Generalization Bounds for Adversarial Metric Learning

Wen Wen¹, Han Li¹, Hong Chen¹,²,³, Rui Wu⁴, Lingjuan Wu¹, Liangxuan Zhu¹
¹College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
²Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan 430070, China
³Key Laboratory of Smart Farming for Agricultural Animals, Wuhan 430070, China
⁴Horizon 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 $O(n^{-\frac{1}{4}})$ 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 $\ell_\infty$ 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 $\ell_r$-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 $O(n^{-\frac{1}{4}})$ 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].
theoretical aspects of adversarial robustness has not been ex-

sibly robust metric learning (see samples, adversarial learning methods are proposed to facil-

samples are limited in their capacity to distinguish ambiguous

Since metric learning methods learn on the original

Huang et al. (2019], since metric learning methods learn on the original

re-identification [Dai et al., 2022] to zero-shot learning [Chen and Deng, 2019;

Xu et al., 2018; Bouniot et al., 2020], and cross-modal retrieval [Xu et al., 2019].

Since metric learning methods learn on the original

samples are affiliated to the same class,

i.e. $y_i = y_j$, and $\tau(y_i, y_j) = -1$, otherwise.

However, in the presence of adversaries, there will be im-

perceptible perturbations on the input samples that lead to

Table 1: Summary of generalization analysis for adversarial learning (•-optimization bound; •-generalization bound in expectation).

<table>
<thead>
<tr>
<th>Perturbation Object</th>
<th>Task</th>
<th>Reference</th>
<th>Perturbation Way</th>
<th>Analysis Tool</th>
<th>Learning Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Single sample</td>
<td>Tu et al. (2019)</td>
<td>$\ell_p$-norm additive</td>
<td>Rademacher complexity</td>
<td>$O(1/\sqrt{n})$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mustafa (2022)</td>
<td>Spatial transformation</td>
<td>Local Rademacher complexity</td>
<td>$O(1/n)$</td>
</tr>
<tr>
<td>Optimization</td>
<td>Optimization</td>
<td>Xiao et al. (2022)</td>
<td>$\ell_p$-norm additive</td>
<td>Algorithmic stability</td>
<td>$O(1/n)$</td>
</tr>
<tr>
<td>Bipartite Ranking</td>
<td>Sample pair</td>
<td>Mo et al. (2022)</td>
<td>$\ell_p$-norm additive</td>
<td>Rademacher complexity</td>
<td>$O(1/\sqrt{n})$</td>
</tr>
</tbody>
</table>

Table 2: Summary of generalization analysis for adversarial learning (•-optimization bound; •-generalization bound in expectation).

<table>
<thead>
<tr>
<th>Perturbation Object</th>
<th>Task</th>
<th>Reference</th>
<th>Perturbation Way</th>
<th>Analysis Tool</th>
<th>Learning Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Single sample</td>
<td>Tu et al. (2019)</td>
<td>$\ell_p$-norm additive</td>
<td>Rademacher complexity</td>
<td>$O(1/\sqrt{n})$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mustafa (2022)</td>
<td>Spatial transformation</td>
<td>Local Rademacher complexity</td>
<td>$O(1/n)$</td>
</tr>
<tr>
<td>Optimization</td>
<td>Optimization</td>
<td>Xiao et al. (2022)</td>
<td>$\ell_p$-norm additive</td>
<td>Algorithmic stability</td>
<td>$O(1/n)$</td>
</tr>
<tr>
<td>Bipartite Ranking</td>
<td>Sample pair</td>
<td>Mo et al. (2022)</td>
<td>$\ell_p$-norm additive</td>
<td>Rademacher complexity</td>
<td>$O(1/\sqrt{n})$</td>
</tr>
</tbody>
</table>

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
$\ell_p$-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 $P$, 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$, 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

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

where $\ell : \mathbb{R} \rightarrow \mathbb{R}_+$ 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 im-
perceptible perturbations on the input samples that lead to
maximizing empirical risk (2). Throughout this paper, we assume that the perturbation $\theta$ is adversarially chosen in the $\ell_r$-ball $B(\varepsilon) \subseteq \mathbb{R}^d$ of radius $\varepsilon$, 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 $\theta_i^*$ and $\theta_j^*$ by [Huai et al., 2022]

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

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

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

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

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

and the adversarial expected risk $\bar{E}(f)$

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

The adversarial empirical risk $\bar{E}_n(f)$ 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 $\bar{E}(f)$ measures how well $f$ generalizes to unseen adversarial samples. In this paper, we are interested in the difference between $\bar{E}_n(f)$ and $\bar{E}(f)$. Our main tool for bounding the generalization error for adversarial metric learning (i.e., $\bar{E}(f) - \bar{E}_n(f)$) is the $\ell_\infty$-covering number defined below.

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

$$
\sup_{a \in A, c \in C} \|a - c\|_\infty \leq v.
$$

Then, the $\ell_\infty$-covering number of $A$ is the minimum cardinality of any subset covers $A$ at scale $v$, denoted as $N_{\ell_\infty}(v, A)$.

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

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

We have the following relationship between the generalization error (i.e., $\bar{E}(f) - \bar{E}_n(f)$) and $\ell_\infty$-covering number of adversarial loss class $\mathcal{L}_{adv}$ on the training sample $S$ (i.e., $N_{\ell_\infty}(\mathcal{L}_{adv}, v, S)$), 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 $f \in \mathcal{F}$

$$
\bar{E}(f) - \bar{E}_n(f) \leq 3 \sqrt{\frac{\log(2/\delta)}{2n}} + \inf_{\alpha > 0} \left\{ \frac{8n}{\sqrt{n}} + \frac{24}{n} \int_{\alpha}^{\sqrt{\pi}} \log N_{\ell_\infty}(\mathcal{L}_{adv}, v, S)dv \right\}.
$$

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

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

and incorporate perturbations $\theta_i, \theta_j$ 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 $\ell : \mathcal{F} \rightarrow \mathbb{R}$ and a distance function $D_f : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ parametrized by $f \in \mathcal{F}$,

1) the loss function $\ell$ is $\xi$-Lipschitz if, for all $f, f' \in \mathcal{F}$

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

2) the distance function $D_f$ is the Multi-variate Lipschitz continuity if, for all $\theta_i, \theta_j, \theta'_i, \theta'_j \in \mathbb{R}^d$

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

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 $\theta_i, \theta_j$ by $\ell((\tau(y_i), y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j))$ 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 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 $\text{C}_{\text{BG}}(3v/4\text{AL})$ be a $3v/4\text{AL}$-cover of $\mathcal{B}(\varepsilon)$, and define the adversarial sample set

$$
\hat{S} = \{(x_i, \theta_i), y_i) : i \in [n], \theta_i \in \text{C}_{\text{BG}}(3v/4\text{AL})\}.
$$

Then, we have

$$
N_{\ell_\infty}(\mathcal{L}_{adv}, v, S) \leq N_{\ell_\infty}(\tilde{\mathcal{L}}_{adv}, v/4, \hat{S}).
$$

Detailed proofs are contained in Appendix C.2. Theorem 1 illustrates that the $\ell_\infty$ covering number of adversarial loss class $\mathcal{L}_{adv}$ on set $S$ can be bounded by the $\ell_\infty$ covering number of class $\tilde{\mathcal{L}}_{adv}$ on adversarial set $\hat{S}$, 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 $\ell_r$-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 Metric Learning Model

We consider the following linear hypothesis class:

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

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

$$D_f(x_i, x_j) = (x_i - x_j)^T W^T W (x_i - x_j). \quad (6)$$

The Lipschitz constant of the function $\theta_i, \theta_j \mapsto \ell(\tau(y_i, y_j))$ is $4\Lambda^2\Psi$-Lipschitz, where $4\Lambda^2\Psi$ is the Multi-variate Lipschitz constant of the function $\theta_i, \theta_j \mapsto \ell(\tau(y_i, y_j))$, and $\Psi$ is max$(1, d^{1-\frac{1}{p} - \frac{1}{2}})$($\|X\|_{r,\infty} + \varepsilon$).

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 $\mathcal{L}_{adv}$ on the sample $\tilde{S}$ in the theorem below.

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

$$\log \mathcal{N}_{\infty}(\mathcal{L}_{adv}, \nu/4, \mathcal{S}) \leq C\frac{\nu^2}{\tilde{\psi}^2} L_{log},$$

where

$$L_{log} = \log \left(4 \left[\frac{32\xi \Lambda^2 \tilde{\psi}^2}{\nu} + 1\right] n \left(\frac{16\xi \Lambda^2 \tilde{\psi}}{\nu}\right)^{d} + 1\right),$$

$$\tilde{\psi} = \max(1, d^{1-\frac{1}{p} - \frac{1}{2}})\left(\|X\|_{r,\infty} + \varepsilon\right)^2$$

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 $f \in \mathcal{L}$, with probability at least $1 - \delta$, we have

$$\mathbb{E}(f) - \mathbb{E}_n(f) \leq 3\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{8}{\nu^{d/2}} + \frac{\xi \Lambda^2 \tilde{\psi} \log(n)}{n} \times \log \left(4 \left[\frac{32\xi \Lambda^2 \|X\|_{2,\infty}^2 + 1}\right] n \left(16\xi \Lambda^2 \tilde{\psi}\right)^{d} + 1\right),$$

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

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

Remark 1. The generalization bound in Theorem 3 suffers from additional dimension dependent terms as compared to its non-adversarial counterpart. The first $d^{1-\frac{1}{p} - \frac{1}{2}}$ dependence in $\tilde{\psi}$ is due to the mismatch between the norm on the input $x$ and the norm in the ball $B(\varepsilon)$. Indeed, we have used the inequality $\|x_i + \theta_i\|_{p^*} \leq \max(1, d^{1-\frac{1}{p} - \frac{1}{2}})[\|x_i\|_{p} + \|\theta_i\|_{p}],$ in Awasthi et al. (2021), where $1 - 1/p = 1/p^*$. If $\frac{1}{p} + \frac{1}{p^*} \geq 1$, $\tilde{\psi}$ 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 Theorem 3, is attributed to the complexity of the perturbation ball $B(\varepsilon)$. For example, if $B(\varepsilon)$ is contained in a low dimensional space $d' < d$, the dependence is reduced to $O(\sqrt{d'})$. 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 $\ell_\infty$-norm perturbations, we consider the general case where the perturbations are measured in $\ell_r$-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 $O(1/\sqrt{n})$ is similar to the generalization bounds in Cao et 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.

Theorem 3 is a high-probability generalization bound for adversarial metric learning in the linear case.

4.2 Deep Metric Learning Model

Let the mapping $f$ be a $L$-layer neural network parameterized 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

$$f(x_i) = W_d^T \rho(W_{L-1}^T \rho(\cdots \rho(W_1^T x_i))),$$

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

$$\mathcal{F} := \{x_i \mapsto f(x_i) : f \in \mathcal{F}, \|W_l\|_{F} \leq b_l, \|W_l\|_{\sigma} \leq s_l\},$$

where $\|\cdot\|_{F}$ represents spectral norm and $\|\cdot\|_{\sigma}$ denotes the Frobenius norm.
As with the linear case, we first establish the Lipschitzness of the function $\theta_i, \theta_j \to \ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j)))$. The results are summarized in the following lemma.

**Lemma 3.** Let $f \in F$ 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 $f \in F$, the function $\ell(\tau(y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j)))$ is $4L_i \prod_{l=1}^{L} s^2_l \Psi$-Lipschitz in $\theta_i, \theta_j$, where $4L_i \prod_{l=1}^{L} s^2_l \Psi$ is the Multi-variate Lipschitz constant of the function $\theta_i, \theta_j \to D_f(x_i + \theta_i, x_j + \theta_j)$, and $\Psi$ is $\max(1, d^{1 - \frac{1}{\gamma} - \frac{1}{\epsilon}})\|X\|_{\infty, \infty} + 3\epsilon$.

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

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

$$\log \mathcal{N}_\infty(\mathcal{L}_{adv}, \nu/4, \tilde{S}) \leq \frac{C \Psi^2 \tilde{\Psi}^2 L^4}{\nu^2} \prod_{l=1}^{L} s^2_l \left( \sum_{l=1}^{L} \frac{b^2_l}{s^2_l} \right)^2 L_{log},$$

where

$$L_{log} = \log \left( \frac{C_1 \Psi^2 \tilde{\Psi}^2}{\nu} + C_2 \hat{h} \left( \frac{16 \Psi L \prod_{l=1}^{L} s^2_l \epsilon \Psi}{\nu} \right)^{\frac{d}{2}} + 1 \right),$$

$\Psi = \max(1, d^{1 - \frac{1}{\gamma} - \frac{1}{\epsilon}}\|X\|_{\infty, \infty} + 3\epsilon)$, $\tilde{\Psi} = \max(1, d^{1 - \frac{1}{\gamma} - \frac{1}{\epsilon}}\|X\|_{\infty, \infty} + 3\epsilon)^2$, $\hat{\Gamma} = \max_{\mathcal{X} \in \tilde{S}} (\prod_{l=1}^{L} s^2_l \epsilon \Psi)$, $\hat{h} = \max_{\mathcal{X} \in \tilde{S}} h$, and $C_1, C_2$ are universal constants.

The proof is given in Appendix F. 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 $f \in F$, with probability at least $1 - \delta$, we have

$$\tilde{\mathcal{E}}(f) - \tilde{\mathcal{E}}_n(f) \leq 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}} + 8 \frac{\log(n)}{\sqrt{n}} + \frac{C \Psi L^2}{n} \prod_{l=1}^{L} s^2_l \left( \sum_{l=1}^{L} \frac{b^2_l}{s^2_l} \right) \log(n)L_{log},$$

where $L_{log}$ is defined by

$$L_{log} = \log \left( \left( \frac{C_1 \Psi^2 \tilde{\Psi}^2}{\nu} + C_2 \hat{h} \left( \frac{16 \Psi L \prod_{l=1}^{L} s^2_l \epsilon \Psi}{\nu} \right)^{\frac{d}{2}} + 1 \right) \right).$$

**Remark 4.** Similar to the linear case, the bound in Theorem 5 has $\max(1, d^{1 - \frac{1}{\gamma} - \frac{1}{\epsilon}}\|X\|_{\infty, \infty})$ and $\sqrt{\hat{\epsilon}}$ dependencies. The first is in $\tilde{\Psi}$, which arises from the mismatch of norms and can be avoided by simply picking the appropriate norm regularization ($\gamma_p$) on the weight matrices ($W$) as discussed above. The second $\sqrt{\hat{\epsilon}}$ dependence in $L_{log}$. 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 $\mathcal{B}(\epsilon)$.

**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 $\epsilon = 0$. The bound (9) is of the order $O(\sqrt{\hat{\epsilon}} \log(\hat{h})/\sqrt{n})$, where $\hat{h}$ is the width of the hidden layer. The generalization bound in Huai et al. (2019) grows as $O(\sqrt{\hat{h}})$, while ours is $O(\log(\hat{h}))$.

$$\mathcal{E}(f) - \mathcal{E}_n(f) \leq 3\sqrt{\frac{\log(2/\delta)}{2n}} + 8 \log(n)\sqrt{n} + \frac{C \log(L_n)}{n} L^2 \prod_{l=1}^{L} s^2_l \left( \sum_{l=1}^{L} \frac{b^2_l}{s^2_l} \right) \log(n) \hat{h} + 1 \log(n).$$

### 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 $\mathcal{H} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 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 $\mathcal{H}$. Let $S$ be sample with cardinality $n$, that is, $\mathcal{R}_S(\mathcal{H}) := \sup_{h \in \mathcal{H}} \mathcal{R}_S(\mathcal{H})$, where $\mathcal{R}_S(\mathcal{H})$ is the empirical Rademacher complexity with respect to sample pairs defined by

$$\mathcal{R}_S(\mathcal{H}) = \frac{1}{[n/2]} \mathcal{E}_n \left( \sup_{h \in \mathcal{H}} \sum_{i=1}^{[n/2]} \sigma_i h(x_i, x_{[n/2]+i}) \right),$$

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

Let $\bar{D} := \{((x_i, \theta_i), (x_j, \theta_j)) \in \mathcal{X}^2 : ((x_i, \theta_i), (x_j, \theta_j)) \in \mathcal{F}, \mathcal{E}(f) \leq \gamma \}$ be a subset of the class of function $D_f \in \hat{D}$ with the adversarial empirical error at most $\gamma$. Similarly, the local adversarial loss class is defined as $\mathcal{L}_{adv}(\gamma) := \{(x_i, \theta_i), (x_j, \theta_j) \in \mathcal{X}^2 : \mathcal{E}(f) \leq \gamma \}$. Let $\mathcal{D} = \{((x_i, \theta_i), (x_j, \theta_j)) \in \mathcal{X}^2 : ((x_i, \theta_i), (x_j, \theta_j)) \in \mathcal{F}, \mathcal{E}(f) \leq \gamma \}$. 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 $O(n^{-1})$ in the realisable case.
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 $\mathcal{L}_{adv}$ be defined as above. Suppose that for any $f \in \mathcal{F}$, $\|f\|_\infty \leq B$, and the loss $\ell$ is $(\lambda, \eta)$-self-bounding Lipschitz bounded by $b$. Further let $\theta_i, \theta_j \mapsto \ell((y_i, y_j)(1 - D_f(x_i + \theta_i, x_j + \theta_j)))$ be $\|\cdot\|$-Lipschitz with constant $L$ and $\bar{S} = \{(x_i, \theta_i) : i \in [n], \theta_i \in C_B(\frac{3b}{L})\}$. Suppose further that $n \rightarrow \sqrt{n}R_{\bar{S}}(\bar{D})$ is non-decreasing. Then, we have

$$R_n(\mathcal{L}_{adv}|\gamma) \leq \lambda(\gamma)^nR_{\bar{S}}(\bar{D})\sqrt{|\bar{S}|}/n\Omega$$

where $\Omega$ grows at the order of

$$O\left(\log^{3/2}\left(\frac{|\bar{S}|}{R_{\bar{S}}(\bar{D})}\right) - \log^{3/2}\left(\frac{B^2|\bar{S}|\lambda}{b^2-\eta}\right)\right)$$

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

$$\mathcal{E}(f) - \mathcal{E}_n(f) \leq 106\lambda^2R_{\bar{S}}(\bar{D})\Omega^2|\bar{S}|/n$$

$$+ \frac{48b}{n}(\log(1/\delta) + \log(\log(n)))$$

$$+ \sqrt{\mathcal{E}_n(f)\left(8\lambda^2R_{\bar{S}}(\bar{D})\Omega^2|\bar{S}|/n + K\right)}$$

where $K = \frac{8b}{n}(\log(1/\delta) + \log(\log(n)))$.

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

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

$$\mathcal{E}(f) \leq \mathcal{E}_n(f) + 106\lambda^2R_{\bar{S}}(\bar{D})\Omega^2 + \frac{48b}{n}(\log(1/\delta))$$

$$+ \log(\log(n))) + \sqrt{\mathcal{E}_n(f)\left(8\lambda^2R_{\bar{S}}(\bar{D})\Omega^2 + K\right)},$$

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

### 5 Experiments

#### 5.1 Experimental Setup

**Datasets.** We adopt the following real-world datasets for experiments: the Wine1, Spambase2, MNIST3 and CIFAR-104 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 non-linear 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 $\ell_\infty$ PGD attack [Madry et al., 2018] adversarial training to minimize the following objective function

$$\min_{f \in W} \sum_{i \neq j, \theta_i, \theta_j \in B(\epsilon)} \ell((y_i, y_j)(1 - D_{f\theta}(x_i + \theta_i, x_j + \theta_j))) + \lambda||W||_1,$$

where $\ell(\cdot)$ is cross entropy loss, $f_{\theta}$ 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 $\lambda \geq 0$ 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

$$|\text{adversarial_train_accuracy} - \text{adversarial_test_accuracy}|.$$

During the training and test phases, the adversarial samples are generated by PGD algorithm with step size $\epsilon/5$, where $\epsilon$ 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size $(n)$</th>
<th>Dimension $(d)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine</td>
<td>178</td>
<td>13</td>
</tr>
<tr>
<td>Spambase</td>
<td>4601</td>
<td>58</td>
</tr>
<tr>
<td>MNIST</td>
<td>70000</td>
<td>784</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>60000</td>
<td>3072</td>
</tr>
</tbody>
</table>

Table 2: The details of the 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, [d/2] and [d/3], 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 $\lambda$ 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 $\ell_p$ 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 $O(n^{-1})$ 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. SYJY2023010.
References


