Enhancing Adversarial Robustness in Low-Label Regime via Adaptively Weighted Regularization and Knowledge Distillation

Dongyoon Yang
Seoul National University
Department of Statistics
ydy0415@gmail.com

Insung Kong
Seoul National University
Department of Statistics
ggong369@snu.ac.kr

Yongdai Kim
Seoul National University
Department of Statistics
ydkim0903@gmail.com

Abstract

Adversarial robustness is a research area that has recently received a lot of attention in the quest for trustworthy artificial intelligence. However, recent works on adversarial robustness have focused on supervised learning where it is assumed that labeled data is plentiful. In this paper, we investigate semi-supervised adversarial training where labeled data is scarce. We derive two upper bounds for the robust risk and propose a regularization term for unlabeled data motivated by these two upper bounds. Then, we develop a semi-supervised adversarial training algorithm that combines the proposed regularization term with knowledge distillation using a semi-supervised teacher (i.e., a teacher model trained using a semi-supervised learning algorithm). Our experiments show that our proposed algorithm achieves state-of-the-art performance with significant margins compared to existing algorithms. In particular, compared to supervised learning algorithms, performance of our proposed algorithm is not much worse even when the amount of labeled data is very small. For example, our algorithm with only 8% labeled data is comparable to supervised adversarial training algorithms that use all labeled data, both in terms of standard and robust accuracies on CIFAR-10.

1. Introduction

Neural network models used for image classification are vulnerable to adversarial perturbations that are imperceptible to humans [28]. These perturbed images are called *adversarial examples*, and they can be generated without any knowledge of the underlying model, leading to security concerns [23, 24, 4, 14, 25]. Adversarial examples can also cause problems in real-world scenarios, where printed images with adversarial perturbations can easily fool the classification model [17]. To defend against adversarial attacks, many adversarial learning algorithms have been proposed, such as [19, 36, 30, 37, 38, 26].

Learning accurate prediction models typically requires a large amount of labeled data, which can be expensive and time-consuming to collect. In contrast, obtaining unlabeled data is relatively easier. Semi-supervised learning is a research area that focuses on effectively utilizing unlabeled data [20, 2, 32, 27, 35]. Because adversarial training algorithms also require labeled data, semi-supervised adversarial training (SS-AT) has become a crucial research area for reliable artificial intelligence [3, 29, 34, 18].

The aim of this paper is to develop a new SS-AT algorithm that is theoretically well motivated and empirically superior to existing competitors when the amount of labeled data is insufficient. We propose an objective function for SS-AT, which consists of three key components - one for generalization of labeled data, a regularization term with unlabeled data for adversarial robustness, and a knowledge distillation term for improving generalization ability of unlabeled data. Our regularization term is motivated by two new upper bounds of the boundary risk. The knowledge distillation term is added to estimate soft pseudo labels of unlabeled data, which contrasts with existing algorithms [29, 3, 34] that use hard pseudo labels. Using soft pseudo labels instead of hard pseudo labels improves the performance significantly.

Our contributions can be summarized as follows:

- We derive two upper bounds of the robust risk for semi-supervised adversarial training, providing theoretical insights into the performance of proposed method.

- We propose a novel semi-supervised adversarial training algorithm, called Semisupervised-Robust-Self-Training with Adaptively Weighted Regularization (SRST-AWR), that combining an adaptively weighted regularization and knowledge distillation with a semi-supervised teacher to put soft pseudo labels in adversarial training.

- We demonstrate the effectiveness of our algorithm through numerical experiments on various benchmark datasets, showing simultaneous improvements in robustness and generalization with significant performance gains over existing state-of-the-art methods.
- Our proposed algorithm exhibits only minor performance degradation even when the amount of labeled data is limited compared to fully supervised methods with a large amount of labeled data.

2. Preliminaries

Let $\mathcal{X} \subset \mathbb{R}^d$ be the input space, $\mathcal{Y} = \{1, \cdots, C\}$ be the set of output labels and $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^C$ be the score function parametrized by parameters $\boldsymbol{\theta}$ such that $\mathbf{p}_\theta(\cdot|\boldsymbol{x}) = \mathrm{softmax}(f_{\boldsymbol{\theta}}(\boldsymbol{x})) \in \mathbb{R}^C$ is the vector of the predictive conditional probabilities. Let $F_{\boldsymbol{\theta}}(\boldsymbol{x}) = \underset{c}{\mathrm{argmax}}[f_{\boldsymbol{\theta}}(\boldsymbol{x})]_c \in \mathbb{R}^C$, $\mathcal{B}_p(\boldsymbol{x}, \varepsilon) = \{\boldsymbol{x}' \in \mathcal{X} : \|\boldsymbol{x} - \boldsymbol{x}'\|_p \leq \varepsilon\}$ and Let $\mathbb{1}\{\cdot\}$ represent the indicator function, which takes the value 1 when the condition $\cdot$ is satisfied, and 0 otherwise.

2.1. Population Robust Risk

The population robust risk used in adversarial training is defined as

$$\mathcal{R}_{\mathrm{rob}}(\theta) = \mathbb{E}_{(\mathbf{X},\mathbf{Y})} \max_{\mathbf{X}' \in \mathcal{B}_p(\mathbf{X},\varepsilon)} \mathbb{1}\left\{F_\theta(\mathbf{X}') \neq \mathbf{Y}\right\}. \quad (1)$$

The objective of adversarial training is to learn $\boldsymbol{\theta}$ that minimizes the population robust risk (1). Most state-of-the-art adversarial training algorithms consist of two steps: a maximization step and a minimization step. In the maximization step, an adversarial example $\boldsymbol{x}' \in \mathcal{B}_p(\boldsymbol{x}, \varepsilon)$ is generated, which is described in Section 2.2. In the minimization step, a certain regularized empirical risk for given adversarial examples is minimized [19, 36, 30, 26].

2.2. Adversarial Attack

An adversarial attack is a method to generate an adversarial example. Adversarial attacks can be categorized into white-box attacks [28, 10, 19] and black-box attacks [23, 24, 1]. In the white-box attack, it is assumed that the adversary can exploit all information about the model architectures and parameters to generate adversarial examples. In a black-box attack, the adversary can only access the outputs of the model.

One of the most popular white-box adversarial attack algorithms is Projected Gradient Descent (PGD), which finds an adversarial example by iteratively updating it by the gradient ascent and projecting onto the ϵ-ball of the original data [19]. The formula of PGD$^{\mathrm{T}}$ is as follows:

$$\boldsymbol{x}^{(T)} = \boldsymbol{\Pi}_{\mathcal{B}_p(\boldsymbol{x},\varepsilon)}\left(\boldsymbol{x}^{(T-1)} + \nu\,\mathrm{sgn}\left(\nabla_{\boldsymbol{x}^{(T-1)}}\eta(\boldsymbol{x}^{(T-1)}|\boldsymbol{\theta}, \boldsymbol{x}, y)\right)\right), \quad (2)$$

where $\boldsymbol{\Pi}_{\mathcal{B}_p(\boldsymbol{x},\varepsilon)}(\cdot)$ is the projection operator to $\mathcal{B}_p(\boldsymbol{x}, \varepsilon)$, $\nu > 0$ is the step size, η is a surrogate loss , $\boldsymbol{x}^{(0)} = \boldsymbol{x}$, T is the number of iterations. The cross-entropy or Kullback–Leibler divergence for η can be used.

2.3. Semi-Supervised Learning

Virtual Adversarial Training (VAT) [20] is a semi-supervised learning algorithm minimizing

$$\frac{1}{n_l}\sum_{i=1}^{n_l}\ell_{\mathrm{ce}}(f_\theta(\boldsymbol{x}_i), y_i) + \lambda\frac{1}{n_{ul}}\sum_{j=1}^{n_{ul}}\mathrm{D}_{\mathrm{KL}}(\mathbf{p}_{\tilde{\theta}}(\cdot|\boldsymbol{x}_j)\|\mathbf{p}_\theta(\cdot|\widehat{\boldsymbol{x}}_j^{\mathrm{adv}})), \quad (3)$$

where $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_l}$ and $\{\boldsymbol{x}_j\}_{j=1}^{n_{ul}}$ are labeled and unlabeled samples, respectively, and $\tilde{\theta}$ is the pretrained parameter and $\widehat{\boldsymbol{x}}_j^{\mathrm{adv}} \in \mathcal{B}_2(\boldsymbol{x}_j, \varepsilon)$ is an adversarial example.

FixMatch [27] is a semi-supervised learning algorithm minimizing

$$\frac{1}{n_l}\sum_{i=1}^{n_l}\ell_{\mathrm{ce}}(f_\theta(\boldsymbol{x}_i), y_i) + \frac{1}{n_\tau}\sum_{j=1}^{n_{ul}}\ell_{\mathrm{ce}}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_j^s), F_{\boldsymbol{\theta}}(\boldsymbol{x}_j^w))\mathbb{1}\{\max_c p_{\boldsymbol{\theta}}(c|\boldsymbol{x}_j^w) > \tau\},$$

where $\tau \in (0, 1)$ is a constant, $n_\tau = \sum_{j=1}^{n_{ul}}\mathbb{1}\{\max_c p_{\boldsymbol{\theta}}(c|\boldsymbol{x}_j^w) > \tau\}$ and $\boldsymbol{x}_j^s$ and $\boldsymbol{x}_j^w$ are strongly and weakly augmented samples [8], respectively.

2.4. Semi-Supervised Adversarial Training

Existing adversarial training algorithms can be categorized into two types: one directly minimizing empirical robust risk (e.g. PGD-AT [19]), and the other decomposing the robust risk into supervised and regularization terms and minimizing the corresponding regularized empirical risk (e.g. TRADES [36]). In algorithms based on PGD-AT, label information for all data is required, and thus it cannot be directly applied in a semi-supervised setting. TRADES can be applied in a semi-supervised setting since the regularization term does not require label information. However, it shows poor performance for semi-supervised learning.

For achieving adversarial robustness in a semi-supervised setting, several SS-AT algorithms have been proposed [29, 3, 34, 18, 11]. RST [3] generates pseudo labels for unlabeled data by predicting the class labels using a teacher model trained only with labeled data. Then, the algorithm minimizes the regularized empirical risk of TRADES [36] using both the labeled data and the unlabeled data with their pseudo labels. That is, it minimizes the fol-

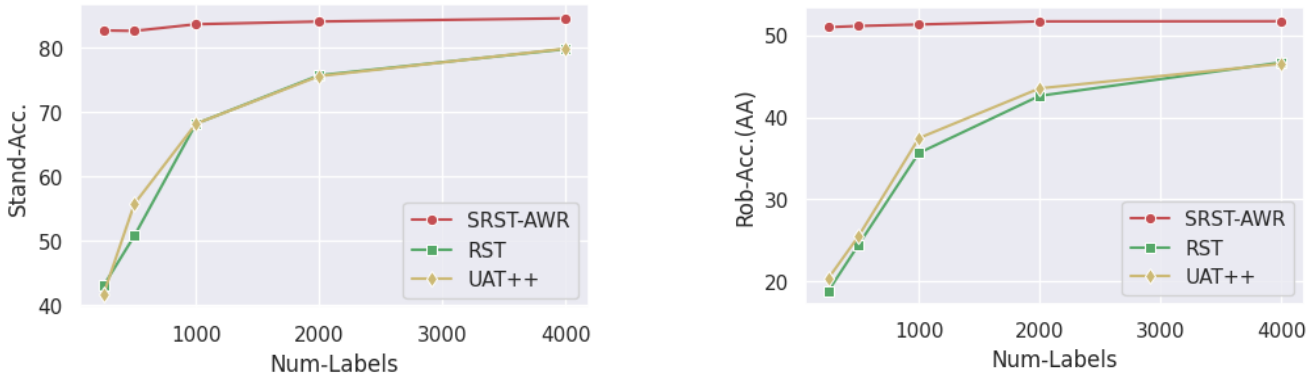

Figure 1: **Performance comparison of SRST-AWR, RST and UAT++ for varying the number of labeled data**. The x-axis is the number of labeled data and y-axis are the standard accuracy and robust accuracy against autoattack, respectively.

lowing regularized empirical risk:

$$\begin{aligned}&\mathcal{R}_{(\mathrm{RST})}(\boldsymbol{\theta}; \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_l}, \{(\boldsymbol{x}_j, \widehat{y}_j)\}_{j=n_l+1}^{n_{n_l}+n_{ul}}, \lambda) \\ &:= \frac{1}{n_l+n_{ul}} \sum_{k=1}^{n_l+n_{ul}} \Big\{ \ell_{\mathrm{ce}}(f_\theta(\boldsymbol{x}_k), y_k \text{ or } \widehat{y}_k) \\ &+ \lambda \, \mathrm{D}_{\mathrm{KL}}(\mathbf{p}_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x}_k) || \mathbf{p}_{\boldsymbol{\theta}}(\cdot|\widehat{\boldsymbol{x}}_k^{\mathrm{pgd}})) \Big\}, \end{aligned} \tag{4}$$

where $\widehat{y}_j = F_{\boldsymbol{\theta}_T}(\boldsymbol{x}_j)$ and $\boldsymbol{\theta}_T$ is the parameter of the teacher model trained only with labeled data.

On the other hand, UAT [29] minimizes the following regularized empirical risk:

$$\begin{aligned}&\mathcal{R}_{(\mathrm{UAT})}(\boldsymbol{\theta}; \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_l}, \{(\boldsymbol{x}_j, \widehat{y}_j)\}_{j=n_l+1}^{n_{n_l}+n_{ul}}, \lambda) \\ &:= \frac{1}{n_l+n_{ul}} \sum_{k=1}^{n_l+n_{ul}} \Big\{ \ell_{\mathrm{ce}}(f_\theta(\widehat{\boldsymbol{x}}_k^{\mathrm{pgd}}), y_k \text{ or } \widehat{y}_k) \\ &+ \lambda \, \mathrm{D}_{\mathrm{KL}}(\mathbf{p}_{\tilde{\boldsymbol{\theta}}}(\cdot|\boldsymbol{x}_k) || \mathbf{p}_{\boldsymbol{\theta}}(\cdot|\widehat{\boldsymbol{x}}_k^{\mathrm{pgd}})) \Big\}. \end{aligned} \tag{5}$$

If $\lambda = 0$, it is called Unsupervised Adversarial Training with Fixed Targets (UAT-FT); otherwise, it is called Unsupervised Adversarial Training Plus Plus (UAT++).

However, as seen in Figure 1, RST and UAT++ show poor performances when the amount of labeled data is insufficient, which is partially the teacher model does not predict well. To resolve this problem, ARMOURED (Adversarially Robust MOdels using Unlabeled data by REgularizing Diversity) [18] combines multi-view learning and diversity regularization. It employs a multi-view ensemble learning approach for selecting high quality pseudo labeled data.

3. Semi-Supervised Robust Self-Training via Adaptively Weighted Regularization and Knowledge Distillation

In this section, we derive two upper bounds of the robust risk and propose a SS-AT algorithm by modifying the upper bounds.

3.1. Upper Bounds of the Population Robust Risk

The population robust risk $\mathcal{R}_{\mathrm{rob}}(\boldsymbol{\theta})$ is decomposed of the two terms - natural risk and boundary risk as follows [36]:

$$\mathcal{R}_{\mathrm{rob}}(\boldsymbol{\theta}) = \mathcal{R}_{\mathrm{nat}}(\boldsymbol{\theta}) + \mathcal{R}_{\mathrm{bdy}}(\boldsymbol{\theta}),$$

where $\mathcal{R}_{\mathrm{nat}}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{X},Y)} \mathbb{1}\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq Y\}$ and $\mathcal{R}_{\mathrm{bdy}}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{X},Y)} \mathbb{1}\{\exists \mathbf{X}' \in \mathcal{B}_p(\mathbf{X}, \varepsilon) : F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(\mathbf{X}'), F_{\boldsymbol{\theta}}(\mathbf{X}) = Y\}$. The objective of adversarial training is to find $\boldsymbol{\theta}$ minimizing the population robust risk.

The following theorems provide two upper bounds of the population robust risk whose proofs are deferred to Appendix A.

Theorem 3.1. *For a given score function $f_{\boldsymbol{\theta}}$, let*

$$z(\boldsymbol{x}) \in \operatorname*{argmax}_{\boldsymbol{x}' \in \mathcal{B}_p(\boldsymbol{x}, \varepsilon)} \mathbb{1}\{F_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq F_{\boldsymbol{\theta}}(\boldsymbol{x}')\}.$$

Then, we have

$$\begin{aligned}&\mathcal{R}_{rob}(\boldsymbol{\theta}) \leq \mathbb{E}_{(\mathbf{X},Y)} \mathbb{1}\{Y \neq F_{\boldsymbol{\theta}}(\mathbf{X})\} \\ &+ \mathbb{E}_{\mathbf{X}} \{\mathbb{1}\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))\} \cdot p(Y \neq F_{\boldsymbol{\theta}}(z(\mathbf{X})) | \mathbf{X})\}. \end{aligned} \tag{6}$$

Theorem 3.2. *For a given score function $f_{\boldsymbol{\theta}}$, let*

$$z(\boldsymbol{x}) \in \operatorname*{argmax}_{\boldsymbol{x}' \in \mathcal{B}_p(\boldsymbol{x}, \varepsilon)} \mathbb{1}\{F_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq F_{\boldsymbol{\theta}}(\boldsymbol{x}')\}.$$

Then, we have

$$\mathcal{R}_{rob}(\boldsymbol{\theta}) \leq \mathbb{E}_{(\mathbf{X},Y)}\mathbb{1}\{Y \neq F_{\boldsymbol{\theta}}(\mathbf{X})\} + \mathbb{E}_{\mathbf{X}}\left\{\mathbb{1}\{F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))\} \cdot p(Y = F_{\boldsymbol{\theta}}(\mathbf{X})|\mathbf{X})\right\}. \tag{7}$$

The key point of Theorems 3.1 and 3.2 is that their second terms on the right-hand side do not depend on label information. Thus, the second terms can be used as regularization term for unlabeled data. In contrast, the regularization terms used in supervised adversarial training algorithms such as MART [30] and ARoW [9] require label information.

Comparison to TRADES [36] For binary classification problems such that $\mathcal{Y} = \{-1, 1\}$, the following upper bounds can be derived using Theorems 3.1 and 3.2:

$$\mathcal{R}_{\text{rob}}(\boldsymbol{\theta}) \leq \mathbb{E}_{(\mathbf{X},Y)}\phi(Y f_{\boldsymbol{\theta}}(\mathbf{X})) + \mathbb{E}_{\mathbf{X}}\{\phi(f_{\boldsymbol{\theta}}(\mathbf{X}) f_{\boldsymbol{\theta}}(z(\mathbf{X}))/\lambda) \cdot p(Y \neq F_{\boldsymbol{\theta}}(z(\mathbf{X}))|\mathbf{X})\}, \tag{8}$$

$$\mathcal{R}_{\text{rob}}(\boldsymbol{\theta}) \leq \mathbb{E}_{(\mathbf{X},Y)}\phi(Y f_{\boldsymbol{\theta}}(\mathbf{X})) + \mathbb{E}_{\mathbf{X}}\{\phi(f_{\boldsymbol{\theta}}(\mathbf{X}) f_{\boldsymbol{\theta}}(z(\mathbf{X}))/\lambda) \cdot p(Y = F_{\boldsymbol{\theta}}(\mathbf{X})|\mathbf{X})\}, \tag{9}$$

where ϕ is the binary cross entropy loss and $\lambda > 0$ is a regularization parameter. The proofs are provided in Appendix A. In contrast, [36] shows that

$$\mathcal{R}_{\text{rob}}(\boldsymbol{\theta}) \leq \mathbb{E}_{(\mathbf{X},Y)}\phi(Y f_{\boldsymbol{\theta}}(\mathbf{X})) + \mathbb{E}_{\mathbf{X}}\phi(f_{\boldsymbol{\theta}}(\mathbf{X}) f_{\boldsymbol{\theta}}(z(\mathbf{X}))/\lambda). \tag{10}$$

Note that the upper bounds (8) and (9) are tighter than TRADES [36].

3.2. Algorithm

By modifying the upper bounds in Theorems 3.1 and 3.2, we propose the corresponding SS-AT algorithm. The modifications are as follows:

- the adversarial example $z(\boldsymbol{x})$ is replaced by $\widehat{\boldsymbol{x}}^{\text{pgd}}$ obtained by the PGD algorithm;
- the term $\mathbb{1}(Y \neq F_{\boldsymbol{\theta}}(\mathbf{X}))$ is replaced by the smooth cross-entropy $\ell^{\text{LS}}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y)$, where $\ell^{\text{LS}}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y) = -\boldsymbol{y}_{\alpha}^{\text{LS}\top} \log \mathbf{p}_{\theta}(\cdot|\boldsymbol{x})$, $\boldsymbol{y}_{\alpha}^{\text{LS}} = (1-\alpha)\mathbf{u}_y + \frac{\alpha}{C}\mathbf{1}_C$, $\mathbf{u}_y \in \mathbb{R}^C$ is the one-hot vector whose the y-th entry is 1 and $\mathbf{1}_C \in \mathbb{R}^C$ is the vector whose entries are all 1;
- the term $\mathbb{1}(F_{\boldsymbol{\theta}}(\mathbf{X}) \neq F_{\boldsymbol{\theta}}(z(\mathbf{X})))$ is replaced by $\lambda \cdot \text{KL}(\mathbf{p}_{\boldsymbol{\theta}}(\cdot|\mathbf{x})||\mathbf{p}_{\boldsymbol{\theta}}(\cdot|\widehat{\mathbf{x}}^{\text{pgd}}))$ for a regularization parameter $\lambda > 0$;
- the terms $p(Y \neq F_{\boldsymbol{\theta}}(\widehat{\mathbf{X}}^{\text{pgd}})|\mathbf{X})$ and $p(Y = F_{\boldsymbol{\theta}}(\mathbf{X})|\mathbf{X})$ are replaced by $1 - \sum_{c=1}^{C} p_{\boldsymbol{\theta}_T}(Y = c|\widehat{\mathbf{x}}^{\text{pgd}}) p_{\boldsymbol{\theta}}(c|\widehat{\mathbf{x}}^{\text{pgd}})$ and $\sum_{c=1}^{C} p_{\boldsymbol{\theta}_T}(Y = c|\mathbf{x}) p_{\boldsymbol{\theta}}(c|\mathbf{x})$, respectively.

The upper bounds in Theorems 3.1 and 3.2 have the conditional probabilities $p(F_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}^{\text{pgd}})|\boldsymbol{x})$ and $p(F_{\boldsymbol{\theta}}(\boldsymbol{x})|\boldsymbol{x})$. We estimate these conditional probabilities by smooth proxies as following:

$$\begin{aligned} p(Y \neq F_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}^{\text{pgd}})|\boldsymbol{x}) &= 1 - p(Y = F_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}^{\text{pgd}})|\boldsymbol{x}) \\ &= 1 - \sum_{c=1}^{C} p(Y = c|\boldsymbol{x})\mathbb{1}(c = F_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}^{\text{pgd}})) \\ &\approx 1 - \sum_{c=1}^{C} p_{\boldsymbol{\theta}_T}(Y = c|\boldsymbol{x}) p_{\boldsymbol{\theta}}(c|\widehat{\boldsymbol{x}}^{\text{pgd}}), \\ p(Y = F_{\boldsymbol{\theta}}(\boldsymbol{x})|\boldsymbol{x}) &= \sum_{c=1}^{C} p(Y = c|\boldsymbol{x})\mathbb{1}(c = F_{\boldsymbol{\theta}}(\boldsymbol{x})) \\ &\approx \sum_{c=1}^{C} p_{\boldsymbol{\theta}_T}(Y = c|\boldsymbol{x}) p_{\boldsymbol{\theta}}(c|\boldsymbol{x}). \end{aligned}$$

where $\boldsymbol{\theta}_T$ is a parameter of the teacher model.

We use the label smooth cross entropy instead of the standard cross entropy during training phase because the use of standard cross-entropy induces overconfident predictions [12]. To accurately estimate the conditional predictive probability $\mathbf{p}_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})$ in modified version of upper bounds in Theorems 3.1 and 3.2, we use label smoothing cross-entropy as a surrogate for $\mathbb{1}(Y \neq F_{\boldsymbol{\theta}}(\mathbf{X}))$ [21].

Furthermore, we introduce a knowledge distillation term using the teacher model trained by a semi-supervised learning algorithm to facilitate soft pseudo-labeling instead of hard labeling which is used exiting works [3, 29].

Algorithm 1 Semi-Supervised Robust-Self-Training with Adaptively Weighted Regularziation Algorithm (SRST-AWR)

Inputs : network $f_{\boldsymbol{\theta}}$, training dataset $\mathcal{D}_l^{n_l} = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{d+1} : i = 1, \cdots, n_l\}$, $\mathcal{D}_{ul}^{n_{ul}} = \{\boldsymbol{x}_j \in \mathbb{R}^{d+1} : j = 1, \cdots, n_{ul}\}$, learning rate η, hyperparameters $(\alpha, \lambda, \beta, \gamma, \tau)$ of (11), number of epochs T, number of batch B, batch size K

Output : adversarially robust model $f_{\boldsymbol{\theta}}$

Train a teacher model $f_{\boldsymbol{\theta}_T}$ using a semi-supervised learning algorithm on $\mathcal{D}_l \cup \mathcal{D}_{ul}$
for $t = 1, \cdots, T$ **do**
  for $b = 1, \cdots, B$ **do**
    for $k = 1, \cdots, K$ **do**
      Generate $\widehat{\boldsymbol{x}}_{b,k}^{\text{pgd}}$ using PGD^{10} in (2) ; $\boldsymbol{x}_{b,k} \in \mathbb{R}^d$
    end for

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \frac{1}{K} \nabla_{\boldsymbol{\theta}} \mathcal{R}_{(\text{SRST-AWR})}(\boldsymbol{\theta}; \mathcal{D}_l^K \cup \mathcal{D}_{ul}^K) \text{ in (11)}$$

  end for
end for
Return $f_{\boldsymbol{\theta}}$

In summary, we combine the adpaptively weighted

Table 1: **Comparison SRST-AWR, RST and UAT++.** $\boldsymbol{\theta}_{\text{sup}}$ and $\boldsymbol{\theta}_{\text{semi}}$ are parameters of the models using supervised and semi-supervised learning algorithms, respectively. $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are labeled sample and unlabeled sample, respetively. x_k can be labeled or unlabeled sample having pseudo label $\widehat{y}_k$.

Method	Pseudo-labeling	Supervised Term	Regularization Term
UAT++	$\widehat{y}_j = F_{\boldsymbol{\theta}_{\text{sup}}}(\boldsymbol{x}_j)$	$\ell_{\text{ce}}(f_\theta(\widehat{\boldsymbol{x}}_k^{\text{pgd}}), y_k \text{ or } \widehat{y}_k)$	$\text{D}_{\text{KL}}(\mathbf{p}_{\bar{\boldsymbol{\theta}}}(\cdot\|\boldsymbol{x}_k)\|\|\mathbf{p}_{\boldsymbol{\theta}}(\cdot\|\widehat{\boldsymbol{x}}_k^{\text{pgd}}))$
RST	$\widehat{y}_j = F_{\boldsymbol{\theta}_{\text{sup}}}(\boldsymbol{x}_j)$	$\ell_{\text{ce}}(f_\theta(\boldsymbol{x}_k), y_k \text{ or } \widehat{y}_k)$	$\text{D}_{\text{KL}}(\mathbf{p}_{\boldsymbol{\theta}}(\cdot\|\boldsymbol{x}_k)\|\|\mathbf{p}_{\boldsymbol{\theta}}(\cdot\|\widehat{\boldsymbol{x}}_k^{\text{pgd}}))$
ARMOURED	Multi-view learning	$\ell_{\text{ce}}(f_\theta(\widehat{\boldsymbol{x}}_k^{\text{pgd}}), y_k \text{ or } \widehat{y}_k)$	$\mathcal{L}_{\text{DPP}}(\boldsymbol{x}_k, y_k \text{ or } \widehat{y}_k), \mathcal{L}_{\text{NEM}}(\boldsymbol{x}_k, y_k \text{ or } \widehat{y}_k)$
SRST-AWR	Soft pseudo-labeling	$\ell_{\text{ce}}(f_\theta(\boldsymbol{x}_i), y_i)$	$\text{D}_{\text{KL}}(\mathbf{p}_{\boldsymbol{\theta}}(\cdot\|\boldsymbol{x}_j)\|\|\mathbf{p}_{\boldsymbol{\theta}}(\cdot\|\widehat{\boldsymbol{x}}_j^{\text{pgd}})) \times w_{\boldsymbol{\theta}}(\boldsymbol{x}_j)$

surrogate loss and knowledge distillation with a semi-supervised teacher. We call the algorithm Semi-Robust-Self-Training with Adaptively Weighted Regularization (SRST-AWR) which minimizes the following regularized empirical risk:

$$\begin{aligned}
&\mathcal{R}_{\text{(SRST-AWR)}}(\boldsymbol{\theta}; \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_l}, \{\boldsymbol{x}_j\}_{j=1}^{n_{ul}}, \alpha, \lambda, \beta, \gamma, \tau, \boldsymbol{\theta}_T) \\
&:= \frac{1}{n_l}\sum_{i=1}^{n_l} \ell_\alpha^{\text{LS}}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) \\
&+ \gamma \cdot \frac{1}{n_{ul}}\sum_{j=1}^{n_{ul}} \text{D}_{\text{KL}}(\mathbf{p}_{\boldsymbol{\theta}_T}^\tau(\cdot|\boldsymbol{x}_j)||\mathbf{p}_{\boldsymbol{\theta}}^\tau(\cdot|\boldsymbol{x}_j)) \\
&+ \lambda \cdot \frac{1}{n_{ul}}\sum_{j=1}^{n_{ul}} \cdot \text{D}_{\text{KL}}(\mathbf{p}_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x}_j)||\mathbf{p}_{\boldsymbol{\theta}}(\cdot|\widehat{\boldsymbol{x}}_j^{\text{pgd}})) \cdot w_{\boldsymbol{\theta}}(\boldsymbol{x}_j; \beta\, \boldsymbol{\theta}_T),
\end{aligned} \tag{11}$$

where

$$\begin{aligned}
w_{\boldsymbol{\theta}}(\boldsymbol{x}_j; \beta, \boldsymbol{\theta}_T) &= \beta \cdot \sum_{c=1}^{C} p_{\boldsymbol{\theta}_T}(Y = c|\boldsymbol{x}_j) p_{\boldsymbol{\theta}}(c|\boldsymbol{x}_j) \\
&+ (1-\beta) \cdot \left(1 - \sum_{c=1}^{C} p_{\boldsymbol{\theta}_T}(Y = c|\boldsymbol{x}_j) p_{\boldsymbol{\theta}}(c|\widehat{\boldsymbol{x}}_j^{\text{pgd}})\right),
\end{aligned} \tag{12}$$

where $\beta \in [0, 1]$ and τ is a temperature for knowledge distillation. The proposed semi-supervised adversarial training algorithm is summarized in Algorithm 1.

Comparison of SRST-AWR to RST [3] The proposed SRST-AWR algorithm differs from RST [3] in three main ways. Firstly, SRST-AWR employs a semi-supervised teacher that is trained with both labeled and unlabeled data, whereas RST utilizes a supervised teacher that is trained only on labeled data. Second, RST uses hard targets (one-hot labels) as pseudo-labels, while SRST-AWR employs soft targets (predictive probabilities) via knowledge distillation. The parameter τ in (11) regulates the smoothness of the pseudo-labels, where smaller values result in harder targets (using one-hot labels) and larger values result in softer targets (using predictive probabilities via knowledge distillation). Finally, SRST-AWR uses a new regularized empirical risk motivated by two upper bounds (6) and (7) for the robust risk. Table 1 provides the comparison of SRST-AWR, RST, UAT++ and ARMOURED.

Knowledge Distillation in Semi-Supervised Learning We have observed that the performance of a student model does not surpass the teacher model in the vanilla semi-supervised setting (i.e., without adversarial robust training). See Appendix C.1 for empirical evidences. However, we have found that applying knowledge distillation with a semi-supervised teacher can improve the adversarial robustness and generalization performance of the student model simultaneously in SS-AT.

Interpretation of $w_{\boldsymbol{\theta}}(\boldsymbol{x}; \beta, \boldsymbol{\theta}_T)$ We set β to 1/2. Then,

$$\begin{aligned}
2w_{\boldsymbol{\theta}}(\boldsymbol{x}; \beta, \boldsymbol{\theta}_T) &= \sum_{c=1}^{C} p_{\boldsymbol{\theta}_T}(Y = c|\boldsymbol{x}) p_{\boldsymbol{\theta}}(c|\boldsymbol{x}) \\
&+ \left(1 - \sum_{c=1}^{C} p_{\boldsymbol{\theta}_T}(Y = c|\boldsymbol{x}_j) p_{\boldsymbol{\theta}}(c|\widehat{\boldsymbol{x}}^{\text{pgd}})\right) \\
&= \langle p_{\boldsymbol{\theta}_T}(\cdot|\boldsymbol{x}), p_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})\rangle + (1 - \langle p_{\boldsymbol{\theta}_T}(\cdot|\boldsymbol{x}), p_{\boldsymbol{\theta}}(\cdot|\widehat{\boldsymbol{x}}^{\text{pgd}})\rangle \\
&\approx \underbrace{\text{corr}(p_{\boldsymbol{\theta}_T}(\cdot|\boldsymbol{x}), p_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x}))}_{:=\text{(a)}} + (1 - \underbrace{\text{corr}(p_{\boldsymbol{\theta}_T}(\cdot|\boldsymbol{x}), p_{\boldsymbol{\theta}}(\cdot|\widehat{\boldsymbol{x}}^{\text{pgd}}))}_{:=\text{(b)}})
\end{aligned}$$

Our $w_{\boldsymbol{\theta}}(\boldsymbol{x}; \beta, \boldsymbol{\theta}_T)$ in (11) is designed to impose more weights when the current prediction on a clean sample is highly correlated to the prediction of the teacher model (i.e. (a) part) and/or the current prediction on adversarial example is lowly correlated to the prediction of the teacher (i.e. (b) part).

4. Experiments

In this section, we report the performance of our algorithm and compare it with other competitors (Section 4.2), investigate how each component of our algorithm affects the performance improvement (Section 4.3), and evaluate the performance of our algorithm in comparison to supervised adversarial training algorithms (Section 4.3.5). The code is available at https://github.com/dyoony/SRST_AWR.

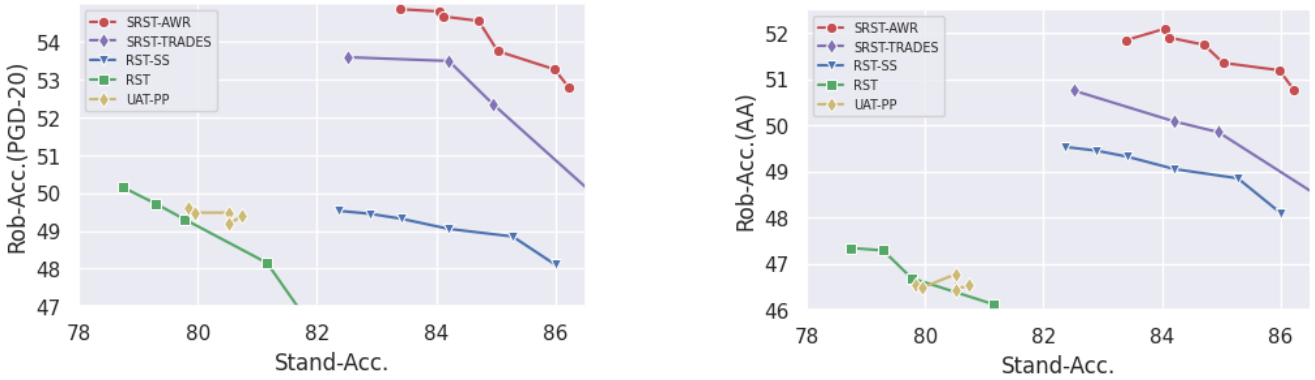

Figure 2: **Comparison of SRST-AWR, SRST-TRADES, Semi-RST, RST and UAT++ for varying** λ. The x-axis and y-axis are the standard and robust accuracies, respectively. The adversarial attacks for robust accuracies are PGD^{20} in the left panel and AutoAttack in right panel.

4.1. Experimental Setup

Training Setup The datasets are normalized into [0, 1]. We consider the three architectures - WideResNet-28-5 (WRN-28-5) [33] for CIFAR-10 [16] and STL-10 [5], WideResNet-28-2 (WRN-28-2) for SVHN [22] and WideResNet-28-8 (WRN-28-8) for CIFAR-100 [16], respectively. The architecture of the teacher network is same as that of the student and the training algorithm for teacher networks is FixMatch [27]. Details for training the teacher model are summarized in Appendix B.3.

We retain 4,000 samples for CIFAR-10 and CIFAR-100 and 1,000 samples for SVHN and STL-10 as labeled data from official train data, and use the remaining data for unlabeled data. For training prediction models, the SGD with momentum 0.9, weight decay 5×10^{-4}, the initial learning rate 0.1 on CIFAR-10 and 0.05 on CIFAR-100, SVHN and STL-10 are used. The total epochs is 200 and the learning rate is multiplied by 0.1 after each 50 and 150 epoch. The batch size of labeled data and unlabeled data are 64 and 128, respectively. For CIFAR-10, CIFAR-100 and STL-10, the random crop and random horizontal flip with probability 0.5 are applied for data augmentation. Stochastic weighting average (SWA) [15] is applied after 50-epochs. We select the models with having a maximum robust accuracy against PGD^{10} on test set. For training SRST-AWR, we set $(\alpha, \gamma, \beta, \tau)$ to (0.2, 4, 0.5, 1.2) and select λ maximizing the robust accuracy.

In maximization step, PGD^{10} with random start, $p = \infty$, $\varepsilon = 8/255$ and $\nu = 2/255$ is used. The final model is set to be the best model against PGD^{10} on the validation set among those obtained until 200 epochs. Experimental details are summarized in Appendix B.

Evaluation Setup For evaluating adversarial robustness, we set the maximum perturbation ε to $8/255$. We implement PGD^{20} and Auto-Attack (AA) [7]. Auto-Attack is ensemble of four attack - APGD, APGD-DLR, FAB [6] and Square Attack [1]. Among them, AGPD-DLR and Square Attack are effective to check the gradient masking [14].

4.2. Performance Evaluation

Comparison of SRST-AWR to RST and UAT++ We compare SRST-AWR to RST and UAT++. Since [29] observes that UAT++ outperforms UAT-FT, we only compare UAT++ among UAT algorithms. Figure 1 shows the performance with varying the number of labeled data. When the labeled data are scarce, SRST-AWR is far superior to the other competitors. SRST-AWR maintains performance even though the size of labeled data is very small. In addition, Figure 2 shows the trade-off between the standard and robust accuracies as the regularization parameter λ varies for SRST-AWR, RST and UAT++. Moreover, SRST-AWR uniformly outperforms the other competitors with large margins. Table 2 shows that SRST-AWR outperforms the other semi-supervised adversarial training algorithms for various benchmark data sets in terms of both standard and robust accuracies. Experimental details of Table 2 are provided in Appendix B.1.

Comparison to ARMOURED [18] We compare SRST-AWR with ARMOURED. As the official code of ARMOURED is not available, we set the architecture of SRST-AWR to be equal to that of ARMOURED and cite the performance measures of ARMOURED (standard accuracies, PGD-7, and AA) reported in the original paper. It is observed that the robust accuracy against PGD of ARMOURED is high, but not against AA, mainly due to gradi-

Table 2: **Comparison to RST and UAT++**. We conduct the experiment three times with different seeds and present the averages of the accuracies with the standard errors in the parenthesis.

Method	CIFAR-10 (WRN-28-5)			CIFAR-100 (WRN-28-8)		
	Stand	**PGD^{20}**	**AA**	**Stand**	**PGD^{20}**	**AA**
RST	79.77(0.06)	49.31(0.09)	46.69(0.02)	47.36(0.06)	17.32(0.06)	14.65(0.19)
UAT++	79.84(0.23)	49.61(0.07)	46.53(0.11)	46.99(0.18)	22.13(0.04)	19.69(0.07)
SRST-AWR	**84.56**(0.06)	**54.41**(0.15)	**51.72**(0.06)	**54.25**(0.04)	**30.90**(0.07)	**25.58**(0.05)
Method	SVHN (WRN-28-2)			STL-10 (WRN-28-5)		
	Stand	**PGD^{20}**	**AA**	**Stand**	**PGD^{20}**	**AA**
RST	86.47(0.51)	52.62(0.44)	42.38(0.48)	60.81(0.32)	35.86(0.37)	34.00(0.22)
UAT++	91.07(0.08)	52.54(0.16)	46.81(0.13)	58.56(0.28)	43.00(0.21)	40.08(0.24)
SRST-AWR	**91.86**(0.19)	**57.02**(0.73)	**50.84**(0.83)	**89.61**(0.21)	**73.93**(0.28)	**69.99**(0.27)

Table 3: **Comparison to ARMOURED**. The performance measures of ARMOURED are cited from original paper i.e. ARMOUREDs are not reimplemented. Maximum perturbation size εs are set to be 8/255 and 4/255 on CIFAR-10 and SVHN, respectively. This settings are identical to ARMOURED [18]. We conduct the experiment three times with different seeds and present the averages of the accuracies with the standard errors in the parenthesis.

Method	CIFAR-10 (WRN-28-2)			SVHN (WRN-28-2)		
	Stand	**PGD^{7}**	**AA**	**Stand**	**PGD^{7}**	**AA**
ARMOURED-F+AT	76.76(1.60)	**55.12**(4.90)	35.24(4.56)	92.44(0.64)	62.10(8.39)	23.35(3.22)
SRST-AWR	**81.78**(0.03)	52.43(0.10)	**47.41**(0.11)	**95.30**(0.05)	**81.19**(0.03)	**78.21**(0.03)

ent masking [24]. Causing gradient masking is considered as a failed defense method in terms of adversarial robustness since adversarial examples are easily generated from a model with gradient masking. Since AA contains several attack algorithms that break down models with gradient masking, it is a more reliable evaluation protocol for adversarial robustness. Table 3 shows that SRST-AWR outperforms ARMOURED with significant margins for CIFAR-10 and SVHN. Experimental details of Table 3 are provided in Appendix B.1.

4.3. Ablation Studies

In this section, we investigate how each component of the objective function (11) for SRST-AWR has influence on performance - (1) effect of the semi-supervised teacher, (2) effect of the $w_{\boldsymbol{\theta}}(\boldsymbol{x}; \beta, \boldsymbol{\theta}_T)$, (3) effect of the knowledge distillation (i.e. soft pseudo-labeling), (4) sensitivity of the regularization parameter λ. Also, we report (5) the results for fully labeled data setting to justify tightness of our bound.

Additionally, we provide the sensitivity analysis of parameter β and τ in Appendix C.3 and C.4, respectively. Unless otherwise stated, the ablation studies are implemented on CIFAR-10 with 4,000 labeled data and $(\lambda, \gamma, \tau, \beta) = (20, 4, 1.2, 0.5)$ in (11) are used.

4.3.1 Effect of Semi-Supervised Teacher

Table 4 shows that using a semi-supervised teacher improves the performance of RST [3] and UAT++ [29] on both standard and robust accuracies. The enhancement is accomplished by assigning labels to the unlabeled data more correctly. When using a supervised teacher, the two methods show comparable performance. However, when using a semi-supervised teacher, RST outperforms UAT++ in terms of robustness, while UAT++ performs better than RST in terms of generalization. However, SRST-AWR still outperforms RST and UAT++ with the semi-supervised teacher.

Table 4: **Effect of Semi-supervised Teacher**. We conduct the experiments three times with different seeds and present the averages of the accuracies with the standard errors in the parenthesis.

Method	CIFAR-10 (WRN-28-5)		
	Stand	**PGD^{20}**	**AA**
RST w/ sup.	79.77(0.06)	49.31(0.09)	46.69(0.02)
RST w/ semi-sup.	83.41(0.08)	51.94(0.04)	49.32(0.03)
UAT++ w/ sup.	79.84(0.23)	49.61(0.07)	46.53(0.11)
UAT++ w/ semi-sup.	84.49(0.15)	51.18(0.10)	48.46(0.07)
RST-AWR	82.40(0.05)	52.16(0.11)	49.34(0.08)
SRST-AWR	**84.56**(0.06)	**54.41**(0.15)	**51.72**(0.06)

4.3.2 Effect of $w_{\boldsymbol{\theta}}(\boldsymbol{x}; \beta, \boldsymbol{\theta}_T)$ in Low-Label Regime

In this subsection, we compare the SRST-TRADES, which is SRST-AWR with $w_{\boldsymbol{\theta}}(\boldsymbol{x}; \beta, \boldsymbol{\theta}_T) = 1$ for all $\boldsymbol{x}$, and SRST-AWR for confirming the role of our proposed weight $w_{\boldsymbol{\theta}}(\boldsymbol{x}; \beta, \boldsymbol{\theta}_T)$ in low-label regime. The number of labeled data is 500, 2,000, 100 and 1,000 on CIFAR-10, CIFAR-

Table 5: **Effect of** $w_{\boldsymbol{\theta}}(\boldsymbol{x}; \beta, \boldsymbol{\theta}_T)$ in Low-Label Regime. We conduct the experiment three times with different seeds and present the averages of the accuracies with the standard errors in the parenthesis.

Method	CIFAR-10 (WRN-28-5)		
	Stand	**PGD^{20}**	**AA**
SRST-TRADES	79.09(0.10)	45.51(0.15)	43.93(0.14)
SRST-AWR	**82.93**(0.06)	**52.54**(0.15)	**50.76**(0.06)
Method	CIFAR-100 (WRN-28-8)		
	Stand	**PGD^{20}**	**AA**
SRST-TRADES	29.74(0.15)	16.51(0.21)	14.43(0.14)
SRST-AWR	**33.65**(0.14)	**18.64**(0.17)	**16.77**(0.15)
Method	SVHN (WRN-28-2)		
	Stand	**PGD^{20}**	**AA**
SRST-TRADES	72.85(0.11)	47.64(0.41)	45.08(0.31)
SRST-AWR	**80.09**(0.19)	**50.19**(0.73)	**48.69**(0.83)
Method	STL-10 (WRN-28-5)		
	Stand	**PGD^{20}**	**AA**
SRST-TRADES	88.19(0.12)	66.71(0.21)	63.41(0.25)
SRST-AWR	**90.61**(0.21)	**75.85**(0.28)	**72.51**(0.27)

Table 6: **Effect of Knowledge Distillation**. We conduct the experiments three times with different seeds and present the averages of the accuracies with the standard errors in the parenthesis.

Method	CIFAR-10 (WRN-28-5)		
	Stand	**PGD^{20}**	**AA**
SRST-AWR w/ KD	**84.56**(0.06)	**54.41**(0.15)	**51.72**(0.06)
SRST-AWR w/o KD	84.37(0.12)	52.77(0.09)	50.02(0.09)

100, SVHN and STL-10, respectively. Table 5 demonstrates that SRST-AWR significantly outperforms SRST-TRADES in terms of both standard and robust accuracies. We also compare the performance of SRST-TRADES and SRST-AWR with varying the number of labeled data on CIFAR-100 and STL-10 in Appendix C.2. Details of the experimental setup are provided in Appendix B.4.

4.3.3 Effect of Knowledge Distillation

To investigate the effect of knowledge distillation, we estimate pseudo labels using a semi-supervised teacher with hard labels. Table 6 shows the effect of Knowledge Distillation (KD) term in our algorithm. KD improves the robustness with retaining generalization performance.

4.3.4 Sensitivity Analysis on λ

Figure 2 shows the trade-off between standard accuracies and robust accuracies with respect to λ for each algorithm. It is observed that SRST-AWR uniformly dominates SRST-TRADES, RST-SS which is RST with the semi-supervised teacher considered in Section 4.3.1, RST [3], and UAT++ [29] regardless of the choice of the regularization parameter λ.

Table 7: **Comparison to Supervised Adversarial Training Algorithms using the whole labeled data.** We conduct the experiment three times with different seeds and present the averages of the accuracies with the standard errors in the parenthesis.

Method	**# Labels (%)**	CIFAR-10 (WRN-28-5)		
		Stand	**PGD^{20}**	**AA**
PGD-AT	100	85.96(0.17)	54.29(0.10)	50.84(0.09)
MART	100	80.98(0.28)	**57.56**(0.14)	51.06(0.03)
TRADES	100	83.90(0.04)	54.74(0.03)	51.72(0.10)
SRST-AWR	8	84.56(0.06)	54.41(0.15)	51.72(0.06)
SRST-AWR	12	**86.06**(0.05)	54.69(0.16)	**51.88**(0.08)

Table 8: **Performance in Fully Labeled Data**. We conduct the experiments three times with different seeds and present the averages of the accuracies with the standard errors in the parenthesis.

Method	CIFAR-10 (WRN-28-5)		
	Stand	**PGD^{20}**	**AA**
TRADES	83.90(0.04)	54.74(0.03)	51.72(0.10)
AWR	**87.01**(0.10)	**55.01**(0.11)	**51.97**(0.09)
TRADES-AWP	84.56(0.06)	54.41(0.15)	51.72(0.06)
AWR-AWP	**85.81**(0.09)	**57.03**(0.14)	**53.90**(0.12)

4.3.5 Comparison to Supervised Adversarial Training Algorithms

Table 7 shows that SRST-AWR with only 12% labeled data outperforms the supervised PGD-AT [19], TRADES [36], and MART [30] (i.e., trained with 100% labeled data) on both standard and robust accuracies against AA, while SRST-AWR with only 8% labeled data achieves comparable robust accuracies to the supervised TRADES, while outperforming on standard accuracy.

4.3.6 Performance on Fully Labeled Data

Table 8 shows the performance of TRADES-based methods - TRADES [36] and AWP [31]. AWR-based methods outperform TRADES-based methods for the given fully labeled data. Details of the experimental setup are provided in Appendix B.6.

5. Conclusion

In this paper, we derived the two upper bounds of the robust risk and developed a new semi-supervised adversarial training algorithm called SRST-AWR. The objective function of SRST-AWR is a combination of a new surrogate version of the robust risk and a knowledge distillation term with a semi-supervised teacher. While existing algorithms

show significant performance degradation as the number of labeled data decreases, our proposed algorithm has shown relatively little performance degradation.

The experiments showed that SRST-AWR outperforms existing algorithms with significant margins in terms of both standard and robust accuracies on various benchmark datasets. Especially, we can achieve standard and robust accuracy that are comparable to supervised adversarial training algorithms when the number of labeled data is around 10% of the whole labeled data.

Acknowledgement This work was supported by a Convergence Research Center (CRC) grant funded by the Korean government (MSIT, No. 2022R1A5A708390811), by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C3A01003550-14) and by Samsung Electronics Co., Ltd.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random searchg. In ECCV, 2020. 2, 6

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In Conference on Neural Information Processing Systems (NeurIPS), 2019. 1

[3] Yair Carmon, Aditi Raghunathan, Schmidt Ludwig, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In Conference on Neural Information Processing Systems (NeurIPS), 2019. 1, 2, 4, 5, 7, 8

[4] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In ACM, 2017. 1

[5] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In AISTATS, 2011. 6

[6] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In International Conference on Machine Learning (ICML), 2020. 6

[7] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In International Conference on Machine Learning (ICML), 2020. 6

[8] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. In Computer Vision Pattern Recognition (CVPR), 2018. 2

[9] Yang Dongyoon, Kong Insung, and Kim Yongdai. Improving adversarial robustness by putting more regularizations on less robust samples. In International Conference on Machine Learning (ICML), 2023. 4, 11

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR), 2015. 2

[11] Sven Gowal, Po-Sen Huang, Aaron van den Oord, Timothy Mann, and Pushmeet Kohli. Self-supervised adversarial robustness for the low-label, high-data regime. In International Conference on Learning Representations (ICLR), 2021. 2

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In International Conference on Machine Learning (ICML), 2017. 4

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In Conference on Neural Information Processing Systems (NeurIPS) Workshop, 2015. 13

[14] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In International Conference on Machine Learning (ICML), 2018. 1, 6

[15] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. Proceedings of the international conference on Uncertainty in Artificial Intelligence, 2018. 6

[16] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 6

[17] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In International Conference on Learning Representations (ICLR), 2017. 1

[18] Kangkang Lu, , Cuong Manh Nguyen, Xun Xu, Kiran Krishnamachari, Yu Jing Goh, and Chuan-Sheng Foo. Armoured: Adversarially robust models using unlabeled data by regularizing diversity. In International Conference on Learning Representations (ICLR), 2021. 1, 2, 3, 6, 7

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations (ICLR), 2018. 1, 2, 8, 13

[20] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. 1, 2

[21] Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In Conference on Neural Information Processing Systems (NeurIPS), 2019. 4

[22] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011. 6

[23] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to

black-box attacks using adversarial samples. arXiv, 2016. 1, 2

[24] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In ACM, 2017. 1, 2, 7

[25] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. 2018 IEEE European Symposium on Security and Privacy (EuroS&P), 2018. 1

[26] Rahul Rade and Seyed-Mohsen Moosavi-Dezfolli. Recuding excessive margin to achieve a better accuracy vs. robustness trade-off. In International Conference on Learning Representations (ICLR), 2022. 1, 2

[27] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In Conference on Neural Information Processing Systems (NeurIPS), 2020. 1, 2, 6

[28] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations (ICLR), 2014. 1, 2

[29] Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Robert Stanforth, Alhussein Fawzi, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In Conference on Neural Information Processing Systems (NeurIPS), 2019. 1, 2, 3, 4, 6, 7, 8

[30] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In International Conference on Learning Representataions(ICLR)), 2020. 1, 2, 4, 8, 13

[31] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In Conference on Neural Information Processing Systems (NeurIPS), 2020. 8

[32] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In Conference on Neural Information Processing Systems (NeurIPS), 2020. 1

[33] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. Proceedings of the British Machine Vision Conference 2016, 2016. 6

[34] Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. In arXiv, 2019. 1, 2

[35] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In Conference on Neural Information Processing Systems (NeurIPS), 2021. 1

[36] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning (ICML), 2019. 1, 2, 3, 4, 8, 13

[37] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan S. Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In International Conference on Machine Learning (ICML), 2020. 1

[38] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan S. Kankanhalli. Geometry-aware instance-reweighted adversarial training. In International Conference on Learning Representations (ICLR), 2021. 1