Outlier Robust Adversarial Training

Shu Hu HU968@PURDUE.EDU
Department of Computer and Information Technology, Purdue School of Engineering and Technology
Indiana University–Purdue University Indianapolis
Indianapolis, IN, 46202. USA

Zhenhuan Yang ZHENHUAN.YANG@HOTMAIL.COM
Etsy, Inc, Brooklyn, New York, USA

Xin Wang XWANG56@ALBANY.EDU
Department of Epidemiology and Biostatistics, School of Public Health
University at Albany, State University of New York
Albany, NY 12222, USA

Yiming Ying YYING@ALBANY.EDU
Department of Mathematics and Statistics
University at Albany, State University of New York
Albany, NY 12222, USA

Siwei Lyu SIWEILYU@BUFFALO.EDU
Department of Computer Science and Engineering
University at Buffalo, State University of New York
Buffalo, NY 14260-2500, USA

Editors: Berrin Yanıkoğlu and Wray Buntine

Abstract

Supervised learning models are challenged by the intrinsic complexities of training data such as outliers and minority subpopulations and intentional attacks at inference time with adversarial samples. While traditional robust learning methods and the recent adversarial training approaches are designed to handle each of the two challenges, to date, no work has been done to develop models that are robust with regard to the low-quality training data and the potential adversarial attack at inference time simultaneously. It is for this reason that we introduce Outlier Robust Adversarial Training (`ORAT`) in this work. `ORAT` is based on a bi-level optimization formulation of adversarial training with a robust rank-based loss function. Theoretically, we show that the learning objective of `ORAT` satisfies the $\mathcal{H}$-consistency (Awasthi et al., 2021) in binary classification, which establishes it as a proper surrogate to adversarial 0/1 loss. Furthermore, we analyze its generalization ability and provide uniform convergence rates in high probability. `ORAT` can be optimized with a simple algorithm. Experimental evaluations on three benchmark datasets demonstrate the effectiveness and robustness of `ORAT` in handling outliers and adversarial attacks. Our code is available at https://github.com/discovershu/ORAT.

Keywords: Robustness; Adversarial Training

1. Introduction

In supervised learning, we obtain a parametric model $f_\theta(\mathbf{x})$ to predict the label (discrete or continuous) y from an input $\mathbf{x}$. The optimal value of the parameter θ is obtained with a set

© 2023 S. Hu, Z. Yang, X. Wang, Y. Ying & S. Lyu.

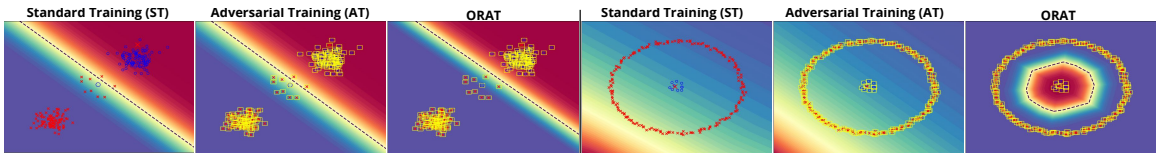

Figure 1: *Illustrative examples of standard training (ST), adversarial training (AT), and* `ORAT` *for binary classification on a balanced but multi-modal synthetic dataset (left panel) with two outliers and an imbalanced synthetic dataset (right panel) with one outlier. outliers in the blue and red classes are shown in* $\times$ *and* $\circ$*, respectively. The yellow squares around data samples represent the samples are perturbed within a* l_∞ *ball. The dash lines are the decision boundaries. In the right figures, using ST and AT without data transfer techniques, there are no decision boundaries since both strategies cannot achieve a classifier that separates the two classes.*

of labeled training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, by minimizing a loss function. However, supervised learning algorithms are challenged by two types of data degradation. First, the quality of training data is affected by erroneous samples due to mistakes in data collection or labeling (often known as the *outliers*) or isolated sub-populations of actual samples (Sukhbaatar et al., 2015; Pu et al., 2022; Guo et al., 2022). In addition, at inference time, an input data point $\mathbf{x}$ could be intentionally modified to create an adversarial sample that misleads $f_\theta(\mathbf{x})$ to make a wrong prediction (Goodfellow et al., 2015; Hu et al., 2021).

To date, the resilience of supervised learning models with regard to unintentional non-ideal training data or intentional adversarial attacks has been studied separately in machine learning. The former is the topic of robust learning methods (Wang et al., 2018; Hu et al., 2020; Zhai et al., 2021; Hu et al., 2022), and the latter is addressed with adversarial training (AT) (Madry et al., 2018; Wang et al., 2020; Zhang et al., 2021). However, in practice, the two issues can occur in tandem. This has been noticed in Sanyal et al. (2021) and Zhu et al. (2021). They have demonstrated that the outlier problem (i.e., label noise) exists in AT and found the model performance degrades with the increase in noise level because outliers hurt the quality of training data. Only Zhu et al. (2021) provide a heuristic algorithm based on a label correction strategy to correct the noise label. However, their approach may introduce more extra noisy labels due to the imperfect classifier and cannot handle outliers with true labels that do not belong to the existing label list.

To reduce the influence of the outliers in AT, one may think of adapting the self-learning (Han et al., 2019) approach (i.e., a robust learning algorithm) to remove examples that are most likely outliers (e.g., examples with larger loss) from data before training and then conducting AT on the cleaner set. However, there are two drawbacks to this simple scheme. First, it is not an end-to-end approach, which will increase the training cost. Second, it is hard to remove outliers precisely, and an imperfect strategy may drop examples of clean data points. This may hurt the final model performance in the AT phase. An alternative solution is that combine robust learning and AT by using robust losses (e.g., Huber loss (Hastie et al., 2009), symmetric cross entropy loss (Wang et al., 2019), etc.) in AT. However, most existing robust losses cannot eliminate the influence of outliers. In addition, constructing the theoretical guarantee is a challenge for using a robust loss in AT, especially for $\mathcal{H}$-calibration and $\mathcal{H}$-consistency properties (Awasthi et al., 2021; Steinwart, 2007) of the designed adversarial loss with respect to the adversarial 0/1 loss in classification.

In this work, we introduce Outlier Robust Adversarial Training (`ORAT`) to combine robust learning and adversarial training. Specifically, we develop an effective adversarial training

algorithm for a rank-based learning objective that can exclude the influence of outliers from the training procedure. Figure 1 shows several illustrative examples. The learning objective in ORAT lends itself to an efficient numerical algorithm based on gradient descent methods after reformulation that removes the explicit ranking operation. To verify whether the optimal minimizers of the ORAT loss are close to or exactly the optimal minimizers of the adversarial 0/1 loss, we show that the ORAT loss satisfies the $\mathcal{H}$-calibration and consistency properties for classification under some moderate conditions, which establishes it as a proper surrogate to adversarial 0/1 loss. The notion of consistency has been studied in Awasthi et al. (2021). However, the ORAT loss is the first adversarial surrogate loss proven to satisfy $\mathcal{H}$-consistency and evaluated on real-world datasets. It encourages the applicability of the $\mathcal{H}$-consistent adversarial surrogate loss in real tasks. We further provide a quantitative error bound of the generalization gap between the training and testing performance of ORAT. Experimental evaluations on three benchmark datasets demonstrate the effectiveness and robustness of ORAT in handling outliers and adversarial attacks. Our contributions can be summarized as follows:

- We present outlier robust adversarial training (ORAT), which can handle outliers in adversarial training jointly.
- We show ORAT loss satisfies $\mathcal{H}$-calibration and $\mathcal{H}$-consistency. To the best of our knowledge, the ORAT loss is the first $\mathcal{H}$-consistent adversarial surrogate loss that is evaluated on real-world datasets.
- We provide a detailed theoretical analysis on the generalization error of training with ORAT loss.

2. Background

Robust Learning. Training accurate machine learning models in the presence of outliers is of great practical importance. To combat outliers, the traditional methods are designed based on label correction (Wang et al., 2018), loss correction (Han et al., 2020), and refined training strategies (Yu et al., 2019). However, they require an extra clean dataset or potentially expensive detection process to estimate the outliers. Recent works Hu et al. (2020, 2023) proposed the average of ranked range (AoRR) loss, which can eliminate the influence of the outliers if their proportion in training data is known. Let $\mathcal{X}$ denote the input feature space, $\mathbf{x} \in \mathcal{X}$ is a training sample, $y \in \mathcal{Y} = \{1, \cdots, C\}$ is its associated label, and $C \geq 2$. $f_\theta(\cdot) : \mathcal{X} \to \mathbb{R}^C$ is the logits output of the predictor, and $\ell : \mathbb{R}^C \times \mathcal{Y} \to \mathbb{R}$ is a loss function. $\ell_{[i]}(f_\theta(\mathbf{x}_j), y_j)_{j=1}^n$ represents the i-th largest loss among the training sample set. For two integers k and m, $0 \leq m < k \leq n$, the AoRR loss is defined as follows,

$$\min_\theta \quad \frac{1}{k-m} \sum_{i=m+1}^{k} \ell_{[i]}(f_\theta(\mathbf{x}_j), y_j)_{j=1}^n. \tag{1}$$

The AoRR loss excludes training samples with the top m-largest loss value, as well as samples with small losses. This is to reduce the influence of the outliers (larger losses) and to enhance the effect of the minority subgroup of the data because the samples with small loss values are most likely from the majority subgroup in the training set.

Adversarial Training. Recent studies have shown some surprising vulnerabilities of advanced supervised learning models, especially those based on deep neural networks, to specially designed adversarial samples (Goodfellow et al., 2015; Carlini and Wagner, 2017).

To mitigate this issue, adversarial training (AT) (Madry et al., 2018) is proposed as a training approach against adversarial attacks. Let $\mathcal{B}_\epsilon(\mathbf{x}) = \{\mathbf{x}' \in \mathcal{X} | \|\mathbf{x} - \mathbf{x}'\|_p \leq \epsilon\}$ be the closed ball of radius $\epsilon > 0$ centered at $\mathbf{x}$, where p is usually chosen as 1, 2, or ∞. Given a training dataset $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ independently drawn from a distribution $\mathcal{D}$, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. The objective function of AT is then

$$\min_\theta \frac{1}{n} \sum_{i=1}^n \Big[\max_{\tilde{\mathbf{x}}_i \in \mathcal{B}_\epsilon(\mathbf{x})} \ell(f_\theta(\tilde{\mathbf{x}}_i), y_i) \Big], \tag{2}$$

where $\tilde{\mathbf{x}}$ is the most extreme adversarial sample within the ϵ-ball centered at $\mathbf{x}$. To generate adversarial data, The original AT applies projected gradient descent (PGD) using a fixed number of iterations P as a stopping criterion, namely the PGD^P algorithm, to approximately solve the inner maximization problem of the Eq.(2). Other methods can also generate adversarial data, e.g., the fast gradient signed method (FGSM) (Goodfellow et al., 2015) and the CW attack (Carlini and Wagner, 2017).

Robust Learning under Adversarial Attacks. Currently, several works have realized the impact of label noise on adversarial training. For example, in Sanyal et al. (2021), the authors identified label noise as one of the causes for adversarial vulnerability but no defense methods are proposed to solve this problem. The work Zhu et al. (2021) empirically studies the efficacy of AT for mitigating the effect of label noise in training data. However, their proposed annotator algorithm (RAA) is based on the label correction strategy, which may introduce more extra noisy labels due to the bottleneck of the selected classifier. Furthermore, both of these methods are only considered label noise problems instead of outliers, especially the error that comes from the sample itself. Therefore, an outlier robust adversarial training method is urgently needed to fill this gap. Several works (Augustin et al., 2020; Bitterwolf et al., 2020) connect adversarial robustness to out-of-distribution (OOD) problems. Note that the notion of outliers is different from OOD points. One assumption for the OOD problem is that the training and test data distributions are mismatched. However, we do not have this assumption in this work. More related works are discussed in Appendix A.

3. Method

In `ORAT`, we combine the rank-based AoRR loss (Eq.(1)) and adversarial training (Eq.(2)) into the same framework. Let $L\left(\{(\mathbf{x}_j, y_j)\}_{j=1}^n\right) := \{\ell(f_\theta(\mathbf{x}_1), y_1), \cdots, \ell(f_\theta(\mathbf{x}_n), y_n)\}$ be the set of all individual losses on the training samples. For notational brevity, we drop the explicit dependence on the individual loss ℓ, the parametric model f_θ, but it should be clear that the set changes with the training dataset $\{(\mathbf{x}_j, y_j)\}_{j=1}^n$. In addition, we denote $\ell_{[i]}\left(\{(\mathbf{x}_j, y_j)\}_{j=1}^n\right)$ as the i-th largest individual loss after sorting the elements in set $L\left(\{(\mathbf{x}_j, y_j)\}_{j=1}^n\right)$ (ties can be broken in any consistent way). With these definitions, we define the learning objective of `ORAT` as a bi-level optimization problem:

$$\min_\theta \frac{1}{k-m} \sum_{i=m+1}^k \ell(f_\theta(\tilde{\mathbf{x}}_{[i]}), y_{[i]}), \text{ s.t. } (\tilde{\mathbf{x}}_{[i]}, y_{[i]}) = \text{argmax}_{\tilde{\mathbf{x}}_j \in \mathcal{B}_\epsilon(\mathbf{x}_j)} \ell_{[i]}(\{\tilde{\mathbf{x}}_j, y_j\}_{j=1}^n), \tag{3}$$

where k and m are two integers such that $0 \leq m < k \leq n$. The notations here require some further explanation. First, $\tilde{\mathbf{x}}_j$ is the result of adversarial perturbation of an original data

point $\mathbf{x}_j$ with a perturbation strength ϵ, and $y_{[i]}$ is the corresponding label of the original $\mathbf{x}_j$. So the maximization sub-problem of Eq.(3) reads as follows. For each original training sample $(\mathbf{x}_j, y_j)$, we find the extreme adversarial input $\tilde{\mathbf{x}}_j$ within the ϵ-ball centered at $\mathbf{x}_j$. Then we calculate the individual loss of the perturbed samples $\ell(f_\theta(\tilde{\mathbf{x}}_j), y_j)$ and sort the losses to find the adversarial sample and label $(\tilde{\mathbf{x}}_{[i]}, y_{[i]})$ corresponding to the top-i individual loss. The outer minimization problem of Eq.(3) is the average of ranked range of $(m, k]$ individual losses found with such a procedure. Note that `ORAT` contains the original AT as a special case with $k = n$ and $m = 0$.

The robustness to outliers and adversarial attacks of the `ORAT` (Eq.(3)) can be more clearly understood as follows. The overall method can be viewed as a sample selection method and can filter incorrect data samples according to the top-k and top-m individual losses. Specifically, the perturbed samples with small losses are most likely come from the clean data. Therefore, it ignores m samples with a large loss (i.e., top-1 to top-m losses) during the training. On the other hand, the perturbed samples with an extremely low loss value are most likely very easy to be learned in the training procedure. They usually come from the majority group in a dataset. On the contrary, the perturbed samples from the minority subgroup can be also viewed as hard samples, which are very hard to be learned. In this case, ignoring the bottom $n - k$ would reduce the influence of the majority subgroup of data, which further prevents the effect of the imbalance data and enhance the impact of the minority subgroup of the data. Figure 1 also demonstrates this influence.

Although `ORAT` is designed by combining AoRR and AT, this combination has not been explored in the existing literature. More importantly, we will demonstrate and validate both theoretically (Section 4) and experimentally (Section 5) that this combination makes sense. To the best of our knowledge, no robust losses are combined with AT that can handle both low-quality training data (with outliers) and adversarial attacks in inference. In addition, the ρ-margin loss (Awasthi et al., 2021), a generalization of the ramp loss, is proved to satisfy the calibration and consistency. Inspired by this loss, we naturally thought that the truncated robust loss could be used to design a new adversarial surrogate loss that satisfies $\mathcal{H}$-calibration and consistency. Therefore, we choose AoRR loss to create `ORAT` because it is a well-defined truncated loss. The ranking operation in `ORAT` is the main obstacle to using Eq.(3) as a learning objective in an efficient way. However, we can substitute the ranking operation by introducing two auxiliary variables λ and $\hat{\lambda}$ and use the equivalent form of `ORAT` in the following result.

Theorem 1 *Denote* $[a]_+ = \max\{0, a\}$ *as the hinge function. Eq.(3) is equivalent to*

$$\begin{aligned} &\frac{1}{k-m}\min_{\theta,\lambda}\max_{\hat{\lambda}} \sum_{i=1}^{n}\hat{\mathcal{L}}(f_\theta,\lambda,\hat{\lambda}) := \Big[\frac{k-m}{n}\lambda + \frac{n-m}{n}\hat{\lambda} - [\hat{\lambda} - [\ell(f_\theta(\tilde{\mathbf{x}}_i), y_i) - \lambda]_+]_+\Big], \\ &\text{s.t.}\quad \tilde{\mathbf{x}}_i = \arg\max_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x}_i)} \ell(f_\theta(\tilde{\mathbf{x}}), y_i). \end{aligned} \tag{4}$$

Furthermore, $\hat{\lambda} > \lambda$*, when the optimal solution is achieved.*

The proof of Theorem 1 can be found in Appendix C.1. Using Theorem 1, we can develop a learning algorithm based on stochastic (mini-batch) gradient descent to optimize Eq.(3). Specifically, with initial choice for the values of $\theta^{(0)}$, $\lambda^{(0)}$, and $\hat{\lambda}^{(0)}$, at the t-th iteration,

Algorithm 1 Outlier Robust Adversarial Training

Input: A training dataset $\mathcal{S}$ of size n

Output: A robust model with parameters θ^*

Initialization: $\theta^{(0)}$, $\lambda^{(0)}$, $\hat{\lambda}^{(0)}$, t=0, η, k, m, ϵ, α, and P
for $e = 1$ *to num_epoch* **do**
  for $b = 1$ *to num_batch* **do**
    Sample a mini-batch $\mathcal{S}_b = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{S}_b|}$ from $\mathcal{S}$
    for $i = 1$ *to batch_size* **do**
      $\tilde{\mathbf{x}}_i \leftarrow \mathbf{x}_i$
      while $P > 0$ **do**
        $\tilde{\mathbf{x}}_i = \Pi_{\mathcal{B}_\epsilon(\mathbf{x}_i)}(\tilde{\mathbf{x}}_i + \alpha \texttt{sign}(\nabla_{\tilde{\mathbf{x}}_i} \ell(f_{\theta^{(t)}}(\tilde{\mathbf{x}}_i), y_i)))$
        $P \leftarrow P - 1$
      end
    end
    $\theta^{(t+1)} \leftarrow \theta^{(t)} - \frac{\eta}{|\mathcal{S}_b|} \sum_{i \in \mathcal{S}_b} \partial_\theta \hat{\mathcal{L}}(f_{\theta^{(t)}}, \lambda^{(t)}, \hat{\lambda}^{(t)})$
    $\lambda^{(t+1)} \leftarrow \lambda^{(t)} - \frac{\eta}{|\mathcal{S}_b|} \sum_{i \in \mathcal{S}_b} \partial_\lambda \hat{\mathcal{L}}(f_{\theta^{(t)}}, \lambda^{(t)}, \hat{\lambda}^{(t)})$
    $\hat{\lambda}^{(t+1)} \leftarrow \hat{\lambda}^{(t)} + \frac{\eta}{|\mathcal{S}_b|} \sum_{i \in \mathcal{S}_b} \partial_{\hat{\lambda}} \hat{\mathcal{L}}(f_{\theta^{(t)}}, \lambda^{(t)}, \hat{\lambda}^{(t)})$
    $t \leftarrow t + 1$
  end
end

a mini-batch set $\mathcal{S}_b$ of training samples is chosen uniformly at random from the training set and used to estimate the (sub)gradient of the objective. Since the optimal solution of parameters of θ, λ, and $\hat{\lambda}$ do not depend on the factor of $\frac{1}{k-m}$, we can replace it to $\frac{1}{|\mathcal{S}_b|}$ for mini-batch optimization, where $|\mathcal{S}_b|$ is the size of $\mathcal{S}_b$. Following the original AT (Madry et al., 2018), we use the projected gradient descent approach to approximately solve the constraint. $\Pi_{\mathcal{B}_\epsilon(\mathbf{x}_i)}(\cdot)$ is the projection function that projects the adversarial data back into the ϵ-ball centered at $\mathbf{x}_i$ if necessary. $\partial_\theta \hat{\mathcal{L}}$, $\partial_\lambda \hat{\mathcal{L}}$, and $\partial_{\hat{\lambda}} \hat{\mathcal{L}}$ are the (sub)gradients of $\hat{\mathcal{L}}$ with respect to θ, λ, and $\hat{\lambda}$, respectively. Their explicit forms can be found in Appendix B. The pseudo-code of optimizing `ORAT` is described in Algorithm 1. Note that our optimization process is similar to the traditional adversarial training algorithms (Madry et al., 2018; Zhang et al., 2021; Liu et al., 2021) with one additional minimization with respect to λ and one additional maximization for $\hat{\lambda}$. Therefore, our algorithm has the same time complexity to original AT. We will show that our algorithm can converge in experiments.

4. Analysis

In this section, we prove that `ORAT` is a $\mathcal{H}$-consistent adversarial surrogate loss and study its generalization property. Subsequently, we focus on the case of binary classification ($\mathcal{Y}$=$\{\pm 1\}$) with margin-based loss function, *i.e.*, $\ell(yf(\mathbf{x}))$=$\ell(f(\mathbf{x}), y)$=$\ell(f, \mathbf{x}, y)$, where we omit model parameter θ for simplicity. We first introduce several notations. Let $\mathcal{H}$ be a Hypothesis function set from $\mathbb{R}^d$ to $\mathbb{R}$. We say $\mathcal{H}$ is symmetric, if for any $f \in \mathcal{H}$, $-f$ is also in $\mathcal{H}$. The 0/1 risk of a classifier $f \in \mathcal{H}$ is $\mathcal{R}_{\ell_0}$=$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell_0(f(\mathbf{x}), y)]$, where $\ell_0(f(\mathbf{x}), y)$=$\mathbb{1}_{yf(\mathbf{x})\leq 0}$

is the 0/1 loss. Denote the adversarial 0/1 risk as $\mathcal{R}_{\widetilde{\ell}_0}(f)=\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\widetilde{\ell}_0(f(\mathbf{x}),y)]$, where $\widetilde{\ell}_0(f(\mathbf{x}),y)=\sup_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})}\mathbb{1}_{yf(\tilde{\mathbf{x}})\leq 0}$ is the adversarial 0/1 loss. We also define the ℓ_s-risk of f for a surrogate loss $\ell_s(f(\mathbf{x}),y)$ as $\mathcal{R}_{\ell_s}(f)=\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell_s(f(\mathbf{x}),y)]$ and the minimal $(\ell_s,\mathcal{H})$-risk, which is defined by $\mathcal{R}^*_{\ell_s,\mathcal{H}}=\inf_{f\in\mathcal{H}}\mathcal{R}_{\ell_s}(f)$.

4.1. Classification Calibration and Consistency

Designing a robust algorithm entails using an appropriate surrogate loss to the standard 0/1 loss since 0/1 loss is very hard to optimize for most hypothesis sets (Awasthi et al., 2021; Bao et al., 2020). We first provide a definition of $\mathcal{H}$-consistency.

Definition 2 *($\mathcal{H}$-Consistency). (Awasthi et al., 2021) Given a hypothesis set $\mathcal{H}$, we say that a loss function ℓ_1 is $\mathcal{H}$-consistent w.r.t. a loss function ℓ_2, if, for all probability distributions and sequences of $\{f_n\}_{n\in\mathbb{N}}\subset\mathcal{H}$, there holds $\mathcal{R}_{\ell_1}(f_n)-\mathcal{R}^*_{\ell_1,\mathcal{H}}\xrightarrow{n\to+\infty}0\Rightarrow\mathcal{R}_{\ell_2}(f_n)-\mathcal{R}^*_{\ell_2,\mathcal{H}}\xrightarrow{n\to+\infty}0$.*

$\mathcal{H}$-consistency guarantees that the optimal minimizers of the surrogate adversarial loss are equal to or near the optimal minimizers of the 0/1 adversarial loss on a restricted hypothesis set $\mathcal{H}$.

For a distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ with random variables X and Y. Let $\eta:=\Pr(Y=1|X=\mathbf{x})\in[0,1]$, for any $\mathbf{x}\in\mathcal{X}$. Using conditional expectation, we can rewrite $\mathcal{R}_{\ell_s}(f)$ as $\mathcal{R}_{\ell_s}(f)=\mathbb{E}_X[\mathcal{C}_{\ell_s}(f,\mathbf{x},\eta)]$, where $\mathcal{C}_{\ell_s}(f,\mathbf{x},\eta):=\eta\ell_s(f,\mathbf{x},1)+(1-\eta)\ell_s(f,\mathbf{x},-1),\forall\mathbf{x}\in\mathcal{X}$. Furthermore, the minimal inner ℓ_s-risk on $\mathcal{H}$ is denoted by $\mathcal{C}^*_{\ell_s,\mathcal{H}}(\mathbf{x},\eta):=\inf_{f\in\mathcal{H}}\mathcal{C}_{\ell_s}(f,\mathbf{x},\eta)$. We now define $\mathcal{H}$-calibration.

Definition 3 *($\mathcal{H}$-Calibration). (Steinwart, 2007; Awasthi et al., 2021) Given a hypothesis set $\mathcal{H}$, we say that a loss function ℓ_1 is $\mathcal{H}$-calibrated with respect to a loss function ℓ_2 if, for any $\tau>0$, $\eta\in[0,1]$, and $\mathbf{x}\in\mathcal{X}$, there exists $\delta>0$ such that, for all $f\in\mathcal{H}$, $\mathcal{C}_{\ell_1}(f,\mathbf{x},\eta)<\mathcal{C}^*_{\ell_1,\mathcal{H}}(\mathbf{x},\eta)+\delta\Rightarrow\mathcal{C}_{\ell_2}(f,\mathbf{x},\eta)<\mathcal{C}^*_{\ell_2,\mathcal{H}}(\mathbf{x},\eta)+\tau$.*

As shown in (Steinwart, 2007), if ℓ_1 is $\mathcal{H}$-calibrated with respect to ℓ_2, then $\mathcal{H}$-consistency holds for any probability distribution verifying the additional conditions, which will be further reflected in the following Theorem 4.

Without considering the adversarial perturbations and omitting the parameter θ, the population version of the objective function of Eq.(4) can be formulated as follows:

$$\min_{\lambda}\max_{\hat{\lambda}}\frac{1}{k-m}\sum_{i=1}^{n}\hat{\mathcal{L}}(f,\lambda,\hat{\lambda})\xrightarrow[n\to\infty]{\frac{k-m}{n}\to\nu,\frac{m}{n}\to\mu}\frac{1}{\nu}\min_{\lambda}\max_{\hat{\lambda}}\mathbb{E}\Big[\hat{\lambda}-[\hat{\lambda}-[\ell(Yf(X))-\lambda]_+]_+\Big]+\nu\lambda-\mu\hat{\lambda}.$$

Throughout the paper, we assume that $\mu>0$ since if $\mu=0$ then it will lead to $\hat{\lambda}=\infty$. As such, the population version of our `ORAT` loss is given by $(f_0^*,\lambda^*,\hat{\lambda}^*)=\arg\inf_{f,\lambda}\sup_{\hat{\lambda}}\Big\{\mathbb{E}\Big[\hat{\lambda}-[\hat{\lambda}-[\ell(Yf(X))-\lambda]_+]_+\Big]+\nu\lambda-\mu\hat{\lambda}\Big\}$. It is difficult to directly work with the optima f_0^* since the above problem is a non-convex minmax problem and the standard minmax theorem does not apply here. Instead, we assume the existence of λ^* and $\hat{\lambda}^*$ and work with the minimizer $f^*=\arg\inf_f\mathcal{L}(f,\lambda^*,\hat{\lambda}^*)$ where $\mathcal{L}(f,\lambda^*,\hat{\lambda}^*):=\mathbb{E}\Big[\hat{\lambda}^*-[\hat{\lambda}^*-[\ell(Yf(X))-\lambda^*]_+]_+\Big]+\nu\lambda^*-\mu\hat{\lambda}^*$. Since the term $\nu\lambda^*-\mu\hat{\lambda}^*$ does not depend on f, we have $f^*=\arg\inf_f\mathbb{E}\Big[\hat{\lambda}^*-[\hat{\lambda}^*-[\ell(Yf(X))-$

$\lambda^*]_+]_+\Big]$. From the above observations, we denote by $\phi_{\texttt{ORAT}}$ and $\tilde{\phi}_{\texttt{ORAT}}$ for ORAT loss without and with the adversarial perturbation, respectively, as follows:

$$\phi_{\texttt{ORAT}}(t) = \hat{\lambda}^* - [\hat{\lambda}^* - [\ell(t) - \lambda^*]_+]_+, \quad \tilde{\phi}_{\texttt{ORAT}}(f, \mathbf{x}, y) = \sup_{\tilde{\mathbf{x}} \in \mathcal{B}_\epsilon(\mathbf{x})} \phi_{\texttt{ORAT}}(yf(\tilde{\mathbf{x}})). \tag{5}$$

We can then obtain the following theorem. Its proof can be found in Appendix C.2.

Theorem 4 *Let $\mathcal{H}$ be a symmetric hypothesis set consisting of the family of all measurable functions $\mathcal{H}_{all}$, suppose $\nu > \min\{\hat{\lambda}^*, \mathcal{R}^*_{\ell,\mathcal{H}}\}$, $0 \le \lambda^* < \hat{\lambda}^*$, λ^* and $\hat{\lambda}^*$ are bounded, and ℓ is a non-negative, continuous, and non-increasing margin-based loss.*

(i) Then $\tilde{\phi}_{ORAT}$ is $\mathcal{H}$-calibrated with respect to $\widetilde{\ell}_0$.

(ii) Furthermore, $\tilde{\phi}_{ORAT}$ is $\mathcal{H}$-consistent with respect to $\widetilde{\ell}_0$ for all distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that satisfy: $\mathcal{R}^*_{\widetilde{\ell}_0,\mathcal{H}} = 0$ and there exists $f^* \in \mathcal{H}$ such that $\mathcal{R}_{\phi_{ORAT}}(f^*) = \mathcal{R}^*_{\phi_{ORAT},\mathcal{H}_{all}} < +\infty$.

$\mathcal{H}$ can be linear models or deep neural networks. The commonly used hinge loss, logistic loss, and cross-entropy loss all satisfy the conditions of ℓ (See Appendix C.3 for details). Furthermore, according to Theorem 1 and assuming $\ell \ge 0$, we have $0 \le \lambda^* < \hat{\lambda}^*$. For ν, it should be larger than the smaller value among $\hat{\lambda}^*$ and minimal (ℓ, $\mathcal{H}$)-risk, which means k and m should be as far away from each other as possible. We call $\mathcal{R}^*_{\widetilde{\ell}_0,\mathcal{H}} = 0$ as the realizability condition. Therefore, we can conclude that ORAT loss satisfies $\mathcal{H}$-consistency. Our proof is mainly inspired by Awasthi et al. (2021). However, our analysis for the $\tilde{\phi}_{\texttt{ORAT}}$ adversarial surrogate loss is more involved since it is a composition function based on ℓ for which we need to consider the different conditions of λ^* and $\hat{\lambda}^*$.

4.2. Generalization Error

In this subsection, we present the generalization error bound for the proposed ORAT loss. Define the adversarial surrogate loss $\widetilde{\ell}(yf(\mathbf{x})) = \max_{\tilde{\mathbf{x}} \in \mathcal{B}_\epsilon(\mathbf{x})} \ell(yf(\tilde{\mathbf{x}}))$ and the composite function class $\widetilde{\ell}_{\mathcal{H}} := \{\widetilde{\ell}(yf(\mathbf{x})) : f \in \mathcal{H}\}$. The generalization error studies the discrepancy between the empirical adversarial risk $\mathcal{R}_{\widetilde{\ell}}(f; \mathcal{S})$ defined on the training data and its population risk $\mathcal{R}_{\widetilde{\ell}}(f)$ measuring the performance on the test data, where $\mathcal{R}_{\widetilde{\ell}}(f; \mathcal{S}) = \inf_\lambda \sup_{\hat{\lambda}} \frac{1}{k-m} \sum_{i=1}^n \hat{\mathcal{L}}(f, \lambda, \hat{\lambda})$. First, we need the Rademacher complexity (Bartlett and Mendelson, 2002) which is defined as follows.

Definition 5 (Rademacher complexity) *For any function class $\mathcal{H}$, given a dataset $\{\mathbf{x}_i\}_{i=1}^n$, the empirical Rademacher complexity is defined as $\mathfrak{R}_n(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\sigma\Big[\sup_{f \in \mathcal{H}} \sum_{i=1}^n \sigma_i f(\mathbf{x}_i)\Big]$, where $\sigma_1, \cdots, \sigma_n$ are i.i.d. Rademacher random variables with $\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = 1/2$.*

With these preparations, we can get the following theorem.

Theorem 6 *Suppose that the range of $\ell(f(\mathbf{x}), y)$ is $[0, M]$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of an i.i.d. training dataset of size n, the following holds for all $f \in \mathcal{H}$,*

$$\mathcal{R}_{\widetilde{\ell}}(f) - \mathcal{R}_{\widetilde{\ell}}(f; \mathcal{S}) \le \frac{2}{\nu}\Big(2\mathfrak{R}_n(\widetilde{\ell}_{\mathcal{H}}) + \frac{M(2\sqrt{2} + 3\sqrt{\log(2/\delta)})}{\sqrt{2n}}\Big).$$

The proof of Theorem 6 is by rewriting the empirical risk and population risk with optimal λ^* and $\hat{\lambda}^*$, and then analytically bound the Rademacher complexity of λ^* and $\hat{\lambda}^*$ by utilizing the boundness condition. See Appendix C.4 for details. Theorem 6 characterizes uniform convergence between the training and testing on `ORAT` given hypothesis set $\mathcal{H}$. The generalization error depends on the limit of ranked range $\frac{k-m}{n} \to \nu$ as well as the Rademacher complexity of the adversarial loss function class. It is worth noting the i.i.d. assumption over the sample is restrictive but required for applying the symmetrization trick in the proof, and how to relax this assumption to a more realistic assumption, for example, $n-m$ i.i.d. inliers with m outliers (Laforgue et al., 2021), is future work for us. Nevertheless, Theorem 6 still highlights the generalization ability of our `ORAT` objective with respect to the ranked range and adversarial loss in this ideal setting. We provide hypothesis set examples of linear classifiers and neural networks in Appendix C.5 for further characterizing the Rademacher complexity $\mathfrak{R}_n(\widetilde{\ell}_{\mathcal{H}})$.

Remark 7 *Theorem 6 together with Theorem 4 indicates that learning with `ORAT` asymptotically converges to zero on the adversarial 0/1 risk. Let* $f^{**} = \arg\inf_{f\in\mathcal{H}} \mathcal{R}_{\widetilde{\ell}}(f)$ *and* $f_S^{**} = \arg\inf_{f\in\mathcal{H}} \mathcal{R}_{\widetilde{\ell}}(f;S)$. *By standard error decomposition,* $\mathcal{R}_{\widetilde{\ell}}(f_S^{**}) - \mathcal{R}_{\widetilde{\ell}}(f^{**}) = \big(\mathcal{R}_{\widetilde{\ell}}(f_S^{**}) - \mathcal{R}_{\widetilde{\ell}}(f_S^{**};S)\big) + \big(\mathcal{R}_{\widetilde{\ell}}(f_S^{**};S) - \mathcal{R}_{\widetilde{\ell}}(f^{**};S)\big) + \big(\mathcal{R}_{\widetilde{\ell}}(f^{**};S) - \mathcal{R}_{\widetilde{\ell}}(f^{**})\big)$. *The first term converges to 0 as n goes to ∞ and $\mathfrak{R}_n(\widetilde{\ell}_{\mathcal{H}})$ is bounded. The second term is always non-positive as f_S^{**} minimizes $\mathcal{R}_{\widetilde{\ell}}(f;S)$. The last term is bounded by $\mathcal{O}(1/\sqrt{n})$ with high probability by Hoeffding's inequality (Hoeffding, 1994). Therefore, $\mathcal{R}_{\widetilde{\ell}}(f_S^{**}) - \mathcal{R}_{\widetilde{\ell}}(f^{**}) \to 0$ as $n \to \infty$. Combined with Theorem 4 (ii) and Definition 2, it shows that $\mathcal{R}_{\widetilde{\ell}_0}(f_S^{**}) - \mathcal{R}_{\widetilde{\ell}_0}(f^{**}) \to 0$ as $n \to \infty$.*

5. Experiments

We evaluate the performance of the proposed `ORAT` with numerical experiments. Due to the limit of the space, we present the most significant information and results of our experiments. More detailed information, additional results, and code are in Appendix D, E, and supplementary files, respectively.

5.1. Experimental Settings

Datasets, Networks, and Baselines. Our experiments are based on three popular datasets, namely MNIST, CIFAR-10, and CIFAR-100. We follow the same training/testing splitting from the original datasets. The pixel values of all images are normalized into the range of [0,1]. We adopt LeNet (LeCun et al., 1998), Small-CNN (Wang et al., 2020), and ResNet-18 (He et al., 2016) for MNIST, CIFAR-10, and CIFAR-100, respectively. Settings of networks are in Appendix D.2. We compare `ORAT` with standard training (ST), adversarial training (AT). In addition, as there are no dedicated adversarial training methods for handling outliers, we compare three instance-reweighted AT methods GAIRAT (Zhang et al., 2021), MAIL (Liu et al., 2021), and WMMR (Zeng et al., 2021a), because they are the most recent works with the best performance against adversarial attacks. We also compare RAA (Zhu et al., 2021) since it considers label correction in training. To support our claims about the drawbacks of using a self-learning strategy to remove outliers before adversarial training in Section 1, we use the AoRR approach to remove examples with larger losses (potential

outliers) from the original training set and obtain a cleaner set for adversarial training. We call this baseline AT w/o and compare it with our method. Appendix E.5 shows more details of this baseline. All methods are designed based on the cross-entropy loss.

Defense Settings and Robustness Evaluation. In training, following Zeng et al. (2021a); Liu et al. (2021), we consider robustness by setting $p = \infty$. For each dataset, we test two different perturbation bounds ϵ, i.e., $\epsilon \in \{0.1, 0.2\}$ for MNIST and $\epsilon \in \{2/255, 8/255\}$ for CIFAR-10 and CIFAR-100. The training attack is PGD^{10} with random start and step size $\epsilon/4$. We evaluate the robustness of our method and baselines using the standard accuracy on natural test data (Natural), FGSM, PGD^{20}, and the CW attack because they are frequently used in the literature of our compared methods. We also evaluate all methods on a very strong attack, AutoAttack (AA) (Croce and Hein, 2020), under specific settings. All of them are constrained by the same perturbation limit ϵ.

Hyperparameters (k and m) and Outliers Generation. Following Hu et al. (2020), we apply a grid search to select the values of k and m. To simulate the outliers in the training dataset, as in the work of Wang et al. (2019), we add the symmetric (uniform) and asymmetric (class-dependent) noises to the labels of the training data. For symmetric noise creation, we randomly choose training samples with a given probability γ and change each of their labels to another random label. For asymmetric noise creation, given γ, flipping labels only within a specific set of classes. More details of the simulation can be found in Appendix D.4. Note that we use 4 different $\gamma \in \{10\%, 20\%, 30\%, 40\%\}$.

5.2. Results

Overall Performance. We report the overall results (accuracy on the testing sets) in Table 1. First, when there is no noise in the datasets, `ORAT` outperforms the AT method in three different attack settings on all datasets by a significant margin (0, 2.14%]. This is probably because the original datasets may contain outliers, as verified in Sanyal et al. (2021).

Second, under the symmetric noise setting, `ORAT` outperforms all compared methods in all attack scenarios in general. We observe the performance gap between noise and noise-free settings is small on MNIST because it is a relatively smaller dataset and LeNet achieves almost perfect performance. However, on a larger dataset such as CIFAR-10 and using a more complex neural network, the outliers have a stronger impact. The instance-reweighted AT methods such as GAIRAT, MAIL, and MAIL are inferior to AT and `ORAT` in 20%, 30%, 40% settings under $\epsilon = 2/255$, because they allocate the loss of the most outliers with higher weight during the training. For CIFAR-100, which is larger and more challenging than CIFAR-10 and the training network ResNet is also larger than Small-CNN. In this case, we find that outliers have a large influence on robust training. Compared to noise-free settings, the performance of AT is reduced by half under the noise settings. However, `ORAT` still outperforms all baselines in all settings. The performance gap between the baselines and `ORAT` is more than 2% and even can be achieved near 10% (in 30% symmetric noise, MAIL, FGSM, and ϵ=2/255 settings). Furthermore, we find RAA is more robust than other compared methods but vulnerable to adversarial samples than `ORAT`. As we discussed before, their label correction approach may correct the wrong labels but inevitably introduces more extra noisy labels during training.

Noise		Defense	MNIST (LeNet)								CIFAR-10 (Small-CNN)								CIFAR-100 (ResNet-18)							
			ϵ=0.1				ϵ=0.2				ϵ=2/255				ϵ=8/255				ϵ=2/255				ϵ=8/255			
			Na	FG	PGD	CW	Na	FG	PGD	CW	Na	FG	PGD	CW	Na	FG	PGD	CW	Na	FG	PGD	CW	Na	FG	PGD	CW
0		ST	99.51	92.43	85.68	86.05	**99.41**	50.19	16.76	18.65	**92.48**	44.27	24.87	25.06	**92.63**	15.03	6.82	7.15	**73.67**	19.64	10.66	11.13	**73.86**	3.58	1.61	1.66
		AT	99.51	98.60	97.93	97.94	99.16	97.33	95.14	95.19	89.78	79.40	73.52	73.64	85.52	54.61	33.34	34.28	67.05	53.70	48.18	47.46	55.13	33.66	25.46	23.28
		Ours	**99.53**	**98.70**	**98.03**	**98.04**	99.30	**97.79**	**96.43**	**96.38**	89.54	**79.53**	**73.96**	**73.86**	85.91	**54.85**	**33.46**	**34.56**	68.00	**55.60**	**50.06**	**49.50**	57.82	**35.51**	**27.22**	**25.26**
Noise	10%	ST	99.32	93.72	89.09	87.17	**99.32**	67.55	36.76	26.02	62.22	29.24	25.72	24.80	**61.80**	14.70	9.72	11.38	29.60	14.17	11.41	9.96	29.41	5.22	2.54	2.19
		AT	99.12	97.98	97.36	97.28	98.39	96.62	95.25	95.11	63.03	51.36	46.02	44.95	59.53	34.17	23.33	22.28	31.02	22.04	20.25	18.97	28.96	15.08	12.42	10.49
		GAIRAT	99.11	97.87	97.33	97.20	98.35	96.57	95.22	95.08	60.41	49.23	44.72	42.02	57.12	35.48	24.22	21.52	31.05	21.31	19.76	18.08	27.73	15.13	12.28	10.26
		MAIL	99.09	98.06	97.50	97.40	98.48	96.69	95.47	95.20	55.14	47.44	44.75	39.64	53.54	33.61	26.18	20.23	29.06	18.90	16.62	14.79	27.65	13.21	10.03	7.26
		WMMR	99.14	97.95	97.32	97.23	98.36	96.58	95.25	95.11	61.83	51.43	46.56	44.97	57.97	34.74	25.15	21.96	29.74	20.11	18.65	16.86	29.23	14.78	11.84	9.96
		RAA	99.22	98.01	97.37	97.31	98.65	96.83	95.32	95.14	62.94	51.32	46.55	45.36	58.28	33.58	23.71	22.89	30.72	21.95	20.34	18.83	28.99	15.60	13.03	11.03
		Ours	**99.52**	**98.45**	**97.78**	**97.79**	99.20	**97.61**	**96.07**	**96.11**	**63.83**	**52.04**	**48.77**	**46.65**	60.42	**36.13**	**26.86**	**23.48**	**35.76**	**25.72**	**22.27**	**21.28**	**35.35**	**20.37**	**15.50**	**14.02**
	20%	ST	99.12	93.35	89.05	86.87	**99.14**	70.18	41.67	29.30	58.89	31.23	25.97	24.13	**57.28**	15.88	12.8	13.26	26.16	14.74	11.98	10.29	26.22	5.09	2.45	1.92
		AT	98.88	97.67	97.03	96.89	97.97	95.98	94.67	94.50	**61.10**	50.67	46.76	45.28	56.96	34.55	25.54	24.34	27.52	20.12	18.49	17.15	26.77	14.39	11.87	9.96
		GAIRAT	98.82	97.59	96.96	96.79	97.86	95.81	94.57	94.30	58.58	49.26	45.44	40.62	53.55	34.40	25.52	22.15	26.77	19.93	18.30	16.78	24.80	14.28	12.03	10.32
		MAIL	98.81	97.82	97.29	97.12	97.98	96.14	94.81	94.58	54.04	47.05	44.78	41.05	52.93	34.28	27.74	21.90	25.13	17.01	16.06	14.32	24.54	12.85	10.02	7.09
		WMMR	98.88	97.59	97.00	96.89	97.85	95.89	94.50	94.36	59.01	49.64	45.96	43.71	56.04	34.65	26.22	23.70	26.72	19.43	18.04	16.53	26.61	13.81	11.94	9.82
		RAA	99.07	97.70	97.07	97.05	98.25	96.28	94.41	94.20	59.53	49.56	45.73	45.60	56.71	33.24	24.51	24.12	28.16	20.73	19.15	17.74	25.34	14.79	12.46	10.25
		Ours	**99.56**	**98.37**	**97.65**	**97.64**	99.06	**97.33**	**95.71**	**95.68**	61.03	**51.85**	**48.48**	**46.16**	54.37	**35.71**	**27.76**	**24.70**	**34.45**	**25.07**	**22.21**	**20.92**	**32.83**	**19.45**	**15.49**	**13.34**
Symmetric	30%	ST	98.91	93.01	88.62	85.73	**98.97**	68.59	41.24	27.73	58.17	30.55	25.71	24.54	**57.55**	14.18	11.50	11.41	23.38	14.26	11.84	10.10	24.29	5.30	2.90	2.35
		AT	98.58	97.21	96.67	96.53	97.26	94.94	93.48	93.17	57.78	49.56	46.22	44.80	54.35	34.21	26.07	24.33	23.79	18.96	17.49	16.26	25.66	13.84	11.78	9.87
		GAIRAT	98.51	97.14	96.58	96.47	97.15	94.70	93.28	92.87	55.41	47.20	44.04	40.09	49.48	33.44	26.26	21.78	24.04	18.93	17.89	16.50	24.41	14.11	11.84	9.73
		MAIL	98.50	97.39	96.88	96.71	97.31	95.35	94.11	93.66	52.94	45.13	42.47	39.01	47.31	32.18	25.78	21.99	21.93	14.55	13.67	12.14	22.61	12.47	9.62	7.09
		WMMR	98.58	97.23	96.69	96.59	97.23	95.10	93.59	93.31	58.21	49.31	46.40	44.17	53.99	34.16	26.83	23.70	23.68	18.27	17.08	15.46	25.45	13.60	11.46	9.32
		RAA	98.84	97.35	96.63	96.59	97.66	95.45	93.74	93.40	58.72	48.78	45.20	44.85	50.81	30.50	24.49	22.60	24.53	20.06	18.54	17.03	22.08	13.95	11.62	9.67
		Ours	**99.55**	**98.30**	**97.51**	**97.53**	98.85	**96.99**	**95.31**	**95.35**	**58.99**	**50.29**	**47.11**	**45.01**	55.34	**34.89**	**27.66**	**25.00**	**31.27**	**23.81**	**21.35**	**19.59**	**31.13**	**18.96**	**15.47**	**13.21**
	40%	ST	98.87	92.37	88.36	84.85	**98.91**	68.43	41.60	29.62	**57.60**	21.03	19.46	18.03	**56.67**	15.22	12.02	11.53	21.25	13.95	12.15	10.21	21.55	6.03	3.77	2.96
		AT	98.33	96.86	96.27	96.11	96.56	94.37	92.55	92.18	47.16	43.45	41.54	40.13	40.65	31.59	26.74	24.98	22.20	18.28	17.33	15.92	23.41	13.75	11.25	9.29
		GAIRAT	98.10	96.70	96.15	95.88	96.17	93.93	92.08	91.59	47.24	40.24	38.08	36.51	40.05	30.33	25.93	23.45	21.87	18.06	17.17	15.67	21.99	12.97	11.26	9.42
		MAIL	98.33	97.06	96.44	96.20	96.50	94.45	93.01	92.54	43.63	39.43	37.97	35.60	36.28	29.91	26.32	23.31	18.83	14.22	13.49	12.08	20.47	12.50	10.33	7.35
		WMMR	98.15	96.79	96.10	95.91	96.50	94.24	92.46	92.04	45.99	41.95	39.64	37.87	30.30	27.14	24.47	22.27	21.04	17.33	16.22	14.76	23.39	12.79	11.13	9.10
		RAA	98.67	96.90	96.07	96.00	97.05	94.47	92.37	92.22	53.37	37.13	35.16	33.82	40.01	28.94	25.16	23.73	22.44	18.59	17.54	16.01	20.74	13.30	11.56	9.56
		Ours	**99.36**	**98.00**	**97.22**	**97.20**	98.63	**96.49**	**94.48**	**94.53**	53.20	**43.75**	**41.92**	**40.60**	41.86	**31.70**	**27.18**	**25.19**	**29.38**	**22.99**	**20.85**	**19.20**	**26.80**	**17.30**	**14.32**	**11.95**
Noise	10%	ST	99.39	93.05	86.09	89.09	**99.39**	65.48	22.17	35.02	**90.30**	40.73	8.24	7.25	**89.88**	18.85	0.27	0.00	31.81	15.10	11.23	10.03	31.71	4.34	1.84	1.71
		AT	99.46	98.14	97.36	97.61	99.08	97.29	95.21	95.64	88.49	78.30	72.22	71.43	75.38	52.31	30.99	30.43	33.21	22.38	20.53	19.09	30.83	15.80	12.49	10.55
		GAIRAT	99.44	98.13	97.34	97.50	99.03	97.20	95.14	95.45	85.13	75.19	68.89	64.18	71.29	52.77	31.62	19.26	33.03	22.73	20.63	19.38	29.88	15.58	12.49	10.57
		MAIL	99.50	98.20	97.38	97.73	99.08	97.52	95.79	96.35	77.93	69.74	65.63	54.36	67.39	49.66	31.62	18.65	31.45	21.36	17.92	16.43	29.66	13.99	10.18	7.64
		WMMR	99.45	98.11	97.38	97.59	99.08	97.28	95.30	95.67	86.68	76.44	70.23	68.39	72.85	53.77	32.57	28.74	32.52	22.08	19.63	17.81	30.89	14.86	11.91	10.01
		RAA	99.42	98.11	97.34	97.51	99.08	97.24	95.17	95.45	88.75	78.20	72.33	69.98	72.85	53.98	32.02	31.69	33.01	23.07	21.07	19.69	30.35	16.09	12.79	10.70
		Ours	**99.50**	**98.58**	**97.81**	**98.04**	99.18	**97.68**	**96.11**	**96.36**	88.77	**78.75**	**72.72**	**71.97**	75.42	**54.62**	**32.70**	**32.17**	**37.09**	**27.07**	**23.65**	**22.59**	**35.72**	**20.51**	**15.46**	**13.83**
	20%	ST	99.19	92.64	86.00	88.63	**99.23**	64.76	20.49	32.18	**87.69**	37.40	8.08	7.52	**87.16**	15.95	0.06	0.00	30.46	14.79	11.57	10.38	30.49	4.92	2.14	2.03
		AT	99.44	98.12	97.29	97.59	99.07	97.11	95.20	95.84	87.51	77.08	71.56	71.38	73.38	52.04	30.84	30.80	31.67	21.15	19.44	18.19	29.96	15.47	12.46	10.51
		GAIRAT	99.41	98.10	97.20	97.54	99.01	96.99	95.16	95.66	74.61	64.33	62.22	59.39	70.03	52.72	31.36	17.68	31.64	21.64	19.80	18.42	29.15	15.45	12.58	10.70
		MAIL	99.46	98.19	97.38	97.78	99.17	97.38	95.67	96.26	74.72	66.63	62.76	50.04	66.52	51.00	30.50	29.09	30.94	20.38	16.98	15.60	28.48	13.32	9.89	7.18
		WMMR	99.44	98.11	97.19	97.60	99.06	97.07	95.22	95.83	70.41	60.90	56.48	51.44	54.06	41.00	25.25	24.82	31.90	20.82	18.96	17.31	30.61	14.28	11.75	9.81
		RAA	99.40	98.14	97.32	97.60	99.13	97.05	95.05	95.44	87.58	77.84	72.10	70.30	71.61	53.56	31.99	31.27	31.91	21.75	20.08	18.75	30.13	15.90	12.83	10.72
		Ours	**99.51**	**98.60**	**97.87**	**98.16**	99.12	**97.56**	**96.11**	**96.55**	88.00	**77.94**	**72.26**	**73.03**	73.90	**54.43**	**32.64**	**32.71**	**36.05**	**25.76**	**22.83**	**21.47**	**34.11**	**19.55**	**15.12**	**13.70**
Asymmetric	30%	ST	99.07	91.12	83.58	85.97	99.12	59.61	20.58	29.17	**84.75**	33.23	8.60	8.05	**85.14**	14.29	0.20	0.03	28.19	14.94	11.91	10.76	28.80	3.97	1.68	1.41
		AT	99.41	98.07	97.14	97.55	98.95	96.90	95.12	95.80	86.53	76.10	70.56	71.59	71.53	51.01	30.78	30.86	30.19	20.77	18.72	17.58	28.42	14.57	11.93	10.15
		GAIRAT	99.37	98.06	97.27	97.65	99.02	97.01	94.90	95.65	73.92	63.94	61.20	58.65	63.39	49.31	31.06	15.88	29.99	20.64	18.69	17.27	27.33	15.00	11.64	9.94
		MAIL	99.44	98.12	97.41	97.56	99.13	97.33	95.48	96.11	68.58	60.40	56.75	44.55	63.48	46.45	29.04	17.50	29.28	19.12	16.09	14.71	27.23	12.99	9.43	6.89
		WMMR	99.40	98.07	97.15	97.63	99.05	97.03	94.98	95.75	69.49	60.25	56.11	50.76	53.57	40.10	23.97	21.27	30.20	19.46	17.70	16.17	29.02	13.98	11.27	9.20
		RAA	99.40	98.07	97.20	97.53	99.04	97.00	95.11	95.62	80.75	72.27	66.87	66.11	64.93	48.97	28.68	29.17	30.58	21.08	19.19	17.98	29.08	15.18	12.28	10.38
		Ours	**99.47**	**98.51**	**97.82**	**98.12**	**99.24**	**97.45**	**95.98**	**96.63**	86.76	**76.29**	**71.15**	**71.60**	71.52	**52.90**	**32.34**	**32.58**	**34.58**	**24.18**	**21.05**	**20.11**	**33.80**	**19.35**	**14.84**	**13.70**
	40%	ST	97.43	81.71	73.74	75.11	97.58	50.69	16.99	22.82	**79.83**	28.98	6.84	6.34	**78.37**	12.44	0.02	0.00	26.99	12.97	10.29	9.56	27.11	4.32	1.91	1.67
		AT	99.33	97.81	96.90	97.22	98.89	96.82	94.63	95.46	85.14	74.66	69.57	69.44	66.68	50.25	30.76	30.40	28.18	20.20	18.52	17.13	27.09	13.90	11.32	9.85
		GAIRAT	99.26	97.70	96.78	97.03	98.85	96.65	94.36	95.24	70.19	60.32	55.63	49.99	62.43	48.98	30.85	15.61	28.42	19.67	17.88	16.75	25.70	14.49	11.55	9.85
		MAIL	99.24	97.74	96.87	96.84	98.79	96.94	94.93	95.26	64.62	57.14	53.99	42.62	53.75	41.72	26.41	15.29	27.68	17.73	15.77	13.91	25.82	11.76	9.00	6.92
		WMMR	99.35	97.84	96.89	97.22	98.91	96.70	94.58	95.30	67.12	57.91	54.26	49.42	47.34	36.84	21.89	20.17	28.65	19.15	17.54	15.94	27.70	13.69	11.33	9.67
		RAA	99.40	98.02	97.14	97.23	98.96	96.91	95.02	95.29	72.65	65.66	61.10	59.58	53.59	43.28	26.02	25.86	28.25	19.97	18.58	17.36	27.68	14.42	11.70	10.21
		Ours	**99.41**	**98.38**	**97.68**	**97.97**	**99.09**	**97.25**	**95.62**	**96.19**	85.10	**75.35**	**69.89**	**69.72**	66.75	**50.70**	**31.52**	**32.47**	**33.65**	**23.25**	**20.76**	**19.46**	**31.69**	**18.47**	**14.16**	**12.81**

Table 1: *Testing accuracy (%) of seven methods on MNIST, CIFAR-10, and CIFAR-100 with different levels of symmetric and asymmetric noisy. The best results are shown in bold. We color the performance of all adversarial training methods on three different attacks. The performance gap between current method and* `ORAT` *are shown in green: ≤2%; yellow: (2%,5%]; orange: (5%,10%]; blue: >10%. According to results,* `ORAT` *outperforms all adversarial training methods on all attack settings. 'Na' represents Natural. 'FG' represents FGSM.*

Third, for asymmetric noise settings, we can get similar observations. Note that `ORAT` can even outperform the baselines over 10% on the CIFAR-10 dataset. All adversarial methods outperform ST in all attack settings since ST does not have a mechanism to handle adversarial samples. In general, increasing the noise percentage, we can find there is a decreasing trend in the performance among all methods. On the other hand, increasing the value of ϵ will

Defense \ Data	MNIST		CIFAR-10		CIFAR-100	
	ϵ=0.1	ϵ=0.2	ϵ=2/255	ϵ=8/255	ϵ=2/255	ϵ=8/255
ST	84.13	23.78	23.54	10.41	10.22	1.72
AT	96.73	93.23	44.56	23.06	17.11	9.89
GAIRAT	96.78	93.57	40.82	20.57	16.54	9.97
MAIL	97.01	93.80	40.39	18.95	13.67	7.29
WMMR	96.88	93.60	43.52	20.68	15.69	8.73
RAA	96.98	93.62	45.07	22.29	17.70	9.78
Ours	**97.67**	**95.05**	**45.59**	**23.35**	**19.74**	**12.22**

Table 2: *Testing accuracy (%) on AutoAttack.*

Noise	Defense	MNIST ($\epsilon = 0.1$)				CIFAR-100 ($\epsilon = 2/255$)			
		Na	FG	PGD	CW	Na	FG	PGD	CW
10%	AT w/o	98.91	98.08	97.61	97.55	29.81	21.09	19.59	18.23
	Ours	**99.52**	**98.45**	**97.78**	**97.79**	**35.76**	**25.72**	**22.27**	**21.28**
30%	AT w/o	97.82	96.97	96.47	96.35	24.18	19.17	17.31	16.71
	Ours	**99.55**	**98.30**	**97.51**	**97.53**	**31.27**	**23.81**	**21.35**	**19.59**

Table 3: *Testing accuracy (%) of AT w/o and* `ORAT` *on symmetric noisy data.*

Noise	Defense	MNIST ($\epsilon = 0.1$)				CIFAR-100 ($\epsilon = 2/255$)			
		Na	FG	PGD	CW	Na	FG	PGD	CW
40% Symmetric Noise	ST	98.38 (0.18)	77.47 (5.79)	59.31 (11.13)	51.47 (12.79)	21.53 (0.71)	13.95 (0.16)	12.12 (0.28)	10.12 (0.28)
	AT	96.60 (0.87)	95.11 (1.17)	94.36 (1.31)	94.16 (1.37)	21.82 (0.30)	18.14 (0.27)	17.02 (0.29)	15.62 (0.32)
	GAIRAT	98.04 (0.11)	96.60 (0.08)	95.98 (0.13)	95.79 (0.12)	21.58 (0.30)	17.94 (0.36)	16.88 (0.36)	15.50 (0.33)
	MAIL	98.06 (0.17)	96.88 (0.16)	96.31 (0.12)	96.11 (0.10)	17.52 (0.74)	14.31 (0.43)	13.46 (0.40)	12.09 (0.37)
	WMMR	98.07 (0.14)	96.69 (0.10)	96.04 (0.08)	95.85 (0.08)	20.64 (0.29)	17.14 (0.33)	16.10 (0.30)	14.76 (0.27)
	RAA	98.58 (0.11)	96.82 (0.08)	95.99 (0.10)	95.91 (0.09)	22.21 (0.32)	18.27 (0.32)	17.22 (0.30)	15.85 (0.33)
	Ours	**99.34 (0.03)**	**98.02 (0.07)**	**97.19 (0.08)**	**97.18 (0.07)**	**30.12 (0.47)**	**23.62 (0.42)**	**21.48 (0.49)**	**20.02 (0.41)**

Table 4: *Mean and standard deviation (in parentheses) of testing accuracy (%) across 10 random runs.*

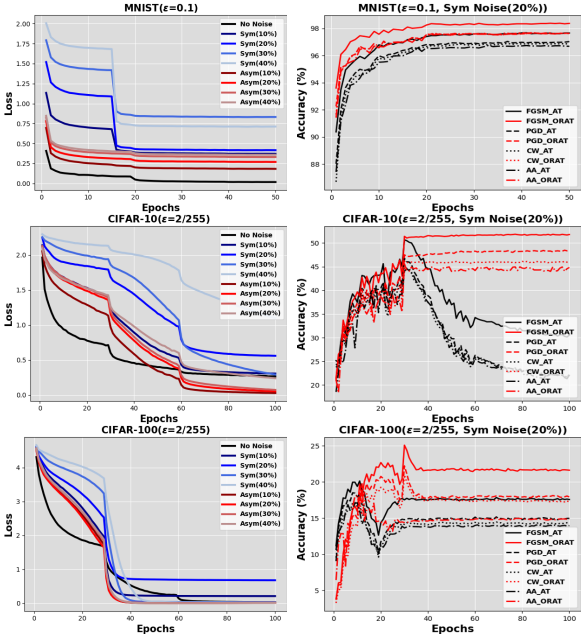

Figure 2: *The tendency curves of training adversarial loss and test accuracy on three datasets.*

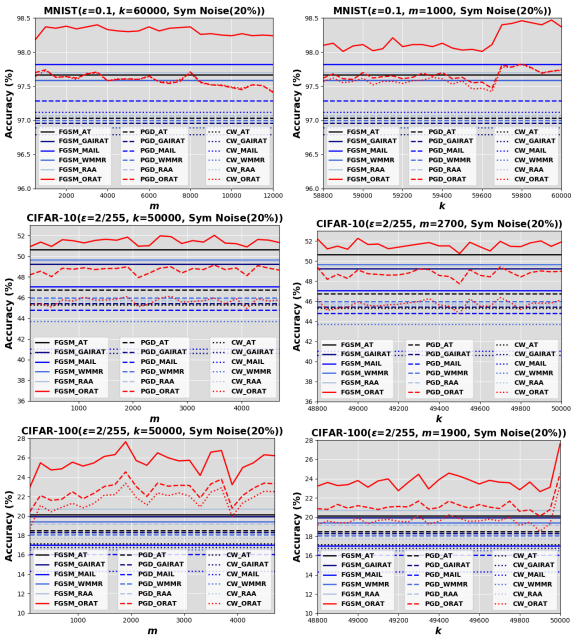

Figure 3: *Effect of k and m on the test accuracy of* `ORAT` *on three datasets.*

also decrease all model performance. Even a small amount of label noise causes classifiers to have significant adversarial errors.

We also report testing accuracy for AA on all datasets with 20% symmetric noise in Table 2. We can find our method outperforms all compared methods even if using a strong attack strategy. In addition, we compare AT w/o with our method and report performance in Table 3. Comparing Table 1 and Table 3 under the same setting, it is evident that simply removing outliers from training data cannot significantly improve AT performance, and our method still achieves the best performance. The stability evaluation results of each method with 10 random runs are shown in Table 4, where we use 40% symmetric noisy data as an example. Comparing Table 1 and Table 4 under the same setting, it is clear that the performance gap becomes larger when we report scores by using mean and standard deviation, and our method shows a stable and stronger ability in handling outliers and adversarial attacks.

Convergence and Robustness Tendency. We show the tendency curves of the training loss on all datasets with different noise when using Algorithm 1 in the left panel of Figure 2. The sharp drops in the curves correspond to decreases in training learning rate. We can observe a steady decrease in the training loss on all datasets with the increase of training epochs when training against adversarial samples, which supports that Algorithm 1 can effectively optimize and solve Eq. (4) even if it is a non-smooth loss. We also compare the robust accuracy of AT and `ORAT` by using four attack strategies on all datasets with 20% symmetric noise in Figure 2 right panel. It is clear that our method outperforms the original AT method in all settings. In general, these plots illustrate that we can consistently reduce the value of the objective function of `ORAT`, thus producing an increasingly robust classifier.

Effect of k and m. We study how the choices of k and m affect the performance of `ORAT` with two types of experiments on all datasets under 20% symmetric noise setting, together with those from other defense methods. In the first set of experiments, we fix k to the total number of training samples and run the algorithm with different values of m. The results are plotted in Figure 3 (left panel). We can see that there is a clear range of m with better performance than all compared methods. In the second set of experiments, we fix m with the best performance from the first set of experiments and run `ORAT` with different values of k. The results are shown in Figure 3 (right panel). Note that there is also a range of k with better performance, in particular, the optimal value of k is less than the number of total training samples. Similar trends are observed on all datasets.

6. Conclusion

In this work, we introduce the outlier robust adversarial training (`ORAT`), which considers both outliers in training data and adversarial attacks in the model training. We provide an optimizing algorithm and analyze the theoretical aspects of `ORAT`. Empirical results showed the effectiveness and robustness of `ORAT` on three benchmark datasets. In the future, we will study the optimization error or convergence rate of our proposed learning algorithm. We will also evaluate our method on large datasets and large deep neural networks. Although achieving fairness (Ju et al., 2023; Hu and Chen, 2022) is not a goal of this work, we have found that our method can benefit the minority subgroup of data (see Figure 1 right panel). Studying fairness with `ORAT` is an interesting direction in the future. Furthermore, we plan to design an efficient method to automatically determine the hyperparameters k and m.

Acknowledgments: This work is supported by an NSF research grant IIS-2008532.

References

Maximilian Augustin et al. Adversarial robustness on in-and out-distribution improves explainability. In *ECCV*, pages 228–245. Springer, 2020.

Pranjal Awasthi et al. Adversarial learning guarantees for linear hypotheses and neural networks. In *ICML*, 2020.

Pranjal Awasthi et al. Calibration and consistency of adversarial surrogate losses. *NeurIPS*, 2021.

Han Bao, Clay Scott, and Masashi Sugiyama. Calibrated surrogate losses for adversarially robust classification. In *Conference on Learning Theory*, pages 408–451. PMLR, 2020.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 2002.

Julian Bitterwolf et al. Certifiably adversarially robust detection of out-of-distribution data. *NeurIPS*, 2020.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

Pratik Chaudhari et al. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, and Hang Su. Adversarial distributional training for robust deep learning. *NeurIPS*, 2020.

Yinpeng Dong et al. Exploring memorization in adversarial training. In *ICLR*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Robust attentive deep neural network for detecting gan-generated faces. *IEEE Access*, 10:32574–32583, 2022.

Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *ICCV*, 2019.

Kai Han et al. Training binary neural networks through learning with noisy supervision. In *ICML*, 2020.

Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Kaiming He et al. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

Dan Hendrycks et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

Chester Holtz, Tsui-Wei Weng, and Gal Mishne. Learning sample reweighting for adversarial robustness. *ICLR Submission*, 2021.

Shu Hu and George H Chen. Distributionally robust survival analysis: A novel fairness loss without demographics. In *Machine Learning for Health*, pages 62–87. PMLR, 2022.

Shu Hu, Yiming Ying, Siwei Lyu, et al. Learning by minimizing the sum of ranked range. *Advances in Neural Information Processing Systems*, 33, 2020.

Shu Hu, Lipeng Ke, Xin Wang, and Siwei Lyu. Tkml-ap: Adversarial attacks to top-k multi-label learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7649–7657, 2021.

Shu Hu, Yiming Ying, Xin Wang, and Siwei Lyu. Sum of ranked range loss for supervised learning. *The Journal of Machine Learning Research*, 23(1):4826–4869, 2022.

Shu Hu, Xin Wang, and Siwei Lyu. Rank-based decomposable losses in machine learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. *Advances in neural information processing systems*, 2020.

Lu Jiang et al. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018.

Yan Ju, Shu Hu, Shan Jia, George H Chen, and Siwei Lyu. Improving fairness in deepfake detection. *arXiv preprint arXiv:2306.16635*, 2023.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *NeurIPS*, 2008.

Pierre Laforgue, Guillaume Staerman, and Stephan Clémençon. Generalization bounds in the presence of outliers: a median-of-means study. In *ICML*. PMLR, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Feng Liu et al. Probabilistic margins for instance reweighting in adversarial training. *Advances in Neural Information Processing Systems*, 34, 2021.

Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

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

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

Wenbo Pu et al. Learning a deep dual-level network for robust deepfake detection. *Pattern Recognition*, 2022.

Hadi Salman et al. Do adversarially robust imagenet models transfer better? *NeurIPS*, 2020.

Amartya Sanyal, Puneet K Dokania, Varun Kanade, and Philip Torr. How benign is benign overfitting? In *International Conference on Learning Representations*, 2021.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

Sainbayar Sukhbaatar et al. Training convolutional networks with noisy labels. In *ICLR workshops*, 2015.

Neeraj Varshney et al. Investigating selective prediction approaches across several tasks in iid, ood, and adversarial settings. *arXiv preprint arXiv:2203.00211*, 2022.

Yisen Wang et al. Iterative learning with open-set noisy labels. In *CVPR*, 2018.

Yisen Wang et al. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019.

Yisen Wang et al. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pages 2691–2699, 2015.

Mingyang Yi et al. Improved ood generalization via adversarial training and pretraing. In *ICML*, 2021.

Dong Yin et al. Rademacher complexity for adversarially robust generalization. In *ICML*, 2019.

Xingrui Yu et al. How does disagreement help generalization against label corruption? In *ICML*, 2019.

Huimin Zeng et al. Are adversarial examples created equal? a learnable weighted minimax risk for robustness under non-uniform attacks. In *AAAI*, 2021a.

Zhiyuan Zeng et al. Adversarial self-supervised learning for out-of-domain detection. In *NAACL*, pages 5631–5639, 2021b.

Runtian Zhai, Chen Dan, J Zico Kolter, and Pradeep Ravikumar. Doro: Distributional and outlier robust optimization. *ICML*, 2021.

Hongyang Zhang et al. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.

Jingfeng Zhang et al. Geometry-aware instance-reweighted adversarial training. In *ICLR*, 2021.

Yuchen Zhang and Percy Liang. Defending against whitebox adversarial attacks via randomized discretization. In *AISTATS*, pages 684–693. PMLR, 2019.

Jianing Zhu et al. Understanding the interaction of adversarial training with noisy labels. *arXiv preprint arXiv:2102.03482*, 2021.