GENERALIZATION ANALYSIS FOR MULTI-DIMENSIONAL CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent studies have experimentally demonstrated that multi-dimensional classification (MDC), where each example is represented by a single instance while associated with multiple class variables, can achieve empirical success in many application scenarios, however, the problem of theoretically understanding MDC remains under-explored. In an attempt to make up for the gap in the generalization theory of MDC, we establish an effective unified theoretical framework, which makes it possible to analyze the generalization of MDC using existing theoretical results. However, the generalization bounds induced by existing theoretical results are vacuous for MDC. To this end, we introduce two decomposition operators and exploit the Lipschitz continuity of loss functions to develop novel vector-contraction inequalities, and derive tight generalization bounds with a logarithmic dependency on the maximum number of classes and with square-root and logarithmic dependency on the number of dimensions, respectively. In addition, we derive generalization bounds for prediction models of different architectures, which reveal the impact of various architectures on generalization bounds. Our tight generalization bounds without strong assumptions explain the good generalization ability of MDC.

1 INTRODUCTION

A common problem in traditional supervised learning is to learn a model from the objects to handle the related task, where each object is represented by a single instance associated with a single class variable, e.g., multi-class classification. However, with the development of real-world applications, the rich semantics of objects need to be characterized from multiple different dimensions. For example, e-commerce websites need to show customers smartphones from different dimensions (e.g., *phone brand, operating system, CPU brand, screen resolution, camera pixels*, etc.) so that customers can fully consider and choose their favorite products. In fact, similar scenarios widely exist in real-word applications such as text classification (Lertnattee & Theeramunkong, 2004; Shatkay et al., 2008; Serafino et al., 2015), computer vision (Song et al., 2018; Lian et al., 2020; Shi et al., 2025), bioinformatics (Rodríguez et al., 2012; Borchani et al., 2013; Fernandez Gonzalez et al., 2015) and other fields (Muktadir et al., 2019). The problems corresponding to these special scenarios can be naturally formalized in the multi-dimensional classification (MDC) learning framework (Read et al., 2014; Jia & Zhang, 2021a; 2023; Huang et al., 2024), where the rich semantics of objects are characterized by multiple class variables, each of which corresponds to a class space that characterizes the semantics of objects from a certain dimension. Although MDC has made impressive progress in many fields, the reasons for the good generalization performance of MDC methods have so far remained completely elusive.

A common idea to solving MDC is to decompose it into q multi-class classification problems. As we all know, for multi-label learning, a common strategy is to decompose the multi-label learning problem into multiple binary classification problems. Therefore, multi-label learning can be regarded as a degenerate case of MDC to some extent. This motivates a natural question about how to theoretically analyze the generalization of MDC:

Can existing theoretical analysis methods be directly used to conduct generalization analysis for MDC? Is it necessary to develop new analysis methods or theoretical tools to obtain a tight generalization bound for MDC?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Here we attempt to answer this question through the systematic theoretical analysis. Based on the unified theoretical framework we established, the answer is that although existing theoretical results can induce bounds, they cannot provide general theoretical guarantees for MDC datasets and methods. We do need to develop new general and effective analysis methods to study the generalization of MDC.

Existing research so far lacks theoretical analysis of MDC, and the study of generalization ability is an important perspective to understand MDC theoretically. Effective generalization analysis of MDC should address two challenges: 1) the establishment of a unified framework that can induce general analysis methods for MDC, and 2) the reduction of the dependency of the bounds on the number of dimensions and the number of classes in each dimension. *First*, since the lack of theoretical analysis, the establishment of a unified framework for MDC requires not only considering the formal definition of multiple dimensions but also considering the formal definition of the multiple classes involved in each dimension. In addition, since the problem setting of MDC is more complex than traditional supervised learning, the theoretical framework also needs to have the potential to help develop general theoretical analysis methods. *Second*, since MDC involves multiple dimensions, and each dimension corresponds to a multi-class classification problem, intuitively, the increase in the number of dimensions and the number of classes in each dimension will affect the difficulty of learning, but the empirical success of MDC methods suggests that these increases have limited impact on the difficulty of learning (Jia & Zhang, 2021a; 2023), which means that the ideal bound should not be strongly dependent on the number of dimensions and the number of classes. Hence, how to reduce the dependency of the bounds on the number of dimensions and the number of classes in each dimension is a critical problem for the theoretical analysis of MDC.

In this paper, we first establish an effective unified theoretical framework and show that the bound with a linear dependency on the number of dimensions (i.e., q) and the maximum number of classes (i.e., k) induced by existing theoretical results is not a tight result. Then, we develop novel analysis methods to derive tighter bounds for MDC. Specifically, we develop a novel vector-contraction inequality for ℓ_2 Lipschitz loss by introducing two decomposition operators, and derive the bound with a square-root dependency on q and no dependency on k, up to logarithmic terms, for general MDC classes. We also develop a novel vector-contraction inequality for ℓ_∞ Lipschitz loss, and derive the bound with no dependency on q and k, up to logarithmic terms, which is the state of the art theoretical result and can provide general theoretical guarantees for MDC. In addition, we also analyze the bounds for different prediction models, and we show that models with various architectures for MDC will affect the bounds.

To our knowledge, this is the first formal attempt to extend generalization analysis to the case of MDC and derive capacity-based bounds for MDC with mild dependency on q and k. Major contributions of the paper include:

- We establish an effective unified theoretical framework for MDC, and reveal that existing theoretical results are insufficient for generalization analysis of MDC.
- We introduce two decomposition operators, and combined with different Lipschitz properties of losses, we derive the state of the art bounds with a weaker dependency on q, k.
- We derive bounds for different prediction models, which reveals the impact of various architectures on the generalization analysis. The theoretical results on FNN and Transformer here serve as promising theoretical tools for the generalization analysis of MDC.

2 PRELIMINARIES

Let $[n] := \{1, \ldots, n\}$ for any natural number n. Let $\mathcal{X} \subseteq \mathbb{R}^d$ denotes the d-dimensional input space and $\mathcal{Y} = \mathcal{C}_1 \times \mathcal{C}_2 \times \cdots \times \mathcal{C}_q$ denotes the output space, where $\mathcal{Y}$ corresponds to the Cartesian product of q class spaces $\mathcal{C}_j = \left\{c_1^j, c_2^j, \ldots, c_{K_j}^j\right\} (j \in [q])$ which consists of K_j possible class labels respectively. Given a dataset $D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ with n examples which are i.i.d. from a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, where $\boldsymbol{y}_i = \left[y_i^1, y_i^2, \ldots, y_i^q\right]^\top \in \mathcal{Y}$ is the q-dimensional class vector associated with $\boldsymbol{x}_i$ with each component y_i^j representing one possible class label in $\mathcal{C}_j$, i.e., $y_i^j \in \mathcal{C}_j$. The task of MDC is to learn a vector-valued prediction function $\boldsymbol{f} = (f_1, \ldots, f_q) : \mathcal{X} \mapsto \mathcal{Y}$ from the dataset D which can return a proper class vector for an unseen instance.

For any function $\boldsymbol{f} : \mathcal{X} \mapsto \mathcal{Y}$, the prediction quality on the example $(\boldsymbol{x}, \boldsymbol{y})$ is measured by a loss function $\ell : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}_+$. The goal of MDC is to find 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: $R(\boldsymbol{f}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim P}[\ell(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y})]$. We denote the empirical risk w.r.t. the training dataset D as $\widehat{R}_D(\boldsymbol{f}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{y}_i)$. In addition, we define the loss function space as $\mathcal{L} = \{\ell(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) : \boldsymbol{f} \in \mathcal{F}\}$. 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.

2.1 Multi-Dimensional Classification

The MDC problem can be solved by training a number of independent multi-class classifiers, one per dimension. In other words, we can decompose the MDC problem into q multi-class classification problems, and the prediction function of each multi-class classification problem is f_j, $j \in [q]$, which consists of K_j class labels. For the convenience of representation in analysis, we denote the maximum number of class labels in q class spaces as $k := \max_j K_j$. Without loss of generality, we can assume that each multi-class classification problem contains k possible class labels. For the class space with $K_j < k$ class labels, the classes from $K_j + 1$ to k can be set to empty.

For each multi-class classification problem, our goal is to learn a vector-valued function $\boldsymbol{h}^j := (h_1^j, \ldots, h_k^j)$ from a function class $\mathcal{H}^j$ for prediction, and the j-th multi-class classification prediction is determined by the rule $\hat{y}^j = c_{s^*}^j$, $s^* = \arg\max_s h_s^j(\boldsymbol{x})$, $s \in [k]$. Hence, we consider the prediction function for each multi-class classification problem of the general form $f_j(\boldsymbol{x}) = \max_{s\in[k]}\{h_s^j(\boldsymbol{x})\} = \max_{s\in[k]}\{{\boldsymbol{w}_s^j}^\top \phi_s^j(\boldsymbol{x})\}$, where ϕ_s^j represents a nonlinear mapping. We define the function class corresponding to g as $\mathcal{G} = \{\boldsymbol{t} \mapsto g(\boldsymbol{t}) : g(\boldsymbol{t}) = \max_{s\in[k]} t_s, \boldsymbol{t} = (t_1, \ldots, t_s, \ldots, t_k) \in \mathbb{R}^k\}$. Then, the general form for each multi-class classification problem can be denoted as $f_j(\boldsymbol{x}) = g(\boldsymbol{h}^j(\boldsymbol{x}))$. We define a vector-valued function class of the MDC as follows:

$$\mathcal{F} = \{\boldsymbol{x} \mapsto \boldsymbol{f}(\boldsymbol{x}) : \boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_q(\boldsymbol{x})), f_j(\boldsymbol{x}) = g(\boldsymbol{h}^j(\boldsymbol{x})) = \max_{s\in[k]}\{h_s^j(\boldsymbol{x})\}, \\ g \in \mathcal{G}, \boldsymbol{h}^j \in \mathcal{H}^j, \boldsymbol{x} \in \mathcal{X}, j \in [q], s \in [k]\}, \tag{1}$$

We also define a function class $\mathcal{H}^j$ for each multi-class classification corresponding to the MDC as:

$$\mathcal{H}^j = \{\boldsymbol{x} \mapsto \boldsymbol{h}^j(\boldsymbol{x}) : \boldsymbol{h}^j(\boldsymbol{x}) = (h_1^j(\boldsymbol{x}), \ldots, h_k^j(\boldsymbol{x})), h_s^j(\boldsymbol{x}) = \langle \boldsymbol{w}_s^j, \phi_s^j(\boldsymbol{x})\rangle, \boldsymbol{x} \in \mathcal{X}, s \in [k], \\ \boldsymbol{w}^j = (\boldsymbol{w}_1^j, \ldots, \boldsymbol{w}_k^j) \in \mathbb{R}^{d\times k}, \alpha(\boldsymbol{w}^j) \le \Lambda, \beta(\phi_s^j(\boldsymbol{x})) \le A, \Lambda > 0, A > 0\}, \tag{2}$$

where α represents a functional that constrains weights, β represents a functional that constrains nonlinear mappings. The functionals α and β here indicate that $\boldsymbol{w}_s^j$ and $\phi_s^j(\boldsymbol{x})$ are norm-bounded. We further show some concrete examples of MDC models corresponding to the definition of classes to provide more motivations, justifications, and practical explanations and to clarify the setup.

For example, the maximun margin MDC method named M^3MDC (Jia & Zhang, 2022), which proposes to use a kernelized linear model with the weight matrix satisfying the F-norm constraint for each dimension to deal with the MDC problem, can be expressed in our function classes, where h_s^j is a kernelized linear model and $\phi_s^j(\boldsymbol{x})$ is the kernel mapping. The method also introduces the column covariance matrix of the whole weight matrix $\boldsymbol{w}$ to model the relationships among all components $\boldsymbol{w}^j$ in $\boldsymbol{w}$, which can be considered as an additional assumption.

For the MDC method named SLEM (Jia & Zhang, 2021a) that models pairwise class variables dependencies, which first considers pairwise dependencies by training a total of $\binom{q}{2}$ classifiers, one per a pair of class spaces, and then each pairwise class space is composed of all possible class labels in the two initial class spaces, i.e., the possible class labels in each pairwise class space are k^2. SLEM learns a multi-output regressor with F-norm bounded weight matrix, i.e., $h_s^j(\boldsymbol{x}) = {\boldsymbol{w}_s^j}^\top \boldsymbol{x}$, and it also introduces an additional sparsity assumption. In addition, the MDC method named PIST (Huang et al., 2024) adopts a similar idea to model pairwise class variables dependencies, but PIST learns a network for classification in each dimension, i.e., $h_s^j(\boldsymbol{x}) = {\boldsymbol{w}_s^j}^\top \mathrm{softmax}({W_s^j}^3 \sigma[{W_s^j}^2(\mathrm{FNN}(\boldsymbol{x}) \odot \sigma({W_s^j}^1 \boldsymbol{l}))])$, where σ is the ReLU activation, $\odot$ is the Hadamard product, $\boldsymbol{l}$ is the latent label embedding vector.

Besides, the function classes here are applicable to the typical Binary Relevance methods for MDC, where different methods correspond to different nonlinear mappings ϕ_s^j, for example, the nonlinear mappings ϕ_s^j in the MDC method named DCOM (Huang et al., 2025) correspond to FNNs.

2.2 Related Evaluation Metrics

Several evaluation metrics are proposed to measure the generalization performance of different MDC methods. Here we focus on two widely used evaluation metrics, i.e., Hamming Score and Exact Match, and give the definitions of their corresponding surrogate losses. The Hamming Score is defined as $\mathrm{HS}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\sum_{j=1}^{q}\mathbb{I}_{y_i^j=\hat{y}_i^j}$, where $\mathbb{I}_\pi$ denotes the indicator function, which returns 1 if π holds and 0 otherwise. Maximizing the Hamming Score is equivalent to optimizing the **Hamming Loss**: $\ell_H^{0\text{-}1}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \frac{1}{q}\sum_{j=1}^{q}\mathbb{I}_{y^j\neq\hat{y}^j}$, and its surrogate loss is defined as $\ell_H(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \frac{1}{q}\sum_{j=1}^{q}\ell_b\left(f_j(\boldsymbol{x}), y^j\right)$, where the base convex surrogate loss ℓ_b can be various popular forms, such as the hinge loss, the logistic loss, and the exponential loss.

The Exact Match is defined as $\mathrm{EM}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{(\sum_{j=1}^{q}\mathbb{I}_{y_i^j=\hat{y}_i^j})=q}$. Maximizing the Exact Match is equivalent to optimizing the **Matching Loss**: $\ell_M^{0\text{-}1}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \max_{j\in[q]}\left\{\mathbb{I}_{\hat{y}^j\neq y^j}\right\}$, and its surrogate loss is defined as $\ell_M(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) = \max_{j\in[q]}\left\{\ell_b\left(f_j(\boldsymbol{x}), y^j\right)\right\}$.

2.3 Related Complexity Measures

Here we introduce the related complexity measures involved in our theoretical results. The Rademacher complexity is used to perform generalization analysis for MDC.

Definition 1 (Rademacher complexity)**.** *Let $\mathcal{F}$ be a real-valued function class 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 $\mathcal{F}$ is defined by $\hat{\mathfrak{R}}_D(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f\left(\boldsymbol{x}_i\right)\right]$, where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables. We define the worst-case Rademacher complexity as $\tilde{\mathfrak{R}}_n(\mathcal{F}) = \sup_{D\in\mathcal{X}^n}\hat{\mathfrak{R}}_D(\mathcal{F})$.*

The vector-valued predictive function class $\mathcal{F}$ of MDC makes traditional analysis methods developed for the Rademacher complexity of scalar-valued function class invalid. According to vector-contraction lemma in Maurer (2016), we can apparently prove that Rademacher complexity of a loss space associated with $\mathcal{F}$ can be bounded by a term with a linear dependency on the product of q and k, and analytically show that such bounds are not relatively tight results. Hence, we need new analysis tools to convert the Rademacher complexity of a loss space into the Rademacher complexity of a tractable scalar-valued function class for MDC. Rademacher complexity can be bounded by other scale-sensitive complexity measures, e.g. covering number and fat-shattering dimension (Srebro et al., 2010; Zhang & Zhang, 2023). **The relevant definitions are provided in the Appendix.**

3 Tight Bounds for MDC with ℓ_2 Lipschitz Loss

In this section, we first introduce the assumptions used in generalization analysis, i.e., the boundedness assumption of the functions and the Lipschitz property assumption of the loss function, and show that, benefiting from the efficient definitions of general function classes for MDC, we can use existing theoretical results (i.e., vector-contraction lemma in Maurer (2016)) to obtain a basic bound with a linear dependency on qk for MDC, and we show that such a bound induced by existing theoretical results is not a relatively tight result. Then, we introduce two decomposition operators and develop a novel vector-contraction inequality for MDC. Finally, with the novel vector-contraction inequality for ℓ_2 Lipschitz loss, we derive the bounds with a square-root dependency on q and no dependency on k, up to logarithmic terms, for general function classes of MDC, which improves the bounds induced by existing theoretical results (e.g., typical vector-contraction lemma in Maurer (2016); Wu & Zhu (2020) or Sudakov's minoration and the relationship between Rademacher and Gaussian complexity in Zhang & Zhang (2024)) by the factor of $\sqrt{q}k$ and $\sqrt{k}$. **The proof sketches and detailed proofs of the theoretical results in this paper are provided in the Appendix.**

Assumption 2. *Assume that the loss function and the prediction function for each multi-class classification are bounded: $\ell(\cdot,\cdot) \leq M$, $|f_j(\cdot)| \leq B$ for $j \in [q]$ where $M, B > 0$ are constants.*

Assumption 2 is a mild assumption. In fact, when we consider the classes (1) and (2) for MDC, we often use the assumptions $\|\boldsymbol{w}_s^j\|_2 \leq \Lambda$, ϕ_s^j is ρ-Lipschitz w.r.t. the ℓ_2 norm and $\|\boldsymbol{x}_i\|_2 \leq R$ for any $j \in [q]$, $s \in [k]$, $i \in [n]$ to replace the boundedness of the prediction function for each multi-class classification. Hence, we have $|h_s^j(\boldsymbol{x})| \leq \Lambda\rho R$ and $|f_j(\boldsymbol{x})| \leq \Lambda\rho R$ for any $j \in [q]$, $s \in [k]$ according to the definitions of classes (1) and (2), i.e., $B := \Lambda\rho R$. The above explanation is further explicitly stated by the following additional assumption:

Assumption 3 (Norm-bounded decomposed prediction)**.** *For each output dimension j and each class s for this dimension, the coordinate prediction $h_s^j(\boldsymbol{x})$ can be written as $h_s^j(\boldsymbol{x}) = \boldsymbol{w}_s^{j\top}\phi_s^j(\boldsymbol{x})$, where: 1) The weight vector satisfies $\|\boldsymbol{w}_s^j\|_2 \leq \Lambda$ for some constant $\Lambda > 0$; 2) The nonlinear mapping vector satisfies $\|\phi_s^j(\boldsymbol{x})\|_2 \leq A$ for all input $\boldsymbol{x}$ and some constant $A > 0$.*

Assumption 3 guarantees the boundedness of the coordinate predictions, and thus the bounded domain of the supremum. Assumption 3 is extremely mild in practice: standard regularization techniques such as weight decay and feature normalization exactly enforce these norm bounds in practical deep models for MDC.

Since multi-label learning is a degenerate case of MDC to some extent, it is natural to try to use existing theoretical results or analysis methods in multi-label learning to analyze the generalization of MDC. We reveal that using existing theoretical results (e.g., typical vector-contraction lemma in Maurer (2016); Wu & Zhu (2020) or Sudakov's minoration and the relationship between Rademacher and Gaussian complexity in Zhang & Zhang (2024)) can only obtain bounds with linearly or square-root dependency on qk. These bound are relatively loose and cannot provide a general theoretical guarantee for existing MDC methods. The use of existing theoretical results is accompanied by the following assumption on the Lipschitz continuity of loss:

Assumption 4. *Assume that the loss ℓ is μ-Lipschitz w.r.t. ℓ_2 norm, i.e., $|\ell(\boldsymbol{f}(\boldsymbol{x}), \cdot) - \ell(\boldsymbol{f}'(\boldsymbol{x}), \cdot)| \leq \mu \|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x})\|_2$, where $\mu > 0$, $\|\boldsymbol{t}\|_2 = \sqrt{\sum_{j=1}^q |t_j|^2}$ for $\boldsymbol{t} = (t_1, \ldots, t_q)$.*

Proposition 5 illustrates that Assumption 4 is very mild since the losses analyzed here all satisfy it.

Proposition 5. *Assume that the base loss ℓ_b is μ-Lipschitz continuous, then the surrogate Hamming Loss is $\frac{\mu}{\sqrt{q}}$-Lipschitz w.r.t. ℓ_2 norm, the surrogate Matching Loss is $\mu\sqrt{q}$-Lipschitz w.r.t. ℓ_2 norm.*

Using the ℓ_2 Lipschitz continuity of the loss function, we have the following basic bound for MDC:

Corollary 6. *Let $\mathcal{F}$ be the class of the MDC defined by (1) and (2). Let Assumptions 2, 3 and 4 hold. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ over a dataset D of size n, the following holds for any $\boldsymbol{f} \in \mathcal{F}$: $R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + \frac{4\mu qkB}{\sqrt{n}} + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$.*

Remark 7. *In Subsection 2.1, we established a unified framework for MDC, i.e., the definitions of classes (1) and (2). Here we find that, benefiting from this effective unified framework, the generalization analysis of MDC can be reduced to the composition of the Rademacher complexities, i.e., the inequality in Maurer (2016) (i.e., $\mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_i \epsilon_i h_i(\boldsymbol{f}(\boldsymbol{x}_i)) \leq \sqrt{2}L\mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_{i,j}\epsilon_{ij}f_j(\boldsymbol{x}_i)$) used twice in succession can induce $\hat{\mathfrak{R}}_D(\mathcal{L}) \leq 2\mu qk\max_{j,s}\hat{\mathfrak{R}}_D(\mathcal{H}_s^j)$.*

Remark 8. *Our analysis in Corollary 6 implies that under the effective unified framework we established, we can use existing theoretical results to obtain bounds that are linearly dependent on qk. However, this kind of bound with a linear dependency on qk is relatively loose and difficult to explain empirical success of MDC, which cannot provide a general theoretical guarantee for MDC. To this end, we will explain this failure from two perspectives. On the one hand, the classic bounds (e.g., a linear dependency on qk) of MDC scale rather unfavorably with qk. This is probably because the classic vector-contraction lemma was not formulated with consideration for the case where a large number of label classes are required. In the era of big data, multi-class learning problems (e.g., text or image classification) may involve tens or even hundreds of thousands of classes (Varma & Langford, 2013; Bhatia et al., 2016). This results in a lack of theoretical results on the generalization of MDC when qk is large. On the other hand, since MDC is a more general case of multi-label learning to some extent, and there may be extreme cases in multi-label learning where the number of labels (i.e., dimensions) is extremely large (Yu et al., 2014; Prabhu & Varma, 2014; Yen et al., 2016; Liu & Shen, 2019). In addition, there may also be extreme cases in multi-class classification corresponding to each dimension where the number of classes is extremely large. When both the number of dimensions and classes are extreme, an extremely large qk will render the bounds with*

a linear dependency on qk completely invalid, resulting in a lack of effective theoretical bounds for MDC. Hence, the above analysis implies that the dependency of an ideal bound on qk should be at least weaker than linear, which also reveals the need to develop new theoretical analysis methods and tools to derive tighter and more effective bounds.

The introduction of two decomposition operators (i.e., outer and inner decomposition operator) is the key to improving the dependency of the bounds on q and k. We first define a class $\mathcal{R}$ consisting of the outer decomposition operators $r_j : \mathbb{R}^q \mapsto \mathbb{R}$ for any $j \in [q]$ which decompose the components corresponding to the j-th coordinate of the q-dimensional vector-valued function $\boldsymbol{f}$, and we have $\mathcal{R}(\mathcal{F}) = \{(j, \boldsymbol{x}) \mapsto r_j(\boldsymbol{f}(\boldsymbol{x})) : r_j(\boldsymbol{f}(\boldsymbol{x})) = f_j(\boldsymbol{x}),\ \ \boldsymbol{f} \in \mathcal{F}, (j, \boldsymbol{x}) \in [q] \times \mathcal{X}\}$. Then, we define a class $\mathcal{T}$ consisting of the inter decomposition operators $t_s : \mathbb{R}^k \mapsto \mathbb{R}$ for any $s \in [k]$ which decompose the components corresponding to the s-th coordinate of the k-dimensional vector-valued function $\boldsymbol{h}^j$, and we have $\mathcal{T}(\mathcal{H}^j) = \left\{(s, j, \boldsymbol{x}) \mapsto t_s(\boldsymbol{h}^j(\boldsymbol{x})) : t_s(\boldsymbol{h}^j(\boldsymbol{x})) = h_s^j(\boldsymbol{x}), \boldsymbol{h}^j \in \mathcal{H}^j, (s, j, \boldsymbol{x}) \in [k] \times [q] \times \mathcal{X}\right\}$. Accordingly, we call $\mathcal{R}(\mathcal{F})$ the outer decomposition class and call $\mathcal{T}(\mathcal{H}^j)$ the inner decomposition class. With the assumption of ℓ_2 Lipschitz loss, we show that the Rademacher complexity of the loss space can be bounded by the worst-case Rademacher complexity of inner decomposition class $\mathcal{T}(\mathcal{H}^j)$. We develop the following novel tight vector-contraction inequality:

Lemma 9. *Let $\mathcal{F}$ be the MDC class, Assumptions 2-4 hold. Given a dataset D of size n, we have*

$$\hat{\mathfrak{R}}_D(\mathcal{L}) \leq \frac{576\mu B\sqrt{q}}{\sqrt{n}} \log^{\frac{1}{2}}(nq) \log \frac{M\sqrt{n}}{\mu B} + 2q\mu 48^2 \sqrt{k} \widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \times \left(1 + \left(1 + \log_2(8en^2q^2k^2) \log \sqrt{nq}\right) \log^{\frac{1}{2}}(nq) \log \frac{M\sqrt{n}}{\mu B}\right),$$

where $\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))$ is the worst-case Rademacher complexity of the inner decomposition class.

Remark 10. *According to existing theoretical results (i.e., Sudakov's minoration & relationship between Rademacher and Gaussian complexity in Zhang & Zhang (2024)), we can convert the Rademacher complexity of loss space $\mathcal{L}$ to that of the outer decomposition function class $\mathcal{R}(\mathcal{F})$. Then, induced by the effective establishment of the unified framework for MDC, we can use a similar method to convert the Rademacher complexity of $\mathcal{R}(\mathcal{F})$ to the worst-case Rademacher complexity of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$. The above proof idea will eventually produce $\widetilde{O}(\sqrt{qk}/\sqrt{n})$ bounds. Here we find that the square-root dependency on q mainly comes from the $\sqrt{q}$ factor in the radius of the empirical ℓ_2 cover of $\mathcal{R}(\mathcal{F})$, which is inevitable for ℓ_2 Lipschitz loss. However, the dependency on k can be further improved, and we no longer use existing theoretical results (i.e., Sudakov's minoration & relationship between Rademacher and Gaussian complexity) to prove the $\widetilde{O}(\sqrt{q}/\sqrt{n})$ bound for MDC with ℓ_2 Lipschitz loss, since they prevent us from obtaining bounds with no dependency on k. The main reason is that we find that a factor of $\sqrt{k}$ in the radius of the empirical ℓ_2 cover of the inner decomposition function class cannot be eliminated by Sudakov's minoration. We can eliminate the $\sqrt{k}$ factor and improve the dependency on k to be independent by exploring the properties of decomposition function classes. To avoid the propagation of the $\sqrt{k}$ factor induced by Sudakov's minoration, we use inequalities between different complexity measures, and the conversion between different complexities will not cause the propagation of the $\sqrt{k}$ factor.*

Remark 11. *The construction of the auxiliary dataset induced by outer and inner decomposition operator can reduce the dependency on q and k. For the construction of other types of auxiliary datasets under different problem settings, please refer to Lei et al. (2019); Wu et al. (2021b); Mustafa et al. (2021); Graf et al. (2022); Lei et al. (2023); Hieu et al. (2024). The use of discretized variant of Dudley's integral inequality in Lei et al. (2023) can achieve more desirable constant.*

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

Theorem 12. *Let $\mathcal{F}$ be the class of the MDC defined by (1) and (2). Let Assumptions 2, 3 and 4 hold. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ over a dataset D of size n, the following holds for any $\boldsymbol{f} \in \mathcal{F}$:*

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + 3M\sqrt{\frac{\log \frac{2}{\delta}}{2n}} + \frac{2 \cdot 24^2 \mu B\sqrt{q}}{\sqrt{n}} \log^{\frac{1}{2}}(nq) \log \frac{M\sqrt{n}}{\mu B} + \frac{96^2 \mu B\sqrt{q}}{\sqrt{n}} \times \left(1 + \left(1 + \log_2(8en^2q^2k^2) \log \sqrt{nq}\right) \log^{\frac{1}{2}}(nq) \log \frac{M\sqrt{n}}{\mu B}\right).$$

Remark 13. *The term $\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \leq B/\sqrt{nqk}$ in Lemma 9, which makes the Rademacher complexity $\hat{\Re}_D(\mathcal{L})$ actually only square-root dependent on q, and results in a tighter bound than the $O(qk/\sqrt{n})$ and $\widetilde{O}(\sqrt{qk}/\sqrt{n})$ bounds with a faster convergence rate $\widetilde{O}(\sqrt{q}/\sqrt{n})$. By introducing the outer and inner decomposition operator, the use of Sudakov's minoration and the relationship between Rademacher and Gaussian complexity can induce the $\widetilde{O}(\sqrt{qk}/\sqrt{n})$ bounds. However, the bounds with a square-root dependency on qk cannot provide theoretical guarantees for all methods of MDC. For example, for MDC methods that model pairwise class dependencies Arias et al. (2016); Jia & Zhang (2020; 2021b), they first considers pairwise dependencies by training a total of $\binom{q}{2}$ classifiers, one per a pair of class spaces, and then each pairwise class space is composed of all possible class labels in the two initial class spaces, i.e., the possible class labels in each pairwise class space are k^2. Hence, for these methods, the product of q and k becomes $\binom{q}{2}k^2$, i.e., its order is q^2k^2. This means that for MDC methods that model pairwise class dependencies, the bounds induce by existing theoretical results with a square-root dependency on qk still suffer from the same failure as the bounds with a linear dependency on qk. However, our tight $\widetilde{O}(\sqrt{q}/\sqrt{n})$ bound with no dependency on k can effectively alleviate this situation.*

We also provide a matching lower bound for MDC with ℓ_2 Lipschitz loss to demonstrate that $\sqrt{q}$ dependency is unavoidable and to support the statement of tightness.

Theorem 14. *Consider the MDC problem, where the loss function is μ-Lipschitz with respect to the ℓ_2 norm, and the prediction function satisfies Assumption 3. Then there exists an input distribution $\mathcal{D}$, such that the ERM solution satisfies $R(\boldsymbol{f}) - \hat{R}_D(\boldsymbol{f}) = \Omega\left(\sqrt{\frac{q}{n}}\right)$.*

Remark 15. *Theorem 14 shows that the lower bound of order $\Omega\left(\sqrt{q}/\sqrt{n}\right)$ matches the $\widetilde{O}(\sqrt{q}/\sqrt{n})$ upper bound in Theorem 12, which indicates that our bound in Theorem 12 is tight.*

4 Tight Bounds for MDC with ℓ_∞ Lipschitz Loss

In this section, we first introduce the assumption of the ℓ_∞ Lipschitz loss. Then, combining the outer and inner decomposition operator, we develop a novel tight vector-contraction inequality for MDC. Finally, with the tight vector-contraction inequality for ℓ_∞ Lipschitz loss, we derive the bounds with no dependency on q and k for general MDC classes, up to logarithmic terms.

Assumption 16. *Assume that loss ℓ is μ-Lipschitz w.r.t. ℓ_∞ norm, i.e., $|\ell(\boldsymbol{f}(\boldsymbol{x}), \cdot) - \ell(\boldsymbol{f}'(\boldsymbol{x}), \cdot)| \leq \mu \left\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x})\right\|_\infty$, where $\mu > 0$, $\|\boldsymbol{t}\|_\infty = \max_{j\in[q]} |t_j|$ for $\boldsymbol{t} = (t_1, \ldots, t_q)$.*

Assumption 16 is very mild, the Lipschitz continuity w.r.t. the ℓ_∞ norm has been considered in several literature (Foster & Rakhlin, 2019; Lei et al., 2019; Wu et al., 2021b; Lei et al., 2023; Zhang & Zhang, 2024). The following Proposition 17 further illustrates that the commonly used loss functions here in MDC actually satisfy Assumption 16.

Proposition 17. *Assume that the base loss ℓ_b is μ-Lipschitz continuous, then the surrogate Hamming Loss is μ-Lipschitz w.r.t. ℓ_∞ norm, the surrogate Matching Loss is μ-Lipschitz w.r.t. ℓ_∞ norm.*

With decomposition operators and ℓ_∞ Lipschitz loss, we have tight vector-contraction inequality:

Lemma 18. *Let $\mathcal{F}$ be the class of the MDC defined by (1) and (2). Let Assumptions 2, 3 and 16 hold. Given a dataset D of size n. Then, we have*

$$\hat{\Re}_D(\mathcal{L}) \leq \frac{12M}{\sqrt{n}} + 96\mu\sqrt{qk}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \times \left(1 + \log_2(8en^2q^2k^2\mu^2) \cdot \ln\frac{M\sqrt{n}}{\mu B}\right),$$

where $\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))$ is the worst-case Rademacher complexity of the inner decomposition class.

Remark 19. *Here we do not follow the idea of the composition of Rademacher complexities to derive a tighter bound, but directly consider the function classes, since intuitively we think that constant terms 576, $2\cdot 48^2$ and logarithmic term $1+\left(1+\log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B}$ in Lemma 9 are not elegant enough, they may be suboptimal. Hence, we directly consider using the empirical ℓ_∞ norm covering number of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$ to bound the empirical ℓ_∞ norm covering number of the loss space $\mathcal{L}$. This actually directly considers the relationship between the function classes themselves, and can improve the constant terms from 576 and $2\cdot 48^2$ to 12 and 96, and the logarithmic term to $1+\log_2(8en^2q^2k^2\mu^2)\cdot\ln\frac{M\sqrt{n}}{\mu B}$.*

With Lemma 18, we can derive the following tight bound:

Theorem 20. *Let $\mathcal{F}$ be the class of MDC defined by (1) and (2). Let Assumptions 2, 3 and 16 hold. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ over a dataset D of size n, the following holds for any $\boldsymbol{f} \in \mathcal{F}$:*

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{24M}{\sqrt{n}} + \frac{192\mu B}{\sqrt{n}}\left(1 + \log_2(8en^2q^2k^2\mu^2)\cdot\ln\frac{M\sqrt{n}}{\mu B}\right).$$

Remark 21. *$\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \leq B/\sqrt{nqk}$ in Lemma 18, which makes $\hat{\Re}_D(\mathcal{L})$ actually independent on q and k, up to logarithmic terms, and results in a tighter bound than $O(qk/\sqrt{n})$, $\widetilde{O}(\sqrt{qk}/\sqrt{n})$ and $\widetilde{O}(\sqrt{q}/\sqrt{n})$ bounds with a faster convergence rate $\widetilde{O}(1/\sqrt{n})$. ℓ_∞ Lipschitz loss assumption, effective unified framework we established and the direct consideration of the relationship between function classes themselves are keys to improving the dependency of bounds on q and k to independency. The first two reduce the dependency on q and k, while the latter improves the dependency on the constant and logarithmic term. Here we use completely different lemmas and scaling techniques from those in Zhang & Zhang (2024), and customized the analysis of the relationship between inner decomposition class and loss class, resulting in important improvements in constants and logarithmic terms. Even for extreme cases mentioned in Remark 8, i.e., $qk \gg n$ or MDC methods that model pairwise class dependencies, i.e., q^2k^2, since $\log qk \ll \sqrt{n}$ and $\log q^2k^2 \ll \sqrt{n}$, tight bounds with no dependency on q and k here, up to logarithmic terms, can provide general theoretical guarantees for MDC.* ***A detailed discussion and clarification regarding the problem setting of MDC, the relationship between our generalization analysis and learnability, the impact of the dependencies among dimensions on the generalization bounds, future ideas for experimental verification, and the differences between our theoretical results and those of other learning theory problems are provided in the Discussion of the Appendix.***

5 Generalization Bounds for different prediction models

In this section, we analyze the bounds of prediction models with various architectures for MDC, i.e., FNN and Transformer. We show that models with various architectures for MDC will lead to significant differences in the constant B of the bounds in Theorem 12 and 20.

5.1 Feedforward Neural Network (FNN)

Consider FNN $f(\boldsymbol{x}) = \boldsymbol{w}^\top\sigma(W_L\sigma(W_{L-1}\cdots\sigma(W_1\boldsymbol{x})))$, where W_l are the parameter metrices, $l \in [L]$, and σ is the ReLU activation. Assume that input features and parameter matrices in FNN are bounded, i.e., $\|\boldsymbol{x}_i\| \leq R$, $\|W_l\| \leq B_{FNN}, \forall i \in [n], l \in [L]$, where $R, B_{FNN} > 0$ are constants.

Corollary 22. *Let $\mathcal{F}$ be the class of MDC defined by (1) and (2), where the prediction model is induced by a three-layer FNN. Let Assumptions 2, 3 and 16 hold. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ over a dataset D of size n, the following holds for any $\boldsymbol{f} \in \mathcal{F}$:*

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{24M}{\sqrt{n}} + \frac{192\mu 2^3\Lambda B_{FNN}^3 R}{\sqrt{n}}\left(1 + \log_2(8en^2q^2k^2\mu^2)\cdot\ln\frac{M\sqrt{n}}{\mu B}\right).$$

Remark 23. *The constant B of bounds in Theorem 12 and 20 corresponds to $2^3\Lambda B_{FNN}^3 R$ here. The generalization analysis of FNN involved here mainly uses peeling argument, which is a commonly used method in capacity-based theoretical analysis for deep models. Peeling argument induces exponential dependency on depth, and it is obvious that $B := 2^L\Lambda B_{FNN}^L R$ for L layer FNN, which may make bounds vacuous Bartlett et al. (2017; 2019). How to develop non-vacuous capacity-based analysis methods for deep models is still an open problem Neyshabur et al. (2018); Golowich et al. (2018); Zhang & Zhang (2023) and we do not focus on exploring this research here. Our goal here is to show how various architectures affect the bounds for MDC.*

5.2 Transformer

We follow the definition of Transformers in Edelman et al. (2022). Let $X := [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T]^\top \in \mathbb{R}^{T\times d}$ be the input, i.e., a sequence of T d-dimensional tokens. Let σ be an element-wise L_σ-Lipschitz activation function with $\sigma(0) = 0$. Let $W_C \in \mathbb{R}^{k\times d}, W_V, W_Q, W_K \in \mathbb{R}^{d\times k}$ be trainable

weight matrices. A Transformer head is defined as $\sigma\left(\mathrm{RowSoftmax}\left(XW_QW_K^\top X^\top\right)XW_V\right)W_C$, where RowSoftmax applies softmax on each row of its input. For multi-head Transformers, each head is summed up at the end of each layer. For the convenience of analysis, we represent $W_QW_K^\top$ with a single matrix $W_{QK}\in\mathbb{R}^{d\times d}$. Here we focus on the scalar output from Transformer. We construct a special extra input in the length-T sequence with a index [CLS], which is a fixed or trainable vector $x_{[\mathrm{CLS}]}\in\mathbb{R}^d$. Finally, the output at index [CLS] is a linear function $\boldsymbol{w}^\top\boldsymbol{y}_{[\mathrm{CLS}]}$, for a trainable parameter $\boldsymbol{w}\in\mathbb{R}^d$. Then, the scalar one layer H-head Transformer is $V(X)=\boldsymbol{w}^\top\boldsymbol{y}_{[\mathrm{CLS}]}\in\mathcal{V}$ where $\boldsymbol{y}_{[\mathrm{CLS}]}=\sum_{u=1}^H\boldsymbol{y}_u=\sum_{u=1}^H W_{u,C}^\top\sigma\left(W_{u,V}^\top X^\top\mathrm{softmax}\left(XW_{u,QK}^\top\boldsymbol{x}_{[\mathrm{CLS}]}\right)\right)$ and we denote $v_u(X)=\boldsymbol{w}^\top\boldsymbol{y}_u$ and $\boldsymbol{v}(X):=(v_1(X),\ldots,v_H(X))$. Assume that the parameter matrices in Transformer are bounded, i.e., $\|W_{u,C}\|\le B_{W_C}$, $\|W_{u,V}\|\le B_{W_V}$, $\|W_{u,QK}\|\le B_{W_{QK}}$ and $\|\boldsymbol{x}_t\|\le R$ for any $t\in T$, $u\in[H]$. The symbol $\|\cdot\|$ denotes the Frobenius norm.

When the prediction model of MDC corresponds to a multi-head Transformer, to reduce the dependency of the bounds for MDC on H, we introduce the multi-head decomposition operator for the multi-head Transformer. We define a class $\mathcal{P}$ consisting of the multi-head decomposition operators $p_u:\mathbb{R}^H\mapsto\mathbb{R}$ for any $u\in[H]$ which decompose the components corresponding to the u-th head of the multi-head Transformer, and we have $\mathcal{P}(\mathcal{V})=\left\{(u,X)\mapsto p_u(\boldsymbol{v}(X)):p_u(\boldsymbol{v}(X))=\boldsymbol{w}^\top\boldsymbol{y}_u,\ V(X)\in\mathcal{V},\ \ (u,X)\in[H]\times\mathcal{X}\right\}$. Then, we have the tight bound:

Theorem 24. *Let $\mathcal{F}$ be the class of MDC defined by (1) and (2), where the prediction model is induced by a one layer H-head Transformer. Let Assumptions 2, 3 and 16 hold. Then, for any $0<\delta<1$, $\boldsymbol{f}\in\mathcal{F}$, the following holds with probability at least $1-\delta$ over a dataset D of size n:*

$$R(\boldsymbol{f})\le\widehat{R}_D(\boldsymbol{f})+3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\frac{24M}{\sqrt{n}}+\frac{192\mu}{\sqrt{n}}\times\left(1+\log_2(8en^2q^2k^2\mu^2)\cdot\ln\frac{M\sqrt{n}}{\mu B}\right)\times$$
$$24H\Lambda B_{W_C}B_{W_V}L_\sigma R\times(1+8(1+\log_2(8en^2q^2k^2H^4)\cdot\ln(H\sqrt{nqk}))).$$

Remark 25. *Theorem 24 implies that the order of the bound for MDC with one layer multi-head Transformer is $\widetilde{O}(\frac{H}{\sqrt{n}})$ with a linear dependency on H and no dependency on the sequence length, which is tight w.r.t. the dependency on the sequence length and matches the recent theoretical bounds for Transformer Edelman et al. (2022); Trauger & Tewari (2024). The key to reducing the dependency on the sequence length is to avoid summation in the norm, where the Lipschitz property of* softmax *is used, i.e., Corollary A.7 in Edelman et al. (2022). In addition, the bound in Theorem 24 with a linear dependency on H is tighter than the bound in Edelman et al. (2022) with a polynomial dependency on H (i.e., $H^{3/2}$). The constant B corresponds to $2\Lambda B_{W_C}B_{W_V}L_\sigma R$ here.*

Remark 26 (Comparison with Related Work)**.** *MDC is closely related to the widely studied multi-label learning, which is obvious from their problem settings, since each dimension in MDC corresponds to a multi-class variable, while each label in multi-label learning corresponds to a binary class variable, so from this perspective, multi-label learning is a degenerate case of MDC. Since the theoretical research of MDC is completely unexplored before, we mainly compare our results with the related work on generalization analysis of multi-label learning.* ***The detailed comparison is provided in the Appendix.***

6 Conclusion

In this paper, we first establish a unified theoretical framework for MDC and show that the bounds induced by existing theoretical results are not tight. Then, we propose several novel vector-contraction inequalities by exploiting the Lipschitz continuity of loss functions, and derive bounds with a logarithmic dependency on k and with square-root and logarithmic dependency on q, respectively. In addition, we study the impact of various architectures on bounds.

In future work, we will explore the impact of different approaches to modeling class dependencies in MDC on the bounds, and develop new theoretical tools to investigate the lower bound to test whether the $\widetilde{O}(1/\sqrt{n})$ bounds here are optimal, and derive sharper bounds for MDC with a faster convergence rate w.r.t. the number of examples.

AI USE STATEMENT

In this work, we did not use any generative AI tools for any tasks.

REFERENCES

Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

Jacinto Arias, Jose A Gamez, Thomas D Nielsen, and Jose M Puerta. A scalable pairwise class interaction framework for multidimensional classification. *International Journal of Approximate Reasoning*, 68:194–210, 2016.

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

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

Kush Bhatia, Kunal Dahiya, Himanshu Jain, Purushottam Kar, Anshul Mittal, Yashoteja Prabhu, and Manik Varma. The extreme classification repository: Multi-label datasets and code. *URL http://manikvarma.org/downloads/XC/XMLRepository. html*, 2016.

Hanen Borchani, Concha Bielza, Carlos Toro, and Pedro Larrañaga. Predicting human immunodeficiency virus inhibitors using multi-dimensional bayesian network classifiers. *Artificial Intelligence in Medicine*, 57(3):219–229, 2013.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the ERM principle. *Journal of Machine Learning Research*, 16:2377–2404, 2015.

Krzysztof Dembczynski, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. 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.

Krzysztof Dembczynski, Wojciech Kotlowski, and Eyke Hüllermeier. Consistent multilabel ranking through univariate losses. *arXiv:1206.6401*, 2012.

Benjamin L. Edelman, Surbhi Goel, Sham M. Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 5793–5831, 2022.

Vitaly Feldman. Generalization of ERM in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3576–3584, 2016.

Pablo Fernandez Gonzalez, Pedro María Larrañaga Múgica, and Maria Concepcion Bielza Lozoya. Multidimensional classifiers for neuroanatomical data. In *Proceedings of the ICML Workshopon Statistics, Machine Learning and Neuroscience*, pp. 1–6, 2015.

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

Santiago Gil-Begue, Concha Bielza, and Pedro Larrañaga. Multi-dimensional bayesian network classifiers: A survey. *Artificial Intelligence Review*, 54(1):519–559, 2021.

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

Florian Graf, Sebastian Zeng, Bastian Rieck, Marc Niethammer, and Roland Kwitt. On measuring excess capacity in neural networks. *Advances in Neural Information Processing Systems 35*, 2022.

Yann Guermeur. Combinatorial and structural results for gamma-psi-dimensions. *arXiv:1809.07310*, 2018.

Nong Minh Hieu, Antoine Ledent, Yunwen Lei, and Cheng Yeaw Ku. Generalization analysis for deep contrastive representation learning. *arXiv preprint arXiv:2412.12014v2*, 2024.

Teng Huang, Bin-Bin Jia, and Min-Ling Zhang. Deep multi-dimensional classification with pairwise dimension-specific features. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, number IJCAI 2024, pp. 4183–4191, 2024.

Teng Huang, Bin-Bin Jia, and Min-Ling Zhang. Towards escaping from class dependency modeling for multi-dimensional classification. In *Proceedings of the 42th International Conference on Machine Learning*, number ICML 2025, 2025.

Bin-Bin Jia and Min-Ling Zhang. Multi-dimensional classification via stacked dependency exploitation. *Science China Information Sciences*, 63(12), 2020.

Bin-Bin Jia and Min-Ling Zhang. Multi-dimensional classification via sparse label encoding. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4917–4926, 2021a.

Bin-Bin Jia and Min-Ling Zhang. Md-knn: An instance-based approach for multi-dimensional classification. In *Proceedings of the 25th International Conference on Pattern Recognition*, number ICPR 2021, pp. 126–133, 2021b.

Bin-Bin Jia and Min-Ling Zhang. Maximum margin multi-dimensional classification. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7185–7198, 2022.

Bin-Bin Jia and Min-Ling Zhang. Multi-dimensional classification via decomposed label encoding. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1844–1856, 2023.

Antoine Ledent, Waleed Mustafa, Yunwen Lei, and Marius Kloft. 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.

Yunwen Lei, Ürün Dogan, Ding-Xuan Zhou, and Marius Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5):2995–3021, 2019.

Yunwen Lei, Tianbao Yang, Yiming Ying, and Ding-Xuan Zhou. Generalization analysis for contrastive representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19200–19227, 2023.

Verayuth Lertnattee and Thanaruk Theeramunkong. Multidimensional text classification for drug information. *IEEE Transactionson Information Technology in Biomedicine*, 8(3):306–312, 2004.

Zheng Lian, Ya Li, Jianhua Tao, Jian Huang, and Mingyue Niu. Expression analysis based on face regions in real-world conditions. *International Journal of Automation and Computing*, 17(1): 96–107, 2020.

Chong Liu, Peng Zhao, Sheng-Jun Huang, Yuan Jiang, and Zhi-Hua Zhou. Dual set multi-label learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, number AAAI 2018, pp. 3635–3642, 2018.

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

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

Roey Magen and Ohad Shamir. Initialization-dependent sample complexity of linear predictors and neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

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

Abu Hena Al Muktadir, Takaya Miyazawa, Pedro Martinez-Julia, Hiroaki Harai, and Ved P. Kafle. Multi-target classification based automatic virtual resource allocation scheme. *IEICE Transactions on Information and Systems*, 102-D(5):898–909, 2019.

Waleed Mustafa, Yunwen Lei, Antoine Ledent, and Marius Kloft. Fine-grained generalization analysis of structured output prediction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, number IJCAI 2021, pp. 2841–2847, 2021.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *Proceedings of 6th International Conference on Learning Representations*, number ICLR 2018, 2018.

Vu-Linh Nguyen, Yang Yang, and Cassio de Campos. Probabilistic multi-dimensional classification. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216, pp. 1522–1533, 2023.

Yashoteja Prabhu and Manik Varma. 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.

Vinod Raman, Unique Subedi, and Ambuj Tewari. A characterization of multioutput learnability. *Journal of Machine Learning Research*, 25:342:1–342:54, 2024.

Jesse Read, Concha Bielza, and Pedro Larrañaga. Multi-dimensional classification with super-classes. *IEEE Transactions on Knowledge and Data Engineering*, 26(7):1720–1733, 2014.

Henry Reeve and Ata Kaban. Optimistic bounds for multi-output learning. In *Proceedings of the 37th International Conference on Machine Learning*, number ICML 2020, pp. 8030–8040, 2020.

Juan Diego Rodríguez, Aritz Pérez Martínez, David Arteta, Diego Tejedor, and José Antonio Lozano. Using multidimensional bayesian network classifiers to assist the treatment of multiple sclerosis. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, 42(6):1705–1715, 2012.

Matan Schliserman and Tomer Koren. Complexity of vector-valued prediction: From linear models to stochastic convex optimization. *arXiv:2412.04274v1*, 2024.

Francesco Serafino, Gianvito Pio, Michelangelo Ceci, and Donato Malerba. Hierarchical multi-dimensional classification of web documents with multiwebclass. In *Proceedings of the 18th International Conference on Discovery Science*, volume 9356, pp. 236–250, 2015.

Hagit Shatkay, Fengxia Pan, Andrey Rzhetsky, and W. John Wilbur. Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093, 2008.

Yi Shi, Han-Jia Ye, Dongliang Man, Xiaoxu Han, De-Chuan Zhan, and Yuan Jiang. Revisiting multi-dimensional classification from a dimension-wise perspective. *Frontiers of Computer Science*, 19 (1):191304, 2025.

Lingyun Song, Jun Liu, Buyue Qian, Mingxuan Sun, Kuan Yang, Meng Sun, and Samar Abbas. A deep multi-modal CNN for multi-instance multi-label image classification. *IEEE Transactions on Image Processing*, 27(12):6025–6038, 2018.

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

Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 1405–1413, 2024.

M Varma and J Langford. Nips workshop on extreme classification. In *NIPS Workshop on eXtreme Classification*, 2013.

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

Guoqiang Wu and Jun Zhu. Multi-label classification: do hamming loss and subset accuracy really conflict with each other? *Advances in Neural Information Processing Systems*, 33(NeurIPS 2020), 2020.

Guoqiang Wu, Chongxuan Li, Kun Xu, and Jun Zhu. 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.

Liang Wu, Antoine Ledent, Yunwen Lei, and Marius Kloft. 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.

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

Ian En-Hsu Yen, Xiangru Huang, Pradeep Ravikumar, Kai Zhong, and Inderjit S. Dhillon. 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.

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

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

Yi-Fan Zhang and Min-Ling Zhang. Generalization analysis for multi-label learning. In *Proceedings of the 41st International Conference on Machine Learning*, number ICML 2024, 2024.

A APPENDIX

A.1 COMPARISON WITH RELATED WORK

MDC is closely related to the widely studied multi-label learning, which is obvious from their problem settings, since each dimension in MDC corresponds to a multi-class variable, while each label in multi-label learning corresponds to a binary class variable, so from this perspective, multi-label learning is a degenerate case of MDC. Since the theoretical research of MDC is completely unexplored before, we mainly compare our results with the related work on generalization analysis of multi-label learning.

Under the assumption of ℓ_2 Lipschitz loss, Wu & Zhu (2020); Wu et al. (2021a) obtained $O(c/\sqrt{n})$ bounds with a linear dependency on the number of labels c (which can be regarded as the dimensions in MDC, i.e., $c := q$) for surrogate Hamming, Subset and Ranking losses, which drew on the analysis of the relationship between evaluation metrics in Dembczynski et al. (2010; 2012). Wu & Zhu (2020); Wu et al. (2021a) also showed that the order of bounds for some losses can be improved to $O(\sqrt{c/n})$ when preserving the coupling among different components for kernel classes. Liu et al. (2018) also obtained a $O(\sqrt{c/n})$ bound for the dual set multi-label learning with the margin loss and kernel classes. Here we derive the $\widetilde{O}(\sqrt{q}/\sqrt{n})$ bounds with a square-root dependency on q and no dependency on k, up to logarithmic terms, for MDC, which improves the $O(qk/\sqrt{n})$ and $O(\sqrt{qk}/\sqrt{n})$ bounds induced by existing theoretical results.

Under the assumption of ℓ_∞ Lipschitz loss, Wu et al. (2021b) derived a $O(\log(nc)/n\sigma)$ bound for norm regularized kernel classes with the additional assumption that the regularizer is σ-strongly convex. Zhang & Zhang (2024) derived a $\widetilde{O}(1/\sqrt{n})$ bound, which decouples the relationship among different components. In addition, Yu et al. (2014) obtained a $O(1/\sqrt{n})$ bound for trace norm regularized linear classes for the decomposable loss. Xu et al. (2016) used the local Rademacher complexity to derive a $\widetilde{O}(1/n)$ bound for trace norm regularized linear classes with the assumption that singular values of $\boldsymbol{w}$ decay exponentially. Here we derive tight $\widetilde{O}(1/\sqrt{n})$ bounds with no dependency on q and k for MDC with ℓ_∞ Lipschitz loss.

A.2 OUTLINE OF THE PROOFS FOR OUR THEORETICAL RESULTS

We give the detailed proofs of our theoretical results in the main paper. Our main proofs include:

- The ℓ_2 Lipschitz continuity of the commonly used losses in MDC (Proposition 5).
- The basic bound for MDC with ℓ_2 Lipschitz loss (Corollary 6).
- The tight vector-contraction inequality for MDC with ℓ_2 Lipschitz loss (Lemma 9).
- The tight bound for MDC with ℓ_2 Lipschitz loss (Theorem 12).
- The lower bound for MDC with ℓ_2 Lipschitz loss (Theorem 14).
- The ℓ_∞ Lipschitz continuity of the commonly used losses in MDC (Proposition 17).
- The tight vector-contraction inequality for MDC with ℓ_∞ Lipschitz loss (Lemma 18).
- The tight bound for MDC with ℓ_∞ Lipschitz loss (Theorem 20).
- The bound for MDC with the prediction model induced by a three-layer FNN (Corollary 22).
- The bound for MDC with the prediction model induced by a one layer multi-head Transformer (Theorem 24).

A.3 PRELIMINARIES

A.3.1 DEFINITIONS OF THE CORRESPONDING COMPLEXITY MEASURES

Definition 27 (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 $\epsilon > 0$ and $p \geq 1$, the empirical ℓ_p*

norm covering number $\mathcal{N}_p(\epsilon, \mathcal{F}, D)$ 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 $\boldsymbol{v}^j$)

$$\left(\frac{1}{n}\sum_{i=1}^{n}\left|f\left(\boldsymbol{x}_i\right)-\boldsymbol{v}_i^j\right|^P\right)^{\frac{1}{p}} \leq \epsilon.$$

In this case, we call $\left\{\boldsymbol{v}^1, \ldots, \boldsymbol{v}^m\right\}$ an (ϵ, ℓ_p)-cover of $\mathcal{F}$ w.r.t. D. We also denote $\mathcal{N}_p(\epsilon, \mathcal{F}, n) = \sup_D \mathcal{N}_p(\epsilon, \mathcal{F}, D)$.

Definition 28 (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 $fat_\epsilon(\mathcal{F})$ 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*

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

A.3.2 The Bound for the loss function space

According to McDiarmid's inequality (McDiarmid et al., 1989) and the symmetrization technique, it is easy to obtain that for any training dataset $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) : i \in [n]\}$, $\ell(\cdot, \cdot) \leq M$, with probability at least $1 - \delta$, the following holds:

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

For simplicity, we slightly abuse $R(\boldsymbol{f})$ to represent $R(\ell(\boldsymbol{f}))$.

A.4 Tight Bounds for MDC with ℓ_2 Lipschitz Loss

A.4.1 Proof of Proposition 5

We first prove that the surrogate Hamming Loss is $\frac{\mu}{\sqrt{q}}$-Lipschitz continuous w.r.t. the ℓ_2 norm.

$$\begin{aligned}
&\left|\ell_H(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) - \ell_H\left(\boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{y}\right)\right| \\
=&\left|\frac{1}{q}\sum_{j=1}^{q}\ell_b\left(f_j(\boldsymbol{x}), y^j\right) - \frac{1}{q}\sum_{j=1}^{q}\ell_b\left(f_j'(\boldsymbol{x}), y^j\right)\right| \\
=&\frac{1}{q}\sum_{j=1}^{q}\left|\ell_b\left(f_j(\boldsymbol{x}), y^j\right) - \ell_b\left(f_j'(\boldsymbol{x}), y^j\right)\right| \\
\leq&\frac{1}{q}\sum_{j=1}^{q}\mu\left|f_j(\boldsymbol{x}) - f_j'(\boldsymbol{x})\right| \\
\leq&\mu\left(\frac{1}{q}\sum_{j=1}^{q}\left|f_j(\boldsymbol{x}) - f_j'(\boldsymbol{x})\right|^2\right)^{\frac{1}{2}} \quad \text{(Use Jensen's Inequality)} \\
=&\frac{\mu}{\sqrt{q}}\left\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x})\right\|_2.
\end{aligned}$$

Second, with the elementary inequality

$$\left|\max\left\{a_1, \ldots, a_q\right\} - \max\left\{b_1, \ldots, b_q\right\}\right| \leq \max\left\{\left|a_1 - b_1\right|, \ldots, \left|a_q - b_q\right|\right\},$$

we prove that the surrogate Matching Loss is $\mu\sqrt{q}$-Lipschitz continuous w.r.t. the ℓ_2 norm.

$$\begin{aligned}
&\left|\ell_M(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) - \ell_M\left(\boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{y}\right)\right| \\
=&\left|\max_{j\in[q]}\{\ell_b\left(f_j(\boldsymbol{x}), y^j\right)\} - \max_{j\in[q]}\{\ell_b\left(f_j'(\boldsymbol{x}), y^j\right)\}\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \max_{j\in[q]} \left|\ell_b\left(f_j(\boldsymbol{x}), y^j\right) - \ell_b\left(f'_j(\boldsymbol{x}), y^j\right)\right| \\
&\leq \mu \max_{j\in[q]} \left|f_j(\boldsymbol{x}) - f'_j(\boldsymbol{x})\right| \\
&\leq \mu \sum_{j=1}^{q} \left|f_j(\boldsymbol{x}) - f'_j(\boldsymbol{x})\right| \\
&\leq \mu q \left(\frac{1}{q}\sum_{j=1}^{q} \left|f_j(\boldsymbol{x}) - f'_j(\boldsymbol{x})\right|^2\right)^{\frac{1}{2}} \qquad \text{(Use Jensen's Inequality)} \\
&= \mu\sqrt{q}\left\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x})\right\|_2 .
\end{aligned}
$$

A.4.2 Proof of Corollary 6

Proof Sketch. By properly scaling vector-contraction lemma in Maurer (2016) (i.e., $\mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_i \epsilon_i h_i\left(\boldsymbol{f}\left(\boldsymbol{x}_i\right)\right) \leq \sqrt{2}L\mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_{i,j}\epsilon_{ij} f_j\left(\boldsymbol{x}_i\right)$), we can see that the Rademacher complexity of the vector-valued function class $\mathcal{F}$ composited with an L-ℓ_2 Lipschitz function h can be bounded by the maximum Rademacher complexity of the restrictions of the function class $\mathcal{F}$ along each coordinate and a multiplicative factor q, i.e., $\hat{\mathfrak{R}}_D(h(\mathcal{F})) \leq \sqrt{2}Lq\max_j \hat{\mathfrak{R}}_D(\mathcal{F}_j)$. Then, according to the definitions of classes (1) and (2), use the above inequality for Rademacher complexity twice, first convert the Rademacher complexity of the loss space $\mathcal{L}$ to the Rademacher complexity of the restriction of the function class $\mathcal{F}$ along the j-th coordinate, i.e., $\mathcal{F}_j$, and then convert the Rademacher complexity of $\mathcal{F}_j$ to the Rademacher complexity of the restriction of the vector-valued function class $\mathcal{H}^j$ along the s-th coordinate, i.e., $\mathcal{H}^j_s$. Hence, the Rademacher complexity of the loss function space $\mathcal{L}$ can be bounded. Finally, We upper bound the Rademacher complexity $\hat{\mathfrak{R}}_D(\mathcal{H}^j_s)$ of the restriction of the vector-valued function class $\mathcal{H}^j$ along the s-th coordinate, the desired bound can be derived. □

We first introduce the following lemmas:

Lemma 29 (Corollary 1 in Maurer (2016)). *Let $\mathcal{X}$ be any set, $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathcal{X}^n$, let $\mathcal{F}$ be a class of vector-valued functions $\boldsymbol{f} : \mathcal{X} \to \ell_2$ and let $h_i : \ell_2 \to \mathbb{R}$ have Lipschitz norm L. Then*

$$
\mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_i \epsilon_i h_i\left(\boldsymbol{f}\left(\boldsymbol{x}_i\right)\right) \leq \sqrt{2}L\mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_{i,j}\epsilon_{ij} f_j\left(\boldsymbol{x}_i\right),
$$

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

Lemma 30 (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

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

We denote the restriction of the vector-valued function class $\mathcal{F}$ along the j-th coordinate as $\mathcal{F}_j$, $f_j \in \mathcal{F}_j$. Then, we have

$$
\mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_{i,j}\epsilon_{ij} f_j\left(\boldsymbol{x}_i\right) \leq \sum_j \mathbb{E}\sup_{f_j\in\mathcal{F}_j}\sum_i \epsilon_i f_j\left(\boldsymbol{x}_i\right) \leq q\max_j \mathbb{E}\sup_{f_j\in\mathcal{F}_j}\sum_i \epsilon_i f_j\left(\boldsymbol{x}_i\right).
$$

Combining with Lemma 29, we have

$$
\mathbb{E}\sup_{\boldsymbol{f}\in\mathcal{F}}\sum_i \epsilon_i h_i\left(\boldsymbol{f}\left(\boldsymbol{x}_i\right)\right) \leq \sqrt{2}Lq\max_j \mathbb{E}\sup_{f_j\in\mathcal{F}_j}\sum_i \epsilon_i f_j\left(\boldsymbol{x}_i\right),
$$

which means that the Rademacher complexity of the vector-valued function class $\mathcal{F}$ composited with an ℓ_2 Lipschitz function can be bounded by the maximum Rademacher complexity of the restrictions of the function class $\mathcal{F}$ along each coordinate and a multiplicative factor q, i.e.,

$$\hat{\mathfrak{R}}_D(h(\mathcal{F})) \leq \sqrt{2}Lq \max_j \hat{\mathfrak{R}}_D(\mathcal{F}_j). \tag{4}$$

According to the above inequality, we have

$$\begin{aligned}
\hat{\mathfrak{R}}_D(\mathcal{L}) \leq & \sqrt{2}\mu q \max_j \hat{\mathfrak{R}}_D(\mathcal{F}_j) \quad \text{(Use inequality (4) and Assumption 6)} \\
= & \sqrt{2}\mu q \max_j \hat{\mathfrak{R}}_D(\mathcal{G}(\mathcal{H}^j)) \quad \text{(Use the definitions of classes (1) and (2))} \\
\leq & \sqrt{2}\mu q \max_j \sqrt{2}k \max_s \hat{\mathfrak{R}}_D(\mathcal{H}_s^j) \quad \text{(Use inequality (4) and } g \text{ is 1-}\ell_2 \text{ Lipschitz)} \\
= & 2\mu q k \max_{j,s} \hat{\mathfrak{R}}_D(\mathcal{H}_s^j).
\end{aligned}$$

Then, we upper bound the Rademacher complexity of the loss function space $\mathcal{L}$ as the following:

$$\begin{aligned}
& \hat{\mathfrak{R}}_D(\mathcal{L}) \\
\leq & 2\mu q k \max_{j,s} \hat{\mathfrak{R}}_D(\mathcal{H}_s^j) \\
= & 2\mu q k \max_{j,s} \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{h_s^j \in \mathcal{H}_s^j} \frac{1}{n} \sum_{i=1}^n \epsilon_i h_s^j(\boldsymbol{x}_i) \right] \\
= & 2\mu q k \max_{\|\phi_s^j(\boldsymbol{x}_i)\| \leq A, s} \frac{1}{n} \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{\|\boldsymbol{w}_s^j\| \leq \Lambda} \sum_{i=1}^n \epsilon_i \langle \boldsymbol{w}_s^j, \phi_s^j(\boldsymbol{x}_i) \rangle \right] \\
\leq & 2\mu q k \max_{\|\phi_s^j(\boldsymbol{x}_i)\| \leq A, s} \frac{\Lambda}{n} \mathbb{E}_{\boldsymbol{\epsilon}} \| \sum_{i=1}^n \epsilon_i \phi_s^j(\boldsymbol{x}_i) \| \\
\leq & 2\mu q k \max_{\|\phi_s^j(\boldsymbol{x}_i)\| \leq A, s} \frac{\Lambda}{n} \left[\mathbb{E}_{\boldsymbol{\epsilon}} \| \sum_{i=1}^n \epsilon_i \phi_s^j(\boldsymbol{x}_i) \|^2 \right]^{\frac{1}{2}} \quad \text{(Use Jensen's Inequality)} \\
\leq & 2\mu q k \max_{\|\phi_s^j(\boldsymbol{x}_i)\| \leq A, s} \frac{\Lambda}{n} \left[\sum_{i=1}^n \| \phi_s^j(\boldsymbol{x}_i) \|^2 \right]^{\frac{1}{2}} \quad \text{(Use Lemma 30)} \\
\leq & \frac{2\mu q k B}{\sqrt{n}}. \quad \text{(Use Assumption 4 and 5)}
\end{aligned}$$

Combining with (3), then we have

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + \frac{4\mu q k B}{\sqrt{n}} + 3M \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

A.4.3 Proof of Lemma 9

Proof Sketch: First, according to the ℓ_2 norm Lipschitz continuity of the loss function, we derive the relationship between the empirical ℓ_2 norm covering number $\mathcal{N}_2(\epsilon, \mathcal{L}, D)$ of the loss space $\mathcal{L}$ and the empirical ℓ_2 norm covering number $\mathcal{N}_2(\epsilon, \mathcal{R}(\mathcal{F}), [q] \times D)$ of the outer decomposition function class $\mathcal{R}(\mathcal{F})$. Second, according to Sudakov's minoration Wainwright (2019) and the relationship between Rademacher and Gaussian complexity, we show that the empirical ℓ_2 norm covering number of the outer decomposition function class $\mathcal{R}(\mathcal{F})$ can be bounded by the Rademacher complexity of the outer decomposition function class $\mathcal{R}(\mathcal{F})$. Third, combining the above results and the refined Dudley's entropy integral inequality, the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded by the Rademacher complexity of the outer decomposition function class $\mathcal{R}(\mathcal{F})$. Fourth, **we find that the square-root dependency on q mainly comes from the**

$\sqrt{q}$ factor in the radius of the empirical ℓ_2 cover of the outer decomposition function class $\mathcal{R}(\mathcal{F})$, which is inevitable for ℓ_2 Lipschitz loss. However, the dependency on k can be further improved. Hence, we no longer use Sudakov's minoration and the relationship between Rademacher and Gaussian complexity to prove the $\widetilde{O}(\sqrt{q}/\sqrt{n})$ bound for MDC with ℓ_2 Lipschitz Loss, because they prevent us from obtaining $\widetilde{O}(\sqrt{q}/\sqrt{n})$ bounds with no dependency on k. The main reason is that we find that a factor of $\sqrt{k}$ in the radius of the empirical ℓ_2 cover of the inner decomposition function class cannot be eliminated by Sudakov's minoration. We can eliminate the $\sqrt{k}$ factor and improve the dependency on k to be independent by exploring the properties of the decomposition function classes. We derive the relationship between the empirical ℓ_2 norm covering number $\mathcal{N}_2(\epsilon, \mathcal{R}(\mathcal{F}), [q] \times D)$ of the outer decomposition function class $\mathcal{R}(\mathcal{F})$ and the empirical ℓ_2 norm covering number $\mathcal{N}_2(\epsilon, \mathcal{T}(\mathcal{H}^j), [k] \times [q] \times D)$ of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$, and we show that the empirical ℓ_2 norm covering number of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$ can be bounded by its worst-case Rademacher complexity. Fifth, we combine the above results and use the refined Dudley's entropy integral inequality again to convert the Rademacher complexity of the outer decomposition function class $\mathcal{R}(\mathcal{F})$ to the worst-case Rademacher complexity of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$. Finally, we estimate the lower bound of the worst-case Rademacher complexity of $\mathcal{T}(\mathcal{H}^j)$, and then combined with the above steps, the Rademacher complexity of the loss function space associated with $\mathcal{F}$ can be bounded.

We first refine the Dudley's entropy integral inequality (i.e., Proposition 22 in Ledent et al. (2021)) as follows, where we replace the function class taking values in $[0, 1]$ with the B-bounded function class, the refinement is obvious.

Lemma 31 (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 $\mathcal{N}_p(\epsilon, \mathcal{F}, S)$.*

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

Then, we introduce the following lemmas:

Lemma 32 (Sudakov's minoration (Wainwright, 2019))**.** *Let $\{Z_f, f \in \mathcal{F}\}$ be a zero-mean Gaussian process. Then*

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

Lemma 33 (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.

Next, our proof is divided into several steps:

Step 1: We first derive the relationship between the empirical ℓ_2 norm covering number $\mathcal{N}_2(\epsilon, \mathcal{L}, D)$ of the loss space $\mathcal{L}$ and the empirical ℓ_2 norm covering number $\mathcal{N}_2(\epsilon, \mathcal{R}(\mathcal{F}), [q] \times D)$ of the outer decomposition function class $\mathcal{R}(\mathcal{F})$.

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

$$\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\ell(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{y}_i) - \ell(\boldsymbol{f}'(\boldsymbol{x}_i), \boldsymbol{y}_i)\right)^2}$$

$$\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\mu^2\|\boldsymbol{f}(\boldsymbol{x}_i)-\boldsymbol{f}'(\boldsymbol{x}_i)\|_2^2} \quad \text{(Use Assumption 6)}$$

$$\leq \mu\sqrt{q\frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\sum_{j=1}^{q}\left(f_j\left(\boldsymbol{x}_i\right)-f_j'\left(\boldsymbol{x}_i\right)\right)^2}$$

$$\leq \mu\sqrt{q}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\sum_{j=1}^{q}\left(r_j(\boldsymbol{f}(\boldsymbol{x}_i))-r_j(\boldsymbol{f}'(\boldsymbol{x}_i))\right)^2}.$$

(The definition of the outer decomposition function class $\mathcal{R}(\mathcal{F})$)

Then, according to the definition of the empirical ℓ_2 norm covering number, we have that an empirical ℓ_2 norm cover of the outer decomposition function class $\mathcal{R}(\mathcal{F})$ at radius $\epsilon/\mu\sqrt{q}$ is also an empirical ℓ_2 norm cover of the loss function space $\mathcal{L}$ at radius ϵ, and we can conclude that:

$$\mathcal{N}_2\left(\epsilon,\mathcal{L},D\right)\leq\mathcal{N}_2\left(\frac{\epsilon}{\mu\sqrt{q}},\mathcal{R}(\mathcal{F}),[q]\times D\right). \tag{5}$$

Step 2: We show that the empirical ℓ_2 norm covering number of the outer decomposition function class $\mathcal{R}(\mathcal{F})$ can be bounded by the Rademacher complexity of the outer decomposition function class $\mathcal{R}(\mathcal{F})$. First, according to Sudakov's minoration (Wainwright, 2019), the ℓ_2 norm covering number of a function class can be bounded by the expectation of a Gaussian process indexed by this class. Then, the expectation of the Gaussian process can be bounded by the Rademacher complexity of the outer decomposition function class $\mathcal{R}(\mathcal{F})$.

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 $Z_f=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sigma_i f\left(\boldsymbol{x}_i\right)$, then according to Lemma 32, we have $\mathbb{E}\left[\sup_{f\in\mathcal{F}}Z_f\right]=\sqrt{n}\hat{\mathfrak{G}}_D(\mathcal{F})$. Then, combining Lemma 32 and Lemma 33, we have

$$\sqrt{\log\mathcal{N}_2(\epsilon,\mathcal{F},D)}\leq\frac{4}{\epsilon}\hat{\mathfrak{R}}_D(\mathcal{F})\sqrt{n\log n}\leq\frac{4}{\epsilon}\widetilde{\mathfrak{R}}_n(\mathcal{F})\sqrt{n\log n}. \tag{6}$$

Hence, for the outer decomposition function class $\mathcal{R}(\mathcal{F})$, we have

$$\sqrt{\log\mathcal{N}_2(\epsilon,\mathcal{R}(\mathcal{F}),[q]\times D)}\leq\frac{4}{\epsilon}\hat{\mathfrak{R}}_{[q]\times D}(\mathcal{R}(\mathcal{F}))\sqrt{nq\log nq}. \tag{7}$$

Step 3: According to Lemma 31 and combined with the above steps, we have

$$\begin{aligned}
\hat{\mathfrak{R}}_D(\mathcal{L}) &\leq \inf_{\alpha>0}\left(4\alpha+\frac{12}{\sqrt{n}}\int_{\alpha}^{M}\sqrt{\log\mathcal{N}_2(\epsilon,\mathcal{L},D)}d\epsilon\right)\\
&\leq \inf_{\alpha>0}\left(4\alpha+\frac{12}{\sqrt{n}}\int_{\alpha}^{M}\sqrt{\log\mathcal{N}_2(\frac{\epsilon}{\mu\sqrt{q}},\mathcal{R}(\mathcal{F}),[q]\times D)}d\epsilon\right) \quad \text{(Use inequality (5))}\\
&\leq \inf_{\alpha>0}\left(4\alpha+48q\mu\hat{\mathfrak{R}}_{[q]\times D}(\mathcal{R}(\mathcal{F}))\log^{\frac{1}{2}}(nq)\int_{\alpha}^{M}\epsilon^{-1}d\epsilon\right) \quad \text{(Use inequality (7))}\\
&\leq \inf_{\alpha>0}\left(4\alpha+48q\mu\hat{\mathfrak{R}}_{[q]\times D}(\mathcal{R}(\mathcal{F}))\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right).
\end{aligned} \tag{8}$$

Step 4: We find that the square-root dependency on q mainly comes from the $\sqrt{q}$ factor in the radius of the empirical ℓ_2 cover of the outer decomposition function class $\mathcal{R}(\mathcal{F})$, which is inevitable for ℓ_2 Lipschitz loss. However, the dependency on k can be further improved. Then, we no longer use Sudakov's minoration and the relationship between Rademacher and Gaussian complexity to prove the $\widetilde{O}(\sqrt{q}/\sqrt{n})$ bound for MDC with ℓ_2 Lipschitz Loss, because they prevent us from obtaining $\widetilde{O}(\sqrt{q}/\sqrt{n})$ bounds with no dependency on k. The main reason is that we find that a factor of $\sqrt{k}$ in the radius of the empirical ℓ_2 cover of the inner decomposition function class cannot be eliminated by Sudakov's minoration. We can eliminate the $\sqrt{k}$ factor and improve the dependency on k to be independent by the following lemmas and the inner decomposition function class.

Lemma 34 (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 $\epsilon > \hat{\mathfrak{R}}_S(\mathcal{F})$, we have that* $\mathrm{fat}_\epsilon(\mathcal{F}) \leq \frac{4n\hat{\mathfrak{R}}_S^2(\mathcal{F})}{\epsilon^2}$.

Lemma 35 (Theorem 12.8 in Anthony & Bartlett (2009)). *For any function in $\mathcal{F}$ takes values in $[-B, B]$ and any S with sample of size n, $\epsilon > 0$, $n \geq d$,*

$$\log \mathcal{N}_\infty(\epsilon, \mathcal{F}, S) \leq 1 + \mathrm{fat}_{\epsilon/4}(\mathcal{F}) \log_2 \frac{4eBn}{d\epsilon} \log \frac{4nB^2}{\epsilon^2}.$$

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

$$\begin{aligned}
&\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\sum_{j=1}^{q}\left(r_j(\boldsymbol{f}(\boldsymbol{x}_i)) - r_j(\boldsymbol{f}'(\boldsymbol{x}_i))\right)^2} \\
=&\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\sum_{j=1}^{q}\left(f_j(\boldsymbol{x}_i) - f_j'(\boldsymbol{x}_i)\right)^2} \\
\leq&\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\sum_{j=1}^{q}\max_{i\in[n],j\in[q]}\left(f_j(\boldsymbol{x}_i) - f_j'(\boldsymbol{x}_i)\right)^2} \\
\leq&\max_{i\in[n],j\in[q]}|f_j(\boldsymbol{x}_i) - f_j'(\boldsymbol{x}_i)| \\
=&\max_i\max_j|g(\boldsymbol{h}^j(\boldsymbol{x}_i)) - g(\boldsymbol{h}^{j'}(\boldsymbol{x}_i))| \\
\leq&\max_i\max_j\|\boldsymbol{h}^j(\boldsymbol{x}_i) - \boldsymbol{h}^{j'}(\boldsymbol{x}_i)\|_\infty \\
=&\max_i\max_j\max_s|h_s^j(\boldsymbol{x}_i) - h_s^{j'}(\boldsymbol{x}_i)| \\
=&\max_i\max_j\max_s|t_s(\boldsymbol{h}^j(\boldsymbol{x}_i)) - t_s(\boldsymbol{h}^{j'}(\boldsymbol{x}_i))|.
\end{aligned}$$

(The definition of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$)

Then, according to the definition of the empirical covering numbers, we have that an empirical ℓ_∞ norm cover of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$ at radius ϵ is also an empirical ℓ_2 norm cover of the outer decomposition function class $\mathcal{R}(\mathcal{F})$ at radius ϵ, and we can conclude that:

$$\mathcal{N}_2(\epsilon, \mathcal{R}(\mathcal{F}), [q] \times D) \leq \mathcal{N}_\infty\left(\epsilon, \mathcal{T}(\mathcal{H}^j), [k] \times [q] \times D\right). \tag{9}$$

Step 5: According to Assumption 4 and 5 in the main paper, we can obtain the lower bound of the worst-case Rademacher complexity $\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))$ by the Khintchine-Kahane inequality with $p = 1$.

First, we define $\phi^{*j}_s : \mathcal{X} \mapsto \mathbb{R}^d$ as:

$$\phi^{*j}_s(\boldsymbol{x}) = (A \cdot |\mathrm{sign}^*(x_1)|, 0, 0, \cdots, 0)^\top, \quad \forall \boldsymbol{x} = (x_1, x_2, \cdots, x_d)^\top \in \mathcal{X},$$

where the modified sign function $\mathrm{sign}^*(t)$ is defined as:

$$\mathrm{sign}^*(t) = \begin{cases} 1, & t \geq 0 \\ -1, & t < 0 \end{cases},$$

i.e.,

$$\phi^{*j}_s(\boldsymbol{x})_t = \begin{cases} A \cdot |\mathrm{sign}^*(x_1)|, & t = 1 \\ 0, & t = 2, \cdots, d \end{cases},$$

for any $\boldsymbol{x} \in \mathcal{X}$, $\|\phi^{*j}_s(\boldsymbol{x})\|_2 = \sqrt{(A \cdot |\mathrm{sign}^*(x_1)|)^2 + 0^2 + \cdots + 0^2} = A \cdot |\mathrm{sign}^*(x_1)| = A$.

Then, we define the weight vector as $\boldsymbol{w}^{*j}_s = (\Lambda, 0, 0, \cdots, 0) = \Lambda \boldsymbol{e}_1$, $\|\boldsymbol{w}^{*j}_s\|_2 = \Lambda$. For any sample $\boldsymbol{x}_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_d})$,

$$h^{*j}_s(\boldsymbol{x}_i) = \langle \boldsymbol{w}^{*j}_s, \phi^{*j}_s(\boldsymbol{x}_i) \rangle = \Lambda A \cdot |\mathrm{sign}^*(x_{i_1})| = \Lambda A := B.$$

We denote the set of functions consisting of ${h^*}^j_s$ as $\mathcal{T}^*$, and we have ${h^*}^j_s \in \mathcal{T}(\mathcal{H}^j)$, $\mathcal{T}^* \subseteq \mathcal{T}(\mathcal{H}^j)$. According to the Khintchine-Kahane inequality with $p = 1$:

$$
\begin{aligned}
&\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \geq \widetilde{\Re}_{nqk}(\mathcal{T}^*) \\
=&\sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n} \hat{\Re}_{[k]\times[q]\times D}(\mathcal{T}^*) \\
=&\sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n} \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{{h^*}^j_s\in\mathcal{T}^*}\frac{1}{nqk}\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}{h^*}^j_s(\boldsymbol{x}_i)\right] \\
=&\sup_{\|\phi^{*j}_s(\boldsymbol{x})\|=A:i,j,k} \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\|\boldsymbol{w}^{*j}_s\|=\Lambda}\frac{1}{nqk}\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}\langle\boldsymbol{w}^{*j}_s,\phi^{*j}_s(\boldsymbol{x}_i)\rangle\right] \\
=&\sup_{\|\phi^{*j}_s(\boldsymbol{x})\|=A:i,j,k} \frac{\Lambda}{nqk}\mathbb{E}_{\boldsymbol{\epsilon}}\|\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}\phi^{*j}_s(\boldsymbol{x}_i)\| \\
\geq&\sup_{\|\phi^{*j}_s(\boldsymbol{x})\|=A:i,j,k} \frac{\Lambda}{nqk}\frac{1}{\sqrt{2}}\left[\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\|\phi^{*j}_s(\boldsymbol{x}_i)\|^2\right]^{\frac{1}{2}}. \quad \text{(Use Lemma 30)} \\
=&\frac{\Lambda}{nqk}\frac{1}{\sqrt{2}}\left[\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}A^2\right]^{\frac{1}{2}} \\
=&\frac{\Lambda A}{\sqrt{2nqk}} := \frac{B}{\sqrt{2nqk}}.
\end{aligned} \tag{10}
$$

Step 6: According to Lemma 35, we have

$$
\begin{aligned}
&\log\mathcal{N}_\infty\left(\epsilon,\mathcal{T}(\mathcal{H}^j),[k]\times[q]\times D\right) \\
\leq&1+\mathrm{fat}_{\epsilon/4}(\mathcal{T}(\mathcal{H}^j))\log_2^2\frac{4eB^2nqk}{\epsilon^2} \\
\leq&1+\frac{64nqk\widetilde{\Re}^2_{nqk}(\mathcal{T}(\mathcal{H}^j))}{\epsilon^2}\log_2^2\frac{4eB^2nqk}{\epsilon^2}. \\
&\text{(Use inequality } \mathrm{fat}_\epsilon(\mathcal{T}(\mathcal{H}^j)) \leq \frac{4nqk\widetilde{\Re}^2_{nqk}(\mathcal{T}(\mathcal{H}^j))}{\epsilon^2}\text{, i.e., Lemma 34)}
\end{aligned} \tag{11}
$$

Then, according to Dudley's entropy integral inequality and combined with inequality (8), we have

$$
\begin{aligned}
&\hat{\Re}_D(\mathcal{L}) \\
\leq&\inf_{\alpha>0}\left(4\alpha+48q\mu\hat{\Re}_{[q]\times D}(\mathcal{R}(\mathcal{F}))\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
\leq&\inf_{\alpha>0}\left(4\alpha+48q\mu\inf_{\beta>0}\left(4\beta+\frac{12}{\sqrt{nq}}\int_\beta^B\sqrt{\log\mathcal{N}_2(\epsilon,\mathcal{R}(\mathcal{F}),[q]\times D)}d\epsilon\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
\leq&\inf_{\alpha>0}\left(4\alpha+48q\mu\inf_{\beta>0}\left(4\beta+\frac{12}{\sqrt{nq}}\int_\beta^B\sqrt{\log\mathcal{N}_\infty\left(\epsilon,\mathcal{T}(\mathcal{H}^j),[k]\times[q]\times D\right)}d\epsilon\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
\leq&\inf_{\alpha>0}\left(4\alpha+48q\mu\inf_{\beta>0}\left(4\beta+\frac{12}{\sqrt{nq}}\int_\beta^B\sqrt{1+\frac{64nqk\widetilde{\Re}^2_{nqk}(\mathcal{T}(\mathcal{H}^j))}{\epsilon^2}\log_2^2\frac{4eB^2nqk}{\widetilde{\Re}^2_{nqk}(\mathcal{T}(\mathcal{H}^j))}}d\epsilon\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right) \\
\leq&\inf_{\alpha>0}\left(4\alpha+48q\mu\inf_{\beta>0}\left(4\beta+\frac{12}{\sqrt{nq}}\int_\beta^B\sqrt{1+\frac{64nqk\widetilde{\Re}^2_{nqk}(\mathcal{T}(\mathcal{H}^j))}{\epsilon^2}\log_2^2(8en^2q^2k^2)}d\epsilon\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \inf_{\alpha>0}\left(4\alpha + 48q\mu \inf_{\beta>0}\left(4\beta + \frac{12B}{\sqrt{nq}} + 96\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))\log_2(8en^2q^2k^2)\int_\beta^B \epsilon^{-1}d\epsilon\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\right)\\
&\leq \inf_{\alpha>0}\Bigg(4\alpha + \frac{576\mu B\sqrt{q}}{\sqrt{n}}\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha} + \\
&\quad 48q\mu \inf_{\beta>0}\left(4\beta + 96\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))\log_2(8en^2q^2k^2)\log\frac{B}{\beta}\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\Bigg)\\
&\leq \inf_{\alpha>0}\Bigg(4\alpha + \frac{576\mu B\sqrt{q}}{\sqrt{n}}\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha} + \\
&\quad 2q\mu 48^2\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))(1+\log_2(8en^2q^2k^2)\log\frac{B}{24\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))})\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\Bigg)\\
&\quad \text{(Choose } \beta = 24\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))\text{)}\\
&\leq \inf_{\alpha>0}\Bigg(4\alpha + \frac{576\mu B\sqrt{q}}{\sqrt{n}}\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha} + \\
&\quad 2q\mu 48^2\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))\left(1+\log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{\alpha}\Bigg) \quad \text{(Use inequality (10))}\\
&\leq \frac{576\mu B\sqrt{q}}{\sqrt{n}}\log^{\frac{1}{2}}(nq)\log\frac{M}{24\cdot 48\mu q\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))} + \\
&\quad 2q\mu 48^2\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))\left(1+\left(1+\log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M}{24\cdot 48\mu q\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))}\right)\\
&\quad \text{(Choose } \alpha = 24\cdot 48\mu q\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))\text{)}\\
&\leq \frac{576\mu B\sqrt{q}}{\sqrt{n}}\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B} + 2q\mu 48^2\sqrt{k}\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))\left(1+\left(1+\log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B}\right).\\
&\quad \text{(Use inequality (10))}
\end{aligned}
$$

A.4.4 Proof of Theorem 12

We upper bound the worst-case Rademacher complexity $\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))$ of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$ as the following:

$$
\begin{aligned}
&\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))\\
&= \sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n}\hat{\Re}_{[k]\times[q]\times D}(\mathcal{T}(\mathcal{H}^j))\\
&= \sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{t_s(\boldsymbol{h}^j(\boldsymbol{x}_i))\in\mathcal{T}(\mathcal{H}^j)}\frac{1}{nqk}\sum_{i=1}^n\sum_{j=1}^q\sum_{s=1}^k\epsilon_{ijs}t_s(\boldsymbol{h}^j(\boldsymbol{x}_i))\right]\\
&= \sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{h_s^j\in\mathcal{H}_s^j}\frac{1}{nqk}\sum_{i=1}^n\sum_{j=1}^q\sum_{s=1}^k\epsilon_{ijs}h_s^j(\boldsymbol{x}_i)\right]\\
&= \sup_{\|\phi_s^j(\boldsymbol{x})\|\leq A:i,j,k}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\|\boldsymbol{w}_s^j\|\leq\Lambda}\frac{1}{nqk}\sum_{i=1}^n\sum_{j=1}^q\sum_{s=1}^k\epsilon_{ijs}\langle\boldsymbol{w}_s^j,\phi_s^j(\boldsymbol{x})\rangle\right]\\
&= \sup_{\|\phi_s^j(\boldsymbol{x})\|\leq A:i,j,k}\frac{\Lambda}{nqk}\mathbb{E}_{\boldsymbol{\epsilon}}\|\sum_{i=1}^n\sum_{j=1}^q\sum_{s=1}^k\epsilon_{ijs}\phi_s^j(\boldsymbol{x})\|\\
&\leq \sup_{\|\phi_s^j(\boldsymbol{x})\|\leq A:i,j,k}\frac{\Lambda}{nqk}\left[\mathbb{E}_{\boldsymbol{\epsilon}}\|\sum_{i=1}^n\sum_{j=1}^q\sum_{s=1}^k\epsilon_{ijs}\phi_s^j(\boldsymbol{x})\|^2\right]^{\frac{1}{2}} \quad \text{(Use Jensen's Inequality)}
\end{aligned}
$$

$$\leq \sup_{\|\phi_s^j(\boldsymbol{x})\|\leq A:i,j,k} \frac{\Lambda}{nqk}\left[\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\|\phi_s^j(\boldsymbol{x})\|^2\right]^{\frac{1}{2}} \quad \text{(Use Lemma 30)}$$
$$\leq \frac{\Lambda A}{\sqrt{nqk}} := \frac{B}{\sqrt{nqk}}. \tag{12}$$

Then, according to Lemma 13 in the main paper, we have

$$\begin{aligned}
&\hat{\mathfrak{R}}_D(\mathcal{L})\\
\leq&\frac{576\mu B\sqrt{q}}{\sqrt{n}}\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B}+2q\mu 48^2\sqrt{k}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))\times\\
&\left(1+\left(1+\log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B}\right)\\
\leq&\frac{576\mu B\sqrt{q}}{\sqrt{n}}\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B}+\\
&\frac{2\sqrt{q}\mu 48^2 B}{\sqrt{n}}\left(1+\left(1+\log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B}\right).
\end{aligned}$$

Combining with (3), then we have

$$\begin{aligned}
R(\boldsymbol{f})\leq&\widehat{R}_D(\boldsymbol{f})+3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\frac{2\cdot 24^2\mu B\sqrt{q}}{\sqrt{n}}\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B}+\\
&\frac{96^2\mu\sqrt{q}B}{\sqrt{n}}\left(1+\left(1+\log_2(8en^2q^2k^2)\log\sqrt{nq}\right)\log^{\frac{1}{2}}(nq)\log\frac{M\sqrt{n}}{\mu B}\right).
\end{aligned}$$

A.4.5 Proof of Theorem 14

We have q dimensions, each with k classes. For each classification dimension $j\in[q]$, and each class $s\in[k]$, we take the nonlinear mapping as the identity transformation, and all weight vectors satisfy the norm bound $\|\boldsymbol{w}_s^j\|_2\leq 1, \forall j\in[q], s\in[k]$. For each dimension, we take the maximum over all class scores as the prediction for that dimension: $f_j(x)=\max_{s\in[k]}\{{\boldsymbol{w}_s^j}^\top\boldsymbol{x}\}$. The final prediction vector is $\boldsymbol{f}(x)=(f_1(\boldsymbol{x}),...,f_q(\boldsymbol{x}))\in\mathbb{R}^q$.

Inspired by Schliserman & Koren (2024), we use the approximate orthogonal vector set (Feldman, 2016) to construct hard instance.

Lemma 36 (Lemma 7 in Schliserman & Koren (2024))**.** *For any $d\geq 100$, there exists a set $U_d\subseteq\mathbb{R}^d$, satisfying:*

- *Size: $|U_d|\geq 2^{d/12}$;*
- *Unit norm: All $\boldsymbol{u}\in U_d$ are unit vectors, i.e., $\|\boldsymbol{u}\|_2=1$;*
- *Approximate orthogonality: For any distinct $\boldsymbol{u},\boldsymbol{v}\in U_d$, we have $\langle\boldsymbol{u},\boldsymbol{v}\rangle\leq 1/2$.*

Lemma 36 allows us to construct an exponentially large set of vectors in an $O(d)$-dimensional space, where the inner product of any two vectors is at most $1/2$.

To introduce the dependency on the precision ε, we define the parameter $J=\frac{12}{\varepsilon^2}$ to scale the base shattering construction by a factor of $1/\varepsilon^2$, so that we can get the required dependency on ε.

According to Lemma 36, we first construct vector sets A and B of size $|A|=2^{\frac{q}{156}}$ and $|B|=2^{\frac{q}{13}}$ respectively, where set A consists of $\frac{q}{13}$-dimensional nearly-orthogonal vectors and uses an arbitrary enumeration of this set $A=\{\boldsymbol{a}_1,\cdots,\boldsymbol{a}_{|A|}\}$, and set B consists of $\frac{12q}{13}$-dimensional nearly-orthogonal vectors and uses an arbitrary enumeration of this set $B=\{\boldsymbol{b}_1,\cdots,\boldsymbol{b}_{|B|}\}$. Let $\frac{\sqrt{q}}{2^{\frac{q}{312}}}\leq\varepsilon\leq\frac{1}{12}$, and we have $\frac{qJ}{13}\leq|A|$, $\frac{J}{12}\leq|B|$.

For every $\boldsymbol{a} \in A \subseteq \mathbb{R}^{\frac{q}{13}}$, we define the vector $\boldsymbol{u} \in \mathbb{R}^{qk}$ where $\boldsymbol{u}[(j-1)k+1] = \boldsymbol{a}[j]$ when $j \in [\frac{q}{13}]$, and all other elements are 0. We denote each vector $\boldsymbol{z} \in \{0,1\}^{\frac{qJ}{13}}$ as the sequence of $\frac{J}{12}$ vectors in $\{0,1\}^{\frac{12q}{13}}$, $\boldsymbol{z}^1, \cdots, \boldsymbol{z}^{\frac{J}{12}}$, where $\boldsymbol{z}^r = \boldsymbol{z}[(r-1)\frac{12q}{13}+1 : \frac{12qr}{13}]$ for each $r \in [\frac{J}{12}]$.

We partition any 0-1 label $\boldsymbol{y} \in \{0,1\}^{qJ/13}$ into $\frac{J}{12}$ independent blocks, each of size $\frac{12q}{13}$, i.e.,

$$\boldsymbol{y} = (\boldsymbol{y}^1, \cdots, \boldsymbol{y}^{\frac{J}{12}}), \quad \boldsymbol{y}^p \in \{0,1\}^{\frac{12q}{13}}.$$

For each block $\boldsymbol{y}^p$, we take the corresponding vector $\boldsymbol{b}_{\boldsymbol{y}^p}$ from the approximate orthogonal vector set B, which is the vector we use to encode the label of this block, and we can identify such labeling $\boldsymbol{y}$ with a sequence $(\boldsymbol{b}_{\boldsymbol{y}^1}, \cdots, \boldsymbol{b}_{\boldsymbol{y}^{\frac{J}{12}}}) \in B^{\frac{J}{12}}$. For each such $\boldsymbol{y}$, we define the vector $\boldsymbol{v} \in \mathbb{R}^{qk}$ where $\boldsymbol{v}[(j-1)k+1] = \boldsymbol{b}[j - \frac{q}{13}]$ when $\frac{q}{13}+1 \le j \le q$, and all other elements are 0.

This construction ensures that the remaining extra rows, over our hard instance distribution, will always output a value of 0, so that the maximum operation will always choose the output of the main row, without interfering with our label encoding.

The construction of our hard instance sample set D is the same as in Schliserman & Koren (2024), for any $j \in [\frac{12q}{13}]$ and $o \in [\frac{J}{12}]$, we define the sample $\boldsymbol{x}_{i,j} = \frac{1}{\sqrt{2}}\boldsymbol{e}_{\frac{12q}{13}(o-1)+j} + \frac{1}{\sqrt{2}}\boldsymbol{e}_{\frac{qJ}{13}+o}$, where $\boldsymbol{e}_t$ is the standard basis vector. This sample construction ensures that each sample has exactly two non-zero components, both equal to $1/\sqrt{2}$, so the input norm is exactly 1, satisfying the input norm constraint. We prove that set $D = \{\boldsymbol{x}_{j,o} : j \in [\frac{12q}{13}], o \in [\frac{J}{12}]\}$ can be shattered by $\mathcal{F}$.

Our loss function is the same as in Schliserman & Koren (2024), it takes the q-dimensional vector after the maximum operation as input:

$$\ell(\hat{\boldsymbol{y}}) = 2\sqrt{8} \max_{(\hat{\boldsymbol{z}},\boldsymbol{z},r,p)\in Z} \left\{ \frac{3}{\sqrt{8}}\varepsilon, \max\left\{ \frac{\varepsilon}{\sqrt{8}}, {\boldsymbol{b}'_{\boldsymbol{z}}}^{\top}\hat{\boldsymbol{y}} \right\} + \max\left\{ \frac{\varepsilon}{\sqrt{8}}, {\boldsymbol{a}'_{r+\frac{12q}{13}(p-1)}}^{\top}\hat{\boldsymbol{y}} \right\} \right\} - 7\varepsilon$$

where $\boldsymbol{a}'$ is the q-dimensional approximate orthogonal vector that embeds $\boldsymbol{a}$ into the first $\frac{q}{13}$ dimensions, with the remaining dimensions filled with 0, and $\boldsymbol{b}'$ is the q-dimensional approximate orthogonal vector that embeds $\boldsymbol{b}$ into the last $\frac{12q}{13}$ dimensions, with the remaining dimensions filled with 0. The set Z is the set of all possible combinations of labels and sample indices, which is defined as:

$$Z = \left\{ \boldsymbol{z} \in \{0,1\}^{\frac{12q}{13}}, \hat{\boldsymbol{z}} \in \{0,1\}^{\frac{qJ}{13}}, r \in \left[\frac{12q}{13}\right], p \in \left[\frac{J}{12}\right] \;\middle|\; \forall 1 \le b \le \frac{J}{12} : \hat{\boldsymbol{z}}^{(b)} = \boldsymbol{z}, \hat{\boldsymbol{z}}^{(p)}(r) = \varepsilon \right\}.$$

ℓ is the maximum of several linear functions, and the maximum of convex functions is still convex, so ℓ is convex. The gradient norm of all linear terms is 1, so the Lipschitz constant of their maximum is also 1.

We then introduce the following definition to show that set $D = \{\boldsymbol{x}_{j,o} : j \in [\frac{12q}{13}], o \in [\frac{J}{12}]\}$ can be shattered by $\mathcal{F}$.

Definition 37 (shattering with margin (Magen & Shamir, 2023; Schliserman & Koren, 2024))**.** *A class of functions $\mathcal{F}$ on an input domain $\mathcal{X}$ shatters m points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m \in \mathcal{X}$ with margin ε, if for all $\boldsymbol{y} \in \{0,1\}^m$ we can find $f \in \mathcal{F}$ such that for all $i \in [m]$, it holds $f(\boldsymbol{x}_i) \le -\varepsilon$ if $y_i = 0$ and $f(\boldsymbol{x}_i) \ge \varepsilon$ if $y_i = 1$.*

Based on the vectors we defined earlier, we construct the corresponding prediction matrix W, where

$$W = \varepsilon(\boldsymbol{u}_1\ \boldsymbol{u}_2\ \cdots\ \boldsymbol{u}_{\frac{qJ}{13}}\ \boldsymbol{v}_{\boldsymbol{y}^1}\ \boldsymbol{v}_{\boldsymbol{y}^2}\ \cdots\ \boldsymbol{v}_{\boldsymbol{y}^{\frac{J}{12}}}).$$

Then, we take any sample $\boldsymbol{x}_{i,j} \in D$, and compute the corresponding loss. Since $f_j(x) = \max_{s\in[k]}\{{\boldsymbol{w}_s^j}^{\top}\boldsymbol{x}\}$,

$$\hat{\boldsymbol{y}} = W'\boldsymbol{x}_{j,o} = \frac{1}{\sqrt{2}}\varepsilon\boldsymbol{a}'_{\frac{12q}{13}(o-1)+j} + \frac{1}{\sqrt{2}}\varepsilon\boldsymbol{b}'_{\boldsymbol{y}^o},$$

where W' is the output after the maximum operation.

Since $\boldsymbol{a}'$ and $\boldsymbol{b}'$ are orthogonal, we have

$$\ell(\hat{\boldsymbol{y}}) = 2\sqrt{8} \max_{(\hat{\boldsymbol{z}},\boldsymbol{z},r,p)\in Z} \left\{ \frac{3}{\sqrt{8}}\varepsilon, \max\left\{ \frac{\varepsilon}{\sqrt{8}}, \frac{\varepsilon}{\sqrt{2}} \boldsymbol{b}_{\boldsymbol{z}}{}^{\top} \boldsymbol{b}_{\boldsymbol{y}^o} \right\} + \max\left\{ \frac{\varepsilon}{\sqrt{8}}, \frac{\varepsilon}{\sqrt{2}} \boldsymbol{a}_{r+\frac{12q}{13}(p-1)}{}^{\top} \boldsymbol{a}_{\frac{12q}{13}(o-1)+j} \right\} \right\} - 7\varepsilon.$$

Next, we discuss two cases:

Case 1: $y_{j,o} = 1$. In this case, $\boldsymbol{y}^o[j] = 1$, we can find the matching term $\boldsymbol{z} = \boldsymbol{y}^o$, $r = j$, $p = o$. At this point, $\boldsymbol{b}_{\boldsymbol{z}}{}^{\top} \boldsymbol{b}_{\boldsymbol{y}^o} = 1$. Meanwhile, $\boldsymbol{a}_{r+\frac{12q}{13}(p-1)}{}^{\top} \boldsymbol{a}_{\frac{12q}{13}(o-1)+j} = \boldsymbol{a}_{j+\frac{12q}{13}(o-1)}{}^{\top} \boldsymbol{a}_{\frac{12q}{13}(o-1)+j} = 1$. Hence, the sum of these two terms is 2. Substituting into the loss function, we have $\ell(\hat{\boldsymbol{y}}) = 2\sqrt{8} \times \sqrt{2}\varepsilon - 7\varepsilon = \varepsilon$. Hence, we exactly get the required ε loss.

Case 2: $y_{j,o} = 0$. In this case, $\boldsymbol{y}^o[j] = 0$, although $\boldsymbol{b}_{\boldsymbol{z}}{}^{\top} \boldsymbol{b}_{\boldsymbol{y}^o}$ is still 1, but $\boldsymbol{a}_{r+\frac{12q}{13}(p-1)}$ and $\boldsymbol{a}_{\frac{12q}{13}(o-1)+j}$ are distinct approximate orthogonal vectors, so their inner product is at most $\frac{1}{2}$, so the term $\boldsymbol{a}_{r+\frac{12q}{13}(p-1)}{}^{\top} \boldsymbol{a}_{\frac{12q}{13}(o-1)+j}$ will be truncated to $\frac{\varepsilon}{\sqrt{8}}$. Hence, the sum of the two terms is $\frac{3\varepsilon}{\sqrt{8}}$. Substituting into the loss function, we have $\ell(\hat{\boldsymbol{y}}) = 2\sqrt{8} \times \frac{3\varepsilon}{\sqrt{8}} - 7\varepsilon = -\varepsilon$. Hence, we exactly get the required $-\varepsilon$ loss.

This proves that for any 0-1 label label $\boldsymbol{y}$, we can find the corresponding prediction, such that the loss of each sample is exactly $\pm\varepsilon$, which means our prediction class can shatter $\frac{qJ}{13} = \Omega(q/\varepsilon^2)$ samples with margin ε.

Finally, we derive the lower bound for the generalization error.

We have a sample set D of size $N = \frac{Cq}{\varepsilon^2}$, satisfying the shattering property: for any 0-1 label $\boldsymbol{y}$, there exists a prediction W, such that for any $\boldsymbol{x} \in D$, we have the population distribution $\mathcal{D}$ is defined as the uniform distribution over D. Take any training set $D' \sim \mathcal{D}^n$, satisfying $n \leq \frac{N}{2} = \frac{Cq}{2\varepsilon^2}$, i.e., the size of the training set is no more than half of the population sample set. We then construct the labels as follows: samples in the training set have label 0, and samples not in the training set have label 1.

According to the shattering property, there exists the corresponding prediction W, we can verify that it is the ERM solution: For samples in the training set $x \in D'$, their loss is all $-\varepsilon$, so the empirical risk is $\hat{R}(\boldsymbol{f}) = \frac{1}{|D'|} \sum_{x\in D'} \ell(\boldsymbol{f}_W(\boldsymbol{x})) = -\varepsilon$. This is the global minimum that the empirical risk can achieve, since the minimum of the loss function is exactly $-\varepsilon$, so W is the output of the ERM algorithm.

Then, we calculate the generalization error. In the population distribution, at least half of the samples are not in the training set (since $n \leq N/2$), the loss of these samples is ε, the remaining half of the samples are in the training set, their loss is $-\varepsilon$. Hence, the population risk is:

$$R(\boldsymbol{f}) = \frac{1}{|D|} \sum_{x\in D} \ell(\boldsymbol{f}_W(\boldsymbol{x})) \geq \frac{1}{2} \cdot \varepsilon + \frac{1}{2} \cdot (-\varepsilon) = 0.$$

Since the empirical risk is $\hat{R}(\boldsymbol{f}) = -\varepsilon$, the generalization error is:

$$R(\boldsymbol{f}) - \hat{R}(\boldsymbol{f}) \geq 0 - (-\varepsilon) = \varepsilon.$$

We have proved that when $n \leq \frac{Cq}{2\varepsilon^2}$, the generalization error of the ERM solution is at least ε. Rearranging this inequality, we have $\varepsilon = \Omega\left(\sqrt{\frac{q}{n}}\right)$.

A.5 Tight Bounds for MDC with ℓ_∞ Lipschitz Loss

A.5.1 Proof of Proposition 17

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

$$\begin{aligned} &|\ell_H(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) - \ell_H(\boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{y})| \\ =& \left| \frac{1}{q} \sum_{j=1}^{q} \ell_b\left(f_j(\boldsymbol{x}), y^j\right) - \frac{1}{q} \sum_{j=1}^{q} \ell_b\left(f'_j(\boldsymbol{x}), y^j\right) \right| \end{aligned}$$

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

Second, with the elementary inequality

$$
\left| \max\left\{a_1, \ldots, a_q\right\} - \max\left\{b_1, \ldots, b_q\right\} \right| \leq \max\left\{ \left|a_1 - b_1\right|, \ldots, \left|a_q - b_q\right| \right\},
$$

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

$$
\begin{aligned}
&\left| \ell_M(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{y}) - \ell_M\left(\boldsymbol{f}'(\boldsymbol{x}), \boldsymbol{y}\right) \right| \\
=& \left| \max_{j \in [q]} \{ \ell_b \left(f_j(\boldsymbol{x}), y^j \right) \} - \max_{j \in [q]} \{ \ell_b \left(f'_j(\boldsymbol{x}), y^j \right) \} \right| \\
\leq& \max_{j \in [q]} \left| \ell_b \left(f_j(\boldsymbol{x}), y^j \right) - \ell_b \left(f'_j(\boldsymbol{x}), y^j \right) \right| \\
\leq& \mu \max_{j \in [q]} \left| f_j(\boldsymbol{x}) - f'_j(\boldsymbol{x}) \right| \\
=& \mu \left\| \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}'(\boldsymbol{x}) \right\|_\infty .
\end{aligned}
$$

A.5.2 Proof of Lemma 18

Proof Sketch: First, according to the ℓ_∞ norm Lipschitz continuity of the loss function, we show that the empirical ℓ_∞ norm covering number $\mathcal{N}_\infty(\epsilon, \mathcal{L}, D)$ of the loss space $\mathcal{L}$ can **be directly bounded** by the empirical ℓ_∞ norm covering number $\mathcal{N}_\infty(\epsilon, \mathcal{T}(\mathcal{H}^j), [k] \times [q] \times D)$ of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$. Then, we show that the empirical ℓ_∞ norm covering number $\mathcal{N}_\infty(\epsilon, \mathcal{T}(\mathcal{H}^j), [k] \times [q] \times D)$ of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$ can be bounded by the fat-shattering dimension, and the fat-shattering dimension can be bounded by the worst-case Rademacher complexity of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$. Third, according to the refined Dudley's entropy integral inequality, the Rademacher complexity of the loss function space $\mathcal{L}$ can be bounded by the worst-case Rademacher complexity of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$. Combined with the above steps, the desired bound can be derived.

We first derive the relationship between the empirical ℓ_∞ norm covering number $\mathcal{N}_\infty(\epsilon, \mathcal{L}, D)$ of the loss space $\mathcal{L}$ and the empirical ℓ_∞ norm covering number $\mathcal{N}_\infty(\epsilon, \mathcal{T}(\mathcal{H}^j), [k] \times [q] \times D)$ of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$.

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

$$
\begin{aligned}
&\max_i \left| \ell(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{y}_i) - \ell(\boldsymbol{f}'(\boldsymbol{x}_i), \boldsymbol{y}_i) \right| \\
\leq& \mu \max_i \left\| \boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{f}'(\boldsymbol{x}_i) \right\|_\infty \quad \text{(Use Assumption 21)} \\
\leq& \mu \max_i \max_j \left| f_j\left(\boldsymbol{x}_i\right) - f'_j\left(\boldsymbol{x}_i\right) \right| \\
=& \mu \max_i \max_j \left| r_j(\boldsymbol{f}(\boldsymbol{x}_i)) - r_j(\boldsymbol{f}'(\boldsymbol{x}_i) \right| \\
& \text{(The definition of the outer decomposition function class } \mathcal{R}(\mathcal{F})) \\
=& \mu \max_i \max_j \left| g(\boldsymbol{h}^j(\boldsymbol{x}_i)) - g(\boldsymbol{h}^{j'}(\boldsymbol{x}_i)) \right| \\
\leq& \mu \max_i \max_j \left\| \boldsymbol{h}^j(\boldsymbol{x}_i) - \boldsymbol{h}^{j'}(\boldsymbol{x}_i) \right\|_\infty \quad \text{(Use } g \text{ is 1-}\ell_\infty \text{ Lipschitz)} \\
=& \mu \max_i \max_j \max_s \left| h^j_s\left(\boldsymbol{x}_i\right) - h^{j'}_s\left(\boldsymbol{x}_i\right) \right| \\
=& \mu \max_i \max_j \max_s \left| t_s(\boldsymbol{h}^j(\boldsymbol{x}_i)) - t_s(\boldsymbol{h}^{j'}(\boldsymbol{x}_i)) \right|. \\
& \text{(The definition of the inner decomposition function class } \mathcal{T}(\mathcal{H}^j))
\end{aligned}
$$

Then, according to the definition of the empirical ℓ_∞ norm covering number, we have that an empirical ℓ_∞ norm cover of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$ at radius ϵ/μ is also an empirical ℓ_∞ norm cover of the loss function space $\mathcal{L}$ at radius ϵ, and we can conclude that:

$$\mathcal{N}_\infty(\epsilon, \mathcal{L}, D) \leq \mathcal{N}_\infty\left(\frac{\epsilon}{\mu}, \mathcal{T}(\mathcal{H}^j), [k] \times [q] \times D\right). \tag{13}$$

Hence, according to Lemma 31, we have

$$\begin{aligned}
&\hat{\mathfrak{R}}_D(\mathcal{L}) \\
\leq &\inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^M \sqrt{\log \mathcal{N}_\infty(\epsilon, \mathcal{L}, D)}d\epsilon\right) \\
\leq &\inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^M \sqrt{\log \mathcal{N}_\infty(\frac{\epsilon}{\mu}, \mathcal{T}(\mathcal{H}^j), [k] \times [q] \times D)}d\epsilon\right) \quad \text{(Use inequality (13))} \\
\leq &\inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^M \sqrt{1 + \frac{64\mu^2 nqk\widetilde{\mathfrak{R}}^2_{nqk}(\mathcal{T}(\mathcal{H}^j))}{\epsilon^2}\log_2^2\frac{4eB^2\mu^2 nqk}{\epsilon^2}}d\epsilon\right) \quad \text{(Use inequality (11))} \\
\leq &\inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^M \sqrt{1 + \frac{64nqk\mu^2\widetilde{\mathfrak{R}}^2_{nqk}(\mathcal{T}(\mathcal{H}^j))}{\epsilon^2}\log_2^2(8en^2q^2k^2\mu^2)}d\epsilon\right) \quad \text{(Use inequality (10))} \\
\leq &\inf_{\alpha>0}\left(4\alpha + \frac{12M}{\sqrt{n}} + 96\mu\sqrt{qk}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))\log_2(8en^2q^2k^2\mu^2)\int_\alpha^M \epsilon^{-1}d\epsilon\right) \\
\leq &\frac{12M}{\sqrt{n}} + \inf_{\alpha>0}\left(4\alpha + 96\mu\sqrt{qk}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))\log_2(8en^2q^2k^2\mu^2)\cdot\ln\frac{M}{\alpha}\right) \\
\leq &\frac{12M}{\sqrt{n}} + 96\mu\sqrt{qk}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))(1 + \log_2(8en^2q^2k^2\mu^2)\cdot\ln\frac{M}{24\mu\sqrt{qk}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j)))}) \\
&\text{(Choose } \alpha = 24\mu\sqrt{qk}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))) \\
\leq &\frac{12M}{\sqrt{n}} + 96\mu\sqrt{qk}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))(1 + \log_2(8en^2q^2k^2\mu^2)\cdot\ln\frac{M\sqrt{n}}{\mu B}). \quad \text{(Use inequality (10))}
\end{aligned}$$

A.5.3 Proof of Theorem 20

According to the upper bound of the worst-case Rademacher complexity $\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))$ of the inner decomposition function class $\mathcal{T}(\mathcal{H}^j)$ in (12), and combined with Lemma 23 in the main paper, we have

$$\begin{aligned}
\hat{\mathfrak{R}}_D(\mathcal{L}) \leq &\frac{12M}{\sqrt{n}} + 96\mu\sqrt{qk}\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))(1 + \log_2(8en^2q^2k^2\mu^2)\cdot\ln\frac{M\sqrt{n}}{\mu B}) \\
\leq &\frac{12M}{\sqrt{n}} + \frac{96\mu B}{\sqrt{n}}(1 + \log_2(8en^2q^2k^2\mu^2)\cdot\ln\frac{M\sqrt{n}}{\mu B})
\end{aligned}$$

Combining with (3), then we have

$$R(\boldsymbol{f}) \leq \widehat{R}_D(\boldsymbol{f}) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{24M}{\sqrt{n}} + \frac{192\mu B}{\sqrt{n}}\left(1 + \log_2(8en^2q^2k^2\mu^2)\cdot\ln\frac{M\sqrt{n}}{\mu B}\right).$$

A.6 Generalization Bounds for different prediction models

A.6.1 Proof of Corollary 22

Since the prediction model is induced by a three-layer FNN, we have $h_s^j(\boldsymbol{x}_i) = \boldsymbol{w}_s^{j\top}\phi_s^j(\boldsymbol{x}_i) = \boldsymbol{w}_s^{j\top}\sigma(W_3^{j,s}\sigma(W_2^{j,s}\sigma(W_1^{j,s}\boldsymbol{x}_i)))$. We upper bound worst-case Rademacher complexity $\widetilde{\mathfrak{R}}_{nqk}(\mathcal{T}(\mathcal{H}^j))$ of the inner decomposition class with the prediction model induced by a three-layer FNN as follows.

According to the upper bound of worst-case Rademacher complexity $\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))$ in (12), we have

$$
\begin{aligned}
&\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \\
=&\sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{h_s^j\in\mathcal{H}_s^j}\frac{1}{nqk}\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}h_s^j\left(\boldsymbol{x}_i\right)\right] \\
=&\sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\|\boldsymbol{w}_s^j\|\leq\Lambda,\sigma}\frac{1}{nqk}\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}\boldsymbol{w}_s^{j\top}\sigma(W_3^{j,s}\sigma(W_2^{j,s}\sigma(W_1^{j,s}\boldsymbol{x}_i)))\right] \\
\leq&\sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n}\frac{\Lambda}{nqk}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{\sigma}\|\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}\sigma(W_3^{j,s}\sigma(W_2^{j,s}\sigma(W_1^{j,s}\boldsymbol{x}_i)))\| \\
\leq&\sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n}\frac{2\Lambda}{nqk}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{\|W_3^{j,s}\|\leq B_{FNN},\sigma}\|\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}W_3^{j,s}\sigma(W_2^{j,s}\sigma(W_1^{j,s}\boldsymbol{x}_i))\| \\
&\text{(ReLU activation is 1-Lipschitz)} \\
\leq&\sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n}\frac{2\Lambda B_{FNN}}{nqk}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{\sigma}\|\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}\sigma(W_2^{j,s}\sigma(W_1^{j,s}\boldsymbol{x}_i))\| \\
\leq&\sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n}\frac{2^2\Lambda B_{FNN}}{nqk}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{\|W_2^{j,s}\|\leq B_{FNN},\sigma}\|\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}W_2^{j,s}\sigma(W_1^{j,s}\boldsymbol{x}_i)\| \\
\leq&\sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n}\frac{2^2\Lambda B_{FNN}^2}{nqk}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{\sigma}\|\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}\sigma(W_1^{j,s}\boldsymbol{x}_i)\| \\
\leq&\sup_{[k]\times[q]\times D\in[k]\times[q]\times\mathcal{X}^n}\frac{2^3\Lambda B_{FNN}^2}{nqk}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{\|W_1^{j,s}\|\leq B_{FNN}}\|\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}W_1^{j,s}\boldsymbol{x}_i\| \\
\leq&\sup_{\|\boldsymbol{x}_i\|\leq R}\frac{2^3\Lambda B_{FNN}^3}{nqk}\mathbb{E}_{\boldsymbol{\epsilon}}\|\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}\boldsymbol{x}_i\| \\
\leq&\sup_{\|\boldsymbol{x}_i\|\leq R}\frac{2^3\Lambda B_{FNN}^3}{nqk}\left[\mathbb{E}_{\boldsymbol{\epsilon}}\|\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\epsilon_{ijs}\boldsymbol{x}_i\|^2\right]^{\frac{1}{2}}\quad\text{(Use Jensen's Inequality)} \\
\leq&\sup_{\|\boldsymbol{x}_i\|\leq R}\frac{2^3\Lambda B_{FNN}^3}{nqk}\left[\sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{s=1}^{k}\|\boldsymbol{x}_i\|^2\right]^{\frac{1}{2}}\quad\text{(Use Lemma 30)} \\
\leq&\frac{2^3\Lambda B_{FNN}^3R}{\sqrt{nqk}}.
\end{aligned}
$$

Combining with (3) and Lemma 23 in the main paper, we have

$$
R(\boldsymbol{f})\leq\widehat{R}_D(\boldsymbol{f})+3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}+\frac{24M}{\sqrt{n}}+\frac{192\mu2^3\Lambda B_{FNN}^3R}{\sqrt{n}}\left(1+\log_2(8en^2q^2k^2\mu^2)\cdot\ln\frac{M\sqrt{n}}{\mu B}\right).
$$

A.6.2 Proof of Theorem 24

Since the prediction model is induced by a one layer multi-head Transformer with H head, we have $h_s^j\left(\boldsymbol{x}\right)=\boldsymbol{w}_s^{j\top}\phi_s^j(\boldsymbol{x})=\boldsymbol{w}_s^{j\top}\boldsymbol{y}_{[\mathrm{CLS}]}^{j,s}=\sum_{u=1}^{H}\boldsymbol{w}_s^{j\top}\boldsymbol{y}_u^{j,s}=\sum_{u=1}^{H}v_u^{j,s}(X)$.

To keep the notation simple, we break down the upper bounding of the worst-case Rademacher complexity $\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))$ with the prediction model induced by a one layer multi-head Transformer

modularly. First, we give the worst-case Rademacher complexity result of a one layer multi-head Transformer independently, which can be used for the subsequent upper bounding of worst-case Rademacher complexity $\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))$.

To reduce the dependency of the worst-case Rademacher complexity for a one layer multi-head Transformer on H, we introduce the multi-head decomposition operator for the multi-head Transformer. We define a function class $\mathcal{P}$ consisting of the multi-head decomposition operators $p_u : \mathbb{R}^H \mapsto \mathbb{R}$ for any $u \in [H]$ which decompose the components corresponding to the u-th head of the multi-head Transformer, and we have $\mathcal{P}(\mathcal{V}) = \{(u, X) \mapsto p_u(\boldsymbol{v}(X)) : p_u(\boldsymbol{v}(X)) = \boldsymbol{w}^\top \boldsymbol{y}_u, V(X) \in \mathcal{V}, (u, X) \in [H] \times \mathcal{X}\}$.

For the dataset $D = \{X_1, \ldots, X_n\}$ with n i.i.d. examples:

$$
\begin{aligned}
&\max_i |V(X_i) - V'(X_i)| \\
\leq &\max_i |\sum_{u=1}^H v_u(X_i) - \sum_{u=1}^H v'_u(X_i)| \\
\leq &\max_i \sum_{u=1}^H |v_u(X_i) - v'_u(X_i)| \\
\leq &H \max_i \max_u |v_u(X_i) - v'_u(X_i)| \\
\leq &H \max_i \max_u |\boldsymbol{w}^\top \boldsymbol{y}_u - \boldsymbol{w}'^\top \boldsymbol{y}_u| \\
\leq &H \max_i \max_u |p_u(\boldsymbol{v}(X_i)) - p_u(\boldsymbol{v}'(X_i)|.
\end{aligned}
$$

(The definition of the multi-head decomposition function class $\mathcal{P}(\mathcal{V})$)

Then, according to the definition of the empirical ℓ_∞ covering number, we have that an empirical ℓ_∞ cover of $\mathcal{P}(\mathcal{V})$ at radius ϵ/H is also an empirical ℓ_∞ cover of $\mathcal{V}$ at radius ϵ, and we can conclude that:

$$
\mathcal{N}_\infty(\epsilon, \mathcal{V}, D) \leq \mathcal{N}_\infty\left(\frac{\epsilon}{H}, \mathcal{P}(\mathcal{V}), [H] \times D\right). \tag{14}
$$

Since $|v_u(X)| \leq B$ for any $u \in [H]$, we have $|V(X)| = |\sum_{u=1}^H v_u(X)| = \sum_{u=1}^H |v_u(X)| \leq HB$. According to Lemma 34 and Lemma 35, we have

$$
\begin{aligned}
&\log \mathcal{N}_\infty(\epsilon, \mathcal{P}(\mathcal{V}), [H] \times D) &(15)\\
\leq &1 + \mathrm{fat}_{\epsilon/4}(\mathcal{P}(\mathcal{V})) \log_2^2 \frac{4eB^2nH}{\epsilon^2} \\
\leq &1 + \frac{64nH\hat{\Re}^2_{[H]\times D}(\mathcal{P}(\mathcal{V}))}{\epsilon^2} \log_2^2 \frac{4eB^2nH}{\epsilon^2} \\
\leq &1 + \frac{64nH\widetilde{\Re}^2_{nH}(\mathcal{P}(\mathcal{V}))}{\epsilon^2} \log_2^2 \frac{4eB^2nH}{\epsilon^2}. &(16)
\end{aligned}
$$

According to Lemma 31 and combined with the above steps, we have

$$
\begin{aligned}
&\hat{\Re}_D(\mathcal{V}) \\
\leq &\inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{HB} \sqrt{\log \mathcal{N}_\infty(\epsilon, \mathcal{V}, D)} d\epsilon\right) \\
\leq &\inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{HB} \sqrt{\log \mathcal{N}_\infty(\frac{\epsilon}{H}, \mathcal{P}(\mathcal{V}), [H] \times D)} d\epsilon\right) \quad \text{(Use inequality (14))} \\
\leq &\inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{HB} \sqrt{1 + \frac{64nH \cdot H^2 \widetilde{\Re}^2_{nH}(\mathcal{P}(\mathcal{V}))}{\epsilon^2} \log_2^2 \frac{4eB^2nH \cdot H^2}{\widetilde{\Re}^2_{nH}(\mathcal{P}(\mathcal{V}))}} d\epsilon\right) \quad \text{(Use inequality (15))}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \inf_{\alpha>0}\left(4\alpha+\frac{12}{\sqrt{n}}\int_{\alpha}^{HB}\sqrt{1+\frac{64nH^3\widetilde{\Re}^2_{nH}(\mathcal{P}(\mathcal{V}))}{\epsilon^2}\log_2^2(8en^2H^4)}d\epsilon\right)\\
&\leq \inf_{\alpha>0}\left(4\alpha+\frac{12HB}{\sqrt{n}}+96H\sqrt{H}\widetilde{\Re}_{nH}(\mathcal{P}(\mathcal{V}))\log_2(8en^2H^4)\int_{\alpha}^{HB}\epsilon^{-1}d\epsilon\right)\\
&\leq \frac{12HB}{\sqrt{n}}+\inf_{\alpha>0}\left(4\alpha+96H\sqrt{H}\widetilde{\Re}_{nH}(\mathcal{P}(\mathcal{V}))\log_2(8en^2H^4)\cdot\ln\frac{HB}{\alpha}\right)\\
&\leq \frac{12HB}{\sqrt{n}}+96H\sqrt{H}\widetilde{\Re}_{nH}(\mathcal{P}(\mathcal{V}))(1+\log_2(8en^2H^4)\cdot\ln\frac{HB}{24H\sqrt{H}\widetilde{\Re}_{nH}(\mathcal{P}(\mathcal{V}))})\\
&\quad \text{(Choose } \alpha=24H\sqrt{H}\widetilde{\Re}_{nH}(\mathcal{P}(\mathcal{V}))\text{)}\\
&\leq \frac{12HB}{\sqrt{n}}+96H\sqrt{H}\widetilde{\Re}_{nH}(\mathcal{P}(\mathcal{V}))(1+\log_2(8en^2H^4)\cdot\ln(H\sqrt{n})). \qquad (17)
\end{aligned}
$$

Then, we can upper bound the worst-case Rademacher complexity $\widetilde{\Re}_{nH}(\mathcal{P}(\mathcal{V}))$ as the following:

$$
\begin{aligned}
&\widetilde{\Re}_{nH}(\mathcal{P}(\mathcal{V}))\\
=&\sup_{[H]\times D\in[H]\times\mathcal{X}^n}\hat{\Re}_{[H]\times D}(\mathcal{P}(\mathcal{V}))\\
=&\sup_{[H]\times D\in[H]\times\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{w}^\top\boldsymbol{y}_u\in\mathcal{V}}\frac{1}{nH}\sum_{i=1}^{n}\sum_{u=1}^{H}\epsilon_{iu}p_u(\boldsymbol{v}(X_i))\right]\\
=&\sup_{[H]\times D\in[H]\times\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{w}^\top\boldsymbol{y}_u\in\mathcal{V}}\frac{1}{nH}\sum_{i=1}^{n}\sum_{u=1}^{H}\epsilon_{iu}\boldsymbol{w}^\top W_{u,C}^\top\sigma\left(W_{u,V}^\top X_i^\top\text{softmax}\left(X_iW_{u,QK}^\top\boldsymbol{x}_{[\text{CLS}]}\right)\right)\right]\\
=&\sup_{[H]\times D\in[H]\times\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\|\boldsymbol{w}\|\leq\Lambda,\sigma}\frac{1}{nH}\sum_{i=1}^{n}\sum_{u=1}^{H}\epsilon_{iu}\langle\boldsymbol{w},W_{u,C}^\top\sigma\left(W_{u,V}^\top X_i^\top\text{softmax}\left(X_iW_{u,QK}^\top\boldsymbol{x}_{[\text{CLS}]}\right)\right)\rangle\right]\\
\leq&\sup_{[H]\times D\in[H]\times\mathcal{X}^n}\Lambda\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{\|W_{u,C}\|\leq B_{W_C},\sigma}\frac{1}{nH}\|\sum_{i=1}^{n}\sum_{u=1}^{H}\epsilon_{iu}W_{u,C}^\top\sigma\left(W_{u,V}^\top X_i^\top\text{softmax}\left(X_iW_{u,QK}^\top\boldsymbol{x}_{[\text{CLS}]}\right)\right)\|\\
\leq&\sup_{[H]\times D\in[H]\times\mathcal{X}^n}\frac{\Lambda B_{W_C}}{nH}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{\sigma}\|\sum_{i=1}^{n}\sum_{u=1}^{H}\epsilon_{iu}\sigma\left(W_{u,V}^\top X_i^\top\text{softmax}\left(X_iW_{u,QK}^\top\boldsymbol{x}_{[\text{CLS}]}\right)\right)\|\\
\leq&\sup_{[H]\times D\in[H]\times\mathcal{X}^n}\frac{2\Lambda B_{W_C}L_\sigma}{nH}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{W_{u,V},W_{u,QK}}\|\sum_{i=1}^{n}\sum_{u=1}^{H}\epsilon_{iu}W_{u,V}^\top X_i^\top\text{softmax}\left(X_iW_{u,QK}^\top\boldsymbol{x}_{[\text{CLS}]}\right)\|\\
\leq&\sup_{[H]\times D\in[H]\times\mathcal{X}^n}\frac{2\Lambda B_{W_C}B_{W_V}L_\sigma}{nH}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{W_{u,QK}}\|\sum_{i=1}^{n}\sum_{u=1}^{H}\epsilon_{iu}X_i^\top\text{softmax}\left(X_iW_{u,QK}^\top\boldsymbol{x}_{[\text{CLS}]}\right)\|\\
\leq&\sup_{[H]\times D\in[H]\times\mathcal{X}^n}\frac{2\Lambda B_{W_C}B_{W_V}L_\sigma}{nH}\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{i,W_{u,QK}}\|\sum_{i=1}^{n}\sum_{u=1}^{H}\epsilon_{iu}X_i^\top\|_{2,\infty}\|\text{softmax}\left(X_iW_{u,QK}^\top\boldsymbol{x}_{[\text{CLS}]}\right)\|_1\\
&\quad\text{(Use } \|P\boldsymbol{x}\|\leq\|P\|_{2,\infty}\|\boldsymbol{x}\|_1\text{)}\\
\leq&\frac{2\Lambda B_{W_C}B_{W_V}L_\sigma}{nH}\sup_{i,W_{u,QK}}\|\text{softmax}\left(X_iW_{u,QK}^\top\boldsymbol{x}_{[\text{CLS}]}\right)\|_1\mathbb{E}_{\boldsymbol{\epsilon}}\|\sum_{i=1}^{n}\sum_{u=1}^{H}\epsilon_{iu}X_i^\top\|_{2,\infty}\\
\leq&\frac{2\Lambda B_{W_C}B_{W_V}L_\sigma}{nH}\mathbb{E}_{\boldsymbol{\epsilon}}\|\sum_{i=1}^{n}\sum_{u=1}^{H}\epsilon_{iu}X_i^\top\|_{2,\infty}\quad\text{(Use } \|\text{softmax}\left(\theta\right)\|_1=1\text{ for }\theta\in\mathbb{R}^p\text{)}\\
=&\frac{2\Lambda B_{W_C}B_{W_V}L_\sigma}{nH}\max_t\mathbb{E}_{\boldsymbol{\epsilon}}\|\sum_{i=1}^{n}\sum_{u=1}^{H}\epsilon_{iu}\boldsymbol{x}_t^{i_u}\|
\end{aligned}
$$

$$\leq \frac{2\Lambda B_{W_C} B_{W_V} L_\sigma}{nH} \max_t \left[\mathbb{E}_{\boldsymbol{\epsilon}} \| \sum_{i=1}^{n} \sum_{u=1}^{H} \epsilon_{iu} \boldsymbol{x}_t^{i_u} \|^2 \right]^{\frac{1}{2}} \quad \text{(Use Jensen's Inequality)}$$

$$\leq \frac{2\Lambda B_{W_C} B_{W_V} L_\sigma}{nH} \max_t \left[\sum_{i=1}^{n} \sum_{u=1}^{H} \| \boldsymbol{x}_t^{i_u} \|^2 \right]^{\frac{1}{2}} \quad \text{(Use Lemma 30)}$$

$$\leq \frac{2\Lambda B_{W_C} B_{W_V} L_\sigma R}{\sqrt{nH}}. \tag{18}$$

Combining with inequality (17) and (18), we have the worst-case Rademacher complexity for a one layer multi-head Transformer as follows:

$$\begin{aligned}\hat{\Re}_D(\mathcal{V}) \leq & \frac{12HB}{\sqrt{n}} + 96H\sqrt{H}\widetilde{\Re}_{nH}(\mathcal{P}(\mathcal{V}))(1 + \log_2(8en^2H^4) \cdot \ln(H\sqrt{n})) \\ \leq & \frac{12HB}{\sqrt{n}} + \frac{96H \cdot 2\Lambda B_{W_C} B_{W_V} L_\sigma R}{\sqrt{n}}(1 + \log_2(8en^2H^4) \cdot \ln(H\sqrt{n})). \end{aligned} \tag{19}$$

With the above results, for the prediction model of MDC induced by a one layer multi-head Transformer with H head, i.e., $h_s^j(\boldsymbol{x}) = \sum_{u=1}^{H} \boldsymbol{w}_s^{j\top} \boldsymbol{y}_u^{j,s} = \sum_{u=1}^{H} v_u^{j,s}(X)$, we only need to slightly refine the definition of the multi-head decomposition function class for the multi-head Transformer to directly upper bound the worst-case Rademacher complexity $\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))$. Therefore, we define the multi-head decomposition function class under the framework for MDC as

$$\begin{aligned}\mathcal{P}(\mathcal{V}^{j,s}) = \{(u,s,j,X) \mapsto & p_u(\boldsymbol{v}^{j,s}(X)) : p_u(\boldsymbol{v}^{j,s}(X)) = v_u^{j,s}(X) = \boldsymbol{w}_s^{j\top} \boldsymbol{y}_u^{j,s}, \\ & h_s^j = V^{j,s}(X) \in \mathcal{V}^{j,s}, (u,s,j,X) \in [H] \times [k] \times [q] \times \mathcal{X}\}.\end{aligned}$$

Then, according to inequality (19), we have the worst-case Rademacher complexity $\widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j))$ for the prediction model of MDC induced by a one layer multi-head Transformer as the following:

$$\begin{aligned}& \widetilde{\Re}_{nqk}(\mathcal{T}(\mathcal{H}^j)) \\ \leq & \frac{12HB}{\sqrt{nqk}} + \frac{96H \cdot 2\Lambda B_{W_C} B_{W_V} L_\sigma R}{\sqrt{nqk}}(1 + \log_2(8e(nqk)^2H^4) \cdot \ln(H\sqrt{nqk})) \\ \leq & \frac{24H\Lambda B_{W_C} B_{W_V} L_\sigma R \times (1 + 8(1 + \log_2(8en^2q^2k^2H^4) \cdot \ln(H\sqrt{nqk})))}{\sqrt{nqk}}.\end{aligned}$$

Combining with Lemma 23 in the main paper, we have the following bound for MDC with the prediction model induced by a one layer multi-head Transformer:

$$\begin{aligned}R(\boldsymbol{f}) \leq & \widehat{R}_D(\boldsymbol{f}) + 3M\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + \frac{24M}{\sqrt{n}} + 192\mu \times \left(1 + \log_2(8en^2q^2k^2\mu^2) \cdot \ln\frac{M\sqrt{n}}{\mu B}\right) \times \\ & \frac{24H\Lambda B_{W_C} B_{W_V} L_\sigma R \times (1 + 8(1 + \log_2(8en^2q^2k^2H^4) \cdot \ln(H\sqrt{nqk})))}{\sqrt{n}}.\end{aligned}$$

A.7 Discussion

In this section, we mainly discuss and clarify the problem setting of MDC, the relationship between our generalization analysis and learnability, the impact of the dependencies among dimensions on the generalization bounds, future ideas for experimental verification, and the differences between our theoretical results and those of other learning theory problems.

A.7.1 Differences between multi-output problem and MDC

The universality of multi-output problems allows MDC to be regarded as a complex type of multi-output problem. However, due to the more complex definition of MDC, the existing theoretical

results on multi-output problems cannot be directly applied to analyze and understand the generalization of MDC. When the multi-class classification problem corresponding to each dimension degenerates to a binary classification problem (i.e., $k = 2$), MDC will degenerate to a multi-label problem, which is a common type of multi-output problem. In this case, the existing theoretical results on multi-output problems can be applied to this degenerate version of MDC, i.e., the multi-label problem.

Moreover, from the perspective of methods and applications, in MDC, the output space consists of multiple heterogeneous class spaces. This is the essential difference between MDC and other related multi-output classification problems (e.g., multi-class/multi-label classification) (Jia & Zhang, 2023), although this may be trivial from the perspective of theoretical analysis.

A.7.2 Connection with learnability

Here we provide a generalization analysis of MDC, and according to Raman et al. (2024), our generalization bound can, to some extent, serve as a sufficient condition for the learnability of MDC. Raman et al. (2024) provides a complete analysis of the learnability of multi-output problems, proving that a class of multi-output functions is learnable if and only if each scalar output restriction of that class is learnable. The relevant theoretical results in Raman et al. (2024) can be used in the learnability analysis of the MDC function class because both the surrogate Hamming loss and the surrogate Matching loss satisfy the assumptions about the loss function. Our generalization analysis provides a tight bound on the Rademacher complexity of the projective function class corresponding to MDC, and since the functions in the projective function class are consistent with those in the restrictions of the MDC function class, this implies that the complexities of the restrictions of the MDC function class are bounded, i.e., the restrictions of the MDC function class are learnable. Based on the characterization of the learnability for multi-output problems in Raman et al. (2024), we can conclude that the MDC function class is learnable. Therefore, the tight generalization bound we derive for MDC can be considered a sufficient condition for the learnability of the MDC function class. We leave a rigorous formal proof for future work.

A.7.3 The impact of dependencies among dimensions on bounds

There may be some potential confounding effects between class variables/spaces, i.e., dependencies among dimensions. Effectively modeling these dependencies can enhance the generalization performance of MDC methods. The dependencies among dimensions induced by different modeling strategies can also affect the generalization bound. We further discuss this as follows:

When it comes to dependencies among class variables, e.g., the graph structure dependencies in Gil-Begue et al. (2021); Nguyen et al. (2023), we need to introduce additional assumptions for the specific optimal graph. How to finely characterize the dependencies induced by the specific graph structure and introduce them into the generalization analysis is a challenge, which requires further exploration. However, our results can be regarded as a general theoretical guarantee for the worst case, since when considering dependencies, this is essentially a specific constraint on the increase in the hypothesis space, which will narrow the hypothesis space and lead to the tighter bounds.

Different types of dependencies have important consequences for generalization analysis, and how to explicitly introduce them in generalization analysis is a crucial open problem. Effective analysis for specific dependencies requires further explicit introduction of valid assumptions to reveal the impact of these dependencies induced factors on the bound. For example, for high-dimensional sparse data (corresponding to a special type of dependency to some extent), one may need to introduce the sparsity assumption into the analysis, thereby inducing bounds that are weakly dependent on the sparsity rate and the number of key (rather than whole) dimensions and labels.

A.7.4 Experimental verification

The main goal of this work is to provide general and efficient theoretical guarantees for empirically successful MDC methods, especially regarding the dependency on q and k. Results based on existing analysis methods show that these theoretical bounds of MDC are linear dependent on q and k, which means that these theoretical results scale rather unfavorably with qk and are difficult to explain empirical success. For example, since MDC is a more general case of multi-label learning to some extent, and there may be extreme cases in multi-label learning where the number of labels

(i.e., dimensions) is extremely large. In addition, there may also be extreme cases in multi-class classification corresponding to each dimension where the number of classes is extremely large. When both the number of dimensions and classes are extreme, effective theoretical bounds for MDC are lacking. Therefore, this suggests that only tighter bounds w.r.t. q and k can provide effective theoretical guarantees for empirically successful existing MDC methods. Our tighter $\widetilde{O}(1/\sqrt{n})$ bounds with no dependency on qk (up to logarithmic terms) solve the limitations of the $O(qk/\sqrt{n})$ results based on existing analysis methods and can provide a general theoretical guarantee for the empirical success of existing methods. Our general capacity-based bounds for MDC do not involve specific assumptions related to the algorithm.

The empirical success of existing MDC methods on datasets of varying sizes q and k suggests that the ideal bound should not be strongly dependent on q and k. The good generalization performance of these existing MDC methods serves as empirical validation of our theoretical results. We leave the work of experimental verification to future work, and we mainly consider experimental verification from two aspects.

On the one hand, verify whether the generalization bound proposed in the paper accurately captures the impact of the number of dimensions and classes on generalization performance. Specifically, check whether the generalization bound is consistent with the experimentally observed trend of the generalization gap (GAP) as q and k vary. GAP is defined as the difference between the generalization error and the empirical error, where the generalization error is estimated on the test set and the empirical error is calculated on the training set. Select benchmark datasets, construct sub-datasets with different numbers of dimensions and classes, repeat training multiple times on each sub-dataset, and average the GAP. Compare the GAP curves and generalization bound curves for different numbers of dimensions and classes.

On the other hand, verify the impact of different dependencies among the dimensions or class variables involved in different methods and the impact of different regularizations involved in different methods on generalization performance, and explore which dimensional dependency modeling methods and which regularization-induced inductive biases are more effective for generalization in practice, and promote further theoretical research.

A.7.5 Comparative analysis with multi-class learning

From the perspective of label space, the label space of MDC can be viewed as the Cartesian product of labels from q dimensions, which is equivalent to the label space of a multi-class classification task with k^q classes.

While Exact Match loss can be trivially reduced to a single multi-class problem over a label set of size k^q, such a reduction does not capture the essence or the challenges of MDC and discards the decomposable structure of MDC. Our upper bound does apply to the Exact Match setting, but for hypothesis classes with a decomposable structure (i.e., each dimension is predicted by an independent f_j). For such decomposable structure classes, we prove Rademacher complexity upper bounds of order $O(\frac{\sqrt{q}\log k}{\sqrt{n}})$ for ℓ_2 loss and order $O(\frac{\log qk}{\sqrt{n}})$ for ℓ_∞ loss. This result does not contradict the Natarajan dimension lower bound, because the Natarajan lower bound applies to unstructured hypothesis classes (arbitrary functions mapping $\mathcal{X}$ to k^q labels). Our decomposable structure hypothesis class is a strict subset, and thus can have much lower complexity. Below we clarify this in detail from three perspectives.

The hypothesis class we analyze is different. If we strictly follow the reduction of MDC to a single multi-class problem, then: the output space size is k^q, the hypothesis class $\mathcal{F}_{\text{flat}}$ maps the input $\mathcal{X}$ directly to $\{1,\dots,k^q\}$. Generalization bounds based on the Natarajan dimension typically scale as $O\left(\sqrt{d_{\text{NAT}}(\mathcal{F}_{\text{flat}})\log k^q/n}\right) = O\left(\sqrt{d_{\text{NAT}}(\mathcal{F}_{\text{flat}})q\log k/n}\right)$ (Daniely et al., 2015). This matches our $\widetilde{O}(\frac{\sqrt{q}\log k}{\sqrt{n}})$ bound for MDC with ℓ_2 Lipschitz loss when we do not consider the possibility that $d_{\text{NAT}}(\mathcal{F}_{\text{flat}})$ may implicitly depend on q, k.

However, we analyze a structured hypothesis class $\mathcal{F}$ defined by (1) and (2) (rather than $\mathcal{F}_{\text{flat}}$) of the form: $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}),\dots,f_q(\boldsymbol{x}))$, $f_j(\boldsymbol{x}) = \max_{s\in[k]} h_s^j(\boldsymbol{x})$, where each h_s^j is a base learner. The key is that $\mathcal{F}$ is a special subset of $\mathcal{F}_{\text{flat}}$, which enforces that the q predicted labels are produced through shared input representations and independent per-dimension max-pooling. Our upper bound

is on the Rademacher complexity of $\mathcal{F}$, not on the Natarajan dimension of $\mathcal{F}_{\text{flat}}$. A more restricted hypothesis class can indeed have a much smaller complexity measure. Hence, there is no logical contradiction between our upper bound and the lower bound.

We will explain from the perspective of proof techniques why the complexity of $\mathcal{F}$ can be lower than that of $\mathcal{F}_{\text{flat}}$. The Natarajan dimension lower bound applies to any function class mapping $\mathcal{X}$ to k^q labels. However, our $\mathcal{F}$ is not arbitrary, which has a decomposable, coordinate-wise structure.

1. The effective number of parameters in $\mathcal{F}$ is roughly $O(q \cdot d_{\text{base}})$, where d_{base} is the complexity of each h_s^j, rather than $O(k^q)$. Our Rademacher complexity analysis precisely captures this parametric complexity, not the combinatorial complexity.
2. The covering number of $\mathcal{F}_{\text{flat}}$ may grow exponentially in k^q. Through our inner and outer decomposition operators, we show that the covering number of $\mathcal{F}$ under the ℓ_∞ metric can be bounded by that of $\mathcal{T}(\mathcal{H}^j)$ (the class of each h_s^j), which scales roughly as $O(qk \cdot$ (base covering number)). This avoids the k^q term.
3. Our upper bound relies on the ℓ_∞ Lipschitz property of the loss. When we apply Dudley's entropy integral, the Lipschitz property of the loss allows us to transfer the analysis from the function value space to the parameter space, further avoiding the curse of dimensionality.

By restricting the hypothesis class structure and exploiting Lipschitz properties of the loss, we can obtain an upper bound with a logarithmic dependency on q and k. This does not violate the Natarajan dimension lower bound, because the Natarajan dimension lower bound applies to unstructured hypothesis classes.

We will further explain this from an empirical and practical perspective. The naive reduction also implies a method for handling MDC problems, i.e., transforming the MDC problem into a multi-class classification problem with q^k classes and solving it using a multi-class classification method. In practical applications, the reduction to k^q labels is computationally infeasible for large q or k (e.g., $q = 10, k = 100$ gives 100^{10} labels). Hence, it is often considered to decompose the MDC problem into q multi-class classification problems, and to try to explicitly or implicitly model the dependencies between the dimensions to better capture the dimensional dependencies of the MDC problem, thereby better solving the practical problem. For example, in practice, MDC methods (e.g., SLEM (Jia & Zhang, 2021a), PIST (Huang et al., 2024), M^3MDC (Jia & Zhang, 2022)) explicitly model per-dimension features and cross-dimension dependencies, which cannot be captured by a flat multi-class classifier. Our theoretical framework also aligns with practical algorithm design, which is a key contribution of our work.

Our upper bound applies to the decomposable structure hypothesis class $\mathcal{F}$, while the Natarajan lower bound applies to the unstructured hypothesis class $\mathcal{F}_{\text{flat}}$. Since $\mathcal{F} \subset \mathcal{F}_{\text{flat}}$, the former can have significantly lower complexity, and there is no contradiction.

A.7.6 Differences between Reeve & Kaban (2020) and our work

Reeve & Kaban (2020) derives the bound that is at worst $\widetilde{O}(1/\sqrt{n})$ (it may be faster in interpolation type settings) for the multi-output problem with a self-bounding loss and the parameter range of the loss is $\theta \in [0, 0.5]$. The self-bounding loss is an interpolation between Lipschitz continuity and smoothness, with $\theta = 0$ being Lipschitz and $\theta = 0.5$ being smooth. When θ is non-zero, the self-bounding loss essentially exploits the smoothness of the loss. Combined with local Rademacher complexity analysis, a rate faster than $\widetilde{O}(1/\sqrt{n})$ is obtained. The essence of the faster rate lies in inducing the empirical risk in the conversion of covering numbers by introducing the loss itself in self-bounding loss, i.e.,

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\ell(\boldsymbol{f}(\boldsymbol{x}_i), \boldsymbol{y}_i) - \ell(\boldsymbol{f}'(\boldsymbol{x}_i), \boldsymbol{y}_i)\right)^2}$$
$$\leq (\lambda \|\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{f}'(\boldsymbol{x}_i)\|_\infty)(\widehat{R}_D(\boldsymbol{f}) + \widehat{R}_D(\boldsymbol{f}'))^\theta.$$

Since the empirical risk is constrained to a small value (i.e., $< r$), the final local Rademacher complexity includes r. Combining the properties of the sub-root function to solve the fixed point, a faster bound can be obtained. When deriving a faster bound for H-smooth loss, the introduction of

empirical risk in the conversion of covering numbers and the introduction of r in local Rademacher complexity are achieved through the property $(f(t) - f(r))^2 \leq 6H(f(t) + f(r))(t - r)^2$ of smooth loss. The remaining proof of the faster bound follows a similar method. The essence of reducing the dependency of the bound on output dimension is the ℓ_∞ norm rather than the smoothness or the loss itself in self-bounding loss. Here we fully consider Lipschitz continuity, and the MDC problem setting is more complicated than the general multi-output problem. In the processing, we consider reducing the dependency on both q and k, not just q.

There is a minor error in Reeve & Kaban (2020): Theorem 3 is unknown. Reeve & Kaban (2020) is an improvement on Lei et al. (2019). Theorem 3 is derived from Lemma 22 in Lei et al. (2019), i.e., $\log \mathcal{N}_\infty(F) \leq O(fat_\epsilon(F)) \log(Bn/\epsilon)$, the result in Lei et al. (2019) citing Srebro et al. (2010) is a miss-citation, as one cannot find a similar result in Srebro et al. (2010). In fact, whether the result that the lower of the logarithmic term is strictly less than 2 holds is a well-known open problem (Alon et al., 1997; Guermeur, 2018). Hence, the convergence rate of the bound will be worsened by a factor $\sqrt{\ln n}$. The phenomenon is a direct consequence of the appearance of the logarithmic function of ϵ^{-1} in the exponent of the upper bound on the metric entropies. We use the correct result, i.e., Theorem 12.8 in Anthony & Bartlett (2009), to derive bounds for MDC, which scales as $\log \mathcal{N}_\infty(F) \leq O(fat_\epsilon(F)) \log^2(Bn/\epsilon)$. Hence, the proof process, scaling method, and results are different.

Since Reeve & Kaban (2020) aims to obtain the faster bound (the more ideal case in interpolation type settings, i.e., when the empirical error is small), it mainly considers the smooth case, i.e., $\theta \neq 0$. When $\theta = 0$, it is not natural to use the relevant results to achieve the bound of MDC. In addition to the different lemmas involved, the theoretical analysis of how to deal with the multi-classification problem corresponding to each dimension is also technical. This leads to a completely different way of constructing the auxiliary dataset. The outer decomposition operator is only a transitional step to solve the problem through the inner decomposition operator, not the whole part.

In Reeve & Kaban (2020), when $\theta = 0$, this case corresponds to a theoretical result for multi-output problem with Lipschitz loss, which cannot be directly applied to MDC. Only when the classification problem corresponding to each dimension degenerates to a binary classification problem, i.e., MDC degenerates to multi-label classification, can the theoretical results for the $\theta = 0$ case in Reeve & Kaban (2020) be used to analyze the generalization of the degenerate version of MDC.