Soften to Defend: Towards Adversarial Robustness via Self-Guided Label Refinement

Supplementary Material

A. Proofs

A.1. Proof of Theorem 1

Theorem 1 (Soft label could reduce the IIW) *Let u be the uniform random variable with p.d.f $p(u)$. By using the composition in Eq. (2), there exists an interpolation ration λ between the clean label distribution and uniform distribution, such that*

$$I(y^*; w|x') \lesssim I(y; w|x') \tag{A.1}$$

where $p(y^*|x', w) = \lambda \cdot p(y|x', w) + (1-\lambda) \cdot p(u)$ and the symbol $\lesssim$ means that the corresponding inequality up to an c-independent constant.

Proof. For this proof, we will use an inequality called the *log-sum inequality*.

Lemma. 1 (Log-sum inequality) *Let $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$ be nonnegative numbers. Denote the sum of all a_is by a and the sum of all b_is by b. The log sum inequality states that*

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b}.$$

with equality if and only if $\frac{a_i}{b_i}$ are equal for all i.

We rewrite the interpolation for simplicity

$$\begin{aligned} p(y^*|x', w) &= \lambda \cdot p(y|x', w) + (1-\lambda) \cdot p(u) \\ p^* &= \lambda \cdot p_1 + (1-\lambda) \cdot p_2 \\ p &= \lambda \cdot p_1 + (1-\lambda) \cdot p_1 \end{aligned} \tag{A.2}$$

Then we could derive the decomposition of cross entropy on different label distribution, i.e., p^* and p.

$$\begin{aligned} \mathcal{H}(p^*, f) &= \mathcal{H}(p^*) - I(w; p^*) + \mathbb{E}_{w\sim Q(w|\mathcal{S})}\mathrm{KL}[p^*\|f] \\ &= \lambda \cdot \mathcal{H}(p_1) + (1-\lambda) \cdot \mathcal{H}(p_2) - I(w; p^*) + \mathbb{E}^* \\ &= \lambda \cdot \mathcal{H}(p) + (1-\lambda) \cdot \mathcal{H}(p_2) - I(w; p^*) + \mathbb{E}^* \end{aligned} \tag{A.3}$$

$$\begin{aligned} \mathcal{H}(p, f) &= \mathcal{H}(p) - I(w; p) + \mathbb{E}_{w\sim Q(w|\mathcal{S})}\mathrm{KL}[p\|f] \\ &= \lambda \cdot \mathcal{H}(p) + (1-\lambda) \cdot \mathcal{H}(p) - I(w; p) + \mathbb{E} \end{aligned} \tag{A.4}$$

We would like to simplify the term $\mathbb{E}^* - \mathbb{E}$. Note that we utilize the important property of KL divergence via log-sum inequality.

$$\begin{aligned} \mathrm{KL}(p^*\|f) &= \sum(\lambda p_1 + (1-\lambda)p_2) \log \frac{\lambda p_1 + (1-\lambda)p_2}{\lambda f + (1-\lambda) f} \\ &\leq \lambda \cdot \mathrm{KL}(p\|f) + (1-\lambda) \cdot \mathrm{KL}(u\|f) \end{aligned}$$

Use the same trick on $\mathrm{KL}(p\|f)$, and we could get

$$\mathbb{E}^* - \mathbb{E} \leq (1-\lambda) \cdot [\mathrm{KL}(u\|f) - \mathrm{KL}(p\|f)]$$

Therefore

$$\begin{aligned} \mathcal{H}(p^*, f) - \mathcal{H}(p, f) &= (1-\lambda)\big(\mathcal{H}(p_2) - \mathcal{H}(p)\big) + \mathcal{Q} + \mathbb{E}^* - \mathbb{E} \\ &\leq (1-\lambda) \cdot [\mathcal{H}^* + \mathcal{R}] + \mathcal{Q} \end{aligned}$$

Here, $\mathcal{H}^* = \mathcal{H}(u) - \mathcal{H}(p)$ is always semi-positive and $\mathcal{R} = \mathrm{KL}(u\|f) - \mathrm{KL}(p\|f)$. The difference of the two entropy could also be $(1-\lambda) \cdot \mathcal{H}(u, f) - \mathcal{H}(p, f)$ and then we could complete the proof, i.e., $\mathcal{Q} \geq 0$.

■

A.2. Proof of Proposition 1

Proposition. 1 *Let ℓ_{sce}, ℓ_{rce} be the symmetric and reverse cross entropy loss function respectively and γ represents their summation, i.e., $\ell_{sce} + \ell_{rce} = \gamma$. When $\gamma \to 1$, then our methods can also be written as:*

$$\ell_{sglr} = \ell_{sce} - \alpha \cdot \ell_{rkl}$$

where ℓ_{rkl} denotes the reverse KL divergence between labels and model predictions, i.e., $D_{KL}(p \,||\, q)$.

Proof. Formally, the conventional SCE loss can be written as:

$$\ell_{sce} = \alpha \cdot \ell_{ce} + \beta \cdot \ell_{rce}$$

Note that $\alpha, \beta = 1$ is a special case of this form. We can still let $\gamma \to 1$, then

$$\ell_{sce} = \alpha \cdot \ell_{ce} + (1-\alpha) \cdot \ell_{rce} \tag{A.5}$$

We take a closer look at the self-guided soft label and write $p(k; x)$ as $p(k)$ for simplicity [1].

$$\begin{aligned} q'(k) &= (1-\alpha) \cdot q(k) + \alpha \cdot p(k) \\ \mathcal{H}(q'(k),\, p(k)) &= -\sum_{k=1}^{K} (1-\alpha) \cdot q(k) \cdot \log p(k) \\ &\quad + \alpha \cdot p(k) \cdot \log p(k) \\ &= (1-\alpha) \cdot \mathcal{H}(q,\, p) + \alpha \cdot \mathcal{H}(p) \end{aligned} \tag{A.6}$$

[1] We omit the adversarial knowledge and historical average prediction temporarily without loss of generality.

where $q(\cdot)$ is the ground truth distribution over the labels and $\mathcal{H}(\cdot)$ denotes the cross entropy loss. Recall that the Kullback-Leibler Divergence could be dubbed as information gain, *i.e.*, $D_{KL}(p\,||\,q) = \mathcal{H}(p,\,q) - \mathcal{H}(p)$. As $\gamma \to 1$, then the cross entropy loss of our method can also be written as:

$$\begin{aligned}\mathcal{H}(q'(k),\,p(k)) =&(1-\alpha)\cdot\mathcal{H}(q,\,p) + \alpha\cdot\mathcal{H}(p,\,q)\\ &-\alpha\cdot D_{KL}(p\,||\,q)\\ =&\ell_{sce} + \alpha\cdot\ell_{rkl}\end{aligned} \tag{A.7}$$

■

A.3. Proof of Proposition 2

Proposition. 2 *Some KD methods, which minimize the distance of the feature map between the teacher and student model, belong to the family of our method. Let p_t be the prediction of the teacher model and then the KD could also be written as $\ell_{KD} = \mathbb{E}_{\widetilde{q}}\left[-\log p\right] = \mathcal{H}(\widetilde{q},\,p)$, where $\widetilde{q} = (1-\alpha)\cdot q + \alpha\cdot p_t$.*

Proof. Firstly, for the self-guided soft label $q' = (1-\alpha)\cdot q + \alpha\cdot p$, if we replace the self-prediction p with the knowldege of teacher model p_t, we have

$$\widetilde{q} = (1-\alpha)\cdot q + \alpha\cdot p_t \tag{A.8}$$

We utilize the special form $\widetilde{q}$ and have

$$\begin{aligned}\mathcal{H}(\widetilde{q},\,p) &= -\sum\widetilde{q}\cdot\log p\\ &= -(1-\alpha)\sum q\cdot\log p - \alpha\sum p_t\cdot\log p\\ &= (1-\alpha)\cdot\mathcal{H}(q,\,p) + \alpha\cdot\mathcal{H}(p_t,\,p)\end{aligned} \tag{A.9}$$

We here apply KL equality again

$$D_{\mathrm{KL}}(p_t\,||\,p) = \mathcal{H}(p_t,\,p) - \mathcal{H}(p_t) \tag{A.10}$$

Note that $\mathcal{H}(p_t)$ represents the entropy of teacher prediction. When teacher is fixed, $\mathcal{H}(p_t)$ is a constant so that we can miss it during loss minimization. Then the loss of special soft label can be written as:

$$\mathcal{H}(\widetilde{q},\,p) = (1-\alpha)\cdot\mathcal{H}(q,\,p) + \alpha\cdot D_{\mathrm{KL}}(p_t\,||\,p) \tag{A.11}$$

Some KD methods minimize this loss function $(1-\alpha)\cdot\mathcal{H}(q,\,p) + \alpha\cdot D_{\mathrm{KL}}(p_t\,||\,p)$ and belong to the family of our method.

■

A.4. Proof of Theorem 2

Theorem 2 *In a K-class classification problem, $\widetilde{\ell}$ is noise-tolerant under symmetric or uniform label noise if noise rate $\eta < 1 - \frac{1}{K}$. And if $R(f^*) = 0$, $\widetilde{l}$ is also noise-tolerant under asymmetric or class-dependent label noise when noise rate $\eta_{y,k} < 1-\eta_y$ with $\sum_{i\neq y}\eta_{y,i} = \eta_y$, then*

$$R^{\eta}_{\mathcal{S}}(f^*) - R^{\eta}_{\mathcal{S}}(f) \simeq (1-\frac{\eta K}{K-1})(R_{\mathcal{S}}(f^*) - R_{\mathcal{S}}(f)) \leq 0$$

Proof. *For symmetric noise:*

$$\begin{aligned}R^{\eta}_{\mathcal{S}}(f) =&\mathbb{E}_{x,y}\widetilde{\ell}(x,y) = \mathbb{E}_x\mathbb{E}_{y|x}\mathbb{E}_{\eta|x,y}\widetilde{\ell}(x,y)\\ =&\mathbb{E}_x\mathbb{E}_{y|x}\left[(1-\eta)\widetilde{\ell}(x,y) + \frac{\eta}{K-1}\sum_{k\neq y}^{K}\widetilde{\ell}(x,k)\right]\\ =&(1-\eta)R_{\mathcal{S}}(f)\\ &+\frac{\eta}{K-1}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\widetilde{\ell}(x_i,k) - R_{\mathcal{S}}(f)\right)\\ =&R_{\mathcal{S}}(f)\left(1-\frac{\eta K}{K-1}\right) - \eta\cdot G(f)\end{aligned}$$

where $G(f) = -\frac{1}{N(K-1)}\sum_{i=1}^{N} f(x_i)\cdot\log f(x_i)$ depends on the performance of classifiers. Then we are more preoccupied with the magitude of the difference between $G(f^*)$ and $G(f)$. Considering that the global minimum is adept at fitting the training data with nearly zero loss and we can make an appropriate assumption that $G(f^*)$ tends to be zero. Besides, we calculate the worst-case of the classifier, which randomly guesses the given data, and the entropy achieve maximum value. The difference could be approximated to $C = \eta\cdot(1-\frac{1}{K})\cdot 0.1$ and thus,

$$R^{\eta}_{\mathcal{S}}(f^*) - R^{\eta}_{\mathcal{S}}(f) = (1-\frac{\eta K}{K-1})(R_{\mathcal{S}}(f^*) - R_{\mathcal{S}}(f)) + C$$

since f^* is the global minimum of $R_{\mathcal{S}}(f)$ undert noise free data and the constant C could be approximately igore as the small magnitude. This proves that f^* is also the the global minimizer of $R^{\eta}_{\mathcal{S}}(f)$ and our method is nearly noise-tolerant.

For asymmetric or class-dependent noise, $1-\eta_y$ is the probability of a label being correct (*i.e.*, $k = y$), and the noise condition $\eta_{y,k} < 1-\eta_y$ generally states that a sample x still has the highest probability of being in the correct class y, though it has probability of $\eta_{y,k}$ being in an arbitrary noisy (incorrect) class $k \neq y$. Following the symmetric case, here we set $C = \max\sum_{i=1}^{N}\sum_{k=1}^{K}\widetilde{\ell}(x_i,k)$ and thus,

$$R^{\eta}(f) =\mathbb{E}_{x,y}\widetilde{\ell}(x,y) = \mathbb{E}_x\mathbb{E}_{y|x}\mathbb{E}_{y|x,y}\widetilde{\ell}(x,y) \tag{A.12}$$

$$=\mathbb{E}_x\mathbb{E}_{y|x}\left[(1-\eta_y)\,\widetilde{\ell}(x,y) + \sum_{k\neq y}\eta_{yk}\widetilde{\ell}(x,k)\right] \tag{A.13}$$

$$\begin{aligned}=&\mathbb{E}_{x,y}\left[(1-\eta_y)\left(\sum_{k=1}^{K}\widetilde{\ell}(x,k) - \sum_{k\neq y}\widetilde{\ell}(x,k)\right)\right]\\ &+\mathbb{E}_{x,y}\left[\sum_{k\neq y}\eta_{yk}\widetilde{\ell}(x,k)\right]\end{aligned} \tag{A.14}$$

$$=\mathbb{E}_{x,y}\left[(1-\eta_y)\left(C - \sum_{k\neq y}\widetilde{\ell}(x,k)\right)\right]$$

$$+\mathbb{E}_{x,y}\left[\sum_{k\neq y}\eta_{yk}\widetilde{\ell}(x,k)\right] \tag{A.15}$$

$$=C\cdot\mathbb{E}_{x,y}\left(1-\eta_y\right) - \mathbb{E}_{x,y}\left[\sum_{k\neq y}\left(1-\eta_y-\eta_{yk}\right)\widetilde{\ell}(x,k)\right]. \tag{A.16}$$

Since $\eta_{y,k} < 1-\eta_y$ and f_η^* is the minimizer of $R^\eta(f)$, $R^\eta(f_\eta^*) - R^\eta(f^*) \leq 0$. So, from Eq. (A.16),

$$\mathrm{E}_{x,y}\left[\sum_{k\neq y}\left(1-\eta_y-\eta_{yk}\right)\left(\widetilde{\ell}\left(f^*(\mathbf{x}),k\right)-\widetilde{\ell}\left(f_\eta^*(\mathbf{x}),k\right)\right)\right]\leq 0. \tag{A.17}$$

This proves that under asymmetric noise setting f^* is also the the global minimizer of $R_{\mathcal{S}}^\eta(f)$ and our method is noise-tolerant.

■

B. More results

Figure B.1. Label noise in CIFAR-10 training dataset.

B.1. Label noise in common datasets

As reported in the study in [23], label noise is surprisingly common in benchmark datasets such as CIFAR-10, which is presented in Fig. B.1. We note that some examples are mislabelling (*e.g.*, dog labeled with cat in Fig. B.1). Such erroneously annotated samples make it hard for models to learn a good decision boundary. Unsurprising, feeding the model with such noisy labels would inevitably exacerbate the problem of robust overfitting and leads to over confidence.

B.2. More results about calibration

A ideally trustworthy and reliable model ought to provide high confidence prediction on correct category and contrariwise. So, we adopt expected calibration error (ECE) for a model f with $0 < m < \infty$, as suggested in [19].

$$\mathrm{ECE}_p = \mathbb{E}\left[\mid \hat{z} - \mathbb{E}[\delta_{y,\hat{y}}\hat{z}] \mid^m\right]^{\frac{1}{m}}$$

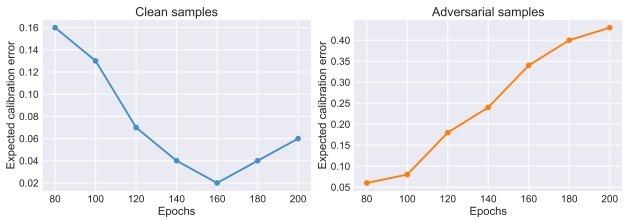

Figure B.2. Expected calibration error for adversarial training on clean and adversarial samples. The learning rate is decayed by a factor of 0.1 at 100-th and 150-th.

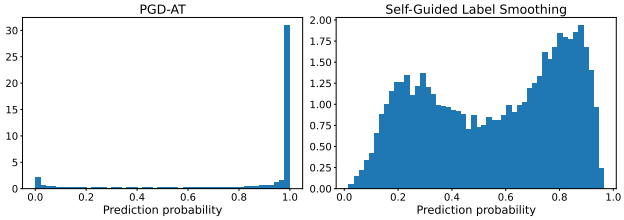

Figure B.3. Samples density *w.r.t.* the prediction probabilities (the softmax outputs on the labeled class).

where $\delta_{i,j}$ is the Kronecker delta, which appears true if the variables are equal. We visualize the ECE during whole adversarial training procedure in Fig. B.2.

From Fig. B.2, we can observe that vanilla adversarial training weakens the over-confident prediction on clean samples, thus achieving a good calibration than standard training. However, as the training progresses, the expected calibration error on adversarial samples shows a rapid ramp-up. It also indicates that the latter predition of the adversarially trained model is not well-calibrated and thus being not able to provide trustworthy knowledge.

Table B.1. Expected calibration error for PGD-AT and our proposed method on white-box (PGD-10) attack and black-box (square) attack.

	White-box (PGD attack)			Black-box (Square attack)		
	Best ↓	Final ↓	Diff ↓	Best ↓	Final ↓	Diff ↓
AT	0.18	0.43	-0.25	0.07	0.39	-0.32
+SGLR	0.11	0.10	0.01	0.20	0.22	-0.02

Further, we also report the expected calibration error for PGD-AT and our method on both best and final checkpoint under different attacks. From Tab. B.1, we can observe that our method effectively decrease the calibration error and thus alliviate the over confident prediction. Additionally, we plot the sample density *w.r.t.* predictions on the labeled class in Fig. B.3. We reach the same conclusion that our proposed method reduce the overconfidence mostly and thus achieve good generalization.

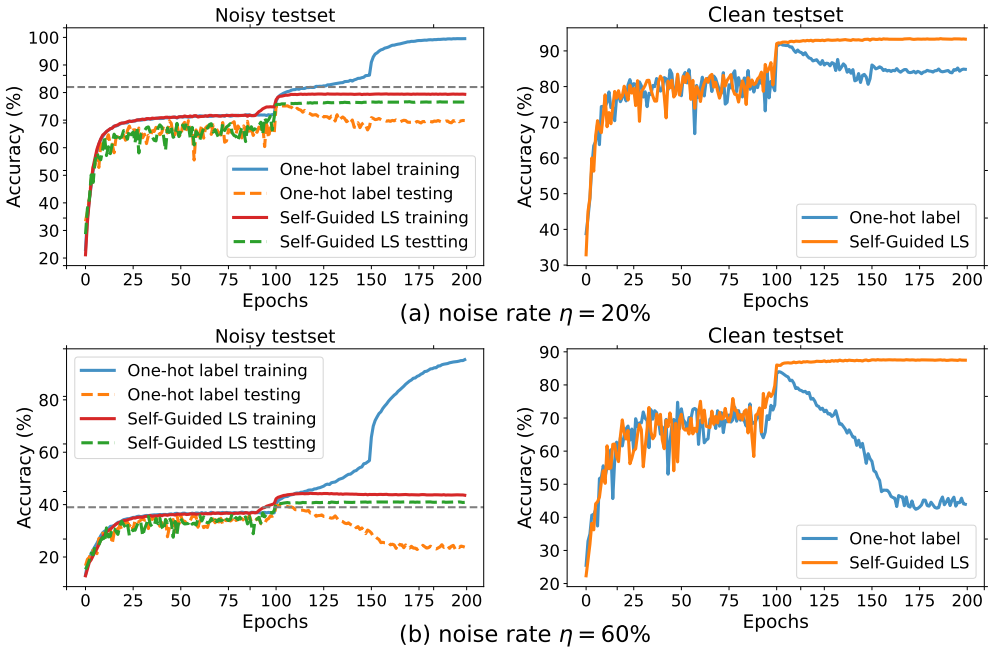

Figure B.4. Test accuracy (%) on clean and different noisy CIFAR-10 testset. The horizontal gray dashed line denotes the portion of correct labels.

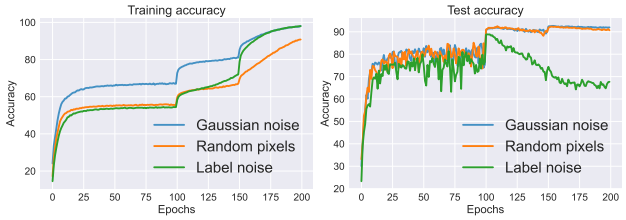

Figure B.5. Training and test accuracy (%) under different noise type in CIFAR-10 training dataset.

B.3. More experiments under different noise settings

As reported in [22], robust overfitting has been prevalent across various datasets and models. However, it may not occur while the strength of attack is relatively weak. We observe that smaller perturbation does not lead to a double-descent test accuracy and even obtains an overall increase, while large perturbation induces robust overfitting.

As discussed in Sec. 2, robust overfitting is similar to label noise in standard training to some extend. Namely, when the noise rete ramps up, it is on the verge of occuring with double descent curves.

From Fig. B.5, we can obverve that other noise types (like Gaussian and random pixels), even with extremely large perturbation, do not incur overfitting. The training accuracy and test accuracy consistently increase as the learning rate decayed though the clean test accuracy oscillates in its infancy.

Besides, we provide the training procedure of hard label and self-guided soft label over different noise rate on CIFAR-10 dataset in Fig. B.4 and Tab. B.2 . It is worth noting that self-guided soft label constantly narrow the training and testing gap as the noise rate ramps up, while the hard label still memorizes the training data and eventually leads to bad testing accuracy.

Table B.2. Evaluating different label strategies at various noise rates.

Rate	0%	20%	60%	80%
Hard Label	26.8	37.8	62.1	81.6
Soft Label	26.4	35.5	59.5	80.4
Self-Guided Soft Label	25.5	33.8	51.3	78.9

B.4. More experiments about black-box attacks and large model architecture

Additionally, we present evaluations on black-box attacks, *i.e.*, adversarial examples generated from a different model (typiclally from a larger model), in Tab. B.3. Here, we

Table B.3. Performance (%) of PGD-AT and our proposed method against different black-box attacks.

ResNet-34 → ResNet-18	PGD-10			CW_∞		
	Best	Final	Diff	Best	Final	Diff
AT	63.9	64.9	-1.0	72.5	70.1	2.4
AT+SGLR	64.0	64.1	-0.1	72.7	72.9	-0.2
ResNet-50 → ResNet-18	**PGD-10**			**CW_∞**		
	Best	Final	Diff	Best	Final	Diff
AT	81.3	82.7	-1.4	83.1	81.6	1.5
AT+SGLR	80.9	80.9	0.0	83.0	82.8	0.2

Table B.4. Clean accuracy and robust accuracy (%) against white-box attacks of networks. All threat models are under ℓ_∞ norm with $\epsilon = 8/255$. The bold indicates the improved performance achieved by the proposed method.

Method	Natrural Accuracy			PGD-20			AutoAttack		
	Best	Final	Diff ↓	Best	Final	Diff ↓	Best	Final	Diff ↓
ResNet-18									
AT	80.7	82.4	-1.6	50.7	41.4	9.3	47.7	40.2	7.5
+SGLR	**82.9**	**83.0**	**0.1**	**56.4**	**55.9**	**0.5**	**51.2**	**50.2**	**1.0**
ResNet-34-10									
AT	**87.6**	86.4	1.2	55.9	50.2	5.7	51.2	45.6	5.6
+SGLR	87.4	**87.2**	**0.2**	**59.5**	**58.0**	**1.5**	**54.3**	**52.3**	**2.0**

test ResNet-18 trained under PGD-AT and our proposed method with crafted adversarial examples from ResNet-34 and ResNet-50 trained with vanilla AT. Results in Tab. B.3 demonstrate that our method indeed close the gap between best and final checkpoint. These results not only show that our method does not suffer from the gradient obfuscation but also show that our method is effective in black-box attack settings.

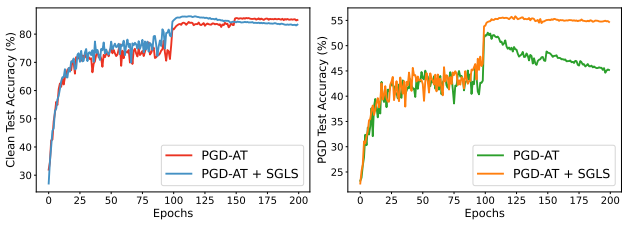

Figure B.6. Result of training and testing accuracy over epochs for ResNet-18 trained on CIFAR-10.

Furthermore, performance against various white-box attacks for large model architecture are shown in Tab. B.4. We similarly found that combing PGD-AT with our method could achieve superior performance even under strong autoattack. Notably, our method can largely reduce the gap between the best and final accuracies and thus effectively prevent robust overfitting.

B.5. Different learning rate strategies

The staircase learning rate schedule (piece-wise) is typiclally applied in adversarial training, which may have negative influence in obtaining robust models. In Fig. B.7, we plot the *test robust accuracy*, *gradient norm* and *trace of hessian*, which is widely used to measure the sharpness. As shown, training with cosine learning rate schedule yields smoother curves compared to that of the piece-wise learning rate schedule. Note that it does not prevent the widening generalization gap and robust overfitting, only influencing the **duration of the Stationary Stage**. The green in Fig. B.7 supplements this with the trace of hessian, to better illustrate the characteristics between the two stages.

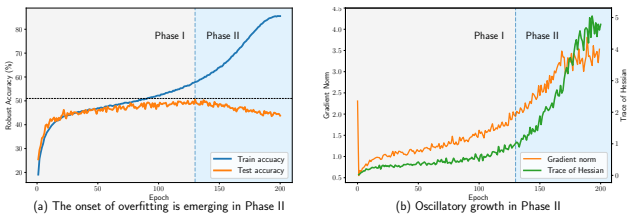

Figure B.7. Different phases of training.

C. Algorithm

Algorithm 1 Self-Guided Label Smoothing

Require: Total Epoch N, Neural Network f_θ with parameters θ, Training Set $\mathcal{D} = \{(x_j, y_j)\}$.

$\widetilde{p_t} = 0$

for all epoch $= 1, \cdots, N$ **do**

  for all $(x_j, y_j) \in \mathcal{D}$ **do**

    * Inner Maximization to update δ

      $\delta \leftarrow \arg\max_{\|\delta\|_p \leq \epsilon} \ell(f(x), y)$

    * Outer Minimization to update θ

      $\widetilde{f}(x, x'; \theta_t) = \lambda \cdot f(x; \theta) + (1 - \lambda) \cdot f(x'; \theta)$ {# Self-Guided Label Refinement}

      $\mathbf{y} = r \cdot \widetilde{p_t} + (1 - r) \cdot \mathbf{y}_{hard}$

      $\widetilde{p_t} = \alpha \cdot \widetilde{p}_{t-1} + (1 - \alpha) \cdot \widetilde{f}(x, x'; \theta_t)$ {# Consensus of the self-distilled models}

      $\ell_{sglr} = \ell(f(x), \widetilde{p_t})$

      $\theta \leftarrow \theta - \eta \cdot (\nabla_\theta \widetilde{L})$

  end for

end for