Towards Sharper Generalization Bounds for Adversarial Contrastive Learning

Wen Wen[1] , **Han Li**[1,2,3*] , **Tieliang Gong**[4] , **Hong Chen**[1,2,3]
[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]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
lihan125@mail.hzau.edu.cn

Abstract

Recently, the enhancement on the adversarial robustness of machine learning algorithms has gained significant attention across various application domains. Given the widespread label scarcity issue in real-world data, adversarial contrastive learning (ACL) has been proposed to adversarially train robust models using unlabeled data. Despite the empirical success, its generalization behavior remains poorly understood and far from being well-characterized. This paper aims to address this issue from a learning theory perspective. We establish novel high-probability generalization bounds for the general Lipschitz loss functions. The derived bounds scale $\mathcal{O}(\log(k))$ with respect to the number of negative samples k, which improves the existing linear dependency bounds. Our results are generally applicable to many prediction models, including linear models and deep neural networks. In particular, we obtain an optimistic generalization bound $\mathcal{O}(1/n)$ under the smoothness assumption of the loss function on the sample size n. To the best of our knowledge, this is the first fast-rate bound valid for ACL. Empirical evaluations on real-world datasets verify our theoretical findings.

1 Introduction

Contrastive learning (CL), which learns generalizable feature representations from unlabeled data [Grill *et al.*, 2020; Chen *et al.*, 2020; Khosla *et al.*, 2020], has demonstrated superior abilities in self-supervised learning [Wu *et al.*, 2018; Misra and Maaten, 2020]. Despite the state-of-the-art performance, the vulnerability of learning models to small input perturbations [Hendrycks and Dietterich, 2019] has raised safety concerns when implementing them to high-risk applications [Lin *et al.*, 2019; Ma *et al.*, 2021]. These perturbations are capable of completely altering the models' decisions, commonly referred to as adversarial perturbations [Xu *et al.*, 2019; Zhang *et al.*, 2023].

Adversarial contrastive learning (ACL) has emerged as a prominent approach to improve the model's robustness against adversarial perturbations. Its fundamental concept involves addressing a min-max optimization problem by injecting adversarial samples into the CL objective [Gowal *et al.*, 2020; Jiang *et al.*, 2020]. The remarkable empirical effectiveness of ACL motivates the need for a theoretical understanding of the adversarially robust generalization of learned representations.

Theoretical analysis of ACL presents more challenges as compared to its non-adversarial counterpart [Arora *et al.*, 2019; Lei *et al.*, 2023]. The maximization operator over perturbations of adversarial learning [Xing *et al.*, 2021; Xiao *et al.*, 2022] introduces difficulties in measuring the generalization capability of the hypothesis class. Several previous work on this problem considers an approximate smoothness [Xiao *et al.*, 2022] or surrogate adversarial loss [Yin *et al.*, 2019; Khim and Loh, 2018], and then provides generalization guarantees. However, these classical results is invalid for ACL due to the inconsistency of analytical framework. They assume that inputs and true labels are known as prior knowledge, while the input label is not available to the learner of ACL, which makes the complexity estimation of hypothesis space more difficult than [Xiao *et al.*, 2022; Yin *et al.*, 2019; Awasthi *et al.*, 2020; Mustafa *et al.*, 2022]. Moreover, the interplay of instance-level perturbations, learned representations, and multi-component loss function renders standard analysis techniques of CL [Arora *et al.*, 2019; Lei *et al.*, 2023] inapplicable in adversarial scenarios.

To tackle these challenges, a recent study [Zou and Liu, 2023] considers the surrogate upper bound on the original adversarial risk, which reduces the analysis problem of the adversarial risk of ACL to that of the general risk using standard learning-theoretic techniques. In particular, they develop generalization bounds scale $\mathcal{O}(k)$ for ACL in terms of the number of negative samples k. However, the bound linearly dependent on the number of negative samples k will not be valid for large values of k [Chen *et al.*, 2020; Lei *et al.*, 2023]. Furthermore, the surrogate bound derived from the average supervised loss in [Zou and Liu, 2023] may overestimate the original adversarial loss, and not yield meaningful generalization guarantees [Awasthi *et al.*, 2020].

This paper provides a comprehensive study of the generalization properties of ACL. In summary, this work makes three key contributions:

- We develop a unified generalization analysis framework

[*] Corresponding author.

Learning Type	Reference	Attack Type	Analysis Tool	Convergence Rate
SL	Yin et al. (2019) Khim and Loh (2018)	ℓ_∞-norm additive	Rademacher complexity	$\mathcal{O}(K/\sqrt{n})$
	Awasthi et al. (2020)	ℓ_r-norm additive	Rademacher complexity	$\mathcal{O}(K/\sqrt{n})$
	Xing et al. (2021) Xiao et al. (2022)	ℓ_r-norm additive	Algorithmic stability	$\sharp\mathcal{O}(1/n)$
	Mustafa et al. (2022)	ℓ_r-norm additive	Covering number	$\mathcal{O}(\log(K)/\sqrt{n})$
		Spatial transformation	Local Rademacher complexity	$\star\mathcal{O}(1/n)$
ACL	Zou and Liu (2023)	ℓ_r-norm additive	Rademacher complexity	$\mathcal{O}(k/\sqrt{n})$
	Section 4.1 (Ours)	ℓ_r-norm additive	Covering number	$\mathcal{O}(\log(k)/\sqrt{n})$
	Section 4.2 (Ours)		Local Rademacher complexity	$\star\mathcal{O}(1/n)$

Table 1: Summary of generalization analysis for adversarial learning (SL-Supervised Learning; $\sharp$-generalization bound in expectation; $\star$-optimization bound; k-the number of negative samples for ACL; K-the number of classes for multi-label classification (especially, $k = K$ in downstream classification tasks)).

for ACL that spans both linear models and deep neural networks in its applicability, which sheds a new light on understanding the generalization properties in the self-supervised setup [Jiang *et al.*, 2020; Fan *et al.*, 2021]. Unlike prior work on optimizing a surrogate upper bound, we work directly with the original adversarial unsupervised risk, yielding a tighter generalization gap compared to [Zou and Liu, 2023].

- We establish generalization bounds scale $\mathcal{O}(\log(k))$ in the number of negative samples k, by leveraging ℓ_∞-covering number and a novel vector concentration estimation technique. Our results hold for a broad range of negative samples, and exhibit a substantial advancement in the current Rademacher complexity-based bounds $\mathcal{O}(k)$ [Yin *et al.*, 2019; Zou and Liu, 2023; Awasthi *et al.*, 2019].
- We provide optimistic generalization bounds $\mathcal{O}(n^{-1})$ for the smooth loss by introducing robust self-bounding property [Reeve and Kaban, 2020] and using the local Rademacher complexity [Bartlett *et al.*, 2005]. To the best of our knowledge, this is the first fast-decaying generalization bounds for ACL. Experimental observations on real-world datasets validate our theoretical findings.

2 Related Work

Adversarial Contrastive Learning. Recently, adversarial robustness has become a crucial requirement when designing learning algorithms. Conventional adversarial learning methods commonly use labeled data to adversarially train robust models [Madry *et al.*, 2018; Zheng *et al.*, 2020]. Due to the difficulty and high cost of acquiring large-scale labeled data, recent work has proposed ACL methods to achieve superior model robustness on unlabeled data, by incorporating the adversarial learning strategy with CL [Jiang *et al.*, 2020; Lee *et al.*, 2021; Kim *et al.*, 2020]. A summary of relevant empirical work is provided in *Supplementary Material B*. Although these work show impressive performance, there is a lack of theoretical guarantees of the generalization of ACL.

Adversarial Robustness. Szegedy et al. (2014) have found that machine learning models are highly susceptible to imperceptible perturbations added to input samples. A recent work focuses on designing models robust to adversarial perturbations [Jiang *et al.*, 2020; Liu *et al.*, 2022; Kim *et al.*, 2020], and investigating adversarially robust generalization from a theoretical perspective [Xing *et al.*, 2021; Xiao *et al.*, 2022; Zou and Liu, 2023].

Yin et al. (2019) study the adversarial generalization under ℓ_∞ attacks by employing Rademacher complexity. They consider a surrogate adversarial loss based on the SDP relaxation. However, the relaxation is quite weak, causing the risk bound to be greatly overestimated. Awasthi et al. (2020) extend the prior work of [Yin *et al.*, 2019] and provide a general analysis of ℓ_r attacks ($r \geq 1$). The studies of Awasthi et al. (2020) and Yin et al. (2019) are both limited to simple models and architectures (e.g., one-hidden-layer neural networks). Khim and Loh (2018) establish the adversarial risk bounds for multi-layer neural networks by introducing the tree-transform, which can be vacuous since they assume that the perturbations propagate through each path in the network independently. Mustafa et al. (2022) obtain the generalization bounds for deep neural networks by leveraging the notion of coverage. Their bounds apply to the original network and grow as $\mathcal{O}(\log(K))$ in the number of label classes K, while the existing bounds mostly grow as $\mathcal{O}(K)$ [Yin *et al.*, 2019; Awasthi *et al.*, 2020; Khim and Loh, 2018].

The aforementioned work is confined to supervised learning paradigms. Theoretical understanding of self-supervised learning under adversarial scenarios remains underdeveloped. A recent work [Zou and Liu, 2023] studies the generalization properties of ACL based on the supervised surrogate risk. Specifically, they derive a surrogate upper bound on the original adversarial unsupervised loss by exploiting the monotonicity of the function and introducing the notion of average supervised risk. By analyzing the Rademacher complexity of the surrogate, they establish the generalization bounds scale $\mathcal{O}(k)$ in the number of negative samples k. In contrast, we consider a general class and a novel adversarial sample set with cardinality nk to approximate the adversarial unsupervised loss class. With these in mind, we develop the bounds scale $\mathcal{O}(\log(k))$ by quantifying the covering number of the general class over adversarial samples. The relevant work is summarized in Table 1.

3 Adversarial Contrastive Learning

Let the vectors are denoted by boldface lowercase letters (e.g., $\mathbf{w}$), and the elements in the vector be denoted by italics letters with subscripts (e.g., w_i). The matrices are denoted by uppercase letter (e.g., W). For a matrix $W = (\mathbf{w}_1, \dots, \mathbf{w}_{d'}) \in \mathbb{R}^{d \times d'}$, the (p, q) matrix norm of W is defined as $\|W\|_{p,q} = \|(\|\mathbf{w}_1\|_p, \dots, \|\mathbf{w}_{d'}\|_p)\|_q$. The Frobenius norm of matrices is denoted by $\|\cdot\|_F$. Let p^* be the number satisfying $1/p + 1/p^* = 1$.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the space of all possible sample points, and $\mathcal{C}$ denote the set of all latent classes. Suppose there exists a probability distribution $\mathcal{D}_c$ over $\mathcal{X}$ that quantifies the relevance of sample $\mathbf{x}$ to the class c, where c is randomly drawn according to the distribution ρ on the classes $\mathcal{C}$. The probability of similar samples $\mathbf{x}$ and $\mathbf{x}^+$ drawn from the same class c is expressed as

$$\mathcal{D}_{sim}(\mathbf{x}, \mathbf{x}^+) = \mathbb{E}_{c \sim \rho}\left[\mathcal{D}_c(\mathbf{x})\mathcal{D}_c(\mathbf{x}^+)\right],$$

and the probability of drawing the negative sample $\mathbf{x}^-$ unrelated to $\mathbf{x}$ is expressed as

$$\mathcal{D}_{neg}(\mathbf{x}^-) = \mathbb{E}_{c \sim \rho}\left[\mathcal{D}_c(\mathbf{x}^-)\right].$$

Let $(\mathbf{x}_i, \mathbf{x}_i^+) \sim \mathcal{D}_{sim}$ and $(\mathbf{x}_{i1}^-, \dots, \mathbf{x}_{ij}^-, \dots, \mathbf{x}_{ik}^-) \sim \mathcal{D}_{neg}$, where k denotes the number of negative samples. Given a training set $S = \left\{(\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_{i1}^-, \dots, \mathbf{x}_{ik}^-)\right\}_{i=1}^n$ of size n, the target of CL is to learn a representation function $f : \mathcal{X} \mapsto \mathbb{R}^{d'}$ on unlabeled samples such that pulls together similar pairs $(\mathbf{x}_i, \mathbf{x}_i^+)$ and pushes apart dissimilar pairs $(\mathbf{x}_i, \mathbf{x}_{ij}^-)$ [Chen *et al.*, 2020; Arora *et al.*, 2019; He *et al.*, 2020]. Let $\mathcal{F}$ be a family of functions mapping from $\mathcal{X}$ to $\mathbb{R}^{d'}$. A popular approach for seeking a good representation $f \in \mathcal{F}$ is to minimize the following empirical risk on S

$$\mathcal{E}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \ell\big(\{f(\mathbf{x}_i)^T(f(\mathbf{x}_i^+) - f(\mathbf{x}_{ij}^-))\}_{j=1}^k\big),$$

where $f(\mathbf{x}_i)^T$ denotes the transpose of $f(\mathbf{x}_i)$, and $\ell : \mathbb{R}^k \mapsto \mathbb{R}_+$ is an unsupervised loss function, such as the hinge loss $\ell(\boldsymbol{\tau}) = \max\{0, 1 + \max_{j \in [k]}\{-\tau_j\}\}$, and the logistic loss $\ell(\boldsymbol{\tau}) = \log(1 + \sum_{j \in [k]} \exp(-\tau_j))$ for $\boldsymbol{\tau} = (\tau_1, \dots, \tau_k) \in \mathbb{R}^k$.

In ACL, an adversary carefully designs adversarial samples to fool the learned model, resulting in performance deterioration [Kim *et al.*, 2020; Fan *et al.*, 2021]. From a representation perspective [Ho and Nvasconcelos, 2020; Kim *et al.*, 2020; Zou and Liu, 2023], one can perturb $\mathbf{x}$ in a manner that its feature representation is as far away from $f(\mathbf{x}^+)$ as possible and as close as possible to $f(\mathbf{x}^-)$. Let $\mathcal{B}_r(\varepsilon)$ be the perturbation space and defined as $\{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_r \leq \varepsilon\}$. Given a sample $(\mathbf{x}, \mathbf{x}^+, \mathbf{x}_1^-, \dots, \mathbf{x}_k^-)$, a learned representation function f, the adversary selects the perturbation parameter by

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta} \in \mathcal{B}_r(\varepsilon)} \ell\big(\{g_f(\mathbf{x} + \boldsymbol{\theta}, \mathbf{x}^+, \mathbf{x}_j^-)\}_{j=1}^k\big).$$

where $g_f(\mathbf{x} + \boldsymbol{\theta}, \mathbf{x}^+, \mathbf{x}^-) := f(\mathbf{x} + \boldsymbol{\theta})^T(f(\mathbf{x}^+) - f(\mathbf{x}^-))$ and its corresponding hypothesis class is defined by

$$\mathcal{G} := \{g_f(\mathbf{x} + \boldsymbol{\theta}, \mathbf{x}^+, \mathbf{x}^-) : f \in \mathcal{F}\}. \tag{1}$$

A common strategy [Madry *et al.*, 2018] for obtaining a robust representation f is to minimize the following *adversarial empirical risk*

$$\widetilde{\mathcal{E}}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \max_{\boldsymbol{\theta}_i \in \mathcal{B}_r(\varepsilon)} \ell\big(\{g_f(\mathbf{x} + \boldsymbol{\theta}, \mathbf{x}^+, \mathbf{x}_j^-)\}_{j=1}^k\big).$$

The risk $\widetilde{\mathcal{E}}_n(f)$ measures the empirical performance of f on training samples subjected to adversarial perturbation. The generalization performance of f in adversarial scenarios is measured by the *adversarial population risk*, defined by

$$\widetilde{\mathcal{E}}(f) := \mathbb{E}\big[\max_{\boldsymbol{\theta} \in \mathcal{B}_r(\varepsilon)} \ell\big(\{g_f(\mathbf{x} + \boldsymbol{\theta}, \mathbf{x}^+, \mathbf{x}_j^-)\}_{j=1}^k\big)\big].$$

In this paper, we are interested in the uniform deviation between adversarial population risk and adversarial empirical risk, denoted as $\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f)$.

Results in learning theory [Bartlett and Mendelson, 2002] show that we can bound $\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f)$ by quantifying the complexity of the adversarial loss class $\widetilde{\mathcal{L}}_{adv}$, where

$$\begin{aligned}\widetilde{\mathcal{L}}_{adv} := \{&(\mathbf{x}, \mathbf{x}^+, \mathbf{x}_1^-, \dots, \mathbf{x}_k^-) \mapsto \\ &\max_{\boldsymbol{\theta} \in \mathcal{B}_r(\varepsilon)} \ell(\{g_f(\mathbf{x} + \boldsymbol{\theta}, \mathbf{x}^+, \mathbf{x}_j^-)\}_{j=1}^k) : f \in \mathcal{F}\}.\end{aligned} \tag{2}$$

One of the main tools used in this paper to measure complexity of hypothesis classes is the ℓ_∞*-covering number* [Zhou, 2002], as defined below.

Definition 1 (ℓ_∞-covering number)**.** *Let $S = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n\} \in \mathcal{Z}^n$, and $\mathcal{H}$ be a function class defined over the space $\mathcal{Z}$. For any $\mu > 0$, the ℓ_∞-covering number of $\mathcal{H}$ w.r.t. S, denote as $\mathcal{N}_\infty(\mathcal{H}, \mu, S)$, is defined as the smallest cardinality m of a collection of vectors $\mathbf{v}_1, \dots, \mathbf{v}_m \in \{(h(\mathbf{z}_1), \dots, h(\mathbf{z}_n)) : h \in \mathcal{H}\}$ such that*

$$\sup_{h \in \mathcal{H}} \min_{j \in [m]} \max_{i \in [n]} |h(\mathbf{z}_i) - v_{ji}| \leq \mu,$$

where v_{ji} is the i-th component of vector $\mathbf{v}_j$.

As pointed out by Mustafa et al. (2022), quantifying ℓ_∞-covering number of class $\widetilde{\mathcal{L}}_{adv}$ (i.e., $\mathcal{N}_\infty(\widetilde{\mathcal{L}}_{adv}, \mu, S)$) is difficult since it takes values in an infinite space. We utilize the following general class to approximate the adversarial loss class, i.e., $\widetilde{\mathcal{L}}_{adv}$:

$$\begin{aligned}\mathcal{L}_{adv} := \big\{&(\mathbf{x} + \boldsymbol{\theta}, \mathbf{x}^+, \mathbf{x}_1^-, \dots, \mathbf{x}_k^-) \mapsto \\ &\ell\big(\{g_f(\mathbf{x} + \boldsymbol{\theta}, \mathbf{x}^+, \mathbf{x}_j^-)\}_{j=1}^k\big) : f \in \mathcal{F}\big\},\end{aligned} \tag{3}$$

which incorporates perturbations into the argument to remove the maximum over $\mathcal{B}_r(\varepsilon)$. We further approximate k components of functions in $\mathcal{L}_{adv}$ by employing an adversarial set of cardinality nk.

To this end, we introduce some necessary Lipschitzness assumptions.

Assumption 1. *Let functions $\ell : \mathbb{R}^k \mapsto \mathbb{R}_+$ and $g_f : \mathcal{X}^3 \mapsto \mathbb{R}$ be Lipschitz continuous:*

1. *The function ℓ is ξ-Lipschitz w.r.t. the $\|\cdot\|_\infty$-norm if, for $\forall \boldsymbol{\tau}_1, \boldsymbol{\tau}_2 \in \mathbb{R}^k$ and $\xi > 0$*
$$|\ell(\boldsymbol{\tau}_1) - \ell(\boldsymbol{\tau}_2)| \leq \xi \|\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2\|_\infty.$$
2. *The function g_f is η-Lipschitz w.r.t. the $\|\cdot\|_r$-norm if, for $\forall \mathbf{x}, \mathbf{x}^+, \mathbf{x}^-, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$, and $\eta > 0$*
$$|g_f(\mathbf{x} + \boldsymbol{\theta}_1, \mathbf{x}^+, \mathbf{x}^-) - g_f(\mathbf{x} + \boldsymbol{\theta}_2, \mathbf{x}^+, \mathbf{x}^-)| \leq \eta \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_r.$$

It is worth noting that the Lipschitz property of ℓ is a fairly standard assumption, encompassing commonly used loss functions such as the hinge loss and the logistic loss [Arora *et al.*, 2019]. Moreover, the Lipschitz condition on $\boldsymbol{\theta} \mapsto g_f(\mathbf{x}+\boldsymbol{\theta}, \mathbf{x}^+, \mathbf{x}^-)$ plays a key role in our analysis and is readily satisfied by a wide range of attacks [Madry *et al.*, 2018; Awasthi *et al.*, 2020; Fan *et al.*, 2021].

4 Theoretical Analysis for ACL

In this section, we present the high probability bound of generalization gap in Section 4.1. We further derive the optimistic generalization bound for the specific case in Section 4.2. Please refer to the Supplementary Material for complete proofs.

4.1 The Generalization Error Bound for ACL

We first derive an upper bound on the ℓ_∞-covering number of the adversarial class $\widetilde{\mathcal{L}}_{adv}$.

Lemma 1. *With definitions of (1), (2), (3), and Assumption 1. Let the set $\mathcal{C}_\mathcal{B}(\mu/2\xi\eta)$ be a $\mu/2\xi\eta$-cover of $\mathcal{B}_r(\varepsilon)$ and define the adversarial sample set*
$$\tilde{S} = \{(\mathbf{x}_i + \boldsymbol{\theta}_i, \mathbf{x}_i^+, \mathbf{x}_{ij}^-), i \in [n], j \in [k], \boldsymbol{\theta}_i \in \mathcal{C}_\mathcal{B}(\mu/2\xi\eta)\}.$$
Then, we have
$$\mathcal{N}_\infty(\widetilde{\mathcal{L}}_{adv}, \mu, S) \leq \mathcal{N}_\infty(\mathcal{L}_{adv}, \mu, S) \leq \mathcal{N}_\infty(\mathcal{G}, \mu/2\xi, \tilde{S}).$$

Remark 1. *The first inequality above allows us to control the ℓ_∞-covering number of $\widetilde{\mathcal{L}}_{adv}$ by the ℓ_∞-covering number of $\mathcal{L}_{adv}$, which extends previous work of [Mustafa* et al.*, 2022] for supervised adversarial learning to unsupervised ACL. By exploiting the Lipschitz continuity of ℓ and covering the perturbation space $\mathcal{B}_r(\varepsilon)$, we reduce the complexity analysis of the loss class $\mathcal{L}_{adv}$ w.r.t. k components to that of the function class $\mathcal{G}$ on $\tilde{S}$ of cardinality nk.*

Motivated by previous work [Srebro *et al.*, 2010; Anthony *et al.*, 1999; Lei *et al.*, 2023], we derive the relationship between the ℓ_∞-covering number and the *worst-case Rademacher complexity*, which serves as the key step in developing the generalization bounds.

Definition 2 (The worst-case Rademacher complexity)**.** *Let $\mathcal{H}$ be a real-valued function class. Given a sample $S = \{\mathbf{z}_i\}_{i=1}^n$, the worst-case Rademacher complexity is defined as $\mathfrak{R}_n(\mathcal{H}) = \sup_{|S| \leq n} \mathfrak{R}_S(\mathcal{H})$, where $|S|$ is the cardinality of S and $\mathfrak{R}_S(\mathcal{H})$ is the empirical Rademacher complexity of $\mathcal{H}$ on S defined as $\mathfrak{R}_S(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\epsilon}}[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \in [n]} \epsilon_i h(\mathbf{z}_i)]$, where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \sim \{\pm 1\}^n$ are i.i.d. Rademacher random variables.*

Theorem 1. *With the same conditions in Lemma 1. Assume that $\|f(\mathbf{x})\|_2 \leq \Lambda$ for any $f \in \mathcal{F}$ and $\mathbf{x} \in \mathcal{X}$. Let the cardinality of cover set be defined as $M_\mu := |\mathcal{C}_\mathcal{B}(\mu/2\xi\eta)|$. Then*
$$\log \mathcal{N}_\infty(\widetilde{L}_{adv}, \mu, S) \leq$$
$$1 + \frac{C_1 nkM_\mu \xi^2 \log^2(C_2 \Lambda^4 \xi^2 nkM_\mu/\mu^2)}{\mu^2} \mathfrak{R}^2_{\tilde{S}, nkM_\mu}(\mathcal{G}),$$
where $C_1 = 256, C_2 = 128e$, and $\mathfrak{R}_{\tilde{S}, nkM_\mu}(\mathcal{G})$ is
$$\frac{1}{nkM_\mu} \mathbb{E}_{\boldsymbol{\epsilon} \sim \{\pm 1\}^{nkM_\mu}} \Big[\sup_{f \in \mathcal{F}} \sum_{i \in [nkM_\mu]} \epsilon_i g_f(\mathbf{x}_i + \boldsymbol{\theta}_i, \mathbf{x}_i^+, \mathbf{x}_i^-) \Big],$$
and denotes the worst-case Rademacher complexity of $\mathcal{G}$ defined on the set $\tilde{S}$.

We then obtain the following high-probability generalization bound.

Theorem 2. *With the same conditions in Theorem 1. Assume that the function ℓ is bounded by B. Then for all $f \in \mathcal{F}$ and $g_f \in \mathcal{G}$, with probability at least $1 - \delta$, we have*
$$\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f) \leq 3B\sqrt{\frac{\log(2/\delta)}{2n}} + 96\xi n^{\frac{1}{2}}(\Lambda^2 + 1) + 192\xi$$
$$(1 + \log(4\Lambda^2 n^{\frac{3}{2}} kM_\mu)\lceil \log_2 \frac{\Lambda^2\sqrt{n}}{12} \rceil)\sqrt{kM_\mu}\mathfrak{R}_{\tilde{S}, nkM_\mu}(\mathcal{G}).$$

Remark 2. *As we can see, the worst-case Rademacher complexity $\mathfrak{R}_{\tilde{S}, nkM_\mu}(\mathcal{G})$ is the key quantity for the generalization of ACL. In a typical case, $\mathfrak{R}_{\tilde{S}, nkM_\mu}(\mathcal{G}) = \mathcal{O}((nkM_\mu)^{-\frac{1}{2}})$, where M_μ is the cardinality of the set covering perturbation space $B_r(\varepsilon)$ and is heavily associated with the feature dimension. In this case, the generalization error bound has the order $\mathcal{O}(\log(nkM_\mu)n^{-\frac{1}{2}})$, similar to the bounds in [Lei* et al.*, 2023; Mustafa* et al.*, 2022]. This result is more favorable for downstream k-category classification tasks with larger k compared to existing bounds scale $\mathcal{O}(k)$ [Zou and Liu, 2023; Awasthi* et al.*, 2020; Yin* et al.*, 2019].*

Remark 3. *Despite Zou and Liu (2023) develop generalization bounds for ACL, they take the label into consideration for analyzing the adversarial risk. In contrast, our result derived from Theorem 1 directly applies to the original adversarial unsupervised risk, which provides a tighter generalization guarantee for ACL in a self-supervised manner [Fan* et al.*, 2021; Lee* et al.*, 2021].*

4.2 The Optimistic Generalization Bound for ACL

We further refine the result of Theorem 2, and then develop the fast-rate bound for the generalization of ACL by using the smooth assumption of the loss function [Srebro *et al.*, 2010; Reeve and Kaban, 2020]. Our analysis is based on the *local Rademacher complexity* [Bartlett *et al.*, 2005] over adversarial samples.

Let the local adversarial loss class $\widetilde{L}_{adv}|^\gamma$ be defined as $\widetilde{L}_{adv}|^\gamma := \{(\mathbf{x}, \mathbf{x}^+, \mathbf{x}_1^-, \ldots, \mathbf{x}_k^-) \mapsto \max_{\boldsymbol{\theta} \in \mathcal{B}_r(\varepsilon)} \ell(\{g_f(\mathbf{x} + \boldsymbol{\theta}, \mathbf{x}^+, \mathbf{x}_j^-)\}_{j=1}^k) : f \in \mathcal{F}, \widetilde{\mathcal{E}}_n(f) \leq \gamma\}$, which is the set of adversarial losses with empirical adversarial training errors at most γ, for $f \in \mathcal{F}$. We consider the smooth loss function with the following *robust self-bounding* Lipschitz property [Reeve and Kaban, 2020; Mustafa *et al.*, 2022].

Assumption 2. *A loss function $\ell : \mathbb{R}^k \mapsto \mathbb{R}_+$ is said to be λ-robust-self-bounding Lipschitz continuous w.r.t. a set $\mathcal{B}_r(\varepsilon)$ if, for any measurable functions $\phi, \psi : \mathcal{B}_r(\varepsilon) \to \mathbb{R}^k$, we have*

$$\begin{aligned} &\big| \max_{\boldsymbol{\theta} \in \mathcal{B}_r(\varepsilon)} \ell(\phi(\boldsymbol{\theta})) - \max_{\boldsymbol{\theta} \in \mathcal{B}_r(\varepsilon)} \ell(\psi(\boldsymbol{\theta})) \big| \\ \leq &\lambda \max\{ \max_{\boldsymbol{\theta} \in \mathcal{B}_r(\varepsilon)} \ell(\phi(\boldsymbol{\theta})), \max_{\boldsymbol{\theta} \in \mathcal{B}_r(\varepsilon)} \ell(\psi(\boldsymbol{\theta})) \}^{\frac{1}{2}} \\ &\times \max_{\boldsymbol{\theta} \in \mathcal{B}_r(\varepsilon)} \|\phi(\boldsymbol{\theta}) - \psi(\boldsymbol{\theta})\|_\infty . \end{aligned}$$

Based on Assumption 2, we show that the covering number of the local adversarial class $\widetilde{L}_{adv}|^\gamma$ can be bounded by the covering number of the function class $\mathcal{G}$.

Lemma 2. *Let $\mathcal{G}$ be defined in (1). Assume Assumption 2 holds. Further let the set $\mathcal{C}_{\mathcal{B}}(\mu/2\sqrt{2\gamma}\lambda\eta)$ be a $\mu/2\sqrt{2\gamma}\lambda\eta$-cover of $\mathcal{B}_r(\varepsilon)$ and the adversarial set be defined as*

$$\hat{S} := \big\{(\mathbf{x}_i + \boldsymbol{\theta}_i, \mathbf{x}_i^+, \mathbf{x}_{ij}^-), i \in [n], j \in [k], \boldsymbol{\theta}_i \in \mathcal{C}_{\mathcal{B}}(\mu/2\sqrt{2\gamma}\lambda\eta)\big\}.$$

Then, we have the upper bound

$$\mathcal{N}_\infty(\widetilde{L}_{adv}|^\gamma, \mu, S) \leq \mathcal{N}_\infty(\mathcal{G}, \mu/2\sqrt{2\gamma}\lambda, \hat{S}).$$

Then, we have the following structural result on the ℓ_∞-covering number of the local adversarial class. It serves as a key step in developing the fast-rate bound by a sub-root bound on the local Rademacher complexity [Bartlett *et al.*, 2005].

Theorem 3. *With the same conditions in Lemma 2. Assume that $\|f(\mathbf{x})\|_2 \leq \Lambda$ for any $f \in \mathcal{F}$ and $\mathbf{x} \in \mathcal{X}$. Let the cardinality of cover set be defined as $\hat{M}_\mu := |\mathcal{C}_{\mathcal{B}}(\mu/2\sqrt{2\gamma}\lambda\eta)|$. Then*

$$\begin{aligned} &\log \mathcal{N}_\infty(\widetilde{L}_{adv}|^\gamma, \mu, S) \leq \\ &1 + \frac{C_3 nk\hat{M}_\mu \gamma\lambda^2 \log^2(C_4\Lambda^4\gamma\lambda^2 nk\hat{M}_\mu/\mu^2)}{\mu^2} \mathfrak{R}^2_{\hat{S}, nk\hat{M}_\mu}(\mathcal{G}), \end{aligned}$$

where $C_3 = 512, C_4 = 256e$, and $\mathfrak{R}_{\hat{S}, nk\hat{M}_\mu}(\mathcal{G})$ is

$$\frac{1}{nk\hat{M}_\mu} \mathbb{E}_{\boldsymbol{\epsilon} \sim \{\pm 1\}^{nk\hat{M}_\mu}} \Big[\sup_{f \in \mathcal{F}} \sum_{i \in [nk\hat{M}_\mu]} \epsilon_i g_f(\mathbf{x}_i + \boldsymbol{\theta}_i, \mathbf{x}_i^+, \mathbf{x}_i^-) \Big],$$

and denotes the worst-case Rademacher complexity of $\mathcal{G}$ defined on the set $\hat{S}$.

In the following theorem, we provide the optimistic generalization bound for the adversarial loss function under the smooth assumption.

Theorem 4. *With the same conditions in Theorem 3. Suppose that the function ℓ is bounded by B. With probability at least $1 - \delta$, for all $f \in \mathcal{F}$ and $g_f \in \mathcal{G}$, we have*

$$\begin{aligned} &\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f) \\ =&\widetilde{\mathcal{O}}\Big((B + \lambda^2\Lambda^4)n^{-1} + \lambda^2 k\hat{M}_\mu \mathfrak{R}^2_{\hat{S}, nk\hat{M}_\mu}(\mathcal{G})\Big) + \\ &\widetilde{\mathcal{O}}\Big((\sqrt{B} + \lambda\Lambda^2)n^{-\frac{1}{2}} + \lambda\sqrt{k\hat{M}_\mu}\mathfrak{R}_{\hat{S}, nk\hat{M}_\mu}(\mathcal{G})\Big)\widetilde{\mathcal{E}}_n^{\frac{1}{2}}(f), \end{aligned}$$

where the logarithmic factors are hidden in $\widetilde{\mathcal{O}}$.

Remark 4. *Theorem 4 provides a fast-rate generalization bound, in the sense that the bound depends on adversarial training errors. Typically, we have $\mathfrak{R}_{\hat{S}, nk\hat{M}_\mu}(\mathcal{G}) = \mathcal{O}((nk\hat{M}_\mu)^{-\frac{1}{2}})$ [Yin* et al.*, 2019; Lei* et al.*, 2023]. Then, the first term grows as $\mathcal{O}(n^{-1})$, while the second term grows at the usual $\mathcal{O}(n^{-\frac{1}{2}})$. However, if $\widetilde{\mathcal{E}}_n(f) = 0$, Theorem 4 implies generalization bounds*

$$\widetilde{\mathcal{E}}(f) = \widetilde{\mathcal{O}}((B + \lambda^2\Lambda^4)n^{-1} + \lambda^2 n^{-1}).$$

This achieves faster rates of convergence than existing generalization bounds [Zou and Liu, 2023; Awasthi et al.*, 2020; Yin* et al.*, 2019].*

5 Explicit Bounds for Hypothesis Classes

Theoretical results in Section 4 indicate that the generalization can be guaranteed in terms of the worst-case Rademacher complexity of hypothesis classes. In this section, we provide clear characterizations of the Rademacher complexity for linear hypotheses and deep neural networks, and establish corresponding generalization error bounds. Please refer to the Supplementary Material for complete proofs.

5.1 Linear Hypothesis Class

Let $\mathcal{F}$ be a class of linear functions from $\mathcal{X}$ to $\mathbb{R}^{d'}$, defined by

$$\mathcal{F} := \{\mathbf{x} \mapsto W\mathbf{x} : W \in \mathbb{R}^{d' \times d}, \|W\|_{2,p} \leq R\}. \quad (4)$$

We derive the following upper bound for linear hypothesis classes.

Theorem 5. *Let the class $\mathcal{G}$ be defined in (1), and the hypothesis class $\mathcal{F}$ be defined in (4). Suppose that the loss $\ell : \mathbb{R}^+ \mapsto \mathbb{R}_+$ is ξ-Lipschitz and bounded by B. Let the perturbation be chosen in ℓ_r norm ball of radius ε. Assume $\|f(\mathbf{x})\| \leq \Lambda$, $\delta \in (0,1)$. Then, we have*

$$\mathfrak{R}_{\tilde{S}, nkM_\mu}(\mathcal{G}) \leq 3\sqrt{12}\Lambda R d'^{1/p^*}\big(B_x^p + B_\varepsilon\big)/\sqrt{nkM_\mu},$$

and with probability at least $1 - \delta$ for all $f \in \mathcal{F}$

$$\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f) = \mathcal{O}\Big(\frac{1}{\sqrt{n}} + \frac{\log(\Lambda nkM_\mu)}{\sqrt{n}}\big(B_x^p + B_\varepsilon\big)\Big)$$

where $M_\mu = (12n\xi R^2 B_x B_\varepsilon)^d$, $B_\varepsilon = \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})\varepsilon$, $B_x = \max\{\|\mathbf{x}_i\|_2, \|\mathbf{x}_i^+\|_2, \|\mathbf{x}_{ij}^-\|_2 : i \in [n], j \in [k]\}$, and $B_x^p = \max(\sqrt{p^* - 1}, 1)\max\{\|\mathbf{x}_i\|_2, \|\mathbf{x}_i^+\|_2, \|\mathbf{x}_{ij}^-\|_2 : i \in [n], j \in [k]\}$.

Remark 5. *The generalization bound has additional dimension dependent terms associated with the adversarial perturbation, i.e., B_ε and M_μ. The dependence $d^{1-\frac{1}{p}-\frac{1}{r}}$ in B_ε arises from the norm mismatch between the input dimension and the weight, which is consistent with Lemma 1 in [Awasthi* et al.*, 2020]. In particular, if $p \in [1, r^*]$, one can avoid a polynomial dimension dependence in B_ε. As discussed in Remark 2, the dependence d in M_μ is due to the complexity of the perturbation space $\mathcal{B}_r(\varepsilon)$. Therefore, we can narrow the generalization gap by reducing the effective dimensionality of the adversarial perturbation through a low-dimensional feature projection ($d' < d$). In this case, the dependence on d is reduced to $\mathcal{O}(d')$.*

Remark 6. *It is worth noting that the bound in [Zou and Liu, 2023] has as an extra polynomial dimension dependence $d^{\max\{\frac{1}{p}-\frac{1}{p^*},\frac{1}{p^*}-\frac{1}{p}\}}$, and grows as $\mathcal{O}(k)$ in the number of negative samples k. In contrast, our result avoids such dependence and only grows as $\mathcal{O}(\log(k))$, which benefits from the application of the vector-concentration inequality [Maurer, 2016] and the covering number [Zhou, 2002], enabling us to obtain an optimistic generalization bound of order $\frac{1}{n}$ by plugging an upper bound of $\Re_{\hat{M}_\mu}$ and $\hat{M}_\mu$ back into Theorem 4.*

5.2 Multi-layer Neural Network

We consider L-layer feed-forward networks with the following hypothesis class

$$\mathcal{F} := \Big\{\mathbf{x} \mapsto W_L\sigma(\cdots\sigma(W_1\mathbf{x})) : \|W_l\|_F \leq B_l, \forall l \in [L]\Big\}, \tag{5}$$

where $\sigma(\cdot)$ is an elementwise 1-Lipschitz activation function with $\sigma(0) = 0$, e.g., the ReLU activation, and $W_l \in \mathbb{R}^{h_l \times h_{l-1}}$, where $h_L = d', h_0 = d$.

By applying the "peeling" argument [Neyshabur *et al.*, 2015; Golowich *et al.*, 2018] and the vector-contraction inequality [Maurer, 2016], we derive the upper bound on the generalization error of deep neural networks. The theoretical results are summarized in the theorem below.

Theorem 6. *Let $\mathcal{G}$ be defined in (1), and $\mathcal{F}$ be a class of neural networks defined in (5). Suppose that the loss $\ell : \mathbb{R}^+ \mapsto \mathbb{R}_+$ be ξ-Lipschitz and bounded by B. Let the perturbation be selected in ℓ_r ball with radius ε. Assume $\|f(\mathbf{x})\| \leq \Lambda$, $\|W_1\|_{2,p} \leq B_1$, and $\delta \in (0,1)$. Then, we have*

$$\mathfrak{R}_{\tilde{S},nkM_\mu}(\mathcal{G}) \leq \tfrac{3\sqrt{12}\Lambda}{\sqrt{nkM_\mu}} h_1^{1/p^*}(\sqrt{L}+1)\textstyle\prod_{l=1}^L B_l^2 B_x(B_x^p + B_\varepsilon),$$

and with probability at least $1-\delta$ for all $f \in \mathcal{F}$

$$\begin{aligned}&\widetilde{\mathcal{E}}(f) - \widetilde{\mathcal{E}}_n(f)\\ =&\mathcal{O}\Big(\frac{1}{\sqrt{n}} + B_x\big(B_x^p + B_\varepsilon\big)\frac{\sqrt{L}\prod_{l=1}^L B_l^2\log(\Lambda nkM_\mu)}{\sqrt{n}}\Big),\end{aligned}$$

where $B_x = \max\{\|\mathbf{x}_i\|_2, \|\mathbf{x}_i^+\|_2, \|\mathbf{x}_{ij}^-\|_2 : i \in [n], j \in [k]\}$, $B_x^p = \max(\sqrt{p^*-1}, 1)\max\{\|\mathbf{x}_i\|_2, \|\mathbf{x}_i^+\|_2, \|\mathbf{x}_{ij}^-\|_2 : i \in [n], j \in [k]\}$, $B_\varepsilon = \max(1, d^{1-\frac{1}{p}-\frac{1}{r}})\varepsilon$, and $M_\mu = (12n\xi\prod_{l=1}^L B_l^2 B_x B_\varepsilon)^d$.

Remark 7. *The generalization bound above suffers from additional dimension dependent terms compared to its non-adversarial counterpart. As discussed in the linear case, one can indeed avoid $d^{1-\frac{1}{p}-\frac{1}{r}}$ dependence in B_ε by choosing an appropriate p-norm regularizer on the weight matrix (W_1) of the first layer, where $p \in [1, r^*]$. A projection on a low-dimensional representation space can help alleviate d dependence in M_μ caused by the complexity of the perturbation. While the bound in Theorem 6 of order $\mathcal{O}(n^{-\frac{1}{2}})$ is similar to the generalization bound in [Zou and Liu, 2023], our result scale $\mathcal{O}(\log(k))$ holds for a large k in the number of negative samples. By applying Theorem 4, we have a faster convergence rate $\mathcal{O}(n^{-1})$, which improves existing results [Zou and Liu, 2023; Yin* et al.*, 2019; Awasthi* et al.*, 2019].*

Dataset	**Size (n)**	**Dimension (d)**
Wine	178	13
A9a	48,842	123
Spambase	4,601	58
Waveform	5,000	21
MNIST	70,000	28x28
CIFAR-10	60,000	32x32x3

Table 2: The details of the adopted datasets.

6 Experimental Evaluation

In this section, we conduct several experiments to validate our theoretical results in Theorem 5&6.

6.1 Experimental Setup

Datasets. We use real-world datasets from UCI Machine Learning Repository[1] for experiments: the Wine, A9a, Spambase, Waveform, CIFAR-10, and MNIST datasets. Statistics of datasets are provided in Table 2.

Model Settings. We adopt a one-layer network architecture without activation function as the linear model. A five-layer feedforward neural network with ReLU [Hahnloser *et al.*, 2000] activation is used as the non-linear model, where the number of units is $(1024, 512, 256, 128, 64)$. All models trained by the Adam [Kingma and Ba, 2015] optimizer with the learning rate $1e-3$. Inspired by theoretical analysis of Theorem 5&6, we train an adversarial robust model by minimizing the objective

$$\max_{\boldsymbol{\theta}_i \in \mathcal{B}_r(\varepsilon)} \ell\big(\{f(\mathbf{x}_i + \boldsymbol{\theta}_i)^T(f(\mathbf{x}_i^+) - f(\mathbf{x}_{ij}^-))\}_{j=1}^k\big) + \lambda\|W_1\|_1, \tag{6}$$

where $\ell(\cdot)$ is contrastive loss defined in [Chen *et al.*, 2020], W_1 denotes the weight matrix of the first layer, and $\lambda \geq 0$ denotes the regularization parameter. The ℓ_∞ PGD algorithm [Madry *et al.*, 2018] with step size $\varepsilon/5$ is used to generate adversarial perturbations, where ε denotes the maximum allowable perturbation.

Evaluation Metric. After adversarial training, we run a PGD attack to check the adversarial test AUC [Ling *et al.*, 2003] for downstream classification tasks. Similar to [Yin *et al.*, 2019], we evaluate the generalization performance of ACL models by calculating the generalization error:

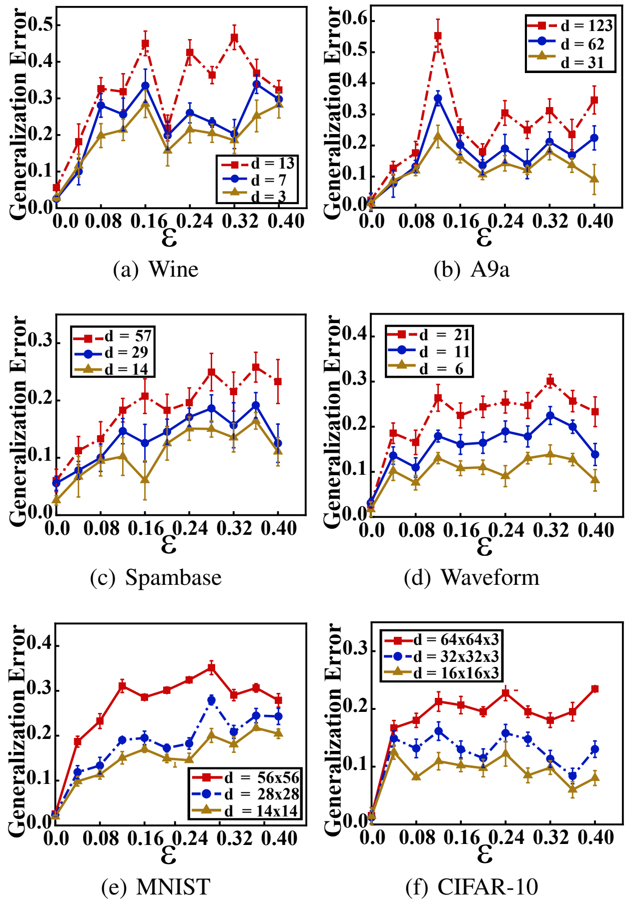
$$|\text{adversarial training AUC} - \text{adversarial test AUC}|\,. \tag{7}$$

Each experiment is independently repeated 10 times, with the mean value and standard deviation of the results across the 10 trials presented in Figure 1 and 2.

6.2 Experiments for Theoretical Observations

Effect of Feature Dimension. We first investigate the effect of feature dimension on the generalization behavior of ACL models. Specifically, we compare the generalization error (7) with different feature dimensions (i.e., d). As shown in Figure 1, the generalization error decreases as the number

[1] https://archive.ics.uci.edu

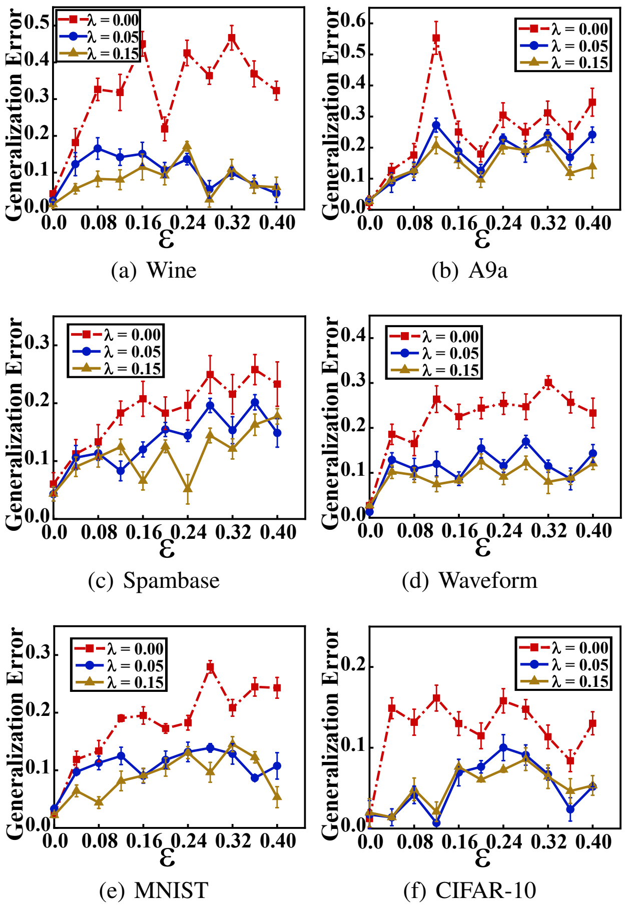


(a) Wine (b) A9a

(c) Spambase (d) Waveform

(e) MNIST (f) CIFAR-10

Figure 1: **The generalization performance (mean value and standard deviation) under different feature dimensions (d).** The dashed and solid lines represent the results of the original and projected dimensions, respectively. The first four subfigures and the last two subfigures show the results of linear and non-linear models, respectively. ε denotes the maximum allowable perturbation.



(a) Wine (b) A9a

(c) Spambase (d) Waveform

(e) MNIST (f) CIFAR-10

Figure 2: **The generalization performance (mean value and standard deviation) under different regularization (λ).** The dashed and solid lines represent the results of models trained without and with regularization, respectively. The first four subfigures and the last two subfigures show the results of linear and non-linear models, respectively. ε denotes the maximum allowable perturbation.

of features d decreases. This empirical observation implies that low-dimensional input can mitigate the impact of perturbations and improve the generalization ability.

Effect of Regularization. We evaluate the effect of the regularization on the generalization ability of ACL models. Here, we consider the objective with ℓ_1 regularization (6), and observe the generalization error (7) of models trained with different regularization parameters λ. As shown in Figure 2, models trained without regularization (i.e., $\lambda = 0$) have larger generalization errors, which suggests that regularization is beneficial for enhancing the robustness of the model.

7 Conclusions

This paper provides a systematic analysis of the generalization properties for ACL. We significantly improve the existing generalization bounds $\mathcal{O}(k)$ to a logical factor $\mathcal{O}(\log k)$ by developing the concentration estimation technique associated with the ℓ_∞-covering number. Under robust self-bounding Lipschitz condition, we further provide the optimistic bound with the fast convergence rate, when adversarial empirical error is zero. Our analysis is easily applicable to various models: it requires simply quantifying the worst-case Rademacher complexity of the hypothesis class. We present the general results for linear models and multi-layer neural networks with ℓ_r additive perturbations. Our results show that the generalization error bounds have additional perturbation terms associated with the feature dimension and the weight matrix, as compared to their non-adversarial counterparts. One can alleviate dimension dependence in the perturbation term by imposing a p-norm regularization on the weight matrix, where $1 \leq p \leq r^*$, and thus reduce the generalization gap. Moreover, projecting the input into the low-dimensional feature space help reduce the effective dimension of adversarial perturbations and improve adversarially robust generalization. Experimental observations on real-world datasets validate these theoretical findings. In future work, we will investigate the generalization behavior of ACL under various adversarial attacks (possibly non-additive).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62376104, 62106191, 12071166, and 11801201), the Fundamental Research Funds for the Central Universities of China (No. 2662023LXPY005), and HZAU-AGIS Cooperation Fund (No. SZYJY2023010).

References

[Anthony *et al.*, 1999] Martin Anthony, Peter L Bartlett, Peter L Bartlett, et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

[Arora *et al.*, 2019] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 9904–9923, 2019.

[Awasthi *et al.*, 2019] Pranjal Awasthi, Abhratanu Dutta, and Aravindan Vijayaraghavan. On robustness to adversarial examples and polynomial optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[Awasthi *et al.*, 2020] Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pages 431–441, 2020.

[Bartlett and Mendelson, 2002] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

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

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[Fan *et al.*, 2021] Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? *Advances in Neural Information Processing Systems*, 34:21480–21492, 2021.

[Golowich *et al.*, 2018] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference on Learning Theory*, pages 297–299. PMLR, 2018.

[Gowal *et al.*, 2020] Sven Gowal, Po-Sen Huang, Aaron van den Oord, Timothy Mann, and Pushmeet Kohli. Self-supervised adversarial robustness for the low-label, high-data regime. In *International Conference on Learning Representations*, 2020.

[Grill *et al.*, 2020] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[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.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[Hendrycks and Dietterich, 2019] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[Ho and Nvasconcelos, 2020] Chih-Hui Ho and Nuno Nvasconcelos. Contrastive learning with adversarial examples. *Advances in Neural Information Processing Systems*, 33:17081–17093, 2020.

[Jiang *et al.*, 2020] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Advances in Neural Information Processing Systems*, 33:16199–16210, 2020.

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

[Khosla *et al.*, 2020] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[Kim *et al.*, 2020] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994, 2020.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[Lee *et al.*, 2021] Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. Contrastive learning with adversarial perturbations for conditional text generation. In *International Conference on Learning Representations*, 2021.

[Lei *et al.*, 2023] Yunwen Lei, Tianbao Yang, Yiming Ying, and Ding-Xuan Zhou. Generalization analysis for contrastive representation learning. *ArXiv Preprint ArXiv:2302.12383*, 2023.

[Lin *et al.*, 2019] Ji Lin, Chuang Gan, and Song Han. Defensive quantization: When efficiency meets robustness. In *International Conference on Learning Representations*, 2019.

[Ling *et al.*, 2003] Charles X Ling, Jin Huang, and Harry Zhang. Auc: a better measure than accuracy in comparing learning algorithms. In *Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003, Halifax, Canada, June 11–13, 2003, Proceedings 16*, pages 329–341. Springer, 2003.

[Liu *et al.*, 2022] Ye Liu, Yaya Cheng, Lianli Gao, Xianglong Liu, Qilong Zhang, and Jingkuan Song. Practical evaluation of adversarial robustness via adaptive auto attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15105–15114, 2022.

[Ma *et al.*, 2021] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021.

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

[Maurer, 2016] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

[Misra and Maaten, 2020] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[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.

[Neyshabur *et al.*, 2015] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401. PMLR, 2015.

[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.

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[Wu *et al.*, 2018] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[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] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured adversarial attack: Towards general implementation and better interpretability. In *International Conference on Learning Representations*, 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.

[Zhang *et al.*, 2023] Yanghao Zhang, Wenjie Ruan, Fu Wang, and Xiaowei Huang. Generalizing universal adversarial perturbations for deep neural networks. *Machine Learning*, 112(5):1597–1626, 2023.

[Zheng *et al.*, 2020] Haizhong Zheng, Ziqi Zhang, Juncheng Gu, Honglak Lee, and Atul Prakash. Efficient adversarial training with transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1181–1190, 2020.

[Zhou, 2002] Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

[Zou and Liu, 2023] Xin Zou and Weiwei Liu. Generalization bounds for adversarial contrastive learning. *Machine Learning Research*, 24:1–54, 2023.