Combining Adversaries with Anti-adversaries in Training

Xiaoling Zhou, Nan Yang, Ou Wu *

Center for Applied Mathematics, Tianjin University, China
{xiaolingzhou, yny, wuou}@tju.edu.cn

Abstract

Adversarial training is an effective learning technique to improve the robustness of deep neural networks. In this study, the influence of adversarial training on deep learning models in terms of fairness, robustness, and generalization is theoretically investigated under more general perturbation scope that different samples can have different perturbation directions (the adversarial and anti-adversarial directions) and varied perturbation bounds. Our theoretical explorations suggest that the combination of adversaries and anti-adversaries (samples with anti-adversarial perturbations) in training can be more effective in achieving better fairness between classes and a better tradeoff between robustness and generalization in some typical learning scenarios (e.g., noisy label learning and imbalance learning) compared with standard adversarial training. On the basis of our theoretical findings, a more general learning objective that combines adversaries and anti-adversaries with varied bounds on each training sample is presented. Meta learning is utilized to optimize the combination weights. Experiments on benchmark datasets under different learning scenarios verify our theoretical findings and the effectiveness of the proposed methodology.

Introduction

Apart from the standard generalization error (also known as natural error), robust generalization error (also known as robust error) has received great attention in recent years. A deep neural network with a low robust error can cope well with adversarial attacks. Adversarial training is an effective technique to reduce the robust error of a model (Wong, Rice, and Kolter 2020; Bai and Luo 2021). Given a model $f(\cdot)$ and a sample $\boldsymbol{x}$ associated with a label y, classical adversarial training methods (Madry et al. 2018; Goodfellow, Shlens, and Szegedy 2014) first generate an adversary (i.e., adversarial example) $\boldsymbol{x}_{\text{adv}}$ for $\boldsymbol{x}$ with the following optimization:

$$\boldsymbol{x}_{\text{adv}} = \boldsymbol{x} + \arg\max_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell(f(\boldsymbol{x}+\boldsymbol{\delta}), y), \tag{1}$$

where $\ell(\cdot,\cdot)$ is a loss function, $\boldsymbol{\delta}$ is the perturbation term, and ϵ is the perturbation bound. Adversaries are then leveraged as the training data to learn a more robust model. A number of variations for adversarial training have been proposed in recent literature. Zhang et al. (2019) decomposed the robust error into the natural and boundary errors. They developed a new method, namely, TRADES, to obtain a better tradeoff between standard generalization and robustness. Wang et al. (2020) proposed a misclassification-aware adversarial training method to focus on the misclassified examples.

In addition to the design of new methods, theoretical studies have been conducted to explore the effectiveness and ineffectiveness of adversarial training (Bai and Luo 2021). Yang et al. (2020) concluded that existing adversarial methods cannot achieve an ideal tradeoff between accuracy and robustness due to the insufficient smoothness (Xie et al. 2020) and generalization properties of classifiers trained by these methods. They pointed out that customized optimization methods or better network architectures should be proposed. Xu et al. (2021) revealed that adversarial training introduces severe unfairness between different categories. Thus, they developed a new method that sets varied perturbation bounds for each class, resulting in better fairness. Different from these studies, we conjectured that one possible reason leading to unsatisfied tradeoff and fairness is that not all training samples should be perturbed adversarially. For instance, adversaries of noisy samples may harm the model performance (Uesato et al. 2019), and these samples should be perturbed in the anti-adversarial direction to reduce their negative influence on model optimization. Zhu et al. (2021) re-annotated pseudo labels for possible noisy samples before generating adversaries for them. The generated adversaries are actually perturbed anti-adversarially in binary classification tasks. In this study, samples with anti-adversarial perturbations are called anti-adversaries[1] ($\boldsymbol{x}_{\text{at-adv}}$)

$$\boldsymbol{x}_{\text{at-adv}} = \boldsymbol{x} + \arg\min_{\|\boldsymbol{\delta}\|\leq\epsilon} \ell(f(\boldsymbol{x}+\boldsymbol{\delta}), y). \tag{2}$$

This study conducts a comprehensive theoretical analysis of adversarial training in the presence of two different perturbation directions (adversarial and anti-adversarial) and varied bounds. Several typical learning scenarios are considered, including classes with different learning difficulties, imbalance learning, and noisy label learning. Our theoretical findings reveal that the perturbation directions and

*Corresponding author.

[1]The anti-adversary defined by Alfarra et al. (2022) is different from ours. They utilize anti-adversaries to deal with attacks, whereas we aim to improve robustness, accuracy, and fairness.

bounds can remarkably influence the model training. The combination of adversaries and anti-adversaries with varied bounds can improve the fairness among classes and achieve a better tradeoff between accuracy and robustness. Accordingly, a general objective that combines adversaries and anti-adversaries is constructed for adversarial training. A meta learning-based method is then proposed to optimize this objective, in which the perturbation direction and bound of each training sample is adjusted in accordance with its learning characteristics during training. Our experimental results show that the combining strategy outperforms state-of-the-art adversarial training methods. Our experimental observations are in accordance with our theoretical findings.

The contributions of our study are as follows:

- To the best of our knowledge, this is the first work that combines adversaries and anti-adversaries in training. A comprehensive theoretical analysis is conducted for the role of the combination strategy with varied perturbation bounds[2] under three typical learning scenarios.
- A new objective is established for adversarial training by combining adversaries and anti-adversaries. Meta learning is utilized to solve the optimization, and the perturbation direction and bound for each training sample are determined in accordance with its learning characteristics, such as training loss and margin.

Related Work

Tradeoff and Fairness in Adversarial Training

Recent studies on adversarial training focus on the tradeoff between accuracy and robustness. Efforts (Raghunathan et al. 2019; Zhang et al. 2019, 2020; Yang et al. 2021) have been made to reduce the natural errors of the adversarially trained models, such as adversarial training with semi/unsupervised learning and robust local feature (Song et al. 2020). Rice et al. (2020) systematically investigated the role of various techniques used in deep learning for achieving a better tradeoff, such as cutout, mixup, and early stopping, where early stopping is found to be the most effective. This investigation was also confirmed by Pang et al. (2021). Unfairness is also a problem caused by adversarial training. Xu et al. (2021) trained a robust classifier to minimize error and stressed it to satisfy two fairness constraints. Several studies (Ding et al. 2020; Cheng et al. 2020; Balaji, Goldstein, and Hoffman 2019) adaptively tune the perturbation bounds for each sample with the inspiration that samples near the decision boundary should have small bounds.

Meta Learning

Meta learning has aroused great interest in recent years. Existing meta learning methods can be divided into three categories, namely, metric-based (Snell, Swersky, and Zemel 2017; Sung et al. 2018), model-based (Santoro et al. 2016), and optimizing-based (Finn, Abbeel, and Levine 2017; Nichol, Achiam, and Schulman 2018) methods. The algorithm we adopted that is inspired by Model-Agnostic Meta-Learning (Finn, Abbeel, and Levine 2017) belongs to the optimizing-based methods. The data-driven manner of meta optimization is always utilized to learn the sample weights or the hyperparameters (Ren et al. 2018; Shu et al. 2019).

[2]Existing theoretical studies presume that the perturbation bounds are identical for all training samples.

Theoretical Investigation

This section conducts theoretical analyses to assess the influence of two different perturbation directions and varied bounds on adversarial training in three typical binary classification cases. Proofs are presented in the online material.

Notation

We denote the sample instance as $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$ as the label, where $\mathcal{X} \subseteq \mathbb{R}^d$ indicates the instance space, and $\mathcal{Y} = \{-1, +1\}$ indicates the label space. The classification model f is a mapping from the input data space $\mathcal{X}$ to the label space $\mathcal{Y}$. It can be parametrized by using linear classifiers or deep neural networks. The overall natural error of f is denoted as $\mathcal{R}_{\text{nat}}(f) := \Pr(f(\boldsymbol{x}) \neq y)$. The overall robust error is denoted as $\mathcal{R}_{\text{rob}}(f) := \Pr(\exists \boldsymbol{\delta} \ ||\boldsymbol{\delta}|| \leq \epsilon, \text{s.t.} f(\boldsymbol{x} + \boldsymbol{\delta}) \neq y)$.

Case I: Classes with Different Difficulties

In this case, the binary setting established by Xu et al. (2021) is followed. The data from each class follow a Gaussian distribution $\mathcal{D}$ that is centered on $\boldsymbol{\theta}$ and $-\boldsymbol{\theta}$, respectively. A K-factor difference is found between two classes' variances: $\sigma_{+1} : \sigma_{-1} = K : 1$ and $K > 1$. The data follow

$$y \overset{u.a.r}{\sim} \{-1, +1\}, \quad \boldsymbol{\theta} = (\eta, \ldots, \eta) \in \mathbb{R}^d, \eta > 0,$$
$$\boldsymbol{x} \sim \begin{cases} \mathcal{N}\left(\boldsymbol{\theta}, \sigma_{+1}^2 \boldsymbol{I}\right), & \text{if } y = +1, \\ \mathcal{N}\left(-\boldsymbol{\theta}, \sigma_{-1}^2 \boldsymbol{I}\right), & \text{if } y = -1. \end{cases} \tag{3}$$

Class "$+1$" is harder because the optimal linear classifier will give a larger error to class "$+1$" than class "-1". Xu et al. (2021) proved that adversarial training with an equal bound will exacerbate the performance gap (including natural and robust errors) between classes and hurt the harder class. We show that adversarial training with unequal bounds on two classes can tune the performance gap and the tradeoff between the robustness and accuracy of the model. Let $\sigma_{-1} = \sigma$. The following theorem is first proposed.

Theorem 1 *For a data distribution $\mathcal{D}$ in Eq. (3), assume that the perturbation bounds of class "-1" and "$+1$" are ϵ and $\rho \times \epsilon$ ($0 \leq \epsilon, \rho\epsilon < \eta$), respectively. The natural errors of the optimal robust linear classifier f_{rob} for two classes are*

$$\mathcal{R}_{nat}(f_{rob}, -1) = \Pr\left\{\mathcal{N}(0,1) \leq B - K \cdot \sqrt{B^2 + q(K)} - \frac{\sqrt{d}}{\sigma}\epsilon\right\},$$
$$\mathcal{R}_{nat}(f_{rob}, +1) = \Pr\left\{\mathcal{N}(0,1) \leq -K \cdot B + \sqrt{B^2 + q(K)} - \frac{\sqrt{d}\rho}{K\sigma}\epsilon\right\}, \tag{4}$$

where $B = \frac{2}{K^2-1} \frac{\sqrt{d}(\eta - \frac{\epsilon(1+\rho)}{2})}{\sigma}$, *and* $q(K) = \frac{2 \log K}{K^2-1}$.

The robust errors are shown in the online material. The natural and robust errors change with different ρ values. A corollary is derived in accordance with Theorem 1.

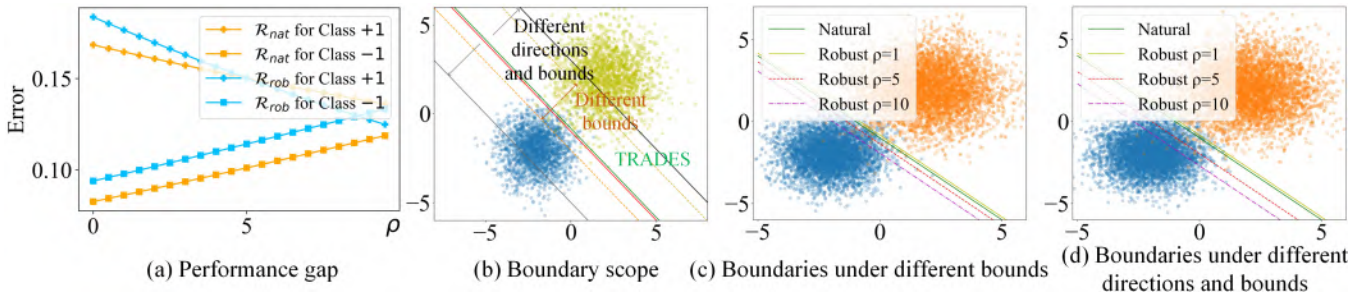

Figure 1: (a) Variation of performance gaps between classes as ρ increases. (b) Scope of the classification boundary of different manners. The values of parameters are $K = 2$, $\eta = 2$, $\epsilon = 0.2$, and $\sigma = 1$. The bounds for class "$+1$" and "-1" are denoted as $\rho_+ \times \epsilon$ and $\rho_- \times \epsilon$ ($-\eta/\epsilon < \rho_+, \rho_- < \eta/\epsilon$), respectively. $\rho_+(\rho_-) < 0$ denotes that class "$+1(-1)$" is anti-adversarially perturbed. The online material provides the formulas of boundaries. (c) Logistic regression classifier boundaries (natural and robust) on simulated data in Eq. (3). (d) Logistic regression classifier boundaries (natural and robust with different directions).

Corollary 1 *The data and perturbations in Theorem 1 are followed. When $K < \exp(d(\eta-\epsilon)^2/2\sigma^2)$, the adversarially trained model will increase and decrease the natural and robust errors of class "-1" and class "$+1$", with the increase in ρ, respectively.*

Accordingly, the performance gaps of $\mathcal{R}_{\text{nat}}$ and $\mathcal{R}_{\text{rob}}$ decrease with the increase in ρ, and better fairness can be achieved, as shown in Fig. 1 (a). In Fig. 1 (c), the boundary shifts toward the easy class "-1". From Fig. 1 (b), adversarial training with varied bounds contributes to larger scope of the boundary compared with TRADES (Zhang et al. 2019). Thus, a better tradeoff can be attained. Therefore, fairness and tradeoff can be tuned with different ρ values. Next, anti-adversaries are considered. Assume that samples in class "-1" perform anti-adversarial perturbation. Similar to Theorem 1, a theorem calculating the natural and robust errors is proposed as shown in the online material. A corollary is then derived, indicating that the combination of adversaries and anti-adversaries can tune the performance gap and tradeoff.

Corollary 2 *For a data distribution $\mathcal{D}$ in Eq. (3), assume that class "-1" is anti-adversarially perturbed with the bound ϵ, and class "$+1$" is adversarially perturbed with the bound $\rho \times \epsilon$ ($0 \le \epsilon, \rho\epsilon < \eta$). When $K < \exp(d(\eta-\epsilon)^2/2\sigma^2)$, the adversarially trained model will increase and decrease the natural and robust errors of class "-1" and class "$+1$", with the increase in ρ, respectively.*

In accordance with Corollaries 1 and 2, the adversarial training and the combination strategy can nearly attain the same performance. However, the combination strategy can contribute to the largest scope of the boundary, as shown in Fig. 1 (b). Thus, the combination strategy is more effective in achieving a better tradeoff and fairness theoretically. As shown in Figs. 1 (c) and (d), the combination strategy has a more pronounced effect under the same bound (i.e., the same ρ), indicating that it needs smaller bounds when the same performance is achieved. Thus, the combination strategy is more efficient than only the adversarial perturbation, indicating that anti-adversaries are valuable.

Case II: Classes with Imbalanced Proportions

In this case, the two variances in Eq. (3) are assumed to be identical[3], that is, $\sigma_{+1} = \sigma_{-1} = \sigma$. However, $p(y=+1)$ (p_+) is no longer equal to $p(y=-1)$ (p_-). Without loss of generality, let $p_+ : p_- = 1 : V$ and $V > 1$.

Class "-1" is the majority category, and an optimal linear classifier will give a smaller natural error for class "-1" than class "$+1$", as proved in the online material. Similarly, we proved that standard adversarial training will exacerbate the performance gap between classes and hurt the smaller class. We then show that adversarial training with unequal bounds on the two classes will tune the performance gap between classes and the tradeoff between robustness and accuracy. The following theorem is first proposed.

Theorem 2 *For a data distribution $\mathcal{D}_V$ described above with the imbalance factor V, assume that the perturbation bounds of classes "-1" and "$+1$" are ϵ and $\rho \times \epsilon$ ($0 \le \epsilon, \rho\epsilon < \eta$), respectively. The natural errors of the optimal robust linear classifier f_{rob} for the two classes are*

$$\begin{aligned} \mathcal{R}_{nat}(f_{rob}, -1) &= \Pr\left\{\mathcal{N}(0,1) \le -A - \frac{\log V}{2A} - \frac{\sqrt{d}}{\sigma}\epsilon\right\}, \\ \mathcal{R}_{nat}(f_{rob}, +1) &= \Pr\left\{\mathcal{N}(0,1) \le -A + \frac{\log V}{2A} - \frac{\sqrt{d}\rho}{\sigma}\epsilon\right\}, \end{aligned} \tag{5}$$

where $A = \sqrt{d}(\eta - \epsilon(1+\rho)/2)/\sigma$.

A corollary is derived on the basis of Theorem 2.

Corollary 3 *The data and perturbations in Theorem 2 are followed. When $V < \exp(d(\eta-\epsilon)^2/2\sigma^2)$, the adversarially trained model will increase and decrease the natural and robust errors of class "-1" and class "$+1$", with the increase in ρ, respectively.*

From Corollary 3, the performance gaps between classes can be decreased with different ρ values. The boundary can be moved within the scope with different ρ values that covers the boundary of standard adversarial training. Therefore, a better tradeoff can be attained by adversarial training

[3]The case with different variances can be explored similarly.

with varied bounds. Next, the anti-adversaries are considered. We assume that samples in class "-1" perform anti-adversarial perturbation. Similar to Theorem 2, a theorem that portrays the training occasion where the adversaries and anti-adversaries are combined is proposed, as shown in the online material. A corollary is then derived.

Corollary 4 *For a data distribution $\mathcal{D}_V$ in Theorem 2, the perturbations in Corollary 2 are followed. When $V < \exp(d(\eta-\epsilon)^2/2\sigma^2)$, the adversarially trained model will increase and decrease the natural and robust errors of class "-1" and class "$+1$", with the increase in ρ, respectively.*

In accordance with Corollary 4, the performance gaps between classes can be tuned by the combination strategy. In addition, it can contribute to the larger scope of the classification boundary compared with only adversaries, and a better tradeoff can be attained. When the same performance is achieved, combining adversaries and anti-adversaries has a smaller bound. Therefore, the combination strategy is more efficient than only the adversarial perturbation. More details are presented in the online material.

Case III: Classes with Noisy Labels

In this case, the two classes' variances and prior probabilities are assumed to be identical, that is, $\sigma_{+1} = \sigma_{-1}$ and $p_+ = p_-$. Without loss of generality, class "-1" is assumed to contain flipped noisy labels. Two main conclusions are obtained. 1) The adversaries of noisy samples will harm the tradeoff and fairness of the robust model. 2) If noisy samples are anti-adversarially perturbed with a bound $\rho \times \epsilon$ and clean samples are adversarially perturbed with a bound ϵ, then the natural and robust errors of class "-1" and class "$+1$" will be decreased and increased with the increase in ρ, respectively. Thus, the combination strategy with varied bounds is effective in achieving a lower performance gap between classes and a better tradeoff between the accuracy and robustness on noisy data. The relevant theorems are shown in the online material.

Summarization

Our theoretical analysis comprehensively reveals that the perturbation directions and bounds remarkably influence the generalization, robustness, and fairness of the robust model under three typical learning scenarios. Adversarial training with different perturbation directions and bounds can better tune the performance gap between classes and the tradeoff between robustness and accuracy. Existing studies ignored anti-adversaries that are valuable. Thus, a new optimized objective considering anti-adversaries is proposed.

Methodology

Illuminated by the theoretical analysis, a new objective function is first established. Accordingly, a meta learning-based method that combines adversaries and anti-adversaries (CAAT) in training with a varied bound for each sample is proposed to solve the optimization, as shown in Fig. 2.

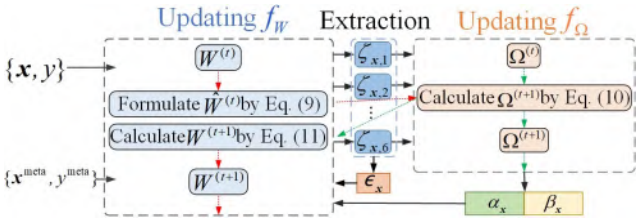

Figure 2: Overall structure of CAAT. The red and green lines represent the learning loops of the classifier network and weighting network, respectively.

Proposed Objective Function

Ideally, the objective function that combines adversaries and anti-adversaries can be formulated as

$$\min_{\boldsymbol{W},\alpha,\beta} \mathbb{E}_{\boldsymbol{x}}\{\alpha_{\boldsymbol{x}}\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_{\text{adv}}), y) + \beta_{\boldsymbol{x}}\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_{\text{at-adv}}), y)\}, \quad \text{s.t. } \alpha_{\boldsymbol{x}} + \beta_{\boldsymbol{x}} = 1 \text{ and } \alpha_{\boldsymbol{x}}, \beta_{\boldsymbol{x}} \in \{0, 1\}, \tag{6}$$

where $\boldsymbol{x}_{\text{adv}}$ and $\boldsymbol{x}_{\text{at-adv}}$ are calculated by using Eqs. (1) and (2) with varied bound $\epsilon_{\boldsymbol{x}}$ for each sample $\boldsymbol{x}$, respectively; $\alpha_{\boldsymbol{x}}$ and $\beta_{\boldsymbol{x}}$ are the combination weights; $f_{\boldsymbol{W}}$ is the classifier network with the parameter $\boldsymbol{W}$. When $\alpha_{\boldsymbol{x}} \equiv 1$, Eq. (6) can be reduced to the objective of standard adversarial training.

To solve Eq. (6), we first assume that the values of $\alpha_{\boldsymbol{x}}$ and $\beta_{\boldsymbol{x}}$ depend on the training characteristics of sample $\boldsymbol{x}$. Accordingly, their values are produced by a weighting network $f_{\boldsymbol{\Omega}}$ (parameterized by $\boldsymbol{\Omega}$), where its input is a series of training characteristics $\boldsymbol{\zeta}_{\boldsymbol{x}}$ of $\boldsymbol{x}$ shown in Fig. 2. $\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_{\text{adv}}), y)$ can be divided into $\ell(f_{\boldsymbol{W}}(\boldsymbol{x}), y)$ and $\ell(f_{\boldsymbol{W}}(\boldsymbol{x}), f_{\boldsymbol{W}}(\boldsymbol{x}_{\text{adv}}))$ to achieve a better tradeoff between the accuracy and robustness (Zhang et al. 2019). To improve the fairness among classes, we further stress f to satisfy two fairness constraints following Ref. Xu et al.(2021). Thus, our adopted objective function is

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{\Omega}} \mathbb{E}_{\boldsymbol{x}}\{&\alpha_{\boldsymbol{x}}[\ell(f_{\boldsymbol{W}}(\boldsymbol{x}), y) + \lambda\ell(f_{\boldsymbol{W}}(\boldsymbol{x}), f_{\boldsymbol{W}}(\boldsymbol{x}_{\text{adv}}))] \\ &+ \beta_{\boldsymbol{x}}\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_{\text{at-adv}}), y)\}, \\ \text{s.t. } &\begin{cases} [\alpha_{\boldsymbol{x}}, \beta_{\boldsymbol{x}}] = f_{\boldsymbol{\Omega}}(\boldsymbol{\zeta}_{\boldsymbol{x}}), \forall \boldsymbol{x} \in \mathcal{X}, \\ \mathcal{R}_{\text{nat}}(f_{\boldsymbol{W}}, c) - \mathcal{R}_{\text{nat}}(f_{\boldsymbol{W}}) \le \tau_1, \forall c \in \mathcal{Y}, \\ \mathcal{R}_{\text{bdy}}(f_{\boldsymbol{W}}, c) - \mathcal{R}_{\text{bdy}}(f_{\boldsymbol{W}}) \le \tau_2, \forall c \in \mathcal{Y}, \end{cases} \end{aligned} \tag{7}$$

where $\mathcal{R}_{\text{bdy}}$ is the boundary error of the model, denoted as $\mathcal{R}_{\text{bdy}}(f_{\boldsymbol{W}}) = \Pr(\exists \boldsymbol{x}_{\text{adv}} \in \mathbb{B}(\boldsymbol{x}, \epsilon), f_{\boldsymbol{W}}(\boldsymbol{x}_{\text{adv}}) \neq f_{\boldsymbol{W}}(\boldsymbol{x}))$; $\mathcal{R}_{\text{nat}}(f_{\boldsymbol{W}}, c) = \Pr(f_{\boldsymbol{W}}(\boldsymbol{x}) \neq y \mid y = c)$; $\mathcal{R}_{\text{bdy}}(f_{\boldsymbol{W}}, c) = \Pr(\exists \boldsymbol{x}_{\text{adv}} \in \mathbb{B}(\boldsymbol{x}, \epsilon), f_{\boldsymbol{W}}(\boldsymbol{x}_{\text{adv}}) \neq f_{\boldsymbol{W}}(\boldsymbol{x}) \mid y = c)$; $f_{\boldsymbol{\Omega}}$ is a multilayer perception (MLP) network with a hidden layer and a τ-softmax layer: Softmax$((\boldsymbol{h}\boldsymbol{\omega} + \boldsymbol{b})/\tau)$; $\lambda > 0$ is a regularization parameter that adjusts the influence of the natural and boundary errors on the model; τ_1 and τ_2 are small and positive predefined parameters. The approach for solving the two fairness constraints is the same as that in Ref. Xu et al.(2021), where a Lagrangian is formed.

Extraction of Training Characteristics ($\boldsymbol{\zeta}_{\boldsymbol{x}}$)

Our theoretical investigation reveals that different training samples can have different perturbation directions. The perturbation direction of a training sample depends on a series

Algorithm 1: CAAT

Input: #Iteration T, step sizes η_0, η_1, and η_2, batch size n, meta batch size m, bound ϵ, #iterations K in inner optimization, classifier network $f_{\boldsymbol{W}}$, weighting network $f_{\boldsymbol{\Omega}}$, D^{train}, D^{meta}.
Output: Trained robust network $f_{\boldsymbol{W}}$.

Initialize networks $f_{\boldsymbol{W}}$ and $f_{\boldsymbol{\Omega}}$;
for $t = 1$ to T **do**
  Sample n and m samples from D^{train} and D^{meta};
  for $i = 1$ to n (in parallel) **do**
    $\boldsymbol{x}_i^{\text{adv}} = \boldsymbol{x}_i + 0.001\mathcal{N}(0, I)$ and $\boldsymbol{x}_i^{\text{at-adv}} = \boldsymbol{x}_i + 0.001\mathcal{N}(0, I)$, where $\mathcal{N}(0, I)$ is the Gaussian distribution;
    Calculate the perturbation bound ϵ_i for sample $\boldsymbol{x}_i$;
    for $k = 1$ to K **do**
      $\boldsymbol{x}_i^{\text{adv}} \leftarrow \Pi_{\mathbb{B}(\boldsymbol{x}_i, \epsilon_i)}(\eta_0 \text{sign}(\nabla_{\boldsymbol{x}_i^{\text{adv}}} \ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i), f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{adv}}))) + \boldsymbol{x}_i^{\text{adv}})$, where Π is the projection operator;
      $\boldsymbol{x}_i^{\text{at-adv}} \leftarrow \Pi_{\mathbb{B}(\boldsymbol{x}_i, \epsilon_i)}(-\eta_0 \text{sign}(\nabla_{\boldsymbol{x}_i^{\text{at-adv}}} \ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{at-adv}}), y_i)) + \boldsymbol{x}_i^{\text{at-adv}})$;
    end for
  end for
  Formulate $\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega})$ by Eq. (9);
  Update $\boldsymbol{\Omega}^{(t+1)}$ by Eq. (10) and update $\boldsymbol{W}^{(t+1)}$ by Eq. (11);
end for

of factors, including learning difficulty, class proportion, and noise degree. Therefore, six training characteristics of each training sample $\boldsymbol{x}$, namely, loss ($\boldsymbol{\zeta}_{\boldsymbol{x},1}$), margin ($\boldsymbol{\zeta}_{\boldsymbol{x},2}$), the norm of loss gradient for the logit vector ($\boldsymbol{\zeta}_{\boldsymbol{x},3}$), the information entropy of the softmax output ($\boldsymbol{\zeta}_{\boldsymbol{x},4}$), class proportion ($\boldsymbol{\zeta}_{\boldsymbol{x},5}$), and the average loss of each class ($\boldsymbol{\zeta}_{\boldsymbol{x},6}$), are extracted, as shown in the extraction module in Fig. 2. The calculation detail of each characteristic is shown in the online material.

Perturbation Bound ($\epsilon_{\boldsymbol{x}}$) Calculation

We employ two types of varied bound in our framework. Following Ref. Xu et al.(2021), the class-wise perturbation bound named ReMargin, which is suitable for imbalanced data, is utilized. A sample-wise bound is proposed to handle noise. It is inspired by the intuition that noisy samples have a large norm of loss gradient in general and these samples should exhibit the greatest degree of anti-adversarial training. Thus, the Grad-Based bound can be calculated as

$$\epsilon_{\boldsymbol{x}} = (\alpha_{\boldsymbol{x}} \overline{\boldsymbol{g}}_{\boldsymbol{x}_{\text{adv}}} + \beta_{\boldsymbol{x}} \overline{\boldsymbol{g}}_{\boldsymbol{x}_{\text{at-adv}}} + \varepsilon) \times \epsilon, \tag{8}$$

where $\overline{\boldsymbol{g}}_{\boldsymbol{x}_{\text{adv}}}$ and $\overline{\boldsymbol{g}}_{\boldsymbol{x}_{\text{at-adv}}}$ are the normalized $||\frac{\partial \ell(f_{\boldsymbol{W}}(\boldsymbol{x}), f_{\boldsymbol{W}}(\boldsymbol{x}_{\text{adv}}))}{\partial \boldsymbol{x}_{\text{adv}}}||_2$ and $||\frac{\partial \ell(f_{\boldsymbol{W}}(\boldsymbol{x}_{\text{at-adv}}), y)}{\partial \boldsymbol{x}_{\text{at-adv}}}||_2$, respectively. ϵ is a predefined perturbation bound, and ε is a hyperparameter that is set to 0.9 in our experiments. This bound is also effective on imbalanced data because samples in tail classes have large norms of loss gradient, and they should do the greatest degree of adversarial training.

Training with Meta-Learning

On the basis of the extracted characteristics and calculated bounds, an online learning strategy is adopted to alternatively update $\boldsymbol{W}$ and $\boldsymbol{\Omega}$ using a single optimization loop, as shown in Fig. 2. Assume that we have a small amount of unbiased meta data $D^{\text{meta}} = \{\boldsymbol{x}_i^{\text{meta}}, y_i^{\text{meta}}\}_{i=1}^{M}$, where $M \ll N$. Even if meta data are lacking, they can be compiled from the training data D^{train} (Zhang and Pfister 2021). The main steps are shown below. Here, we ignore the regularization terms introduced by the fairness constraints, while the online material provides the complete formulas.

$\boldsymbol{\Omega}$ is treated as the to-be-updated parameter, and the parameter of the updated classifier $\boldsymbol{W}$, which is a function of $\boldsymbol{\Omega}$, is formulated. Stochastic gradient descent (SGD) is utilized to optimize the training loss. Specifically, a minibatch of training samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$ is selected in each iteration, where n is the size of the mini-batch. The updating of $\boldsymbol{W}$ can be formulated as

$$\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega}) = \boldsymbol{W}^{(t)} - \eta_1 \frac{1}{n} \sum\nolimits_{i=1}^{n} \nabla_{\boldsymbol{W}} \{\alpha_i [\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i), y_i) + \lambda \ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i), f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{adv}}))] + \beta_i \ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{at-adv}}), y_i)\}|_{\boldsymbol{W}^{(t)}}, \tag{9}$$

where η_1 is the step size. The parameter of the weighting network $\boldsymbol{\Omega}$ after receiving feedback from the classifier network can be updated on a minibatch of meta data as follows:

$$\boldsymbol{\Omega}^{(t+1)} = \boldsymbol{\Omega}^{(t)} - \eta_2 \frac{1}{m} \sum\nolimits_{i=1}^{m} \nabla_{\boldsymbol{\Omega}} [\ell^{\text{meta}}(f_{\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega})}(\boldsymbol{x}_i), y_i) + \lambda \ell^{\text{meta}}(f_{\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega})}(\boldsymbol{x}_i), f_{\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega})}(\boldsymbol{x}_i^{\text{adv}})) + \ell^{\text{meta}}(f_{\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega})}(\boldsymbol{x}_i^{\text{at-adv}}), y_i)]|_{\boldsymbol{\Omega}^{(t)}}, \tag{10}$$

where m and η_2 are the minibatch size of meta data and the step size, respectively. The parameters of the classifier network are updated with the obtained weights by fixing the parameters of the weighting network as $\boldsymbol{\Omega}^{(t+1)}$:

$$\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta_1 \frac{1}{n} \sum\nolimits_{i=1}^{n} \nabla_{\boldsymbol{W}} \{\alpha_i [\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i), y_i) + \lambda \ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i), f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{adv}}))] + \beta_i \ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{at-adv}}), y_i)\}|_{\boldsymbol{W}^{(t)}}. \tag{11}$$

The steps of our CAAT method are shown in Algorithm 1.

Experiments

Experiments are conducted to verify our theoretical findings and the effectiveness of the proposed CAAT in improving the accuracy, robustness, and fairness of the robust models.

Experimental Settings

Benchmark adversarial learning datasets: CIFAR10 (Krizhevsky 2009) and SVHN (Netzer et al. 2011) are adopted in our experiments, including the noisy and imbalanced versions of the CIFAR data (Shu et al.

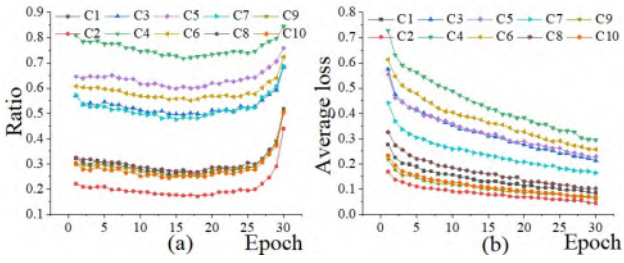

Figure 3: (a): Ratio of adversaries in each class during training on standard CIFAR10. (b): Average loss of each class during training on standard CIFAR10.

	Avg. Nat.	Worst Nat.	Avg. Bdy.	Worst Bdy.	Avg. Rob.	Worst Rob.
PGD Adv. Training	15.5	33.8	40.9	55.9	56.4	82.7
TRADES ($1/\lambda = 1$)	14.6	31.2	43.1	64.6	57.7	84.7
TRADES ($1/\lambda = 6$)	19.6	39.1	29.9	49.5	49.3	77.6
Baseline ReWeight	19.2	28.3	39.2	53.7	58.2	80.1
FRL (ReWeight)	16.0	**22.5**	41.6	54.2	57.6	73.3
FRL (ReMargin)	16.9	24.9	35.0	50.6	51.9	75.5
FRL (ReWeight+ReMargin)	17.0	26.8	35.7	44.5	52.7	69.5
CAAT (Grad-Based)	14.6	23.6	**14.4**	**23.3**	**28.6**	48.1
CAAT (ReMargin)	**13.9**	24.3	15.4	24.9	29.3	**44.4**

Table 1: Average and worstclass natural, boundary, and robust errors (%) for various algorithms on CIFAR10.

2019). For the two datasets, PreAct-ResNet18 (He et al. 2016) and Wide-ResNet28-10 (WRN28-10) (Zagoruyko and Komodakis 2016) are adopted as the backbone network. This section only represents the results of PreAct-ResNet18. Others are presented in the online material. The compared methods include three popular adversarial training algorithms, namely, PGD (Madry et al. 2018), TRADES (Zhang et al. 2019), and FRL (Xu et al. 2021). A debiasing method (Agarwal et al. 2018) is also compared which is to upweight the loss of the class with the largest robust error in the training data. The results of TRADES and FRL are calculated by using the codes in their official repositories.

The training and testing configurations used in Ref. Xu et al. (2021) are followed. The number of iterations in an adversarial attack is set to 10. Following Xu et al. (2021), 300 samples in each class with clean labels are selected as the meta dataset, which helps us tune the hyperparameters and train the weighting network. Adversarial training is trained on PGD attack setting $\epsilon = 8/255$ with cross-entropy loss. For our method and FRL (ReMargin), the predefined perturbation bound is also set to $8/255$. All the models are trained by using SGD with momentum 0.9 and weight decay 5×10^{-4}. The value of λ is selected in $\{2/3, 1, 1.5, 6\}$. During the evaluation phase, we report each model's average and worstclass natural, boundary, and robust error rates.

Experiments on Standard Dataset

Tables 1 shows the performance of our proposed CAAT and the compared methods on standard CIFAR10. Those on SVHN are shown in the online material. Considering that our training/testing configuration is the same as that in Ref. Xu et al.(2021), the results of the above competing methods reported in the FRL (Xu et al. 2021) paper are directly presented.

From the results, our methods with two types of bound reduce the average natural and robust errors under different degrees, indicating that CAAT obtains better accuracy and robustness of the model. Compared with other methods, CAAT decreases the average and worst robust error rates by 21% and 25% on CIFAR10. Baseline ReWeight can only decrease the worst intraclass natural error but cannot equalize boundary or robust errors. FRL (Xu et al. 2021) has only a limited ability to reduce the worst boundary and robust errors, resulting in limited fairness between classes. Our method more effectively decreases the worst intraclass errors. Thus, CAAT achieves better fairness among classes compared with other methods. Although FRL (ReWeight) obtains the lowest worst natural error, it has large average and worst robust errors, which is inferior to CAAT. Hard classes (classes with a large average loss) have a higher ratio of adversaries than easy ones, as shown in Fig. 3, which helps improve the performance of hard classes and effectively enhances the fairness among classes. The same conclusions can also be obtained on the SVHN dataset.

	Avg. Nat.	Avg. Bdy.	Avg. Rob.
PGD Adv. Training	15.6	37.1	52.8
TRADES ($1/\lambda = 1$)	15.6	31.0	46.5
TRADES ($1/\lambda = 6$)	16.4	21.0	37.4
FRL (ReWeight)	15.3	36.0	51.4
FRL (ReMargin)	15.2	36.0	51.1
FRL (ReWeight+ReMargin)	15.7	34.3	50.0
CAAT (Grad-Based)	**14.6**	**13.9**	**28.5**
CAAT (ReMargin)	14.7	14.7	29.4

Table 2: Average natural, boundary, and robust errors (%) for various algorithms on CIFAR10 with 20% pair-flip noise.

Experiments of Noisy Classification

Two settings of corrupted labels, including uniform and pair-flip noises, are adopted (Shu et al. 2019). The values of the noise ratio are set to 20% and 40%. CIFAR10 dataset,

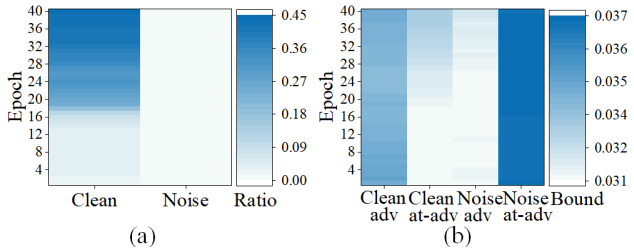

Figure 4: (a): Ratio of adversaries for noisy and clean samples on CIFAR10 with 20% uniform noise during training. (b): Average adversarial and anti-adversarial perturbation bounds for clean and noisy samples during training.

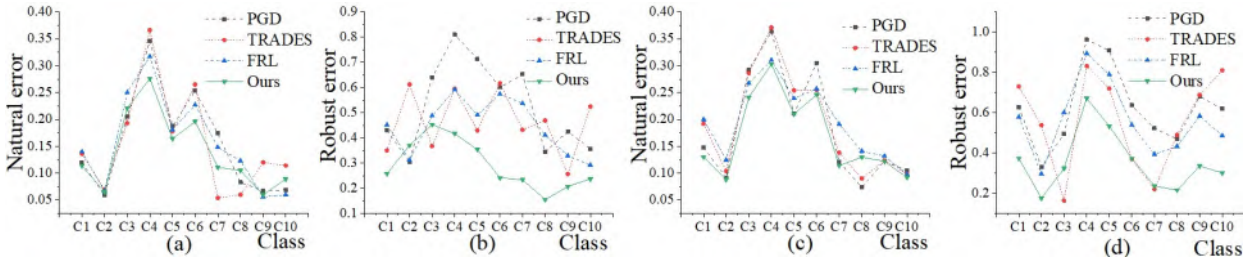

Figure 5: (a) and (b): Natural and robust errors for each class of different methods on CIFAR10 with imbalance factor 10. (c) and (d): Natural and robust errors for each class of different methods on CIFAR10 with imbalance factor 100.

	Avg. Nat.	Avg. Bdy.	Avg. Rob.
PGD Adv. Training	20.1	42.8	62.9
TRADES ($1/\lambda = 1$)	16.8	32.3	49.1
TRADES ($1/\lambda = 6$)	23.6	23.8	47.4
FRL (ReWeight)	16.9	38.1	55.0
FRL (ReMargin)	17.5	35.6	53.1
FRL (ReWeight+ReMargin)	17.2	35.1	52.3
CAAT (Grad-Based)	**15.8**	14.2	30.0
CAAT (ReMargin)	16.2	**13.7**	**29.9**

Table 3: Average and worstclass natural, boundary, and robust errors (%) on CIFAR10 with imbalance factor 10.

which is popularly used for the evaluation of noisy labels, is adopted. Here, we only show the average errors of CIFAR10 with 20% pair-flip noise. Others are presented in the online material. From the results in Table 2 and the online material, CAAT achieves the lowest average and worst natural and robust errors, indicating that it obtains the best generalization, robustness, and fairness compared with other methods.

As shown in Fig. 4 (a), most of the noisy samples are anti-adversarially perturbed during training, which is in accordance with our theoretical findings. From Fig. 4 (b), the average anti-adversarial perturbation bound for noisy samples is the largest, implying that noisy samples exhibit the largest degree of anti-adversarial training. Thus, the negative influence of noisy samples can be decreased. The ratio of adversaries for clean samples increases with the progress of training, demonstrating that clean samples are playing a more important role than noisy ones during training.

Experiments of Imbalanced Classification

The long-tailed version of CIFAR10 compiled by Cui et al. (2019) is utilized. The values of the imbalance factor are set to 10 and 100. Here, we only show the average results when the imbalance factor equals 10. Others are presented in the online material. Compared with other methods, CAAT achieves the minimum average and worst natural and robust errors, as shown in Table 3. As shown in Fig. 5, CAAT decreases the natural and robust errors for most classes and achieves the lowest performance gap among different classes. We also verify that the first head class has the lowest ratio of adversaries and tail classes have a high ratio of adversaries, which is consistent with our theoretical findings.

	Avg. Nat. (%)	Avg. Bdy. (%)	Avg. Rob. (%)
Setting I	16.0	41.6	57.6
Setting II	16.1	35.8	51.9
Setting III	14.9	**13.8**	**28.7**
Setting IV	**13.9**	15.4	29.3

Table 4: Ablation studies of CAAT on standard CIFAR10.

The details are presented in the online material.

Ablation Studies

Four variations of CAAT are considered, including adversarial training with the same perturbation direction and bound (Setting I), adversarial training with the same perturbation direction and different bounds (Setting II), adversarial training with different perturbation directions (adversaries and anti-adversaries) and the same bound (Setting III), and adversarial training with different perturbation directions and bounds (Setting IV). PreAct-ResNet18 is used. The results are shown in Table 4. Settings III and IV obtain better performance compared with Settings I and II. Thus, the combination strategy is more effective. Compared with Setting III, Setting IV further decreases the average natural error, indicating that the varied bound is more valid in some cases. The worst errors are shown in the online material.

Conclusions

This study theoretically investigates the role of adversarial training with different directions (adversarial and anti-adversarial) and bounds for the robust model. Three typical occasions are considered, including classes with different difficulties, imbalance learning, and noisy label learning. A series of theoretical findings are obtained, illuminating a new objective function that combines adversaries and anti-adversaries in training. Consequently, an adversarial training framework (CAAT) is proposed to solve the objective, in which meta learning is utilized to optimize the combined weights of the adversary and anti-adversary for each sample in accordance with its learning characteristics. Extensive experiments verify the rationality of our theoretical findings and the effectiveness of CAAT in achieving better accuracy, robustness, and fairness of the robust models compared with other adversarial training methods.

Acknowledgments

This study is partially supported by NSFC 62076178, TJF 22ZYYYJC00020, and 19ZXAZNGX00050.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach to Fair Classification. In *ICML*, 102–119.

Alfarra, M.; Pérez, J. C.; Thabet, A.; Bibi, A.; Torr, P. H. S.; and Ghanem, B. 2022. Combating Adversaries with Anti-Adversaries. In *AAAI*, 1–13.

Bai, T.; and Luo, J. 2021. Recent Advances in Adversarial Training for Adversarial Robustness. In *IJCAI*, 4312–4321.

Balaji, Y.; Goldstein, T.; and Hoffman, J. 2019. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. arXiv:1910.08051.

Cheng, M.; Lei, Q.; Chen, P.-Y.; Dhillon, I.; and Hsieh, C.-J. 2020. CAT: Customized Adversarial Training for Improved Robustness. arXiv:2002.06789.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *CVPR*, 9260–9270.

Ding, G. W.; Sharma, Y.; Lui, K. Y. C.; and Huang, R. 2020. MMA Training: Direct Input Space Margin Maximization through Adversarial Training. In *ICLR*, 1–34.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML*, 1856–1868.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Scandinavian Journal of Statistics*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*, 1–18.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, R.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. *Scandinavian Journal of Statistics*.

Nichol, A.; Achiam, J.; and Schulman, J. 2018. On First-Order Meta-Learning Algorithms. arXiv:1803.02999.

Pang, T.; Yang, X.; Dong, Y.; Su, H.; and Zhu, J. 2021. Bag of Tricks for Adversarial Training. In *ICLR*, 1–21.

Raghunathan, A.; Xie, S. M.; Yang, F.; Duchi, J. C.; and Liang, P. 2019. Adversarial Training Can Hurt Generalization. arXiv:1906.06032.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to Reweight Examples for Robust Deep Learning. In *ICML*, 6900–6909.

Rice, L.; Wong, E.; and Kolter, J. Z. 2020. Overfitting in adversarially robust deep learning. In *ICML*, 8093–8104.

Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; and Wierstra, D. 2016. Meta Learning With Memory-Augmented Neural Networks. In *ICML*, 2740–2751.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-Weight-Net: Learning an Explicit Mapping For Sample Weighting. In *NeurIPS*, 1917–1928.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks For Few Shot Learning. In *NeurIPS*, 4078–4088.

Song, C.; He, K.; Lin, J.; Wang, L.; and Hopcroft, J. E. 2020. Robust Local Features for Improving the Generalization of Adversarial Training. In *ICLR*, 1–12.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning To Compare: Relation Network For Few-Shot Learning. In *CVPR*, 1199–1208.

Uesato, J.; Alayrac, J.-B.; Huang, P.-S.; Stanforth, R.; Fawzi, A.; and Kohli, P. 2019. Are Labels Required for Improving Adversarial Robustness? In *NeurIPS*, 12214–12223.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2020. Improving Adversarial Robustness Requires Revisiting Misclassified Examples. In *ICLR*, 1–14.

Wong, E.; Rice, L.; and Kolter, J. Z. 2020. Fast is better than free: Revisiting adversarial training. In *ICLR*, 1–17.

Xie, C.; Tan, M.; Gong, B.; Yuille, A.; and Le, Q. V. 2020. Smooth Adversarial Training. arXiv:2002.11242.

Xu, H.; Liu, X.; Li, Y.; Jain, A. K.; and Tang, J. 2021. To be Robust or to be Fair: Towards Fairness in Adversarial Training. In *ICML*, 11492–11501.

Yang, S.; Guo, T.; Wang, Y.; and Xu, C. 2021. Adversarial Robustness through Disentangled Representations. In *AAAI*, 3145–3153.

Yang, Y.-Y.; Rashtchian, C.; Zhang, H.; Salakhutdinov, R.; and Chaudhuri, K. 2020. A Closer Look at Accuracy vs. Robustness. In *NeurIPS*, 8588–8601.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *BMVC*, 1–12.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *ICML*, 12907–12929.

Zhang, J.; Xu, X.; Han, B.; Niu, G.; Cui, L.; Sugiyama, M.; and Kankanhalli, M. 2020. Attacks Which Do Not Kill Training Make Adversarial Learning Stronger. In *ICML*, 1–15.

Zhang, Z.; and Pfister, T. 2021. Learning Fast Sample Reweighting Without Reward Data. In *ICCV*, 705–714.

Zhu, J.; Zhang, J.; Han, B.; Liu, T.; Niu, G.; Yang, H.; Kankanhalli, M.; and Sugiyama, M. 2021. Understanding the Interaction of Adversarial Training with Noisy Labels. arXiv:2102.03482.

Supplementary Material for Combining Adversaries with Anti-adversaries in Training

Xiaoling Zhou, Nan Yang, Ou Wu

Center for Applied Mathematics, Tianjin University

Theoretical Proof for Section 3 (Theoretical Investigation)

In this section, we present the proofs and discussions omitted in Section 3. Following Xu et al. (2021), we consider a binary classification task under a mixture Gaussian distribution. Xu et al. (2021) proved that adversarial training would exacerbate the performance gap between classes and hurt the harder class. Compared with previous theoretical investigations, our analysis involves more typical learning scenarios, including classes with different learning difficulties, imbalance learning, noisy label learning, and classes with skewed training distributions. In addition, previous theoretical findings only considered adversarial training with an identical bound for all training samples, whereas both adversaries and anti-adversaries with varied perturbation bounds are considered in our theoretical investigation. The natural and robust errors of the optimal robust linear classifier are calculated. We prove that combining adversaries and anti-adversaries in training with varied perturbation bounds can effectively tune the performance gap between classes and the tradeoff between the robustness and accuracy of the model.

Case I: Classes with Different Difficulties

In this subsection, we focus on the occasion when the two classes have different learning difficulties. The binary classification setting established by Xu et al. (2021) is followed. The data are from two classes $\mathcal{Y} = \{-1, +1\}$ and the data from each class follow a Gaussian distribution $\mathcal{D}$ which is centered on $\boldsymbol{\theta}$ and $-\boldsymbol{\theta}$, respectively. There is a K-factor difference between the two classes' variances: $\sigma_{+1} : \sigma_{-1} = K : 1$ and $K > 1$. In our following proof, the data follow

$$\begin{aligned} &y \overset{u.a.r}{\sim} \{-1, +1\}, \quad \boldsymbol{\theta} = (\overbrace{\eta, \ldots, \eta}^{\dim=d}), \\ &\boldsymbol{x} \sim \begin{cases} \mathcal{N}\left(\boldsymbol{\theta}, \sigma_{+1}^2 \boldsymbol{I}\right), & \text{if } y = +1, \\ \mathcal{N}\left(-\boldsymbol{\theta}, \sigma_{-1}^2 \boldsymbol{I}\right), & \text{if } y = -1. \end{cases} \end{aligned} \tag{A.1}$$

Intuitively, class "$+1$" is harder than class "-1" because it is less compacted in the data space, and the optimal linear classifier will give a larger error to class "$+1$" than class "-1", as proved by Xu et al. (2021). Here, we assume that $\sigma_{-1} = \sigma$. Then, we will prove that adversarial training with different perturbation bounds can tune the fairness between classes and the tradeoff between the accuracy and robustness of the model.

Proof of Theorem A.1 (Theorem 1 in Section 3) Xu et al. (2021) proved that, compared with natural training, adversarial training with an identical bound exacerbated the performance gap between classes and hurt the harder class. In this subsection, we prove that adversarial training with unequal perturbation bounds for the two classes can tune the performance gap between classes and improve the model performance on the harder class. Besides, a better tradeoff between the robustness and accuracy of the model can be achieved. Theorem A.1 calculates the natural and robust errors of the classifier that is adversarially trained with unequal bounds.

Theorem A.1. (Theorem 1 in Section 3) *For a data distribution $\mathcal{D}$ in Eq. (A.1), assume that the perturbation bounds of class "-1" and class "$+1$" are ϵ and $\rho \times \epsilon$ ($0 \leq \epsilon, \rho\epsilon < \eta$), respectively. The optimal robust linear classifier f_{rob} which minimizes the average robust error is*

$$\begin{aligned} f_{rob} = \arg\min_f \{&\Pr(\exists \|\boldsymbol{\delta}\| \leq \epsilon,\ s.t. f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid y = -1) \\ &+ \Pr(\exists \|\boldsymbol{\delta}\| \leq \rho \times \epsilon,\ s.t. f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid y = +1)\}. \end{aligned} \tag{A.2}$$

It has the natural errors for the two classes:

$$\begin{aligned} &\mathcal{R}_{nat}(f_{rob}, -1) \\ &= \Pr\left\{\mathcal{N}(0,1) \leq B - K \cdot \sqrt{B^2 + q(K)} - \frac{\sqrt{d}}{\sigma}\epsilon\right\}, \\ &\mathcal{R}_{nat}(f_{rob}, +1) \\ &= \Pr\left\{\mathcal{N}(0,1) \leq -K \cdot B + \sqrt{B^2 + q(K)} - \frac{\sqrt{d}\rho}{K\sigma}\epsilon\right\}, \end{aligned} \tag{A.3}$$

where $B = \frac{2}{K^2-1} \frac{\sqrt{d}(\eta - \frac{\epsilon(1+\rho)}{2})}{\sigma}$, and $q(K) = \frac{2\log K}{K^2-1}$.

Proof. Assume that the perturbation bounds of class "$+1$" and class "-1" are $\rho \times \epsilon$ ($0 \leq \rho < \frac{\eta}{\epsilon}$) and ϵ, respectively. To guarantee that the robust optimization gives a reasonable classification for the data in Eq. (A.1), the perturbation bound ϵ is limited in the region of $[0, \eta)$ and $\rho \times \epsilon$ is also in the region of $[0, \eta)$.

For a data distribution $\mathcal{D}$ in Eq. (A.1), the optimal linear classifier f_{nat} which minimizes the average natural classification error is

$$f_{\text{nat}} = \arg\min_f \Pr(f(\boldsymbol{x}) \neq y). \tag{A.4}$$

In accordance with *Lemma 1* in Ref. (Xu et al. 2021), the optimal linear classifier can be denoted as $f_{\text{nat}}(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{d} \boldsymbol{x}_i + b_{\text{nat}}\right)$ and the optimal b_{nat} is

$$b_{\text{nat}} = \frac{K^2+1}{K^2-1} \cdot d\eta - K\sqrt{\frac{4d^2\eta^2}{(K^2-1)^2} + q(K)d\sigma^2}, \quad \text{(A.5)}$$

where $q(K) = \frac{2\log K}{K^2-1}$ is a positive constant and only depends on K. In accordance with *Lemma 2* in Ref. (Xu et al. 2021), the optimal robust linear classifier can be denoted as $f_{\text{rob}}(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{d} \boldsymbol{x}_i + b_{\text{rob}}\right)$. Now, we calculate the optimal b_{rob} when the model is adversarially trained with varied bounds. The average robust error is

$$\begin{aligned}
\mathcal{R}_{\text{rob}}(f) &= \Pr(\exists\|\boldsymbol{\delta}\| \le \epsilon, \quad f(\boldsymbol{x}+\boldsymbol{\delta}) \neq -1 \mid y=-1) \\
&+ \Pr(\exists\|\boldsymbol{\delta}\| \le \rho\times\epsilon, \quad f(\boldsymbol{x}+\boldsymbol{\delta}) \neq +1 \mid y=+1) \\
&= \max_{\|\boldsymbol{\delta}\|\le\epsilon} \Pr(f(\boldsymbol{x}+\boldsymbol{\delta}) \neq -1 \mid y=-1) \\
&+ \max_{\|\boldsymbol{\delta}\|\le\rho\times\epsilon} \Pr(f(\boldsymbol{x}+\boldsymbol{\delta}) \neq +1 \mid y=+1) \\
&= \frac{1}{2}\Pr(f(\boldsymbol{x}+(\overbrace{\epsilon,\cdots,\epsilon}^{\dim=d})) \neq -1 \mid y=-1) \\
&+ \frac{1}{2}\Pr(f(\boldsymbol{x}-(\overbrace{\rho\times\epsilon,\cdots,\rho\times\epsilon}^{\dim=d})) \neq +1 \mid y=+1) \\
&= \Pr\left\{\sum_{i=1}^{d}(\boldsymbol{x}_i+\epsilon) + b_{\text{rob}} > 0 \mid y=-1\right\} \\
&+ \Pr\left\{\sum_{i=1}^{d}(\boldsymbol{x}_i-\rho\epsilon) + b_{\text{rob}} < 0 \mid y=+1\right\} \\
&= \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} + \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\} \\
&+ \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\rho\epsilon)}{K\sigma} - \frac{1}{K\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}.
\end{aligned} \quad \text{(A.6)}$$

The optimal b_{rob} to minimize $\mathcal{R}_{\text{rob}}(f)$ is achieved at the point that $\frac{\partial\mathcal{R}_{\text{rob}}(f)}{\partial b_{\text{rob}}} = 0$. Thus, the optimal b_{rob} is calculated as

$$\begin{aligned}
b_{\text{rob}} &= \frac{K^2+1}{K^2-1}\cdot d(\eta - \frac{\epsilon(1+\rho)}{2}) \\
&- K\sqrt{\frac{4d^2(\eta-\frac{\epsilon(1+\rho)}{2})^2}{(K^2-1)^2} + q(K)d\sigma^2} + \frac{(\rho-1)d\epsilon}{2}.
\end{aligned} \quad \text{(A.7)}$$

In accordance with the definition of robust error, by incorporating the optimal b_{rob} into the following formula

$$\begin{aligned}
&\Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} + \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\} \\
&+ \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{K\sigma} - \frac{1}{K\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\},
\end{aligned} \quad \text{(A.8)}$$

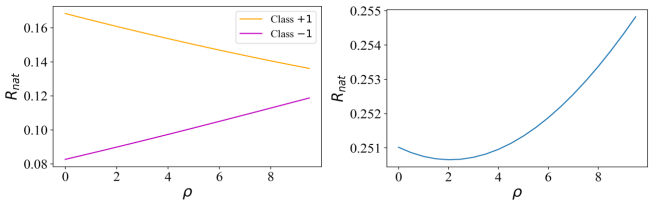

Figure A-1: Left: Natural errors for the two classes of the robust classifier trained with varied perturbation bounds. Right: Total natural error for the two classes of the robust classifier trained with varied perturbation bounds.

we can get the class-wise robust errors for the two classes

$$\begin{aligned}
&\mathcal{R}_{\text{rob}}(f_{\text{rob}}, -1) \\
&= \Pr\left\{\mathcal{N}(0,1) \le B - K\cdot\sqrt{B^2+q(K)}\right\}, \\
&\mathcal{R}_{\text{rob}}(f_{\text{rob}}, +1) \\
&= \Pr\left\{\mathcal{N}(0,1) \le -K\cdot B + \sqrt{B^2+q(K)} + \frac{\sqrt{d}(1-\rho)}{K\sigma}\epsilon\right\},
\end{aligned} \quad \text{(A.9)}$$

where $B = \frac{2}{K^2-1}\frac{\sqrt{d}(\eta-\frac{\epsilon(1+\rho)}{2})}{\sigma}$. In accordance with the definition of natural error, by incorporating the optimal b_{rob} into the following formula

$$\begin{aligned}
&\Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}\eta}{\sigma} + \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\} \\
&+ \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}\eta}{K\sigma} - \frac{1}{K\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\},
\end{aligned} \quad \text{(A.10)}$$

we obtain that the class-wise natural errors for the two classes are

$$\begin{aligned}
&\mathcal{R}_{\text{nat}}(f_{\text{rob}}, -1) \\
&= \Pr\left\{\mathcal{N}(0,1) \le B - K\cdot\sqrt{B^2+q(K)} - \frac{\sqrt{d}}{\sigma}\epsilon\right\}, \\
&\mathcal{R}_{\text{nat}}(f_{\text{rob}}, +1) \\
&= \Pr\left\{\mathcal{N}(0,1) \le -K\cdot B + \sqrt{B^2+q(K)} - \frac{\sqrt{d}\rho}{K\sigma}\epsilon\right\}.
\end{aligned} \quad \text{(A.11)}$$

□

The robust and natural errors of the two classes change with the increase in ρ. Thus, adversarial training with unequal perturbation bounds on the two classes can tune the performance gap between the two classes. Then, we show how the natural and robust errors of the two classes change with the increase in ρ.

Proof of Corollary A.1 (Corollary 1 in Section 3) In this subsection, we demonstrate how the natural and robust errors of the two classes change as ρ increases when the model is adversarially trained with unequal perturbation bounds.

Corollary A.1. (Corollary 1 in Section 3) *For a data distribution $\mathcal{D}$ in Eq. (A.1), assume that the perturbation bounds*

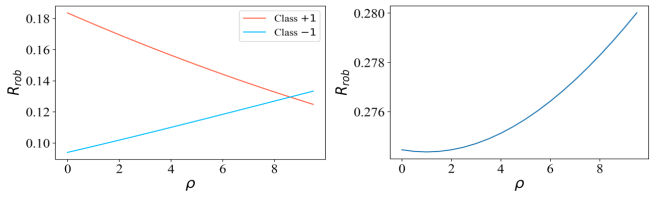

Figure A-2: Left: Robust errors for the two classes of the robust classifier trained with varied perturbation bounds. Right: Total robust error for the two classes of the robust classifier trained with varied perturbation bounds.

of class "−1" and "+1" are ϵ and $\rho \times \epsilon$ ($0 \leq \epsilon, \rho\epsilon < \eta$), respectively. When $K < e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}$, the adversarially trained model will increase and decrease the natural and robust errors of class "−1" and class "+1", with the increase in ρ, respectively.

Proof. From the definition of the data distribution $\mathcal{D}$, the only difference between the two classifiers is their interception terms: b_{nat} and b_{rob}. Following Xu et al. (2021), we know

$$b_{\text{nat}} = \frac{K^2+1}{K^2-1} \cdot d\eta - K\sqrt{\frac{4d^2\eta^2}{(K^2-1)^2} + q(K)d\sigma^2} := g(\eta). \tag{A.12}$$

When the classifier is adversarially trained with varied bounds, the robust classifier f_{rob} directly minimizes the natural error of the adversaries $(\boldsymbol{x} - (\overbrace{\rho \times \epsilon, \cdots, \rho \times \epsilon}^{\dim=d}))$ for samples $\boldsymbol{x}$ in class "+1", and the error of adversaries $(\boldsymbol{x} + (\overbrace{\epsilon, \cdots, \epsilon}^{\dim=d}))$ for samples $\boldsymbol{x}$ in class "−1". As a consequence, the robust model f_{rob} minimizes the error of samples in class "+1" with the new center $\boldsymbol{\theta}' = (\overbrace{\eta - \frac{\epsilon(1+\rho)}{2}, \ldots, \eta - \frac{\epsilon(1+\rho)}{2}}^{\dim=d})$ and minimizes the error of samples in class "−1" with the new center $-\boldsymbol{\theta}' = -(\overbrace{\eta - \frac{\epsilon(1+\rho)}{2}, \ldots, \eta - \frac{\epsilon(1+\rho)}{2}}^{\dim=d})$. Both centers are at a distance of $\frac{\epsilon(1+\rho)}{2}$ from the original center $\pm\boldsymbol{\theta}$. Thus, we can get the interception term of the robust model which is adversarially trained with varied perturbation bounds by replacing η in b_{nat} by $\eta - \frac{\epsilon(1+\rho)}{2}$. Besides, in accordance with the coordinate transformation, the distance that the coordinate system moves is $\frac{(\rho-1)d\epsilon}{2}$. Thus, b_{rob} can be expressed as

$$b_{\text{rob}} = g(\eta - \frac{\epsilon(1+\rho)}{2}) + \frac{(\rho-1)d\epsilon}{2}. \tag{A.13}$$

Then, we show that when $K < e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}$, b_{rob} is a monotone increasing function of ρ. The derivative of b_{rob} with respect to ρ is

$$\frac{\partial b_{\text{rob}}}{\partial \rho} = g'(\eta - \frac{\epsilon(1+\rho)}{2}) \cdot (-\frac{\epsilon}{2}) + \frac{d\epsilon}{2}. \tag{A.14}$$

Then, to ensure $\frac{\partial b_{\text{rob}}}{\partial \rho} > 0$, we just need that $g'(\eta - \frac{\epsilon(1+\rho)}{2}) < d$. $g'(\eta)$ is

$$\frac{dg(\eta)}{d\eta} = \frac{K^2+1}{K^2-1}d - K\frac{\frac{4}{(K^2-1)^2}d^2 \cdot 2\eta}{2\sqrt{\frac{4}{(K^2-1)^2}d^2\eta^2 + q(k)d\sigma^2}}. \tag{A.15}$$

Thus, the following inequality needs to be satisfied:

$$K\frac{\frac{4}{(K^2-1)^2}d^2 \cdot 2(\eta - \frac{\epsilon(1+\rho)}{2})}{2\sqrt{\frac{4}{(K^2-1)^2}d^2(\eta - \frac{\epsilon(1+\rho)}{2})^2 + q(k)d\sigma^2}} > \frac{2d}{K^2-1}. \tag{A.16}$$

Then, we obtain that the only condition for b_{rob} to be a monotonically increasing function with respect to ρ is

$$K < e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}. \tag{A.17}$$

Thus, with the increase in ρ, b_{rob} will increase. Therefore, the classification boundary gradually moves toward class "−1". The definitions of the robust and natural errors of classes "+1" and "−1" are shown below:

$$\mathcal{R}_{\text{rob}}(f_{\text{rob}}, -1) = \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} + \frac{1}{\sqrt{d}\sigma} \cdot b_{\text{rob}}\right\}, \tag{A.18}$$

$$\mathcal{R}_{\text{rob}}(f_{\text{rob}}, +1) = \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{K\sigma} - \frac{1}{K\sqrt{d}\sigma} \cdot b_{\text{rob}}\right\}, \tag{A.19}$$

$$\mathcal{R}_{\text{nat}}(f_{\text{rob}}, -1) = \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}\eta}{\sigma} + \frac{1}{\sqrt{d}\sigma} \cdot b_{\text{rob}}\right\}, \tag{A.20}$$

$$i\mathcal{R}_{\text{nat}}(f_{\text{rob}}, +1) = \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}\eta}{K\sigma} - \frac{1}{K\sqrt{d}\sigma} \cdot b_{\text{rob}}\right\}. \tag{A.21}$$

As we can see, the natural and robust errors of class "−1" are positively related to the value of b_{rob}, and the natural and robust errors of class "+1" are negatively related to the value of b_{rob}. Thus, with the increase in ρ, the natural and robust errors for class "−1" will be increased and those for class "+1" will be decreased. □

Figs. A-1 and A-2 show the variation of the natural and robust errors as ρ increases. For the parameters, σ and K are set to 1 and 1.5, respectively. The dimension d is set to 2. ϵ and η are set to 0.05 and 1, respectively. The natural and robust errors of class "+1" are decreased and those of class "−1" are increased. Thus, the fairness between the two classes can be tuned with different values of ρ and the model performance on the harder class can be improved. When the errors of the two classes are the same, the best fairness between classes is obtained. Moreover, adversarial training with varied bounds can contribute to larger scope of the classification boundary, which covers the boundary of the standard adversarial training. Thus, a better tradeoff between the accuracy and robustness of the model can be attained.

Proof of Theorem A.2 In this subsection, we assume that samples in class "-1" perform anti-adversarial perturbation. Then, we prove that combining adversaries with anti-adversaries in training with varied bounds can tune the performance gap between classes and the tradeoff between robustness and accuracy. Theorem A.2 calculates the natural and robust errors of the two classes when adversaries and anti-adversaries are combined in training.

Theorem A.2. *For a data distribution $\mathcal{D}$ in Eq. (A.1), assume that class "-1" is anti-adversarially perturbed with the perturbation bound ϵ, and class "$+1$" is adversarially perturbed with the bound $\rho \times \epsilon$ ($0 \leq \epsilon, \rho\epsilon < \eta$). The optimal robust linear classifier f_{rob} which minimizes the average robust error is*

$$f_{rob} = \arg\min_{f} \{\Pr(\exists \|\boldsymbol{\delta}\| \leq \epsilon,\ s.t. f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid y=-1) + \Pr(\exists \|\boldsymbol{\delta}\| \leq \rho\times\epsilon,\ s.t. f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid y=+1)\}. \tag{A.22}$$

It has the natural errors for the two classes:

$$\begin{aligned} &\mathcal{R}_{nat}\left(f_{rob}, -1\right) \\ &= \Pr\left\{\mathcal{N}(0,1) \leq B - K\cdot\sqrt{B^2+q(K)} + \frac{\sqrt{d}}{\sigma}\epsilon\right\}, \\ &\mathcal{R}_{nat}\left(f_{rob}, +1\right) \\ &= \Pr\left\{\mathcal{N}(0,1) \leq -K\cdot B + \sqrt{B^2+q(K)} - \frac{\sqrt{d}\rho}{K\sigma}\epsilon\right\}, \end{aligned} \tag{A.23}$$

where $B = \frac{2}{K^2-1}\frac{\sqrt{d}(\eta-\frac{\epsilon(\rho-1))}{2}}{\sigma}$, and $q(K) = \frac{2\log K}{K^2-1}$.

Proof. The perturbation bound ϵ is limited in the region of $[0,\eta)$ and $\rho\times\epsilon$ is also in the region of $[0,\eta)$. Similar to the proof of Theorem A.1, the optimal linear robust classifier is defined as $f_{\text{rob}}(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{d}\boldsymbol{x}_i + b_{\text{rob}}\right)$. Now we calculate the optimal b_{rob} when adversaries and anti-adversaries are combined in training. The average robust error is

$$\begin{aligned} \mathcal{R}_{\text{rob}}(f) &= \Pr(\exists\|\boldsymbol{\delta}\| \leq \epsilon, \quad f(\boldsymbol{x}+\boldsymbol{\delta}) \neq -1 \mid y=-1) \\ &\quad + \Pr(\exists\|\boldsymbol{\delta}\| \leq \rho\times\epsilon, \quad f(\boldsymbol{x}+\boldsymbol{\delta}) \neq +1 \mid y=+1) \\ &= \min_{\|\boldsymbol{\delta}\|\leq\epsilon} \Pr(f(\boldsymbol{x}+\boldsymbol{\delta}) \neq -1 \mid y=-1) \\ &\quad + \max_{\|\boldsymbol{\delta}\|\leq\rho\times\epsilon} \Pr(f(\boldsymbol{x}+\boldsymbol{\delta}) \neq +1 \mid y=+1) \\ &= \frac{1}{2}\Pr(f(\boldsymbol{x} - (\overbrace{\epsilon,\cdots,\epsilon}^{\dim=d})) \neq -1 \mid y=-1) \\ &\quad + \frac{1}{2}\Pr(f(\boldsymbol{x} - (\overbrace{\rho\times\epsilon,\cdots,\rho\times\epsilon}^{\dim=d})) \neq +1 \mid y=+1) \\ &= \Pr\left\{\sum_{i=1}^{d}(\boldsymbol{x}_i - \epsilon) + b_{\text{rob}} > 0 \mid y=-1\right\} \\ &\quad + \Pr\left\{\sum_{i=1}^{d}(\boldsymbol{x}_i - \rho\epsilon) + b_{\text{rob}} < 0 \mid y=+1\right\} \\ &= \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta+\epsilon)}{\sigma} + \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\} \\ &\quad + \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\rho\epsilon)}{K\sigma} - \frac{1}{K\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}. \end{aligned} \tag{A.24}$$

The optimal b_{rob} to minimize $\mathcal{R}_{\text{rob}}(f)$ is achieved at the point that $\frac{\partial\mathcal{R}_{\text{rob}}(f)}{\partial b_{\text{rob}}} = 0$. Thus, the optimal b_{rob} is calculated as

$$\begin{aligned} b_{\text{rob}} &= \frac{K^2+1}{K^2-1}\cdot d(\eta - \frac{\epsilon(\rho-1)}{2}) \\ &\quad - K\sqrt{\frac{4d^2(\eta-\frac{\epsilon(\rho-1)}{2})^2}{(K^2-1)^2} + q(K)d\sigma^2} + \frac{(\rho+1)d\epsilon}{2}. \end{aligned} \tag{A.25}$$

Similar to the proof of Theorem A.1, we get the class-wise robust errors for the two classes are

$$\begin{aligned} &\mathcal{R}_{\text{rob}}\left(f_{\text{rob}}, -1\right) \\ &= \Pr\left\{\mathcal{N}(0,1) \leq B - K\cdot\sqrt{B^2+q(K)} + \frac{2\sqrt{d}}{\sigma}\epsilon\right\}, \\ &\mathcal{R}_{\text{rob}}\left(f_{\text{rob}}, +1\right) \\ &= \Pr\left\{\mathcal{N}(0,1) \leq -K\cdot B + \sqrt{B^2+q(K)} + \frac{\sqrt{d}(1-\rho)}{K\sigma}\epsilon\right\}, \end{aligned} \tag{A.26}$$

where $B = \frac{2}{K^2-1}\frac{\sqrt{d}(\eta-\frac{\epsilon(\rho-1)}{2})}{\sigma})$. Similar to the manner in the proof of Theorem A.1, it is easy to obtain that the class-wise natural errors for the two classes are

$$\begin{aligned} &\mathcal{R}_{\text{nat}}\left(f_{\text{rob}}, -1\right) \\ &= \Pr\left\{\mathcal{N}(0,1) \leq B - K\cdot\sqrt{B^2+q(K)} + \frac{\sqrt{d}}{\sigma}\epsilon\right\}, \\ &\mathcal{R}_{\text{nat}}\left(f_{\text{rob}}, +1\right) \\ &= \Pr\left\{\mathcal{N}(0,1) \leq -K\cdot B + \sqrt{B^2+q(K)} - \frac{\sqrt{d}\rho}{K\sigma}\epsilon\right\}. \end{aligned} \tag{A.27}$$

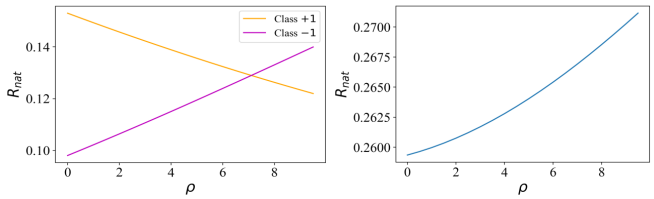

Figure A-3: Left: Natural errors for the two classes of the robust classifier trained with adversaries and anti-adversaries. Right: Total natural error for the two classes of the robust classifier trained with adversaries and anti-adversaries.

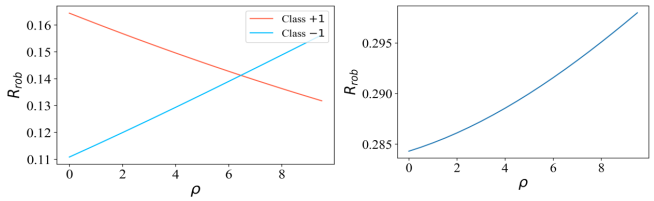

Figure A-4: Left: Robust errors for the two classes of the robust classifier trained with adversaries and anti-adversaries. Right: Total robust error for the two classes of the robust classifier trained with adversaries and anti-adversaries.

□

From Theorem A.2, we can see that both the robust and natural errors of the two classes change as ρ changes. Then, we show how the natural and robust errors of the two classes change as ρ increases.

Proof of Corollary A.2 (Corollary 2 in Section 3) In this subsection, we show how the natural and robust errors of the two classes change with the increase in ρ when the model is trained with adversaries and anti-adversaries.

Corollary A.2. (Corollary 2 in Section 3) *For a data distribution $\mathcal{D}$ in Eq. (A.1), assume that class "-1" is anti-adversarially perturbed with the perturbation bound ϵ, and class "$+1$" is adversarially perturbed with the bound $\rho \times \epsilon$ ($0 \leq \epsilon, \rho\epsilon < \eta$). When $K < e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}$, the adversarially trained model on two classes will increase and decrease the natural and robust errors of class "-1" and class "$+1$", with the increase in ρ, respectively.*

Proof. The value of b_{nat} is shown in the proof of Corollary A.1. For the classifier trained with adversaries and anti-adversaries, b_{rob} is calculated in Eq. (A.25), which can be expressed as

$$b_{\text{rob}} = g(\eta - \frac{\epsilon(\rho-1)}{2}) + \frac{(\rho+1)d\epsilon}{2}. \tag{A.28}$$

Similar to the manner of the proof of Corollary A.1, we can verify that b_{rob} is a monotone increasing function of ρ when $K < e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}$.

Thus, the classification boundary moves toward class "-1" as ρ increases. Similar to the proof of Corollary A.1, we can obtain that the natural and robust errors of class "$+1$" will be decreased and those of class "-1" will be increased as ρ increases. □

Figure A-5: Left: Logistic Regression classifiers (natural and robust) on simulated binary data in Eq. (A.1). Right: Logistic Regression classifiers (natural and robust with both adversaries and anti-adversaries) on simulated binary data in Eq. (A.1).

Figs. A-3 and A-4 show the variation of the natural and robust errors when adversaries and anti-adversaries are combined in training. The values of the parameters are the same as those in the last subsection. The natural and robust errors for class "-1" are increased, and those for class "$+1$" are decreased, with the increase in ρ. When the errors of the two classes are the same, the best fairness between the two classes is obtained. Thus, the performance gap between classes can be tuned with different ρ values.

Next, we show that when the same performance (same natural error) is achieved, the combination of adversaries and anti-adversaries has smaller perturbation bounds compared with only the adversaries. In accordance with our calculation, if the perturbation bound of combining adversaries and anti-adversaries is ϵ, then the perturbation bound of using only adversaries needs to be $10.419 \times \epsilon$ to achieve the same performance. Thus, combining adversaries and anti-adversaries in training is more efficient than using only adversaries, indicating that anti-adversaries are valuable. Fig. A-5 shows the data distribution and the classification boundaries which are naturally trained and adversarially trained with different directions and perturbation bounds. Adversarial training with the same bound exacerbates the performance gap, while both adversarial training and combining adversaries and anti-adversaries in training with varied bounds can decrease the performance gap. Besides, under the same bound (i.e., the same ρ), the combination strategy has a more pronounced effect.

Then, we calculate the scope of the classification boundary of the adversarial training with varied bounds and the combination of adversaries and anti-adversaries with varied bounds. Assume that the perturbation bounds for class "-1" and class "$+1$" are $\rho_+ \times \epsilon$ and $\rho_- \times \epsilon$ ($-\eta/\epsilon < \rho_+, \rho_- < \eta/\epsilon$), respectively. $\rho_+ < 0$ means that samples in class "$+1$" are anti-adversarial perturbed. $\rho_- < 0$ means that samples in class "-1" are anti-adversarial perturbed. By calculating, when $\rho_+ = \eta/\epsilon$ and $\rho_- = 0$, then b_{rob} is

$$b_{\text{rob}} = g(\frac{\eta}{2}) + \frac{d\eta}{2}. \tag{A.29}$$

Figure A-6: Scope of the classification boundary under different manners, including natural training (red line), standard adversarial training (green line), TRADES (scope between the red and green lines), adversarial training with different perturbation bounds, and adversarial training with different perturbation directions and bounds. The values of parameters are $K=2$, $\eta=2$, $\epsilon=0.2$, and $\sigma=1$. The bounds for class "$+1$" and "-1" are denoted as $\rho_+ \times \epsilon$ and $\rho_- \times \epsilon$ ($-\eta/\epsilon < \rho_+, \rho_- < \eta/\epsilon$), respectively. $\rho_+(\rho_-) < 0$ means that class "$+1(-1)$" is anti-adversarially perturbed.

When $\rho_+ = 0$ and $\rho_- = \eta/\epsilon$, then b_{rob} is

$$b_{\text{rob}} = g(\frac{\eta}{2}) - \frac{d\eta}{2}. \tag{A.30}$$

Then, we consider the occasion where adversaries and anti-adversaries are combined in training. When $\rho_+ = \eta/\epsilon$ and $\rho_- = -\eta/\epsilon$, then b_{rob} is

$$b_{\text{rob}} = g(\eta) + d\eta. \tag{A.31}$$

When $\rho_+ = -\eta/\epsilon$ and $\rho_- = \eta/\epsilon$, then b_{rob} is

$$b_{\text{rob}} = g(\eta) - d\eta. \tag{A.32}$$

The scope of the classification boundary under different manners is shown in Fig. A-6. The scope of the classification boundary of adversarial training with varied bounds is larger than that of TRADES (Zhang et al. 2019). Combining adversaries and anti-adversaries with a varied bound accounts for the biggest scope of the classification boundary. Thus, the combination strategy can achieve a better tradeoff between the robustness and the accuracy of the model.

Case II: Classes with Imbalanced Proportions

In this subsection, we focus on the occasion when the category distribution of the two classes is imbalanced. The data are from two classes $\mathcal{Y} = \{-1, +1\}$ and the data from each class follow a Gaussian distribution $\mathcal{D}_V$ which is centered on $\boldsymbol{\theta}$ and $-\boldsymbol{\theta}$ respectively. The two variances of the two classes are assumed to be identical, i.e., $\sigma_{+1} = \sigma_{-1} = \sigma$. Nevertheless, a V-factor difference is found between the two classes' prior probabilities: $p_+ : p_- = 1 : V$ and $V > 1$. In our following proof, the data follow

$$\begin{aligned} &\Pr(y=+1) = p_+, \quad \Pr(y=-1) = p_-, \\ &\boldsymbol{\theta} = (\overbrace{\eta, \ldots, \eta}^{\dim=d}), \\ &\boldsymbol{x} \sim \begin{cases} \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2 I\right), & \text{if } y=+1, \\ \mathcal{N}\left(-\boldsymbol{\theta}, \sigma^2 I\right), & \text{if } y=-1. \end{cases} \end{aligned} \tag{A.33}$$

Intuitively, class "$+1$" is harder than the dominant class "-1", because it has a small number of samples.

Proof of Theorem A.3 In this subsection, we prove that the error of the dominant class is smaller than that of the other class under natural training. The theorem is shown below.

Theorem A.3. *For a data distribution $\mathcal{D}_V$ in Eq. (A.33) with the imbalance factor V, the optimal linear classifier f_{nat} which minimizes the average natural classification error is*

$$f_{nat} = \arg\min_f \Pr(f(\boldsymbol{x}) \neq y). \tag{A.34}$$

It has the natural errors for the two classes:

$$\begin{aligned} R_{nat}\left(f_{nat}, -1\right) &= \Pr\left\{\mathcal{N}(0,1) \leq -A - \frac{\log V}{2A}\right\}, \\ R_{nat}\left(f_{nat}, +1\right) &= \Pr\left\{\mathcal{N}(0,1) \leq -A + \frac{\log V}{2A}\right\}, \end{aligned} \tag{A.35}$$

where $A = \frac{\sqrt{d}\eta}{\sigma}$. As a result, class "$+1$" has a larger natural error:

$$\mathcal{R}_{nat}\left(f_{nat}, -1\right) < \mathcal{R}_{nat}\left(f_{nat}, +1\right). \tag{A.36}$$

Proof. We define the optimal linear classifier as $f_{\text{nat}}(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^d \boldsymbol{x}_i + b_{\text{nat}}\right)$. Now, we calculate the optimal b_{nat} under natural training. The average natural error is

$$\begin{aligned} R_{\text{nat}}(f) &= \Pr\{f(\boldsymbol{x}) \neq y\} \\ &\propto V \Pr\{f(\boldsymbol{x}) = +1 \mid y = -1\} \\ &\quad + \Pr\{f(\boldsymbol{x}) = -1 \mid y = +1\} \\ &= V \Pr\left\{\sum_{i=1}^d \boldsymbol{x}_i + b_{\text{nat}} > 0 \mid y = -1\right\} \\ &\quad + \Pr\left\{\sum_{i=1}^d \boldsymbol{x}_i + b_{\text{nat}} < 0 \mid y = +1\right\} \\ &= V \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}\eta}{\sigma} + \frac{1}{\sqrt{d}\sigma} \cdot b_{\text{nat}}\right\} \\ &\quad + \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}\eta}{\sigma} - \frac{1}{\sqrt{d}\sigma} \cdot b_{\text{nat}}\right\}. \end{aligned} \tag{A.37}$$

The optimal b_{nat} to minimize $R_{\text{nat}}(f)$ is achieved at the point that $\frac{\partial \mathcal{R}_{\text{nat}}(f)}{\partial b_{\text{nat}}} = 0$. Thus, the optimal b_{nat} is calculated as

$$b_{\text{nat}} = -\frac{\sigma^2 \log V}{2\eta}. \tag{A.38}$$

By incorporating the optimal b_{nat} into Eq. (A.37), we can get the class-wise natural errors for the two classes:

$$\begin{aligned}\mathcal{R}_{\text{nat}}\left(f_{\text{nat}},-1\right)&=\Pr\left\{\mathcal{N}(0,1)\leq -A-\frac{\log V}{2A}\right\},\\ \mathcal{R}_{\text{nat}}\left(f_{\text{nat}},+1\right)&=\Pr\left\{\mathcal{N}(0,1)\leq -A+\frac{\log V}{2A}\right\}.\end{aligned}\tag{A.39}$$

Since $V > 1$, we have the direct conclusion that $\mathcal{R}_{\text{nat}}\left(f_{\text{nat}},-1\right) < \mathcal{R}_{\text{nat}}\left(f_{\text{nat}},+1\right)$. □

Theorem A.3 demonstrates that class "+1" which has a small prior probability is harder to be classified than the dominant class "−1" under natural training. The class-wise difference is due to the prior probability ratio V. If the two classes' prior probabilities are equal, i.e., $V = 1$, the natural errors for the two classes are the same. Next, we will show that adversarial training with the same perturbation direction and bound will exacerbate the performance gap, while adversarial training with different perturbation directions and bounds can tune the performance gap between classes and the tradeoff between the robustness and accuracy of the model.

Proof of Theorem A.4 (Theorem 2 in Section 3) In this subsection, we calculate the natural and robust errors of the two classes under adversarial training.

Theorem A.4. (Theorem 2 in Section 3) *For a data distribution $\mathcal{D}_V$ in Eq. (A.33) with the imbalance factor V, assume that the perturbation bounds of class "−1" and class "+1" are ϵ and $\rho\times\epsilon$ ($0 \leq \epsilon, \rho\epsilon < \eta$), respectively. The optimal robust linear classifier f_{rob} which minimizes the average robust error is*

$$\begin{aligned}f_{rob}=\arg\min_{f}\{&\Pr(\exists\|\boldsymbol{\delta}\|\leq\epsilon,\ s.t. f(\boldsymbol{x}+\boldsymbol{\delta})\neq y\mid y=-1)\\&+\Pr(\exists\|\boldsymbol{\delta}\|\leq\rho\times\epsilon,\ s.t. f(\boldsymbol{x}+\boldsymbol{\delta})\neq y\mid y=+1)\}.\end{aligned}\tag{A.40}$$

It has the natural errors for the two classes:

$$\begin{aligned}&\mathcal{R}_{nat}\left(f_{rob},-1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A-\frac{\log V}{2A}-\frac{\sqrt{d}}{\sigma}\epsilon\right\},\\&\mathcal{R}_{nat}\left(f_{rob},+1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A+\frac{\log V}{2A}-\frac{\sqrt{d}\rho}{\sigma}\epsilon\right\},\end{aligned}\tag{A.41}$$

where $A=\frac{\sqrt{d}(\eta-\frac{\epsilon(1+\rho)}{2})}{\sigma}$.

Proof. As stated before, the optimal robust linear classifier can be defined as $f_{\text{rob}}(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{d}\boldsymbol{x}_i + b_{\text{rob}}\right)$. Now, we calculate the optimal b_{rob} which can minimize the average robust error:

$$\begin{aligned}\mathcal{R}_{\text{rob}}(f)&=\Pr(\exists\|\boldsymbol{\delta}\|\leq\epsilon,\quad f(\boldsymbol{x}+\boldsymbol{\delta})\neq -1\mid y=-1)\\&+\Pr(\exists\|\boldsymbol{\delta}\|\leq\rho\times\epsilon,\quad f(\boldsymbol{x}+\boldsymbol{\delta})\neq +1\mid y=+1)\\&=\max_{\|\boldsymbol{\delta}_1\|\leq\epsilon}\Pr(f(\boldsymbol{x}+\boldsymbol{\delta})\neq -1\mid y=-1)\\&+\max_{\|\boldsymbol{\delta}_2\|\leq\rho\times\epsilon}\Pr(f(\boldsymbol{x}+\boldsymbol{\delta})\neq +1\mid y=+1)\\&=V\Pr(f(\boldsymbol{x}+(\overbrace{\epsilon,\cdots,\epsilon}^{\dim=d}))\neq -1\mid y=-1)\\&+\Pr(f(\boldsymbol{x}-(\overbrace{\rho\times\epsilon,\cdots,\rho\times\epsilon}^{\dim=d}))\neq +1\mid y=+1)\\&=V\Pr\left\{\sum_{i=1}^{d}(\boldsymbol{x}_i+\epsilon)+b_{\text{rob}}>0\mid y=-1\right\}\\&+\Pr\left\{\sum_{i=1}^{d}(\boldsymbol{x}_i-\rho\epsilon)+b_{\text{rob}}<0\mid y=+1\right\}\\&=V\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}(\eta-\epsilon)}{\sigma}+\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}\\&+\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}(\eta-\rho\epsilon)}{\sigma}-\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}.\end{aligned}\tag{A.42}$$

The optimal b_{rob} to minimize $\mathcal{R}_{\text{rob}}(f)$ is achieved at the point that $\frac{\partial\mathcal{R}_{\text{rob}}(f)}{\partial b_{\text{rob}}}=0$. Thus, the optimal b_{rob} is calculated as

$$b_{\text{rob}}=-\frac{\sigma^2\log V}{2(\eta-\frac{\epsilon(1+\rho)}{2})}+\frac{(\rho-1)d\epsilon}{2}.\tag{A.43}$$

In accordance with the definition of robust error, by incorporating the optimal b_{rob} into the following formula

$$\begin{aligned}&V\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}(\eta-\epsilon)}{\sigma}+\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}\\&+\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}(\eta-\epsilon)}{\sigma}-\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\},\end{aligned}\tag{A.44}$$

we can get the class-wise robust errors for the two classes

$$\begin{aligned}&\mathcal{R}_{\text{rob}}\left(f_{\text{rob}},-1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A-\frac{\log V}{2A}\right\},\\&\mathcal{R}_{\text{rob}}\left(f_{\text{rob}},+1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A+\frac{\log V}{2A}+\frac{\sqrt{d}(1-\rho)}{\sigma}\epsilon\right\},\end{aligned}\tag{A.45}$$

where $A=\frac{\sqrt{d}(\eta-\frac{\epsilon(1+\rho)}{2})}{\sigma}$. In accordance with the definition of natural error, by incorporating the optimal b_{rob} into the following formula

$$\begin{aligned}&V\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}\eta}{\sigma}+\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}\\&+\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}\eta}{\sigma}-\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\},\end{aligned}\tag{A.46}$$

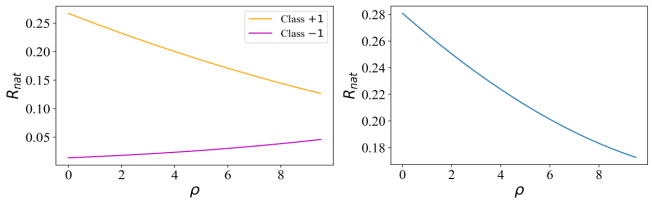

Figure A-7: Left: Natural errors for the two classes of the robust classifier trained with varied perturbation bounds. Right: Total natural error for the two classes of the robust classifier trained with varied perturbation bounds.

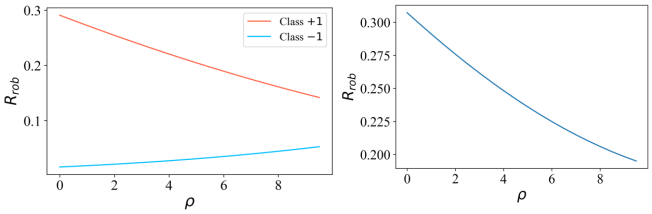

Figure A-8: Left: Robust errors for the two classes of the robust classifier trained with varied perturbation bounds. Right: Total robust error for the two classes of the robust classifier trained with varied perturbation bounds.

the class-wise natural errors for the two classes are

$$\begin{aligned}&\mathcal{R}_{\text{nat}}\left(f_{\text{rob}},-1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A-\frac{\log V}{2A}-\frac{\sqrt{d}}{\sigma}\epsilon\right\},\\&\mathcal{R}_{\text{nat}}\left(f_{\text{rob}},+1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A+\frac{\log V}{2A}-\frac{\sqrt{d}\rho}{\sigma}\epsilon\right\}.\end{aligned}\tag{A.47}$$

□

Then, we show that standard adversarial training which has an equal bound ($\rho = 1$) for all training samples exacerbates the performance gap between classes, and adversarial training with unequal perturbation bounds can tune the performance gap and the tradeoff. Corollary 3 shows how the natural and robust errors of the two classes change as ρ increases.

Proof of Corollary A.3 (Corollary 3 in Section 3) In this subsection, we show how the natural and robust errors of the two classes change as ρ increases when the model is adversarially trained with unequal perturbation bounds.

Corollary A.3. (Corollary 3 in Section 3) *For a data distribution $\mathcal{D}_V$ in Eq. (A.33), assume that the perturbation bounds for class "-1" and class "$+1$" are ϵ and $\rho \times \epsilon$ ($0 \leq \epsilon, \rho\epsilon < \eta$), respectively. When $V < e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}$, the adversarially trained model will increase and decrease the natural and robust errors of class "-1" and class "$+1$", with the increase in ρ, respectively.*

Proof. As stated before, the only difference between f_{nat} and f_{rob} is their interception terms b_{nat} and b_{rob}. We already have

$$b_{\text{nat}} = -\frac{\sigma^2 \log V}{2\eta} := g_{\text{m}}(\eta). \tag{A.48}$$

The robust classifier f_{rob} directly minimizes the natural error of adversaries $(\boldsymbol{x} - (\overbrace{\rho\times\epsilon,\cdots,\rho\times\epsilon}^{\dim=d}))$ for samples $\boldsymbol{x}$ in class "$+1$", and it minimizes the error of adversaries $(\boldsymbol{x} + (\overbrace{\epsilon,\cdots,\epsilon}^{\dim=d}))$ for samples $\boldsymbol{x}$ in class "-1". As a consequence, the robust model f_{rob} minimizes the errors of samples in two classes whose new centers are $\pm\boldsymbol{\theta}' = \pm(\overbrace{\eta-\frac{\epsilon(1+\rho)}{2},\cdots,\eta-\frac{\epsilon(1+\rho)}{2}}^{\dim=d})$ which are both at a distance $\frac{\epsilon(1+\rho)}{2}$ from the original center $\pm\boldsymbol{\theta}$. Thus, we can get the interception term of the robust model by replacing η in b_{nat} by $(\eta - \frac{\epsilon(1+\rho)}{2})$. Besides, in accordance with the coordinate translation, the coordinate system is shifted by a distance of $\frac{(\rho-1)d\epsilon}{2}$, so b_{rob} can be expressed as

$$b_{\text{rob}} = g_{\text{m}}(\eta - \frac{\epsilon(1+\rho)}{2}) + \frac{(\rho-1)d\epsilon}{2}. \tag{A.49}$$

Then, we show that when $V < e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}$, b_{rob} is a monotone increasing function of ρ. The derivative of b_{rob} with respect to ρ is

$$\frac{\partial b_{\text{rob}}}{\partial \rho} = g'_{\text{m}}(\eta - \frac{\epsilon(1+\rho)}{2}) \cdot (-\frac{\epsilon}{2}) + \frac{d\epsilon}{2}. \tag{A.50}$$

Then, to ensure $\frac{\partial b_{\text{rob}}}{\partial \rho} > 0$, we just need that $g'_{\text{m}}(\eta - \frac{\epsilon(1+\rho)}{2}) < d$. $g'_{\text{m}}(\eta)$ is

$$\frac{dg_{\text{m}}(\eta)}{d\eta} = \frac{\sigma^2 \log V}{2\eta^2}. \tag{A.51}$$

Thus, for our robust classifier, the following inequality needs to be satisfied:

$$\log V < \frac{2d(\eta - \frac{\epsilon(1+\rho)}{2})^2}{\sigma^2}. \tag{A.52}$$

Then, we obtain that the only condition for b_{rob} to be a monotonically increasing function with respect to ρ is

$$V < e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}. \tag{A.53}$$

Thus, the classification boundary moves toward class "-1", with the increase in ρ. The following formulas show the natural and robust errors for class "-1" and class "$+1$":

$$\mathcal{R}_{\text{rob}}\left(f_{\text{rob}},-1\right) = V\Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} + \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}, \tag{A.54}$$

$$\mathcal{R}_{\text{rob}}\left(f_{\text{rob}},+1\right) = \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} - \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}, \tag{A.55}$$

$$\mathcal{R}_{\text{nat}}\left(f_{\text{rob}},-1\right)=V\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}\eta}{\sigma}+\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}, \tag{A.56}$$

$$\mathcal{R}_{\text{nat}}\left(f_{\text{rob}},+1\right)=\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}\eta}{\sigma}-\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}. \tag{A.57}$$

As we can see, the natural and robust errors of class "-1" are positively related to b_{rob} and those of class "$+1$" are negatively related to b_{rob}. Thus, the natural and robust errors for the easy class "-1" will be increased and those of class "$+1$" will be decreased.

We notice that when $\rho=1$, there is $b_{\text{rob}}=g_{\text{m}}(\eta-\epsilon)$. Then, we show that g_{m} is a monotone increasing function from 0 to η. The deviate of g_{m} is

$$\frac{dg_{\text{m}}(\eta)}{d\eta}=\frac{\sigma^2\log V}{2\eta^2}>0. \tag{A.58}$$

Thus, we have the relation that $0<b_{\text{rob}}<b_{\text{nat}}$, which indicates that the classification boundary is more close to class "$+1$". Therefore, standard adversarial training will harm the performance of the model on class "$+1$". We can get the conclusion that an adversarially trained model with the identical bound on two classes will exacerbate the performance gap, including natural and robust errors. □

Figs. A-7 and A-8 show the variation of the natural and robust errors with the increase in ρ, when the classifier is adversarially trained with unequal bounds. For the parameters, the imbalance factor V is set to 3. d and ϵ are set to 2 and 0.05, respectively. The value of η is 1. As for the variances of the two classes, we set $\sigma=1$. The adversarially trained model will increase the natural and robust errors of the dominant class "-1" and decrease the two errors of class "$+1$", with the increase in ρ. When the natural and robust errors of the two classes are the same, the best fairness is obtained. Moreover, the scope of the classification boundary of adversarial training with varied bounds covers the boundary which is adversarially trained with an equal bound. Thus, a better tradeoff between robustness and accuracy can be attained by adversarial training with varied bounds.

Proof of Theorem A.5 In this subsection, we prove that combining adversaries and anti-adversaries in training can more effectively tune the performance gap between the two classes and the tradeoff between the robustness and the accuracy of the model. First, we calculate the natural and robust errors of the two classes when the model is trained with adversaries and anti-adversaries.

Theorem A.5. *For a data distribution $\mathcal{D}_V$ in Eq. (A.33), assume that class "-1" is anti-adversarially perturbed with the perturbation bound ϵ, and class "$+1$" is adversarially perturbed with the bound $\rho\times\epsilon$ ($0\le\epsilon,\rho\epsilon<\eta$). The optimal robust linear classifier f_{rob} which minimizes the average robust error is*

$$\begin{aligned}f_{rob}=\arg\min_{f}\{&\Pr(\exists\|\boldsymbol{\delta}\|\le\epsilon,\ s.t.f(\boldsymbol{x}+\boldsymbol{\delta})\neq y\mid y=-1)\\&+\Pr(\exists\|\boldsymbol{\delta}\|\le\rho\times\epsilon,\ s.t.f(\boldsymbol{x}+\boldsymbol{\delta})\neq y\mid y=+1)\}.\end{aligned} \tag{A.59}$$

It has the natural errors for the two classes:

$$\begin{aligned}&\mathcal{R}_{nat}\left(f_{rob},-1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\le-A-\frac{\log V}{2A}+\frac{\sqrt{d}}{\sigma}\epsilon\right\},\\&\mathcal{R}_{nat}\left(f_{rob},+1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\le-A+\frac{\log V}{2A}-\frac{\sqrt{d}\rho}{\sigma}\epsilon\right\},\end{aligned} \tag{A.60}$$

where $A=\frac{\sqrt{d}(\eta-\frac{\epsilon(\rho-1)}{2})}{\sigma}$.

Proof. We consider that the two classes have different perturbation directions. Samples in class "$+1$" are adversarially perturbed with the perturbation bound $\rho\times\epsilon$, and those of class "-1" are anti-adversarially perturbed with the bound ϵ. As before, the perturbation bound ϵ is limited in the region of $[0,\eta)$ and $\rho\times\epsilon$ is also in the region of $[0,\eta)$.

The optimal linear robust classifier is $f_{\text{rob}}(\boldsymbol{x})=\text{sign}\left(\sum_{i=1}^{d}\boldsymbol{x}_i+b_{\text{rob}}\right)$. Now, we calculate the optimal b_{rob} which is trained with adversaries and anti-adversaries on imbalanced data. The average robust error is

$$\begin{aligned}\mathcal{R}_{\text{rob}}(f)&=\Pr(\exists\|\boldsymbol{\delta}\|\le\epsilon,\quad f(\boldsymbol{x}+\boldsymbol{\delta})\neq-1\mid y=-1)\\&+\Pr(\exists\|\boldsymbol{\delta}\|\le\rho\times\epsilon,\quad f(\boldsymbol{x}+\boldsymbol{\delta})\neq+1\mid y=+1)\\&=\min_{\|\boldsymbol{\delta}\|\le\epsilon}\Pr(f(\boldsymbol{x}+\boldsymbol{\delta})\neq-1\mid y=-1)\\&+\max_{\|\boldsymbol{\delta}\|\le\rho\times\epsilon}\Pr(f(\boldsymbol{x}+\boldsymbol{\delta})\neq+1\mid y=+1)\\&=V\Pr(f(\boldsymbol{x}-(\overbrace{\epsilon,\cdots,\epsilon}^{\dim=d}))\neq-1\mid y=-1)\\&+\Pr(f(\boldsymbol{x}-(\overbrace{\rho\times\epsilon,\cdots,\rho\times\epsilon}^{\dim=d}))\neq+1\mid y=+1)\\&=V\Pr\left\{\sum_{i=1}^{d}(\boldsymbol{x}_i-\epsilon)+b_{\text{rob}}>0\mid y=-1\right\}\\&+\Pr\left\{\sum_{i=1}^{d}(\boldsymbol{x}_i-\rho\epsilon)+b_{\text{rob}}<0\mid y=+1\right\}\\&=V\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}(\eta+\epsilon)}{\sigma}+\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}\\&+\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}(\eta-\rho\times\epsilon)}{\sigma}-\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}.\end{aligned} \tag{A.61}$$

The optimal b_{rob} to minimize $\mathcal{R}_{\text{rob}}(f)$ is achieved at the point that $\frac{\partial\mathcal{R}_{\text{rob}}(f)}{\partial b_{\text{rob}}}=0$. Thus, the optimal b_{rob} is solved as

$$b_{\text{rob}}=-\frac{\sigma^2\log V}{2(\eta-\frac{\epsilon(\rho-1)}{2})}+\frac{(\rho+1)d\epsilon}{2}. \tag{A.62}$$

Similar to the proof of Theorem A.4, we can get the class-

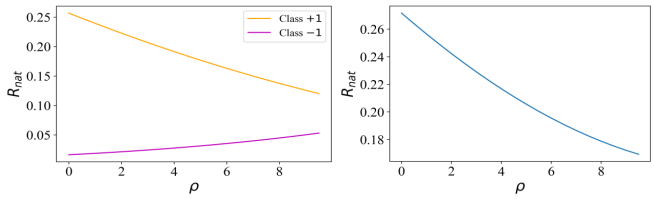

Figure A-9: Left: Natural errors for the two imbalanced classes of the robust classifier trained with adversaries and anti-adversaries. Right: Total natural error for the two imbalanced classes of the robust classifier trained with adversaries and anti-adversaries.

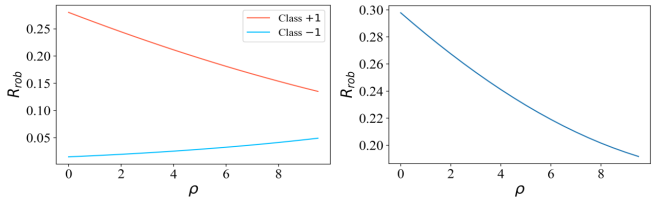

Figure A-10: Left: Robust errors for the two imbalanced classes of the robust classifier trained with adversaries and anti-adversaries. Right: Total robust error for the two imbalanced classes of the robust classifier trained with adversaries and anti-adversaries.

wise robust errors for the two classes

$$\begin{aligned}&\mathcal{R}_{\text{rob}}\left(f_{\text{rob}},-1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A-\frac{\log V}{2A}+\frac{2\sqrt{d}}{\sigma}\epsilon\right\},\\&\mathcal{R}_{\text{rob}}\left(f_{\text{rob}},+1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A+\frac{\log V}{2A}+\frac{\sqrt{d}(1-\rho)}{\sigma}\epsilon\right\},\end{aligned}\tag{A.63}$$

where $A=\frac{\sqrt{d}(\eta-\frac{\epsilon(\rho-1)}{2})}{\sigma}$. Similar to the manner of the proof of Theorem A.4, the class-wise natural errors for the two classes are

$$\begin{aligned}&\mathcal{R}_{\text{nat}}\left(f_{\text{rob}},-1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A-\frac{\log V}{2A}+\frac{\sqrt{d}}{\sigma}\epsilon\right\},\\&\mathcal{R}_{\text{nat}}\left(f_{\text{rob}},+1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A+\frac{\log V}{2A}-\frac{\sqrt{d}\rho}{\sigma}\epsilon\right\}.\end{aligned}\tag{A.64}$$

□

Thus, the natural and robust errors of the two classes change as ρ changes. Then, we show how the natural and robust errors of the two classes change with the increase in ρ, when the adversaries and anti-adversaries are combined in training.

Proof of Corollary A.4 (Corollary 4 in Section 3) In this subsection, we show how the natural and robust errors of the

Corollary A.4. (Corollary 4 in Section 3) *For a data distribution $\mathcal{D}_V$ in Eq. (A.33), assume that class "-1" is anti-adversarially perturbed with the perturbation bound ϵ, and class "$+1$" is adversarially perturbed with the bound $\rho\times\epsilon$ ($0\leq\epsilon,\rho\epsilon<\eta$). When $V<e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}$, the adversarially trained model will increase and decrease the natural and robust errors of class "-1" and class "$+1$", with the increase in ρ, respectively.*

Proof. The value of b_{nat} is shown in the proof of Corollary A.3. For the classifier trained with adversaries and anti-adversaries, b_{rob} is calculated in Eq. (A.62), which can be represented as

$$b_{\text{rob}}=g_{\text{m}}(\eta-\frac{\epsilon(\rho-1)}{2})+\frac{(\rho+1)d\epsilon}{2}.\tag{A.65}$$

Similar to the proof of Corollary A.3, we can verify that b_{rob} is a monotone increasing function of ρ when $V<e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}$. Thus, the classification boundary moves toward class "-1", with the increase in ρ. Similar to the proof of Corollary A.3, we prove that the adversarially trained model will increase the natural and robust errors of the easy class "-1" and decrease the natural and robust errors of the hard class "$+1$", with the increase in ρ. □

Figs. A-9 and A-10 show the variation of the natural errors with the increase in ρ. The values of the parameters are the same as those in the last subsection. The natural and robust errors of class "-1" are increased, and the two errors of class "$+1$" are decreased, with the increase in ρ. When the natural and robust errors of the two classes are the same, the best fairness between classes can be obtained.

Then, we show that when the same performance (same natural error) is achieved, combining adversaries and anti-adversaries requires a smaller perturbation bound compared with only adversaries. In accordance with our calculation, if the bound of combining adversaries with anti-adversaries is ϵ, then the bound required for models trained with only adversaries needs to be $10.425\times\epsilon$. Thus, the combination strategy is more effective than using only adversaries, indicating that the anti-adversaries are valuable. Fig. A-11 shows the data distribution and the classification boundaries which are naturally trained and adversarially trained with different directions and bounds. Adversarial training with the same bound exacerbates the performance gap between classes, while both adversarial training with varied bounds and combining adversaries and anti-adversaries with varied bounds can tune the performance gap. In addition, under the same bound (i.e., the same ρ), combining adversaries and anti-adversaries has a more pronounced effect. Thus, the combination strategy is more efficient than only the adversarial perturbation.

In a similar manner in Case I, we can calculate the scope of the classification boundary when different manners are adopted. Assume that the perturbation bounds for class "-1" and class "$+1$" are $\rho_+\times\epsilon$ and $\rho_-\times\epsilon$ ($-\eta/\epsilon<\rho_+,\rho_-<$

Figure A-11: Left: Logistic Regression classifiers (natural and robust) on simulated binary data in Eq. (A.33). Right: Logistic Regression classifiers (natural and robust with adversaries and anti-adversaries) on simulated binary data in Eq. (A.33). Here, the imbalance factor V is set to 50.

η/ϵ), respectively. $\rho_+ < 0$ means that samples in class "+1" are anti-adversarial perturbed. $\rho_- < 0$ means that samples in class "−1" are anti-adversarial perturbed. By calculating, when $\rho_+ = \eta/\epsilon$ and $\rho_- = 0$, then b_{rob} is

$$b_{\text{rob}} = g_{\text{m}}(\frac{\eta}{2}) + \frac{d\eta}{2}. \tag{A.66}$$

When $\rho_+ = 0$ and $\rho_- = \eta/\epsilon$, then b_{rob} is

$$b_{\text{rob}} = g_{\text{m}}(\frac{\eta}{2}) - \frac{d\eta}{2}. \tag{A.67}$$

Then, we consider the occasion where adversaries and anti-adversaries are combined in training. When $\rho_+ = \eta/\epsilon$ and $\rho_- = -\eta/\epsilon$, then b_{rob} is

$$b_{\text{rob}} = g_{\text{m}}(\eta) + d\eta. \tag{A.68}$$

When $\rho_+ = -\eta/\epsilon$ and $\rho_- = \eta/\epsilon$, then b_{rob} is

$$b_{\text{rob}} = g_{\text{m}}(\eta) - d\eta. \tag{A.69}$$

Fig. A-12 shows the scope of the classification boundary under different manners. Adversarial training with varied bounds has larger scope of the classification boundary compared with TRADES (Zhang et al. 2019). Besides, the scope of the classification boundary under the combination strategy is the largest. Thus, the combination strategy can achieve a better robustness-accuracy tradeoff.

Case III: Classes with Noisy Labels

In this case, the two classes' variances and prior probabilities are assumed to be identical, i.e., $\sigma_{+1} = \sigma_{-1} = \sigma$ and $p_+ = p_-$. Without loss of generality, class "−1" is assumed to contain flipped noise labels. That is, the data are generated as follows:

$$\begin{aligned}
&\tilde{y} \overset{u.a.r}{\sim} \{-1, +1\},\\
&y = \begin{cases} +1 & \tilde{y} = +1,\\ +1 & \text{with a probability } p \text{ and } \tilde{y} = -1,\\ -1 & \text{with a probability } 1-p \text{ and } \tilde{y} = -1, \end{cases}\\
&\boldsymbol{\theta} = (\overbrace{\eta, \ldots, \eta}^{\dim=d}),\\
&\boldsymbol{x} \sim \begin{cases} \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2 I\right), & \text{if } y = +1,\\ \mathcal{N}\left(-\boldsymbol{\theta}, \sigma^2 I\right), & \text{if } y = -1. \end{cases}
\end{aligned} \tag{A.70}$$

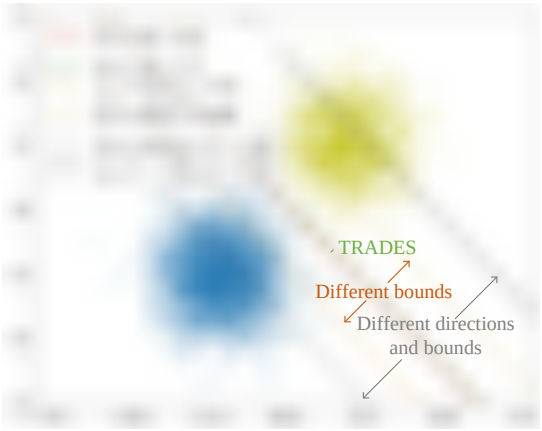

Figure A-12: Scope of the classification boundary under different manners on imbalanced data, including natural training (red line), standard adversarial training (green line), TRADES (scope between the red line and green line), adversarial training with different bounds, and adversarial training with different directions and bounds. The values of parameters are $V = 2$, $\eta = 2$, $\epsilon = 0.8$, and $\sigma = 1$. The bounds for class "+1" and "−1" are denoted as $\rho_+ \times \epsilon$ and $\rho_- \times \epsilon$ ($-\eta/\epsilon < \rho_+, \rho_- < \eta/\epsilon$), respectively. $\rho_+(\rho_-) < 0$ means that class "$+1(-1)$" is anti-adversarially perturbed.

where p (< 1) is the flipping rate for class "−1". Intuitively, class "−1" is harder than class "+1" as it contains noisy labels. We will show that the error of class "−1" is larger than that of class "+1" under natural training.

Proof of Theorem A.6 In this subsection, we first prove that the error of class "−1" is larger than that of class "+1" under natural training. The theorem is shown below.

Theorem A.6. *For a data distribution $\mathcal{D}_N$ in Eq. (A.70) with the flipping rate p, the optimal linear classifier f_{nat} which minimizes the average natural classification error is*

$$f_{nat} = \arg\min_f \Pr(f(\boldsymbol{x}) \neq y). \tag{A.71}$$

It has the natural errors for the two classes:

$$\begin{aligned}
\mathcal{R}_{nat}\left(f_{nat}, -1\right) &= \Pr\left\{\mathcal{N}(0,1) \leq -A + \frac{\log\sqrt{\frac{1+p}{1-p}}}{A}\right\},\\
\mathcal{R}_{nat}\left(f_{nat}, +1\right) &= \Pr\left\{\mathcal{N}(0,1) \leq -A - \frac{\log\sqrt{\frac{1+p}{1-p}}}{A}\right\},
\end{aligned} \tag{A.72}$$

where $A = \frac{\sqrt{d}\eta}{\sigma}$. As a result, class "−1" has a larger natural error:

$$\mathcal{R}_{nat}\left(f_{nat}, -1\right) > \mathcal{R}_{nat}\left(f_{nat}, +1\right). \tag{A.73}$$

Proof. The optimal linear classifier is denoted as $f_{\text{nat}}(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^d \boldsymbol{x}_i + b_{\text{nat}}\right)$. Now, we calculate the optimal b_{nat}

under natural training. The average natural error is

$$\begin{aligned}
R_{\text{nat}}(f) &= \Pr\{f(\boldsymbol{x}) \neq y\} \\
&\propto \frac{1-p}{2}\Pr\{f(\boldsymbol{x}) = +1 \mid y = -1\} \\
&+ \frac{1+p}{2}\Pr\{f(\boldsymbol{x}) = -1 \mid y = +1\} \\
&= \frac{1-p}{2}\Pr\left\{\sum_{i=1}^{d} \boldsymbol{x}_i + b_{\text{nat}} > 0 \mid y = -1\right\} \\
&+ \frac{1+p}{2}\Pr\left\{\sum_{i=1}^{d} \boldsymbol{x}_i + b_{\text{nat}} < 0 \mid y = +1\right\} \\
&= \frac{1-p}{2}\Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}\eta}{\sigma} + \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{nat}}\right\} \\
&+ \frac{1+p}{2}\Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}\eta}{\sigma} - \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{nat}}\right\}.
\end{aligned} \tag{A.74}$$

The optimal b_{nat} to minimize $R_{\text{nat}}(f)$ is achieved at the point that $\frac{\partial \mathcal{R}_{\text{nat}}(f)}{\partial b_{\text{nat}}} = 0$. Thus, the optimal b_{nat} is calculated as

$$b_{\text{nat}} = \frac{\sigma^2 \log\sqrt{\frac{1+p}{1-p}}}{\eta} := g_{\text{n}}(\eta). \tag{A.75}$$

By incorporating the optimal b_{nat} into Eq. (A.74), we can get the class-wise natural errors for the two classes:

$$\begin{aligned}
\mathcal{R}_{\text{nat}}(f_{\text{nat}}, -1) &= \Pr\left\{\mathcal{N}(0,1) \leq -A + \frac{\log\sqrt{\frac{1+p}{1-p}}}{A}\right\}, \\
\mathcal{R}_{\text{nat}}(f_{\text{nat}}, +1) &= \Pr\left\{\mathcal{N}(0,1) \leq -A - \frac{\log\sqrt{\frac{1+p}{1-p}}}{A}\right\}.
\end{aligned} \tag{A.76}$$

Since $0 < p < 1$, we have the direct conclusion that $\mathcal{R}_{\text{nat}}(f_{\text{nat}}, -1) > \mathcal{R}_{\text{nat}}(f_{\text{nat}}, +1)$. □

Theorem A.6 demonstrates that class "-1" which contains flipped noise labels is harder than class "$+1$" under natural training. The class-wise difference is due to the flipping rate p. If $p = 0$, the natural errors for the two classes are the same. Next, we will show that standard adversarial training which has the same perturbation direction and bound for all training samples will exacerbate the performance gap, while adversarial training with different perturbation directions and bounds can tune the performance gap between the two classes, as well as the tradeoff between the robustness and the accuracy of the robust model.

Proof of Theorem A.7 In this subsection, we prove that standard adversarial training will exacerbate the performance gap between the two classes.

Theorem A.7. *For a data distribution $\mathcal{D}_N$ in Eq. (A.70) with the flipping rate p, assume that the perturbation bound for all samples is ϵ ($0 < \epsilon < \eta$). Then, the optimal robust linear classifier f_{rob} which minimizes the average robust error is*

$$f_{rob} = \arg\min_{f} \Pr(\exists \|\boldsymbol{\delta}\| \leq \epsilon, \quad s.t. f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y). \tag{A.77}$$

It has the natural errors for the two classes:

$$\begin{aligned}
&\mathcal{R}_{rob}(f_{nat}, -1) \\
&= \Pr\left\{\mathcal{N}(0,1) \leq -A + \frac{\log\sqrt{\frac{1+p}{1-p}}}{A} + \frac{(3p-2)\sqrt{d}}{2\sigma}\epsilon\right\}, \\
&\mathcal{R}_{rob}(f_{nat}, +1) \\
&= \Pr\left\{\mathcal{N}(0,1) \leq -A - \frac{\log\sqrt{\frac{1+p}{1-p}}}{A} - \frac{(p+2)\sqrt{d}}{2\sigma}\epsilon\right\},
\end{aligned} \tag{A.78}$$

where $A = \frac{\sqrt{d}(\eta - (1-\frac{p}{2}))\epsilon}{\sigma}$.

Proof. The optimal linear robust classifier is defined as $f_{\text{rob}}(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{d} \boldsymbol{x}_i + b_{\text{rob}}\right)$. Then, we calculate the optimal b_{rob}. The average robust error is

$$\begin{aligned}
\mathcal{R}_{\text{rob}}(f) &= \Pr(\exists \|\boldsymbol{\delta}\| \leq \epsilon, \quad f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y) \\
&= \max_{\|\boldsymbol{\delta}\| \leq \epsilon} \Pr(f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y) \\
&= \frac{1-p}{2}\Pr(f(\boldsymbol{x} + (\overbrace{\epsilon, \cdots, \epsilon}^{\dim=d})) \neq -1 \mid y = -1) \\
&+ \frac{1+p}{2}\Pr(f(\boldsymbol{x} - (\overbrace{\epsilon, \cdots, \epsilon}^{\dim=d})) \neq +1 \mid y = +1) \\
&= \frac{1-p}{2}\Pr\left\{\sum_{i=1}^{d} (\boldsymbol{x}_i + \epsilon) + b_{\text{rob}} > 0 \mid y = -1\right\} \\
&+ \frac{1+p}{2}\Pr\left\{\sum_{i=1}^{d} (\boldsymbol{x}_i - \epsilon) + b_{\text{rob}} < 0 \mid y = +1\right\} \\
&= \frac{1-p}{2}\Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} + \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\} \\
&+ \frac{1+p}{2}\Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} - \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}.
\end{aligned} \tag{A.79}$$

The optimal b_{rob} to minimize $\mathcal{R}_{\text{rob}}(f)$ is achieved at the point that $\frac{\partial \mathcal{R}_{\text{rob}}(f)}{\partial b_{\text{rob}}} = 0$. Thus, the optimal b_{rob} is solved as

$$b_{\text{rob}} = \frac{\sigma^2 \log\sqrt{\frac{1+p}{1-p}}}{\eta - (1-\frac{p}{2})\epsilon} + pd\epsilon. \tag{A.80}$$

By incorporating the optimal b_{rob} into Eq. (A.79), we can get

the class-wise robust errors for the two classes

$$\begin{aligned}&\mathcal{R}_{\text{rob}}\left(f_{\text{rob}},-1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A+\frac{\log\sqrt{\frac{1+p}{1-p}}}{A}+\frac{3p\sqrt{d}}{2\sigma}\epsilon\right\},\\&\mathcal{R}_{\text{rob}}\left(f_{\text{rob}},+1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A-\frac{\log\sqrt{\frac{1+p}{1-p}}}{A}-\frac{p\sqrt{d}}{2\sigma}\epsilon\right\},\end{aligned}\tag{A.81}$$

where $A=\frac{\sqrt{d}(\eta-(1-\frac{p}{2}))\epsilon}{\sigma}$. In accordance with the definition of natural error, by incorporating the optimal b_{rob} into the following formula

$$\begin{aligned}&\frac{1-p}{2}\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}\eta}{\sigma}+\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}\\&+\frac{1+p}{2}\Pr\left\{\mathcal{N}(0,1)<-\frac{\sqrt{d}\eta}{\sigma}-\frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\},\end{aligned}\tag{A.82}$$

we obtain the class-wise natural errors for the two classes

$$\begin{aligned}&\mathcal{R}_{\text{rob}}\left(f_{\text{nat}},-1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A+\frac{\log\sqrt{\frac{1+p}{1-p}}}{A}+\frac{(3p-2)\sqrt{d}}{2\sigma}\epsilon\right\},\\&\mathcal{R}_{\text{rob}}\left(f_{\text{nat}},+1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A-\frac{\log\sqrt{\frac{1+p}{1-p}}}{A}-\frac{(p+2)\sqrt{d}}{2\sigma}\epsilon\right\}.\end{aligned}\tag{A.83}$$

b_{rob} in Eq. (A.80) can be written as

$$b_{\text{rob}}=g_{\text{n}}(\eta-(1-\frac{p}{2})\epsilon)+pd\epsilon.\tag{A.84}$$

Then, we show that g_{n} is a monotone decreasing function from 0 to η. The deviate of g_{n} in terms of η is

$$\frac{dg_{\text{n}}(\eta)}{d\eta}=-\frac{\sigma^2\log\sqrt{\frac{1+p}{1-p}}}{\eta^2}<0.\tag{A.85}$$

Thus, we have the relation that $b_{\text{nat}}<b_{\text{rob}}<0$. It indicates that the classification boundary further moves toward class "-1" when the model is adversarially trained with the same perturbation bound for all samples. Thus, the performance gap between the two classes is further exacerbated. Then, we show that adversarial training with unequal perturbation bounds (0 for noisy samples and ϵ for clean samples) can tune the performance gap and the tradeoff between the robustness and accuracy of the model on noisy data. □

Proof of Theorem A.8 In this subsection, we prove that adversarial training with unequal perturbation bounds can tune the performance gap between classes and the tradeoff between the robustness and accuracy of the model. Specifically, the perturbation bound for noisy samples is 0, and the adversarial perturbation bound for clean samples is ϵ.

Theorem A.8. *For a data distribution $\mathcal{D}_N$ in Eq. (A.70) with the flipping rate p, assume that the perturbation bound for noisy samples and clean samples are 0 and ϵ ($0<\epsilon<\eta$), respectively. Then, the optimal robust linear classifier f_{rob} which minimizes the average robust error is*

$$\begin{aligned}f_{rob}=\arg\min_f\{&\Pr(\exists\|\boldsymbol{\delta}\|\leq\epsilon,\quad s.t.f(\boldsymbol{x}+\boldsymbol{\delta})\neq y\mid\boldsymbol{x}\in\boldsymbol{X}^c)\\&+\Pr(f(\boldsymbol{x})\neq y\mid\boldsymbol{x}\in\boldsymbol{X}^n)\},\end{aligned}\tag{A.86}$$

where $\boldsymbol{X}^c$ and $\boldsymbol{X}^n$ refer to the set of clean and noisy samples, respectively. It has the natural errors for the two classes:

$$\begin{aligned}&\mathcal{R}_{rob}\left(f_{nat},-1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A+\frac{\log\sqrt{\frac{1+p}{1-p}}}{A}+\frac{(p-1)\sqrt{d}}{\sigma}\epsilon\right\},\\&\mathcal{R}_{rob}\left(f_{nat},+1\right)\\&=\Pr\left\{\mathcal{N}(0,1)\leq -A-\frac{\log\sqrt{\frac{1+p}{1-p}}}{A}-\frac{\sqrt{d}}{\sigma}\epsilon\right\},\end{aligned}\tag{A.87}$$

where $A=\frac{\sqrt{d}(\eta-(1-\frac{p}{2}))\epsilon}{\sigma}$.

Proof. The perturbation bound ϵ is also limited in the region of $[0,\eta)$. The optimal linear robust classifier is $f_{\text{rob}}(\boldsymbol{x})=\text{sign}\left(\sum_{i=1}^{d}\boldsymbol{x}_i+b_{\text{rob}}\right)$. Now, we calculate the optimal b_{rob}.

The average robust error is

$$\begin{aligned}
\mathcal{R}_{\text{rob}}(f) &= \Pr(\exists\|\boldsymbol{\delta}\| \le \epsilon, \quad \text{s.t.} f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid \boldsymbol{x} \in \boldsymbol{X}^c) \\
&\quad + \Pr(f(\boldsymbol{x}) \neq y \mid \boldsymbol{x} \in \boldsymbol{X}^n) \\
&= \max_{\|\boldsymbol{\delta}\| \le \epsilon} \Pr(f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid \boldsymbol{x} \in \boldsymbol{X}^c) \\
&\quad + \Pr(f(\boldsymbol{x}) \neq y \mid \boldsymbol{x} \in \boldsymbol{X}^n) \\
&= \frac{1-p}{2} \Pr(f(\boldsymbol{x}+(\overbrace{\epsilon,\cdots,\epsilon}^{\dim=d})) \neq -1 \,|\, y=-1, x \in \boldsymbol{X}^c) \\
&\quad + \frac{1}{2} \Pr(f(\boldsymbol{x}-(\overbrace{\epsilon,\cdots,\epsilon}^{\dim=d})) \neq +1 \,|\, y=+1, \boldsymbol{x} \in \boldsymbol{X}^c) \\
&\quad + \frac{p}{2} \Pr(f(\boldsymbol{x}) \neq +1 \,|\, y=+1, \boldsymbol{x} \in \boldsymbol{X}^n) \\
&= \frac{1-p}{2} \Pr\left\{ \sum_{i=1}^{d} (\boldsymbol{x}_i + \epsilon) + b_{\text{rob}} > 0 \mid y=-1 \right\} \\
&\quad + \frac{1}{2} \Pr\left\{ \sum_{i=1}^{d} (\boldsymbol{x}_i - \epsilon) + b_{\text{rob}} < 0 \mid y=+1, \boldsymbol{x} \in \boldsymbol{X}^c \right\} \\
&\quad + \frac{p}{2} \Pr\left\{ \sum_{i=1}^{d} \boldsymbol{x}_i + b_{\text{rob}} < 0 \mid y=+1, \boldsymbol{x} \in \boldsymbol{X}^n \right\} \\
&= \frac{1-p}{2} \Pr\left\{ \mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} + \frac{1}{\sqrt{d}\sigma} \cdot b_{\text{rob}} \right\} \\
&\quad + \frac{1}{2} \Pr\left\{ \mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} - \frac{1}{\sqrt{d}\sigma} \cdot b_{\text{rob}} \right\} \\
&\quad + \frac{p}{2} \Pr\left\{ \mathcal{N}(0,1) < -\frac{\sqrt{d}\eta}{\sigma} - \frac{1}{\sqrt{d}\sigma} \cdot b_{\text{rob}} \right\}.
\end{aligned} \tag{A.88}$$

The optimal b_{rob} to minimize $\mathcal{R}_{\text{rob}}(f)$ is achieved at the point that $\frac{\partial \mathcal{R}_{\text{rob}}(f)}{\partial b_{\text{rob}}} = 0$. Thus, the optimal b_{rob} is calculated as

$$b_{\text{rob}} = \frac{\sigma^2 \log \sqrt{\frac{1+p}{1-p}}}{\eta - (1-\frac{p}{2})\epsilon} + \frac{pd}{2}\epsilon. \tag{A.89}$$

Similar to the proof of Theorem A.7, we can get the class-wise robust errors for the two classes

$$\begin{aligned}
&\mathcal{R}_{\text{rob}}(f_{\text{rob}}, -1) \\
&= \Pr\left\{ \mathcal{N}(0,1) \le -A + \frac{\log \sqrt{\frac{1+p}{1-p}}}{A} + \frac{p\sqrt{d}}{\sigma}\epsilon \right\}, \\
&\mathcal{R}_{\text{rob}}(f_{\text{rob}}, +1) \\
&= \Pr\left\{ \mathcal{N}(0,1) \le -A - \frac{\log \sqrt{\frac{1+p}{1-p}}}{A} \right\},
\end{aligned} \tag{A.90}$$

where $A = \frac{\sqrt{d}(\eta-(1-\frac{p}{2}))\epsilon}{\sigma}$. Following the manner of the proof of Theorem A.7, it is easy to obtain that the class-wise natural errors for the two classes are

$$\begin{aligned}
&\mathcal{R}_{\text{rob}}(f_{\text{nat}}, -1) \\
&= \Pr\left\{ \mathcal{N}(0,1) \le -A + \frac{\log \sqrt{\frac{1+p}{1-p}}}{A} + \frac{(p-1)\sqrt{d}}{\sigma}\epsilon \right\}, \\
&\mathcal{R}_{\text{rob}}(f_{\text{nat}}, +1) \\
&= \Pr\left\{ \mathcal{N}(0,1) \le -A - \frac{\log \sqrt{\frac{1+p}{1-p}}}{A} - \frac{\sqrt{d}}{\sigma}\epsilon \right\}.
\end{aligned} \tag{A.91}$$

b_{rob} in Eq. (A.89) can be written as

$$b_{\text{rob}} = g_{\text{n}}(\eta - (1-\frac{p}{2})\epsilon) + \frac{pd}{2}\epsilon. \tag{A.92}$$

Obviously, b_{rob} in Eq. (A.92) is smaller than that in Eq. (A.84), indicating that the classification boundary moves toward class "+1" when the noisy samples are not adversarially perturbed. Thus, the error of class "+1" is increased and that of class "−1" is decreased. Consequently, the performance gap between the two classes can be decreased when the model is adversarially trained with varied bounds. □

Proof of Theorem A.9 In this subsection, we prove that combining adversaries and anti-adversaries in training can tune the performance gap between classes and the tradeoff between the robustness and accuracy of the model.

Theorem A.9. *For a data distribution $\mathcal{D}_N$ in Eq. (A.70) with the flipping rate p, assume that clean samples $\boldsymbol{x} \in \boldsymbol{X}^c$ are adversarially perturbed with the perturbation bound ϵ, and noisy samples $\boldsymbol{x} \in \boldsymbol{X}^n$ are anti-adversarially perturbed with the bound $\rho \times \epsilon$ ($0 \le \epsilon, \rho\epsilon < \eta$). The optimal robust linear classifier f_{rob} which minimizes the average robust error is*

$$\begin{aligned}
f_{rob} = \arg\min_f \{ &\Pr(\exists\|\boldsymbol{\delta}\| \le \epsilon, \ s.t. f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid \boldsymbol{x} \in \boldsymbol{X}^c) \\
&+ \Pr(\exists\|\boldsymbol{\delta}\| \le \rho\times\epsilon, \ s.t. f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid \boldsymbol{x} \in \boldsymbol{X}^n) \},
\end{aligned} \tag{A.93}$$

It has the natural errors for the two classes:

$$\begin{aligned}
&\mathcal{R}_{rob}(f_{nat}, -1) \\
&= \Pr\left\{ \mathcal{N}(0,1) \le -A + \frac{\log \sqrt{\frac{1+p}{1-p}}}{A} - \frac{p(\rho-1)\sqrt{d}}{\sigma}\epsilon - \frac{\sqrt{d}\epsilon}{\sigma} \right\}, \\
&\mathcal{R}_{rob}(f_{nat}, +1) \\
&= \Pr\left\{ \mathcal{N}(0,1) \le -A - \frac{\log \sqrt{\frac{1+p}{1-p}}}{A} - \frac{\sqrt{d}\epsilon}{\sigma} \right\},
\end{aligned} \tag{A.94}$$

where $A = \frac{\sqrt{d}(\eta-\epsilon-\frac{p(\rho-1)}{2}\epsilon)}{\sigma}$.

Proof. We consider that clean and noisy samples have different perturbation directions. Clean samples $\boldsymbol{x} \in \boldsymbol{X}^c$ are adversarially perturbed with the bound ϵ, and noisy samples $\boldsymbol{x} \in \boldsymbol{X}^n$ are anti-adversarially perturbed with the bound

$\rho \times \epsilon$. ϵ is limited in the region of $[0, \eta)$, and $\rho \times \epsilon$ is also in the region of $[0, \eta)$.

The optimal linear robust classifier is $f_{\text{rob}}(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{d} \boldsymbol{x}_i + b_{\text{rob}}\right)$. Then, the optimal b_{rob} which is trained with adversaries and anti-adversaries can be calculated. The average robust error is

$$
\begin{aligned}
\mathcal{R}_{\text{rob}}(f) &= \Pr(\exists \|\boldsymbol{\delta}\| \leq \epsilon, \quad f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid \boldsymbol{x} \in \boldsymbol{X}^c) \\
&\quad + \Pr(\exists \|\boldsymbol{\delta}\| \leq \rho \times \epsilon, \quad f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid x \in \boldsymbol{X}^n) \\
&= \max_{\|\boldsymbol{\delta}\| \leq \epsilon} \Pr(f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid x \in \boldsymbol{X}^c) \\
&\quad + \min_{\|\boldsymbol{\delta}\| \leq \rho \times \epsilon} \Pr(f(\boldsymbol{x}+\boldsymbol{\delta}) \neq y \mid x \in \boldsymbol{X}^n) \\
&= \frac{(1-p)}{2} \Pr(f(\boldsymbol{x} + (\overbrace{\epsilon, \cdots, \epsilon}^{\dim=d})) \neq -1 \mid y=-1, x \in \boldsymbol{X}^c) \\
&\quad + \frac{p}{2} \Pr(f(\boldsymbol{x} + (\overbrace{\rho \times \epsilon, \cdots, \rho \times \epsilon}^{\dim=d})) \neq +1 \mid y=+1, x \in \boldsymbol{X}^n) \\
&\quad + \frac{1}{2} \Pr(f(\boldsymbol{x} - (\overbrace{\epsilon, \cdots, \epsilon}^{\dim=d})) \neq +1 \mid y=+1, x \in \boldsymbol{X}^c) \\
&= \frac{(1-p)}{2} \Pr\left\{\sum_{i=1}^{d} (\boldsymbol{x}_i + \epsilon) + b_{\text{rob}} > 0 \mid y=-1, \boldsymbol{x} \in \boldsymbol{X}^c\right\} \\
&\quad + \frac{p}{2} \Pr\left\{\sum_{i=1}^{d} (\boldsymbol{x}_i + \rho\epsilon) + b_{\text{rob}} < 0 \mid y=+1, x \in \boldsymbol{X}^n\right\} \\
&\quad + \frac{1}{2} \Pr\left\{\sum_{i=1}^{d} (\boldsymbol{x}_i - \epsilon) + b_{\text{rob}} < 0 \mid y=+1, x \in \boldsymbol{X}^c\right\} \\
&= \frac{(1-p)}{2} \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} + \frac{1}{\sqrt{d}\sigma} \cdot b_{\text{rob}}\right\} \\
&\quad + \frac{p}{2} \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta+\rho\times\epsilon)}{\sigma} - \frac{1}{\sqrt{d}\sigma} \cdot b_{\text{rob}}\right\} \\
&\quad + \frac{1}{2} \Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} - \frac{1}{\sqrt{d}\sigma} \cdot b_{\text{rob}}\right\}.
\end{aligned}
\tag{A.95}
$$

The optimal b_{rob} to minimize $\mathcal{R}_{\text{rob}}(f)$ is achieved at the point that $\frac{\partial \mathcal{R}_{\text{rob}}(f)}{\partial b_{\text{rob}}} = 0$. Thus, the optimal b_{rob} is solved as

$$
b_{\text{rob}} = \frac{\sigma^2 \log \sqrt{\frac{1+p}{1-p}}}{\eta - \epsilon - \frac{p(\rho-1)}{2}\epsilon} - \frac{p(\rho-1)d\epsilon}{2}. \tag{A.96}
$$

Similar to the proof of Theorem A.7, we can get the class-wise robust errors for the two classes

$$
\begin{aligned}
&\mathcal{R}_{\text{rob}}(f_{\text{rob}}, -1) \\
&= \Pr\left\{\mathcal{N}(0,1) \leq -A + \frac{\log \sqrt{\frac{1+p}{1-p}}}{A} - \frac{p(\rho-1)\sqrt{d}}{\sigma}\epsilon\right\}, \\
&\mathcal{R}_{\text{rob}}(f_{\text{rob}}, +1) \\
&= \Pr\left\{\mathcal{N}(0,1) \leq -A - \frac{\log \sqrt{\frac{1+p}{1-p}}}{A}\right\},
\end{aligned}
\tag{A.97}
$$

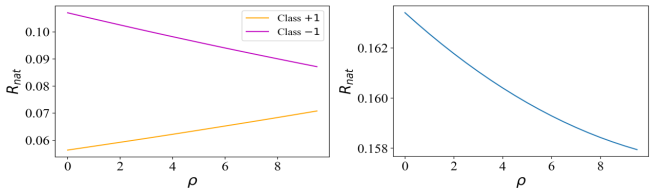

Figure A-13: Left: Natural errors for the two classes of the robust classifier trained with adversaries and anti-adversaries. Right: Total natural error for the two classes of the robust classifier trained with adversaries and anti-adversaries.

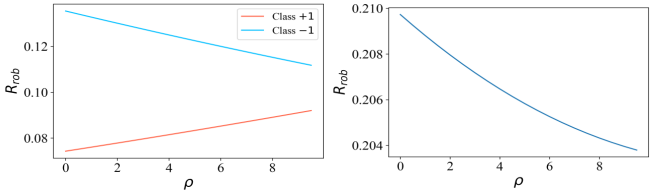

Figure A-14: Left: Robust errors for the two classes of the robust classifier trained with adversaries and anti-adversaries. Right: Total robust error for the two classes of the robust classifier trained with adversaries and anti-adversaries.

where $A = \frac{\sqrt{d}(\eta - \epsilon - \frac{p(\rho-1)}{2}\epsilon)}{\sigma}$. Following the manner of the proof of Theorem A.7, it is easy to obtain that the class-wise natural errors for the two classes are

$$
\begin{aligned}
&\mathcal{R}_{\text{rob}}(f_{\text{nat}}, -1) \\
&= \Pr\left\{\mathcal{N}(0,1) \leq -A + \frac{\log \sqrt{\frac{1+p}{1-p}}}{A} - \frac{p(\rho-1)\sqrt{d}}{\sigma}\epsilon - \frac{\sqrt{d}\epsilon}{\sigma}\right\}, \\
&\mathcal{R}_{\text{rob}}(f_{\text{nat}}, +1) \\
&= \Pr\left\{\mathcal{N}(0,1) \leq -A - \frac{\log \sqrt{\frac{1+p}{1-p}}}{A} - \frac{\sqrt{d}\epsilon}{\sigma}\right\}.
\end{aligned}
\tag{A.98}
$$

□

Then, we prove that adversarial training with different perturbation directions on noisy and clean samples can tune the performance gap between classes and the tradeoff between the robustness and the accuracy of the model on noisy data. Corollary A.5 shows how the natural and robust errors of the two classes change with the increase in ρ.

Proof of Corollary A.5 In this subsection, we show how the natural and robust errors of the two classes change with the increase in ρ when the model is trained with adversaries and anti-adversaries.

Corollary A.5. *For a data distribution $\mathcal{D}_N$ in Eq. (A.70) with the flipping rate p, assume that clean samples $\boldsymbol{x} \in \boldsymbol{X}^c$ are adversarially perturbed with the perturbation bound ϵ, and noisy samples $\boldsymbol{x} \in \boldsymbol{X}^n$ are anti-adversarially perturbed with the bound $\rho \times \epsilon$ ($0 \leq \epsilon, \rho\epsilon < \eta$). When*

$p < e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}} - 1/e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}} + 1$, *the adversarially trained model will increase and decrease the natural and robust errors of class "$+1$" and class "-1", with the increase in ρ, respectively.*

Proof. The value of b_{nat} is shown in the proof of Theorem A.6. For the classifier trained with adversaries and anti-adversaries, b_{rob} is calculated in Eq. (A.96), which can be represented as

$$b_{\text{rob}} = g_{\text{n}}(\eta - \epsilon - \frac{p(\rho-1)}{2}\epsilon) - \frac{p(\rho-1)d\epsilon}{2}. \quad \text{(A.99)}$$

Then, we show that when $p < \frac{e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}-1}{e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}}+1}$, b_{rob} is a monotone decreasing function of ρ. The derivative of b_{rob} with respect to ρ is

$$\frac{\partial b_{\text{rob}}}{\partial \rho} = g'_{\text{n}}(\eta - \epsilon - \frac{p(\rho-1)}{2}\epsilon) \cdot (-\frac{p\epsilon}{2}) - \frac{pd\epsilon}{2}. \quad \text{(A.100)}$$

Then, to ensure $\frac{\partial b_{\text{rob}}}{\partial \rho} < 0$, we just need that $g'_{\text{n}}(\eta - \epsilon - \frac{p(\rho-1)}{2}\epsilon) > -d$. $g'_{\text{n}}(\eta)$ is

$$\frac{dg_{\text{n}}(\eta)}{d\eta} = -\frac{\sigma^2 \log\sqrt{\frac{1+p}{1-p}}}{\eta^2}. \quad \text{(A.101)}$$

Thus, for the robust classifier which is adversarially trained with different directions and bounds, the following inequality needs to be satisfied:

$$\log\sqrt{\frac{1+p}{1-p}} < \frac{d(\eta - \epsilon - \frac{p(\rho-1)}{2}\epsilon)^2}{\sigma^2}. \quad \text{(A.102)}$$

Then, we obtain that the only condition for b_{rob} to be a monotonically decreasing function with respect to ρ is

$$p < \frac{e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}} - 1}{e^{\frac{d(\eta-\epsilon)^2}{2\sigma^2}} + 1}. \quad \text{(A.103)}$$

Thus, the classification boundary moves toward class "$+1$" with the increase in ρ. The following formulas show the natural and robust errors for class "-1" and class "$+1$":

$$\mathcal{R}_{\text{rob}}(f_{\text{rob}}, -1) = \frac{1-p}{2}\Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} + \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}, \quad \text{(A.104)}$$

$$\mathcal{R}_{\text{rob}}(f_{\text{rob}}, +1) = \frac{1+p}{2}\Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}(\eta-\epsilon)}{\sigma} - \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}, \quad \text{(A.105)}$$

$$\mathcal{R}_{\text{nat}}(f_{\text{rob}}, -1) = \frac{1-p}{2}\Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}\eta}{\sigma} + \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}, \quad \text{(A.106)}$$

$$\mathcal{R}_{\text{nat}}(f_{\text{rob}}, +1) = \frac{1+p}{2}\Pr\left\{\mathcal{N}(0,1) < -\frac{\sqrt{d}\eta}{\sigma} - \frac{1}{\sqrt{d}\sigma}\cdot b_{\text{rob}}\right\}. \quad \text{(A.107)}$$

As we can see, the natural and robust errors of class "-1" are positively related to the value of b_{rob} and those of class "$+1$" are negatively related to the value of b_{rob}. Therefore, as ρ increases, b_{rob} will be decreased. Thus, the natural and robust errors of class "-1" and class "$+1$" will be decreased and increased, respectively. □

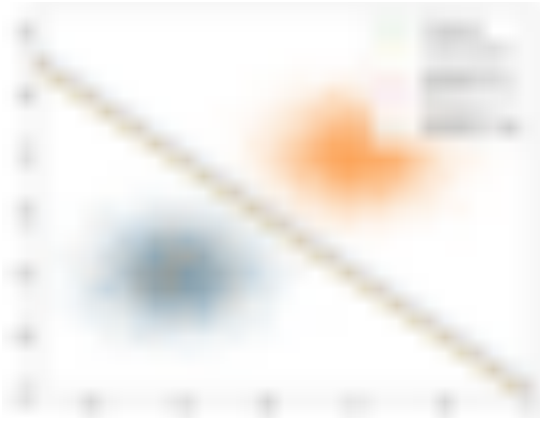

Figure A-15: Logistic Regression classifiers (natural and robust with adversaries and anti-adversaries) on simulated binary data in Eq. (A.70). "Robust $\rho = 0$" means that noisy samples are not perturbed and clean samples are adversarially perturbed with the perturbation bound ϵ. "Robust $\rho = 5$ and $\rho = 10$" means that noisy samples are anti-adversarially perturbed with the perturbation bound $\rho \times \epsilon$ and clean samples are adversarially perturbed with the bound ϵ. Here, the flipping rate p is set to 0.2.

Figs. A-13 and A-14 show the variation of the natural and robust errors as ρ increases. For the parameters, the flipping rate p is set to 0.2. d and ϵ are set to 2 and 0.1, respectively. σ and η are both equal to 1. The natural and robust errors for class "-1" will be decreased, and the two errors for class "$+1$" will be increased, with the increase in ρ. When the natural and robust errors of the two classes are the same, the best fairness between the two categories is obtained.

As shown in Fig. A-15, when both noisy and clean samples are adversarially perturbed, the classification boundary is closer to the harder class "-1". Thus, the performance gap between classes is increased compared with that of natural training. When noisy samples are not perturbed and clean samples are adversarially perturbed, the classification boundary moves toward class "$+1$" compared with standard adversarial training. Thus, the performance gap is decreased when noisy samples are not adversarially perturbed. Thus, the adversaries of noisy samples may harm the model performance. Furthermore, combining adversaries and anti-adversaries (noisy samples are anti-adversarially perturbed and clean samples are adversarially perturbed) in training can further decrease the performance gap.

Then we calculate the scope of the classification boundary when the classifier is adversarially trained and the adversaries and anti-adversaries are combined in training with varied bounds. Assume that the perturbation bounds for samples in class "$+1$", clean samples in class "-1", noisy samples in class "-1" are $\rho_+ \times \epsilon$, $\rho_-^{\text{c}} \times \epsilon$, and $\rho_-^{\text{n}} \times \epsilon$

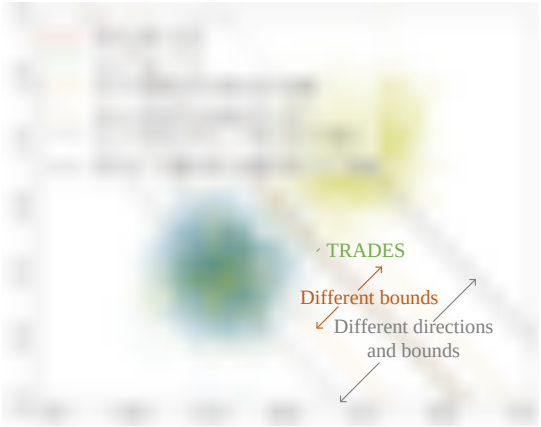

Figure A-16: Scope of the classification boundary under different manners on noisy data, including natural training (red line), standard adversarial training (green line), TRADES (scope between the red line and green line), adversarial training with different bounds, and adversarial training with different directions and bounds. The values of parameters are $p = 0.2$, $\eta = 2$, $\epsilon = 0.2$, and $\sigma = 1$. The bounds for samples in class "$+1$", clean samples in class "-1", and noisy samples in class "-1" are denoted as $\rho_+ \times \epsilon$, $\rho_-^{\mathrm{c}} \times \epsilon$, $\rho_-^{\mathrm{n}} \times \epsilon$ ($-\eta/\epsilon < \rho_+, \rho_-^{\mathrm{c}}, \rho_-^{\mathrm{n}} < \eta/\epsilon$), respectively. $\rho_+(\rho_-^{\mathrm{c}}, \rho_-^{\mathrm{n}}) < 0$ means that samples in class "$+1$" (clean samples in class "-1", noisy samples in class "-1") are anti-adversarially perturbed.

($-\eta/\epsilon < \rho_+, \rho_-^{\mathrm{c}}, \rho_-^{\mathrm{n}} < \eta/\epsilon$), respectively. $\rho_-^{\mathrm{c}} < 0$ means that clean samples in class "-1" are anti-adversarial perturbed. $\rho_-^{\mathrm{n}} < 0$ means that noisy samples in class "-1" are anti-adversarial perturbed. $\rho_+ < 0$ means that samples in class "$+1$" are anti-adversarial perturbed. By calculating, when $\rho_+ = \eta/\epsilon$, $\rho_-^{\mathrm{c}} = 0$, and $\rho_-^{\mathrm{n}} = \eta/\epsilon$, then b_{rob} is

$$b_{\mathrm{rob}} = g_{\mathrm{n}}(\frac{(1+p)\eta}{2}) + \frac{d(p+1)\eta}{2}. \tag{A.108}$$

When $\rho_+ = 0$, $\rho_-^{\mathrm{c}} = \eta/\epsilon$, and $\rho_-^{\mathrm{n}} = 0$, then b_{rob} is

$$b_{\mathrm{rob}} = g_{\mathrm{n}}(\frac{(1+p)\eta}{2}) - \frac{d(1-p)\eta}{2}. \tag{A.109}$$

Then, we consider the occasion where adversaries and anti-adversaries are combined in training. When $\rho_+ = \eta/\epsilon$, $\rho_-^{\mathrm{c}} = -\eta/\epsilon$, and $\rho_-^{\mathrm{n}} = \eta/\epsilon$, then b_{rob} is

$$b_{\mathrm{rob}} = g_{\mathrm{n}}(\eta) + d\eta. \tag{A.110}$$

When $\rho_+ = -\eta/\epsilon$, $\rho_-^{\mathrm{c}} = \eta/\epsilon$, and $\rho_-^{\mathrm{n}} = -\eta/\epsilon$, then b_{rob} is

$$b_{\mathrm{rob}} = g_{\mathrm{n}}(\eta) - d\eta. \tag{A.111}$$

The scope of the classification boundary under different manners is shown in Fig. A-16. Adversarial training with varied perturbation bounds contributes to larger scope of the classification boundary compared with TRADES (Zhang et al. 2019). The scope of the classification boundary under the combination strategy is the largest. Thus, the combination strategy can achieve a better robustness-accuracy trade-off on noisy data.

Case IV: Classes with Skewed Distribution

In this case, the two classes' variances and prior probabilities are assumed to be identical, i.e., $\sigma_{+1} = \sigma_{-1} = \sigma$ and $p_+ = p_-$. Besides, there is no noisy sample in both the two classes. To simplify the problem, we consider that the data are one-dimensional, i.e., $d = 1$, on this occasion. The data are still assumed to be from two classes $\{-1, +1\}$. And the data in class "-1" follow a Gaussian distribution $\mathcal{N}(-\theta, \sigma^2)$. However, the data in class "$+1$" are no longer following the Gaussian distribution. Due to some reasons such as improper data pre-processing or sampling, the training data of class "$+1$" follow a skewed distribution which is denoted as $S\mathcal{N}(\theta, \sigma^2, \alpha)$ (Azzalini 1985), where α is the skew coefficient and the distribution is reduced to the normal distribution when $\alpha = 0$. The probability density function of $S\mathcal{N}(\theta, \sigma^2, \alpha)$ (Azzalini 1985) is $f(x; \theta, \sigma) = 2\phi\left(x; \theta, \sigma\right) \Phi\left(\alpha(x - \theta)\right)$, where $\phi(x; \theta, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}$ and $\Phi(x; \theta, \sigma) = \int_{-\infty}^{x} \phi(t; \theta, \sigma) \mathrm{d}t$. Thus, the data distribution of the two classes with skewed distribution is as follows:

$$\begin{aligned} &y \overset{u.a.r}{\sim} \{-1, +1\}, \quad \theta = \eta, \\ &\boldsymbol{x} \sim \begin{cases} S\mathcal{N}(\theta, \sigma^2, \alpha) & \text{if } y = +1, \\ \mathcal{N}\left(-\theta, \sigma^2\right) & \text{if } y = -1. \end{cases} \end{aligned} \tag{A.112}$$

We consider two cases, including $\alpha > 0$ and $\alpha < 0$. Intuitively, class "$+1$" is harder than class "-1" when $\alpha > 0$, and class "$+1$" is easier than class "-1" when $\alpha < 0$.

Proof of Theorem A.10 In this subsection, we prove that when $\alpha < 0$ ($\alpha > 0$), the error of class "$+1$" is smaller (larger) than that of class "-1" under natural training. The theorem is shown below.

Theorem A.10. *For a data distribution $\mathcal{D}_\alpha$ in Eq. (A.112) which is one-dimensional with the skew coefficient α, assume that the optimal classifier boundary of the two classes is $-\eta < x = x^* < \eta$. When $\alpha < 0$, then the optimal classification boundary $x = x^* < 0$ under natural training. When $\alpha > 0$, then the optimal classification boundary $x = x^* > 0$ under natural training.*

Proof. We assume that the two probability density functions of the two classes have only one intersection. This assumption is reasonable because if the two density functions have two intersections, then the two distributions will overlap and the classification task will fail. In this case, it is reasonable to assume that the optimal classifier boundary is $-\eta < x = x^* < \eta$. Consequently, the optimal classifier boundary $x = x^*$ is obtained where the two probability density functions are equal, which is

$$\begin{aligned} 2\phi\left(x^*; \eta, \sigma^2\right) \Phi\left(\alpha(x^* - \eta)\right) &= \phi\left(x^*; -\eta, \sigma^2\right), \\ 2\Phi\left(\alpha(x^* - \eta)\right) &= \mathrm{e}^{-\frac{(x^*+\eta)^2}{2\sigma^2} + \frac{(x^*-\eta)^2}{2\sigma^2}}, \\ 2\Phi\left(\alpha(x^* - \eta)\right) &= \mathrm{e}^{-\frac{2x^*\eta}{\sigma^2}}. \end{aligned} \tag{A.113}$$

As we can see, when $\alpha = 0$, there is $x^* = 0$. When $\alpha > 0$, if $x^* < 0$, then a contradiction occurs. Thus, we know that the optimal classifier x^* should be $x^* > 0$. When $\alpha < 0$, if

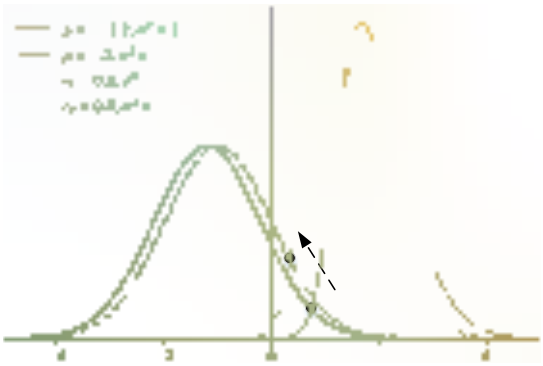

Figure A-17: The occasion when the skewed distribution of class "+1" is far from the decision boundary ($\alpha > 0$). The solid lines represent the distributions of the training set, and the dashed lines represent the perturbed distributions of the training set. For the parameters, there are $\alpha = 3$, $\eta = 1.2$, and $\sigma = 1$.

Figure A-18: The occasion when the skewed distribution of class "+1" is close to the decision boundary ($\alpha < 0$). For the parameters, there are $\alpha = -3$, $\eta = 2$, and σ =1.

$x^* > 0$, then a contradiction occurs. Thus, we know that the optimal classifier x^* should be $x^* < 0$. □

Theorem A.10 demonstrates that when $\alpha < 0$, the error of the optimal classifier for class "+1" is smaller than that for class "−1" under natural training as the optimal classification boundary is biased toward class "−1". When $\alpha > 0$, the error of the classifier for class "+1" is larger than that for class "+1" under natural training as the optimal classification boundary is biased toward class "+1". The class-wise difference is only due to the skew coefficient α. If $\alpha = 0$, then the natural errors for the two classes are the same. Next, we show that if $\alpha > 0$, then the adversaries of samples in class "+1" can tune the performance gap between classes and the tradeoff between the accuracy and robustness of the model. In addition, if $\alpha < 0$, then the anti-adversaries of class "+1" can tune the performance gap and tradeoff.

Proof of Corollary A.6 In this subsection, we show that when the skewed distribution of class "+1" is far from the classification boundary ($\alpha > 0$), the adversaries of samples in class "+1" can tune the performance gap between classes and the tradeoff between the accuracy and robustness of the model.

Corollary A.6. *For a data distribution $\mathcal{D}_\alpha$ in Eq. (A.112) which is one-dimensional with the skew coefficient $\alpha > 0$, assume that the adversarial perturbation bounds for class "+1" and class "−1" are $\rho \times \epsilon$ and ϵ ($0 \leq \epsilon, \rho\epsilon < \eta$), respectively. The adversarially trained model will increase and decrease the errors of class "−1" and class "+1", with the increase in ρ, respectively.*

Proof. When only one intersection x^* exists between the two probability density functions of the two distributions, then the optimal classifier for the two classes is $x = x^*$. If x^* moves left, then the errors of classes "−1" and "+1" are increased and decreased, respectively. We can easily obtain that the intersection moves left as ρ increases. Fig. A-17 shows the variation of the position of the optimal classification boundary. As ρ increases, the intersection of the two probability density functions moves left. Thus, the errors of classes "+1" and "−1" are decreased and increased, respectively. Therefore, the performance gap between classes can be tuned with different values of ρ. In addition, adversarial training with varied bounds contributes to larger scope of the classification boundary which covers the boundary of the standard adversarial training. Through this method, a better tradeoff between the accuracy and robustness of the model can be obtained. □

Proof of Corollary A.7 In this subsection, we show that when the skewed distribution of class "+1" is closer to the classification boundary ($\alpha < 0$), the anti-adversaries of samples in class "+1" can tune the performance gap between classes and the tradeoff between the accuracy and robustness of the model.

Corollary A.7. *For a data distribution $\mathcal{D}_\alpha$ in Eq. (A.112) which is one-dimensional with the skew coefficient $\alpha < 0$, assume that class "+1" is anti-adversarially perturbed with the perturbation bound $\rho \times \epsilon$, and class "−1" is adversarially perturbed with the bound ϵ ($0 \leq \epsilon, \rho\epsilon < \eta$). The adversarially trained model will increase and decrease the errors of class "+1" and class "−1", with the increase in ρ, respectively.*

Proof. As with the proof of Corollary A.6, when there is only one intersection x^* between the two probability density functions, then the optimal classifier is $x = x^*$. If x^* moves right, the errors of classes "+1" and "−1" will be increased and decreased, respectively. It is easy to know that x^* moves right as ρ increases. Fig. A-18 shows the variation of the position of the optimal classification boundary. When samples in class "+1" are anti-adversarially perturbed and those in class "−1" are adversarially perturbed, the intersection of two probability density functions moves right, as ρ increases, which indicates that errors of class "+1" and class "−1" will be increased and decreased, respectively. Thus, the performance gap between classes can be tuned with different ρ values. In addition, combining adversaries and anti-

adversaries can contribute to larger scope of the classification boundary than that of adversarial training with both equal and unequal bounds as stated before. Thus, a better tradeoff between the accuracy and robustness of the model can be achieved by the combination strategy. □

Supplementary Material for Section 4 (Methodology)

Extraction of Training Characteristics ($\zeta_{\boldsymbol{x}}$)

During the training process, the perturbation direction of each training sample $\boldsymbol{x}$ should be determined by its training characteristics. A total of six characteristics $(\zeta_{\boldsymbol{x},1}, \zeta_{\boldsymbol{x},2}, \cdots, \zeta_{\boldsymbol{x},6})$ for sample $\boldsymbol{x}$ are considered and extracted.

(1) Loss ($\zeta_{\boldsymbol{x},1}$) is the most widely used factor to reflect the training behavior of samples (Shu et al. 2019).

(2) Margin ($\zeta_{\boldsymbol{x},2}$) refers to the distance from the sample to the classification boundary (Zhang et al. 2021), which is always used to measure the learning difficulty of samples. It is calculated by

$$\zeta_{\boldsymbol{x},2} = f_{\boldsymbol{W}}(\boldsymbol{x})_{y_{\boldsymbol{x}}} - \max_{j \neq y_{\boldsymbol{x}}}(f_{\boldsymbol{W}}(\boldsymbol{x})_j), \tag{A.114}$$

where $f_{\boldsymbol{W}}(\boldsymbol{x})$ is the output of the classifier after Softmax.

(3)The norm of loss gradient ($\zeta_{\boldsymbol{x},3}$) of $f_{\boldsymbol{W}}(\boldsymbol{x})$ is another commonly used characteristic (Santiago et al. 2021). As the Cross-Entropy loss is adopted, it can be calculated by

$$\zeta_{\boldsymbol{x},3} = ||\boldsymbol{y}_{\boldsymbol{x}} - f_{\boldsymbol{W}}(\boldsymbol{x})||_2, \tag{A.115}$$

where $\boldsymbol{y}_{\boldsymbol{x}}$ is the one-hot label vector of sample $\boldsymbol{x}$.

(4) Information entropy ($\zeta_{\boldsymbol{x},4}$) of $f_{\boldsymbol{W}}(\boldsymbol{x})$ is used to measure the uncertainty of training samples (Wang 2008). Its calculation is

$$\zeta_{\boldsymbol{x},4} = -\sum_{j=1}^{|\mathcal{Y}|} f_{\boldsymbol{W}}(\boldsymbol{x})_j \log_2(f_{\boldsymbol{W}}(\boldsymbol{x})_j), \tag{A.116}$$

where $\mathcal{Y}$ refers to the label set, and $f_{\boldsymbol{W}}(\boldsymbol{x})$ is the output of the classifier after Softmax.

(5) Class proportion ($\zeta_{\boldsymbol{x},5}$) is commonly used to handle imbalanced class distribution (Cui et al. 2019). Its calculation is

$$\zeta_{\boldsymbol{x},5} = N_{y_{\boldsymbol{x}}}/N, \tag{A.117}$$

where $N_{y_{\boldsymbol{x}}}$ and N are the numbers of samples in class $y_{\boldsymbol{x}}$ and in the entire training set, respectively.

(6) Average loss of each category ($\zeta_{\boldsymbol{x},6}$) is another class-level factor indicating the average learning difficulty of samples in each class. Its calculation is

$$\zeta_{\boldsymbol{x},6} = \bar{\ell}_{y_{\boldsymbol{x}}}, \tag{A.118}$$

where $\bar{\ell}_{y_{\boldsymbol{x}}}$ is the average loss of samples in class $y_{\boldsymbol{x}}$.

The above six characteristics will be input into the weighting network to generate the weights of the adversary and anti-adversary of each sample.

Training with Meta Learning

Here, we show the complete formulas which consider the regulation terms introduced by the two fairness constraints. The regularization terms introduced by fairness constraints are denoted as $\gamma(\cdot)$. First, $\boldsymbol{\Omega}$ is treated as the to-be-updated parameter, and the parameter of the updated classifier $\boldsymbol{W}$, is formulated. The updating of $\boldsymbol{W}$ can be formulated as

$$\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega}) = \boldsymbol{W}^{(t)} - \eta_1 \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{W}}\{\alpha_i[\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i), y_i) + \lambda\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i), f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{adv}}))] + \beta_i\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{at-adv}}), y_i) + \gamma(f_{\boldsymbol{W}}(\boldsymbol{x}_i), f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{adv}}))\}|_{\boldsymbol{W}^{(t)}}, \tag{A.119}$$

where η_1 is the step size. After receiving the feedback of the classifier network, the parameter of the weighting network $\boldsymbol{\Omega}$ can be updated on a mini-batch of meta data as follows:

$$\begin{aligned}\boldsymbol{\Omega}^{(t+1)} = \boldsymbol{\Omega}^{(t)} - \eta_2 \frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{\Omega}}[&\ell^{\text{meta}}(f_{\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega})}(\boldsymbol{x}_i), y_i) \\ &+ \lambda\ell^{\text{meta}}(f_{\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega})}(\boldsymbol{x}_i), f_{\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega})}(\boldsymbol{x}_i^{\text{adv}})) + \ell^{\text{meta}}(f_{\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega})}(\boldsymbol{x}_i^{\text{at-adv}}), y_i) \\ &+ \gamma(f_{\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega})}(\boldsymbol{x}_i), f_{\hat{\boldsymbol{W}}^{(t)}(\boldsymbol{\Omega})}(\boldsymbol{x}_i^{\text{adv}}))]|_{\boldsymbol{\Omega}^{(t)}},\end{aligned} \tag{A.120}$$

where m and η_2 are the mini-batch size of meta data and the step size, respectively. Then, by fixing the parameters of the weighting network as $\boldsymbol{\Omega}^{(t+1)}$, the parameters of the classifier network are finally updated with the obtained weights:

$$\boldsymbol{W}^{(t+1)} = \boldsymbol{W}^{(t)} - \eta_1 \frac{1}{n}\sum_{i=1}^{n} \nabla_{\boldsymbol{W}}\{\alpha_i[\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i), y_i) + \lambda\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i), f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{adv}}))] + \beta_i\ell(f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{at-adv}}), y_i) + \gamma(f_{\boldsymbol{W}}(\boldsymbol{x}_i), f_{\boldsymbol{W}}(\boldsymbol{x}_i^{\text{adv}}))\}|_{\boldsymbol{W}^{(t)}}. \tag{A.121}$$

Supplementary Material for Section 5 (Experiments)

In this section, we present more experimental results and discussions of some typical scenarios for our method. The results verify the effectiveness of our method and the rationality of our theoretical findings. Our code is placed in the supplements.

Further Results on Standard Dataset

The comparison results of Wide-ResNet28-10 (WRN28-10) on standard CIFAR10 and SVHN datasets are shown in Tables A-1 and A-2.

The results indicate that CAAT decreases the natural and robust errors, as well as the worst intraclass errors. Although TRADES ($1/\lambda = 6$) achieves the lowest average boundary error, its average natural and robust errors are high. Nevertheless, CAAT achieves a better tradeoff between the accuracy and robustness of the model and enhances the fairness of the model. Thus, combining adversaries and anti-adversaries in training attains better generalization, robustness, and fairness of the robust model compared with other methods.

Further Results of Noisy Classification

The compared results on CIFAR10 with 20% and 40% pair-flip and uniform noise are reported in Tables A-3, A-4, A-5, and A-6. The PreAct-ResNet18 model is utilized.

The experimental results show that our methods with the two types of varied bounds reduce the average and the worstclass natural and robust errors under different degrees

	Avg. Nat.	Worst Nat.	Avg. Bdy.	Worst Bdy.	Avg. Rob.	Worst Rob.
PGD Adv. Training	14.0	29.3	38.1	53.0	52.2	78.8
TRADES($1/\lambda = 1$)	12.6	25.2	40.2	58.7	52.8	76.7
TRADES($1/\lambda = 6$)	15.5	29.1	**31.8**	45.7	47.3	71.8
Baseline ReWeight	14.2	26.3	38.6	53.7	52.8	77.9
FRL(ReWeight)	14.5	23.2	40.0	53.3	54.4	76.8
FRL(ReMargin)	15.4	24.9	38.1	49.6	53.5	70.5
FRL(ReWeight+ReMargin)	15.4	25.0	37.8	46.7	53.2	67.1
CAAT (Grad-Based)	**12.4**	**22.7**	34.5	45.5	46.9	**65.0**
CAAT (ReMargin)	13.1	23.1	32.4	**42.8**	**45.5**	65.3

Table A-1: Average and worstclass natural, boundary, and robust errors (%) for various algorithms of WRN28-10 on CIFAR10.

	Avg. Nat.	Worst Nat.	Avg. Bdy.	Worst Bdy.	Avg. Rob.	Worst Rob.
PGD Adv. Training	8.1	16.8	38.5	57.3	46.7	71.2
TRADES($1/\lambda = 1$)	8.0	19.6	40.1	60.0	48.1	73.3
TRADES($1/\lambda = 6$)	10.6	23.1	**32.1**	52.5	**42.7**	70.6
Baseline ReWeight	8.5	16.2	40.3	57.8	48.8	71.1
FRL(ReWeight)	7.8	13.4	38.9	56.9	46.7	70.7
FRL(ReMargin)	8.4	13.4	40.8	52.1	49.2	65.5
FRL(ReWeight+ReMargin)	8.4	13.2	38.4	52.1	46.8	63.1
CAAT (Grad-Based)	7.3	12.1	37.4	**46.4**	44.7	59.5
CAAT (ReMargin)	**6.2**	**11.7**	36.6	48.8	42.8	**56.0**

Table A-2: Average and worstclass natural, boundary, and robust errors (%) for various algorithms of WRN28-10 on SVHN.

on noisy data. For CIFAR10 with 20% and 40% pair-flip noise, our method can decrease the average robust error up to 9% and 12% compared with other methods. For CIFAR10 with 20% and 40% uniform noise, our method can decrease the average robust error up to 6% and 16% compared with other methods. FRL (Xu et al. 2021) can not effectively decrease the average and worst boundary and robust errors on noisy data. Compared with FRL (Xu et al. 2021), our method effectively reduces the worst intraclass boundary and robust errors. Thus, it achieves better fairness among different classes. Therefore, our method achieves the best generalization, robustness, and fairness, which makes it superior to comparable methods. Figs. A-19 (a) - (d) indicate the natural and robust errors of each class on CIFAR10 with 20% and 40% pair-flip noise. Figs. A-20 (a) - (d) indicate the natural and robust errors of each class on CIFAR10 with 20% and 40% uniform noise. As we can see, our method decreases the robust and the natural errors of most categories. The gaps between the largest and smallest errors are also decreased, which means that our method achieves the best fairness compared with other methods.

Fig. A-21 (a) indicates the ratio of the adversaries of clean and noisy samples on CIFAR10 with 40% pair-flip noise. The ratio of adversaries for clean samples is considerably higher than that for noisy samples. As shown in Fig. A-21 (b), the average adversarial bound for clean samples is large. Thus, clean samples play a more important role in model training than noisy samples. Fig. A-21 (b) indicates that the average anti-adversarial perturbation bound of noisy samples is large. Thus, the negative influence of noisy samples on the model can be decreased as noisy samples do the highest degree of anti-adversarial training. Therefore, CAAT performs well on noisy data.

Further Results of Imbalanced Classification

The comparison results on CIFAR10 with imbalance factor 100 are shown in Table A-7. The PreAct-ResNet18 model is utilized.

We can see that our methods with two types of varied bounds reduce the average and the worstclass natural and robust errors under different degrees on CIFAR10 with imbalance factor 100. CAAT remarkably decreases the average and worstclass robust errors. Thus, our method achieves a better tradeoff between the accuracy and robustness of the model. Compared with other methods, CAAT decreases the average and worst robust errors up to 20% and 17%, respectively. As with the results on noisy data, FRL can not effectively decrease the average and worst boundary and robust errors on imbalanced data. Compared with FRL (Xu et al. 2021), our method remarkably decreases the worst natural, boundary, and robust errors. Thus, our method achieves better fairness among different categories. In conclusion, our method performs well on imbalanced data. Figs. A-21 (c) and (d) illustrate the ratio of adversaries and the average perturbation bound of each class during the training process. The first head class with the largest number of samples has the lowest ratio of adversaries, while the tail classes have a high proportion of adversaries which is in accordance with our theoretical analysis. In addition, the head and tail categories have a large perturbation bound, which means that the head class does the greatest degree of anti-adversarial perturbation, and the tail classes do the greatest degree of

	Avg. Nat.	Worst Nat.	Avg. Bdy.	Worst Bdy.	Avg. Rob.	Worst Rob.
PGD Adv. Training	15.6	34.6	37.1	52.5	52.8	81.0
TRADES ($1/\lambda = 1$)	15.6	36.6	31.0	54.2	46.5	61.7
TRADES ($1/\lambda = 6$)	16.4	34.2	21.0	38.9	37.4	59.9
FRL (ReWeight)	15.3	31.6	36.0	50.7	51.4	61.4
FRL (ReMargin)	15.2	31.3	36.0	51.9	51.1	79.1
FRL (ReWeight+ReMargin)	15.7	31.7	34.3	48.2	50.0	59.1
CAAT (Grad-Based)	**14.6**	**25.2**	**13.9**	24.5	**28.5**	**45.4**
CAAT (ReMargin)	14.7	30.6	14.7	**24.0**	29.4	52.3

Table A-3: Average and worstclass natural, boundary, and robust errors (%) for various algorithms on CIFAR10 with 20% pair-flip noise.

	Avg. Nat.	Worst Nat.	Avg. Bdy.	Worst Bdy.	Avg. Rob.	Worst Rob.
PGD Adv. Training	18.3	36.3	34.7	47.2	52.0	77.8
TRADES($1/\lambda = 1$)	18.5	37.1	25.6	45.5	44.2	62.9
TRADES($1/\lambda = 6$)	18.3	35.2	27.4	51.9	45.7	65.3
FRL(ReWeight)	18.1	33.6	36.4	52.2	54.5	78.7
FRL(ReMargin)	18.4	38.7	37.0	51.9	55.4	83.6
FRL(ReWeight+ReMargin)	18.4	30.8	36.3	49.3	54.7	74.6
CAAT (Grad-Based)	**17.0**	**29.1**	15.2	**22.1**	**32.2**	**51.2**
CAAT (ReMargin)	17.3	29.6	**15.1**	22.8	32.4	**51.7**

Table A-4: Average and worstclass natural, boundary, and robust errors (%) for various algorithms on CIFAR10 with 40% pair-flip noise.

adversarial perturbation.

Further Results of Ablation Studies

In this subsection, we show more detailed results of our ablation studies. Four variations of our method are considered, including adversarial training with the same perturbation direction and bound (Setting I), adversarial training with the same perturbation direction and different bounds (Setting II), adversarial training with different perturbation directions (adversaries and anti-adversaries) and the same bound (Setting III), and adversarial training with different perturbation directions and bounds (Setting IV). The PreAct-ResNet18 model is utilized in this experiment. The average and worst natural, boundary, and robust errors are all presented, which are shown in Table A-8.

Settings III and IV obtain better performances compared with those of Settings I and II. Thus, combining adversaries with anti-adversaries in training is more effective than using only adversaries. Compared with Setting III, Setting IV further decreases the average and worst natural errors and the worst intraclass robust error, indicating that the varied bound is more effective in some cases.

References

Azzalini, A. 1985. A Class of Distributions Which Includes the Normal Ones. *Scandinavian Journal of Statistics*, 12: 171–178.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *CVPR*, 9260–9270.

Santiago, C.; Barata, C.; Sasdelli, M.; Carneiro, G.; and C.Nascimentoa, J. 2021. LOW: Training Deep Neural Networks by Learning Optimal Sample Weights. *Pattern Recognition*, 110: 130–141.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-Weight-Net: Learning an Explicit Mapping For Sample Weighting. In *NeurIPS*, 1917–1928.

Wang, Q. A. 2008. Probability distribution and entropy as a measure of uncertainty. *Journal of Physics A: Mathematical and Theoretical*, 41: 1–8.

Xu, H.; Liu, X.; Li, Y.; Jain, A. K.; and Tang, J. 2021. To be Robust or to be Fair: Towards Fairness in Adversarial Training. In *ICML*, 11492–11501.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *ICML*, 12907–12929.

Zhang, J.; Zhu, J.; Niu, G.; Han, B.; Sugiyama, M.; and Kankanhalli, M. 2021. Geometry-aware Instance-reweighted Adversarial Training. In *ICLR*, 1–29.

	Avg. Nat.	Worst Nat.	Avg. Bdy.	Worst Bdy.	Avg. Rob.	Worst Rob.
PGD Adv. Training	16.3	40.8	36.8	55.1	53.0	82.6
TRADES($1/\lambda = 1$)	15.7	30.5	29.4	44.8	45.1	57.7
TRADES($1/\lambda = 6$)	18.7	37.7	19.7	40.0	38.4	53.6
FRL(ReWeight)	15.0	26.0	36.3	52.7	51.3	74.6
FRL(ReMargin)	14.7	29.2	36.0	55.6	50.7	78.0
FRL(ReWeight+ReMargin)	15.3	29.4	34.1	49.8	49.5	73.9
CAAT (Grad-Based)	**14.3**	25.6	**17.4**	31.5	**31.7**	52.2
CAAT (ReMargin)	14.9	**24.0**	18.6	**27.8**	33.5	**50.6**

Table A-5: Average and worstclass natural, boundary, and robust errors (%) for various algorithms on CIFAR10 with 20% uniform noise.

	Avg. Nat.	Worst Nat.	Avg. Bdy.	Worst Bdy.	Avg. Rob.	Worst Rob.
PGD Adv. Training	17.8	35.3	32.9	51.3	50.7	74.0
TRADES($1/\lambda = 1$)	18.6	35.3	33.9	56.8	52.5	72.1
TRADES($1/\lambda = 6$)	23.6	38.8	28.7	55.8	52.3	69.7
FRL(ReWeight)	15.5	29.5	34.8	49.7	50.3	75.5
FRL(ReMargin)	15.7	32.1	33.3	50.0	49.0	75.8
FRL(ReWeight+ReMargin)	15.9	29.9	33.2	50.4	49.0	72.9
CAAT (Grad-Based)	**14.6**	**28.6**	**18.3**	**31.3**	**32.9**	**57.9**
CAAT (ReMargin)	15.1	28.9	18.8	31.8	33.9	60.4

Table A-6: Average and worstclass natural, boundary, and robust errors (%) for various algorithms on CIFAR10 with 40% uniform noise.

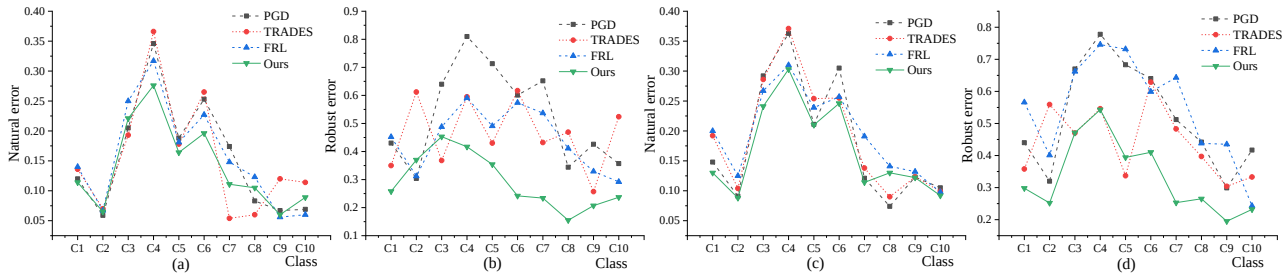

Figure A-19: (a) and (b): Natural and robust errors of four methods for each class in CIFAR10 with 20% pair-flip noise. (c) and (d): Natural and robust errors of four methods for each class in CIFAR10 with 40% pair-flip noise.

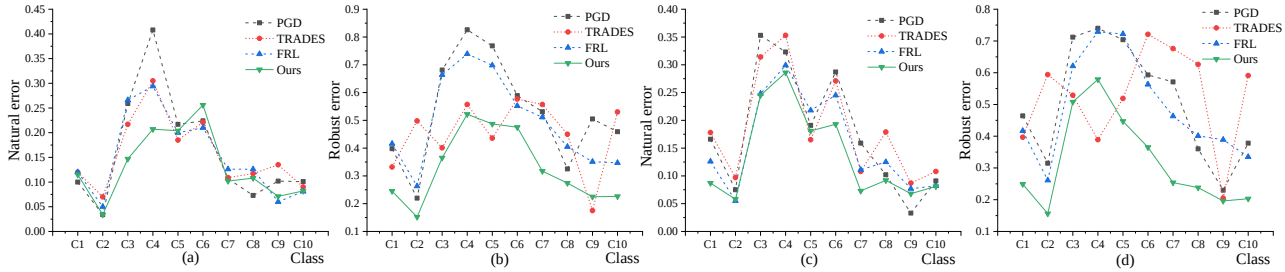

Figure A-20: (a) and (b): Natural and robust errors of four methods for each class in CIFAR10 with 20% uniform noise. (c) and (d): Natural and robust errors of four methods for each class in CIFAR10 with 40% uniform noise.

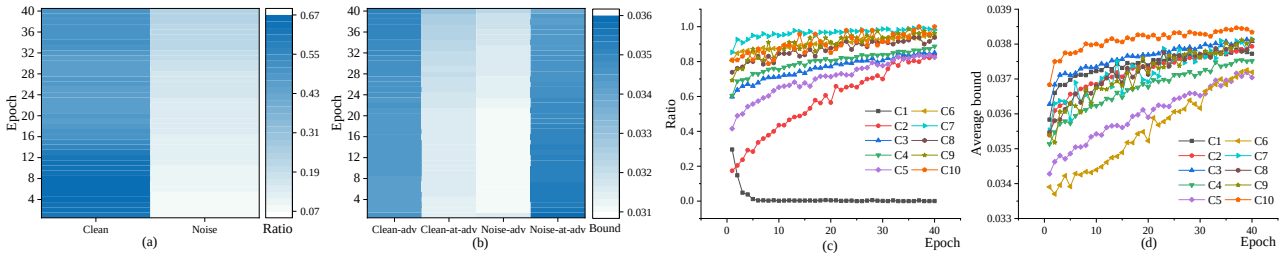

Figure A-21: (a): Ratio of adversaries of noisy and clean samples during the training procedure on CIFAR10 with 40% pair-flip noise. (b): Average adversarial and anti-adversarial perturbation bounds of clean and noisy samples during the training process on CIFAR10 with 40% pair-flip noise. (c): Ratio of adversaries in each class during training on CIFAR10 with imbalance factor 100. (d): Average perturbation bound of each class during training on CIFAR10 with imbalance factor 100. Here, "C1" to "C10" are from the first head category to the last tail category.

	Avg. Nat.	Worst Nat.	Avg. Bdy.	Worst Bdy.	Avg. Rob.	Worst Rob.
PGD Adv. Training	24.5	74.8	38.0	47.5	62.5	96.2
TRADES($1/\lambda = 1$)	30.8	68.3	29.9	64.2	60.7	83.8
TRADES($1/\lambda = 6$)	39.2	83.2	**16.7**	34.3	55.9	86.1
FRL(ReWeight)	19.8	42.6	36.9	50.8	56.7	86.9
FRL(ReMargin)	23.2	69.6	34.2	43.8	57.4	94.3
FRL(ReWeight+ReMargin)	19.2	48.3	36.8	52.5	56.0	89.4
CAAT (Grad-Based)	18.8	**39.3**	**16.8**	27.5	**35.6**	**66.8**
CAAT (ReMargin)	**18.7**	41.5	17.6	**25.8**	36.3	72.9

Table A-7: Average and worstclass natural, boundary, and robust errors (%) for various algorithms on CIFAR10 with imbalance factor 100.

	Avg. Nat.	Worst Nat.	Avg. Bdy.	Worst Bdy.	Avg. Rob.	Worst Rob.
Setting I	16.0	**22.5**	41.6	54.2	57.6	73.3
Setting II	16.9	24.7	35.0	50.8	51.9	75.2
Setting III	14.9	24.6	**13.8**	**22.0**	**28.7**	44.5
Setting IV	**13.9**	24.3	15.4	24.9	29.3	**44.4**

Table A-8: Average and worstclass natural, boundary, and robust errors (%) for four variations of CAAT on standard CIFAR10.