Appendix for "Outlier Robust Adversarial Training"

Appendix A. Related Works

Traditional Robust Learning. Training accurate machine learning models in the presence of noisy data is of great practical importance (Sukhbaatar et al., 2015). However, a degradation in the performance of classification models is inevitable when there are outliers in the training data. To combat outliers, the traditional robust learning methods are designed from four directions. 1) The *label correction methods* (Wang et al., 2018) improve the quality of the raw labels by correcting wrong labels into correct ones. However, it requires an extra clean dataset or potentially expensive detection process to estimate the outliers. 2) The *loss correction methods* (Han et al., 2020) improve the robustness by modifying the loss function based on an estimated noise transition matrix that defines the probability of mislabeling one class with another. However, these methods are sensitive to the noise transition matrix, which is also hard to be estimated. 3) The *refined training strategies* such as Co-teaching (Yu et al., 2019), MentorNet (Jiang et al., 2018) are robust to outliers. These studies all rely on an auxiliary network for sample weighting or learning supervision, which is hard to adapt and tune. 4) Some simpler and arguably generic *robust loss functions* are also designed for robust learning. For example, a recent work Hu et al. (2020) proposed AoRR loss, which can mitigate the influence of the outliers if their proportion in training data is known. Furthermore, Some smoothing methods are proposed in Chaudhari et al. (2019) and have been proven to be effective in solving the problems of label and data noise. However, none of these methods are related to adversarial robust learning.

Adversarial Robust Learning. The omnipotent DNN models are surprisingly vulnerable to adversarial examples (Goodfellow et al., 2015), which can easily mislead a DNN model to make erroneous predictions. To mitigate this issue, the adversarial training (AT) (Madry et al., 2018) is first proposed as one of the most effective robust learning methodologies against adversarial attacks. To improve adversarial robustness, instance-reweighted AT methods are studied by considering the unequal importance of the adversarial data in several recent works. Intuitively, the samples assigned a low weight to correspond to samples on which the classifier is already sufficiently robust. Specifically, the reweight mechanism in WMMR (Zeng et al., 2021a) and MAIL (Liu et al., 2021) is based on the multi-class probabilistic margin of the model outputs (Zhang and Liang, 2019). The reweighting method in work GAIRAT (Zhang et al., 2021) identifies non-robust (easily be-attacked) data by estimating how many steps the PGD method needs to attack natural data successfully. The most recent work BiLAW (Holtz et al., 2021) uses a validation set to learn weights based on bi-level optimization and meta-learning. The most significant assumption in these works is that the natural dataset is clean. However, the performance of the model based on these methods will be degraded if the training dataset contains outliers. In Sanyal et al. (2021), the authors identified label noise as one of the causes of adversarial vulnerability. However, no defense methods are proposed to solve this problem. The work Zhu et al. (2021) empirically studies the efficacy of AT for mitigating the effect of label noise in training data. However, their proposed annotator algorithm is based on the label correction strategy, which inevitably introduces more extra noisy labels due to the bottleneck of the classifier. In Dong et al. (2020), the authors proposed an adversarial distributional training. They focus on the distribution shift of adversarial samples but they do not consider the outliers problem. Several works (Augustin et al., 2020; Bitterwolf et al., 2020) connect adversarial robustness to out-of-distribution (OOD) problems. However, they are in different settings from ours because the notion of outliers is different from OOD points. Dong et al. (Dong et al.) also discuss the effect of the label noise. However, they focus on the memorization effect in AT. We focus on outlier problems in AT. Huang et al. (Huang et al., 2020) created a self-adaptive method for robust learning with noisy labels or adversarial examples, but did not consider both present simultaneously. This is also mentioned in Zhu et al. (2021).

Appendix B. Explicit Forms of (sub)gradients

From Eq.(4), we have $\hat{\mathcal{L}}(f_\theta, \lambda, \hat{\lambda}) := \frac{k-m}{n}\lambda + \frac{n-m}{n}\hat{\lambda} - [\hat{\lambda} - [\ell(f_\theta(\tilde{\mathbf{x}}_i), y_i) - \lambda]_+]_+$. Denote $\mathbb{I}_{[a]}$ as an indicator function with $\mathbb{I}_{[a]} = 1$ if a is true and 0 otherwise. Then we can get

$$\partial_\theta \hat{\mathcal{L}}(f_{\theta^{(t)}}, \lambda^{(t)}, \hat{\lambda}^{(t)}) = \partial \ell(f_{\theta^{(t)}}(\tilde{\mathbf{x}}_i), y_i) \cdot \mathbb{I}_{[\hat{\lambda}^{(t)} > [\ell(f_{\theta^{(t)}}(\tilde{\mathbf{x}}_i), y_i) - \lambda^{(t)}]_+]} \cdot \mathbb{I}_{[\ell(f_{\theta^{(t)}}(\tilde{\mathbf{x}}_i), y_i) > \lambda^{(t)}]},$$

$$\partial_\lambda \hat{\mathcal{L}}(f_{\theta^{(t)}}, \lambda^{(t)}, \hat{\lambda}^{(t)}) = \frac{k-m}{n} - \mathbb{I}_{[\hat{\lambda}^{(t)} > [\ell(f_{\theta^{(t)}}(\tilde{\mathbf{x}}_i), y_i) - \lambda^{(t)}]_+]} \cdot \mathbb{I}_{[\ell(f_{\theta^{(t)}}(\tilde{\mathbf{x}}_i), y_i) > \lambda^{(t)}]},$$

$$\partial_{\hat{\lambda}} \hat{\mathcal{L}}(f_{\theta^{(t)}}, \lambda^{(t)}, \hat{\lambda}^{(t)}) = \frac{n-m}{n} - \mathbb{I}_{[\hat{\lambda}^{(t)} > [\ell(f_{\theta^{(t)}}(\tilde{\mathbf{x}}_i), y_i) - \lambda^{(t)}]_+]}.$$

Appendix C. Proofs

C.1. Proof of Theorem 1

Denote $[a]_+ = \max\{0, a\}$ as the hinge function. First, we introduce two Lemmas as follows,

Lemma C.1 *(Hu et al., 2020) For a set of real numbers $S = \{s_1, \cdots, s_n\}$, $s_i \in \mathbb{R}$, and $s_{[i]}$ represents the i-th largest value after sorting the elements in S, we have*

$$\sum_{i=1}^{k} s_{[i]} = \min_{\lambda \in \mathbb{R}} \left\{ k\lambda + \sum_{i=1}^{n} [s_i - \lambda]_+ \right\}.$$

Furthermore, $s_{[k]} \in \arg\min_{\lambda \in \mathbb{R}} \{k\lambda + \sum_{i=1}^{n} [s_i - \lambda]_+\}$.

Proof We know $\sum_{i=1}^{k} s_{[i]}$ is the solution of

$$\max_{\mathbf{p}} \mathbf{p}^\top S, \text{ s.t. } \mathbf{p}^\top \mathbf{1} = k, \mathbf{0} \le \mathbf{p} \le \mathbf{1}.$$

We apply Lagrangian to this equation and get

$$L = -\mathbf{p}^\top S - \mathbf{v}^\top \mathbf{p} + \mathbf{u}^\top (\mathbf{p} - 1) + \lambda(\mathbf{p}^\top \mathbf{1} - k)$$

where $\mathbf{u} \ge \mathbf{0}$, $\mathbf{v} \ge \mathbf{0}$ and $\lambda \in \mathbb{R}$ are Lagrangian multipliers. Taking its derivative w.r.t. $\mathbf{p}$ and set it to 0, we have $\mathbf{v} = \mathbf{u} - S + \lambda \mathbf{1}$. Substituting it back into the Lagrangian, we get

$$\min_{\mathbf{u}, \lambda} \mathbf{u}^\top \mathbf{1} + k\lambda, \text{ s.t. } \mathbf{u} \ge \mathbf{0}, \mathbf{u} + \lambda \mathbf{1} - S \ge 0.$$

This means

$$\sum_{i=1}^{k} s_{[i]} = \min_{\lambda} \Big\{ k\lambda + \sum_{i=1}^{n} [s_i - \lambda]_+ \Big\}. \tag{C.1}$$

Furthermore, we can see that $\lambda = s_{[k]}$ is always one optimal solution for Eq.(C.1). So

$$s_{[k]} \in \arg\min_{\lambda} \Big\{ k\lambda + \sum_{i=1}^{n} [s_i - \lambda]_+ \Big\}.$$

■

Lemma C.2 *For a set of real numbers $S = \{s_1, \cdots, s_n\}$, $s_i \in \mathbb{R}$, we have*

$$\sum_{i=m+1}^{n} s_{[i]} = \max_{\lambda \in \mathbb{R}} \Big\{ (n-m)\lambda - \sum_{i=1}^{n} [\lambda - s_i]_+ \Big\}.$$

Furthermore, $s_{[m]} \in \arg\max_{\lambda\in\mathbb{R}} \{(n-m)\lambda - \sum_{i=1}^{n} [\lambda - s_i]_+\}$.

Proof

$$\begin{aligned}
\sum_{i=m+1}^{n} s_{[i]} &= \sum_{i=1}^{n} s_i - \sum_{i=1}^{m} s_{[i]} \\
&= \sum_{i=1}^{n} s_i - \min_{\lambda} \Big\{ m\lambda + \sum_{i=1}^{n} [s_i - \lambda]_+ \Big\} \\
&= -\min_{\lambda} \Big\{ -\sum_{i=1}^{n} (s_i - \lambda) - (n-m)\lambda + \sum_{i=1}^{n} [s_i - \lambda]_+ \Big\}. \\
&= -\min_{\lambda} \Big\{ -(n-m)\lambda + \sum_{i=1}^{n} [\lambda - s_i]_+ \Big\} \\
&= \max_{\lambda} \Big\{ (n-m)\lambda - \sum_{i=1}^{n} [\lambda - s_i]_+ \Big\}
\end{aligned}$$

The second equation holds because of Lemma C.1. The fourth equation holds because the fact of $[a]_+ - a = [-a]_+$. Furthermore, we can see that $\lambda = s_{[m]}$ is always one optimal solution. So

$$s_{[m]} \in \arg\max_{\lambda \in \mathbb{R}} \Big\{ (n-m)\lambda - \sum_{i=1}^{n} [\lambda - s_i]_+ \Big\}.$$

■

Theorem C.3 *(Theorem 1 restated) Suppose $\lambda \in \mathbb{R}$, $\hat{\lambda} \in \mathbb{R}$, then Eq.(3) is equivalent to*

$$\begin{aligned}
\min_{\theta,\lambda} \max_{\hat{\lambda}} \quad & \frac{1}{k-m} \sum_{i=1}^{n} \Big[\frac{k-m}{n}\lambda + \frac{n-m}{n}\hat{\lambda} - [\hat{\lambda} - [\ell(f_\theta(\tilde{\mathbf{x}}_i), y_i) - \lambda]_+]_+ \Big] \\
s.t. \quad & \tilde{\mathbf{x}}_i = \arg \max_{\tilde{\mathbf{x}} \in \mathcal{B}_\epsilon(\mathbf{x}_i)} \ell(f_\theta(\tilde{\mathbf{x}}), y_i)
\end{aligned} \tag{C.2}$$

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

Proof To extract the sum of (m,k)-ranked range individual losses, we can first select a subset, which contains the bottom $n-m$ losses from the ranked list of $L\left(\{(\mathbf{x}_j,y_j)\}_{j=1}^n\right)$. Then we select top-$(k-m)$ individual losses from this subset as the finalized (m,k)-ranked range. Therefore, We sum the bottom $n-m$ individual losses as follows,

$$\sum_{i=m+1}^{n}\ell(f_\theta(\tilde{\mathbf{x}}_{[i]}),y_{[i]})=\min_q\sum_{i=1}^{n}q_i\ell(f_\theta(\tilde{\mathbf{x}}_{[i]}),y_{[i]})\quad\text{s.t. } q_i\in\{0,1\},\ ||q||_0=n-m,$$

where $q=\{q_1,\cdots,q_n\}\in\{0,1\}^n$, and q_i is an indicator. When $q_i=0$, it indicates that the i-th individual loss is not included in the objective function. Otherwise, the objective function should include this individual loss. Next, we sum the top-$(k-m)$ individual losses from the bottom $n-m$ individual losses as follows,

$$\begin{aligned}
&\min_q\sum_{i=1}^{k-m}(q\ell(f_\theta(\tilde{\mathbf{x}}),y))_{[i]}\quad\text{s.t. } q_i\in\{0,1\},\ ||q||_0=n-m\\
=&\min_{\lambda,q}(k-m)\lambda+\sum_{i=1}^{n}[q_i\ell(f_\theta(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+\quad\text{s.t. } q_i\in\{0,1\},\ ||q||_0=n-m\\
=&\min_{\lambda,q}(k-m)\lambda+\sum_{i=1}^{n}q_i[\ell(f_\theta(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+\quad\text{s.t. } q_i\in[0,1],\ ||q||_0=n-m\qquad\text{(C.3)}\\
=&\min_{\lambda}(k-m)\lambda+\sum_{i=m+1}^{n}[[\ell(f_\theta(\tilde{\mathbf{x}}),y)-\lambda]_+]_{[i]}\\
=&\min_{\lambda}(k-m)\lambda+\max_{\hat{\lambda}}\Big\{(n-m)\hat{\lambda}-\sum_{i=1}^{n}[\hat{\lambda}-[\ell(f_\theta(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+]_+\Big\},
\end{aligned}$$

where $q\ell(f_\theta(\tilde{\mathbf{x}}),y)=\{q_1\ell(f_\theta(\tilde{\mathbf{x}}_1),y_1),\cdots,q_n\ell(f_\theta(\tilde{\mathbf{x}}_n),y_n)\}$. The first equation holds because of Lemma C.1. Since $q_i\ell(f_\theta(\tilde{\mathbf{x}}_i),y_i)\geq 0$, we know the optimal $\lambda^*\geq 0$ from Lemma C.1. If $q_i=0$, $[q_i\ell(f_\theta(\tilde{\mathbf{x}}_i),y_i)-\lambda^*]_+=0=q_i[\ell(f_\theta(\tilde{\mathbf{x}}_i),y_i)-\lambda^*]_+$. If $q_i=1$, $[q_i\ell(f_\theta(\tilde{\mathbf{x}}_i),y_i)-\lambda^*]_+=[\ell(f_\theta(\tilde{\mathbf{x}}_i),y_i)-\lambda^*]_+=q_i[\ell(f_\theta(\tilde{\mathbf{x}}_i),y_i)-\lambda^*]_+$. Thus the second equation holds. It should be mentioned that the discrete indicator q_i can be replaced by a continue one, which means $q_i\in[0,1]$. The third equation holds because we take the optimal q^* into the objective function and remove the constraints. The fourth equation can be obtained by applying Lemma C.2.

Therefore,

$$\begin{aligned}
&\min_\theta\frac{1}{k-m}\sum_{i=m+1}^{k}\ell(f_\theta(\tilde{x}_{[i]}),y_{[i]})\\
=&\min_\theta\frac{1}{k-m}\Bigg\{\min_\lambda(k-m)\lambda+\max_{\hat{\lambda}}\Big\{(n-m)\hat{\lambda}-\sum_{i=1}^{n}[\hat{\lambda}-[\ell(f_\theta(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+]_+\Big\}\Bigg\}\qquad\text{(C.4)}\\
=&\min_{\theta,\lambda}\max_{\hat{\lambda}}\frac{1}{k-m}\sum_{i=1}^{n}\Big[\frac{k-m}{n}\lambda+\frac{n-m}{n}\hat{\lambda}-[\hat{\lambda}-[\ell(f_\theta(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+]_+\Big].
\end{aligned}$$

Furthermore, according to Lemma C.1 and C.2, we know the optimal λ^* and $\hat{\lambda}^*$ can be obtained at the top-k and top-m values of loss ℓ, respectively. Since $m < k$, we have $\lambda^* < \hat{\lambda}^*$. Therefore, $\hat{\lambda} > \lambda$, when the optimal solution is achieved. ■

C.2. Proof of Theorem 4

To prove Theorem 4, we first introduce the calibration function as follows,

Definition C.4 *(Calibration function). (Awasthi et al., 2021) Given a hypothesis set $\mathcal{H}$, we define the calibration function $\delta_{\max}$ for a pair of losses (ℓ_1, ℓ_2) as follows: for all $\mathbf{x} \in \mathcal{X}$, $\eta \in [0,1]$ and $\tau > 0$,*

$$\delta_{\max}(\tau, \mathbf{x}, \eta) = \inf_{f\in\mathcal{H}}\{\mathcal{C}_{\ell_1}(f, \mathbf{x}, \eta) - \mathcal{C}^*_{\ell_1,\mathcal{H}}(\mathbf{x}, \eta) | \mathcal{C}_{\ell_2}(f, \mathbf{x}, \eta) - \mathcal{C}^*_{\ell_2,\mathcal{H}}(\mathbf{x}, \eta) \geq \tau\}. \tag{C.5}$$

The calibration function gives the maximal δ satisfying the calibration condition (Definition 3). The following proposition is an important result from Steinwart (2007).

Proposition C.5 *(Steinwart, 2007). Given a hypothesis set $\mathcal{H}$, loss ℓ_1 is $\mathcal{H}$-calibrated with respect to ℓ_2 if and only if its calibration function $\delta_{\max}$ satisfies $\delta_{\max}(\tau, \mathbf{x}, \eta) > 0$ for all $\mathbf{x} \in \mathcal{X}$, $\eta \in [0,1]$, and $\tau > 0$.*

Next, we define the adversarial loss of $f \in \mathcal{H}$ at $(\mathbf{x}, y)$ as

$$\widetilde{\ell}_s(f, \mathbf{x}, y) = \sup_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} \ell_s(yf(\tilde{\mathbf{x}})). \tag{C.6}$$

The above naturally motivates supremum-based surrogate losses that are commonly used to optimize the adversarial 0/1 loss (Goodfellow et al., 2015; Madry et al., 2018; Zhang et al., 2019). When ℓ_s is non-increasing, the following equality holds (Yin et al., 2019):

$$\sup_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} \ell_s(yf(\tilde{\mathbf{x}})) = \ell_s\Big(\inf_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} yf(\tilde{\mathbf{x}})\Big). \tag{C.7}$$

Therefore, the adversarial 0/1 loss $\widetilde{\ell}_0$ has the equivalent form

$$\widetilde{\ell}_0(f, \mathbf{x}, y) := \sup_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} \mathbb{1}_{yf(\tilde{\mathbf{x}})\leq 0} = \mathbb{1}_{\inf_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} yf(\tilde{\mathbf{x}})\leq 0}. \tag{C.8}$$

In this paper, we aim to characterize surrogate losses ℓ_1 satisfying $\mathcal{H}$-calibration (Definition 3) with $\ell_2 = \widetilde{\ell}_0$ and for the hypothesis sets $\mathcal{H}$ which are regular for adversarial calibration.

For convenience, let $\underline{M}(f, \mathbf{x}, \epsilon) := \inf_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} f(\tilde{\mathbf{x}})$ and $\overline{M}(f, \mathbf{x}, \epsilon) := -\inf_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} -f(\tilde{\mathbf{x}}) = \sup_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} f(\tilde{\mathbf{x}})$. Then we provide three useful Lemmas as follows,

Lemma C.6 *(Awasthi et al. (2021), Lemma 28). Let $\mathcal{H}$ be a symmetric hypothesis set, ℓ be a surrogate loss function, and $\mathcal{X}_2 =\{\mathbf{x} \in \mathcal{X}$: there exists $f' \in \mathcal{H}$ such that $\underline{M}(f', \mathbf{x}, \epsilon) > 0$ $\}$. If $\mathcal{X}_2 = \emptyset$, any loss ℓ is $\mathcal{H}$-calibrated with respect to $\widetilde{\ell}_0$. If $\mathcal{X}_2 \neq \emptyset$, then ℓ is $\mathcal{H}$-calibrated with respect to $\widetilde{\ell}_0$ if and only if for any $\mathbf{x} \in \mathcal{X}_2$,*

$$\begin{aligned} &\inf_{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\epsilon)\leq 0\leq\overline{M}(f,\mathbf{x},\epsilon)} \mathcal{C}_\ell(f, \mathbf{x}, \tfrac{1}{2}) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f, \mathbf{x}, \tfrac{1}{2}), \textit{and} \\ &\inf_{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\epsilon)\leq 0} \mathcal{C}_\ell(f, \mathbf{x}, \eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f, \mathbf{x}, \eta), \quad \forall\eta \in (\tfrac{1}{2}, 1], \textit{and} \\ &\inf_{f\in\mathcal{H}:0\leq\overline{M}(f,\mathbf{x},\epsilon)} \mathcal{C}_\ell(f, \mathbf{x}, \eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_\ell(f, \mathbf{x}, \eta), \quad \forall\eta \in [0, \tfrac{1}{2}). \end{aligned} \tag{C.9}$$

Lemma C.7 *(Awasthi et al. (2021), Theorem 23 and Theorem 24). Let $\mathcal{H}$ be a symmetric hypothesis set consisting of the family of all measurable functions $\mathcal{H}_{all}$, ϕ be a non-increasing margin-based loss, and $\widetilde{\phi}(f, \mathbf{x}, y) = \sup_{\tilde{\mathbf{x}} \in \mathcal{B}_\epsilon(\mathbf{x})} \phi(yf(\tilde{\mathbf{x}}))$. If $\widetilde{\phi}$ is $\mathcal{H}$-calibrated with respect to $\widetilde{\ell}_0$, then $\widetilde{\phi}$ is $\mathcal{H}$-consistent with respect to $\widetilde{\ell}_0$ for all distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that satisfy: $\mathcal{R}^*_{\widetilde{\ell}_0, \mathcal{H}} = 0$ and there exists $f^* \in \mathcal{H}$ such that $\mathcal{R}_\phi(f^*) = \mathcal{R}^*_{\phi, \mathcal{H}_{all}} < +\infty$.*

The proofs of the above two Lemmas can be found in Awasthi et al. (2021).

Lemma C.8 *Let $\mathcal{H}$ be a symmetric hypothesis set and $f \in \mathcal{H}$. Suppose $0 \le \lambda^* < \hat{\lambda}^*$, $\nu > \min\{\hat{\lambda}^*, \mathcal{R}^*_{\ell,\mathcal{H}}\}$, $\ell(yf(\mathbf{x})) \ge 0\ \forall \mathbf{x}$, and λ^* is bounded, then $\lambda^* < \ell(0)$.*

Proof Based on the definition of $(f_0^*, \lambda^*, \hat{\lambda}^*) = \arg\inf_{f,\lambda} \sup_{\hat{\lambda}} \Big\{ \mathbb{E}\Big[\hat{\lambda} - [\hat{\lambda} - [\ell(Yf(X)) - \lambda]_+]_+\Big] + \nu\lambda - \mu\hat{\lambda} \Big\}$. We choose $f = 0$, $\lambda = \ell(0)$ and $\hat{\lambda} = \hat{\lambda}^*$ there holds

$$\begin{aligned} \nu\lambda^* - \mu\hat{\lambda}^* &\le \mathbb{E}\Big[\hat{\lambda}^* - [\hat{\lambda}^* - [\ell(Yf_0^*(X)) - \lambda^*]_+]_+\Big] + \nu\lambda^* - \mu\hat{\lambda}^* \\ &\le \mathbb{E}\Big[\hat{\lambda}^* - [\hat{\lambda}^* - [\ell(0) - \ell(0)]_+]_+\Big] + \nu\ell(0) - \mu\hat{\lambda}^* \\ &= \nu\ell(0) - \mu\hat{\lambda}^* \end{aligned}$$

Thus $\nu\lambda^* \le \nu\ell(0)$ which shows that $\lambda^* \le \ell(0)$. Let $\beta = \ell(0) - \lambda$ which implies

$$(f_0^*, \beta^*, \hat{\lambda}^*) = \arg\inf_{f,\lambda} \sup_{\hat{\lambda}} \left\{ \mathbb{E}\Big[\hat{\lambda} - [\hat{\lambda} - [\ell(Yf(X)) + \beta - \ell(0)]_+]_+\Big] - \nu\beta - \mu\hat{\lambda} \right\}.$$

Let $(f_0^*, \beta^*, \hat{\lambda}^*)$ be the minimizer. we have, for any f and choosing $\beta = \ell(0)$, that

$$\begin{aligned} -\nu\beta^* - \mu\hat{\lambda}^* &\le \mathbb{E}\Big[\hat{\lambda}^* - [\hat{\lambda}^* - [\ell(Yf(X)) + \beta^* - \ell(0)]_+]_+\Big] - \nu\beta^* - \mu\hat{\lambda}^* \\ &\le \mathbb{E}\Big[\hat{\lambda}^* - [\hat{\lambda}^* - [\ell(Yf(X)) + \ell(0) - \ell(0)]_+]_+\Big] - \nu\ell(0) - \mu\hat{\lambda}^*. \end{aligned}$$

Therefore, we have

$$-\nu\beta^* \le \mathbb{E}\Big[\hat{\lambda}^* - [\hat{\lambda}^* - [\ell(Yf(X))]_+]_+\Big] - \nu\ell(0).$$

Since f is arbitrary, $\beta^* \ge \frac{\nu - \mathbb{E}\Big[\hat{\lambda}^* - [\hat{\lambda}^* - [\ell(Yf^*(X))]_+]_+\Big]}{\nu}$. Since $\ell(Yf^*(X)) \ge 0$, we have

$$0 \le \mathbb{E}\Big[\hat{\lambda}^* - [\hat{\lambda}^* - [\ell(Yf^*(X))]_+]_+\Big] = \mathbb{E}\Big[\hat{\lambda}^* - [\hat{\lambda}^* - \ell(Yf^*(X))]_+\Big] \le \min\left\{\hat{\lambda}^*, \inf_f \mathbb{E}[\ell(yf(\mathbf{x}))]\right\}.$$

By using the assumption $\nu > \min\{\hat{\lambda}^*, \mathcal{R}^*_{\ell,\mathcal{H}}\}$=$\min\left\{\hat{\lambda}^*, \inf_f \mathbb{E}[\ell(yf(\mathbf{x}))]\right\}$, we get $\beta^* \ge \frac{\nu - \mathbb{E}\Big[\hat{\lambda}^* - [\hat{\lambda}^* - [\ell(Yf^*(X))]_+]_+\Big]}{\nu} > 0$. Consequently, the above arguments show that $0 \le \lambda^* = \ell(0) - \beta^* < \ell(0)$ if $\nu > \min\left\{\hat{\lambda}^*, \inf_f \mathbb{E}[\ell(yf(\mathbf{x}))]\right\}$. ■

Theorem C.9 *(Theorem 4 restated) Let $\mathcal{H}$ be a symmetric hypothesis set consisting of the family of all measurable functions $\mathcal{H}_{all}$, suppose $\nu > \min\{\hat{\lambda}^*, \mathcal{R}^*_{\ell,\mathcal{H}}\}$, $0 \leq \lambda^* < \hat{\lambda}^*$, λ^* and $\hat{\lambda}^*$ are bounded, and ℓ is a non-negative, continuous, and non-increasing margin-based loss.*

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

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

Proof Below we will prove the theorem using Lemma C.6 which is from Awasthi et al. (2021). Recall that, from the definition of $\phi_{\texttt{ORAT}}(t)$ in Eq.(5), $\ell(t)$ is a continuous and non-increasing function, and λ^* and $\hat{\lambda}^*$ are bounded, we can conclude $\phi_{\texttt{ORAT}}(t)$ is bounded, continuous, non-increasing.

By Lemma C.6, if $\mathcal{X}_2 = \emptyset$, $\tilde{\phi}_{\texttt{ORAT}}$ is $\mathcal{H}$-calibrated with respect to $\widetilde{\ell}_0$. Consequently, it suffices to consider the case where $\mathcal{X}_2 \neq \emptyset$. In this case, in order to show $\tilde{\phi}_{\texttt{ORAT}}$ is $\mathcal{H}$-calibrated with respect to $\widetilde{\ell}_0$, from Lemma C.6 we only need to show, $\forall \mathbf{x} \in \mathcal{X}_2$, that

$$\inf_{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\epsilon)\leq 0\leq \overline{M}(f,\mathbf{x},\epsilon)} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\frac{1}{2}) > \inf_{f\in\mathcal{H}} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\frac{1}{2}), \textit{and}$$

$$\inf_{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\epsilon)\leq 0} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta), \quad \forall \eta \in (\frac{1}{2},1], \textit{and}$$

$$\inf_{f\in\mathcal{H}:0\leq \overline{M}(f,\mathbf{x},\epsilon)} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta) > \inf_{f\in\mathcal{H}} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta), \quad \forall \eta \in [0,\frac{1}{2}).$$

To this end, recall that, by the definition of inner ℓ_s-risk, the inner $\tilde{\phi}_{\texttt{ORAT}}$-risk is given by

$$\begin{aligned}\mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,x,\eta) &= \eta\tilde{\phi}_{\texttt{ORAT}}(f,\mathbf{x},+1) + (1-\eta)\tilde{\phi}_{\texttt{ORAT}}(f,\mathbf{x},-1)\\ &= \eta\phi_{\texttt{ORAT}}\Big(\inf_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} f(\tilde{\mathbf{x}})\Big) + (1-\eta)\phi_{\texttt{ORAT}}\Big(\inf_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} -f(\tilde{\mathbf{x}})\Big)\\ &= \eta\phi_{\texttt{ORAT}}\Big(\underline{M}(f,\mathbf{x},\epsilon)\Big) + (1-\eta)\phi_{\texttt{ORAT}}\Big(-\overline{M}(f,\mathbf{x},\epsilon)\Big).\end{aligned}$$

For any $\mathbf{x} \in \mathcal{X}_2$, let $M_{\mathbf{x}} = \sup_{f\in\mathcal{H}} \underline{M}(f,\mathbf{x},\epsilon) > 0$. Since $\mathcal{H}$ is symmetric consisting of all measurable functions, we have $-M_{\mathbf{x}} = \inf_{f\in\mathcal{H}} \overline{M}(f,\mathbf{x},\epsilon) < 0$. Since $\phi_{\texttt{ORAT}}(\cdot)$ is continuous, for any $\mathbf{x} \in \mathcal{X}_2$ and $\tau > 0$, there exists $f^\tau_{\mathbf{x}} \in \mathcal{H}$ such that $\phi_{\texttt{ORAT}}(\underline{M}(f^\tau_{\mathbf{x}},\mathbf{x},\epsilon)) < \phi_{\texttt{ORAT}}(M_{\mathbf{x}}) + \tau$, $\phi_{\texttt{ORAT}}(-\overline{M}(f^\tau_{\mathbf{x}},\mathbf{x},\epsilon)) < \phi_{\texttt{ORAT}}(0) + \tau$, $\overline{M}(f^\tau_{\mathbf{x}},\mathbf{x},\epsilon) \geq \underline{M}(f^\tau_{\mathbf{x}},\mathbf{x},\epsilon) > 0$, $\underline{M}(-f^\tau_{\mathbf{x}},\mathbf{x},\epsilon) \leq \overline{M}(-f^\tau_{\mathbf{x}},\mathbf{x},\epsilon) = -\underline{M}(f^\tau_{\mathbf{x}},\mathbf{x},\epsilon) < 0$, and $\phi_{\texttt{ORAT}}(\underline{M}(-f^\tau_{\mathbf{x}},\mathbf{x},\epsilon)) < \phi_{\texttt{ORAT}}(0) + \tau$. Next we analyze three cases:

1. When $\eta = \frac{1}{2}$, since $\phi_{\texttt{ORAT}}$ is non-increasing,

$$\begin{aligned}&\inf_{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\epsilon)\leq 0\leq \overline{M}(f,\mathbf{x},\epsilon)} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\frac{1}{2})\\ &= \inf_{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\epsilon)\leq 0\leq \overline{M}(f,\mathbf{x},\epsilon)} \frac{1}{2}\phi_{\texttt{ORAT}}\Big(\underline{M}(f,\mathbf{x},\epsilon)\Big) + \frac{1}{2}\phi_{\texttt{ORAT}}\Big(-\overline{M}(f,\mathbf{x},\epsilon)\Big)\\ &\geq \frac{1}{2}\phi_{\texttt{ORAT}}(0) + \frac{1}{2}\phi_{\texttt{ORAT}}(0)\\ &= \phi_{\texttt{ORAT}}(0)\\ &= \hat{\lambda}^* - [\hat{\lambda}^* - [\ell(0) - \lambda^*]_+]_+.\end{aligned}$$

For any $\mathbf{x} \in \mathcal{X}_2$, there exists $f' \in \mathcal{H}$ such that $\underline{M}(f', \mathbf{x}, \epsilon) > 0$ and $-\overline{M}(f', \mathbf{x}, \epsilon) \leq -\underline{M}(f', \mathbf{x}, \epsilon) < 0$, we obtain

$$\mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f', \mathbf{x}, \frac{1}{2}) = \frac{1}{2}\phi_{\texttt{ORAT}}\Big(\underline{M}(f', \mathbf{x}, \epsilon)\Big) + \frac{1}{2}\phi_{\texttt{ORAT}}\Big(-\overline{M}(f', \mathbf{x}, \epsilon)\Big)$$

According to Lemma C.8, we have $0 \leq \lambda^* < \hat{\lambda}^*$ and $\lambda^* < \ell(0)$. Therefore, we also analyze two cases:

(a) If $0 < \lambda^* + \hat{\lambda}^* \leq \ell(0)$, then we have

$$\inf_{f \in \mathcal{H}: \underline{M}(f, \mathbf{x}, \epsilon) \leq 0 \leq \overline{M}(f, \mathbf{x}, \epsilon)} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f, \mathbf{x}, \frac{1}{2}) \geq \hat{\lambda}^* - [\hat{\lambda}^* - [\ell(0) - \lambda^*]_+]_+ = \hat{\lambda}^*.$$

On the other hand, since $\phi_{\texttt{ORAT}}$ is continuous, there exists $f' \in \mathcal{H}$ and $t = \underline{M}(f', \mathbf{x}, \epsilon)$, then $0 \leq \phi_{\texttt{ORAT}}\Big(\underline{M}(f', \mathbf{x}, \epsilon)\Big) < \hat{\lambda}^*$. Thus,

$$\begin{aligned}\mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f', \mathbf{x}, \frac{1}{2}) &= \frac{1}{2}\phi_{\texttt{ORAT}}\Big(\underline{M}(f', \mathbf{x}, \epsilon)\Big) + \frac{1}{2}\phi_{\texttt{ORAT}}\Big(-\overline{M}(f', \mathbf{x}, \epsilon)\Big) \\ &\leq \frac{1}{2}\phi_{\texttt{ORAT}}\Big(\underline{M}(f', \mathbf{x}, \epsilon)\Big) + \frac{1}{2}\hat{\lambda}^* \\ &< \frac{1}{2}\hat{\lambda}^* + \frac{1}{2}\hat{\lambda}^* = \hat{\lambda}^*.\end{aligned}$$

Therefore, for any $\mathbf{x} \in \mathcal{X}_2$,

$$\inf_{f \in \mathcal{H}} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f, \mathbf{x}, \frac{1}{2}) \leq \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f', \mathbf{x}, \frac{1}{2}) < \hat{\lambda}^* \leq \inf_{f \in \mathcal{H}: \underline{M}(f, \mathbf{x}, \epsilon) \leq 0 \leq \overline{M}(f, \mathbf{x}, \epsilon)} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f, \mathbf{x}, \frac{1}{2}). \tag{C.10}$$

(b) If $\lambda^* + \hat{\lambda}^* > \ell(0)$,

$$\inf_{f \in \mathcal{H}: \underline{M}(f, \mathbf{x}, \epsilon) \leq 0 \leq \overline{M}(f, \mathbf{x}, \epsilon)} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f, \mathbf{x}, \frac{1}{2}) \geq \hat{\lambda}^* - [\hat{\lambda}^* - [\ell(0) - \lambda^*]_+]_+ = \ell(0) - \lambda^*.$$

On the other hand, recall both $\phi_{\texttt{ORAT}}(\cdot)$ and $\ell(\cdot)$ are continuous and non-increasing and $\ell(0) > \lambda^*$ from Lemma 5. Therefore, we can find $f' \in \mathcal{H}$ such that $\ell(0) > \ell(\underline{M}(f', \mathbf{x}, \epsilon)) > \lambda^*$, $\lambda^* + \hat{\lambda}^* > \ell(-\overline{M}(f', \mathbf{x}, \epsilon)) > \ell(0) > \lambda^*$, and $\ell(\underline{M}(f', \mathbf{x}, \epsilon)) + \ell(-\overline{M}(f', \mathbf{x}, \epsilon)) < 2\ell(0)$. Consequently, there holds

$$\begin{aligned}&\mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f', \mathbf{x}, \frac{1}{2}) \\ &= \frac{1}{2}\phi_{\texttt{ORAT}}\Big(\underline{M}(f', \mathbf{x}, \epsilon)\Big) + \frac{1}{2}\phi_{\texttt{ORAT}}\Big(-\overline{M}(f', \mathbf{x}, \epsilon)\Big) \\ &= \frac{1}{2}\Big[\hat{\lambda}^* - [\hat{\lambda}^* - [\ell(\underline{M}(f', \mathbf{x}, \epsilon)) - \lambda^*]_+]_+\Big] + \frac{1}{2}\Big[\hat{\lambda}^* - [\hat{\lambda}^* - [\ell(-\overline{M}(f', \mathbf{x}, \epsilon)) - \lambda^*]_+]_+\Big] \\ &= \frac{1}{2}[\ell(\underline{M}(f', \mathbf{x}, \epsilon)) - \lambda^*] + \frac{1}{2}[\ell(-\overline{M}(f', \mathbf{x}, \epsilon)) - \lambda^*] \\ &= \frac{1}{2}[\ell(\underline{M}(f', \mathbf{x}, \epsilon)) + \ell(-\overline{M}(f', \mathbf{x}, \epsilon))] - \lambda^* \\ &< \frac{1}{2} \times 2\ell(0) - \lambda^* = \ell(0) - \lambda^*.\end{aligned}$$

Therefore, for any $\mathbf{x} \in \mathcal{X}_2$,

$$\inf_{f\in\mathcal{H}} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\frac{1}{2}) \leq \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f',\mathbf{x},\frac{1}{2}) < \ell(0) - \lambda^* \leq \inf_{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\epsilon)\leq 0\leq \overline{M}(f,\mathbf{x},\epsilon)} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\frac{1}{2}). \tag{C.11}$$

2. When $\eta \in (\frac{1}{2}, 1]$, since $\phi_{\texttt{ORAT}}$ is non-increasing, for any $\mathbf{x} \in \mathcal{X}_2$,

$$\begin{aligned}\inf_{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\epsilon)\leq 0} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta) &= \inf_{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\epsilon)\leq 0} \eta\phi_{\texttt{ORAT}}\Big(\underline{M}(f,\mathbf{x},\epsilon)\Big) + (1-\eta)\phi_{\texttt{ORAT}}\Big(-\overline{M}(f,\mathbf{x},\epsilon)\Big)\\ &\geq \eta\phi_{\texttt{ORAT}}(0) + (1-\eta)\phi_{\texttt{ORAT}}(M_{\mathbf{x}}).\end{aligned}$$

On the other hand, for any $\mathbf{x} \in \mathcal{X}_2$ and $\tau > 0$,

$$\begin{aligned}\mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f_{\mathbf{x}}^{\tau},\mathbf{x},\eta) &= \eta\phi_{\texttt{ORAT}}\Big(\underline{M}(f_{\mathbf{x}}^{\tau},\mathbf{x},\epsilon)\Big) + (1-\eta)\phi_{\texttt{ORAT}}\Big(-\overline{M}(f_{\mathbf{x}}^{\tau},\mathbf{x},\epsilon)\Big)\\ &< \eta[\phi_{\texttt{ORAT}}(M_{\mathbf{x}}) + \tau] + (1-\eta)[\phi_{\texttt{ORAT}}(0) + \tau]\\ &= \eta\phi_{\texttt{ORAT}}(M_{\mathbf{x}}) + (1-\eta)\phi_{\texttt{ORAT}}(0) + \tau.\end{aligned}$$

Since $\eta > \frac{1}{2}$ and $M_{\mathbf{x}} > 0$, we have

$$\begin{aligned}&\inf_{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\epsilon)\leq 0} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta) - \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f_{\mathbf{x}}^{\tau},\mathbf{x},\eta)\\ &> [\eta\phi_{\texttt{ORAT}}(0) + (1-\eta)\phi_{\texttt{ORAT}}(M_{\mathbf{x}})] - [\eta\phi_{\texttt{ORAT}}(M_{\mathbf{x}}) + (1-\eta)\phi_{\texttt{ORAT}}(0) + \tau]\\ &= [2\eta - 1][\phi_{\texttt{ORAT}}(0) - \phi_{\texttt{ORAT}}(M_{\mathbf{x}})] - \tau\\ &> 0,\end{aligned}$$

where we take $0 < \tau < [2\eta - 1][\phi_{\texttt{ORAT}}(0) - \phi_{\texttt{ORAT}}(M_{\mathbf{x}})]$. Therefore, for any $\eta \in (\frac{1}{2}, 1]$ and $\mathbf{x} \in \mathcal{X}_2$, there exists $0 < \tau < [2\eta - 1][\phi_{\texttt{ORAT}}(0) - \phi_{\texttt{ORAT}}(M_{\mathbf{x}})]$ such that

$$\inf_{f\in\mathcal{H}} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta) \leq \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f_{\mathbf{x}}^{\tau},\mathbf{x},\eta) < \inf_{f\in\mathcal{H}:\underline{M}(f,\mathbf{x},\epsilon)\leq 0} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta). \tag{C.12}$$

3. When $\eta \in [0, \frac{1}{2})$, since $\phi_{\texttt{ORAT}}$ is non-increasing, for any $\mathbf{x} \in \mathcal{X}_2$,

$$\begin{aligned}\inf_{f\in\mathcal{H}:\overline{M}(f,\mathbf{x},\epsilon)\geq 0} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta) &= \inf_{f\in\mathcal{H}:\overline{M}(f,\mathbf{x},\epsilon)\geq 0} \eta\phi_{\texttt{ORAT}}\Big(\underline{M}(f,\mathbf{x},\epsilon)\Big) + (1-\eta)\phi_{\texttt{ORAT}}\Big(-\overline{M}(f,\mathbf{x},\epsilon)\Big)\\ &\geq (1-\eta)\phi_{\texttt{ORAT}}(0) + \inf_{f\in\mathcal{H}:\overline{M}(f,\mathbf{x},\epsilon)\geq 0} \eta\phi_{\texttt{ORAT}}\Big(\underline{M}(f,\mathbf{x},\epsilon)\Big)\\ &\geq (1-\eta)\phi_{\texttt{ORAT}}(0) + \eta\phi_{\texttt{ORAT}}\Big(M_{\mathbf{x}}\Big).\end{aligned}$$

On the other hand, for any $\mathbf{x} \in \mathcal{X}_2$ and $\tau > 0$,

$$\begin{aligned}\mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(-f_{\mathbf{x}}^{\tau},\mathbf{x},\eta) &= \eta\phi_{\texttt{ORAT}}\Big(\underline{M}(-f_{\mathbf{x}}^{\tau},\mathbf{x},\epsilon)\Big) + (1-\eta)\phi_{\texttt{ORAT}}\Big(-\overline{M}(-f_{\mathbf{x}}^{\tau},\mathbf{x},\epsilon)\Big)\\ &= \eta\phi_{\texttt{ORAT}}\Big(\underline{M}(-f_{\mathbf{x}}^{\tau},\mathbf{x},\epsilon)\Big) + (1-\eta)\phi_{\texttt{ORAT}}\Big(\underline{M}(f_{\mathbf{x}}^{\tau},\mathbf{x},\epsilon)\Big)\\ &< \eta[\phi_{\texttt{ORAT}}(0) + \tau] + (1-\eta)\phi_{\texttt{ORAT}}\Big(\underline{M}(f_{\mathbf{x}}^{\tau},\mathbf{x},\epsilon)\Big)\\ &< \eta[\phi_{\texttt{ORAT}}(0) + \tau] + (1-\eta)[\phi_{\texttt{ORAT}}(M_{\mathbf{x}}) + \tau]\\ &= \eta\phi_{\texttt{ORAT}}(0) + (1-\eta)\phi_{\texttt{ORAT}}(M_{\mathbf{x}}) + \tau.\end{aligned}$$

Since $\eta < \frac{1}{2}$ and $M_{\mathbf{x}} > 0$, we have

$$\begin{aligned}
&\inf_{f\in\mathcal{H}:\overline{M}(f,\mathbf{x},\epsilon)\geq 0} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta) - \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(-f^{\tau}_{\mathbf{x}},\mathbf{x},\eta) \\
&> [(1-\eta)\phi_{\texttt{ORAT}}(0) + \eta\phi_{\texttt{ORAT}}(M_{\mathbf{x}})] - [\eta\phi_{\texttt{ORAT}}(0) + (1-\eta)\phi_{\texttt{ORAT}}(M_{\mathbf{x}}) + \tau] \\
&= (1-2\eta)[\phi_{\texttt{ORAT}}(0) - \phi_{\texttt{ORAT}}(M_{\mathbf{x}})] - \tau,
\end{aligned}$$

where we take $0 < \tau < (1-2\eta)[\phi_{\texttt{ORAT}}(0) - \phi_{\texttt{ORAT}}(M_{\mathbf{x}})]$. Therefore for any $\eta \in [0, \frac{1}{2})$ and $\mathbf{x} \in \mathcal{X}_2$, there exists $0 < \tau < (1-2\eta)[\phi_{\texttt{ORAT}}(0) - \phi_{\texttt{ORAT}}(M_{\mathbf{x}})]$ such that

$$\inf_{f\in\mathcal{H}} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta) \leq \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(-f^{\tau}_{\mathbf{x}},\mathbf{x},\eta) < \inf_{f\in\mathcal{H}:\overline{M}(f,\mathbf{x},\epsilon)\geq 0} \mathcal{C}_{\tilde{\phi}_{\texttt{ORAT}}}(f,\mathbf{x},\eta) \tag{C.13}$$

From (C.10), (C.11), (C.12), (C.13), we conclude that $\tilde{\phi}_{\texttt{ORAT}}$ is $\mathcal{H}$-calibrated with respect to $\widetilde{\ell}_0$. Thus, (i) holds.

According to Lemma C.7, we can conclude that the $\tilde{\phi}_{\texttt{ORAT}}$ is $\mathcal{H}$-consistent with respect to $\widetilde{\ell}_0$ for all distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that satisfy: $\mathcal{R}^*_{\widetilde{\ell}_0,\mathcal{H}} = 0$ and there exists $f^* \in \mathcal{H}$ such that $\mathcal{R}_{\phi_{\texttt{ORAT}}}(f^*) = \mathcal{R}^*_{\phi_{\texttt{ORAT}},\mathcal{H}_{all}} < +\infty$. Therefore, (ii) holds. ■

C.3. Cross-entropy as A Margin-based Loss

The cross-entropy loss can be rewritten as a margin-based loss. For example, in binary classification, the conventional binary cross-entropy (bce) loss is given by $bce = -(ylog(\sigma(f(\mathbf{x}))) + (1-y)log(1-\sigma(f(\mathbf{x}))))$ when $y = \{0, 1\}$. Here σ is the sigmoid function. It is clear that this conventional bce loss is not a margin-based loss. However, we can transfer the negative label 0 to -1. In this case, by the property of the sigmoid function $1 - \sigma(\mathbf{x}) = \sigma(-\mathbf{x})$, the original bce loss can be rewritten as $bce = -log(\sigma(yf(\mathbf{x})))$ when $y = \{-1, 1\}$. This is in fact a non-negative, continuous, and non-increasing margin-based loss.

C.4. Proof of Theorem 6

To get the generalization error bound, we need an equivalent formulation of (4) which is stated in the following lemma.

Lemma C.10 *Suppose $\lambda \in \mathbb{R}, \hat{\lambda} \in \mathbb{R}$, then the empirical risk $\mathcal{R}_{\widetilde{\ell}}(f;\mathcal{S})$ defined by Eq. (4) is equivalent to*

$$\mathcal{R}_{\widetilde{\ell}}(f;\mathcal{S}) = \frac{1}{k-m}\left(\min_{\lambda\in\mathbb{R}}\left\{k\lambda + \sum_{i=1}^{n}[\ell(f(\tilde{\mathbf{x}}_i),y_i) - \lambda]_+\right\} - \min_{\hat{\lambda}\in\mathbb{R}}\left\{m\hat{\lambda} + \sum_{i=1}^{n}[\ell(f(\tilde{\mathbf{x}}_i),y_i) - \hat{\lambda}]_+\right\}\right) \tag{C.14}$$

Proof According to Eq.(C.3), we have

$$\begin{aligned}
&\min_{\lambda}(k-m)\lambda+\max_{\hat{\lambda}}\Big\{(n-m)\hat{\lambda}-\sum_{i=1}^{n}[\hat{\lambda}-[\ell(f(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+]_+\Big\}\\
=&\min_{\lambda}(k-m)\lambda+\sum_{i=m+1}^{n}[[\ell(f(\tilde{\mathbf{x}}),y)-\lambda]_+]_{[i]}\\
=&\min_{\lambda,q}(k-m)\lambda+\sum_{i=1}^{n}q_i[\ell(f(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+ \quad \text{s.t. } q_i\in[0,1],\ ||q||_0=n-m.
\end{aligned}$$

Under the constraints, we can rewrite the last formula as

$$\begin{aligned}
&(k-m)\lambda+\sum_{i=1}^{n}q_i[\ell(f(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+\\
=&(k-m)\lambda+\sum_{i=1}^{n}[\ell(f(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+-\sum_{i=1}^{n}(1-q_i)[\ell(f(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+\\
=&k\lambda+\sum_{i=1}^{n}[\ell(f(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+-\sum_{i=1}^{n}(1-q_i)\{[\ell(f(\tilde{\mathbf{x}}_i),y_i)-\lambda]_++\lambda\}.
\end{aligned}$$

The last equality holds because $\sum_{i=1}^{n}(1-q_i)=n-(n-m)=m$.

For the term $\sum_{i=1}^{n}(1-q_i)\{[\ell(f(\tilde{\mathbf{x}}_i),y_i)-\lambda]_++\lambda\}$, we assume $\ell(f^*(\tilde{\mathbf{x}}_i),y_i)$, $\forall i$, are sorted in descending order when getting the optimal model f^*. For example, $\ell(f^*(\tilde{\mathbf{x}}_1),y_1)\geq\ell(f^*(\tilde{\mathbf{x}}_2),y_2)\geq\cdots\geq\ell(f^*(\tilde{\mathbf{x}}_n),y_n)$. Since $\lambda^*\geq 0$, the optimal q^* should be $q_1^*=\cdots=q_m^*=0$, $q_{m+1}^*=\cdots=q_n^*=1$. Note that λ^* must be an optimal solution of the problem

$$\min_{\lambda}(k-m)\lambda+\sum_{i=m+1}^{n}q_i^*[\ell(f^*(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+.$$

From Lemma C.1, we know $\ell(f^*(\tilde{\mathbf{x}}_{m+1}),y_{m+1})\geq\lambda^*$, which implies that $\ell(f^*(\tilde{\mathbf{x}}_i),y_i)-\lambda^*\geq 0$ holds for $q_i<1$. Therefore, $\sum_{i=1}^{n}(1-q_i)\{[\ell(f(\tilde{\mathbf{x}}_i),y_i)-\lambda]_++\lambda\}=\sum_{i=1}^{n}(1-q_i)\ell(f(\tilde{\mathbf{x}}_i),y_i)$. Furthermore, we know

$$\min_{\hat{\lambda}}\Big\{m\hat{\lambda}+\sum_{i=1}^{n}[\ell(f(\tilde{\mathbf{x}}_i),y_i)-\hat{\lambda}]_+\Big\}=\max_{q}\Big\{\sum_{i=1}^{n}(1-q_i)\ell(f(\tilde{\mathbf{x}}_i),y_i)\Big|\ q_i\in[0,1],\ ||q||_0=n-m\Big\}.$$

Then we get

$$\begin{aligned}
&\frac{1}{k-m}\left(\min_{\lambda}(k-m)\lambda+\max_{\hat{\lambda}}\Big\{(n-m)\hat{\lambda}-\sum_{i=1}^{n}[\hat{\lambda}-[\ell(f(\tilde{\mathbf{x}}_i),y_i)-\lambda]_+]_+\Big\}\right)\\
=&\frac{1}{k-m}\left(\min_{\lambda}\Big\{k\lambda+\sum_{i=1}^{n}[\ell(f(\tilde{\mathbf{x}}_{[i]}),y_{[i]})-\lambda]_+\Big\}-\min_{\hat{\lambda}}\Big\{m\hat{\lambda}+\sum_{i=1}^{n}[\ell(f(\tilde{\mathbf{x}}_{[i]}),y_{[i]})-\hat{\lambda}]_+\Big\}\right).
\end{aligned}$$

The proof is complete. ■

Considering the limit case of (C.14), the population risk $\mathcal{R}_{\widetilde{\ell}}(f)$ can be written as

$$\frac{1}{k-m}\left(\min_{\lambda\in\mathbb{R}}\left\{k\lambda+\sum_{i=1}^{n}[\ell(f(\tilde{\mathbf{x}}_{[i]}),y_{[i]})-\lambda]_+\right\}-\min_{\hat{\lambda}\in\mathbb{R}}\left\{m\hat{\lambda}+\sum_{i=1}^{n}[\ell(f(\tilde{\mathbf{x}}_{[i]}),y_{[i]})-\hat{\lambda}]_+\right\}\right)$$
$$\xrightarrow[n\to\infty]{\frac{k-m}{n}\to\nu,\frac{m}{n}\to\mu}\frac{1}{\nu}\left(\min_{\lambda\in\mathbb{R}}\left\{(\nu+\mu)\lambda+\mathbb{E}[\widetilde{\ell}(f(\mathbf{x}),y)-\lambda]_+\right\}-\min_{\hat{\lambda}\in\mathbb{R}}\left\{\mu\hat{\lambda}+\mathbb{E}[\widetilde{\ell}(f(\mathbf{x}),y)-\hat{\lambda}]_+\right\}\right)=\mathcal{R}_{\widetilde{\ell}}(f).$$

The next Lemma tells us if the loss function is bounded, we can constrain the problem of $\mathcal{R}_{\widetilde{\ell}}(f)$ and $\mathcal{R}_{\widetilde{\ell}}(f;\mathcal{S})$ in the bounded range as well.

Lemma C.11 *Suppose that the range of ℓ is $[0,M]$. Then we have*

$$\mathcal{R}_{\widetilde{\ell}}(f)=\frac{1}{\nu}\left(\min_{\lambda\in[0,M]}\left\{(\nu+\mu)\lambda+\mathbb{E}[\widetilde{\ell}(f(\mathbf{x}),y)-\lambda]_+\right\}-\min_{\hat{\lambda}\in[0,M]}\left\{\mu\hat{\lambda}+\mathbb{E}[\widetilde{\ell}(f(\mathbf{x}),y)-\hat{\lambda}]_+\right\}\right),\tag{C.15}$$

and so does the empirical risk

$$\mathcal{R}_{\widetilde{\ell}}(f;\mathcal{S})=\frac{1}{k-m}\left(\min_{\lambda\in[0,M]}\left\{k\lambda+\sum_{i=1}^{n}[\widetilde{\ell}(f(\mathbf{x}_i),y)-\lambda]_+\right\}-\min_{\hat{\lambda}\in[0,M]}\left\{m\hat{\lambda}+\sum_{i=1}^{n}[\widetilde{\ell}(f(\mathbf{x}_i),y)-\hat{\lambda}]_+\right\}\right).\tag{C.16}$$

Proof The proof of (C.16) is straight forward. By Lemma C.1 and Lemma C.2, we know $\lambda^*_{\mathcal{S}}=\widetilde{\ell}(f(\mathbf{x}_{[k]}),y_{[k]})$ and $\hat{\lambda}^*_{\mathcal{S}}=\widetilde{\ell}(f(\mathbf{x}_{[m]}),y_{[m]})$ are a pair of solution of (C.16). Since $\widetilde{\ell}(f(\mathbf{x}),y)=\max_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})}\ell(f(\tilde{\mathbf{x}}),y)\in[0,M]$ for any $\mathbf{x},y$, we have $\lambda^*_{\mathcal{S}},\hat{\lambda}^*_{\mathcal{S}}\in[0,M]$.

Next we move on to (C.15). Let λ^* and $\hat{\lambda}^*$ be a pair of solution of (C.15). Let $\lambda=M$, then we have

$$\begin{aligned}(\nu+\mu)\lambda^*&\le(\nu+\mu)\lambda^*+\mathbb{E}[\widetilde{\ell}(f(\mathbf{x}),y)-\lambda^*]_+\\&\le(\nu+\mu)M+\mathbb{E}[\widetilde{\ell}(f(\mathbf{x}),y)-M]_+\\&\le(\nu+\mu)M+\mathbb{E}[M-M]_+=(\nu+\mu)M,\end{aligned}$$

which implies $\lambda^*\le M$. On the other hand, assume $\lambda^*=-\varepsilon$ for some $\varepsilon>0$. Let $\lambda=0$, then we have

$$-(\nu+\mu)\varepsilon+\mathbb{E}[\widetilde{\ell}(f(\mathbf{x}),y)+\varepsilon]_+=(1-(\nu+\mu))\varepsilon+\mathbb{E}[\widetilde{\ell}(f(\mathbf{x}),y)]_+\le\mathbb{E}[\widetilde{\ell}(f(\mathbf{x}),y)]_+,$$

which contradicts with $(1-(\nu+\mu))\varepsilon>0$. Therefore we have $\lambda^*\ge 0$. Similarly, we can show that $\hat{\lambda}^*\in[0,M]$. The proof is complete. ■

The next lemma shows the uniform convergence of learning with `ORAT` without using perturbation.

Lemma C.12 *Suppose that the range of $\ell(f(\mathbf{x}),y)$ is $[0,M]$. Then, for any $\delta\in(0,1)$, with probability at least $1-\delta$ over the draw of an i.i.d. training dataset of size n, the following holds for all $\ell_f\in\ell_{\mathcal{H}}$,*

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

Proof By the subadditivity of max operator, for any $\ell_f \in \ell_{\mathcal{H}}$, we have

$$\begin{aligned}
&\mathcal{R}_\ell(f) - \mathcal{R}_\ell(f;\mathcal{S})\\
=&\frac{1}{\nu}\min_{\lambda\in[0,M]}\Big\{(\nu+\mu)\lambda + \mathbb{E}[\ell(f(\mathbf{x}),y)-\lambda]_+\Big\} - \frac{1}{\nu}\min_{\lambda\in[0,M]}\Big\{(\nu+\mu)\lambda + \frac{1}{n}\sum_{i=1}^n(\ell(f(\mathbf{x}),y)-\lambda)_+\Big\}\\
&+\frac{1}{\nu}\min_{\hat\lambda\in[0,M]}\Big\{\mu\hat\lambda + \frac{1}{n}\sum_{i=1}^n(\ell(f(\mathbf{x}),y)-\hat\lambda)_+\Big\} - \frac{1}{\nu}\min_{\hat\lambda\in[0,M]}\Big\{\mu\hat\lambda + \mathbb{E}[\ell(f(\mathbf{x}),y)-\hat\lambda]_+\Big\}\\
\le&\max_{\lambda\in[0,M]}\Big\{\frac{1}{\nu}\mathbb{E}[\ell(f(\mathbf{x}),y)-\lambda]_+ - \frac{1}{n\nu}\sum_{i=1}^n(\ell(f(\mathbf{x}),y)-\lambda)_+\Big\} := L_1(f,\ell) \qquad \text{(C.17)}\\
&+\max_{\hat\lambda\in[0,M]}\Big\{\frac{1}{n\nu}\sum_{i=1}^n(\ell(f(\mathbf{x}),y)-\hat\lambda)_+ - \frac{1}{\nu}\mathbb{E}[\ell(f(\mathbf{x}),y)-\hat\lambda]_+\Big\} := L_2(f,\ell). \qquad \text{(C.18)}
\end{aligned}$$

Without loss of generality, we consider (C.17), the bound for (C.18) can be derived in a similar manner. Taking supremum on both sides, we have

$$\sup_{\ell_f\in\ell_{\mathcal{H}}} L_1(f,\ell) \le \sup_{\ell_f\in\ell_{\mathcal{H}},\lambda\in[0,M]}\Big\{\frac{1}{\nu}\mathbb{E}[\ell(f(\mathbf{x}),y)-\lambda]_+ - \frac{1}{n\nu}\sum_{i=1}^n(\ell(f(\mathbf{x}),y)-\lambda)_+\Big\} := \Phi(\mathcal{S}).$$

It is standard to verify that $\Phi(\mathcal{S})$ satisfies the bounded differences condition with parameter $\frac{M}{\nu}$ and one can apply McDiarmid's inequality (McDiarmid et al., 1989) so that with probability at least $1-\delta/4$, there holds

$$\Phi(\mathcal{S}) \le \mathbb{E}[\Phi(\mathcal{S})] + \frac{M}{\nu}\sqrt{\frac{\log(4/\delta)}{2n}}.$$

By further standard reduction from the expectation to Rademacher complexity (Theorem 3.3 Mohri et al. (2018)), with probability at least $1-\delta/2$, there holds

$$\Phi(\mathcal{S}) \le 2\mathfrak{R}_n\Big(\frac{1}{\nu}(\mathcal{G})_+\Big) + \frac{3M}{\nu}\sqrt{\frac{\log(4/\delta)}{2n}}, \qquad \text{(C.19)}$$

where $\mathcal{G} = \{\ell_f - \lambda | \ell_f \in \ell_{\mathcal{H}}, \lambda \in [0,M]\}$ and $(\cdot)_+ = \max(\cdot, 0)$. Since the ramp function $(\cdot)_+$ is 1-Lipschitz and $(0)_+ = 0$, by Ledoux-Talagrand contraction inequality (Ledoux and Talagrand, 1991) we have

$$\begin{aligned}
\mathfrak{R}_n\Big(\frac{1}{\nu}(\mathcal{G})_+\Big) \le&\frac{1}{\nu}\mathfrak{R}_n(\mathcal{G}) = \frac{1}{\nu}\mathbb{E}_\sigma\Big[\sup_{\ell_f\in\ell_{\mathcal{H}},\lambda\in[0,M]}\Big(\frac{1}{n}\sum_{i=1}^n\sigma_i\ell(f(\mathbf{x}_i),y_i) - \frac{1}{n}\sum_{i=1}^n\sigma_i\lambda\Big)\Big]\\
\le&\frac{1}{\nu}\Big(\mathbb{E}_\sigma\Big[\sup_{\ell_f\in\ell_{\mathcal{H}}}\frac{1}{n}\sum_{i=1}^n\sigma_i\ell(f(\mathbf{x}_i),y_i)\Big] + \mathbb{E}_\sigma\Big[\sup_{\lambda\in[0,M]}\frac{1}{n}\sum_{i=1}^n\sigma_i\lambda\Big]\Big)\\
\le&\frac{1}{\nu}\Big(\mathfrak{R}_n(\ell_{\mathcal{H}}) + \frac{M}{n}\mathbb{E}_\sigma\Big|\sum_{i=1}^n\sigma_i\Big|\Big)\\
\le&\frac{1}{\nu}\Big(\mathfrak{R}_n(\ell_{\mathcal{H}}) + \frac{M}{\sqrt{n}}\Big), \qquad \text{(C.20)}
\end{aligned}$$

where the last inequality follows by $\left(\mathbb{E}_\sigma\left[\sum_{i=1}^n \sigma_i\right]\right)^2 \le \mathbb{E}_\sigma\left(\sum_{i=1}^n \sigma_i\right)^2 = n$. By putting (C.20) into (C.19), we have

$$\sup_{\ell_f \in \ell_{\mathcal{H}}} L_1(f,\ell) \le \frac{1}{\nu}\Big(2\mathfrak{R}_n(\ell_{\mathcal{H}}) + \frac{M(2\sqrt{2}+3\sqrt{\log(4/\delta)})}{\sqrt{2n}}\Big)$$

with probability at least $1-\delta/2$. The lemma holds by noting $\sup_{\ell_f\in\ell_{\mathcal{H}}}\{R(f,\ell)-R_n(f,\ell)\} \le \sup_{\ell_f\in\ell_{\mathcal{H}}} L_1(f,\ell) + \sup_{\ell_f\in\ell_{\mathcal{H}}} L_2(f,\ell)$. ■

The next corollary is straight-forward from Lemma C.12 by replacing ℓ with $\widetilde{\ell}$.

Corollary C.13 (Theorem 6 restated) *Suppose that the range of $\ell(f(\mathbf{x}), y)$ is $[0, M]$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of an i.i.d. training dataset of size n, the following holds for all $\ell_f \in \ell_{\mathcal{H}}$,*

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

C.5. Examples of Hypothesis Sets

We give two examples of hypothesis sets: linear classifiers and nonlinear neural networks, that satisfy the condition in Theorem 4 and 6. Suppose $\ell : \mathbb{R} \to [0, M]$ is monotonically non-increasing and L-Lipschitz continuous. In this case, the adversarial loss can be written (Yin et al., 2019) as $\widetilde{\ell}(f_\theta(\mathbf{x}), y) := \max_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} \ell(f_\theta(\tilde{\mathbf{x}}), y) = \ell\big(\min_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} y f_\theta(\mathbf{x})\big)$. Therefore, given any function class $\mathcal{H}$, we can define the function class $\widetilde{\mathcal{H}} \subseteq \mathbb{R}^{\mathcal{X}\times\{\pm1\}}$, such that $\widetilde{\mathcal{H}} = \{\min_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(\mathbf{x})} y f_\theta(\mathbf{x}) : f_\theta \in \mathcal{H}\}$. By the Ledoux-Talagrand contraction inequality (Ledoux and Talagrand, 1991), we have $\mathfrak{R}_n(\widetilde{\ell}_{\mathcal{H}}) \le L\mathfrak{R}_n(\widetilde{\mathcal{H}})$. Hence we only need to characterize the Rademacher complexity of $\widetilde{\mathcal{H}}$ given $\mathcal{H}$.

Linear Classifiers. Let the hypothesis set $\mathcal{H}_{lin} \subseteq \mathbb{R}^{\mathcal{X}}$ be a set of linear functions of $\mathbf{x} \in \mathcal{X}$. More specifically, we consider prediction vector θ with l_p ($p \ge 1$) norm constraint, i.e. $\mathcal{H}_{lin} = \{f_\theta(\mathbf{x}) = \theta^\top \mathbf{x} : \|\theta\|_p \le r\}$. Then it is straight-forward to check $\mathcal{H}_{lin}$ is a symmetric hypothesis set. Furthermore, let $\tilde{d} = d^{1-1/p}$. Awasthi et al. (2020) showed that $\max\left\{\mathfrak{R}_n(\mathcal{H}_{lin}), \epsilon r \frac{\max\{\tilde{d},1\}}{2\sqrt{2n}}\right\} \le \mathfrak{R}_n(\widetilde{\mathcal{H}}_{lin}) \le \mathfrak{R}_n(\mathcal{H}_{lin}) + \epsilon r \frac{\max\{\tilde{d},1\}}{2\sqrt{n}}$. Combined with Theorem 6 with probability at least $1-\delta$, we have $\mathcal{R}_{\widetilde{\ell}}(f) - \mathcal{R}_{\widetilde{\ell}}(f;\mathcal{S}) \le \frac{2}{\nu}\Big(2L\mathfrak{R}_n(\mathcal{H}_{lin}) + L\epsilon r \frac{\max\{\tilde{d},1\}}{2\sqrt{n}} + \frac{M(2\sqrt{2}+3\sqrt{\log(2/\delta)})}{\sqrt{2n}}\Big)$, where $\mathfrak{R}_n(\mathcal{H}_{lin})$ is given by classical result in Kakade et al. (2008).

Neural Networks. We consider feedforward neural networks with ReLU activation function ρ, i.e. $\rho(t) = \max\{0, t\}$. In particular, if the hypothesis set is one-layer neural networks defined as $\mathcal{H}_{one} = \{f_\theta(\mathbf{x}) = \theta_0^\top \rho(\Theta\mathbf{x}) : \|\theta_0\|_1 \le r_0, \|\Theta_i\|_p \le r\}$ where $\Theta_i \in \mathbb{R}^d$ is the i-th row of $\Theta \in \mathbb{R}^{d'\times d}$. This is a symmetric hypothesis set. Furthermore, the Rademacher complexity can be upper bounded (Awasthi et al., 2020) as $\mathfrak{R}_n(\widetilde{\mathcal{H}}_{one}) \le \frac{rr_0 \max\{1,\tilde{d}(\|\mathbf{X}\|_\infty+\epsilon)\}}{\sqrt{n}}(1+\sqrt{d(d'+1)\log(36)})$, where $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$. Combined with Theorem 6 with probability at least $1-\delta$, we have $\mathcal{R}_{\widetilde{\ell}}(f) - \mathcal{R}_{\widetilde{\ell}}(f;\mathcal{S}) \le \frac{2}{\nu}\Big(\frac{2Lrr_0 \max\{1,\tilde{d}(\|\mathbf{X}\|_\infty+\epsilon)\}}{\sqrt{n}} \times (1+\sqrt{d(d'+1)\log(36)}) + \frac{M(2\sqrt{2}+3\sqrt{\log(2/\delta)})}{\sqrt{2n}}\Big)$. Such bound implies the generalization error depends on the perturbation size ϵ, which demonstrates the intrinsic complexity of adversarial training.

Appendix D. Additional Experimental Details

D.1. Source Code

For the purpose of review, the source code is accessible in the supplementary file.

D.2. Settings of Networks and Computing Infrastructure Description

For all networks, we training them by using (mini-batch) stochastic gradient descent with momentum 0.9, weight decay 2e-4, batch size 128, epochs 50 (for LeNet) / 100 (for Small-CNN) / 100 (for ResNet-18), and initial learning rate 0.03 (for LeNet) / 0.1 (for Small-CNN) / 0.1 (for ResNet-18) which is divided by the 10 at 20-th and 40-th epoch for LeNet / 30-th and 60-th epoch for Small-CNN and ResNet-18.

All algorithms are implemented in Python 3.6 and trained and tested on an Intel(R) Xeon(R) CPU W5590 @3.33GHz with 48GB of RAM and an NVIDIA Quadro RTX 6000 GPU with 24GB memory.

D.3. Training Settings on Toy Examples

In this section, we provide more details about how to generate synthetic datasets in Figure 1.

We generate two sets of 2D synthetic data (Figure 1). Each dataset contains 200 samples from Gaussian distributions with different means and variances. We consider both the case of the balanced (Figure 1 left) and the imbalanced (Figure 1 right) data distributions, in the former, the training data for the two classes are approximately equal while in the latter one class has a dominating number of samples in comparison to the other. In the balanced dataset (Figure 1 left), we create two outliers. One is in the blue class (shown as red $\times$), the other is in the red class (shown as blue $\circ$). In the imbalanced dataset, we create one outlier in the blue class (shown as red $\times$). For both datasets, the yellow squares around data samples represent the samples are perturbed within a ℓ_∞ ball.

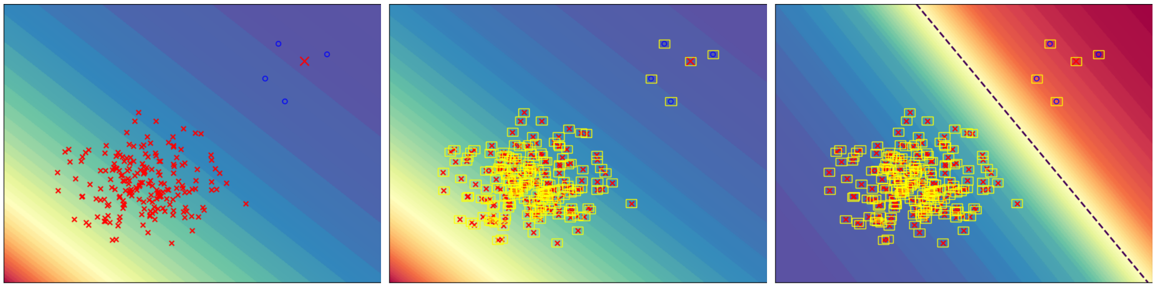
For the balanced dataset (Figure 1 left), we build a simple network contains two linear layers and one ReLU layer (Nair and Hinton, 2010). The number of hidden units is three. For the imbalanced dataset (Figure 1 right), the network contains four linear layers and three ReLU layers. The number of hidden units is 20. We train these networks using SGD with 0.9 momentum for 3000 (balanced dataset) / 100,000 (imbalanced dataset) iterations with the learning rate of 0.02. We set $k = 20$ and $m = 2$ for balanced dataset, and $k = 20$ and $m = 1$ for imbalanced dataset when run our `ORAT` algorithm. In AT and `ORAT` , the training attack is PGD^{10} and we set the perturbation bound $\epsilon = 0.01$ and the PGD step size $\epsilon/4$.

D.4. Details of Outliers Generation by Using Asymmetric Noise

In asymmetric noise generation procedure, for MNIST, flipping 2→7, 3→8, 5↔6 and 7→1; for CIFAR-10, flipping TRUCK→AUTOMOBILE, BIRD→AIRPLANE, DEER→HORSE, CAT↔DOG; for CIFAR-100, the 100 classes are grouped into 20 super-classes with each having 5 sub-classes, then flipping between two randomly selected sub-classes within each super-class.

Noise		MNIST				CIFAR-10				CIFAR-100			
		ϵ=0.1		ϵ=0.2		ϵ=2/255		ϵ=8/255		ϵ=2/255		ϵ=8/255	
		k	m	k	m	k	m	k	m	k	m	k	m
0		60000	1	60000	1	45000	1	45000	1	49950	1	49950	1
Symmetric Noise	10%	59950	2000	60000	2000	50000	300	50000	500	50000	300	50000	500
	20%	59950	6000	60000	3000	50000	300	50000	300	50000	500	49800	500
	30%	59950	5000	60000	5000	50000	10	50000	200	49800	100	50000	500
	40%	59950	11000	60000	11000	50000	100	50000	50	49950	100	49950	500
Asymmetric Noise	10%	60000	100	60000	10	49950	300	50000	500	49900	100	49950	500
	20%	59950	100	59950	100	49950	500	50000	300	50000	100	50000	500
	30%	59950	10	60000	100	49950	200	50000	300	50000	100	50000	500
	40%	59950	10	60000	10	50000	450	50000	500	50000	100	49800	500

Table D.1: *The k and m settings of* `ORAT` *on real datasets in different noise.*



(a) Standard Training (ST)

(b) Adversarial Training (AT)

(c) `ORAT`

Figure E.1: *An additional illustrative example of standard training (ST), adversarial training (AT), and ATRR for binary classification on an imbalanced synthetic dataset with one outlier (shown as red $\times$) in the blue class. The yellow squares around data samples represent the samples are perturbed within a ℓ_∞ ball. The dashed line is the decision boundary. The figure is better viewed in color.*

D.5. k and m Settings on Real Datasets

We provide a reference for setting k and m to reproduce our `ORAT` experimental results (Table 1) on real datasets in Table D.1.

Appendix E. Additional Experimental Results

E.1. More Experiments on Toy Example

We generate additional 2D synthetic data as shown in Figure E.1 to demonstrate the performance of our `ORAT` method. This imbalanced dataset contains 200 samples from Gaussian distribution with different means and variances. For this dataset, we create one outlier in the blue class (shown as red $\times$). In order to train this dataset, we build a network, which contains two linear layers and one ReLU layer. The number of hidden units is 64. We train this network using SGD with 0.9 momentum for 100,000 iterations with a learning rate of 0.1. We set $k = 5$ and $m = 1$ for `ORAT` . Similarly, in AT and `ORAT` , the training attack is PGD^{10}, the perturbation bound $\epsilon = 0.01$, and the PGD step size is $\epsilon/4$.

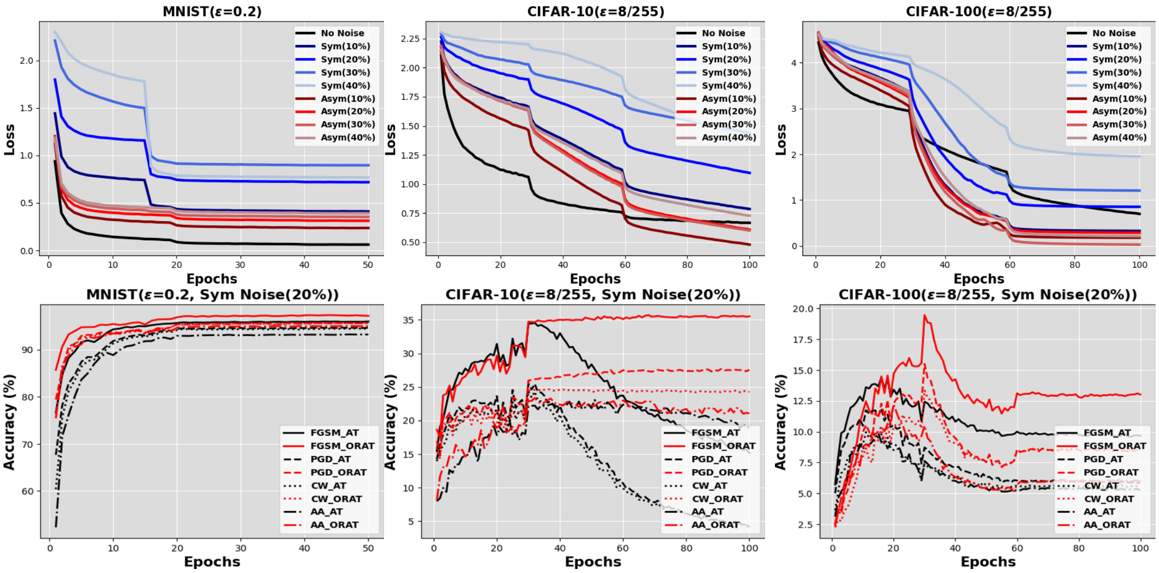


Figure E.2: *The tendency curves of training adversarial loss and test accuracy on three datasets. The sharp drops in the curves correspond to decreases in training learning rate.*

From Figure E.1, we can find the classifiers are trained from ST (a) and AT (b) cannot separate two classes in the training data. However, we find the classifier is training by using our proposed ORAT can separate these two classes. The results demonstrate that our ORAT can eliminate the influence of the outliers when doing the adversarial training.

E.2. More Experiments on Real Datasets

In the main paper, we only show the tendency curves for MNIST when ϵ=0.1 and CIFAR-10 and CIFAR-100 when ϵ=2/255. In this section, we show more results on three datasets with 20% symmetric noise by setting a big value of ϵ in Figure E.2. Similar to the observations in Figure 2, we can find the losses are dramatically decreased in the first row of Figure E.2, which means Algorithm 1 can be successfully applied to solve ORAT optimization problem. From the second row of Figure E.2, it is obvious that the performance of our method is higher than the original AT approach on all attacks.

E.3. More Experiments on the Effect of k and m

We conduct more experiments to study the effect of hyperparameters k and m with using 20% symmetric noise on all datasets by setting a big value of ϵ. The results are shown in Figure E.3. Similar to the results that we get in Figure 3, we can see that there is a clear range of m with better performance than all compared methods. Fix m and test various k, we can find the performance can be improved by using some specific k values.

E.4. Connection with Adversarial Training on Out-of-Distribution Problems

Out-of-Distribution (OOD) problem exists due to the training and test data distributions mismatching (Hendrycks et al., 2021). Although the OOD problem setting is different from our outliers problem setting, some similarities exist between OOD data and outliers. For

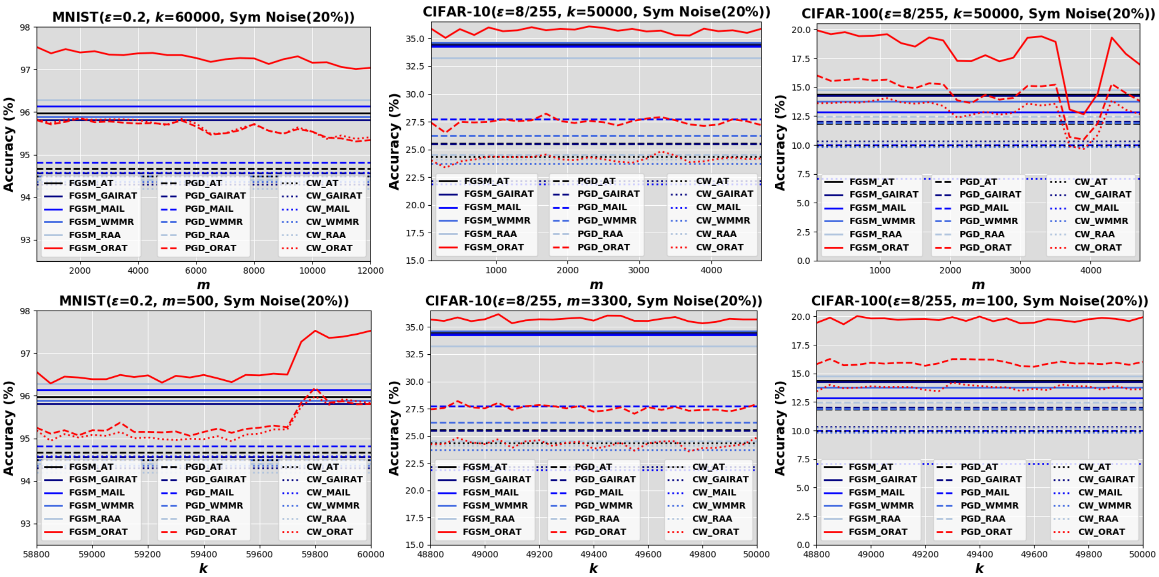


Figure E.3: *Effect of k and m on the test accuracy of* `ORAT` *on three datasets.*

example, both of them are not from the data generating distribution. Therefore, whether the OOD methods can directly apply to solving our outlier problem in adversarial training is a question. Some works such as Zeng et al. (2021b); Varshney et al. (2022); Yi et al. (2021) connect adversarial robustness to out-of-distribution (OOD) problems.

Specifically, Zeng et al. (2021b) focuses on OOD detection. The problem in Zeng et al. (2021b) is that not enough labeled OOD samples can be used for training the OOD detection model. To improve the diversity of the unlabeled data augmentation, they apply an adversarial attack technique on unlabeled data to generate pseudo-positive samples. Then use these pseudo-positive samples with labeled data to improve the performance of the OOD detection model. However, their approach cannot directly apply to our setting since we only focus on supervised learning. All training data points are labeled in our setting, and the adversarial training works on labeled data. The authors in Varshney et al. (2022) test different selective prediction approaches for Natural Language Processing systems in in-domain, OOD, and adversarial settings. They regard several existing datasets as adversarial datasets for testing. However, no adversarial training approach is proposed and involved in Varshney et al. (2022). For Yi et al. (2021), the authors theoretically and experimentally show that a model (original AT (Madry et al., 2018) or pre-trained AT (Salman et al., 2020)) robust to input perturbation generalizes well on OOD data.

Therefore, we test whether the pre-trained AT method (Salman et al., 2020) can solve outlier problems in adversarial training. Following the experimental setting from Yi et al. (2021), we download the ImageNet-based adversarially pre-trained robust ResNet-18 model in the setting of L_∞ and $\epsilon = 2/255$ from the public repository [1]. Then fine-tune it on our noisy training datasets. We report pre-trained AT testing accuracy (%) on CIFAR-100 in Table E.1. To make the comparison explicit, we also attach our method performance. From Table E.1, we can find our method outperforms pre-trained AT under all settings. Most of the performance gaps between pre-trained AT and our method in Table E.1 are more than

[1] `https://github.com/microsoft/robust-models-transfer`

Noise		Defense	CIFAR-100($\epsilon = 2/255$)			
			Na	FG	PGD	CW
Symmetric Noise	10%	pre-trained AT	24.20	17.09	14.83	14.17
		Ours	**35.76**	**25.72**	**22.27**	**21.28**
	20%	pre-trained AT	19.53	16.64	14.01	13.26
		Ours	**34.45**	**25.07**	**22.21**	**20.92**
	30%	pre-trained AT	19.41	16.22	13.06	13.04
		Ours	**31.27**	**23.81**	**21.35**	**19.59**
	40%	pre-trained AT	18.86	15.63	12.78	11.91
		Ours	**29.38**	**22.99**	**20.85**	**19.20**
Asymmetric Noise	10%	pre-trained AT	20.15	19.66	17.78	16.9
		Ours	**37.09**	**27.07**	**23.65**	**22.59**
	20%	pre-trained AT	24.60	17.43	16.67	15.15
		Ours	**36.05**	**25.76**	**22.83**	**21.47**
	30%	pre-trained AT	22.86	16.92	15.94	14.89
		Ours	**34.58**	**24.18**	**21.05**	**20.11**
	40%	pre-trained AT	21.58	16.18	15.32	14.48
		Ours	**33.65**	**23.35**	**20.76**	**19.46**

Table E.1: *Testing accuracy (%) of pre-trained AT and our method (ORAT) on CIFAR-100 ($\epsilon = 2/255$) with different levels of symmetric and asymmetric noise. The best results are shown in bold.*

5%. One reason is that the pre-trained AT is not designed to handle outliers. According to these results, it is clear that pre-trained AT cannot directly apply to solving our problem even if it has a good performance on OOD data.

E.5. Extension of Table 3

Self-learning (Han et al., 2019) is a useful strategy for learning model on noise data. For example, we can use AoRR to filter examples with larger loss (potential outliers), then conducting adversarial training on the cleaner set. We call this method AT w/o. However, it is not an end-to-end training approach. In contrast, our method is an end-to-end method, which means it is very easy to be conducted. To compare the effectiveness of `ORAT` and this AT w/o approach, we conduct experiments on MNIST with symmetric noise and CIFAR-100 with symmetric noise as follows.

In the first stage, for each dataset, we apply a grid search to select the values of k and m for training the model using the AoRR approach that can return a good testing accuracy. Then we use the trained model to test the loss for each sample from the training set. Therefore, we can obtain a training sample loss list. Next, we delete data points for the m largest losses in the training set to construct a clean set. This is because the AoRR uses m to determine how many examples (potential outliers) with the largest losses are ignored during each training epoch.

In the second stage, after we get a clean set, we use the conventional AT approach to train the model on the clean set and test the trained model on the testing set.

We report the testing accuracy (%) of the AT w/o approach on MNIST (symmetric noise, $\epsilon = 0.1$) and CIFAR-100 (symmetric noise, $\epsilon = 2/255$) in Table E.2. To make the comparison

Noise		Defense	MNIST ($\epsilon = 0.1$)				CIFAR-100 ($\epsilon = 2/255$)			
			Na	FG	PGD	CW	Na	FG	PGD	CW
Symmetric Noise	10%	AT w/o	98.91	98.08	97.61	97.55	29.81	21.09	19.59	18.23
		Ours	**99.52**	**98.45**	**97.78**	**97.79**	**35.76**	**25.72**	**22.27**	**21.28**
	20%	AT w/o	98.77	97.76	97.20	97.12	27.81	20.92	19.15	17.95
		Ours	**99.56**	**98.37**	**97.65**	**97.64**	**34.45**	**25.07**	**22.21**	**20.92**
	30%	AT w/o	97.82	96.97	96.47	96.35	24.18	19.17	17.31	16.71
		Ours	**99.55**	**98.30**	**97.51**	**97.53**	**31.27**	**23.81**	**21.35**	**19.59**
	40%	AT w/o	97.03	95.85	95.22	95.05	21.17	17.81	16.85	15.48
		Ours	**99.36**	**98.00**	**97.22**	**97.20**	**29.38**	**22.99**	**20.85**	**19.20**

Table E.2: *Testing accuracy (%) of self-learning based method (Self-learning) and our method (ORAT) on MNIST ($\epsilon = 0.1$) and CIFAR-100 ($\epsilon = 2/255$) with different levels of symmetric noise. The best results are shown in bold.*

explicit, we also attach our method performance. From Table E.2, we can find our method outperforms the AT w/o approach under all settings. For example, the performance gap between the AT w/o approach and our method (ORAT) on MNIST can achieve more than 2% under the 40% symmetric noise setting. Most of the performance gaps on CIFAR-100 can achieve more than 4%.

One reason for low performance from the self-learning approach is that the training data points ignored by AoRR may contain clean data points. In this case, the constructed clean set is smaller than the original dataset. This may hurt the final model performance. Moreover, removing the examples with the largest losses before the adversarial training may lose the important feature information from the original training dataset. In other words, this compromises the richness and representational power of the data. In contrast, our ORAT method considers all examples during adversarial training. According to these results, it is clear that our approach (ORAT) gives a better solution than the self-learning approach for solving outlier problems in adversarial training either in the algorithm efficiency or effectiveness.

E.6. More Analysis on Stability of ORAT

To evaluate the stability of each method, we report the the mean and standard deviation of testing accuracy (%) of all methods on MNIST (40% symmetric noise, $\epsilon = 0.1$) and CIFAR-100 (40% symmetric noise, $\epsilon = 2/255$) in Table 4. For each method, the reported performance is obtained by averaging the testing accuracy according to 10 random seeds. From Table 4, we can find our method still outperforms the compared methods in both datasets. For MNIST, our method can even outperform AT by more than 2%. Most importantly, we can find that the standard deviation in our method is less than or equal to that of other compared methods. For CIFAR-100, we can find the mean value of our method ORAT even higher than the reported performance in our submission. The standard deviation of the performance of our method differs from the comparison methods by at most 0.26% (compared to ST on FGSM attack). Comparing Table 4 and Table 1, it is clear that the performance gap becomes larger when we report scores by using mean and standard deviation, and our method shows a stable and stronger ability in handling outliers and adversarial attacks.

Noise	Defense	CIFAR-10 ($\epsilon = 2/255$)				CIFAR-10 ($\epsilon = 8/255$)			
		Na	FG	PGD	CW	Na	FG	PGD	CW
20% Sym Noise	ST	91.31	53.25	27.66	26.47	91.09	35.36	7.11	7.34
	AT	90.74	82.86	78.97	78.96	81.80	62.43	50.84	50.92
	GAIRAT	88.17	84.00	81.72	76.05	78.55	67.95	61.94	47.81
	MAIL	73.31	65.94	62.85	58.05	69.01	52.09	44.33	38.85
	WMMR	87.75	79.83	76.21	75.48	80.97	61.70	51.17	49.82
	RAA	90.55	82.44	78.38	78.56	77.62	60.05	48.87	48.99
	Ours	90.72	**85.47**	**82.30**	**80.34**	81.99	**69.08**	**63.87**	**53.53**

Table E.3: *Testing accuracy (%) using Wide ResNet.*

E.7. Evaluation on Wide ResNet

We evaluate all methods using Wide ResNet on CIFAR-10 dataset with 20% symmetric noise. The Wide ResNet framework is WRN-32-10, which is the same as Madry et al. (2018). Results in Table E.3 show our approach outperforms others when using a large model.

E.8. Experiments on Clothing1M

To demonstrate the effectiveness of our method ORAT on a more real scenario, we conduct experiments on the Clothing1M dataset Xiao et al. (2015). This dataset contains roughly one million clothing images crawled from the Internet. Most of them have noisy labels extracted from their surrounding texts. A few of them have clean labels, which are manually annotated by Xiao et al. Xiao et al. (2015). Specifically, we extract 30000 clean labeled images as the clean training set and 10000 clean labeled images as the test set. To create a noise training set, we select 80% images from the clean training set and extract 30000×20%=6000 images from the original noise labeled images. Therefore, we can obtain a noise training set with the same sample size as the clean training set. Then we use AT to train the Small-CNN model on the clean training set (named AT (No noise)) and noise training set (named AT (20% noise)), respectively. For our method ORAT, we use it to train the same model on the noise training set, named ORAT (20% noise). We show the tendency curves of the test accuracy in Figure E.4. From the table, comparing AT (No noise) and AT (20% noise), we can see that AT performance is decreased if the data contains noise, which means outliers affect the performance of AT. In addition, our ORAT method outperforms AT on the noise data, which means our method can reduce the influence of outliers on adversarial training.

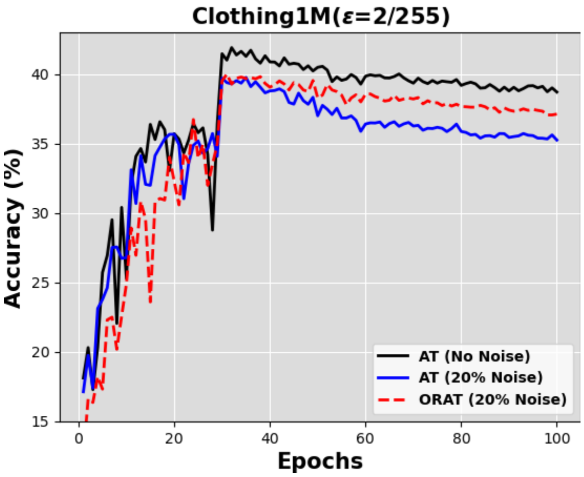


Figure E.4: *The tendency curves of testing accuracy on the Clothing1M dataset.*