Data-dependent Algorithmic Robustness Analysis of Pairwise Learning

Donglai Wu u3596228@connect.hku.hk
Runqiu Wu* u3598039@connect.hku.hk
Yunwen Lei leiyw@hku.hk
Department of Mathematics, The University of Hong Kong

Editors: Hung-yi Lee and Tongliang Liu

Abstract

This paper develops a new framework for understanding generalization in pairwise learning problems like metric learning and ranking. By integrating robust optimization principles with pairwise loss structures, we establish data-dependent generalization bounds that significantly improve over existing approaches. Our method overcomes key limitations of prior work by leveraging observable training data properties rather than restrictive theoretical assumptions. This results in tighter performance guarantees that better reflect real-world learning behavior, particularly for complex datasets with dependent training pairs. The framework provides both theoretical advances and practical foundations for more reliable machine learning applications.

Keywords: Pairwise learning, Algorithmic robustness, Generalization analysis

1. Introduction

In recent years, the intersection of robustness and generalization in machine learning has garnered significant attention (Bertsimas et al., 2011; Christmann and Zhou, 2016; Xiao et al., 2022). Robust optimization techniques have proven effective in addressing the challenges posed by noisy and uncertain data, enabling models to perform reliably across varying conditions (Bertsimas et al., 2011; Ju et al., 2022). Building on this foundation, a seminal work (Xu and Mannor, 2012) introduced the concept of algorithmic robustness, and showed its connection to generalization. This connection builds a theoretical foundation to safeguard the learning performance of robustness-inducing methods. Recently, a significant improvement is achieved in Kawaguchi et al. (2022), who provided a data-dependent generalization bound based on algorithmic robustness.

The above robustness analysis is conducted in a pointwise learning setting, where the loss function depends only on a single example, e.g., the classification and regression. However, in practical machine learning problems we often encounter pairwise learning problems with the associated loss functions dependent on a pair of training examples. We refer to such problems as pairwise learning problems. Notable examples of pairwise learning include AUC maximization (Cortes and Mohri, 2004; Gao et al., 2013; Ying et al., 2016; Liu et al., 2018; Yang et al., 2022), metric learning (Cao et al., 2016; Ye et al., 2016), ranking (Clémençon et al., 2008; Agarwal and Niyogi, 2009), and learning with minimum error entropy loss functions (Hu et al., 2015).

* The first two authors contributed equally

The popularity of pairwise learning motivates a lot of theoretical analysis to understand their practical success. Notably, the work (Bellet and Habrard, 2015) extended the robustness analysis to the specific metric learning, and showed that robustness is necessary and sufficient for a metric learning algorithm to generalize. While this work introduces several techniques to handle the challenges associated with metric learning, it suffers from two drawbacks. First, the generalization bound there is not data-dependent and can only imply somewhat crude bounds. Then, the existing analysis does not fully fill the potential of robustness analysis. Second, it only considers the specific metric learning problem, leaving other popular pairwise learning problems untouched.

In this work, we aim to address the above issues by developing data-dependent generalization bounds for general pairwise learning algorithms. By applying the principles of robust optimization to the tailored metrics, we propose a novel approach that not only retains the benefits of improved generalization but also addresses the unique challenges associated with non-IID training pairs, which is a distinct difference between pointwise learning and pairwise learning. Our approach leverages observable properties of training samples, minimizing reliance on strict assumptions about data distributions while ensuring strong performance across various scenarios. Through this integration, we provide a comprehensive analysis of how the robustness principles can enhance pairwise learning, ultimately contributing to a more reliable and effective framework for machine learning applications. This synergy between robust optimization and pairwise learning paves the way for future research and practical implementations that can better handle the complexities of real-world data.

2. Related Work

2.1. Related work on algorithmic robustness

We first discuss the related work on algorithmic robustness and generalization. Xu and Mannor (2012) established a framework that connects robustness to generalization through data-dependent bounds, demonstrating that robust algorithms can significantly enhance performance in various machine learning contexts. Kawaguchi et al. (2022) improved existing generalization bounds in two significant ways: reducing the dependence on the covering number, which measures function approximation, and eliminating reliance on specific hypothesis spaces. Importantly, these enhancements do not require additional assumptions about data distributions; instead, they leverage observable properties of training samples. Algorithmic robustness has found wide applications in SVMs, principle component analysis (Xu and Mannor, 2012) and deep learning (Gouk et al., 2021). A closely related concept is the algorithmic stability, which instead measures the robustness of an algorithm up to a perturbation of training dataset (Bousquet and Elisseeff, 2002), which has found wide applications to study the generalization behavior of stochastic optimization algorithms (Hardt et al., 2016; Nikolakakis et al., 2022; Zhu et al., 2024; Deora et al., 2024; Kuzborskij and Orabona, 2013; Schliserman and Koren, 2022; Lei, 2025; Liu et al., 2017; Fan and Lei, 2024; Lei et al., 2025).

2.2. Related work on generalization of pairwise learning

We now discuss the related work on the generalization analysis pairwise learning. The pioneering work considered the generalization of pairwise learning work in an online learning

setting (Kar et al., 2013; Wang et al., 2012), where the training examples come in a sequential manner. A popular approach to study the performance of pairwise learning is to consider the uniform deviation between training and generalization based on U-statistics and U-process (Clémençon et al., 2008; Cao et al., 2016; Rejchel, 2012; Zhou et al., 2023; Hieu and Ledent, 2025; Hieu et al., 2025). The work (Bellet and Habrard, 2015) applied the algorithmic robustness to metric learning, and show that robustness has close connection to learnability. Recently, there is a growing interest in studying the generalization of pairwise learning via algorithmic stability (Lei et al., 2020; Chen et al., 2023, 2025), which, however, often imposes a convexity assumption. Several work considered more learning settings beyond pairwise learning, where the loss function can depend on more than two training examples (Papa et al., 2015).

3. Preliminaries

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be a sample space, where $\mathcal{X} \subseteq \mathbb{R}^d$ is an input space and $\mathcal{Y} \subseteq \mathbb{R}$ is an output space. Let μ be an unknown probability distribution defined on $\mathcal{Z}$, according to which we independently draw n training examples $\mathbf{s} = (z_1, \ldots, z_n)$. Based on $\mathbf{s}$, we aim to learn a model f in a model space $\mathcal{F}$ for further prediction, where $f : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ or $f : \mathcal{X} \mapsto \mathbb{R}$. In this paper, we consider a class of learning problems where the performance of f depends on a pair of examples, which we refer to pairwise learning. Specifically, let $\ell : \mathcal{F} \times \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}^+$ be a loss function with $\ell(f, z, z')$ quantifying the performance of $f \in \mathcal{F}$ on an example pair (z, z'). The population and empirical risk of f are then defined as

$$\mathcal{L}(f) = \mathbb{E}_{z,z' \sim \mu} \ell(f, z, z') \quad \text{and} \quad \mathcal{L}_{\mathbf{emp}}(f) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \ell(f, z_i, z_j).$$

As a convention of pairwise learning, we set $\ell(f, z, z) = 0$, i.e., we do not suffer a loss if the pair is built from the same example. We now give representative pairwise learning problems.

Example 1 Ranking is a pairwise learning problem which determines the order a set of items based on their relevance or score (Cortes and Mohri, 2004; Qiu et al., 2022). Given a set of items $X = \{x_1, x_2, \ldots, x_n\}$, ranking aims to learn a function $f : \mathcal{X} \to \mathbb{R}$ such that for any pair of items (x_i, x_j), the function satisfies: $f(x_i) > f(x_j)$, if x_i is more relevant than x_j. Common ranking methods include RankNet, RankBoost, and RankSVM.

Example 2 Area Under the Curve (AUC) maximization is a common objective in binary classification tasks for imbalanced labels, i.e., there is a large difference between the number of positive and negative examples. AUC represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. The AUC maximization problem can be formulated as a pairwise learning problem with the objective

$$\text{AUC} = \frac{1}{|P||N|} \sum_{p \in P} \sum_{n \in N} \mathbb{I}(f(p) > f(n)),$$

where P is the set of positive instances, N is the set of negative instances, and $\mathbb{I}$ is the indicator function.

We often apply an optimization algorithm to (approximately) minimize the empirical risk to get a model in $\mathcal{F}$. We denote by $\mathcal{A}_{\mathbf{s}}$ the model learned by applying an algorithm $\mathcal{A}$ to the dataset $\mathbf{s}$. However, a small empirical risk does not necessarily mean that $\mathcal{A}_{\mathbf{s}}$ has a good generalization behavior as there is a bias of a model to the training dataset. We refer to the difference between the population risk and empirical risk as the generalization gap.

4. Algorithmic Robustness and Generalization

4.1. Robustness

In this paper, we are interested in the generalization gap of $\mathcal{A}_{\mathbf{s}}$ to understand how the empirical and population risk differ at the output model $\mathcal{A}_{\mathbf{s}}$, i.e., $\mathcal{L}(\mathcal{A}_{\mathbf{s}}) - \mathcal{L}_{\mathbf{emp}}(\mathcal{A}_{\mathbf{s}})$. We will leverage an important concept called robustness to study the generalization gap. For a finite set $\mathcal{B}$, we let $|\mathcal{B}|$ represent the number of elements in $\mathcal{B}$. For an integer n, we use $[n]$ to denote the set of integers $\{1, \ldots, n\}$.

Definition 1 (Robustness (Bellet and Habrard, 2015)) We say an algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$-robust for $K \in \mathbb{N}$ and $\epsilon : \mathcal{Z}^n \mapsto \mathbb{R}_+$ if $\mathcal{Z}$ can be partitioned into K disjoints sets, denoted by $\{C_k\}_{k=1}^K$, such that for all sample $\mathbf{s} \in \mathcal{Z}^n$, the following holds for any $z_p, z_q \in \mathbf{s}$ and $z, z' \in \mathcal{Z}$ if z_p and z are from the same C_i, and z_q, z' are from the same C_j for some $i, j \in [n]$

$$\left|\ell\left(\mathcal{A}_{\mathbf{s}}, z_p, z_q\right) - \ell\left(\mathcal{A}_{\mathbf{s}}, z, z'\right)\right| \leq \epsilon(\mathbf{s}). \tag{4.1}$$

For a given partition $\{C_k\}_{k=1}^K$, we denote $p_k = \mathbb{P}\left(z \in C_k\right)$. Let $p = (p_1, \ldots, p_K)$. Define

$$\mathcal{T}_{\mathbf{s}} := \{k \in [K] : |\mathcal{I}_k^{\mathbf{s}}| \geq 1\} \quad \text{with} \quad \mathcal{I}_k^{\mathbf{s}} := \{i \in [n] : z_i \in C_k\}, \tag{4.2}$$

i.e., $\mathcal{T}_{\mathbf{s}}$ contains the indices of the sets C_k which are not empty. Below, we present our main result on connecting the robustness and generalization, which is data-dependent since $\xi(A(\mathbf{s})), \epsilon(\mathbf{s})$ and $\mathcal{T}_{\mathbf{s}}$ depend on $\mathbf{s}$.

Theorem 2 (Main result) *If a pairwise learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$-robust (with $\{C_k\}_{k=1}^K$), then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have*

$$\left|\mathcal{L}\left(\mathcal{A}_{\mathbf{s}}\right) - \mathcal{L}_{emp}\left(\mathcal{A}_{\mathbf{s}}\right)\right| \leq \epsilon(\mathbf{s}) + \xi(\mathcal{A}_{\mathbf{s}})\left((2\sqrt{2}+2)\sqrt{\frac{|\mathcal{T}_{\mathbf{s}}|\ln(2K/\delta)}{n}} + \frac{2|\mathcal{T}_{\mathbf{s}}|\ln(2K/\delta)}{n}\right),$$

where

$$\xi\left(\mathcal{A}_{\mathbf{s}}\right) := \max_{i,j \in [K]} \mathbb{E}_{z,z' \sim \mu}\left[\ell\left(\mathcal{A}_{\mathbf{s}}, z, z'\right) \mid z \in \mathcal{C}_i,\ z' \in \mathcal{C}_j\right]. \tag{4.3}$$

We now compare our result with the existing robustness of pairwise learning. Specifically, the following proposition was derived for pairwise learning.

Proposition 3 (Bellet and Habrard 2015) *If a learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$-robust, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$*

$$\left|\mathcal{L}\left(\mathcal{A}_{\mathbf{s}}\right) - \mathcal{L}_{emp}\left(\mathcal{A}_{\mathbf{s}}\right)\right| \leq \epsilon\left(\mathbf{s}\right) + 2B\sqrt{\frac{2K\ln 2 + 2\ln 1/\delta}{n}},$$

where $B = \sup_{f,z,z'} \ell(f, z, z')$ is an upper bound of the loss.

Remark 4 (Comparison) Our robustness analysis outperforms the existing results in the following two aspects.

First, Theorem 2 gives a refined dependence on the upper bound of the loss value. Note that the generalization bounds in Proposition 3 depend on an uniform upper bound of the loss function over all $f \in \mathcal{F}, z, z' \in \mathcal{Z}$. As a comparison, $\xi(\mathcal{A}_{\mathbf{s}})$ in our bound depends only on the single actual hypothesis, $\mathcal{A}_{\mathbf{s}}$, returned by the specific algorithm applied to $\mathbf{s}$, which can be significantly smaller than B. Furthermore, $\xi(\mathcal{A}_{\mathbf{s}})$ involves a conditional expectation instead of a supremum.

Second, Proposition 3 depends on K, which is the cardinality of the partition and can be extremely large (Kawaguchi et al., 2022). Theorem 2 replaces K by $|\mathcal{T}_{\mathbf{s}}|$ (the dependency of our bound on K is logarithmic, which can be ignored), which is the number of sets in the partition containing at least a single training example. As verified in Kawaguchi et al. (2022), $|\mathcal{T}_{\mathbf{s}}|$ can be of smaller orders of magnitude as compared to K. Intuitively, $|\mathcal{T}_{\mathbf{s}}|$ is likely to be significantly less than K when there are many sparsely populated classes C_k.it is improbable that many of these classes are represented in the sample data.

With a refined analysis, we also prove a stronger (yet more complicated) version of Theorem 2. For any $k \in [K]$, define $\alpha_k : \mathcal{F} \mapsto \mathbb{R}$ by

$$\alpha_k(f) := \max_{j\in[n]} \mathbb{E}_{z,z'\sim\mu}[\ell(f,z,z')|z \in C_k, z' \in C_j]. \tag{4.4}$$

Define

$$\alpha_{\mathcal{T}_{\mathbf{s}}}(f) = \max_{k\in\mathcal{T}_{\mathbf{s}}} \alpha_k(f) \quad \text{and} \quad \alpha_{\mathcal{T}_{\mathbf{s}}^c}(f) = \max_{k\in\mathcal{T}_{\mathbf{s}}^c} \alpha_k(f),$$

which are smaller than $\xi(\mathcal{A}_{\mathbf{s}})$.

Theorem 5 *If the learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot))$-robust (with $\{C_k\}_{k=1}^K$), then for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\left|\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')] - \frac{1}{n^2}\sum_{i,j\in[n]} \ell(\mathcal{A}_{\mathbf{s}},z_i,z_j)\right| \le \epsilon(\mathbf{s}) + \mathcal{Q}_1\sqrt{\frac{\ln(2K/\delta)}{n}} + \frac{2\mathcal{Q}_2\ln(2K/\delta)}{n},$$

$$\text{where} \quad \mathcal{Q}_1 := \sum_{k\in\mathcal{T}_{\mathbf{s}}} (\alpha_{\mathcal{T}_{\mathbf{s}}^c}(\mathcal{A}_{\mathbf{s}}) + \sqrt{2}\alpha_k(\mathcal{A}_{\mathbf{s}}))\sqrt{\frac{|\mathcal{I}_k^{\mathbf{s}}|}{n}}, \quad \mathcal{Q}_2 := \alpha_{\mathcal{T}_{\mathbf{s}}^c}(\mathcal{A}_{\mathbf{s}})|\mathcal{T}_{\mathbf{s}}| + \sum_{k\in\mathcal{T}_{\mathbf{s}}} \alpha_k(\mathcal{A}_{\mathbf{s}}). \tag{4.5}$$

Remark 6 Theorem 5 is a stronger form of Theorem 2. This can be verified as follows.

First, since $\sum_{k\in\mathcal{T}_{\mathbf{s}}} \alpha_k(\mathcal{A}_{\mathbf{s}}) \le |\mathcal{T}_{\mathbf{s}}|\,\xi(\mathcal{A}_{\mathbf{s}})$ and $\sum_{k\in\mathcal{T}_s} \sqrt{|\mathcal{I}_k^{\mathbf{s}}|/n} \le \sqrt{|\mathcal{T}_{\mathbf{s}}|}$, Theorem 5 significantly upgrades Theorem 2 approximately when

$$\sum_{k\in\mathcal{T}_s} \alpha_k(\mathcal{A}_{\mathbf{s}}) \ll |\mathcal{T}_{\mathbf{s}}|\,\xi(\mathcal{A}_{\mathbf{s}}) \text{ or } \sum_{k\in\mathcal{T}_s} \sqrt{|\mathcal{I}_k^{\mathbf{s}}|/n} \ll \sqrt{|\mathcal{T}_{\mathbf{s}}|}.$$

Second, the bound in Theorem 5 depends on $\alpha_k(\mathcal{A}_{\mathbf{s}})$ which provides a more delicate estimate as compared to Theorem 2 dependent on $\xi(\mathcal{A}_{\mathbf{s}})$. Indeed, if the maximum expected loss of the classes is significantly greater than the typical expected loss, or if the distribution of samples among the classes is skewed, then Theorem 5 will provide an even tighter bound.

4.2. Pseudo-Robustness

In the previous section, we required that every training pair satisfy the robustness property. Knowing that this is a strong condition, we explored whether we could relax this condition so that it only needs to hold for a subset of the training pairs, while still ensuring generalization guarantees. Here we use the definition of Pseudo-robustness proposed by Bellet and Habrard (2015). For brevity, we write $\mathbf{s} = (z_1, \ldots, z_n)$,and $\mathbf{s}^2 = \{(z_1, z_1), \ldots, (z_1, z_n), \ldots, (z_n, z_n)\}$.

Definition 7 (Pseudo-robustness) Algorithm $\mathcal{A}$ is $(K,\ \epsilon(\cdot),\ \hat{p}_n(\cdot))$ pseudo-robust for $K \in \mathbb{N}$, $\epsilon(\cdot) : (\mathcal{Z} \times \mathcal{Z})^n \mapsto \mathbb{R}$ and $\hat{p}_n(\cdot) : (\mathcal{Z} \times \mathcal{Z})^n \mapsto \left\{1, \ldots, n^2\right\}$, if $\mathcal{Z}$ can be partitioned into K disjoints sets, denoted by $\{C_i\}_{i=1}^{K}$, such that for all $\mathbf{s} \in \mathcal{Z}^n$, there exists a subset of training pair samples $\hat{\mathbf{s}}^2 \subseteq \mathbf{s}^2$, with $\left|\hat{\mathbf{s}}^2\right| = \hat{p}_n\left(\mathbf{s}^2\right)$, satisfying: $\forall\, (z_p, z_q) \in \hat{\mathbf{s}}^2, \quad \forall z, z' \in \mathcal{Z}, \quad \forall i, j = 1, \ldots, K$: if $z_p, z \in C_i$ and $z_q, z' \in C_j$ then

$$\left|\ell\left(\mathcal{A}_{\mathbf{s}}, z_p, z_q\right) - \ell\left(\mathcal{A}_{\mathbf{s}}, z, z'\right)\right| \leq \epsilon\left(\mathbf{s}^2\right). \tag{4.6}$$

Proposition 8 *If a learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot), \hat{p}_n(\cdot))$ pseudo-robust, the training pairs come from a sample generated by n IID draws from μ, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\left|\mathcal{L}\left(\mathcal{A}_{\mathbf{s}}\right) - \mathcal{L}_{emp}\left(\mathcal{A}_{\mathbf{s}}\right)\right| \leq \frac{\hat{p}_n\left(\mathbf{s}^2\right)}{n^2}\epsilon\left(\mathbf{s}\right) + B\left(\frac{n^2 - \hat{p}_n\left(\mathbf{s}^2\right)}{n^2} + 2\sqrt{\frac{2K\ln 2 + 2\ln 1/\delta}{n}}\right). \tag{4.7}$$

The following theorem gives a generalization guarantee related to pseudo-robustness which is improved from Proposition 8.

Theorem 9 (Analog of Theorem 2) *If a learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot), \hat{p}_n(\cdot))$ pseudo-robust, the training pairs come from a sample generated by n IID draws from μ, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have:*

$$\left|\mathcal{L}\left(\mathcal{A}_{\mathbf{s}}\right) - \mathcal{L}_{emp}\left(\mathcal{A}_{\mathbf{s}}\right)\right| \leq \frac{\hat{p}_n\left(\mathbf{s}^2\right)}{n^2}\epsilon(\mathbf{s}) + \xi\left(\mathcal{A}_{\mathbf{s}}\right)\left(\frac{n^2 - \hat{p}_n(\mathbf{s}^2)}{n^2} + \frac{5}{2}\sqrt{\frac{\left|\mathcal{T}_{\mathbf{s}}\right|\ln(2K/\delta)}{n}} + \frac{2\left|\mathcal{T}_{\mathbf{s}}\right|\ln(2K/\delta)}{n}\right),$$

where $\xi\left(\mathcal{A}_{\mathbf{s}}\right)$ is defined in Eq. (4.3) and $\mathcal{T}_{\mathbf{s}}$ is defined in Eq. (4.2).

By a refined analysis, we also prove a stronger (yet more complicated) version of Theorem 9.

Theorem 10 (Analog of Theorem 5) *If a learning algorithm $\mathcal{A}$ is $(K, \epsilon(\cdot), \hat{p}_n(\cdot))$ pseudo-robust, the training pairs come from a sample generated by n IID draws from μ, then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have:*

$$\left|\mathcal{L}\left(\mathcal{A}_{\mathbf{s}}\right) - \mathcal{L}_{emp}\left(\mathcal{A}_{\mathbf{s}}\right)\right| \leq \frac{\hat{p}_n(\mathbf{s}^2)}{n^2}\epsilon(\mathbf{s}) + \xi\left(\mathcal{A}_{\mathbf{s}}\right)\left(\frac{n^2 - \hat{p}_n(\mathbf{s}^2)}{n^2}\right) + \mathcal{Q}_1\sqrt{\frac{\ln(2K/\delta)}{n}} + \frac{2\mathcal{Q}_2\ln(2K/\delta)}{n},$$

where $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are defined in Eq. (4.5).

These theorems provide corresponding enhancements to the pseudo-robustness bounds established by Bellet and Habrard (2015).

Remark 11 In practice, we found that optimizing the metric for all possible pairs will be challenging and almost impossible. However, concentrating on some selected pairs may lead to significantly better generalization performance. This means that some properties based on pseudo-robustness are more important for metric learning.

5. Applications to Metric Learning

5.1. Robust Metric learning Algorithms

While our Theorems 2 and 5 are applicable to a wide range of applications, we provide a few simple examples that demonstrate how various metric learning algorithms inherently satisfy the robustness conditions of Bellet and Habrard (2015) to which Theorems 2 and 5 can be applied. The core idea utilizes partition-based robustness at the pair level: when a test pair (z_1', z_2') is "close" to a training pair (z_p, z_q) in the sense that z_p and z_1' fall within the same partition subset $C_i \subseteq \mathcal{Z}$, while z_q and z_2' fall within the same partition subset $C_j \subseteq \mathcal{Z}$, then their loss values must satisfy $|\ell(\mathcal{A}_\mathbf{s}, z_p, z_q) - \ell(\mathcal{A}_\mathbf{s}, z_1', z_2')| \leq \epsilon(\mathbf{s})$.

Definition 12 (Covering number) For a set $\mathbf{s}$ equipped with metric ρ, we define $\hat{\mathbf{s}}$ as an ε-cover of $\mathbf{s}$ if for all $s \in \mathbf{s}$, there exists $\hat{s} \in \hat{\mathbf{s}}$ such that $\rho(s, \hat{s}) \leq \varepsilon$. We then define the ε-covering number as

$$\mathcal{N}(\varepsilon, \mathbf{s}, \rho) = \min\{|\hat{\mathbf{s}}| : \hat{\mathbf{s}} \text{ is an } \varepsilon\text{-cover of } \mathbf{s}\}.$$

We use $\|\cdot\|_p$ to denote the standard p-norm for a vector.

The following lemma presents sufficient conditions for an algorithm to be robust.

Lemma 13 (Bellet and Habrard 2015) *Fix $\gamma > 0$ and a metric ρ of $\mathcal{Z}$. Assume*

$$|\ell(\mathcal{A}_\mathbf{s}, z_1, z_2) - \ell(\mathcal{A}_\mathbf{s}, z_1', z_2')| \leq \epsilon(\mathbf{s}),\ \forall z_1, z_2, z_1', z_2' : z_1, z_2 \in \mathbf{s}, \quad \rho(z_1, z_1') \leq \gamma, \quad \rho(z_2, z_2') \leq \gamma,$$

and $\mathcal{N}(\gamma/2, \mathcal{Z}, \rho) < \infty$. Then $\mathcal{A}$ is $(\mathcal{N}(\gamma/2, \mathcal{Z}, \rho), \epsilon(\mathbf{s}))$-robust.

The examples below demonstrate how different regularization norms induce distinct robustness parameters:

Consider Mahalanobis distance learning algorithms with matrix norm regularization

$$\min_{\mathbf{M} \succeq 0}\ c\|\mathbf{M}\| + \frac{1}{n^2} \sum_{(s_i, s_j) \in \mathcal{P}_s} g(y_{ij}[1 - f(\mathbf{M}, \mathbf{x}_i, \mathbf{x}_j)]), \tag{5.1}$$

where $y_{ij} = 1$ if $y_i = y_j$ and -1 otherwise, $f(\mathbf{M}, \mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)$ is the Mahalanobis distance parameterized by the $d \times d$ PSD matrix $\mathbf{M}$, $\|\cdot\|$ some matrix norm and c a regularization parameter. We consider a loss function $l(f, s_i, s_j) = g(y_{ij}[1 - f(\mathbf{M}, \mathbf{x}_i, \mathbf{x}_j)])$, which outputs a small value when its input is large positive and a large value when it is large negative. We assume g to be nonnegative and Lipschitz continuous with Lipschitz constant U. Lastly, $g_0 = \sup_{s_i, s_j} g(y_{ij}[1 - f(0, \mathbf{x}_i, \mathbf{x}_j)])$ is the largest loss when $\mathbf{M} = 0$. We assume that $\forall \mathbf{x} \in \mathbf{X}$, $\|\mathbf{x}\| \leq R$. Now we specify the norms in Eq. (5.1) to get specific robust algorithms (Bellet and Habrard, 2015).

Example 3 (Frobenius norm) Consider Eq. (5.1) with the Frobenius norm $\|\mathbf{M}\| = \|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^d \sum_{j=1}^d m_{ij}^2}$ is $(|\mathcal{Y}|\mathcal{N}(\gamma/2, \mathbf{X}, \|\cdot\|_2), 8URc\gamma g_0/c)$-robust.

Example 4 (ℓ_1 norm) Algorithm defined in Eq. (5.1) with $\|\mathbf{M}\| = \|\mathbf{M}\|_1$ is $(|\mathcal{Y}|\mathcal{N}(\gamma, \mathbf{X}, \|\cdot\|_1), 8URc\gamma g_0/c)$-robust.

Example 5 ($\ell_{2,1}$ norm and trace norm) Algorithm defined in Eq. (5.1) with $\|\mathbf{M}\| = \|\mathbf{M}\|_{2,1} = \sum_{i=1}^{d} \|\mathbf{m}^i\|_2$, where $\mathbf{m}^i$ is the i-th column of $\mathbf{M}$. This algorithm is $(|\mathcal{Y}|\mathcal{N}(\gamma, \mathbf{X}, \|\cdot\|_2), 8UR\gamma g_0/c)$-robust. The same holds for the trace norm $\|\mathbf{M}\|_*$, which is the sum of the singular values of $\mathbf{M}$.

Example 6 (Kernelization) Let $\mathcal{X}$ be compact. Consider the kernelized version of Eq. (5.1):

$$\min_{\mathbf{M}\succeq 0} c\|\mathbf{M}\|_{\mathcal{H}} + \frac{1}{n^2} \sum_{(s_i,s_j)\in\mathbf{s}^2} g(y_{ij}[1 - f(\mathbf{M}, \phi(x_i), \phi(x_j))]). \tag{5.2}$$

Here $\phi(\cdot)$ is a feature mapping to a kernel space $\mathcal{H}$, $\|\cdot\|_{\mathcal{H}}$ is the norm function of $\mathcal{H}$, and $k(\cdot,\cdot)$ is the kernel function. We assume $\mathcal{H}$ to be a Hilbert space, equipped with an inner product operator $\langle\cdot,\cdot\rangle$. A feature mapping $\phi(\cdot)$ is a continuous mapping from $\mathcal{X}$ to $\mathcal{H}$. The norm $\|\mathbf{w}\|_{\mathcal{H}} : \mathcal{H} \mapsto \mathbb{R}$ is defined as $\|\mathbf{w}\|_{\mathcal{H}} = \langle \mathbf{w}, \mathbf{w}\rangle$, for all $\mathbf{w} \in \mathcal{H}$. The kernel function $k : \mathcal{X}\times\mathcal{X} \mapsto \mathbb{R}$ is defined as $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2)\rangle$. Define $f_{\mathcal{H}}(\gamma) \triangleq \max_{a,b\in\mathcal{X}, \|a-b\|_2\leq\gamma}(k(a,a) + k(b,b) - 2k(a,b))$ and $B_y = \max_{x\in X}\sqrt{k(x,x)}$. If $k(\cdot,\cdot)$ is continuous, by the compactness of $\mathcal{X}$ we consider a cover of $\mathcal{X}$, then for any $\gamma > 0$, B_y and $f_{\mathcal{H}}$ are finite for any $\gamma > 0$, and algorithm (5.2) is $(|\mathcal{N}(\gamma, X, \|\cdot\|_2)|, 8UB_y\sqrt{f_{\mathcal{H}}g_0/c})$-robust.

Now we derive bounds for Bilinear Similarity Metric Learning.

Example 7 Replace Mahalanobis distance defined in Eq. (5.1) by bilinear similarity $x_i^T\mathbf{M}x_j$:

$$\min_{\mathbf{M}\succeq 0} \; c\|\mathbf{M}\| + \frac{1}{n^2} \sum_{(s_i,s_j)\in\mathcal{P}_s} g(y_{ij}[1 - x_i^T\mathbf{M}x_j]) \tag{5.3}$$

For the regularizers considered in Examples 3, 4, 5, robustness can be improved to $2UR_y g_0/c$.

The improved robustness constant $2UR_y g_0/c$ in bilinear similarity learning ((5.3)) stems from fundamental differences in how perturbations affect the similarity measure compared to Mahalanobis distance. The Mahalanobis distance $f(\mathbf{M}, \mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)$ amplifies perturbation effects via difference operations, while the bilinear similarity $x_i^T\mathbf{M}x_j$ exhibits simpler perturbation behavior. Specifically, for perturbations $\|\delta_i\|, \|\delta_j\| \leq \gamma$:

$$\left|x_i^T\mathbf{M}x_j - (x_i + \delta_i)^T\mathbf{M}(x_j + \delta_j)\right| \leq \|\mathbf{M}\|(\|x_i\|\|\delta_j\| + \|\delta_i\|\|x_j\| + \|\delta_i\|\|\delta_j\|)$$

yielding a tighter Lipschitz constant of $2R_y\|\mathbf{M}\|$ (vs. $8R\|\mathbf{M}\|$ for Mahalanobis). This directly translates to the improved robustness constant $2UR_y g_0/c$ under identical regularization frameworks. The constant R_y here specifically denotes $\max_i \|x_i\|$, maintaining consistency with input norm assumptions.

5.2. Theoretical Comparisons

Here, we further demonstrate that when the data are embedded with high probability on a low-dimensional manifold in the data space, and our bound is much stronger than that of Bellet and Habrard (2015). The following proposition demonstrates that $|\mathcal{T}_{\mathbf{s}}|$ is indeed independent of K and only scales logarithmically with n under a mild condition on p_k, proving that we have $|\mathcal{T}_{\mathbf{s}}| \ll K$ and $|\mathcal{T}_{\mathbf{s}}| \ll n$ in a general case.

Proposition 14 (Kawaguchi et al. 2022) *Under the assumptions of Theorem 2, we denote* $p_k = \mathbb{P}(z \in C_k)$ *where* $p_1 \geq p_2 \geq \cdots \geq p_K$. *If there are some constants* $\alpha, \beta, C > 0$ *such that* p_k *decays as* $p_k \leq Ce^{-(k/\beta)^\alpha}$, *and* $\ln n \geq \max\{1, 1/\alpha\}$ *then with probability at least* $1-\delta$, *the following holds*

$$|\mathcal{T}_{\mathbf{s}}| \leq \begin{cases} \beta(\ln n)^{\frac{1}{\alpha}} + C(e-1)\frac{\beta}{\alpha} + \log(1/\delta), & \text{if } \alpha \geq 1, \\ (1 + 2C(e-1))\beta(\ln n)^{\frac{1}{\alpha}} + \log(1/\delta), & \text{if } \alpha < 1. \end{cases}$$

In Proposition 14, the parameter α controls how rapidly p_k decays. For real-world datasets, the data distribution often concentrates on a lower-dimensional manifold or around a small number of modes. In such settings, it is expected that probability p_k (arranged in decreasing order) exhibits fast decays. If $\alpha = \infty$, p_k concentrates on unknown β bins, therefore we have $|\mathcal{T}_{\mathbf{s}}| \leq \beta$. If $\alpha < \infty$, we have $p_k \neq 0$ for all $k \in [K]$, but $|\mathcal{T}_{\mathbf{s}}|$ remains bounded above up to a logarithmic factor and does not depend on K.

Proposition 14 also highlights that even with complete prior knowledge of the data distribution, $|\mathcal{T}_{\mathbf{s}}|$ can be much smaller than K as $|\mathcal{T}_{\mathbf{s}}|$ is more adaptive according to the training data while K cells need to cover all regions with positive mass in the distribution. Without the perfect knowledge, $|\mathcal{T}_{\mathbf{s}}|$ can be more significantly smaller than K. A crucial aspect of Theorem 2 is that $\mathcal{T}_{\mathbf{s}}$ is determined solely by the training sample data instead of the true background distribution. As a result, this outcome is particularly valuable in statistical learning scenarios, where knowledge of the true distribution is limited to what can be inferred from the training data.

6. Experimental Verification

In this section, we present experimental results to verify the superiority of our bound compared to Bellet and Habrard (2015). We used five real-world datasets from LIBSVM (Chang and Lin, 2011), whose details are provided in the following table

Dataset	cod-rna	poker	svmguide	w8a	ijcnn1
n	59535	25010	7089	49749	49990
d	8	10	4	10	10

We normalized these datsets so that $\mathcal{X} \subset [0,1]^d$. Following the suggestions in the literature (Bellet and Habrard, 2015; Kawaguchi et al., 2022), we used an ϵ-cover (with respect to the infinity norm) of the original input space to choose the partitions and set $\epsilon = 0.1$. We conducted an empirical comparison between Theorem 5 (our result) and Proposition 3 for pairwise learning. Specifically, Proposition 3 shows that $|\mathcal{L}(\mathcal{A}_{\mathbf{s}}) - \mathcal{L}_{\mathbf{emp}}(\mathcal{A}_{\mathbf{s}}) - \epsilon(\mathbf{s})| \leq G_1 := 2B\big((2K\ln 2 + 2\ln(1/\delta))/n\big)^{\frac{1}{2}}$, while Theorem 5 shows that $|\mathcal{L}(\mathcal{A}_{\mathbf{s}}) - \mathcal{L}_{\mathbf{emp}}(\mathcal{A}_{\mathbf{s}}) - \epsilon(\mathbf{s})| \leq G_2 := \mathcal{Q}_1\big(\ln(2K/\delta)/n\big)^{\frac{1}{2}} + 2\mathcal{Q}_2\ln(2K/\delta)/n$, where $\mathcal{Q}_1, \mathcal{Q}_2$ are defined in Eq. (4.5). Since $\epsilon(\mathbf{s})$ are the same for both Proposition 3 and Theorem 5, we only compare G_1 versus G_2 in the following table to illustrate our generalization bound relative to the existing bound in Bellet and Habrard (2015). We set $\delta = 0.1$, and ignore the difference between B and α_k and $\alpha_k(\mathcal{A}_{\mathbf{s}})$. This constitutes a favorable setting for the bound in Bellet and Habrard (2015) since by definition we know $B \geq \alpha_k$ and $B \geq \alpha_k(\mathcal{A}_{\mathbf{s}})$. The comparisons are summarized

in Table 1. The experimental results in Table 1 demonstrate that our data-dependent generalization bounds are significantly tighter than the existing bounds in Bellet and Habrard (2015). The ratio in the improvement can be as large as $1053/0.06 \approx 17,550$.

Dataset	cod-rna	poker	svmguide	w8a	ijcnn1
G_1	96.5	1489	2.8	1055.8	1053
G_2	0.14	0.36	0.47	0.10	0.06

Table 1: Comparison between G_1 and G_2

7. Proof

In this section we will give the proof of the main theorem. The main theorem is an improvement based on Proposition 3. This improvement is largely based on the two upgraded bounds for the multinomial distribution proposed by Kawaguchi et al. (2022). Recall that $a_{\mathcal{T}_\mathbf{s}}(X) := \max_{i\in\mathcal{T}_\mathbf{s}} a_i(X)$ and $a_{\mathcal{T}_\mathbf{s}^c}(X) := \max_{i\in\mathcal{T}_\mathbf{s}^c} a_i(X)$ where $T_\mathbf{s}^c = [K] - T_\mathbf{s}$ (Definition 1).

Proposition 15 (Kawaguchi et al. 2022) *Let the vector $X = (X_1, \ldots, X_K)$ follow the multinomial distribution with parameter n and $p = (p_1, \ldots, p_K)$, where $p_k = \mathbb{P}(z \in C_k)$ and $X_k = \sum_{i=1}^n \mathbb{1}\{z_i \in C_k\}$. For any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\sum_{i=1}^K a_i(X)(p_i - \frac{X_i}{n}) \leq \left((\sqrt{2}a_{\mathcal{T}_\mathbf{s}}(X) + a_{\mathcal{T}_\mathbf{s}^c}(X))\sqrt{\frac{|\mathcal{T}_\mathbf{s}|\ln(2K/\delta)}{n}} + a_{\mathcal{T}_\mathbf{s}^c}(X)\frac{2|\mathcal{T}_\mathbf{s}|\ln(2K/\delta)}{n}\right),$$

where a_i is an arbitrary function with $a_i(X) \geq 0$ for all $i \in \{1, \ldots, K\}$.

Proposition 16 (Kawaguchi et al. 2022) *For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds:*

$$\sum_{i=1}^K a_i(X)(p_i - \frac{X_i}{n}) \leq \left(\frac{\ln(2K/\delta)}{n}\right)^{\frac{1}{2}}\left(\sum_{i\in\mathcal{T}_\mathbf{s}}\left(a_{\mathcal{T}_\mathbf{s}^c}(X) + \sqrt{2}a_i(X)\right)\sqrt{\frac{X_i}{n}}\right) + \frac{2\ln(2K/\delta)(a_{\mathcal{T}_\mathbf{s}^c}(X)|\mathcal{T}_\mathbf{s}| + \sum_{i\in\mathcal{T}_\mathbf{s}} a_i(X))}{n}.$$

7.1. Proof of Theorem 2

We start with the proof of the following lemma that relates the gap to the concentration of the multinomial distributions. We use the abbreviation $\mathcal{I}_i := \mathcal{I}_i(\mathbf{s})$ in Definition 1. The proofs of Lemma 17 and Lemma 18 are given in the appendix.

Lemma 17 *For any $\mathbf{s} \in \mathcal{Z}^n$, we have*

$$|\mathcal{L}(\mathcal{A}_\mathbf{s}) - \mathcal{L}_{emp}(\mathcal{A}_\mathbf{s})| \leq \sum_{i,j\in[K]} \mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_\mathbf{s}, z, z')|z \in C_i, z' \in C_j](p_j + \frac{|\mathcal{I}_j|}{n})\left|p_i - \frac{|\mathcal{I}_i|}{n}\right|$$
$$+ \left|\sum_{i,j\in[K]} \mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_\mathbf{s}, z, z')|z \in C_i, z' \in C_j]\frac{|\mathcal{I}_i||\mathcal{I}_j|}{n^2} - \frac{1}{n^2}\sum_{i,j=1}^n \ell(\mathcal{A}_\mathbf{s}, z_i, z_j)\right|.$$

The second term in Lemma 17 is bounded by the following lemma.

Lemma 18 *For any* $\mathbf{s} \in \mathcal{Z}^n$, *we have*

$$\Big| \sum_{i,j\in[K]} \mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_\mathbf{s}, z, z')|z \in C_i, z' \in C_j]\frac{|\mathcal{I}_i||\mathcal{I}_j|}{n^2} - \frac{1}{n^2}\sum_{i,j=1}^{n} \ell(\mathcal{A}_\mathbf{s}, z_i, z_j)\Big| \leq \max_{i,j\in[n]} \max_{z_p,z\in C_i,z_q,z'\in C_j} \big|\ell(\mathcal{A}_\mathbf{s}, z, z') - \ell(\mathcal{A}_\mathbf{s}, z_p, z_q)\big|.$$

Combining these lemmas and the concentration bounds from Proposition 15, we can finish the proof of Theorem 2 as follows.

Proof [Proof of Theorem 2] From Lemma 17, we get

$$\Big|\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_\mathbf{s}, z, z')] - \frac{1}{n^2}\sum_{i,j\in[n]} \ell(\mathcal{A}_\mathbf{s}, z_i, z_j)\Big| \leq \sum_{i,j\in[K]} \mathbb{E}_{z,z'}[\ell(\mathcal{A}_\mathbf{s}, z, z')|z \in C_i, z' \in C_j](p_j + \frac{|\mathcal{I}_j|}{n})\Big|p_i - \frac{|\mathcal{I}_i|}{n}\Big|$$
$$+ \Big| \sum_{i,j\in[K]} \mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_\mathbf{s}, z, z')|z \in C_i, z' \in C_j]\frac{|\mathcal{I}_i||\mathcal{I}_j|}{n^2} - \frac{1}{n^2}\sum_{i,j=1}^{n} \ell(\mathcal{A}_\mathbf{s}, z_i, z_j)\Big|.$$

By introducing $\xi(\mathcal{A}_\mathbf{s})$, we further get

$$|\mathcal{L}(\mathcal{A}_\mathbf{s}) - \mathcal{L}_\mathbf{emp}(\mathcal{A}_\mathbf{s})| \leq \xi(\mathcal{A}_\mathbf{s})\sum_{j=1}^{K}(p_j + \frac{|\mathcal{I}_j|}{n})\sum_{i=1}^{K}\big|(p_i - \frac{|\mathcal{I}_i|}{n})\big|$$
$$+ \Big| \sum_{i,j\in[K]} \mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_\mathbf{s}, z, z')|z \in C_i, z' \in C_j]\frac{|\mathcal{I}_i||\mathcal{I}_j|}{n^2} - \frac{1}{n^2}\sum_{i,j=1}^{n} \ell(\mathcal{A}_\mathbf{s}, z_i, z_j)\Big|.$$

By the definition of p_i and $\mathcal{I}_i$, we have $\sum_{i=1}^{K}(p_i + \frac{|\mathcal{I}_i|}{n}) = 2$ and

$$|\mathcal{L}(\mathcal{A}_\mathbf{s}) - \mathcal{L}_\mathbf{emp}(\mathcal{A}_\mathbf{s})| \leq 2\xi(\mathcal{A}_\mathbf{s})\sum_{i=1}^{K}|p_i - \frac{|\mathcal{I}_i|}{n}|$$
$$+ \Big| \sum_{i,j\in[K]} \mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_\mathbf{s}, z, z')|z \in C_i, z' \in C_j]\frac{|\mathcal{I}_i||\mathcal{I}_j|}{n^2} - \frac{1}{n^2}\sum_{i,j=1}^{n} \ell(\mathcal{A}_\mathbf{s}, z_i, z_j)\Big|.$$

Applying Proposition 15 with $a_k(X) = 1$, we get

$$|\mathcal{L}(\mathcal{A}_\mathbf{s}) - \mathcal{L}_\mathbf{emp}(\mathcal{A}_\mathbf{s})| \leq \xi(\mathcal{A}_\mathbf{s})\left((2\sqrt{2}+2)\sqrt{\frac{|\mathcal{T}_\mathbf{s}|\ln(2K/\delta)}{n}} + \frac{2|\mathcal{T}_\mathbf{s}|\ln(2K/\delta)}{n}\right)$$
$$+ \frac{1}{n^2}\Big| \sum_{i,j\in[K]} \mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_\mathbf{s}, z, z')|z \in C_i, z' \in C_j]|\mathcal{I}_i||\mathcal{I}_j| - \sum_{i,j=1}^{n} \ell(\mathcal{A}_\mathbf{s}, z_i, z_j)\Big|.$$

It then follows from Lemma 18 that

$$|\mathcal{L}(\mathcal{A}_\mathbf{s}) - \mathcal{L}_\mathbf{emp}(\mathcal{A}_\mathbf{s})| \leq \xi(\mathcal{A}_\mathbf{s})\left((2\sqrt{2}+2)\sqrt{\frac{|\mathcal{T}_\mathbf{s}|\ln(2K/\delta)}{n}} + \frac{2|\mathcal{T}_\mathbf{s}|\ln(2K/\delta)}{n}\right)$$
$$+ \max_{i,j\in[n]} \max_{z_p,z\in C_i,z_q,z'\in C_j} \big|\ell(\mathcal{A}_\mathbf{s}, z, z') - \ell(\mathcal{A}_\mathbf{s}, z_p, z_q)\big|.$$

Using the robustness hypothesis that (A is $(K, \epsilon(\cdot))$-robust), we know

$$|\mathcal{L}\left(\mathcal{A}_{\mathbf{s}}\right) - \mathcal{L}_{\mathbf{emp}}\left(\mathcal{A}_{\mathbf{s}}\right)| \leq \xi(\mathcal{A}_{\mathbf{s}})\left((2\sqrt{2}+2)\sqrt{\frac{|\mathcal{T}_{\mathbf{s}}|\ln(2K/\delta)}{n}} + \frac{2|\mathcal{T}_{\mathbf{s}}|\ln(2K/\delta)}{n}\right) + \epsilon(\mathbf{s}).$$

The proof is completed. ■

7.2. Proof of Theorem 5

In this section, we give the proof of Theorem 5 to obtain a tighter bound by using a new concentration bounds on multinomial distributions from Proposition 16.

Proof [Proof of Theorem 5] From Lemma 17:

$$\begin{aligned}
&\left|\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')] - \frac{1}{n^2}\sum_{i,j\in[n]}\ell(\mathcal{A}_{\mathbf{s}},z_i,z_j)\right| \\
&\leq \sum_{i,j\in[K]}\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')|z\in C_i,z'\in C_j](p_j+\frac{|\mathcal{I}_j|}{n})\Big|p_i-\frac{|\mathcal{I}_i|}{n}\Big| \\
&+\left|\sum_{i,j\in[K]}\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')|z\in C_i,z'\in C_j]\frac{|\mathcal{I}_i||\mathcal{I}_j|}{n^2} - \frac{1}{n^2}\sum_{i,j=1}^{n}\ell(\mathcal{A}_{\mathbf{s}},z_i,z_j)\right| \\
&\leq \sum_{i\in[K]}\max_{j\in[K]}\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')|z\in C_i,z'\in C_j]\sum_{j\in[K]}(p_j+\frac{|\mathcal{I}_j|}{n})\Big|p_i-\frac{|\mathcal{I}_i|}{n}\Big| \\
&+\left|\sum_{i,j\in[K]}\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')|z\in C_i,z'\in C_j]\frac{|\mathcal{I}_i||\mathcal{I}_j|}{n^2} - \frac{1}{n^2}\sum_{i,j=1}^{n}\ell(\mathcal{A}_{\mathbf{s}},z_i,z_j)\right| \\
&\leq \sum_{i=1}^{K}\max_{j\in[K]}\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')|z\in C_i,z'\in C_j]\Big|p_i-\frac{|\mathcal{I}_i|}{n}\Big|\sum_{j=1}^{K}(p_j+\frac{|\mathcal{I}_j|}{n}) \\
&+\left|\sum_{i,j\in[K]}\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')|z\in C_i,z'\in C_j]\frac{|\mathcal{I}_i||\mathcal{I}_j|}{n^2} - \frac{1}{n^2}\sum_{i,j=1}^{n}\ell(\mathcal{A}_{\mathbf{s}},z_i,z_j)\right|.
\end{aligned}$$

By definition, we have $\sum_{i=1}^{K}|\mathcal{I}_i| = n$ and $\sum_{i=1}^{K}p_i = 1$. Then

$$\begin{aligned}
\left|\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')] - \frac{1}{n^2}\sum_{i,j\in[n]}\ell(\mathcal{A}_{\mathbf{s}},z_i,z_j)\right| &\leq 2\sum_{i=1}^{K}\alpha_i(\mathcal{A}_{\mathbf{s}})|p_i-\frac{|\mathcal{I}_i|}{n}| \\
&+\left|\sum_{i,j\in[K]}\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')|z\in C_i,z'\in C_j]\frac{|\mathcal{I}_i||\mathcal{I}_j|}{n^2} - \frac{1}{n^2}\sum_{i,j=1}^{n}\ell(\mathcal{A}_{\mathbf{s}},z_i,z_j)\right|,
\end{aligned}$$

where we have used the definition of $\alpha_i(\mathcal{A}_{\mathbf{s}})$ in Eq. (4.3). Invoking Lemma 18, we get

$$\begin{aligned}
&2\sum_{i=1}^{K}\alpha_i(\mathcal{A}_{\mathbf{s}})|p_i-\frac{|\mathcal{I}_i|}{n}| + \left|\sum_{i,j\in[K]}\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')|z\in C_i,z'\in C_j]\frac{|\mathcal{I}_i||\mathcal{I}_j|}{n^2} - \frac{1}{n^2}\sum_{i,j=1}^{n}\ell(\mathcal{A}_{\mathbf{s}},z_i,z_j)\right| \\
&\leq 2\sum_{i=1}^{K}\alpha_i(\mathcal{A}_{\mathbf{s}})|p_i-\frac{|\mathcal{I}_i|}{n}| + \max_{i,j\in[n]}\max_{z_p,z\in C_i,z_q,z'\in C_j}\left|\ell(\mathcal{A}_{\mathbf{s}},z,z') - \ell(\mathcal{A}_{\mathbf{s}},z_p,z_q)\right|.
\end{aligned}$$

By Definition 1, we have

$$\max_{i,j\in[n]}\max_{z_p,z\in C_i,z_q,z'\in C_j}\left|\ell(\mathcal{A}_{\mathbf{s}},z,z')-\ell(\mathcal{A}_{\mathbf{s}},z_p,z_q)\right|<\epsilon(\mathbf{s}).$$

Invoking Proposition 16 with $a_k(X)=\alpha_k(\mathcal{A}_{\mathbf{s}})$, $a_{\mathcal{T}_{\mathbf{s}}}(X)=\alpha_{\mathcal{T}_{\mathbf{s}}}(\mathcal{A}_{\mathbf{s}})$ and $a_{\mathcal{T}_{\mathbf{s}}^c}(X)=\alpha_{\mathcal{T}_{\mathbf{s}}^c}(\mathcal{A}_{\mathbf{s}})$, we have that for any $\delta\in(0,1)$, with probability at least $1-\delta$, the following holds

$$\left|\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')]-\frac{1}{n^2}\sum_{i,j\in[n]}\ell(\mathcal{A}_{\mathbf{s}},z_i,z_j)\right|\leq\epsilon(\mathbf{s})+$$
$$\sum_{k\in\mathcal{T}_{\mathbf{s}}}(\alpha_{\mathcal{T}_{\mathbf{s}}^c}(\mathcal{A}_{\mathbf{s}})+\sqrt{2}\alpha_k(\mathcal{A}_{\mathbf{s}})\sqrt{\frac{|\mathcal{I}_k^{\mathbf{s}}|}{n}}\sqrt{\frac{\ln(2K/\delta)}{n}}+\frac{2(\alpha_{\mathcal{T}_{\mathbf{s}}^c}(\mathcal{A}_{\mathbf{s}})|\mathcal{T}_{\mathbf{s}}|+\sum_{k\in\mathcal{T}_{\mathbf{s}}}\alpha_k(\mathcal{A}_{\mathbf{s}}))\ln(2K/\delta)}{n}.$$

The proof is completed. ■

7.3. Proof of Theorem 9

Under the pseudo-robustness setting we can easily deduce the following result from the Lemma 18. The proof is given in the appendix.

Lemma 19 *If a learning algorithm $\mathcal{A}$ is $(K,\epsilon(\cdot),\hat{p}_n(\cdot))$ pseudo-robust, the training pairs come from a sample generated by n IID draws from μ, then we have:*

$$\left|\sum_{i,j\in[K]}\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')|z\in C_i,z'\in C_j]\frac{|\mathcal{I}_i||\mathcal{I}_j|}{n^2}-\frac{1}{n^2}\sum_{i,j=1}^{n}\ell(\mathcal{A}_{\mathbf{s}},z_i,z_j)\right|\leq\frac{|\hat{\mathbf{s}}^2|}{n^2}+\frac{n^2-|\hat{\mathbf{s}}^2|}{n^2}\xi\left(\mathcal{A}_{\mathbf{s}}\right).$$

The second term in Lemma 17 is now bounded by the Lemma 19. Then, analyzing similarly to the proof of Theorem 2 gives Theorem 9.

Proof [Proof of Theorem 9] By Lemma 17 and Lemma 19, we know

$$\begin{aligned}\left|\mathcal{L}\left(\mathcal{A}_{\mathbf{s}}\right)-\mathcal{L}_{\mathbf{emp}}\left(\mathcal{A}_{\mathbf{s}}\right)\right|\leq&\frac{|\hat{\mathbf{s}}^2|}{n^2}+\frac{n^2-|\hat{\mathbf{s}}^2|}{n^2}\xi\left(\mathcal{A}_{\mathbf{s}}\right)\\&+\sum_{j=1}^{K}\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')|z\in C_i,z'\in C_j](p_j+\frac{|\mathcal{I}_j|}{n})\sum_{i=1}^{K}\left|p_i-\frac{|\mathcal{I}_i|}{n}\right|.\end{aligned}\tag{7.1}$$

Then we treat the last term in the same way as in Theorem 2, and use Proposition 8 to get

$$\begin{aligned}&\sum_{j=1}^{K}\mathbb{E}_{z,z'\sim\mu}[\ell(\mathcal{A}_{\mathbf{s}},z,z')|z\in C_i,z'\in C_j](p_j+\frac{|\mathcal{I}_j|}{n})\sum_{i=1}^{K}\left|p_i-\frac{|\mathcal{I}_i|}{n}\right|\\&\leq\xi\left(\mathcal{A}_{\mathbf{s}}\right)\left((\sqrt{2}+1)\sqrt{\frac{|\mathcal{T}_{\mathbf{s}}|\ln(2K/\delta)}{n}}+\frac{2\,|\mathcal{T}_{\mathbf{s}}|\ln(2K/\delta)}{n}\right).\end{aligned}\tag{7.2}$$

By combining the above two inequalities, we obtain the desired statement. ■

8. Conclusion

Our work bridges a critical gap in robustness theory by establishing the first data-dependent generalization bounds for pairwise learning algorithms. Moving beyond the limitations of prior analyses, we introduce a novel robust optimization framework that fundamentally recalibrates the theoretical foundations of pairwise learning through three advances. First, we replace uniform supremum bounds with hypotheses-sensitive quantities $\xi(\mathcal{A}_{\mathbf{s}})$ and $\xi(\mathcal{A}_{\mathbf{s}})$, using actual algorithm output and conditional expectations to eliminate crude uniform bounding. Second, we substitute combinatorial partition cardinality K with the effective partition size $|\mathcal{T}_{\mathbf{s}}|$, achieving logarithmic dependence on K and orders-of-magnitude tightness gains for sparse distributions. Third, we formalize robustness guarantees for non-IID training pairs through conditional partition expectations and adaptive error weighting, overcoming the IID assumption bottleneck that limited prior pointwise-to-pairwise transfers.

These advances collectively resolve the "relative looseness" limitation noted by Bellet and Habrard (2015) while preserving the geometric interpretability of robustness frameworks. By demonstrating exponentially tighter bounds than in the prior art through Theorems 2 and 9, our approach unifies robust optimization with pairwise loss structures to establish a new theoretical foundation. The minimal distributional assumptions of this framework make it particularly suitable for noisy real-world scenarios in high-impact applications that include metric learning, ranking, and AUC maximization. Future work will extend this foundation to adversarial pairwise learning and deep metric architectures.

Acknowledgement

The corresponding author is Yunwen Lei. The work of Yunwen Lei is partially supported by the Research Grants Council of Hong Kong [Project Nos. 22303723, 17302624].

References

Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *JMLR*, 10(Feb):441–474, 2009.

Aurélien Bellet and Amaury Habrard. Robustness and generalization for metric learning. *Neurocomputing*, 151:259–267, 2015.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

Qiong Cao, Zheng-Chu Guo, and Yiming Ying. Generalization bounds for metric and similarity learning. *Machine Learning*, 102(1):115–132, 2016.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

Jun Chen, Hong Chen, Xue Jiang, Bin Gu, Weifu Li, Tieliang Gong, and Feng Zheng. On the stability and generalization of triplet learning. In *AAAI*, pages 7033–7041, 2023.

Jun Chen, Hong Chen, Bin Gu, Guodong Liu, Yingjie Wang, and Weifu Li. Error analysis affected by heavy-tailed gradients for non-convex pairwise stochastic gradient descent. In *AAAI*, volume 39, pages 15803–15811, 2025.

Andreas Christmann and Ding-Xuan Zhou. On the robustness of regularized pairwise learning methods based on kernels. *Journal of Complexity*, 37:1–33, 2016.

Stéphan Clémençon, Gabor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, pages 844–874, 2008.

Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. In *NeurIPS*, pages 313–320, 2004.

Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. On the optimization and generalization of multi-head attention. *Transactions on Machine Learning Research*, 2024.

Jun Fan and Yunwen Lei. High-probability generalization bounds for pointwise uniformly stable algorithms. *Applied and Computational Harmonic Analysis*, 70:101632, 2024.

Wei Gao, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. One-pass AUC optimization. In *ICML*, pages 906–914, 2013.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110:393–416, 2021.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, pages 1225–1234, 2016.

Nong Minh Hieu and Antoine Ledent. Generalization analysis for supervised contrastive representation learning under non-IID settings. In *ICML*, 2025.

Nong Minh Hieu, Antoine Ledent, Yunwen Lei, and Cheng Yeaw Ku. Generalization analysis for deep contrastive representation learning. In *AAAI*, pages 17186–17194, 2025.

Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Regularization schemes for minimum error entropy principle. *Analysis and Applications*, 13(04):437–455, 2015.

Haotian Ju, Dongyue Li, and Hongyang R Zhang. Robust fine-tuning of deep neural networks with hessian-based generalization guarantees. In *ICML*, pages 10431–10461, 2022.

Purushottam Kar, Bharath Sriperumbudur, Prateek Jain, and Harish Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. In *ICML*, pages 441–449, 2013.

Kenji Kawaguchi, Zhun Deng, Kyle Luh, and Jiaoyang Huang. Robustness implies generalization via data-dependent generalization bounds. In *ICML*, pages 10866–10894, 2022.

Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *ICML*, pages 942–950. PMLR, 2013.

Yunwen Lei. Towards better generalization bounds of stochastic optimization for nonconvex learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3621591.

Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. In *NeurIPS*, volume 33, pages 21236–21246, 2020.

Yunwen Lei, Tao Sun, and Mingrui Liu. Minibatch and local sgd: Algorithmic stability and linear speedup in generalization. *Applied and Computational Harmonic Analysis*, page 101795, 2025.

Mingrui Liu, Xiaoxuan Zhang, Zaiyi Chen, Xiaoyu Wang, and Tianbao Yang. Fast stochastic AUC maximization with O(1/n)-convergence rate. In *ICML*, pages 3195–3203, 2018.

Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *ICML*, pages 2159–2167, 2017.

Konstantinos Nikolakakis, Farzin Haddadpour, Amin Karbasi, and Dionysios Kalogerias. Beyond lipschitz: Sharp generalization and excess risk bounds for full-batch gd. In *ICLR*, 2022.

Guillaume Papa, Stéphan Clémençon, and Aurélien Bellet. Sgd algorithms based on incomplete u-statistics: large-scale minimization of empirical risk. *NeurIPS*, 28, 2015.

Zi-Hao Qiu, Quanqi Hu, Yongjian Zhong, Lijun Zhang, and Tianbao Yang. Large-scale stochastic optimization of ndcg surrogates for deep learning with provable convergence. In *ICML*, pages 18122–18152, 2022.

Wojciech Rejchel. On ranking and generalization bounds. *JMLR*, 13(May):1373–1392, 2012.

Matan Schliserman and Tomer Koren. Stability vs implicit bias of gradient methods on separable data and beyond. In *COLT*, pages 3380–3394, 2022.

Yuyang Wang, Roni Khardon, Dmitry Pechyony, and Rosie Jones. Generalization bounds for online learning algorithms with pairwise loss functions. In *COLT*, pages 13–1, 2012.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. *NeurIPS*, 35:15446–15459, 2022.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine Learning*, 86:391–423, 2012.

Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He, Xiaochun Cao, and Qingming Huang. Optimizing two-way partial auc with an end-to-end framework. *TPAMI*, 45(8):10228–10246, 2022.

Han-Jia Ye, De-Chuan Zhan, Xue-Min Si, Yuan Jiang, and Zhi-Hua Zhou. What makes objects similar: A unified multi-metric learning approach. *NeurIPS*, 29, 2016.

Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online AUC maximization. In *NeurIPS*, pages 451–459, 2016.

Junyu Zhou, Shuo Huang, Han Feng, Puyu Wang, and Ding-Xuan Zhou. Fine-grained analysis of non-parametric estimation for pairwise learning. *arXiv preprint arXiv:2305.19640*, 2023.

Miaoxi Zhu, Yan Sun, Li Shen, Bo Du, and Dacheng Tao. Stability and generalization for distributed sgda. *arXiv preprint arXiv:2411.09365*, 2024.