Generalization Analysis for Multi-Label Learning

Yi-Fan Zhang [1 2] **Min-Ling Zhang** [2 3]

Abstract

Despite great advances in algorithms for multi-label learning, research on the theoretical analysis of generalization is still in the early stage. Some recent theoretical results has investigated the generalization performance of multi-label learning under several evaluation metrics, however, how to reduce the dependency on the number of labels, explicitly introduce label correlations, and quantitatively analyze the impact of various inductive biases in the generalization analysis of multi-label learning is still a crucial and open problem. In an attempt to make up for the gap in the generalization theory of multi-label learning, we develop several novel vector-contraction inequalities, which exploit the Lipschitz continuity of loss functions, and derive generalization bounds with a weaker dependency on the number of labels than the state of the art in the case of decoupling the relationship among different components, which serves as theoretical guarantees for the generalization of multi-label learning. In addition, we derive the generalization bound for Macro-Averaged AUC and analyze its relationship with class-imbalance. The mild bounds without strong assumptions explain the good generalization ability of multi-label learning with first-order label correlations and high-order label correlations induced by norm regularizers.

1. Introduction

Multi-label learning is one of the most studied and important machine learning paradigms in practice, in which each object is represented by a single instance while being associated with a set of labels instead of a single label. The goal of multi-label learning is to learn a hypothesis which can predict the proper sets of labels for unseen instances. It has made important advances in text categorization (Schapire & Singer, 2000; Rubin et al., 2012), multimedia content annotation (Boutell et al., 2004; Cabral et al., 2011), bioinformatics (Barutçuoglu et al., 2006; Cesa-Bianchi et al., 2012) and other fields (Yu et al., 2005). Although multi-label learning has achieved impressive empirical advances across a wide range of tasks (Zhang & Zhou, 2014), the problem of understanding multi-label learning theoretically remains relatively under-explored.

As we all know, the generalization ability, i.e., the performance of learning machines trained on certain datasets on unseen data, of learning machines is an important question of theoretical research in machine learning, and it is no exception for multi-label learning. Efforts to explain why multi-label models generalize well is an important open problem in multi-label learning community. Uniform convergence is a powerful tool in learning theory for understanding the generalization ability of learners, and it is also used in the generalization analysis of multi-label learning (Yu et al., 2014; Xu et al., 2016; Wu & Zhu, 2020; Wu et al., 2021b;a). However, the progress on the generalization analysis of multi-label learning appears to be severely scarce. A satisfactory and complete study of the generalization analysis for multi-label learning should include three aspects: 1) the reduction of the dependency on the number of labels of the generalization bounds, 2) the explicit introduction of label correlations in the generalization analysis, and 3) the impact of various inductive biases on the generalization performance. First of all, the generalization analysis of multi-label learning is more difficult than that of traditional supervised learning since their difference in problem settings. In particular, the vector-valued output of multi-label learning makes the typical theoretical results not applicable to multi-label learning (Maurer, 2016; Wu & Zhu, 2020). Hence, how to reduce the dependency on the number of labels of the generalization bounds is a very critical problem. Secondly, the consideration of label correlations can often effectively improve the generalization performance of multi-label learning (Zhang & Zhou, 2014), so it is very necessary to explicitly and formally introduce label correlations in the generalization analysis. Finally, the impressive em-

[1]School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China [2]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China [3]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China. Correspondence to: Min-Ling Zhang <zhangml@seu.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

pirical success of multi-label learning algorithms motivates us to further investigate the inductive biases induced by these algorithms, e.g., the inductive bias induced by various norm regularizers (Huang et al., 2016) or label-specific features (Zhang & Wu, 2015; Hang & Zhang, 2022; Jia et al., 2023). Therefore, theoretical research on generalization can promote a better understanding of multi-label learning.

In this paper, we derive novel and tighter bounds based on the Rademacher complexity for multi-label learning. Specifically, for ℓ_2 Lipschitz loss, we improve the basic linear dependent bound to be independent on the number of labels, which decouples the relationship among different components. For ℓ_∞ Lipschitz loss, we improve the square-root dependent bound to be independent on the number of labels, which also decouples the relationship among different components. These bounds are tighter than the state of the art. We also give several tight bounds for the coupling case to study the impact of different types of label correlations on the generalization analysis. Finally, we derive the generalization bound based on the label-based ranking multi-label Rademacher complexity for Macro-Averaged AUC, and analyze the relationship between Macro-Averaged AUC and class-imbalance.

Our generalization bounds reduce the dependency on the number of labels and account for different types of label correlations. Major contributions of the paper include:

- We prove several novel vector-contraction inequalities for the generalization analysis of multi-label learning, which exploits the Lipschitz continuity of the loss function with respect to the ℓ_2 and ℓ_∞ norm and decouples the relationship among different components.
- We derive generalization bounds for general function classes with a weaker dependency on the number of labels than the state of the art, which provides general theoretical guarantees for multi-label learning with different types of label correlations.
- We introduce the label-based ranking multi-label Rademacher complexity and analyze the relationship between Macro-Averaged AUC and class-imbalance according to the generalization bound.

We structure our work as follows. We first introduce an overview of the problem setting for multi-label learning, the definitions of the related evaluation metrics and complexity measures in Section 2. We then present our main results in Sections 3, where we develop several novel vector-contraction inequalities and derive bounds with a weaker dependency on the number of labels than the state of the art for ℓ_2 and ℓ_∞ Lipschitz loss, and study the impact of different types of label correlations. In Section 4, we derive the bound for Macro-Averaged AUC and analyze its relationship with class-imbalance. In Section 5, we provide a comparison of our theoretical results with the related works. Finally, we give a conclusion of our work in Section 6.

2. Preliminaries

In this section, we first present the problem setting for multi-label learning. Secondly, we give the definitions of commonly used surrogate losses. Finally, we introduce the related complexity measures involved in the main results.

2.1. Multi-Label Learning

In the context of multi-label learning, given a dataset $D = \{(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)\}$ with n examples which are identically and independently distributed (i.i.d.) from a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ denotes the d-dimensional input space and $\mathcal{Y} = \{y_1, \ldots, y_c\}$ denotes the label space with c class labels, $\boldsymbol{x}_i \in \mathcal{X}$, $Y_i \subseteq \mathcal{Y}$. Let $[n] := \{1, \ldots, n\}$ for any natural number n and $i \in [n]$.

Let $\mathcal{Y} = \{-1, +1\}^c$, i.e., each $\boldsymbol{y} = (y_1, \ldots, y_c)$ is a binary vector and $y_j = 1$ ($y_j = -1$) denotes that the j-th label is relevant (irrelevant), $j \in [c]$. The task of multi-label learning is to learn a multi-label classifier $\boldsymbol{h} \in \mathcal{H} : \mathcal{X} \mapsto \{-1, +1\}^c$ which assigns each instance with a set of relevant labels. A common strategy is to learn a vector-valued function $\boldsymbol{f} = (f_1, \ldots, f_c) : \mathcal{X} \mapsto \mathbb{R}^c$ and derive the classifier by a thresholding function which dichotomizes the label space into relevant and irrelevant label sets.

We consider the prediction function for each label of the general form $f_j(\boldsymbol{x}) = \langle \boldsymbol{w}_j, \phi(\boldsymbol{x}) \rangle$, where ϕ represents a nonlinear mapping. We define a vector-valued function class of the multi-label learning as follows:

$$\begin{aligned}\mathcal{F} = \{\boldsymbol{x} \mapsto \boldsymbol{f}(\boldsymbol{x}) : &\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_c(\boldsymbol{x})), \\ &f_j(\boldsymbol{x}) = \langle \boldsymbol{w}_j, \phi(\boldsymbol{x}) \rangle, \boldsymbol{x} \in \mathcal{X}, j \in [c] \\ &\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_c) \in \mathbb{R}^{d \times c}, \alpha(\boldsymbol{w}) \leq \Lambda, \\ &\beta(\phi(\boldsymbol{x})) \leq A, \Lambda > 0, A > 0\}, \end{aligned} \tag{1}$$

where α represents a functional that constrains weights, β represents a functional that constrains nonlinear mappings.

For any function $\boldsymbol{f} : \mathcal{X} \mapsto \mathbb{R}^c$, the quality of a prediction on a single example $(\boldsymbol{x}, \boldsymbol{y})$ is measured by a loss function $L : \mathbb{R}^c \times \{-1, +1\}^c \mapsto \mathbb{R}_+$. The goal is to learn a hypothesis $\boldsymbol{f} \in \mathcal{F}$ with good generalization performance from the dataset D by optimizing the loss L. The generalization performance is measured by the expected risk: $R(\boldsymbol{f}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim P}[L(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y})]$. We denote the empirical risk w.r.t. the training dataset D as $\widehat{R}_D(\boldsymbol{f}) = \frac{1}{n}\sum_{i=1}^n L(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{y}_i)$. In addition, we denote the optimal risk as $R^* = \inf_{\boldsymbol{f}\in\mathcal{F}} R(\boldsymbol{f})$ and denote the minimizer of the empirical risk as $\hat{\boldsymbol{f}}^* = \arg\min_{\boldsymbol{f}\in\mathcal{F}} \widehat{R}_D(\boldsymbol{f})$.

The above definitions apply to Hamming loss, Subset loss and Ranking loss.

For Macro-Averaged AUC, it involves the pairwise loss, so we additionally define the corresponding risks. Maximizing Macro-Averaged AUC is equivalent to minimizing the following empirical risk w.r.t. Macro-Averaged AUC:

$$\widehat{R}_D(\boldsymbol{f}) = \frac{1}{c}\sum_{j=1}^{c}\frac{1}{|X_j^+|\,|X_j^-|}\sum_{\boldsymbol{x}_i\in X_j^+}\sum_{\boldsymbol{x}_i'\in X_j^-}\ell_{0/1}\left(f_j(\boldsymbol{x}_i)-f_j(\boldsymbol{x}_i')\right), \tag{2}$$

where $X_j^+ = \{\boldsymbol{x}_i \mid y_j = +1, i \in [n]\}$ ($X_j^- = \{\boldsymbol{x}_i' \mid y_j = -1, i \in [n]\}$) corresponds to the set of test instances that are relevant (irrelevant) to the j-th label. The expected risk w.r.t. Macro-Averaged AUC is defined as $R(\boldsymbol{f}) = \mathbb{E}_D[\widehat{R}_D(\boldsymbol{f})]$.

However, the above mentioned loss is typically the $0-1$ loss, which is hard to handle in practice. Hence, one usually consider its surrogate losses.

2.2. Related Evaluation Metrics

A number of evaluation metrics are proposed to measure the generalization performance of different approaches for multi-label learning. Here we focus on commonly used evaluation metrics, i.e., Hamming loss, Subset loss, Ranking loss and Macro-Averaged AUC, and their surrogate losses are defined as follows:

Hamming Loss: $L_H(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \frac{1}{c}\sum_{j=1}^{c}\ell\left(y_j f_j(\boldsymbol{x})\right)$,

where the base convex surrogate loss ℓ can be various popular forms, such as the hinge, logistic and exponential loss.

Subset Loss: $L_S(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \max_{j\in[c]}\left\{\ell\left(y_j f_j(\boldsymbol{x})\right)\right\}$.

Ranking Loss:

$$L_R(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \frac{1}{|Y^+|\,|Y^-|}\sum_{p\in Y^+}\sum_{q\in Y^-}\ell\left(f_p(\boldsymbol{x})-f_q(\boldsymbol{x})\right),$$

where Y^+ (Y^-) denotes the relevant (irrelevant) label index set induced by $\boldsymbol{y}$, and $|\cdot|$ denotes the cardinality of a set.

The surrogate loss for **Macro-Averaged AUC**:

$$L_M(\boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{x}_i'), \boldsymbol{y}) = \frac{1}{c}\sum_{j=1}^{c}\ell\left(f_j(\boldsymbol{x}_i)-f_j(\boldsymbol{x}_i')\right), \tag{3}$$

where $\boldsymbol{x}_i$ ($\boldsymbol{x}_i'$) corresponds to the instances that are relevant (irrelevant) to the j-th label.

2.3. Related Complexity Measures

Here we use the Rademacher complexity to perform generalization analysis for multi-label learning.

Definition 2.1 (Rademacher complexity)**.** Let $\mathcal{G}$ be a class of real-valued functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Let $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set with n i.i.d. samples. The empirical **Rademacher complexity** over $\mathcal{G}$ is defined by

$$\hat{\mathfrak{R}}_D(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g\left(\boldsymbol{x}_i\right)\right],$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables. In addition, we define the worst-case Rademacher complexity as $\tilde{\mathfrak{R}}_n(\mathcal{G}) = \sup_{D\in\mathcal{X}^n}\hat{\mathfrak{R}}_D(\mathcal{G})$.

In multi-label learning, $\mathcal{F}$ is a class of vector-valued functions, which makes traditional Rademacher complexity analysis methods invalid. A common practice is to use the multi-label Rademacher complexity to bound the Rademacher complexity of a loss function space associated with the vector-valued function class $\mathcal{F}$ according to the vector-contraction inequality in (Maurer, 2016).

Definition 2.2 (Multi-label Rademacher complexity)**.** Let $\mathcal{F}$ be a class of vector-valued functions mapping from $\mathcal{X}$ to $\mathbb{R}^c$. Let $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set with n i.i.d. samples. The empirical **multi-label Rademacher complexity** over $\mathcal{F}$ is defined by

$$\hat{\mathfrak{R}}_D(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\epsilon_{ij} f_j\left(\boldsymbol{x}_i\right)\right],$$

where each ϵ_{ij} is an independent doubly indexed Rademacher random variable, and $f_j\left(\boldsymbol{x}_i\right)$ is the j-th component of $\boldsymbol{f}\left(\boldsymbol{x}_i\right)$.

Here we use the covering number to bound the Rademacher complexity for multi-label learning. The covering number can be bounded by the fat-shattering dimension (Srebro et al., 2010; Lei et al., 2019; Zhang & Zhang, 2023):

Definition 2.3 (Covering number)**.** Let $\mathcal{F}$ be a class of real-valued functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Let $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set with n i.i.d. samples. For any $\epsilon > 0$, the empirical ℓ_2 (or ℓ_∞) norm covering number $\mathcal{N}_2(\epsilon, \mathcal{F}, D)$ (or $\mathcal{N}_\infty(\epsilon, \mathcal{F}, D)$) w.r.t. D is defined as the minimal number m of a collection of vectors $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m \in \mathbb{R}^n$ such that ($\boldsymbol{v}_i^j$ is the i-th component of the vector $\boldsymbol{v}^j$)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f\left(\boldsymbol{x}_i\right)-\boldsymbol{v}_i^j)^2} \le \epsilon. \left(\text{or } \max_{i=1,\ldots,n}\left|f\left(\boldsymbol{x}_i\right)-\boldsymbol{v}_i^j\right| \le \epsilon\right)$$

In this case, we call $\left\{\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m\right\}$ an (ϵ, ℓ_2)-cover (or (ϵ, ℓ_∞)-cover) of $\mathcal{F}$ with respect to D. We also denote $\mathcal{N}_2(\epsilon, \mathcal{F}, n) = \sup_D \mathcal{N}_2(\epsilon, \mathcal{F}, D)$ (or $\mathcal{N}_\infty(\epsilon, \mathcal{F}, n) = \sup_D \mathcal{N}_\infty(\epsilon, \mathcal{F}, D)$).

Definition 2.4 (Fat-shattering dimension)**.** Let $\mathcal{F}$ be a class of real-valued functions mapping from $\mathcal{X}$ to $\mathbb{R}$. We

define the fat-shattering dimension $\text{fat}_\epsilon(\mathcal{F})$ at scale $\epsilon > 0$ as the largest $p \in \mathbb{N}$ such that there exist p points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \in \mathcal{X}$ and witnesses $s_1, \ldots, s_p \in \mathbb{R}$ satisfying: for any $\delta_1, \ldots, \delta_p \in \{-1, +1\}$ there exists $f \in \mathcal{F}$ with

$$\delta_i \left(f(\boldsymbol{x}_i) - s_i\right) \geq \frac{\epsilon}{2}, \quad \forall i = 1, \ldots, p.$$

3. Generalization Bounds Based on the Rademacher Complexity

In this section, we derive several novel bounds under the assumption that the loss is Lipschitz continuous w.r.t. ℓ_2 norm or ℓ_∞ norm. For ℓ_2 Lipschitz loss, we first introduce a basic bound with a linear dependency on c, then we develop a novel vector-contraction inequality and improve the bound with no dependency on c, which is tighter than the state of the art. For ℓ_∞ Lipschitz loss, we first introduce a basic bound with a square-root dependency on c, then we develop a novel vector-contraction inequality and improve the bound with no dependency on c, which is tighter than the state of the art in the decoupling case. We also give several tight bounds for the coupling case to study the impact of different types of label correlations on the generalization analysis. The theoretical results in this section hold for surrogate Hamming loss, surrogate Subset loss and surrogate Ranking loss. The detailed proofs of the theoretical results in this paper are provided in the appendix.

3.1. Generalization Bounds for ℓ_2 Lipschitz Loss

We first introduce the assumptions used, and we show that for general function classes, Lipschitz continuity of the loss function w.r.t. the ℓ_2 norm combined with the multi-label Rademacher complexity yields basic generalization bounds with a linear dependency on the number of labels, which exploits the typical vector-contraction inequality (Maurer, 2016). Second, we develop a novel vector-contraction inequality and derive a tighter bound with no dependency on the number of labels for ℓ_2 Lipschitz loss.

Assumption 3.1. Assume that the loss function and the components of the vector-valued function are bounded: $L(\cdot, \cdot) \leq M$, $|f_j(\cdot)| \leq B$ for $j \in [c]$ where $M > 0$ and $B > 0$ are constants.

Assumption 3.2. Assume that the loss function is μ-Lipschitz continuous w.r.t. the ℓ_2 norm.

Assumption 3.1 is a relatively mild assumption. In fact, when we consider the function class (1) for multi-label learning, we often use the assumptions $\|\boldsymbol{w}_j\|_2 \leq \Lambda$, ϕ_j is ρ-Lipschitz w.r.t. the ℓ_2 norm and $\|\boldsymbol{x}_i\|_2 \leq A$ for any $j \in [c]$, $i \in [n]$ to replace the boundedness of the components of the vector-valued function. The following Proposition 3.3 also illustrates that Assumption 3.2 is very mild.

Proposition 3.3 (Lemma 1 in (Wu & Zhu, 2020))**.** *Assume that the base loss function ℓ defined in Subsection 2.2 is μ-Lipschitz continuous, then the surrogate Hamming Loss is $\frac{\mu}{\sqrt{c}}$-Lipschitz w.r.t. the ℓ_2 norm, the surrogate Subset Loss is μ-Lipschitz w.r.t. the ℓ_2 norm, and the surrogate Ranking Loss is μ-Lipschitz w.r.t. the ℓ_2 norm.*

3.1.1. A Basic Bound for ℓ_2 Lipschitz Loss

Using the ℓ_2 Lipschitz continuity of loss and the multi-label Rademacher complexity, we have the following theorem:

Theorem 3.4. *Let $\mathcal{F}$ be a vector-valued function class of the multi-label learning defined by (1). Let Assumptions 3.1 and 3.2 hold. Given a dataset D of size n. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, the following holds for any $\boldsymbol{f} \in \mathcal{F}$:*

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + \frac{2\sqrt{2}\mu c B}{\sqrt{n}} + 3M\sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

Proof Sketch. We first use the multi-label Rademacher complexity to bound the Rademacher complexity of the loss function space associated with the vector-valued function class $\mathcal{F}$ according to the typical vector-contraction inequality (Maurer, 2016), and then complete the proof with the McDiarmid's inequality and the symmetrization technique. □

Remark 3.5. The above bound with a linear dependency on the number of labels indicates that good generalization performance will be obtained when the number of examples ($\sqrt{n}$) is larger than the number of labels (c), but in practice, it is often encountered in the case of an extremely large number of labels, that is, extreme multi-label learning (Yu et al., 2014; Prabhu & Varma, 2014; Yen et al., 2016; Liu & Shen, 2019). At this time, the number of labels will probably be more than the number of examples, so the bound in Theorem 3.4 will not be able to provide theoretical guarantees, thus prompting us to develop the bound that is tighter on the number of labels. Our analysis in Theorem 3.4 implies the following inequality: $\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\epsilon_{ij} f_j(\boldsymbol{x}_i)\right] \leq c \max_j \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f_j\in\mathcal{F}_j} \frac{1}{n}\sum_{i=1}^{n}\epsilon_{ij} f_j(\boldsymbol{x}_i)\right]$, which shows that decoupling the relationship among different components (since the maximization over $j \in [c]$ is outside of the expectation operator) will lead to bounds with a linear dependency on c for general function classes. When considering kernel function classes, the dependency of the bounds on c can be improved to square-root (Maurer, 2016; Wu & Zhu, 2020; Wu et al., 2021a). Such improvements essentially come from preserving the coupling among different components reflected by the constraint, i.e., $\|\boldsymbol{w}\| \leq \Lambda$. As a comparison, when $\|\boldsymbol{w}_j\|_2 \leq \Lambda$ for any $j \in [c]$, if we consider the group norm $\|\cdot\|_{2,2}$, we have $\|\boldsymbol{w}\|_{2,2} \leq \sqrt{c}\Lambda$,

which means that these improved bounds still suffer from a linear dependency on the number of labels. Hence, the improvement in the preservation of coupling by a factor of $\sqrt{c}$ benefits from replacing Λ with $\sqrt{c}\Lambda$ in the constraint to some extent. This reveals that in order to improve the existing bounds, we need to improve the dependency on the number of labels in the decoupling case.

3.1.2. Tighter Bounds for ℓ_2 Lipschitz Loss

We develop a novel vector-contraction inequality for ℓ_2 Lipschitz loss, which decouples the relationship among different components and guarantees that the derived generalization bounds are tighter than the state of the art.

We show that the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded by the worst-case Rademacher complexity of the projection function class $\mathcal{P}(\mathcal{F})$. We first define a function class $\mathcal{P}$ consisting of projection operators $p_j : \mathbb{R}^c \mapsto \mathbb{R}$ for any $j \in [c]$ which project the c-dimensional vector onto the j-th coordinate. Then, we have $\mathcal{P}(\mathcal{F}) = \{(j, \boldsymbol{x}) \mapsto p_j(\boldsymbol{f}(\boldsymbol{x})) : p_j(\boldsymbol{f}(\boldsymbol{x})) = f_j(\boldsymbol{x}), \boldsymbol{f} \in \mathcal{F}, (j, \boldsymbol{x}) \in [c] \times \mathcal{X}\}$. With the above definitions, we develop the following vector-contraction inequality:

Lemma 3.6. *Let $\mathcal{F}$ be a vector-valued function class of the multi-label learning defined by (1). Let Assumptions 3.1 and 3.2 hold. Given a dataset D of size n. Then, we have*

$$\hat{\mathfrak{R}}_D(\mathcal{F}) \leq 48\mu\sqrt{c}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\left(1 + \log^{\frac{1}{2}}(nc) \cdot \log\frac{M\sqrt{n}}{\mu B}\right),$$

where $\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))$ is the worst-case Rademacher complexity of the projection function class.

Proof Sketch. First, the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded by the empirical ℓ_2 norm covering number with the refined Dudley's entropy integral inequality. Second, according to the Lipschitz continuity w.r.t the ℓ_2 norm, the empirical ℓ_2 norm covering number of $\mathcal{F}$ can be bounded by the empirical ℓ_2 norm covering number of $\mathcal{P}(\mathcal{F})$. Third, the empirical ℓ_2 norm covering number of $\mathcal{P}(\mathcal{F})$ can be bounded by using Sudakov's minoration (Wainwright, 2019), which bounds the ℓ_2 norm covering number of a function class by the expectation of a Gaussian process indexed by the function class, and the expectation of the Gaussian process can be bounded by the worst-case Rademacher complexity of the projection function class $\mathcal{P}(\mathcal{F})$. Hence, the problem is transferred to the estimation of the worst-case Rademacher complexity. Finally, we estimate the lower bound of the worst-case Rademacher complexity of $\mathcal{P}(\mathcal{F})$, and then combined with the above steps, the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded. □

With the vector-contraction inequality above, we can derive the following tight bound for ℓ_2 Lipschitz loss:

Theorem 3.7. *Let $\mathcal{F}$ be a vector-valued function class of the multi-label learning defined by (1). Let Assumptions 3.2 and 3.1 hold. Given a dataset D of size n. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, the following holds for any $\boldsymbol{f} \in \mathcal{F}$:*

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{96B\mu\left(1 + \log^{\frac{1}{2}}(nc) \cdot \log\frac{M\sqrt{n}}{\mu B}\right)}{\sqrt{n}}.$$

Proof Sketch. We first upper bound the worst-case Rademacher complexity $\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))$, and then combined with Lemma 3.6, the desired bound can be derived. □

Remark 3.8. Although Lemma 3.6 shows a factor of $\sqrt{c}$, the term $\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F})) \leq \frac{B}{\sqrt{nc}}$, which makes the Rademacher complexity of the loss function space associated with $\mathcal{F}$ (i.e., $\hat{\mathfrak{R}}_D(\mathcal{F})$) actually independent on c, and results in a tighter bound than the $O(c/\sqrt{n})$ bound in Theorem 3.4 with a faster convergence rate $\widetilde{O}(1/\sqrt{n})$. Lemma 3.6 decouples the relationship among different components (since the supremum over $j \in [c]$ is outside of the expectation operator by the definition of $\mathcal{P}(\mathcal{F})$). Hence, the bound with no dependency on c in Theorem 3.7 is clearly tighter than the state of the art with square-root dependency on c in (Lei et al., 2015; Maurer, 2016; Wu & Zhu, 2020; Wu et al., 2021a) for ℓ_2 Lipschitz loss assumption, where the analyzes all preserve the coupling among different components, not to mention, as discussed in Remark 3.5, the constraint on preserving the coupling ($\|\boldsymbol{w}\| \leq \Lambda$) directly implies the improvement by a factor of $\sqrt{c}$. In fact, decoupling the relationship or preserving the coupling among different components corresponds to different types of label correlations in multi-label learning (Zhang & Zhou, 2014). The former corresponds to first-order label correlations (which tackle multi-label learning problem by decomposing it into a number of independent binary classification problems, i.e., ignorance of label correlations), and the latter corresponds to high-order label correlations (which tackle multi-label learning problem by exploiting high-order relationships among labels) induced by norm regularizers. The assumption of the coupling case that holds for many learning scenarios (e.g., multi-class learning) does not hold in multi-label learning methods with first-order label correlations. This means that we need to develop new vector-contraction inequalities to handle the assumption of the decoupling case. The projection function class is used to help handle the generalization analysis in the decoupling case. The above bounds provide theoretical guarantees for first-order label correlations methods, e.g., Binary Relevance methods (Boutell et al., 2004; Zhang & Zhou, 2014; Zhang & Wu, 2015; Hang & Zhang, 2022). Furthermore, according to the Proposition 3.3, we

can obtain a $\widetilde{O}(1/\sqrt{nc})$ bound for ℓ_2 Lipschitz surrogate Hamming Loss in the decoupling case.

3.2. Generalization Bounds for ℓ_∞ Lipschitz Loss

We first introduce the assumption of Lipschitz continuity w.r.t. the ℓ_∞ norm, and we derive a basic bound with a square-root dependency on c for general function classes by refining Theorem 1 in (Foster & Rakhlin, 2019). Second, we develop a novel vector-contraction inequality and derive a tighter bound with no dependency on c for ℓ_∞ Lipschitz loss in the decoupling case, up to logarithmic terms, and we also give several bounds with no dependency on c for ℓ_∞ Lipschitz loss in the coupling case.

Assumption 3.9. Assume that the loss function L is μ-Lipschitz continuous w.r.t. the ℓ_∞ norm, that is:

$$\left|L(\boldsymbol{f}(\boldsymbol{x}),\cdot)-L(\boldsymbol{f}'(\boldsymbol{x}),\cdot)\right|\leq\mu\left\|\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{f}'(\boldsymbol{x})\right\|_\infty,$$

where $\mu>0$, $\|\boldsymbol{t}\|_\infty=\max_{j\in[c]}|t_j|$ for $\boldsymbol{t}=(t_1,\ldots,t_c)$.

In fact, the commonly used loss functions in multi-label learning actually satisfy the Lipschitz continuity w.r.t. the ℓ_∞ norm, and it has been considered in some literature (Lei et al., 2019; Wu et al., 2021b). The following Proposition 3.10 further illustrates that Assumption 3.9 is very mild.

Proposition 3.10. *Assume that the base loss ℓ defined in Subsection 2.2 is μ-Lipschitz continuous, then the surrogate Hamming Loss is μ-Lipschitz w.r.t. the ℓ_∞ norm, the surrogate Subset Loss is μ-Lipschitz w.r.t. the ℓ_∞ norm, and the surrogate Ranking Loss is 2μ-Lipschitz w.r.t. the ℓ_∞ norm.*

3.2.1. A Basic Bound for ℓ_∞ Lipschitz Loss

Using the Lipschitz continuity w.r.t. the ℓ_∞ norm, we first show that the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded by the worst-case Rademacher complexity of the restriction of the function class along each coordinate with timing a factor of $\widetilde{O}(\sqrt{c})$. Then, we derive the basic bound with a square-root dependency on the number of labels in the decoupling case.

Lemma 3.11. *Let $\mathcal{F}$ be a vector-valued function class of the multi-label learning defined by (1). Let Assumptions 3.1 and 3.9 hold. Given a dataset D of size n. Then, for any $0<\eta<1$, $0<a<1$, we have*

$$\hat{\mathfrak{R}}_D(\mathcal{F})\leq\frac{48\sqrt{2E}\mu}{a}\sqrt{c}\max_j\widetilde{\mathfrak{R}}_n\left(\mathcal{F}_j\right)\times\left(1+\log^{\frac{1+\eta}{2}}(\frac{2\sqrt{2n}}{B})\cdot\log(\frac{M}{\sqrt{E}\mu B}\sqrt{\frac{n}{c}})\right),$$

where $\widetilde{\mathfrak{R}}_n\left(\mathcal{F}_j\right)$ is the worst-case Rademacher complexity, $\mathcal{F}_j$ is the restriction of the function class along the j-th coordinate, $f_j\in\mathcal{F}_j$ and $E>0$ is an absolute constant.

Proof Sketch. We obtain the lower bound of the worst-case Rademacher complexity $\widetilde{\mathfrak{R}}_n\left(\mathcal{F}_j\right)$ through the Khintchine-Kahane inequality, combined with the refined Theorem 1 in (Foster & Rakhlin, 2019), the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded. □

With the vector-contraction inequality above, we can derive the following basic bound for ℓ_∞ Lipschitz loss:

Theorem 3.12. *Let $\mathcal{F}$ be a vector-valued function class of the multi-label learning defined by (1). Let Assumptions 3.1 and 3.9 hold. Given a dataset D of size n. Then, for any $0<\delta,\eta,a<1$, there exists an absolute constant $E>0$ such that with probability at least $1-\delta$, the following holds for any $\boldsymbol{f}\in\mathcal{F}$:*

$$R(\boldsymbol{f})\leq\widehat{R}_D(\boldsymbol{f})+3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\frac{96\sqrt{2E}\mu B}{a}\sqrt{\frac{c}{n}}\times\left(1+\log^{\frac{1+\eta}{2}}(\frac{2\sqrt{2n}}{B})\cdot\log(\frac{M}{\sqrt{E}\mu B}\sqrt{\frac{n}{c}})\right).$$

Proof Sketch. We first upper bound the worst-case Rademacher complexity $\widetilde{\mathfrak{R}}_n\left(\mathcal{F}_j\right)$, and then combined with Lemma 3.11, the desired bound can be derived. □

Remark 3.13. Lemma 3.11 is not a direct application of Theorem 1 in (Foster & Rakhlin, 2019) which absorbs all terms independent of c and n into an unspecified numerical constant and yields a logarithmic term of order $O(\log^{\frac{3+\eta}{2}}(\sqrt{n}))$. We refine the proof of Theorem 1 in (Foster & Rakhlin, 2019), bound the relevant constant terms and imporve the order of the logarithmic term to $O(\log^{\frac{1+\eta}{2}}(\sqrt{n})\cdot\log(\sqrt{\frac{n}{c}}))$. The term $\max_j\widetilde{\mathfrak{R}}_n\left(\mathcal{F}_j\right)\leq\frac{B}{\sqrt{n}}$ means that Lemma 3.11 decouples the relationship among different components, which results in a tighter bound than the $O(c/\sqrt{n})$ bound in Theorem 3.4 with a faster convergence rate $\widetilde{O}(\sqrt{c/n})$.

3.2.2. Tighter Bounds for ℓ_∞ Lipschitz Loss

We first give a corollary with no dependency on c, which preserves the coupling among different components.

Corollary 3.14. *Let $\mathcal{F}$ be a vector-valued function class of the multi-label learning defined by (1). Let Assumptions 3.1 and 3.9 hold. Given a dataset D of size n. Then, for any $0<\delta<1$, with probability at least $1-\delta$, the following holds for any $\boldsymbol{f}\in\mathcal{F}$:*

***1.)** If $\alpha(\boldsymbol{w}):=\|\boldsymbol{w}\|$, $\beta(\phi(\boldsymbol{x})):=\sup_{\boldsymbol{x}\in\mathcal{X},j\in[c]}\|\tilde{\boldsymbol{\phi}}_j(\boldsymbol{x})\|_*$, and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, we have:*

$$R(\boldsymbol{f})\leq\widehat{R}_D(\boldsymbol{f})+\frac{36\mu\Lambda A\log_2^{\frac{3}{2}}(\sqrt{2}n^{\frac{3}{2}}c)}{\sqrt{n}}+3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

2.) *If* $\alpha(\boldsymbol{w}) := \|\boldsymbol{w}\|_{S_p}$, $\beta(\phi(\boldsymbol{x})) := \max_{i\in[n]} \|\phi(\boldsymbol{x}_i)\|_2$, *when* $1 \leq p \leq 2$, *we have:*

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + \frac{36\mu\Lambda A \log_2^{\frac{3}{2}}(\sqrt{2}n^{\frac{3}{2}}c)}{\sqrt{n}} + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}},$$

when $p > 2$, *we have:*

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{36\mu\Lambda A \min\{c,d\}^{\frac{1}{2}-\frac{1}{p}} \log_2^{\frac{3}{2}}(\sqrt{2}n^{\frac{3}{2}}c)}{\sqrt{n}}.$$

The Schatten-p *norm is defined as the* ℓ_p *norm of the singular value vector of a matrix, i.e.,* $\|\boldsymbol{w}\|_{S_p} = \|\sigma(\boldsymbol{w})\|_p$, *where the singular values are sorted in non-increasing order. For any* $\boldsymbol{x} \in \mathcal{X}$ *and* $j \in [c]$, *the notation* $\tilde{\boldsymbol{\phi}}_j(\boldsymbol{x})$ *is defined by* $\tilde{\boldsymbol{\phi}}_j(\boldsymbol{x}) := (\underbrace{0,\ldots,0}_{j-1}, \phi(\boldsymbol{x}), \underbrace{0,\ldots,0}_{c-j}) \in \mathbb{R}^{d\times c}$.

Proof Sketch. We first show that the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded by the worst-case Rademacher complexity of $\widetilde{\mathcal{F}}$, which refines Theorem 5 in (Lei et al., 2019) (i.e., $\hat{\mathfrak{R}}_D(\mathcal{F}) \leq 16\sqrt{\log 2}\mu\sqrt{c}\widetilde{\mathfrak{R}}_{nc}(\widetilde{\mathcal{F}})(1 + \log_2^{\frac{3}{2}} \frac{\Lambda A n\sqrt{c}}{\widetilde{\mathfrak{R}}_{nc}(\widetilde{\mathcal{F}})})$), where $\widetilde{\mathcal{F}} := \left\{\boldsymbol{v} \mapsto \langle \boldsymbol{w}, \boldsymbol{v}\rangle : \boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^{d\times c}, \alpha(\boldsymbol{w}) \leq \Lambda, \beta(\boldsymbol{v}) \leq A, \boldsymbol{v} \in \widetilde{S}\right\}$ and $\widetilde{S} = \{\tilde{\boldsymbol{\phi}}_j(\boldsymbol{x}_i) : j \in [c], i \in [n]\}$. Then, we upper bound $\widetilde{\mathfrak{R}}_{nc}(\widetilde{\mathcal{F}})$ for different norm regularizers. Combining these results, the desired bounds can be derived. □

Remark 3.15. Corollary 3.14 for preserving the coupling case has less novelty, which has the same order $\widetilde{O}(1/\sqrt{n})$ as the results in (Lei et al., 2019; Wu et al., 2021b), since these results all use the same vector-contraction inequality, i.e., Theorem 5 in (Lei et al., 2019). However, here we are more concerned with investigating the impact of the label correlation induced by the norm regularizer on the generalization analysis. The bounds here involve a constraint on the overall weight $\boldsymbol{w}$, which consider that the components share some constraint properties with each other, that is, the label correlation induced by the norm regularizer. The above bounds involve label correlations induced by the general norm regularizer (case 1) and the Schatten-p norm regularizer (case 2), respectively. Trace norm regularizer, corresponding to Schatten-p norm regularizer with $p = 1$, is a common practice to consider the label correlation in multi-label learning, which imposes a low-rank constraint on the spectrum of $\boldsymbol{w}$. The bounds here also explain the good generalization ability of multi-label learning with the label correlation induced by the trace norm regularizer.

Then, we develop a novel vector-contraction inequality, which guarantees that the derived bounds are tighter than the state of the art in the decoupling case.

Lemma 3.16. *Let* $\mathcal{F}$ *be a vector-valued function class of the multi-label learning defined by (1). Let Assumptions 3.1 and 3.9 hold. Given a dataset* D *of size* n. *Then, for any* $0 < \eta < 1$, $0 < a < 1$, $E > 0$, *we have*

$$\hat{\mathfrak{R}}_D(\mathcal{F}) \leq \frac{96\sqrt{E}\mu}{a}\sqrt{c}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\times \left(1 + \log^{\frac{1+\eta}{2}}(\frac{8}{5B}\sqrt{nc})\cdot\log(\frac{M\sqrt{n}}{\sqrt{E}\mu B})\right),$$

where $\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))$ *is the worst-case Rademacher complexity of the projection function class.*

Proof Sketch. The overall proof idea is similar to Lemma 3.6, but there are two obvious differences. First of all, the covering numbers involved in the proof here are all ℓ_∞ norm covering numbers instead of ℓ_2 norm covering numbers, which makes the proof techniques involved completely different. Secondly, in the third step, instead of using Sudakov's minoration inequality, we use the fat-shattering dimension to bound the empirical ℓ_∞ norm covering number of $\mathcal{P}(\mathcal{F})$, and the fat-shattering dimension can be bounded by the worst-case Rademacher complexity of $\mathcal{P}(\mathcal{F})$. □

Theorem 3.17. *Let* $\mathcal{F}$ *be a vector-valued function class of the multi-label learning defined by (1). Let Assumptions 3.1 and 3.9 hold. Given a dataset* D *of size* n. *Then, for any* $0 < \delta, \eta, a < 1$, *there exists an absolute constant* $E > 0$ *such that with probability at least* $1-\delta$, *the following holds for any* $\boldsymbol{f} \in \mathcal{F}$:

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{192\sqrt{E}\mu B}{a}\frac{1}{\sqrt{n}}\times \left(1 + \log^{\frac{1+\eta}{2}}(\frac{8}{5B}\sqrt{nc})\cdot\log(\frac{M\sqrt{n}}{\sqrt{E}\mu B})\right).$$

Proof Sketch. We first upper bound $\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))$, then combined with Lemma 3.16, the bound is immediate. □

Remark 3.18. Lemma 3.16 decouples the relationship among different components, and the term $\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F})) \leq \frac{B}{\sqrt{nc}}$, which makes the Rademacher complexity $\hat{\mathfrak{R}}_D(\mathcal{F})$ actually independent on c, and results in a tighter bound than the $\widetilde{O}(\sqrt{c/n})$ bound in Theorem 3.12 with a faster convergence rate $\widetilde{O}(1/\sqrt{n})$. How to develop novel vector-contraction inequalities that can induce $\widetilde{O}(1/\sqrt{n})$ bounds and deal with the assumption of the decoupling case are the two most critical difficulties in deriving tighter bounds. The introduction of the projection function class plays an important role in solving these two difficulties. It improves the vector-contraction inequalities by a factor of $\sqrt{c}$ and handles the assumption of the decoupling case indirectly. Theorem 3.17 improves the bounds from $\widetilde{O}(\sqrt{c/n})$ to $\widetilde{O}(1/\sqrt{n})$ for

ℓ_∞ Lipschitz Loss in the decoupling case, and explains the good generalization ability of multi-label learning with first-order label correlations.

4. Generalization Analysis for Macro-Averaged AUC

Macro-Averaged AUC is a typical label-based ranking metric (Zhang & Zhou, 2014). Here we analyze the relationship between Macro-Averaged AUC and class-imbalance according to the generalization bound. We use Lemma 3.11 for the generalization analysis, since improving the dependency on the number of labels is not our main concern here.

Rademacher complexity has proved to be a powerful data-dependent measure of hypothesis space complexity (Bartlett & Mendelson, 2002; Koltchinskii & Panchenko, 2002). However, a sequence of pairs of i.i.d. individual observation in (2) is no longer independent, which makes standard techniques in the i.i.d case for traditional Rademacher complexity inapplicable for Macro-Averaged AUC. We convert the non-sum-of-i.i.d pairwise function to a sum-of-i.i.d form by using permutations in U-process (Clémençon et al., 2008). We denote $\left|X_j^+\right|$ and $\left|X_j^-\right|$ in (2) as s_j and t_j, $r_j = \min\{s_j, t_j\}$, $r_0 = \min_{j\in[c]}\{r_j\}$, and $s_j + t_j = n$ for any $j \in [c]$. Then, we have the following definition:

Definition 4.1. Let $\mathcal{F}$ be a class of vector-valued functions mapping from $\mathcal{X}$ to $\mathbb{R}^c$ and ℓ be the base loss defined in (3). Let $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set with n i.i.d. samples. The empirical **label-based ranking multi-label Rademacher complexity** of a loss function space associated with the vector-valued function class $\mathcal{F}$ is defined by

$$\hat{\mathfrak{R}}_D(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}} \frac{1}{c}\sum_{j=1}^{c}\frac{1}{r_j}\sum_{i=1}^{r_j}\epsilon_{ij}\ell\left(f_j(\boldsymbol{x}_i) - f_j(\boldsymbol{x}_i')\right)\right],$$

where each ϵ_{ij} is an independent doubly indexed Rademacher random variable, and $f_j(\boldsymbol{x}_i)$ is the j-th component of $\boldsymbol{f}(\boldsymbol{x}_i)$. We refer to the expectation $\mathfrak{R}(\mathcal{F}) = \mathbb{E}_D[\hat{\mathfrak{R}}_D(\mathcal{F})]$ as the label-based ranking multi-label Rademacher complexity of $\mathcal{F}$.

With this definition, we then derive the bound as follows:

Theorem 4.2. *Let $\mathcal{F}$ be a vector-valued function class of the multi-label learning defined by (1) and the loss be Macro-Averaged AUC. Let Assumptions 3.1 and 3.9 hold. Given a dataset D of size n. Then, for any $0 < \delta < 1$, with probability at least $1-\delta$, the following holds for any $\boldsymbol{f} \in \mathcal{F}$:*

$$R(\hat{\boldsymbol{f}}^*) - R^* = \widetilde{O}(\sqrt{c/r_0}).$$

Proof Sketch. First, by using the U-process technique, we bound $\mathbb{E}[R(\hat{\boldsymbol{f}}^*)] - R^*$ with the label-based ranking multi-label Rademacher complexity. Second, combining with Lemma 3.11 and bounding the worst-case Rademacher complexity, we can upper bound the label-based ranking multi-label Rademacher complexity. Finally, with the McDiarmid's inequality, the desired bound can be derived. □

Remark 4.3. Theorem 4.2 shows that when class-imbalance occurs, r_0 will be smaller than $\frac{n}{2}$. When class-imbalance is more serious, r_0 will be smaller, which will lead to a looser bound for Macro-Averaged AUC. This means that when class-imbalance becomes more and more serious, if the learned classifier cannot handle the problem of class imbalance well, then its performance on Macro-Averaged AUC will be worse. Wu et al. (2023) also obtained similar conclusions, but the methods used were completely different. Wu et al. (2023) transformed the macro-averaged maximization problem in multi-label learning into the problem of learning multiple tasks with graph-dependent examples-which is hard to verify in practice, then proposed a new McDiarmid-type inequality to develop $O(\frac{1}{\sqrt{n}})$ bound in the balanced case. Our method is simpler and can also yield bounds with no dependency on c by combining Lemma 3.16.

5. Comparison with Related Work

The generalization analysis of multi-label learning originated from (Dembczynski et al., 2010), which performed regret analysis on Hamming and Subset loss, and derived the relationship between the expectations of Hamming and Subset loss. Dembczynski et al. (2012) performed regret analysis on Ranking loss. These analyses laid the foundation for research in (Wu & Zhu, 2020; Wu et al., 2021a).

Theorem 3.4 shows that when using the typical vector-contraction inequality (Maurer, 2016) (i.e., ℓ_2 norm Lipschitz loss) and the multi-label Rademacher complexity, one can only derive bounds of order $O(c/\sqrt{n})$ for general function classes in the decoupling case. Wu & Zhu (2020); Wu et al. (2021a) also obtained similar results for surrogate Hamming, Subset and Ranking losses, which mainly exploited the relationship between losses. Wu & Zhu (2020); Wu et al. (2021a) also showed that for kernel function classes, the order of the bounds for some losses can be improved to $O(\sqrt{c/n})$ when preserving the coupling. Lei et al. (2015) first derived a $O(\sqrt{c/n})$ bound for multi-class SVM with ℓ_p norm regularized kernel function classes under the assumption of ℓ_2 Lipschitz loss. Li et al. (2018) used the local Rademacher complexity to derive a $O(\log^2 c/n)$ bound for multi-class classification with ℓ_p norm regularized kernel function classes under the assumption of ℓ_2 Lipschitz and smooth loss. Theorem 3.7 improves the bounds to $O(1/\sqrt{n})$ for ℓ_2 Lipschitz loss even in the decoupling case.

Theorem 3.12 shows that when using ℓ_∞ norm Lipschitz loss, one can derive bounds of order $O(\sqrt{c/n})$ for general

function classes. Liu et al. (2018) also obtained a bound of order $O(\sqrt{c/n})$ for the dual set multi-label learning for margin loss and kernel function classes. Lei et al. (2019); Wu et al. (2021b) improved the dependency on c in the coupling case, where Lei et al. (2019) derived a $\widetilde{O}(1/\sqrt{n})$ bound for multi-class classification with norm regularized function classes under ℓ_∞ Lipschitz loss, and Wu et al. (2021b) derived a $O(\log^3(nc)/n\sigma)$ bound for vector-valued learning with norm regularized kernel function classes under the assumption of ℓ_∞ Lipschitz and σ-strongly convex loss. Here Theorem 3.17 shows a $\widetilde{O}(1/\sqrt{n})$ bound for general function classes with ℓ_∞ Lipschitz loss in the decoupling case. Yu et al. (2014) obtained a $O(1/\sqrt{n})$ bound for trace norm regularized linear function classes with the decomposable loss. In addition, Xu et al. (2016) used the local Rademacher complexity to derive a $\widetilde{O}(1/n)$ bound for trace norm regularized linear function classes with the assumption that the singular values of $\boldsymbol{w}$ decay exponentially. Compared with these works, we obtain tighter bounds with the state-of-the-art dependency on c for general function classes under the mild assumptions in both decoupling and coupling cases.

6. Discussion

The main goal of our capacity-based generalization bounds is to provide general and efficient theoretical guarantees for empirically successful multi-label learning methods, especially regarding the dependency on the number of labels. The generalization differences of different multi-label learning models or algorithms are mainly reflected in two aspects. On the one hand, the differences are reflected in the Lipschitz constant of the loss functions, as we showed in Proposition 3.3 and 3.10, different loss functions have different μ values in our bounds. On the other hand, the differences are reflected in the nonlinear mappings corresponding to the specific models used. In fact, when we analyze the generalization of the specific models, the constraint on nonlinear mappings $\beta(\phi(\boldsymbol{x})) \leq A$ is actually refined as $\|\phi(\boldsymbol{x})\| \leq A$ in our analysis, and we will further have $\|\phi(\boldsymbol{x})\| \leq \rho\|\boldsymbol{x}\|$ (ρ is the Lipschitz constant of the nonlinear mappings) to take into account the differences or characteristics of different models. The generalization differences are further reflected in the corresponding Lipschitz constants ρ. Compared with the Lipschitz constants of deep models and shallow models, their differences are particularly obvious (Bartlett et al., 2017; Golowich et al., 2018; Bartlett et al., 2019; Zhang & Liao, 2020; Ledent et al., 2021; Zhang & Zhang, 2023). In order to provide general theoretical guarantees for multi-label learning, here we only make the most general assumptions (e.g., for nonlinear mappings) and do not specify specific models, so the generalization differences of different multi-label learning models or algorithms are not explicitly shown. And if a specific model or algorithm is specified or refined, the results obtained will lose their generality and will not be able to provide theoretical guarantees for all or most multi-label learning models or algorithms.

The analysis of the lower bound will greatly promote our theoretical understanding of multi-label learning and is the most powerful criterion for testing whether the given upper bound is the tightest. However, the effective lower bound remains relatively under-explored. We believe that such a lower bound $\Omega(1/\sqrt{n})$ is still not tight enough. The main evidence is as follows: if we regard the classification problem corresponding to each label as a task, then multi-label learning can be regarded as multi-task learning. Since tasks share some constraint or generative properties with each other, the typical bound for multi-task learning is $O(1/\sqrt{nt})$, where t is the number of tasks. Hence, if we consider that the components share some constraint properties or information with each other, i.e., the label correlations, then appropriate label correlations will improve the generalization performance of multi-label learning. This means that some constraint properties shared between labels (i.e., label correlations) will facilitate the learning effect of multiple labels. In such a case, we should have the lower bound $\Omega(1/\sqrt{nc})$. A $\widetilde{O}(1/\sqrt{nc})$ bound for ℓ_2 Lipschitz surrogate Hamming loss in the decoupling case (as we discussed in Remark 3.8) provides evidence for this analysis. Hence, although our bounds are tighter than the state of the art, the above analysis actually motivates us to develop new theories to investigate the lower bound and break through the theoretical limitations of current methodologies.

7. Conclusion

In this paper, we propose several novel vector-contraction inequalities and derive bounds with a weaker dependency on c, and study the impact of different label correlations on the generalization analysis. In addition, with the label-based ranking multi-label Rademacher complexity, we derive the bound for Macro-Averaged AUC and analyze the relationship between Macro-Averaged AUC and class-imbalance.

In future work, we will extend our bounds to more general settings (especially for label correlations induced by label-specific features and other norm regularizers), and derive tighter bounds for multi-label learning with a faster convergence rate with respect to the number of observations, and further design efficient algorithms to construct multi-label models with good generalization performance.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation of China (62225602), the Fundamental Research Funds for the Cen-

tral Universities (2242024K30035), and the Big Data Computing Center of Southeast University.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 30 (NIPS 2017):6240–6249, 2017.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Barutçuoglu, Z., Schapire, R. E., and Troyanskaya, O. G. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

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

Cabral, R. S., la Torre, F. D., Costeira, J. P., and Bernardino, A. Matrix completion for multi-label image classification. *Advances in Neural Information Processing Systems*, 24 (NIPS 2011):190–198, 2011.

Cesa-Bianchi, N., Re, M., and Valentini, G. Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference. *Machine Learning*, 88(1-2):209–241, 2012.

Clémençon, S., Lugosi, G., and Vayatis, N. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Dembczynski, K., Waegeman, W., Cheng, W., and Hüllermeier, E. Regret analysis for performance metrics in multi-label classification: The case of hamming and subset zero-one loss. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, volume 6321, pp. 280–295, 2010.

Dembczynski, K., Kotlowski, W., and Hüllermeier, E. Consistent multilabel ranking through univariate losses. *arXiv:1206.6401*, 2012.

Foster, D. J. and Rakhlin, A. ℓ_∞ vector contraction for rademacher complexity. *arXiv:1911.06468v1*, 2019.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. *International Conference on Computational Learning Theory*, 75(COLT 2018):297–299, 2018.

Hang, J.-Y. and Zhang, M.-L. Collaborative learning of label semantics and deep label-specific features for multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9860–9871, 2022.

Huang, J., Li, G., Huang, Q., and Wu, X. Learning label-specific features and class-dependent labels for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3309–3323, 2016.

Jia, B., Liu, J., Hang, J., and Zhang, M. Learning label-specific features for decomposition-based multi-class classification. *Frontiers of Computer Science*, 17(6): 176348, 2023.

Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Ledent, A., Mustafa, W., Lei, Y., and Kloft, M. Norm-based generalisation bounds for deep multi-class convolutional neural networks. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35, pp. 8279–8287, 2021.

Lei, Y., Dogan, Ü., Binder, A., and Kloft, M. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2035–2043, 2015.

Lei, Y., Dogan, Ü., Zhou, D., and Kloft, M. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5):2995–3021, 2019.

Li, J., Liu, Y., Yin, R., Zhang, H., Ding, L., and Wang, W. Multi-class learning: From theory to algorithm. *Advances in Neural Information Processing Systems*, 31(NeurIPS 2018):1593–1602, 2018.

Liu, C., Zhao, P., Huang, S., Jiang, Y., and Zhou, Z. Dual set multi-label learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, number AAAI 2018, pp. 3635–3642, 2018.

Liu, W. and Shen, X. Sparse extreme multi-label learning with oracle property. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 4032–4041, 2019.

Lust-Piquard, F. and Pisier, G. Non commutative khintchine and paley inequalities. *Arkiv för matematik*, 29:241–260, 1991.

Maurer, A. A vector-contraction inequality for rademacher complexities. In *Proceedings of the 27th International Conference on Algorithmic Learning Theory*, volume 9925, pp. 3–17, 2016.

Prabhu, Y. and Varma, M. Fastxml: a fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, number KDD 2014, pp. 263–272, 2014.

Rubin, T. N., Chambers, A., Smyth, P., and Steyvers, M. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012.

Rudelson, M. and Vershynin, R. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pp. 603–648, 2006.

Schapire, R. E. and Singer, Y. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

Srebro, N., Sridharan, K., and Tewari, A. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, volume 23, pp. 2199–2207, 2010.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

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 (NeurIPS 2020), 2020.

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(NeurIPS 2021):14332–14344, 2021a.

Wu, G., Li, C., and Yin, Y. Towards understanding generalization of macro-auc in multi-label learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 37540–37570, 2023.

Wu, L., Ledent, A., Lei, Y., and Kloft, M. Fine-grained generalization analysis of vector-valued learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, number AAAI 2021, pp. 10338–10346, 2021b.

Xu, C., Liu, T., Tao, D., and Xu, C. Local rademacher complexity for multi-label learning. *IEEE Transactions on Image Processing*, 25(3):1495–1507, 2016.

Yen, I. E., Huang, X., Ravikumar, P., Zhong, K., and Dhillon, I. S. Pd-sparse : A primal and dual sparse approach to extreme multiclass and multilabel classification. In *Proceedings of the 33nd International Conference on Machine Learning*, volume 48, pp. 3069–3077, 2016.

Yu, H., Jain, P., Kar, P., and Dhillon, I. S. Large-scale multi-label learning with missing labels. In *Proceedings of the 31th International Conference on Machine Learning*, volume 32, pp. 593–601, 2014.

Yu, K., Yu, S., and Tresp, V. Multi-label informed latent semantic indexing. In Baeza-Yates, R. A., Ziviani, N., Marchionini, G., Moffat, A., and Tait, J. (eds.), *Proceedings of the 28th Annual International Conference on Research and Development in Information Retrieval*, number SIGIR 2005, pp. 258–265, 2005.

Zhang, M.-L. and Wu, L. Lift: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):107–120, 2015.

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, Y. and Liao, S. A kernel perspective for the decision boundary of deep neural networks. In *Proceedings of the 32nd IEEE International Conference on Tools with Artificial Intelligence*, number ICTAI 2020, pp. 653–660, 2020.

Zhang, Y. and Zhang, M. Nearly-tight bounds for deep kernel learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 41861–41879, 2023.

A. Appendix Outline

In the appendix, we give the detailed proofs of those theoretical results which we only provide proof sketches in the main paper. Our main proofs include:

- The basic generalization bound for ℓ_2 Lipschitz loss (Theorem 3.4).
- The novel vector-contraction inequality for for ℓ_2 Lipschitz loss (Lemma 3.6).
- The generalization bound with a square-root dependency on c for ℓ_2 Lipschitz loss (Theorem 3.7).
- The ℓ_∞ Lipschitz continuity of the commonly used loss for multi-label learning (Proposition 3.10).
- The novel vector-contraction inequality for ℓ_∞ Lipschitz loss with a square-root dependency on c (Lemma 3.11).
- The generalization bound with a square-root dependency on c for ℓ_∞ Lipschitz loss (Theorem 3.12).
- The generalization bounds with no dependency on c for ℓ_∞ Lipschitz loss in the coupling case (Corollary 3.14).
- The novel vector-contraction inequality for ℓ_∞ Lipschitz loss with no dependency on c (Lemma 3.16).
- The generalization bound with no dependency on c for ℓ_∞ Lipschitz loss in the decoupling case (Theorem 3.17).
- The generalization bound for Macro-Averaged AUC w.r.t. ℓ_∞ Lipschitz loss (Theorem 4.2).

B. Preliminaries

First, we define the loss function space as follows:

$$\mathcal{L} = \{L(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) : \boldsymbol{f} \in \mathcal{F}\},$$

where $\mathcal{F}$ is the vector-valued function class (1) of multi-label learning defined in the main paper:

$$\begin{aligned}\mathcal{F} = \{\boldsymbol{x} \mapsto \boldsymbol{f}(\boldsymbol{x}) :& \boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_c(\boldsymbol{x})), f_j(\boldsymbol{x}) = \langle \boldsymbol{w}_j, \phi(\boldsymbol{x})\rangle, \\ & \boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_c) \in \mathbb{R}^{d\times c}, \alpha(\boldsymbol{w}) \leq \Lambda, \beta(\phi(\boldsymbol{x})) \leq A, \boldsymbol{x} \in \mathcal{X}, j \in [c]\},\end{aligned}$$

For any training dataset $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) : i \in [n]\}$, let $D' = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) : i \in [n]\}$ be the training dataset with only one sample different from D, where the k-th sample is replaced by $(\boldsymbol{x}'_k, \boldsymbol{y}'_k)$. Let $\Phi(D) = \sup_{\boldsymbol{f}\in\mathcal{F}}[\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim P}[L(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y})] - \frac{1}{n}\sum_{i=1}^n L(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{y}_i)] = \sup_{\boldsymbol{f}\in\mathcal{F}}[R(\boldsymbol{f}) - \widehat{R}_D(\boldsymbol{f})]$, then

$$\begin{aligned}&\Phi(D') - \Phi(D) \\ =& \sup_{\boldsymbol{f}\in\mathcal{F}}[R(\boldsymbol{f}) - \widehat{R}_{D'}(\boldsymbol{f})] - \sup_{\boldsymbol{f}\in\mathcal{F}}[R(\boldsymbol{f}) - \widehat{R}_D(\boldsymbol{f})] \\ \leq& \sup_{\boldsymbol{f}\in\mathcal{F}}[\widehat{R}_D(\boldsymbol{f}) - \widehat{R}_{D'}(\boldsymbol{f})] \\ =& \sup_{\boldsymbol{f}\in\mathcal{F}} \frac{[L(\boldsymbol{f}(\boldsymbol{x}_k), \boldsymbol{y}_k) - L(\boldsymbol{f}(\boldsymbol{x}'_k), \boldsymbol{y}'_k)]}{n} \\ \leq& \frac{M}{n}.\end{aligned}$$

According to McDiarmid's inequality, for any $0 < \delta < 1$, with probability at least $1 - \frac{\delta}{2}$ over the training dataset D, the following holds:

$$\Phi(D) \leq \mathbb{E}_D[\Phi(D)] + M\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{4}$$

Then, we will estimate the upper bound of $\mathbb{E}_D[\Phi(D)]$.

$$\begin{aligned}
&\mathbb{E}_D[\Phi(D)] \\
=&\mathbb{E}_D\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\left[\mathbb{E}_{D'}\left[\widehat{R}_{D'}(\boldsymbol{f})-\widehat{R}_D(\boldsymbol{f})\right]\right]\right] \\
\leq&\mathbb{E}_{D,D'}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\left[\widehat{R}_{D'}(\boldsymbol{f})-\widehat{R}_D(\boldsymbol{f})\right]\right] \\
=&\mathbb{E}_{D,D'}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\left[\sum_{i=1}^{n}L(\boldsymbol{f}(\boldsymbol{x}_i'),\boldsymbol{y}_i')-L(\boldsymbol{f}(\boldsymbol{x}_i),\boldsymbol{y}_i)\right]\right] \\
=&\mathbb{E}_{\boldsymbol{\epsilon},D,D'}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\left[\sum_{i=1}^{n}\epsilon_i\left(L(\boldsymbol{f}(\boldsymbol{x}_i'),\boldsymbol{y}_i')\right)-L(\boldsymbol{f}(\boldsymbol{x}_i),\boldsymbol{y}_i)\right)\right]\right] \\
\leq&\mathbb{E}_{\boldsymbol{\epsilon},D'}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i L(\boldsymbol{f}(\boldsymbol{x}_i'),\boldsymbol{y}_i')\right] \\
+&\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}-\epsilon_i L(\boldsymbol{f}(\boldsymbol{x}_i),\boldsymbol{y}_i)\right] \\
=&2\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i L(\boldsymbol{f}(\boldsymbol{x}_i),\boldsymbol{y}_i)\right].
\end{aligned}\tag{5}$$

Then apply McDiarmid's inequality to $\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i L(\boldsymbol{f}(\boldsymbol{x}_i),\boldsymbol{y}_i)\right]$, we have

$$\mathbb{E}_{\boldsymbol{\epsilon},D}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i L(\boldsymbol{f}(\boldsymbol{x}_i),\boldsymbol{y}_i)\right]\leq\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i L(\boldsymbol{f}(\boldsymbol{x}_i),\boldsymbol{y}_i)\right]+M\sqrt{\frac{\ln(2/\delta)}{2n}},$$

i.e.,

$$\Re(\mathcal{L})\leq\hat{\Re}_D(\mathcal{L})+M\sqrt{\frac{\ln(2/\delta)}{2n}}.\tag{6}$$

Combining with (4), (5) and (6), then

$$R(\boldsymbol{f})\leq\widehat{R}_D(\boldsymbol{f})+2\hat{\Re}_D(\mathcal{L})+3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.\tag{7}$$

C. Generalization Bounds for ℓ_2 Lipschitz Loss

C.1. Proof of Theorem 3.4

We first introduce the following lemmas:

Lemma C.1 (Corollary 1 in (Maurer, 2016)). *Let $\mathcal{F}$ be a vector-valued function class. Given a dataset D of size n. Assume that the loss function is μ-Lipschitz continuous with respect to the ℓ_2 norm. Then*

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i L(\boldsymbol{f}(\boldsymbol{x}_i))\right]\leq\sqrt{2}\mu\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\epsilon_{ij}f_j\left(\boldsymbol{x}_i\right)\right],$$

where each ϵ_{ij} is an independent doubly indexed Rademacher random variable, and $f_j\left(\boldsymbol{x}_i\right)$ is the j-th component of $\boldsymbol{f}\left(\boldsymbol{x}_i\right)$.

Lemma C.2 (Khintchine-Kahane inequality (Lust-Piquard & Pisier, 1991)). *Let $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n\in\mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with $\|\cdot\|$ being the associated p-th norm. Let $\epsilon_1,\ldots,\epsilon_n$ be a sequence of independent Rademacher variables. Then,*

for any $p \geq 1$ there holds

$$\min(\sqrt{p-1},1)\left[\sum_{i=1}^{n}\|\boldsymbol{v}_i\|^2\right]^{\frac{1}{2}} \leq \left[\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum_{i=1}^{n}\epsilon_i\boldsymbol{v}_i\right\|^p\right]^{\frac{1}{p}} \leq \max(\sqrt{p-1},1)\left[\sum_{i=1}^{n}\|\boldsymbol{v}_i\|^2\right]^{\frac{1}{2}},$$

and

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum_{i=1}^{n}\epsilon_i\boldsymbol{v}_i\right\| \geq 2^{-\frac{1}{2}}\left[\sum_{i=1}^{n}\|\boldsymbol{v}_i\|^2\right]^{\frac{1}{2}}.$$

According to Lemma C.1, we then have

$$\begin{aligned}
&\hat{\mathfrak{R}}_D(\mathcal{L})\\
\leq&\sqrt{2}\mu\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\epsilon_{ij}f_j\left(\boldsymbol{x}_i\right)\right]\\
\leq&\sqrt{2}\mu c\max_j\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f_j\in\mathcal{F}_j}\frac{1}{n}\sum_{i=1}^{n}\epsilon_{ij}f_j\left(\boldsymbol{x}_i\right)\right]\\
\leq&\sqrt{2}\mu c\max_j\sup_{f_j\in\mathcal{F}_j}\frac{1}{n}\left(\sum_{i=1}^{n}\left(f_j\left(\boldsymbol{x}_i\right)\right)^2\right)^{\frac{1}{2}} \quad \text{(Use Lemma C.2)}\\
\leq&\frac{\sqrt{2}\mu cB}{\sqrt{n}}. \quad \text{(Use Assumption 3.1 in the main paper)}
\end{aligned}$$

Combining with (7), then

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + \frac{2\sqrt{2}\mu cB}{\sqrt{n}} + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

C.2. Proof of Lemma 3.6

Proof Sketch: First, the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded by the empirical ℓ_2 norm covering number with the refined Dudley's entropy integral inequality. Second, according to the Lipschitz continuity w.r.t the ℓ_2 norm, the empirical ℓ_2 norm covering number of $\mathcal{F}$ can be bounded by the empirical ℓ_2 norm covering number of $\mathcal{P}(\mathcal{F})$. Third, the empirical ℓ_2 norm covering number of $\mathcal{P}(\mathcal{F})$ can be bounded by using Sudakov's minoration (Wainwright, 2019), which bounds the ℓ_2 norm covering number of a function class by the expectation of a Gaussian process indexed by the function class, and the expectation of the Gaussian process can be bounded by the worst-case Rademacher complexity of the projection function class $\mathcal{P}(\mathcal{F})$. Hence, the problem is transferred to the estimation of the worst-case Rademacher complexity. Finally, we estimate the lower bound of the worst-case Rademacher complexity of $\mathcal{P}(\mathcal{F})$, and then combined with the above steps, the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded.

We first introduce the following lemmas:

Lemma C.3 (Sudakov's minoration (Wainwright, 2019)). *Let $\{Z_f, f\in\mathcal{F}\}$ be a zero-mean Gaussian process. Then*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}Z_f\right] \geq \sup_{\epsilon>0}\frac{\epsilon}{2}\sqrt{\log\mathcal{M}_2(\epsilon,\mathcal{F},D)} \geq \sup_{\epsilon>0}\frac{\epsilon}{2}\sqrt{\log\mathcal{N}_2(\epsilon,\mathcal{F},D)}.$$

Lemma C.4 (Relationship between Rademacher and Gaussian complexity (Wainwright, 2019)). *Let $\mathcal{G}$ be a class of real-valued functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Let $D=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ be a set with n i.i.d. samples. Then*

$$\sqrt{\frac{2}{\pi}}\hat{\mathfrak{R}}_D(\mathcal{G}) \leq \hat{\mathfrak{G}}_D(\mathcal{G}) \leq 2\hat{\mathfrak{R}}_D(\mathcal{G})\sqrt{\log n},$$

where $\hat{\mathfrak{G}}_D(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i g\left(\boldsymbol{x}_i\right)\right]$, and $\sigma_1,\ldots,\sigma_n$ are i.i.d. random variables obeying the normal distribution.

The following lemma is a refined result of Proposition 22 in (Ledent et al., 2021), where we replace the function class taking values in $[0,1]$ with the b-bounded function class, the refinement is obvious.

Lemma C.5 (Refined Dudley's entropy integral inequality)**.** *Let $\mathcal{F}$ be a real-valued function class with $f \leq b$, $f \in \mathcal{F}$, $b > 0$, and assume that $0 \in \mathcal{F}$. Let S be a finite sample of size n. For any $2 \leq p \leq \infty$, we have the following relationship between the Rademacher complexity $\hat{\mathfrak{R}}_S(\mathcal{F})$ and the covering number $\mathcal{N}_p(\epsilon, \mathcal{F}, S)$.*

$$\hat{\mathfrak{R}}_S(\mathcal{F}) \leq \inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_{\alpha}^{b}\sqrt{\log \mathcal{N}_p(\epsilon, \mathcal{F}, S)}d\epsilon\right).$$

Step 1: We first derive the relationship between the empirical ℓ_2 norm covering number $\mathcal{N}_2(\epsilon, \mathcal{L}, D)$ and the empirical ℓ_2 norm covering number $\mathcal{N}_2(\epsilon, \mathcal{P}(\mathcal{F}), [c] \times D)$.

For the dataset $D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ with n i.i.d. examples:

$$\begin{aligned}
&\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(L(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{y}_i) - L(\boldsymbol{f}'(\boldsymbol{x}_i), \boldsymbol{y}_i)\right)^2} \\
\leq &\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mu^2\|\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{f}'(\boldsymbol{x}_i)\|_2^2} \quad \text{(Use Assumption 3.2)} \\
\leq &\mu\sqrt{c\frac{1}{n}\sum_{i=1}^{n}\frac{1}{c}\sum_{j=1}^{c}\left(f_j(\boldsymbol{x}_i) - f_j'(\boldsymbol{x}_i)\right)^2} \\
\leq &\mu\sqrt{c}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{c}\sum_{j=1}^{c}\left(p_j(\boldsymbol{f}(\boldsymbol{x}_i)) - p_j(\boldsymbol{f}'(\boldsymbol{x}_i))\right)^2}. \quad \text{(The definition of the projection function class } \mathcal{P}(\mathcal{F}))
\end{aligned}$$

Then, according to the definition of the empirical ℓ_2 covering number, we have that an empirical ℓ_2 cover of $\mathcal{P}(\mathcal{F})$ at radius $\epsilon/\mu\sqrt{c}$ is also an empirical ℓ_2 cover of the loss function space associated with $\mathcal{F}$ at radius ϵ, and we can conclude that:

$$\mathcal{N}_2(\epsilon, \mathcal{L}, D) \leq \mathcal{N}_2\left(\frac{\epsilon}{\mu\sqrt{c}}, \mathcal{P}(\mathcal{F}), [c] \times D\right). \tag{8}$$

Step 2: We show that the empirical ℓ_2 norm covering number of $\mathcal{P}(\mathcal{F})$ can be bounded by using Sudakov's minoration (Wainwright, 2019), which bounds the ℓ_2 norm covering number of a function class by the expectation of a Gaussian process indexed by the function class, and the expectation of the Gaussian process can be bounded by the worst-case Rademacher complexity of the projection function class $\mathcal{P}(\mathcal{F})$.

Let $\mathcal{G}$ be a class of real-valued functions. Let $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set with n i.i.d. samples. We consider the Gaussian process $Z_f = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sigma_i f(\boldsymbol{x}_i)$, then according to Lemma C.3, we have $\mathbb{E}\left[\sup_{f\in\mathcal{F}} Z_f\right] = \sqrt{n}\hat{\mathfrak{G}}_D(\mathcal{F})$. Then, combining Lemma C.3 and Lemma C.4, we can get $\sqrt{\log \mathcal{N}_2(\epsilon, \mathcal{F}, D)} \leq \frac{4}{\epsilon}\hat{\mathfrak{R}}_D(\mathcal{F})\sqrt{n\log n}$.

Hence, for the projection function class $\mathcal{P}(\mathcal{F})$, we have

$$\sqrt{\log \mathcal{N}_2(\epsilon, \mathcal{P}(\mathcal{F}), [c] \times D)} \leq \frac{4}{\epsilon}\hat{\mathfrak{R}}_{[c]\times D}(\mathcal{P}(\mathcal{F}))\sqrt{n\log n}. \tag{9}$$

Step 3: According to Assumption 3.1 in the main paper, we can obtain the lower bound of the worst-case Rademacher complexity $\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))$ by the Khintchine-Kahane inequality with $p = 1$:

$$\begin{aligned}
\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F})) &= \sup_{[c]\times D\in[c]\times\mathcal{X}^n}\hat{\mathfrak{R}}_{[c]\times D}(\mathcal{P}(\mathcal{F})) = \sup_{[c]\times D\in[c]\times\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{p_j(\boldsymbol{f}(\boldsymbol{x}_i))\in\mathcal{P}(\mathcal{F})}\frac{1}{nc}\sum_{i=1}^{n}\sum_{j=1}^{c}\epsilon_i p_j(\boldsymbol{f}(\boldsymbol{x}_i))\right] \\
&= \sup_{D\in\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f_j\in\mathcal{F}_j}\frac{1}{nc}\sum_{i=1}^{n}\sum_{j=1}^{c}\epsilon_i f_j(\boldsymbol{x}_i)\right] \geq \sup_{D\in\mathcal{X}^n}\frac{1}{nc}\sup_{f_j\in\mathcal{F}_j}\frac{1}{\sqrt{2}}\left[\sum_{i=1}^{n}\sum_{j=1}^{c}\left(f_j(\boldsymbol{x}_i)\right)^2\right]^{\frac{1}{2}}.
\end{aligned}$$

Since $|f_j(\cdot)| \leq B$, we set $\sup_{D\in\mathcal{X}^n} \frac{1}{nc} \sup_{f_j\in\mathcal{F}_j} \left[\sum_{i=1}^n \sum_{j=1}^c \left(f_j\left(\boldsymbol{x}_i\right)\right)^2\right]^{\frac{1}{2}} = \frac{B}{\sqrt{nc}}$. So,

$$\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F})) \geq \frac{B}{\sqrt{2nc}}. \tag{10}$$

Then, according to Lemma C.5 and combined with the above steps, we have

$$\begin{aligned}
&\hat{\mathfrak{R}}_D(\mathcal{L}) \\
\leq &\inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^M \sqrt{\log \mathcal{N}_2(\epsilon, \mathcal{L}, D)}d\epsilon\right) \\
\leq &\inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^M \sqrt{\log \mathcal{N}_2(\frac{\epsilon}{\mu\sqrt{c}}, \mathcal{P}(\mathcal{F}), [c]\times D)}d\epsilon\right) \quad \text{(Use inequality (8))} \\
\leq &\inf_{\alpha>0}\left(4\alpha + 48\sqrt{c}\mu\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\log^{\frac{1}{2}}(nc)\int_\alpha^M \epsilon^{-1}d\epsilon\right) \\
&\text{(Use inequality (9) and the definition of the worst-case Rademacher complexity)} \\
\leq &48\mu\sqrt{c}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F})) + 48\mu\sqrt{c}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\log^{\frac{1}{2}}(nc)\cdot\log\frac{M}{12\sqrt{c}\mu\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))} \quad \text{(Choose } \alpha = 12\sqrt{c}\mu\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\text{)} \\
\leq &48\mu\sqrt{c}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\left(1 + \log^{\frac{1}{2}}(nc)\cdot\log\frac{M\sqrt{n}}{\mu B}\right). \quad \text{(Use inequality (10))}
\end{aligned}$$

C.3. Proof of Theorem 3.7

We upper bound the worst-case Rademacher complexity $\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))$ as the following:

$$\begin{aligned}
&\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F})) \\
= &\sup_{[c]\times D\in[c]\times\mathcal{X}^n} \hat{\mathfrak{R}}_{[c]\times D}(\mathcal{P}(\mathcal{F})) \\
= &\sup_{[c]\times D\in[c]\times\mathcal{X}^n} \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{p_j(\boldsymbol{f}(\boldsymbol{x}_i))\in\mathcal{P}(\mathcal{F})} \frac{1}{nc}\sum_{i=1}^n\sum_{j=1}^c \epsilon_i p_j(\boldsymbol{f}(\boldsymbol{x}_i))\right] \\
= &\sup_{D\in\mathcal{X}^n} \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f_j\in\mathcal{F}_j} \frac{1}{nc}\sum_{i=1}^n\sum_{j=1}^c \epsilon_i f_j\left(\boldsymbol{x}_i\right)\right] \\
\leq &\sup_{D\in\mathcal{X}^n}\sup_{f_j\in\mathcal{F}_j} \frac{1}{nc}\left(\sum_{i=1}^n\sum_{j=1}^c \left(f_j\left(\boldsymbol{x}_i\right)\right)^2\right)^{\frac{1}{2}} \quad \text{(Use Lemma C.2)} \\
\leq &\frac{B}{\sqrt{nc}}. \quad \text{(Use Assumption 3.1 in the main paper)}
\end{aligned} \tag{11}$$

Then, we have

$$\begin{aligned}
\hat{\mathfrak{R}}_D(\mathcal{F}) \leq &48\mu\sqrt{c}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\left(1 + \log^{\frac{1}{2}}(nc)\cdot\log\frac{M\sqrt{n}}{\mu B}\right) \\
\leq &\frac{48B\mu\left(1 + \log^{\frac{1}{2}}(nc)\cdot\log\frac{M\sqrt{n}}{\mu B}\right)}{\sqrt{n}}.
\end{aligned}$$

Combining with (7), then

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + \frac{96B\mu\left(1 + \log^{\frac{1}{2}}(nc)\cdot\log\frac{M\sqrt{n}}{\mu B}\right)}{\sqrt{n}} + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

D. Generalization Bounds for ℓ_∞ Lipschitz Loss

D.1. Proof of Proposition 3.10

We first prove that the surrogate Hamming Loss is μ-Lipschitz continuous with respect to the ℓ_∞ norm.

$$\begin{aligned}
&\left|L_H(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) - L_H\left(\boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{y}\right)\right| \\
=&\left|\frac{1}{c}\sum_{j=1}^{c}\ell\left(y_j f_j(\boldsymbol{x})\right) - \frac{1}{c}\sum_{j=1}^{c}\ell\left(y_j f_j'(\boldsymbol{x})\right)\right| \\
=&\frac{1}{c}\sum_{j=1}^{c}\left|\ell\left(y_j f_j(\boldsymbol{x})\right) - \ell\left(y_j f_j'(\boldsymbol{x})\right)\right| \\
\leq&\frac{1}{c}\sum_{j=1}^{c}\mu\left|f_j(\boldsymbol{x}) - f_j'(\boldsymbol{x})\right| \\
\leq&\frac{1}{c}\mu c\max_{j\in[c]}\left|f_j(\boldsymbol{x}) - f_j'(\boldsymbol{x})\right| \\
=&\mu\left\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x})\right\|_\infty.
\end{aligned}$$

Then, with the elementary inequality

$$\left|\max\left\{a_1, \ldots, a_c\right\} - \max\left\{b_1, \ldots, b_c\right\}\right| \leq \max\left\{\left|a_1 - b_1\right|, \ldots, \left|a_c - b_c\right|\right\},$$

we proof that the surrogate Subset Loss is μ-Lipschitz continuous with respect to the ℓ_∞ norm.

$$\begin{aligned}
&\left|L_S(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) - L_S\left(\boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{y}\right)\right| \\
=&\left|\max_{j\in[c]}\ell\left(y_j f_j(\boldsymbol{x})\right) - \max_{j\in[c]}\ell\left(y_j f_j'(\boldsymbol{x})\right)\right| \\
\leq&\max_{j\in[c]}\left|\ell\left(y_j f_j(\boldsymbol{x})\right) - \ell\left(y_j f_j'(\boldsymbol{x})\right)\right| \\
\leq&\mu\max_{j\in[c]}\left|f_j(\boldsymbol{x}) - f_j'(\boldsymbol{x})\right| \\
=&\mu\left\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x})\right\|_\infty.
\end{aligned}$$

Finally, we proof that the surrogate Ranking Loss is 2μ-Lipschitz continuous with respect to the ℓ_∞ norm.

$$\begin{aligned}
&\left|L_R(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) - L_R\left(\boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{y}\right)\right| \\
=&\frac{1}{|Y^+|\,|Y^-|}\left|\sum_{p\in Y^+}\sum_{q\in Y^-}\left(\ell\left(f_p(\boldsymbol{x}) - f_q(\boldsymbol{x})\right) - \ell\left(f_p'(\boldsymbol{x}) - f_q'(\boldsymbol{x})\right)\right)\right| \\
\leq&\max_{p\in Y^+, q\in Y^-}\left|\ell\left(f_p(\boldsymbol{x}) - f_q(\boldsymbol{x})\right) - \ell\left(f_p'(\boldsymbol{x}) - f_q'(\boldsymbol{x})\right)\right| \\
\leq&\mu\max_{p\in Y^+, q\in Y^-}\left|\left(f_p(\boldsymbol{x}) - f_q(\boldsymbol{x})\right) - \left(f_p'(\boldsymbol{x}) - f_q'(\boldsymbol{x})\right)\right| \\
\leq&\mu\left(\max_{p\in Y^+}\left|f_p(\boldsymbol{x}) - f_p'(\boldsymbol{x})\right| + \max_{q\in Y^-}\left|f_q(\boldsymbol{x}) - f_q'(\boldsymbol{x})\right|\right) \\
\leq&2\mu\max_{j\in[c]}\left|f_j(\boldsymbol{x}) - f_j'(\boldsymbol{x})\right| \\
=&2\mu\left\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x})\right\|_\infty.
\end{aligned}$$

D.2. Proof of Lemma 3.11

We first introduce the following vector-contraction inequality in (Foster & Rakhlin, 2019):

Lemma D.1 (Theorem 1 in (Foster & Rakhlin, 2019)). *Let $\mathcal{F} \subseteq \{\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^c\}$, and let $\phi : \mathbb{R}^c \to \mathbb{R}$ be L-lipschitz with respect to the ℓ_∞ norm. For any $1 > \eta > 0$, there exists a constant $C > 0$ such that if $|\phi(\boldsymbol{f}(x))| \vee \|\boldsymbol{f}(x)\|_\infty \leq \beta$, then*

$$\hat{\mathfrak{R}}_D(\phi \circ \mathcal{F}) \leq CL\sqrt{c} \max_j \widetilde{\mathfrak{R}}_n(\mathcal{F}_j) \log^{\frac{3+\eta}{2}} \left(\frac{\beta n}{\max_j \widetilde{\mathfrak{R}}_n(\mathcal{F}_j)} \right),$$

where $\hat{\mathfrak{R}}_D(\phi \circ \mathcal{F}) = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \phi(\boldsymbol{f}(\boldsymbol{x}_i)) \right], \widetilde{\mathfrak{R}}_n(\mathcal{F}_j) = \sup_{D \in \mathcal{X}^n} \hat{\mathfrak{R}}_D(\mathcal{F}_j)$.

Next we will refine the proof of Lemma D.1. First, we introduce the result in Theorem 1 in (Foster & Rakhlin, 2019): $\log \mathcal{N}_\infty(\epsilon, \phi \circ \mathcal{F}, D) \leq \max_i EcLr_i \log^{1+\eta} \frac{e^{2+\eta} n}{r_i \epsilon_0}$, where $r_i = 8n(\frac{\widetilde{\mathfrak{R}}_n(\mathcal{F}_j)}{a\epsilon_0})^2$, $E > 0$, $0 < a < 1$, $\epsilon_0 = \frac{\epsilon}{L}$.

Then, by direct calculation and proper scaling, we have

$$\begin{aligned}
&\log \mathcal{N}_2(\epsilon, \phi \circ \mathcal{F}, D) \\
\leq &\max_j Ecr_i \log^{1+\eta} \frac{e^{2+\eta} n}{r_i \epsilon_0} \\
\leq &\max_j Ec8n(\frac{\widetilde{\mathfrak{R}}_n(\mathcal{F}_j)}{a\epsilon_0})^2 \log^{1+\eta} \frac{4}{(\widetilde{\mathfrak{R}}_n(\mathcal{F}_j))^2} \quad \text{(Use inequality } 2\widetilde{\mathfrak{R}}_n(\mathcal{F}_j) < \epsilon_0 < 1 \text{)} \\
\leq &\max_j Ec32n(\frac{\widetilde{\mathfrak{R}}_n(\mathcal{F}_j)}{a\epsilon_0})^2 \log^{1+\eta} \frac{2}{\widetilde{\mathfrak{R}}_n(\mathcal{F}_j)} \\
\leq &\max_j Ec32n(\frac{\widetilde{\mathfrak{R}}_n(\mathcal{F}_j)}{a\epsilon_0})^2 \log^{1+\eta} \frac{2\sqrt{2n}}{B} \qquad (12) \\
&\text{(Use the similar technique to the proof of Lemma 3.6, the lower bound of } \widetilde{\mathfrak{R}}_n(\mathcal{F}_j) \geq \frac{B}{\sqrt{2n}} \text{)}
\end{aligned}$$

According to Lemma C.5 and combined with (12), we have

$$\begin{aligned}
&\hat{\mathfrak{R}}_D(\phi \circ \mathcal{F}) \\
\leq &\inf_{\alpha > 0} \left(4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^M \sqrt{\log \mathcal{N}_2(\epsilon, \phi \circ \mathcal{F}, D)} d\epsilon \right) \\
\leq &\inf_{\alpha > 0} \left(4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^M \sqrt{\max_j Ec32n(\frac{\widetilde{\mathfrak{R}}_n(\mathcal{F}_j)}{a\epsilon_0})^2 \log^{1+\eta} \frac{2\sqrt{2n}}{B}} d\epsilon \right) \quad \text{(Use inequality (12))} \\
\leq &\inf_{\alpha > 0} \left(4\alpha + \frac{48\sqrt{2Ec}L}{a} \max_j \widetilde{\mathfrak{R}}_n(\mathcal{F}_j) \log^{\frac{1+\eta}{2}} (\frac{2\sqrt{2n}}{B}) \int_\alpha^M \epsilon^{-1} d\epsilon \right) \\
\leq &\frac{48\sqrt{2Ec}L}{a} \max_j \widetilde{\mathfrak{R}}_n(\mathcal{F}_j) + \frac{48\sqrt{2Ec}L}{a} \max_j \widetilde{\mathfrak{R}}_n(\mathcal{F}_j) \log^{\frac{1+\eta}{2}} (\frac{2\sqrt{2n}}{B}) \cdot \log \frac{aM}{12\sqrt{2Ec}L\widetilde{\mathfrak{R}}_n(\mathcal{F}_j)} \\
&\text{(Choose } \alpha = \frac{12\sqrt{2Ec}L}{a} \max_j \widetilde{\mathfrak{R}}_n(\mathcal{F}_j)) \\
\leq &\frac{48\sqrt{2Ec}L}{a} \max_j \widetilde{\mathfrak{R}}_n(\mathcal{F}_j) + \frac{48\sqrt{2Ec}L}{a} \max_j \widetilde{\mathfrak{R}}_n(\mathcal{F}_j) \log^{\frac{1+\eta}{2}} (\frac{2\sqrt{2n}}{B}) \cdot \log \frac{M\sqrt{n}}{\sqrt{Ec}LB} \quad \text{(Use } 0 < a < 1 \text{ and } \widetilde{\mathfrak{R}}_n(\mathcal{F}_j) \geq \frac{B}{\sqrt{2n}} \text{)}
\end{aligned}$$

Finally, combined with our problem setting, we have

$$\hat{\mathfrak{R}}_D(\mathcal{F}) \leq \frac{48\sqrt{2E}\mu}{a} \sqrt{c} \max_j \widetilde{\mathfrak{R}}_n(\mathcal{F}_j) \times \left(1 + \log^{\frac{1+\eta}{2}} (\frac{2\sqrt{2n}}{B}) \cdot \log(\frac{M}{\sqrt{E}\mu B} \sqrt{\frac{n}{c}}) \right).$$

D.3. Proof of Theorem 3.12

Using the similar technique to the proof of Theorem 3.4, $\widetilde{\mathfrak{R}}_n(\mathcal{F}_j)$ can be upper bounded by $\frac{B}{\sqrt{n}}$, we then have

$$\hat{\Re}_D(\mathcal{F}) \leq \frac{48\sqrt{2E}\mu B}{a}\sqrt{\frac{c}{n}} \times \left(1 + \log^{\frac{1+\eta}{2}}(\frac{2\sqrt{2n}}{B}) \cdot \log(\frac{M}{\sqrt{E}\mu B}\sqrt{\frac{n}{c}})\right).$$

Combining with (7), then

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + \frac{96\sqrt{2E}\mu B}{a}\sqrt{\frac{c}{n}} \times \left(1 + \log^{\frac{1+\eta}{2}}(\frac{2\sqrt{2n}}{B}) \cdot \log(\frac{M}{\sqrt{E}\mu B}\sqrt{\frac{n}{c}})\right) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

D.4. Proof of Corollary 3.14

We first introduce the following lemmas:

Lemma D.2 (Theorem 5 in (Lei et al., 2019)). *Let $\mathcal{F}$ be a vector-valued function class of the multi-label learning defined by (1). Let Assumptions 3.1 and 3.9 hold. Given a dataset D of size n. Suppose that the loss function is L-Lipschitz continuous w.r.t. ℓ_∞ norm. Then the $\hat{\Re}_D(\mathcal{L})$ can be bounded by*

$$\hat{\Re}_D(\mathcal{L}) \leq 16L\sqrt{c\log 2}\widetilde{\Re}_{nc}(\widetilde{\mathcal{F}})\left(1 + \log_2^{\frac{3}{2}}\left(\sqrt{2}n^{\frac{3}{2}}c\right)\right).$$

Since $\log_2^{\frac{3}{2}}\left(\sqrt{2}n^{\frac{3}{2}}c\right) \geq 1$ and $16\sqrt{\log 2} < 8.8$, we have

$$\hat{\Re}_D(\mathcal{L}) \leq 32L\sqrt{c\log 2}\widetilde{\Re}_{nc}(\widetilde{\mathcal{F}})\log_2^{\frac{3}{2}}\left(\sqrt{2}n^{\frac{3}{2}}c\right) \leq 18L\sqrt{c}\widetilde{\Re}_{nc}(\widetilde{\mathcal{F}})\log_2^{\frac{3}{2}}\left(\sqrt{2}n^{\frac{3}{2}}c\right).$$

Using the similar technique to the proof of Theorem 3.4, $\widetilde{\Re}_{nc}(\widetilde{\mathcal{F}})$ can be upper bounded by $\frac{\Lambda A}{\sqrt{nc}}$. Then combining with (7) and our problem setting, we have

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + \frac{36\mu\Lambda A\log_2^{\frac{3}{2}}\left(\sqrt{2}n^{\frac{3}{2}}c\right)}{\sqrt{n}} + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Lemma D.3 (Corollary 10 in (Lei et al., 2019)). *Let $\mathcal{F}$ be a vector-valued function class defined by (1), where $\alpha(\boldsymbol{w}) := \|\boldsymbol{w}\|_{S_p}$, $\beta(\phi(\boldsymbol{x})) := \max_{i\in[n]}\|\boldsymbol{x}_i\|_2$, i.e., ϕ be the identity map, and $p \geq 1$. Assume that the loss function is L-Lipschitz continuous w.r.t. ℓ_∞ norm. Then, if $1 \leq p \leq 2$, for any $0 < \delta < 1$ with probability of $1-\delta$, we have*

$$\hat{\Re}_D(\mathcal{L}) \leq \frac{16L\sqrt{\log 2}\Lambda\max_{i\in[n]}\|\boldsymbol{x}_i\|_2}{\sqrt{n}}\left(1 + \log_2^{\frac{3}{2}}\left(\sqrt{2}n^{\frac{3}{2}}c\right)\right) \leq \frac{18L\Lambda\max_{i\in[n]}\|\boldsymbol{x}_i\|_2\log_2^{\frac{3}{2}}(\sqrt{2}n^{\frac{3}{2}}c)}{\sqrt{n}}$$

If $p > 2$,

$$\begin{aligned}\hat{\Re}_D(\mathcal{L}) \leq& \frac{16L\sqrt{\log 2}\Lambda\max_{i\in[n]}\|\boldsymbol{x}_i\|_2\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{n}}\left(1 + \log_2^{\frac{3}{2}}\left(\sqrt{2}n^{\frac{3}{2}}c\right)\right)\\ \leq& \frac{18L\Lambda\max_{i\in[n]}\|\boldsymbol{x}_i\|_2\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}\log_2^{\frac{3}{2}}(\sqrt{2}n^{\frac{3}{2}}c)}{\sqrt{n}},\end{aligned}$$

where we scale them by $\log_2^{\frac{3}{2}}\left(\sqrt{2}n^{\frac{3}{2}}c\right) \geq 1$ and $16\sqrt{\log 2} < 8.8$.

This lemma holds for the class of linear functions and is logarithmically dependent on the number of labels. We adjust the constraint on nonlinear mappings such that our bound holds for the general class of nonlinear functions. Then combining with (7) and our problem setting, when $1 \leq p \leq 2$, we have:

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + \frac{36\mu\Lambda A\log_2^{\frac{3}{2}}(\sqrt{2}n^{\frac{3}{2}}c)}{\sqrt{n}} + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}},$$

when $p > 2$, we have:

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + \frac{36\mu\Lambda A\min\{c,d\}^{\frac{1}{2}-\frac{1}{p}}\log_2^{\frac{3}{2}}(\sqrt{2}n^{\frac{3}{2}}c)}{\sqrt{n}} + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

D.5. Proof of Lemma 3.16

Proof Sketch: First, the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded by the empirical ℓ_∞ norm covering number with the refined Dudley's entropy integral inequality. Second, according to the Lipschitz continuity w.r.t the ℓ_∞ norm, the empirical ℓ_∞ norm covering number of $\mathcal{F}$ can be bounded by the empirical ℓ_∞ norm covering number of $\mathcal{P}(\mathcal{F})$. Third, the empirical ℓ_∞ norm covering number of $\mathcal{P}(\mathcal{F})$ can be bounded by the fat-shattering dimension, and the fat-shattering dimension can be bounded by the worst-case Rademacher complexity of $\mathcal{P}(\mathcal{F})$. Hence, the problem is transferred to the estimation of the worst-case Rademacher complexity. Finally, we estimate the lower bound of the worst-case Rademacher complexity of $\mathcal{P}(\mathcal{F})$, and then combined with the above steps, the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded.

We first introduce the following lemmas:

Lemma D.4 (Lemma A.2 in (Srebro et al., 2010)). *For any function class $\mathcal{F}$, any S with a finite sample of size n and any $\epsilon > 2\hat{\mathfrak{R}}_S(\mathcal{F})$, we have that*

$$\mathrm{fat}_\epsilon(\mathcal{F}) \leq \frac{16n\hat{\mathfrak{R}}_S^2(\mathcal{F})}{\epsilon^2}.$$

Lemma D.5 (Theorem 4.4 in (Rudelson & Vershynin, 2006)). *For any function class $\mathcal{F}$, any S with a finite sample of size n, and any $\eta \in (0,1)$ there exist constants $0 < a < 1$ and $E > 0$ such that for all $\epsilon \in (0,1)$,*

$$\log \mathcal{N}_\infty(\epsilon, \mathcal{F}, S) \leq Ev \log(en/v\epsilon) \log^\eta(en/v),$$

where $v = \mathrm{fat}_{a\epsilon}(\mathcal{F})$.

Step 1: We first derive the relationship between the empirical ℓ_∞ norm covering number $\mathcal{N}_\infty(\epsilon, \mathcal{L}, D)$ and the empirical ℓ_∞ norm covering number $\mathcal{N}_\infty(\epsilon, \mathcal{P}(\mathcal{F}), [c] \times D)$.

For the dataset $D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ with n i.i.d. examples:

$$\begin{aligned}
&\max_i |L(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{y}_i) - L(\boldsymbol{f}'(\boldsymbol{x}_i), \boldsymbol{y}_i)| \\
\leq &\mu \max_i \|\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{f}'(\boldsymbol{x}_i)\|_\infty \quad \text{(Use Assumption 3.9)} \\
\leq &\mu \max_i \max_j |f_j(\boldsymbol{x}_i) - f_j'(\boldsymbol{x}_i)| \\
\leq &\mu \max_i \max_j |p_j(\boldsymbol{f}(\boldsymbol{x}_i)) - p_j(\boldsymbol{f}'(\boldsymbol{x}_i)|. \quad \text{(The definition of the projection function class } \mathcal{P}(\mathcal{F}))
\end{aligned}$$

Then, according to the definition of the empirical ℓ_∞ covering number, we have that an empirical ℓ_∞ cover of $\mathcal{P}(\mathcal{F})$ at radius ϵ/μ is also an empirical ℓ_∞ cover of the loss function space associated with $\mathcal{F}$ at radius ϵ, and we can conclude that:

$$\mathcal{N}_\infty(\epsilon, \mathcal{L}, D) \leq \mathcal{N}_\infty\left(\frac{\epsilon}{\mu}, \mathcal{P}(\mathcal{F}), [c] \times D\right). \tag{13}$$

Step 2: We show that the empirical ℓ_∞ norm covering number of $\mathcal{P}(\mathcal{F})$ can be bounded by the fat-shattering dimension, and the fat-shattering dimension can be bounded by the worst-case Rademacher complexity of $\mathcal{P}(\mathcal{F})$.

According to Lemma D.4, for any $\epsilon > 2\hat{\mathfrak{R}}_{[c]\times D}(\mathcal{P}(\mathcal{F}))$, we have

$$\mathrm{fat}_\epsilon(\mathcal{P}(\mathcal{F})) \leq \frac{16nc\hat{\mathfrak{R}}_{[c]\times D}^2(\mathcal{P}(\mathcal{F}))}{\epsilon^2}.$$

Then, combining with Lemma D.5, for any $\eta \in (0,1)$ there exist constants $0 < a < 1$ and $E > 0$ such that for all $\epsilon \in (0,1)$, we have

$$\log \mathcal{N}_\infty(\epsilon, \mathcal{P}(\mathcal{F}), [c] \times D) \leq Ev \log(enc/v\epsilon) \log^\eta(enc/v),$$

where $v = \mathrm{fat}_{a\epsilon}(\mathcal{P}(\mathcal{F})) \leq \frac{16nc\hat{\mathfrak{R}}^2_{[c]\times D}(\mathcal{P}(\mathcal{F}))}{a^2\epsilon^2}$, and we set $d = \frac{16nc\hat{\mathfrak{R}}^2_{[c]\times D}(\mathcal{P}(\mathcal{F}))}{a^2\epsilon^2}$. Hence,

$$\begin{aligned}
&\log \mathcal{N}_\infty(\epsilon, \mathcal{P}(\mathcal{F}), [c]\times D)\\
\leq& Ev \log \frac{enc}{v\epsilon} \log^\eta \frac{enc}{v}\\
\leq& Ed \log \frac{e^{2+\eta}nc}{d\epsilon} \log^\eta \frac{e^{2+\eta}nc}{d}\\
\leq& Ed \log^{1+\eta} \frac{e^{2+\eta}nc}{d\epsilon}\\
\leq& E\frac{16nc\hat{\mathfrak{R}}^2_{[c]\times D}(\mathcal{P}(\mathcal{F}))}{a^2\epsilon^2} \log^{1+\eta}\left(\frac{1.2}{\hat{\mathfrak{R}}_{[c]\times D}(\mathcal{P}(\mathcal{F}))}\right)^2, \quad \text{(Use inequality } 2\hat{\mathfrak{R}}_{[c]\times D}(\mathcal{P}(\mathcal{F})) < \epsilon < 1\text{)}
\end{aligned} \tag{14}$$

where we use that for any $s, t > 0$, the function $x \mapsto x \log \frac{s}{x} \log^\eta \frac{t}{x}$ is non-decreasing as long as $s > t > e^{1+\eta}x$ and that $v \leq nc$ in the second inequality, the third and fourth inequalities are obtained by direct calculation and proper scaling.

Then, we have

$$\begin{aligned}
&\log \mathcal{N}_\infty(\epsilon, \mathcal{P}(\mathcal{F}), [c]\times D)\\
\leq& E\frac{16nc\hat{\mathfrak{R}}^2_{[c]\times D}(\mathcal{P}(\mathcal{F}))}{a^2\epsilon^2} \log^{1+\eta}\left(\frac{1.2}{\hat{\mathfrak{R}}_{[c]\times D}(\mathcal{P}(\mathcal{F}))}\right)^2\\
\leq& E\frac{64nc\hat{\mathfrak{R}}^2_{[c]\times D}(\mathcal{P}(\mathcal{F}))}{a^2\epsilon^2} \log^{1+\eta}\left(\frac{1.2}{\hat{\mathfrak{R}}_{[c]\times D}(\mathcal{P}(\mathcal{F}))}\right)\\
\leq& E\frac{64nc\hat{\mathfrak{R}}^2_{[c]\times D}(\mathcal{P}(\mathcal{F}))}{a^2\epsilon^2} \log^{1+\eta}\frac{8\sqrt{nc}}{5B} \quad \text{(Use inequality (10))}\\
\leq& E\frac{64nc\widetilde{\mathfrak{R}}^2_{nc}(\mathcal{P}(\mathcal{F}))}{a^2\epsilon^2} \log^{1+\eta}\frac{8\sqrt{nc}}{5B} \quad \text{(The definition of the worst-case Rademacher complexity)}
\end{aligned} \tag{15}$$

Step 3: According to Lemma C.5 and combined with the above steps, we have

$$\begin{aligned}
&\hat{\mathfrak{R}}_D(\mathcal{L})\\
\leq& \inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^M \sqrt{\log \mathcal{N}_\infty(\epsilon, \mathcal{L}, D)}d\epsilon\right)\\
\leq& \inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^M \sqrt{\log \mathcal{N}_\infty(\frac{\epsilon}{\mu}, \mathcal{P}(\mathcal{F}), [c]\times D)}d\epsilon\right) \quad \text{(Use inequality (14))}\\
\leq& \inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^M \sqrt{\frac{64E\mu^2nc\widetilde{\mathfrak{R}}^2_{nc}(\mathcal{P}(\mathcal{F}))}{a^2\epsilon^2} \log^{1+\eta}\frac{8\sqrt{nc}}{5B}}d\epsilon\right) \quad \text{(Use inequality (15))}\\
\leq& \inf_{\alpha>0}\left(4\alpha + \frac{96\sqrt{Ec}\mu}{a}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\log^{\frac{1+\eta}{2}}(\frac{8\sqrt{nc}}{5B})\int_\alpha^M \epsilon^{-1}d\epsilon\right)\\
\leq& \frac{96\sqrt{Ec}\mu}{a}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F})) + \frac{96\sqrt{Ec}\mu}{a}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\log^{\frac{1+\eta}{2}}(\frac{8\sqrt{nc}}{5B})\cdot\log\frac{aM}{24\sqrt{Ec}\mu\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))}\\
&\text{(Choose } \alpha = \frac{24\sqrt{Ec}\mu}{a}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\text{)}\\
\leq& \frac{96\sqrt{Ec}\mu}{a}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F})) + \frac{96\sqrt{Ec}\mu}{a}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\log^{\frac{1+\eta}{2}}(\frac{8\sqrt{nc}}{5B})\cdot\log\frac{M\sqrt{n}}{\sqrt{E}\mu B} \quad \text{(Use } 0<a<1 \text{ and inequality (10))}\\
=& \frac{96\sqrt{Ec}\mu}{a}\widetilde{\mathfrak{R}}_{nc}(\mathcal{P}(\mathcal{F}))\left(1 + \log^{\frac{1+\eta}{2}}(\frac{8\sqrt{nc}}{5B})\cdot\log\frac{M\sqrt{n}}{\sqrt{E}\mu B}\right).
\end{aligned}$$

D.6. Proof of Theorem 3.17

According to the inequality (11), we have $\widetilde{\Re}_{nc}(\mathcal{P}(\mathcal{F})) \leq \frac{B}{\sqrt{nc}}$, then

$$\hat{\Re}_D(\mathcal{F}) \leq \frac{96\sqrt{E}\mu B}{a}\frac{1}{\sqrt{n}} \times \left(1 + \log^{\frac{1+\eta}{2}}(\frac{8\sqrt{nc}}{5B}) \cdot \log\frac{M\sqrt{n}}{\sqrt{E}\mu B}\right).$$

Combining with (7), then

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + \frac{192\sqrt{E}\mu B}{a}\frac{1}{\sqrt{n}} \times \left(1 + \log^{\frac{1+\eta}{2}}(\frac{8}{5B}\sqrt{nc}) \cdot \log(\frac{M\sqrt{n}}{\sqrt{E}\mu B})\right) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

E. Generalization Analysis for Macro-Averaged AUC

E.1. Proof of Theorem 4.2

We first proof the following lemma:

Lemma E.1. *Let $q_\tau : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be real-valued functions indexed by $\tau \in T$ where T is some set. If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$ and $\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_t$ are i.i.d., $r = \min\{s, t\}$, then for any convex non-decreasing function ψ,*

$$\mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{st}\sum_{i=1}^{s}\sum_{j=1}^{t} q_\tau\left(\boldsymbol{x}_i, \boldsymbol{x}'_j\right)\right) \leq \mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{r}\sum_{i=1}^{r} q_\tau\left(\boldsymbol{x}_i, \boldsymbol{x}'_i\right)\right).$$

Proof. The proof of this lemma is inspired by (Clémençon et al., 2008).

$$\begin{aligned}
&\mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{st}\sum_{i=1}^{s}\sum_{j=1}^{t} q_\tau\left(\boldsymbol{x}_i, \boldsymbol{x}'_j\right)\right)\\
=&\mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{s!}\sum_{\pi_{\boldsymbol{x}}}\frac{1}{t!}\sum_{\pi_{\boldsymbol{x}'}}\frac{1}{r}\sum_{i=1}^{r} q_\tau\left(\boldsymbol{x}_{\pi(i)}, \boldsymbol{x}'_{\pi(i)}\right)\right)\\
\leq&\mathbb{E}\psi\left(\frac{1}{s!}\sum_{\pi_{\boldsymbol{x}}}\frac{1}{t!}\sum_{\pi_{\boldsymbol{x}'}}\sup_{\tau\in T}\frac{1}{r}\sum_{i=1}^{r} q_\tau\left(\boldsymbol{x}_{\pi(i)}, \boldsymbol{x}'_{\pi(i)}\right)\right) \quad (\psi \text{ is nondecreasing})\\
\leq&\frac{1}{s!}\sum_{\pi_{\boldsymbol{x}}}\frac{1}{t!}\sum_{\pi_{\boldsymbol{x}'}}\mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{r}\sum_{i=1}^{r} q_\tau\left(\boldsymbol{x}_{\pi(i)}, \boldsymbol{x}'_{\pi(i)}\right)\right) \quad (\text{Jensen's inequality})\\
=&\mathbb{E}\psi\left(\sup_{\tau\in T}\frac{1}{r}\sum_{i=1}^{r} q_\tau\left(\boldsymbol{x}_i, \boldsymbol{x}'_i\right)\right).
\end{aligned}$$

□

With this lemma, we first prove that $\mathbb{E}\left[R\left(\hat{\boldsymbol{f}}^*\right)\right] - R^* \leq \frac{384\sqrt{2E}\mu B}{a}\sqrt{\frac{c}{r_0}}\left(1 + \log^{\frac{1+\eta}{2}}(\frac{2\sqrt{2n}}{B}) \cdot \log(\frac{M}{\sqrt{E}\mu B}\sqrt{\frac{n}{c}})\right)$.

$$\begin{aligned}
&\mathbb{E}\left[R\left(\hat{\boldsymbol{f}}^*\right)\right] - R^* = \mathbb{E}\left[R\left(\hat{\boldsymbol{f}}^*\right) - \widehat{R}_D(\hat{\boldsymbol{f}}^*) + \widehat{R}_D(\hat{\boldsymbol{f}}^*) - R^*\right]\\
=&\mathbb{E}\left[R\left(\hat{\boldsymbol{f}}^*\right) - \widehat{R}_D(\hat{\boldsymbol{f}}^*)\right] + \mathbb{E}\left[\widehat{R}_D(\hat{\boldsymbol{f}}^*) - R^*\right]\\
\leq&\mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\left|R\left(\boldsymbol{f}\right) - \widehat{R}_D(\boldsymbol{f})\right| + \mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\left|\widehat{R}_D(\boldsymbol{f}) - R\left(\boldsymbol{f}\right)\right| = 2\mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\left|R\left(\boldsymbol{f}\right) - \widehat{R}_D(\boldsymbol{f})\right|\\
\leq&2\mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\left|R\left(\boldsymbol{f}\right) - \frac{1}{c}\sum_{j=1}^{c}\frac{1}{r_j}\sum_{i=1}^{r_j}\ell\left(f_j(\boldsymbol{x}_i) - f_j(\boldsymbol{x}'_i)\right)\right|. \quad (\text{Use Lemma E.1})
\end{aligned}$$

Let $\overline{D} = \{\overline{\boldsymbol{x}}_1, \ldots, \overline{\boldsymbol{x}}_n\}$ be an independent copy of $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, then by the standard symmetrization technique and the Jensen's inequality similar to Preliminaries in Section B, the above inequality can be bounded by:

$$
\begin{aligned}
&2\mathbb{E}_{D,\overline{D}} \sup_{\boldsymbol{f}\in\mathcal{F}} \left| \frac{1}{c}\sum_{j=1}^{c}\frac{1}{r_j}\sum_{i=1}^{r_j} \ell\left(f_j(\overline{\boldsymbol{x}}_i) - f_j(\overline{\boldsymbol{x}}_i')\right) - \frac{1}{c}\sum_{j=1}^{c}\frac{1}{r_j}\sum_{i=1}^{r_j} \ell\left(f_j(\boldsymbol{x}_i) - f_j(\boldsymbol{x}_i')\right) \right| \\
=&2\mathbb{E}_{D,\overline{D},\boldsymbol{\epsilon}} \sup_{\boldsymbol{f}\in\mathcal{F}} \left| \frac{1}{c}\sum_{j=1}^{c}\frac{1}{r_j}\sum_{i=1}^{r_j} \epsilon_{ij}\left[\ell\left(f_j(\overline{\boldsymbol{x}}_i) - f_j(\overline{\boldsymbol{x}}_i')\right) - \ell\left(f_j(\boldsymbol{x}_i) - f_j(\boldsymbol{x}_i')\right)\right] \right| \\
=&4\mathbb{E}_{D,\boldsymbol{\epsilon}} \sup_{\boldsymbol{f}\in\mathcal{F}} \frac{1}{c}\sum_{j=1}^{c}\frac{1}{r_j}\sum_{i=1}^{r_j} \epsilon_{ij}\ell\left(f_j(\boldsymbol{x}_i) - f_j(\boldsymbol{x}_i')\right) \\
=&4\Re(\mathcal{F}) \\
\leq&\frac{192\sqrt{2E}\mu}{a}\sqrt{c}\max_j \widetilde{\Re}\left(\mathcal{F}_j\right) \times \left(1 + \log^{\frac{1+\eta}{2}}\left(\frac{2\sqrt{2r_j}}{B}\right)\cdot\log\left(\frac{M}{\sqrt{E}\mu B}\sqrt{\frac{r_j}{c}}\right)\right) \quad \text{(Use Lemma 3.11)} \\
=&\frac{192\sqrt{2E}\mu}{a}\sqrt{c}\max_j \left(1 + \log^{\frac{1+\eta}{2}}\left(\frac{2\sqrt{2r_j}}{B}\right)\cdot\log\left(\frac{M}{\sqrt{E}\mu B}\sqrt{\frac{r_j}{c}}\right)\right) \mathbb{E}_D \sup_{D\in\mathcal{X}^n} \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f_j\in\mathcal{F}_j}\frac{1}{r_j}\sum_{i=1}^{r_j}\epsilon_i\left(f_j(\boldsymbol{x}_i) - f_j(\boldsymbol{x}_i')\right)\right] \\
\leq&\frac{192\sqrt{2E}\mu}{a}\sqrt{c}\max_j \left(1 + \log^{\frac{1+\eta}{2}}\left(\frac{2\sqrt{2r_j}}{B}\right)\cdot\log\left(\frac{M}{\sqrt{E}\mu B}\sqrt{\frac{r_j}{c}}\right)\right) \mathbb{E}_D \sup_{D\in\mathcal{X}^n}\sup_{f_j\in\mathcal{F}_j}\frac{1}{r_j}\left(\sum_{i=1}^{r_j}\left(f_j(\boldsymbol{x}_i) - f_j(\boldsymbol{x}_i')\right)^2\right)^{\frac{1}{2}} \\
&\text{(Use Lemma C.2)} \\
\leq&\frac{384\sqrt{2E}\mu B}{a}\max_j\sqrt{\frac{c}{r_j}}\left(1 + \log^{\frac{1+\eta}{2}}\left(\frac{2\sqrt{2r_j}}{B}\right)\cdot\log\left(\frac{M}{\sqrt{E}\mu B}\sqrt{\frac{r_j}{c}}\right)\right) \\
\leq&\frac{384\sqrt{2E}\mu B}{a}\sqrt{\frac{c}{r_0}}\left(1 + \log^{\frac{1+\eta}{2}}\left(\frac{2\sqrt{2n}}{B}\right)\cdot\log\left(\frac{M}{\sqrt{E}\mu B}\sqrt{\frac{n}{c}}\right)\right).
\end{aligned}
$$

Similarly, we can derive that

$$
\begin{aligned}
&R\left(\hat{\boldsymbol{f}}^*\right) - R^* \\
=&R\left(\hat{\boldsymbol{f}}^*\right) - \widehat{R}_D(\hat{\boldsymbol{f}}^*) + \widehat{R}_D(\hat{\boldsymbol{f}}^*) - R^* \\
\leq&\sup_{\boldsymbol{f}\in\mathcal{F}}[R\left(\boldsymbol{f}\right) - \widehat{R}_D(\boldsymbol{f})] + \sup_{\boldsymbol{f}\in\mathcal{F}}[\widehat{R}_D(\boldsymbol{f}) - R\left(\boldsymbol{f}\right)] \\
=&2\sup_{\boldsymbol{f}\in\mathcal{F}}[R\left(\boldsymbol{f}\right) - \widehat{R}_D(\boldsymbol{f})].
\end{aligned}
$$

Let $D' = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k', \ldots, \boldsymbol{x}_n\}$ be the training dataset with only one sample different from D, where the k-th sample is replaced by $\boldsymbol{x}_k'$. Then

$$
\begin{aligned}
&\sup_{\boldsymbol{f}\in\mathcal{F}}[R(\boldsymbol{f}) - \widehat{R}_{D'}(\boldsymbol{f})] - \sup_{\boldsymbol{f}\in\mathcal{F}}[R(\boldsymbol{f}) - \widehat{R}_D(\boldsymbol{f})] \\
\leq&\sup_{\boldsymbol{f}\in\mathcal{F}}[\widehat{R}_D(\boldsymbol{f}) - \widehat{R}_{D'}(\boldsymbol{f})] \\
\leq&\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{c}\sum_{j=1}^{c}\frac{1}{\left|X_j^+\right|\left|X_j^-\right|}\sum_{\boldsymbol{x}_i'\in X_j^-}\left(\ell\left(f_j(\boldsymbol{x}_k) - f_j(\boldsymbol{x}_i')\right) - \ell\left(f_j(\boldsymbol{x}_k) - f_j(\boldsymbol{x}_i')\right)\right) \\
\leq&\frac{M}{r_0}.
\end{aligned}
$$

According to McDiarmid's inequality, for any $0 < \delta < 1$, with probability at least $1 - \delta$ over the training dataset D, the following holds:

$$R(\hat{\boldsymbol{f}}^*) - R^* \leq \frac{384\sqrt{2E}\mu B}{a}\sqrt{\frac{c}{r_0}}\left(1 + \log^{\frac{1+\eta}{2}}(\frac{2\sqrt{2n}}{B}) \cdot \log(\frac{M}{\sqrt{E}\mu B}\sqrt{\frac{n}{c}})\right) + 2M\sqrt{\frac{\log\frac{1}{\delta}}{r_0}}.$$

Note that when we analyze class-imbalance, we focus on whether s_j and t_j are seriously imbalanced for a fixed number of n examples. Hence, the order of the bound is $\widetilde{O}(\sqrt{\frac{c}{r_0}})$.