Boosting Adversarial Training via Fisher-Rao Norm-based Regularization

Xiangyu Yin Wenjie Ruan*
University of Liverpool, UK
x.yin22@liverpool.ac.uk, w.ruan@trustai.uk

Abstract

Adversarial training is extensively utilized to improve the adversarial robustness of deep neural networks. Yet, mitigating the degradation of standard generalization performance in adversarial-trained models remains an open problem. This paper attempts to resolve this issue through the lens of model complexity. First, We leverage the Fisher-Rao norm, a geometrically invariant metric for model complexity, to establish the non-trivial bounds of the Cross-Entropy Loss-based Rademacher complexity for a ReLU-activated Multi-Layer Perceptron. Then we generalize a complexity-related variable, which is sensitive to the changes in model width and the trade-off factors in adversarial training. Moreover, intensive empirical evidence validates that this variable highly correlates with the generalization gap of Cross-Entropy loss between adversarial-trained and standard-trained models, especially during the initial and final phases of the training process. Building upon this observation, we propose a novel regularization framework, called Logit-Oriented Adversarial Training (LOAT), which can mitigate the trade-off between robustness and accuracy while imposing only a negligible increase in computational overhead. Our extensive experiments demonstrate that the proposed regularization strategy can boost the performance of the prevalent adversarial training algorithms, including PGD-AT, TRADES, TRADES (LSE), MART, and DM-AT, across various network architectures. Our code will be available at https://github.com/TrustAI/LOAT.

1. Introduction

Deep Neural Networks (DNNs) are extensively applied in a variety of safety-critical systems, such as surveillance systems, drones, autonomous vehicles, and malware detection [4, 7, 10, 36, 41]. Despite their pervasiveness, considerable evidence shows that the standard-trained deep learning models can be easily fooled by subtly modified data points, leading to inaccurate predictions [6, 11, 19, 20, 28, 38]. To make models more resistant to such imperceptible perturbations, numerous adversarial training algorithms have been proposed in recent years, such as PGD-AT [17], TRADES [43], MART [32], TRADES (LSE) [24], DM (Diffusion Model)-AT [34], etc. These methods aim to ensure the model's feature representation consistency across clean and adversarial inputs [9, 30]. Nonetheless, it is observed that these adversarial training algorithms commonly lead to a compromise in the standard generalization performance. This phenomenon, often referred to as '*trade-off between robustness and accuracy*', continues to be a subject of vigorous discussion in recent years.

Specifically, several studies have attributed the deterioration of standard accuracy to the bias introduced by adversarial training algorithms [24, 40]. Other researchers have suggested that sufficient training data is required to bridge the feature gap between adversarial and standard-trained models [25, 29]. Another viewpoint relates this trade-off with the generalizability of the network using a measure known as local Lipschitzness [37]. However, most of these studies tend to focus on a *single* specific perspective, such as *training objective*, *learning data*, or *model architecture*, which fails to consider the comprehensive integration of these various factors. This context compels us to explore an important yet challenging question: *Can we interpret the degradation of standard accuracy in a unified and principled way?*

Affirmatively, as elucidated by [8], all the factors above that contribute to the degradation of standard generalization in adversarial training ultimately influence a central concept: *model complexity*, which offers a potential approach to analyze this trade-off fundamentally. Considering the diversity of DNNs, identifying a universal framework for model complexity remains an elusive goal. For the sake of simplicity, we conceptualize models with varying widths and depths as Multi-Layer Perceptrons (MLP). Specifically, given a ReLU-activated L-layer MLP, we examine its Rademacher complexity concerning the Cross-Entropy loss out of a set of hypotheses. The range of complexity is captured by Fisher-Rao norm [15], a geometric invariant measure for model complexity. Deriving from this, we can establish bounds for the Rademacher complex-

*Corresponding Author

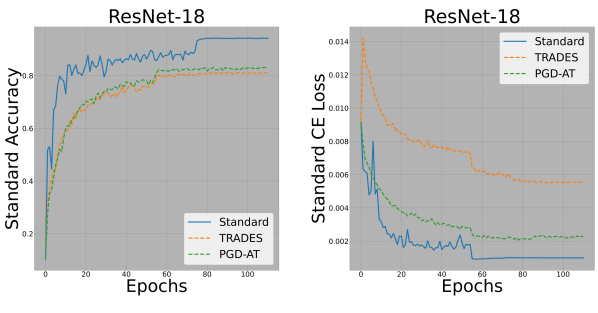

Figure 1. Standard generalization performance on CIFAR10.

ity based on the Fisher-Rao norm. Moreover, we observe that the upper and lower bounds of the Cross-Entropy Loss-based Rademacher complexity are significantly influenced by a variable Γ_{ce}, which is determined by the classification performance on clean training samples. Empirical studies reveal that adversarially-trained MLPs, with different model widths and trade-off factors, have *unique* values of Γ_{ce}. Furthermore, through adjustments to the model width and trade-off factor, Γ_{ce} demonstrates a *positive* correlation with the generalization gap of Cross-Entropy loss between adversarial-trained and standard-trained models in the *early* epochs, and a *negative* correlation in the *later* epochs.

Finally, capitalizing on the empirical link between Γ_{ce} and the generalization gap, we propose a novel *epoch-specific* framework for adversarial training, named Logit-Oriented Adversarial Training (LOAT). It uniquely combines two regularization tactics: the standard logit-oriented regularizer and the adaptive adversarial logit pairing strategy. Unlike other adversarial training regularizers such as [12, 31, 35], LOAT operates as a *black-box* solution, preventing the prior knowledge about the model's weights. Notably, LOAT focuses exclusively on the initial and final stages of the adversarial training process, thereby maintaining a *minimal* increase in computational demand. We summarize our main contributions as follows:

- For a ReLU-activated MLP, we introduce the Fisher-Rao norm to capture its Rademacher complexity for Cross-Entropy loss. We empirically and theoretically demonstrate that the logit-based variable Γ_{ce} notably influences both the upper and lower bounds.
- Empirical analysis reveals a non-trivial link between Γ_{ce} in an adversarial-trained MLP and critical parameters such as model width and other various trade-off factors. Notably, Γ_{ce}'s correlation with the generalization gap of Cross-Entropy loss is found to be epoch-dependent, varying across different stages of the training.
- We propose a new regularization methodology, Logit-Oriented Adversarial Training (LOAT), which can seamlessly integrate with current adversarial training algorithms and, more importantly, boost their performances without substantially increasing computational overhead.

2. Related works

2.1. Trade-off Between Robustness and Accuracy

Various factors influence the effectiveness of adversarial training [21, 33]. For instance, [21] highlights the need for increased model capacity to achieve robustness against adversarial examples. Other methods utilize data augmentation or mixup to avert robust overfittings [26, 39]. Additionally, recent studies, including [16, 18, 27], focus on loss perspectives, with strategies like regularization during training to smoothen the input loss landscapes. Conversely, [12, 35] explore the interplay between weight loss landscape and adversarial robustness.

The above methodologies underscore various means to enhance adversarial robustness. However, as indicated by [29, 42], there appears to be a natural conflict between adversarial robustness and standard accuracy. [40] shows that the bias from adversarial training increases with perturbation radius, significantly impacting overall risk. [29] observes that the feature representations in adversarial-trained models differ from those in standard-trained models. Alternative perspectives suggest that this trade-off is not inherent and can be mitigated by increasing training data size or theoretically bounding natural and boundary errors. Furthermore, [24, 37] attribute the trade-off to current adversarial training algorithms. Our paper dissects this trade-off, mainly focusing on the influence of model complexity towards the degradation of standard generalization in adversarial training. This phenomenon is depicted in Fig. 1.

2.2. Fisher-Rao Norm

The study of DNN's capacity has been extensive over the last decade. The Vapnik-Chervonenkis dimension is a classical complexity metric, as discussed in [2]. Yet, its applicability to over-parameterized models might be too broad to explain their generalization. Norm-based measures, as a form of capacity control, have been a focus of recent research [1, 14, 23], although they may not fully encapsulate the distinct variances across various architectures. To address this, the Fisher-Rao norm, introduced in [15], emerges as a crucial geometric complexity measure. It encompasses existing norm-based capacity metrics and is particularly valued for its geometric invariance, a feature enriched by its connections to information geometry and nonlinear local transformations.

3. Preliminaries

3.1. Basic Notions

Consider a typical image classification task, where the objective is to categorize an input image into one of K classes.

Let $\mathbf{x} \in [0,1]^d$ represent a normalized input image, and let $y \in \{0, \ldots, K-1\}$ denote its corresponding ground truth label. These are assumed to be jointly sampled from a $\mathcal{D}$ distribution. Focusing on a specific hypothesis set $\mathcal{F}$ and a loss function $\mathcal{L}$, the objective of adversarial training is broadly defined as follows:

$$\min_{f \in \mathcal{F}} \left[(1-\lambda)\mathcal{L}(f(\mathbf{x}), y) + \lambda \mathcal{L}(f(\mathcal{A}_p^\epsilon(\mathbf{x})), y) \right] \quad (1)$$

Here, $\mathcal{A}_p^\epsilon(\mathbf{x}) = \arg\max_{\mathbf{x}' \in \mathcal{B}_p^\epsilon(\mathbf{x})} \mathcal{L}(f(\mathbf{x}'), y)$, where $\mathcal{B}_p^\epsilon(\mathbf{x}) = \{\mathbf{x}' | \, \|\mathbf{x}' - \mathbf{x}\|_p \leq \epsilon\}$, λ signifies the trade-off factor, and ϵ represents the radius of the ℓ_p-norm ball. In particular, when λ=1.0, it indicates the training objective of PGD-AT. Eq. 1 outlines the process for identifying the optimal hypothesis within $\mathcal{F}$. This typically translates to training a neural network that approximates this objective in practical applications.

Assumption 1. *For a neural network-based set of hypotheses $\mathcal{F}$, every $f \in \mathcal{F}$ has an identical architecture and is trained on the same dataset.*

Given the recent proliferation of DNN architectures, conducting a theoretical analysis of various hypothesis sets $\mathcal{F}$ within a universal framework poses a considerable challenge. To address this, our paper simplifies the problem by focusing on MLPs. We analyze the properties of DNNs by varying the depth and width of MLPs, providing a more manageable scope for in-depth study. Detailed definitions are provided below.

Definition 1 (L-layer MLP)**.** *Given a set of weight matrices $\mathcal{W} = \{\mathbf{W}_1, \cdots, \mathbf{W}_L\}$, and an activation function $\phi(\cdot)$. We define the hypothesis $f(\mathbf{x})$ as an approximation using L layers of matrix multiplication, which is expressed as:*

$$f(\mathbf{x}) \approx f_{\mathcal{W}}^L(\mathbf{x}) = \mathbf{W}_L \phi(\cdots \phi(\mathbf{W}_1 \mathbf{x})) \quad (2)$$

where $\mathbf{W}_l \in \mathbb{R}^{H_l \times H_{l-1}}$ for $1 \leq l \leq L$, and H_l denotes the number of hidden units in the l-th layer. Specifically, $H_0 = d$ represents the input dimension, and H_L is the number of output classes.

Upon processing by the softmax function $\sigma(\cdot)$, $f_{\mathcal{W}}^L(\mathbf{x})$ yields probabilities for the K classes. Following Eq. 1 and Def. 1, we now present a detailed definition of the risk for clean samples.

Definition 2 (Standard Risk)**.** *Consider a specific hypothesis $f_{\mathcal{W}}^L$. Empirical risk and population risk for clean samples are defined as follows:*

$$\begin{aligned} \tilde{R}_{N_{tr}}\left(\mathcal{L} \circ f_{\mathcal{W}}^L\right) &= \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \mathcal{L}(f_{\mathcal{W}}^L(\mathbf{x}_i), y_i) \\ R\left(\mathcal{L} \circ f_{\mathcal{W}}^L\right) &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathcal{L}(f_{\mathcal{W}}^L(\mathbf{x}), y)\right] \end{aligned} \quad (3)$$

N_{tr} denotes the number of clean training samples.

3.2. Generalization Gap between Algorithms

Given two distinct training algorithms, a_1 and a_2, let $\mathcal{F}_{a_1}$ and $\mathcal{F}_{a_2}$ represent the sets of hypotheses generated by each, respectively. Utilizing Def. 2, we define the generalization gap of standard risk concerning the loss $\mathcal{L}$ between $\mathcal{F}_{a_1}$ and $\mathcal{F}_{a_2}$ as follows:

$$G_{\mathcal{L}}^{\langle \mathcal{F}_{a_1}, \mathcal{F}_{a_2} \rangle} = \min_{f_1 \in \mathcal{F}_{a_1}} R(\mathcal{L} \circ f_1) - \min_{f_2 \in \mathcal{F}_{a_2}} R(\mathcal{L} \circ f_2) \quad (4)$$

It is important to note that the architecture and training set for $\mathcal{F}_{a_1}$ are identical to those of $\mathcal{F}_{a_2}$, and $G_{\mathcal{L}}^{\langle \mathcal{F}_{a_1}, \mathcal{F}_{a_2} \rangle}$ is determined by training networks in $\mathcal{F}_{a_1}$ and $\mathcal{F}_{a_2}$ for the same number of epochs. For instance, as depicted in Fig. 1, the generalization gap of Cross-Entropy (CE) loss between adversarial-trained and standard-trained models is observable. This is further characterized by a significant reduction in standard test accuracy across the three adversarial training methods compared to standard training. Specifically, we define the hypothesis sets trained under standard conditions as $\mathcal{F}_{std}$, and those trained using the objective outlined in Eq. 1 as $\mathcal{F}_{at}$.

4. Proposed Methods

4.1. Rademacher Complexity via CE Loss

As depicted in Fig. 1, $\mathcal{L}_{ce}$ is a key metric for standard test accuracy and is widely used in adversarial training approaches.

In this section, we aim to understand how adversarial training impacts the generalization performance as measured by $\mathcal{L}_{ce}$. To achieve this, we explore the concept of Rademacher Complexity [3], which affects the ability of the hypothesis set $\mathcal{F}$ to fit random noise and, consequently, its generalization performance.

Definition 3. *Given a neural network-based set of hypotheses $\mathcal{F}$, the Rademacher complexity via $\mathcal{L}$ can be written as:*

$$\mathcal{R}_{N_{tr}}(\mathcal{L} \circ \mathcal{F}) = \mathbb{E}_\xi \frac{1}{N_{tr}} \left[\sup_{f_{\mathcal{W}}^L \in \mathcal{F}} \sum_{i=1}^{N_{tr}} \xi_i \mathcal{L}\left(f_{\mathcal{W}}^L(\mathbf{x}_i), y_i\right) \right] \quad (5)$$

where ξ_i takes values in $\{-1, 1\}$ with equal probability. Furthermore, suppose $\mathcal{L} \circ \mathcal{F} \in [0, B]$, then given $\delta \in (0,1)$, the following generalization bound holds for any $f_{\mathcal{W}}^L \in \mathcal{F}$ with probability 1-δ:

$$R(\mathcal{L} \circ f_{\mathcal{W}}^L) \leq \tilde{R}_{N_{tr}}(\mathcal{L} \circ f_{\mathcal{W}}^L) + 2B\mathcal{R}_{N_{tr}}(\mathcal{L} \circ \mathcal{F}) + 3B\sqrt{\frac{\ln \frac{2}{\delta}}{2N_{tr}}} \quad (6)$$

Although it is not feasible to deduce the upper bound for $R(\mathcal{L}_{ce} \circ f_{\mathcal{W}}^L)$ from Eq. 6, due to the absence of an upper bound for $\mathcal{L}_{ce}$, we can still reference the Rademacher complexity relative to $\mathcal{L}_{ce}$. It is important to note that a higher

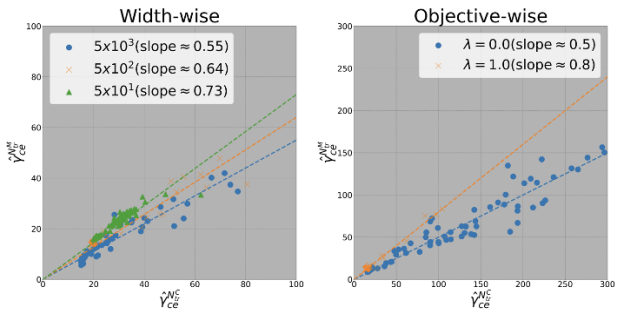

Figure 2. The correlation between $\hat{\gamma}_{ce}^{N_{tr}^C}$ on the x-axis and $\hat{\gamma}_{ce}^{N_{tr}^M}$ on the y-axis within 1-layer MLPs. Each data point corresponds to different epochs during the training process.

Rademacher complexity implies weaker generalization capability, while a lower complexity suggests stronger generalization. Hence, in the ensuing sections, our focus will be on exploring the bounds of $\mathcal{R}_{N_{tr}}(\mathcal{L}_{ce} \circ \mathcal{F})$.

4.2. Bounds of complexity via Fisher-Rao Norm

As detailed in Def. 3, defining the range of $\mathcal{F}$ is crucial for establishing the bounds of Rademacher Complexity in the context of CE loss. Previous research, such as [1, 22, 23], has relied on norm-based complexity measures that encompass all possible candidates within $\mathcal{F}$. However, the varied nature of models used in adversarial training, each with its unique architecture, poses a challenge in formulating a standard measure of model complexity. To address this, we adopt the Fisher-Rao norm, which effectively bypasses the inherent disparities between different models, focusing instead on the variability in output logits.

Lemma 1 (Fisher-Rao Norm [15]). *Given an L-layer MLP-approximated hypothesis $f_{\mathcal{W}}^L$ as defined in Def. 1, if $\mathcal{L}$ is smooth with respect to $f_{\mathcal{W}}^L$, the following identity holds:*

$$\|\mathcal{W}\|_{FR\circ\mathcal{L}}^2 = L^2 \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\left\langle \frac{\partial \mathcal{L}(f_{\mathcal{W}}^L(\mathbf{x}), y)}{\partial f_{\mathcal{W}}^L(\mathbf{x})}, f_{\mathcal{W}}^L(\mathbf{x}) \right\rangle^2 \right] \quad (7)$$

We then incorporate the CE loss, which exhibits smoothness with respect to $f_{\mathcal{W}}^L$, into Lemma 1. This substitution allows us to establish an upper bound for the radius of the Fisher-Rao norm ball with respect to the CE loss.

Lemma 2 ($\mathcal{L}_{ce}$-based Fisher-Rao Norm Ball). *Let the radius of the Fisher-Rao norm ball with respect to $\mathcal{L}_{ce}$ be denoted as $\gamma_{ce} = \frac{1}{L}\|\mathcal{W}\|_{FR\circ\mathcal{L}_{ce}}$. It can be upper bounded by:*

$$\gamma_{ce} \leq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\max_{k\neq y} \left| f_{\mathcal{W}}^L(\mathbf{x})_k - f_{\mathcal{W}}^L(\mathbf{x})_y \right| \right] \quad (8)$$

The right side of the above inequality is denoted as $\hat{\gamma}_{ce}$.

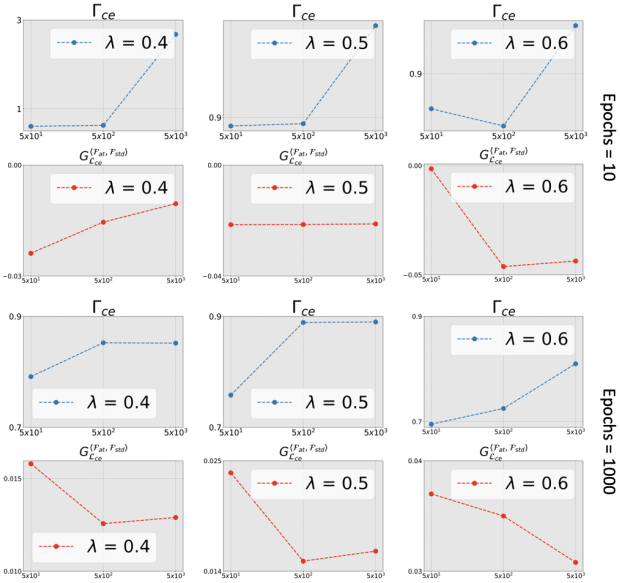

Figure 3. Depicting Γ_{ce} and $G_{\mathcal{L}_{ce}}^{\langle \mathcal{F}_{at}, \mathcal{F}_{std} \rangle}$ in 1-layer MLPs with respect to various trade-off factors λ ranging from 0.1 to 1.0 in $\mathcal{F}_{at}$. The x-axis represents the number of hidden units from 50 to 5000.

The comprehensive proof of Lemma 2 is provided in the Appendix. Both Eq. 7 and Eq. 8 can be empirically estimated using the training or the test set. Consequently, the Fisher-Rao norm-based complexity measure for the hypothesis $f_{\mathcal{W}}^L$ may exhibit variation due to discrepancies between these two datasets. Nevertheless, our analysis is confined to the training set exclusively for the empirical determination of Rademacher complexity bounds. Building on Lemma 2, we proceed to empirically approximate $\hat{\gamma}_{ce}$.

Definition 4. *Let N_{tr}^C represent the number of correctly classified clean samples and N_{tr}^M the number of misclassified clean samples. Then, the radii of the Fisher-Rao norm balls concerning $\mathcal{L}_{ce}$ for N_{tr}, N_{tr}^C, and N_{tr}^M can be estimated as follows:*

$$\begin{aligned}
\hat{\gamma}_{ce} &\approx \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \max_{k\neq y_i} \left| f_{\mathcal{W}}^L(\mathbf{x}_i)_k - f_{\mathcal{W}}^L(\mathbf{x}_i)_{y_i} \right| \\
\hat{\gamma}_{ce}^{N_{tr}^C} &\approx \frac{1}{N_{tr}^C} \sum_{i=1}^{N_{tr}} \mathcal{I}_C \max_{k\neq y_i} \left| f_{\mathcal{W}}^L(\mathbf{x}_i)_k - f_{\mathcal{W}}^L(\mathbf{x}_i)_{y_i} \right| \quad (9) \\
\hat{\gamma}_{ce}^{N_{tr}^M} &\approx \frac{1}{N_{tr}^M} \sum_{i=1}^{N_{tr}} \mathcal{I}_M \max_{k\neq y_i} \left| f_{\mathcal{W}}^L(\mathbf{x}_i)_k - f_{\mathcal{W}}^L(\mathbf{x}_i)_{y_i} \right|
\end{aligned}$$

where $\mathcal{I}_C = \mathbb{1}(\arg\max \sigma(f_{\mathcal{W}}^L(\mathbf{x}_i)) = y_i)$, $\mathcal{I}_M = \mathbb{1}(\arg\max \sigma(f_{\mathcal{W}}^L(\mathbf{x}_i)) \neq y_i)$

In line with Lemma 2 and Def. 4, we proceed to examine the Rademacher complexity constrained by the Fisher-Rao norm with respect to $\mathcal{L}_{ce}$.

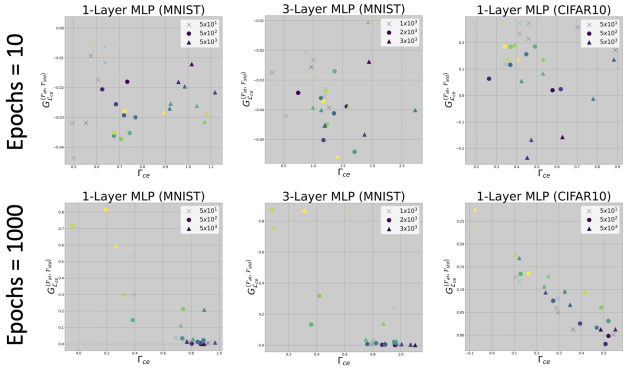

Figure 4. Assessment of $G^{\langle \mathcal{F}_{at}, \mathcal{F}_{std} \rangle}_{\mathcal{L}_{ce}}$ over three 3 distinct architecture-dataset combinations. Diverse symbols such as $\times$, $\blacktriangle$ and $\bullet$ represent different numbers of hidden units. Additionally, varying shades indicate a range of trade-off factors λ within $\mathcal{F}_{at}$, specifically from 0.1 to 1.0.

Theorem 1. *Given a set of hypotheses* $\mathcal{F}^{\hat{\gamma}_{ce}} = \{f^L_{\mathcal{W}} | \, \|\mathcal{W}\|_{FR \circ \mathcal{L}_{ce}} \leq L\hat{\gamma}_{ce}\}$, *if we denote* $\Gamma_{ce} = \frac{\hat{\gamma}^{N^C_{tr}}_{ce} - \hat{\gamma}^{N^M_{tr}}_{ce}}{\hat{\gamma}^{N^M_{tr}}_{ce}}$, $\mathcal{C}_C = \frac{N_{tr}}{N^C_{tr}}$, $\mathcal{C}_M = \frac{N_{tr}}{N^M_{tr}}$, $\mathcal{C}_{MC} = \frac{\sqrt{N^M_{tr}} + \sqrt{N^C_{tr}}}{N_{tr}}$, *then we can provide bounds for the Rademacher complexity w.r.t* $\mathcal{L}_{ce}$ *as follows:*

$$\mathcal{R}_{N_{tr}}\left(\mathcal{L}_{ce} \circ \mathcal{F}^{\hat{\gamma}_{ce}}\right) \gtrsim \mathcal{C}_{MC} \ln K - \frac{\hat{\gamma}^{N^M_{tr}}_{ce} \mathcal{C}_C^{-0.5} (\mathcal{C}_C^{-1} \Gamma_{ce} + 1)}{N_{tr}^{0.5}}$$
$$\mathcal{R}_{N_{tr}}\left(\mathcal{L}_{ce} \circ \mathcal{F}^{\hat{\gamma}_{ce}}\right) \lesssim \mathcal{C}_{MC} \ln K + \frac{\hat{\gamma}^{N^M_{tr}}_{ce} \mathcal{C}_M^{-0.5} (\mathcal{C}_C^{-1} \Gamma_{ce} + 1)}{N_{tr}^{0.5}} \tag{10}$$

Remark 1. *It is noteworthy that the bounds for* $\mathcal{R}(\mathcal{L}_{ce} \circ \mathcal{F}^{\hat{\gamma}_{ce}})$ *are intricately linked to* $\mathcal{C}_M^{-0.5}$ *and* $\mathcal{C}_C^{-0.5}$. *Theoretically, irrespective of* $\mathcal{C}_{MC} \log K$, *the rate of change between the upper and lower bounds with respect to* Γ_{ce} *can be approximated by* $\mathcal{O}(\sqrt{\frac{N^M_{tr}}{N^C_{tr}}})$. *Comprehensive proofs of Thm. 1 are detailed in the Appendix.*

4.3. Sensitivity to Complexity-Related Factors

As noted in Remark 1, Eq. 10 provides the rate of change between the upper and lower bounds of Rademacher complexity with respect to Γ_{ce} within a single hypothesis set, yet the variation of Γ_{ce} with respect to model complexity-related factors, which result in diverse hypothesis sets preserving apparently different generalization performance, still remains unknown. Therefore, we select two representative model complexity-related factors, model width and training objective, to test the sensitivity of Γ_{ce} across different hypothesis sets. Specifically, we train models within various $\mathcal{F}_{at}$ until $\mathcal{L}_{ce}$ reaches convergence.

Width-wise. In the left panel of Fig. 2, we investigate the impact of varying the number of hidden units in a 1-layer MLP, ranging from 5×10^1 to 5×10^3. An intriguing observation emerges as we increase the number of hidden units: the slope of the dashed line, used to approximate the relationship between $\hat{\gamma}^{N^M_{tr}}_{ce}$ and $\hat{\gamma}^{N^C_{tr}}_{ce}$, gradually decreases from 0.73 to 0.55.

Objective-wise. In the right panel of Fig. 2. As λ transitions from 0.0 (indicative of standard training) to 1.0 (corresponding to PGD-AT), the slope of the dashed line shifts from 0.5 to 0.8. Specifically, an increase in Γ_{ce}, represented by a rising ratio of $\hat{\gamma}^{N^C_{tr}}_{ce} / \hat{\gamma}^{N^M_{tr}}_{ce}$, is consistent with the trajectory spanning from PGD-AT (λ=1.0) to standard training (λ=0.0).

4.4. Influence on Generalization Gap of CE Loss

As shown in Thm. 1, during the early stages of adversarial training on single $\mathcal{F}^{\hat{\gamma}_{ce}}$, characterized by $N^M_{tr} \gg N^C_{tr}$, an increase in Γ_{ce} might lead to a reduction in the lower bound of $\mathcal{R}_{N_{tr}}(\mathcal{L}_{ce} \circ \mathcal{F}^{\hat{\gamma}_{ce}})$. Conversely, a decrease in Γ_{ce} tends to have a more pronounced effect on lowering its upper bound. In contrast, when $N^M_{tr} \ll N^C_{tr}$, although a reduction in Γ_{ce} can lower the upper bound of $\mathcal{R}(\mathcal{L}_{ce} \circ \mathcal{F}^{\hat{\gamma}_{ce}})$, an increase in Γ_{ce} may cause a more rapid decline in the lower bound. It is commonly known that $G^{\langle \mathcal{F}_{at}, \mathcal{F}_{std} \rangle}_{\mathcal{L}_{ce}}$ is significantly influenced by both the model width and the trade-off factor λ. Therefore, the empirical relationship observed between Γ_{ce} and the model width/trade-off factor λ as detailed in Sec. 4.3 leads us to naturally infer that there should also be an inherent correlation between Γ_{ce} and $G^{\langle \mathcal{F}_{at}, \mathcal{F}_{std} \rangle}_{\mathcal{L}_{ce}}$.

Objective-wise. Fig. 3 depicts the relationship between Γ_{ce} and $G^{\langle \mathcal{F}_{at}, \mathcal{F}_{std} \rangle}_{\mathcal{L}_{ce}}$ of 1-layer MLPs in relation to the trade-off factor λ. For models trained with 10 epochs, Γ_{ce} is roughly in positive correlation with $G^{\langle \mathcal{F}_{at}, \mathcal{F}_{std} \rangle}_{\mathcal{L}_{ce}}$, whilst for models trained with 1000 epochs, Γ_{ce} is roughly in negative correlation with $G^{\langle \mathcal{F}_{at}, \mathcal{F}_{std} \rangle}_{\mathcal{L}_{ce}}$, which is consistent with the Thm. 1.

Width-wise. Fig. 4 indicates the relationship between Γ_{ce} and $G^{\langle \mathcal{F}_{at}, \mathcal{F}_{std} \rangle}_{\mathcal{L}_{ce}}$ of different MLPs in relation to the number of hidden units. For models trained over 10 epochs, Γ_{ce} appears to be broadly in positive correlation with $G^{\langle \mathcal{F}_{at}, \mathcal{F}_{std} \rangle}_{\mathcal{L}_{ce}}$, whereas for models subjected to 1000 epochs of training, Γ_{ce} tends to exhibit a general negative correlation with $G^{\langle \mathcal{F}_{at}, \mathcal{F}_{std} \rangle}_{\mathcal{L}_{ce}}$.

4.5. Logit-Oriented Adversarial Training

In Section 4.4, we report a noteworthy trend: Γ_{ce} initially correlates positively with the generalization gap of CE loss, but this correlation turns negative as training progresses. To counteract this, we introduce an epoch-dependent regularization strategy. Specifically, we initially penalize the disparity $\hat{\gamma}^{N^C_{tr}}_{ce} - \hat{\gamma}^{N^M_{tr}}_{ce}$ and later reverse this to penalize $\hat{\gamma}^{N^M_{tr}}_{ce} - \hat{\gamma}^{N^C_{tr}}_{ce}$, which is in the opposite direction. Con-

currently, we employ an adaptive technique for enhanced model resilience to synchronize the distributions of standard and adversarial logits. These complementary regularization methods construct our proposed Logit-Oriented Adversarial Training (LOAT) framework.

4.5.1 Standard Logit-Oriented Regularization

For broader applicability, we represent the adversarial-trained model $f_{\mathcal{W}}^{L}(\cdot)$ by a more general notation, $\mathcal{M}(\cdot)$, and denote the number of training samples, which could be a batch or the entire dataset, as N. It is important to note that directly optimizing $\hat{\gamma}_{ce}^{N^M}$ and $\hat{\gamma}_{ce}^{N^C}$ is infeasible due to inherent computational complexities and challenges. Therefore, in line with Eq. 9, we opt to utilize their lower bounds as the subject to be optimized.

$$\begin{aligned}\overline{\gamma}_{ce}^{N^M} &\approx \frac{1}{N^M}\sum_{i=1}^{N}\mathcal{I}_M\left[\frac{1}{K-1}\sum_{k\neq y_i}\left|\mathcal{M}(\mathbf{x}_i)_k-\mathcal{M}(\mathbf{x}_i)_{y_i}\right|\right]\leq\hat{\gamma}_{ce}^{N^M}\\ \overline{\gamma}_{ce}^{N^C} &\approx \frac{1}{N^C}\sum_{i=1}^{N}\mathcal{I}_C\left[\frac{1}{K-1}\sum_{k\neq y_i}\left|\mathcal{M}(\mathbf{x}_i)_k-\mathcal{M}(\mathbf{x}_i)_{y_i}\right|\right]\leq\hat{\gamma}_{ce}^{N^C}\end{aligned} \tag{11}$$

where $\mathcal{I}_C = \mathbb{1}(\arg\max\sigma(\mathcal{M}(\mathbf{x}_i)) = y_i)$, $\mathcal{I}_M = \mathbb{1}(\arg\max\sigma(\mathcal{M}(\mathbf{x}_i)) \neq y_i)$. Echoing the TRADES (LSE) approach [24], we use softmax-transformed surrogates instead of raw logits throughout the regularization phase. Moreover, given the non-smooth properties of the ℓ_1 norm discussed in Eq. 11 and its inclination towards generating sparse solutions, we prefer the ℓ_2 norm as a more appropriate option. Consequently, following these adjustments, we define our surrogate optimization term $\mathcal{P}_C^N$ for $\overline{\gamma}_{ce}^{N^C}$ as follows:

$$\frac{1}{N^C}\sum_{i=1}^{N}\mathcal{I}_C\frac{1}{K-1}\sum_{k\neq y_i}\left(\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)_k-\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)_{y_i}\right)^2 \tag{12}$$

Notice that for correctly classified samples, it is trivial to prove that $\sigma(\mathcal{M}(\mathbf{x}_i)_{y_i}) \geq \frac{1}{K-1}\sum_{k\neq y_i}\sigma(\mathcal{M}(\mathbf{x}_i))_k$, then the lower bound $\check{\mathcal{P}}_C^N$ for $\mathcal{P}_C^N$ can be written as:

$$\frac{1}{N^C}\sum_{i=1}^{N}\mathcal{I}_C\frac{1}{K-1}\sum_{k\neq y_i}\left(\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)_k-\frac{1}{K-1}\sum_{k\neq y_i}\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)_k\right)^2 \tag{13}$$

In a parallel manner, we designate the surrogate regularization term for $\overline{\gamma}_{ce}^{N^M}$ as $\mathcal{P}_M^N$. In this context, the ensuing inequalities can be established

$$\begin{aligned}\mathcal{P}_M^N &= \frac{1}{N^M}\sum_{i=1}^{N}\mathcal{I}_M\frac{1}{K-1}\sum_{k\neq y_i}\left(\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)_k-\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)_{y_i}\right)^2\\ &\overset{(a)}{\geq}\frac{1}{N^M}\sum_{i=1}^{N}\mathcal{I}_M\frac{1}{K-1}\sum_{k\neq y_i}\left(\frac{1}{K-1}\sum_{k\neq y_i}\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)_k-\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)_{y_i}\right)^2\\ &\overset{(b)}{\approx}\frac{1}{N^M}\sum_{i=1}^{N}\mathcal{I}_M\frac{1}{K-1}\sum_{k\neq y_i}\left(\frac{1}{K-1}\sum_{k\neq y_i}\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)_k-\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)_k\right)^2\end{aligned} \tag{14}$$

Given the convexity of the mean square loss function, the inference of (a) can be efficiently deduced using Jensen's inequality. Moreover, when considering misclassified samples, assuming a uniform distribution of logits for correct labels is plausible relative to other logits. This assumption enables us to approximate $\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)y_i$ with $\sigma\left(\mathcal{M}(\mathbf{x}_i)\right)k$. Therefore, under these premises, the validity of (b) is affirmed, and the expression on its right side can be represented as $\check{\mathcal{P}}_M^N$.

Throughout the implementation of the SLORE procedure, effective regularization of $\mathcal{P}_C^N$ and $\mathcal{P}_M^N$ is achieved by concentrating on their respective lower bounds, $\check{\mathcal{P}}_C^N$ and $\check{\mathcal{P}}_M^N$. This approach facilitates the efficient optimization of the lower bounds $\overline{\gamma}_{ce}^{N^M}$ and $\overline{\gamma}_{ce}^{N^C}$ in relation to $\hat{\gamma}_{ce}^{N^M}$ and $\hat{\gamma}_{ce}^{N^C}$.

4.5.2 Adaptive Adversarial Logit Pairing

The surrogate terms $\check{\mathcal{P}}_M^N$ and $\check{\mathcal{P}}_C^N$ serve as an innovative method for regularizing $\hat{\gamma}_{ce}^{N^M}$ and $\hat{\gamma}_{ce}^{N^C}$. However, as explicated in Eq.13 and Eq.14, penalizing $\mathcal{P}_M^N$ and $\mathcal{P}_C^N$ may inadvertently lead to the standard logits being paired with logits from an unknown distribution. This potential misalignment poses a risk to the model's overall robustness.

Algorithm 1 Logit-Oriented Adversarial Training (LOAT)

Input: Victim model $\mathcal{M}$, training set (X, Y), $\mathcal{A}_p^\epsilon(\cdot)$-based adversarial training objective $\mathcal{L}_{\mathcal{A}_p^\epsilon}(\cdot)$, epoch-wise breakpoints $\mathcal{E}_1$ and $\mathcal{E}_2$, number of training epochs $\hat{\mathcal{E}}$, hyperparameter τ and γ, LOAT type T, and updated loss L.
Output: LOAT-boosted adversarial-trained model $\mathcal{M}$.
for $\mathcal{E}$ = 1, 2, ..., $\hat{\mathcal{E}}$ **do**
  for (X^j, Y^j) in batch j **do**
    if $\mathcal{E} \leq \mathcal{E}_1$ **then**
      if T = SLORE **then**
        L = $\mathcal{L}_{\mathcal{A}_p^\epsilon}((X^j, Y^j), \mathcal{M}) + \tau \cdot (\check{\mathcal{P}}_C^{N_j} - \check{\mathcal{P}}_M^{N_j})$
      else if T=LORE **then**
        L = $\mathcal{L}_{\mathcal{A}_p^\epsilon}((X^j, Y^j), \mathcal{M}) + \tau \cdot (\check{\mathcal{P}}_C^{N_j} - \check{\mathcal{P}}_M^{N_j}) + \gamma * \mathcal{AP}_M^{N_j}$
      end if
    else if $\mathcal{E} \geq \mathcal{E}_2$ **then**
      if T = SLORE **then**
        L = $\mathcal{L}_{\mathcal{A}_p^\epsilon}((X^j, Y^j), \mathcal{M}) + \tau \cdot (\check{\mathcal{P}}_M^{N_j} - \check{\mathcal{P}}_C^{N_j})$
      else if T=LORE **then**
        L = $\mathcal{L}_{\mathcal{A}_p^\epsilon}((X^j, Y^j), \mathcal{M}) + \tau \cdot (\check{\mathcal{P}}_M^{N_j} - \check{\mathcal{P}}_C^{N_j}) + \gamma * \mathcal{AP}_C^{N_j}$
      end if
    else
      L = $\mathcal{L}_{\mathcal{A}_p^\epsilon}((X^i, Y^i), \mathcal{M})$
    end if
    $\mathcal{M} \leftarrow$ Update Model With Loss L.
  end for
end for
return $\mathcal{M}$

To mitigate the potential drawbacks identified, we incorporate Adversarial Logit Pairing (ALP) as described in [13], tailoring an adversary-focused regularization strategy. In the early stages of training, where standard penalization is applied for the decrease of $\check{\mathcal{P}}_M^N$, we introduce an adversarial penalty, $\mathcal{AP}_M^N$. This penalty is designed to diminish the disparity between the logits of misclassified clean samples and

Table 1. Classification Accuracy on Cifar10 (%). **Bold** and *italic* texts represent the highest and the second-highest accuracy in each block respectively.

Models	Defense	Clean_{tr}	Clean_{te}	FGSM	PGD^7	PGD^{20}	T/E
ResNet18	TRADES	93.39	82.23	57.89	*53.68*	**52.00**	216s
	TRADES+SLORE	*94.77*	*82.29*	*62.19*	53.32	49.77	218s
	TRADES+LORE	**95.35**	**82.51**	**62.47**	**53.96**	*50.43*	219s
	TRADES(S)	*92.92*	*82.95*	58.14	54.01	**52.71**	206s
	TRADES(S)+SLORE	**93.25**	**83.01**	*62.32*	**54.99**	*51.96*	208s
	TRADES(S)+LORE	92.79	82.73	**62.76**	*54.06*	51.03	209s
	MART	87.85	80.61	62.41	54.86	51.65	98s
	MART+SLORE	*88.07*	*80.81*	*62.61*	*55.06*	*51.89*	100s
	MART+LORE	**89.04**	**81.53**	**63.35**	**55.58**	**52.15**	100s
	PGD-AT	**97.62**	83.19	59.99	49.86	44.94	158s
	PGD-AT+SLORE	97.13	**83.41**	**61.80**	**51.46**	**46.50**	160s
	PGD-AT+LORE	*97.38*	*83.25*	*60.94*	*50.28*	*45.30*	161s
ResNet50	TRADES	93.89	83.38	*64.53*	*56.23*	**52.39**	674s
	TRADES+SLORE	*94.74*	**84.07**	64.32	55.78	52.15	676s
	TRADES+LORE	**94.83**	*83.81*	**64.69**	**56.30**	*52.34*	679s
	MART	87.66	80.49	*62.66*	*55.19*	*51.62*	209s
	MART+SLORE	*88.25*	**81.29**	62.61	54.76	51.23	213s
	MART+LORE	**88.29**	*80.54*	**62.90**	**55.43**	**52.00**	214s
	PGD-AT	*94.77*	82.92	60.79	50.74	46.03	174s
	PGD-AT+SLORE	93.51	*82.92*	*62.34*	*52.44*	48.14	184s
	PGD-AT+LORE	**95.09**	**83.05**	**62.94**	**53.37**	**48.98**	185s
WRN-34-10	TRADES	98.98	84.57	*63.45*	52.60	**48.15**	755s
	TRADES+SLORE	99.65	**85.22**	63.34	*52.66*	47.60	763s
	TRADES+LORE	**99.69**	*85.14*	**63.49**	**52.89**	*47.84*	768s
	MART	90.21	84.36	68.26	71.55	58.34	369s
	MART+SLORE	**91.11**	**85.17**	**69.35**	**72.87**	**59.10**	374s
	MART+LORE	*90.89*	*85.00*	*68.92*	*72.32*	*59.64*	374s
	PGD-AT	**99.90**	84.47	59.89	48.48	43.68	520s
	PGD-AT+SLORE	*99.88*	*84.68*	*60.18*	*49.49*	*44.79*	528s
	PGD-AT+LORE	99.81	**85.29**	**61.04**	**49.53**	**44.79**	529s

their adversarial counterparts. In contrast, during the latter epochs, we apply a similar adversarial penalty, $\mathcal{AP}_C^N$, but focus on reducing the logit distance for correctly classified clean samples compared to their adversarial equivalents. By employing the adversarial generator $\mathcal{A}_p^\epsilon(\cdot)$, we can formally express $\mathcal{AP}_M^N$ and $\mathcal{AP}_C^N$ following the derivations of $\check{\mathcal{P}}_M^N$ and $\check{\mathcal{P}}_C^N$:

$$\mathcal{AP}_M^N = \frac{1}{N^M}\sum_{i=1}^{N}\mathcal{I}_M\sum_{k=1}^{K}(\sigma(\mathcal{M}(\mathbf{x}_i))_k - \sigma(\mathcal{A}_p^\epsilon(\mathcal{M}(\mathbf{x}_i)))_k)^2$$
$$\mathcal{AP}_C^N = \frac{1}{N^C}\sum_{i=1}^{N}\mathcal{I}_C\sum_{k=1}^{K}(\sigma(\mathcal{M}(\mathbf{x}_i))_k - \sigma(\mathcal{A}_p^\epsilon(\mathcal{M}(\mathbf{x}_i)))_k)^2 \tag{15}$$

More precisely, the integration of standard logit-oriented regularization with adaptive adversarial logit pairing forms the basis of what we call Logit-Oriented REgularization (LORE). Comprehensive details and the procedural steps of the corresponding training algorithm are delineated in Alg. 1.

Table 2. Classification Accuracy of models trained by 1×10^6 EDM-generated images-augmented Cifar10 (%).

Models	Defense	Clean_{tr}	Clean_{te}	PGD^{40}	AA	T/E
PreResNet18 (Swish)	TRADES	92.54	86.17	58.47	**54.65**	744s
	TRADES+SLORE	*94.48*	*88.35*	*59.00*	53.76	749s
	TRADES+LORE	**94.69**	**88.46**	**59.00**	*53.87*	750s
	TRADES(S)	92.64	86.52	*59.74*	*54.61*	743s
	TRADES(S)+SLORE	**94.91**	**88.94**	58.87	52.41	745s
	TRADES(S)+LORE	*93.30*	*86.85*	**60.08**	**55.06**	745s
	MART	85.52	84.86	58.51	51.70	745s
	MART+SLORE	**89.25**	*86.29*	**62.11**	**51.96**	747s
	MART+LORE	*87.75*	**87.05**	*61.67*	*51.92*	747s
WRN-28-10 (Swish)	TRADES	94.80	88.87	*63.16*	**59.44**	754s
	TRADES+SLORE	*96.39*	*90.36*	62.95	58.12	756s
	TRADES+LORE	**96.58**	**90.67**	**63.57**	*58.74*	757s
	TRADES(S)	94.76	88.66	*63.78*	*58.85*	784s
	TRADES(S)+SLORE	**96.40**	**90.82**	63.09	56.86	798s
	TRADES(S)+LORE	*95.30*	*89.45*	**64.17**	**59.54**	798s
	MART	87.22	88.83	62.91	57.08	755s
	MART+SLORE	**90.50**	*89.82*	*65.78*	*57.44*	757s
	MART+LORE	*89.75*	**90.18**	**66.03**	**58.07**	758s

5. Experiments

In this section, we comprehensively evaluate the performance enhancement of LOAT regarding standard accuracy and adversarial robustness. We test against a spectrum of attacks, including FGSM [6], PGD-7/20/40 [17], and AutoAttack (AA) [5].

Experimental Setup. Our experiments are conducted under the ℓ_∞ threat model. We set the perturbation radius (ϵ) to 8/255, the number of perturbation steps (k) to 10, and the step size (α) to 2/255. All experiments utilize dual GeForce RTX 3090 GPUs, with each experimental run replicated thrice to ensure reliability and consistency.

For the CIFAR10 dataset, our models include ResNet-18, ResNet-50, and WideResNet-34-10, each employing PGD-AT, TRADES, TRADES (LSE) (also referred to as TRADES (S)), and MART, respectively. The trade-off factor (β) for TRADES, TRADES (LSE), and MART is uniformly set at 6.0. The training duration is set for 110 epochs. As outlined in Alg. 1, we establish epoch-wise breakpoints, $\mathcal{E}_1$ and $\mathcal{E}_2$, at 1 and 100 for ResNet-18 and ResNet-50, and 1 and 85 for WideResNet-34-10, respectively.

For experiments on the Elucidating Diffusion Model-augmented CIFAR10 (DM-AT), we use ResNet-18 and WideResNet-34-10 as the backbones, conducting TRADES, TRADES (S), and MART individually. Here, β is set at 5.0, with the training spanning 100 epochs. The breakpoints $\mathcal{E}_1$ and $\mathcal{E}_2$ are set at 1 and 90, respectively.

Boost in Standard Accuracy. As detailed in Tab. 1, MART+LORE achieves a significant 0.92% increase in accuracy for ResNet-18. For ResNet-50 and WideResNet-34-10, MART+SLORE leads to improvements of 0.80% and 0.81%, respectively. In the case of TRADES+SLORE,

Table 3. Ablation study for $\leftarrow \mathcal{E}_1$ and $\mathcal{E}_2 \rightarrow$ on ResNet-18 (%) .

PGD-AT	SLORE					LORE				
	Clean_{tr}	Clean_{te}	FGSM	PGD^7	PGD^{20}	Clean_{tr}	Clean_{te}	FGSM	PGD^7	PGD^{20}
$\leftarrow \mathcal{E}_1$	96.84	83.39	59.80	48.31	43.86	96.50	83.07	59.75	48.80	44.09
$\mathcal{E}_2 \rightarrow$	96.34	82.65	59.68	49.33	45.18	96.65	82.67	60.46	**50.85**	**45.54**
$\leftarrow \mathcal{E}_1, \mathcal{E}_2 \rightarrow$	**97.13**	**83.41**	**61.80**	**51.46**	**46.50**	**97.38**	**83.25**	**60.94**	50.28	45.30

MART	SLORE					LORE				
	Clean_{tr}	Clean_{te}	FGSM	PGD^7	PGD^{20}	Clean_{tr}	Clean_{te}	FGSM	PGD^7	PGD^{20}
$\leftarrow \mathcal{E}_1$	88.03	80.76	62.60	**55.18**	51.64	88.01	80.62	62.81	55.25	51.75
$\mathcal{E}_2 \rightarrow$	88.04	80.79	62.58	55.13	51.67	88.28	80.89	62.92	55.60	**52.57**
$\leftarrow \mathcal{E}_1, \mathcal{E}_2 \rightarrow$	**88.07**	**80.81**	**62.61**	55.06	**51.89**	**89.04**	**81.53**	**63.35**	**55.58**	52.15

Table 4. Ablation study for $\check{\mathcal{P}}_C$ and $\check{\mathcal{P}}_M$ on ResNet-18 (%).

PGD-AT	SLORE					LORE				
	Clean_{tr}	Clean_{te}	FGSM	PGD^7	PGD^{20}	Clean_{tr}	Clean_{te}	FGSM	PGD^7	PGD^{20}
$\check{\mathcal{P}}_C$	96.49	83.01	59.71	49.62	44.71	96.15	82.80	60.02	49.40	44.77
$\check{\mathcal{P}}_M$	96.78	82.92	60.79	50.19	45.52	97.09	82.72	60.19	49.68	**45.44**
$\check{\mathcal{P}}_C, \check{\mathcal{P}}_M$	**97.13**	**83.41**	**61.80**	**51.46**	**46.50**	**97.38**	**83.25**	**60.94**	**50.28**	45.30

MART	SLORE					LORE				
	Clean_{tr}	Clean_{te}	FGSM	PGD^7	PGD^{20}	Clean_{tr}	Clean_{te}	FGSM	PGD^7	PGD^{20}
$\check{\mathcal{P}}_C$	87.98	80.22	62.77	55.29	52.24	88.01	80.62	62.81	55.25	51.75
$\check{\mathcal{P}}_M$	**88.34**	**80.88**	**62.99**	**55.24**	**51.94**	88.28	80.89	62.92	**55.60**	**52.57**
$\check{\mathcal{P}}_C, \check{\mathcal{P}}_M$	88.07	80.81	62.61	55.06	51.89	**89.04**	**81.53**	**63.35**	55.58	52.15

Table 5. Evaluation of LOAT-boosted S2O on ResNet-18 (%).

S2O	TRADES				MART				PGD-AT			
	Clean_{te}	PGD^7	PGD^{20}	PGD^{40}	Clean_{te}	PGD^7	PGD^{20}	PGD^{40}	Clean_{te}	PGD^7	PGD^{20}	PGD^{40}
Vanilla	83.90	54.87	52.11	51.98	82.08	54.84	*51.66*	*51.38*	84.41	50.98	47.05	46.57
SLORE	*84.30*	*55.06*	**52.31**	*52.05*	**82.45**	*54.84*	51.53	51.23	**84.90**	**52.39**	**47.92**	**47.54**
LORE	**84.90**	**55.08**	*52.26*	**52.19**	*82.32*	**55.05**	**52.37**	**52.14**	*84.86*	*51.31*	*47.24*	*46.74*

there is an enhancement in clean accuracy by 0.69% for ResNet-50 and 0.65% for WideResNet-34-10. Additionally, PGD-AT+LORE boosts WideResNet-34-10 performance by 0.82%.

These performance gains are even more pronounced in settings augmented with EDM-generated data. As Tab. 2 indicates, TRADES+LORE registers a substantial increase of 2.29% and 1.80% in standard accuracy for Swish-activated PreAct ResNet-18 (abbreviated as PreResNet18 (Swish)) and Swish-activated WideResNet-28-10 (abbreviated as WRN-28-10 (Swish)), respectively. Meanwhile, MART achieves boosts of 2.19% and 1.35% in these models. Further, TRADES (S)+SLORE records increases of 2.42% and 2.16% in PreResNet18 (Swish) and WRN-28-10 (Swish), respectively.

Boost in Adversarial Accuracy. In the case of CIFAR10 trained models (referenced in Tab. 1), LOAT significantly bolsters the robustness of MART and PGD-AT against all evaluated adversarial attacks. Although its impact on TRADES and TRADES (S) is less pronounced, it still ensures effective defence, especially in ResNet-18. For instance, TRADES's accuracy against the FGSM attack is improved by 4.58%, and TRADES (S)'s by 4.62%. In scenarios involving DM-AT, our defence strategy demonstrates even greater effectiveness. Notably, against the highly potent AutoAttack, our approach distinctly improves the robustness of TRADES (S) and MART. Specifically, the accuracy of TRADES (S) against AutoAttack is increased by 0.45% for PreResNet18 (Swish) and by 0.69% for WideResNet-28-10 (Swish). Similarly, MART's accuracy against AutoAttack is elevated by 0.26% for PreResNet18 (Swish) and 0.99% for WideResNet-28-10 (Swish).

Ablation Study. In accordance with Alg. 1, we assess the effectiveness of applying regularizations before $\mathcal{E}_1$, after $\mathcal{E}_2$, as well as their combination. Additionally, we examine the impacts of regularizations of $\check{\mathcal{P}}_C$ and $\check{\mathcal{P}}_M$ individually and in combination. Detailed experimental results are shown in Tab. 3, 4. Our findings indicate that for MART, which has been regularized, the sole application of $\check{\mathcal{P}}_M$ yields the most favourable results. We hypothesize that this is due to MART's already robust standard generalization performance, which renders the influence of $\check{\mathcal{P}}_C$ less pronounced.

Adaptibility to Weight Regularizations. LOAT is architecture-agnostic and algorithm-agnostic, suggesting that it can be integrated seamlessly into other weight-oriented regularization algorithms to enhance adversarial training. To empirically validate this hypothesis, we apply our framework to train algorithms boosted by Second-Order Statistics Optimization (S2O). The extensive results of this integration are displayed in Tab. 5.

Time Efficiency. We measure the time taken for each epoch (Time/Epochs) and present the detailed findings in Tab. 1 and 2, which confirm that LOAT introduces minimal computational overhead, maintaining time efficiency.

More Experiments. To solidify the credibility of LOAT, we extended our experimentation to CIFAR100 and SVHN. Additional results can be found in the Appendix.

6. Conclusions

In this paper, we address the challenge of maintaining standard generalization performance in adversarial-trained deep neural networks. We explore the Rademacher complexity of ReLU-activated MLPs through the lens of the Fisher-Rao norm. This perspective enables us to introduce a logit-based variable, which exhibits sensitivity to various factors related to model complexity, including model width and training objective. Also, it shows a strong correlation with the generalization gap of CE loss between adversarial-trained and standard-trained models. We leverage these insights to develop the Logit-Oriented Adversarial Training (LOAT) framework, enhancing adversarial training without significant computational cost. Our comprehensive experiments on prevalent adversarial training algorithms and diverse network architectures confirm the effectiveness of LOAT in mitigating the trade-off between robustness and accuracy.

7. Acknowledgements

This work is supported by the University of Liverpool and the China Scholarship Council (CSC).

References

[1] Peter Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks, 2017. 2, 4

[2] Peter L Bartlett and Wolfgang Maass. Vapnik-chervonenkis dimension of neural nets. *The handbook of brain theory and neural networks*, pages 1188–1192, 2003. 2

[3] 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. 3

[4] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars, 2016. 1

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 7

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 7

[7] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. Adversarial examples for malware detection. In *Computer Security–ESORICS 2017: 22nd European Symposium on Research in Computer Security, Oslo, Norway, September 11-15, 2017, Proceedings, Part II 22*, pages 62–79. Springer, 2017. 1

[8] Xia Hu, Lingyang Chu, Jian Pei, Weiqing Liu, and Jiang Bian. Model complexity of deep learning: A survey. *Knowledge and Information Systems*, 63:2585–2619, 2021. 1

[9] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020. 1

[10] Xiaowei Huang, Gaojie Jin, and Wenjie Ruan. Deep reinforcement learning. In *Machine Learning Safety*, pages 219–235. Springer, 2023. 1

[11] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *arXiv preprint arXiv:2305.11391*, 2023. 1

[12] Gaojie Jin, Xinping Yi, Wei Huang, Sven Schewe, and Xiaowei Huang. Enhancing adversarial training with second-order statistics of weights, 2022. 2

[13] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing, 2018. 6

[14] Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991. 2

[15] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd international conference on artificial intelligence and statistics*, pages 888–896. PMLR, 2019. 1, 2, 4

[16] Chunchuan Lyu, Kaizhu Huang, and Hai-Ning Liang. A unified gradient regularization family for adversarial examples, 2015. 2

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 7

[18] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa, 2018. 2

[19] Ronghui Mu, Wenjie Ruan, Leandro Soriano Marcolino, and Qiang Ni. Sparse adversarial video attacks with spatial transformations. In *The 32nd British Machine Vision Conference (BMVC'21)*, 2021. 1

[20] Ronghui Mu, Wenjie Ruan, Leandro Soriano Marcolino, Gaojie Jin, and Qiang Ni. Certified policy smoothing for cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'23)*, 2023. 1

[21] Preetum Nakkiran. Adversarial robustness may be at odds with simplicity, 2019. 2

[22] Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks, 2015. 4

[23] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks, 2015. 2, 4

[24] Tianyu Pang, Min Lin, Xiao Yang, Junyi Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, 2022. 1, 2, 6

[25] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018. 1

[26] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning, 2020. 2

[27] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, 2017. 2

[28] 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. 1

[29] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy, 2019. 1, 2

[30] Fu Wang, Chi Zhang, Peipei Xu, and Wenjie Ruan. Deep learning and its adversarial robustness: A brief introduction. In *HANDBOOK ON COMPUTER LEARNING AND INTELLIGENCE: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation*, pages 547–584. 2022. 1

[31] Fu Wang, Zeyu Fu, Yanghao Zhang, and Wenjie Ruan. Self-adaptive adversarial training for robust medical segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'23)*, pages 725–735. Springer, 2023. 2

[32] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020. 1

[33] Zheng Wang and Wenjie Ruan. Understanding adversarial robustness of vision transformers via cauchy problem. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD'22)*, 2022. 2

[34] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. *arXiv preprint arXiv:2302.04638*, 2023. 1

[35] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020. 2

[36] Han Wu, Syed Yunas, Sareh Rowlands, Wenjie Ruan, and Johan Wahlström. Adversarial driving: Attacking end-to-end autonomous driving. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pages 1–7. IEEE, 2023. 1

[37] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Ruslan Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness, 2020. 1, 2

[38] Xiangyu Yin, Wenjie Ruan, and Jonathan Fieldsend. Dimba: discretely masked black-box attack in single object tracking. *Machine Learning*, pages 1–19, 2022. 1

[39] Xiangyu Yin, Sihao Wu, Jiaxu Liu, Meng Fang, Xingyu Zhao, Xiaowei Huang, and Wenjie Ruan. Rerogcrl: Representation-based robustness in goal-conditioned reinforcement learning. *arXiv preprint arXiv:2312.07392*, 2023. 2

[40] Yaodong Yu, Zitong Yang, Edgar Dobriban, Jacob Steinhardt, and Yi Ma. Understanding generalization in adversarial training via the bias-variance decomposition, 2021. 1, 2

[41] Chi Zhang, Wenjie Ruan, and Peipei Xu. Reachability analysis of neural network control systems. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'23)*, 2023. 1

[42] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. *ArXiv*, abs/1901.08573, 2019. 2

[43] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy, 2019. 1