A. Details of the illustration Example

In Sec. 2.3, we use an illustration example to draw the theoretical results. Here we show the implementation detail of this toy example.

Note that for the case $d = 1$, the data distribution is

$$x_1 = \begin{cases} +y, & \text{w.p. } p_y \\ -y, & \text{w.p. } 1 - p_y \end{cases} \quad \text{and} \quad x_2 \overset{\text{i.i.d}}{\sim} \mathcal{N}(\eta y, \sigma^2). \tag{9}$$

In this toy model, we select $p_{+1} = 0.85 > 0.7 = p_{-1}$ and $\eta = 0.4$. The variance σ^2 is set to be 0.6 for better visualization in this toy model, and in the following theoretical analysis, we set $\sigma^2 = 1$ for simplicity. In Fig. 1(a), we randomly sample 100 pairs of (x_1, x_2) for each class $y \in \{+1, -1\}$. In Fig. 1(b), the robustness is evaluated under perturbation bound $\epsilon = 2\eta = 0.8$, which is consistent to the evaluation in [22].

B. Proofs for Theorems in Sec. 2.3

B.1. Preliminaries

We denote the distribution function and the probability density function of the *normal distribution* $\mathcal{N}(0, 1)$ as $\phi(x)$ and $\Phi(x)$:

$$\begin{aligned} \Phi(x) &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}t = \mathrm{Pr.}(\mathcal{N}(0,1) < x), \\ \phi(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \Phi'(x). \end{aligned} \tag{10}$$

Recall that the data distribution is

$$\begin{aligned} x_1 &= \begin{cases} +y, & \text{w.p. } p_y, \\ -y, & \text{w.p. } 1 - p_y, \end{cases} \\ x_2, \cdots, x_{d+1} &\overset{\text{i.i.d}}{\sim} \mathcal{N}(\eta y, 1), \end{aligned} \tag{11}$$

where $1 > p_{+1} > p_{-1} > \frac{1}{2}$. First we calculate the clean accuracy $\mathcal{A}_y(f_w)$ and the robust accuracy $\mathcal{R}_y(f_w)$ for any class $y \in \{+1, -1\}$ and $w > 0$. Also recall that the classifier

$$f_w = \text{sign}(x_1 + \frac{x_2 + \cdots + x_{d+1}}{w}) \tag{12}$$

Note that $w > 0$, we have

$$\begin{aligned} \mathcal{A}_{+1}(f_w) &= \mathrm{Pr.}(\text{sign}(f_w) = 1) \\ &= \mathrm{Pr.}(x_1 + \frac{x_2 + \cdots + x_{d+1}}{w} > 0) \\ &= p_{+1} \cdot \mathrm{Pr.}(1 + \frac{x_2 + \cdots + x_{d+1}}{w} > 0) \\ &+ (1 - p_{+1}) \cdot \mathrm{Pr.}(-1 + \frac{x_2 + \cdots + x_{d+1}}{w} > 0) \\ &= p_{+1} \cdot \mathrm{Pr.}(x_2 + \cdots + x_{d+1} > -w) \\ &+ (1 - p_{+1}) \cdot \mathrm{Pr.}(x_2 + \cdots + x_{d+1} > w) \\ &= p_{+1} \cdot \mathrm{Pr.}(\mathcal{N}(d\eta, d) > -w) \\ &+ (1 - p_{+1}) \cdot \mathrm{Pr.}(\mathcal{N}(d\eta, d) > w) \\ &= p_{+1} \cdot \mathrm{Pr.}(\mathcal{N}(0, d) > -d\eta - w) \\ &+ (1 - p_{+1}) \cdot \mathrm{Pr.}(\mathcal{N}(0, d) > -d\eta + w) \\ &= p_{+1} \cdot \mathrm{Pr.}(\mathcal{N}(0, 1) < \frac{d\eta + w}{\sqrt{d}}) \\ &+ (1 - p_{+1}) \cdot \mathrm{Pr.}(\mathcal{N}(0, 1) < \frac{d\eta - w}{\sqrt{d}}) \\ &= p_{+1}\Phi(\frac{d\eta + w}{\sqrt{d}}) + (1 - p_{+1})\Phi(\frac{d\eta - w}{\sqrt{d}}). \end{aligned} \tag{13}$$

Similarly, we have

$$\mathcal{A}_{-1}(f_w) = p_{-1}\Phi(\frac{d\eta + w}{\sqrt{d}}) + (1 - p_{-1})\Phi(\frac{d\eta - w}{\sqrt{d}}). \tag{14}$$

For the robustness, following the evaluation in the original model [22], we evaluate the robustness $\mathcal{R}_y$ under l_∞-norm perturbation bound $\epsilon = 2\eta < 1$. Consider the distribution of adversarial examples $\hat{x} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_{d+1})$. Since we restrict the robust feature $x_1 \in \{-1, +1\}$ and $\epsilon < 1$, we have $\hat{x}_1 = x_1$. For the non-robust features $x_i \sim \mathcal{N}(\eta y, 1)$, the corresponding adversarial example has $\hat{x}_i \sim \mathcal{N}(-\eta y, 1)$ under the perturbation bound $\epsilon = 2\eta$. Therefore, the distribution of adversarial examples is

$$\hat{x}_1 = \begin{cases} +y, & \text{w.p. } p_y \\ -y, & \text{w.p. } 1 - p_y \end{cases} \quad \text{and} \quad \hat{x}_2, \cdots, \hat{x}_{d+1} \overset{\text{i.i.d}}{\sim} \mathcal{N}(-\eta y, 1). \tag{15}$$

By simply replacing η with $-\eta$ in derivative process of (13), for any $w > 0$, we have

$$\begin{aligned} \mathcal{R}_{+1}(f_w) &= p_{+1}\Phi(\frac{-d\eta + w}{\sqrt{d}}) + (1 - p_{+1})\Phi(\frac{-d\eta - w}{\sqrt{d}}), \\ \mathcal{R}_{-1}(f_w) &= p_{-1}\Phi(\frac{-d\eta + w}{\sqrt{d}}) + (1 - p_{-1})\Phi(\frac{-d\eta - w}{\sqrt{d}}). \end{aligned} \tag{16}$$

B.2. Proof of Theorem 1

The theorem 1 shows the class $y = -1$ is intrinsically difficult to learn than class $y = +1$:

Theorem 1 For any $w > 0$ and the classifier $f_w = \text{sign}(x_1 + \frac{x_2+\cdots+x_{d+1}}{w})$, we have $\mathcal{A}_{+1}(f_w) > \mathcal{A}_{-1}(f_w)$ and $\mathcal{R}_{+1}(f_w) > R_{-1}(f_w)$.

Proof. Note that $p_{+1} > p_{-1}$, and $\Phi(\frac{d\eta+w}{\sqrt{d}}) > \Phi(\frac{d\eta-w}{\sqrt{d}})$, we have

$$\begin{aligned}\mathcal{A}_{+1}(f_w) &= p_{+1}\Phi(\frac{d\eta+w}{\sqrt{d}}) + (1-p_{+1})\Phi(\frac{d\eta-w}{\sqrt{d}})\\ &= p_{+1}(\Phi(\frac{d\eta+w}{\sqrt{d}}) - \Phi(\frac{d\eta-w}{\sqrt{d}})) + \Phi(\frac{d\eta-w}{\sqrt{d}})\\ &> p_{-1}(\Phi(\frac{d\eta+w}{\sqrt{d}}) - \Phi(\frac{d\eta-w}{\sqrt{d}})) + \Phi(\frac{d\eta-w}{\sqrt{d}})\\ &= \mathcal{A}_{-1}(f_w).\end{aligned} \tag{17}$$

B.3. Proof of Theorem 2

The theorem 2 shows the relation between the parameter w and the attack strength (perturbation bound ϵ) in adversarial training:

Theorem 2 For any $0 \le \epsilon \le \eta$, the optimal parameter w for adversarial training with perturbation bound ϵ is monotone increasing at ϵ.

Proof. Similar to the adversarial example distribution analysis (15), under the perturbation bound ϵ, the data distribution of the crafted adversarial example for training is

$$\begin{aligned}&\tilde{x}_1 = \begin{cases} +y, & \text{w.p. } p_y \\ -y, & \text{w.p. } 1-p_y \end{cases},\\ &\tilde{x}_2, \cdots, \tilde{x}_{d+1} \overset{\text{i.i.d}}{\sim} \mathcal{N}((\eta-\epsilon)y, 1).\end{aligned} \tag{18}$$

We use $\tilde{\mathcal{A}}(f_w)$, $\tilde{\mathcal{A}}_y(f_w)$ to denote the overall and class-wise *train accuracy* of the classifier f_w on training data distribution (18). Let $p = p_{+1} + p{-}1$. Then the overall train accuracy of f_w is

$$\begin{aligned}&\tilde{\mathcal{A}}(f_w) = \frac{1}{2}(\tilde{\mathcal{A}}_{+1}(f_w) + \tilde{\mathcal{A}}_{-1}(f_w))\\ &= \frac{1}{2}(p_{+1}\Phi(\frac{d(\eta-\epsilon)+w}{\sqrt{d}}) + (1-p_{+1})\Phi(\frac{d(\eta-\epsilon)-w}{\sqrt{d}})\\ &+ p_{-1}\Phi(\frac{d(\eta-\epsilon)+w}{\sqrt{d}}) + (1-p_{-1})\Phi(\frac{d(\eta-\epsilon)-w}{\sqrt{d}}))\\ &= \frac{1}{2}(p\Phi(\frac{d(\eta-\epsilon)+w}{\sqrt{d}}) + (2-p)\Phi(\frac{d(\eta-\epsilon)-w}{\sqrt{d}})).\end{aligned} \tag{19}$$

Now we calculate the best parameter w for $\tilde{\mathcal{A}}(f_w)$. Note that $\Phi'(x) = \phi(x)$, we have

$$\begin{aligned}&\frac{\partial\tilde{\mathcal{A}}(f_w)}{\partial w} = \frac{1}{2\sqrt{d}}(p\phi(\frac{d(\eta-\epsilon)+w}{\sqrt{d}}) - (2-p)\phi(\frac{d(\eta-\epsilon)-w}{\sqrt{d}}))\\ &= \frac{1}{2\sqrt{2\pi d}}\{p\exp[-\frac{1}{2}(\frac{d(\eta-\epsilon)+w}{\sqrt{d}})^2]\\ &- (2-p)\exp[-\frac{1}{2}(\frac{d(\eta-\epsilon)-w}{\sqrt{d}})^2]\}\end{aligned} \tag{20}$$

Therefore, $\frac{\partial\tilde{\mathcal{A}}(f_w)}{\partial w} > 0$ is equivalent to

$$\begin{aligned}&p\exp[-\frac{1}{2}(\frac{d(\eta-\epsilon)+w}{\sqrt{d}})^2] > (2-p)\exp[-\frac{1}{2}(\frac{d(\eta-\epsilon)-w}{\sqrt{d}})^2]\\ &\iff \exp[-\frac{1}{2}((\frac{d(\eta-\epsilon)+w}{\sqrt{d}})^2 - (\frac{d(\eta-\epsilon)-w}{\sqrt{d}})^2)] > \frac{2-p}{p}\\ &\iff \exp[-\frac{1}{2d}\cdot(4d(\eta-\epsilon)w)] > \frac{2-p}{p}\\ &\iff \exp[-2(\eta-\epsilon)w] > \frac{2-p}{p}\\ &\iff -2(\eta-\epsilon)w > \ln(\frac{2-p}{p})\\ &\iff w < \frac{1}{2(\eta-\epsilon)}\ln(\frac{p}{2-p}) := \hat{w}_\epsilon.\end{aligned} \tag{21}$$

Recall that we assume $p_{+1}, p_{-1} > \frac{1}{2}$, thus $p = p_{+1}+p_{-1} > 1$ and $\frac{p}{2-p} > 1$. Therefore, $\frac{\partial\tilde{\mathcal{A}}(f_w)}{\partial w} > 0$ when $w < \hat{w}_\epsilon$, and $\frac{\partial\tilde{\mathcal{A}}(f_w)}{\partial w} < 0$ when $w > \hat{w}_\epsilon$. We can conclude that f_w obtains the optimal parameter w, *i.e.*, w achieves the highest train accuracy, when $w = \hat{w}_\epsilon = \frac{1}{2(\eta-\epsilon)}\ln(\frac{p}{2-p})$, which is monotone increasing at ϵ.

B.4. Proof of Theorem 3

Theorem 3 shows the clean accuracy of the hard class $y = -1$ drops earlier than class $y = +1$ as the attack strength increases:

Theorem 3 Let $w_y^* = \arg\max_w \mathcal{A}_y(f_w)$ be the parameter for the best clean accuracy of class y, then $w_{+1}^* > w_{-1}^*$.

Proof. As calculated in (13) and (14), we have $\mathcal{A}_y(f_w) = p_y\Phi(\frac{d\eta+w}{\sqrt{d}}) + (1-p_y)\Phi(\frac{d\eta-w}{\sqrt{d}})$ and

$$\frac{\partial\mathcal{A}(f_w)}{\partial w} = \frac{1}{\sqrt{d}}(p_y\phi(\frac{d\eta+w}{\sqrt{d}}) - (1-p_y)\phi(\frac{d\eta-w}{\sqrt{d}})). \tag{22}$$

Therefore, $\frac{\partial \mathcal{A}(f_w)}{\partial w} > 0$ is equivalent to

$$\begin{aligned} &\exp\{-\frac{1}{2}[(\frac{d\eta+w}{\sqrt{d}})^2-(\frac{d\eta-w}{\sqrt{d}})^2]\} > \frac{1-p_y}{p_y} \\ \iff &\exp\{-2\eta w\} > \frac{1-p_y}{p_y} \\ \iff & -2\eta w > \ln(\frac{1-p_y}{p_y}) \\ \iff & w < \frac{1}{2\eta}\ln(\frac{p_y}{1-p_y}). \end{aligned} \tag{23}$$

Similar to the proof of Theorem 2, we have $w_y^* = \arg\max \mathcal{A}_y(f_w) = \frac{1}{2\eta}\ln(\frac{p_y}{1-p_y})$. Since $1 > p_{+1} > p_{-1} > \frac{1}{2}$, we have $\frac{p_{+1}}{1-p_{+1}} > \frac{p_{-1}}{1-p_{-1}} > 1$ and hence $w_{+1}^* > w_{-1}^*$.

B.5. Proof of Theorem 4

Theorem 4 shows how strong attack in adversarial training hurts the hard class $y = -1$:

Theorem 4 Suppose $\Delta_w > 0$, then for $\forall w > w_{+1}^*$, $\mathcal{A}_{-1}(f_{w+\Delta_w}) - \mathcal{A}_{-1}(f_w) < \mathcal{A}_{+1}(f_{w+\Delta_w}) - \mathcal{A}_{+1}(f_w) < 0$, and for $\forall w > 0$, $0 < \mathcal{R}_{-1}(f_{w+\Delta_w}) - \mathcal{R}_{-1}(f_w) < \mathcal{R}_{+1}(f_{w+\Delta_w}) - \mathcal{R}_{+1}(f_w)$.

Proof. First we prove for $u > w_{+1}^*$,

$$\mathcal{A}_{-1}(f_{w+\Delta_w}) - \mathcal{A}_{-1}(f_w) < \mathcal{A}_{+1}(f_{w+\Delta_w}) - \mathcal{A}_{+1}(f_w) < 0. \tag{24}$$

Since we have

$$\mathcal{A}_y(f_{w+\Delta_w}) - \mathcal{A}_y(f_w) = \int_w^{w+\Delta w} \frac{\partial \mathcal{A}_y(f_u)}{\partial u} \mathrm{d}u, \tag{25}$$

It's suffice to show that

$$\frac{\partial \mathcal{A}_{-1}(f_u)}{\partial u} < \frac{\partial \mathcal{A}_{+1}(f_u)}{\partial u} < 0, \quad \forall u > w_{+1}^*. \tag{26}$$

Recall that in the proof of Theorem 3, we have shown

$$\begin{aligned} \frac{\partial \mathcal{A}(f_u)}{\partial w} &= \frac{1}{\sqrt{d}}(p_y\phi(\frac{d\eta+w}{\sqrt{d}}) - (1-p_y)\phi(\frac{d\eta-w}{\sqrt{d}})) \\ &= \frac{1}{\sqrt{d}}\{p_y[\phi(\frac{d\eta+w}{\sqrt{d}}) + \phi(\frac{d\eta-w}{\sqrt{d}})] - \phi(\frac{d\eta-w}{\sqrt{d}})\}. \end{aligned} \tag{27}$$

Therefore, since $p_{-1} < p_{+1}$ and $\phi(\frac{d\eta+w}{\sqrt{d}}) + \phi(\frac{d\eta-w}{\sqrt{d}}) > 0$, we have

$$\frac{\partial \mathcal{A}_{-1}(f_u)}{\partial u} < \frac{\partial \mathcal{A}_{+1}(f_u)}{\partial u}. \tag{28}$$

Further, since $u > w_{+1}^*$, we have $\frac{\partial \mathcal{A}_{+1}(f_u)}{\partial u} < 0$ as shown in the proof of Theorem 3.

Next, we prove that for $\forall w > 0$,

$$0 < \mathcal{R}_{-1}(f_{w+\Delta_w}) - \mathcal{R}_{-1}(f_w) < \mathcal{R}_{+1}(f_{w+\Delta_w}) - \mathcal{R}_{+1}(f_w). \tag{29}$$

Similarly, it suffice to show

$$0 < \frac{\partial \mathcal{R}_{-1}(f_u)}{\partial u} < \frac{\partial \mathcal{R}_{+1}(f_u)}{\partial u}, \quad \forall u > 0. \tag{30}$$

Recall the expression (16), we have

$$\mathcal{R}_y = p_y\Phi(\frac{-d\eta+w}{\sqrt{d}}) + (1-p_y)\Phi(\frac{-d\eta-w}{\sqrt{d}}), \tag{31}$$

hence

$$\begin{aligned} \frac{\partial \mathcal{R}_y(f_w)}{\partial w} &= \frac{1}{\sqrt{d}}\{p_y\phi(\frac{-d\eta+w}{\sqrt{d}}) - (1-p_y)\phi(\frac{-d\eta-w}{\sqrt{d}})\} \\ &= \frac{1}{\sqrt{d}}\{p_y[\phi(\frac{-d\eta+w}{\sqrt{d}}) + \phi(\frac{-d\eta-w}{\sqrt{d}})] - \phi(\frac{-d\eta-w}{\sqrt{d}})\} \end{aligned} \tag{32}$$

Since $p_{+1} > p_{-1}$ and $\phi(\frac{-d\eta+w}{\sqrt{d}}) + \phi(\frac{-d\eta-w}{\sqrt{d}}) > 0$, we have

$$\frac{\partial \mathcal{R}_{-1}(f_u)}{\partial u} < \frac{\partial \mathcal{R}_{+1}(f_u)}{\partial u}. \tag{33}$$

Finally, as $d, \eta, w > 0$, we have $(\frac{-d\eta+w}{\sqrt{d}})^2 < (\frac{-d\eta-w}{\sqrt{d}})^2$ by comparing their absolute value. This indicates $\phi(\frac{-d\eta+w}{\sqrt{d}}) > \phi(\frac{-d\eta-w}{\sqrt{d}})$. Also note that $p_{-1} > \frac{1}{2}$ and $p_{-1} > (1-p_{-1})$, we have

$$\frac{1}{\sqrt{d}}\{p_{-1}\phi(\frac{-d\eta+w}{\sqrt{d}}) - (1-p_{-1})\phi(\frac{-d\eta-w}{\sqrt{d}})\} > 0, \tag{34}$$

which completes our proof.

C. More Experiments

Here we present additional experimental results.

C.1. Experiment on Tiny-ImageNet

Besides CIFAR-10, we additionally compare CFA with baseline+EMA on **Tiny-ImageNet** with ResNet-18 under ℓ_∞-norm bound $\epsilon = 4/255$. Since the worst class robustness is extremely low and there are only 50 images for each class in the test set, we report the average of the worst-20% class robustness. The threshold of FAWA is also set on the average robustness of these classes on validation set. The results in Table 4 shows that our CFA framework still outperforms baseline+EMA on Tiny-ImageNet.

C.2. Class-wise robustness comparison

To evaluate class-wise robustness, we present a comparison between CFA and EMA on CIFAR-10, as shown in Fig. 6. The results show that CFA significantly outperforms EMA on classes {2,3,4,6}, while slightly dropping on classes {7,8}. Compared with the improvements, the decreases are very slight. Moreover, the variance of class-wise robustness, which measures differences between classes, is also lower for CFA (0.15) compared to EMA (0.17). This indicates that CFA indeed reduces the difference among class-wise robustness and improves the fairness without harming other classes.

Table 4. Overall comparison of experiment on Tiny-ImageNet.

Tiny-ImageNet	Best (Avg. / Worst-20%)		Last (Avg. / Worst-20%)	
Method	Clean	AA.	Clean	AA.
AT	41.1/**16.4**	20.2/3.6	39.4/17.3	14.9/1.6
AT + EMA	40.7/14.7	21.8/4.2	41.6/19.8	17.4/3.9
AT + CFA	**41.2**/16.2	**22.4/5.2**	**42.3/20.0**	**19.9/4.8**
TRADES	**43.2**/18.4	20.9/3.7	42.5/18.7	18.8/3.4
TRADES + EMA	41.2/19.5	21.6/4.1	**43.3/19.8**	19.9/3.8
TRADES + CFA	41.7/**20.0**	**22.3/5.5**	42.4/19.6	**21.2/5.2**
FAT	43.6/19.2	19.2/2.6	39.7/17.8	14.3/1.7
FAT + EMA	43.4/18.6	21.0/4.1	42.9/19.9	17.0/2.6
FAT + CFA	**43.7/19.6**	**21.6/4.9**	**43.6/21.3**	**19.1/3.4**

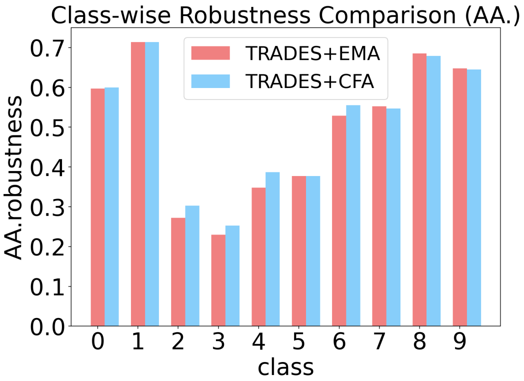


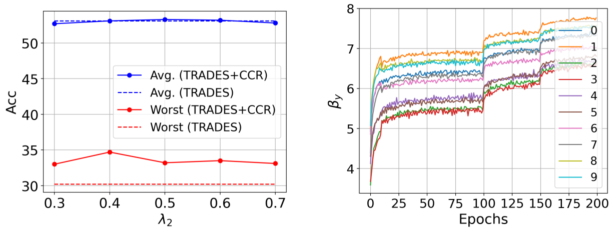
Figure 6. Class-wise robustness comparison between TRADES+EMA and TRADES+CFA on CIFAR-10 dataset at the best checkpoint. Robustness evaluated under AutoAttack.

C.3. Selection of λ_2

Following our analysis on the selection of the perturbation budget λ_1 for AT+CCM in Sec. 5.3.1, we conduct a similar analysis on the influence of regularization budget λ_2 for TRADES + CCM + CCR in Fig. 7.

In Fig. 7(a), we compare the selection of λ_2 from 0.3 to 0.7. The robustness is evaluated under PGD-10. The base perturbation budget λ_1 of CCM is still selected as 0.3. Comparing to vanilla TRADES, our TRADES+CCM+CCR outperforms in the worst class robustness significantly, and the overall robustness is marginal higher than TRADES for $\lambda_2 = 0.4$, 0.5 and 0.6.

Fig. 7(b) shows the β_y used in the case $\lambda_2 = 0.4$. We can see that the hard classes use $\beta_y \approx 6$, while the easy classes use higher β_y. This is consistent to our analysis on class-wise robustness under different regularization β in Sec 3.2.



(a) (b)

Figure 7. Analysis on the base regularization budget λ_2. (a): Average and the worst class robustness of models trained with different λ_2 (solid) and vanilla TRADES (dotted). (b): Class-wise calibrated regularization β_y in the training phase of $\lambda_2 = 0.4$.