Generalization Bounds for Adversarial Metric Learning

Wen Wen[1] , **Han Li**[1,*] , **Hong Chen**[1,2,3] , **Rui Wu**[4] , **Lingjuan Wu**[1] , **Liangxuan Zhu**[1]
[1]College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[2]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan 430070, China
[3]Key Laboratory of Smart Farming for Agricultural Animals, Wuhan 430070, China
[4]Horizon Robotics, Haidian District, BeiJing 100190, China
lihan125@mail.hzau.edu.cn

Abstract

Recently, adversarial metric learning has been proposed to enhance the robustness of the learned distance metric against adversarial perturbations. Despite rapid progress in validating its effectiveness empirically, theoretical guarantees on adversarial robustness and generalization are far less understood. To fill this gap, this paper focuses on unveiling the generalization properties of adversarial metric learning by developing the uniform convergence analysis techniques. Based on the capacity estimation of covering numbers, we establish the first high-probability generalization bounds with order $O(n^{-\frac{1}{2}})$ for adversarial metric learning with pairwise perturbations and general losses, where n is the number of training samples. Moreover, we obtain the refined generalization bounds with order $O(n^{-1})$ for the smooth loss by using local Rademacher complexity, which is faster than the previous result of adversarial pairwise learning, e.g., adversarial bipartite ranking. Experimental evaluation on real-world datasets validates our theoretical findings.

1 Introduction

The robustness of metric learning against adversarial perturbations has attracted increasing attention in the machine learning literature, where abundant adversarial algorithms have been proposed from various application motivations, *e.g.*, [Huang *et al.*, 2019; Bouniot *et al.*, 2020; Liu *et al.*, 2022]. Despite the previous adversarial metric learning enjoys the adversarial robustness [Madry *et al.*, 2018; Kurakin *et al.*, 2018; Carlini and Wagner, 2017] empirically, its generalization guarantee is touched scarcely in theory. In this paper, our goal is to fill this theoretical gap and provide the sharper high-probability generalization bounds of adversarial metric learning from the lens of statistical learning theory [Vapnik, 1999; Mohri *et al.*, 2018].

Although theoretical foundations of metric learning have been well understood in [Huai *et al.*, 2019; Lei *et al.*, 2020; Ye *et al.*, 2019], there are two-fold challenges in establishing generalization analysis for adversarial counterparts. The first one is caused by the joint perturbations on sample pairs [Huai *et al.*, 2022], which is more complicated than the case of the single-sample perturbation [Yin *et al.*, 2019; Xing *et al.*, 2021; Mustafa *et al.*, 2022]. The other arises from the non-smoothness and non-differentiable optimization objective associated with the adversarial loss function [Xing *et al.*, 2021; Xiao *et al.*, 2022], which leads to the standard analysis techniques (*e.g.*, [Cao *et al.*, 2016]) inapplicable.

To surmount the above challenges, we introduce the ℓ_∞ covering number [Reeve and Kaban, 2020; Mustafa *et al.*, 2022] to measure the complexity of function space with pairwise perturbations and employ a general loss class to approximate the adversarial loss class on training samples to tackle the non-smoothness problem. In addition to providing generalization guarantees for adversarial metric learning, we also validate our theoretical findings through experimental analysis on real-world datasets. In summary, the main contributions of this paper are listed as follows:

- We establish the high-probability generalization bounds with order $O(n^{-\frac{1}{2}})$ for adversarial metric learning with pairwise perturbations, where n is the sample size. Indeed, our high probability bounds are beneficial to understand the robustness of optimization algorithms [Bousquet *et al.*, 2020; Klochkov and Zhivotovskiy, 2021; Li and Liu, 2021] and are different from the existing bounds in expectation [Xing *et al.*, 2021; Farnia and Ozdaglar, 2021; Xiao *et al.*, 2022]. These developed learning bounds are valid for general adversarial perturbations measured by ℓ_r-norm $(r \geq 1)$, and adapt to linear metric learning models and deep metric learning models simultaneously. To the best of our knowledge, this is the first-ever-known generalization bounds for metric learning with pairwise perturbations.
- Under the self-bounding Lipschitz assumption [Reeve and Kaban, 2020] of loss function, we provide the sharper generalization bound with the order $\mathcal{O}(n^{-1})$ by developing the concentration estimation technique associated with the local Rademacher complexity [Bartlett *et al.*, 2005]. As a by-product, the current generalization bounds with respect to the non-adversarial metric learning assure faster rates than the previous generalization analysis in [Huai *et al.*, 2019; Lei *et al.*, 2020].

*Corresponding author.

Perturbation Object	Task	Reference	Perturbation Way	Analysis Tool	Learning Bound
Single sample	Classification	Yin et al. (2019) Khim and Loh (2018)	ℓ_∞-norm additive	Rademacher complexity	$\mathcal{O}(1/\sqrt{n})$
		Tu et al. (2019)	ℓ_r-norm additive	Rademacher complexity	$\mathcal{O}(1/\sqrt{n})$
		Mustafa (2022)	ℓ_r-norm additive	Covering number	$\mathcal{O}(1/\sqrt{n})$
			Spatial transformation	Local Rademacher complexity	$*\mathcal{O}(1/n)$
	Optimization	Xing et al. (2021) Xiao et al. (2022)	ℓ_r-norm additive	Algorithmic stability	$\star\mathcal{O}(1/n)$
	Bipartite Ranking	Mo et al. (2022)	ℓ_r-norm additive	Rademacher complexity	$\mathcal{O}(1/\sqrt{n})$
Sample pair	Metric Learning	Ours	ℓ_r-norm additive	Covering number	$\mathcal{O}(1/\sqrt{n})$
				Local Rademacher complexity	$*\mathcal{O}(1/n)$

Table 1: Summary of generalization analysis for adversarial learning ($*$-optimization bound; $\star$-generalization bound in expectation).

2 Related Work

Adversarial Metric Learning. Adversarial metric learning plays a vital role in applications ranging from person re-identification [Dai *et al.*, 2018; Bouniot *et al.*, 2020; Liu *et al.*, 2022] to zero-shot learning [Chen and Deng, 2019; Huang *et al.*, 2019] and cross-modal retrieval [Xu *et al.*, 2019]. Since metric learning methods learn on the original samples are limited in their capacity to distinguish ambiguous samples, adversarial learning methods are proposed to facilitate robust metric learning (see *Appendix A*). Although adversarial metric learning has shown empirical effectiveness, theoretical aspects of adversarial robustness has not been exhaustively studied.

Adversarial Generalization. The nonsmoothness and non-differentiability with respect to adversarial loss are central difficulties in the generalization analysis of adversarial learning. To overcome these obstacles, Xing et al. (2021) propose a noise injection method to avoid non-differentiability, and establish stability upper bound and lower bound for a generic adversarial training algorithm. Xiao et al. (2022) tackle the non-smooth problem by considering the η-approximate smoothness on adversarial loss. Based on this, they derive stability-based generalization bounds for stochastic gradient descent on the general class, which covers the adversarial loss. Tu et al. (2019) fit the adversarial learning problem into the minimax framework by introducing a transport map between distributions. They derive a new risk bound for the minimax problem through the lens of covering numbers under the Lipschitz assumption. Mo et al. (2022) extend the prior work of Tu et al. (2019) to pairwise learning and establish the generalization bounds for adversarial bipartite ranking. Yin et al. (2019) and Khim et al. (2018) derive a surrogate upper bound on the adversarial loss, and then show upper bounds for the adversarial Rademacher complexity of the surrogate. All the above generalization analysis are limited to adversarial learning with pointwise perturbation. To the best of our knowledge, there is no the related generalization analysis for adversarial learning with pairwise perturbation. This paper try to fill this gap. Table 1 summarizes the related work on adversarial learning.

3 Preliminaries

This section introduces the main notations used in this paper, the necessary backgrounds on adversarial metric learning [Wang *et al.*, 2020; Liu *et al.*, 2022; Yang *et al.*, 2021], and some theoretical techniques and structural results used for the generalization analysis.

3.1 Notations

We denote vectors as lowercase letters (*e.g.*, x) and matrices as uppercase letters (*e.g.*, X). We write $\|w\|_p$ to denote the ℓ_p-norm of a vector $w \in \mathbb{R}^n$. The dual norm of w is denoted by a star (*i.e.*, $\|w\|_{p^*}$). For a matrix $W \in \mathbb{R}^{n\times n}$ with columns W_i, $i \in [n]$, the matrix (p,q)-norm is defined by $\|W\|_{p,q} = \|(\|W_1\|_p, \ldots, \|W_n\|_p)\|_q$.

3.2 Adversarial Metric Learning

Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a set of training samples drawn according to an unknown distribution $\mathcal{P}$, where $x_i \in \mathbb{R}^d$ is the d dimensional feature vector and $y_i \in \mathbb{R}$ is the class label. The $d \times n$ input feature matrix is denoted by $X = (x_i : i \in [n])$. Given a mapping $f : \mathbb{R}^d \to \mathbb{R}^{d'}$, which maps the d-dimensional input into an embedding space with d'-dimension, then the distance between samples x_i and x_j is measured by

$$D_f(x_i, x_j) := (f(x_i) - f(x_j))^T (f(x_i) - f(x_j)). \quad (1)$$

The target of metric learning is to learn an adequate f such that reflects the similarity between sample pairs [Wang *et al.*, 2020; Huai *et al.*, 2022]. The widely adopted method of seeking such f is to minimize the following empirical risk over the given training samples

$$\mathcal{E}_n(f) = \frac{1}{n(n-1)} \sum_{i\neq j} \ell(\tau(y_i, y_j)(1 - D_f(x_i, x_j))), \quad (2)$$

where $\ell : \mathbb{R} \to \mathbb{R}_+$ is a given loss function such as the hinge loss function, and $\tau(y_i, y_j) \in \{-1, 1\}$ indicates whether two samples are affiliated to the same class, *i.e.*, $\tau(y_i, y_j) = 1$ if $y_i = y_j$, and $\tau(y_i, y_j) = -1$ otherwise.

However, in the presence of adversaries, there will be imperceptible perturbations on the input samples that lead to maximizaing empirical risk (2). Throughout this paper, we assume that the perturbation θ is adversarially chosen in the ℓ_r-ball $\mathcal{B}(\varepsilon) \subseteq \mathbb{R}^d$ of radius ε, for an arbitrary $r \geq 1$. Given a sample pair $(x_i, y_i), (x_j, y_j)$ and a learned mapping f, the adversary selects valid perturbations θ_i^* and θ_j^* by [Huai *et al.*, 2022]

$$\theta_i^*, \theta_j^* = \arg \max_{\theta_i, \theta_j \in \mathcal{B}(\varepsilon)} \ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j))),$$

and the *adversarial loss* $\ell_{adv}((x_i, y_i), (x_j, y_j); f)$ of f at $(x_i, y_i), (x_j, y_j)$ can be written as

$$\max_{\theta_i, \theta_j \in \mathcal{B}(\varepsilon)} \ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j))). \tag{3}$$

We then have the following *adversarial empirical risk* $\widetilde{\mathcal{E}}_n(f)$

$$\frac{1}{n(n-1)} \sum_{i \neq j} \max_{\theta_i, \theta_j \in \mathcal{B}(\varepsilon)} \ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j))),$$

and the *adversarial expected risk* $\widetilde{\mathcal{E}}(f)$

$$\mathbb{E}_{\mathcal{P}} \Big[\max_{\theta_i, \theta_j \in \mathcal{B}(\varepsilon)} \ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j))) \Big].$$

The adversarial empirical risk $\widetilde{\mathcal{E}}_n(f)$ measures the ability of f to place similar samples nearby and separate dissimilar samples on the training samples with adversarial perturbations. The adversarial expected risk $\widetilde{\mathcal{E}}(f)$ measures how well f generalizes to unseen adversarial samples. In this paper, we are interested in the difference between $\widetilde{\mathcal{E}}_n(f)$ and $\widetilde{\mathcal{E}}(f)$. Our main tool for bounding the generalization error for adversarial metric learning (*i.e.*, $\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f)$) is the ℓ_∞-covering number defined below.

Definition 1 (ℓ_∞-covering number)**.** *Let $\upsilon > 0$ and let $(\mathcal{A}, \|\cdot\|_\infty)$ be a metric space. We say that $\mathcal{C} \subseteq \mathcal{A}$ is an $(\upsilon, \|\cdot\|_\infty)$-covering of $\mathcal{A}$ if*

$$\sup_{a \in \mathcal{A}} \inf_{c \in \mathcal{C}} \|a - c\|_\infty \leq \upsilon.$$

Then, the ℓ_∞-covering number of $\mathcal{A}$ is the minimum cardinality of any subset covers $\mathcal{A}$ at scale υ, denoted as $\mathcal{N}_\infty(\upsilon, \mathcal{A})$.

As a special case of Zhang et al. (2002), Definition 1 generally characterizes the complexity of the function space measured by the infinite norm [Reeve and Kaban, 2020; Mustafa *et al.*, 2022].

Let the mapping f be selected from the hypothesis class $\mathcal{F} := \{x \mapsto f_W(x) : x \in \mathbb{R}^d, W \in \mathbb{R}^{d \times d'}\}$. The class of adversarial loss functions (3) is written as

$$\mathcal{L}_{adv} := \{(x_i, y_i), (x_j, y_j) \mapsto \ell_{adv}((x_i, y_i), (x_j, y_j); f) : f \in \mathcal{F}\}. \tag{4}$$

We have the following relationship between the generalization error (*i.e.*, $\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f)$) and ℓ_∞-covering number of adversarial loss class $\mathcal{L}_{adv}$ on the training sample S (*i.e.*, $\mathcal{N}_\infty(\mathcal{L}_{adv}, \upsilon, S)$), which extends the previous results of Bartlett et al. (2017) to adversarial learning.

Lemma 1. *Let $\mathcal{L}_{adv}$ be the adversarial loss class defined in (4) and bounded by* 1*. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over a sample S of size n, the following holds for all $f \in \mathcal{F}$*

$$\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f) \leq 3\sqrt{\frac{\log(2/\delta)}{2n}} + \inf_{\alpha > 0} \Big(\frac{8\alpha}{\sqrt{n}} + \frac{24}{n} \int_\alpha^{\sqrt{n}} \sqrt{\log \mathcal{N}_\infty(\mathcal{L}_{adv}, \upsilon, S)} d\upsilon \Big).$$

Lemma 1 allows us to control the generalization error by bounding the ℓ_∞-covering number of the adversarial loss class on training samples. However, deriving an upper bound on $\mathcal{N}_\infty(\mathcal{L}_{adv}, \upsilon, S)$ is intractable due to the outer maximization of loss functions in class $\mathcal{L}_{adv}$ and the joint action of pairwise perturbations θ_i, θ_j. Our approach is to approximate the loss class $\mathcal{L}_{adv}$ on sample S by the following class $\widetilde{\mathcal{L}}_{adv}$

$$\widetilde{\mathcal{L}}_{adv} := \{((x_i, \theta_i), y_i), ((x_j, \theta_j), y_j) \mapsto \ell(\tau(y_i, y_j) (1 - D_f(x_i + \theta_i, x_j + \theta_j))) : f \in \mathcal{F}\}, \tag{5}$$

and incorporate perturbations θ_i, θ_j into the argument. Based on this, we reduce the problem of measuring the complexity of the adversarial loss class on training samples to measuring the complexity of a general loss class. Some necessary Lipschitz conditions are introduced for our theoretical analysis.

Definition 2. *Let $\|\cdot\|$ denote a norm metric, and $\xi, \zeta \geq 0$. For a loss function $\ell : \mathcal{F} \to \mathbb{R}$ and a distance function $D_f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ parametrized by $f \in \mathcal{F}$,*

1) the loss function ℓ is ξ-Lipschitz if, for $\forall f, f' \in \mathcal{F}$

$$|\ell(f) - \ell(f')| \leq \xi \|f - f'\|,$$

2) the distance function D_f is the Muti-variate Lipschitz continuity if, for $\forall \theta_i, \theta_j, \theta_i', \theta_j' \in \mathbb{R}^d$

$$|D_f(\theta_i, \theta_j) - D_f(\theta_i', \theta_j')| \leq \zeta \|(\theta_i, \theta_j) - (\theta_i', \theta_j')\|.$$

The Lipschitzness on the loss function in Definition 2 is a mild condition, which is satisfied by some common losses, *e.g.*, the hinge loss and logistic loss [Yin *et al.*, 2019; Tu *et al.*, 2019; Lei *et al.*, 2020]. By utilizing this Lipschitzness and the notion of Multi-variate Lipschitz continuity [Zantedeschi *et al.*, 2016], we have the Lipschitzness on the functions $\theta_i, \theta_j \mapsto \ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j)))$ for $f \in \mathcal{F}$, which is necessary and fulfilled by most attacks [Madry *et al.*, 2018; Awasthi *et al.*, 2021].

We now present our first main result as follows.

Theorem 1. *Let $\theta_i, \theta_j \mapsto \ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j)))$ be the Lipschitzness with constant L. Let $C_{\mathcal{B}}(3\upsilon/4L)$ be a $3\upsilon/4L$-cover of $\mathcal{B}(\varepsilon)$, and define the adversarial sample set*

$$\tilde{S} = \{((x_i, \theta_i), y_i) : i \in [n], \theta_i \in C_{\mathcal{B}}(3\upsilon/4L)\}.$$

Then, we have

$$\mathcal{N}_\infty(\mathcal{L}_{adv}, \upsilon, S) \leq \mathcal{N}_\infty(\widetilde{\mathcal{L}}_{adv}, \upsilon/4, \tilde{S}).$$

Detailed proofs are contained in *Appendix C.2.* Theorem 1 illustrates that the ℓ_∞ covering number of adversarial loss class $\mathcal{L}_{adv}$ on set S can be bounded by the ℓ_∞ covering number of class $\widetilde{\mathcal{L}}_{adv}$ on adversarial set $\tilde{S}$, which extends the Lemma 4.4 of Mustafa, Lei and Kloft (2022) for adversarial pointwise learning to adversarial pairwise learning. It will be served to derive high-probability generalization bounds of adversarial metric learning.

4 The Generalization Bounds for Adversarial Metric Learning

In this section, we provide a sharp characterization of the generalization for two commonly-used adversarial metric learning models: the linear and deep metric learning models. The adversarial perturbation is measured in ℓ_r-norm. Moreover, we establish fast generalization bounds for adversarial metric learning through the local Rademacher complexity under the smooth Lipschitz assumption on loss functions.

4.1 Linear Metirc Learning Model

We consider the following linear hypothesis class:

$$\mathcal{F} := \{x_i \mapsto Wx_i : W \in \mathbb{R}^{d\times d'}, \|W\|_{p,1} \le \Lambda\}.$$

For any linear mapping $f \in \mathcal{F}$, the distance metric function is defined by

$$D_f(x_i, x_j) = (x_i - x_j)^T W^T W(x_i - x_j). \tag{6}$$

The Lipschitz constant of the function $\theta_i, \theta_j \mapsto \ell(\tau(y_i, y_j)(1-D_f(x_i+\theta_i, x_j+\theta_j)))$ required in Theorem 1 is provided in the following lemma.

Lemma 2. *Let D_f be the linear distance function defined in (6). Then, for any sample pair (x_i, y_i) and (x_j, y_j), the function $\theta_i, \theta_j \mapsto \ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j)))$ is $\mathbf{4\xi\Lambda^2\Psi}$-Lipschitz, where $\mathbf{4\Lambda^2\Psi}$ is the Multi-variate Lipschitz constant of the function $\theta_i, \theta_j \mapsto D_f(x_i + \theta_i, x_j + \theta_j)$, and $\mathbf{\Psi}$ is $\max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|X\|_{r,\infty} + \varepsilon)$.*

The proof is provided in *Appendix.D.1.* Based on the Lipschitzness of the adversarial loss function in Lemma 2, we obtain an upper bound on the covering number of the loss class $\widetilde{\mathcal{L}}_{adv}$ on the sample $\tilde{S}$ in the theorem below.

Theorem 2. *With the notation in Lemma 2. Let $\widetilde{\mathcal{L}}_{adv}$ be defined in (5) and $\tilde{S}$ be defined in Theorem 1. Then, for $\upsilon > 0$, we have*

$$\log \mathcal{N}_\infty(\widetilde{\mathcal{L}}_{adv}, \upsilon/4, \tilde{S}) \le C\frac{\xi^2\Lambda^4\widehat{\Psi}^2}{\upsilon^2}L_{\log},$$

where

$$L_{log} = \log\left(4\left\lceil\frac{32\xi\Lambda^2\widehat{\Psi}}{\upsilon} + 1\right\rceil n\Big(\frac{16\xi\Lambda^2\varepsilon\Psi}{\upsilon}\Big)^d + 1\right),$$

$\Psi = \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|X\|_{r,\infty} + \varepsilon)$, $\widehat{\Psi} = \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|X\|_{r,\infty} + \varepsilon)^2$ *and C is a constant.*

The detailed proof is contained in *Appendix D.1.* Based on Theorem 2 and Lemma 1, we establish the following high-probability generalization bound.

Theorem 3. *With the notation above. For any fixed $\xi > 0$ and all $f \in \mathcal{F}$, with probability at least $1 - \delta$, we have*

$$\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f) \le 3\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{8}{n^{3/2}} + C\frac{\xi\Lambda^2\widehat{\Psi}\log(n)}{n} \times \sqrt{\log\left(4\left\lceil 32n\xi\Lambda^2\widehat{\Psi} + 1\right\rceil n\Big(16n\xi\Lambda^2\varepsilon\Psi\Big)^d + 1\right)}.$$

where $\widehat{\Psi} = \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|X\|_{r,\infty} + \varepsilon)^2$ and C is a constant.

The proof of this theorem is provided in *Appendix D.2.*

Remark 1. *The generalization bound in Therorem 3 suffers from additional dimension dependent terms as compared to its non-adversarial counterpart. The first $d^{1-\frac{1}{p}-\frac{1}{r}}$ dependence in $\widehat{\Psi}$ is due to the mismatch between the norm on the input x and the norm in the ball $\mathcal{B}(\varepsilon)$. Indeed, we have used the inequality $\|x_i + \theta_i\|_{p^*} \le \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})\|x_i + \theta_i\|_r$ in Awasthi et al. (2021), where $1 - 1/p = 1/p^*$. If $\frac{1}{p} + \frac{1}{r} \ge 1$, $\widehat{\Psi}$ is dimension independent, which implies that one should choose a p-norm regularizer on W (i.e., the weight matrix), where $p \in [1, r^*]$. The second $\sqrt{d}$ dependence in square root of the third term on the right side of Therorem 3, is attributed to the complexity of the perturbation ball $\mathcal{B}(\varepsilon)$. For example, if $\mathcal{B}(\varepsilon)$ is contained in a low dimensional space $d' < d$, the dependence is reduced to $\mathcal{O}(\sqrt{d'})$. This motivates the mapping f to project the input $x \in \mathbb{R}^d$ into a low-dimensional subspace to reduce the effective dimensionality of adversarial perturbations.*

Remark 2. *Theorem 3 is a high-probability generalization bound for adversarial metric learning in the linear case, motivated by the recent analyses in the adversarial pointwise learning (Awasthi et al. 2021; Mustafa, Lei and Kloft 2022). In contrast with prior work of Mustafa et al. (2022) that studies ℓ_∞-norm perturbations, we consider the general case where the perturbations are measured in ℓ_r-norm. Moreover, our theoretical analysis is novel since it is the first touch for adversarial pairwise learning with pairwise perturbations than existing work [Yin* et al.*, 2019; Mo* et al.*, 2022; Mustafa* et al.*, 2022].*

Remark 3. *Setting $\varepsilon = 0$, we obtain a standard risk bound for linear metric learning in non-adversarial case; see (7). Although the bound (7) with order $\mathcal{O}(1/\sqrt{n})$ is similar to the generalization bounds in Cao at al. (2016), Ye et al. (2019), and Let et al. (2020), our result applies to a wider range of loss functions such as hinge loss and logistic loss.*

$$\mathcal{E}(f) - \mathcal{E}_n(f) \le 3\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{8}{n^{3/2}} + C\frac{\xi\Lambda^2\|X\|^2_{p^*,\infty}}{n} \times \sqrt{\log\left(4\left\lceil 32n\xi\Lambda^2\|X\|^2_{p^*,\infty} + 1\right\rceil n + 1\right)}\log(n). \tag{7}$$

4.2 Deep Metirc Learning Model

Let the mapping f be a L-layer neural network parametrized by the weights $W = \{W^l \in \mathbb{R}^{h_l\times h_{l-1}}\}_{l=1}^L$, where h_l is the number of neurons in the l-th layer of the network and $h_0 = d$. Given the input sample $x_i \in \mathbb{R}^d$, the output of the final layer in the network can be written as

$$f(x_i) = W^L\rho(W^{L-1}\rho(\cdots\rho(W^1x_i))), \tag{8}$$

where $\rho(\cdot)$ denotes the non-linear 1-Lipschitz activation function. We consider norm-bounded networks with the following hypothesis class

$$\mathcal{F} := \{x_i \mapsto f(x_i) : f \in \mathcal{F}, \|W^l\|_F \le b_l, \|W^l\|_\sigma \le s_l\},$$

where $\|\cdot\|_\sigma$ represents spectral norm and $\|\cdot\|_F$ denotes the Frobenius norm.

As with the linear case, we first establish the Lipschitzness of the function $\theta_i, \theta_j \mapsto \ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j)))$. The results are summarized in the following lemma.

Lemma 3. *Let $f \in \mathcal{F}$ be the neural network defined in (8) and D_f be the distance metric function defined as (1). For all $(x_i, y_i), (x_j, y_j)$ and $f \in \mathcal{F}$, the function $\ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j)))$ is $\mathbf{4\xi \prod_{l=1}^{L} s_l^2 \Psi}$-Lipschitz in θ_i, θ_j, where $\mathbf{4 \prod_{l=1}^{L} s_l^2 \Psi}$ is the Multi-variate Lipschitz constant of the function $\theta_i, \theta_j \mapsto D_f(x_i + \theta_i, x_j + \theta_j)$, and Ψ is* $\max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|X\|_{r,\infty} + \varepsilon)$.

The proof is given in *Appendix E.* With this lemma, we establish the following upper bound on the ℓ_∞-covering number of the class $\widetilde{\mathcal{L}}_{adv}$ in nonlinear cases.

Theorem 4. *With the notation in Lemma 3. Let $\widetilde{\mathcal{L}}_{adv}$ be defined in (5) and $\tilde{S}$ be defined in Theorem 1. Then, for $\epsilon > 0$, we have*

$$\log \mathcal{N}_\infty(\widetilde{\mathcal{L}}_{adv}, \upsilon/4, \tilde{S}) \leq \frac{C\xi^2 \widehat{\Psi}^2 L^4}{\upsilon^2} \prod_{l=1}^{L} s_l^4 \left(\sum_{l=1}^{L} \frac{b_l^2}{s_l^2} \right)^2 L_{\log},$$

where

$$L_{log} = \log \left(\Big\lceil \frac{C_1 \xi \widehat{\Psi} \Gamma^2}{\upsilon} + C_2 \Big\rceil n\hat{h} \Big(\frac{16\xi \prod_{l=1}^{L} s_l^2 \varepsilon \Psi}{\upsilon} \Big)^d + 1 \right),$$

$\Psi = \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})(\|X\|_{r,\infty} + \varepsilon)$, $\widehat{\Psi} = \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})$ $(\|X\|_{r,\infty} + \varepsilon)^2$, $\Gamma = \max_{l\in[L]}(\prod_{i=1}^{L} s_i) b_l / s_l$, $\hat{h} = \max_{l\in[L]} h_l$, *and C, C_1, C_2 are universal constants.*

The proof is given in *Appendix E.* By combining Theorem 4 with Lemma 1, we obtain the following sharp bound with high probability.

Theorem 5. *With the notation in Theorem 4. For any fixed $\xi > 0$ and all $f \in \mathcal{F}$, with probability at least $1 - \delta$, we have*

$$\begin{aligned} &\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f) \\ &\leq 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}} + \frac{8}{n^{3/2}} + \frac{C\xi \widehat{\Psi} L^2}{n} \prod_{l=1}^{L} s_l^2 \sum_{l=1}^{L} \frac{b_l^2}{s_l^2} \log(n) \widetilde{L}_{\log}, \end{aligned}$$

where $\widetilde{L}_{\log}$ is defined by

$$\sqrt{\log \left(\Big\lceil C_1 n \xi \widehat{\Psi} \Gamma^2 + C_2 \Big\rceil n\hat{h} \Big(16 n \xi \prod_{l=1}^{L} s_l^2 \varepsilon \Psi \Big)^d + 1 \right)}.$$

Remark 4. *Similar to the linear case, the bound in Theorem 5 has $\max(1, d^{1-\frac{1}{p}-\frac{1}{r}})$ and $\sqrt{d}$ dependencies. The first is in $\widehat{\Psi}$, which arises from the mismatch of norms and can be avoided by simply picking the appropriate norm regularization (ℓ_p) on the weight matrices (W) as discussed above. The second $\sqrt{d}$ dependence in $\widetilde{L}_{\log}$. As discussed in the linear case, a projection on a low-dimensional represent space can help alleviate such dependence incurred by the complexity of the adversarial perturbation ball $\mathcal{B}(\varepsilon)$.*

Remark 5. *Theorem 5 provides generalization guarantees for adversarial metric learning in nonlinear case. The bounds in Yin et al. (2019) and Awasthi et al. (2021) apply only to a one-hidden-layer neural network. This contrasts with our bound, which applies to multi-layer networks. While the bounds in Khim and Loh (2018) and Mustafa, Lei and Klof (2022) apply to multi-layer networks, they are only applicable to pointwise learning and the single-sample perturbation case.*

Remark 6. *Similar to the linear case, Theorem 5 can recover the non-adversarial generalization bound (9) by setting $\varepsilon = 0$. The bound (9) is of the order $\mathcal{O}(\sqrt{d}\log(\hat{h})/\sqrt{n})$, where $\hat{h}$ is the width of the hidden layer. The generalization bound in Huai et al. (2019) grows as $\mathcal{O}(\sqrt{\hat{h}})$, while ours is $\mathcal{O}(\log(\hat{h}))$.*

$$\begin{aligned} &\mathcal{E}(f) - \mathcal{E}_n(f) \\ &\leq 3\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{8}{n^{3/2}} + \frac{C\xi \|X\|_{p^*,\infty}^2 L^2}{n} \prod_{l=1}^{L} s_l^2 \sum_{l=1}^{L} \frac{b_l^2}{s_l^2} \\ &\times \sqrt{\log \left(\lceil C_1 n \xi \Gamma^2 \|X\|_{p^*,\infty}^2 + C_2 \rceil n\hat{h} + 1 \right)} \log(n). \end{aligned} \tag{9}$$

4.3 Optimistic Bounds

Optimistic bounds have been studied in [Srebro *et al.*, 2010; Reeve and Kaban, 2020], where they have resulted in fast-rate generalization bounds for smooth losses under low-noise conditions. We aim to extend these approaches to adversarial metric learning. Our results are based on the *local Rademacher complexity* [Bartlett *et al.*, 2005] with respect to sample pairs [Cao *et al.*, 2016].

Definition 3 (Local Rademacher complexity)**.** *Let $\mathcal{H} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a hypothesis class. Given a sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of size n, The local Rademacher complexity is the worst-case Rademacher complexity of $\mathcal{H}$ on S of cardinality n, that is, $\mathfrak{R}_n(\mathcal{H}) := \sup_{|S|\leq n} \mathfrak{R}_S(\mathcal{H})$, where $\mathfrak{R}_S(\mathcal{H})$ is the empirical Rademacher complexity with respect to sample pairs defined by*

$$\mathfrak{R}_S(\mathcal{H}) = \frac{1}{\lfloor n/2 \rfloor} \mathbb{E}_\sigma \Big[\sup_{h\in\mathcal{H}} \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i h(x_i, x_{\lfloor n/2 \rfloor + i}) \Big],$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d Rademacher random variables with $\mathbb{P}\{\sigma_i = 1\} = \mathbb{P}\{\sigma_i = -1\} = \frac{1}{2}$.

Let $\widetilde{\mathcal{D}} := \{((x_i, \theta_i), y_i), ((x_j, \theta_j), y_j) \mapsto D_f(x_i + \theta_i, x_j + \theta_j) : f \in \mathcal{F}\}$. The local class $\widetilde{\mathcal{D}}|_\gamma = \{((x_i, \theta_i), y_i), ((x_j, \theta_j), y_j) \mapsto D_f(x_i + \theta_i, x_j + \theta_j) : f \in \mathcal{F}, \widetilde{\mathcal{E}}_n(f) \leq \gamma\} \subset \widetilde{\mathcal{D}}$ is defined as the set of function $D_f \in \widetilde{\mathcal{D}}$ with the adversarial empirical error at most γ. Similarly, the local adversarial loss class is defined as $\mathcal{L}_{adv}|_\gamma := \{(x_i, y_i), (x_j, y_j) \mapsto \max_{\theta_i, \theta_j \in \mathcal{B}(\varepsilon)} \ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j))) : f \in \mathcal{F}, \widetilde{\mathcal{E}}_n(f) \leq \gamma\}$.

We introduce the *self-bounding Lipschitz* [Reeve and Kaban, 2020] for the loss function, which assumes that the loss function is smooth. Srebro et al. (2010) show that such smoothness condition can give rise to an optimistic bound having a fast rate $\mathcal{O}(n^{-1})$ in the realisable case. Under smoothness assumption, we derive an upper bound on the local Rademacher complexity of adversarial loss class, which serves as a key step in developing fast-rate bounds.

Lemma 4. *Let $\mathcal{L}_{adv}|_\gamma$ and $\widetilde{\mathcal{D}}$ be defined as above. Suppose that for any $f \in \mathcal{F}$, $\|f\|_\infty \le B$, and the loss ℓ is (λ, η)-self-bounding Lipschitz bounded by b. Further let $\theta_i, \theta_j \mapsto \ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j)))$ be $\|\cdot\|$-Lipschitz with constant L and $\tilde{S} = \{((x_i, \theta_i), y_i) : i \in [n], \theta_i \in C_{\mathcal{B}}(\frac{3v}{\lambda(2\gamma)^\eta 8L})\}$. Suppose further that $n \mapsto \sqrt{n}\mathfrak{R}_{|\tilde{S}|}(\widetilde{\mathcal{D}})$ is non-decreasing. Then, we have*

$$\mathfrak{R}_n(\mathcal{L}_{adv}|_\gamma) \le \lambda(\gamma)^\eta \mathfrak{R}_{|\tilde{S}|}(\widetilde{\mathcal{D}})\sqrt{|\tilde{S}|/n}\Omega$$

where Ω grows at the order of

$$\mathcal{O}\Big(\log^{3/2}\Big(\frac{|\tilde{S}|}{\mathfrak{R}_{|\tilde{S}|}(\widetilde{\mathcal{D}})}\Big) - \log^{3/2}\Big(\frac{B^2|\tilde{S}|\lambda}{b^{1-\eta}}\Big)\Big)$$

The detailed proof is provided in *Appendix F.1*. Based on Lemma 4 and the sub-root property in [Bartlett *et al.*, 2005], fast generalization bounds adversarial metric learning with smooth losses are given in the following theorem.

Theorem 6. *With the above notation and assumption of Lemma 4, for all $f \in \mathcal{F}$, with probability at least $1 - \delta$, we have*

$$\begin{aligned}\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f) \le\ & 106\lambda^2\mathfrak{R}^2_{|\tilde{S}|}(\widetilde{\mathcal{D}})\Omega^2|\tilde{S}|/n \\ & + \frac{48b}{n}(\log(1/\delta) + \log(\log(n))) \\ & + \sqrt{\widetilde{\mathcal{E}}_n(f)\Big(8\lambda^2\mathfrak{R}^2_{|\tilde{S}|}(\widetilde{\mathcal{D}})\Omega^2|\tilde{S}|/n + K\Big)}\end{aligned}$$

where $K = \frac{4b}{n}(\log(1/\delta) + \log(\log(n)))$.

Remark 7. *The convergence rate of the generalization bound in Theorem 6 grows as $\mathfrak{R}^2_{|\tilde{S}|}(\widetilde{\mathcal{D}})$. For the majority of function classes (e.g., linear models [Yin* et al.*, 2019]), the Rademacher complexity is at least $\mathcal{O}(n^{-1/2})$. The second term would then grow as $\mathcal{O}(n^{-1})$, while the fourth term would grow at the usual $\mathcal{O}(n^{-1/2})$ rate. However, if $\widetilde{\mathcal{E}}_n(f) = 0$, the fourth term vanishes, thus achieving a fast rate of convergence at least $\mathcal{O}(n^{-1})$.*

Remark 8. *We establish the fast-rate generalization bound with high probability for metric learning in non-adversarial case,*

$$\begin{aligned}\mathcal{E}(f) \le\ & \mathcal{E}_n(f) + 106\lambda^2\mathfrak{R}^2_n(\mathcal{D})\hat{\Omega}^2 + \frac{48b}{n}(\log(1/\delta) \\ & + \log(\log(n))) + \sqrt{\mathcal{E}_n(f)\Big(8\lambda^2\mathfrak{R}^2_n(\mathcal{D})\hat{\Omega}^2 + K\Big)},\end{aligned}$$

where $\mathcal{D}$ is the class of distance function and $\hat{\Omega}$ grows at the order $\mathcal{O}(\log^{3/2}(\frac{n}{\mathfrak{R}_n(\mathcal{D})}) - \log^{3/2}(\frac{B^2 n\lambda}{b^{1-\eta}}))$. It extends the previous optimistic results [Srebro et al.*, 2010] of pointwise learning to pairwise learning. In contrast with the bounds of order $\mathcal{O}(n^{-1/2})$ [Cao* et al.*, 2016; Huai* et al.*, 2019; Lei* et al.*, 2020], this is the improved result.*

Dataset	Size (n)	Dimension (d)
Wine	178	13
Spambase	4601	58
MNIST	70000	784
CIFAR-10	60000	3072

Table 2: The details of the adopted datasets.

5 Experiments

5.1 Experimental Setup

Datasets. We adopt the following real-world datasets for experiments: the Wine[1], Spambase[2], MNIST[3] and CIFAR-10[4] datasets. Table 2 provides details of dimension (d) and size (n). Note that the input for adversarial metric learning models is a set of sample pairs rather than single samples. For the original dataset with single samples, we make one-by-one matching to form $n(n-1)$ sample pairs, and then randomly select n pairs to construct the new dataset. The pair composed of samples with the same category is assigned to label 1, and the other is assigned to 0. We randomly split the new dataset into training, validation and test sets with a ratio of $6:2:2$, where the validation set is used for early stopping to prevent overfitting of model.

Model and Attack Settings. We use one-layer neural networks without nonlinear activation as the linear model. Denote the number of units in the output layer by d'. We utilize five-layer feed-forward neural networks with ReLU activation [Hahnloser *et al.*, 2000] as the non-linear model, where the number of the units in each layer is $(512, 256, 128, 64, d')$. All models trained with the Adam optimizer. The learning rates of the linear model and the nonlinear model are set as $1e-2$ and $1e-3$, respectively.

We apply ℓ_∞ PGD attack [Madry *et al.*, 2018] adversarial training to minimize the following objective function

$$\min_{f_W}\sum_{i\ne j}\max_{\theta_i,\theta_j\in\mathcal{B}(\varepsilon)}\ell(\tau(y_i,y_j)(1-D_{f_W}(x_i+\theta_i,x_j+\theta_j)))+\lambda\|W^1\|_1,$$

where $\ell(\cdot)$ is cross entropy loss, f_W is the mapping function parameterized by $W = \{W^l \in \mathbb{R}^{h_l\times h_{l-1}}\}_{l=1}^L$, where h_l is the number of neurons in the l-th layer of the network (especially, $h_0 = d, h_L = d'$), and $\lambda \ge 0$ is the regularization parameter. Then, we run PGD attack to check the generalization error. Similar to Yin et al. (2019), the generalization error is approximately calculated by

$$|\text{adversarial\_train\_accuracy} - \text{adversarial\_test\_accuracy}|. \tag{10}$$

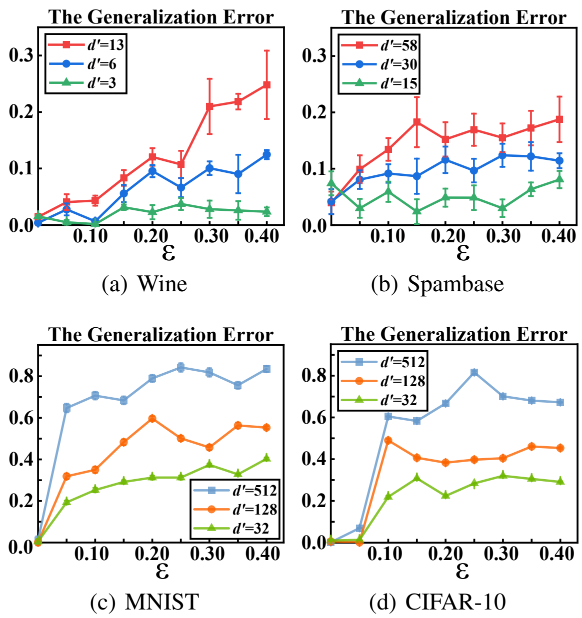
During the training and test phases, the adversarial samples are generated by PGD algorithm with step size $\varepsilon/5$, where ε is the maximum magnitude of the allowed perturbations that varies in $\{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$. Overall, each experiment is independently repeated 10 times, and average generalization error with standard deviation of adversarial metric learning models are reported.

[1]https://archive.ics.uci.edu/ml/datasets/wine/

[2]https://archive.ics.uci.edu/ml/datasets/spambase/

[3]http://yann.lecun.com/exdb/mnist/

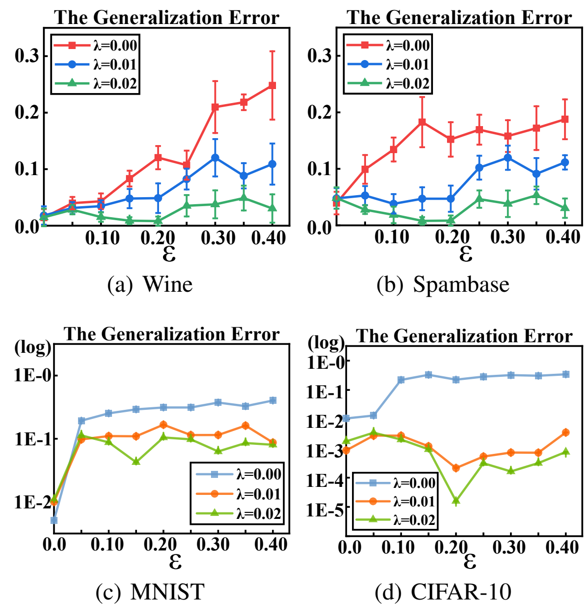
[4]https://www.kaggle.com/competitions/cifar-10/data



(a) Wine (b) Spambase

(c) MNIST (d) CIFAR-10

Figure 1: **The generalization error (10) (mean and standard deviation over 10 runs) for different dimensional outputs (d') of adversarial metric learning models.** ε denotes the perturbation bound. Subfigures (a) and (b) present results for linear models. Subfigures (c) and (d) present results for nonlinear models on adopted datasets.



(a) Wine (b) Spambase

(c) MNIST (d) CIFAR-10

Figure 2: **The generalization error (10) (mean and standard deviation over 10 runs) for adversarial metric learning models with ($\lambda \neq 0$) and without ($\lambda = 0$) L_1 regularization.** λ denotes regularization parameter, and ε denotes the perturbation bound. Subfigures (a) and (b) present results for linear models. Subfigures (c) and (d) present results for nonlinear models on adopted datasets.

5.2 Experiment Results

Theorem 3 and 5 suggest that projecting the input feature to low-dimensional output and applying appropriate regularization to the weights of models, can reduce the generalization error of adversarial metric learning models. Here, we conduct linear and non-linear experiments to validate these theoretical findings.

The Effect of the Output Dimension. To investigate the effect of the output dimension on the generalization performance, we train models with different dimensional outputs. In the linear case, we consider three cases where the output dimension (*i.e.*, d') is set as d, $\lceil d/2 \rceil$ and $\lceil d/3 \rceil$, respectively. For the nonlinear case, the output dimension of the final layer of neural networks is set as 512, 128 and 32, respectively. Figure 1 plots the generalization errors (10) of linear and nonlinear models on the adopted datasets. As we can see, the fewer the output features, the smaller the generalization error, which suggests that projecting input into the low-dimension feature space can potentially reduce the generalization gap of adversarial metric learning models.

The Effect of Regularization. We evaluate the effect of the weight parameters on the generalization of adversarial metric learning models by comparing the performance of the models with and without regularization. We employ linear and nonlinear models with output dimensions d and 32, respectively, and apply the L_1 regularization to W^1 (*i.e.*, the weights of the first layer of models). The regularization parameter λ is set as 0, 0.01 and 0.02. Note that $\lambda = 0$ indicates the model trained without regularization. The generalization errors (10) of linear and nonlinear models on the adopted datasets are presented in Figure 2. We can see that generalization gap of the model with regularization is smaller than that of the model without regularization, thus we conclude that applying L_1-norm regularization to adversarial metric learning models is helpful for reducing generalization error.

6 Conclusions

This paper presents a detailed study of the generalization properties of adversarial metric learning under ℓ_r adversarial perturbations. We derive the high-probability generalization bounds for adversarial metric learning with pairwise perturbations by developing the uniform convergence analysis techniques. Our results apply to both linear and deep metric learning models, as well as to various loss functions. To our knowledge, this is the first generalization analysis for adversarial pairwise learning with pairwise perturbations. We further extended our analysis to the case of smooth losses, and establish a fast generalization bound at a rate of $\mathcal{O}(n^{-1})$ by the local Rademacher complexity. In future work, we will investigate the generalization properties of models under non-additive attacks.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant No. 12071166, the Fundamental Research Funds for the Central Universities of China under Grant Nos. 2662021JC008, 2662022XXYJ005, and HZAU-AGIS Cooperation Fund No. SZYJY2023010.

References

[Awasthi *et al.*, 2021] Pranjal Awasthi, George Yu, Chun-Sung Ferng, Andrew Tomkins, and Da-Cheng Juan. Adversarial robustness across representation spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7608–7616, 2021.

[Bartlett *et al.*, 2005] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[Bartlett *et al.*, 2017] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

[Bouniot *et al.*, 2020] Quentin Bouniot, Romaric Audigier, and Angelique Loesch. Vulnerability of person re-identification models to metric adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 794–795, 2020.

[Bousquet *et al.*, 2020] Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626, 2020.

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

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proceedings of 2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[Chen and Deng, 2019] Binghui Chen and Weihong Deng. Energy confused adversarial metric learning for zero-shot image retrieval and clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8134–8141, 2019.

[Dai *et al.*, 2018] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *Proceedings of International Joint Conference on Artificial Intelligence*, volume 1, page 6, 2018.

[Farnia and Ozdaglar, 2021] Farzan Farnia and Asuman Ozdaglar. Train simultaneously, generalize better: Stability of gradient-based minimax learners. In *International Conference on Machine Learning*, pages 3174–3185. PMLR, 2021.

[Hahnloser *et al.*, 2000] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, 2000.

[Huai *et al.*, 2019] Mengdi Huai, Hongfei Xue, Chenglin Miao, Liuyi Yao, Lu Su, Changyou Chen, and Aidong Zhang. Deep metric learning: The generalization analysis and an adaptive algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2535–2541, 2019.

[Huai *et al.*, 2022] Mengdi Huai, Tianhang Zheng, Chenglin Miao, Liuyi Yao, and Aidong Zhang. On the robustness of metric learning: An adversarial perspective. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(5):1–25, 2022.

[Huang *et al.*, 2019] He Huang, Changhu Wang, Philip S Yu, and Chang-Dong Wang. Generative dual adversarial network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 801–810, 2019.

[Khim and Loh, 2018] Justin Khim and Po-Ling Loh. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

[Klochkov and Zhivotovskiy, 2021] Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34:5065–5076, 2021.

[Kurakin *et al.*, 2018] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. 2018.

[Lei *et al.*, 2020] Yunwen Lei, Antoine Ledent, and Marius Kloft. Sharper generalization bounds for pairwise learning. *Advances in Neural Information Processing Systems*, 33:21236–21246, 2020.

[Li and Liu, 2021] Shaojie Li and Yong Liu. High probability generalization bounds with fast rates for minimax problems. In *International Conference on Learning Representations*, 2021.

[Liu *et al.*, 2022] Deyin Liu, Lin Wu, Richang Hong, Zongyuan Ge, Jialie Shen, Farid Boussaid, and Mohammed Bennamoun. Generative metric learning for adversarially robust open-world person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2022.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of International Conference on Learning Representations*, 2018.

[Mo *et al.*, 2022] Yingxiang Mo, Hong Chen, Yuxiang Han, and Hao Deng. Error bounds of adversarial bipartite ranking. *Neurocomputing*, 478:81–88, 2022.

[Mohri *et al.*, 2018] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[Mustafa *et al.*, 2022] Waleed Mustafa, Yunwen Lei, and Marius Kloft. On the generalization analysis of adversarial learning. In *International Conference on Machine Learning*, pages 16174–16196, 2022.

[Reeve and Kaban, 2020] Henry Reeve and Ata Kaban. Optimistic bounds for multi-output learning. In *International Conference on Machine Learning*, pages 8030–8040, 2020.

[Srebro *et al.*, 2010] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in Neural Information Processing Systems*, 23, 2010.

[Tu *et al.*, 2019] Zhuozhuo Tu, Jingwei Zhang, and Dacheng Tao. Theoretical analysis of adversarial learning: A minimax approach. *Advances in Neural Information Processing Systems*, 32, 2019.

[Vapnik, 1999] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 1999.

[Wang *et al.*, 2020] Zhuoyi Wang, Yigong Wang, Bo Dong, Sahoo Pracheta, Kevin Hamlen, and Latifur Khan. Adaptive margin based deep adversarial metric learning. In *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 100–108, 2020.

[Xiao *et al.*, 2022] Jiancong Xiao, Yanbo Fan, Ruoyu Sun, Jue Wang, and Zhi-Quan Luo. Stability analysis and generalization bounds of adversarial training. In *Advances in Neural Information Processing Systems*, 2022.

[Xing *et al.*, 2021] Yue Xing, Qifan Song, and Guang Cheng. On the algorithmic stability of adversarial training. *Advances in Neural Information Processing Systems*, 34:26523–26535, 2021.

[Xu *et al.*, 2019] Xing Xu, Li He, Huimin Lu, Lianli Gao, and Yanli Ji. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web*, 22(2):657–672, 2019.

[Yang *et al.*, 2021] Xiaochen Yang, Mingzhi Dong, Yiwen Guo, and Jing-Hao Xue. Metric learning for categorical and ambiguous features: An adversarial method. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 223–238, 2021.

[Ye *et al.*, 2019] Han-Jia Ye, De-Chuan Zhan, and Yuan Jiang. Fast generalization rates for distance metric learning. *Machine Learning*, 108(2):267–295, 2019.

[Yin *et al.*, 2019] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pages 7085–7094, 2019.

[Zantedeschi *et al.*, 2016] Valentina Zantedeschi, Rémi Emonet, and Marc Sebban. Metric learning as convex combinations of local models with generalization guarantees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1478–1486, 2016.

[Zhang, 2002] Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.