PGD-2 CAN BE BETTER THAN FGSM + GRADALIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

One major issue of adversarial training (AT) with the fast gradient sign method (FGSM AT) is the phenomenon of catastrophic overfitting, meaning that the trained model suddenly loses its robustness over a single epoch. In addition to FGSM AT, Andriushchenko & Flammarion (2020) observed that two-step projected gradient descent adversarial training (PGD-2 AT) also suffers from catastrophic overfitting for large ℓ_∞ perturbations. To prevent catastrophic overfitting, Andriushchenko & Flammarion (2020) proposed a gradient alignment regularization method (GradAlign) and claimed that GradAlign can prevent catastrophic overfitting in FGSM AT and PGD-2 AT. In this paper, we show that PGD-2 AT with random initialization (PGD-2-RS AT) and attack step size $\alpha = 1.25\epsilon/2$ only needs approximately a half computational cost of FGSM + GradAlign AT and actually can avoid catastrophic overfitting for large ℓ_∞ perturbations. We hypothesize that, if FGSM-RS AT with $\alpha = 1.25\epsilon/2$ can avoid catastrophic overfitting for ℓ_∞ perturbation size $\epsilon/2$, then PGD-2-RS AT with $\alpha = 1.25\epsilon/2$ may be able to avoid catastrophic overfitting for ℓ_∞ perturbation size ϵ. Our intuitions to justify this empirical hypothesis induce a more unexpected finding: If we apply random noise from the uniform distribution $\mathcal{U}(-\epsilon/2, \epsilon/2)$ to the perturbations before each step of PGD-2 with $\alpha = 1.25\epsilon/2$, instead of initializing the perturbations with random noise from $\mathcal{U}(-\epsilon, \epsilon)$ at the beginning (*i.e.,* the conventional random initialization scheme), the corresponding AT method can also avoid catastrophic overfitting and even achieve better robust accuracy in most cases. We refer to this AT method as Qusai-PGD-2-RS AT. Extensive evaluations demonstrate that PGD-2-RS AT and Qusai-PGD-2-RS AT with $\alpha = 1.25\epsilon/2$ achieve better performance and efficiency than FGSM + GradAlign AT. Notably, Qusai-PGD-2-RS AT achieves comparable robust accuracy against PGD-50-10 as PGD-3-RS AT on CIFAR10 and SVHN, and it also achieves approximately 18% top-1 and 38% top-5 robust accuracy against PGD-50-10 at $\epsilon = 8/255$ on ImageNet.

1 INTRODUCTION

The past decade has witnessed tremendous achievements of deep learning in many application domains, such as computer vision Krizhevsky et al. (2012), natural language processing (Cho et al., 2014), and human-level control (Mnih et al., 2015). However, deep learning is also demonstrated to be vulnerable to adversarial examples, which are hardly distinguishable from natural samples according to human perception, but can mislead deep neural networks (DNNs) to make incorrect predictions with high confidence (Szegedy et al., 2013; Goodfellow et al., 2014). The vulnerability to adversarial examples is considered as a significant obstacle to the deployment of deep learning techniques in security-critical applications. Thus, the community has developed many defensive techniques against adversarial examples.

Among the existing defenses, adversarial training has survived the battles against many strong attacks (Madry et al., 2017; Athalye et al., 2018; Andriushchenko et al., 2019; Croce & Hein, 2019; 2020), achieving the overall best empirical performance. Although the community has also developed some certified defenses (Wong & Kolter, 2018; Cohen et al., 2019; Mirman et al., 2018) against adversarial examples, the empirical performance of those certified defenses is usually not comparable to adversarial training. To this end, many recent studies focus on adversarial training (Zhang et al., 2019b; Carmon et al., 2019; Ding et al., 2019; Wang et al., 2020; Rice et al., 2020; Zhang et al., 2020). In those studies, adversarial training is consistently formulated as a min-max problem:

The inner maximization problem aims to find adversarial examples that maximize a surrogate loss, and the outer minimization problem is to minimize the surrogate loss on the adversarial examples by optimizing the model parameters. Projected gradient descent (PGD) (Madry et al., 2017) is currently the most popular and effective approximation method to solve the inner problem. We refer to adversarial training with k-step PGD to solve the inner maximization problem as PGD-k adversarial training (PGD-k AT). However, since conventional PGD-k AT always requires many additional forward and backward propagations (*e.g.,* 10 additional propagations in PGD-10 AT) to generate adversarial examples in each training step, which induces an order of magnitude more computational cost than standard training and thus impedes the broader application of adversarial training.

To improve the scalability of adversarial training, some recent works focus on accelerating adversarial training (Shafahi et al., 2019; Zhang et al., 2019a; Wong et al., 2019; Andriushchenko & Flammarion, 2020). Notably, Wong et al. (2019) found that FGSM with random initialization and a proper step size, which only needs one additional forward and backward propagation to generate training adversarial examples, can achieve comparable defensive performance as PGD adversarial training. We refer to this fast adversarial training method as FGSM-RS AT. Although FGSM-RS AT has a very low cost, it may suffer from the phenomenon of catastrophic overfitting, where the trained model suddenly loses robustness against PGD in the training process. Specifically, Andriushchenko & Flammarion (2020) (NeurIPS 2020) observed that for large ℓ_∞ perturbations, both FGSM-RS AT and PGD-2 AT suffer from catastrophic overfitting. Andriushchenko & Flammarion (2020) conjectured that local non-linearity is the cause of catastrophic overfitting and proposed a regularization method to enhance the trained model's local linearity by maximizing the alignment of the gradients in the perturbation sets, referred to as gradient alignment regularization (GradAlign). Andriushchenko & Flammarion (2020) claimed that GradAlign can prevent catastrophic overfitting in both FGSM AT and PGD-2 AT (without random initialization). Besides, Kim et al. (2021) (AAAI 2021) proposed to search for the appropriate step size for each data sample to improve fast adversarial training, and this method is referred to as stable single-step adversarial training.

In this paper, we show that PGD-2 AT with random initialization (PGD-2-RS AT) and a proper attack step size α actually can avoid catastrophic overfitting for large ℓ_∞ perturbations (*e.g.,* $\epsilon = 16/255$ on CIFAR10 and $\epsilon = 12/255$ on SVHN). Notably, compared to FGSM + GradAlign AT, PGD-2-RS AT with $\alpha = 1.25\epsilon/2$ only needs approximately a half computational cost but can achieve better standard accuracy and comparable or better robust accuracy. Furthermore, we propose and justify an empirical hypothesis with intuitions and extensive evaluations to shed light on why PGD-2-RS AT with $\alpha = 1.25\epsilon/2$ can avoid catastrophic overfitting for large ℓ_∞ perturbations. Specifically, we hypothesize that, *if FGSM-RS AT with attack step size $\alpha = 1.25\epsilon/2$ can avoid catastrophic overfitting for ℓ_∞ perturbation size $\epsilon/2$, then under appropriate training settings, PGD-2-RS AT with $\alpha = 1.25\epsilon/2$ may be able to avoid catastrophic overfitting for ℓ_∞ perturbation size ϵ.* Our main intuition to propose this empirical hypothesis is that, if FGSM-RS with $\alpha = 1.25\epsilon/2$ can generate *qualified* adversarial examples for adversarial training with ℓ_∞ perturbation size $\epsilon/2$, then PGD-2-RS with $\alpha = 1.25\epsilon/2$ may be able to generate *qualified* adversarial examples for adversarial training with ℓ_∞ perturbation size ϵ. This is because PGD-2-RS can be *approximately* viewed as two-step FGSM-RS, with each step generating qualified adversarial examples of the previous samples for adversarial training with ℓ_∞ perturbation size $\epsilon/2$. Then, the adversarial examples generated by the second step may be qualified adversarial examples of the original samples for adversarial training with a total ℓ_∞ perturbation size ϵ [*]. Note that if the gap between the loss of the adversarial examples (generated by FGSM-RS or PGD-2-RS) and the loss of the adversarial examples generated by the PGD attack used for conventional PGD adversarial training (*e.g.,* PGD-10-RS) is small, we refer to those adversarial examples as *qualified* adversarial examples for adversarial training.

We verify our hypothesis by extensive evaluations on CIFAR10, SVHN, and ImageNet. Given the fact that FGSM-RS AT can avoid catastrophic overfitting at $\epsilon = 8/255$ on CIFAR10, $\epsilon = 6/255$ on SVHN, and $\epsilon = 4/255$ on ImageNet **with the training settings in (Wong et al., 2019; Andriushchenko & Flammarion, 2020)**, we show PGD-2-RS AT with $\alpha = 1.25\epsilon/2$ can avoid catastrophic overfitting and train robust models at $\epsilon = 16/255$ on CIFAR10, $\epsilon = 12/255$ on SVHN, and $\epsilon = 8/255$ on ImageNet. In addition, our intuitions induce a more unexpected finding: If we apply random noise from $\mathcal{U}(-\epsilon/2, \epsilon/2)$ to the perturbations for each step of PGD-2 with $\alpha = 1.25\epsilon/2$, instead of initializing the perturbations with random noise from $\mathcal{U}(-\epsilon, \epsilon)$ at the beginning,

[*]The ℓ_∞ perturbation *size* is "addable": $\| \boldsymbol{\delta}_1 + \boldsymbol{\delta}_2 \|_\infty \leq \epsilon$ if $\| \boldsymbol{\delta}_1 \|_\infty \leq \epsilon/2$ and $\| \boldsymbol{\delta}_2 \|_\infty \leq \epsilon/2$.

we can also prevent catastrophic overfitting and train models with better robust accuracy for large ℓ_∞ perturbations. We refer to this modified PGD-2-RS AT method as *Qusai-PGD-2-RS AT*.

We conduct an extensive array of experiments to compare PGD-2-RS AT and Qusai-PGD-2-RS AT with FGSM-RS AT (Wong et al., 2019), FGSM + GradAlign AT (Andriushchenko & Flammarion, 2020), stable single-step adversarial training (Kim et al., 2021), and free adversarial training (AT for Free) (Shafahi et al., 2019). We show that PGD-2-RS AT and Qusai-PGD-2-RS AT achieve consistently better performance than the other efficient adversarial training methods. Notably, PGD-2-RS AT and Qusai-PGD-2-RS AT can avoid catastrophic overfitting and train robust models at $\epsilon = 8/255$ on ImageNet. In contrast, Andriushchenko & Flammarion (2020) only reported the results of FGSM + GradAlign AT at $\epsilon = 6/255$ on ImageNet; Shafahi et al. (2019); Wong et al. (2019) only reported the results of free adversarial training and FGSM-RS AT at $\epsilon = 2/255$ and $\epsilon = 4/255$. In our testbed, FGSM + GradAlign AT suffers from catastrophic overfitting, and AT for Free achieves only 7.3% top-1 robust accuracy against PGD-50-10, at $\epsilon = 8/255$ on ImageNet.

2 Problem Overview

In this paper, we mainly consider the task of classification over $(\mathbf{x}, y) \in \mathcal{D}$, where $\mathbf{x}$ denotes a data sample, and y denotes the ground-truth label. $\mathcal{D}$ is the data-generating distribution. Given an input $\mathbf{x} \in R^d$ with dimension d, we denote a neural network by $\mathbf{f}(\mathbf{x}) \triangleq \{\mathbf{f}_{\boldsymbol{\theta},k}(\mathbf{x})\}_{k=1,2,\dots K}$, where $\boldsymbol{\theta}$ and $\mathbf{f}_{\boldsymbol{\theta},k}(\mathbf{x})$ represent the network parameters and the logit output for the k-th class, respectively. The network predicts the label as $\arg_k \max \mathbf{f}_{\boldsymbol{\theta},k}(\mathbf{x})$. Let $\ell(\mathbf{x}, y, \boldsymbol{\theta})$ denote the loss of the model on $(\mathbf{x}, y)$. For simplicity, we may rewrite $\ell(\mathbf{x}, y, \boldsymbol{\theta})$ as $\ell(\mathbf{x})$ or $\ell(\mathbf{x}, y)$. By default, we use cross entropy as the loss. A data sample $\mathbf{x}'$ is considered as an adversarial example if the prediction on $\mathbf{x}'$ is wrong, *i.e.,* $\arg_k \max \mathbf{f}_{\boldsymbol{\theta},k}(\mathbf{x}') \neq y$, and $\mathbf{x}'$ is close to the original sample $\mathbf{x}$ according to certain distance metric $d(\cdot, \cdot)$, *i.e.,* $d(\mathbf{x}, \mathbf{x}') \leq \epsilon$. Previous work usually employs ℓ_p-norm as the distance metric, and in this regard, the above constraint could be denoted by $\mathbf{x}' \in \mathbb{B}_\epsilon^p(\mathbf{x})$, where $\mathbb{B}_\epsilon^p(\mathbf{x})$ is an ℓ_p-norm ball centered at $\mathbf{x}$ with radius ϵ. In this paper, we mainly focus on the ℓ_∞-norm metric. Also, we denote the uniform distribution that takes the value within the range a to b by $\mathcal{U}(a, b)$ or $Uniform(a, b)$. In general, adversarial training can be formulated as a min-max problem:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\max_{\mathbf{x}'\in\mathbb{B}_\epsilon^p(\mathbf{x})} \ell(\mathbf{x}', y, \boldsymbol{\theta})]. \tag{1}$$

The inner maximization problem is to find adversarial examples that can maximize the loss, which is solved by the fast gradient sign method (FGSM) in fast adversarial training, or projected gradient descent (PGD) in conventional PGD adversarial training. Specifically, FGSM solves the inner problem by updating $\mathbf{x}$ along the sign direction of the gradient of the loss w.r.t. $\mathbf{x}$, *i.e.,*

$$\mathbf{x}' = \Pi_{\mathbb{B}_\epsilon^p(\mathbf{x})}\{\mathbf{x} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \ell(\mathbf{x}, y, \boldsymbol{\theta}))\}\ , \tag{2}$$

where $\Pi_{\mathbb{B}_\epsilon^p(\mathbf{x})}$ projects the input into $\mathbb{B}_\epsilon^p(\mathbf{x})$ (and also the valid data range); α denotes the attack step size; sign function outputs the $sign$ of the input. PGD-k solves the inner problem by the iterative execution of (2) for k steps with a smaller α. The outer minimization problem aims to minimize the loss w.r.t. the network parameters, which can be solved by several optimization methods.

3 Computational Complexity Analysis

We first compare the computational complexity of FGSM + AT, PGD-2 AT (PGD-2-RS AT), FGSM + GradAlign AT to illustrate why PGD-2 AT has approximately a half computational cost of FGSM + GradAlign AT. **We mainly consider the forward and backward propagation operations to estimate the computational cost.** Compared to those two operations, the other operations have less computational demand. Suppose the computational cost of a forward and a backward propagation operation on one data sample are respectively c_1 and c_2, and the size of a minibatch is N. In each training step, FGSM AT needs one forward and backward propagation to generate adversarial examples, and one for updating the model parameters. So for FGSM AT, the cost of one training step is approximately $2(c_1 + c_2)N$. Likewise, for PGD-2 AT, the cost of one training step is approximately $3(c_1 + c_2)N$. Thus, PGD-2 AT consumes about $1.5\times$ computational cost of FGSM AT. Due to double backpropagation (Etmann, 2019), FGSM + GradAlign AT consumes approximately $3\times$ computational cost of FGSM AT on a GPU (Andriushchenko & Flammarion, 2020). We refer

Methods on CIFAR10	$\epsilon = 8/255$		$\epsilon = 16/255$	
	Standard	PGD-50-10	Standard	PGD-50-10
PGD-2-NRS AT ($\alpha = \epsilon/2$)	$80.78\% \pm 0.39\%$	$49.37\% \pm 0.25\%$	$66.73\% \pm 3.30\%$	$12.74\% \pm 13.62\%$
PGD-2-NRS AT ($\alpha = 1.25\epsilon/2$)	$80.71\% \pm 0.52\%$	$49.47\% \pm 0.41\%$	$53.75\% \pm 16.71\%$	$22.70\% \pm 11.49\%$
PGD-2-RS AT ($\alpha = \epsilon/2$)	$84.72\% \pm 0.25\%$	$45.63\% \pm 0.42\%$	$71.44\% \pm 0.38\%$	$24.11\% \pm 0.43\%$
PGD-2-RS AT ($\alpha = 1.25\epsilon/2$)	$82.98\% \pm 0.32\%$	$47.79\% \pm 0.23\%$	$67.77\% \pm 0.52\%$	$26.22\% \pm 0.37\%$
FGSM-RS AT ($\alpha = 1.25\epsilon$)	$83.47\% \pm 0.25\%$	$46.08\% \pm 0.45\%$	$46.88\% \pm 24.86\%$	$0.00\% \pm 0.00\%$
FGSM + GradAlign AT	$81.21\% \pm 0.52\%$	$47.60\% \pm 0.34\%$	$60.08\% \pm 0.40\%$	$25.80\% \pm 0.74\%$

Table 1: For the sake of clarity, we refer to PGD-2 AT without random initialization as PGD-2-NRS AT and PGD-2 AT with random initialization as PGD-2-RS AT. For PGD-2-RS AT, the random noise used for initializing the perturbation is sampled from uniform distribution $\mathcal{U}(-\epsilon, \epsilon)$. Different from (Andriushchenko & Flammarion, 2020), which evaluates the models' adversarial (robust) accuracy against PGD-50-10 on 1000 random points, we evaluate models' robust accuracy on the whole test set. The experimental settings are detailed in Section 5.1 and Section C.

the interested readers to (Etmann, 2019; Andriushchenko & Flammarion, 2020) for more details. Thus, the computational cost of PGD-2 AT is about 50% less than that of FGSM + GradAlign AT. Actually, PGD-5-RS AT has a similar cost as FGSM + GradAlign AT on a GPU. Besides, although AT for Free (Shafahi et al., 2019) takes less time than FGSM-RS AT and PGD-2-RS AT per training step, it usually needs much more training steps (mainly due to minibatch replays) to achieve comparable performance (if training the model with mixed precision/half precision) (Wong et al., 2019; Andriushchenko & Flammarion, 2020). Thus, AT for Free usually needs more computational time than FGSM-RS AT and PGD-2-RS AT in practice. We note that the empirical results on CIFAR10 and SVHN in Table 4 have verified the above analysis.

4 PGD-2 Adversarial Training (AT)

4.1 Revisiting PGD-2 AT

In general, PGD-2 AT updates the adversarial examples for two steps following (2) with attack step size α and then updates the model parameters on the adversarial examples in each training step. We refer PGD-2 with random initialization as PGD-2-RS and PGD-2 without random initialization as PGD-2-NRS. Andriushchenko & Flammarion (2020) evaluated PGD-2 AT on CIFAR10 and found that PGD-2-NRS AT with $\alpha = \epsilon/2$ achieves better performance than PGD-2-RS AT with $\alpha = \epsilon/2$ or $\alpha = \epsilon$ at $\epsilon = 8/255$ on CIFAR10. Thus, Andriushchenko & Flammarion (2020) selected PGD-2-NRS AT with $\alpha = \epsilon/2$ as the PGD-2 AT baseline throughout their paper. An observation from (Andriushchenko & Flammarion, 2020) is that this PGD-2 AT baseline also suffers from catastrophic overfitting for large ℓ_∞ perturbations (*e.g.,* $\epsilon = 16/255$ on CIFAR10 and $\epsilon = 12/255$ on SVHN).

We confirm Andriushchenko & Flammarion (2020)'s observation by evaluating PGD-2-NRS AT on CIFAR10: As shown in Table 1, PGD-2-NRS AT losses robustness against PGD-50-10[†] at $\epsilon = 16/255$ on CIFAR10. However, we also observe that PGD-2-RS AT actually can prevent catastrophic overfitting for large ℓ_∞ perturbations. We compare PGD-2-RS AT with FGSM + GradAlign AT, and an unexpected finding is that PGD-2-RS AT with attack step size $\alpha = 1.25\epsilon/2$ achieves overall better performance than FGSM + GradAlign AT, with approximately 50% less computational cost. We note that on CIFAR10, Andriushchenko & Flammarion (2020) reported 58.5% standard accuracy and over 28% robust accuracy against PGD-50-10 with $\lambda = 0.1$ at $\epsilon = 16/255$ (λ: the regularization hyperparameter). However, as shown in Fig. 15 in (Andriushchenko & Flammarion, 2020), the setting of $\lambda = 0.1$ lies in/near the unstable zone. In our testbed, FGSM + GradAlign AT with $\lambda = 0.1$ suffers from catastrophic overfitting at $\epsilon = 16/255$[‡]. Thus, we set $\lambda = 0.2$ at $\epsilon = 8/255$ and $\lambda = 2.0$ at $\epsilon = 16/255$ for GradAlign on CIFAR10 (as in Table 2 in (Andriushchenko & Flammarion, 2020)). Besides, Andriushchenko & Flammarion (2020) evaluate robust accuracy on 1000 random points, while we evaluate robust accuracy on the whole test set.

[†]A commonly-used attack for evaluating AT methods: 50-step PGD with 10 random starts.

[‡]**Surprisingly, we also ran Andriushchenko & Flammarion (2020)'s Github code in our conda environment and found Andriushchenko & Flammarion (2020)'s implementation of FGSM + GradAlign AT also suffers from catastrophic overfitting with the setting of $\lambda = 0.1$ at $\epsilon = 16/255$ on CIFAR10.**

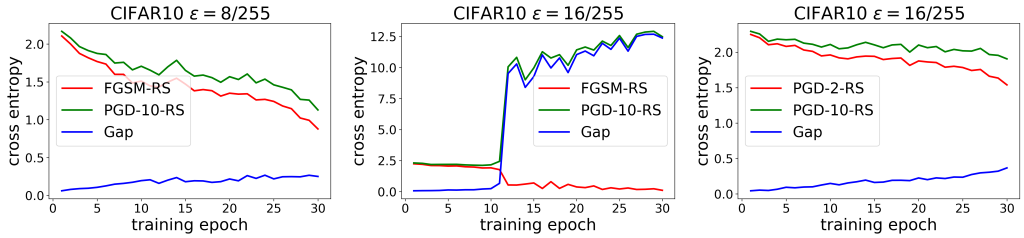

Figure 1: The gap between the averaged loss of FGSM-RS/PGD-2-RS generated training adversarial examples and the averaged loss of PGD-10-RS generated training adversarial examples in the 30-epoch training process with the cyclic learning schedule: The left two figures show the training process of FGSM-RS AT ($\alpha = 1.25\epsilon$); The right figure shows the training process of PGD-2-RS AT ($\alpha = 1.25\epsilon/2$). The loss is computed and averaged over the first training minibatch.

4.2 WHY PGD-2-RS WITH $\alpha = 1.25\epsilon/2$ WORKS?

Our observations in the previous subsection naturally raise a question:

Why PGD-2-RS AT with $\alpha = 1.25\epsilon/2$ can avoid catastrophic overfitting for large perturbations?

In the following, we propose and justify an empirical hypothesis to shed light on this question. In general, PGD-2 updates the adversarial examples for two steps: $\mathbf{x} \to \mathbf{x}_1 \to \mathbf{x}_2$. Apparently, we can treat $\mathbf{x}_1$ as the FGSM adversarial example of $\mathbf{x}$ and $\mathbf{x}_2$ as the FGSM adversarial example of $\mathbf{x}_1$. Based on the above fact, we propose the following **empirical** hypothesis:

Hypothesis *If FGSM-RS AT with attack step size $\alpha = 1.25\epsilon/2$ can avoid catastrophic overfitting for ℓ_∞ perturbation size $\epsilon/2$, then **under appropriate training settings**, PGD-2-RS AT with $\alpha = 1.25\epsilon/2$ **may be able to** avoid catastrophic overfitting for ℓ_∞ perturbation size ϵ.*

It is challenging to give a strict proof for the above empirical hypothesis, but we can provide some intuitions and extensive experimental results to justify this hypothesis. Specifically, if FGSM-RS AT with attack step size $\alpha = 1.25\epsilon/2$ can avoid catastrophic overfitting for ℓ_∞ perturbation size $\epsilon/2$, it implies that FGSM-RS (FGSM) can find qualified adversarial examples for adversarial training with ℓ_∞ perturbation size $\epsilon/2$. As shown in Fig. 1, the gap between the loss of FGSM-RS and PGD-10-RS generated training adversarial examples is small at $\epsilon = 8/255$ on CIFAR10, which indicates that FGSM-RS indeed can craft qualified adversarial examples for adversarial training at $\epsilon = 8/255$ on CIFAR10. PGD-2-RS with $\alpha = 1.25\epsilon/2$ can be **approximately** viewed as two step FGSM-RS with $\alpha = 1.25\epsilon/2$ ($\mathbf{x} \to \mathbf{x}_1 \to \mathbf{x}_2$). That is to say, $\mathbf{x}_1$ can be viewed as a qualified FGSM adversarial example of $\mathbf{x}$ for adversarial training with ℓ_∞ perturbation size $\epsilon/2$, and $\mathbf{x}_2$ can be viewed as a qualified FGSM adversarial example of $\mathbf{x}_1$ for adversarial training with an additional ℓ_∞ perturbation size $\epsilon/2$. Thus, we conjecture that $\mathbf{x}_2$ may be a qualified adversarial example of $\mathbf{x}$ for adversarial training with a total ℓ_∞ perturbation size ϵ. As shown in Fig. 1, the gap between the loss of PGD-2-RS and PGD-10-RS generated training adversarial examples is also small at $\epsilon = 16/255$ on CIFAR10, indicating that PGD-2-RS indeed can generate qualified adversarial examples for adversarial training at $\epsilon = 16/255$ on CIFAR10 (twice of $\epsilon = 8/255$).

According to the above intuitions, we note that a better way to generate qualified adversarial examples for large-perturbation adversarial training is executing FGSM-RS with random noise from $\mathcal{U}(-\epsilon/2, \epsilon/2)$ and step size $\alpha = 1.25\epsilon/2$ for *two steps* to craft the adversarial examples (namely Qusai-PGD-2-RS). *We refer to this attack method as Qusai-PGD-2-RS since it is similar to but not the same as conventional PGD-2-RS.* The main difference is that Qusai-PGD-2-RS adopts an unconventional random initialization scheme. Intuitively, Qusai-PGD-2-RS can be represented by

$$\text{Qusai-PGD-2-RS:}\quad \mathbf{x} \xrightarrow[FGSM\ \alpha=1.25\epsilon/2]{+\,\boldsymbol{\eta}\sim\mathcal{U}(-\epsilon/2,\epsilon/2)} \mathbf{x}_1 \xrightarrow[FGSM\ \alpha=1.25\epsilon/2]{+\,\boldsymbol{\eta}_1\sim\mathcal{U}(-\epsilon/2,\epsilon/2)} \mathbf{x}_2\ (\mathbf{x}'), \tag{3}$$

where $\boldsymbol{\eta}$ ($\boldsymbol{\eta}_1$) is the random noise with same dimension as $\mathbf{x}$, and each component of $\boldsymbol{\eta}$ ($\boldsymbol{\eta}_1$) is sampled from the uniform distribution $\mathcal{U}(-\epsilon/2, \epsilon/2)$. After adding $\boldsymbol{\eta}$ ($\boldsymbol{\eta}_1$) to $\mathbf{x}$ ($\mathbf{x}_1$), we execute FGSM (2) on $\mathbf{x}+\boldsymbol{\eta}$ ($\mathbf{x}_1+\boldsymbol{\eta}_1$) and clip the perturbations into the range of $[-\epsilon, \epsilon]$ and $[\mathbf{0} - \mathbf{x}, \mathbf{1} -$

Methods	CIFAR10 $\epsilon = 16/255$		SVHN $\epsilon = 12/255$	
	Standard	PGD-50-10	Standard	PGD-50-10
PGD-2-RS AT	$67.77\% \pm 0.52\%$	$26.22\% \pm 0.37\%$	$88.97\% \pm 0.25\%$	$32.68\% \pm 0.42\%$
Qusai-PGD-2-RS AT (3)	$65.49\% \pm 0.50\%$	$28.37\% \pm 0.26\%$	$88.11\% \pm 0.19\%$	$33.53\% \pm 0.52\%$

Table 2: PGD-2-RS AT and Qusai-PGD-2-RS AT (with $\alpha = 1.25\epsilon/2$) for large perturbations.

$\mathbf{x}$]. Finally, we use $\mathbf{x}_2$ ($\mathbf{x}'$) to train the model. We note that, if our intuitions are correct, then Qusai-PGD-2-RS AT may also be able to prevent catastrophic overfitting for large ℓ_∞ perturbations. Surprisingly, despite using an unconventional random initialization scheme, Qusai-PGD-2-RS AT achieves better robust accuracy than PGD-2-RS AT in most cases.

In addition to the above intuitions, we also verify our empirical hypothesis by extensive evaluations. We first verify our hypothesis by experiments on CIFAR10 and SVHN. According to (Wong et al., 2019; Andriushchenko & Flammarion, 2020), FGSM-RS AT with attack step size $\alpha = 1.25\epsilon$ is effective for medium perturbations, *e.g.,* $\epsilon = 8/255$ *on CIFAR10 and* $\epsilon = 6/255$ *on SVHN*. However, FGSM-RS AT may encounter (catastrophic) overfitting starting from $\epsilon = 9/255$ on CIFAR10 and $\epsilon = 7/255$ on SVHN with the training settings of (Andriushchenko & Flammarion, 2020). If our empirical hypothesis is correct, PGD-2-RS AT and Qusai-PGD-2-RS AT with $\alpha = 1.25\epsilon/2$ may be able to avoid catastrophic overfitting at $\epsilon = 16/255$ on CIFAR10 and $\epsilon = 12/255$ on SVHN. As shown in Table 3, both PGD-2-RS AT and Qusai-PGD-2-RS AT indeed can avoid catastrophic overfitting and train robust models at $\epsilon = 16/255$ on CIFAR10 and $\epsilon = 12/255$ on SVHN. An unexpected finding is that Qusai-PGD-2-RS AT (3), which uses an unconventional random initialization scheme, even achieves better robust accuracy against PGD-50-10 than PGD-2-RS AT on CIFAR10 and SVHN. The performance of those methods against AutoAttack (Croce & Hein, 2020) is shown in Section 5.2. Moreover, we also verify our hypothesis on ImageNet. According to (Wong et al., 2019; Andriushchenko & Flammarion, 2020), FGSM-RS AT is able to train robust models at $\epsilon = 4/255$ (Wong et al., 2019) but will encounter catastrophic overfitting at $\epsilon = 6/255$ (Andriushchenko & Flammarion, 2020). If our hypothesis is correct, then PGD-2-RS AT and Qusai-PGD-2-RS AT may be able to avoid catastrophic overfitting at $\epsilon = 8/255$ on ImageNet. We detail the experimental results on ImageNet in Section 5.3, which indicate that PGD-2-RS AT and Qusai-PGD-2-RS AT can avoid catastrophic overfitting at $\epsilon = 8/255$ on ImageNet. Note that (Shafahi et al., 2019; Andriushchenko & Flammarion, 2020; Kim et al., 2021) did not report any result at $\epsilon = 8/255$ on ImageNet. In our testbed, FGSM + GradAlign AT (with mixed precision) experiences severe catastrophic overfitting at $\epsilon = 8/255$ on ImageNet. AT for Free experiences relatively mild overfitting but can only achieves about 7% top-1 robust accuracy against PGD-50-10 at $\epsilon = 8/255$ on ImageNet. All in all, it is very likely that our hypothesis is correct in most cases.

5 Experiments

5.1 Experimental Setup

We mainly follow the settings of (Wong et al., 2019; Andriushchenko & Flammarion, 2020), and our implementation is based on Wong et al. (2019)'s code. **We include our code for the experiments in the supplementary material.** We compare FGSM-RS AT (Wong et al., 2019), FGSM + GradAlign AT (Andriushchenko & Flammarion, 2020), free adversarial training (AT for free) (Shafahi et al., 2019), stable single step AT (Kim et al., 2021), PGD-2-RS AT, and Qusai-PGD-2-RS AT on CIFAR10 and SVHN. We also evaluate those efficient AT methods on ImageNet at $\epsilon = 8/255$. Different from (Andriushchenko & Flammarion, 2020), which evaluates the robust accuracy on 1000 random points for each dataset, we evaluate the standard and robust accuracy on the whole test set for CIFAR10 and SVHN and the whole validation set for ImageNet. Following (Wong et al., 2019), we conduct the experiments on PreAct ResNet-18 (He et al., 2016b) for CIFAR10 and SVHN. For ImageNet, we conduct the experiments on ResNet-50 (He et al., 2016a).

We follow the initialization schemes in (Shafahi et al., 2019; Wong et al., 2019; Andriushchenko & Flammarion, 2020) to initialize the perturbations for the corresponding AT methods. For FGSM-RS AT, FGSM + GradAlign AT, and stable single-step AT, we set the attack step size as $\alpha = 1.25\epsilon$. For PGD-2-RS AT and Qusai-PGD-2-RS AT, we set the attack step size as $\alpha = 1.25\epsilon/2$. For AT for Free, we set the number of minibatch replays as $m = 8$ on CIFAR10 & SVHN and $m = 4$ on

Methods	CIFAR10		SVHN	
	Standard	PGD-50-10	Standard	PGD-50-10
	$\epsilon = 8/255$			
FGSM-RS AT	$83.47\% \pm 0.25\%$	$46.08\% \pm 0.45\%$	$95.73\% \pm 0.32\%$	$0.01\% \pm 0.01\%$
PGD-2-RS AT	$82.98\% \pm 0.32\%$	$47.79\% \pm 0.23\%$	$93.43\% \pm 0.19\%$	$50.14\% \pm 0.47\%$
Qusai-PGD-2-RS AT	$81.92\% \pm 0.41\%$	$\mathbf{48.77\% \pm 0.24\%}$	$92.90\% \pm 0.23\%$	$\mathbf{51.16\% \pm 0.45\%}$
FGSM + GradAlign AT	$81.21\% \pm 0.52\%$	$47.60\% \pm 0.34\%$	$92.93\% \pm 0.18\%$	$45.87\% \pm 0.45\%$
AT for Free	$82.21\% \pm 0.46\%$	$47.87\% \pm 0.57\%$	$\mathbf{95.97\% \pm 0.37\%}$	$0.02\% \pm 0.03\%$
Stable single-step AT	$\mathbf{87.66\% \pm 0.50\%}$	$38.06\% \pm 0.51\%$	$95.04\% \pm 0.60\%$	$24.60\% \pm 3.53\%$
	$\epsilon = 16/255$		$\epsilon = 12/255$	
FGSM-RS AT	$46.88\% \pm 24.86\%$	$0.00\% \pm 0.00\%$	$94.49\% \pm 0.24\%$	$0.00\% \pm 0.00\%$
PGD-2-RS AT	$\mathbf{67.77\% \pm 0.52\%}$	$26.22\% \pm 0.37\%$	$88.97\% \pm 0.25\%$	$32.68\% \pm 0.42\%$
Qusai-PGD-2-RS AT	$65.49\% \pm 0.50\%$	$\mathbf{28.37\% \pm 0.26\%}$	$88.11\% \pm 0.19\%$	$\mathbf{33.53\% \pm 0.52\%}$
FGSM + GradAlign AT	$60.08\% \pm 0.40\%$	$25.80\% \pm 0.74\%$	$88.19\% \pm 0.38\%$	$24.10\% \pm 0.53\%$
AT for Free	$36.80\% \pm 24.96\%$	$0.00\% \pm 0.00\%$	$\mathbf{95.10\% \pm 0.45\%}$	$0.00\% \pm 0.00\%$
Stable single-step AT	$41.76\% \pm 7.79\%$	$0.00\% \pm 0.00\%$	$75.99\% \pm 7.66\%$	$4.97\% \pm 2.35\%$

Table 3: Performance of different efficient AT methods on CIFAR10 and SVHN. All the results are reported with the average and the standard deviation averaged over 5 random seeds.

AutoAttack	CIFAR10		SVHN	
	$\epsilon = 8/255$	$\epsilon = 16/255$	$\epsilon = 8/255$	$\epsilon = 12/255$
FGSM-RS AT	$43.41\% \pm 0.40\%$	$0.00\% \pm 0.00\%$	$0.00\% \pm 0.00\%$	$0.00\% \pm 0.00\%$
PGD-2-RS AT	$44.89\% \pm 0.28\%$	$19.93\% \pm 0.40\%$	$46.14\% \pm 0.48\%$	$26.76\% \pm 0.31\%$
Qusai-PGD-2-RS AT	$\mathbf{45.66\% \pm 0.28\%}$	$\mathbf{20.34\% \pm 0.44\%}$	$\mathbf{46.91\% \pm 0.47\%}$	$\mathbf{27.33\% \pm 0.19\%}$
FGSM + GradAlign AT	$44.40\% \pm 0.35\%$	$18.17\% \pm 0.64\%$	$42.15\% \pm 0.40\%$	$19.67\% \pm 0.38\%$
AT for Free	$45.13\% \pm 0.65\%$	$0.00\% \pm 0.00\%$	$0.00\% \pm 0.00\%$	$0.00\% \pm 0.00\%$
Stable single-step AT	$35.59\% \pm 0.40\%$	$0.00\% \pm 0.00\%$	$4.42\% \pm 1.81\%$	$0.01\% \pm 0.03\%$

Table 5: Performance of different efficient AT methods against AutoAttack. All the results are reported with the average and the standard deviation averaged over 5 random seeds.

ImageNet. For stable single-step AT, we set the number of checkpoints as $c = 3$. For all the AT methods except AT for Free, we train the models with the cyclic learning schedule (Smith, 2017; Wong et al., 2019) on CIFAR10 for 30 epochs with the maximum learning rate 0.3 and on SVHN for 15 epochs with the maximum learning rate 0.05 (Andriushchenko & Flammarion, 2020). For AT for Free, we set the learning rate as 0.04 on CIFAR10 and 0.01 on SVHN following (Wong et al., 2019; Andriushchenko & Flammarion, 2020). The Apex amp package is incorporated for speedup with mixed precision following (Wong et al., 2019; Andriushchenko & Flammarion, 2020). For reliable evaluation, we evaluate the trained models against 50-step PGD with 10 random restarts (PGD-50-10) and AutoAttack (Croce & Hein, 2020) with single precision. The attack step size for PGD-50-10 is set as $\epsilon/4$ by default. PGD-50-10 is the attack baseline used by Wong et al. (2019); Andriushchenko & Flammarion (2020), and AutoAttack is a commonly-used strong attack baseline in recent works on adversarial training. Note that all the evaluation results (except the training time) are obtained on the whole test set for CIFAR10 and SVHN or the whole validation set for ImageNet. **The interested readers can refer to the appendix (Section C) and our code in the supplementary material for more details on the training or evaluation settings.**

5.2 CIFAR10 & SVHN

We report the standard and robust accuracy achieved by different efficient AT methods on CIFAR10 and SVHN in Table 3 & 5. On CIFAR10, PGD-2-RS AT achieves better standard accuracy and comparable or better robust accuracy against PGD-50-10 and AutoAttack, compared to FGSM + GradAlign AT. Qusai-PGD-2-RS AT achieves the best robust accuracy in all the cases. On SVHN, both PGD-2-RS AT and Qusai-PGD-2-RS AT outperforms FGSM + GradAlign AT by a non-negligible margin in robust accuracy against PGD-50-10 and AutoAttack. Our analysis on the drawbacks of FGSM + GradAlign AT in Section B in the appendix sheds light on why PGD-2-RS AT and Qusai-PGD-2-RS AT can outperform FGSM + GradAlign AT. To investigate the gap between the results of PGD-50-10 and AutoAttack, we train a model using PGD-10-RS at $\epsilon = 16/255$ on CIFAR10. **The gap between the attack results by PGD-50-10 and AutoAttack is about 6% for**

the PGD-10-RS AT model. Thus, a large gap is a common phenomenon, especially for large ℓ_∞ perturbations. We also conduct an ablation study on the training settings on CIFAR10 in Section D in the appendix, which shows that PGD-2-RS AT and Qusai-PGD-2-RS AT have consistently better performance than FGSM + GradAlign AT under all our evaluated training settings on CIFAR10.

Besides, in our testbed, AT for Free fails to train robust models at $\epsilon = 16/255$ on CIFAR10 and at $\epsilon = 8/255$ and $\epsilon = 12/255$ on SVHN. Only for AT for Free, Andriushchenko & Flammarion (2020) train the models for 96 epochs on CIFAR10 and for 45 epochs on SVHN. In their testbed, AT for Free suffers from high variance at $\epsilon = 8/255$ on SVHN and catastrophic overfitting at $\epsilon = 12/255$ on SVHN and $\epsilon = 16/255$ on CIFAR10. For stable single-step AT, (Kim et al., 2021) reported 33.9% robust accuracy against PGD-50 at $\epsilon = 8/255$ with a piecewise learning schedule on CIFAR10. Our implementation of stable single-step AT achieves averagely 38.06% robust accuracy against PGD-50-10 with the cyclic learning schedule at $\epsilon = 8/255$ on CIFAR10. (Kim et al., 2021) did not report any result on SVHN and at $\epsilon = 16/255$ on CIFAR10. In our testbed, stable single-step AT suffers from catastrophic overfitting at $\epsilon = 16/255$ on CIFAR10 and $\epsilon = 8/255$ on SVHN, as shown in Table 3. **Note that we also plot the standard and robust accuracy of these efficient AT methods for different perturbation sizes in Fig. 2 in the appendix.**

Methods	CIFAR10	SVHN
FGSM-RS AT	9.58 min	6.99 min
FGSM + GradAlign AT	29.33 min	21.60 min
PGD-2-RS AT	13.65 min	10.21 min
Adaptive PGD-2-RS AT	13.74 min	10.27 min
Stable Single-Step AT	12.24 min	8.97 min
AT for Free	28.32 min	20.83 min

Table 4: The training time measured on a Tesla V100 GPU with seed 0 (30 epochs on CIFAR10 and 15 epochs on SVHN).

As shown in Table 4, the computational time of PGD-2-RS AT or Qusai-PGD-2-RS AT is slightly less than $1.5\times$ FGSM-RS AT's computational time in practice. This is because the computational time for the other operations (except forward and backward propagations) is similar for FGSM-RS AT and PGD-2-RS AT. We denote the additional cost of the other operations as $c_3 N$. Then the cost of FGSM-RS AT is approximately $2(c_1 + c_2)N + c_3 N$, and the cost of PGD-2-RS AT is approximately $3(c_1 + c_2)N + c_3 N$. The ratio $\frac{3(c_1+c_2)N+c_3N}{2(c_1+c_2)N+c_3N}$ is slightly less than 1.5. Compared to FGSM-RS AT, FGSM + GradAlign AT leads to $3\times$ slowdown in our testbed, which is consistent with the result of (Andriushchenko & Flammarion, 2020). As shown in Table 4, PGD-2-RS AT or Qusai-PGD-2-RS AT only takes approximately a half computational time of FGSM + GradAlign AT.

5.3 IMAGENET

We follow the training scheme of (Wong et al., 2019) to train the ImageNet models on ResNet-50, and we evaluate the trained models on the whole validation set of ImageNet. The validation set is not used in the training stage. Notably, Qusai-PGD-2-RS AT can achieve approximately 18% top-1 and 38% top-5 robust accuracy against PGD-50-10, as shown in Table 6. In contrast, both FGSM-RS AT and FGSM + GradAlign AT suffer from catastrophic overfitting at $\epsilon = 8/255$ on ImageNet in our testbed, *i.e.,* their trained models' robust accuracy against PGD-50-10 drops to nearly 0 at a certain point in the training process. Compared to FGSM-RS AT and FGSM + GradAlign AT, the overfitting issue of AT for Free ($m = 4$) is relatively mild at $\epsilon = 8/255$, but its final top-1 robust accuracy against PGD-50-10 on the validation set is still less than $1/2$ of the top-1 robust accuracy achieved by PGD-2-RS AT and Qusai-PGD-2-RS AT.

ImageNet $\epsilon = 8/255$	Top 1 accuracy		Top 5 accuracy	
	Standard	PGD-50-10	Standard	PGD-50-10
AT for Free	54.728%	7.304%	77.422%	30.450%
PGD-2-RS AT	46.192%	17.586%	69.728%	38.638%
Qusai-PGD-2-RS AT (3)	44.492%	17.750%	66.988%	38.184%

Table 6: Performance of PGD-2-RS AT and Qusai-PGD-2-RS AT on ImageNet. The final-epoch **training** top-1 accuracy against PGD-2-RS or Qusai-PGD-2-RS is around 23%, and the gap between training accuracy and the testing robust accuracy is only around 5%. Thus, we can say PGD-2-RS AT and Qusai-PGD-2-RS AT do not suffer from catastrophic overfitting at $\epsilon = 8/255$ on ImageNet.

Methods	CIFAR10		SVHN	
	Standard	PGD-50-10	Standard	PGD-50-10
	$\epsilon = 8/255$			
PGD-3-RS AT	$82.92\% \pm 0.42\%$	$48.05\% \pm 0.37\%$	$93.33\% \pm 0.22\%$	$50.88\% \pm 0.59\%$
PGD-4-RS AT	$82.62\% \pm 0.41\%$	$48.20\% \pm 0.34\%$	$93.29\% \pm 0.25\%$	$51.19\% \pm 0.68\%$
PGD-5-RS AT	$82.68\% \pm 0.36\%$	$48.51\% \pm 0.48\%$	$93.23\% \pm 0.25\%$	$51.36\% \pm 0.81\%$
	$\epsilon = 16/255$		$\epsilon = 12/255$	
PGD-3-RS AT	$67.11\% \pm 0.50\%$	$28.25\% \pm 0.46\%$	$88.57\% \pm 0.30\%$	$33.80\% \pm 0.49\%$
PGD-4-RS AT	$66.97\% \pm 0.56\%$	$28.70\% \pm 0.41\%$	$88.44\% \pm 0.36\%$	$34.20\% \pm 0.62\%$
PGD-5-RS AT	$66.78\% \pm 0.61\%$	$28.97\% \pm 0.29\%$	$88.29\% \pm 0.32\%$	$34.41\% \pm 0.63\%$

Table 7: Performance of PGD-k-RS AT on CIFAR10 and SVHN. All the results are reported with the average and the standard deviation averaged over 5 random seeds.

5.4 PGD-K-RS AT BASELINES

We also study the performance of PGD adversarial training with differing numbers of attack steps k. As analyzed in Section 3, the computational cost of FGSM + GradAlign AT is similar to the cost of PGD-5-RS AT (5-step PGD adversarial training with random initialization). Thus, as long as $k < 5$, PGD-k-RS AT has a lower cost than FGSM + GradAlign AT. For the PGD-k-RS AT here, we set α as $1.25\epsilon/k$ and initialize the perturbations with the random noise from $\mathcal{U}(-\epsilon, \epsilon)$ (the conventional random initialization scheme). We report the standard accuracy and robust accuracy against PGD-50-10 of PGD-k-RS AT in Table 7. Increasing k leads to additional computational cost on forward and backward propagations. However, it seems that the benefits coming along with the increasing k do not merit the additional cost: The increasing trend of robust accuracy due to the increase of k after $k \geq 2$ becomes slow, and Qusai-PGD-2-RS AT achieves comparable robust accuracy as PGD-3-RS AT. Thus, we hold the opinion that for most cases, Qusai-PGD-2-RS (or PGD-2-RS) can provide an efficient and acceptable solution to the inner problem for ℓ_∞-norm adversarial training with large ℓ_∞ perturbations under proper training settings.

6 ADDITIONAL DISCUSSIONS

The previous works observed that the training settings, such as the learning schedule, can have significant impacts (Kim et al., 2021). For instance, FGSM-RS AT with a long-time (*e.g,* 100 or 200 epochs) piecewise learning schedule and learning rate decay suffers from catastrophic overfitting at $\epsilon = 8/255$ on CIFAR10. In our testbed, we observe that catastrophic overfitting due to long-time training with the piecewise learning schedule can be mitigated by simply increasing the weight decay and stopping model training after the first epoch when the learning rate decays to $O(10^{-3})$ (0.001 or 0.002). This observation is detailed in the appendix. We also identify two potential drawbacks of gradient-based regularization in adversarial training, inducing an open question: *Is gradient-based regularization worthy of its additional cost in adversarial training?* We discuss the **potential** drawbacks of gradient-based regularization in Section B in the appendix. Also, due to the space limit, we detail the related work in the appendix.

7 CONCLUSION

In this paper, we show that PGD-2-RS AT with attack step size $\alpha = 1.25\epsilon/2$ only has approximately a half computational cost of FGSM + GradAlign AT but actually can avoid catastrophic overfitting for large ℓ_∞ perturbations. We hypothesize that, if FGSM-RS AT with attack step size $\alpha = 1.25\epsilon/2$ can avoid catastrophic overfitting for ℓ_∞ perturbation size $\epsilon/2$, then PGD-2-RS AT with $\alpha = 1.25\epsilon/2$ may be able to avoid catastrophic overfitting for ℓ_∞ perturbation size ϵ under proper settings. Inspired by this hypothesis, we propose to execute PGD-2-RS with an unconventional two-step initialization scheme and refer to the corresponding AT method as Qusai-PGD-2-RS AT. Through extensive evaluations, we verify the empirical hypothesis and demonstrate that PGD-2-RS AT and Qusai-PGD-2-RS AT with $\alpha = 1.25\epsilon/2$ can achieve overall better performance and efficiency than FGSM + GradAlign AT. Notably, Qusai-PGD-2 AT achieves comparable robust accuracy as PGD-3-RS AT on CIFAR10 and SVHN, and it also achieves approximately 18% top-1 robust accuracy against PGD-50-10 at $\epsilon = 8/255$ on ImageNet.

REFERENCES

Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33, 2020.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. *arXiv preprint arXiv:1912.00049*, 2019.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pp. 11192–11203, 2019.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. *arXiv preprint arXiv:1907.02044*, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, 2020.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2019.

Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, and Hang Su. Adversarial distributional training for robust deep learning. In *Advances in Neural Information Processing Systems*, 2020.

Christian Etmann. A closer look at double backpropagation. *arXiv preprint arXiv:1906.06637*, 2019.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b.

Hoki Kim, Woojin Lee, and Jaewook Lee. Understanding catastrophic overfitting in single-step adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8119–8127, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

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

Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pp. 3575–3583, 2018.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Hang Su, and Jun Zhu. Boosting adversarial training with hypersphere embedding. In *Advances in Neural Information Processing Systems*, 2020.

Leslie Rice, Eric Wong, and J Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, 2020.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pp. 3358–3369, 2019.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472. IEEE, 2017.

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

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

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

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pp. 5283–5292, 2018.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2019.

Dongxian Wu, Shutao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, 2020.

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems*, pp. 227–238, 2019a.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482, 2019b.

Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International Conference on Machine Learning*, 2020.

A Additional Experimental Results

We plot the standard accuracy and the robust accuracy against PGD-50-10 for different perturbation sizes in Fig. 2. The fluctuation for FGSM-RS AT at $\epsilon = 14/255$ (robust accuracy) is because FGSM-RS AT happens to train a robust model at $\epsilon = 14/255$ with seed 0 on CIFAR10. FGSM-RS AT and AT for Free suffer from catastrophic overfitting for large ℓ_∞ perturbations, *e.g.,* at $\epsilon = 12/255$ on CIFAR10 and $\epsilon = 8/255$ on SVHN. On CIFAR10, stable single-step AT also encounters catastrophic overfitting at $\epsilon = 12/255$. On SVHN, stable single-step AT encounters relatively mild overfitting for large ℓ_∞ perturbations, *i.e.,* the robust accuracy against PGD-50-10 drops quickly but does not drop to 0 when the perturbation size ϵ is larger than $6/255$. FGSM + GradAlign AT does not have the issue of catastrophic overfitting for large ℓ_∞ perturbations on CIFAR10 and SVHN, but Qusai-PGD-2-RS AT can outperform FGSM + GradAlign AT by a non-negligible margin in robust accuracy for large ℓ_∞ perturbations with approximately 50% less computational cost.

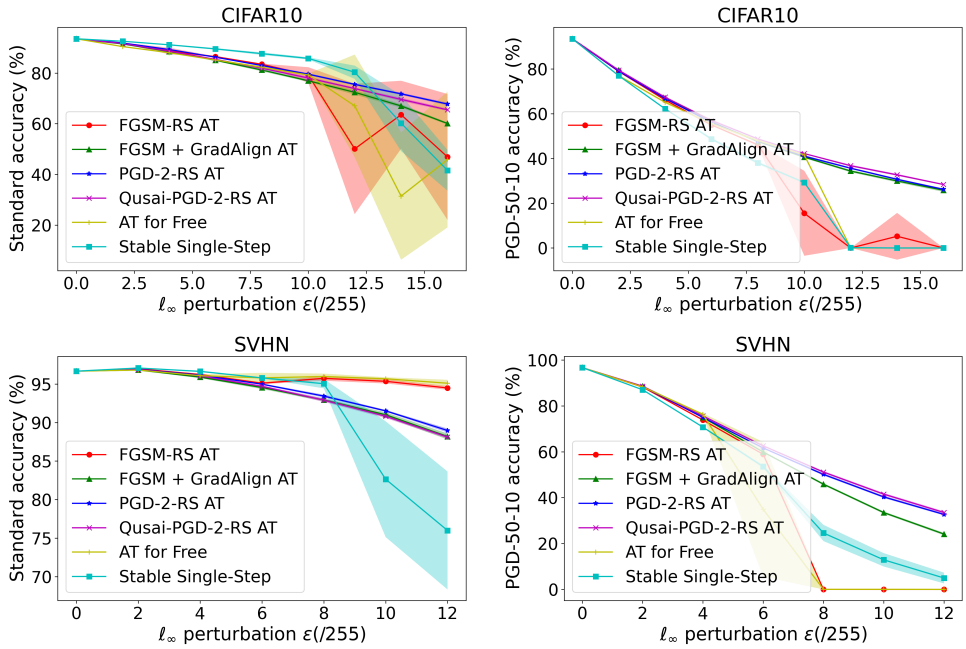

Figure 2: Standard and robust accuracy of different efficient AT methods on CIFAR-10 and SVHN with PreAct ResNet-18 trained and evaluated with different perturbation sizes. The results are obtained by averaging over 5 random seeds used for training and reported with the standard deviation.

B Rethinking Gradient-based Regularization

Our observations in this paper naturally raise an open question:

Is gradient-based regularization worthy of its additional cost in adversarial training?

We identify two potential drawbacks of gradient-based regularization, which makes us prefer to answer "No" to this open question. The two potential drawbacks can be summarized as

1. Gradient-based regularization may promote an artificial/odd change to the loss topology, which may lead to a sub-optimal loss topology shape in some cases.
2. Fine-tuning the hyperparameter before the gradient-based regularizer may need a fair amount of additional cost on a new dataset.

Gradient-based regularization may promote an artificial change to the model gradients and loss topology. For instance, a gradient-based linearity regularizer will render the loss topology to be

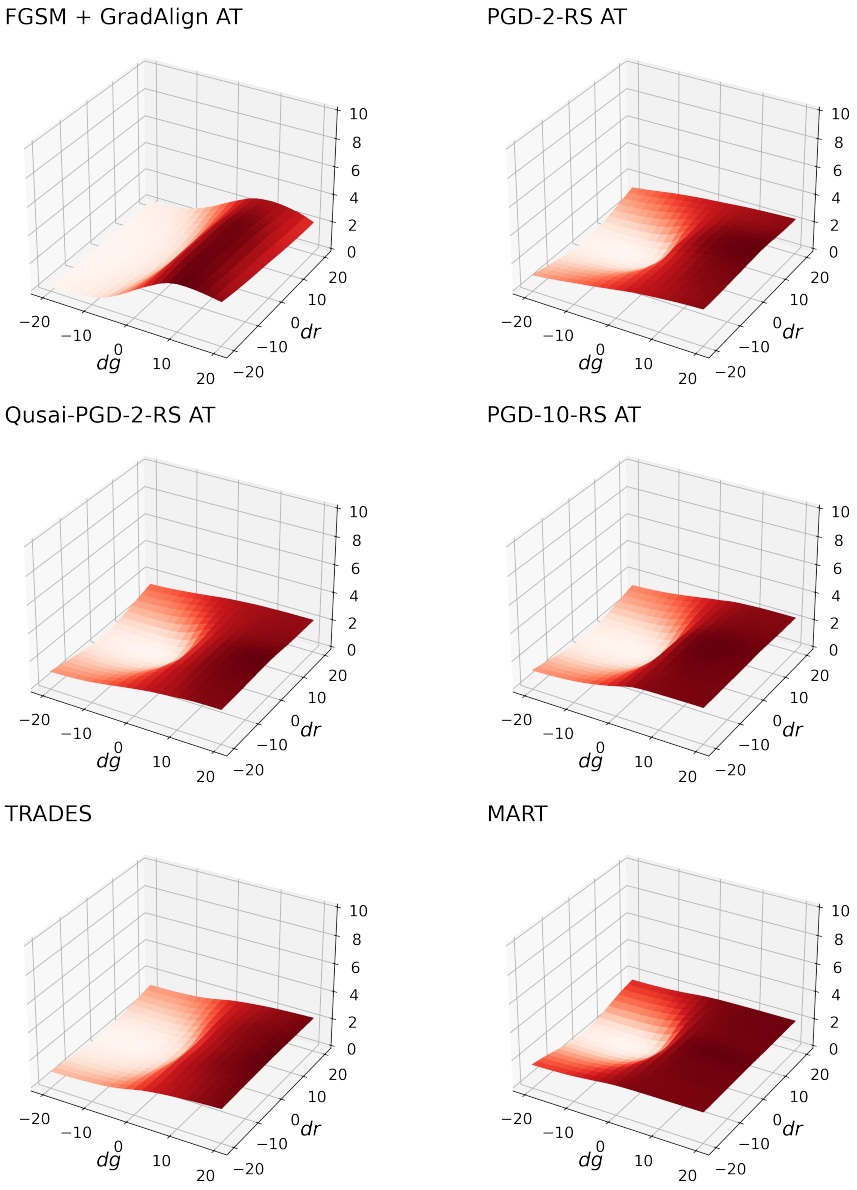

Figure 3: CIFAR10: cross-entropy loss topology (in the vicinity of a testing input) of the models trained by different AT methods on CIFAR10 ($\epsilon = 16/255$). dg denotes the unit direction vector of the gradient, *i.e.,* $\|dg\|_2 = 1$, and dr denotes a random (orthogonal) unit direction vector. Deeper red refers to higher loss. Note that we also train robust models by TRADES (Zhang et al., 2019b) and MART (Wang et al., 2020) and plot the model loss topology. Apparently, the loss topology of the model trained by FGSM + GradAlign is the odd one, and the loss topologies of the models trained the other AT methods are similar.

more linear. Our intuition is that such changes to the loss topology may be artificial since they are promoted by manually-designed regularizers not really learned by the neural network itself. Those changes may force the loss topology to shape in a sub-optimal pattern. We use GradAlign as an example to illustrate the above intuition. We randomly select a sample from the test set and plot the loss topology in the vicinity of the testing sample for FGSM + GradAlign AT, PGD-2-RS AT, Qusai-PGD-2-RS AT, PGD-10-RS AT, TRADES, and MART. All the models are trained with random seed 0. We show the loss topology of those models in Fig. 3. Apparently, the loss topology of FGSM + GradAlign AT trained model is somehow artificial and different from the loss topology of the models trained by the other AT methods, including TRADES and MART. Since PGD-10-RS (with

cross-entropy loss), TRADES, and MART all achieve over 30% robust accuracy against PGD-50-10 at $\epsilon = 16/255$ on CIFAR10, we can say the pattern of the loss topology of FGSM + GradAlign AT trained model is very likely to be a sub-optimal one.

In addition, a fair amount of additional cost is needed to fine-tune the hyperparameter before the regularizer when we use gradient-based regularization on a new dataset. Given GradAlign as an example, the optimal regularization hyperparameters λ for CIFAR10 and SVHN are quite different: On CIFAR10, Andriushchenko & Flammarion (2020) set $\lambda = 0.2$ at $\epsilon = 8/255$ and $\lambda = 2.0$ at $\epsilon = 16/255$, while on SVHN, Andriushchenko & Flammarion (2020) set $\lambda = 1.0$ at $\epsilon = 8/255$ and $\lambda = 2.0$ at $\epsilon = 12/255$ on SVHN. Although Andriushchenko & Flammarion (2020) claim that GradAlign is not sensitive to the change of λ, we can still observe slight decrease in the robust accuracy as λ increases after $\lambda = 2.0$ at $\epsilon = 16/255$ on CIFAR10, as shown in (Andriushchenko & Flammarion, 2020). Also, we note that in our testbed, $\lambda = 2.0$ is not a good setting at $\epsilon = 12/255$ on SVHN, which trains the model to be a majority classifier, with the random seed being set as 4. In practice, we have tried five hyperparameter settings at $\epsilon = 12/255$ on SVHN (*i.e.,* $\lambda = 1.2, 1.5, 1.6, 1.8, 2.0$) and finally set $\lambda = 1.6$ at $\epsilon = 12/255$ on SVHN. Apparently, this hyperparameter fine-tuning process requires considerable additional cost on a new dataset. In contrast, PGD-2-RS AT and Qusai-PGD-2-RS AT are free of such a costly fine-tuning process on a regularization hyperparameter like λ in FGSM + GradAlign AT.

C Detailed Training Settings

On CIFAR10 and SVHN, we follow (Wong et al., 2019; Andriushchenko & Flammarion, 2020) to use a batch size of 128, an SGD optimizer with momentum 0.9, cyclic learning schedules (introduced in Section 5.1), and weight decay 5e-4 by default. For the experiments with piecewise learning schedules, we increase the weight decay to 2e-3 to avoid catastrophic overfitting, as detailed in Section D. On ImageNet, we follow (Wong et al., 2019) to divide model training into three phases: 6 epochs, 6 epochs, and 3 epochs. The batch size for the three phases is set as 512, 224, 128, and the crop size is set as 128, 224, 288. The weight decay is set as 0.0001. Those settings are same as the settings of (Wong et al., 2019)'s code[§]. We set the minibatch replay as $m = 4$ for AT for Free. For more details about the training settings on ImageNet, the interested readers can refer to the files in ImageNet/configs in our code repository in the supplementary material.

D Ablation Study on Training Settings

In this section, we study the impacts of the training settings, such as the learning schedule, weight decay, and learning rate, on the performance of those efficient AT methods. We mainly conduct this ablation study on CIFAR10, due to the limited computational resource. We first consider the following two piecewise learning schedules: (1) We use SGD with the initial learning rate as 0.1 and momentum 0.9; The learning rate decays with a factor of 0.1 at 50 and 75 epochs. (2) We use SGD with the initial learning rate as 0.01 and momentum 0.9; The learning rate decays with a factor of 0.2 at 60, 120, 160 epochs (Kim et al., 2021). We confirm that FGSM-RS AT suffers from catastrophic overfitting at $\epsilon = 8/255$ on CIFAR10 under the above two long-time learning schedules with weight decay 5e-4 in our testbed. To avoid catastrophic overfitting under these two long-time learning schedules, we propose two remedies: (1) increasing the weight decay and (2) stopping model training after the first epoch when the learning rate decays to $O(10^{-3})$. Note that increasing weight decay is a commonly-used remedy to mitigate overfitting. Also, training the model with a small learning rate (*e.g.,* 0.001 *or* 0.002) for a long time is a common cause for overfitting in practice in both standard and adversarial training. Thus, we stop after the first epoch with the learning rate 0.001 or 0.002. We show the experimental results after applying the proposed two remedies in Table 8. As we can see, Qusai-PGD-2-RS AT achieves the best robust accuracy against PGD-50-10 under both learning schedules. Also, we note that AT for Free with a large learning rate may not be able to converge in some cases, *e.g.,* AT for Free can not converge under the second learning schedule. Thus, the previous works (Wong et al., 2019; Andriushchenko & Flammarion, 2020) use a relatively small learning rate in AT for Free. Also, we note that either of the above two remedies is essential to avoiding catastrophic overfitting. If we only stop after the first epoch with a small

[§]https://github.com/locuslab/fast_adversarial/tree/master/ImageNet

CIFAR10 $\epsilon = 8/255$	Schedule 1		Schedule 2	
	Standard	PGD-50-10	Standard	PGD-50-10
FGSM-RS AT	$83.11\% \pm 0.23\%$	$44.51\% \pm 0.37\%$	$84.43\% \pm 0.25\%$	$44.24\% \pm 0.48\%$
PGD-2-RS AT	$82.79\% \pm 0.23\%$	$46.34\% \pm 0.37\%$	$84.14\% \pm 0.29\%$	$46.42\% \pm 0.46\%$
Qusai-PGD-2-RS AT	$81.61\% \pm 0.26\%$	$47.26\% \pm 0.34\%$	$83.06\% \pm 0.30\%$	$47.45\% \pm 0.32\%$
FGSM + GradAlign AT	$80.93\% \pm 0.24\%$	$46.26\% \pm 0.14\%$	$82.59\% \pm 0.41\%$	$46.11\% \pm 0.44\%$
AT for Free	$75.60\% \pm 0.60\%$	$42.65\% \pm 0.73\%$	$25.10\% \pm 7.56\%$	$17.86\% \pm 8.93\%$
Stable single-step AT	$87.03\% \pm 0.21\%$	$36.16\% \pm 0.31\%$	$87.80\% \pm 0.19\%$	$36.77\% \pm 0.41\%$

Table 8: Performance of different adversarial training methods on CIFAR10 under piecewise learning schedules. All the results are reported with the average and the standard deviation averaged over 5 random seeds. The best results are marked in bold. **We set the weight decay as 2e-3 instead of 5e-4 and stop training after the first epoch with the learning rate** 0.001 **or** 0.002**.**

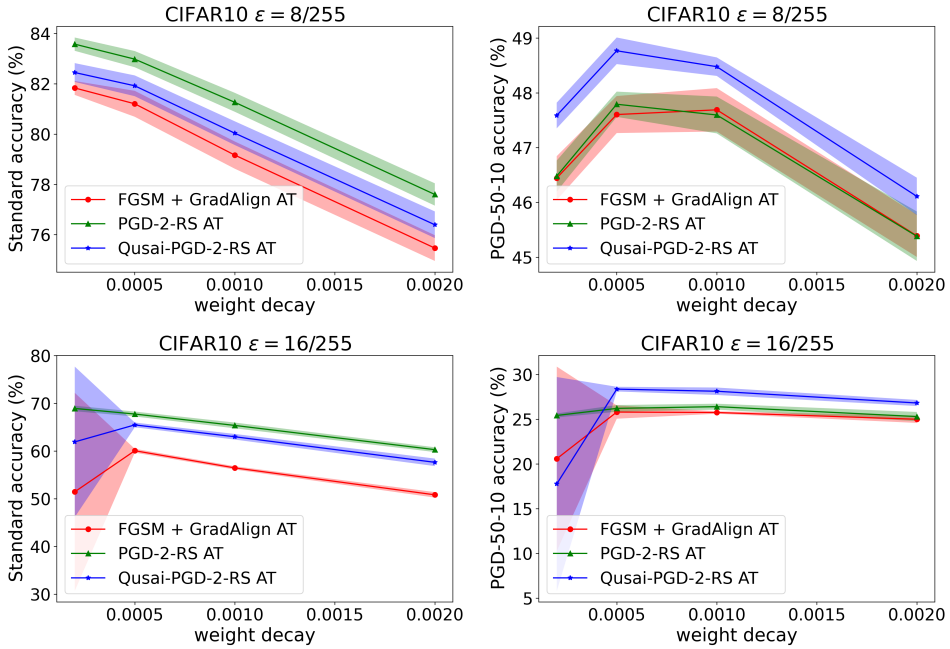

Figure 4: The impacts of weight decay on FGSM + GradAlign AT, PGD-2-RS AT, and Qusai-PGD-2-RS AT. The results are obtained by averaging over 5 random seeds used for training and reported with the standard deviation. The other hyperparameter settings are same as the settings described in Section 5.1 and Section C.

learning rate, without increasing the weight decay, FGSM-RS AT still suffers from catastrophic overfitting with certain random seeds. Even if we increase weight decay to 1.5e-3, FGSM-RS AT still has the issue of catastrophic overfitting, with the random seed being set as 2. On the other hand, if we only increase the weight decay, then FGSM-RS AT either suffers from catastrophic overfitting with a medium weight decay or exhibits a very poor performance with a large weight decay.

We also study the impacts of the weight decay and (maximum) learning rate on the performance of FGSM + GradAlign AT, PGD-2-RS AT, and Qusai-PGD-2-RS AT under the cyclic learning schedule. We set the weight decay as 2e-4, 5e-4, 1e-3, 2e-3 (the maximum learning rate is fixed as 0.3), and plot the corresponding standard and robust accuracy in Fig 4. As shown in Fig 4, small weight decay may lead to overfitting (or even catastrophic overfitting), while large weight decay substantially reduces the model compacity and thus may lead to limited model performance. 5e-4 is a proper setting for the weight decay. Also, we set the maximum learning rate as 0.1, 0.2, 0.3, 0.4, 0.5 (the weight decay is fixed as 5e-4), and plot the corresponding standard and robust accuracy of FGSM + GradAlign AT, PGD-2-RS AT, and Qusai-PGD-2-RS AT in Fig. 5. Though with different settings of

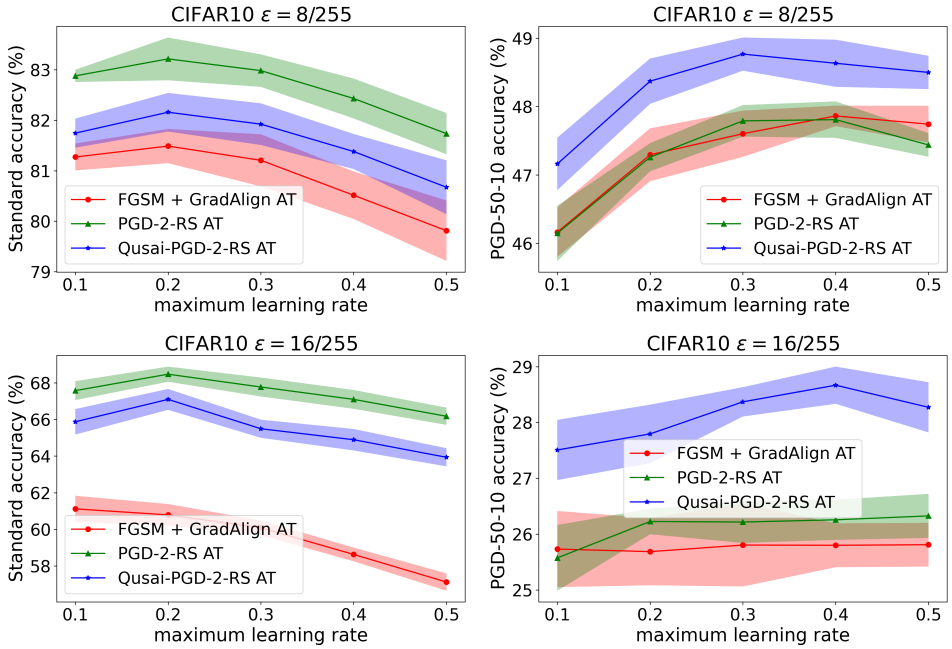

Figure 5: The impacts of maximum learning rate on FGSM + GradAlign AT, PGD-2-RS AT, and Qusai-PGD-2-RS AT under the cyclic learning schedule. The results are obtained by averaging over 5 random seeds used for training and reported with the standard deviation. The other hyperparameter settings are same as the settings described in Section 5.1 and Section C.

the maximum learning rate, PGD-2-RS AT always has better standard accuracy and comparable or better robust accuracy compared to FGSM + GradAlign AT, and Qusai-PGD-2-RS AT outperforms FGSM + GradAlign AT in robust accuracy in all the cases. All in all, under all our evaluated training schemes, both PGD-2-RS AT and Qusai-PGD-2-RS AT have consistently better performance than FGSM + GradAlign AT.

E Related Work

The vulnerability of deep learning models to adversarial examples was first reported in (Szegedy et al., 2013). Since then, the community has proposed many approaches to generate or defend against adversarial examples. Goodfellow et al. (2014) first proposed FGSM to generate adversarial examples and adversarially train robust models against the adversarial examples. However, the FGSM adversarial training method proposed by Goodfellow et al. (2014) trains a model to fool FGSM by gradient masking (Tramèr et al., 2018), and thus cannot defend against stronger multi-step attacks (Kurakin et al., 2016). Madry et al. (2017) then proposed to adversarially train robust models by PGD adversarial examples, namely PGD adversarial training. In the past few years, PGD adversarial training has survived the battles against many strong attacks Athalye et al. (2018); Andriushchenko et al. (2019); Croce & Hein (2020); thus, it is widely known as the most effective empirical defense. Following Madry et al. (2017), the community developed many variants of PGD adversarial training and plug-in methods to further improve the defensive performance (Zhang et al., 2019b; Wang et al., 2020; Dong et al., 2020; Pang et al., 2020; Wu et al., 2020). Despite its effectiveness, PGD adversarial training suffers from high computational overhead due to the multiple adversary updates for one model update in each training step. Therefore, a great deal of research has been devoted to accelerating PGD adversarial training. Shafahi et al. (2019) proposed to simultaneously update adversarial examples and model parameters in each training step (AT for Free). Since AT for Free makes full use of every adversary update, it can achieve comparable performance with PGD adversarial training with fewer adversary updates. Zhang et al. (2019a) proposed to reduce the costs of the adversary updates by restricting most of the forward and backward propagations within the first layer of the network during the adversary updates.

Contrary to the previous popular belief, Wong et al. (2019) found that using FGSM with appropriate random initialization and attack step size as the adversary can also train a robust model against multi-step attacks, referred to as FGSM-RS AT. Wong et al. (2019) also identified failure modes called "catastrophic overfitting" in the previous FGSM adversarial training methods, meaning that the trained model's robust accuracy against PGD suddenly drops to 0% over a single epoch. Wong et al. (2019) owed "catastrophic overfitting" to the inappropriate settings on initialization, step size, etc. FGSM-RS AT only executes a one-step update on the adversarial examples in each training step, so it is more efficient than the previous accelerated adversarial training methods (Shafahi et al., 2019; Zhang et al., 2019a), opening up a promising direction for accelerating adversarial training.

In this line of research, Andriushchenko & Flammarion (2020) observed that even with proper settings, FGSM-RS AT is still prone to "catastrophic overfitting" for larger ℓ_∞ perturbations. For instance, on CIFAR10, when the ℓ_∞ perturbation size increases to $10/255$, FGSM-RS AT will encounter catastrophic overfitting in the training process with robust accuracy against PGD-50-10 dropping to 0%. Andriushchenko & Flammarion (2020) conjectured that the model's local nonlinearity is the cause of catastrophic overfitting. Thus, to prevent catastrophic overfitting, Andriushchenko & Flammarion (2020) proposed to enhance the trained model's local linearity by a regularization method, which maximizes the gradient alignment inside the perturbation sets, namely gradient alignment regularization (GradAlign). Kim et al. (2021) proposed to search for the appropriate step size for each data sample to improve FGSM-RS AT and avoid catastrophic overfitting under the long-time piecewise learning schedules.