Generalization Bounds for Adversarial Contrastive Learning

Anonymous authors
Paper under double-blind review

Abstract

Contrastive learning has emerged as a powerful technique for learning meaningful representations from unlabeled data. However, its vulnerability to adversarial attacks remains a significant challenge. Adversarial contrastive learning, which combines contrastive learning with adversarial training, offers a promising avenue for achieving robust representations that withstand diverse attacks. Despite empirical evidence supporting its effectiveness, a comprehensive theoretical framework for adversarial contrastive learning has been lacking. In this paper, we address this gap by providing novel generalization bounds and theoretical insights. Leveraging the covering number technique and Lipschitzness of loss functions, we derive generalization bounds on the unsupervised adversarial contrastive learning function class. Our results apply to both linear and non-linear representations, including those obtained using deep neural networks (DNNs), which have improved the generalization bounds.

1 Introduction

Learning meaningful representations from unlabeled data plays a crucial role in enhancing the performance of machine learning models. Representation learning has shown great success in fields such as computer vision (Chen et al., 2020b; He et al., 2020; Caron et al., 2020) and natural language processing (Brown et al., 2020; Gao et al., 2021; Radford et al., 2021). Among various representation learning techniques, self-supervised contrastive learning, popularized by the SimCLR framework (Chen et al., 2020b), stands out. The core idea behind contrastive learning is to bring similar pairs (x, x^+) closer together in the embedding space while pushing apart negative samples from x, (denoted as $(x, x_1^-, \cdots, x_k^-)$). These learned representations can then be leveraged for downstream tasks, whether supervised or unsupervised (Chen et al., 2020b; He et al., 2020; Khosla et al., 2020). Notably, extensive research in contrastive learning has revealed that a sufficient number of negative samples is essential for achieving high-quality representations (Chen et al., 2020b; Khosla et al., 2020; Henaff, 2020; Tian et al., 2020).

Despite significant progress in representation learning, these representations remain susceptible to adversarial examples (Szegedy et al., 2013; Biggio et al., 2013), which are subtly perturbed samples carefully crafted to manipulate a model's predictions. Specifically, adversarial attacks aim to maximize the model's loss by slightly perturbing input samples. To mitigate this vulnerability, researchers have proposed adversarial training (Chen et al., 2020a; Tramer & Boneh, 2019). This technique employs a min-max optimization approach, where the model simultaneously minimizes its loss while facing maximally perturbed examples. By doing so, adversarial training enhances the robustness of the learned representations against adversarial attacks.

Adversarial contrastive learning (ACL) emerges from applying adversarial training to contrastive learning. In this paradigm, adversarial training enhances the robustness of representations learned from unlabeled data during unsupervised training. Empirical evidence supports the effectiveness of ACL in improving the quality of these robust representations (Kim et al., 2020; Ho & Nvasconcelos, 2020; Jiang et al., 2020). Despite its empirical success, the theoretical foundations of ACL remain somewhat limited. Recent work by Zou & Liu (2023) leverages Rademacher complexity to show the connection between unsupervised contrastive learning and the downstream classification task and claimed that the average adversarial risk of downstream tasks can be upper bounded by the adversarial unsupervised risk of the upstream task. In case of a linear representation and ℓ_∞ attack, they demonstrate that the bound depends on factors such as ℓ_2-norm of the input, the Frobenius norm of the weights, and the input dimension. For non-linear cases, the bound is influenced by the square root of the input dimension. However, their study only considers scenarios with a

single negative sample, overlooking the impact of a large number of negative samples, which is crucial for achieving optimal generalization performance (Chen et al., 2020b; Tian et al., 2020; Henaff, 2020; Khosla et al., 2020). Our work extends this analysis to a general case with a large number of negative samples, and our aim isto improve the dependency on this number.
In this paper, we present the following contributions:

- We apply the ℓ_∞-Lipschitz property of loss functions to derive generalization error bounds for ACL. These bounds incorporate the covering number of feature classes and show improved dependency on the number of negative examples, resulting in tighter bounds compared to existing literature.
- Our general results are applied to two specific scenarios of unsupervised representation learning: learning linear features and learning non-linear features via DNNs. In both cases, the bounds show a logarithmic dependence on the number of negative samples. For the non-linear case, the bound is also depends on the depth of the DNNs.

The remainder of the paper is organized as follows: Section 2 reviews related work and state-of-the-art approaches in this area. Section 3 defines the problem formulation. Our main theorem on the generalization error bound of ACL is presented in Section 4. Section 5 applies our theorem and demonstrates the generalization bounds for both linear and nonlinear feature representations. Section 6 contains all the proofs of our lemmas and theorems, as well as an additional theorem used in our proofs. Finally, Section 7 provides the conclusion.

2 Related work

Contrastive Learning Theoretical generalization analysis in contrastive learning is limited to a few works such as Arora et al. (2019); Ji et al. (2023); Lei et al. (2023). Arora et al. (2019) shows that the upstream unsupervised errors of representation learning bound the downstream classification task. They used Rademacher complexity to develop a generalization bound for the unsupervised CRL of the representation function class. However, their generalization bounds depend linearly on k, the number of negative samples, which is not beneficial if k is large, which is mostly the case in CRL. Motivated by this, Lei et al. (2023) improved the dependency on the number of negative samples regarding Rademacher complexity. For ℓ_2 Lipschitz loss, their bound doesn't depend on k; for ℓ_∞ Lipschitz loss, it is improved by a factor of k. Nozawa et al. (2020) derived PAC-Bayesian bounds on the posterior distribution of representation functions. Negative examples are typically selected randomly, which may result in them having the same label as the point of interest. This can introduce bias in the CRL loss function and potentially reduce performance in practice. Chuang et al. (2020) derived an approximation for the unbiased error of CRL loss and developed a generalization bound for the downstream task.

Adversarial Robustness Since Szegedy et al. (2013) demonstrated the vulnerability of neural networks to input perturbations, numerous generalization bounds for adversarial learning have been proposed. Montasser et al. (2019) established the PAC learnability of adversarial robust learning and Xu & Liu (2022) extended these results to the multi-class setting.
Additionally, other researchers have employed Rademacher complexity to generalize the ℓ_p-additive perturbation attack (Yin et al., 2019; Awasthi et al., 2020; Khim & Loh, 2018; Xiao et al., 2022; Mustafa et al., 2022). Yin et al. (2019) used the Rademacher complexity-based bound for linear models and one-layer neural networks based on surrogate loss, whereas Awasthi et al. (2020) introduced a bound based on the direct loss for linear models and two-layer neural networks. Mustafa et al. (2022) broadened this approach to encompass a wider array of attacks, applying their bounds directly to the loss function, resulting in a growth rate of $\mathcal{O}(\log K)$. Conversely, Khim & Loh (2018) proposed a tree-transform technique to propagate adversarial noise through the network, leading to a bound that grows exponentially as $\mathcal{O}(K)$, where K is the number of classes.

Adversarial Contrastive Learning Recently, numerous studies have applied adversarial training to contrastive learning loss to enhance model robustness. However, the theoretical foundations of ACL remain underexplored. Zou & Liu (2023) utilized Rademacher complexity to establish a bound on the ACL loss

for linear models and multi-layer neural networks under ℓ_p-attack and with only one negative sample. Their findings indicate that the average adversarial risk of downstream tasks can be upper-bounded by the adversarial unsupervised risk of the upstream task.

3 Problem Formulation

3.1 Contrastive Representation Learning

Let $\mathcal{X}$ be some input space (e.g., a set of input images). Given pairs of similar data, represented as (x, x^+) drawn from some unknown distribution $\mathcal{D}_{\text{sim}}$, and negative data samples $x_1^-, x_2^-, \cdots, x_K^-$ drawn from a negative distribution $\mathcal{D}_{\text{neg}}$. The distributions $\mathcal{D}_{\text{sim}}$ and $\mathcal{D}_{\text{neg}}$ are generally characterized through a set of latent classes $\mathcal{C}$ and an associated probability distribution ρ over these classes (Arora et al., 2019). For each latent class $c \in \mathcal{C}$, let $\mathcal{D}_c$ be the conditional distribution of the inputs given the latent class c. $\mathcal{D}_{\text{sim}}$ and $\mathcal{D}_{\text{neg}}$ are defined as:

$$\mathcal{D}_{\text{sim}}(x, x^+) = \mathbb{E}_{c\sim\rho}[\mathcal{D}_c(x)\mathcal{D}_c(x^+)],$$
$$\mathcal{D}_{\text{neg}}(x^-) = \mathbb{E}_{c\sim\rho}[\mathcal{D}_c(x^-)].$$

That is, $\mathcal{D}_{\text{sim}}(x, x^+)$ measures the probability of drawing x and x^+ from the same class $c \sim \rho$, while $\mathcal{D}_{\text{neg}}(x^-)$ measures the probability of drawing x^- that is irrelevant to x and x^+. The objective of CRL is to select a feature map $f : \mathcal{X} \to \mathbb{R}^d$ from a class of representation functions $\mathcal{F} = \{f : \|f(\cdot)\|_1 \leq R\}$, for some $R > 0$, where $\|\cdot\|_1$ denotes the ℓ_1-norm, and $d \in \mathbb{N}$ represents the dimensionality of the feature space. This is achieved using the training set

$$S = \{(x_1, x_1^+, x_{11}^-, \cdots, x_{1K}^-), (x_2, x_2^+, x_{21}^-, \cdots, x_{2K}^-), \cdots, (x_n, x_n^+, x_{n1}^-, \cdots, x_{nK}^-)\},$$

where $(x_j, x_j^+) \sim \mathcal{D}_{\text{sim}}$ and $(x_{j1}^-, \cdots, x_{jK}^-) \sim \mathcal{D}_{\text{neg}}^K$, with $j \in [n] := \{1, \cdots, n\}$ and K indicating the number of negative samples. The quality of the representation f is evaluated using the loss $\ell\left(\{f(x)^T(f(x^+) - f(x_k^-)\}_{k=1}^K\right)$, where $\ell : \mathbb{R}^K \to [0, B]$ is some loss function and $f(x)^T$ is the transpose of $f(x)$. The population and empirical risks are then defined as follows.

Definition 3.1 (Upstream unsupervised risk)**.** The population unsupervised risk is defined as:

$$L_{\text{un}}(f) = \mathbb{E}_D[\ell(\{f(x)^T(f(x^+) - f(x_k^-))\}_{k=1}^K)],$$

and the empirical unsupervised risk on S is defined as:

$$\widehat{L}_{\text{un}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(\{f(x_i)^T(f(x_i^+) - f(x_{ik}^-))\}_{k=1}^K)].$$

3.2 Adversarial Contrastive Representation Learning

In this paper, we examine adversarial settings where an attacker employs a noise function $A : \mathcal{X} \times \mathcal{B} \to \mathcal{X}$, with $\mathcal{B}$ being a noise set, to subtly introduce noise $\delta \in \mathcal{B}$ to an input $x \in \mathcal{X}$ in order to maximize the loss. For example, in the L_p-additive attack, $A(x, \delta) = x + \delta$ and $\mathcal{B}$ is the ℓ_p-ball $\{\delta : \|\delta\|_p \leq \beta\}$. The attacker's objective is to select $\delta^* \in \mathcal{B}$ that maximizes the loss:

$$\delta^* = \arg\max_{\delta\in\mathcal{B}} \ell(\{f(A(x, \delta))^T(f(x^+) - f(x_k^-))\}_{k=1}^k).$$

The adversarial and empirical risks are subsequently defined as follows.

Definition 3.2. (Unsupervised adversarial contrastive risk). The population unsupervised adversarial contrastive risk is defined as:

$$L_{\text{un}}^{\text{adv}}(f) = \mathbb{E}[\max_{\delta\in\mathcal{B}} \ell(\{f(A(x, \delta))^T(f(x^+) - f(x_k^-))\}_{k=1}^K)]$$

and the empirical unsupervised adversarial contrastive risk is defined as:

$$\hat{L}_{\text{un}}^{\text{adv}}(f) = \frac{1}{n}\sum_{i=1}^{n} \max_{\delta\in\mathcal{B}} \ell(\{f(A(x_i, \delta))^T(f(x_i^+) - f(x_{ik}^-))\}_{k=1}^K)]$$

Our goal is to derive a generalization bound for ACL, that is a bound on $L_{\text{un}}^{\text{adv}}(f) - \widehat{L}_{\text{un}}^{\text{adv}}(f)$.

4 Generalization Error Bounds

In this section, we establish a generalization bound for ACL. Our technique relies on the concept of *covering numbers* of the adversarial contrastive loss class.

Definition 4.1 (Covering number)**.** Let $\mathcal{F} := \{(f(x_1), \cdots, f(x_n)) : f \in \mathcal{F}\}$ be a real-valued function class defined over a vector space $\mathcal{V}$, and let $S := \{x_1, \cdots, x_n\} \subset \mathcal{X}^n$ be a dataset. For any $\epsilon > 0$, the ℓ_p-norm covering number, denoted as $\mathcal{N}_p(\epsilon, \mathcal{F}, S)$, is defined as the size of the smallest set of vectors $v_1, \cdots, v_m$ that covers $\mathcal{F}$. Specifically, it satisfies:

$$\sup_{f \in \mathcal{F}} \min_{j \in [m]} \Big(\frac{1}{n} \sum_{i \in [n]} |f(x_i) - v_j^i|^p \Big)^{\frac{1}{p}} \leq \epsilon,$$

where $v_1, \cdots, v_m$ forms the (ϵ, ℓ_p)-cover of $\mathcal{F}$ with respect to S. We define the worst-case covering number as $\mathcal{N}_p(\epsilon, \mathcal{F}, S) = \max_{S \in \mathcal{X}} \mathcal{N}_p(\epsilon, \mathcal{F}, S)$.

Our generalization analysis relies on the Lipschitz continuity of the loss. We consider a general Lipschitz continuity w.r.t. a p-norm.

Definition 4.2 (Lipschitz continuity)**.** A function $\ell : \mathbb{R}^C \to \mathbb{R}_+$ is said to be λ-Lipschitz continuous with respect to the ℓ_p-norm ($p \geq 1$) if for any $v, v' \in \mathbb{R}^C$, the following inequality holds:

$$|\ell(v) - \ell(v')| \leq \lambda \|v - v'\|_p.$$

In other words, a Lipschitz continuous function exhibits only a small change in its output when its inputs slightly change. We now proceed to derive the generalization bounds for ACL. The loss function class of interest is defined as follows:

$$\mathcal{G}_{\text{adv}} = \big\{(x, x^+, x_1^-, \ldots, x_K^-) \mapsto \max_{\delta \in \mathcal{B}} \ell(\{f(A(x, \delta)^T (f(x^+) - f(x_k^-))\}_{k=1}^K) : f \in \mathcal{F}\big\}.$$

The main challenge for the analysis of that class is due to the $\max_{\delta \in \mathcal{B}}$ operator. The following lemma shows a bound on the covering number of $\mathcal{G}_{\text{adv}}$ in terms of the covering number of an extended function class that does not contain the $\max_{\delta \in \mathcal{B}}$-operator and the loss function ℓ on f. The extended function class is defined as:

$$\mathcal{H} = \Big\{(x, x^+, x^-, \tilde{\delta}) \mapsto f(A(x, \tilde{\delta}))^T (f(x^+) - f(x^-)) : f \in \mathcal{F}\Big\},$$

That is, the functions in $\mathcal{H}$ are explicitly parameterized by the adversarial noise δ. Consequently, the data set is extended to:

$$S_{\mathcal{H}} = \big\{(x_i, x_i^+, x_{ik}^-, \tilde{\delta}) : i \in [n], k \in [K], \tilde{\delta} \in \mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})\big\},$$

where $\mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})$ is an $(\frac{\epsilon}{2\lambda_1}, \ell_\infty)$-cover of $\mathcal{B}$, for some $\epsilon, \lambda_1 > 0$. Specifically, for any $\delta \in \mathcal{B}$, there exists $\tilde{\delta} \in \mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})$ such that $\|\delta - \tilde{\delta}\|_\infty \leq \frac{\epsilon}{2\lambda_1}$. We now introduce our next lemma, which establishes a relationship between the covering number of $\mathcal{G}_{\text{adv}}$ on S and that of $\mathcal{H}$ on $S_{\mathcal{H}}$.

Lemma 4.1. Let $\delta \mapsto \ell(\{f(A(x, \delta))^T (f(x^+) - f(x_k^-))\}_{k=1}^K)$ be λ_1-Lipschitz and ℓ be λ_2-Lipschitz with respect to the ℓ_∞-norm, for all $(x, x^+, x_1^-, \ldots, x_K^-) \in \mathcal{X}^{K+2}$ and $f \in \mathcal{F}$. Then, we have:

$$\mathcal{N}_\infty(\epsilon, \mathcal{G}_{\text{adv}}, S) \leq \mathcal{N}_\infty\Big(\frac{\epsilon}{2\lambda_2}, \mathcal{H}, S_{\mathcal{H}}\Big).$$

The lemma shows that we can upper-bound the ℓ_∞ covering number of the ACL function class by that of the class $\mathcal{H}$ with the extended training set $S_{\mathcal{H}}$. Notably, the class $\mathcal{H}$ does not include the $\max_{\delta \in \mathcal{B}}$ operator, significantly simplifying the analysis. Furthermore, it shifts the dependence on the number of negative samples from the dimensionality of the function class's output to the size of the training set. For most classes, the dependence of the covering numbers on the size of the training set is only logarithmic (Zhang, 2002). Consequently, our bound will lead to a generalization bound that only has a logarithmic dependence on the number of negative samples.

While Lemma 4.1 provides an upper bound on the ACL class $\mathcal{G}_{\text{adv}}$ in terms of the non-adversarial class $\mathcal{H}$, the class $\mathcal{H}$ is not directly the representation function class $\mathcal{F}$. This makes it challenging to utilize existing covering number results for typical models (e.g., linear models (Zhang, 2002) or DNNs (Ledent et al., 2021b)). The following lemma establishes a relationship between the covering numbers of $\mathcal{H}$ and those of $\mathcal{F}$.

Lemma 4.2. Assume the previous conditions hold and that $\|f(\tilde{x})\|_1 \leq R$. We define the function class $\tilde{\mathcal{F}}$ as follows:

$$\tilde{\mathcal{F}} = \big\{(x, j) \mapsto f_j(x) : f \in \mathcal{F}, x \in \mathcal{X}, j \in [d]\big\}$$

over the training set $S_{\tilde{\mathcal{F}}}$:

$$S_{\tilde{\mathcal{F}}} = \{(A(x_i, \tilde{\delta}), j) : i \in [n], j \in [d], \tilde{\delta} \in \mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})\} \cup \{(x_{ik}^-, j) : i \in [n], k \in [K], j \in [d]\} \cup \{(x_i^+, j) : i \in [n], j \in [d]\}.$$

Then, we have:

$$\mathcal{N}_\infty\Big(\frac{\epsilon}{2\lambda_2}, \mathcal{H}, S_{\mathcal{H}}\Big) \leq \mathcal{N}_\infty\Big(\frac{\epsilon}{8R\lambda_2}, \tilde{\mathcal{F}}, S_{\tilde{\mathcal{F}}}\Big).$$

The lemma upper-bounds the covering number of $\mathcal{H}$ by the covering number of the class $\tilde{\mathcal{F}}$. Notably, $\tilde{\mathcal{F}}$ is a class of scalar-valued functions of the same form as the representation function class $\mathcal{F}$. This simplifies the analysis by (1) reducing the form of the functions in $\mathcal{H}$ to that of the representation class, and (2) simplifying the analysis from vector-valued functions to scalar functions of the same form. Note that the number of dimensions contributes only through the size of the dataset $S_{\tilde{\mathcal{F}}}$, and for many typical function classes, this contribution is only logarithmic. This achieves the best known rate for vector-valued functions (Lei et al., 2019). Combining Lemmas 4.2 and 4.1 with Dudley's integral gives our main result.

Theorem 4.1. Let $\delta \in (0, 1)$, and $\mathcal{F}$ and $\tilde{\mathcal{F}}$ be defined as above. With probability at least $1 - \delta$ over the randomness of the training data S with size n, we have for all $f \in \mathcal{F}$:

$$L_{\text{un}}^{\text{adv}}(f) \leq \hat{L}_{\text{un}}^{\text{adv}}(f) + \frac{3B\sqrt{\log(\frac{2}{\delta})}}{\sqrt{2n}} + \inf_{a>0}\Big(8a + \frac{24}{\sqrt{n}}\int_a^B \log^{\frac{1}{2}} \mathcal{N}_\infty\big(\frac{\epsilon}{8R\lambda_2}, \tilde{\mathcal{F}}, S_{\tilde{\mathcal{F}}}\big)d\epsilon\Big).$$

The theorem demonstrates that we can control the generalization error of ACL by controlling the covering number of the class $\tilde{\mathcal{F}}$. The covering numbers of many classes $\tilde{\mathcal{F}}$ (e.g., linear models (Zhang, 2002), MLPs (Bartlett et al., 2017), CNNs (Ledent et al., 2021b), and structured learning models (Mustafa et al., 2021)) can be directly applied here to derive generalization bounds for ACL across a large family of models.

5 Applications

To find the representations f in an unsupervised manner, we employ an unsupervised loss function $\ell : \mathbb{R}^K \mapsto \mathbb{R}_+$ which can be chosen to be a hinge loss. This loss function can be chosen as a hinge loss. Given a vector u, the hinge loss is defined as:

$$\ell(u) = \max\{0, 1 + \max_{i\in[K]}\{-u_i\}\}.$$

For simplicity, we assume the loss function is bounded by B for any $f \in \mathcal{F}$. Specifically, this means:

$$\ell(\{f(A(x, \delta)^T(f(x^+) - f(x_k^-))\}_{k=1}^K) \leq B, \quad \forall f \in \mathcal{F}.$$

This assumption is valid because we can impose constraints on the norms of the model's weights and inputs, ensuring the loss function remains bounded.

In this section, we instantiate our bound (Theorem 4.1) for two models: linear and DNN-based features. Throughout the section, we consider feature extractors of the form $x \mapsto U\mathbf{v}(x)$, where $U \in \mathbb{R}^{d\times d'}$ is a transformation matrix, and $\mathbf{v} : \mathcal{X} \mapsto \mathbb{R}^{d'}$ is a map from the original data $x \in \mathcal{X}$ to some intermediate embedding space in $\mathbb{R}^{d'}$. We consider the linear feature extractor in Section 5.1, while in Section 5.2, we explore features from a DNN.

5.1 Linear Features

First, we focus on the linear features. That is, we assume that $\mathbf{v}$ is the identity map.

We consider the ℓ_∞-attack, in which the attacker uses an additive noise function $A(x, \delta) = x + \delta$, for $x \in \mathcal{X}$ and $\delta \in \mathcal{B}$, where the noise set, $\mathcal{B}$, is the ℓ_∞-ball

$$\mathcal{B} = \left\{\delta : \|\delta\|_\infty \leq \beta\right\} \subset \mathbb{R}^D.$$

We begin by showing that the function $\delta \mapsto \ell((U\mathbf{v}(x+\delta))^T(U\mathbf{v}(x^+) - U\mathbf{v}(x_k^-))_{k=1}^K)$ is indeed Lipschitz. The following lemma establishes and quantifies the upper bound on the Lipschitz constant of $\delta \mapsto \ell((U(x+\delta))^T(U(x^+) - U(x_k^-))_{k=1}^K)$.

Lemma 5.1. Consider the function $g_U(x, \delta) = \ell((U(x+\delta))^T(U(x^+) - U(x_k^-))_{k=1}^K)$ and assume $\|U\|_2 \leq \Lambda_1$. Then, for any x, the function $\delta \mapsto g_U(x, \delta)$ is $\|\cdot\|_\infty$-Lipschitz with the Lipschitzness constant $2\Lambda_1^2\|x\|_2\sqrt{D}$.

Now that we have the Lipschitzness of δ on the loss function ℓ, we can bound the covering number of our linear scalar-valued feature class in the following lemma.

Lemma 5.2. Let $\tilde{\mathcal{F}}$ be the linear feature class and $S_{\tilde{\mathcal{F}}}$ be a given dataset in equation 4.2 with $\|x\|_2 \leq \Psi$, for all $x \in \mathcal{X}$, and $\|U\|_{2,2} \leq \Lambda$, then for all $\epsilon > 0$, we have

$$\log \mathcal{N}_\infty\Big(\frac{\epsilon}{8R\lambda_2}, \tilde{\mathcal{F}}, S_{\tilde{\mathcal{F}}}\Big) \leq \frac{CR^2\lambda_2^2\Lambda^2(\Psi + \sqrt{D}\beta)^2 L_{\log}}{\epsilon^2},$$

where C is an absolute constant, $m = |\mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})|$, $\Psi' = \Psi + \sqrt{D}\beta$ and

$$L_{\log} := \log\left(2\left\lceil\frac{4\Lambda\Psi'}{\epsilon} + 2\right\rceil(nmd + ndK + nd)\left(\frac{12\beta 2\Lambda_1^2\Psi\sqrt{D}}{\epsilon}\right)^D + 1\right).$$

If we plug Lemma 5.2 back into the Theorem 4.1, we get the following corollary.

Corollary 5.1. Assuming the above assumptions, for all $f \in \mathcal{F}$, with probability at least $1 - \delta$ over the training data, we have

$$L_{\text{un}}^{\text{adv}}(f) \leq \hat{L}_{\text{un}}^{\text{adv}}(f) + \frac{8}{n} + 3B\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{\sqrt{C}R\lambda_2\Lambda\Psi'\tilde{L}_{\log}}{\sqrt{n}},$$

where C is a constant, $\Psi' = \Psi + \sqrt{D}\beta$, $m = |\mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})|$, $N = nmd + ndK + nd$ and

$$\tilde{L}_{\log} := \log^{\frac{1}{2}}\left(4\left\lceil 2\Lambda\Psi' N + 1\right\rceil N\Big(12\beta 2\Lambda_1^2\|x\|_2\sqrt{D}N\Big)^D + 1\right)(\log(n) + \log(B)).$$

Remark. Our bound has a dependency on the square root of the input dimension, $\sqrt{D}$, in the term Ψ'. This arises due to the mismatch between the ℓ_2-norm of the input and the ℓ_∞-norm in the ball β, as encapsulated by the inequality $\|\delta\|_2 \leq \sqrt{D}\|\delta\|_\infty$. Additionally, the bound has a logarithmic dependence on the negative samples, K. The logarithmic dependency shows the appealing behavior of ACL for learning with a large number of negative examples.

5.2 Nonlinear Features

Now, we consider the covering numbers for learning the nonlinear features by DNNs. We say an activation function $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is positive-homogeneous if $\sigma(ax) = a\sigma(x)$ for $a > 0$, and is contracting if $|\sigma(x) - \sigma(x')| \leq |x - x'|$. The ReLU activation function $\sigma(x) = \max\{x, 0\}$ is both positive-homogeneous and contractive. Now assume the DNN feature map is defined as (removing matrix U for now),

$$\mathcal{V} = \{x \mapsto \mathbf{v}(x) = \sigma(V_L\sigma(V_{L-1}\cdots\sigma(V_1 x))) : \forall l \in [L]\}$$

Each layer $l \in [L]$ has the width of w_l, where $w_0 = D$ (the input dimension) and $w_L = d$ (the number of feature dimension).
Let $V \in \mathbb{V}$ be the weight of the network. Suppose that $\mathbb{V}$ is such that, for all $V \in \mathbb{V}$, $\|V_l\|_2 \leq a_l$ and $\|V_l\|_\sigma \leq s_l$ for all $l \in [L-1]$. Further, suppose that, for all $V \in \mathbb{V}$, $\|V_L\|_2 \leq a_L$, $\|V_L\|_{2,\infty} \leq s_L$ and $\|V_1\|_{1,\infty} \leq s_1'$.

We now consider the ℓ_∞-additive perturbation applied to the DNN. As with the linear case, we first establish the Lipschitzness of the function $\delta \mapsto \ell((U\mathbf{v}(x+\delta))^T(U\mathbf{v}(x^+) - U\mathbf{v}(x_k^-))_{k=1}^K)$ w.r.t. $\|\cdot\|_\infty$-norm. The following lemma establishes the Lipschitz continuity of the loss as a function of δ.

Lemma 5.3. Consider the function $g_{UV}(x,\delta) = \ell((U\mathbf{v}(x+\delta))^T(U\mathbf{v}(x^+) - U\mathbf{v}(x^-))_{k=1}^K)$ and assume $\|U\|_2 \leq \Lambda_1$ and $\mathbf{v}(\cdot)$ is the neural network. Then, for any $x \in \mathcal{X}$ and $V \in \mathbb{V}$, the function $\delta \mapsto g_{UV}\mathbf{v}(x,\delta)$ is $\|\cdot\|_\infty$-Lipschitz with constant $2\Lambda_1^2\sqrt{w_L}\prod_{l=2}^L s_l\sqrt{w_1}s_1'\prod_{l=1}^L s_l\|x\|_2$.

Now, we can get the upper bound of the covering number of the neural network scalar-valued feature class w.r.t. $\|.\|_\infty$-norm.

Lemma 5.4. Let $\tilde{\mathcal{F}}$ be the DNN (nonlinear) feature class on the extended dataset $S_{\tilde{\mathcal{F}}}$, defined as before. Let $\mathcal{B} := \{\delta : \|\delta\|_\infty \leq \beta\}$. Assume the previous assumptions on the weights of DNN, and $\|x\|_2 \leq \Psi$. Then, for $S_{\tilde{\mathcal{F}}}$ and $\epsilon > 0$, we have

$$\log \mathcal{N}_\infty\Big(\frac{\epsilon}{8R\lambda_2}, \tilde{\mathcal{F}}, S_{\tilde{\mathcal{F}}}\Big) \leq \frac{CR^2\lambda_2^2L^2\Psi'^2}{\epsilon^2}\prod_{l=1}^L s_l^2\left(\sum_{l=1}^L \frac{a_l^2}{s_l^2}\right)L_{\log},$$

where

$$L_{\log} := \log\left(\left(C_1\Psi'\Gamma/\epsilon + C_2\bar{w}\right)(nmd + ndK + nd)\left(\frac{12\beta\Lambda_1^2\sqrt{w_L}\prod_{l=2}^L s_l\sqrt{w_1}s_1'\prod_{l=1}^L s_l\|x\|_2}{\epsilon}\right)^D + 1\right),$$

$\Psi' = (\Psi + \sqrt{D}\beta)$, $\Gamma = \max_{l\in[L]}(\prod_{i=1}^L s_i)a_lw_l/s_l$, $\bar{w} = \max_{l\in[L]} w_l$, $m = |\mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})|$, and C, C_1, C_2 are universal constants.

Plugging Lemma 5.4 into Theorem 4.1, we get the following corollary.

Corollary 5.2. Under the previous assumptions, for all $f \in \mathcal{F}$, with probability at least $1-\delta$ over the training data, we have

$$L_{\text{un}}^{\text{adv}}(f) \leq \hat{L}_{\text{un}}^{\text{adv}}(f) + \frac{8}{n} + 3\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{CR\lambda_2L\Psi'}{\sqrt{n}}\prod_{l=1}^L s_l\sqrt{\left(\sum_{l=1}^L\frac{a_l^2}{s_l^2}\right)}\tilde{L}_{\log},$$

where C is an absolute constant, $\Psi' = \Psi + \sqrt{D}\beta$, $m = |\mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})|$, $N = nmd + ndK + nd$ and

$$\tilde{L}_{\log} = \log^{\frac{1}{2}}\left(\left(C_1\Psi'\Gamma N + C_2\bar{w}\right)N\left(12\beta 2\Lambda_1^2\sqrt{w_L}\prod_{l=2}^L s_l\sqrt{w_1}s_1'\prod_{l=1}^L s_l\|x\|_2N\right)^D + 1\right)(\log(n) + \log(B)).$$

Remark. Similar to the results for the linear case, our analysis reveals a dependency on the square root of the input dimension. This issue can be resolved if we assume the ℓ_2 attack, where $\mathcal{B} = \{\delta : \|\delta\|_2 \leq \beta\}$. As in the linear case, our results maintain a logarithmic dependence on the negative samples, K.

6 Proofs

6.1 Proofs of Results in Section 4

To discretize the function space, we assume $\delta \mapsto \ell(\{f(A(x,\cdot))^T(f(x^+) - f(x_k^-))\}_{k=1}^K)$ is λ_1-Lipschitz for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$. Let $\mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})$ be an $(\frac{\epsilon}{2\lambda_1}, \ell_\infty)$-cover of $\mathcal{B}$. Assume the discretized class of loss function is $\tilde{\mathcal{G}}_{adv}$, defined as below:

$$\tilde{\mathcal{G}}_{\text{adv}} = \Big\{(x, x^+, x_1^-, \ldots, x_K^-, \delta) \to \ell(\{f(A(x,\delta))^T(f(x^+) - f(x_k^-))\}_{k=1}^K) : f \in \mathcal{F}\Big\}$$

with an extended training set $\tilde{S}$:

$$\tilde{S} = \Big\{(x_i, x_i^+, x_{i1}^-, \ldots, x_{iK}^-, \tilde{\delta}) : i \in [n], \tilde{\delta} \in \mathcal{C}_{\mathcal{B}}\big(\epsilon/(2\lambda_1)\big)\Big\}.$$

Now, we review a lemma from Mustafa et al. (2022) to relate the ℓ_∞-covering number of class $\mathcal{G}_{\text{adv}}$ to the covering number of the discretized version $\tilde{\mathcal{G}}_{adv}$.

Lemma 6.1 (Mustafa et al. 2022)**.** Let $\tilde{\mathcal{G}}_{\text{adv}}$ be defined as above. Then, the following holds:

$$\mathcal{N}_\infty(\epsilon, \mathcal{G}_{\text{adv}}, S) \leq \mathcal{N}_\infty(\epsilon/2, \tilde{\mathcal{G}}_{\text{adv}}, \tilde{S}).$$

This lemma simplifies the complexity of our function class by discretizing the loss function according to δ.

Now, we can prove Lemma 4.1:

Proof of Lemma 4.1. According to Lemma 6.1, it suffices to show that

$$\mathcal{N}_\infty(\epsilon/2, \tilde{\mathcal{G}}_{\text{adv}}, \tilde{S}) \leq \mathcal{N}_\infty(\epsilon/(2\lambda_2), \mathcal{H}, S_{\mathcal{H}}).$$

The observation here is that we can construct a cover for the function class $\tilde{\mathcal{G}}_{\text{adv}}$ on the training set $\tilde{S}$ from the elements of the cover of the function class $\mathcal{H}$. Additionally, from the Definition 4.1, we know that the covering number of a set is the cardinality of the smallest cover for a set.
For any f, define h_f as

$$h_f(x, x^+, x^-, \tilde{\delta}) = f(A(x, \tilde{\delta}))^T (f(x^+) - f(x^-)).$$

Let $\mathcal{C}_B(\epsilon/(2\lambda_1)) = \{\delta_1, \ldots, \delta_m\}$. The projection of $\mathcal{H}$ onto the set $S_{\mathcal{H}}$ is

$$\mathcal{H}_{S_{\mathcal{H}}} := \left\{ \begin{bmatrix} h_f(x_1, x_1^+, x_{11}^-, \delta_1) & \ldots & h_f(x_1, x_1^+, x_{11}^-, \delta_m) \\ \vdots & \ddots & \vdots \\ h_f(x_n, x_n^+, x_{nK}^-, \delta_1) & \ldots & h_f(x_n, x_n^+, x_{nK}^-, \delta_m) \end{bmatrix} : f \in \mathcal{F} \right\} \subset \mathbb{R}^{nK \times m}.$$

Let

$$\mathcal{C}_{\mathcal{H}} := \left\{ \begin{bmatrix} c_{i'}^{11}(\delta_1) & \ldots & c_{i'}^{11}(\delta_m) \\ \vdots & \ddots & \vdots \\ c_{i'}^{nK}(\delta_1) & \ldots & c_{i'}^{nK}(\delta_m) \end{bmatrix} : i' = 1, \ldots, M \right\} \subset \mathbb{R}^{nK \times m}$$

be an $(\epsilon/(2\lambda_2), \ell_\infty)$-cover of $\mathcal{H}_{S_{\mathcal{H}}}$. This means, for all $f \in \mathcal{F}$, there exists an $r \in [M]$, such that:

$$\max_{i \in [n]} \max_{k \in [K]} \max_{\delta \in \mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})} \left| h_f(x_i, x_i^+, x_{ik}^-, \delta) - c_r^{ik}(\delta) \right| \leq \frac{\epsilon}{2\lambda_2}.$$

Now we show the following set is an $(\epsilon/2, \ell_\infty)$-cover of $\tilde{\mathcal{G}}_{adv}$ w.r.t. $\widetilde{S}$:

$$\mathcal{C}_{\tilde{\mathcal{G}}_{adv}} := \left\{ \begin{bmatrix} \ell(\{c_{i'}^{1k}(\delta_1)\}_{k=1}^K) & \ldots & \ell(\{c_{i'}^{1k}(\delta_m)\}_{k=1}^K) \\ \vdots & \ddots & \vdots \\ \ell(\{c_{i'}^{nk}(\delta_1)\}_{k=1}^K) & \ldots & \ell(\{c_{i'}^{nk}(\delta_m)\}_{k=1}^K) \end{bmatrix} : i' = 1, \ldots, M \right\} \subset \mathbb{R}^{n \times m}.$$

Indeed, for any $f \in \mathcal{F}$, we know

$$\begin{aligned}
&\max_{i \in [n]} \max_{\delta \in \mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})} \left| \ell\big(\{f(A(x_i, \delta)^T (f(x_i^+) - f(x_{i,k}^-)))\}_{k=1}^K\big) - \ell(\{c_r^{ik}(\delta)\}_{k=1}^K) \right| \\
&= \max_{i \in [n]} \max_{\delta \in \mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})} \left| \ell(\{h_f(x_i, x_i^+, x_{ik}^-, \delta)\}_{k=1}^K) - \ell(\{c_r^{ik}(\delta)\}_{k=1}^K) \right| \\
&\leq \max_{i \in [n]} \max_{\delta \in \mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})} \left| \max_{k \in [K]} \ell(h_f(x_i, x_i^+, x_{ik}^-, \delta)) - \max_{k \in [K]} \ell(c_r^{ik}(\delta)) \right| \\
&\leq \max_{i \in [n]} \max_{\delta \in \mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})} \max_{k \in [K]} \left| \ell(h_f(x_i, x_i^+, x_{ik}^-, \delta)) - \ell(c_r^{ik}(\delta)) \right| \\
&\leq \lambda_2 \max_{i \in [n]} \max_{\delta \in \mathcal{C}_{\mathcal{B}}(\frac{\epsilon}{2\lambda_1})} \max_{k \in [K]} \left| h_f(x_i, x_i^+, x_{ik}^-, \delta) - c_r^{ik}(\delta) \right| \leq \lambda_2 \frac{\epsilon}{2\lambda_2} = \frac{\epsilon}{2}.
\end{aligned}$$

The second inequality comes from $|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$ and the third inequality follows from the λ_2-Lipschitzness of the loss function ℓ. Since the cardinality of $\mathcal{C}_{\mathcal{H}}$ and $\mathcal{C}_{\tilde{\mathcal{G}}_{adv}}$ are the same, we have $\mathcal{N}_\infty(\frac{\epsilon}{2}, \tilde{\mathcal{G}}_{adv}, \tilde{S}) \leq \mathcal{N}_\infty(\frac{\epsilon}{2\lambda_2}, \mathcal{H}, S_{\mathcal{H}})$. □

Now, we are going to prove Lemma 4.2.

Proof of Lemma 4.2. Our goal is to control the ℓ_∞-covering number of the function class $\mathcal{H}$ using the covering number of the representation function class $\mathcal{F}$. We will prove the lemma in two parts.
First, we claim to show:

$$\mathcal{N}_\infty\Big(\frac{\epsilon}{2\lambda_2}, \mathcal{H}, S_{\mathcal{H}}\Big) \leq \mathcal{N}_\infty\Big(\frac{\epsilon}{8R\lambda_2}, \mathcal{F}, S_{\mathcal{F}}\Big),$$

where we introduce $S_{\mathcal{F}}$ as follows:

$$S_{\mathcal{F}} = \{\tilde{x}_j : j \in [nm+nK+n]\} = \{A(x_i, \tilde{\delta}) : i \in [n], \tilde{\delta} \in \mathcal{C}_{\mathcal{B}}(\epsilon/(2\lambda_1))\} \cup \{x_{ik}^- : i \in [n], k \in [K]\} \cup \{x_i^+ : i \in [n]\}$$

and $m = |\mathcal{C}_{\mathcal{B}}(\epsilon/2\lambda_1)|$. Consider the following class of functions defined on $S_{\mathcal{F}}$:

$$\mathcal{F}_{S_{\mathcal{F}}} := \Big\{\big(f(\tilde{x}_1), \ldots, f(\tilde{x}_{nm+nK+n})\big) : f \in \mathcal{F}\Big\} \subset \mathbb{R}^{nm+nK+n},$$

which can be expanded as:

$$\{(f(A(x_1, \delta_1)), \ldots, f(A(x_n, \delta_m)), f(x_1^+), \ldots, f(x_n^+), f(x_{11}^-), \ldots, f(x_{1K}^-), f(x_{n1}^-), \ldots, f(x_{nK}^-))\}.$$

Suppose the function class $\mathcal{F}_{S_{\mathcal{F}}}$ has a proper $(\epsilon/(8R\lambda_2), \ell_\infty)$-cover as below

$$\mathcal{C}_{\mathcal{F}} := \big\{(\tilde{c}_{i'}^1(\tilde{\delta}_1), \ldots, \tilde{c}_{i'}^n(\tilde{\delta}_m), \tilde{c}_{i'}^{1+}, \ldots, \tilde{c}_{i'}^{n+}, \tilde{c}_{i'}^{11-}, \ldots, \tilde{c}_{i'}^{nK-}) : i' \in [M]\big\} \subset \mathbb{R}^{nm+nK+n}.$$

Then for all $f \in \mathcal{F}$, there exists an $r \in [M]$ such that:

$$\begin{aligned}
\max_{i\in[n]} \max_{a\in[m]} \big|f(A(x_i, \delta_a)) - \tilde{c}_r^i(\tilde{\delta}_a)\big| &\leq \frac{\epsilon}{8R\lambda_2}, \\
\max_{i\in[n]} \big|f(x_i^+) - \tilde{c}_r^{i+}\big| &\leq \frac{\epsilon}{8R\lambda_2}, \\
\max_{i\in[n]} \max_{k\in[K]} \big|f(x_{ik}^-) - \tilde{c}_r^{ik-}\big| &\leq \frac{\epsilon}{8R\lambda_2}.
\end{aligned}$$

Now, for the following function class

$$\mathcal{H}_{S_{\mathcal{H}}} := \Big\{(f(A(x_1, \delta_1))^T(f(x_1^+) - f(x_{11}^-)), \ldots, f(A(x_n, \delta_m))^T(f(x_n^+) - f(x_{nK}^-)))\Big\} \subset \mathbb{R}^{nKm}$$

projected onto the dataset $S_{\mathcal{F}}$, we construct a cover as follows

$$\mathcal{C}_{\mathcal{H}} := \Big\{(c_{i'}^1(\tilde{\delta}_1), \ldots, c_{i'}^{nK}(\tilde{\delta}_m)) : i' \in [M]\Big\} \subset \mathbb{R}^{nKm},$$

where $c_{i'}^{ik}(\tilde{\delta}_a) = \tilde{c}_{i'}^{i}(\tilde{\delta}_a)^T(\tilde{c}_{i'}^{i+} - \tilde{c}_{i'}^{ik-})$. We then have:

$$
\begin{aligned}
&\max_{i\in[n]}\max_{k\in[K]}\max_{a\in[m]}\left|f(A(x_i,\tilde{\delta}_a))^T(f(x_i^+)-f(x_{ik}^-)) - c_r^{ik}(\tilde{\delta}_a)\right| \\
&= \max_{i\in[n]}\max_{k\in[K]}\max_{a\in[m]}\left|f(A(x_i,\tilde{\delta}_a))^T(f(x_i^+)-f(x_{ik}^-)) - \tilde{c}_r^{iT}(\tilde{\delta}_a)(\tilde{c}_r^{i+} - \tilde{c}_r^{ik-})\right| \\
&= \max_{i\in[n]}\max_{k\in[K]}\max_{a\in[m]}\left|f(A(x_i,\tilde{\delta}_a))^T(f(x_i^+)-f(x_{ik}^-)) - f(A(x_i,\tilde{\delta}_a))^T(\tilde{c}_r^{i+} - \tilde{c}_r^{ik-})\right. \\
&\qquad \left. + f(A(x_i,\tilde{\delta}_a))^T(\tilde{c}_r^{i+} - \tilde{c}_r^{ik-}) - \tilde{c}_r^{iT}(\tilde{\delta}_a)(\tilde{c}_r^{i+} - \tilde{c}_r^{i-})\right| \\
&\le \max_{i\in[n]}\max_{k\in[K]}\max_{a\in[m]}\Big(\left|f(A(x_i,\tilde{\delta}_a))^T(f(x_i^+)-f(x_{ik}^-)) - f(A(x_i,\tilde{\delta}_a))^T(\tilde{c}_r^{i+} - \tilde{c}_r^{ik-})\right| \\
&\qquad + \left|f(A(x_i,\tilde{\delta}_a))^T(\tilde{c}_r^{i+} - \tilde{c}_r^{ik-}) - \tilde{c}_r^{iT}(\tilde{\delta}_a)(\tilde{c}_r^{i+} - \tilde{c}_r^{ik-})\right|\Big) \\
&\le \max_{i\in[n]}\max_{k\in[K]}\max_{a\in[m]}\left|f(A(x_i,\tilde{\delta}_a))^T(f(x_i^+) - \tilde{c}_r^{i+} - f(x_{ik}^-) + \tilde{c}_r^{ik-})\right| \\
&\qquad + \max_{i\in[n]}\max_{k\in[K]}\max_{a\in[m]}\left|(\tilde{c}_r^{i+} - \tilde{c}_r^{ik-})^T(f(A(x_i,\tilde{\delta}_a)) - \tilde{c}_r^{i}(\tilde{\delta}_a))\right| \\
&\le \max_{i\in[n]}\max_{a\in[m]}\|f(A(x_i,\tilde{\delta}_a))\|_1 \max_{i\in[n]}\max_{k\in[K]}\|f(x_i^+) - \tilde{c}_r^{i+} - f(x_{ik}^-) + \tilde{c}_r^{ik-}\|_\infty \\
&\qquad + \max_{i\in[n]}\max_{k\in[K]}\|\tilde{c}_r^{i+} - \tilde{c}_r^{ik-}\|_1 \max_{i\in[n]}\max_{a\in[m]}\|f(A(x_i,\tilde{\delta}_a)) - \tilde{c}_r^{i}(\tilde{\delta}_a)\|_\infty \\
&\le 2R\frac{\epsilon}{8R\lambda_2} + 2R\frac{\epsilon}{8R\lambda_2} = 4R\frac{\epsilon}{8R\lambda_2} = \frac{\epsilon}{2\lambda_2}
\end{aligned}
$$

Here, the third inequality uses the property that $\left|x^T y\right| \le \|x\|_1\|y\|_\infty$. Since the cardinality of $\mathcal{C}_\mathcal{F}$ and $\mathcal{C}_\mathcal{H}$ are the same, we have: $\mathcal{N}_\infty(\frac{\epsilon}{2\lambda_2}, \mathcal{H}, S_\mathcal{H}) \le \mathcal{N}_\infty(\frac{\epsilon}{8R\lambda_2}, \mathcal{F}, S_\mathcal{F})$.

For the second part of the proof, we need to show:

$$
\mathcal{N}_\infty\Big(\frac{\epsilon}{8R\lambda_2}, \mathcal{F}, S_\mathcal{F}\Big) \le \mathcal{N}_\infty\Big(\frac{\epsilon}{8R\lambda_2}, \tilde{\mathcal{F}}, S_{\tilde{\mathcal{F}}}\Big).
$$

We introduce $S_{\tilde{\mathcal{F}}}$:

$$
\begin{aligned}
S_{\tilde{\mathcal{F}}} &= \{(\tilde{x}, l) : l \in [ndm + ndK + nd]\} \\
&= \big\{(A(x_i,\tilde{\delta}), j) : i \in [n], j \in [d], \tilde{\delta} \in \mathcal{C}_\mathcal{B}(\frac{\epsilon}{2\lambda_1})\} \cup \{(x_{ik}^-, j) : i \in [n], k \in [K], j \in [d]\} \cup \{(x_i^+, j) : i \in [n], j \in [d]\big\}.
\end{aligned}
$$

Assume

$$
\tilde{\mathcal{F}} = \{(x, j) \mapsto f_j(x) : f \in \mathcal{F}, x \in \mathcal{X}, j \in [d]\}
$$

over $S_{\tilde{\mathcal{F}}}$ has a $(\epsilon/(8R\lambda_2), \ell_\infty)$-cover defined as below:

$$
\mathcal{C}_{\tilde{\mathcal{F}}} := \big\{(\tilde{c}_{i'}^{11}(\tilde{\delta}_1), \ldots, \tilde{c}_{i'}^{nd}(\tilde{\delta}_m), \tilde{c}_{i'}^{11+}, \ldots, \tilde{c}_{i'}^{nd+}, \tilde{c}_{i'}^{111-}, \ldots, \tilde{c}_{i'}^{nKd-}) : i' \in [M]\big\} \subset \mathbb{R}^{ndm+ndK+nd}
$$

This means the projection of $\tilde{\mathcal{F}}$ on the extended dataset $S_{\tilde{\mathcal{F}}}$ is:

$$
\tilde{\mathcal{F}}_{S_{\tilde{\mathcal{F}}}} = \{(f_1(A(x_1,\tilde{\delta}_1)), \ldots, f_d(A(x_n,\tilde{\delta}_m)), f_1(x_1^+), \ldots, f_d(x_n^+), f_1(x_{11}^-), \ldots, f_d(x_{nK}^-))\} \subset \mathbb{R}^{ndm+ndK+nd}
$$

Then for all $f \in \tilde{\mathcal{F}}$, there exists an $r \in [M]$, such that:

$$
\begin{aligned}
\max_{i\in[n]}\max_{a\in[m]}\max_{j\in[d]}\left|f_j(A(x_i,\delta_a)) - \tilde{c}_r^{ij}(\tilde{\delta}_a)\right| &\le \frac{\epsilon}{8R\lambda_2}, \\
\max_{i\in[n]}\max_{j\in[d]}\left|f_j(x_i^+) - \tilde{c}_r^{ij+}\right| &\le \frac{\epsilon}{8R\lambda_2}, \\
\max_{i\in[n]}\max_{k\in[K]}\max_{j\in[d]}\left|f_j(x_{ik}^-) - \tilde{c}_r^{ikj-}\right| &\le \frac{\epsilon}{8R\lambda_2}.
\end{aligned}
$$

Now, for the function class,

$$\mathcal{F}_{S_{\mathcal{F}}} := \{(f(A(x_1,\delta_1)),\dots,f(A(x_n,\delta_m)),f(x_1^+),\dots,f(x_n^+),f(x_{11}^-),\dots,f(x_{1K}^-),f(x_{n1}^-),\dots,f(x_{nK}^-))\},$$

we construct a cover:

$$\mathcal{C}_{\mathcal{F}} := \left\{(c_{i'}^1(\tilde{\delta}_1),\dots,c_{i'}^n(\tilde{\delta}_1),c_{i'}^{1+},\dots,c_{i'}^{n+},c_{i'}^{11-},\dots,c_{i'}^{nK-}) : i' \in [M]\right\} \subset \mathbb{R}^{nm+nK+n},$$

where $c_{i'}^i(\tilde{\delta}_a) = (\tilde{c}_{i'}^{i1}(\tilde{\delta}_a),\dots,\tilde{c}_{i'}^{id}(\tilde{\delta}_a))^T$, $c_{i'}^{i+} = (\tilde{c}_{i'}^{i1+},\dots,\tilde{c}_{i'}^{id+})^T$, and $c_{i'}^{ik-} = (\tilde{c}_{i'}^{ik1-},\dots,\tilde{c}_{i'}^{ikd-})^T$. Therefore,

$$\begin{aligned}\max_{i\in[n]}\max_{a\in[m]}\|f(A(x_i,\tilde{\delta}_a)) - c_r^i(\tilde{\delta}_a)\|_\infty &= \max_{i\in[n]}\max_{a\in[m]}\left|\max_{j\in[d]} f_j(A(x_i,\tilde{\delta}_a)) - \max_{j\in[d]}\tilde{c}_r^{ij}(\tilde{\delta}_a)\right| \\ &\le \max_{i\in[n]}\max_{a\in[m]}\max_{j\in[d]}\left|f_j(A(x_i,\tilde{\delta}_a)) - \tilde{c}_r^{ij}\right| \le \frac{\epsilon}{8R\lambda_2},\end{aligned}$$

$$\begin{aligned}\max_{i\in[n]}\|f(x_i^+) - c_r^{i+}\|_\infty &= \max_{i\in[n]}\left|\max_{j\in[d]} f_j(x_i^+) - \max_{j\in[d]}\tilde{c}_r^{ij+}\right| \\ &\le \max_{i\in[n]}\max_{j\in[d]}\left|f_j(x_i^+) - \tilde{c}_r^{ij}\right| \le \frac{\epsilon}{8R\lambda_2},\end{aligned}$$

$$\begin{aligned}\max_{i\in[n]}\max_{k\in[K]}\|f(x_{ik}^-) - c_r^{ik-}\|_\infty &= \max_{i\in[n]}\max_{k\in[K]}\left|\max_{j\in[d]} f_j(x_{ik}^-) - \max_{j\in[d]}\tilde{c}_r^{ikj}\right| \\ &\le \max_{i\in[n]}\max_{k\in[K]}\max_{j\in[d]}\left|f_j(x_{ik}^-) - \tilde{c}_r^{ikj}\right| \le \frac{\epsilon}{8R\lambda_2}.\end{aligned}$$

The second inequality for all the three equations comes from $|\max_x f(x) - \max_x g(x)| \le \max_x |f(x) - g(x)|$ and the cardinality of $\mathcal{C}_{\tilde{\mathcal{F}}}$ and $\mathcal{C}_{\mathcal{F}}$ are the same. It then follows that $\mathcal{C}_{\mathcal{F}}$ is an $(\epsilon/(8R\lambda_2), \ell_\infty)$-cover to $\mathcal{F}$. Thus, we have:

$$\mathcal{N}_\infty\Big(\frac{\epsilon}{8R\lambda_2}, \mathcal{F}, S_{\mathcal{F}}\Big) \le \mathcal{N}_\infty\Big(\frac{\epsilon}{8R\lambda_2}, \tilde{\mathcal{F}}, S_{\tilde{\mathcal{F}}}\Big).$$

The proof is completed. □

To prove Theorem 4.1, we define the Rademacher complexity and review the theorem 6.1 from Mohri et al. (2018), which controls generalization of learning algorithms by Rademacher complexity of function classes.

Definition 6.1 (Rademacher complexity)**.** Given a class of real-valued functions $\mathcal{F}$ and dataset $S = \{z_i\}_{i=1}^m$ drawn from the distribution $\mathcal{D}$ over a space $\mathcal{Z}$, the empirical Rademacher complexity of $\mathcal{F}$ w.r.t. S is defined as $\mathfrak{R}_S = \mathbb{E}_\epsilon[\sup_{f\in\mathcal{F}} \frac{1}{m}\Sigma_{i\in[m]}\epsilon_i f(z_i)]$, where each ϵ_i is an independent Rademacher variable, uniformly distributed over $\{+1,-1\}^m$. The worse-case Rademacher complexity is then $\mathfrak{R}_{\mathcal{Z},m} = \sup_{S\subset\mathcal{Z}:|S|=m}[\mathfrak{R}_S(\mathcal{F})]$.

Theorem 6.1 (Mohri et al. 2018)**.** Let $S = \{z_i\}_{i=1}^m$ be i.i.d. random sample from a distribution $\mathcal{D}$ defined over $\mathcal{Z}$. Further let $\mathcal{F} \subset [0,1]^{\mathcal{Z}}$ be a loss class. Then for all $\delta \in (0,1)$, we have with probability at least $1-\delta$ over the draw of the sample S, for all $f \in \mathcal{F}$ that

$$R(f) \le \hat{R}(f) + 2\mathfrak{R}_S(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Our approach relies, however, on another complexity measure, namely ℓ_∞-covering numbers. The following classical result of Dudley's entropy integral (Boucheron et al., 2003; Bartlett et al., 2017; Ledent et al., 2021a; Srebro et al., 2010) gives a relationship between the Rademacher complexity and ℓ_∞-covering number. We apply the version by Srebro et al. (2010).

Theorem 6.2 (Srebro et al. 2010)**.** Let $\mathcal{F}$ be a class of functions mapping from a space $\mathcal{Z}$ and taking values in $[0,b]$, and assume that $0 \in \mathcal{F}$. Let S be a finite sample of size m and $\hat{\mathbb{E}}[f(z)^2] := \frac{1}{m}\sum_{i=1}^m f(z_i)^2$. Then

$$\mathfrak{R}(\mathcal{F}) \le \inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^{\sup_{f\in\mathcal{F}}\sqrt{\hat{\mathbb{E}}[f(z)^2]}} \sqrt{\log \mathcal{N}_2(\epsilon,\mathcal{F},S)}d\epsilon\right).$$

We are now ready to present the proof of Theorem 4.1.

Proof of Theorem 4.1. The proof is a direct application of Theorems 6.2 and 6.1. With probability at least $1-\delta$, for all $f \in \mathcal{F}$ and $\delta \in (0,1)$, we have,

$$\begin{aligned}
L_{\text{un}}^{\text{adv}}(f) &\leq \hat{L}_{\text{un}}^{\text{adv}}(f) + 2\mathfrak{R}_S(\mathcal{G}_{\text{adv}}) + 3B\sqrt{\frac{\log(2/\delta)}{2n}} \\
&\leq \hat{L}_{\text{un}}^{\text{adv}}(f) + \inf_{\alpha>0}\left(8\alpha + \frac{24}{\sqrt{n}}\int_{\alpha}^{B}\sqrt{\log \mathcal{N}_\infty(\epsilon, \mathcal{G}_{\text{adv}}, S)}d\epsilon\right) + 3B\sqrt{\frac{\log(2/\delta)}{2n}} \\
&\leq \hat{L}_{\text{un}}^{\text{adv}}(f) + \inf_{\alpha>0}\left(8\alpha + \frac{24}{\sqrt{n}}\int_{\alpha}^{B}\sqrt{\log \mathcal{N}_\infty(\frac{\epsilon}{8R\lambda_2}, \tilde{\mathcal{F}}, S_{\tilde{\mathcal{F}}})}d\epsilon\right) + 3B\sqrt{\frac{\log(2/\delta)}{2n}}.
\end{aligned}$$

The first and the second inequality follow from Theorem 6.1 and Theorem 6.2, respectively. The final inequality is derived from Lemmas 4.1 and 4.2. □

6.2 Proofs of Results in Section 5.1

In this subsection, we present the omitted proofs of section 5.1 when the features are linear. Our approach relies on ℓ_∞-covering numbers for the feature classes. First, we show that the loss function is ℓ_∞-Lipschtiz with respect to the noise parameter δ. Next, we derive a bound on the size of the set $\mathcal{C}_{\mathcal{B}}(\epsilon/2\lambda_1)$. Finally, we establish a bound on the ℓ_∞-covering number of the feature class $\tilde{\mathcal{F}}$ on the extended data set $S_{\tilde{\mathcal{F}}}$.

Initially, we prove the bounds of ℓ_∞-additive attacks applied to linear models as stated in Lemma 5.1. Our first step is to derive the $\|\cdot\|_\infty$-Lipschitz constant of the function $\delta \mapsto \ell(\{(U(x+\delta))^T(U(x^+) - U(x_k^-))\}_{k=1}^K)$. The following is the proof of Lemma 5.1.

Proof of Lemma 5.1. The proof is a direct derivation. For all $x \in \mathcal{X}$, $\|U\|_2 \leq \Lambda_1$, $\delta, \delta' \in \mathcal{B}$, we have:

$$\begin{aligned}
&|\ell(\{(U(x+\delta))^T(U(x^+) - U(x_k^-))\}_{k=1}^K) - \ell(\{(U(x+\delta'))^T(U(x^+) - U(x_k^-))\}_{k=1}^K)| \\
&\leq |\ell((U(x+\delta))^T(U(x^+) - U(x^-))) - \ell((U(x+\delta'))^T(U(x^+) - U(x^-)))| \\
&\leq |(U(x+\delta))^T(U(x^+) - U(x^-)) - (U(x+\delta'))^T(U(x^+) - U(x^-))| \\
&\leq \|U(x+\delta) - U(x+\delta')\|_2\|U(x^+) - U(x^-)\|_2 \leq \|U(\delta-\delta')\|_2\|U(x^+ - x^-)\|_2 \\
&\leq 2\|U\|_2\sqrt{D}\|\delta-\delta'\|_\infty\|U\|_2\|x\|_2 \leq 2\Lambda_1^2\|x\|_2\sqrt{D}\|\delta-\delta'\|_\infty.
\end{aligned}$$

The second inequality is derived from the fact that hinge loss ℓ is ℓ_∞-Lipschitz with constant 1. The last inequality follows from $\|\delta-\delta'\|_2 \leq \sqrt{D}\|\delta-\delta'\|_\infty$, for all $\delta, \delta' \in \mathbb{R}^D$. □

In this paper, we need upper bounds on covering numbers of bounded balls in $\mathbb{R}^D$. We start by reviewing a result that provides an upper bound on the size of the set $\mathcal{C}_{\mathcal{B}}(\epsilon)$ defined w.r.t. a general norm $\|\cdot\|$.

Lemma 6.2 (Long & Sedghi 2019)**.** Let d be a positive integer, $\|\cdot\|$ be a norm, ρ be the metric induced by it, and $\kappa, \epsilon > 0$. A ball of radius κ in $\mathbb{R}^d$ w.r.t. ρ can be covered by $(\frac{3\kappa}{\epsilon})^d$ balls of radius ϵ.

We now review the upper bounds on the ℓ_∞-covering numbers of linear models.

Lemma 6.3 (Zhang 2002)**.** Let $\mathcal{L}$ be a class of linear functions on a set of size n. That is, $\mathcal{L} = \{\langle w, x\rangle, x, w \in \mathbb{R}^N\}$. If $\|x\|_q \leq b$ and $\|w\|_p \leq a$, where $2 \leq q < \infty$ and $1/p + 1/q = 1$, then for any $\epsilon > 0$, we have

$$\log \mathcal{N}_\infty(\epsilon, \mathcal{L}, n) \leq 36(q-1)\frac{a^2b^2}{\epsilon^2}\log[2\lceil 4ab/\epsilon + 2\rceil n + 1],$$

where $\mathcal{N}_\infty(\epsilon, \mathcal{L}, n)$ is the worst case covering number of the class $\mathcal{L}$ on a dataset of size n.

In the following, we present the proof of Lemma 5.2.

Proof of Lemma 5.2. First, we consider the ℓ_∞-norm on the set $\mathcal{B}$. According to Lemma 5.1, the function $\delta \mapsto \ell(\{(U(x+\delta))^T(U(x^+) - U(x_k^-))\}_{k=1}^K)$ is $\|\cdot\|_\infty$-Lipschitz with a constant of $2\Lambda_1^2\|x\|_2\sqrt{D}$. Nest, consider the set $\mathcal{C}_{\mathcal{B}}(\epsilon/4\Lambda_1^2\|x\|_2\sqrt{D})$. By applying Lemma 6.2, and noting that $\|\delta\|_\infty \leq \beta$, we have for all $\delta \in \mathcal{B}$:

$$\left|\mathcal{C}_{\mathcal{B}}(\epsilon/4\Lambda_1^2\|x\|_2\sqrt{D})\right| \leq \left(\frac{12\Lambda_1^2\|x\|_2\sqrt{D}\beta}{\epsilon}\right)^D.$$

Thus, the size of our dataset is:

$$|S_{\tilde{\mathcal{F}}}| = n\left(\frac{12\Lambda_1^2\|x\|_2\sqrt{D}\beta}{\epsilon}\right)^D d + ndK + nd.$$

For $\tilde{x} \in S_{\tilde{\mathcal{F}}}$, where $\tilde{x} = (x, \tilde{\delta})$, we have: $\|\tilde{x}\|_2 \leq \|x\|_2 + \|\tilde{\delta}\|_2 \leq \Psi + \sqrt{D}\|\delta\|_\infty = \Psi'$. Therefore, the result follows from Lemma 6.3. □

Below, we provide the proof for Corollary 5.1.

Proof of Corollary 5.1. The proof follows directly from Theorem 4.1 by setting α to $\frac{1}{n}$. Therefore, consider the following integral

$$\int_a^B \sqrt{\log \mathcal{N}_\infty\big(\frac{\epsilon}{8R\lambda_2}, \tilde{\mathcal{F}}, S_{\tilde{\mathcal{F}}}\big)}d\epsilon \leq \int_{\frac{1}{n}}^B \sqrt{\frac{CR^2\lambda_2^2\Lambda^2(\Psi+\sqrt{D}\beta)^2 L_{\log}}{\epsilon^2}}d\epsilon \leq \sqrt{C}R\lambda_2\Lambda\Psi'\frac{\tilde{L}_{\log}}{\log(n)}\int_{\frac{1}{n}}^B \frac{1}{\epsilon}d\epsilon$$
$$\leq \sqrt{C}R\lambda_2\Lambda\Psi'\frac{\tilde{L}_{\log}}{\log(n)}\left[\log(\epsilon)\right]_{\frac{1}{n}}^B \leq \sqrt{C}R\lambda_2\Lambda\Psi'\frac{\tilde{L}_{\log}}{\log(n)}(\log(B) + \log(n)).$$

The first inequality follows from the monotonicity property of integrals. The second inequality derives from the observation that replacing ϵ by $\frac{1}{n}$ in $\tilde{L}_{\log}$ can only increase its value. Substituting this into Theorem 4.1 yields the desired result. □

6.3 Proofs of Results in Section 5.2

In this subsection, we provide the omitted proofs from section 5.2 for the case when the features are non-linear. As in the linear case, we begin by showing that the loss function is ℓ_∞-Lipschitz with respect to the noise parameter δ. We then establish a bound on the set $\mathcal{C}_{\mathcal{B}}(\epsilon/2\lambda_1)$ and apply the ℓ_∞-covering number results of the non-linear feature class $\tilde{\mathcal{F}}$ on the extended dataset $S_{\tilde{\mathcal{F}}}$.

First, we prove the bounds of the ℓ_∞-additive attacks applied to non-linear models as stated in Lemma 5.3. The first step is to derive the ℓ_∞-Lipschitz constant of the function $\delta \mapsto \ell(\{(U\mathbf{v}(x+\delta))^T(U\mathbf{v}(x^+) - U\mathbf{v}(x_k^-))\}_{k=1}^K)$. The following is the proof of Lemma 5.3.

Proof of Lemma 5.3. The proof is a direct derivation. For all $x, x^+, x_k^- \in \mathcal{X}$, $\|U\|_2 \leq \Lambda_1$, $\delta, \delta' \in \mathcal{B}$, we have:

$$
\begin{aligned}
&|\ell(\{(U\mathbf{v}(x+\delta))^T(U\mathbf{v}(x^+) - U\mathbf{v}(x_k^-))\}_{k=1}^K) - \ell(\{(U\mathbf{v}(x+\delta'))^T(U\mathbf{v}(x^+) - U\mathbf{v}(x_k^-))\}_{k=1}^K)| \\
&\leq |\ell((U\mathbf{v}(x+\delta))^T(U\mathbf{v}(x^+) - U\mathbf{v}(x^-))) - \ell((U\mathbf{v}(x+\delta'))^T(U\mathbf{v}(x^+) - U\mathbf{v}(x^-)))| \\
&\leq |(U\mathbf{v}(x+\delta))^T(U\mathbf{v}(x^+) - U\mathbf{v}(x^-)) - (U\mathbf{v}(x+\delta'))^T(U\mathbf{v}(x^+) - U\mathbf{v}(x^-))| \\
&\leq \|U\mathbf{v}(x+\delta) - U\mathbf{v}(x+\delta')\|_2 \|U\mathbf{v}(x^+) - U\mathbf{v}(x^-)\|_2 \leq \|U(\mathbf{v}(x+\delta) - \mathbf{v}(x+\delta'))\|_2 \|U(\mathbf{v}(x^+) - \mathbf{v}(x^-))\|_2 \\
&\leq \|U\|_2 \|\mathbf{v}(x+\delta) - \mathbf{v}(x+\delta')\|_2 \|U\|_2 \|\mathbf{v}(x^+) - \mathbf{v}(x^-)\|_2 \\
&\leq \|U\|_2^2 \|V^L(\mathbf{v}_V^{L-1}(x+\delta) - \mathbf{v}_V^{L-1}(x+\delta'))\|_2 \|V^L(\mathbf{v}_V^{L-1}(x^+) - \mathbf{v}_V^{L-1}(x^-))\|_2 \\
&\leq \|U\|_2^2 \sqrt{w_L} \|V^L(\mathbf{v}_V^{L-1}(x+\delta) - \mathbf{v}_V^{L-1}(x+\delta'))\|_\infty \|V^L(\mathbf{v}_V^{L-1}(x^+) - \mathbf{v}_V^{L-1}(x^-))\|_2 \\
&\leq \Lambda_1^2 \sqrt{w_L} \max_{i \in [w_L]} \|V_{i,\cdot}^L\|_2 \|(\mathbf{v}_V^{L-1}(x+\delta) - \mathbf{v}_V^{L-1}(x+\delta'))\|_2 \|V^L\|_2 \|(\mathbf{v}_V^{L-1}(x^+) - \mathbf{v}_V^{L-1}(x^-))\|_2 \\
&\leq \Lambda_1^2 \sqrt{w_L} \prod_{l=2}^L s_l \|V^1(x+\delta) - V^1(x+\delta')\|_2 \prod_{l=1}^L s_l \|x^+ - x^-\|_2 \\
&\leq \Lambda_1^2 \sqrt{w_L} \prod_{l=2}^L s_l \sqrt{w_1} \|V^1(x+\delta) - V^1(x+\delta')\|_\infty \prod_{l=1}^L s_l \|x^+ - x^-\|_2 \\
&\leq \Lambda_1^2 \sqrt{w_L} \prod_{l=2}^L s_l \sqrt{w_1} \max_{i \in [w_1]} \|V_{i,\cdot}^1\|_1 \|x + \delta - x - \delta'\|_\infty \prod_{l=1}^L s_l \|x^+ - x^-\|_2 \\
&\leq 2\Lambda_1^2 \sqrt{w_L} \prod_{l=2}^L s_l \sqrt{w_1} s_1' \prod_{l=1}^L s_l \|x\|_2 \|\delta - \delta'\|_\infty.
\end{aligned}
$$

The second inequality is derived from the 1-Lipschitz property of the loss function ℓ. The seventh inequality results from converting the $\|.\|_2$-norm to $\|.\|_\infty$-norm. The eighth inequality stems from the definition of ℓ_∞-norm and Hölder inequality. The ninth inequality follows from the 1-Lipschitz property of the non-linearity and an induction over the layers. Finally, the last inequality is based on the fact that $\|x^+ - x^-\|_2 \leq 2\|x\|_2$, for all $x \in \mathbb{R}^D$. □

We now review the upper bounds on the ℓ_∞-covering numbers of norm-bounded neural networks (non-linear) function classes.

Lemma 6.4 (Ledent et al. 2021b)**.** Let $\mathcal{V}$ be the class of neural networks, that is, $\mathcal{V} = \{x \mapsto \mathbf{v}(x)\}$, where $V = (V^1, \ldots, V^L)$ are a set of weights the DNN $v(\cdot)$ and σ is defined as above. Suppose that $\|V^l\|_{2,1} \leq a_l$ and $\|V^l\|_\sigma \leq s_l$ for all $l \in [L-1]$, $\|V^L\|_2 \leq a_L$, $\|V^L\|_{2,\infty} \leq s_L$, $\|x\|_2 \leq b$, and w_l is the width of the l'th layer. Then given a data set S with n elements and $\epsilon > 0$, we have

$$
\log \mathcal{N}_\infty(\epsilon, \mathcal{V}, S) \leq \frac{CL^2b^2}{\epsilon^2} \prod_{l=2}^L s_l^2 \left(\sum_{l=1}^L \frac{a_l^2}{s_l^2} \right) \log\left((\, C_1 b\Gamma/\epsilon + C_2 \bar{w}\,)n + 1 \right),
$$

where $\Gamma = \max_{l \in [L]} (\prod_{i=1}^L s_i) a_l m_l / s_l$, $\bar{w} = \max_{l \in [L]} w_l$, and C, C_1, C_2 are universal constants.

In the following, we present the proof of Lemma 5.4.

Proof of Lemma 5.4. First, consider the ℓ_∞-norm on the set $\mathcal{B}$. By Lemma 5.3, we have the function $\delta \mapsto \ell(\{(U\mathbf{v}(x+\delta))^T(U\mathbf{v}(x^+) - U\mathbf{v}(x_k^-))\}_{k=1}^K)$ is $\|\cdot\|_\infty$-Lipschitz with constant $2\Lambda_1^2\sqrt{w_L}\prod_{l=2}^L s_l\sqrt{w_1}s_1'\prod_{l=1}^L s_l\|x\|_2$. Consider the set $\mathcal{C}_{\mathcal{B}}(\epsilon/4\Lambda_1^2\sqrt{w_L}\prod_{l=2}^L s_l\sqrt{w_1}s_1'\prod_{l=1}^L s_l\|x\|_2)$. By Lemma 6.2, and that $\|\delta\|_\infty \leq \beta$, we have for all $\delta \in \mathcal{B}$:

$$
\left| \mathcal{C}_{\mathcal{B}}(\epsilon/4\Lambda_1^2\sqrt{w_L}\prod_{l=2}^L s_l\sqrt{w_1}s_1'\prod_{l=1}^L s_l\|x\|_2) \right| \leq \left(\frac{12\Lambda_1^2\sqrt{w_L}\prod_{l=2}^L s_l\sqrt{w_1}s_1'\prod_{l=1}^L s_l\|x\|_2\beta}{\epsilon} \right)^D.
$$

Thus, the size of our dataset is:

$$|S_{\tilde{\mathcal{F}}}| = n\left(\frac{12\Lambda_1^2\sqrt{w_L}\prod_{l=2}^{L} s_l\sqrt{w_1}s_1'\prod_{l=1}^{L} s_l\|x\|_2\beta}{\epsilon}\right)^D d + ndK + nd,$$

For $\tilde{x} \in S_{\tilde{\mathcal{F}}}$, where $\tilde{x} = (x, \tilde{\delta})$, we have:

$$\|\tilde{x}\|_2 \leq \|x\|_2 + \|\tilde{\delta}\|_2 \leq \Psi + \sqrt{D}\|\delta\|_\infty = \Psi'.$$

Therefore, the result follows from Lemma 6.4. □

In the following, we present the proof of Corollary 5.2.

Proof of Corollary 5.2. The proof is similar to the proof of Corollary 5.1. It is a direct application of Theorem 4.1 by setting α to $\frac{1}{n}$. □

7 Conclusion

We conducted a generalization analysis of ACL, showing that the generalization error is bounded by the covering number of the feature class. Our results leverage the Lipschitz continuity and boundedness of the hinge loss as our unsupervised loss function, given the constraints on the model's weights and inputs. We applied this bound on both linear and non-linear features, subject to ℓ_∞-additive attacks. Our analysis reveals a logarithmic dependence on the number of negative samples. However, these are algorithm-independent bounds, which could be extended to algorithm-dependent bounds to see how the optimization process affects the generalization error. In this paper, we have applied only ℓ_∞-additive attacks; however, other kinds of adversarial attacks can also be tested, especially non-additive attacks.

Broader Impact Statement

In this optional section, TMLR encourages authors to discuss possible repercussions of their work, notably any potential negative impact that a user of this research should be aware of. Authors should consult the TMLR Ethics Guidelines available on the TMLR website for guidance on how to approach this subject.

References

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

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

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pp. 387–402. Springer, 2013.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pp. 208–240. Springer, 2003.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Kejiang Chen, Yuefeng Chen, Hang Zhou, Xiaofeng Mao, Yuhong Li, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Self-supervised adversarial training. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2218–2222. IEEE, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

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*, pp. 9729–9738, 2020.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pp. 4182–4192. PMLR, 2020.

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

Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: A theoretical analysis. *Journal of Machine Learning Research*, 24(330):1–78, 2023.

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.

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

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.

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

Antoine Ledent, Rodrigo Alves, Yunwen Lei, and Marius Kloft. Fine-grained generalization analysis of inductive matrix completion. In *Advances in Neural Information Processing Systems*, volume 34, pp. 25540–25552, 2021a.

Antoine Ledent, Waleed Mustafa, Yunwen Lei, and Marius Kloft. Norm-based generalisation bounds for deep multi-class convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8279–8287, 2021b.

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

Yunwen Lei, Tianbao Yang, Yiming Ying, and Ding-Xuan Zhou. Generalization analysis for contrastive representation learning. In *International Conference on Machine Learning*, pp. 19200–19227. PMLR, 2023.

Philip M Long and Hanie Sedghi. Size-free generalization bounds for convolutional neural networks. *arXiv preprint arXiv:1905.12600*, 35:4053–4061, 2019.

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

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pp. 2512–2530. PMLR, 2019.

Waleed Mustafa, Yunwen Lei, Antoine Ledent, and Marius Kloft. Fine-grained generalization analysis of structured output prediction. *arXiv preprint arXiv:2106.00115*, 2021.

Waleed Mustafa, Yunwen Lei, and Marius Kloft. On the generalization analysis of adversarial learning. In *International Conference on Machine Learning*, pp. 16174–16196. PMLR, 2022.

Kento Nozawa, Pascal Germain, and Benjamin Guedj. Pac-bayesian contrastive unsupervised representation learning. In *Conference on Uncertainty in Artificial Intelligence*, pp. 21–30. PMLR, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Florian Tramer and Dan Boneh. Adversarial training and robustness for multiple perturbations. *Advances in neural information processing systems*, 32, 2019.

Jiancong Xiao, Yanbo Fan, Ruoyu Sun, and Zhi-Quan Luo. Adversarial rademacher complexity of deep neural networks. *arXiv preprint arXiv:2211.14966*, 2022.

Jingyuan Xu and Weiwei Liu. On robust multiclass learnability. *Advances in Neural Information Processing Systems*, 35:32412–32423, 2022.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pp. 7085–7094. PMLR, 2019.

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

Xin Zou and Weiwei Liu. Generalization bounds for adversarial contrastive learning. *Journal of Machine Learning Research*, 24(114):1–54, 2023.