given a training set $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ of n i.i.d. samples drawn from a distribution P over $\mathcal{X} \times \mathcal{Y}$, the original goal of MLC is to learn a multi-label classifier $H : \mathbb{R}^d \to \{-1,+1\}^K$.

To solve MLC, a standard approach is first to learn a vector-based *score function* (or predictor) $f = [f_1,\dots,f_K] : \mathcal{X} \to \mathbb{R}^K$ from a hypothesis space $\mathcal{F}$ and then get the classifier H by a thresholding function. A typical goal in MLC is to learn the best predictor from the finite training data in terms of some ranking-based measure, which is usually called Multi-label Ranking (Dembczynski et al., 2012; Wu et al., 2021), and this is our focus in this paper.

2.2. Evaluation Measure

Many evaluation measures have been developed to evaluate the performance of different algorithms. Here we focus on the common measure Macro-AUC, which *macro-averages* the AUC measure across all class labels. Given a dataset S and a predictor $f \in \mathcal{F}$, Macro-AUC is defined as follows:[2]

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

where S_k^+ (or S_k^-) denotes the relevant (or irrelevant) instance index set for the label k.

Maximizing Macro-AUC is equivalent to minimizing the following objective (i.e., one minus Macro-AUC):

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

where the $0/1$ loss function is defined as

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

in which $\mathbf{x}^+$ (or $\mathbf{x}^-$) denotes a relevant (or irrelevant) input for the label k.

[1] Note that one may use the union bound to combine the original bounds in BR to get the desired bound w.r.t. Macro-AUC in MLC, which would lead to a loosely bound involving a term $\log(\frac{K}{\delta})$.

[2] Note that here we do not adopt another common form w.r.t. the equality (i.e., $[\![f_k(\mathbf{x}_p) \ge f_k(\mathbf{x}_q)]\!]$), in order to avoid the trial zero hypothesis f. Besides, these two forms are nearly the same practically in evaluating algorithms.

2.3. Risk

Since Macro-AUC (or the $0/1$ loss function) is discontinuous and non-convex, optimizing it directly would lead to NP-hard problem (Arora & Barak, 2009). Practically, one often seeks (convex) surrogate losses for computational efficiency. Let $L_\phi : \mathcal{X} \times \mathcal{X} \times \mathcal{F}_k \to \mathbb{R}_+$ denote a surrogate loss function where $\mathcal{F}_k = \{f_k : \mathcal{X} \to \mathbb{R}\}$ and we will discuss its specific form in the next section. For a predictor $f \in \mathcal{F}$, the true $(0/1)$ generalization (or expected) risk w.r.t. Macro-AUC is defined as

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

where the conditional distribution $P_k^+ = P(\mathbf{x}|y_k = 1)$ and $P_k^- = P(\mathbf{x}|y_k = -1)$. Besides, the surrogate empirical and generalization risks w.r.t. L_ϕ are defined as follows, respectively:

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

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

Note that we do not define the surrogate generalization risk as the following common form

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

This is because that Eq.(3) is more general than Eq.(4) where Eq.(3) can cover the surrogate loss L_ϕ depending on the training dataset S while Eq.(4) cannot.[3] Besides, they are equal for certain losses independent of S.

3. Methods

In this section, we present the considered surrogate losses and their corresponding learning algorithms.

3.1. Surrogate Losses

To optimize Macro-AUC, it is natural to use the following (surrogate) pairwise loss:

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

where the base loss function $\ell(t)$ could be many popular (margin-based) loss functions, e.g., the hinge loss $\ell(t) = \max(0, 1-t)$, the logistic function $\ell(t) = \log_2(1+\exp(-t))$, and so on. A natural property of the base loss is that it is an upper bound of the original $0/1$ loss, i.e., $\ell(t) \geq [\![t \leq 0]\!]$. Note that minimizing this pairwise loss-based risk leads to a computational complexity of $O(n^2)$, which could be prohibitively costly when the sample size n is large.[4]

The widely-used univariate loss that originally aims to optimize the Hamming Loss measure (Boutell et al., 2004; Wu & Zhu, 2020), could be also viewed as a surrogate loss L_{u_1} for Macro-AUC. Its original empirical risk can be written as:

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

Thus, we can define this surrogate univariate loss L_{u_1} w.r.t. Macro-AUC as follows:

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

Note that this surrogate loss cannot strictly upper bound the $0/1$ loss, i.e., $L_{0/1} \nleq L_{u_1}$. The upper bound property of the surrogate loss w.r.t. the $0/1$ loss is critical to provide its generalization analysis w.r.t. the $0/1$ loss and we discuss it in detail later. Besides, note that we cannot define the generalization risk w.r.t. L_{u_1} by Eq.(4) due to its dependency on the dataset S.

To upper bound the $0/1$ loss (i.e., $L_{0/1}$), here we propose a new reweighted univariate surrogate loss L_{u_2} with computational efficiency, which is defined as below:

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

Then, we can write its empirical risk as

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

We can see that, computationally, minimizing the empirical risk w.r.t. L_{u_2} could lead to a complexity of $O(n)$, which is the same as L_{u_1}. Intuitively, this reweighted loss could be seen as a cost-sensitive loss that optimizes for balanced accuracy.

There are some relationships between these losses and we will discuss them thoroughly in the subsequent section.

[3] Note that, the aim we define Eq.(3) is for the convenience of analyses for L_{u_1} and finally to get the bounds where terms only depend on the dataset. If we define the common form, we will eventually get the bounds involving the term depending on the distribution.

[4] Note that there are possible ways to accelerate it for certain base losses.

3.2. Learning Algorithm

In the subsequent analyses, we focus on the kernel-based learning algorithms, which have been widely used in practice (Elisseeff & Weston, 2001; Boutell et al., 2004; Hariharan et al., 2010; Tan et al., 2020; Wu et al., 2020a) and in theory (Wu & Zhu, 2020; Wu et al., 2021) in MLC. Note that our subsequent analyses can be extended to other forms of the hypothesis space, e.g., neural networks (Anthony & Bartlett, 1999). Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Positive Definite Symmetric (PDS) kernel and denote its induced reproducing kernel Hilbert space (RKHS) as $\mathbb{H}$. Let $\Phi : \mathcal{X} \to \mathbb{H}$ be a feature mapping associated with κ. The considered kernel-based hypothesis class can be defined as

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

where $\|\mathbf{w}_k\|$ denotes $\|\mathbf{w}_k\|_{\mathbb{H}}$ for convenience.

Here we consider the following three regularized learning algorithms with the aforementioned corresponding surrogate losses:

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

where λ denotes a trade-off hyper-parameter and $\|\mathbf{W}\|$ denotes $\|\mathbf{W}\|_{\mathbb{H},2} = (\sum_{j=1}^c \|\mathbf{w}_j\|_{\mathbb{H}}^2)^{1/2}$ for convenience.

4. Theoretical Results

In this section, we mainly introduce the generalization results of the aforementioned learning algorithms with different surrogates w.r.t. the Macro-AUC measure, where the proofs of related lemmas, theorems, and corollaries are in Appendix B.

Technically, to establish it, we propose new techniques including a new McDiarmid-type inequality. (Please see Section 4.5 for the proof sketch and Appendix A for details).

Firstly, we give the following definition to characterize the label-wise class imbalance in MLC.

Definition 1 (**Label-wise class imbalance**). *Given a dataset S, define the following factor to characterize the label-wise class imbalance level for each label $k \in [K]$:*

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

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

From the above definition, we can see, the smaller τ_k, the higher the label-wise class imbalance level. Besides, for the convenience of following discussions, we give the following definition.[5]

Definition 2 (**Label-wise class balanced and extremely imbalanced dataset**). *Given a dataset S, we say that it is label-wise class balanced (or extremely class imbalanced) if $\forall k \in [K], \tau_k = \frac{1}{2}$ (or $\tau_k = \frac{1}{n}$) holds.*[6]

Then, we introduce the common mild assumptions for the subsequent analyses.

Assumption 1 (**The common assumptions**).

(1) The training dataset $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ is an i.i.d. sample drawn from the distribution P, where $\exists\, r > 0$, it satisfies $\kappa(\mathbf{x}, \mathbf{x}) \le r^2$ for all $\mathbf{x} \in \mathcal{X}$.

(2) The hypothesis class is defined in Eq.(8).

(3) The base (convex) loss $\ell(z)$ is ρ-Lipschitz continuous and bounded by B.[7]

Here we give the definition of the fractional Rademacher complexity of the loss space.

Definition 3 (**The fractional Rademacher complexity of the loss space**). *For each label $k \in [K]$, construct the dataset $\tilde{S}_k = \{(\tilde{\mathbf{x}}_{ki}, \tilde{y}_{ki})\}_{i=1}^{m_k} = \{((\tilde{\mathbf{x}}_{ki}^+, \tilde{\mathbf{x}}_{ki}^-), 1)\}_{i=1}^{m_k}$ based on the original dataset S_k, where $(\tilde{\mathbf{x}}_{ki}^+, \tilde{\mathbf{x}}_{ki}^-) \in S_k^+ \times S_k^-$, and let $\{(I_{kj}, \omega_{kj})\}_{j \in [J_k]}$ be a fractional independent vertex cover of the dependence graph G_k constructed over $\tilde{S}_k$ with $\sum_{j \in [J_k]} \omega_{kj} = \chi_f(G_k)$, where $\chi_f(G_k)$ is the fractional chromatic number of G_k. For the hypothesis space $\mathcal{F}$ and loss function $L : \mathcal{X} \times \mathcal{X} \times \mathcal{F}_k \to \mathbb{R}_+$, the empirical fractional Rademacher complexity of the loss space is defined as*

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

Then, we give the base theorem of Macro-AUC used in the subsequent generalization analyses.

Theorem 1 (**The base theorem of Macro-AUC**). *Assume the loss function $L_\phi : \mathcal{X} \times \mathcal{X} \times \mathcal{F}_k \to \mathbb{R}_+$ is bounded by M. Then, for any $\delta > 0$, the following generalization bound holds with probability at least $1 - \delta$ over the draw of an i.i.d.*

[5] Note that multi-label datasets can also be imbalanced in an inter-label way, i.e. some labels having very few positives, and other labels having many.

[6] In this paper we call it balanced or extremely imbalanced for simplicity.

[7] Note that, the widely-used hinge and logistic loss are both 1-Lipschitz continuous. Although the exponential and squared hinge losses are not globally Lipschitz continuous, they are locally Lipschitz continuous.

sample S of size n:

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

Then, we analyze the relationship between the surrogate and true losses as follows.

Lemma 1 (**The relationship between the surrogate and true losses**). *Assume the base loss function upper bounds the original $0/1$ loss, i.e., $\ell(t) \ge [\![t \le 0]\!]$. Then, for any $f_k \in \mathcal{F}_k$ and $(\mathbf{x}^+, \mathbf{x}^-) \in S_k^+ \times S_k^-$, the following inequalities hold:*

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

Remark. *From this lemma, we can observe that when minimizing L_{u_1}, it also minimizes an upper bound of $L_{0/1}$ depending on $\frac{1}{\tau_k}$. Besides, for the second inequality involving L_{u_1} and L_{u_2}, the bound is tight since the equality holds when $\tau_k = \frac{1}{2}$.*

Based on Lemma 1, we can get the relationship between the surrogate and true risks as follows, which is critical for the generalization analyses.

Lemma 2 (**The relationship between the surrogate and true risks**). *Assume the base loss function upper bounds the original $0/1$ loss, i.e., $\ell(t) \ge [\![t \le 0]\!]$. Then, for any $f \in \mathcal{F}$ and any sample $S \overset{i.i.d.}{\sim} P$, the following inequalities hold:*

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

Remark. *For the second inequality involving the generalization risk w.r.t. L_{u_1} and L_{u_2}, the bound is tight since the equality holds when $\tau_k = \frac{1}{2}$.*

Next, for clear discussions, we introduce the generalization results of algorithms w.r.t. the label-wise class imbalance: general, balanced, and extremely imbalanced cases.

4.1. General Case

Here we introduce the generalization results of algorithms w.r.t. general datasets which cover the subsequent balanced and extremely imbalanced datasets.

Theorem 2 (**Learning guarantee of $\mathcal{A}^{pa}$ in general case**). *Assume the loss $L_\phi = L_{pa}$, where L_{pa} is defined in Eq.(5). Besides,* Assumption 1 *holds. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

From this theorem, we can observe that $\mathcal{A}^{pa}$ has a label-wise class imbalance-aware learning guarantee of $O\left(\frac{1}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)\right) \approx O\left(\frac{1}{\sqrt{n}}\left(\sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k}}\right)\right)$ w.r.t. Macro-AUC.

Theorem 3 (**Learning guarantee of $\mathcal{A}^{u_1}$ in general case**). *Assume the loss $L_\phi = \frac{1}{\tau_S^*}L_{u_1}$, where L_{u_1} is defined in Eq.(6). Besides,* Assumption 1 *holds. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

From this theorem, we can see that $\mathcal{A}^{u_1}$ has an imbalance-aware learning guarantee of $O\left(\frac{1}{\tau_S^*\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)\right)$ w.r.t. Macro-AUC.

Theorem 4 (**Learning guarantee of $\mathcal{A}^{u_2}$ in general case**). *Assume the loss $L_\phi = L_{u_2}$, where L_{u_2} is defined in Eq.(7). Besides,* Assumption 1 *holds. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

From this theorem, we can see that $\mathcal{A}^{u_2}$ has an imbalance-aware learning guarantee of $O\left(\frac{1}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)\right)$ w.r.t. Macro-AUC, which is nearly the same as $\mathcal{A}^{pa}$.

4.2. Balanced Case

Here we consider the balanced dataset. Note that in this case, algorithms $\mathcal{A}^{u_1}$ and $\mathcal{A}^{u_2}$ are exactly the same, which should share the same learning guarantee and it is confirmed by the following corollary.

Corollary 1 (Learning guarantee of $\mathcal{A}^{u_1}$ and $\mathcal{A}^{u_2}$ in balanced case). *Assume the loss $L_\phi = 2L_{u_1} = L_{u_2}$, where L_{u_1} and L_{u_2} are defined in Eq.(6) and Eq.(7), respectively. Besides,* Assumption 1 *holds and suppose S is balanced. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

where $\widehat{R}_{u_2}(f) = 2\widehat{R}_{u_1}(f)$.

Note that in this case, the same error bound of $\mathcal{A}^{u_1}$ and $\mathcal{A}^{u_2}$ confirms the validity of our analyses. From this corollary, we can see $\mathcal{A}^{u_1}$ (or $\mathcal{A}^{u_2}$) has an error bound of $O(\sqrt{\frac{1}{n}})$, which is the same as $\mathcal{A}^{pa}$ (see Corollary 4 in Appendix B.3.4).

4.3. Extremely Imbalanced Case

Here we consider the extremely imbalanced datasets. In this case, the generalization results are as follows.

Corollary 2 (Learning guarantee of $\mathcal{A}^{pa}$ in extremely imbalanced case). *Assume the loss $L_\phi = L_{pa}$, where L_{pa} is defined in Eq.(5). Besides,* Assumption 1 *holds and suppose S is extremely imbalanced. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

Corollary 3 (Learning guarantee of $\mathcal{A}^{u_1}$ in extremely imbalanced case). *Assume the loss $L_\phi = nL_{u_1}$, where L_{u_1} is defined in Eq.(6). Besides,* Assumption 1 *holds and suppose S is extremely imbalanced. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

From the above corollaries, we can see that $\mathcal{A}^{pa}$ has an error bound of $O(1)$ w.r.t. n, while $\mathcal{A}^{u_1}$ depends on $O(n)$. Besides, $\mathcal{A}^{u_2}$ has a similar error bound to $\mathcal{A}^{pa}$ (see Corollary 5 in Appendix B.4.7). One may notice that these bounds all diverge when $n \to \infty$. This may be due to the following two reasons. On one hand, learning in the extremely imbalanced case is indeed difficult. On the other hand, our analysis techniques might not be optimal w.r.t. n, and advanced techniques (e.g., local Rademacher-type complexity (Bartlett et al., 2005)) might be used to improve it. However, this is not our focus in this paper and we mainly focus on the generalization effect of label-wise class imbalance factors under the same framework, and the orders of different algorithms can still provide valuable insights.

4.4. Comparison and Discussion

For generalization analyses, a tighter upper bound usually implies probably better performance (Mohri et al., 2018).[8] In this paper, all algorithms are analyzed under the same framework and inequalities between the surrogate and true risks (or losses) are tight. Therefore, it is relatively safe to evaluate the performance of the algorithms theoretically by comparing their upper bounds. We now compare these algorithms as follows.

- **$\mathcal{A}^{pa}$ vs $\mathcal{A}^{u_1}$.** $\mathcal{A}^{pa}$ usually has a tighter bound than $\mathcal{A}^{u_1}$. Specifically, given the same hypothesis space, it is usually easier to train $\widehat{R}_S^{pa}$ than other risks, making $\widehat{R}_S^{pa}$ smaller than $\frac{1}{\tau_S^*}\widehat{R}_S^{u_1}$.[9] Besides, for the model complexity terms (i.e., the last two terms), $\mathcal{A}^{pa}$ has an error bound of $O\left(\frac{1}{K}\sum_{k=1}^K\sqrt{\frac{1}{\tau_k}}\right)$ while $\mathcal{A}^{u_1}$ depends on $O\left(\frac{1}{\tau_S^*}\left(\frac{1}{K}\sum_{k=1}^K\sqrt{\frac{1}{\tau_k}}\right)\right)$.
- **$\mathcal{A}^{u_2}$ vs $\mathcal{A}^{u_1}$.** Similarly, we argue that $\mathcal{A}^{u_2}$ usually has a tighter bound than $\mathcal{A}^{u_1}$. For the first risk term, $\frac{1}{\tau_S^*}\widehat{R}_S^{u_1}$ is usually comparable or even larger than $\widehat{R}_S^{u_2}$.[10] For the model complexity term, $\mathcal{A}^{u_2}$ has an error bound of $O\left(\frac{1}{K}\sum_{k=1}^K\sqrt{\frac{1}{\tau_k}}\right)$ while $\mathcal{A}^{u_1}$ depends on $O\left(\frac{1}{\tau_S^*}\left(\frac{1}{K}\sum_{k=1}^K\sqrt{\frac{1}{\tau_k}}\right)\right)$.
- **$\mathcal{A}^{pa}$ vs $\mathcal{A}^{u_2}$.** $\mathcal{A}^{pa}$ and $\mathcal{A}^{u_2}$ have similar or comparable learning guarantees. Specifically, for the first risk term, $\widehat{R}_S^{pa}$ is usually comparable to $\widehat{R}_S^{u_2}$.[11] For the model term, $\mathcal{A}^{pa}$ and $\mathcal{A}^{u_2}$ are nearly the same (with only different constant) w.r.t. the label-wise class imbalance.

[8] Note that when comparing bounds, it is usually more reasonable to compare the order of dependent variables rather than the absolute values.

[9] Although we can not formally express this claim, we empirically observed it in experiments.

[10] In some cases, the first risk term may be bigger than 1 but we can still take insights from the error bound through the dependent variables of the model complexity.

[11] Note that although we cannot formally express this, the experiments verify it.

Overall, the tighter bound level of $\mathcal{A}^{pa}$ (and $\mathcal{A}^{u_2}$) over $\mathcal{A}^{u_1}$ heavily depend on $\frac{1}{\tau_S^*}$. Thus, when $\frac{1}{\tau_S^*}$ is large, $\mathcal{A}^{pa}$ and $\mathcal{A}^{u_2}$ would probably perform significantly better than $\mathcal{A}^{u_1}$. In contrast, when $\frac{1}{\tau_S^*}$ is small, $\mathcal{A}^{pa}$ and $\mathcal{A}^{u_2}$ would probably perform slightly better than or nearly comparably to $\mathcal{A}^{u_1}$. Besides, $\mathcal{A}^{pa}$ and $\mathcal{A}^{u_2}$ have nearly the same learning guarantee, thus they would probably perform comparably. Experimental results on benchmark datasets in Table 3 confirm our analyses.

Note that there may be another way to analyze the learning guarantees of $\mathcal{A}^{u_1}$ and $\mathcal{A}^{u_2}$ w.r.t. the Macro-AUC under the analytical framework of prior work (Wu & Zhu, 2020; Wu et al., 2021) and here we focus on analyzing three algorithms under the same framework, leaving it as future work.

Implications of the theory in real-world applications. While the MLC datasets are usually highly (label-wise) class imbalanced in real-world applications, our theory can have valuable implications in practice. Specifically, our theoretical results on the imbalance-aware bounds show that the imbalance-aware loss-based algorithm $\mathcal{A}^{u_2}$ has a better learning guarantee w.r.t. the label-wise class imbalance than the algorithm $\mathcal{A}^{u_1}$ with the original univariate loss (e.g., cross-entropy loss), which probably implies its performance superiority. This can provide valuable insights to explain why the existing imbalance-aware reweighting losses (Ridnik et al., 2021; Wu et al., 2020b) can have promising performance w.r.t. ranking-based measures (e.g., mean average precision (mAP)) similar to Macro-AUC in practice. Further, how to design more effective imbalance-aware loss w.r.t. specific measures and how to make these bounds tighter would inspire more effective algorithms.

4.5. Proof Sketch

Here we mainly summarize the proof sketch of the generalization results of the general case in Section 4.1 as follows.

Proof sketch: Overall, the proof can be mainly divided into the following two steps.

Step 1: Construct general techniques in need (see Appendix A for details). First, we propose a new (and more general) McDiarmid-type inequality (i.e., Theorem 5). Then, based on it, we propose a general generalization bound (i.e., Theorem 6) for the problem setting of learning multiple tasks with graph-dependent examples, which involves the fractional Rademacher complexity of the loss space.

Step 2: Get the results of the Macro-AUC maximization (MaAUCM) in MLC by applying the generalization bound obtained in Step 1 (see Appendix B for details). Firstly we transform the MaAUCM problem into the problem setting of learning multiple tasks with graph-dependent examples in Step 1 and then we get the base error bound of Macro-AUC (i.e., Theorem 1). Next, we analyze the relationships between surrogate and true risks. Then, for each algorithm, define its specific fractional Rademacher complexity of the hypothesis space, upper bound the kernel-based one (e.g., Lemma 4), and get the specific contraction inequality (e.g., Lemma 5) to connect the complexity of the loss space with the complexity of the hypothesis space. Finally, we can get the desired results based on the above intermediate results.

4.6. Consistency of Surrogate Losses

Except for the finite-sample generalization guarantee, the consistency of surrogates is also important. Following (Gao & Zhou, 2015; Kotlowski et al., 2011), here we consider the consistency of L_{pa}, L_{u_1} and L_{u_2} w.r.t. the Macro-AUC with $L_{0/1}(\mathbf{x}^+, \mathbf{x}^-, f_k) = [\![f_k(\mathbf{x}^+) < f_k(\mathbf{x}^-)]\!] + \frac{1}{2}[\![f_k(\mathbf{x}^+) = f_k(\mathbf{x}^-)]\!]$. The Macro-AUC maximization task of MLC can be decomposed into K AUC maximization tasks of binary classification. Thus, we can investigate the consistency of these surrogates based on the previous well-studied consistency results of AUC in binary classification (or, equivalently bipartite ranking) (Gao & Zhou, 2015; Kotlowski et al., 2011). Specifically, we can get the following results about these surrogates:

- L_{pa}: Based on the previous result in binary classification with AUC maximization (Gao & Zhou, 2015) (i.e., Corollary 1 on Page 4), we can get that L_{pa} is consistent w.r.t. Macro-AUC with the (base) logistic loss and exponential loss. Besides, based on the previous result (Gao & Zhou, 2015) (i.e., Lemma 3 on Page 3), we can get that L_{pa} is inconsistent w.r.t. Macro-AUC with the (base) hinge loss and absolute loss.

- L_{u_1}: Based on the previous result (Gao & Zhou, 2015) (i.e., Theorem 7 on Page 6), we can get that L_{u_1} is consistent with the (base) exponential loss.

- L_{u_2}: As a reweighting univariate loss, L_{u_2} involves reweighting factors depending on the dataset (i.e., $\frac{1}{|S_k^+|}$, $\frac{1}{|S_k^-|}$ for the positive and negative instances, respectively). Thus, in the infinite-sample (i.e., population) setting, it could be regarded to be dependent on the distribution, where the reweighting factors of positive and negative instances are propositional to $\frac{1}{P(y_k=1)}$ and $\frac{1}{1-P(y_k=1)}$, respectively. In this case, based on the previous result of bipartite ranking (Kotlowski et al., 2011) (i.e., Theorem 4.1 on Page 4), we can get that L_{u_2} is consistent w.r.t. Macro-AUC with the (base) logistic loss and exponential loss.

As for the surrogates L_{u_1} and L_{u_2} with other base losses, we left it as future work.

5. Related Work

Consistency. Gao & Zhou (2013) studied the consistency of surrogate losses w.r.t. the Hamming and (partial) ranking measures in general. Besides, Dembczynski et al. (2012) presented an explicit regret bound for a surrogate univariate loss under the partial ranking measure. Notably, for the F-measure in binary classification, Ye et al. (2012) justified and connected the empirical utility maximization (EUM) framework and the decision-theoretic approach (DTA), with applications to optimizing the macro-F measure in MLC.[12] For these two approach frameworks, Dembczyński et al. (2017) revisited the consistency analysis for binary classification with complex metrics, where they chose the more descriptive names Population Utility (PU) and Expected Test Utility (ETU). Further, for the F-measure in MLC, Waegeman et al. (2014); Zhang et al. (2020a) studied the consistency in the perspective of DTA via estimating the conditional distribution $P(\mathbf{y}|\mathbf{x})$ differently. Further, Koyejo et al. (2015) studied consistent MLC approaches w.r.t. various measures in the EUM framework and Menon et al. (2019) investigated the multi-label consistency of various reduction methods w.r.t. precision@k and recall@k measures.

Generalization. Wu & Zhu (2020) studied the generalization of learning algorithms with surrogates aiming to optimize Hamming loss and subset accuracy w.r.t. these two measures, and found that the label size played an important role in the generalization bounds, which explains the empirical phenomena that when the label size is not large, optimizing Hamming loss with its surrogate can have promising performance w.r.t. subset accuracy. Further, Wu et al. (2021) revisited the consistency and generalization of many surrogate loss-based algorithms w.r.t. the ranking loss measure and identified the *instance-wise class imbalance* of the dataset (or distribution) plays a critical role in the generalization bounds, which could explain the empirical phenomena better than consistency.

We mention that Wu & Zhou (2017) also proposed a pairwise loss (similar to Eq. (5)), which omits the reweighting factor $\frac{1}{|S_k^+||S_k^-|}$ and lacks formal generalization analyses. Besides, please see Appendix D for detailed discussions about comparisons between a recent McDiarmid-type concentration inequality (Zhang et al., 2019) and ours for data with graph dependence.

6. Experiments

As a theoretical work, the primary goal of experiments is to verify our theory findings rather than illustrate the superior performance of the proposed method. Therefore, we evaluate the aforementioned three learning algorithms in Section 3.2 in terms of Macro-AUC on 10 widely-used benchmark datasets with various domains and sizes of labels and data. The detailed statistics of the datasets are summarized in Table 2, including four label-wise class imbalance-related factors.[13] Besides, the label-wise class imbalance levels of three representative datasets are illustrated in Figure 1. (See Figure 2 in Appendix C.1 for all datasets). For all algorithms, we take linear models with the base logistic loss for simplicity and fair comparison. Besides, we utilize the same efficient stochastic optimization algorithm (i.e., SVRG-BB (Tan et al., 2016)) to solve these convex optimization problems. Moreover, we search the hyper-parameter λ for all algorithms on all datasets in a wide range of $\{10^{-6}, 10^{-5}, \ldots, 10^{2}\}$ using 3-fold cross-validation.[14]

The experimental results are summarized in Table 3. Overall, we can observe that algorithms $\mathcal{A}^{pa}$ and $\mathcal{A}^{u_2}$ performs better than the algorithm $\mathcal{A}^{u_1}$, which confirms our theoretical results that $\mathcal{A}^{pa}$ and $\mathcal{A}^{u_2}$ have better learning guarantees w.r.t. the label-wise class imbalance than $\mathcal{A}^{u_1}$. Besides, $\mathcal{A}^{pa}$ performs comparably to $\mathcal{A}^{u_2}$, which also verifies our theoretical results that they share the learning guarantee w.r.t. the label-wise class imbalance.

Further, from Table 2 and 3, we can carefully study the effects of the label-wise class imbalance on the performance. Recall that the learning guarantees of $\mathcal{A}^{pa}$ and $\mathcal{A}^{u_2}$ both depends on the factor Imb_1, while the one of $\mathcal{A}^{u_1}$ depends on the factor Imb_4. For the datasets CAL500, enron, rcv-s1, bibtex, corel5k and delicious, factors Imb_1 and Imb_4 have a large order gap (or equivalently Imb_3 is large), and $\mathcal{A}^{u_2}$ (or $\mathcal{A}^{pa}$) performs significantly better than $\mathcal{A}^{u_1}$. In contrast, for the datasets emotions, image, and scene, factors Imb_1 and Imb_4 have a small gap (or equivalently Imb_3 is small), and $\mathcal{A}^{u_2}$ (or $\mathcal{A}^{pa}$) performs slightly better than or is nearly comparable to $\mathcal{A}^{u_1}$. This also confirms our theoretical findings of these algorithms on the label-wise class imbalance.

Furthermore, similarly to previous theoretical results (Wu et al., 2021) for the Ranking Loss measure in MLC, our generalization upper bound absolute values might not reflect the true generalization error reasonably well (i.e., bigger than 1). However, they can still offer valuable insight into these learning algorithms under the same analytical framework. (See Table 4 in Appendix C.2 for details). Advanced techniques (e.g., local Radermacher-type complexity) can refine the results, left as future work.

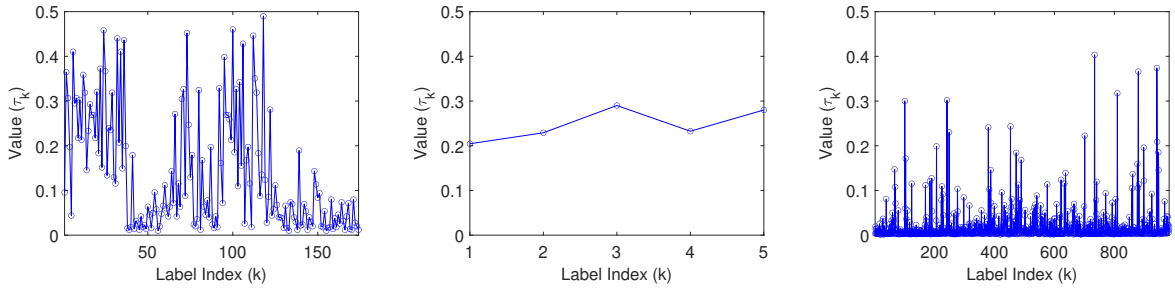
[12] Note that our generalization analyses of learning algorithms w.r.t. Macro-AUC is in the EUM framework.

[13] These datasets can be downloaded from http://mulan.sourceforge.net/datasets-mlc.html and http://palm.seu.edu.cn/zhangml/.

[14] Our code is available at https://github.com/GuoqiangWoodrowWu/Macro-AUC-Theory

Table 2. Basic statistics of the benchmark datasets. Denote the label-wise class imbalance-related factors $\text{Imb}_1 = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}$, $\text{Imb}_2 = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\tau_k}}$, $\text{Imb}_3 = \frac{1}{\tau_S^*}$ and $\text{Imb}_4 = \frac{1}{\tau_S^*}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right)$, respectively.

Dataset	#Instance	#Feature	#Label	Domain	Imb_1	Imb_2	Imb_3	Imb_4
CAL500	502	68	174	music	4.2	4.8	100.4	421.1
emotions	593	72	6	music	1.8	1.8	4.0	7.3
image	2000	294	5	images	2.0	2.0	4.9	9.9
scene	2407	294	6	images	2.4	2.4	6.6	15.7
yeast	2417	103	14	biology	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	images	23.4	29.1	5000	117000
delicious	16105	500	983	text(web)	12.2	13.3	766.9	9344



(a) CAL500 (b) image (c) delicious

Figure 1. Illustration of the label-wise class imbalance of three representative datasets.

Table 3. Macro-AUC (mean ± std, the symbol . means 0.) of all three algorithms on benchmark datasets. On each dataset, the top two algorithms are highlighted in bold and the top one is labeled with †. Besides, "-" means that $\mathcal{A}^{pa}$ takes more than one week by using a 16-core CPU server on the corresponding datasets.

Dataset	$\mathcal{A}^{pa}$	$\mathcal{A}^{u_1}$	$\mathcal{A}^{u_2}$
CAL500	**.5735 ± .0186**†	.5571 ± .0102	**.5717 ± .0177**
emotions	**.8372 ± .0172**†	.8346 ± .0223	**.8348 ± .0189**
image	**.8383 ± .0073**†	**.8359 ± .0121**	.8314 ± .0094
scene	**.9319 ± .0013**†	.9271 ± .0067	**.9285 ± .0035**
yeast	**.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**†

7. Conclusion

Towards understanding the generalization of Macro-AUC in MLC, this paper takes an initial step by analyzing the generalization bounds of the algorithms with various surrogates including the widely-used univariate one. Our results show that the label-wise class imbalance of the dataset plays a critical role in these bounds. The algorithms with the proposed pairwise and reweighted univariate loss have better learning guarantees than the original univariate-based algorithm, which probably implies their superior performance. Experimental results also confirm our theoretical findings.

Social Impact: As a theoretical research, this work will help understand and potentially develop better algorithms for multi-label learning, while without explicit negative consequences to society.

Acknowledgements

This work was supported by NSF of China (Nos. 62206159, 62076145); Shandong Provincial Natural Science Foundation (Nos. ZR2022QF117, ZR2021ZD15); Beijing Outstanding Young Scientist Program (NO. BJJWZYJH012019100020098); the Fundamental Research Funds of Shandong University; Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China; the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (22XNKJ13). C. Li was also sponsored by Beijing Nova Program.

References

Amini, M.-R. and Usunier, N. *Learning with partially labeled and interdependent data*. Springer, 2015.

Anthony, M. and Bartlett, P. L. *Neural network learning: Theoretical foundations*. Cambridge University Press, 1999.

Arora, S. and Barak, B. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

Carneiro, G., Chan, A. B., Moreno, P. J., and Vasconcelos, N. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

Dembczynski, K., Kotłowski, W., and Hüllermeier, E. Consistent multilabel ranking through univariate loss minimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1347–1354, 2012.

Dembczyński, K., Kotłowski, W., Koyejo, O., and Natarajan, N. Consistency analysis for binary classification revisited. In *International Conference on Machine Learning*, pp. 961–969. PMLR, 2017.

Elisseeff, A. and Weston, J. A kernel method for multi-labelled classification. *Advances in neural information processing systems*, 14, 2001.

Gao, W. and Zhou, Z.-H. On the consistency of multi-label learning. *Artificial Intelligence*, 199(1):22–44, 2013.

Gao, W. and Zhou, Z.-H. On the consistency of auc pairwise optimization. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Hariharan, B., Zelnik-Manor, L., Varma, M., and Vishwanathan, S. Large scale max-margin multi-label classification with priors. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 423–430, 2010.

Janson, S. Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms*, 24 (3):234–248, 2004.

Kotlowski, W., Dembczynski, K. J., and Huellermeier, E. Bipartite ranking through minimization of univariate loss. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1113–1120. Citeseer, 2011.

Koyejo, O., Natarajan, N., Ravikumar, P., and Dhillon, I. S. Consistent multilabel classification. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3321–3329, 2015.

McCallum, A. K. Multi-label text classification with a mixture model trained by EM. In *AAAI 99 workshop on Text Learning*. Citeseer, 1999.

McDiarmid, C. et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Menon, A. K., Rawat, A. S., Reddi, S., and Kumar, S. Multilabel reductions: what is my loss optimising? In *Advances in Neural Information Processing Systems*, pp. 10599–10610, 2019.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91, 2021.

Schapire, R. E. and Singer, Y. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2): 135–168, 2000.

Tan, C., Ma, S., Dai, Y.-H., and Qian, Y. Barzilai-borwein step size for stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 685–693, 2016.

Tan, Z.-H., Tan, P., Jiang, Y., and Zhou, Z.-H. Multi-label optimal margin distribution machine. *Machine Learning*, 109(3):623–642, 2020.

Tarekegn, A. N., Giacobini, M., and Michalak, K. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965, 2021.

Usunier, N., Amini, M. R., and Gallinari, P. Generalization error bounds for classifiers trained with interdependent data. *Advances in neural information processing systems*, 18, 2005.

Waegeman, W., Dembczyński, K., Jachnik, A., Cheng, W., and Hüllermeier, E. On the bayes-optimality of f-measure maximizers. *Journal of Machine Learning Research*, 15: 3333–3388, 2014.

Wu, G. and Zhu, J. Multi-label classification: do hamming loss and subset accuracy really conflict with each other? *Advances in Neural Information Processing Systems*, 33: 3130–3140, 2020.

Wu, G., Zheng, R., Tian, Y., and Liu, D. Joint ranking svm and binary relevance with robust low-rank learning for multi-label classification. *Neural Networks*, 122:24–39, 2020a.

Wu, G., Li, C., Xu, K., and Zhu, J. Rethinking and reweighting the univariate losses for multi-label ranking: Consistency and generalization. *Advances in Neural Information Processing Systems*, 34:14332–14344, 2021.

Wu, T., Huang, Q., Liu, Z., Wang, Y., and Lin, D. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, pp. 162–178, 2020b.

Wu, X.-Z. and Zhou, Z.-H. A unified view of multi-label performance measures. In *international conference on machine learning*, pp. 3780–3788. PMLR, 2017.

Ye, N., Chai, K. M., Lee, W. S., and Chieu, H. L. Optimizing f-measures: a tale of two approaches. In *International Conference on Machine Learning*, pp. 289–296. Omnipress, 2012.

Zhang, M., Ramaswamy, H. G., and Agarwal, S. Convex calibrated surrogates for the multi-label f-measure. In *International Conference on Machine Learning*, pp. 11246–11255. PMLR, 2020a.

Zhang, M.-L. and Zhou, Z.-H. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

Zhang, M.-L., Li, Y.-K., Yang, H., and Liu, X.-Y. Towards class-imbalance aware multi-label learning. *IEEE Transactions on Cybernetics*, 2020b.

Zhang, R.-R. and Amini, M.-R. Generalization bounds for learning under graph-dependence: A survey. *arXiv preprint arXiv:2203.13534*, 2022.

Zhang, R. R., Liu, X., Wang, Y., and Wang, L. Mcdiarmid-type inequalities for graph-dependent variables and stability bounds. *Advances in Neural Information Processing Systems*, 32, 2019.

Contents of Appendix

A General Techniques **13**

A.1 A new McDiarmid-type concentration inequality . . . 13

A.1.1 Backgrounds . . . 13

A.1.2 Proof of the new McDiarmid-type concentration inequality . . . 14

A.2 Learning multiple tasks with graph-dependent examples . . . 16

A.2.1 Problem setting . . . 16

A.2.2 The fractional Rademacher complexity of the loss space . . . 17

A.2.3 Proof of the general generalization bound of learning multiple tasks with graph-dependent examples 17

B Macro-AUC Maximization in MLC **19**

B.1 Proof of Theorem 1 . . . 19

B.1.1 Problem transformation . . . 19

B.1.2 Proof of Theorem 1 . . . 19

B.2 Proof of Lemma 1 and 2 . . . 20

B.2.1 Proof of Lemma 1 . . . 20

B.2.2 Proof of Lemma 2 . . . 20

B.3 Proof of Theorem 2, Corollary 4 and 2 . . . 22

B.3.1 The fractional Rademacher complexity of the hypothesis space . . . 22

B.3.2 The contraction inequality . . . 23

B.3.3 Proof of Theorem 2 . . . 25

B.3.4 Proof of Corollary 4 . . . 25

B.3.5 Proof of Corollary 2 . . . 25

B.4 Proof of Theorem 3 and 4, Corollary 1, 3 and 5 . . . 25

B.4.1 The fractional Rademacher complexity of the hypothesis space . . . 25

B.4.2 The contraction inequality . . . 27

B.4.3 Proof of Theorem 3 . . . 29

B.4.4 Proof of Theorem 4 . . . 29

B.4.5 Proof of Corollary 1 . . . 30

B.4.6 Proof of Corollary 3 . . . 30

B.4.7 Proof of Corollary 5 . . . 30

C Additional Experimental Results **30**

C.1 Label-wise class imbalance illustrations of benchmark datasets . . . 30

C.2 Experimental results about the absolute value of bounds . . . 30

D Additional Related Work **30**

A. General Techniques

In this section, we introduce the general techniques, which mainly consist of a new McDiarmid-type concentration inequality and a general generalization bound of learning multiple tasks with graph-dependent examples.

A.1. A new McDiarmid-type concentration inequality

A.1.1. Backgrounds

First, we introduce the bounded differences property and a lemma for the proof of the subsequent theorem.

Definition 4 (**The bounded differences property** (McDiarmid et al., 1989))**.** *Let $x_1, x_2, \ldots, x_m \in \mathcal{X}$, and function $f : \mathcal{X}^m \to \mathbb{R}$. Then, f is said to have bounded differences property if there exist $c_1, \ldots, c_m > 0$ such that*

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

for all $i \in [m]$ and any points $x_1, \ldots, x_m, x_i' \in \mathcal{X}$.

Lemma 3 ((McDiarmid et al., 1989))**.** *Let $\mathbf{X} = (X_1, \ldots, X_m) \in \mathcal{X}^m$ be a vector of m independent random variables and function $f : \mathcal{X}^m \to \mathbb{R}$ satisfies the bounded differences property with c_i ($i \in [m]$), then for any $s > 0$,*

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

Here we introduce some necessary notions of graph theory in this paper, and we refer readers to (Janson, 2004; Amini & Usunier, 2015) and recent survey (Zhang & Amini, 2022).

Given a graph $G = (V, E)$, we introduce the following notions.

Definition 5 (**Fractional independent vertex cover, and fractional chromatic number** (Zhang & Amini, 2022))**.**

*(1) A family $\{(F_j, \omega_j)\}_j$ of pairs (F_j, ω_j), where $F_j \subseteq V(G)$ and $\omega_j \in (0, 1]$ is a **fractional vertex cover** of G if $\sum_{j:v\in F_j} \omega_j = 1$ for every $v \in V(G)$.*

*(2) An **independent set** of G is a set of vertices in G such that no two them are adjacent. The set of independent sets of G is denoted by $\mathcal{I}(G)$.*

*(3) A **fractional independent vertex cover** $\{(I_j, \omega_j)\}_j$ of G is fractional vertex cover such that $I_j \in \mathcal{I}(G)$ for every j.*

*(4) A **fractional coloring** of a graph G is a mapping g from $\mathcal{I}(G)$ to $(0, 1]$ such that $\sum_{I\in\mathcal{I}(G):v\in I} g(I) \ge 1$ for every vertex $v \in V(G)$. The **fractional chromatic number** $\chi_f(G)$ is the minimum of the value $\sum_{I\in\mathcal{I}(G)} g(I)$ over fractional colorings of G.*

Note that the fractional chromatic number $\chi_f(G)$ of a graph G is the minimum of $\sum_j \omega_j$ over all fractional independent vertex covers $\{(I_j, \omega_j)\}_j$ of G.

Next, we introduce the notion of dependency graph as follows.

Definition 6 (**Dependency graph** (Janson, 2004))**.** *An undirected graph $G = (V, E)$ is called a dependency graph of a random vector $\mathbf{X} = (X_1, \ldots, X_m)$ if*

(1) $V(G) = [m]$.

(2) For all disjoint $I, J \in [m]$, if I, J are not adjacent in G, then random variables $\{X_i\}_{i\in I}$ and $\{X_j\}_{j\in J}$ are independent.

Then, we say that random vector $\mathbf{X}$ is G-dependent with a dependency graph G.

An important property of the dependency graph, combined with the notion of fractional independent covers, is Janson's decomposition property (Janson, 2004). Specifically, suppose interdependent random variables $(X_i)_{i\in[m]}$ is G-dependent with a dependency graph G, and $\{(I_j, \omega_j)\}_{j\in[J]}$ is a fractional independent vertex cover of G. Then, we can decompose the sum of interdependent variables into a weighted sum of sums of independent variables, i.e.,

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

A.1.2. Proof of the new McDiarmid-type concentration inequality

Here we propose a new and more general McDiarmid-type inequality as follows, which mainly follows the work (Janson, 2004; Usunier et al., 2005; Amini & Usunier, 2015) and we refer to a recent related survey (Zhang & Amini, 2022).

Theorem 5 (A new and more general McDiarmid-type inequality). *Let $\mathbf{X}_1 = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{1m_1}) \in \mathcal{X}^{m_1}, \ldots, \mathbf{X}_K = (\mathbf{x}_{K1}, \ldots, \mathbf{x}_{Km_K}) \in \mathcal{X}^{m_K}$ be vectors of random variables and $\mathbf{X}$ denote $(\mathbf{X}_1, \ldots, \mathbf{X}_K) = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{Km_K})$ for convenience. Let $f_1 : \mathcal{X}^{m_1} \to \mathbb{R}, \ldots, f_K : \mathcal{X}^{m_K} \to \mathbb{R}$ and $f : \mathcal{X}^m \to \mathbb{R}$ be functions with $\sum_{k=1}^K m_k = m$. Assume each $\mathbf{X}_k$ ($k \in [K]$) is G_k-dependent with a dependency graph G_k.[15] Besides, assume the function f satisfies the following constraints:*

(1) $f(\mathbf{X}) = \sum_{k\in[K]} f_k(\mathbf{X}_k)$;
(2) $f_k(\mathbf{X}_k)$ has the decomposability constraint with the bounded difference property w.r.t. the graph G_k, i.e., for all $\mathbf{x}_k \in \mathcal{X}^{m_k}$ and the minimal fractional independent vertex covers $\{(I_{kj}, \omega_{kj})\}_{j\in[J_k]}$ of G_k, there exists functions $\{f_{kj} : \mathcal{X}^{|I_{kj}|} \to \mathbb{R}\}_{j\in[J_k]}$ such that f_{kj} satisfies the bounded difference property with c_{ki} ($i \in I_{kj}$) and

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

where $\mathbf{x}_{I_{kj}}$ denotes $(\mathbf{x}_{ki})_{i\in I_{kj}}$.

Then, for any $t > 0$,

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

where $\chi_f(G_k)$ is the fractional chromatic number of G_k.

Remark. *The McDiarmid-type inequality in prior work (Usunier et al., 2005; Amini & Usunier, 2015) can be viewed as a special case of the above one by setting $K = 1$. Thus, our proposed new McDiarmid-type inequality is more general.*

Proof. Following the Cramér-Chernoff method (Boucheron et al., 2013), we have for any $s > 0$ and $t > 0$,

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

For the dependency graph, the k-th sub-graph G_k has m_k vertexes. Further, let I_k be the vertex set of G_k and $\{(I_{kj}, \omega_{kj})\}_{j\in[J_k]}$ be a minimal fractional independent vertex cover of G_k with $\sum_{j\in[J_k]} \omega_{kj} = \chi_f(G_k)$. Utilizing the decomposition property $f(\mathbf{x}) = \sum_{k\in[K]} f_k(\mathbf{x}_k) = \sum_{k\in[K]}\sum_{j\in[J_k]} \omega_{kj} f_{kj}(I_{kj})$ where $\mathbf{x}_{I_{kj}}$ is denoted by I_{kj} for notation simplicity, we have for the expectation term on the right-hand side of the above inequality:

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

Let $\{p_1, p_2, ..., p_K\}$ be any set of K strictly positive real numbers that sum to 1. Similarly, for each $k \in [K]$, let $\{q_{k1}, q_{k2}, ..., q_{kJ_k}\}$ be any set of J_k strictly positive real numbers that sum to 1. Then, based on the convexity of the exponential function, we can have the following:

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

[15] Note that here we only make the dependency assumptions within each $\mathbf{X}_k$ but have no assumptions between different $\mathbf{X}_k$s, where $\mathbf{X}_k$s can be independent or dependent, regardless of independence.

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

Here we can observe that for the summation term $\clubsuit_k$, the random variables associated with each term $j \in [J_k]$ are independent. Thus, applying the Lemma 3, we can get

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

By rearranging terms in the exponential of right hand side of the inequality above and by setting

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

we have:

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

Till now, we have the following:

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

Next, similarly to the above proof idea w.r.t. the q_{kj}, we set p_k as follows:

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

Then, it comes:

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

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

For ①, it is based on the inequality $\left(\sum_{i=1}^n a_i\right)^2 \leq n\sum_{i=1}^n a_i^2$. For ②, it is due to the Cauchy-Schwarz inequality. For ③, it is due to the definition of the fractional chromatic number, i.e., $\sum_{j\in[J_k]}\omega_{kj} = \chi_f(G_k)$, and the decomposition property of fractional independent vertex covers of dependency graph (Janson, 2004), i.e., for a fractional independent vertex cover $\{(I_{kj},\omega_{kj})\}_{j\in[J_k]}$ of G_k, then the sum of interdependent variables can be decomposed into a weighted sum of sums of independent variables as follows:

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

Since $\mathbf{X}_k = [\mathbf{x}_{k1},\ldots,\mathbf{x}_{km_k}]$ is a random vector, we can take the specific values to get the equation ③ based on the inequality ②. Specifically, if we take $\mathbf{x}_{ki} = c_{ki}^2$ for each $i\in[m_k]$, then we can get

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

Thus, we have obtained

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

Combining the inequality (14), we can get

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

We can obtain the final result by minimizing the right-hand side of the above inequality over s. □

A.2. Learning multiple tasks with graph-dependent examples

A.2.1. Problem setting

Here we consider learning with multiple tasks where each task might contain dependent training examples and the dependency relationship is characterized by a dependency graph. Formally, given a training dataset $\tilde{S} = \{(\tilde{\mathbf{x}},\tilde{y})\}_{i=1}^m$ that is composed of K blocks (or tasks), i.e., $\tilde{S} = (\tilde{S}_1,\ldots,\tilde{S}_K)$ with each $\tilde{S}_k = \{(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki})\}_{i=1}^{m_k}$ drawn from the distribution D_k ($k\in[K]$) over $\widetilde{\mathcal{X}}\times\widetilde{\mathcal{Y}}$ with a dependency graph G_k and $\sum_{k\in[K]}m_k = m$. The goal is to learn a mapping $\tilde{h} = (\tilde{h}_1,\ldots,\tilde{h}_K)$, where $\tilde{h}_k:\widetilde{\mathcal{X}}\to\widetilde{\mathcal{Y}}$ for each $k\in[K]$.

Let $\widetilde{\mathcal{F}} = \left\{\tilde{f} = (\tilde{f}_1,\ldots,\tilde{f}_K) \mid \tilde{f}_k:\widetilde{\mathcal{X}}\to\widetilde{\mathcal{Y}}, k\in[K]\right\}$ be the hypothesis space, and denote $\widetilde{\mathcal{F}}_k = \left\{\tilde{f}_k \mid \tilde{f}_k:\widetilde{\mathcal{X}}\to\widetilde{\mathcal{Y}}\right\}$ for each $k\in[K]$. Consider a loss function $L:\widetilde{\mathcal{X}}\times\widetilde{\mathcal{Y}}\times\widetilde{\mathcal{F}}_k\to\mathbb{R}_+$. For a hypothesis $\tilde{f}\in\widetilde{\mathcal{F}}$ and a training set $\tilde{S}$, the empirical risk of $\tilde{f}$ is defined as

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

and the generalization (or expected) risk is defined as

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

Note that we do not define the generalization risk as the following usual form

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

This is because the definition in Eq.(15) is more general than Eq.(16). Specifically, Eq.(15) can cover the loss function dependent on the training set $\tilde{S}$ while Eq.(16) cannot. Besides, they are equal for certain losses independent of $\tilde{S}$.

A.2.2. The fractional Rademacher complexity of the loss space

Here we give the definition of the fractional Rademacher complexity of the loss space as follows.

Definition 7 (**The fractional Rademacher complexity of the loss space**). *For each $k \in [K]$, let $\{(I_{kj},\omega_{kj})\}_{j\in[J_k]}$ be a fractional independent vertex cover of the dependence graph G_k constructed over $\tilde{S}_k$ with $\sum_{j\in[J_k]}\omega_{kj} = \chi_f(G_k)$. Let $\widetilde{\mathcal{F}} = \left\{\tilde{f} = (\tilde{f}_1,\ldots,\tilde{f}_K) \mid \tilde{f}_k : \widetilde{\mathcal{X}} \to \widetilde{\mathcal{Y}}, k \in [K]\right\}$ be the hypothesis space. Then, the empirical fractional Rademacher complexity of $\widetilde{\mathcal{F}}$ given $\tilde{S}$ is defined by*

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

where $\boldsymbol{\sigma} = (\sigma_{ki})_{k\in[K],i\in[m_k]}$ denotes m independent Rademacher variables, that is, $\mathbb{P}(\sigma_{ki}=+1) = \mathbb{P}(\sigma_{ki}=-1) = 1/2$ for all variables. Furthermore, the fractional Rademacher complexity of $\widetilde{\mathcal{F}}$ over all samples of size m is defined by

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

where $\tilde{S}\sim D^{m}_{[K]}$ denotes $\tilde{S}_1\sim D_1^{m_1},\ldots,\tilde{S}_K\sim D_K^{m_K}$ for simplicity.

A.2.3. Proof of the general generalization bound of learning multiple tasks with graph-dependent examples

Here we give a general generalization bound of learning multiple tasks with graph-dependent examples as follows.

Theorem 6 (**A general generalization bound of learning multiple tasks with graph-dependent examples**). *Give a sample $\tilde{S} = \{\tilde{S}_1,\ldots,\tilde{S}_K\}$ where each $\tilde{S}_{k\in[K]}$ is of size m_k with dependency graph G_k and a loss function $L : \widetilde{\mathcal{X}}\times\widetilde{\mathcal{Y}}\times\widetilde{\mathcal{F}}_k \to [0,M]$. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$, we have*

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

and

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

Proof. The proof can be divided into three major steps as follows.

Step 1: link the supremum of $R(\tilde{f}) - \widehat{R}_{\tilde{S}}(\tilde{f})$ on $\widetilde{\mathcal{F}}$ with its expectation.

For any $\tilde{f} \in \widetilde{\mathcal{F}}$, we have $\widehat{R}(\tilde{f})$ is an unbiased estimator of $R(\tilde{f})$ because the data points in the sample $\tilde{S}_k$ are assumed to be G-dependent and have the same marginal distribution. Hence considering an independent ghost sample $\tilde{S}'$ with the same generation process as $\tilde{S}$, we have

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

For each $k \in [K]$, let $\{(I_{kj}, \omega_{kj})\}_{j\in[J_k]}$ be a fractional independent vertex cover of the dependence graph G_k with $\sum_{j\in[J_k]}\omega_{kj} = \chi_f(G_k)$. Since the supremum of the expectation is lower than the expectation of the supremum, we can have

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

where the inequality ① is due to the Janson's decomposition (Janson, 2004), and ② is due to the sub-additivity of the supremum function (i.e., $\sup(a+b) \le \sup(a) + \sup(b)$) and the linearity of the expectation.

By defining $g(\tilde{S}) = \sum_{k=1}^{K} g_k(\tilde{S}_k)$ with each $g_k : \tilde{S}_k \mapsto \sum_{j\in[J_k]}\omega_{kj} g_{kj}(I_{kj})$ where each

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

have differences bounded by $\frac{M}{Km_k}$ in the sense of the condition of Theorem 5; then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

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

Step 2: bound ★ with respect to the fractional Rademacher complexity.

Next, taking the symmetrization technique by introduction of Rademacher variables, we have

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

For a fixed pair (k,i), $\sigma_{ki}=1$ does not change anything but $\sigma_{ki}=-1$ consists in swapping both examples $(\tilde{\mathbf{x}}'_{ki},\tilde{y}'_{ki})$ and $(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki})$. Thus, when taking the expectations over $\tilde{S}_k$ and $\tilde{S}'_k$, the introduction of Rademacher variables does not change the value. For ③, it is due to the sub-additivity of the supremum function and the linearity of the expectation.

Thus, we can obtain

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

Besides, based on the definition of supremum of functions, we have

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

Then, we can obtain the desired first bound (17).

Step 3: bound the fractional Rademacher complexity with the empirical one.

By defining $g(\tilde{S})=\sum_{k=1}^{K}g_k(\tilde{S}_k)$ with each $g_k:\tilde{S}_k\mapsto\sum_{j\in[J_k]}\omega_{kj}g_{kj}(I_{kj})$ where each

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

having differences bounded by $\frac{M}{Km_k}$ in the sense of the condition of Theorem 5; then for any $\delta\in(0,1)$, with probability at least $1-\delta$, we have

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

Then, we can get the desired second bound (18) by using the union bound with the first bound (17). □

B. Macro-AUC Maximization in MLC

B.1. Proof of Theorem 1

B.1.1. Problem transformation

For the Macro-AUC maximization problem in multi-label learning, we can transform it into the problem of learning multiple tasks with graph-dependent examples which is considered in Section A.2.1.

Specifically, construct the training dataset $\tilde{S}$ based on the original training set S as follows. For each label $k\in[K]$, based on the original dataset S_k, construct the dataset $\tilde{S}_k=\{(\tilde{\mathbf{x}}_{ki},\tilde{y}_{ki})\}_{i=1}^{m_k}$, where $\tilde{\mathbf{x}}_{ki}=(\tilde{\mathbf{x}}^{+}_{ki},\tilde{\mathbf{x}}^{-}_{ki})$, $\tilde{y}_{ki}=1$, and $(\tilde{\mathbf{x}}^{+}_{ki},\tilde{\mathbf{x}}^{-}_{ki})\in S^{+}_k\times S^{-}_k$, $m_k=|S^{+}_k||S^{-}_k|=n^2\tau_k(1-\tau_k)$ and let $\{(I_{kj},\omega_{kj})\}_{j\in[J_k]}$ be a fractional independent vertex cover of the dependence graph G_k constructed over $\tilde{S}_k$ with $\sum_{j\in[J_k]}\omega_{kj}=\chi_f(G_k)$, where $\chi_f(G_k)$ is the fractional chromatic number of G_k. From previous results in bipartite ranking (Usunier et al., 2005; Amini & Usunier, 2015), we know that

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

Besides, $\tilde{f}_k(\tilde{\mathbf{x}}_i)=f_k(\mathbf{x}^{+}_i)-f_k(\mathbf{x}^{-}_i)$ for each label $k\in[K]$.

B.1.2. Proof of Theorem 1

Theorem 1 (**The base theorem of Macro-AUC**). *Assume the loss function $L_\phi:\mathcal{X}\times\mathcal{X}\times\mathcal{F}_k\to\mathbb{R}_+$ is bounded by M. Then, for any $\delta>0$, the following generalization bound holds with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n:*

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

Proof. Based on the problem transformation in Section B.1.1, we can straightforwardly get this theorem by applying Theorem 6. □

B.2. Proof of Lemma 1 and 2

B.2.1. Proof of Lemma 1

Lemma 1 (The relationship between the surrogate and true losses). *Assume the base loss function upper bounds the original* $0/1$ *loss, i.e.,* $\ell(t) \geq [\![t \leq 0]\!]$. *Then, for any* $f_k \in \mathcal{F}_k$ *and* $(\mathbf{x}^+, \mathbf{x}^-) \in S_k^+ \times S_k^-$, *the following inequalities hold:*

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

Proof. For the first inequality, the following holds:

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

For the second inequality, the following holds:

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

Thus, the inequalities hold. □

B.2.2. Proof of Lemma 2

Lemma 2 (The relationship between the surrogate and true risks). *Assume the base loss function upper bounds the original* $0/1$ *loss, i.e.,* $\ell(t) \geq [\![t \leq 0]\!]$. *Then, for any* $f \in \mathcal{F}$ *and any sample* $S \overset{i.i.d.}{\sim} P$, *the following inequalities hold:*

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

Proof. For the first inequality, the following holds:

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

For the second inequality, we first have the following

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

Then, we can get

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

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

Thus, the second inequality holds. □

B.3. Proof of Theorem 2, Corollary 4 and 2

B.3.1. The fractional Rademacher complexity of the hypothesis space

Definition 8 (**The fractional Rademacher complexity of the hypothesis space for L_{pa}**). *Given a dataset S (and its corresponding constructed dataset $\tilde{S}$), define the empirical fractional Rademacher complexity of the hypothesis space $\widetilde{\mathcal{F}}$ w.r.t. $\tilde{S}$ as follows:*

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

Lemma 4 (**The fractional Rademacher complexity of the kernel-based hypothesis space for L_{pa}**). *Suppose (1) and (2) in* Assumption 1 *hold. Then, for the kernel-based hypothesis space* (8)*, its empirical fractional Rademacher complexity w.r.t. the dataset $\tilde{S}$, can be bounded as bellow:*

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

Proof. By the definition of $\widehat{\mathfrak{R}}_{\tilde{S}}^*(\widetilde{\mathcal{F}})$, we have

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

$$= \frac{1}{K}\sum_{k=1}^{K}\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\sum_{i\in I_{kj}}\langle\Phi(\tilde{\mathbf{x}}_{ki}),\Phi(\tilde{\mathbf{x}}_{ki})\rangle\right)^{\frac{1}{2}} \qquad (\forall p\neq q, \mathbb{E}[\sigma_{kp}\sigma_{kq}]=0 \text{ and } \mathbb{E}[\sigma_{ki}\sigma_{ki}]=1)$$

$$= \frac{1}{K}\sum_{k=1}^{K}\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\sum_{i\in I_{kj}}\langle\Phi(\mathbf{x}^+_{ki})-\Phi(\mathbf{x}^-_{ki}),\Phi(\mathbf{x}^+_{ki})-\Phi(\mathbf{x}^-_{ki})\rangle\right)^{\frac{1}{2}} \qquad (\Phi(\tilde{\mathbf{x}})=\Phi(\mathbf{x}^+)-\Phi(\mathbf{x}^-))$$

$$\leq \frac{1}{K}\sum_{k=1}^{K}\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\sum_{i\in I_{kj}}(\langle\Phi(\mathbf{x}^+_{ki}),\Phi(\mathbf{x}^+_{ki})\rangle+\langle\Phi(\mathbf{x}^-_{ki}),\Phi(\mathbf{x}^-_{ki})\rangle)\right)^{\frac{1}{2}} \qquad (\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2)$$

$$= \frac{1}{K}\sum_{k=1}^{K}\frac{B}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\sum_{i\in I_{kj}}2\left(\kappa(\mathbf{x}^+_{ki},\mathbf{x}^+_{ki})+\kappa(\mathbf{x}^-_{ki},\mathbf{x}^-_{ki})\right)\right)^{\frac{1}{2}} \qquad (\kappa(\mathbf{x},\mathbf{x})=\langle\Phi(\mathbf{x}),\Phi(\mathbf{x})\rangle)$$

$$\leq \frac{1}{K}\sum_{k=1}^{K}\frac{2Br}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sqrt{m_{kj}} \qquad (\kappa(\mathbf{x},\mathbf{x})\leq r^2 \text{ and let } |I_{kj}|=m_{kj})$$

$$= \frac{1}{K}\sum_{k=1}^{K}\frac{2Br\chi_f(G_k)}{m_k}\sum_{j\in[J_k]}\frac{\omega_{kj}}{\chi_f(G_k)}\sqrt{m_{kj}}$$

$$\leq \frac{1}{K}\sum_{k=1}^{K}\frac{2Br\sqrt{\chi_f(G_k)}}{m_k}\sqrt{\sum_{j\in[J_k]}\omega_{kj}m_{kj}} \qquad \left(\sum_{j\in[J_k]}\frac{\omega_{kj}}{\chi_f(G_k)}=1 \text{ and Jensen's inequality}\right)$$

$$= \frac{1}{K}\sum_{k=1}^{K}2Br\sqrt{\frac{\chi_f(G_k)}{m_k}} \qquad \left(\sum_{j\in[J_k]}\omega_{kj}m_{kj}=m_k\right)$$

Since for Macro-AUC optimization in multi-label learning, $\chi_f(G_k)=\max\{|S_k^+|,|S_k^-|\}$, $m_k=|S_k^+||S_k^-|$ and $\tau_k=\min\{|S_k^+|,|S_k^-|\}$ hold, then we can get

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

Thus, we can get the desired result by combining the above equality and previous inequality. □

B.3.2. The contraction inequality

Lemma 5 (**Contraction inequality for** L_{pa})**.** *Assume the loss function $L_\phi = L_1(y,\tilde{f}_k(\tilde{\mathbf{x}}))$ is μ-Lipschitz continuous w.r.t. the second argument where L_1 denotes the loss function of two inputs for convenience. Then, the following holds:*

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

Proof. Since $\widehat{\mathfrak{R}}^*_{\tilde{S}}(L_\phi\circ\mathcal{F})=\frac{1}{K}\sum_{k=1}^{K}\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(L_\phi\circ\mathcal{F}_k)$ and $\widehat{\mathfrak{R}}^*_{\tilde{S}}(\mathcal{F})=\frac{1}{K}\sum_{k=1}^{K}\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(\mathcal{F}_k)$, we first prove $\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(L_\phi\circ\mathcal{F}_k)\leq\mu\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(\mathcal{F}_k)$ and then can get the desired result.

Here we prove the inequality $\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(L_\phi\circ\mathcal{F}_k)\leq\mu\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(\mathcal{F}_k)$ following the idea in Mohri et al. (2018) (Lemma 5.7, p.93), and we omit the index k and the symbol $\tilde{S}_k$ for notation simplicity in the following.

First we fix a sample $(\tilde{\mathbf{x}}_1,\ldots,\tilde{\mathbf{x}}_m)$, then by defintion,

$$\widehat{\mathfrak{R}}^*(L_\phi\circ f)=\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{j\in[J]}w_j\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}\left(\sum_{i\in I_j}\sigma_iL_1(y_i,\tilde{f}(\tilde{\mathbf{x}}_i))\right)\right]$$

$$=\frac{1}{m}\sum_{j\in[J]}w_j\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_{n_j-1}}\left[\mathop{\mathbb{E}}_{\sigma_{n_j}}\left[\sup_{\tilde{f}\in\widetilde{\mathcal{F}}}u_{n_j-1}(\tilde{f})+\sigma_{n_j}L_1(y_{n_j},\tilde{f}(\tilde{\mathbf{x}}_{n_j}))\right]\right], \qquad (\text{denote } n_j=|I_j| \text{ for simplicity})$$

where $u_{n_j-1}(\tilde{f}) = \sum_{i=1}^{n_j} \sigma_i L_1(y_i, \tilde{f}(\tilde{\mathbf{x}}_i))$. By the definition of the supremum, for any $\epsilon > 0$, there exists $\tilde{f}^1, \tilde{f}^2 \in \widetilde{\mathcal{F}}$ such that

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

and

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

Thus, for any $\epsilon > 0$, by definition of $\mathop{\mathbb{E}}_{\sigma_{n_j}}$,

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

Let $s = \operatorname{sgn}(\tilde{f}^1(\tilde{\mathbf{x}}_{n_j}) - \tilde{f}^2(\tilde{\mathbf{x}}_{n_j}))$. Then, the previous inequality implies

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

Since the inequality holds for any $\epsilon > 0$, we have

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

Proceeding in the same way for all other σ_i ($i \in [I_j], i \neq n_j$) proves that

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

By proceeding other $j \in [J]$, we can obtain the following

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

□

B.3.3. PROOF OF THEOREM 2

Theorem 2 (Learning guarantee of $\mathcal{A}^{pa}$ in general case). *Assume the loss $L_\phi = L_{pa}$, where L_{pa} is defined in Eq.(5). Besides,* Assumption 1 *holds. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

Proof. Since the base loss $\ell(z)$ is bounded by B and the loss $L_{pa}(\mathbf{x}^+, \mathbf{x}^-, f_k) = \ell\left(f_k(\mathbf{x}^+) - f_k(\mathbf{x}^-)\right)$, the loss $L_\phi = L_{pa}$ is bounded by B. Then, applying Theorem 1, and combining Lemma 5 and Lemma 4, we can obtain that for any $\delta > 0$, the following generalization bound holds with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n:

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

Finally, we can get the desired result by combining the above inequality and Lemma 2 (i.e., $R_{0/1}(f) \leq R_{pa}(f)$). □

B.3.4. PROOF OF COROLLARY 4

Corollary 4 (Learning guarantee of $\mathcal{A}^{pa}$ in balanced case). *Assume the loss $L_\phi = L_{pa}$, where L_{pa} is defined in Eq.(5). Besides,* Assumption 1 *holds and suppose S is balanced. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

Proof. It is straightforward to get the result by applying the Theorem 2 by plugging $\tau_k = \frac{1}{2}$. □

B.3.5. PROOF OF COROLLARY 2

Corollary 2 (Learning guarantee of $\mathcal{A}^{pa}$ in extremely imbalanced case). *Assume the loss $L_\phi = L_{pa}$, where L_{pa} is defined in Eq.(5). Besides,* Assumption 1 *holds and suppose S is extremely imbalanced. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

Proof. It is straightforward to get the result by applying the Theorem 2 by plugging $\tau_k = \frac{1}{n}$. □

B.4. Proof of Theorem 3 and 4, Corollary 1, 3 and 5

B.4.1. THE FRACTIONAL RADEMACHER COMPLEXITY OF THE HYPOTHESIS SPACE

Definition 9 (The fractional Rademacher complexity of the hypothesis space w.r.t. (reweighted) univariate losses). *Given a dataset S (and its corresponding constructed dataset $\tilde{S}$), assume the loss function $L_\phi = L_u(\mathbf{x}^+, \mathbf{x}^-, f_k) = a^+(S_k)\ell(f_k(\mathbf{x}^+)) + a^-(S_k)\ell(-f_k(\mathbf{x}^-))$ for each label $k \in [K]$, where $\ell(t)$ is the base loss function and the reweighting*

function $a^+(S_k)$ (or $a^-(S_k)$) indicates its dependency on S_k. Then, define the empirical fractional Rademacher complexity of the hypothesis space $\mathcal{F}$ w.r.t. $\tilde{S}$ and L_u as follows:

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

Lemma 6 (The fractional Rademacher complexity of kernel-based hypothesis space w.r.t. (reweighted) univariate losses). *Suppose (1) and (2) in* Assumption 1 *hold and the loss function $L_u(\mathbf{x}^+,\mathbf{x}^-,f_k) = a^+(S_k)\ell(f_k(\mathbf{x}^+)) + a^-(S_k)\ell(-f_k(\mathbf{x}^-))$. Then, for the kernel-based hypothesis space* (8)*, its empirical fractional Rademacher complexity w.r.t. the dataset $\tilde{S}$ and loss function L_u, can be bounded as bellow:*

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

Proof. By the definition of $\widehat{\mathfrak{R}}^*_{\tilde{S},L_u}(\mathcal{F})$ and let a^+_k (or a^-_k) denote $a^+(S_k)$ (or $a^-(S_k)$) for notation simplicity, we can have

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

where the inequality ① is due to sub-additivity of the supremum function (i.e., $\sup(a+b)\le\sup a+\sup b$), and

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

Next, we will (upper) bound $\clubsuit$ and $\spadesuit$, respectively.

Firstly, for $\clubsuit$, we can have

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

$$= \frac{1}{K}\sum_{k=1}^{K}\frac{Ba_k^+}{m_k}\sum_{j\in[J_k]}\omega_{kj}\left(\sum_{i\in I_{kj}}\langle\Phi(\mathbf{x}_{ki}^+),\Phi(\mathbf{x}_{ki}^+)\rangle\right)^{\frac{1}{2}} \qquad (\forall p\neq q, \mathbb{E}[\sigma_{kp}\sigma_{kq}]=0 \text{ and } \mathbb{E}[\sigma_{ki}\sigma_{ki}]=1)$$

$$\leq \frac{1}{K}\sum_{k=1}^{K}\frac{Bra_k^+}{m_k}\sum_{j\in[J_k]}\omega_{kj}\sqrt{m_{kj}} \qquad (\langle\Phi(\mathbf{x}_{ki}^+),\Phi(\mathbf{x}_{ki}^+)\rangle=\kappa(\mathbf{x}_{ki}^+,\mathbf{x}_{ki}^+)\leq r^2 \text{ and let } m_{kj}=|I_{kj}|)$$

$$= \frac{1}{K}\sum_{k=1}^{K}\frac{Bra_k^+\chi_f(G_k)}{m_k}\sum_{j\in[J_k]}\frac{\omega_{kj}}{\chi_f(G_k)}\sqrt{m_{kj}}$$

$$\leq \frac{1}{K}\sum_{k=1}^{K}\frac{Bra_k^+\sqrt{\chi_f(G_k)}}{m_k}\sqrt{\sum_{j\in[J_k]}\omega_{kj}m_{kj}} \qquad (\sum_{j\in[J_k]}\frac{\omega_{kj}}{\chi_f(G_k)}=1 \text{ and Jensen's inequality}).$$

Since $\sum_{j\in[J_k]}\omega_{kj}m_{kj}=m_k, \chi_f(G_k)=\max\{|S_k^+|,|S_k^-|\}, m_k=|S_k^+||S_k^-|$ and $\tau_k=\min\{|S_k^+|,|S_k^-|\}$ hold, it comes

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

Similarly to the proof of $\clubsuit$, we can obtain the upper bound for $\spadesuit$:

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

Thus, we can obtain the final result:

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

□

Proposition 1 (The fractional Rademacher complexity of the kernel-based hypothesis space w.r.t. L_{u_1} and L_{u_2})**.** *For the surrogate loss functions L_{u_1} and L_{u_2}, which are defined in Eq.(6) and Eq.(7) respectively, we have*

$$\widehat{\mathfrak{R}}^*_{\tilde{S},L_{u_1}}(\mathcal{F}) \leq \frac{Br}{\sqrt{n}}\left(\frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{\tau_k}}\right),$$

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

Proof. The proof is straightforward based on Lemma 6 by plugging in the specific reweighted values (i.e., $a^+(S_k)$ and $a^-(S_k)$) of the surrogate univariate loss functions. □

B.4.2. The contraction inequality

Lemma 7 (**Contraction inequality for the (reweighted) univariate loss L_u**)**.** *For a dataset S (and its corresponding constructed dataset $\tilde{S}$), assume the loss function $L_\phi = L_u(\mathbf{x}^+,\mathbf{x}^-,f_k) = a^+(S_k)\ell(f_k(\mathbf{x}^+)) + a^-(S_k)\ell(-f_k(\mathbf{x}^-))$ for each $k\in[K]$, where the base loss function $\ell(t)$ is ρ-Lipschitz and the reweighting function $a^+(S_k)$ (or $a^-(S_k)$) indicates its dependency on S_k. Then, the following inequality holds*

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

Proof. Since $\widehat{\mathfrak{R}}^*_{\tilde{S}}(L_\phi\circ\mathcal{F}) = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(L_\phi\circ\mathcal{F}_k)$ and $\widehat{\mathfrak{R}}^*_{\tilde{S}}(\mathcal{F},L_u) = \frac{1}{K}\sum_{k=1}^{K}\widehat{\mathfrak{R}}^*_{\tilde{S},L_u}(\mathcal{F})$, we first prove $\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(L_\phi\circ\mathcal{F}_k) \leq 2\rho\widehat{\mathfrak{R}}^*_{\tilde{S},L_u}(\mathcal{F})$ and then can get the desired result.

Here we prove the inequality $\widehat{\mathfrak{R}}^*_{\tilde{S}_k}(L_\phi\circ\mathcal{F}_k) \leq \rho\widehat{\mathfrak{R}}^*_{\tilde{S}_k,L_u}(\mathcal{F}_k)$ following the idea in Mohri et al. (2018) (Lemma 5.7, p.93), and we omit the index k and the symbol $\tilde{S}_k$ or S for notation simplicity in the following.

First we fix a sample $(\tilde{\mathbf{x}}_1 = (\mathbf{x}_1^+, \mathbf{x}_1^-), \ldots, \tilde{\mathbf{x}}_m = (\mathbf{x}_m^+, \mathbf{x}_m^-))$, then by defintion,

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

where $u_{n_j-1}(f) = \sum_{i=1}^{n_j}\sigma_i L_u(\mathbf{x}_i^+, \mathbf{x}_i^-, f)$. By the definition of the supremum, for any $\epsilon > 0$, there exists $f^1, f^2 \in \mathcal{F}$ such that

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

and

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

Thus, for any $\epsilon > 0$, by definition of $\mathop{\mathbb{E}}_{\sigma_{n_j}}$,

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

Let $s^+ = \operatorname{sgn}(f^1(\mathbf{x}_{n_j}^+) - f^2(\mathbf{x}_{n_j}^+))$ and $s^- = \operatorname{sgn}(f^1(\mathbf{x}_{n_j}^-) - f^2(\mathbf{x}_{n_j}^-))$. Then, the previous inequality implies

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

Since the inequality holds for any $\epsilon > 0$, we have

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

Proceeding in the same way for all other σ_i $(i \in [I_j], i \ne n_j)$ proves that

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

By proceeding other $j \in [J]$, we can obtain the following

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

□

B.4.3. Proof of Theorem 3

Theorem 3 (**Learning guarantee of $\mathcal{A}^{u_1}$ in general case**). *Assume the loss $L_\phi = \frac{1}{\tau_S^*}L_{u_1}$, where L_{u_1} is defined in Eq.(6). Besides,* Assumption 1 *holds. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

Proof. Since the base loss $\ell(z)$ is bounded by B and the loss $L_{u_1}(\mathbf{x}^+, \mathbf{x}^-, f_k) = \frac{|S_k^+|}{n}\ell(f_k(\mathbf{x}^+)) + \frac{|S_k^-|}{n}\ell(-f_k(\mathbf{x}^-))$, the loss $L_\phi = \frac{1}{\tau_S^*}L_{u_1}$ is bounded by $\frac{B}{\tau_S^*}$. Then, applying Theorem 1, and combining Lemma 7 and Proposition 1, we can obtain that for any $\delta > 0$, the following generalization bound holds with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n:

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

Finally, we can get the desired result by combining the above inequality and Lemma 2 (i.e., $R_{0/1}(f) \leq \mathop{\mathbb{E}}_S\left[\frac{1}{\tau_S^*}\widehat{R}_{u_1}(f)\right]$). □

B.4.4. Proof of Theorem 4

Theorem 4 (**Learning guarantee of $\mathcal{A}^{u_2}$ in general case**). *Assume the loss $L_\phi = L_{u_2}$, where L_{u_2} is defined in Eq.(7). Besides,* Assumption 1 *holds. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

Proof. Since the base loss $\ell(z)$ is bounded by B and the loss $L_{u_2}(\mathbf{x}^+, \mathbf{x}^-, f_k) = \ell(f_k(\mathbf{x}^+)) + \ell(-f_k(\mathbf{x}^-))$, the loss $L_\phi = L_{u_2}$ is bounded by $2B$. Then, applying Theorem 1, and combining Lemma 7 and Proposition 1, we can obtain that for any $\delta > 0$, the following generalization bound holds with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n:

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

Finally, we can get the desired result by combining the above inequality and Lemma 2 (i.e., $R_{0/1}(f) \leq R_{u_2}(f)$). □

B.4.5. Proof of Corollary 1

Corollary 1 (Learning guarantee of $\mathcal{A}^{u_1}$ and $\mathcal{A}^{u_2}$ in balanced case). *Assume the loss $L_\phi = 2L_{u_1} = L_{u_2}$, where L_{u_1} and L_{u_2} are defined in Eq.(6) and Eq.(7), respectively. Besides,* Assumption 1 *holds and suppose S is balanced. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

where $\widehat{R}_{u_2}(f) = 2\widehat{R}_{u_1}(f)$.

Proof. It is straightforward to get the result by applying the Theorem 3 and Theorem 4 by plugging $\tau_k = \frac{1}{2}$. □

B.4.6. Proof of Corollary 3

Corollary 3 (Learning guarantee of $\mathcal{A}^{u_1}$ in extremely imbalanced case). *Assume the loss $L_\phi = nL_{u_1}$, where L_{u_1} is defined in Eq.(6). Besides,* Assumption 1 *holds and suppose S is extremely imbalanced. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

Proof. It is straightforward to get the result by applying the Theorem 3 by plugging $\tau_k = \frac{1}{n}$. □

B.4.7. Proof of Corollary 5

Corollary 5 (Learning guarantee of $\mathcal{A}^{u_2}$ in extremely imbalanced case). *Assume the loss $L_\phi = L_{u_2}$, where L_{u_2} is defined in Eq.(7). Besides,* Assumption 1 *holds and suppose S is extremely imbalanced. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size n, the following generalization bound holds for any $f \in \mathcal{F}$:*

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

Proof. It is straightforward to get the result by applying the Theorem 4 and Theorem 4 by plugging $\tau_k = \frac{1}{n}$. □

C. Additional Experimental Results

C.1. Label-wise class imbalance illustrations of benchmark datasets

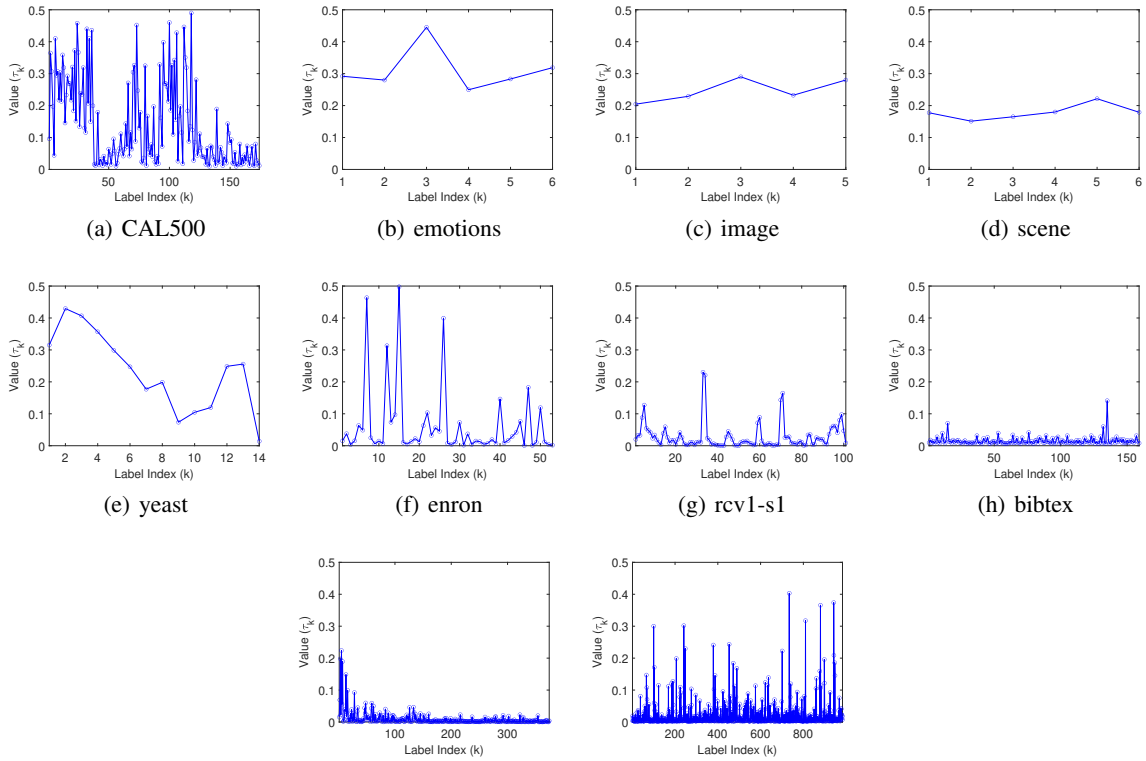
The label-wise class imbalance levels of all the benchmark datasets are illustrated in Figure 2.

C.2. Experimental results about the absolute value of bounds

Here we report the mean upper bound values of three algorithms for the benchmark datasets in Table 4. From the experimental results, we can observe that the absolute values might not reflect the true generalization risk (i.e., bigger than 1), but they can still offer valuable insights into these algorithms by comparing the order of dependent factors under the same framework.

D. Additional Related Work

Recently, there have been some works on the McDiarmid-type concentration inequality for data with graph dependence in quite general settings, e.g., Zhang et al. (2019). However, to the best of our knowledge, it is not clear to apply the conclusion of the existing work (Zhang et al., 2019) (i.e., Theorem 3.6 for the general dependency graph on Page 5) to get the explicit



(a) CAL500 (b) emotions (c) image (d) scene

(e) yeast (f) enron (g) rcv1-s1 (h) bibtex

(i) corel5k (j) delicious

Figure 2. The illustration of the label-wise class imbalance of each benchmark dataset.

Table 4. The mean upper bound values of three algorithms for the benchmark datasets. We set $\delta = 0.01$.

Dataset	$\mathcal{A}^{pa}$	$\mathcal{A}^{u_1}$	$\mathcal{A}^{u_2}$
CAL500	13.0	7082.4	36.3
emotions	13.6	56.8	33.7
image	13.1	191.0	30.7
scene	7.4	146.2	19.6
yeast	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

generalization bound in the Macro-AUC maximization of MLC, as the forest complexity of the dependency graph might be non-trivial to estimate in this case. In contrast, our proposed new McDiarmid-type concentration inequality (i.e., Theorem 5) is easy to apply in this case.

Note that our proposed concentration inequality (i.e., Theorem 5) is not a corollary of the existing one (i.e., Theorem 3.6 in (Zhang et al., 2019)). They are complementary to each other with different assumptions. Although Zhang et al. (2019) is general for the general dependency graph, ours consider the particular case with additional assumptions by constraining the function and the dependency graph (i.e., the assumptions (1) and (2) in Theorem 5), which cannot be induced by Zhang et al. (2019).